House Price Prediction in King Country, USA – Task 1

1. Introduction

The analysis was aimed to predict the price of a house based on eighteen features such as; number of bed rooms, bathrooms and floors etc. using multiple linear regression model. Predicting the house price and capturing the patterns of the house price based on the features is an essential thing for house sellers nowadays.

2. Methodology

Multiple linear regression was used to predict the house price in the city. This method is a well-known supervise statistical method to predict a continuous variable based on other independent variables.

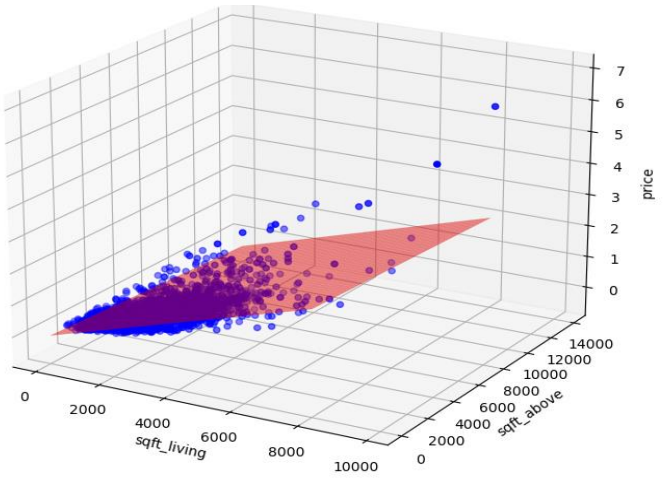
$$price = c + b_0\,(number\ of\ bed\ rooms) + b_1\,(\,bathrooms\,) + \cdots + error \qquad (1)$$

(1) equation is represented the model equation for the price prediction. Furthermore, c represents the intercept of the model, b values represent the coefficient of each independent variables of the model and model is performed to reduce the residual sum.The performance of the model is calculated from the coefficient of determination method (COD) and it is a widely used interpretable method to calculate the effectiveness of the regression model.

3. Results and Discussion

First and foremost, the analysis was tried to capture the patterns of the dataset using exploratory data analysis and data preprocessing methods. Initially, the analysis was tried to capture the missing values, duplicate values, data distribution using density plots and correlation between each variable using heatmap and correlation with the response variable using bar chart of correlation matrix. There are not missing values and analysis was removed the duplicate values. Basically, it was selected only two independent variables based on the correlation which is the linear relationship between the variables with price and data distribution.

Figure 1: Plane which represents the regression model



The initial model was built using ‘sqft lifting’ and ‘sqft above’ with ‘price’ variable. This model was given the effectiveness of the model as 48% of COD. It means 48% of the variability in house price is explained by the regression model. [Task 1.1 & Task 1.2]

Secondly, the model was taken all the independent variables without choosing them based on different techniques and that model describes 70% of the variability in house price. The model improved rather using only two independent variables [Task 1.3]. Next, the analysis was improved its data preprocessing techniques and detected the extreme outliers based on the box plots and interquartile range (IQR) method. All the variables have the outliers except ‘floors’, ‘yr_built’, ‘zipcode’ and ‘lat’ vaiables. Then those

outliers were replaced from the median of relevant variables. The analysis was used median instead of mean to replace the outliers because none of the variables do not follow the normal distribution according to the density plots. Therefore, variable transformation for the response variable was done using log transformation, boxcox transformation and quantile transformation and choosed only quantile transformation one because it was given the normal distribution rather than others. The response variable with normal distribution is helpful to improve the accuracy of the regression model according to the theory. Furthermore, the model was built under normalizing and standardizing scaling methods. After applying those techniques standardizing data model was given 62% of COD, normalizing model was given 63% of COD and unscaled data under these preprocessing techniques was given 63% of COD.

Then, the same preprocessing was done after removing the outliers and standardizing data model was given 64% of COD, normalizing model was given 63% of COD and unscaled data under these preprocessing techniques was given 64% of COD. [Task 1.4]

As the last step the analysis was done by without handling the outliers but doing the variable transformation for scaled and non-scaled data. TABLE I represents the results of them.

TABLE II: RESULTS AFTER NON-HANDLED OUTLIERS

	Log Transformation	Boxcox Transformation	Quantile Transformation
Unscaled Data	77%	76%	76%
Standardized Data	77%	75%	75%
Normalized Data	77%	75%	75%

4. Conclusion and Recommendations

TABLE III : OVERALL RESULTS BASED ON COD

	Initial Model	Simple Model with All Variables	Outliers Replace; Median			Outlier Removal			Not handling outliers		
			Unscaled	Stand	Nor	Unscaled	Stand	Nor	Unscaled	Stand	Nor
Original Distribution	48%	70%									
Log Transformation									77%	77%	77%
Boxcox Transformation									76%	75%	75%
Quantile Transformation			63%	62%	63%	64%	65%	64%	76%	75%	75%

By comparing the TABLE IV results, the analysis can be concluded that 77% of the variability in house price is explained from the best multiple linear regression model. We have to follow the following process which is handling missing values, removing duplicate values, checking distribution of the variables and transform the response variable from log transformation, checking correlations and multicollinearity, not handling outliers and unscaling or standardizing or normalizing the data to get the best model. For the future work, if we can add more data to the model, the model may be provided better results than this.