**BOSTON HOUSE PRICE PREDICTION USING MACHINE LEARNING**

**ABSTRACT**

Everyone wishes to buy and live in a house which suits their lifestyle and which provides amenities according to their needs. There are many factors that are to be taken into consideration like area, location, view etc. for prediction of house price. It is very difficult to predict house price as it is constantly changing and quite often the prices are exaggerated for which people who want to buy houses, and various real estate agencies who want to invest in properties, find it difficult to buy or sell houses.For this reason, in this project an advanced automated Machine Learning model is created using Linear Regression, Random Forest, XGBoost and Support Vector Machine using the Boston house dataset to predict house price in boston future accurately, and to measure the accuracy of these models various measuring metrics like R-Squared, Root Mean Square Error (RMSE), Mean Square Error(MSE) and Mean Absolute Error(MAE). This project also studies the correlation of various attributes of the Boston dataset using the heat map to see which attributes actually impact the prediction of the models. In this project it is observed that XGBoost performs better in all the measuring matrices whereas Support Vector Machine performs poor in all of the measuring matrices.

**INTRODUCTION**

Accurately predicting the value of a plot or house is an important task for many house owners, house buyers, plot owners, plot buyers or stake holders. Real estate agencies and people buy and sell houses all the time, people buy houses to live in or as an investment whereas real estate agencies buy it to run a business. But the problem arises in evaluation of the cost of the property. Over-validation / Under-validation have always been the issues faced in house markets due to lack of proper detection measures. It is also very difficult task. We know that features like size, area, location etc affect the price of the property but there are many other features also which affect the property such as inflation rates in market, age of the property etc. In order to overcome these problems a throw analysis is done using Machine Learning (ML) which is a branch of Artificial Intelligence (AI).

**PROJECT STATEMENT**

The idea behind this project is to build a model that will predict the price of house in boston. It is based on per capita crime rate by town, average number of rooms per dwelling, the proportion of owner-occupied units built prior to 1940, The median value of owner-occupied, etc.

### APPROACH

In this research Linear Regression, Random Forest, XGBoost and Support Vector Machine has been used to predict house price in boston using 14 features of homes from various suburbs located in Boston.

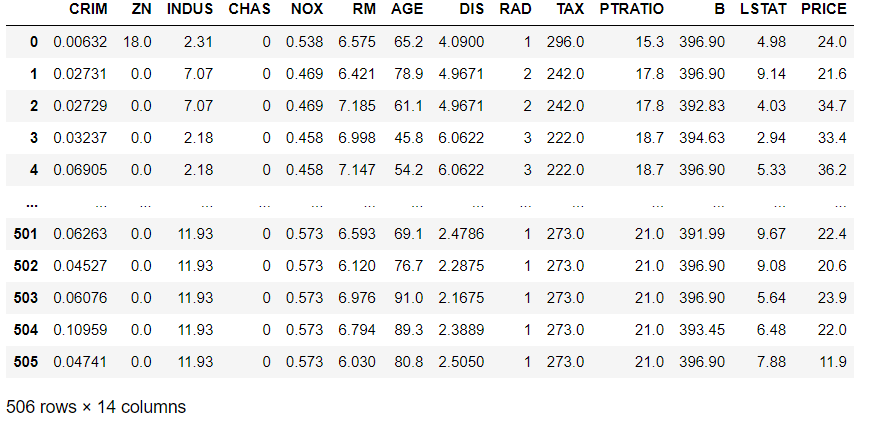
**DATA DESCRIPTION**

The dataset used in this project comes from the UCI Machine Learning Repository which concerns housing values in the suburbs of Boston. This data was collected in 1978 and contains 506 entries which give information about 14 attributes of homes from various suburbs located in Boston and one “target” attribute. The attribute description of this dataset is given below:

* CRIM: This is the per capita crime rate by town
* ZN: This is the proportion of residential land zoned for lots larger than 25,000 sq.ft.
* INDUS: This is the proportion of non-retail business acres per town.
* CHAS: This is the Charles River dummy variable (this is equal to 1 if tract bounds river; 0 otherwise)
* NOX: This is the nitric oxides concentration (parts per 10 million)
* RM: This is the average number of rooms per dwelling
* AGE: This is the proportion of owner-occupied units built prior to 1940
* DIS: This is the weighted distances to five Boston employment centers
* RAD: This is the index of accessibility to radial highways
* TAX: This is the full-value property-tax rate per $10,000
* PTRATIO: This is the pupil-teacher ratio by town
* B: This is calculated as 1000(Bk — 0.63)², where Bk is the proportion of people of African American descent by town
* LSTAT: This is the percentage lower status of the population
* PRICE: This is the median value of owner-occupied homes in $1000s

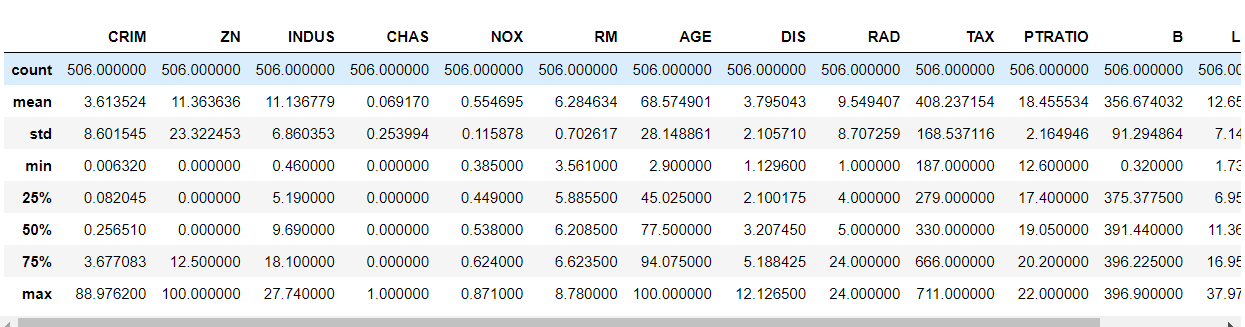
The “target” variable in this dataset is the PRICE variable on which will be predicted by the ML models. The rest of the variables are used for training the models. Statistics of the features are described in the table below:

This is an overview of the original dataset, with its original features:



The dataset doesn’t contain any null value nor does it contain any duplicate row. The dataset contain 13 numerical values and only one categorical value which is “CHAS”.

The statistics of dataset are as follows:



**METHODOLOGY**

Machine Learning has various analyses like classification, regression, clustering and association and this analysis use various algorithms like logistic regression, Simple linear regression, support vector machine, Naïve Bayes etc. The analysis used here is regression, as regression is used for predicting continuous variable and house price prediction is prediction of continuous value. Various statistical methods are used for regression analysis to model the relationship between a dependent/ target variable and the independent or the predictor variable with one or more independent variables. The change of a dependent variable corresponding to an independent variable when other variables are kept fixed is understood using regression analysis. Certain features such as temperature, age, salary, house price etc. continuous/real values are predicted using regression.

**MODEL SELECTION**

Model selection is one of the most important tasks in ML for doing accurate prediction. Correct models must be selected to get good accuracy. There are various models available under regression analysis but for this paper four regression models are used which are Linear Regression, Random Forest, XGBoost and Support Vector Machine on the Boston house dataset. After implementation of these models we measure the accuracy by splitting the dataset into two parts which are training dataset and test dataset. We use 70% of the dataset as training data and 30% is used as test data. The fitting of our models is done using the training dataset and evaluation of the model is done in test dataset.

**ALGORITHMS USED**

**1:LOGISTIC REGRESSION:**

Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable.

The linear regression model provides a sloped straight line representing the relationship between the variables.

**2:RANDOM FOREST**

 It is based on the concept of **ensemble learning,** which is a process of *combining multiple classifiers to solve a complex problem and to improve the performance of the model.*

As the name suggests, ***"Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset."*** Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

**The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.**

**3:XGBOOST**

Gradient boosted decision trees are implemented by the XGBoost library of Python, intended for speed and execution, which is the most important aspect of ML (machine learning).

**XgBoost**: XgBoost (Extreme Gradient Boosting) library of Python was introduced at the University of Washington by scholars. It is a module of Python written in C++, which helps ML model algorithms by the training for Gradient Boosting.

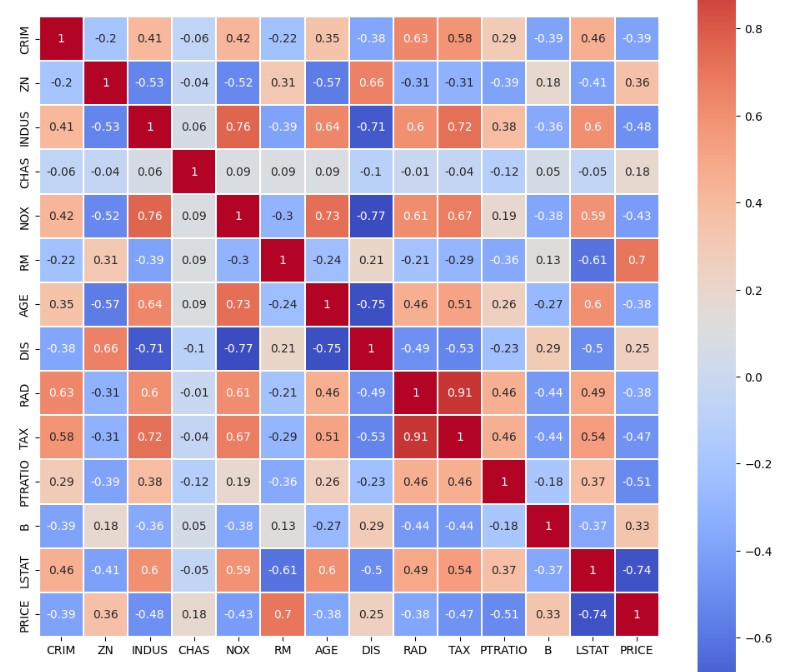
**Gradient boosting:** This is an AI method utilized in classification and regression assignments, among others. It gives an expectation model as a troupe of feeble forecast models, commonly called decision trees.

**4:SUPPORT VECTOR MACHINE**

The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

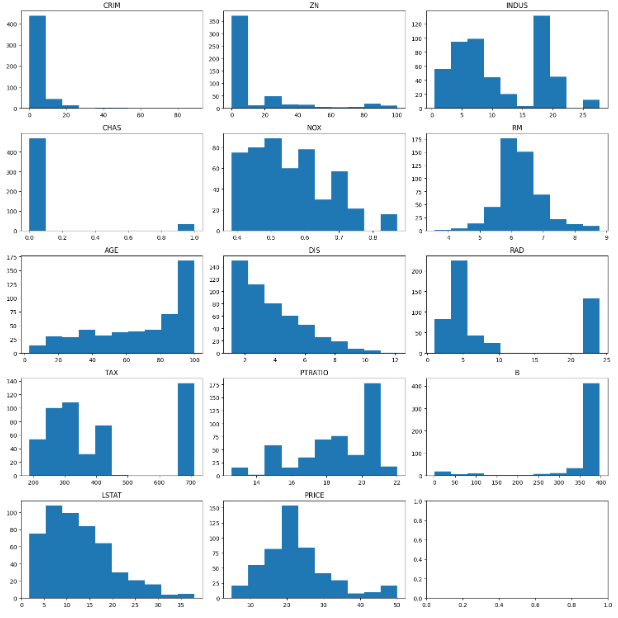
SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine.

**DATA ANALYSIS:**



From heat map it is observed that the attributes “TAX” and “RAD” are highly correlated. As both of them are highly correlated they are having similar behavior and will also have similar impact while doing prediction calculation.

Histograms of each feature to see the distribution of the data and identify any outliers.



**RESULTS:**

### LINEAR REGRESSION:

After fitting the training data in linear regression model **we print the values of coefficients. These coefficients can be used to find the value of house price.**



For **Model Evaluation,** we will print the R^2 Score, Adjusted R^2 Score, MAE, MSE and RMSE Score.

**𝑅^2** : It is a measure of the linear relationship between X and Y. It is interpreted as the proportion of the variance in the dependent variable that is predictable from the independent variable.

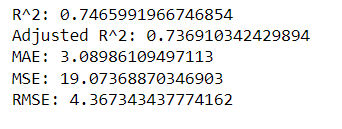
**Adjusted 𝑅^2** :The adjusted R-squared compares the explanatory power of regression models that contain different numbers of predictors.

**MAE** : It is the mean of the absolute value of the errors. It measures the difference between two continuous variables, here actual and predicted values of y.

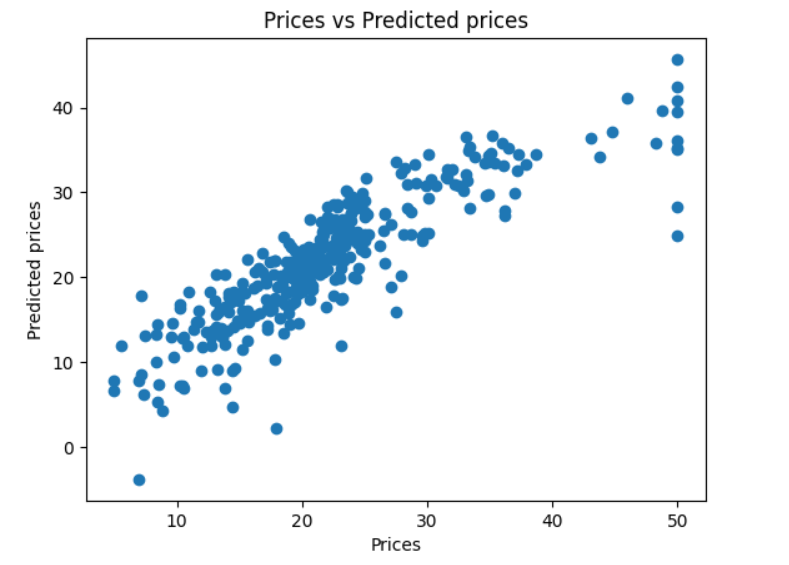
**MSE**: The mean square error (MSE) is just like the MAE, but squares the difference before summing them all instead of using the absolute value.

**RMSE**: The mean square error (MSE) is just like the MAE, but squares the difference before summing them all instead of using the absolute value.

The R^2, Adjusted R^2, MAE, MSE and RMSE Score of linear regression model are as follows:



Visualizing the differences between actual prices and predicted values.

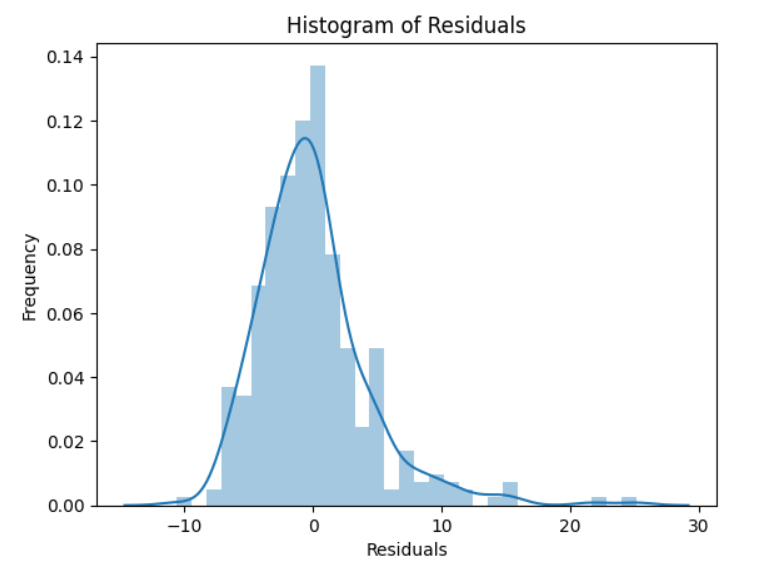


Checking residuals with predicted values

### 

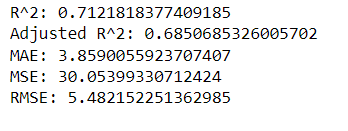
There is no pattern visible in this plot and values are distributed equally around zero. So Linearity assumption is satisfied.

**Normality of Errors**



Here the residuals are normally distributed. So normality assumption is satisfied.

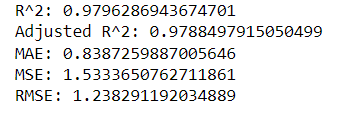
**MODEL EVALUATION for Testing Data**



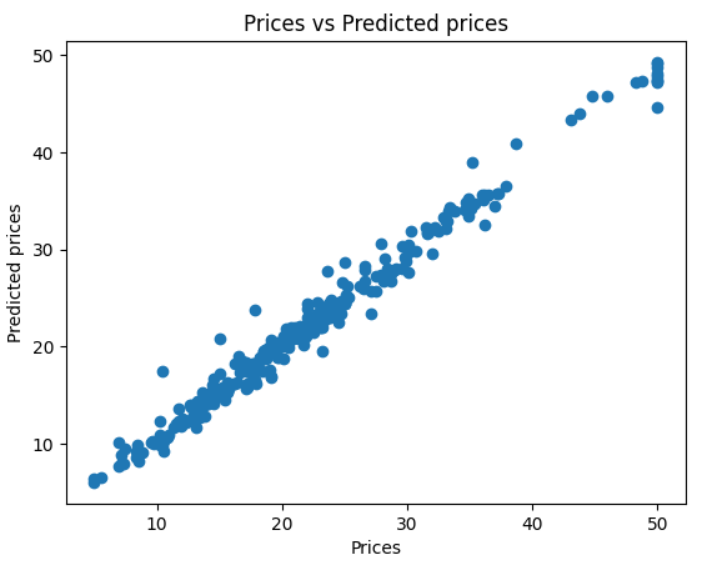
Here the model evaluations scores are almost matching with that of train data. So the model is not overfitting.

**RANDOM FOREST**

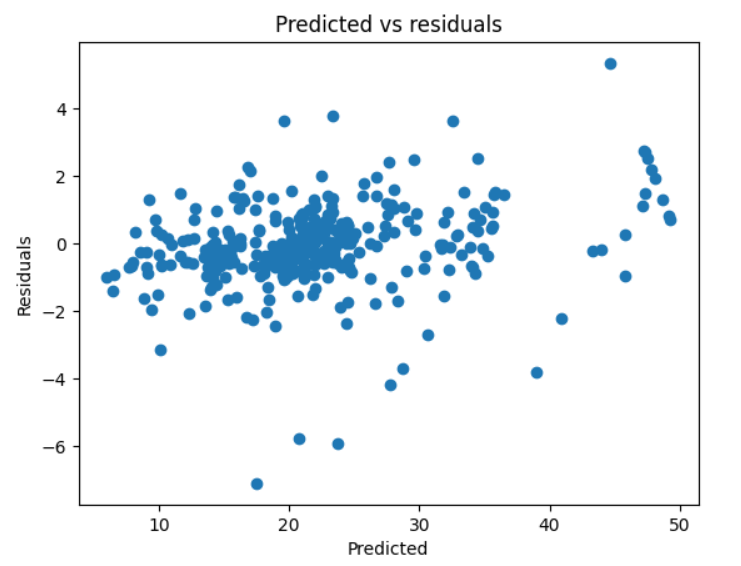
Model Evaluation of Random Forest



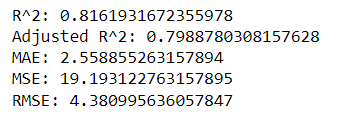
Visualizing the differences between actual prices and predicted values.



Checking residuals with predicted values

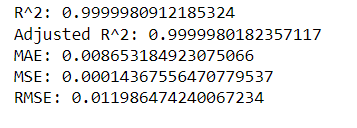


Model Evaluation Score for Testing data:



**XGBOOST REGRESSOR**

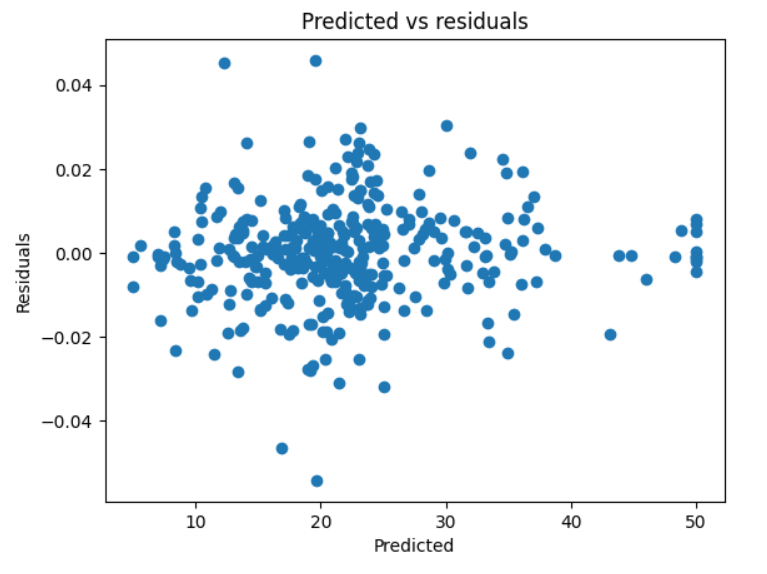
Model Evaluation of Random Forest



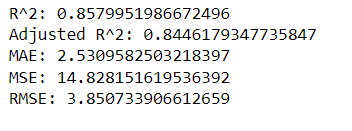
Visualizing the differences between actual prices and predicted values



Checking residuals with predicted values

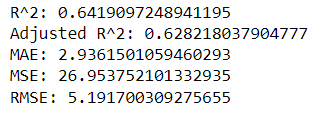


Model Evaluation Score for Testing data:

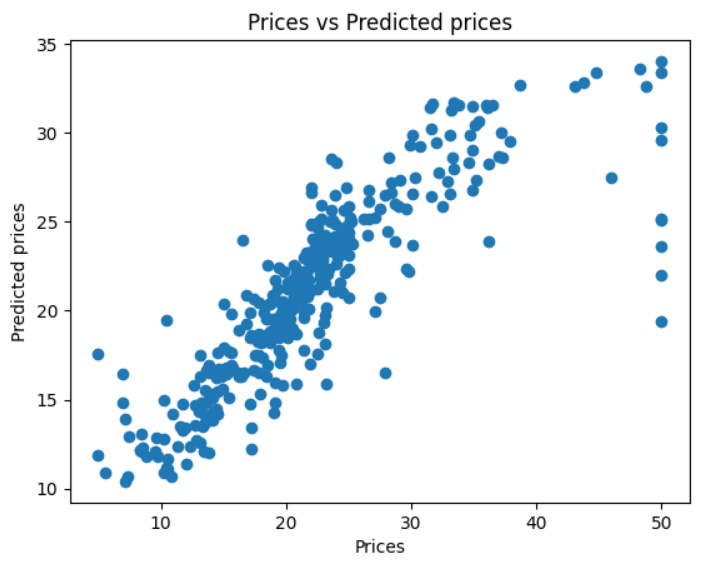


**SUPPORT VECTOR MACHINE**

Model Evaluation of SVM

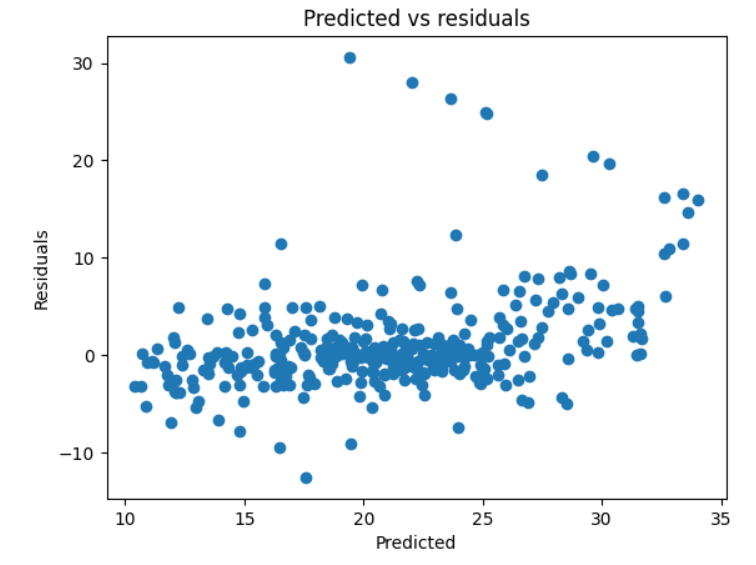


Visualizing the differences between actual prices and predicted values

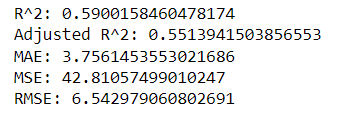


There is big difference in actual and predicted prices.

Checking residuals with predicted values

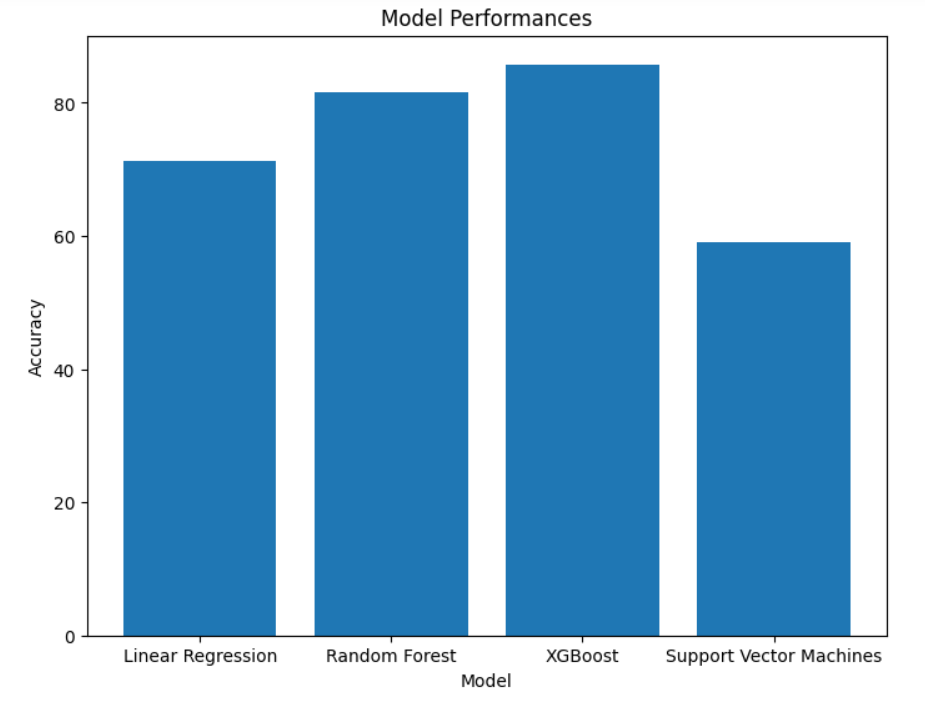


Model Evaluation Score for Testing data:



# Evaluation and Comparison of all the Models

Comparing the R^2 Score of all the models by plotting histogram and finding the best model out of all.



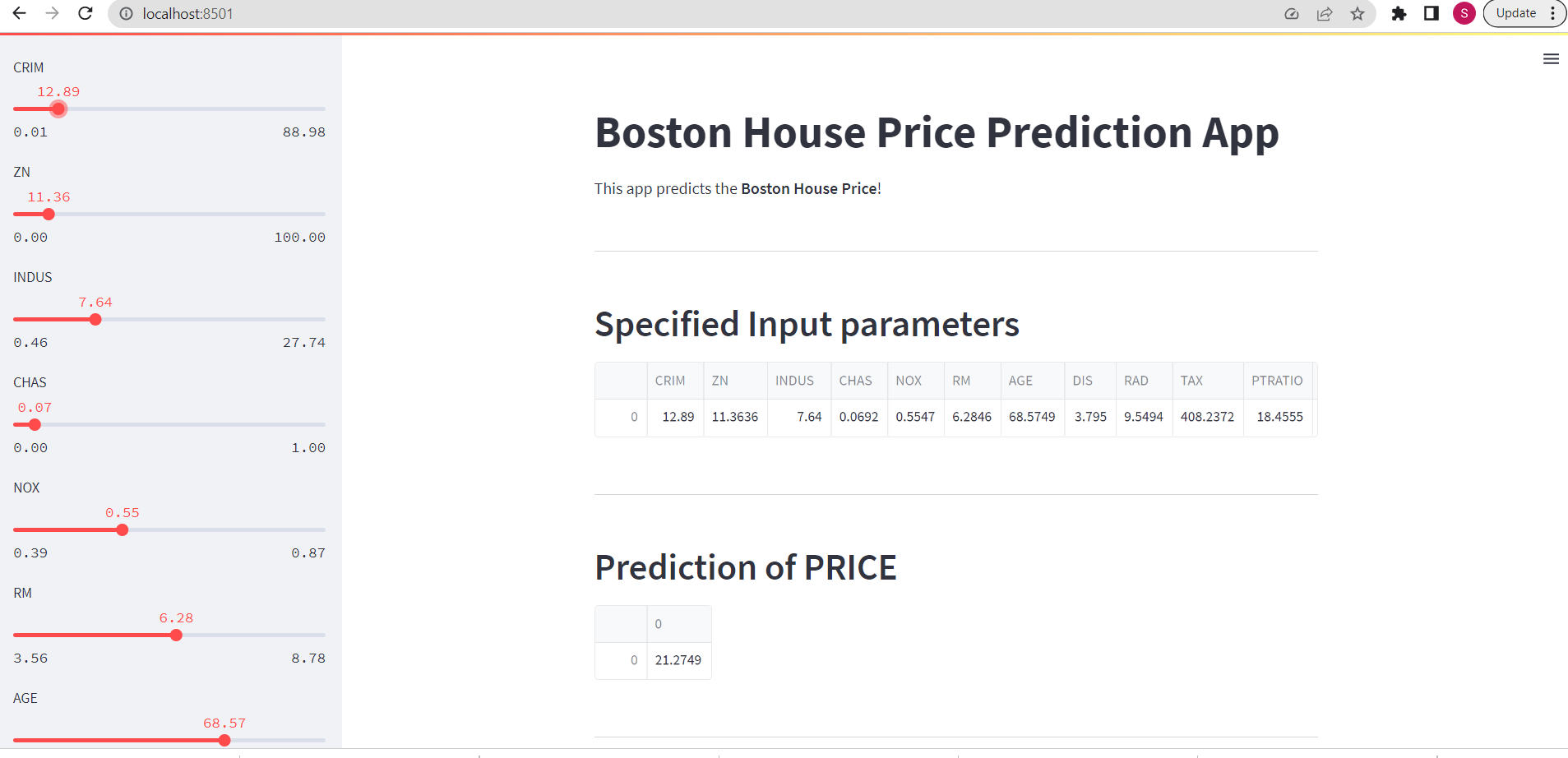
From the plot, we can say that XGBoost has the highest R^2 Score.

## Hence XGBoost Regression works the best for this dataset.

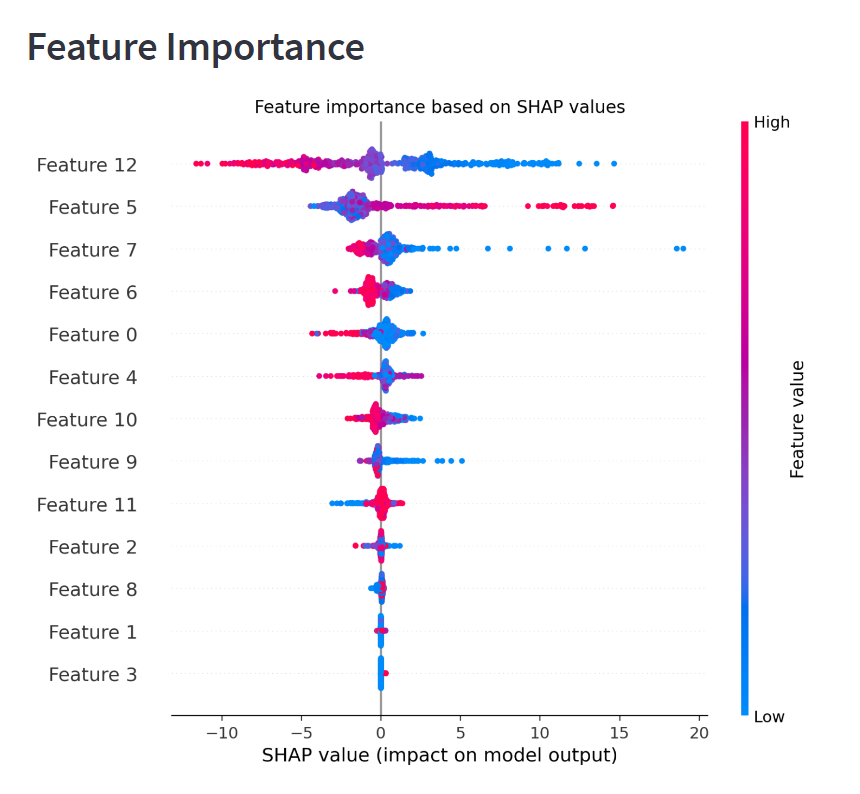
**WEB APPLICATION**

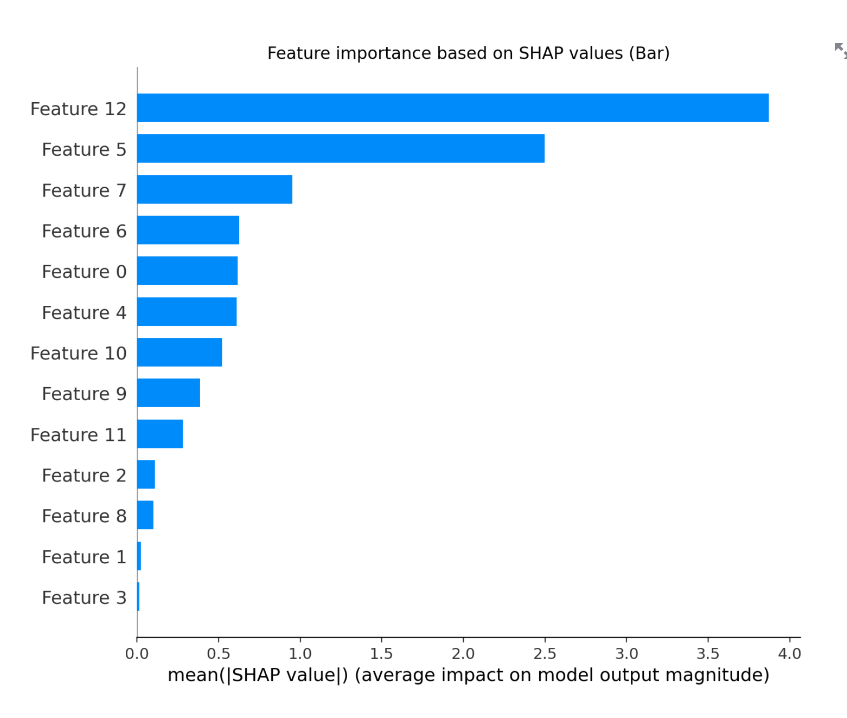
A web application has been created with the best model i.e XGBoost Regressor for prediction of house prices in Boston using **Streamlit** library.

In this application, user can input their values and based on that values a price value has been predicted using XGBoost model. The input variables are the training features used while building the model.



**Feature importance graphs have been created based on SHAP values.**





**FUTURE WORK**

This model can be considered as the baseline for predicting house price. Further evaluation can be done here by increasing the data. More data can be collected and more attributes can be increased for getting a much better evaluation of the model. The data collected in Boston house dataset is from 1978 which is almost 50 years old and since then a lot of changes have occurred in house price due to inflation rate. Thus, new data can be collected and further evaluation can be made on the new collected data.

**CONCLUSION**

It is very important to predict house price accurately. To accurately predict house price various variables must be taken into consideration like location of the house, the views that are visible from the house, crime rate around that area etc. A lot of time people pay overprice from the actual market price for a real estate property, similarly a lot of time sellers get very low price compare to the actual market price of the property. Not only people, various estate agencies also face the same problem where they are not sure whether to invest toward a certain property or not. They are confused as they are not able to predict what the price of the house can be in future. The main purpose of this paper is to help people who are facing these issues to predict the house price in future years. In this paper an intelligent system is made using the Regressor models which are Simple Linear Regression, Random Forest, XGBoost and SVM on the Boston House Dataset to predict the house price. In this paper it is observed that using the Boston house dataset, and implementing various data preprocessing techniques which are needed on the dataset and then splitting the dataset into 70-30, 70% for training and 30% for testing, XGBoost Regression performs the best with R^2 score of 0.98. The best performance is given by all the methods which use regularization techniques. This may change if we use different dataset or different pre-processing approaches but in this case this is the results that we get.

**REFERENCES**

[**https://www.kaggle.com/code/mastmustu/boston-house-price-prediction**](https://www.kaggle.com/code/mastmustu/boston-house-price-prediction)

[**https://www.kaggle.com/code/anurag629/boston-house-prices-regression-model**](https://www.kaggle.com/code/anurag629/boston-house-prices-regression-model)

[**https://www.enjoyalgorithms.com/blog/boston-house-price-prediction-using-machine-learning**](https://www.enjoyalgorithms.com/blog/boston-house-price-prediction-using-machine-learning)

[**https://towardsdatascience.com/machine-learning-project-predicting-boston-house-prices-with-regression-b4e47493633d?gi=090502dd58d2**](https://towardsdatascience.com/machine-learning-project-predicting-boston-house-prices-with-regression-b4e47493633d?gi=090502dd58d2)