**Minor Project Report**

**Introduction:**

Boston House Prices detection.

This report is based on EDA, Pre-processing and model optimization of the given dataset by running the files using jupyter notebook installed with the essential libraries.

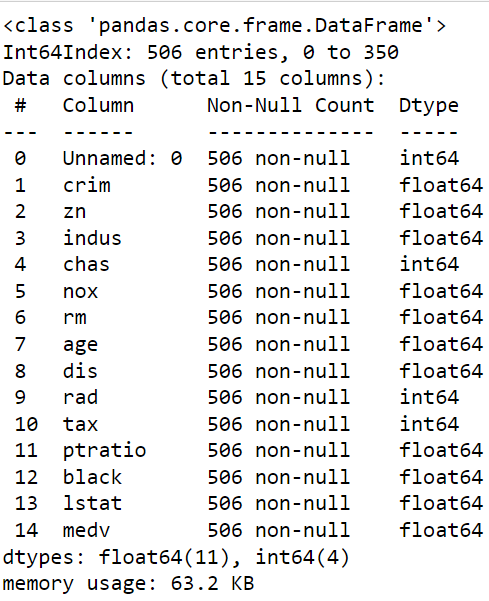
This report has been prepared by:

Name: Divyesh Cirikonda

E-mail: ciri.div@gmail.com

**Data Description:**

Given both train data set and test dataset have been combined. The complete dataset contained 506 indexes and 15 columns. Data has then been divided into X, Y where X contains all the features columns and Y contained the label column. Below is the picture of the complete information about the combined dataset. From the information tab we can also see that there are no null values.

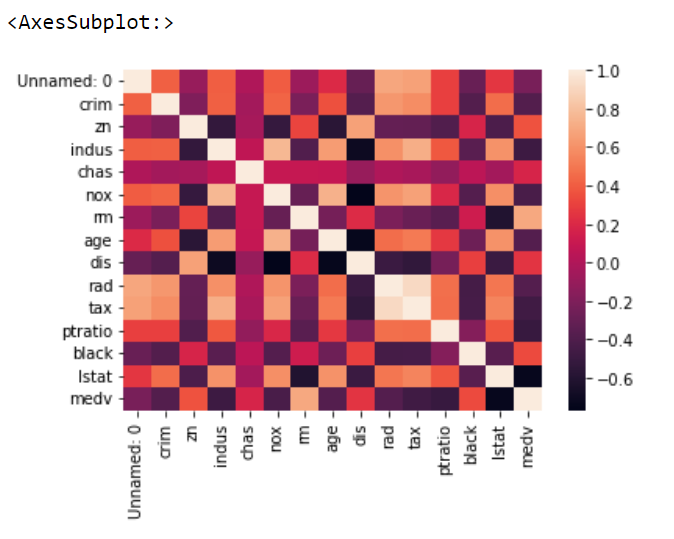


**Approach:**

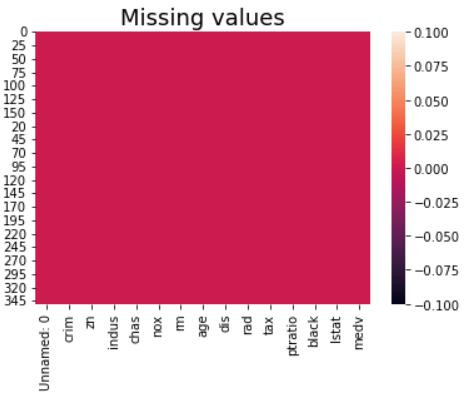
Looking at the correlation matrix of the dataset, I found out the major feature which can be used for regression model. Based on that I have separated the train and test data and then finally applied models and optimized them to get the maximum r2 score.

**Visualization:**

Below is the heat map of correlation matrix of the combined dataset.



We can also look at the graph for null values. From the dataset there are no null values.

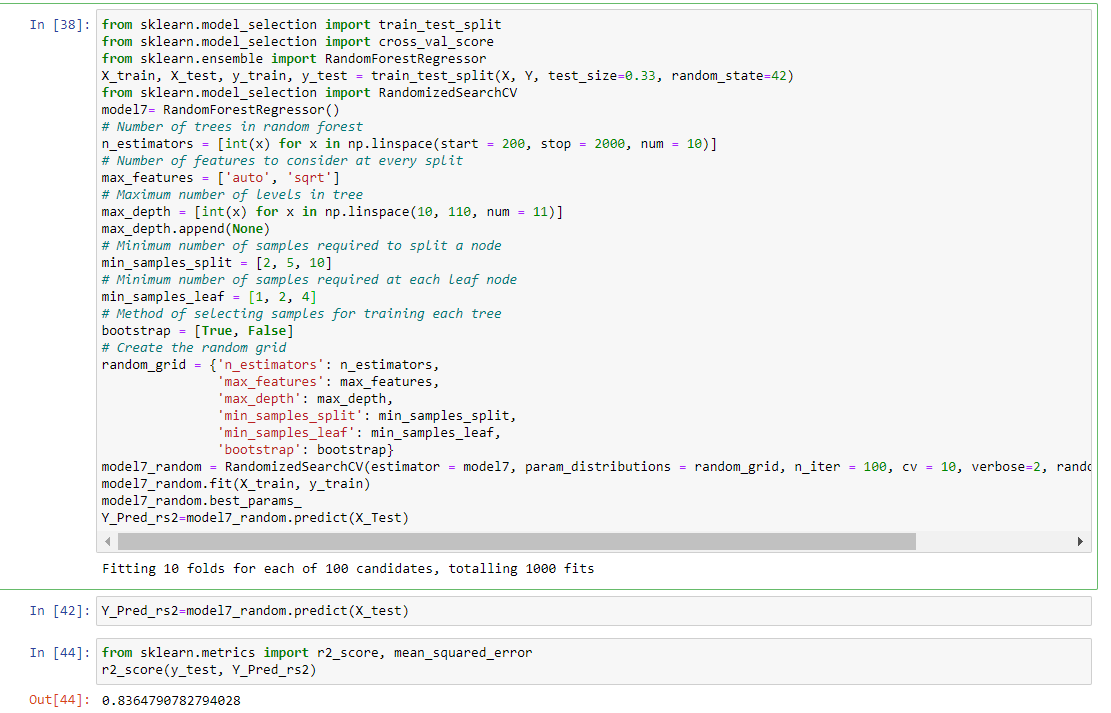


**Algorithms:**

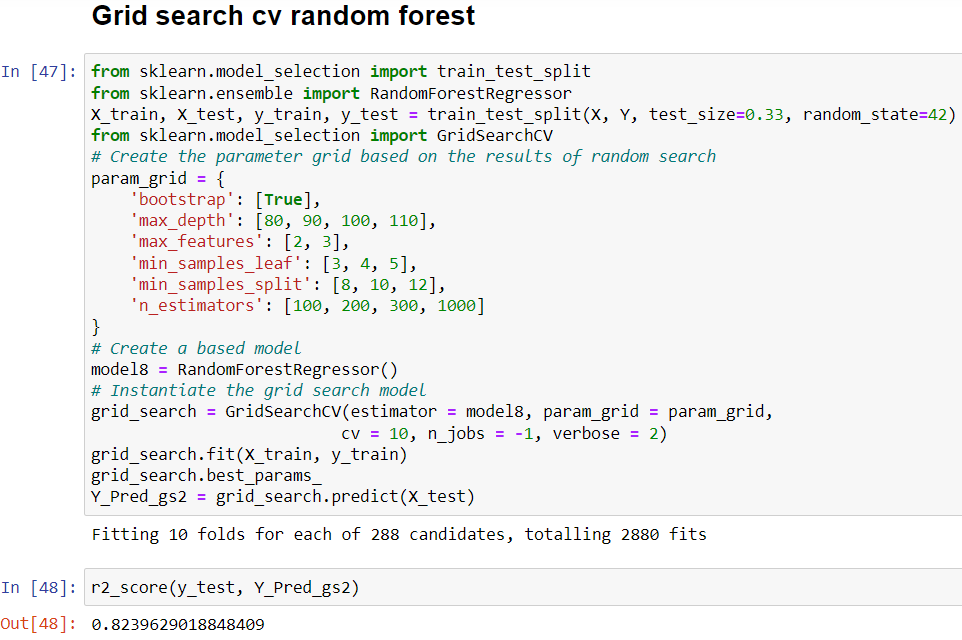
Implemented RandomForestRegressor and modified/optimized the model using GridSearchCV as well as RandomizedSearchCV.

Below are the codes and also the results obtained for both the models.

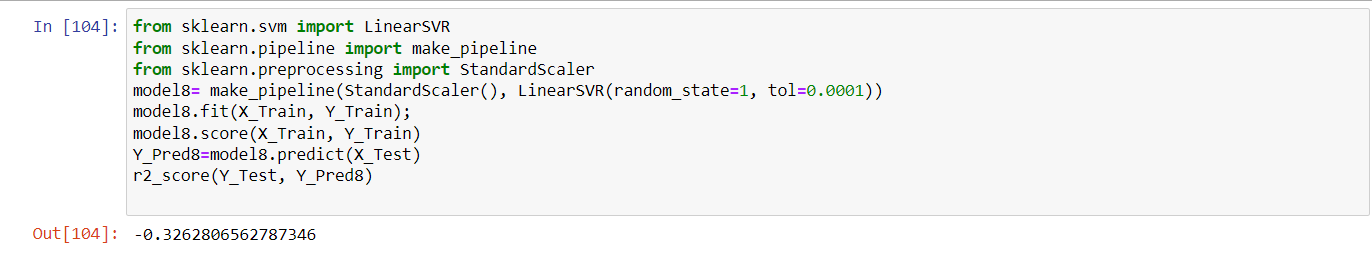
1. Using RandomizedSearchCV



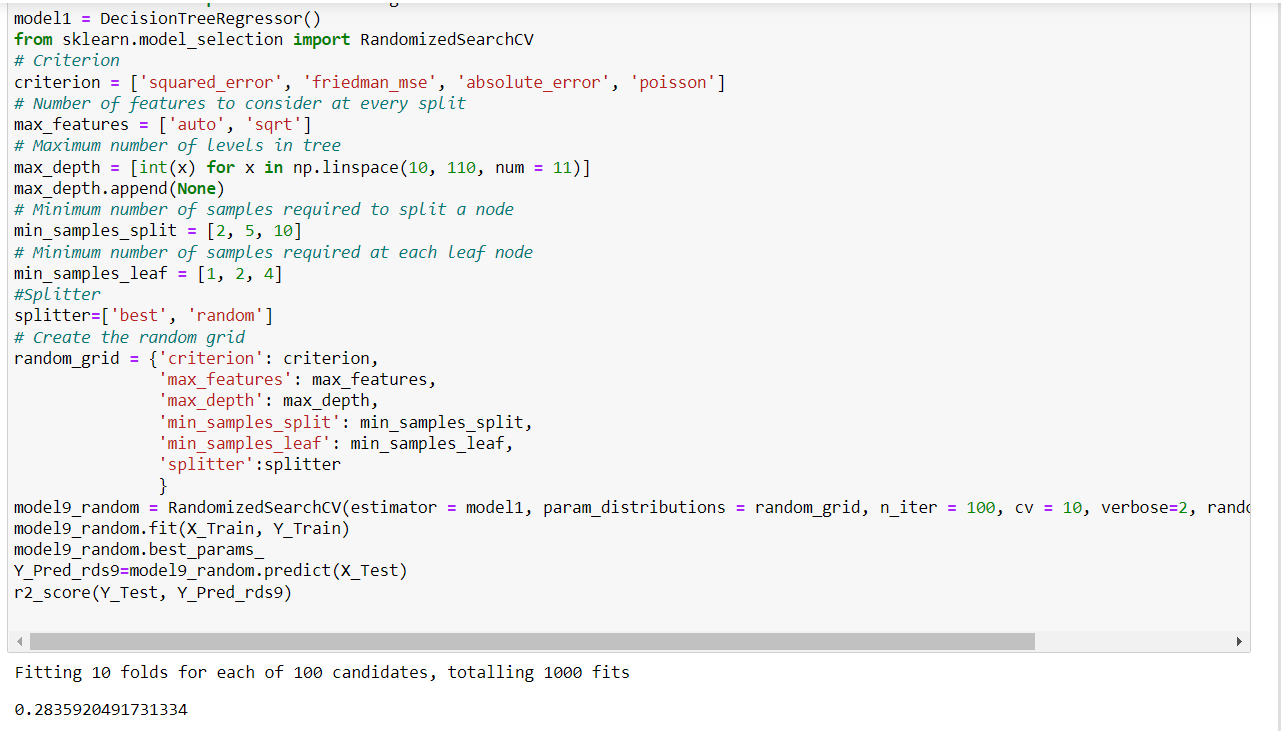
1. Using GridSearchCV



1. SVM



1. Decision Tree



**Result and discussion:**

For combined data I used SVM, Decision Tree, Random Forest Regression models for calculating the r2 score of the model. From the above figures we can see that the r2 score of Randon Forest Regressor optimized using Randomized Search CV gave the best r2 score from all the models tried.

**Conclusion:**

From this project we were able to predict the house prices in boston where Randon Forest Regressor optimized using Randomized Search CV gave the best r2 score from all the applied models. This project also gives insights about the data parameters being taken under consideration for predictions.

**Future Work**:

More of the models for regression can be tried out and optimized to find the best r2 score from all the models. The model can be deployed into real world by using flash. Data interpretation can be made user-friendly using streamlit application. There is always scope of improvement for any kind of machine learning. With experience on can truly gain a lot of knowledge and built good machine learning models.

**References:**

1. Scikit-learn documentation
2. Anaconda documentation
3. <https://stackoverflow.com/>