Predicting House Prices in King County, USA.

Introduction

This report is on using machine learning algorithms to predict the price of a house using a dataset that contains information about previous house sales in King County, USA. The dataset used has 19 features which includes price, number of bedrooms, bathrooms, grade square foot living etc. Linear regression(simple and multiple) was used to develop models, and the best effective model to predict the house was evaluated. Find below a summary of steps used for the analysis.

Exploratory Data Analysis

Python was used for our data analysis and several libraries relevant to the analysis were imported.

The data was loaded, read and explored using several functions like data.head, data.info, data.describe to get adequate summary and understanding of the data.

The data. shape function was used to get the dimensions of the data. This revealed that the data had 21613 rows and 19 columns.

Data Cleaning

To ensure data accuracy and consistency, the following data cleaning processes were done.

Checked for null values: The data didn't have any null values, so there were no null values to be filled.

Checked for outliers using the Interquartile range and visualised with box plots of the variables. The outliers are attributed to some of the houses that are considered to be luxurious because they have very high prices.

Standardization of the data was done using Standard Scaler. This involves transforming the features of a dataset so that they have a mean of 0 and a standard deviation of 1. This is because different features of the dataset have different scales, and it is to prevent features with larger magnitudes from dominating the algorithm.

Data Analysis

Data Visualisation : A correlation plot was done to find out the relationship between price and other variables. Price is the target variable.

Feature Selection: The features to be used for the model were selected with 'price' being the y since it is the target variable.

Training the Model: The dataset was divided into training and test sets using a 1:3 ratio. The training set was 1/3 of the test sets. The model was then fit into the training set.

Predictions: Based on the trained model, the test set was used to make predictions and the model's performance was evaluated.

Results and Discussion

From the correlation plot, price has relatively strong relationship with sqft_living (0.7) grade(0.67) and sqft_above (0.61). It has a slightly strong relationship with sqft_living15(0.59) and bathroom(0.53).

A simple linear regression was done at first using the variables that had the strongest relationship with price which is sqft_living and grade. Looking at fig 1 and 2, it can be seen that models using price with sqft_living and grade did not give the best fit. Bathroom was also used to test the model because it is likely that a house price is affected by the number of bathrooms it has.

As seen in Table 1, these variables on a simple linear regression model did not give an effective model. The table also displays the results gotten when several other features where combined using multiple linear regression. Rationally, one would think that bathroom and bedroom will be a great determinant of price, but our model refuted this assumption as these variables had a very small coefficient of determination.

From the analysis, the best model was gotten when all the features were used against price. The coefficient of determination was 0.69 which is the highest gotten from all the other models used. This means that about 69% of the variance in house prices is explained by the 18 features included in the model. Even though, the mean squared error is quite high indicating a substantial amount of variability not explained by the model, and 31% of the variance is unexplained, the model is a good fit, and still the best compared to other models, however, there is room for improvement. Thus, the best model to predict the house price is a multiple regression model where all the features are combined.

Variables	Mean Squared Error	Coefficient of Determination
Price vs all the other 18 features	45432134835.745	0.69
Price vs square_ft_living	72251932678.752	0.50
Price vs grade	80176587956.507	0.45
Price vs bathroom	103316332478.788	0.29
Price vs bedroom	131633238349.300	0.09
Price vs sqft_living, sqft_above	72211798793.042	0.50
Price vs sqft_living vs grade	66651553549.949	0.54
Price vs grade, sqft_living, sqft_above	65772005464.491	0.54
Price vs grade, sqft_living, sqft_above,	65890394607.029	0.54

sqft_living15		
Price vs bathrooms , sqft_living, grade, sqft_above , sqft_living15	65634041423.542	0.55

Table 1: table showing different models with different features.

Fig 1: Price vs Sqft_living

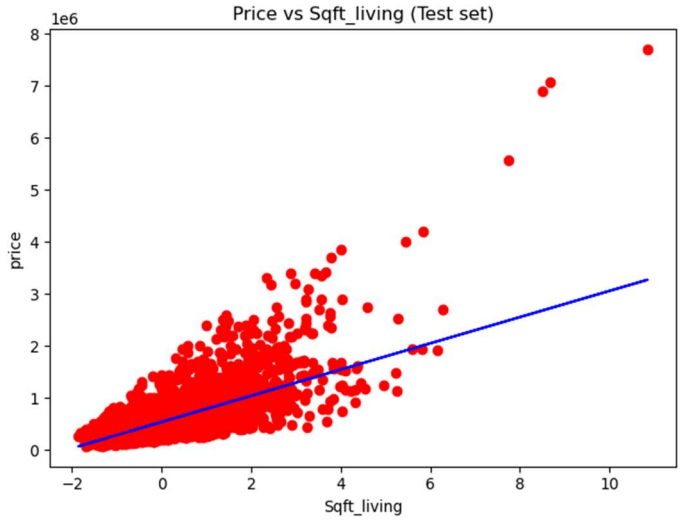
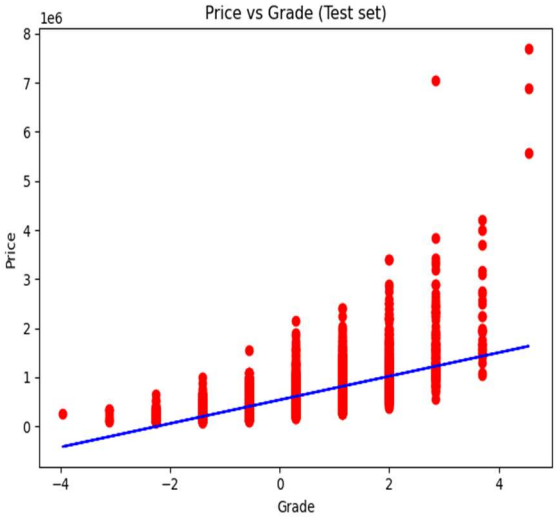


Fig 2: Price vs Grade



Conclusion

Using the simple and the multiple linear regression model, the multiple linear regression using all the data features appears to be the best model to predict the house price in Kings County,USA. Even though there is still room for improvement, as it as a coefficient of determination of 0.69 and a very high mean squared error, exploring more advanced techniques might help to capture underlying patterns in predicting the house prices.