

**House price prediction with analysis and visualization**

A Data Visualization(CSE3020) project report

**Submitted to:**

**Dr. Parvathi.R**

**Submitted by:**

Alpanshu Kataria (18BCE1267)

In partial fulfilment for the award of the degree of

**Bachelor of Technology**

in

**Computer Science and Engineering**

*Feburary- June 2021*

**Abstract:**

We know that the price of a house depends on numerous factors. Therefore, it sometimes becomes very difficult for a person to estimate the price of the house before buying it, as a result the seller can ask for whatever price he wants and the buyer will never know if he got a good deal or he just got cheated. The scenario can be seen from the other side also, as sometimes sellers are also not able to ask for the legitimate price of the house as they do not know what is its actual worth. Therefore, I decided to use Machine Learning and data visualization which will help to estimate the actual worth of the house based on various factors. Machine learning plays a major role from past years in image detection, spam reorganization, normal speech command, product recommendation and medical diagnosis. Present machine learning algorithm helps us in enhancing security alerts, ensuring public safety and improve medical enhancements. Machine learning system also provides better customer service and safer automobile systems. In this project I have discussed about the prediction of future housing prices that is generated by machine learning algorithm. For prediction we have applied and compared various machine learning algorithms such as Linear Regression, Random Forest, SVM, XGBoost. Our result exhibit that our approach of the issue need to be successful, and has the ability to process predictions that would be comparative with other house cost prediction models. This project utilizes machine learning algorithms as a research method that develops housing price prediction models. We in that point recommend a housing cost prediction model to support a house vender or a real estate agent for better information based on the valuation of house. Those examinations exhibit that Random Forest algorithm, in view of accuracy, reliably outperforms alternate models in the execution of housing cost prediction.

**Introduction:**

What Is Learning? Rats Learning to Avoid Poisonous Baits: Rats normally stumble upon food items by its look and smell and start eating in small amounts and later depending on food and physiological effect the feeding of food goes on. If the rat notices the illness of the food, the rat will not touch that food. Similarly, the machine learning mechanism plays a vital role same as animal usage of past experience for acquiring and expertise in detecting the food safety. By mistake if the knowledge with the food is negatively labelled, the prediction of the animal will also will be negatively affected and encountered in the future. With the inspiration of the previous example of successful learning we demonstrate a typical machine learning algorithm. Likewise, we would want to program a machine that learns how to filter spam e-mails. A credulous result might be apparently comparable of the lifestyle of rats that, how to keep away from poisonous baits. The machine will basically remember the past e-mails that needed been named similarly as spam e-mails by the human user. When another email arrives, the machine will look for it in the past set about spam e-mails. Though it matches among them, it will be trashed. Otherwise, it will make moved of the user’s inbox organizer. At the same time the first “learning by memorization” methodology may be useful, it fails to offer an important aspect known as learning systems – the capacity to mark unseen email messages. A fruitful learner ought to have the ability which will be the advancement from distinctive samples to more extensive generalization. This may be Likewise as inductive thinking or inductive induction. With an ever increasing amount accessible digitally recorded data, the scope of machine learning is increasing day by day. In this project we have tried to apply machine learning to predict the price of a house based on various parameters. The solution would prove to be very useful for both buyer as well as seller in determining the true price of the house.

**About the dataset:**

For this project I have used The Boston housing dataset.

Link: <http://lib.stat.cmu.edu/datasets/boston>

The Boston Housing Dataset is a derived from information collected by the U.S. Census Service concerning housing in the area of Boston MA. The following describes the dataset columns:

ZN - proportion of residential land zoned for lots over 25,000 sq.ft.

INDUS - proportion of non-retail business acres per town.

CHAS - Charles River dummy variable (1 if tract bounds river; 0 otherwise)

NOX - nitric oxides concentration (parts per 10 million)

RM - average number of rooms per dwelling

AGE - proportion of owner-occupied units built prior to 1940

DIS - weighted distances to five Boston employment centres

RAD - index of accessibility to radial highways

TAX - full-value property-tax rate per $10,000

PTRATIO - pupil-teacher ratio by town

B - 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town

LSTAT - % lower status of the population

MEDV - Median value of owner-occupied homes in $1000's

**Language and libraries used:**

I have done this project using python language. Python is widely used in scientific and numeric computing. Libraries used for this project are:

* Pandas
* NumPy
* Matplotlib
* Seaborn
* Scikit Learn
* XG Boost

**Models Used:**

**Linear Regression:**

* Linear Regression is a machine learning algorithm based on supervised learning.
* It performs a regression task. Regression models a target prediction value based on independent variables.
* It is mostly used for finding out the relationship between variables and forecasting**.**

**SVM:**

* Support Vector Machine (SVM) is a relatively simple **Supervised Machine Learning Algorithm** used for classification and/or regression.
* SVM finds a hyper-plane that creates a boundary between the types of data. In 2-dimensional space, this hyper-plane is nothing but a line.
* SVM works very well without any modifications for linearly separable data.

**Random forest Regression Model:**

* A Random Forest is an ensemble technique capable of performing both regression and classification tasks with the use of multiple decision trees and a technique called Bootstrap Aggregation, commonly known as bagging.
* Bagging, in the Random Forest method, involves training each decision tree on a different data sample where sampling is done with replacement.
* The basic idea behind this is to combine multiple decision trees in determining the final output rather than relying on individual decision trees.

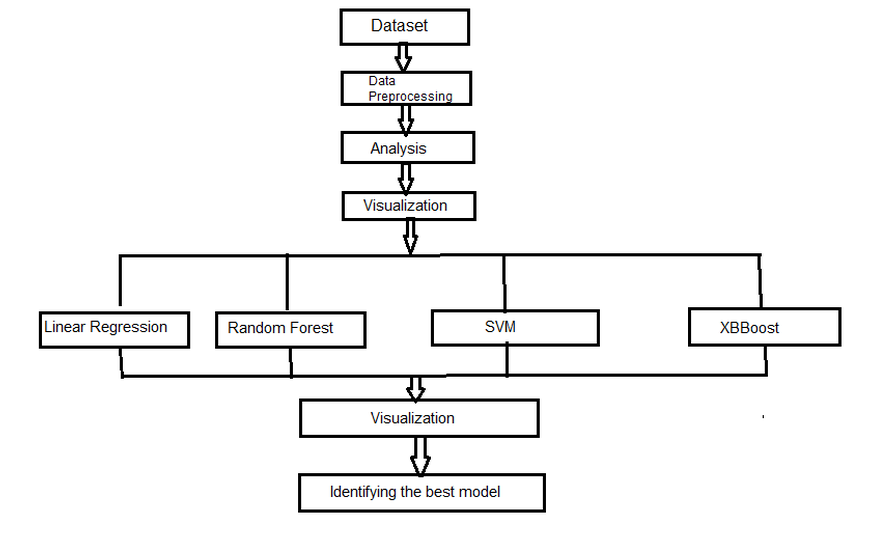
**XGBoost Regressor model:**

* XG Boost stands for eXtreme Gradient Boosting.
* The XG Boost library implements the gradient boosting decision tree algorithm.
* Boosting is an ensemble technique where new models are added to correct the errors made by existing models.
* Models are added sequentially until no further improvements can be made.

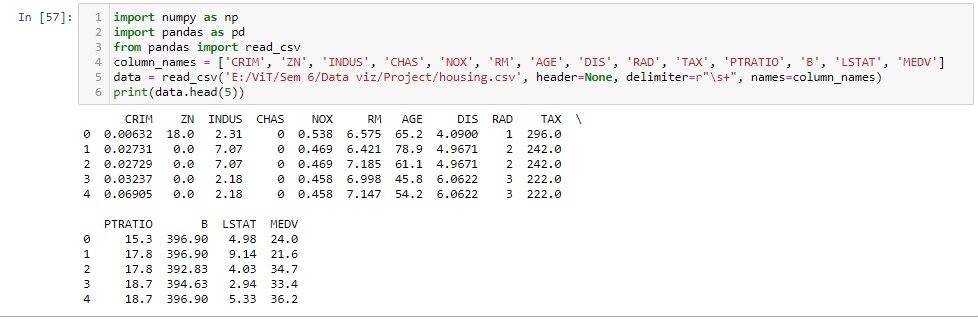
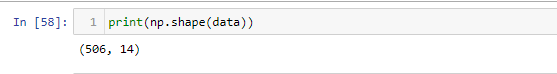
**Methodology:**

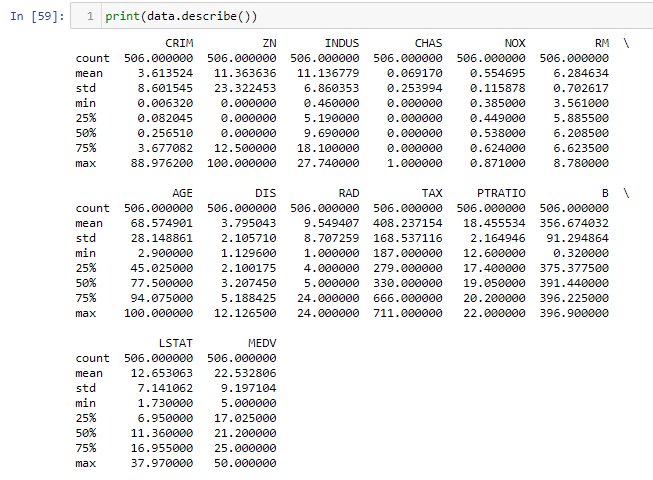
First of all, all the necessary libraries are imported. Then we start analysing the dataset. By using describe() method we can know about which of the variables are categorical so that they can be eliminated during regression. We can also analyse the outliers for each column so that they can also be eliminated, because they will not help us in prediction. Boxplots will help us in this scenario. To check if any of the columns is showing skewed distribution, we can plot the histograms. Heat maps can be used to visualize the correlation among various variables. All the variables which are categorical they can be specifically visualized and analysed using strip plot and violin plot. Width of the violin plot will tell us the approximate frequency of data in a particular region. Then we’ll apply various machine learning models on our dataset. The outputs of these models will be analysed using scatter plots and histogram. Actual price vs predicted price and price vs residual, these plots are very important to analyse the performance of the model. The accuracies and performance of each model will be analysed and finally the best performing model will be identified.

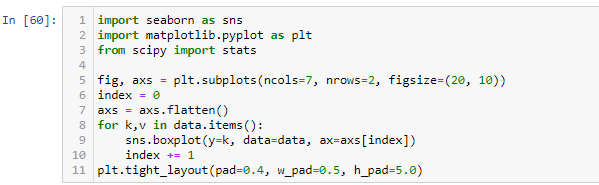
**Architecture Diagram:**

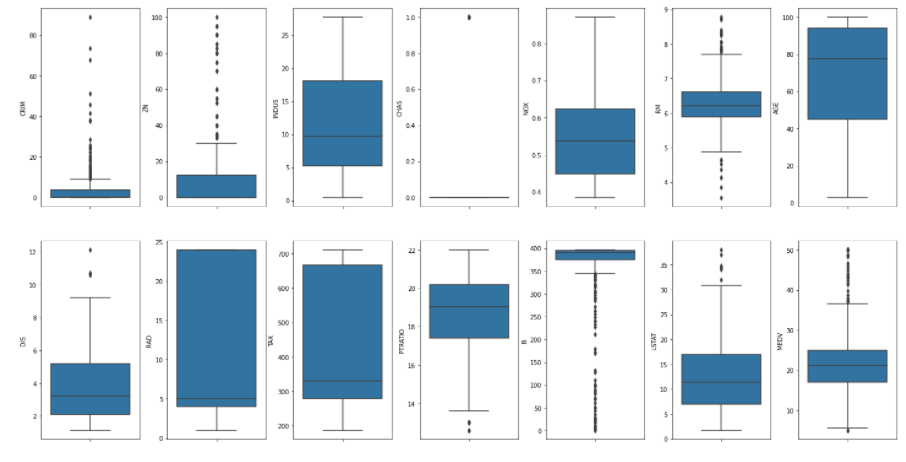


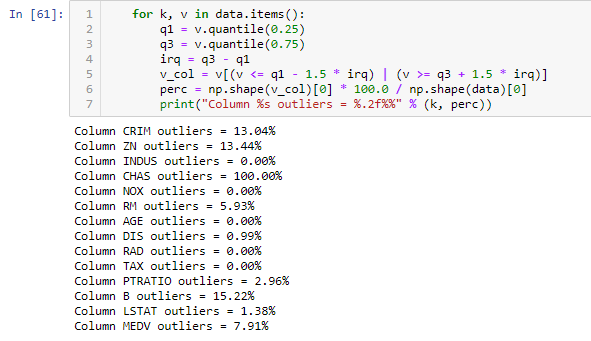
**Code, Implementation and Outputs:**

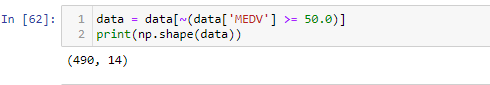


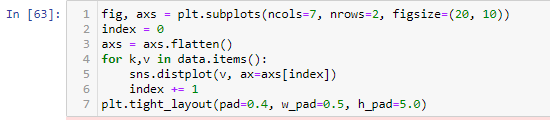


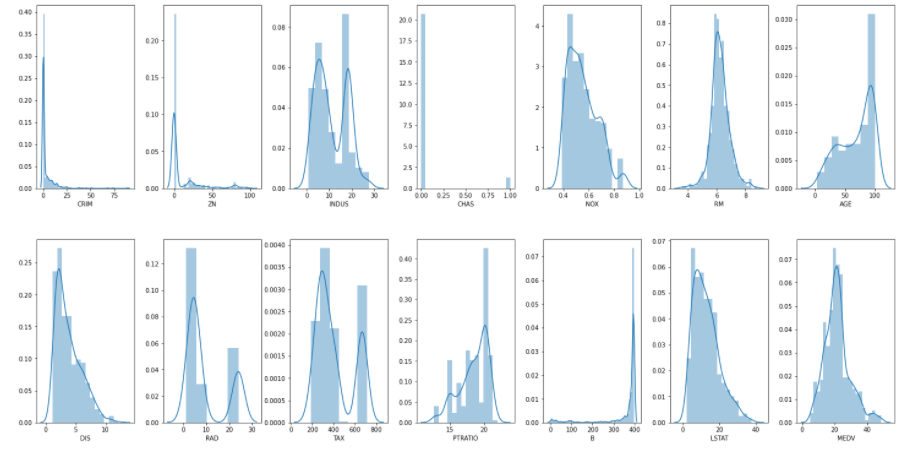


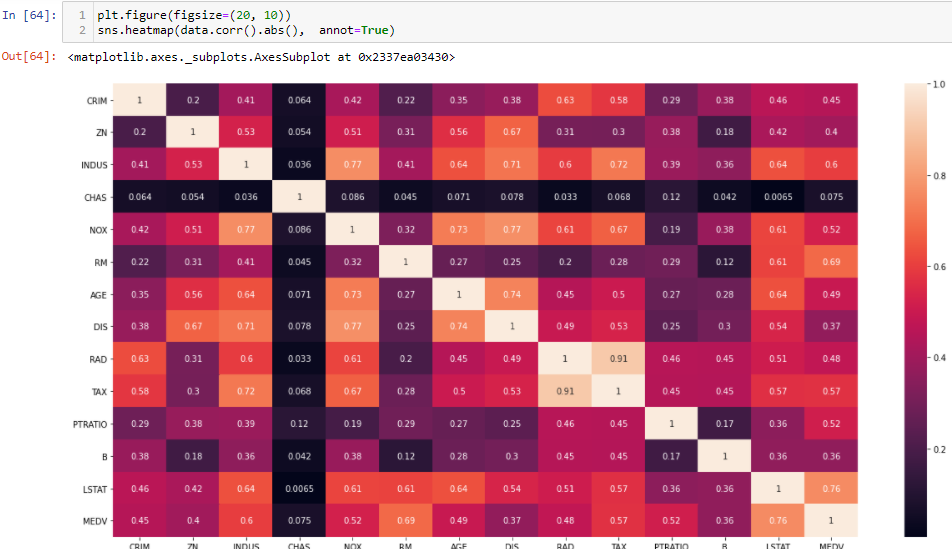




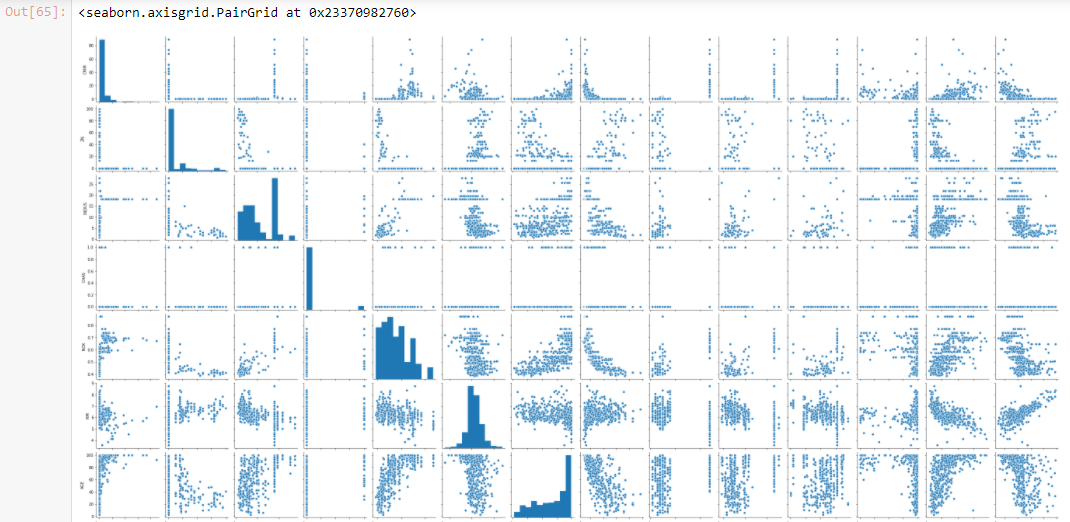


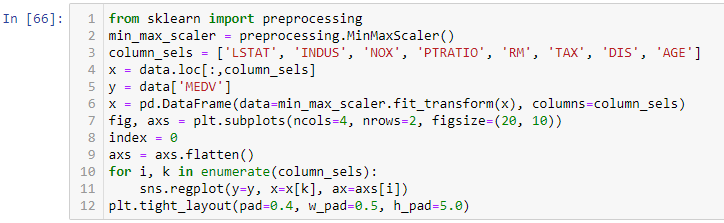


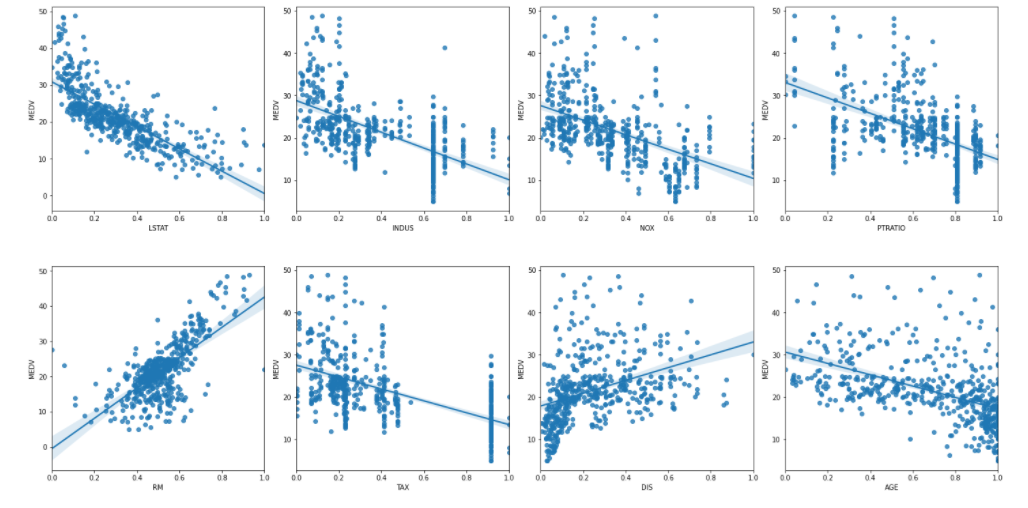


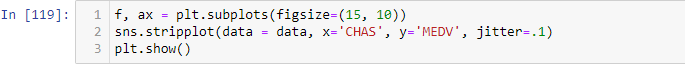


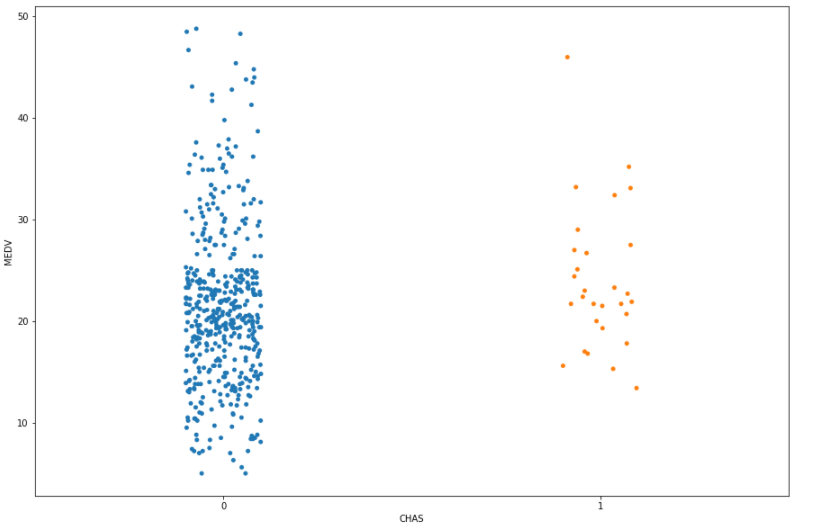




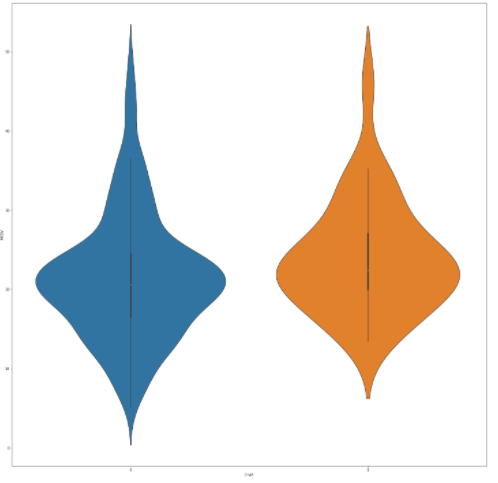


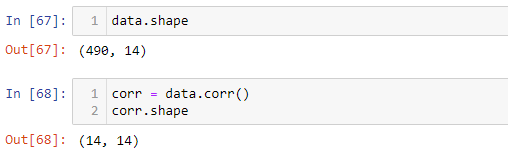


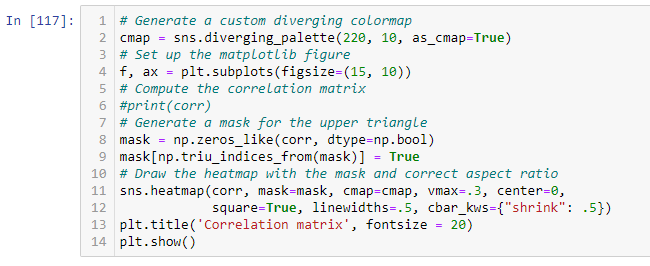


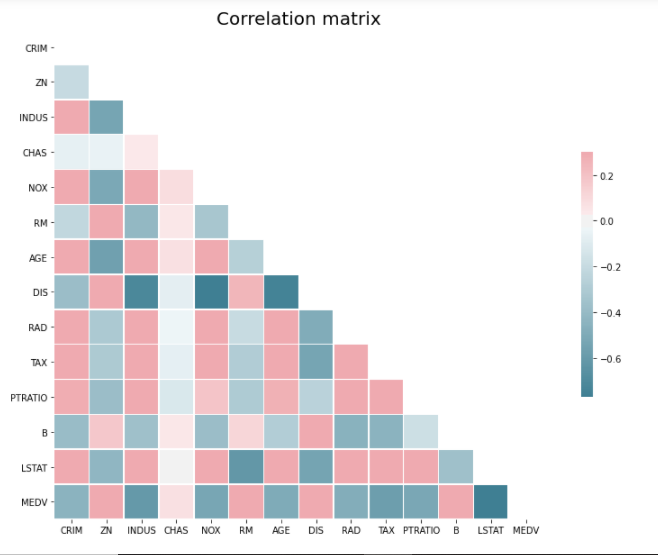


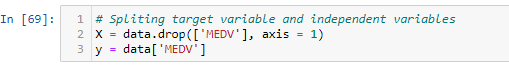


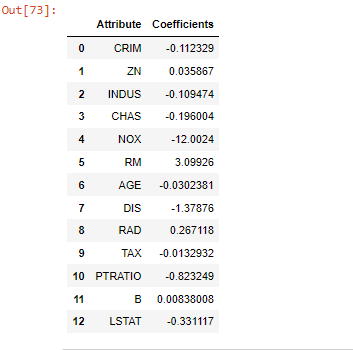




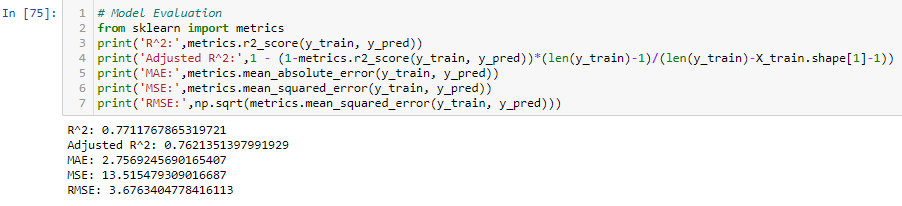


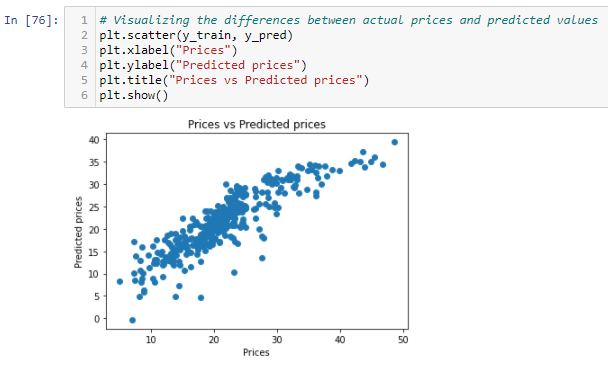


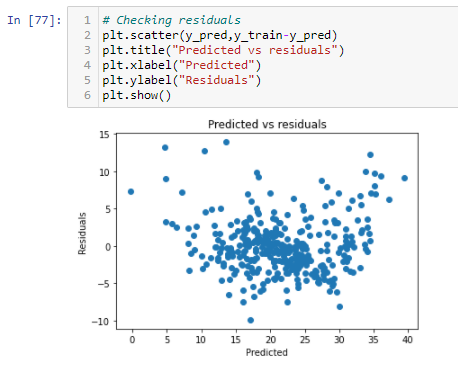


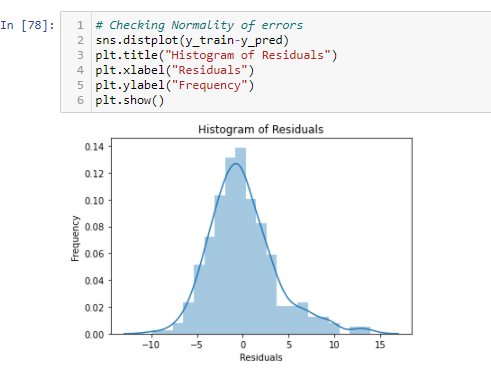




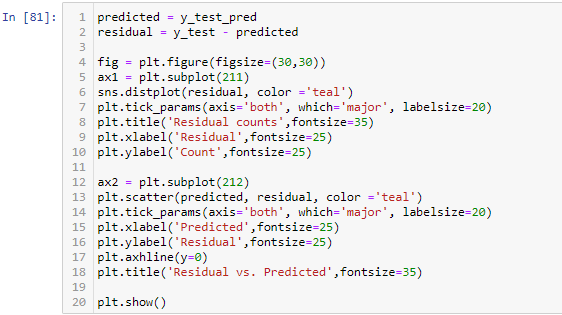


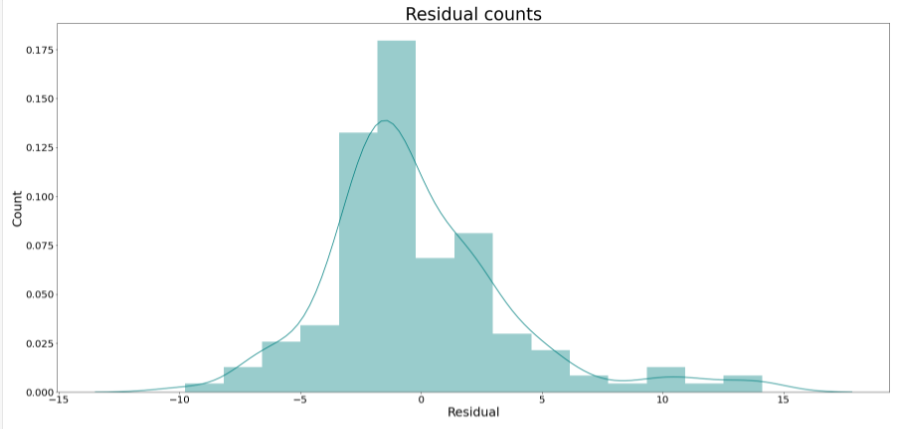


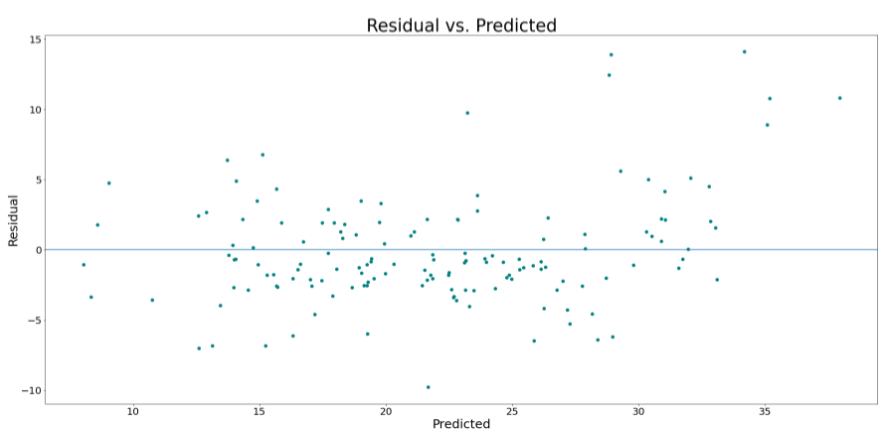




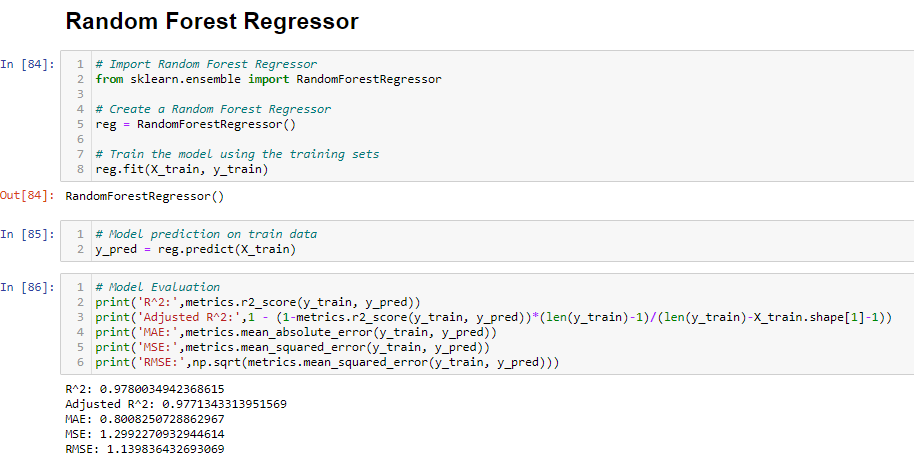


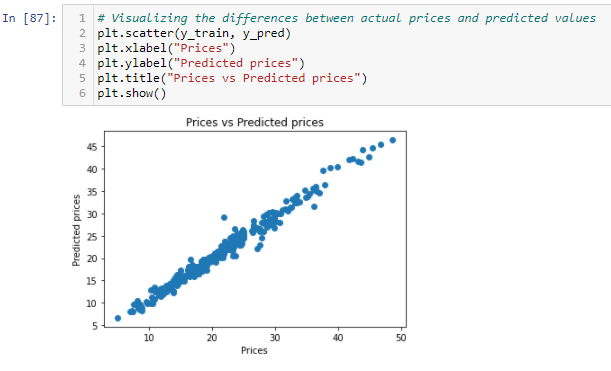


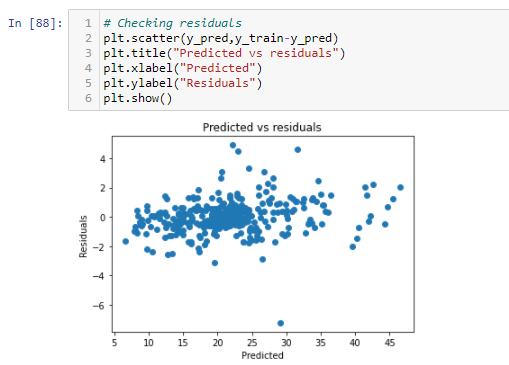


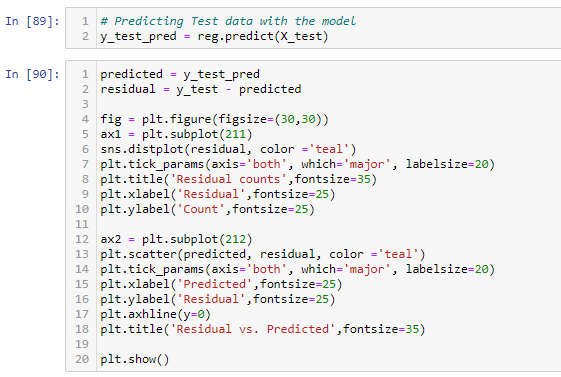


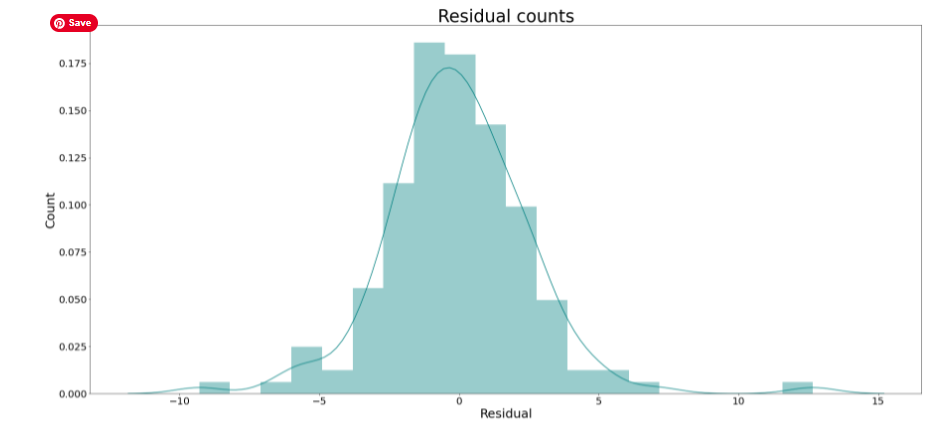


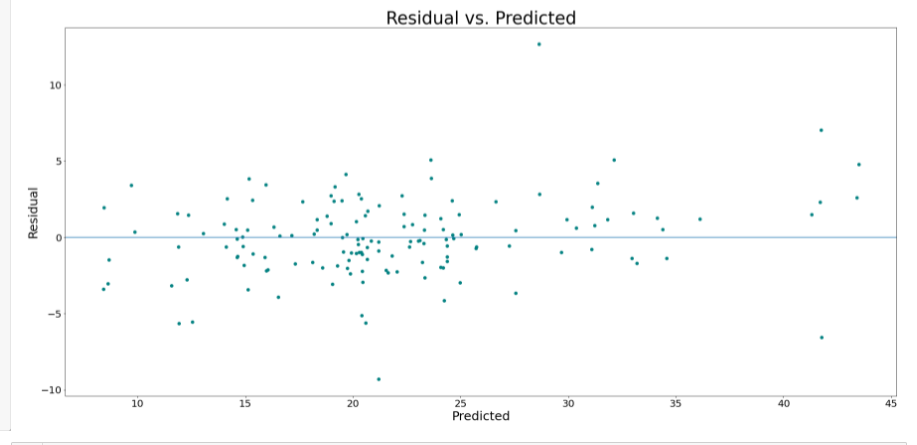




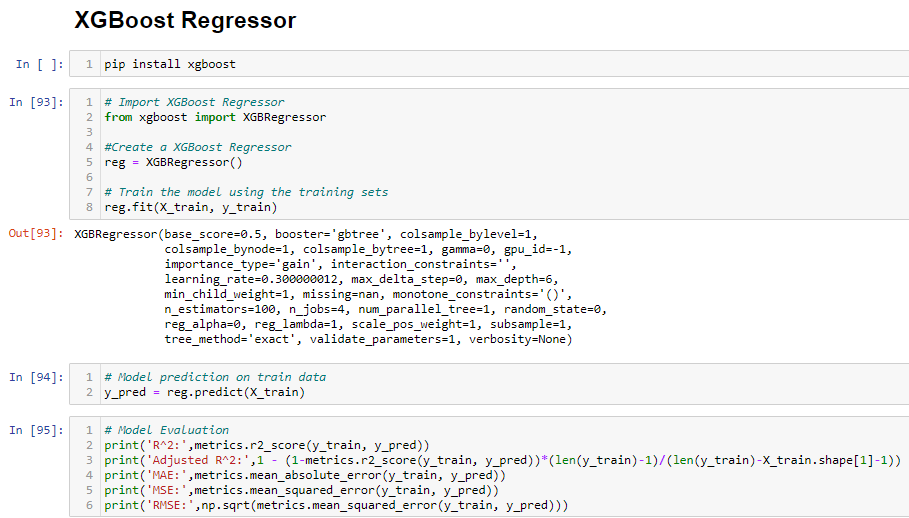


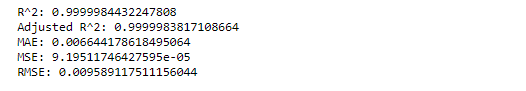


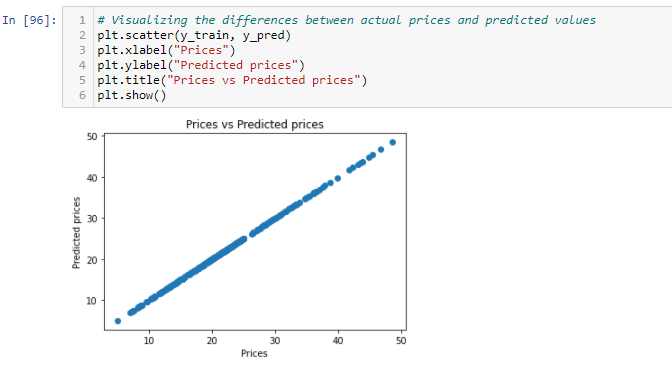


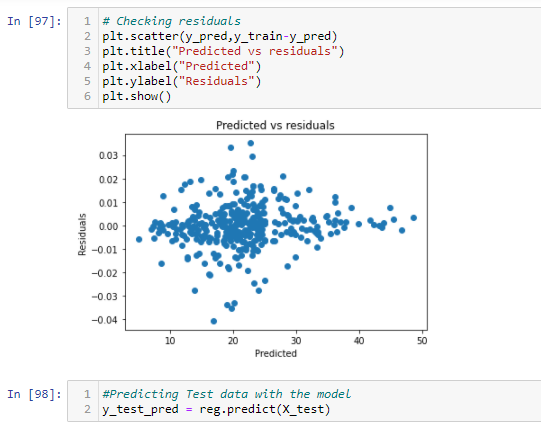


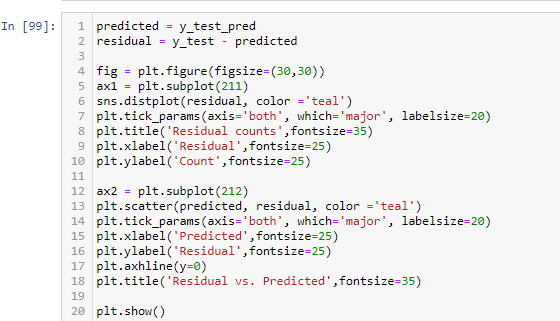


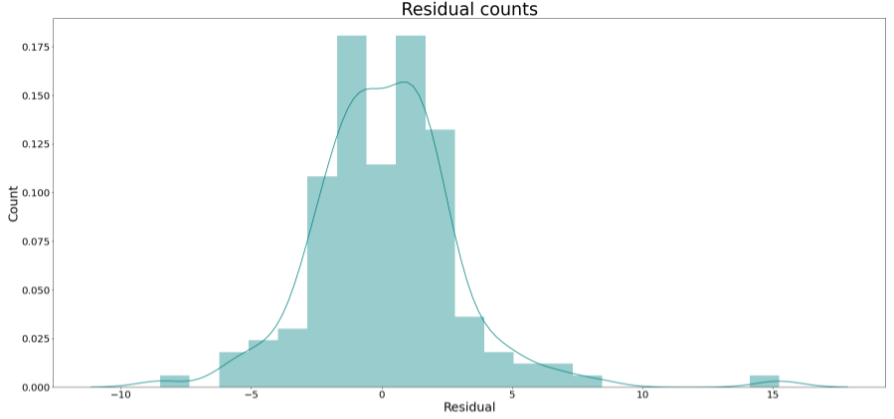


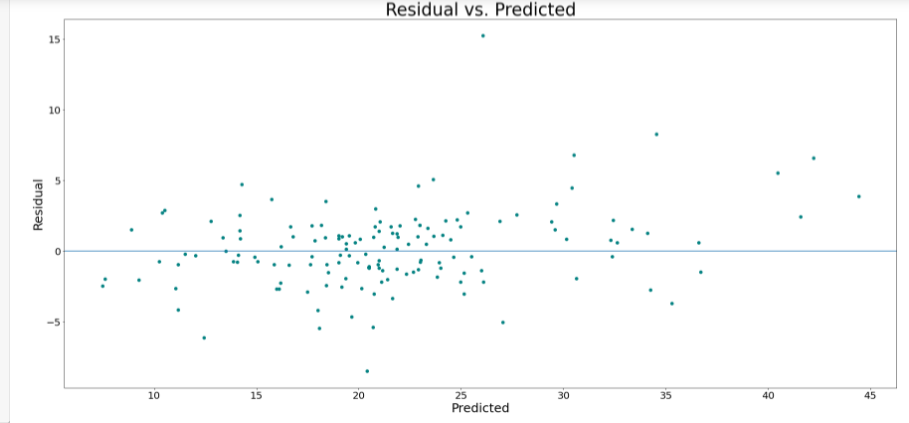


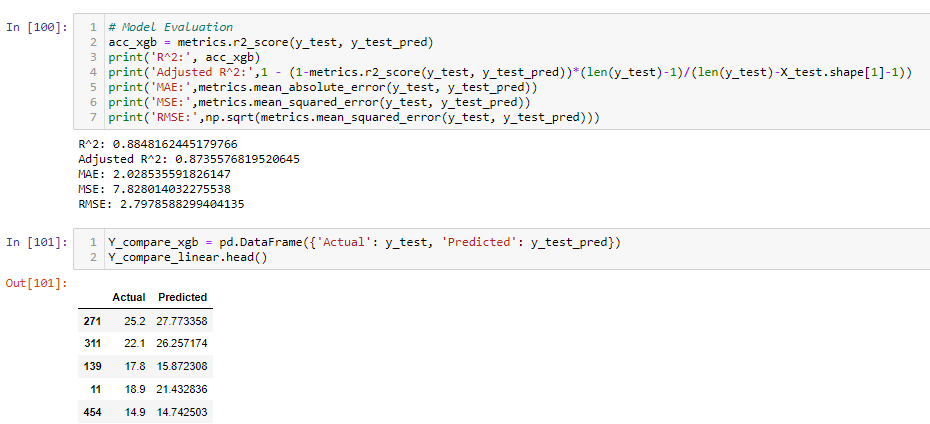


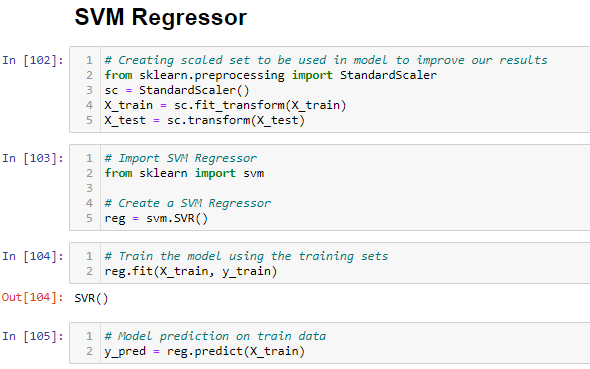


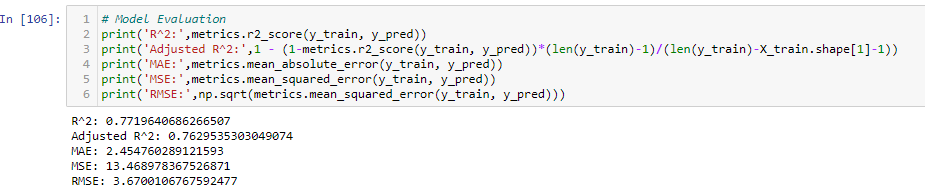


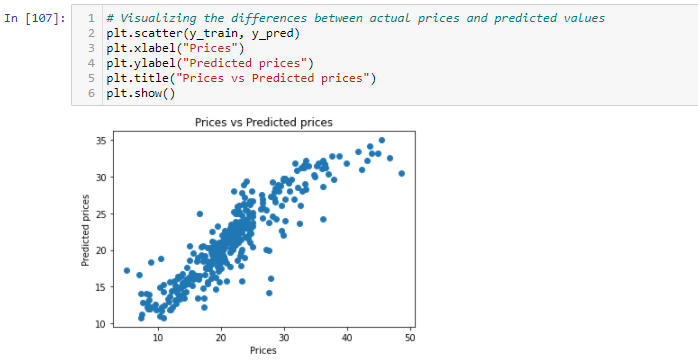


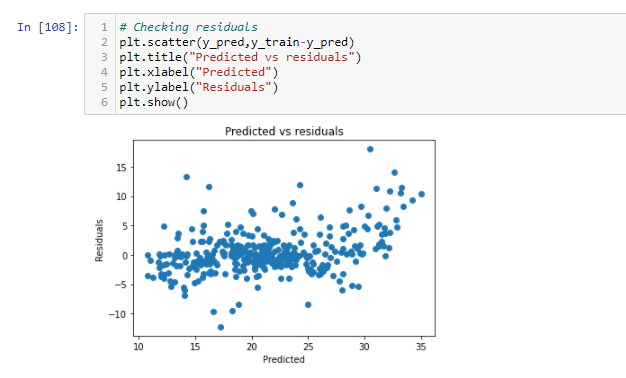


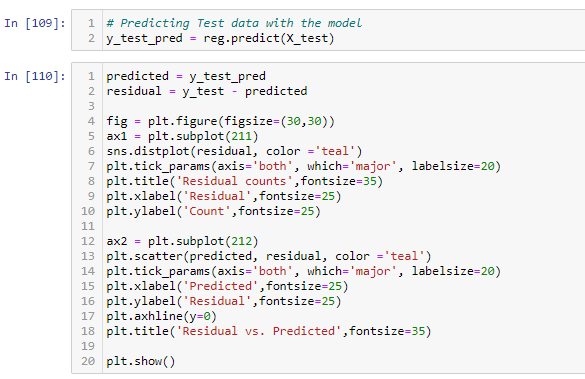


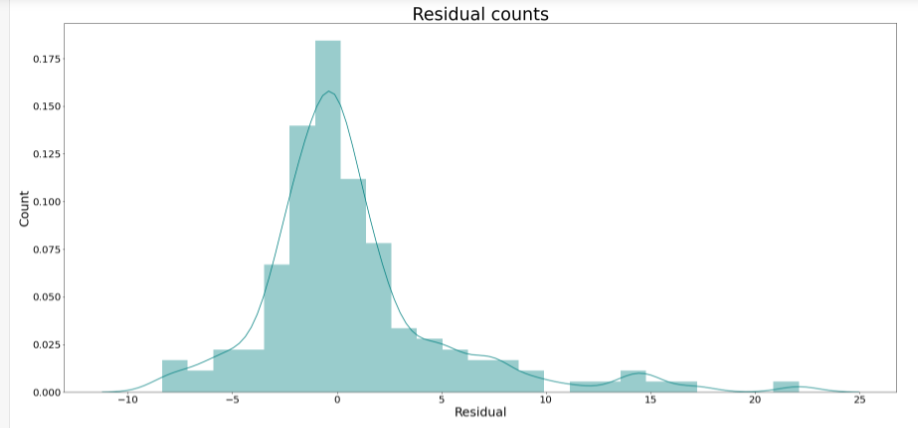


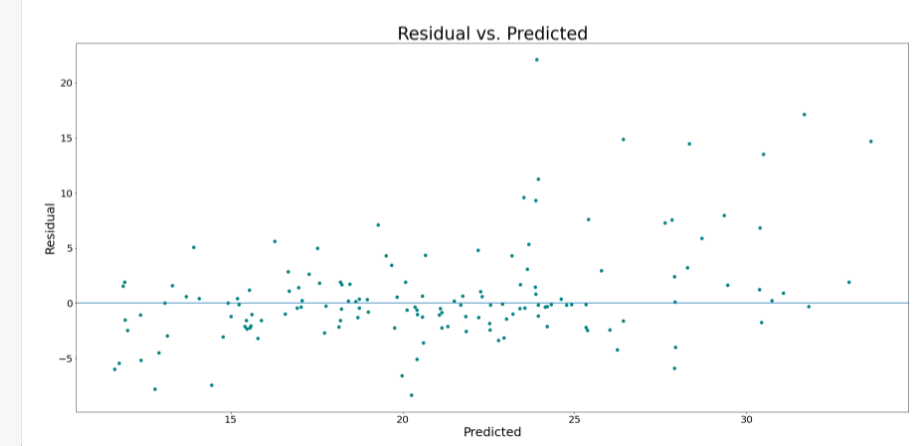


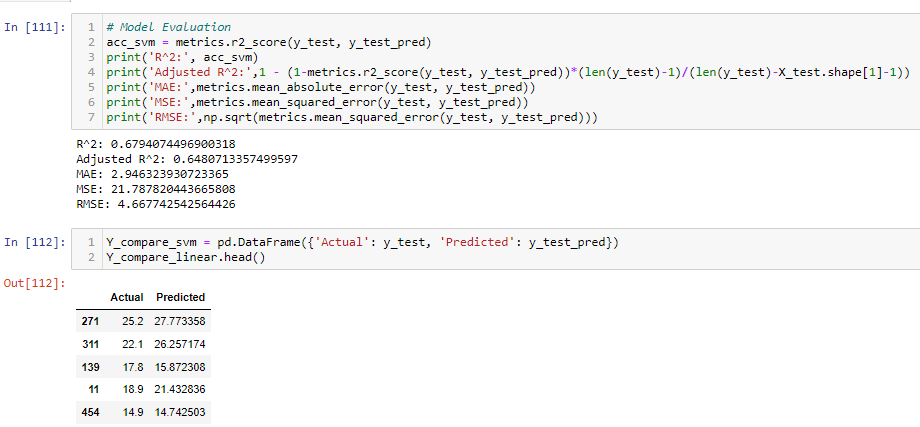






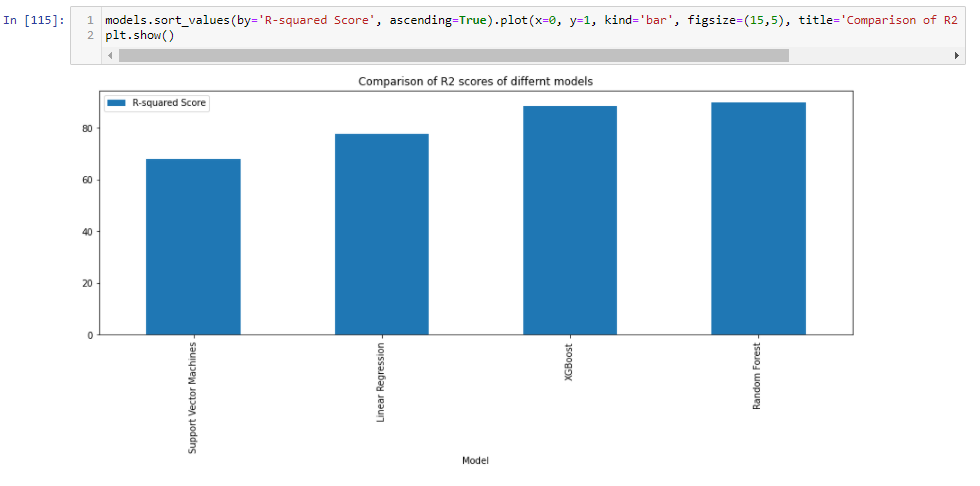






**Comparison of models and their accuracies:**





**Comparison of accuracies:**

|  |  |
| --- | --- |
| **Model** | **Accuracy(%)** |
| Random Forest | 89.96 |
| XGBoost | 88.48 |
| Linear Regression | 77.60 |
| Support Vector Machines | 67.94 |

**Conclusion and future work:**

In my solution I have tried to predict the prices of houses using various parameters. Moreover, the predictions are done not only using a single model. I’ve applied four of the very famous machine learning models to do the prediction, namely, linear regression, random forest, support vector machine and XGBoost. Then, I have compared their accuracies and finally the best performing model is chosen. Since, I also wanted to analyse and visualize the dataset as well as model performances, therefore a lot of visualization techniques have been applied to the dataset, so that the interpretation becomes very easy. In future I’m planning to build a proper interface for this model in the form of a website or an android app so that it can be used by anyone, anywhere and anytime.

**References:**

[1] <https://www.geeksforgeeks.org/ml-linear-regression>

[2] <https://www.geeksforgeeks.org/introduction-to-support-vector-machines-svm/>

[3] <https://towardsdatascience.com/understanding-random-forest-58381e0602d2>

[4] [https://www.analyticsvidhya.com/blog/2018/09/an-end-to-end-guide-to-understand-the- math-behind-xgboost/](https://www.analyticsvidhya.com/blog/2018/09/an-end-to-end-guide-to-understand-the-%20%20%20math-behind-xgboost/)

[5] <https://towardsdatascience.com/data-visualization-for-machine-learning-and-data-science-a45178970be7>

[6] <http://lib.stat.cmu.edu/datasets/boston>