



## **HOUSE - SALE PRICE PREDICTION**

SUBMITTED BY:

*SHIKHA KAMAL*

## **ACKNOWLEDGMENT**

This includes mentioning of all the references, research papers, data sources, professionals and other resources that helped you and guided you in completion of the project.

# **INTRODUCTION**

- **Business Problem Framing**

As the peoples are searching house to live due to many reasons, some are transferred to any other places, some are shifting to another cities for some reasons, some are renting house. These are a several reasons for which a person searching the houses but sometimes are unaware to many environmental factors and having histories, etc.

So this project includes various factors ,including various varities of house details so that one can predict their price and know the general factors.

- **Conceptual Background of the Domain Problem**

This project mostly included all the factors of house that must be included for every detailing. Houses are one of the necessary need of each and every person around the globe and therefore housing and real estate market is one of the markets which is one of the major contributors in the world's economy.

- **Review of Literature**

In this project all the data are being provided in datasets which we execute in jupyter and analyse the data and fit into multiple models. The various examine of datasets are being done in analysing and various visualization is also done to have visual observation. The models are being trying to fitted with all data values.

- **Motivation for the Problem Undertaken**

Data Science comes as a very important tool to solve problems in the domain to help the companies increase their overall revenue, profits, improving their marketing strategies and focusing on changing trends in house sales and

purchases. Predictive modelling , Market mix modelling , recommendation systems are some of the machine learning techniques used for achieving the business goals for housing companies.

We require to build a model using Machine Learning in order to predict the actual value of the prospective properties and decide whether to invest in them or not.

## **Analytical Problem Framing**

- **Mathematical/ Analytical Modeling of the Problem**

The data values are being executed in DataFrame and then various kind of functions are being run to have an overlook of datasets and their values. Like head(), tail(), sample(),info(), describe(), shape(), isnull() various functions used .

- The head function by default execute the first 5 rows in datasets
- The tail function by default execute the last 5 rows in datasets
- The sample function by default randomly pick any row with all columns
- The info function execute the basic kind of detailing of datasets execute with number of rows, columns, null values and data type
- The describe function executes the statistical mathematical calculation of all related rows and columns, with their mean, standard deviation, min value, max value, and different stages percentile
- The shape function execute the number of rows and columns
- The is null function executes any null values in datasets

- **Data Sources and their formats**

While analysing the data , one function of correlation is used, the correlated values execute the every related row or column and find their correlated values by which they can inter-relate.

The data value which is highly lighter in shade predict the most correlated value, where we can find necessities factors or columns to have major role in predicting the target variable.

- **Data Preprocessing Done**

For data cleaning, we can use zscore , power transform method also.

First we use Label Encoder that convert the string data type into numeric data type, and imputer used to fill the nan values present in the datasets.

Secondly, we check for the outliers , which means data values present outside the zscore value which generally lie between -3 to +3 in normally distributed graph. This is the most data cleaning process used only for numeric data type, here in this house project datasets we have outliers but they are in string data types , so it can be ignored. We don't need to remove the outliers from string/categorical data type values or from target variable column

Thirdly, we check for the skewness , which means the spreading of data values is much wide which go beyond normally distributed . the range lie in -3 to +3 ,which is best fit . the same reason apply for skewness, it is not removed for string/ categorical data type and target variable , only removed in numeric data type.

This is the major cleaning process done and check for every project and removed process also applied.

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

Here , in this we can visualize many datasets values by plotting variety of graph according to their data type also. The plotting can be done univariant, bivariant, trivariant.

The univariant is plotting the graph by considering the one by one column and observe their plotting points.

The bivariate plotting can consider by taking 2 variables and comparing their relation with each other. We can compare the target variable with each column and compare the columns relation with target variable.

The trivariate plotting can consider taking 3 variable or 4 variable simultaneously to plot the graph and compare their correlation with each other. In this categorical variable should be used is necessary.

- State the set of assumptions (if any) related to the problem under consideration

The zscore and skewness is not removed as the data type is generally object type which contains the string and categorical data values.

## **Model/s Development and Evaluation**

- Identification of possible problem-solving approaches (methods)

In this house project, we need to predict the sale price with including so many factors mentioned. So first check for data types whether it is continuous or categorical data type. Here sale price is continuous data type so accordingly we select the models to training process and so on.

- Testing of Identified Approaches (Algorithms)

As here , the data type is continuous as target variable, there are many models which can fit in continuous data type like :

- Linear Regression
- Decision Tree Classifier

- Random Forest Regressor
- Support vector machines
- Lasso and Ridge Regressor
- KNN models
- Etc

The continuous data type used the suffix as regressor whereas categorical data type used the suffix as classifier.

## • Run and Evaluate selected models

I run all models in one loop and have their prediction at one go.

Firstly , we split the datasets into x and y variable dataframe , then we fit the x in model and trained the data .

Secondly, we need to find the best random state for which a separate loop is run where we can execute the max random state with their max percent.

Thirdly, we run all models in one loop and execute the output with score, how fit the model is trained

```

1167      183200
Name: SalePrice, Length: 1168, dtype: int64

checking for max random state

In [283]: maxAcc=0
          maxRS=0

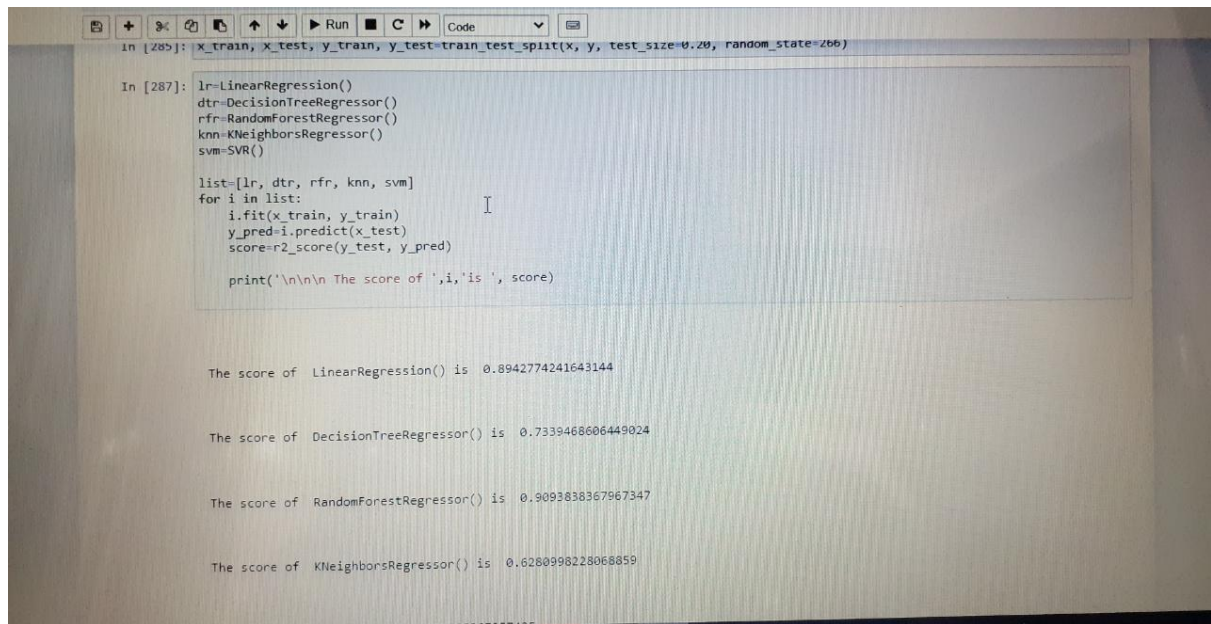
          for i in range(1,300):
              x_train, x_test, y_train, y_test=train_test_split(x, y, test_size=0.20, random_state=i)
              lr=LinearRegression()
              lr.fit(x_train, y_train)
              y_pred=lr.predict(x_test)
              score=r2_score(y_test, y_pred)
              print('score is ',score,' random state is ',i)

          if score>maxAcc:
              maxAcc=score
              maxRS=i
              print('max score is ', maxAcc,' at random state ',maxRS)

score is  0.8124524945291326  random state is  1
max score is  0.8124524945291326  at random state  1
score is  0.7444292513679549  random state is  2
score is  0.8619551425382294  random state is  3
max score is  0.8619551425382294  at random state  3
score is  0.5672562307767804  random state is  4
score is  0.5877247234179062  random state is  5
score is  0.8431624000146117  random state is  6
score is  0.7558331036446564  random state is  7
score is  0.8547837933686167  random state is  8
score is  0.8428329609811486  random state is  9
score is  0.7493840633202011  random state is  10
score is  0.15566163728898952  random state is  11
score is  0.8242965295021343  random state is  12

```

The above image is finding the max random state .



```
In [285]: x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.20, random_state=266)

In [287]: lr=LinearRegression()
          dtr=DecisionTreeRegressor()
          rfr=RandomForestRegressor()
          knn=KNeighborsRegressor()
          svm=SVR()

          list=[lr, dtr, rfr, knn, svm]
          for i in list:
              i.fit(x_train, y_train)
              y_pred=i.predict(x_test)
              score=r2_score(y_test, y_pred)

              print('\n\n The score of ',i,' is ', score)

The score of LinearRegression() is 0.8942774241643144

The score of DecisionTreeRegressor() is 0.7339468606449024

The score of RandomForestRegressor() is 0.9093838367967347

The score of KNeighborsRegressor() is 0.6280998228068859
```

The above image is showing the all 5 models run in one loop.

- Key Metrics for success in solving problem under consideration

Here the target variable is continuous data type , so only we need to calculate the r2 score as a metrics. The library used is metrics importing r2 score.

R2 score is only calculating the accuracy percentage of how much the data value is fitted in the particular model. The r2 score as calculated, we pick any best model for prediction.

- Visualizations

Here plotting is done , by having various graph in order to have visual observation. We use the count plot , scatter plot and outliers graph and distribution graph.

Count plot is generally used for any variable type numeric or categorical type. But best for categorical type, it made counts of every category .



Scatter plot is plotting dotting points , wherever density is high that region contains more relatable data range.

Outliers plotting is detecting any points which lie outside the plot box, and it is removed by using zscore method.

Distribution plot is detecting spreading of data values in columns , and it is removed by power transform method, boxcox method, log transform method.

- **Interpretation of the Results**

Here we do visualization, model training, in which we observe that there is presence of outliers and skewness but we don't remove it, as most of the columns are object type which contains the string or categorical data value.

After having fitting the model , the cross validation score is cross check to test the training model, whether the model is giving accuracy model score or not.

After cross validation, random forest regressor model is having the most efficient score compare to all other models.

So we continue with random forest and we search for grid search cv and again we cross verify for random forest.

After fitting the model in grid search cv , we find for the best parameters for random forest and again fit the model in random forest and having the prediction and execute the accuracy score.

## **CONCLUSION**

- **Key Findings and Conclusions of the Study**

First we need to search what the problem statement is and where to proceed and have prediction. We search for target variable and then observe its data type , hence it come continuous variable column.

Then we need to select the models for continuous target variable to have prediction and best fitted model to proceed for best results.

- **Learning Outcomes of the Study in respect of Data Science**

For analysing the datasets, visualization of datasets is compulsory to have and plotted in graph. There are various kinds of graph like count plot, scatter plot, distribution plot, violin plot, box plot, etc.

Count plot used for categorical type, scatter plot generally used to find the relatable columns relation, box plot used to find outliers and distribution plot used to check for skewness.

Violin plot used to have density or spreading of data at particular region. Many more graphs are also there which can be used and can predict.