HOUSING PRICE PREDICTION

MACHINE LEARNING-1

SATISH KUMAR

ROLL NO: E20029

PRAXIS BUSINESS SCHOOL, BANGALORE

1.1. Introduction

Housing is an important aspect of any national economy of a country. It is estimated that housing and other related activities account for 5 – 10 percent of the GDP (Singh, 2013). A number of studies have pointed out that the key determinant for housing prices are income levels, interest rates, supply conditions demographic changes, number and size of the household, maintenance costs, property taxes and speculative pressures (Singh, 2013). The financial reform of 1991 lead to the rise in income levels among the middle and educated class in India, faster urbanization and higher demand houses in urban areas.

1.1.1 The Problem Statement

Buying a house in Indian cities is a tedious and long drawn process. The price of houses vary and depends on a lot of factors such the how big the house is in terms of square feet, which area is it located in, the number of bedrooms in the house etc. It is very difficult for a prospective buyer to gather information regarding house prices based on their budget and requirements. The prospective buyer has to depend on brokers and other local consultancies to scout for houses which are in their budget and fulfils their requirement making it even more cumbersome. This is coupled with other tedious activities such as loan processing, registration etc adding to the difficulty of the buyer.

1.1.2 Objective

The objective of this project is to use Machine Learning Algorithms to most accurately predict house prices based on various requirements of a prospective home buyer such as locality, area of the house, number of bedrooms etc. The models developed in the project can be used by infrastructure consultancy firms to better service their customers and help them to find the right house within their requirement and budget.

1.2 Exploratory Data Analysis and Anomaly Detection

1.2.1 The Dataset

We have used the American Housing Data. The data has been split into training, validation and test data. The training data consists of 9761 rows and 21 columns. The validation data consists of 9635 observations and the test data consists of 2217 observations. The list of variables is given below.

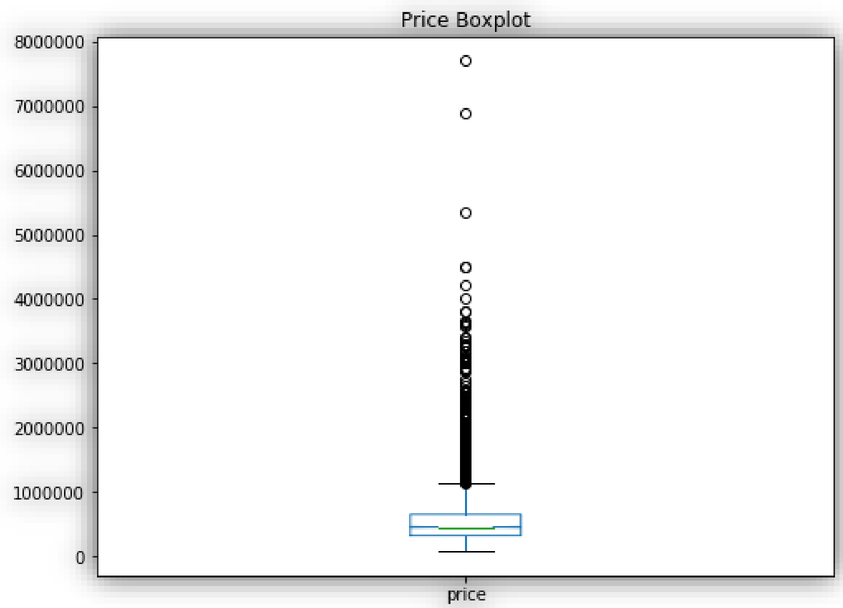
```
Index(['id', 'date', 'price', 'bedrooms', 'bathrooms', 'sqft_living',
       'sqft_lot', 'floors', 'waterfront', 'view', 'condition', 'grade',
       'sqft_above', 'sqft_basement', 'yr_built', 'yr_renovated', 'zipcode',
       'lat', 'long', 'sqft_living15', 'sqft_lot15'],
      dtype='object')
```

The data did not have any null values. There were a number of numeric discrete variables in the data set (Number of bedrooms, Number of bathrooms, Number of floors, view, condition, grade of the house). The variable waterfront was a binary variable which was zero for houses which did not have a waterfront and one for houses which had a waterfront. The sqft_living, sqft_lot, sqft_above, sqft_basement, sqft_living15, sqft_lot15 were numeric variables indicating various measures of area of the house. The Zip code variable contains the Zip code in which the house falls.

1.2.2 Data Analysis

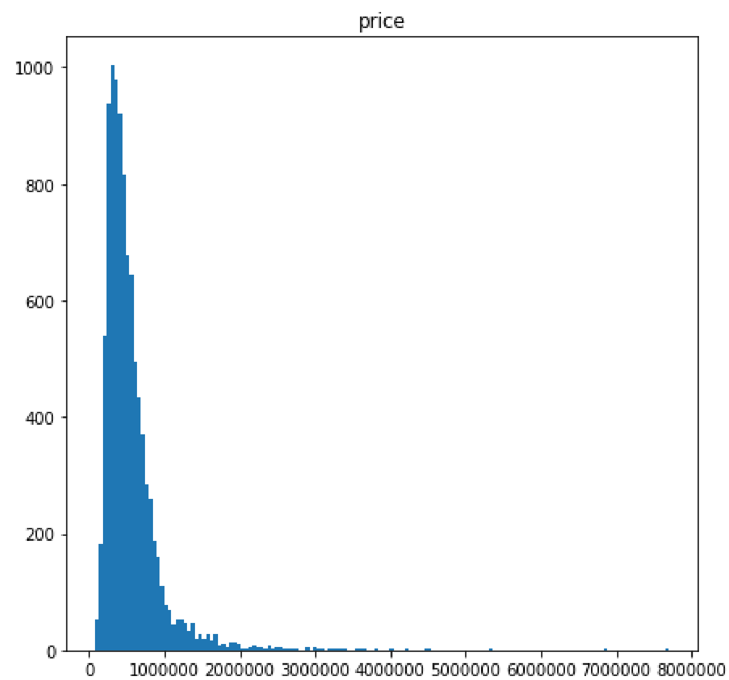
The price variable in the data was highly skewed with the mean price of 5, 42,735 USD. Some of the statistics for the price variable is given below. The minimum price was 80,000 USD. The maximum price was 77, 00,000 USD. Seventy five percent of houses had a price below 6, 49,000 USD. The median price was 4, 50,000 USD. The variable had a lot of outlier values.

```
count        9761.0
mean       542735.0
std        379528.0
min         80000.0
25%        320000.0
50%        450000.0
75%        649000.0
max       7700000.0
Name: price, dtype: float64
```



Price

The box plot and the histogram below shows the distribution of the variable. It is very evidently right skewed and contains a number of outliers



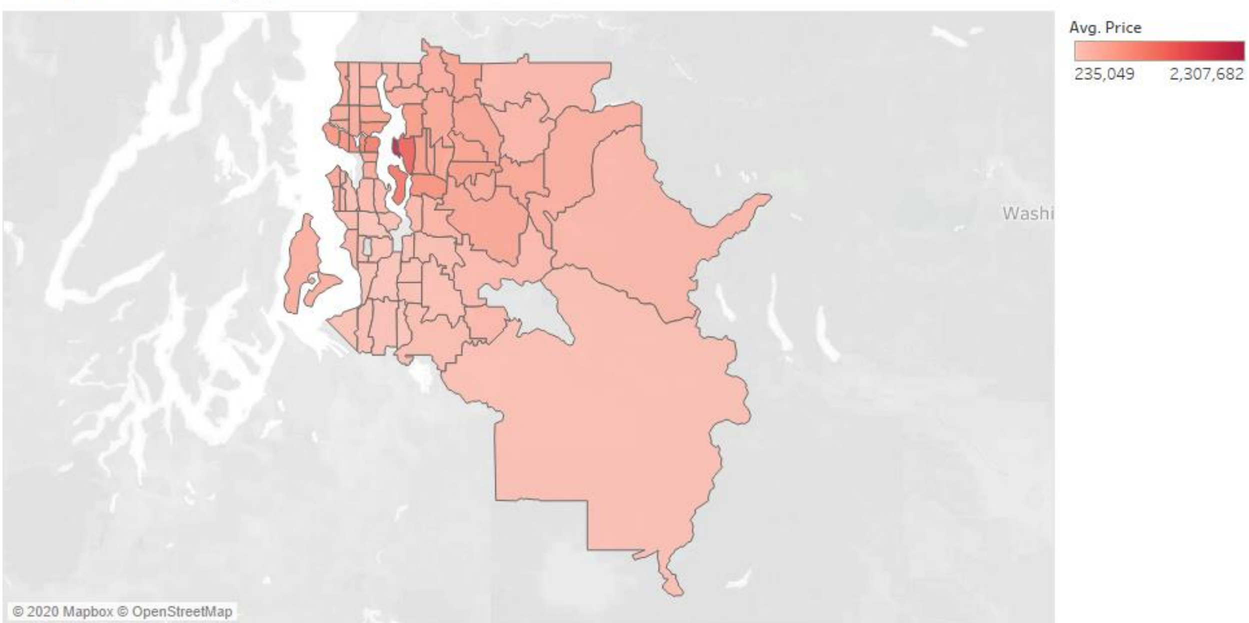
Since price was our target variable we generated a correlation matrix to find out association between the variables in the data set. We used the Pearson correlation coefficient. The price variable had high correlation with sqft_living (0.705), number of bathrooms (0.527), grade (0.665), sqft_above (0.611), sqft_living15 (0.584). These variables can be our prospective predictor variables. The zip code variable had 70 unique values. The unique values for some discreet variables are given below.

```
Unique Bedrooms--> [0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 33]
Unique Bathrooms--> [0.0, 0.75, 1.0, 1.25, 1.5, 1.75, 2.0, 2.25, 2.5, 2.75, 3.0, 3.25, 3.5, 3.75, 4.0, 4.25, 4.5, 4.75, 5.0, 5.
25, 5.5, 5.75, 6.0, 6.25, 6.5, 7.5, 7.75, 8.0]
Unique Floors--> [1.0, 1.5, 2.0, 2.5, 3.0, 3.5]
Unique Waterfront--> [0 1]
View--> [0, 1, 2, 3, 4]
Condition--> [1, 2, 3, 4, 5]
Grade--> [1, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13]
```

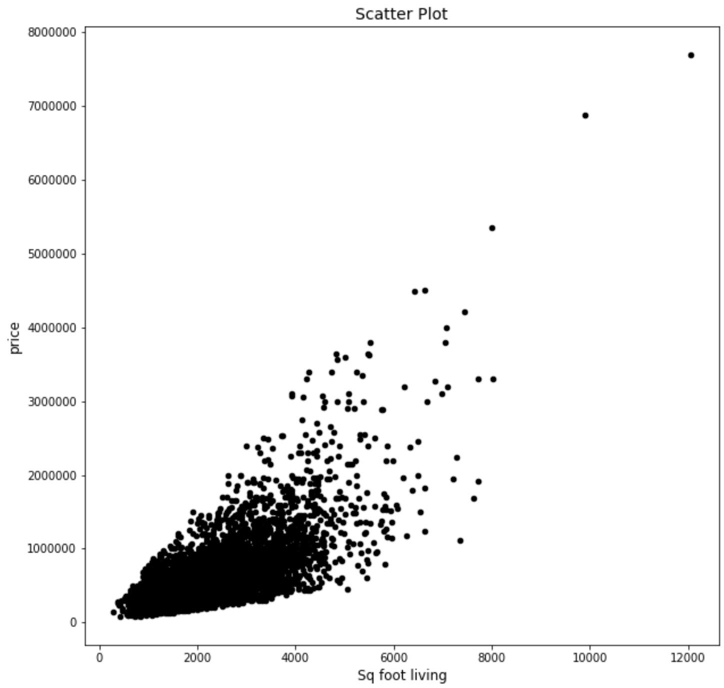
1.2.3 Multivariate Analysis and Anomaly Detection

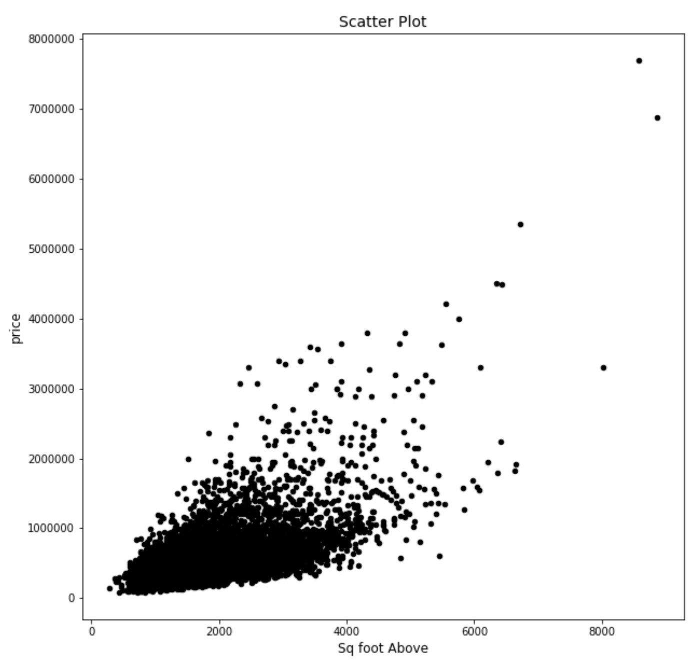
In the bedrooms variable, a value of 33 bedrooms seems highly unlikely. It had only one observation. The houses with 10 and 11 bedrooms also had only one observation each. Most houses did not have a waterfront, only 82 houses had a waterfront and the mean price of these house were almost 3 times the mean price of houses which did not have a waterfront. In terms of the number of floors two houses had 3.5 floors and these two houses had the highest mean price. A plot of the mean price against the Zip code revealed that in certain Zip Codes the average prices were very large. These were concentrated around the bay area. The zip codes with highest mean prices were 98039 (mean price of 23, 07,682 USD), 98004 (mean price of 14, 33,854 USD), 98040 (mean price of 11, 86,892 USD) and 98112 (mean price of 11, 41,247 USD). In terms of the variable view the category which had the value 4 had the highest mean price 15, 05,171. In terms of the variable grade the houses with grade 13 had the mean price of a staggering 42, 21,429 with only 7 houses in this category. The grade 13 houses consists of a house with the highest price of 77, 00,000 USD. All of these houses came in different areas. The mean prices for houses belonging to the grades 10, 11, and 12 were in excess of a million USD. The grade 12 had 45 houses with a mean price of 21, 66,210. Both the variables sqft_living and above have are concentrated between 1000 to 5000 square feet.

Average House Prices by Zip Code



Map based on Longitude (generated) and Latitude (generated). Color shows average of Price. Details are shown for Zipcode.


Scatter Plot
price
Sq foot living
8000000
7000000
6000000
5000000
4000000
3000000
2000000
1000000
0
0
2000
4000
6000
8000
10000
12000


Scatter Plot
price
Sq foot Above
8000000
7000000
6000000
5000000
4000000
3000000
2000000
1000000
0
0
2000
4000
6000
8000

1.3 Model Building

Since we are dealing with labelled data here the most obvious choice would be to use a supervised learning algorithm here to solve the problem. Since the target variable in this problem is numerical, this would be considered as a regression problem. We can either use the Multiple Linear Regression or the Regression Tree algorithm.

The choice of predictors for the model is an important step before fitting a model. The obvious choice for predictors were those numerical variables which had a high correlation coefficient in terms of the target variable (price). The Linear Regression model was used here using the sklearn package in python.

Some predictors were tried out at the beginning of the model building process. The root mean square values for the initial set of predictors using Linear Regression was very high (close to 4, 40,000). One major improvement came to RMSE value with the introduction of the zip code variable as a predictor variable. It reduced the RMSE value by almost half. Since zip code is a categorical variable we had to use dummy variables to introduce it into the linear regression model. After trying out various permutations of predictor variables the variables in table 1 yielded the lowest value of RMSE. The RMSE values increased when winsorization technique was used for outlier treatment for some of the variables like sqft_living, sqft_lot, sqft_above which was an interesting observation.

Further the Regression Tree model was also used to see if it predicted the price values more accurately. The RMSE values for the Regression Tree model were quit higher than the Linear Regression RMSE values. Since the objective of the project was predicting house prices accurately the Linear Regression model performed better at predicting house prices more accurately than the Regression Tree model. The RMSE values for the regression tree model is provided in table 2.

Linear Regression		
Variables	RMSE_Validation	RMSE_Test
Zip code	158435.4508	152974.761
grade		
bathrooms		
sqft_above		
sqft_living		
view		
bedrooms		
sqft_living15		
sqft_basement		
waterfront		
floors		
sqft_lot		
Condition		

Table 1

Regression Tree		
Variables	RMSE_Validation	RMSE_Test
Zip code	259611.412	260194.534
grade		
bathrooms		
sqft_above		
sqft_living		
view		
bedrooms		
sqft_living15		
sqft_basement		
waterfront		
floors		
sqft_lot		
Condition		

Table 2

1.4 Conclusion

In conclusion given the problem at hand, which was predicting the house prices given various variables, the linear regression model proved to be the best model in terms of the accuracy among the models tested and used.

References

Almaliki, Z. A. (2019). *https://towardsdatascience.com/.* Retrieved from https://towardsdatascience.com/do-you-know-how-to-choose-the-right-machine-learning-algorithm-among-7-different-types-295d0b0c7f60

Singh, C. (2013). Housing market in India: A Comparison with the US and Spain. *Singh, Charan, Housing Market in India: A Comparison with the US and Spain (May 1, 2013). IIM Bangalore Research Paper No. 406*.

Codes, Data Set and other Related Documents (GitHub Repository):

https://github.com/SatishFaction/Housing-Price-Prediction