House Price Prediction

School of Computing, Engineering and Digital Technologies

Module: Artificial Intelligence

Author Name: Keyhan Azarjoo

Email: keyhanazarjoo@gmail.com

May 2022

Contents

House Price Prediction by Two Different Pre-Processing Methods

Abstract

In this report, a dataset related to house prices in Tehran, the capital of Iran, has been examined. This dataset has been trained by two different pre-processing methods, Training 1 and Training 2, with different models and different parameters. In Training 1, we convert all non-integer data to integer and in Training 2 we use the "get.dummy" function to convert all non-integer data to Boolean. The results have been compared to each other. The results show which method for pre-processing and which model and which parameters are the best to predict prices with the highest accuracy.

By Keyhan Azarjoo

Introduction

Technology has advanced in recent years. You can get any information in a second if you have an internet connection. The advancement of data storage and AI provided an opportunity to improve businesses. Furthermore, companies are struggling to computerize their processes. On the other hand, Individuals' economic status is defined by their property, and since people started trading, they attempted to acquire more properties and estate.

Houses are a significant part of most people's property and getting a fair price when purchasing or selling one is crucial. In this research, we used some Machin learning algorithms for predicting house prices and measured the accuracy of each model.

Background

Housing, as a part of the real estate market, is crucial to the economy's long-term viability [1].

People who want to buy a house do so for a variety of reasons, but they may benefit from increases in the value of their homes [2,3]. Web sites and technological use have an impact on housing prices in the real estate industry [4,5,6]. Machine learning is one topic that is expanding at a significant rate right now [7].

In some datasets like house-related data, some factors, such as the number of bedrooms, the availability of (Parking, elevator, or warehouse), the location, and so on, have a direct impact on the price of a home. [8, 9].

After a series of international sanctions were imposed on Iran, everything became more expensive, and estimating housing prices without using technology is a risky decision for people, especially for those who are unfamiliar with real estate [10].

Pre-Processing

About dataset

The dataset used has been downloaded from Kaggle and it is about house prices in different parts of Tehran, Iran. 187 regions have been gathered in this dataset.

In this dataset we can see:

- Area in square meters
- Number of bedrooms
- Has Parking or not
- Has elevator or not
- Has warehouse or not
- The region where the house is placed
- Price in Toman and USD (Every USD is equal to 30,000 Tomans)

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3479 entries, 0 to 3478
Data columns (total 7 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   Area        3479 non-null   object
 1   Room        3479 non-null   int64
 2   Parking     3479 non-null   bool
 3   Warehouse   3479 non-null   bool
 4   Elevator    3479 non-null   bool
 5   Address     3456 non-null   object
 6   Price       3479 non-null   float64
dtypes: bool(3), float64(1), int64(1), object(2)
memory usage: 119.0+ KB
```

Figure 1: A brief of the Dataset

In this research, we use two different pre-processing methods and showed the accuracy of each of them.

The first method, Train 1, is converting all object values to integers while in the second method, Train 2, we use dummy variables for object values and convert them to Boolean.

Finding Errors and Outliers

In the first step, we checked the data for null values, and 23 null values in the Address column have been found.

Then we eliminated all null values and in the second step we search for mistakes in the data, we find some string

values in the Area column, and we convert them to integers. After that, as we have two price columns, we deleted the Toman prices and used USD Prices as our target. Then we checked Area and Price columns for finding the outlier and rare information. The reason was that in any data, some rows are rare or mistakes. And as those data affect our model, we should delete them

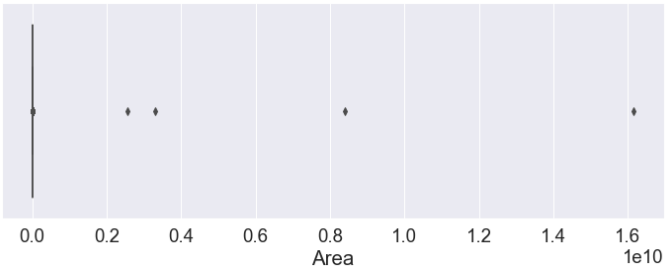


Figure 2

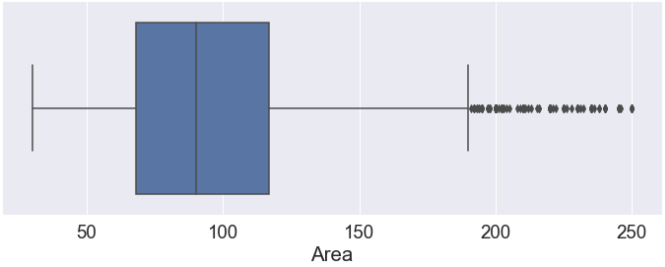


Figure 3

In figures 2 and 3 we can see how variance changed after deleting outliers. We check outliers for Price as well (figures 4 and 5).

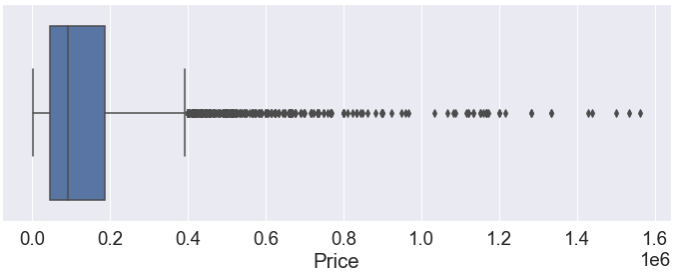


Figure 4

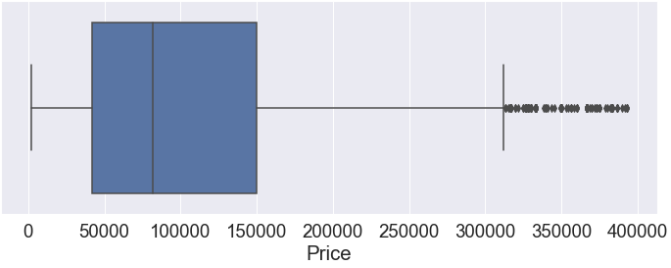


Figure 5

Virtualizing data

In figure 6 we can see the amount value of 3 columns, Parking, Warehouse, and Elevator.

Figure 6

Then we check the distribution for Rooms related to different Prices and Areas and you can see the result in figure 7.

This picture shows that the size of houses that have one room is usually between 20 to 100 square meters with the prices up to $150,000 while these numbers for houses with 2 rooms are 50 to almost 150 squares in meter. This range for houses that have 3 rooms is between 100 to 200 meters. Although the number of houses with 4 and 5 room are rare, we can see some of them usually when the Area became more than 150.

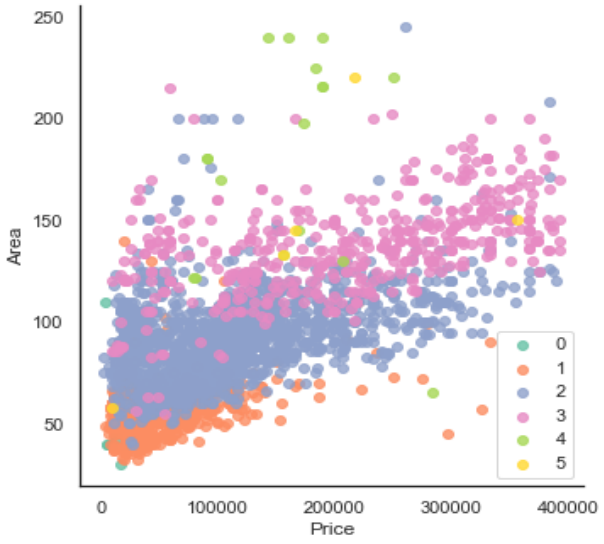


Figure 7

In figure 8, we show the distribution for parking, warehouse, and elevator in comparison with Price and Area.

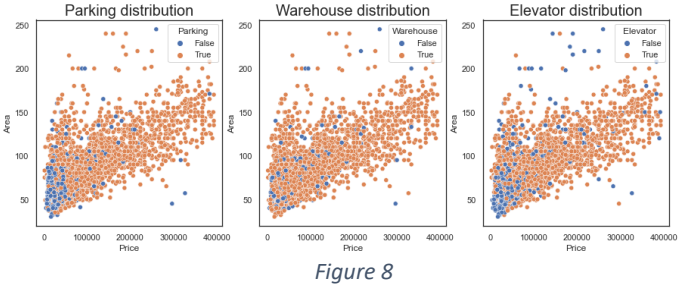


Figure 8

The False points for all 3 figures are usually when the price is less than $100,000 and the area is less than 150 square meters.

In figure 9, you can see the relation between Area and Prices which shows that by increasing the area, the prices will increase, and in figure 10 you can see the relation between Address and Prices, there is no linear relation.

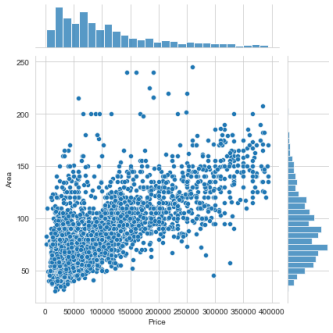


Figure 9

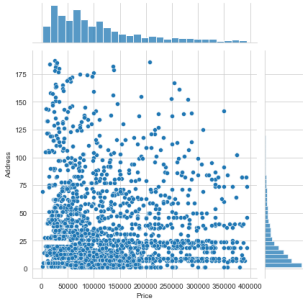


Figure 10

Different Views on Dataset

For having a better understanding of the data, we grouped data based on their address and put the mean of rows, in the same group, for each column. The result showed that there are strong linear relations between some of the features (figure 11).

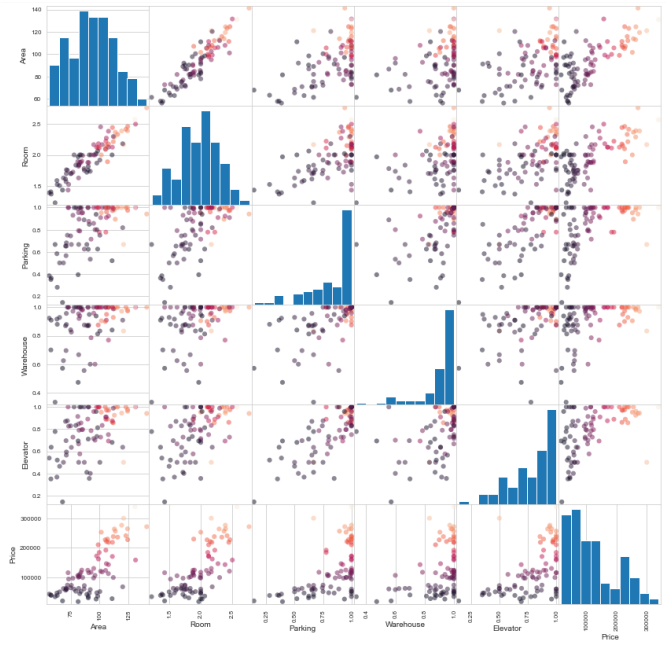


Figure 11

To be more accurate in Area Room and Price as an example, we can see, the growing the area, the number of rooms, and price increase (Figures 12 and 13).

Figure 12

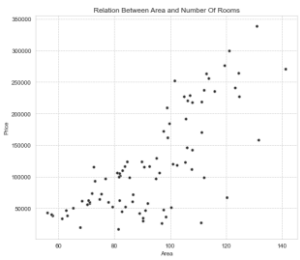


Figure 13

Converting objects to integers

As the Addresses are String and in the first method, Train 1, we need integer values, we convert all object and Boolean values to int.

As you can see in the distribution of Prices and Addresses (Figure 14), prices are mostly less than $200,000 and most of the addresses are in the area between 0 to 50 And you can see these numbers for Area and room in figure 15.

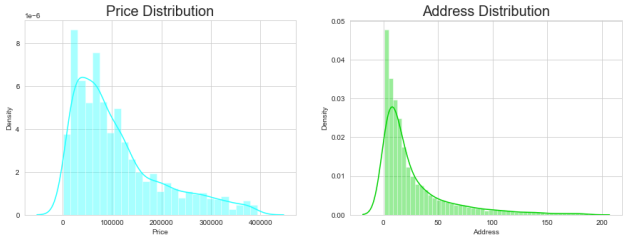


Figure 14

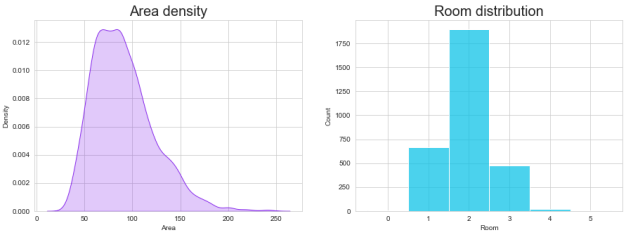


Figure 15

Statistic Information

Pearson coefficient

You can see the Pearson coefficient for all columns in figure 16. Those numbers have been shown by Heatmap to express the strength of the relationship between all variables (Figure 17).

	Area	Room	Parking	Warehouse	Elevator	Address	Price
Area	1.000000	0.776401	0.295796	0.143153	0.267082	0.070559	0.660709
Room	0.776401	1.000000	0.265140	0.117202	0.236295	0.021437	0.502769
Parking	0.295796	0.265140	1.000000	0.412933	0.432311	-0.109955	0.333021
Warehouse	0.143153	0.117202	0.412933	1.000000	0.178952	-0.063955	0.189137
Elevator	0.267082	0.236295	0.432311	0.178952	1.000000	-0.056262	0.288933
Address	0.070559	0.021437	-0.109955	-0.063955	-0.056262	1.000000	0.060803
Price	0.660709	0.502769	0.333021	0.189137	0.288933	0.060803	1.000000

Figure 16

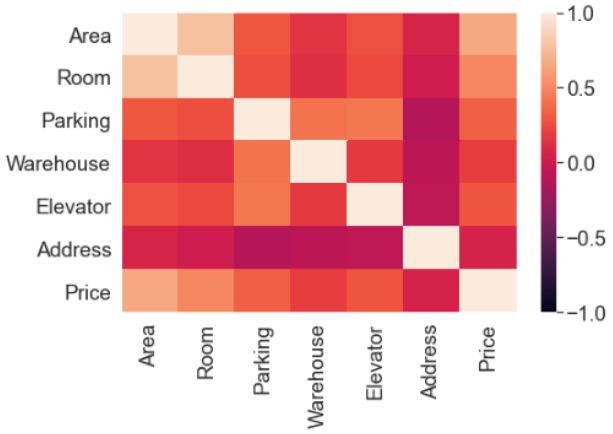


Figure 17

As you can see there are some relations between room and area, prices and area, and prices and room.

Pearson coefficient formula

$$\frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

Empirical Distribution Function:

The stat models Python library provides the ECDF class for fitting an empirical cumulative distribution function and calculating the cumulative probabilities for specific observations from the domain.

We calculated the ECDF for two Area and Price columns and virtualized them as you can see in Figure 18.

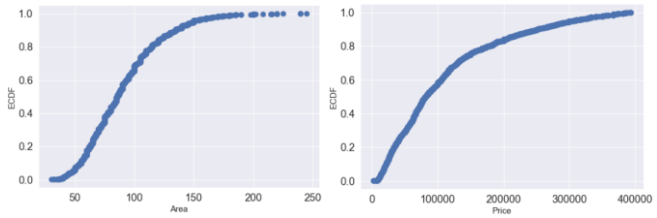


Figure 18

Spearman's rank correlation

The Spearman correlation between two variables is equal to the Pearson correlation between the rank values of those two variables. while Pearson's correlation assesses linear relationships, Spearman's correlation assesses monotonic relationships.

The result of Spearman for two columns, Room and Area, became 0.78 which shows that there is a positive correlation between these two columns.

Chi-square

But we use Chi-square for two Room and Parking columns, as an example, and the result shows that there is not any correlation between these two columns (Figure 19).

Room	0	1	2	3	4	5
Parking						
0	7	241	241	21	1	3
1	2	424	1657	453	16	2

```
Chi2              :  287.24399018818104
p_value           :  5.52691954182817e-60
degreeoffreedom :  5
```

Figure 19

The degree of freedom is 5 and based on the Freedom Table the 0.05 for DOF 5 is 11.07.

H0 has been rejected as 11.07 < 287.24

so we can find that there is no relation between Parking and Room.

TRAINING 1

In this step, we convert all values in Address columns to int, ordered by the number of cases in each group we gave them a code.

Then by using the “GridSearchCV” function, we use some parameters and find the best parameter for each model. The scales and results are as below:

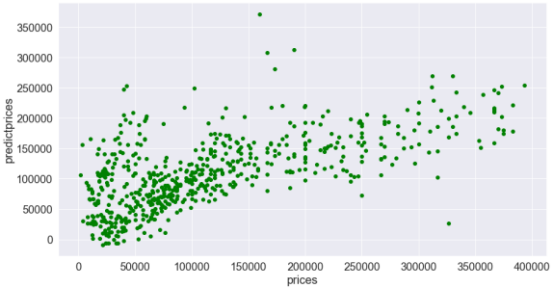
The tables in the first part of each model shown

The best parameters for the model, are Training Coefficient of determination (R2 score), Test Coefficient of determination (R2 score), Explain Variance Score (EVS), Mean Absolute Error (MAB), Mean Squared Error (MSE), Square-Root of MSE (RMSE), Median Absolute Error (MAE), Runtime of the program.

The picture in the second part shows the difference between Prices and Predicted Prices in each model.

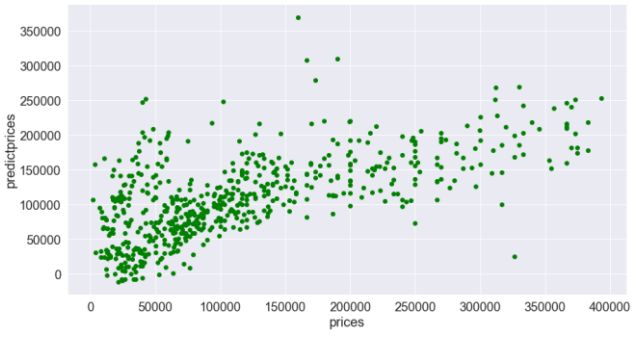
Elastic Net Regression

```
The best parameters for ElasticNet model is:
 {'alpha': 0.1, 'l1_ratio': 0.8}
------------------------------------------------------------
Training Info :

  Coefficient of determination (R2 score)  :  47.87%.

------------------------------------------------------------
Testing Info :

  Coefficient of determination (R2 score)  :  40.99%.
------------------------------------------------------------
  Explain Variance Score                   :  0.42
------------------------------------------------------------
  Mean Absolute Error                      :  51177.48
------------------------------------------------------------
  MSE (Mean Squared Error)                 :  5028209439.299366
------------------------------------------------------------
  Square-Root of MSE                       :  70910
------------------------------------------------------------
  Median Absolute Error                    :  32169.78
------------------------------------------------------------
 Runtime of the program                   :  0.34
------------------------------------------------------------
```



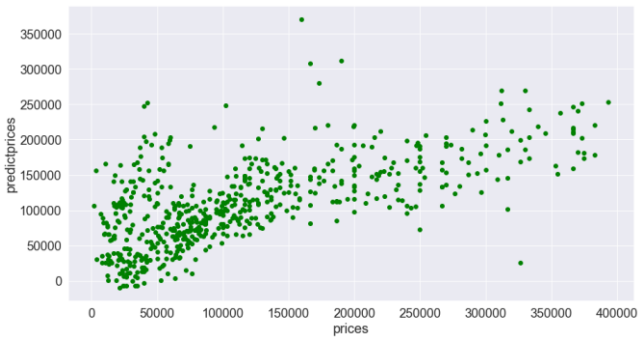
linear Regression

```
Training Info :

  Coefficient of determination (R2 score)  :  47.89%.

------------------------------------------------------------
Testing Info :

  Coefficient of determination (R2 score)  :  41.05%.
------------------------------------------------------------
  Explain Variance Score                   :  0.42
------------------------------------------------------------
  Mean Absolute Error                      :  51169.24
------------------------------------------------------------
  MSE (Mean Squared Error)                 :  5023045074.842355
------------------------------------------------------------
  Square-Root of MSE                       :  70873
------------------------------------------------------------
  Median Absolute Error                    :  32556.11
------------------------------------------------------------
 Runtime of the program                   :  0.05
------------------------------------------------------------
```



Ridge Regression

```
The best parameters for Ridge model is:
 {'alpha': 30}
------------------------------------------------------------
Training Info :

  Coefficient of determination (R2 score)  :  47.88%.

------------------------------------------------------------
Testing Info :

  Coefficient of determination (R2 score)  :  41.02%.
------------------------------------------------------------
  Explain Variance Score                   :  0.42
------------------------------------------------------------
  Mean Absolute Error                      :  51173.42
------------------------------------------------------------
  MSE (Mean Squared Error)                 :  5025977842.190939
------------------------------------------------------------
  Square-Root of MSE                       :  70894
------------------------------------------------------------
  Median Absolute Error                    :  32341.44
------------------------------------------------------------
 Runtime of the program                   :  0.14
------------------------------------------------------------
```

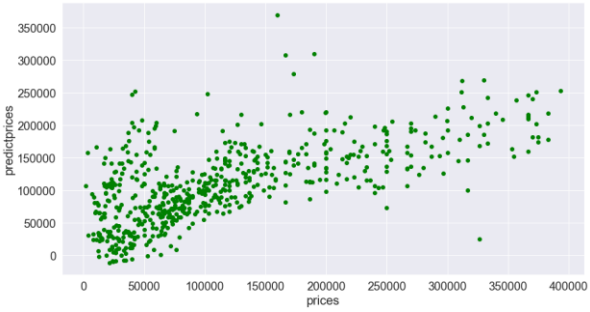


Lasso Regression

The best parameters for Lasso model is:
{'alpha': 10}

Training Info :	
Coefficient of determination (R2 score)	47.89%.

Testing Info :	
Coefficient of determination (R2 score)	41.05%.
Explain Variance Score	0.42
Mean Absolute Error	51169.66
MSE (Mean Squared Error)	5023245003.642378
Square-Root of MSE	70875
Median Absolute Error	32511.61
Runtime of the program	0.07

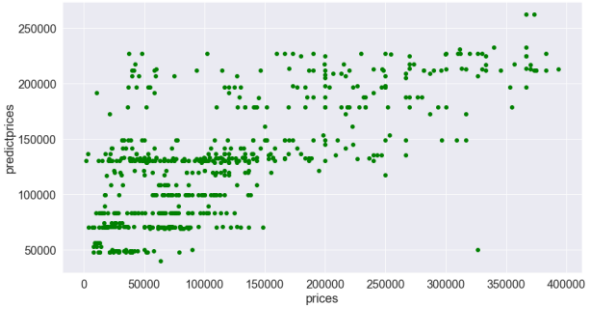


Ada Boost Regression

The best parameters for AdaBoostRegressor model is:
{'learning_rate': 2, 'n_estimators': 40}

Training Info :	
Coefficient of determination (R2 score)	50.90%.

Testing Info :	
Coefficient of determination (R2 score)	46.34%.
Explain Variance Score	0.48
Mean Absolute Error	51307.7
MSE (Mean Squared Error)	4572317363.869867
Square-Root of MSE	67619
Median Absolute Error	40791.53
Runtime of the program	8.52

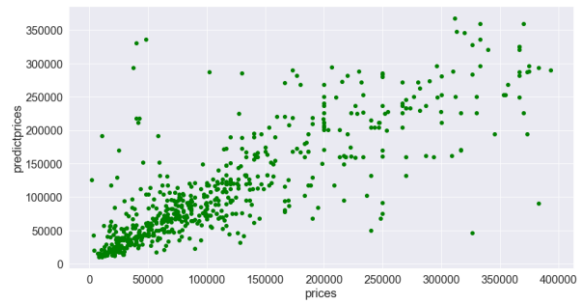


Decision Tree Regression

The best parameters for DecisionTreeRegressor model is:
{'min_samples_leaf': 4, 'min_samples_split': 2}

Training Info :	
Coefficient of determination (R2 score)	86.15%.

Testing Info :	
Coefficient of determination (R2 score)	66.14%.
Explain Variance Score	0.66
Mean Absolute Error	32913.88
MSE (Mean Squared Error)	2885105339.546384
Square-Root of MSE	53713
Median Absolute Error	18472.23
Runtime of the program	0.94

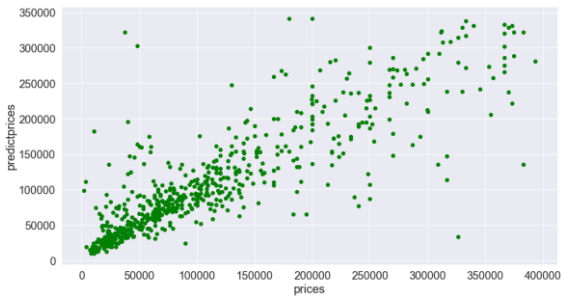


Random Forest Regression

The best parameters for RandomForestRegressor model is:
{'min_samples_leaf': 2, 'min_samples_split': 2}

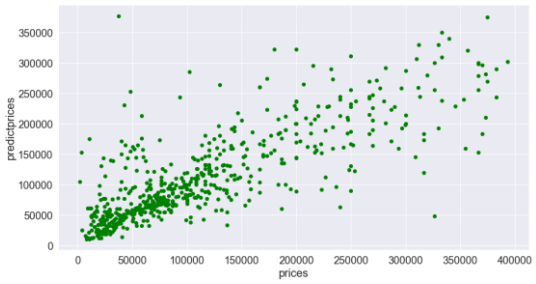
Training Info :	
Coefficient of determination (R2 score)	91.84%.

Testing Info :	
Coefficient of determination (R2 score)	72.10%.
Explain Variance Score	0.72
Mean Absolute Error	29533.14
MSE (Mean Squared Error)	2377904689.861629
Square-Root of MSE	48764
Median Absolute Error	15186.51
Runtime of the program	37.92



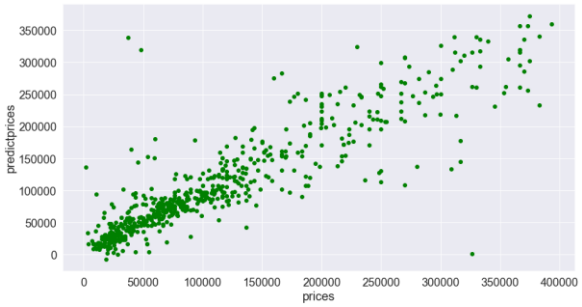
K-Neighbors Regression

```
The best parameters for KNeighborsRegressor model is:
 {'n_neighbors': 15, 'weights': 'distance'}
------------------------------------------------------------
Training Info :

  Coefficient of determination (R2 score)  :  98.75%.

------------------------------------------------------------
Testing Info :

  Coefficient of determination (R2 score)  :  63.84%.
------------------------------------------------------------
  Explain Variance Score                   :  0.64
------------------------------------------------------------
  Mean Absolute Error                      :  35131.23
------------------------------------------------------------
  MSE (Mean Squared Error)                 :  3081780861.392797
------------------------------------------------------------
  Square-Root of MSE                       :  55514
------------------------------------------------------------
  Median Absolute Error                    :  20000.0
------------------------------------------------------------
 Runtime of the program                   :  2.75
------------------------------------------------------------
```



Gradient Boosting Regression

```
The best parameters for GradientBoostingRegressor model is:
 {'alpha': 0.5, 'learning_rate': 0.4, 'max_depth': 4}
------------------------------------------------------------
Training Info :

  Coefficient of determination (R2 score)  :  94.99%.

------------------------------------------------------------
Testing Info :

  Coefficient of determination (R2 score)  :  79.40%.
------------------------------------------------------------
  Explain Variance Score                   :  0.8
------------------------------------------------------------
  Mean Absolute Error                      :  24553.19
------------------------------------------------------------
  MSE (Mean Squared Error)                 :  1755170625.7187567
------------------------------------------------------------
  Square-Root of MSE                       :  41895
------------------------------------------------------------
  Median Absolute Error                    :  12757.79
------------------------------------------------------------
 Runtime of the program                   :  79.57
------------------------------------------------------------
```



Total results and comparison

	Model Name	Train R2	Test R2	EVS	MAE	MSE	RMSE	MedAE
0	LinearRegression	0.4790	0.4107	0.42	51165.91	5.021564e+09	70863	32558.65
1	ElasticNet	0.4787	0.4101	0.42	51173.37	5.026763e+09	70900	32186.27
2	Ridge	0.4789	0.4104	0.42	51169.49	5.024517e+09	70884	32426.76
3	Lasso	0.4790	0.4107	0.42	51166.43	5.021763e+09	70864	32543.82
4	DecisionTreeRegressor	0.8774	0.6321	0.63	33477.73	3.134734e+09	55989	16469.05
5	AdaBoostRegressor	0.5593	0.4754	0.48	49306.89	4.470434e+09	66861	37096.91
6	RandomForestRegressor	0.9187	0.7166	0.72	29641.97	2.415256e+09	49145	15356.48
7	KNeighborsRegressor	0.9874	0.6320	0.63	35371.29	3.135685e+09	55997	19717.00
8	GradientBoostingRegressor	0.9232	0.8014	0.80	23965.29	1.692602e+09	41141	11924.59

Figure 20 (Train 1 comparison)

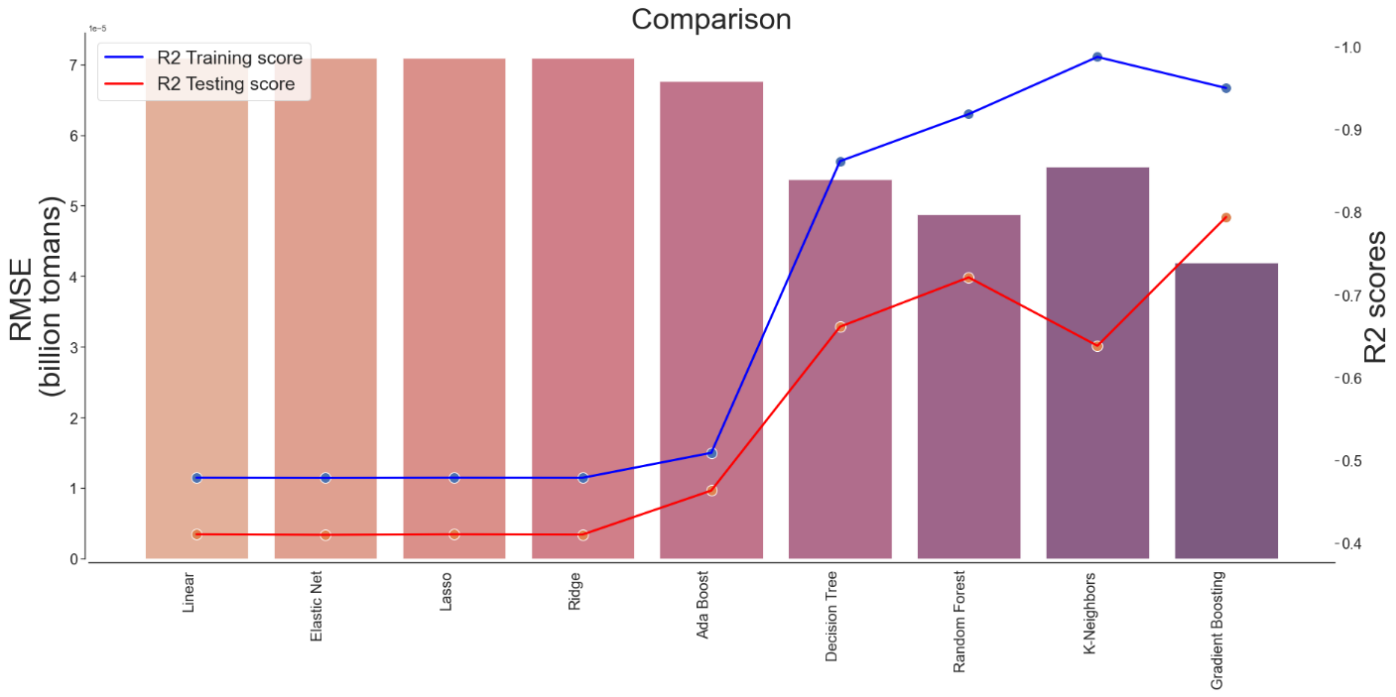


Figure 21 (Train 1 comparison)

As you can see, the best model was Gradient Boosting Regression and the best parameters for the Gradient Boosting Regressor model are as below

Alpha: 0.5, learning rate: 0.2, max depth: 5

TRAINING 2

Here we used the “get.dummies” function and converted Address, Parking, Warehouse, and Elevator to 0 and 1 in separated columns.

A dummy variable is a binary variable that indicates whether a separate categorical variable takes on a specific value.

Then we wrote a function to train our data and find the accuracy results like Training r2 score, Testing r2 score, Explain Variance Score, Mean Absolute Error, MSE (Mean Squared Error), Square-Root of MSE, and Median Absolute Error.

Before finding the best parameters, we check the Cross-Validation Score with different CVs, and we showed the CVs and results in linear graphs then showed the best CV. This is to check how much parameters can affect the training model.

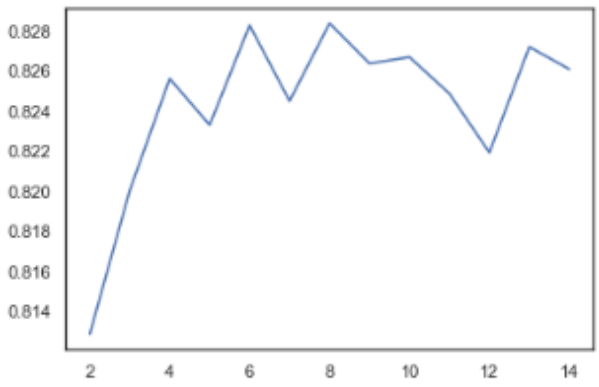
Cross-validation is primarily used in applied machine learning to estimate the skill of a machine learning model on unseen data. That is, to use a limited sample to estimate how the model is expected to perform in general when used to make predictions on data not used during the training of the model.

As CVS is the mean of MSE (Mean Squared errors) with different train and test parts, it shows the best result for training.

Then we train models with different parameters, by giving each parameter separately and comparing the r2 score for each of them then showing the best parameters. After that, we compared the real data with predicted data for each model and showed it in a graph.

Ridge Regression

Cross-validation Scores




```
Cross Validation score is 0.828378 with CV = 8.0.
```

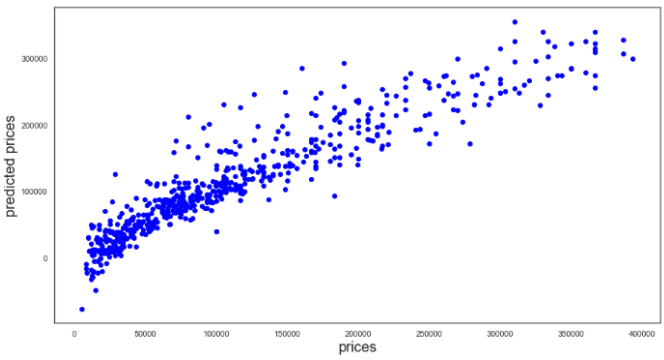
The best CV is 8 with 82% accuracy in Cross Val Score.

The model has been trained with below parameters:

alpha = [0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9,1]

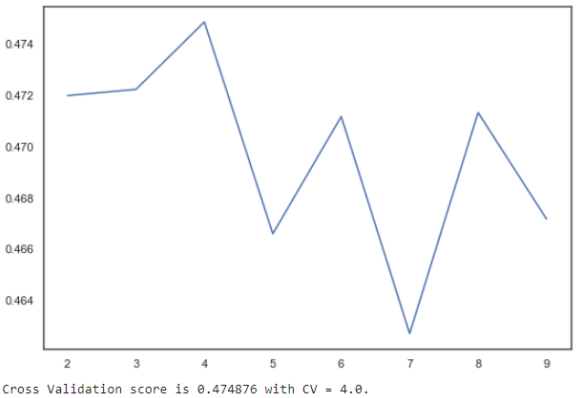
The function shows that the best alpha is 0.7.

```
Training Info :

  Coefficient of determination (R2 score)  :  86.33%.

-----------------------------------------------------------------
Testing Info :

  Coefficient of determination (R2 score)  :  87.03%.
-----------------------------------------------------------------
  Explain Variance Score                   :  0.87
-----------------------------------------------------------------
  Mean Absolute Error                      :  22109.75
-----------------------------------------------------------------
  MSE (Mean Squared Error)                 :  1024274444.9114076
-----------------------------------------------------------------
  Square-Root of MSE                       :  32004
-----------------------------------------------------------------
  Median Absolute Error                    :  14456.5
-----------------------------------------------------------------
 Runtime of the program                   :  0.02
-----------------------------------------------------------------
```



Elastic Net Regression

Cross-validation Scores




```
Cross Validation score is 0.474876 with CV = 4.0.
```

The best CV is 4 with 47% accuracy in Cross Val Score.

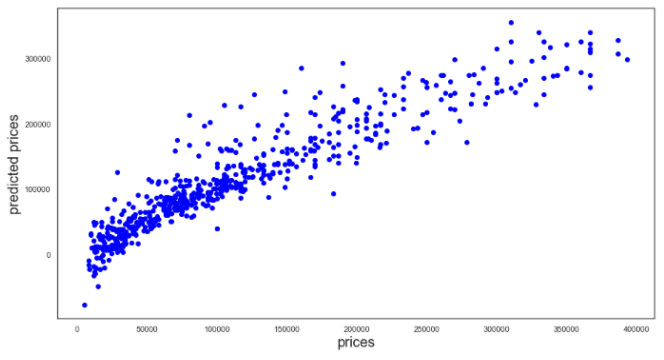
The model has been trained with below parameters:

alpha=[0.001, 0.01, 0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9,1]

l1_ratio = [0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9]

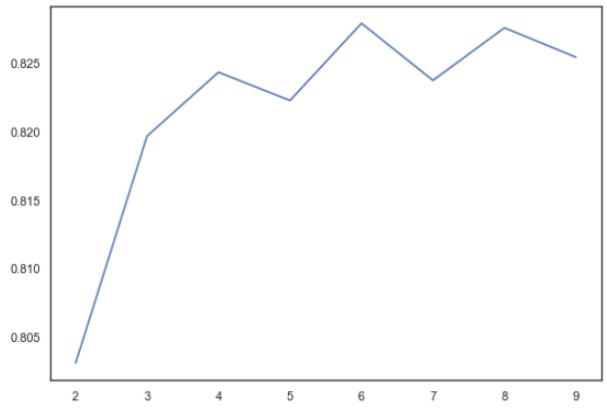
the function show that the best alpha is 0.001 and the best l1_ratio is 0.7.

```
Training Info :

  Coefficient of determination (R2 score)  :  86.07%.

-----------------------------------------------------------------
Testing Info :

  Coefficient of determination (R2 score)  :  87.09%.
-----------------------------------------------------------------
  Explain Variance Score                   :  0.87
-----------------------------------------------------------------
  Mean Absolute Error                      :  22145.26
-----------------------------------------------------------------
  MSE (Mean Squared Error)                 :  1019683670.9782329
-----------------------------------------------------------------
  Square-Root of MSE                       :  31932
-----------------------------------------------------------------
  Median Absolute Error                    :  14426.17
-----------------------------------------------------------------
 Runtime of the program                   :  0.32
-----------------------------------------------------------------
```



Lasso Regression

Cross-validation Scores




```
Cross Validation score is 0.827905 with CV = 6.0.
```

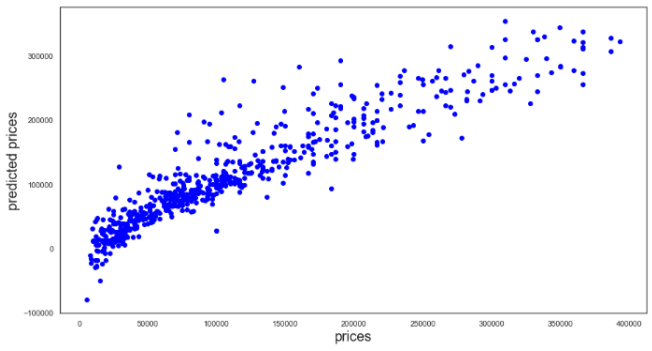
The best CV is 6 with 82% accuracy in Cross Val Score.

The model has been trained with below parameters:

alpha = [0.001, 0.01, 0.1, 0.2 ,0.3 ,0.4 ,0.5 ,0.6 ,0.7 ,0.8 ,0.9 ,1 ,5 ,10]

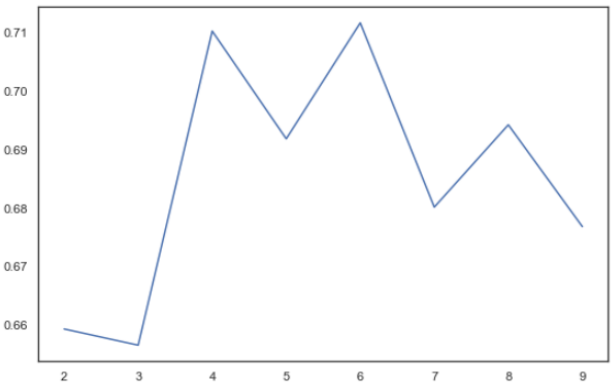
the function show that the best alpha is 5.

```
Training Info :

  Coefficient of determination (R2 score)  :  86.36%.

-----------------------------------------------------------------
Testing Info :

  Coefficient of determination (R2 score)  :  86.75%.
-----------------------------------------------------------------
  Explain Variance Score                   :  0.87
-----------------------------------------------------------------
  Mean Absolute Error                      :  22063.4
-----------------------------------------------------------------
  MSE (Mean Squared Error)                 :  1046191681.527863
-----------------------------------------------------------------
  Square-Root of MSE                       :  32345
-----------------------------------------------------------------
  Median Absolute Error                    :  13845.09
-----------------------------------------------------------------
 Runtime of the program                    :  0.05
-----------------------------------------------------------------
```



Decision Tree Regressor

Cross-validation Scores




```
Cross Validation score is 0.711702 with CV = 6.0.
```

The best CV is 6 with 71% accuracy in Cross Val Score.

The model has been trained with below parameters:

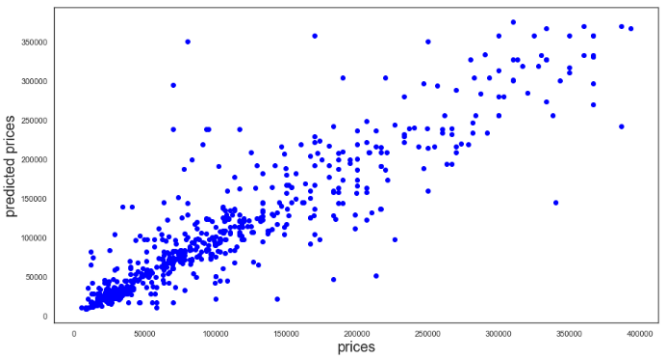
max_depth = [2 ,3 ,5 ,10 ,15 ,20 ,50 ,60 ,70 ,80 ,90 ,100 ,110 ,150]

min_samples_split = range (2,50)

random_state = 1

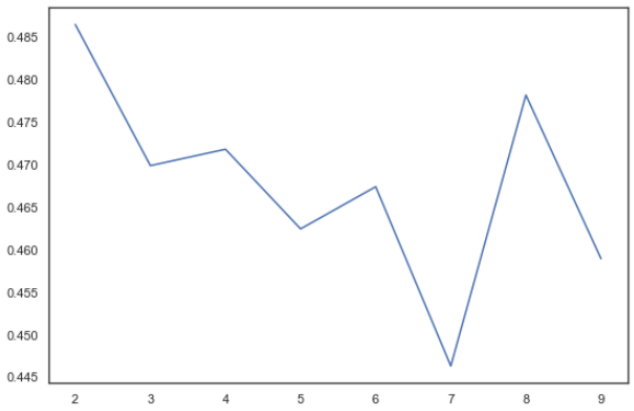
the function show that the best parameterS ARE min_samples_split = 6 and max_depth = 150.

```
Training Info :

  Coefficient of determination (R2 score)  :  96.42%.

-----------------------------------------------------------------
Testing Info :

  Coefficient of determination (R2 score)  :  80.97%.
-----------------------------------------------------------------
  Explain Variance Score                   :  0.81
-----------------------------------------------------------------
  Mean Absolute Error                      :  23018.9
-----------------------------------------------------------------
  MSE (Mean Squared Error)                 :  1503129619.5360315
-----------------------------------------------------------------
  Square-Root of MSE                       :  38770
-----------------------------------------------------------------
  Median Absolute Error                    :  11111.11
-----------------------------------------------------------------
 Runtime of the program                    :  0.05
-----------------------------------------------------------------
```



Ada Boost Regressor

Cross-Validation Score




```
Cross Validation score is 0.486509 with CV = 2.0.
```

The best CV is 2 with 48% accuracy in Cross Val Score.

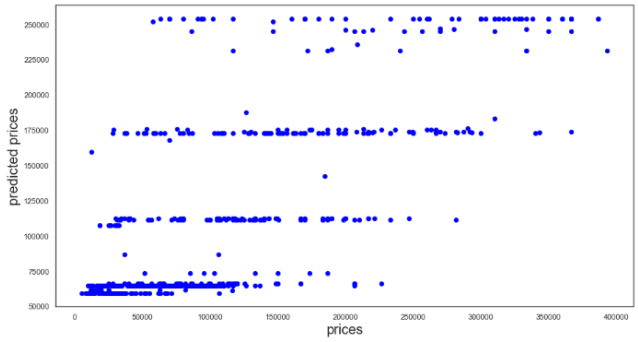
The model has been trained with below parameters:

n_estimators = [5, 10,15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95, 100]

learning_rate = [0.01, 0.1, 1, 2, 3, 4, 5, 10]

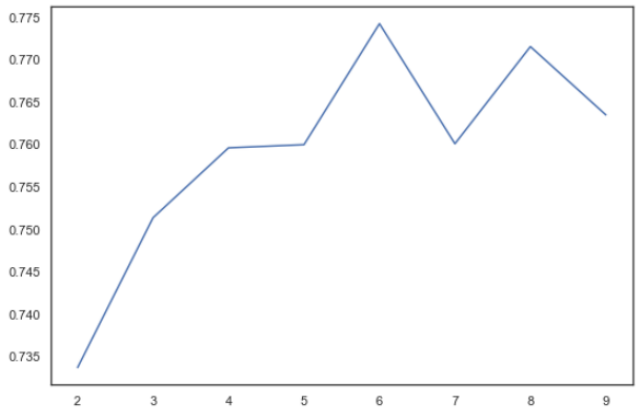
the function show that the best parameter for n_estimators = 30 and learning_rate = 0.01. however, these parameters change every time each run.

```
Training Info :

  Coefficient of determination (R2 score)  :  53.34%.

-----------------------------------------------------------------
Testing Info :

  Coefficient of determination (R2 score)  :  54.92%.
-----------------------------------------------------------------
  Explain Variance Score                   :  0.55
-----------------------------------------------------------------
  Mean Absolute Error                      :  45364.1
-----------------------------------------------------------------
  MSE (Mean Squared Error)                 :  3560559403.1359224
-----------------------------------------------------------------
  Square-Root of MSE                       :  59670
-----------------------------------------------------------------
  Median Absolute Error                    :  36083.34
-----------------------------------------------------------------
 Runtime of the program                   :  0.43
-----------------------------------------------------------------
```



Random Forest Regressor

Cross-Validation Score




```
Cross Validation score is 0.774304 with CV = 6.0.
```

The best CV is 6 with 77% accuracy in Cross Val Score.

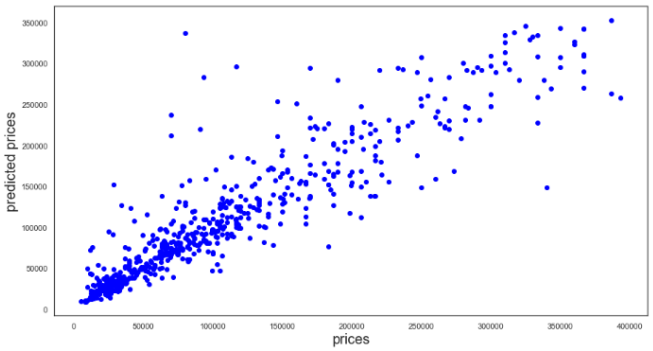
The model has been trained with below parameters:

n_estimators = [5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100,150,200]

min_samples_split = [2, 3, 4, 5,6,7,8,9, 10, 15, 20, 25, 30, 40, 50]

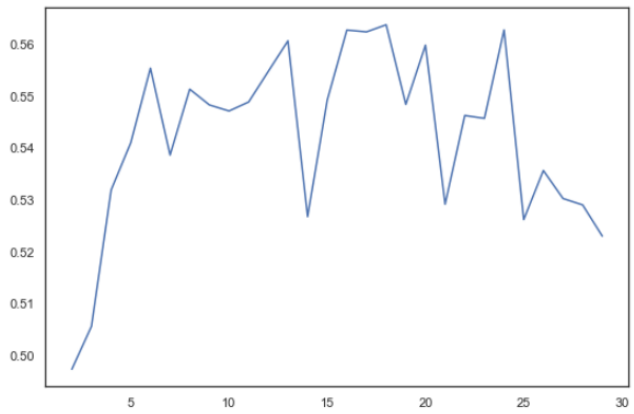
the function show that the best parameter for n_estimators = 10 and min_samples_split = 6.

```
Training Info :

  Coefficient of determination (R2 score)  :  93.32%.

-----------------------------------------------------------------
Testing Info :

  Coefficient of determination (R2 score)  :  83.55%.
-----------------------------------------------------------------
  Explain Variance Score                   :  0.84
-----------------------------------------------------------------
  Mean Absolute Error                      :  21705.98
-----------------------------------------------------------------
  MSE (Mean Squared Error)                 :  1299096029.1120207
-----------------------------------------------------------------
  Square-Root of MSE                       :  36043
-----------------------------------------------------------------
  Median Absolute Error                    :  11789.58
-----------------------------------------------------------------
 Runtime of the program                   :  0.11
-----------------------------------------------------------------
```



K-Neighbors Regressor

Cross-Validation Score




```
Cross Validation score is 0.563755 with CV = 18.0.
```

The best CV is 18 with 56% accuracy in Cross Val Score.

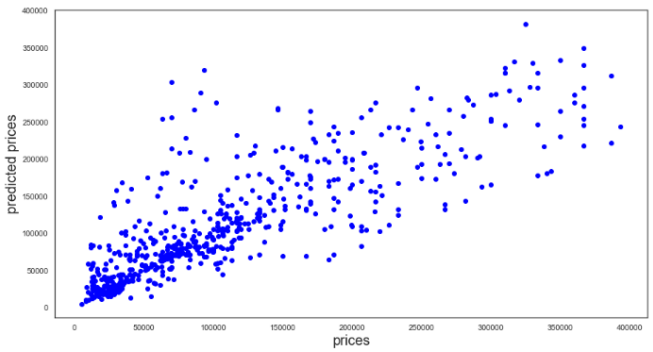
The model has been trained with below parameters:

weights = ['uniform', 'distance']

n_neighbors = range(2,50)

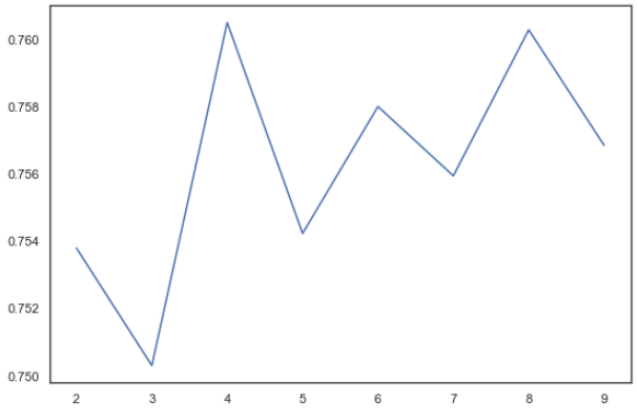
the function show that the best parameter for n_neighbors = 9 and weights = 'distance'.

```
Training Info :

  Coefficient of determination (R2 score)  :  98.28%.

-------------------------------------------------------------
Testing Info :

  Coefficient of determination (R2 score)  :  67.78%.
-------------------------------------------------------------
  Explain Variance Score                   :  0.68
-------------------------------------------------------------
  Mean Absolute Error                      :  31696.84
-------------------------------------------------------------
  MSE (Mean Squared Error)                 :  2544795994.79584
-------------------------------------------------------------
  Square-Root of MSE                       :  50446
-------------------------------------------------------------
  Median Absolute Error                    :  16666.66
-------------------------------------------------------------
 Runtime of the program                   :  0.27
-------------------------------------------------------------
```



Gradient Boosting Regressor

Cross-Validation Score




```
Cross Validation score is 0.760498 with CV = 4.0.
```

The best CV is 4 with 76% accuracy in Cross Val Score.

The model has been trained with below parameters:

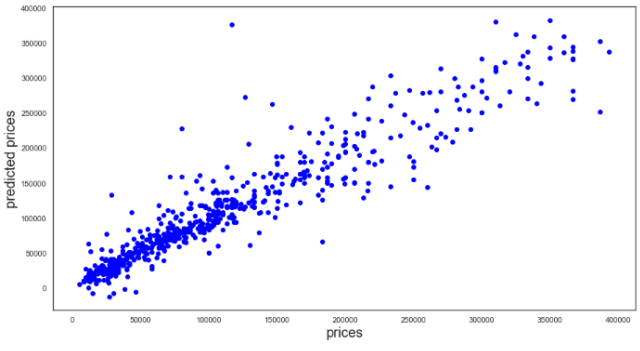
learning_rate = [0.01, 0.1,0.2, 0.3, 0.4, 0.5]

alpha = [0.01, 0.1,0.2, 0.3, 0.4, 0.5]

max_depth = [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]

the function show that the best parameter for learning_rate = 0.40, alpha = 0.1, and max_depth = 9.

```
Training Info :

  Coefficient of determination (R2 score)  :  97.75%.

-------------------------------------------------------------
Testing Info :

  Coefficient of determination (R2 score)  :  88.49%.
-------------------------------------------------------------
  Explain Variance Score                   :  0.89
-------------------------------------------------------------
  Mean Absolute Error                      :  18183.25
-------------------------------------------------------------
  MSE (Mean Squared Error)                 :  908829023.6292182
-------------------------------------------------------------
  Square-Root of MSE                       :  30147
-------------------------------------------------------------
  Median Absolute Error                    :  9605.4
-------------------------------------------------------------
 Runtime of the program                   :  1.65
-------------------------------------------------------------
```



Total results and comparison

	Model Name	Train R2	Test R2	EVS	MAE	MSE	RMSE	MedAE
0	Ridge	0.8610	0.8709	0.87	22139.81	1.019670e+09	31932	14542.95
1	ElasticNet	0.8607	0.8709	0.87	22145.26	1.019684e+09	31932	14426.17
2	Lasso	0.8636	0.8675	0.87	22063.40	1.046192e+09	32345	13845.09
3	DecisionTreeRegressor	0.9642	0.8097	0.81	23018.90	1.503130e+09	38770	11111.11
4	AdaBoostRegressor	0.5328	0.5510	0.55	45156.76	3.546861e+09	59556	35146.34
5	RandomForestRegressor	0.9332	0.8355	0.84	21705.98	1.299096e+09	36043	11789.58
6	KNeighborsRegressor	0.9828	0.6778	0.68	31696.84	2.544796e+09	50446	16666.66
7	GradientBoostingRegressor	0.9775	0.8815	0.88	18403.55	9.358017e+08	30591	9634.30

Figure 22 (Train 2 comparison)

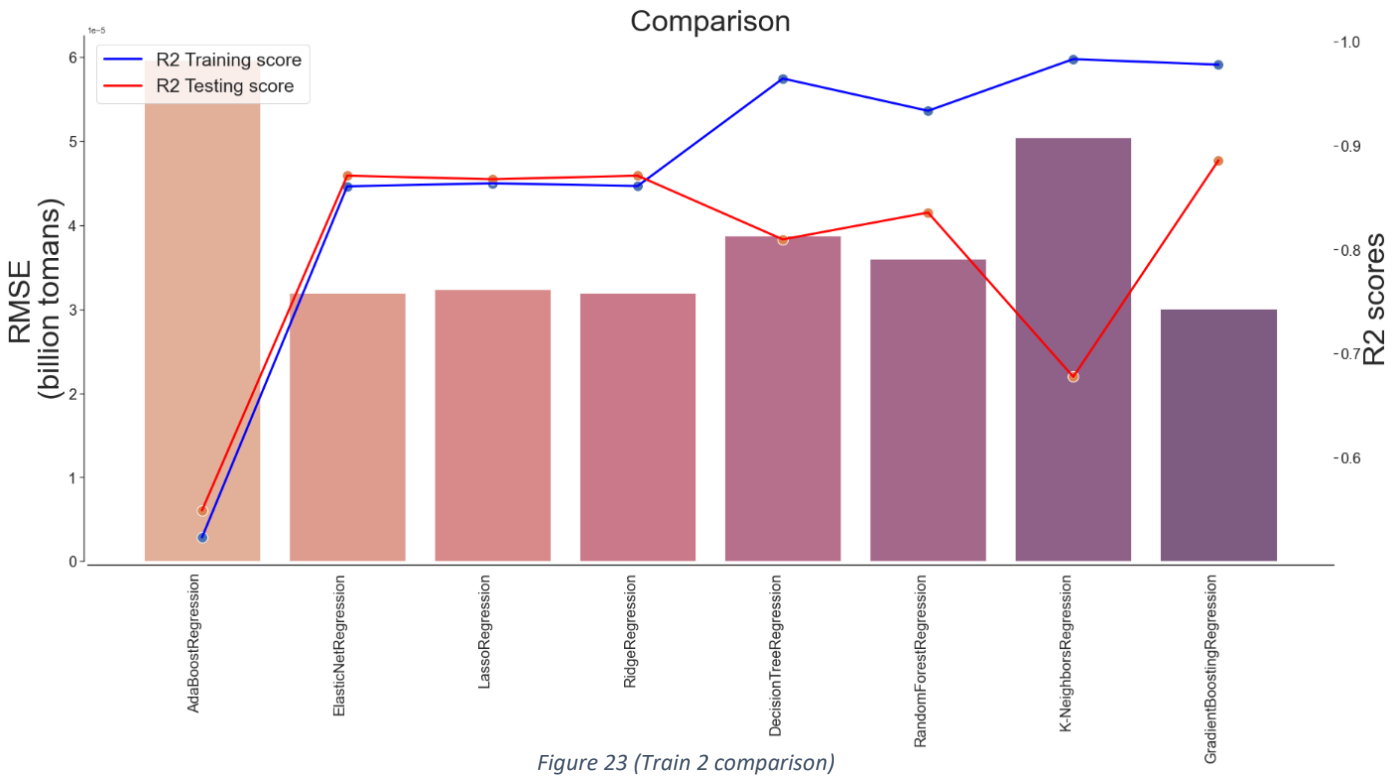


Figure 23 (Train 2 comparison)

As you can see, the best model for training our data is Gradient Boosting Regression with the below scores.

Train r2 squared: %86

Test r2 squared = %88

EVS = %88

In other scores, in the figure, this model has lees Errors among others.

Ethical Issues

The most important ethical issue with this prediction may be the wrong prediction that may happen and cause pricing wrong about a house, resulting in losing money for the buyer or seller. However, by recording any information, ethical issues will appear. In this research, as you can see, we have some information related to houses, and these houses belong to people, and it means that the real estate may know a lot of information about your house. Not having a boiler, for example, they may sell your contact details to a boiler producer company for advertising you, resulting in disturbing you.

Conclusion

We train a dataset related to housing price with some linear Regression algorithms in two different pre-processing ways. We converted data to int in Training 1 while we used the “get.dummies” function in Training 2, and finally, we check the results for finding the best pre-processing, model, and parameters.

As the best model in both training 1 and 2 was Gradient Boosting Regression, we chose this model to train our data in the final step.

By comparing the results for this model in training 1 and Train 2 we can see the result below:

Based on r2 score, Train 2 has better result. As r2 for test in ‘Train 1’ = %79.40 and for ‘Train 2’ = %88.49 () while r2 for train in ‘Train 1’ = %94.99 and for ‘Train 2’ = %97.75.

Therefore, based on the results, we can find that the best way for pre-processing data is using the get.dummies() function to convert classified data to Boolean.

Then we wrote a function to prepare raw data in an acceptable format for the model and predicted a sample.

References

1. GEBEŞOĞLU, P.F., 2019. Housing price index dynamics in Turkey. *Journal of Yaşar University*, *14*, pp.100-107.

2. Nuuter, T.I.I.N.A., Lill, I.R.E.N.E. and Tupenaite, L.A.U.R.A., 2014. Ranking of housing market sustainability in selected European Countries. *WSEAS Trans Bus Econ*, *11*, pp.778-786.

3. Wang, Z., Hoon, J. and Lim, B., 2012. The impacts of housing affordability on social and economic sustainability in Beijing. In *Australasian Journal of Construction Economics and Building-Conference Series* (Vol. 1, No. 1, pp. 47-55).

4. Ford, J.S., Rutherford, R.C. and Yavas, A., 2005. The effects of the internet on marketing residential real estate. *Journal of Housing Economics*, *14*(2), pp.92-108.

5. Kulkarni, R., Haynes, K.E., Stough, R.R. and Paelinck, J.H., 2009. Forecasting housing prices with Google econometrics. *GMU School of public policy research paper*, (2009-10).

6. Raftery, A.E., Kárný, M. and Ettler, P., 2010. Online prediction under model uncertainty via dynamic model averaging: Application to a cold rolling mill. *Technometrics*, *52*(1), pp.52-66.

7. Varma, A., Sarma, A., Doshi, S. and Nair, R., 2018, April. House price prediction using machine learning and neural networks. In *2018 second international conference on inventive communication and computational technologies (ICICCT)* (pp. 1936-1939). IEEE.

8. James, G., Witten, D., Hastie, T. and Tibshirani, R., 2013. *An introduction to statistical learning* (Vol. 112, p. 18). New York: springer.

9. Abdulhafedh, A., 2017. Road traffic crash data: an overview on sources, problems, and collection methods. *Journal of transportation technologies*, *7*(2), pp.206-219.

10. Amini, A., Nafari, K. and Singh, R., 2022. Effect of air pollution on house prices: Evidence from sanctions on Iran. *Regional Science and Urban Economics*, *93*, p.103720.