Murali b

CAPSTONE PROJECT

CONTENTS Page numbers

Figures:

1.1 Histogram 4-5  
1.2 Variable price 6  
1.3 CID variable analysis 6  
1.4 Room bed analysis 7  
1.5 Bathroom analysis 8  
1.6 CID and price 9  
1.7 Dayhours and price 9  
1.8 Pairplot 10  
1.10 Heatmap 11  
1.14 Before fixing outliers 13  
1.15 After fixing outliers 14  
1.56 Dendogram 21  
1.57 Elbow method 22  
1.63 RMSE value plot 26

Output:

1.9 Correlation 11  
1.11 Duplicate values 12  
1.12 Checking for Null values 12  
1.13 Fixing Null values 13  
1.16 cid variable unique values 14  
1.17 cid variable shape 14  
1.18 dayhours variable unique values 15  
1.19 dayhours variable shape 15  
1.20 price variable unique values 15  
1.21 price variable shape 15  
1.22 room\_bed variable variable unique values 15  
1.23 room\_bed variable shape 15  
1.24 room\_bath variable unique values 15  
1.25 room\_bath variable shape 15  
1.26 living measure variable shape 16  
1.27 living measure variable values 16  
1.28 lot\_measure variable unique values 16  
1.29 lot\_measure variable shape 16  
1.30 ceil variable unique values 16  
1.31 coast variable shape 16  
1.32 sight variable unique values 16  
1.33 sight variable shape 17  
1.34 condition variable unique shape 17  
1.35 quality variable unique values 17  
1.36 quality variable shape 17  
1.37 ceil\_measure variable unique values 17  
1.38 ceil\_measure variable shape 17  
1.39 basement variabl unique values 18  
1.40 basement variable shape 18  
1.41 year\_built variable unique values 18  
1.42 year\_renovated variable unique values 18  
1.43 year\_renovated variable shape 19  
1.44 zipcode variable shape 19  
1.45 lat variable unique values 19  
1.46 lat variable shape 19  
1.47 long variable unique values 19  
1.48 living\_measure15 variable unique values 20  
1.49 living\_measure15 variable shape 20  
1.50 lot\_measure15 variable unique values 20  
1.51 lot\_measure15 variable shape 20  
1.52 furnished variable unique values 20  
1.53 furnished variable shape 20  
1.54 total\_area variable unique values 20  
1.55 new variable details 21  
1.58 OLS Regression results 23  
1.59 train set 24  
1.60 test set 24  
1.61 LDA model coefficient 25  
1.62 RMSE values 25  
1.63 RMSE values plot 26  
1.64 Feature Importance 26  
1.65 Feature importance 27  
1.66 Feature Importance 28

1. Brief Introduction about the problem statement and the need of solving it.
2. Defining problem statement

Basically, any house is valued based on the area in which it is located and its dimensions which is measured in square feet. (For ex: 30\*40 = 1200sqft). So, any layman or common individual would obviously wish to know all the features related to housing value properties. Hence if in case if we wish to buy any house or sell any house then we will be not knowing the exact amount to which it's being valued whether the cost involved is more or less so in order to address these kind of issues we tend to take samples information which is already existing such as finding same pattern in the same locality and with that data we can try to predict the house value.

1. Need of the study/project

The basic need of the study of this project is to determine the right correct price of the housing value. So, when any person tries to buy or sell a house, usually they won't know the exact price to which it should be sold or bought. Hence the house value might be undervalued or overvalued. Based on the data such as the available information of the house property in that geographical location we can determine the price. Primarily we should ensure to make better understanding of the various important factors and features which helps in determining the right house pricing.

1. Understanding business/social opportunity

Here main ideology is to avoid the loss and time so many people will not be aware of the features which determines the house price. So, to help them further we can provide them a detailed information and guidelines based on house buy/sell services so that people can buy the property with feasible prices so that losses can be minimized and waiting time can be reduced.

EDA and Business Implication   
  
Univariate Analysis:

A picture containing diagram, text, plot, line

Description automatically generatedA picture containing diagram, text, plot, line

Description automatically generated  
Fig 1.1 Histogram

We can see most of the data representation shows as left skewed and the data is uniformly distributed.

Variable: Price   
A picture containing text, diagram, screenshot, line

Description automatically generated  
Fig 1.2 Variable price

Variable: cid   
  
A screenshot of a graph

Description automatically generated with low confidence  
Fig 1.3 CID variable analysis

Variable: room\_bed   
A picture containing text, screenshot, plot, diagram

Description automatically generated  
Fig 1.4 Room bed analysis

We see 9767 houses has 3 bedrooms, 6854 houses have 4 bedrooms and 2747 houses has 2 bedrooms.

Variable: room\_bath   
A picture containing text, screenshot, plot, line

Description automatically generated  
Fig 1.5 Bathroom analysis   
  
We see more than 5358 houses has more than 2 bathroom and 3829 houses has 1 bathroom.

Bivariate analysis

Variable used: cid and price   
A picture containing text, screenshot, diagram

Description automatically generated  
Fig 1.6 cid and price   
  
We see all the points are scattered at the bottom of the graph, so the value keeps decreasing.

Variable used: dayhours and price   
A screen shot of a graph

Description automatically generated with low confidence  
Fig 1.7 dayhours and price   
  
Here, we see there is more scattered relationship between dayhours and price.

Pair plot :   
A picture containing text, screenshot, handwriting, rectangle

Description automatically generated

Fig 1.8 Pairplot   
  
We see all the variable data representation through pairplot which shows unique distribution of data points when they are related to each other.

Correlation:   
A picture containing text, screenshot, number, font

Description automatically generated  
Output 1.9 Correlation

Heatmap:   
A picture containing screenshot, rectangle, square, colorfulness

Description automatically generated  
Fig 1.10 Heatmap

From the heatmap, we can see that there is dense correlation which exists between most of the variables.

The data seems to be unbalanced from the data analysis, hence we so we try to balance the data. We see that more people are interested in buying houses which has more bedrooms and more bathrooms so we should build more houses which has more bathrooms and bedrooms. We can see that houses with basement have higher prices than compared to houses which are very low in number. Therefore, we must build houses which has basement in them and, we can go publicity like posting ads in the sites having basement so that many people will be aware whether any house is for sale or not. People likely to tend to buy houses which has 2 floors so we can focus more on building houses with 2 floors. We see many houses are not furnished so if we do fully furnish or semi furnished then we can attract more people. As per the quality wise, we see many houses are not with good quality so if we improve the quality of the houses, we can demand more prices for houses.   
3) Data cleaning and pre-processing.   
  
Missing Value treatment (if applicable)   
  
Duplicate values:   
  
Output 1.11 duplicate values   
  
We can see that there are no duplicate values present in the dataset.

Checking for Null values:   
A screenshot of a computer

Description automatically generated with low confidence  
Output 1.12 checking for null values

We see that there are 108 null values in room\_bed , room\_bath , 17 in living\_measure , 42 in lot\_measure , 1 in coast & quality & ceil\_measure & basement & year\_built, 57 in sight and condition , 166 in living\_measure15 , 29 in lot\_measure15 & furnished & total\_area so let’s fix all this Null values.

After fixing Null values:   
  
We treated Null values with mode and median so now we can see that Null values has been fixed.   
A picture containing text, screenshot, font, menu

Description automatically generated  
Output 1.13 fixing Null values   
  
  
Now we can see that there are no Null values present in the dataset.

Outlier treatment  
A picture containing diagram, line, plan, plot

Description automatically generated  
Fig 1.14 Before fixing outliers.

We see many outliers present in the dataset except for lot\_measure, zipcode and lat so let’s fix the outliers.

After fixing outliers:   
A screenshot of a computer

Description automatically generated with low confidence  
Fig 1.15 After fixing outliers

Now we can that there are no outliers present in the dataset and it has been taking care already.

Variable transformation (if applicable)

Cid variable unique values:   
  
Output 1.16 cid variable unique values

Cid variable shape:   
  
Output 1.17 Cid variable shape

Dayhours variable unique values:   
  
A screenshot of a computer code

Description automatically generated  
Output 1.18 Dayhours variable unique values

Dayhours variable shape:   
   
Output 1.19 Dayhours variable shape  
  
Price variable unique values:   
  
Output 1.20 Price variable unique values  
  
Price variable shape:   
  
Output 1.21 Price variable shape  
  
Room\_bed variable unique values:   
  
Output 1.22 Room\_bed variable unique values

Room\_bed variable shape:   
  
Output 1.23 Room\_bed variable shape

Room\_bath variable unique values :   
A picture containing text, font, line, screenshot

Description automatically generated   
Output 1.24 Room\_bath variable unique values  
  
Room\_bath variable shape :   
  
Output 1.25 Room\_bath variable shape  
  
living\_measure variable unique values :   
  
Output 1.26 living\_measure variable unique values  
  
living\_measure variable shape :   
  
Output 1.27 living\_measure variable shape  
  
lot\_measure variable unique values :   
   
Output 1.28 lot\_measure variable unique values  
  
lot\_measure variable shape :   
  
Output 1.29 lot\_measure variable shape  
  
ceil variable unique values:   
  
Output 1.30 ceil variable unique values  
  
coast variable shape :   
  
Output 1.31 coast variable shape  
  
sight variable unique values :   
  
Output 1.32 sight variable unique values  
  
sight variable shape :   
  
Output 1.33 sight variable shape

condition variable unique values:   
  
Output 1.34 condition variable unique values  
  
quality variable unique values:   
   
Output 1.35 quality variable unique values  
  
quality variable shape :   
  
Output 1.36 quality variable shape  
  
ceil\_measure variable unique values :   
A picture containing text, screenshot, number, font

Description automatically generated  
Output 1.37 ceil\_measure variable unique values

Ceil\_measure variable shape:   
  
  
Output 1.38 Ceil\_measure variable shape

basement variable unique values:   
A picture containing text, screenshot, font, number

Description automatically generated  
Output 1.39 basement variable unique values   
  
basement variable shape:   
  
Output 1.40 basement variable shape   
  
year\_built variable unique values:   
A picture containing text, screenshot, font, number

Description automatically generated  
Output 1.41 year\_built variable unique values  
  
year\_renovated variable unique values :   
A picture containing text, screenshot, font, number

Description automatically generated  
Output 1.42 year\_renovated variable unique values

year\_renovated variable shape:   
  
Output 1.43 year\_renovated variable shape  
  
zipcode variable shape:   
  
Output 1.44 zipcode variable shape  
  
lat variable unique values:   
  
Output 1.45 lat variable unique values  
  
lat variable shape :   
  
Output 1.46 lat variable shape  
  
long variable unique values :   
A screenshot of a computer code

Description automatically generated  
Output 1.47 long variable unique values

living\_measure15 variable unique values:   
A picture containing text, screenshot, font, number

Description automatically generated  
Output 1.48 living\_measure15 variable unique values  
  
living\_measure15 variable shape:   
  
Output 1.49 living\_measure15 variable shape  
  
lot\_measure15 variable unique values :   
  
Output 1.50 lot\_measure15 variable unique values  
  
lot\_measure15 variable shape :   
  
Output 1.51 lot\_measure15 variable shape  
  
furnished variable unique values :   
  
Output 1.52 furnished variable unique values  
  
furnished variable shape :   
  
Output 1.53 furnished variable shape   
  
total\_area variable unique values :   
  
  
Output 1.54 total\_area variable unique values

So we have converted the variable type: ceil, coast, condition, year\_built , total\_area and long to string type. We have replaced all “$” values with nan values and then imputes all nan values with KNNImputer.

Addition of new variables:   
  
We have added new variable called house\_land\_ratio by taking the round values of living\_measure divided by total\_area multiplied by 100. The output shows as:   
  
A picture containing text, screenshot, font, line

Description automatically generated  
Output 1.55 new variable details.

4) Model building:

Dendogram with p value as 10:   
  
A picture containing diagram, rectangle, plan, line

Description automatically generated  
Fig 1.56 Dendogram with p value as 10

Elbow Method:   
A graph with a red line

Description automatically generated with medium confidence  
Fig 1.57 Elbow Method

OLS Regression Results: A screenshot of a computer

Description automatically generated with low confidence  
Output 1.58 OLS Regression results

Train set:   
  
A picture containing text, screenshot, number, font

Description automatically generated  
Output 1.59 train set

Test set:   
  
A picture containing text, screenshot, number, font

Description automatically generated  
Output 1.60 test set

Ensemble methods are really performing better than linear models. Of all the ensemble models, Gradient boosting regressor is giving better R2 score. The important features that mostly drive the price of the property are: 'furnished', 'year\_built', 'living\_measure','quality', 'lot\_measure15', 'ceil\_measure', 'total\_area'. Best of all models is Xtreme Gradient boost which is an enhanced version of gradient boost. It includes regularization and is faster too. Going forward this model can be improved further as we don’t have much data for very high-priced houses. So, when more data comes in, we can revisit our model and make necessary changes to accommodate more variation in data to deliver better results by decreasing RMSE values. For further improvisation, the datasets can be made by treating outliers in different ways and hyper tuning the ensemble models.One needs to thoroughly check its property on parameters suggested and list its price accordingly, similarly if one wants to buy house, they need to check the features suggested for the house and calculate the predicted price. The same can then be compared to listed price. Making polynomial features and improvising the model performance can also be explored further.

5) Model validation:   
  
LDA model coefficient:   
A picture containing text, screenshot, font, number

Description automatically generated  
Output 1.61 LDA model coefficient

RMSE values:   
A picture containing text, font, screenshot, number

Description automatically generated  
  
Output 1.62 RMSE values   
  
RMSE value plot:   
A picture containing text, screenshot, plot, line

Description automatically generated  
Graph 1.63 RMSE value plot   
  
Feature Importance:   
A picture containing text, screenshot, font, line

Description automatically generated  
Output 1.64 Feature Importance  
  
A screenshot of a computer code

Description automatically generated with low confidence  
Output 1.65 Feature Importance   
  
From the above output, we can see top 5 important features are quality, living\_measure, lat , long and coast.

Feature importance was mainly performed to reduce the overall dimensions, but there was no improvement found in the modeling accuracy. Each of the outlier strategy was evaluated for the various algorithms, and the best performing algorithm was subjected to hyper parameter tuning. We can look at the co-efficient and interpret values based on positive and negative. variables co-efficient is positive meaning for example every increase in value of living\_measure there is price increase we can apply the logic in other variable which is positive and whichever variable co-efficient is negative for example if there are negative co-efficient variable tends to lose the price or can say there is decrease in price. P -Values whichever variable has a higher value then 0.05 that is insignificant and the variables which are under or equal to 0.05 are significant value and important variables for prediction on our data.

Feature Importance:   
A screenshot of a computer

Description automatically generated with low confidence  
Output 1.66 Feature Importance  
  
We can see top 5 most important features are quality, living measure, lat , long , furnished.

The best model is Random Forest Model because other models are either overfitted or have low value and based on RMSE value If the noise is small, which means model is good at predicting observed data, and if RMSE is large, this means model is failing to account for important features. These models will help us analyze the best way to determine correct prices and areas and all the variables which are available in our data.

6) Final Interpretation/Recommendations:   
  
In hierarchical clustering with the help of Dendogram, we can see that there are 3 group of clusters. So, we can observe that the first group of clusters which shows the highest price range values. The second group shows the lowest price range and 3rd group shows the medium price range values. We can see that most of the properties belongs to lowest price.   
  
In KNN clustering, we can see here that three groups are created. First group is 1 which has 1030 properties and group 2 has 12743 properties then group 3 has 7840 properties. We can observe that most of the properties lies between range of group 2 then group 3 followed by group 1. We see that group 2 has price range between group 1 and 3 and group 1 has maximum price range followed by group 3.

When compared to other variables, so we can apply a new strategy to educate the people about the information which most of them are not aware of it. Since many people do not have interest to know about all things so we make sure that they understand the details related to their houses and guiding them which in turn can improve the business. We can create a user-friendly customized website related to all information or a mobile app so that they can check the details about the details of their houses so based on their choice of houses we can make the sales.

We can see that properties with higher price value have a greater number of sites when compared to houses with lesser price. So once the data is cleaned then it’s stable for model building. We can see that there is a greater increase in the upward trend in price values with respect to ceil\_measure. We can see that smaller house are in a better condition but with more higher prices.

We can observe that the properties which are coming under in any of the 3 clusters based on that we can predict the price of the properties and we can see if the prices get increased in the future or not.

The house can be sold based on the cluster by choosing the price average price but if the house is situated in good location and having more bedrooms and bathrooms with fully furnished having coastal view then price can be increased.

So, I would suggest if you wanted to sell your house, please keep in mind the value which we generated from cluster. I’m going to choose the price around 350,000$ which is in average price, and you can increase the charge around 550,000$ - 600,000$.

From the exploratory analysis, we can conclude that the overall quality of the house effects the house price. Other important features that every homeowner considers are square footage of the house, garage capacity, neighborhood, exterior condition, HVAC system, basement, and kitchen quality. Some more additional information on neighborhood like schools in the neighborhood, access to shopping, transport, and details about traffic around the area would have been more helpful in making the model.

Houses with full bath in basement, good condition, more low-quality finished area (sqft), bigger garage are (sqft), more square footage in 2nd floor, lesser age, a greater number of fireplaces, recent remodeling, a greater number of half baths above basement/ground floor (for houses without basement) are priced high.

Houses with bigger front yard (more than back yard), bigger lot area, a greater number of rooms above basement/ground floor (for houses without basement), kitchen above ground floor, bigger finished square footage of second basement, bigger area in 1st floor (sqft), garage capacity, bigger area in wooden deck, yer sold and enclosed porch decreases the house price.   
  
Businesses/homeowners can quote a price for based on some of the important features such as the overall condition of the house and basement (if any), bigger garage, extra square footage in 2nd floor, recently built/remodeled house, and a greater number of bathrooms.