A

Capstone Project

in

Data Science

HOUSE PRICE PREDICTION

By:
Ganesh Bhandari

@

January 2021

1. Problem, background and data:

A realtor of a Real Estate Company, in King County, Washington, is having trouble in deciding suitable prices of houses in the county. The realtor would really appreciate it if someone could predict house price for them so that he can put suitable price for the houses in the sale.

I would like to help the realtor to price the houses. The realtor gave me a data set of houses sold between May 2014 and May 2015. This data set contains 21613 observations with 21 features. The description of the data is as follows:

- **id:** Unique ID for each home sold
- **date:** Date of the home sold
- **price:** Price of each home sold
- **bedrooms:** Number of bedrooms
- **bathrooms:** Number of bathrooms, where .5 accounts for a room with a toilet but no shower
- **sqft_living:** Square footage of the apartments interior living space
- **sqft_lot:** Square footage of the land space
- **floors:** Number of floors
- **waterfront:** A dummy variable for whether the apartment was overlooking the waterfront or not
- **view:** An index from 0 to 4 of how good the view of the property was
- **condition:** An index from 1 to 5 on the condition of the apartment,
- **grade:** An index from 1 to 13, where 1-3 falls short of building construction and design, 7 has an average level of construction and design, and 11-13 have a high quality level of construction and design.
- **sqft_above:** The square footage of the interior housing space that is above ground level
- **sqft_basement:** The square footage of the interior housing space that is below ground level
- **yr_built:** The year the house was initially built
- **yr_renovated:** The year of the house's last renovation
- **zipcode:** What zipcode area the house is in
- **lat:** Lattitude
- **long:** Longitude
- **sqft_living15:** The square footage of interior housing living space for the nearest 15 neighbors
- **sqft_lot15:** The square footage of the land lots of the nearest 15 neighbors

2. Methods used:

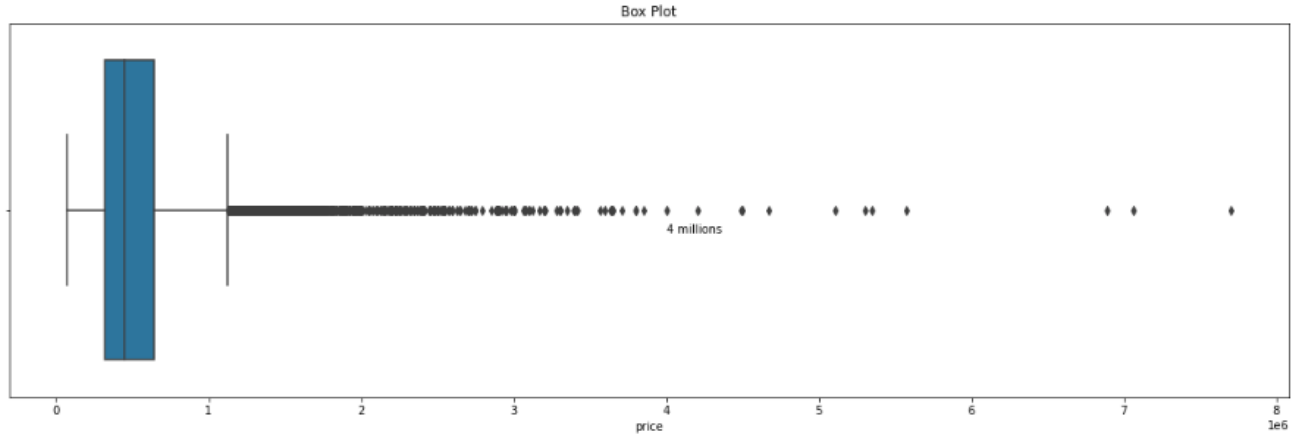
This problem is a regression problem. So to solve this problem I have created a model based on regression algorithms . In this project, I used regression algorithms like Multiple Linear Regression, Ridge Regression, Polynomial Regression, Random Forest Regressor, Extra Trees Regressor, XGBRegressor and chose the best model based on their accuracy performance. I used the best selected model to predict the house prices. The steps that I used to get the result are as follows:

- Loading data
- Cleaning and organizing data
- Performing Exploratory Data Analysis (EDA) to understand the data and important features.

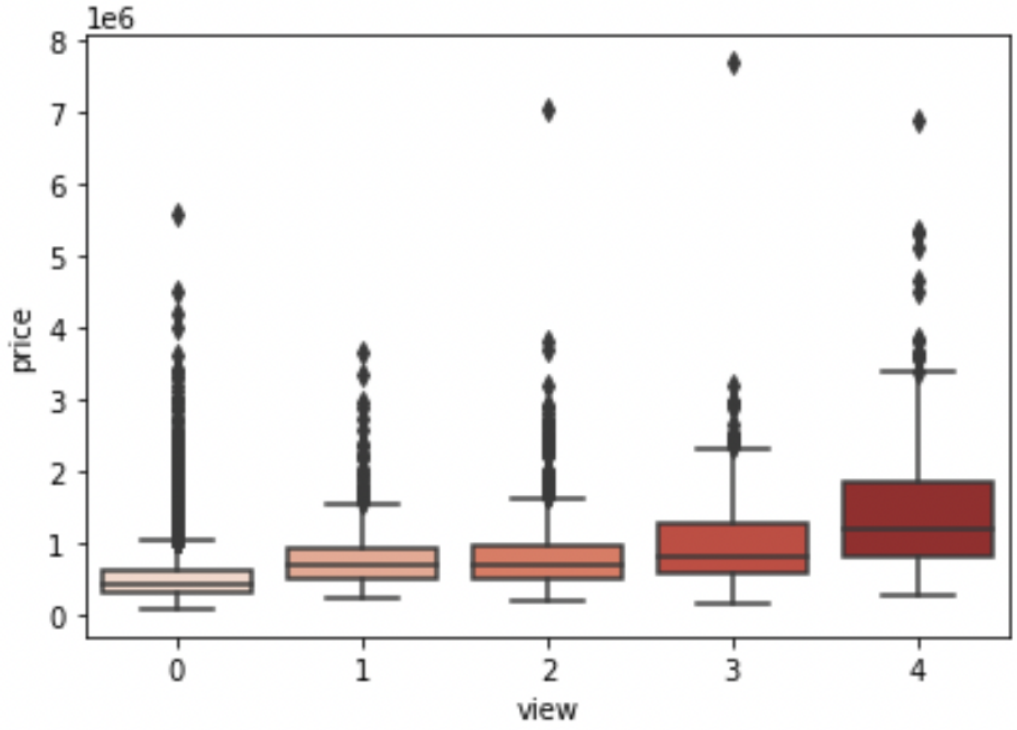
- Creating model and training the model
- Using suitable metric to check the performance of the model
- Deploying the model to get the results

3. Current Situation and Findings:

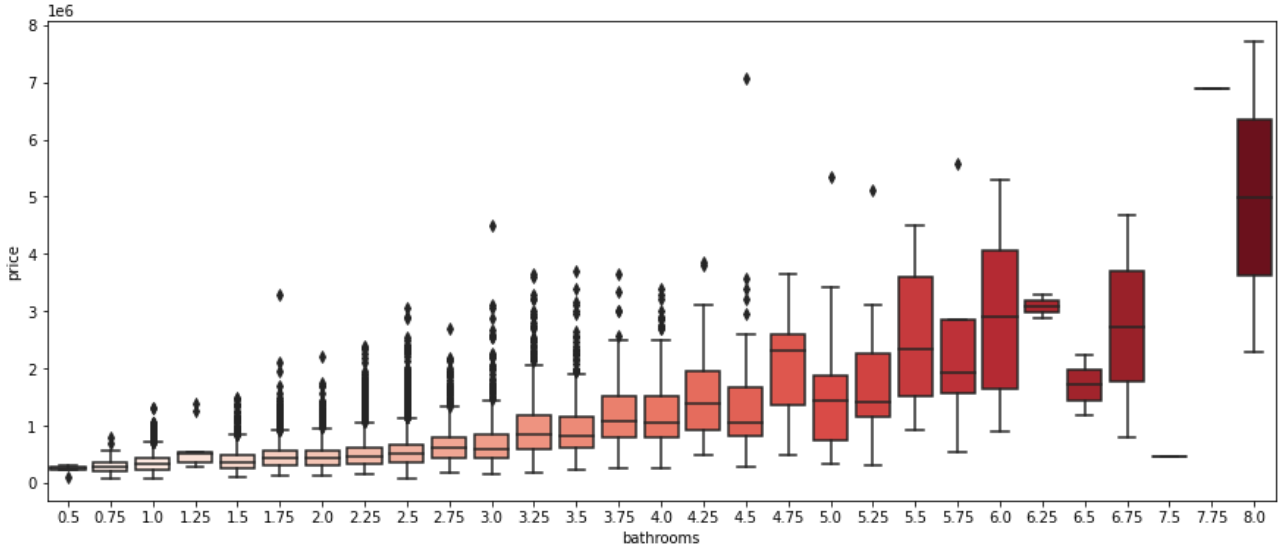
There are houses having prices more than 4 millions. The distribution of the price in the data is shown below.



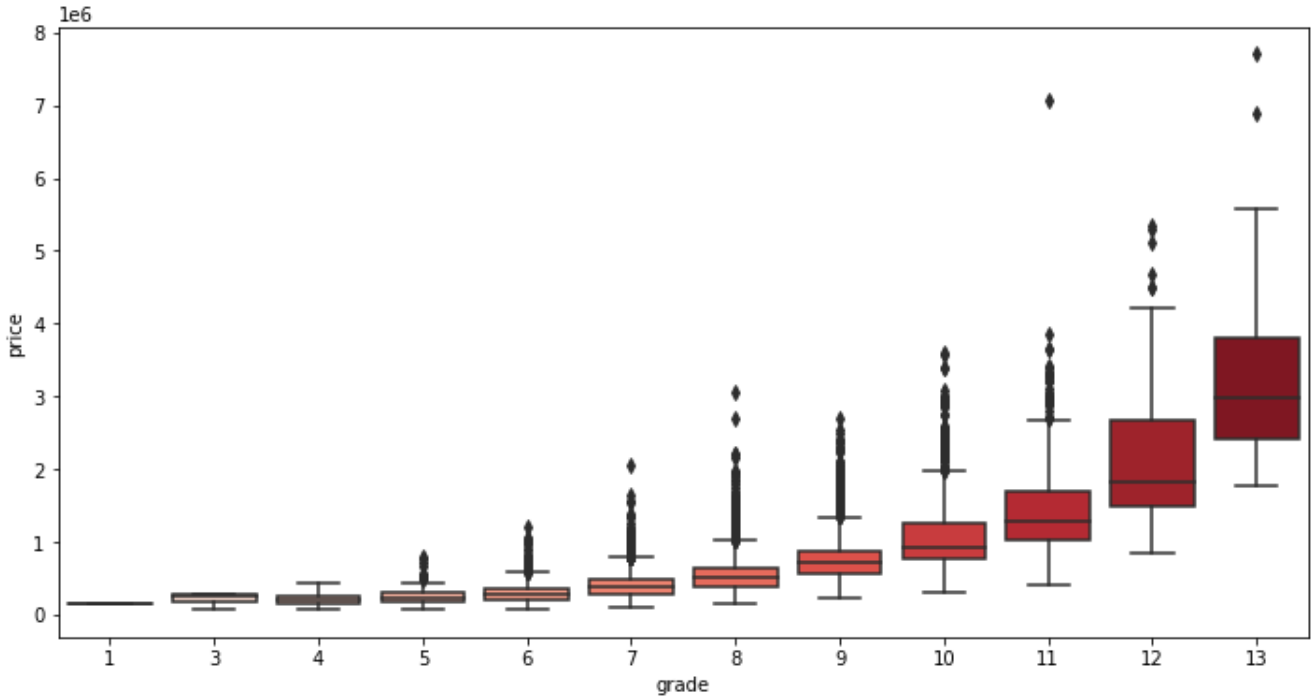
People would love to pay more for the house with better views.



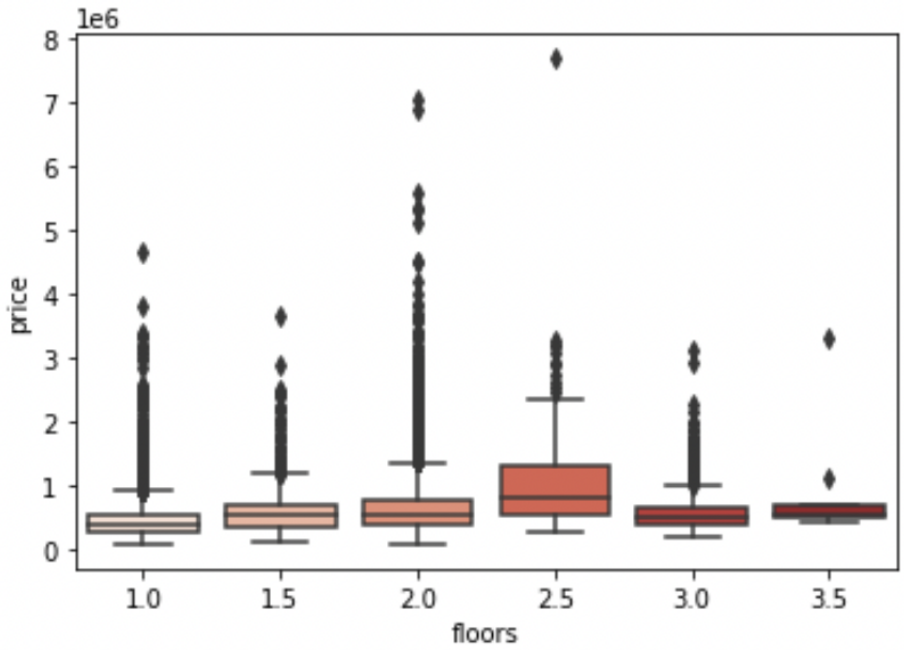
The data shows that the price of the houses increase as the number of bathrooms increases.



Also the price of the houses increases as grade increases.



House price increases as number of floors increases up to 2.5.



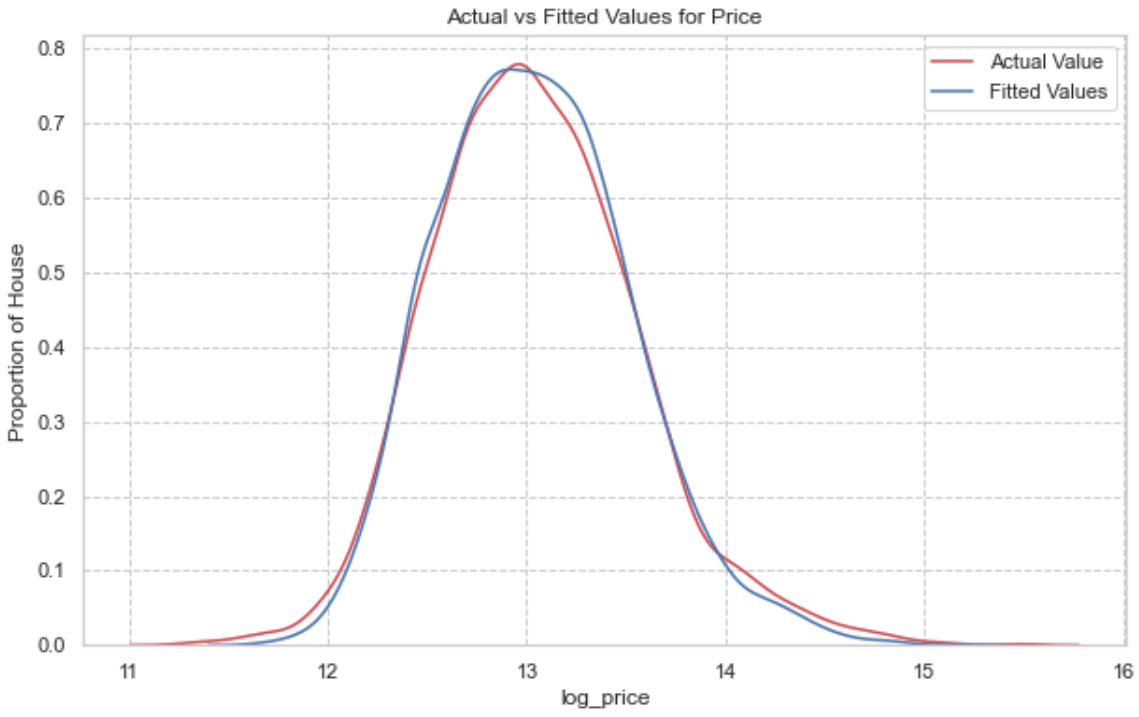
The distribution of the price over time is shown below. This graph shows that the average price of houses remains high during the summer season.

4. Model:

In this project, I used regression algorithms like Multiple Linear Regression, Ridge Regression, Polynomial Regression, Random Forest Regressor, Extra Trees Regressor, XGBRegressor to create a suitable model. The performance report of these model on the data is as follows:

Algorithm	R-Squared	MSE
Multiple Linear Regression	0.76	0.25
Pipe Line with Linear Regression	0.81	0.23
Ridge Regression	0.76	0.25
Polynomial Regression of degree 2	0.79	0.24
Polynomial Regression of degree 3	0.78	0.24
Random Forest Regressor	0.86	0.19
Extra Tree Regressor	0.86	0.19
XGB Regressor	0.85	0.20

The report table shows that, among all these algorithm, the Extra Trees Regressor is more consistent and works better than the others. So, I chose the Extra Trees Regressor model as the final model for my project. The fitting of the model is given below:



5. Prediction of House Prices using the model:

The prediction of House Prices using the model for the first 10 houses is:

```
([331048.85882255, 409051.99854484, 541324.75509907, 763926.69341889,
  742667.40941887, 628523.2242349 , 863242.55764499, 504158.53939873,
  886345.19588571, 800063.6028555 ])
```

6. Conclusion:

The Extra Trees Regressor is more consistent and works better than the others. It has 86 % of accuracy, which is very good level of accuracy. One apply the model to predict the price of houses.

7. Acknowledgments:

I am very grateful to my mentor for his valuable suggestions while completing this project. Also I am thankful to the bank authority for providing valuable information.