PROJECT TITLE :
HOUSE PRICE DATA SET USING MACHINE LEARNING ALGORIHTMS

Project Title :
House price dataset using machine learning algorithms
Mail-id : moezzema10@gmail.com
GitHub: https://github.com/sayyedamoezzema

ABSTRACT : The real state market is one more of the most competitive in term or price and trends to vary significally based on a lot of factors,hence it becomes one of the prime field to apply the concept of machine learning.There in this project,we present various algorithms while predicting house price with good accuracy.We tested a regression models such as simple Linear Regression,Random Forest Regression, Support Vector Regression and K-Nearest Neighbors Regression and selected the best fir among the algorithms.This project direct us that it can be best application of machine learning models in order to optimize the result.

1.INTRODUCTION : Machine learning has been used for many years to offer image recognition, spam detection, natural speech comprehension, product recommendations and medical diagnoses. Today, machine learning algorithms can help us to enhance cyber security, ensure public safety, and improve medical outcomes. In this project we used a machine learning concept, For example, if we're going to sell a house, we need to know what price tag to put on it. Here the machine learning algorithm can give us an accurate estimation or prediction. Predicting housing prices has always been a challenge for many machine learning engineers.

Several researchers have tried to come with a model to accurately predict housing prices with high accuracy and least error. Our goal for this project was to use regression models and classification techniques in order to predict the sale price of a house.

These models are created using various features such as square feet of the house, number of room, parking, warehouse, elevator, address, price,price(USD). In this project we tested a regression models like Simple Linear Regression,Random Forest Regression, Support Vector Regression and K-Nearest Neighbors Regression and will choose the best fit

To evaluate the utility of machine learning models to estimate prices on samples of their housing dataset.

To develop user friendly house price predicting system which reduces the man power. House price prediction can help the developer to determine the selling price of a house and can help the customer to arrange the right time to purchase a house.

1.1 About the data : The house price dataset has 34479 entries. Each entry contains the following information about an individual:

- Area : The area in square meter.
- Room : The room of an individual Integer greater than 0 .
- Parking : In this data set warehouse is present in Boolean form either true or false.
- Warehouse : In this data set warehouse is present in Boolean form either true or false.
- Elevator : In this data set warehouse is present in Boolean form either true or false.
- Address : In this data set the column os address is based on the Tehran iran. House address is different region Shahran, Pardis, Shahrake Qods, Shahrake Gharb, Southern Janatabad, Niavaran, Parand and Dorous etc.

●Price : In this data set price in the form of toman rial.

●Price(USD) : In this data set price in the form of USD (ubite state dollar).

Preprocessing : The collected data may conatin missing values that may lead to in consistency.To gain better result data need to be preprocessing and so it will better the effectiveness of the algorithms.

Data collection : Collect data on house price, including their quality characteristics such as area,room, elevator,address ,price and parking etc .

Feature selection : Identify the most important features that can help predict the house price. This can be done using statistical analysis or machine learning techniques.

Model training : Select an appropriate machine learning algorithm, such as Linear Regression, Random Forest Regression, Support Vector Regression and K-Nearest Neighbors Regression and train the model on the data. The model is trained to predict the house price of samples based on their characteristics.

Model evaluation : Model evaluation: Evaluate the performance of the model on a separate test dataset to determine its error MSE,RMSE, MAE, and R-Squared(R2).

Model deployment : Once the model has been trained and evaluated, it can be deployed in a house price to predict the house price.

2.Data exploration and analysis : In this data set we done some exploratory data analysis .we have plotted some boxplot,barplot and lineplot of price(USD).

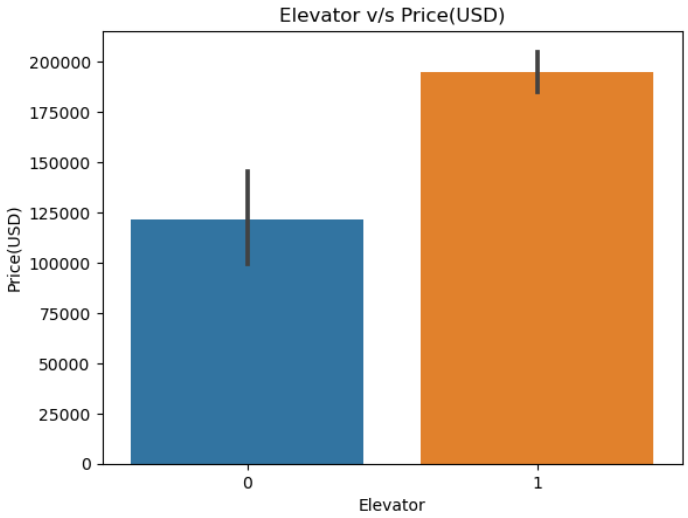


Fig 2.1 : Elevator Vs Price(USD)

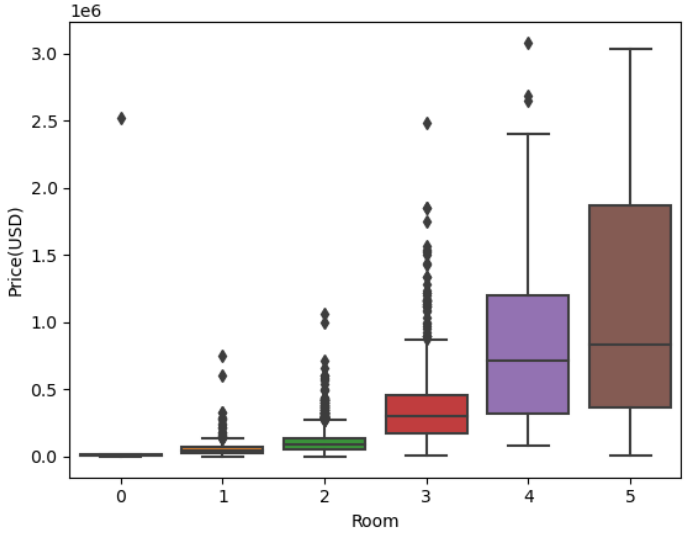


Fig 2.2 : Room Vs Price(USD)

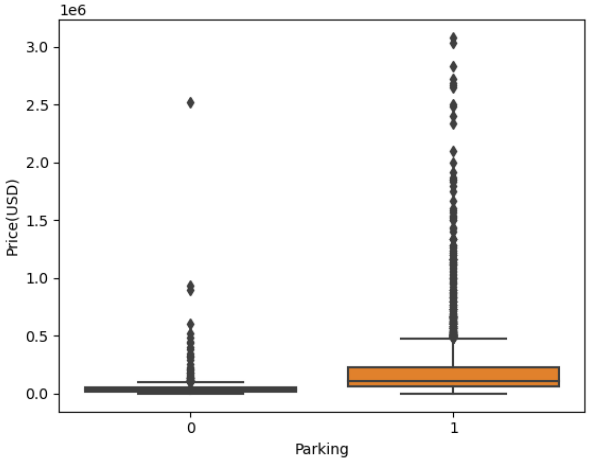


Fig 2.3 : Parking Vs Price(USD)

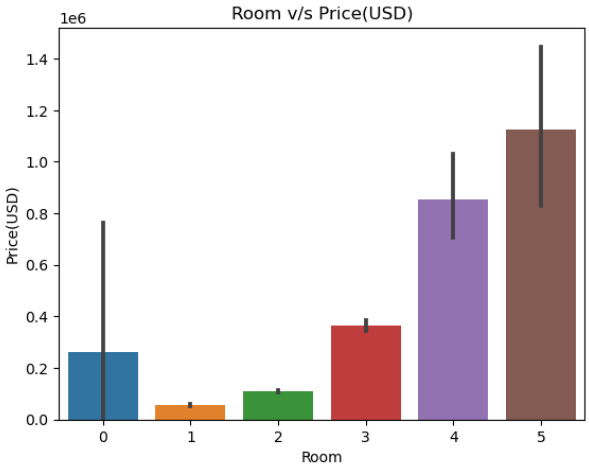


Fig 2.4 : Room Vs Price(USD)

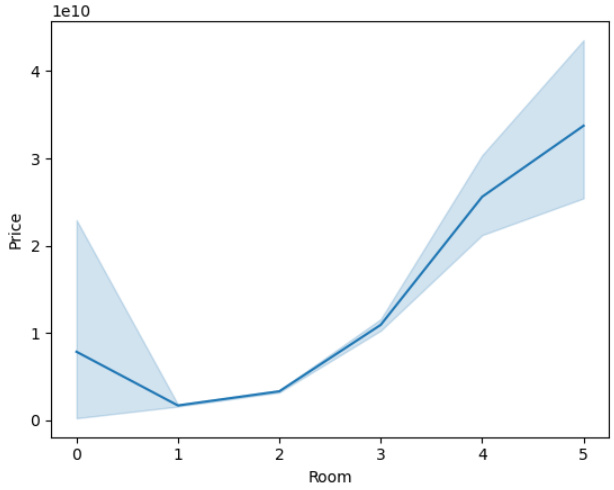


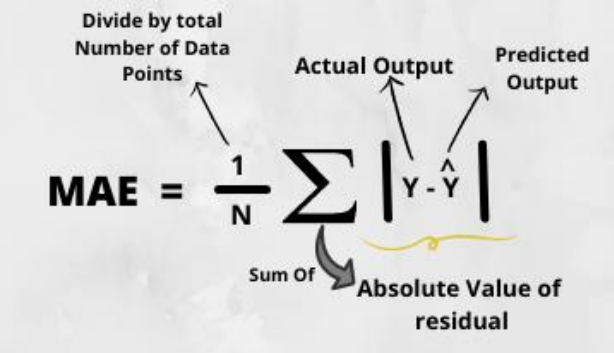
Fig 2.5 : Room Vs Price(USD)

Fig 2.6 : Heatmap of dataset

3 . Experiments : The dataset contains the continuous data which we have to predict which is the dependent feature 'Price(USD)'. So the problem we have is a Supervised Regression.

Regression is a type of Machine learning which helps in finding the relationship between independent and dependent variable.

3.1.1 Mean Absolute Error(MAE) : MAE is a very simple metric which calculates the absolute difference between actual and predicted values.

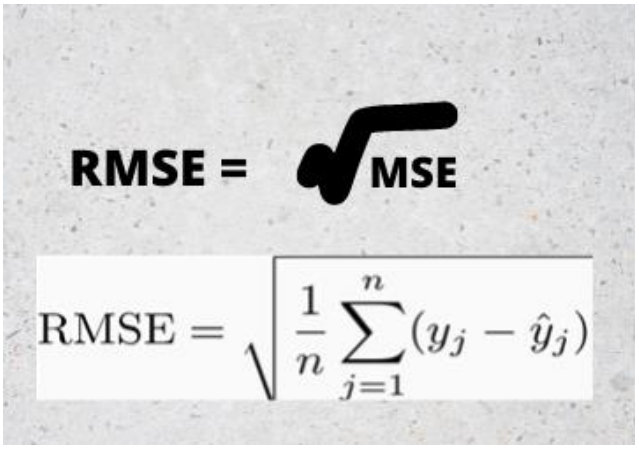


3.1.2 Mean Squared Error(MSE) : MSE is a most used and very simple metric with a little bit of change in mean absolute error. Mean squared error states that finding the squared difference between actual and predicted value.

So, above we are finding the absolute difference and here we are finding the squared difference.

$$MSE = \frac{1}{n}\Sigma\underbrace{\left(y - \hat{y}\right)^2}_{\text{The square of the difference between actual and predicted}}$$

3.1.3 Root Mean Squared Error(RMSE) : As RMSE is clear by the name itself, that it is a simple square root of mean squared error.



3.1.4 Root Mean Squared Log Error(RMSLE) : Taking the log of the RMSE metric slows down the scale of error. The metric is very helpful when you are developing a model without calling the inputs. In that case, the output will vary on a large scale.

To control this situation of RMSE we take the log of calculated RMSE error and resultant we get as RMSLE.To perform RMSLE we have to use the NumPy log function over RMSE.

print("RMSE",np.log(np.sqrt(mean_squared_error(y_test,y_pred))))

It is a very simple metric that is used by most of the datasets hosted for Machine Learning competitions.

3.1.5 R Squared (R2) : R2 score is a metric that tells the performance of your model, not the loss in an absolute sense thathow many wells did your model perform.

$$\textbf{R2 Squared} = 1 - \frac{SSr}{SSm}$$

SSr = Squared sum error of regression line

SSm = Squared sum error of mean line

3.1.6 Adjusted R Squared : The disadvantage of the R2 score is while adding new features in data the R2 score starts increasing or remains constant but it never decreases because It assumes that while adding more data variance of data increases.

But the problem is when we add an irrelevant feature in the dataset then at that time R2 sometimes starts increasing which is incorrect.

Hence, To control this situation Adjusted R Squared came into existence.

$$R_a^2 = 1 - \left[\left(\frac{n-1}{n-k-1}\right) \times (1 - R^2)\right]$$

where:
n = number of observations
k = number of independent variables
R_a^2 = adjusted R^2

3.2 procedure

3.2.1 any further processing?

Further data processing such as to use standardscaler is to improve the performance of the machine learning models. Many machine learning algorithms such as regression are sensitive to the scale of input feature. when some features have a larger scale than other so it can lead to biased results and reduced performance. By StanderScaler, we ensure that all feature have similar scale and thus avoid this issue. StandardScaler is typically used as a preprocessing step before training a machine learning model. It involves fitting the StandardScaler on the training data and then transforming both the training and test data using the same scaler.

This ensures that the scaling is consistent across the training and test data. The main use of StandardScaler is to standardize the features of a dataset to improve the performance of machine learning models, especially those that are sensitive to the scale of the input features.

3.2.2 Models used :

There are several models used in machine learning supervised regression, but for this use case we have continuous data i.e. target column contains house price (USD). Some of the popular one is Linear Regression,Random Forest Regression, Support Vector Regression and K-Nearest Neighbors Regression and check which model gives best suitable accuracy and which model we can deploy further.

4.Result : In the below table we compare of the multiple algorithms like Linear Regression,Random Forest Regression, Support Vector Regression and K-Nearest Neighbors Regression that I used.In our dataset we have used different algorithms that squared error as below.

Regressor	Dataset	RMSE
Linear Regression	House Price using ML algorithm	166307.310628
RandomForest Regression	House Price using ML algorithm	201882.849857
SVR (Support vector Regression)	House Price using ML algorithm	132040.277844
K-Neighbors Regression	House Price using ML algorithm	116725.996798

5. DISCUSSION

Justify four algorithms :

Regression is a supervised machine learning technique which is used to predict continuous values.The ultimate goal of the regression algorithm is to plot a best-fit line or a curve between the data.The three main metrics that are used for evaluating the trained regression model are variance, bias and error. If the variance is high, it leads to overfitting and when the bias is high, it leads to underfitting. Based on the number of input features and output labels, regression is classified as linear (one input and one output), multiple (many inputs and one output) and multivariate (many outputs).Linear regression allows us to plot a linear equation, i.e., a straight line. We need to tune the coefficient and bias of the linear equation over the training data for accurate predictions.The tuning of coefficient and bias is achieved through gradient descent or a cost function — least squares method.

5.1 Linear regression : Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.

This form of analysis estimates the coefficients of the linear equation,involving one or more independent variables that best predict the value of the dependent variable. Linear regression fits a straight line or surface that minimizes the discrepancies between predicted and actual output values. There are simple linear regression calculators that use a “least squares” method to discover the best-fit line for a set of paired data.You then estimate the value of X (dependent variable) from Y (independent variable).

5.2 RandomForest Regression : Random Forest Regression is a supervised learning algorithm that uses ensemble learning method for regression. Ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model.

5.3 SVR (Support vector Regression) :

Support Vector Regression is a supervised learning algorithm that is used to predict discrete values. Support Vector Regression uses the same principle as the SVMs. The basic idea behind SVR is to find the best fit line. In SVR, the best fit line is the hyper plane that has the maximum number of points.

Unlike other Regression models that try to minimize the error between the real and predicted value, the SVR tries to fit the best line within a threshold value.The threshold value is the distance between the hyperplane and boundary line. The fit time complexity of SVR is more than quadratic with the number of samples which makes it hard to scale to datasets with more than a couple of 10000 samples.

5.4 K-Neighbors Regression : KNN

regression is a non-parametric method that, in an intuitive manner, approximates the association between independent variables and the continuous outcome by averaging the observations in the same neighbour.

The size of the neighbourhood needs to be set by the analyst or can be chosen using cross-validation (we will see this later) to select the size that minimises the mean-squared error.While the method is quite appealing, it quickly becomes impractical when the dimension increases, i.e., when there are many independent variables.

6. Conclusion : The fundamental algorithm based on the multiple linear regression method to predict housing prices and combines it with the correlation coefficient to determine the influential factors affecting housing prices. To train and test the parameters of this multiple linear regression model, applies the data set of the housing prices is for model construction. From the simulation results shown above, it can be concluded that the proposed multiple linear regression model can effectively analyze and predict the housing price to some extent.

References :

1.Dataset from

https://www.kaggle.com/datasets/mokar2001/house-price-tehran-iran/code

2.
https://www.analyticsvidhya.com/blog/2021/05/know-the-best-evaluation-metrics-for-your-regression-model/

3.Linear Regression

https://www.ibm.com/topics/linear-regression#:~:text=Resources-,What%20is%20linear%20regression%3F,is%20called%20the%20independent%20variable.

4.RandomForestRegression

https://levelup.gitconnected.com/random-forest-regression-209c0f354c84#:~:text=Random%20Forest%20Regression%20is%20a%20supervised%20learning%20algorithm%20that%20uses,prediction%20than%20a%20single%20model.

5.KNN Regression

https://bookdown.org/tpinto_home/Regression-and-Classification/k-nearest-neighbours-