House Price Prediction using Linear Regression

Manish Bharti
1947235, 4MCA
Christ (Deemed to be University),
Bangalore
manish.bharti@mca.christuniversity.in

Mareea Mehbin
1947236, 4MCA
Christ (Deemed to be University),
Bangalore
mareea.mehbin@mca.christuniversity.in

***Abstract*— Buying a house is a crucial concern for many people as they increase every year and there is an uncertainty related to it. People are very cautious as they have to match all the things with their budgets and their needs. The objective of the paper is to forecast the house prices for non-house holders based on their financial provisions and their aspirations. The paper focuses on the Linear Regression algorithm which is a type of predictive algorithm. The paper briefs about the regression and further details. The dataset that has been referred to in this paper is House Sales in King County, USA[6] which illustrates the prediction of house price based on various attributes of house data. The linear regression model has been trained on the training data and after training, the model is evaluated for test data.**

Keywords*— *linear regression, prediction, machine learning, training, testing

I. Introduction

Analytics focuses on inference by statistical and mathematical analysis of data. The analysis helps to identify the problem from the collected data source. The solutions or other decisions can be provided with a data analytics tool like Online Analytical Processing (OLAP). Later it uses various tools and algorithms for better outcomes of data.

There are many technologies used in data analytics but predictive analytics is the one that uses machine learning algorithms and statistical analysis for future prediction. Here, we are going to use a linear regression algorithm for house price prediction. House price is a prediction problem which relies on several different contributing factors which include the location of house, residing area, view, count of bedrooms, bathroom availability and so on.

II. Problem Statement

To create a linear regression model that is provided with more than 20 influencing factors for the price prediction of a house. The goal of this statistical analysis is to help us understand the relationship between house features and how these variables are used to predict house prices.

III. Dataset Description

Online property companies offer valuations of houses using machine learning techniques. The aim of this report is to predict the house sales in King County, Washington State, USA using Multiple Linear Regression (MLR). The dataset consisted of historic data of houses sold between May 2014 to May 2015.

We will predict the sales of houses in King County with an accuracy of at least 75-80% and understand which factors are responsible for higher property value - $650K and above.

The dataset consists of house prices from King County, an area in the US State of Washington, this data also covers Seattle. The dataset was obtained from Kaggle. This data was published/released under CC0: Public Domain.

The dataset consists of 21 variables and 21,613 observations.

Attributes information:

- id - Unique ID for each home sold
- date - Date of the home sale
- 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 zip code area the house is in
- lat - Latitude
- 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.

IV. Data Visualization

Data visualization is the representation of data or information in a graph, chart, or another visual format. It communicates the relationships of the data with images.

This is important because it allows trends and patterns to be more easily seen.

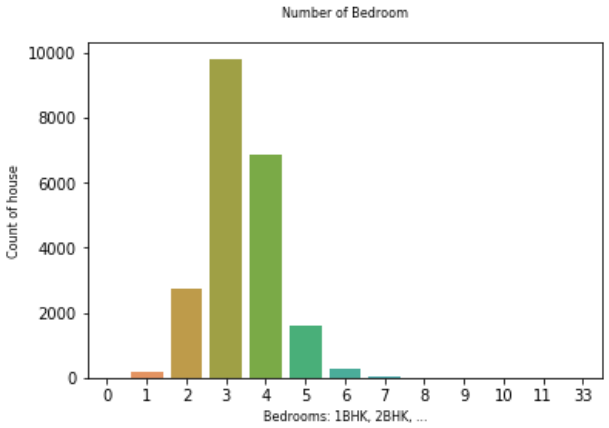

Fig. 4.1: Barchart - Number of bedrooms

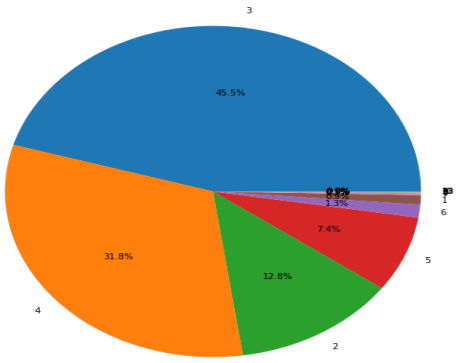

Fig. 4.2: Pie Chart - Number of bedrooms

These two visualizations show that houses having 3-rooms are the highest sold in numbers followed by 4 BHK and 2 BHK.

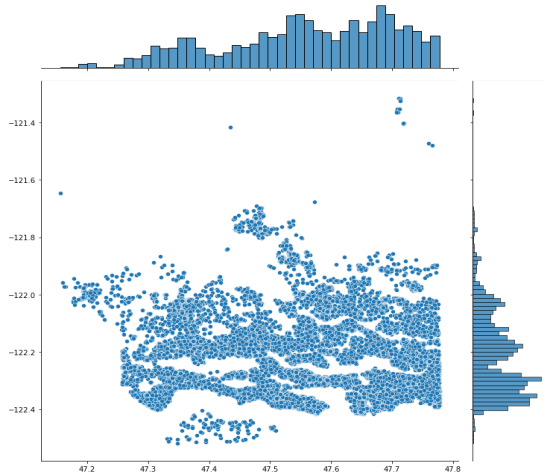

Fig. 4.3: JoinPlot - Longitude and Latitude

The Join Plot function helps us see the concentration of data and placement of data and can be really useful. For latitude between -47.7 and -48.8, there are many houses, which would mean that maybe it's an ideal location and for longitude, we can see that concentration is high between -122.2 to -122.4. Which would mean that most of the buy's have been for this particular location[4].

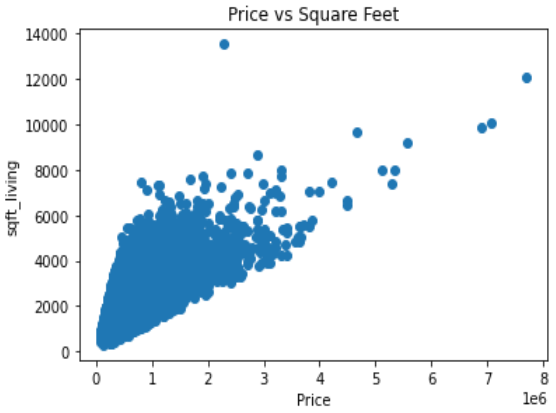

Fig. 4.4: ScatterPlot - price and sqrt_living

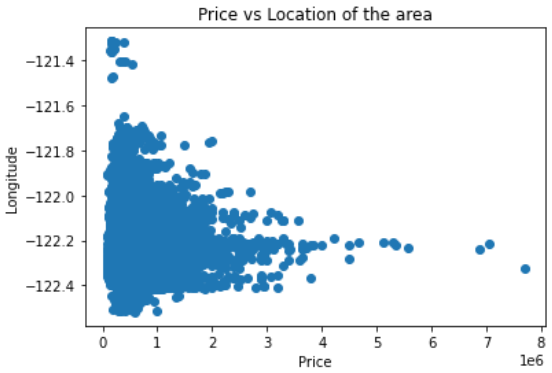

Fig. 4.5: ScatterPlot - price and sqrt_living

A scatter plot is a chart type that is normally used to observe and visually display the relationship between variables. From the figure 4.4 we can see that more the living area, more the price though data is concentrated towards a particular price zone, but from the figure 4.5 we can see that the data points seem to be in linear direction.

V. Model Evaluation

Regression is a statistical analysis method to identify the relationship between the variables. The relationship can be identified between the dependent and independent variables[1].

It can be described using probability distribution functions:

$$y = f(X, \beta)$$

Here, Y is a dependent variable, X is the independent variable and β is an unknown parameter.

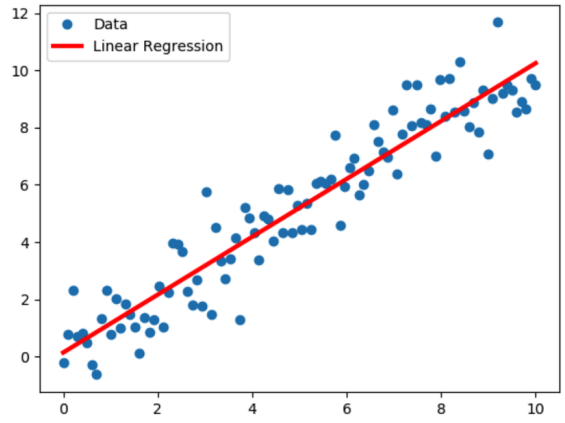

Fig. 5.1: Linear Regression

Linear Regression is the most common predictive model which can be used in one of two ways - to establish if there is a relation between two variables or to see if there is a statistically significant relationship between the two variables, which helps in predicting the future. Apart from univariate or multivariate data types the concept is linear. Linear regression can be considered as a line of best fit and it can be either simple linear or multiple linear regression. Commonly recognised as y = mx + b, the equation for linear regression is similar to the equation of a straight line, where 'y' is the unknown variable, 'x' is the known variable, 'b' is the intercept, which when increased move up or down along the y axis, and 'm' is the slope or gradient, if it changes, then the line rotates along the intercept.

Data is actually a series of X and Y observations as shown on the scatter plot. They don't follow exactly a straight line, however, they do follow a linear pattern, hence, the term linear regression. Assuming that we already have the best fit line, we can calculate the error term epsilon (ε) also known as the residual and this is the term that we would like to minimise along all the points in the data series. So, this linear equation in statistical notation is represented as follows:

$$y = x\beta + \varepsilon$$

This equation or trend line is set through the data points to predict the outcome.

As for our model, it was seen in the various scatter plots in our visualisation that the data is scattered in a linear direction which gave a positive sign for us to proceed with linear regression as our prediction model. We used the python library 'sklearn' for this purpose. Prices are to be predicted, thus, we set labels (output) as price columns and we also converted dates to 1's and 0's so that they don't influence the data much, 0 for houses which are new i.e. built after 2014. To split the data into training data set and testing data set, we imported another dependency from sklearn library and we split the data as 90% train and 10% test. We used random_state in order to randomise the splitting of data. After fitting the model using our training data, we found its accuracy score or prediction to be 73.2%, making it a weak learner or weak prediction model.

To improve the performance of a weak prediction model gradient boosting techniques are used for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models typically decision trees For our model, we used the gradient boosting regressor and set the parameters as follows:

- n_estimator - 400
- max_depth -5
- min_sample_split - 2
- learning_rate - 0.08
- loss - 'ls'

We set the n_estimator i.e. the number of boosting stages to perform as 400, not too high to avoid overfitting the model and the loss function used is 'least squares regression'. Fitting the gradient boosting model with our training data we got an accuracy of 92.5%.

VI. Inferences and Conclusion

This paper presented a machine learning model for forecasting the prices of houses in King County, USA. We used Linear Regression and Gradient Boosting Regressor for the prediction of the prices and found that using a gradient boosting technique for our model in particular increased the accuracy rate by 19.3%. Our initial aim was to achieve an accuracy between 75% to 80% and with a simple linear regression model achieved an accuracy of 73.2% and using a gradient boosting regressor fetched an accuracy of 92.5%, an increase of 19.3%. The learning rate of the gradient boosting model affects the accuracy, when set as 0.1, it gave an accuracy of 91.9% and when we changed the learning rate to 0.08, accuracy increased to 92.5%. The figure 6.1 describes the small deviance between the train set and test set, which shows that until around the 25th boosting iteration, the training and testing scores were almost the same, as can be seen overlapping in the following figure.

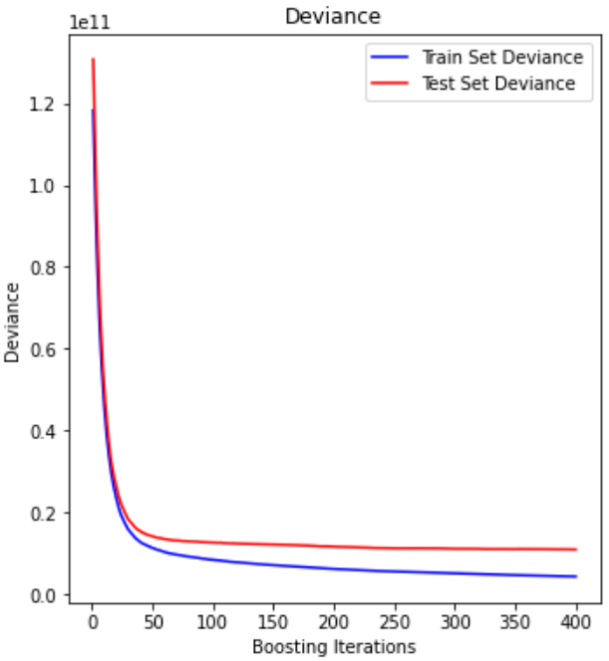

Fig. 6.1: Deviance - train vs test

However, after that they started to deviate and the gap between their deviance formed a plateau from the 100th

iteration until the end. We ran impurity-based feature importance and permutation importance and found that the features or variables - view, bedrooms, yr_renovated, zipcode - in a different order are the top 4 features which are important for the prediction of prices. Out of the 4 features, 'bedrooms' is the most influential feature that affects the forecasting in this case.

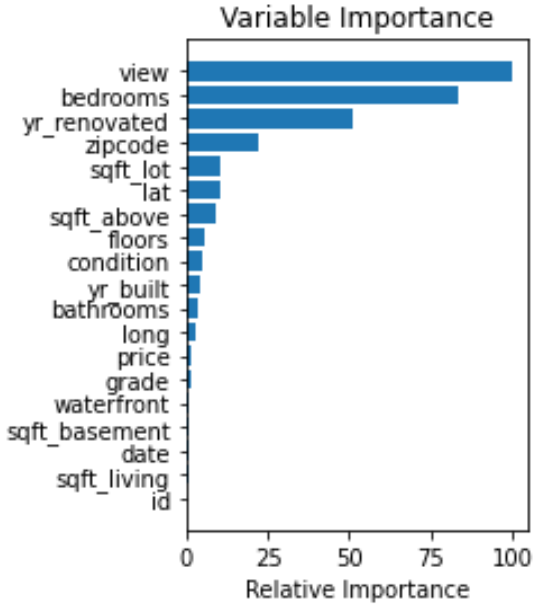

Fig. 6.2: Impurity-based Feature Importance

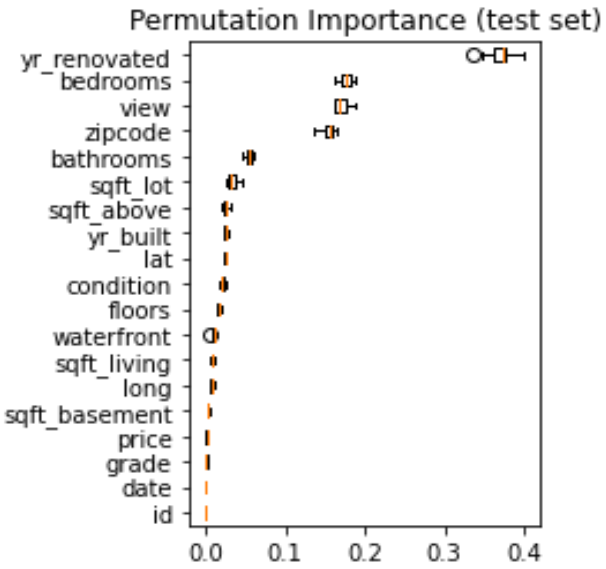

Fig. 6.3: Permutation Importance

References

[1] Kavitha, S., S. Varuna, and R. Ramya. "A comparative analysis on linear regression and support vector regression." 2016 online international conference on green engineering and technologies (IC-GET). IEEE, 2016.

[2] Alfiyatin, Adyan Nur, et al. "Modeling house price prediction using regression analysis and particle swarm optimization." International Journal of Advanced Computer Science and Applications 8 (2017).

[3] Madhuri, CH Raga, G. Anuradha, and M. Vani Pujitha. "House price prediction using regression techniques: A comparative study." 2019 International Conference on Smart Structures and Systems (ICSSS). IEEE, 2019.

[4] https://towardsdatascience.com/create-a-model-to-predict-house-prices-using-python-d34fe8fad88f

[5] https://www.kaggle.com/mohammadhy/eda-statistic-ml

[6] https://www.kaggle.com/shivachandel/kc-house-data/

[7] https://www.kaggle.com/harlfoxem/housesalesprediction