**HOUSE PRICE PREDICTION**

Final Report

Harshini Nimmagadda

Date: 06/01/2023

Contents

[**1)** **Introduction** 1](#_Toc123754026)

[**a)** **Defining problem statement** 1](#_Toc123754027)

[**b)** **Need of the study/project** 1](#_Toc123754028)

[**c)** **Understanding how data was collected in terms of time, frequency, and methodology** 1](#_Toc123754029)

[**d)** **Visual inspection of data (rows, columns, descriptive details)** 2](#_Toc123754030)

[**e)** **Understanding of attributes (variable info, renaming if required)** 2](#_Toc123754031)

[**2) EDA and Business Implication** 3](#_Toc123754032)

[**a)** **Univariate analysis** 3](#_Toc123754033)

[**b)** **Bivariate analysis** 7](#_Toc123754034)

[**3) Data Cleaning and Pre-processing** 11](#_Toc123754035)

[**a) Missing Value Treatment** 11](#_Toc123754036)

[**b) Outlier Treatment** 12](#_Toc123754037)

[**c) Variable Transformation** 13](#_Toc123754038)

[**d) Addition of New Variables** 13](#_Toc123754039)

[**e) Removal of Unwanted Variables** 13](#_Toc123754040)

[**4) Model building** 14](#_Toc123754041)

[**a)** **Build various models** 14](#_Toc123754042)

[**Linear Regression** 14](#_Toc123754043)

[**KNN Model** 14](#_Toc123754044)

[**Linear Regression – Lasso** 14](#_Toc123754045)

[**Linear Regression – Ridge** 14](#_Toc123754046)

[**Decision Tree** 15](#_Toc123754047)

[**Gradient Boosting** 15](#_Toc123754048)

[**Bagging** 16](#_Toc123754049)

[**Random Forest** 16](#_Toc123754050)

[**5) Model validation** 17](#_Toc123754051)

[**6. Final interpretation / recommendation** 18](#_Toc123754052)

|  |  |  |
| --- | --- | --- |
| **Figure No** | **Figure Name** | **Page No** |
| 1 | Missing Values | 14 |
| 2 | After Imputing Missing Values | 14 |
| 3 | After treating outliers | 15 |
| 4 | Addition of new column – City | 15 |
| 5 | Linear Regression | 16 |
| 6 | KNN Model | 16 |
| 7 | Liner Regression - Lasso | 16 |
| 8 | Linear Regression – Ridge | 17 |
| 9 | Decision Tree | 17 |
| 10 | Decision Tree with Max depths | 17 |
| 11 | Gradient Boosting | 18 |
| 12 | Bagging | 18 |
| 13 | Random Forest | 18 |
| 14 | Model Performance | 19 |
| 15 | Boosting Imp Features | 20 |

# **Introduction**

# **Defining problem statement**

A house value is simply more than location and square footage. Like the features that make up a person, an educated party would want to know all aspects that give a house its value. For example, you want to sell a house and you don’t know the price which you may expect — it can’t be too low or too high. To find house price you usually try to find similar properties in your neighbourhood and based on gathered data you will try to assess your house price.

# **Need of the study/project**

Real Estate is a much talked about industry across the global. This is sheerly due to the impact it has on the economy. It not only generates substantial revenue for the government but also provides direct and indirect employment to thousands of people. For the developing countries and cities, the first and foremost need that they generate during development is to have ample space to live, work and function. This is where real estate comes of very high importance. Real estate is and will always be in demand for many more coming decades or even forever. Because it is one of the fundamental needs of human.

It is essential to do this study because: the main one is to have information that allows you to invest and buy safely. Other reasons also include:

* To know if you should invest in one city instead of another.
* To identify which elements, hinder investment in certain places.
* To know demographic aspects to determine the evolution of the area where you want to invest.
* To have information about investment projects in certain areas and to know if in the future they will have a development that will increase the value of the properties.

# **Understanding how data was collected in terms of time, frequency, and methodology**

This data might have been derived in multiple ways. Firstly, this data can be downloaded from the official websites, or the details recorded by the agents or brokers. Secondly, there is a chance that the data can be obtained through surveys conducted by the owners living in a common society. There are chances that an individual can collect this historical information for his/her own interest while purchasing or reselling the property. As, this project is part of Capstone project, I got this dataset from Great Learning Data Science Academics team.

# Graphical user interface, application, table, Excel Description automatically generated**Visual inspection of data (rows, columns, descriptive details)**

* The intercity dataset has 21613 rows and 23 columns. The size of the dataset is 497099.
* The price of the house is depending on various factors available in the dataset.
* Cid is the unique identifier in the dataset.
* Few columns contains special characters, which needs to be treated.
* There are missing values and outliers, which needs to be treated before performing the EDA.
* We have mixed set of data types in the given dataset. There are no duplicates.

# **Understanding of attributes (variable info, renaming if required)**

* + cid: a notation for a house.
  + dayhours: Date house was sold.
  + price: Price is prediction target.
  + room\_bed: Number of Bedrooms/House.
  + **Table

    Description automatically generated** room\_bath: Number of bathrooms/bedrooms.
  + living\_measure: square footage of the home.
  + lot\_measure: quare footage of the lot.
  + ceil: Total floors (levels) in house.
  + coast: House which has a view to a waterfront.
  + sight: Has been viewed.
  + condition: How good the condition is (Overall).
  + quality: grade given to the housing unit, based on grading system.
  + ceil\_measure: square footage of house apart from basement.
  + basement: square footage of the basement.
  + yr\_built: Built Year.
  + yr\_renovated: Year when house was renovated.
  + zipcode: zip.
  + lat: Latitude coordinate.
  + long: Longitude coordinate.
  + living\_measure15: Living room area in 2015(implies-- some renovations) This might or might not have affected the lotsize area.
  + lot\_measure15: lotSize area in 2015(implies-- some renovations).
  + furnished: Based on the quality of room.
  + total\_area: Measure of both living and lot.

# **2) EDA and Business Implication**

# **Univariate analysis**

|  |  |  |
| --- | --- | --- |
| Variable | Plot | Inference |
| Price | Chart, box and whisker chart  Description automatically generated | * There are no outliers, and the distribution is Right Skewed. |
| Room Bed | Chart  Description automatically generated | * There are outliers, and the distribution is Slightly Right Skewed. |
| Room Bath | Chart, box and whisker chart  Description automatically generated | * There are no outliers, and the distribution is Slightly Left Skewed. |
| Living Measure | Chart, histogram  Description automatically generated | * There are outliers, and the distribution is Right Skewed. |
| Lot Measure | Chart, histogram  Description automatically generated | * There are outliers, and the distribution is Right Skewed. |
| Sight | Chart, shape, histogram, rectangle  Description automatically generated | * There are no outliers, and the distribution is Right Skewed. |
| Ceil Measure | Chart, box and whisker chart  Description automatically generated | * There are outliers, and the distribution is Right Skewed. |
| Basement | Chart  Description automatically generated | * There are outliers, and the distribution is Right Skewed. |
| Living Measure 15 | Chart, histogram  Description automatically generated | * There are outliers, and the distribution is Right Skewed. |
| Lot Measure 15 | Chart  Description automatically generated | * There are outliers, and the distribution is Right Skewed. |
| Furnished | Chart, histogram  Description automatically generated | * There are outliers, and the distribution is Right Skewed. |
| Ceil | Chart, histogram  Description automatically generated | * There are no outliers, and the distribution is Right Skewed. |
| Coast | Chart, shape  Description automatically generated | * There are no outliers, and the distribution is Right Skewed. |
| Condition | Chart, histogram, box and whisker chart  Description automatically generated | * There are no outliers, and the distribution is Right Skewed. |
| Yr\_built | Chart, histogram  Description automatically generated | * There are no outliers, and the distribution is Left Skewed. |
| Quality | Chart, histogram, box and whisker chart  Description automatically generated | * There are no outliers, and the distribution is Slightly Right Skewed. |

# **Bivariate analysis**

Chart, timeline

Description automatically generated

* price: Price is the dependent variable.
* room\_bed: Has no linear relationship with price.
* room\_bath: Has linear relationship with price.
* living\_measure: Has a strong relationship with price.
* lot\_measure: Has relationship with price.
* ceil: Has no clear relationship with price.
* coast: Has no clear relationship with price.
* sight: No clear relationship with price.
* condition: No clear relationship with price.
* quality: Has slight linear relationship with price.
* ceil\_measure: Strong linear relationship with price.
* basement: No clear relationship with price.
* yr\_built: No clear relationship with price.
* yr\_renovated: No clear relationship with price.
* zipcode, lat, long: No relationship with price.
* living\_measure15: Has slight linear relationship with price.
* lot\_measure15: Has no clear relationship with price.
* furnished: Has no clear relationship with price.
* total\_area: Has no clear relationship with price.

|  |  |  |
| --- | --- | --- |
| Variable | Plot | Inference |
| Month Year | Chart, line chart  Description automatically generated | * The dayhours has been converted to Month Year for the better understanding. * The price of the house is high in March, June, May. The price is very less in Feb. |
| Room Bed | Chart, line chart  Description automatically generated | * The price of the house increases with the more no of the bedrooms. House with 4.4 bedrooms has the highest price. |
| Room Bath | Chart  Description automatically generated | * It is evident that house with more no of bedrooms will also have more no of bathrooms. So, the price of the house has an increasing trend for Room bath as well. |
| Living Measure - Sight | Chart, scatter chart  Description automatically generated | * There is an increasing trend in the price as the Living Measure increases. However, the houses with sight are very less and the houses which has sight view, the living measure is more and the price as well. |
| Living Measure - Condition | Chart, scatter chart  Description automatically generated | * Most of the houses are in an average condition. However, the price of the house is still high. |
| Lot Measure | Chart, scatter chart  Description automatically generated | * There is no relation between Price and Lot Measure. |
| Ceil | Chart, line chart  Description automatically generated | * There is a visible increase trend between Price and Ceil. |
| Coast | Chart, line chart  Description automatically generated | * The price is high for the properties which has waterfront. |
| Quality | Chart, line chart  Description automatically generated | * The price of the house increases with the quality. |
| Ceil Measure | Chart, scatter chart  Description automatically generated | * There is no relationship between Price and Ceil Measure. |
| Basement | Chart, line chart  Description automatically generated | * The price is high for the properties which has Basement. |
| Living Measure - Renovated | Chart, scatter chart  Description automatically generated | * There is an increasing trend in the price as the Living Measure increases. However, there are quite few houses which are renovated. |

# **3) Data Cleaning and Pre-processing**

# **a) Missing Value Treatment**

* Missing value Treatment has been performed before EDA.
* Before, performing Missing Value Treatment, special characters imputation has been taken care. So, the columns where the special characters are present, those are replaced with “np.NAN”
* KNN Imputer is used to treat the missing values. KNN imputer is the best approach to be used for Missing Value Treatment.

A picture containing text, receipt, screenshot

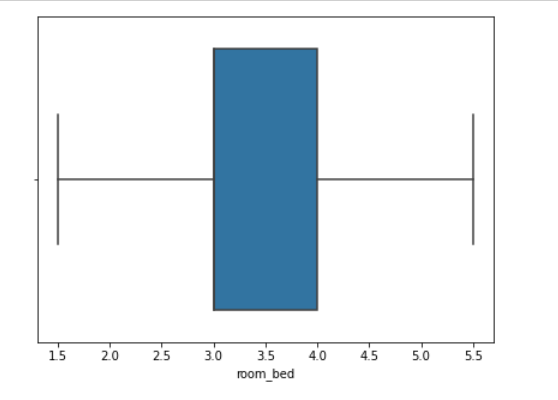
Description automatically generated Table

Description automatically generated

**Fig 1: Missing Values Fig 2: After Imputing Missing Values**

# **b) Outlier Treatment**

* As seen above in Univariate analysis, there are outliers present for few variables.
* Outlier Treatment has been performed after Univariate analysis.
* Ceil Measure, Basement, Living Measure, Lot Measure, and Room bed has outliers.
* The outliers in these variables are imputed with IQR method.
* IQR method is the best way used for Outliers treatment.

 Chart, histogram

Description automatically generated

Chart

Description automatically generated Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

**Fig 3: After treating outliers**

# **c) Variable Transformation**

* 3 variables has been transformed from the original dataset
* Day month, basement and renovated.
* Day month has been converted to Month year for the better visualization.
* Basement has been converted to categorial variable. Value – 0 has been converted to No: which mean “No basement” and the value – 1 has been converted to Yes: which mean “The property has basement”.
* Renovated has been converted to categorial variable. Value – 0 has been converted to No: which mean “No renovation” and the value – 1 has been converted to Yes: which mean “The property has renovated”.

# **d) Addition of New Variables**

* New column – “City has been added to the dataset”.
* City column is derived from zipcode.
* From Uszipcode package, searchEngine has been imported. This invocated all cities with the provided zipcodes.

Table

Description automatically generated with medium confidence

**Fig 4: Addition of new column – City**

# **e) Removal of Unwanted Variables**

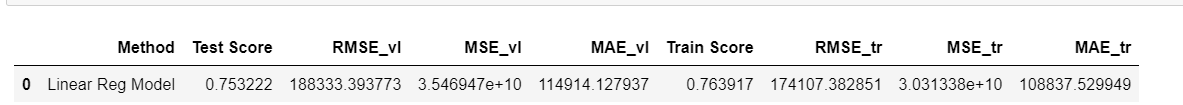
* For model building, few columns are not required and hence these columns are dropped from the dataset.
* Cid, dayhours, yr\_renovated, zipcode, lat, and long columns has been dropped from the dataset.

# **4) Model building**

# **Build various models**

## **Linear Regression**

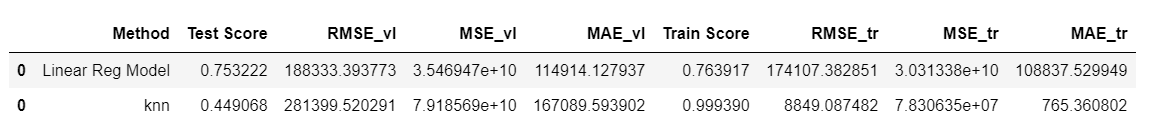
Linear regression is a supervised machine learning method that is used by the Train Using AutoML tool and finds a linear equation that best describes the correlation of the explanatory variables with the dependent variable.



**Fig 5: Linear Regression**

The Regression model performed well for both Train & Test datasets.

## **KNN Model**



**Fig 6: KNN Model**

KNN regressor performed well in training set, but the performance score in test set is very less. This is evident that the model is overfitting for training set.

## **Linear Regression – Lasso**

Table

Description automatically generated

**Fig 7: Linear Regression – Lasso**

The Regression- Lasso model performed well for both Train & Test datasets.

## **Linear Regression – Ridge**

Table

Description automatically generated

**Fig 8: Linear Regression – Ridge**

The Regression- Ridge model performed well for both Train & Test datasets, and it is like Linear Regression & Lasso models.

## **Decision Tree**

Table

Description automatically generated

**Fig 9: Decision Tree**

The initial Decision tree model shows overfit in training set with 0.99 score and low performance in test dataset.

Table

Description automatically generated

**Fig 10: Decision Tree with Max depths**

The decision tree model with modified parameter has better performed on the training set and test set compared to initial decision tree. But overall decision tree has not performed well than linear regression models.

## **Gradient Boosting**

Table

Description automatically generated

**Fig 11: Gradient Boosting**

Gradient Boosting has performed good for both Test & Training datasets.

# **Bagging**

Table

Description automatically generated

**Fig 12: Bagging**

Bagging method has performed good for Test dataset but for Training dataset, the performance is overfit.

# **Random Forest**

Table

Description automatically generated

**Fig 13: Random Forest**

Random Forest method has performed good for Test dataset but for Training dataset, the performance is overfit.

# **5) Model validation**

The best model is validated based on low RMSE as the dependent variable is continuous variable in case of binary dependent variable best accuracy will be best model.

**Table

Description automatically generated**

**Fig 14: Model Performance**

Gradient Boosting has a better performance compared to all the models with Train Score – 86% and Test Score – 79%.

Table

Description automatically generated

A picture containing chart

Description automatically generated

**Fig 15: Boosting Imp Features**

In gradient boosting model top 30 features are covering 98% variance.

# **6. Final interpretation / recommendation**

* The price of the house/property depends on the locality. So, a new column “City” has been derived from zipcode.
* The price of the house is dependent on various factors. For example: Living Measure, Room bed, Room bath etc.
* The cost of the price has an increasing trend with increase in Living Measure. So, it mean the property is sold or rented based on the square feet.
* The price of the house has an increasing trend with more no of bedrooms. The no of bathrooms is equivalent to bedrooms. So, an ideal revenue can be produced for 3- and 4-bedroom houses.
* Most of the houses are not furnished. To increase the revenue, furnished 3 – and 4- bedroom furniture houses are ideal.
* The ensemble models has performed well when compared to the other models.
* The best performance is given by Gradient Boosting, with Training score of 86% and Test score of 79%.
* The top & Imp features that increase the price of the property: furnished\_1.0, living\_measure, quality, ceil\_measure, sight, basement, yr\_built.
* The above data is also reinforced by the analysis done during bivariate analysis.
* For further improvisation, the datasets can be made by treating outliers in different ways and hyper tuning the ensemble models.