House Price Prediction Using Machine Learning

- **Vishakha D. Warudkar**

Date: 15-Jan-2022

The focus of this project is to develop Machine Learning Model that can accurately estimate the price of a House based on its Features. We will predict the pricing of house using Supervised machine learning algorithm. The prediction is totally based on past data. In this case I have collected this data from UCI Repository of Machine Learning. We are going to use Linear Regression, Random Forest Regressor, and SVM for making predictions. Here we have to evaluate and compare the prediction in order to find those which provide the best performance.

Problem Statement:

The aim is to predict the efficient house pricing for real estate customers with respect to their budgets and priorities. By analyzing previous market trends and price ranges, and also upcoming developments future prices will be predicted.

Context:

As we all know that House prices are increasing day by day. It is not possible to purchase same by everyone in actual price. There are so many brokers taking their commission while doing same. In this era of world, we don't need to assume or decide the price of house. There is superfast machine learning technique which will help us to predict the price of house on the basis of its features like number of rooms, cafeteria etc.

Agents can also use this application for making their business easy. This will be helpful for selling new or old house even they can give house on rent too with actual price.
We can implement this technique anywhere in the world to predict the pricing of the house. This Machine Learning technique includes Algorithm which will analyze data, understanding the data, prepare it for applying algorithm. On the basis of features, after knowing correlation between the each feature. Model trained and gets prepared for predicting price of the house.

External Search:

https://archive.ics.uci.edu/ml/machine-learning-databases/housing/

https://www.kaggle.com/vikrishnan/boston-house-prices

Characterization and Target Specification:

An individual wants to sell or buy a house due to various reasons such as: desire to get the new house, condition of the house etc. So, in the present scenario it is difficult to get the proper approximation of the price of the house. This model serves the purpose of predicting an appropriate value for the house so as the customers can get proper value for money in case of both selling and buying. For this case many features are used like Location, connectivity & Transport, Basic Infrastructure.

This Data Concerns housing values in suburbs of Boston.

Number of Instances: 506

Number of Attributes: 13 continuous attributes (including "class" attribute "MEDV"), 1 binary-valued attribute.

Attribute Information:

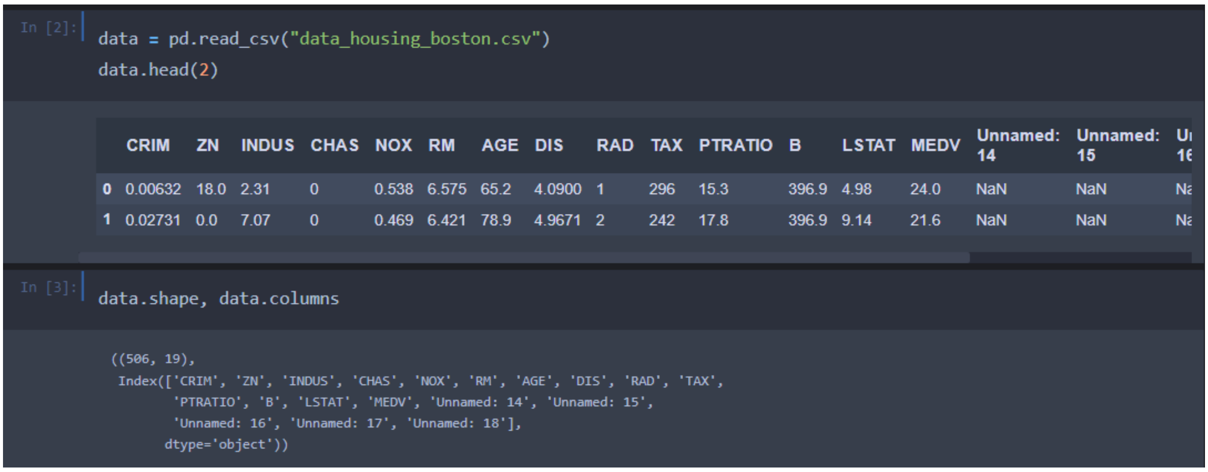
1. **CRIM** - per capita crime rate by town
2. **ZN** - proportion of residential land zoned for lots over 25,000 sq.ft.
3. **INDUS** - proportion of non-retail business acres per town
4. **CHAS** - Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
5. **NOX** - nitric oxides concentration (parts per 10 million)
6. **RM** - average number of rooms per dwelling
7. **AGE** - proportion of owner-occupied units built prior to 1940
8. **DIS** - weighted distances to five Boston employment centres
9. **RAD** - index of accessibility to radial highways
10. **TAX** - full-value property-tax rate per $10,000
11. **PTRATIO** - pupil-teacher ratio by town
12. **B** - 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
13. **LSTAT** - % lower status of the population
14. **MEDV** - Median value of owner-occupied homes in $1000's

Missing Attribute Values: None.

Benchmark:

Link: https://www.kaggle.com/altavish/boston-housing-dataset

I am using the above link from kaggle. As a Benchmark, this is helpful for developing application of machine learning model. There are so many users using this technique for their business development like Housing.com, 99 Acre.



```
In [2]: data = pd.read_csv("data_housing_boston.csv")
        data.head(2)
```

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV	Unnamed: 14	Unnamed: 15	Ui 16
0	0.00632	18.0	2.31	0	0.538	6.575	65.2	4.0900	1	296	15.3	396.9	4.98	24.0	NaN	NaN	Na
1	0.02731	0.0	7.07	0	0.469	6.421	78.9	4.9671	2	242	17.8	396.9	9.14	21.6	NaN	NaN	Na

```
In [3]: data.shape, data.columns
```

```
((506, 19),
 Index(['CRIM', 'ZN', 'INDUS', 'CHAS', 'NOX', 'RM', 'AGE', 'DIS', 'RAD', 'TAX',
        'PTRATIO', 'B', 'LSTAT', 'MEDV', 'Unnamed: 14', 'Unnamed: 15',
        'Unnamed: 16', 'Unnamed: 17', 'Unnamed: 18'],
       dtype='object'))
```

Model Concept generation:

The project is primarily about developing a system that will help people or a company to give a house, the right price. The result will be saving time and money and it is difficult to guess the actual price according to the feature, scenario and various factors of the house.

Business Opportunity:

Business opportunities nowadays many people are buying and selling new or old houses, there are many organizations 99-area, Housing.com is also attaching people who want to buy or sell their house. One of the world's leading and incoming housing marketplaces is Munich, which saw housing leaders increase by 11 per cent annually between 2018 and 2019. Although many consumers want to build their own home, others either like it or have no choice. They live in a house of surrogate.

The Covid-19 pandemic had minimal impact on the industry. With the increasing number of people preferring personal mobility and the inclusion of more finance options in the house market, the market is set to grow significantly.

These e-commerce websites do not have home price estimation facility. Also, it can be used for the purpose of predicting the price of the house that a person wants to buy.

We are going to use Machine Learning for this project. We will use python for implementation of the model,

1. **Data Preparation**:
   We need to import data with the help of pandas. And prepare it for training the model. Preparation includes processing missing values, normalization. This is very important method if there is improper format of data. Here the separation of Dependent (Target) and Independent variable (Features) is mandatory for applying Machine Learning Algorithm. Such a various technique is used for implementation of the model.

2. **EDA**:
   For better understand of data, we need to perform EDA (Exploratory Data Analysis). It provides insights that can be useful when creating later Models, as well as insights that are independently interesting.

3. **Analysing and Building the model:**
   Here for the House price prediction, Model is developed by LinearRegression, DecisionTreeRegressor, Random Forest Regressor, SVM, and XG Boost.
   In this dataset, except MEDV all are the independent variable. And MEDV is the dependent variable. For applying desired model, we will split the data into training and testing set. Model will only be trained on Training set of data. Testing set needs to keep for testing purpose. This will show the correctness and completeness of the trained model.

   Support Vector Machine (**SVM**) gives best result for predicting the pricing of the house.

Model Output:

- LinearRegression:
  Mean: 4.806501319130534,
  Standard Deviation: 1.1704297637633674

- DecisionTreeRegressor :
  Mean: 4.518179323644875,
  Standard Deviation: 2.4806075618751002

- Random Forest Regressor:
  Mean: 3.3688847279815577,
  Standard Deviation: 1.5304354547864867

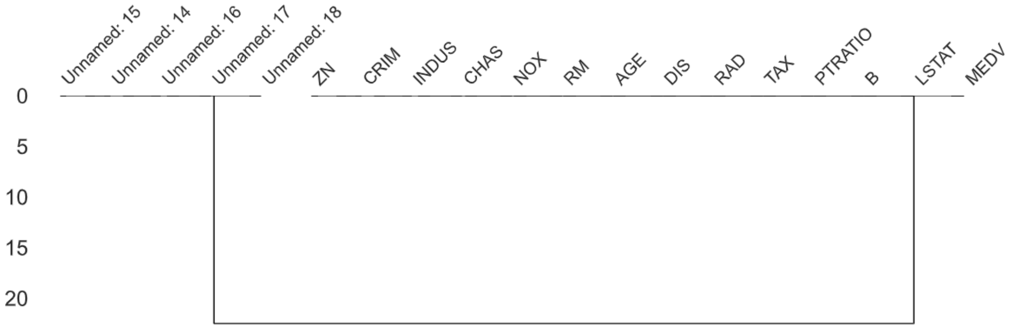
- SVM:
  Mean: 7.708261736098047,
  Standard Deviation: 2.4004880820482315

- XG Boost:
  Mean: 3.0359095119041912,
  Standard Deviation: 1.2188007473795597

Data Visualization:

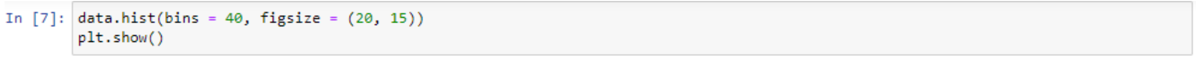
The resulting visual representation of data makes it easier to identify and share real-time trends, outliers, and new insights about the information represented in the data.

The dendrogram allows you to more fully correlate variable completion, revealing trends deeper than the pair wise ones visible in the correlation heat map.

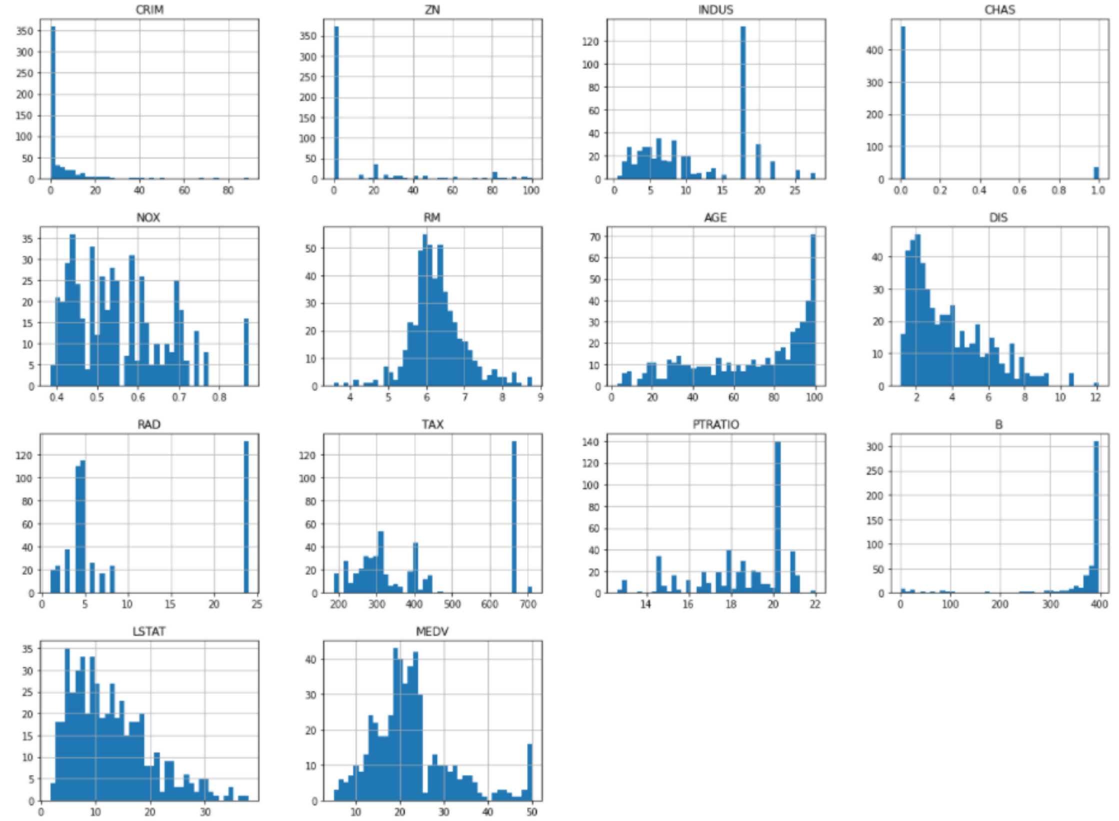


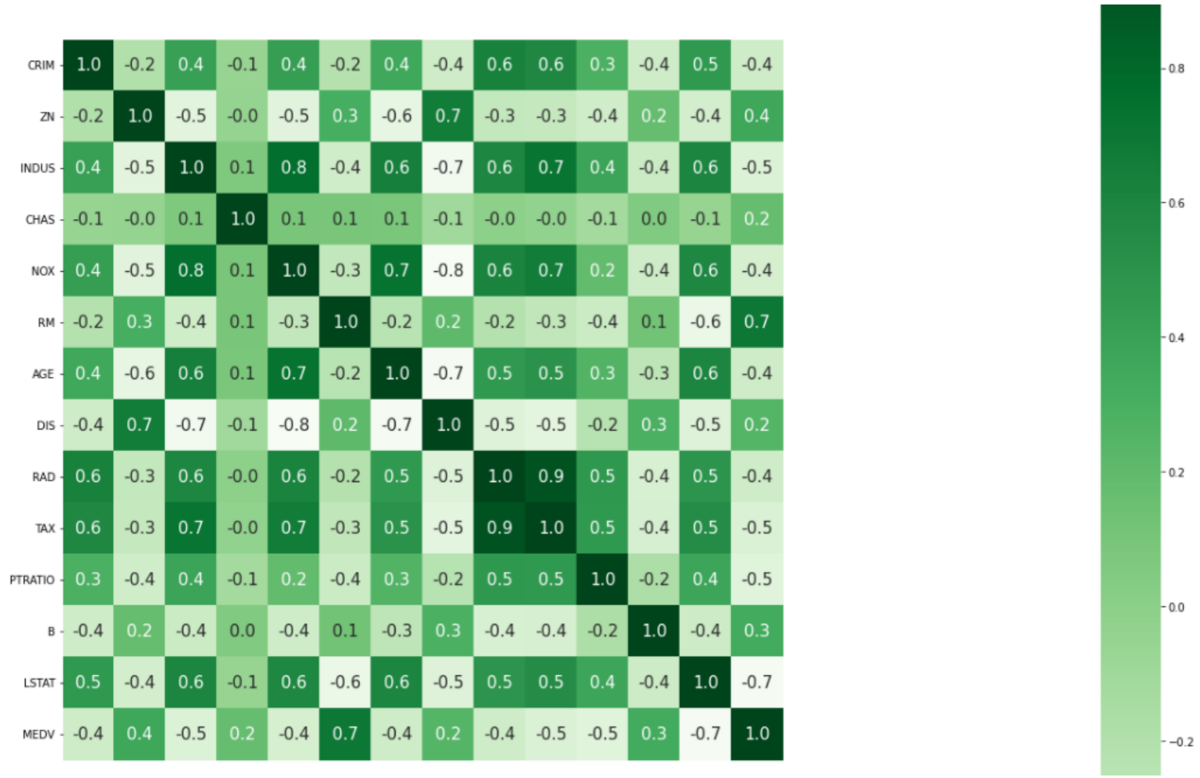
HeatMap & Histogram Visualization:

Checking the correlation between the features.



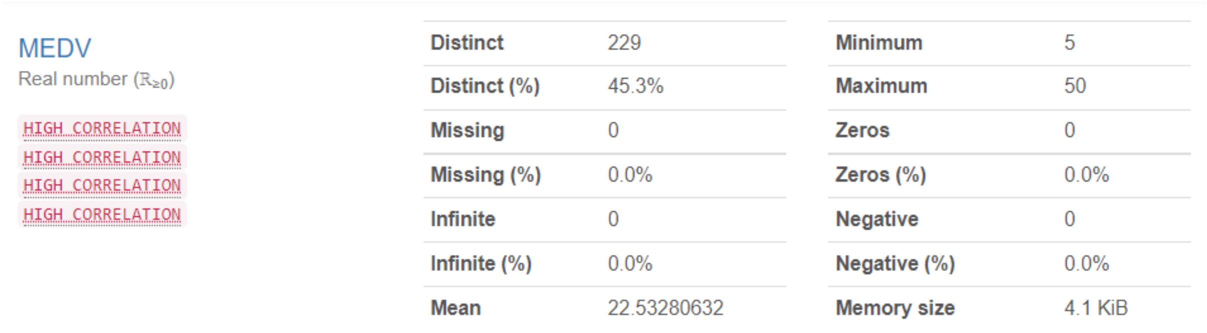
```
In [7]: data.hist(bins = 40, figsize = (20, 15))
        plt.show()
```





CRIM	1.0	-0.2	0.4	-0.1	0.4	-0.2	0.4	-0.4	0.6	0.6	0.3	-0.4	0.5	-0.4
ZN	-0.2	1.0	-0.5	-0.0	-0.5	0.3	-0.6	0.7	-0.3	-0.3	-0.4	0.2	-0.4	0.4
INDUS	0.4	-0.5	1.0	0.1	0.8	-0.4	0.6	-0.7	0.6	0.7	0.4	-0.4	0.6	-0.5
CHAS	-0.1	-0.0	0.1	1.0	0.1	0.1	0.1	-0.1	-0.0	-0.0	-0.1	0.0	-0.1	0.2
NOX	0.4	-0.5	0.8	0.1	1.0	-0.3	0.7	-0.8	0.6	0.7	0.2	-0.4	0.6	-0.4
RM	-0.2	0.3	-0.4	0.1	-0.3	1.0	-0.2	0.2	-0.2	-0.3	-0.4	0.1	-0.6	0.7
AGE	0.4	-0.6	0.6	0.1	0.7	-0.2	1.0	-0.7	0.5	0.5	0.3	-0.3	0.6	-0.4
DIS	-0.4	0.7	-0.7	-0.1	-0.8	0.2	-0.7	1.0	-0.5	-0.5	-0.2	0.3	-0.5	0.2
RAD	0.6	-0.3	0.6	-0.0	0.6	-0.2	0.5	-0.5	1.0	0.9	0.5	-0.4	0.5	-0.4
TAX	0.6	-0.3	0.7	-0.0	0.7	-0.3	0.5	-0.5	0.9	1.0	0.5	-0.4	0.5	-0.5
PTRATIO	0.3	-0.4	0.4	-0.1	0.2	-0.4	0.3	-0.2	0.5	0.5	1.0	-0.2	0.4	-0.5
B	-0.4	0.2	-0.4	0.0	-0.4	0.1	-0.3	0.3	-0.4	-0.4	-0.2	1.0	-0.4	0.3
LSTAT	0.5	-0.4	0.6	-0.1	0.6	-0.6	0.6	-0.5	0.5	0.5	0.4	-0.4	1.0	-0.7
MEDV	-0.4	0.4	-0.5	0.2	-0.4	0.7	-0.4	0.2	-0.4	-0.5	-0.5	0.3	-0.7	1.0

Visualization of Target (Dependent Variable) :



MEDV

Real number ($\mathbb{R}_{\geq 0}$)

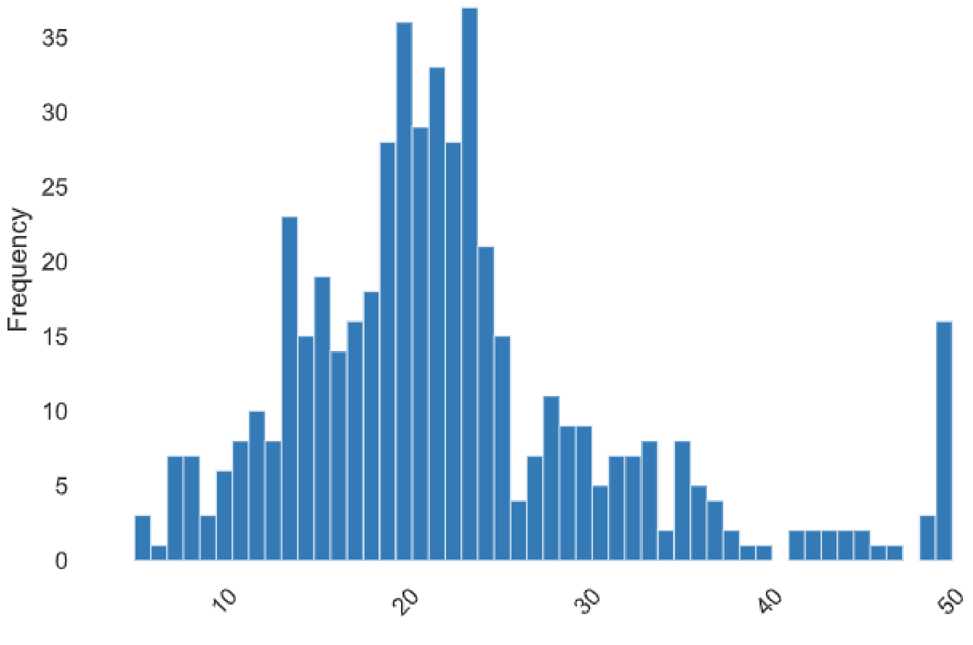
HIGH CORRELATION

HIGH CORRELATION

HIGH CORRELATION

HIGH CORRELATION

Distinct	229	Minimum	5
Distinct (%)	45.3%	Maximum	50
Missing	0	Zeros	0
Missing (%)	0.0%	Zeros (%)	0.0%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	22.53280632	Memory size	4.1 KiB



Histogram with fixed size bins (bins=50)

Model development:

This algorithmic training and optimization needs to be completed to reduce over-fitting of the model and hyper parameter tuning

Performing EDA will give us exact scenario of independent and dependent variable.

Model development is very much important before releasing the service.

For optimizing automation task, lot of supervised and unsupervised machine learning algorithm needs to be performed.

Libraries: NumPy, pandas, pandas_profiling, joblib, seaborn, matplotlib, scikit-learn

Data Collection and ML Model:

This data is available in UCI Machine Learning Repository. SVM giving best result for prediction of house. SVM is a supervised machine learning algorithm which can be used for classification or regression problems. It uses a technique called the kernel trick to transform your data and then based on these transformations it finds an optimal boundary between the possible outputs. For this we need expert team for the development of this complete model.

Code Implementation & EDA Report:

https://github.com/vishakhawarudkar96/HousePricePrediction.git

https://drive.google.com/file/d/1rOEWzv7Sba4DEE2gZbWBGVRnOvr-BZqi/view?usp=sharing

Conclusion:

If implemented properly, managing team and also guiding them properly. We can achieve the expected result by using **SVM** Algorithm. By analyzing previous market trends and price ranges, and also upcoming future development price will be predicted. When compared to all other algorithms, The Support Vector Machine (**SVM)** Algorithm achieved the highest Mean and the standard variation

References:

1. W3 School https://www.w3schools.com/python/python_ml_standard_deviation.asp#:~:text=A%20low%20standard%20deviation%20means,out%20over%20a%20wider%20range.&text=37.85-,Meaning%20that%20most%20of%20the%20values%20are%20within%20the%20range,mean%20value%2C%20which%20is%2077.4.
2. Pandas Profiling https://pandas-profiling.github.io/pandas-profiling/docs/master/rtd/
3. SKlearn SVM Documentation https://scikit-learn.org/stable/modules/svm.html