PROJECT DONE ON HOUSING PRICE PREDICTION CASE STUDY

SUBMITTED BY :
JUHI MISHRA

ACKNOWLEDGMENT

I started doing project on Housing Price prediction for a US based company named Surprise Housing who wants to repurchase and sale the houses in the market to have profit and establish themselves as renowned real estate company in US. In other word they wants to enter to real estate business and to have good profit along with good customer base in market, for that as a data scientist we need to predict on the basis of data collected from the sale of houses from Australia, whether investing in property which company wants to purchase at lower price and resale the same at higher price will be profitable for them or not and what is the trend that the respective market follows where this company wants to invest . To complete this project I had taken help from various sites which were YouTube , Kaggle, geeksforgeeks, Google search and at last went through whole case study provided by Flip Robo what exactly the data is and how to process the same to build the model and predict for what our client require. I am thankful to all the resources and site with the help which able to complete this project on time.

INTRODUCTION

- ## BUSINESS PROBLEM FRAMING :

Real Estate business is one of the leading business in market with good revenue and customer base as we know that with time most of the population wants to invest and have their own house or property either for resident purpose or for investing purpose , This market is such that never went through a recession where they may face a big loss. Most of the companies wants to enter to this domain to have good profit and establish themselves as renowned real estate company. Now what the problem here is as lot many competitors are there in the market so to establish, one should have better knowledge what's the trend going on regardless price of purchase and sale of house in a market where a company wants to establish themselves.Also should gather information regarding what the preference of customers and how to deal with the same. To invest in any property one should have good knowledge of what property to purchase and how much return they can earn by selling the same. In real world without having information of market trend it is almost impossible to establish themselves into this business, as these type of business require good analysis so that once they enter should not face loss or stuck with properties in which they already invested , which may turn to dead asset and will have to face a big loss.

- ## CONCEPTUAL BACKGROUND OF THE DOMAIN PROBLEM :

Here we had taken the case study of real estate company named surprise Housing who wants to enter to US market to have investment in purchase of properties at low price and resale the same at higher price, for which they hired Data scientist whose role is to identify whether to invest in perspective properties or not. For this they have been provided data from the sale of Houses in Australia to know the trends going on in market and analyse whether to invest and if yes then how much? As previous data gives insight about what's the trend a particular market follows and what are the pros and cons for the one who wants to invest. As a data scientist we need to go through the previous dataset provided

by company to have insight and gather information using domain knowledge whether investing in properties will be profitable or not , if yes then how much revenue a company can generate from this business and what market strategy to follow as there are lot many other companies who are already established. Using machine learning we can build a model which can predict what will be the actual vale of the perspective property and whether to invest in it or not.

• REVIEW OF LITERATURE :

This dataset is all about prediction of house price for further purchase and resale to enter into real estate business in US market. With the help of model prediction using machine learning algorithm came across whether company should invest in perspective properties or not. While researching came across that this dataset is all about trend followed by market and what type of customer do invest in which type of properties. With the help of visualisation technique using matlotlib and seaborn came across some insights mentioned below :

1st we came across MSsubclass which means type of dwelling involved in sale. Here type of dwelling means type of living quarters in which people lives. Which we came across that type 25 and 75 are higher in demand, as the selling price for these two properties are higher , we can assume that most population wants to have properties of these two type so the demand is high , simultaneously price are higher. Zone wise seen that FV is the most expensive zone as sales price for this we can see is higher almost above 2 lac. While purchasing property people not only focus on how the property is and price even lot many other concerns are there, say for an example near to the house what is the connectivity of road, how wide the road adjacent to property, what is the shape of property, configuration, slope nearby to property, neighbourhood, what is the condition of property as if any one who invests in repurchasing, condition of property matters a lot as no one wants to invest in such properties which are in bad condition, reason behind is again they have to reinvest in renovation which will be double cost and may be a loss in buying such houses. Buying a house is a dream of almost all people in the globe so they look everything what they want in their house and wants to have best for them. House style which brings looks to any of the houses and as every person do have their own choice and likes so they wants to have their property as per their choice and convenient. In repurchasing as people do not know what quality of material being used

during construction, so verify the same whether the material used are of good quality or not . And this can be assumed by seeing the condition and how old the house is. coming to interior of houses what are the trend we observed through insight with the help of visualisation that mostly people even looks how the roof style and quality of material for the same being used. Exterior of house which is also one of the important aspect which people looks while investing in properties. How good the exterior depending upon which quality and what material being used while construction. This was all about property looks, condition and accessibility , now will go through some other necessary aspects which is one of the important requirement for any buyer.

Basement is mainly the floor of a building which is partly or entirely below ground label. Basement quality, condition, exposure, finishing, in what area does it cover, means in how many sft do the basement being constructed, unfinished surfaced area , total area covered. These are the information collected regarding basement of property .Use of basement we all know and how much it is important, as most of the people uses basement for car or bike parking. Even for some other purpose if the place is bigger can be used. What type of heating being used and what is the quality of heating matters while anyone wants to have a particular property. Now a days AC is the requirement for everyone and as we observed through our visualisation technique that properties which are having central ac are more costly. And we know that price varies with demand. Here we can assume that demand for centralised AC in US are more so the price is high. What sort of electricity being used in a building or apartment matters as electricity being considered as necessity and no one wants to compromise with this now a days. Here we can see that property with SB rkr are more in demand as the price for such properties are higher.

Now coming to the building construction type which means in how many sft the 1st floor being constructed then 2nd floor and so on and how many floor one house do have. Here in this dataset we seen that on 1st floor from 500 - 1500 sft being sold at higher price while in 2nd floor even 250 sft been sold and maximum upto 1250 are having higher price. Here we can assume that may be top floor do have higher demand so even with less sft people do prefer and purchase with almost same amount but less in area. Because we already noticed that 2nd floor even 250 sft being sold at 1 lac while 1st floor 500 sft being sold at 1 lac too. And same for 1250 sft on 2nd floor having equivalent price with respect to 1st floor 1500 sft.

Came across quality of material being used for exterior, interior,

basement of houses. On the basis of these what is the price of house and how much they are in demand. With this we come to know that what is the preference of people living in US and how do they invest with what parameters matters for them. Floor quality is also one of the parameter which matters a lot while purchasing property where we seen that quality 572 for floor is having higher selling price . Ground living area where we seen that from 500 sft to 2000 sft are in demand as the price ranges from 2 lac to 4 lac. Even showing positive relation between price vs sft. Here people are preferring bigger space sometimes price doesn't matter for them instead space matters. Above we seen basement area and quality, but what type of bathroom do these basement should have and what is the trend going for these. Here we came across that basement with 2 full bathroom are having higher price followed by 1 bathroom. Definitely if the area covered will be bigger, higher the price fr the same which we seen here too that with 2 full bathroom the price is above 2 lac. And basement with half bathroom is below 1.75 lac as the Rea covered is less. Same for full bathroom inside the house we seen that with no of 3 price almost reached upto 3.5 lac and with half bathroom upto 2 lac. There are few people who doesn't wants to have bedroom above basement , that we also came across via insight that mostly bedroom which are in 0 that means not above basement are having higher price. Every house without kitchen is incomplete. Same as other things kitchen quality also matters. And definitely if a house being constricted with a particular material so most of the area of that house will be constructed using same material and that only we came across through our insight that everywhere Ex quality being considered superior so the price for the same is higher in every aspect. Now coming to Garage area which is now a days mandatory in every house construction as almost all the population in US do use their own vehicle to travel. Here again what area being used for garage means how many sft being used in garage construction, how many vehicles capacity do the garage have, what is the quality, condition , finishing of the garage is, accordingly the price being fixed for that and one of the important deciding factor to sale the house.

Coming to road that is adjacent to property is how good for driving purpose also matters. If the surface of road is smooth and flat price for that property is higher. House construction style differs from country to country as everywhere architect being used are not same. So what is the trend going on in that particular country where an investor wants to invest or wants to enter to real estate business should know so that accordingly they will come to know the customers choice and whether to invest or not will get clarity. Here in US we can see that some of the floor

being covered by wooden area which we can say is the trend there and the price also varies accordingly. As we seen that not fully but partial coverage of wooden area being preferred by people there.

Now coming to other areas which are in demand are open porch, enclosed porch, 3 Ssn porch, screen porch. These are all being used to enhance the beauty of house and to have sitting area either for guest or for themselves to have relax time . While screen porch mainly used to cover the windows or exterior of house so that no insect or any undesirable object could enter the house. As we know the beauty of house and the space matters a lot accordingly price and demand raises. Here we noticed the same Houses with these facilities having larger area are more costly. Pool area is one of the most relaxing and lavishing part of any house. So properties with no.555 pool area are more costly which we came across through our insight.

Now coming to the most important insight which is,in which year the sale is high and whether the trend is going upward or downward and what we came across is after 2007 the sale being decreased and become constant means no increase in sale after 2007. As we can clearly see that in 2008, 2009, 2010 sale is constant , neither increased nor decreased. Here we can assume that with the time rate of properties been diminished and not showing boom after 2007. Now coming to sale type and condition mostly seen that newly constructed houses are having higher price which is obvious as if the house is newly constructed demand for the same will be higher. Condition wise more preferred we seen is partial condition being sold at higher price. Which means houses with less requirement of renovation are in more demand. As investing in old property and again reinvesting in renovation will be more costly and not so much affordable for everyone.

Here with the above insight we can say that Real estate investor should have knowledge of what is the market trend going on and if the price for any property are higher then what are the features being considered, and what type of customer do invest. If customer are more in number who invests in such properties which are having high price and high valuable property then company should think to enter to that particular market and can start investing. But if the property which are for resale are not in much demand and to return that much of revenue which being expected are not being fulfilled, then investing in such property or market may turn loss for a company.

- # Motivation for the Problem

Undertaken :

To start this project my main object was to know the market trend for Real Estate business in US. Using previous dataset of properties being sold and purchased came across, what strategy to follow and whether the company for which I am doing this project i.e., Surprise Housing should invest and start business in US or not in real estate field. Also to predict the purchasing price of old houses which a company can buy at lowest price and can sell at higher price to have good revenue and profit . So as a data scientist with the help of machine learning algorithm built such a model which can predict whether to invest and if yes then how much and in which area, even what sort of customer are preferring what kind of properties. This way will able to sort down in which kind of property to invest and what customer base to concentrate. Also how much revenue a company can generate and what profit they can earn entering to a new market. While doing this project my main motivation was to know the US market trend for real estate business and as a data scientist we should explore in each domain what our data wants to say and how we can provide better solution to any business house.

ANALYTICAL PROBLEM FRAMING :

- ## Mathematical / Analytical Modeling of the Problem :

This dataset is all about Housing price prediction in US market. To start this project first I imported all the important libraries which were required for my analysis and model building. While going through thr dataset came across that my training dataset was consisting total no of rows 1168 and column 81 while test dataset was having 292 rows and 80 columns. Here in test data 1 column was less because it doesn't have target or label column. Then to have the information of how many

column do have datatypes object and how many are in integer or float format. Where I came across that train dataset do have total 43 object datatypes and 38 integer/float datatypes. Once we got the information about object columns now need to change these into integer or float so that our machine algorithm could understand these columns as machine doesn't understand string type of data. Used describe method to know statistical analysis of data so that we can get better insight for the same. With the help of describe method came across few of the points mentioned below :

1.There are missing values in some of the columns, which we have to treat as can't process data with missing values.

2. There were 1 Id column which is not required for our prediction so will going to drop this column.

3. By looking after mean and deviation statistical data we can say that there are skewness in most of the columns and we need to bring t to normal distribution that is within <30 which is considered as normal distribution or range between 0 -1

4. Quantile stats which means from min to max I.e, starting from minimum value of data to maximum value and in between there were more deviation I.e., 25%, 50% and 75% . Now the difference between these stats should not be much . If it's higher that means skewness or outliers are present in data which we do have in our dataset.

5. There are few columns which are discrete columns means though they are in float / integer but considered under categorical columns only.

With the help of describe method we can only have the analysis of continuous columns not for categorical column so to find whether there are any missing values present in categorical column or not we need to check through isna method.

And with the help of isna method came across that even in categorical column too there are missing values and as we know that for categorical column mainly we use mode to fix the missing values because we can't treat it mathematically or can have average. For continuous column we can use mean or median to treat the missing values but mean will be more effective than median as mean takes the average of all and then provide the value while median do take middle value of the data which we can't consider much for continuous columns.

Checked using Unique method whether any missing values are still

present in data or not and how many categories do each column have so that we can use visualisation technique accordingly to have better insight and more clarity.

- **Data Sources and their Formats :**

This dataset about housing price prediction taken from FilpRobo company in CSV format consist of both train and test data. Along with these two also got the case study about Real Estate business, what this business is about, how much it's contribution globally towards economy , what role do data scientist have to perform and what is the expectation from a data scientist by a company , all these been cleared in the case study provided in the form of PDF. Even data description about each column what exactly the terminology mentioned in column been described in text format and what to mention in PDF file for that a word sample documentation been provided by the company. Snapshot of dataset what exactly this dataset is all about been given below:

Train Dataset :

```
In [4]: df_train.head()
```

Out[4]:

F	OpenPorchSF	EnclosedPorch	3SsnPorch	ScreenPorch	PoolArea	PoolQC	Fence	MiscFeature	MiscVal	MoSold	YrSold	SaleType	SaleCondition	SalePrice
0	205	0	0	0	0	NaN	NaN	NaN	0	2	2007	WD	Normal	128000
1	207	0	0	224	0	NaN	NaN	NaN	0	10	2007	WD	Normal	268000
0	130	0	0	0	0	NaN	NaN	NaN	0	6	2007	WD	Normal	269790
0	122	0	0	0	0	NaN	MnPrv	NaN	0	1	2010	COD	Normal	190000
0	0	0	0	0	0	NaN	NaN	NaN	0	6	2009	WD	Normal	215000

```
In [6]: df_test.head()
```

Out[6]:

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	LotConfig	LandSlope	Neighborhood	Condition1	Conc
0	337	20	RL	86.0	14157	Pave	NaN	IR1	HLS	AllPub	Corner	Gtl	StoneBr	Norm	
1	1018	120	RL	NaN	5814	Pave	NaN	IR1	Lvl	AllPub	CulDSac	Gtl	StoneBr	Norm	
2	929	20	RL	NaN	11838	Pave	NaN	Reg	Lvl	AllPub	Inside	Gtl	CollgCr	Norm	
3	1148	70	RL	75.0	12000	Pave	NaN	Reg	Bnk	AllPub	Inside	Gtl	Crawfor	Norm	
4	1227	60	RL	86.0	14598	Pave	NaN	IR1	Lvl	AllPub	CulDSac	Gtl	Somerst	Feedr	

Just mentioned above the glimpse of dataset for both train and test.

- **Data Preprocessing Done :**

While analysing the data we came across that there were missing values

present in most of the columns either it be categorical or continous. We can opt drop option too to fix these missing data but dropping all such rows having missing value may lead us to loose some of the information which may affect our model performance, as we know that our model gets trained on previous dataset provided and if some of the information get deleted, there may be a chance that model may not get better training on data provided which may affect our prediction result. For this reason usually we don't apply drop unless and until the missing values are more than 50% in any of the particular row. Here intros dataset get some columns which were having more than 50% missing values and even more than 90% too , so treating these column may not give us a proper information as lot many data are missing from that particular rows. So I deleted 5 columns which were :

1. Allay which were almost having 1091 rows information missing and our dataset total was only 1168 . That means more than 90% data was missing from this column.
2. Fireplace QU which were also having more that 50% data missing.
3. Pool QC
4. Fence
5. MiscFeature
6. ID - this column was not required for our prediction purpose so dropped the same.

All the columns mentioned above were having missing information in both train and test dataset more than 50% and even 90% so deleted these columns to avoid multicollinearity problem.

As we know that we need to predict for test data too after doing all the pre-processing and data cleaning steps. So here along with train data applied data cleaning on test data too.

After deleting the columns mentioned above used mean method on all the continuous columns assuming that it will take mean of nearest rows and will place value in null rows as we don't know the exact information what to put there so for safer side taken the mean of all the continuous column. Used Mode method for categorical column assuming that mode usually consider that which being repeated more no of times so here that value will be placed which being used more no of time in other rows. This way using mean and mode fixed all the missing values.

Next used Label encoder to convert all the object form to integer or float format so that machine algorithm could understand and can predict in a better way by understanding each and every information gathered via dataset. One hot encoder being not used as no of columns increases as many times as no of categories are present in column. So just to avoid having bigger data use label encoder.

- **Data Inputs - Logic - Output Relationships :**

This dataset is all about prediction of price for houses being purchased at a very low price and to resale the same at higher prices for business purpose. Here a company wants to know the market strategy
to be followed to get success in real estate business, what revenue and profit they can earn from the mentioned market . If anyone who wants to enter to a new market they should have the knowledge of what trend being followed by that particular market and what is the demand of customer, what type of customer available etc. In this dataset we almost have 81 columns in which 80 columns were independent variable and one column was dependent variable i.e., sales price. Which means this column is dependent on all the other 80 columns . So here logic behind this ,was to have the insight of all information being gathered via dataset of sale of houses provided by Australian company. Now the trend we seen here is sale for those properties are higher which is not very old and the condition of the house should not be such bad else demand for that is low either considering pool garage, kitchen, basement, floor, exterior, interior etc. Price being varied as per the preference and demand of customer.
Considering all the factor we can say that demand for those houses are higher which is in good condition and not much old , also have other facilities such as road connectivity, how wide the road is, how's
the neighbours there, and other facilities which may enhance the beauty of house.

Models Development and Evaluation

- **Identification of Possible Problem -**

Solving Approaches :

Approaches that I followed to solve this problem analytically was first to come across what the shape of my data,means how many rows and column do my data consist. After knowing the shape came across if any missing values are there and if yes then in how many columns and what amount of missing values are present. Then observed how many columns are categorical and how many are continuous columns, among integer datatype there were few discrete columns which may be considered as categorical only though the data are having numbers but they are fixed in nature means nothing can come between those numbers. Continuous columns means they can have data which are continue in nature means they can be either in float or integer doesn't matter. Once identified this, used label encoder after replacing all the missing values using mean and mode for continuous and categorical column respectively. Also deleted few of the columns which were having missing values more than 50%.
Later coming to statistical analysis we came across that there were most of the columns which were skewed and having outliers. And as we know that we treat skewness and outliers for continuous columns only not for categorical. Skewed data are those which are not symmetrical in nature means not normally distributed which usually ranges between 0 -1. And for skewness we generally consider those which are above +/-0.5. Using df.skew found skewed columns and treated the same using log1p which is considered to be near to 0 and good for our model performance. Removed skewness from both train and test data. Later found outliers in both the train and test dataset and removed the same using z-score method. Outliers are generally those data point which are unusual compare to other data points and are out of the box. So for better performance of our model we have to remove these datapoints which are not in range.
Later scaled all the feature columns to bring them to same units so that our model can understand each and every column and can predict more appropriately. Scaled only training dataset as for test data we just need to predict after all the process followed above from data cleaning to feature engineering.

- # Testing of Identified Approaches :

After scaling all the feature now we have to train our model on the

dataset provided using train test split process. For which we require few of the algorithm with the help of which our model will be trained
and will predict the desired output. Algorithms which I used for training and prediction purpose summarised below:

1. Train Test Split
2. Decision Tree Regressor
3. Linear Regression
4. Random Forest Regressor
5. Ridge, Lassor, LassoCV, RidgeCV

Metrices used are :

1. R2 Score

Validation score used are :

Cross Validation Score

Hyperparameter Tuning used are :

Grid Search CV

With the help of these algorithm my model able to predict the desired output.

Here I used different algorithm mentioned above that is decision treee, linear regression , random forest regressor and ridge , lasso to get the prediction and also to see on which algorithm do my model performs best .

Train test split being used for training and testing purpose means here I splitted on what percentage my model will get trained and what percentage do model get tested. Provided 70% for training purpose and 30% for testing purpose . How well the model get trained , the prediction will be that much accurate.

After this used metrics known as R2 score as this is my regression problem so will use R2 score for prediction and later on used CV score so that can validate which model perform better by getting the
difference between r2 score and CV score. Though the r2 score for any model be high, but we will consider best model those whose difference between R2 and CV score will be less.

And at last did hyperparameter tuning using GridSearch CV to improve more accuracy score of model.

- **Run and Evaluate Selected Models :**

1st found the best random state using decision tree regressor and r2 score. Using best random state we can train and test our model .

```
maxscore = 0
maxrs = 0

for i in range(1,1000):
    x_train,x_test,y_train,y_test = train_test_split(x_scaler,y,test_size = 0.30,random_state = i)
    dt = DecisionTreeRegressor()
    dt.fit(x_train,y_train)
    pred = dt.predict(x_test)
    rsc = r2_score(y_test,pred)
    if rsc>maxscore:
        maxscore=rsc
        maxrs=i
print("Best r2 score is:",maxscore,"On Random state: ",maxrs)
```

```
Best r2 score is: 0.8279445905614323 On Random state:  215
```

Best random state is 215 on which will train and test the model.

```
x_train,x_test,y_train,y_test=train_test_split(x_scaler,y,test_size=0.30,random_state=i)
```

Train test split done using random state that we got after finding best random state and splitted training data to 70% and test data to 30%

```
lr = LinearRegression()
lr.fit(x_train,y_train)
pred = lr.predict(x_test)
print(r2_score(y_test,pred))
```

```
0.8605602894869785
```

```
print(cross_val_score(lr,x_scaler,y,cv=5).mean())
```

```
0.8429006614677291
```

Using Linear Regression got r2 score as 86% and CV score 84% . So the difference between these two are approximately 2%

```
dt = DecisionTreeRegressor()
dt.fit(x_train,y_train)
pred = dt.predict(x_test)
print(r2_score(y_test,pred))
```

```
0.6097497649114122
```

```
print(cross_val_score(dt,x_scaler,y,cv=5).mean())
```

```
0.7142887797906264
```

Here clearly we can see that decision tree algorithm is not so good performing on this dataset. Both the R2 score and CV score compare to linear regression are very low. So we will not consider this for prediction further.

```
rf = RandomForestRegressor()
rf.fit(x_train,y_train)
pred = rf.predict(x_test)
print(r2_score(y_test,pred))
```

```
0.8852057661311514
```

```
print(cross_val_score(rf,x_scaler,y,cv=5).mean())
```

```
0.8694358298573228
```

As far as in performance we can see that Random forest is the best performing model with highest R2 and CV score which 88% and 86% respectively . Though the difference between linear and Random is not much still Random forest being considered as best performing model .

Using Lasso and Ridge checking whether model is overfitted or not

```
lassocv = LassoCV(alphas = None, max_iter=1000, normalize = True)
lassocv.fit(x_train,y_train)
```

```
LassoCV(normalize=True)
```

```
alpha = lassocv.alpha_
alpha
```

```
32.192325622877476
```

```
lasso_reg = Lasso(alpha)
lasso_reg.fit(x_train,y_train)
```

```
Lasso(alpha=32.192325622877476)
```

```
lasso_reg.score(x_test,y_test)
```

```
0.862425562607965
```

```
ridgecv = RidgeCV(alphas =(0.1,1.0,10.0),normalize = True)
ridgecv.fit(x_train,y_train)
```

```
RidgeCV(alphas=array([ 0.1,  1. , 10. ]), normalize=True)
```

```
alpha = ridgecv.alpha_
alpha
```

```
0.1
```

```
ridge_reg = Ridge(alpha)
ridge_reg.fit(x_train,y_train)
```

```
Ridge(alpha=0.1)
```

```
ridge_reg.score(x_test,y_test)
```

```
0.8606079607931862
```

Both lasso and Ridge score we can see that almost same as Linear regression score that is 86% approximately. With this came across conclusion that our model is not na overfitting model . Overfitting model means no duplicate or multicollinearity present in dataset.

After getting the best performing model which is Random Forest Regressor did hyperparameter tuning using Grid Search CV to get more accuracy and to increase the score.

```
# RandomForestRegressor
param = {'n_estimators':[300,500,700],
         'criterion':['mse','mae'],
         'max_depth':[2,4,6,8,10],
         'max_features':['auto','sqrt','log2']}
```

Given all the required parameters on which our model will get trained for random forest regressor. Here I used 4 parameters which were n_estmators, Criterion, Max_depth and max_features. Accordingly will fit to my model.

```
GC = GridSearchCV(rf,param,cv=5)
```

Total 5 cross validation I used here for random forest regressor. No of cross validation be more the higher the chances to get more score. But as it will take more time to run so I had taken here 5.

Trained the model using fit. Method in Grid search CV.

```
GC.fit(x_train,y_train)
```

```
GridSearchCV(cv=5, estimator=RandomForestRegressor(),
             param_grid={'criterion': ['mse', 'mae'],
                         'max_depth': [2, 4, 6, 8, 10],
                         'max_features': ['auto', 'sqrt', 'log2'],
                         'n_estimators': [300, 500, 700]})
```

Now finding the best parameters which we got after training fitting my model.

```
GC.best_params_
```

```
{'criterion': 'mae',
 'max_depth': 10,
 'max_features': 'auto',
 'n_estimators': 500}
```

So here we got the best parameters which we need to feed to our best performing model so that will get some improved score using this.

```
final_rfc = RandomForestRegressor(criterion='mae',max_depth = 10, max_features = 'auto',n_estimators = 500)
final_rfc.fit(x_train,y_train)
pred = final_rfc.predict(x_test)
score = r2_score(y_test,pred)
print(score*100)
```

```
88.9872351155201
```

Here I got only 1% improvement in score using Hyperparameter tuning. This may be because of I used only 4 parameters. There might be a chance of improving more by using more parameters.

- **Key Metrics for success in Solving problem under consideration :**

This dataset is all about Housing price prediction in US and as we know that, we need to predict on the basis of target or dependent column whether to invest in that particular business or not. Here our target column was continuous in nature which means we need to solve using Regression models as this problem is a Regression bases problem. After using pandas library for getting the dataset on which I was suppose to work analysed what were the techniques and further process to proceed so that our model can perform better and can give more accurate result. Once gone through the dataset came across that there were missing data in few of the columns. Also almost half of the columns were having object data type. Object data type is difficult for machine to understand so we need to convert those to integer form either using label encoder or one hot encoder whichever required. Skewness and outliers were also present in few of the continuous columns , we don't treat skewness and outliers for categorical columns. Removed the same for our model to understand and perform for best prediction with higher score. Due to multicollinearity issue and and as few of the columns were having null values more than 50%, dropped it. After data cleaning and feature engineering for both train and test dataset, scaled the train data feature using standard scaler so that all the units come to same label. After all this found best random state and using 5 algorithm checked which algorithm perform best for this dataset and and more score. What we found is Random Forest classifier was giving highest score compare to others , didi hyperparameter tuning using Grid Search CV for Random Forest Classifier to improve the score if possible. Finally saved the model and reloaded to check the prediction on test dataset. These were some of the important factors mentioned above, which helped me in

solving this problem. As per my knowledge data cleaning and feature engineering played an important role if this would not have been done properly may be my model would not come with a score of approx 89%. Even skewness and outliers which if presented in data will not allow model to perform in a better way so it's important to remove these using proper technique. Visualisation presentation done in ppt with detail analysis.

- **Interpretation of the Results :**

Using visualisations came across insight of dataset, what is the trend followed in market, what is the choice and demand of customer, on what basis the price of property gets higher and what are the parameters customers usually prefer at higher prices they purchase the property.Even investing in such market, a company will get that much of revenue which they expected along with good profit, as entering to new market will always be a risk , so before investing any company should first analyse and gather all the necessary information required to step to that particular market. In which visualisation technique helps in showing glimpse about market trend and helps in building strategy that a business person should follow to get success.with preprocessing cleaned data by analysing what were the steps required to have more cleaned data. And once data cleaning and feature engineering done trained model to give higher score with more accurate prediction.

- **Conclusion :**

Everybody around the globe wants to have their own house, so the real estate business is one of demanding and high generating revenue towards the economy of country. And as more demanding and high generating revenue business lot many companies are in this domain. Competition is too high due to lot many competitors are there in the market. Companies uses data scientist to have knowledge of trend going on in market, what strategy to follow, how to increase overall revenue and earn good profit, what is the customer preference and how prices varies with the time. These are the few factor which every company focuses to establish as a successful real estate

business. Here data scientist by using their domain knowledge can predict whether investing in that particular market is profitable or not.We are doing project for one of the real estate company I.e., surprise housing who wants to enter to US country to establish their real estate business, and for this they gathered data from sale of houses in Australia , with the help of which they can come to conclusion what trend going on in market by looking after the previous sales information regarding other properties in that particular country. By looking after the previous data came across that property rate varies to lot many features , such as exterior and interior quality of house, condition of house, how old the property is, condition and qualityof over all extra facilities such as garage, porch, pool, fire , electricity, heating, what kind of road and how wide is the road adjacent to property, floor wise also rate varies, what sft being used as wooden floor, how much area been used as open space, etc. If any customer do have floor preference there also price varies. What sft been preferred by the customer mostly , and whether price for those area are same whenever sold. All these factors were important while selling the property. Also we came across that what is the demand of customer by looking after the price of property being sold mostly. As we know that buying property is one of the important and sentimental decision for most population so they go through every aspect to have the best for themselves. As a property consultant company should always consider what a customer demand and what is their budget, and if both are eligible then how to convince them to purchase from a particular agent. With the help of machine learning we came across that this data was having lot many missing values, skewness and outliers which we were suppose to treat. Now as a data scientist treated all the missing values using mean and mode for continuous and categorical column respectively. Later done feature engineering as there were few column which was in object form. Using label encoding converted all the object data type to integer form. After fixing everything trained model to have prediction using 5 algorithm , just to check which algorithm performs best. In which Random Forest classifier were the best performing model with highest score of 88% . Later did hyperparameter tuning to improve more accuracy, 1% increment being observed in score.

- **Learning Outcomes of the Study in Respect of Data Science :**

While doing this project came across that this dataset was not a cleaned dataset and lot many work needed to be done on this. First came across that both in train and test dataset there were missing values in most of the columns. And as almost half of the columns were categorical one so needed to treat the same accordingly. Where I preferred mode for categorical column, as per my knowledge mode can replace nan using those data which being repeated more number of times. And for continuous column used mean , mean I used because mean takes the average of all the values and replace nan , While median could also be used as there were few discrete columns but as we know that median returns the middle value which may not be accurate sometimes and model performance may be affected. Next came across that most of the columns were in object data type which were suppose to be converted to integer form otherwise our machine algorithm will not understand the string / object form. So used label encoder to convert all the object form to integer form. Instead we could have used One hot encoder as there were lot many categorical columns but I didn't used that as it increases the number of columns which may turn my data to be too big. So to avoid huge number of columns used label encoder. After doing data cleaning and feature engineering removed skewness and outliers present in few of the continuous columns. Then choosen few of the regression based algorithm to make my model trained and predict the score. And what I observed is Random Forest regression was the best model among all as we know that Random Forest group all the decision tree due to which chances are more to have higher score in this algorithm. Atlast did hyperparameter tuning to increase the score and only 1% increased this may be because of used only 4 parameters in grid search cv , if would have used more, chances to get more improved score might be higher as model would have got more parameters to learn and predict on the basis of that. Then saved the model and later on loaded to get the friction on test dataset.

- **Limitations of this Work And Scope for Future Work :**

Most of the informations were provided by the dataset gathered from sale of House from Australia. So I don't think that there were any problem for the solution required. Future scope I can say is as Real estate is one of the leading industry in business world so having knowledge in this domain and solving issue for this may result in having good knowledge of what type of strategy do these type of industry follows and what sort of customer this markets having. May be some of the columns what I felt was having multicollinearity chances would have been improved by using more techniques and using more parameters for hype parameter tuning would have improved my score. At the end just want to add on that we can even use Random Search hyperparameter tuning and XG Boost to have more improved score which may result in better prediction and better model performance, And what sort of customer this market having. May be some of the columns what I felt was having multicollinearity chances would have been improved by using more techniques and using more parameters for hyper parameter tuning would have improved my score. At the end just want to add on that we can even use Random Search hyperparameter tuning and XG Boost to have more improved score which may result in better prediction and better model performance.

Real estate is one of the leading industry in business world so having knowledge in this domain and solving issue for this may result in having good knowledge of what type of strategy do these type of industry follows and what sort of customer these markets having. May be some of the columns what I felt was having multicollinearity chances would have been improved by using more techniques and using more parameters for hype parameter tuning would have improved my score. At the end just want to add on that we can even use Random Search hyperparameter tuning and XG Boost to have more

improved score which may result in better prediction and better model performance.