PROG8420 – Programming for Big Data

Project

Predict House Prices

Himali Hemantkumar Thaker (8638034)

Shubhangini Manoharsinh Zala(8679920)

Professor Norbert Mika

Due Date: April 20th, 2020

Background:

The purpose of the project is to predict the price of the house by applying liner regression on the dataset and try to get good accuracy. The threshold of accuracy is up to 85%, but for this dataset we tried to get 91% approx accuracy. The dataset that we are using for predicting the house price is kc_house_data.csv in csv format, where this file includes all the details that are essential for predicting the price such as id, price, number of bedroom, number of bathroom, sqft_living, sqft_lot, sqft_basement, date, floor, waterfront, view, year, sqft_living15, sqft_lot15 and many other. Along with that visual.py file is used which has function of model learning and for getting decision tree.

Explanation:

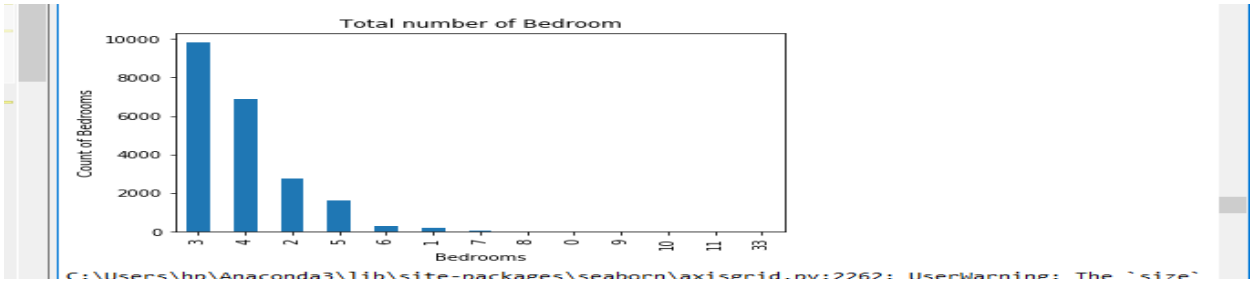
The important packages are imported and csv is read. Here the important factor is price which is to be predicted based on rest all other factors. So dataset is created according to that.

```
 _cost = _housedata['price']

_mainfeatures = _housedata.drop('price', axis = 1)

_details=_housedata.describe()

print("All the Data: ",_details)
```

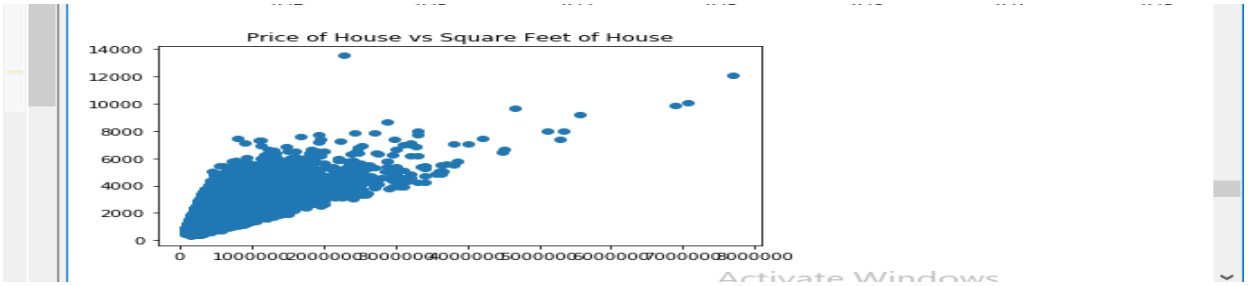
After getting total count, min, max and all details it was figured out that there are 33 max rooms with 13540 sqrf and many more information.

It can be initiated by considering how many number of bedrooms by bought by the people and by plotting graph it showed house with 3 bedrooms is preferred most by the people.

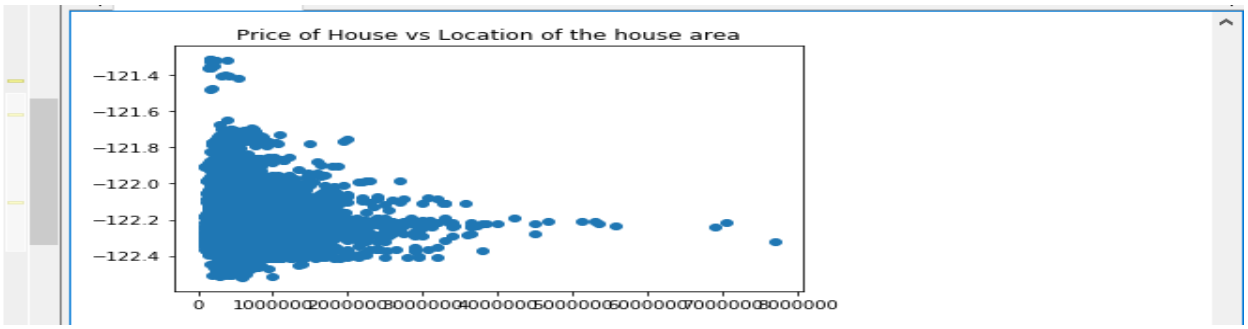
```
_housedata['bedrooms'].value_counts().plot(kind='bar')

plt.title('Total number of Bedroom')

plt.xlabel('Bedrooms')

plt.ylabel('Count of Bedrooms')

plt.show()
```



After that if we consider other common factor that might affect while predicting the price can be square feet, location, square feet living, basement



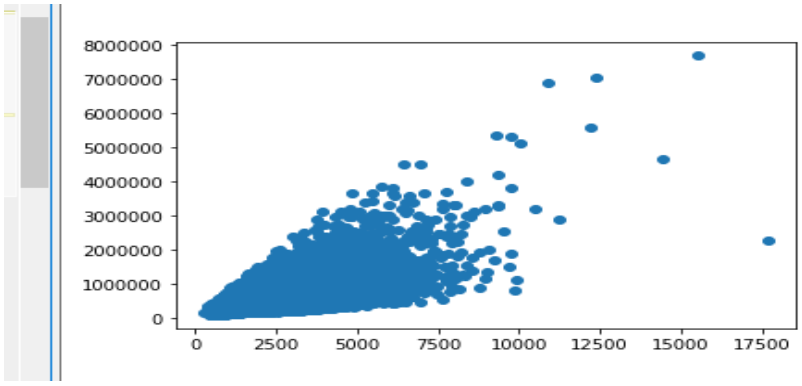
From the above figure it can be seen that more the living area the price will also be high and it is clearly seen that they are linearly distributed.



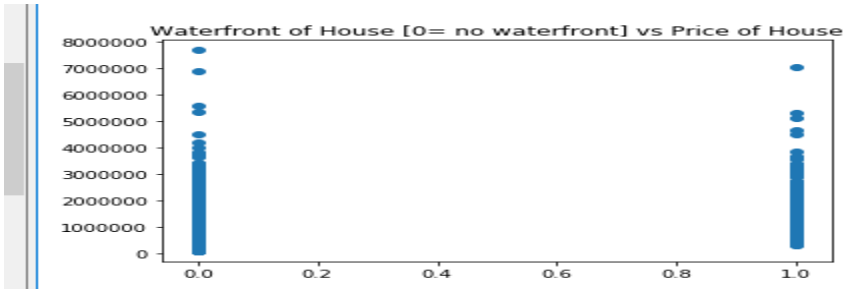
From the above figure it can be stated that at -122.2 to -122.4 longitudes houses are sold high rate.

Other factors that can affect are shown below:

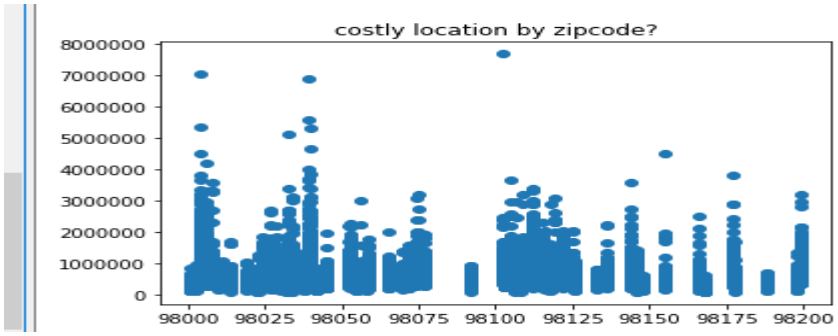
Square feet living and basement:



House with waterfront, here waterfront =0 means there is no waterfront in house vice versa:



Costly location by zipcode:

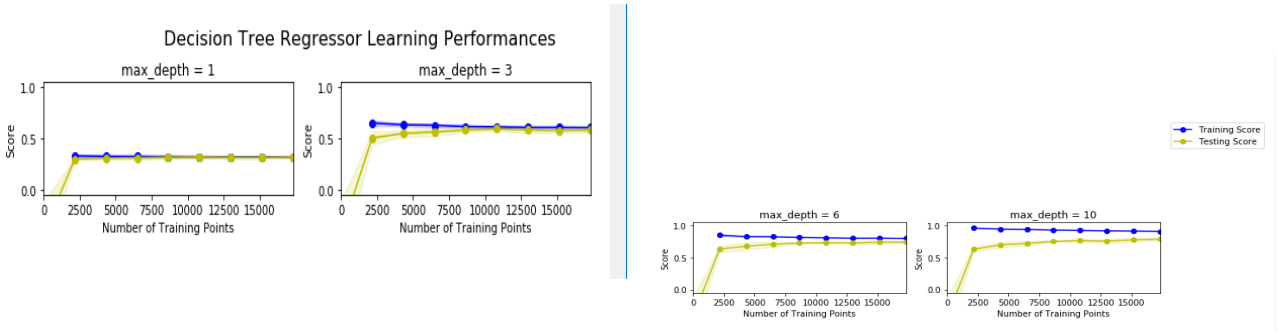


Linear Regression:

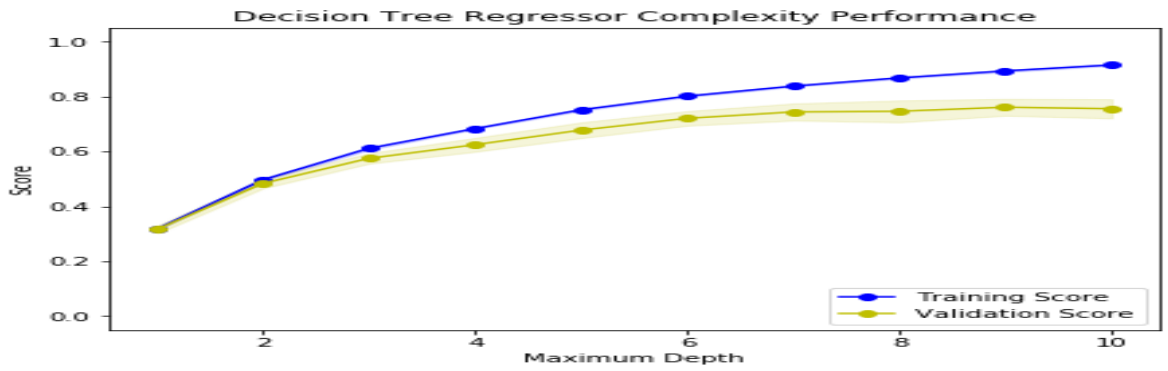
For liner regression the data in divided in to two sets i.e training data and testing data. For linear regression sklearn is used and liner regression is imported. Predicting the price is main purpose so linlable is set for price and dates are converted to 0 and 1 so it will not create any issues in predicting. Further splitting is performed on data resulting in to train data (90%) and test data (total 10%). As a result, after considering all these data liner regression model can be performed, and accuracy achieved is 73%, which is low compared to threshold accuracy (85%). So we made use of Gradient Boosting Regression. The required package is imported, variables

is created that defines GBR and training and testing data are set as their parameter and accuracy achieved is 91%, which is more than the threshold accuracy.

Decision Tree: By making use of visual.py file decision tree is created.



From the figure it can be stated that as training points increases score decreases. As a result, when columns are in large number model will get more data by which prediction can be done properly.



References:

Roman, V. (2019, March 10). Machine Learning Project: Predicting Boston House Prices with Regression. Retrieved from https://towardsdatascience.com/machine-learning-project-predicting-boston-house-prices-with-regression-b4e47493633d

Raghavan, S. (2017, June 20). Create a model to predict house prices using Python. Retrieved from https://towardsdatascience.com/create-a-model-to-predict-house-prices-using-python-d34fe8fad88f