FLIP ROBO

Housing Price Prediction

Submitted By

HariharaSudhan

Acknowledgement

I would like to express my deepest gratitude to my SME (Subject Matter Expert) Shwetank Mishra as well as Flip Robo Technologies who gave me the opportunity to do this project on Surprise Housing Price Prediction, which also helped me in doing lots of research wherein I came to know about so many new things.

Also, I have utilized a few external resources that helped me to complete the project.

INTRODUCTION

Business Problem Framing

Houses are one of the necessary needs 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. It is a very large market and there are various companies working in the domain.

Conceptual BackGround of Domain Problem

A US-based housing company named Surprise Housing has decided to enter the Australian market. The company uses data analytics to purchase houses at a price below their actual values and flip them at a higher price. For the same purpose, the company has collected a data set from the sale of houses in Australia. The data is provided in the CSV file below.

Review of Literature

Based on the sample data provided to us from our client database where we have understood that the company is looking at prospective properties to buy houses to enter the market. The data set

explains it is a regression problem as we need 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. Also, we have other independent features that would help to decide which all variables are important to predict the price of the variable and how do these variables describe the price of the house.

Motivation for the Problem Undertaken

Our main objective of doing this project is to build a model to predict the house prices with the help of other supporting features. We are going to predict by using Machine Learning algorithms.

The sample data is provided to us from our client database. In order to improve the selection of customers, the client wants some predictions that could help them in further investment and improvement in selection of customers.

House Price Index is commonly used to estimate the changes in housing price. Since housing price is strongly correlated to other factors such as location, area, population, it requires other information apart from HPI to predict individual housing price.

As a result, to explore various impacts of features on prediction methods, this paper will apply both traditional and advanced machine learning approaches to investigate the difference among several advanced models. This paper will also comprehensively validate multiple techniques in model implementation on regression and provide an optimistic result for housing price prediction.

Analytical Problem Framing

Mathematical/ Analytical Modeling of the Problem

We are building a model in Machine Learning to predict the actual value of the prospective properties and decide whether to invest in them or not. So, this model will help us to determine which variables are important to predict the price of variables & also how do these variables describe the price of the house. This will help to determine the price of houses with the available independent variables. They can accordingly manipulate the strategy of the firm and concentrate on areas that will yield high returns. Regression analysis is a set of statistical processes for estimating the relationships between a dependent variable (often called the 'outcome variable') and one or more independent variables (often called 'predictors',

'covariates', or 'features'). The most common form of regression analysis is linear regression, in which one finds the line (or a more complex linear combination) that most closely fits the data according to a specific mathematical criterion. For specific mathematical reasons this allows the researcher to estimate the conditional expectation of the dependent variable when the independent variables take on a given set of values.
Regression analysis is also a form of predictive modelling technique which investigates the relationship between a dependent (target) and independent variable (predictor). This technique is used for forecasting, time series modelling and finding the causal effect relationship between the variables.

Data Sources and their formats

Data set provided by Flip Robo was in the format of CSV (Comma Separated Values). The dimension of data is 1168 rows and 81 columns. There are 2 data sets that are given. One is training data and one is testing data.

1. Train file will be used for training the model, i.e., the model will learn from this file. It contains all the independent variables and the target variable. Size of training set: 1168 records.

2. Test file contains all the independent variables, but not the target variable. We will apply the model to predict the target variable for the test data. Size of test set: 292 records.

Data Preprocessing Done

Data pre-processing in Machine Learning refers to the technique of preparing (cleaning and organizing) the raw data to make it suitable for a building and training Machine Learning models. In other words, whenever the data is gathered from different sources it is collected in raw format which is not feasible for the analysis. Data pre-processing is an integral step in Machine Learning as the quality of data and the useful information that can be derived from it directly affects the ability of our model to learn; therefore, it is extremely important that we pre- process our data before feeding it into our model. Therefore, it is the first and crucial step while creating a machine learning model. I have used some following pre-processing steps:

- a. Loading the training dataset as a dataframe.
- b. Used pandas to set display I ensuring we do not see any truncated information
- c. Checked the number of rows and columns present in our training dataset
- d. Checked for missing data and the number of rows with null values
- e. Verified the percentage of missing data in each column and decided to discard the one's that have more than 50% of null values
- f. Dropped all the unwanted columns and duplicate data present in our dataframe
- g. Separated categorical column names and numeric column names in separate list variables for ease in visualization
- h. Checked the unique values information in each column to get a gist for categorical data

- Performed imputation to fill missing data using mean on numeric data and mode for categorical data column
- j. Used Pandas Profiling during the visualization phase along with pie plot, count plot, scatter plot and the others
- k. With the help of ordinal encoding technique converted all object datatype columns to numeric datatype
- l. Thoroughly checked for outliers and skewness information
- m. With the help of heatmap, correlation bar graph was able to understand the Feature vs Label relativity and insights on multicollinearity amongst the feature columns
- n. Separated feature and label data to ensure feature scaling is performed avoiding any kind of biasness
- o. Checked for the best random state to be used on our Regression Machine Learning model pertaining to the feature importance details
- p. Finally created a regression model function along with evaluation metrics to pass through various model formats

• Data Inputs- Logic- Output Relationships

When we loaded the training dataset, we had to go through various data pre processing steps to understand what was given to us and what we were expected to predict for the project. When it comes to logical part the domain expertise of understanding how real estate works and how we are supposed to cater to the customers came in handy to train the model with the modified input data. In Data Science community there is a saying "Garbage In Garbage Out" therefore we had to be very cautious and spent almost 80% of our project building time in understanding each and every aspect of the data how they were related to each other as well as our target label.

With the objective of predicting hosing sale prices accurately we had to make sure that a model was built that understood the customer priorities trending in the market imposing those norms when a relevant price tag was generated. I tried my best to retain as much data possible that was collected but I feel discarding columns that had lots of missing data was good. I did not want to impute data and then cause a biasness in the machine learning model from values that did not come from real people.

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

The assumption part for me was relying strictly on the data provided to me and taking into consideration that the separate training and testing datasets were obtained from real people surveyed for their preferences and how reasonable a price for a house with various features inclining to them were.

• Hardware and Software Requirements and Tools Used

Hardware Used:

i i. RAM: 4 GB

ii ii. CPU: Intel(R) Core(Duo) i2

Software Used:

i i. Programming language: Python

ii ii. Distribution: Anaconda Navigator

iii iii. Browser based language shell: Jupyter Notebook

Libraries/Packages Used:
Pandas, NumPy, matplotlib, seaborn and scikit-learn

Model/s Development and Evaluation

- Identification of possible problem-solving approaches (methods)

I have used both statistical and analytical approaches to solve the problem which mainly includes the pre-processing of the data and EDA to check the correlation of independent and dependent features. Also, before building the model, I made sure that the input data is cleaned and scaled before it was fed into the machine learning models.

For this project we need to predict the sale price of houses, means our target column is continuous so this is a regression problem. I have used various regression algorithms and tested for the prediction. By doing various evaluations I have selected Extra Trees Regressor as best suitable algorithm for our final model as it is giving good r2-score and least difference in r2-score and CV-score among all the algorithms used. Other regression algorithms are also giving me good accuracy but some are over-fitting and some are with under-fitting the results which may be because of less amount of data.

In order to get good performance as well as accuracy and to check my model from over-fitting and under-fitting I have made use of the K-Fold cross validation and then hyper parameter tuned the final model.

Once I was able to get my desired final model I ensured to save that model before I loaded the testing data and started performing the data pre-processing as the training dataset and obtaining the predicted sale price values out of the Regression Machine Learning Model.

- Testing of Identified Approaches (Algorithms)

The algorithms used on training and test data are as follows:

➢ Ridge Regularization Regression Model
➢ Lasso Regularization Regression Model
➢ Decision Tree Regression Model
➢ Random Forest Regression Model
➢ Gradient Boosting Regression Model

Ridge Regression

```
In [126]: ridge.score(X,y_train)
Out[126]: 0.905595593171787

In [127]: ridge.score(X_test[X.columns],y_test)
Out[127]: 0.8754334796131733
```

Lasso Regression

```
('SaleCondition_Normal', 36871.961),
('SaleCondition_Partial', 42641.13)]

In [140]: print('r2_score for ridge:')
          print('Train dataset:', round(r2_score(y_train, y_pred_ridge_train), 4))
          print('Test dataset:', round(r2_score(y_test, y_pred_ridge_test), 4))

          r2_score for ridge:
          Train dataset: 0.8468
          Test dataset: 0.8533
```

RandomForestRegrtession

```
print('R2_Score:',r2_score(y_test,predRF))

# Metric evaluation
print('MAE:',metrics.mean_absolute_error(y_test, predRF))
print('MSE:',metrics.mean_squared_error(y_test, predRF))
print("RMSE:",np.sqrt(metrics.mean_squared_error(y_test, predRF)))

Out[142]: RandomForestRegressor()

R2_Score: 0.8640978429974712
MAE: 16519.0512209302 3
MSE: 593122440.1956494
RMSE: 24354.10520211427
```

DecisionTreeRegression

```
# Metric evaluation
print('MAE:',metrics.mean_absolute_error(y_test, predDT))
print('MSE:',metrics.mean_squared_error(y_test, predDT))
print("RMSE:",np.sqrt(metrics.mean_squared_error(y_test, predDT)))

Out[143]: DecisionTreeRegressor()

R2_Score: 0.6111271780617631
MAE: 28057.43895348837
MSE: 1697170980.6598837
RMSE: 41196.735072817166
```

GradientBosstRegrtessor

```
predGRA=GRA.predict(X_test)
print('R2_Score:',r2_score(y_test,predGRA))

# Metric Evaluation
print('MAE:',metrics.mean_absolute_error(y_test, predGRA))
print('MSE:',metrics.mean_squared_error(y_test, predGRA))
print("RMSE:",np.sqrt(metrics.mean_squared_error(y_test, predGRA)))

Out[146]: GradientBoostingRegressor()

R2_Score: 0.8756403494767638
MAE: 16074.764356441401
MSE: 542747083.6893891
RMSE: 23296.93292451582
```

BaggingRegressor

```
In [148]: BAG=BaggingRegressor()
          BAG.fit(X_train,y_train)

          # prediction
          predBAG=BAG.predict(X_test)
          print('R2_Score:',r2_score(y_test,predBAG))

          # Metric Evaluation
          print('MAE:',metrics.mean_absolute_error(y_test, predBAG))
          print('MSE:',metrics.mean_squared_error(y_test, predBAG))
          print("RMSE:",np.sqrt(metrics.mean_squared_error(y_test, predBAG)))

Out[148]: BaggingRegressor()

          R2_Score: 0.802214839467612
          MAE: 19861.121220930232
          MSE: 863200552.7864826
          RMSE: 29380.27489296999
```

Key Metrics for success in solving problem under consideration

The key metrics used here were r2_score, cross_val_score, MAE, MSE and RMSE. We tried to find out the best parameters and also to increase our scores by using Hyperparameter Tuning and we will be using GridSearchCV method.

Cross Validation:

Cross-validation helps to find out the over fitting and under fitting of the model. In the cross validation the model is made to run on different subsets of the dataset which will get multiple measures of the model. If we take 5 folds, the data will be divided into 5 pieces where each part being 20% of full dataset. While running the Cross-validation the 1st part (20%) of the 5 parts will be kept out as a holdout set for validation and everything else is used for training data. This way we will get the first estimate of the model quality of the dataset.

In the similar way further iterations are made for the second 20% of the dataset is held as a holdout set and remaining 4 parts are used for training data during process. This way we will get the second estimate of the model quality of the dataset. These steps are repeated during the cross-validation process to get the remaining estimate of the model quality.

R2 Score:

It is a statistical measure that represents the goodness of fit of a regression model. The ideal value for r-square is 1. The closer the value of r-square to 1, the better is the model fitted.

Mean Squared Error (MSE):
MSE of an estimator (of a procedure for estimating an unobserved quantity) measures the average of the squares of the errors — that is, the average squared difference between the estimated values and what is estimated. MSE is a risk function, corresponding to the
expected value of the squared error loss. RMSE is the Root Mean Squared Error.

Mean Absolute Error (MAE):
MAE measures the average magnitude of the errors in a set of predictions, without considering their direction. It's the average over the test sample of the absolute differences between prediction and actual observation where all individual differences have equal weight.

Hyperparameter Tuning:
There is a list of different machine learning models. They all are different in some way or the other, but what makes them different is nothing but input parameters for the model. These input parameters are named as Hyperparameters. These hyperparameters will define the architecture of the model, and the best part about these is that you get a choice to select these for your model. You must select from a specific list of hyperparameters for a given model as it varies from model to model.
We are not aware of optimal values for hyperparameters which would generate the best model output. So, what we tell the model is to explore and select the optimal model architecture automatically. This selection procedure for hyperparameter is known as Hyperparameter Tuning. We can do tuning by using GridSearchCV. GridSearchCV is a function that comes in Scikit-learn (or SK-learn) model selection package. An important point here to note is that we need to have Scikit-learn library installed on the computer. This function helps to loop through predefined hyperparameters and fit your estimator (model) on your training set. So, in the end, we can select the best parameters from the listed hyperparameters.

```
[121]: GridSearchCV(cv=5, estimator=Ridge(),
                    param_grid={'alpha': [0.0001, 0.001, 0.01, 0.05, 0.1, 0.2, 0.3,
                                          0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0, 2.0, 3.0,
                                          4.0, 5.0, 6.0, 7.0, 8.0, 9.0, 10.0, 20, 50,
                                          100, 500, 1000]},
                    return_train_score=True, scoring='r2', verbose=1)
```

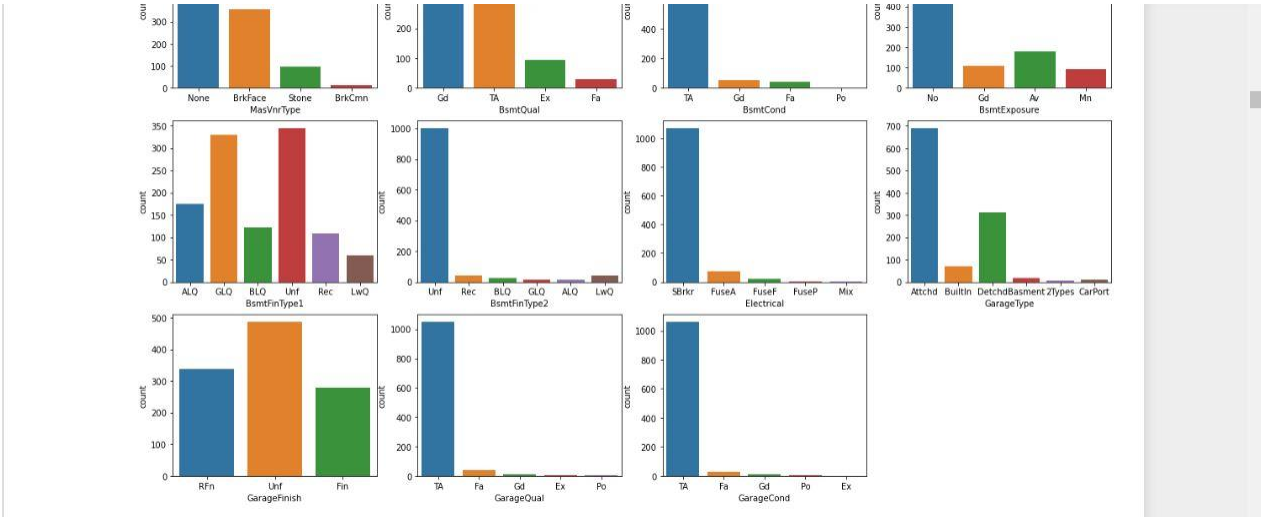
```
model_cv.fit(X_train,y_train)
```

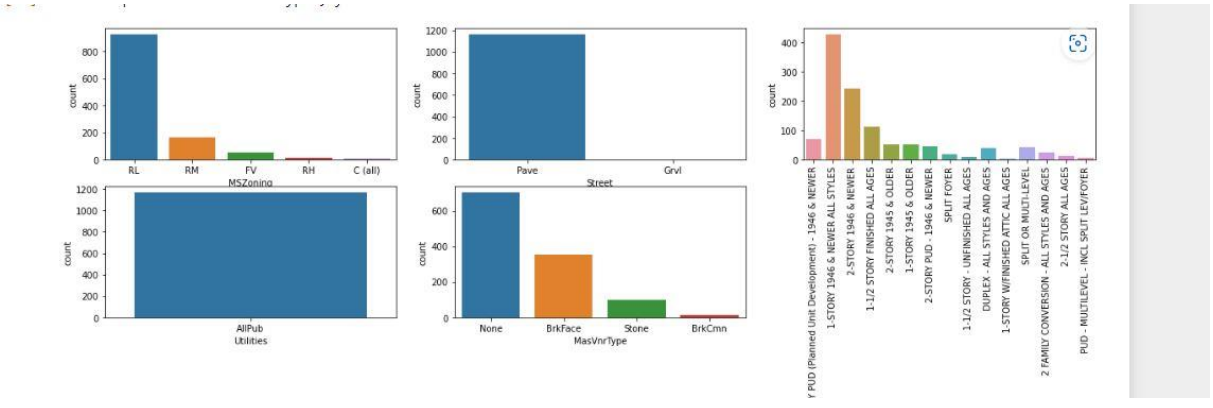
```
Fitting 5 folds for each of 6 candidates, totalling 30 fits
```

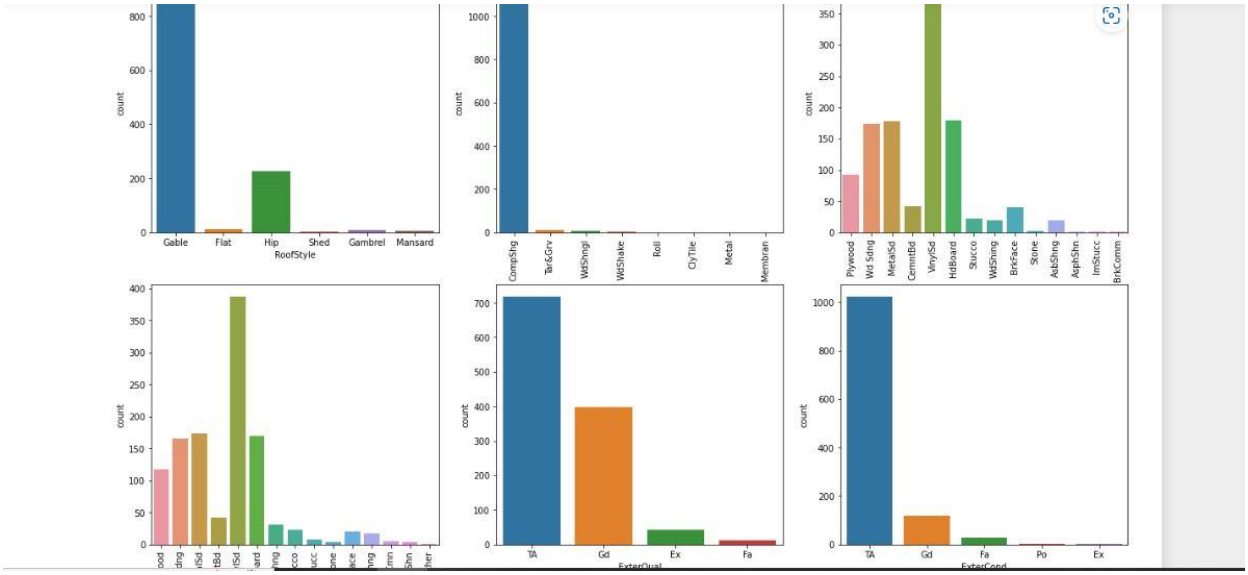
```
Out[129]: GridSearchCV(cv=5, estimator=Lasso(),
                       param_grid={'alpha': [0.0001, 0.0002, 0.0003, 0.0004, 0.0005,
                                             0.01]},
                       return_train_score=True, scoring='neg_mean_absolute_error',
                       verbose=1)
```

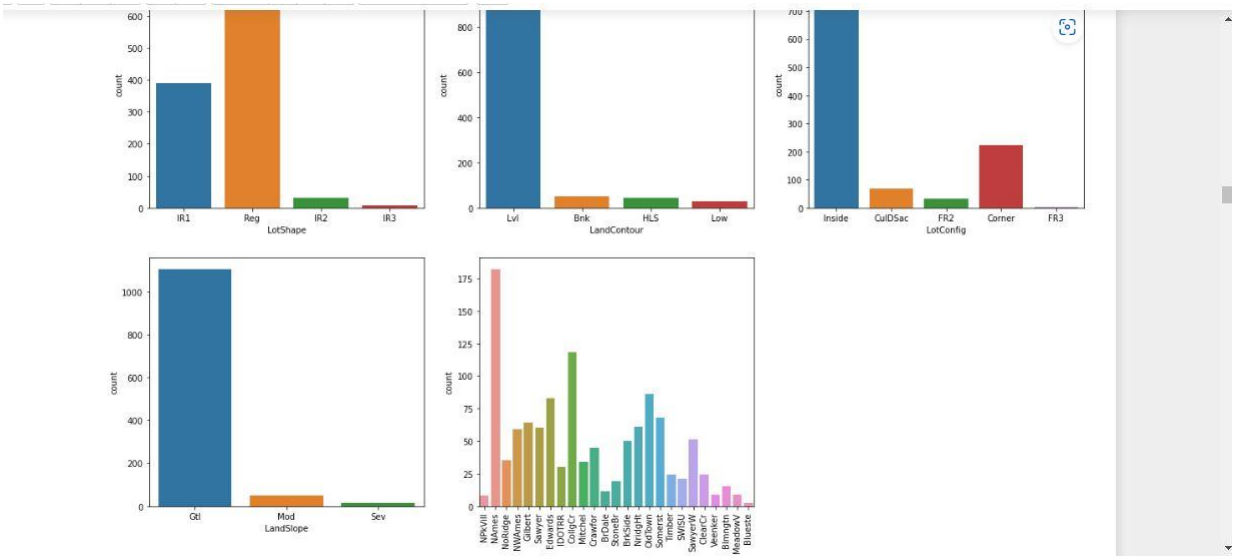
It is possible that there are times when the default parameters perform better than the parameters list obtained from the tuning and it only indicates that there are more permutations and combinations that one needs to go through for obtaining better results.

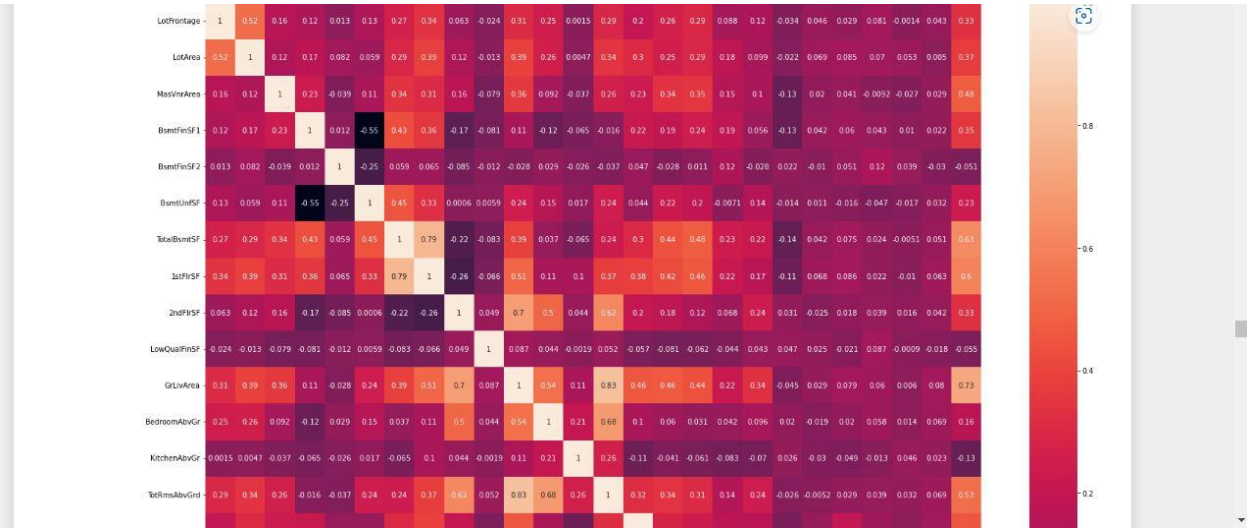
Visualizations :

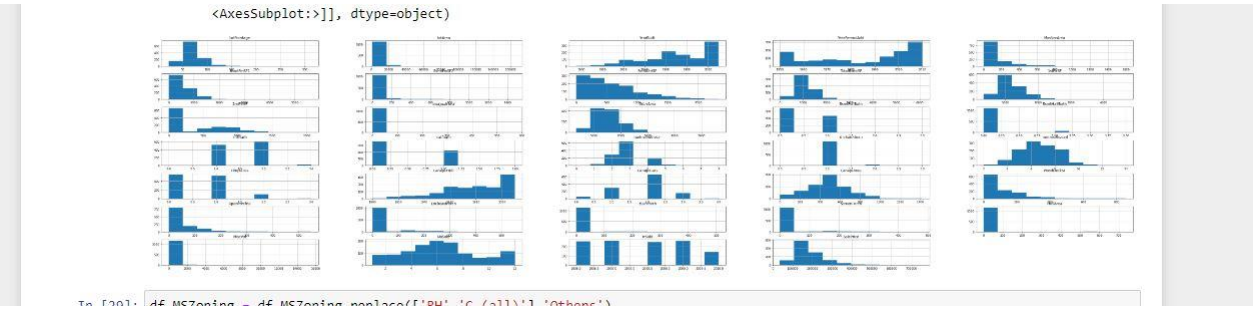





count
RoofStyle
Gable
Flat
Hip
Shed
Gambrel
Mansard
CompShg
Tar&Grv
WdShngl
WdShake
Roll
ClyTile
Metal
Membran
Plywood
Wd Sdng
MetalSd
CemntBd
VinylSd
HdBoard
Stucco
WdShing
BrkFace
Stone
AsbShng
AsphShn
ImStucc
BrkComm
ExterQual
TA
Gd
Ex
Fa
ExterCond
Po


count
LotShape
IR1
Reg
IR2
IR3
LandContour
Lvl
Bnk
HLS
Low
LotConfig
Inside
CulDSac
FR2
Corner
FR3
LandSlope
Gtl
Mod
Sev
NPkVill
NAmes
NoRidge
NWAmes
Gilbert
Sawyer
Edwards
IDOTRR
CollgCr
Mitchel
Crawfor
BrDale
StoneBr
BrkSide
NridgHt
OldTown
Somerst
Timber
SWISU
Sawyer
ClearCr
Veenker
Blmngtn
MeadowV
Blueste





Visualizations: It helped me to understand the correlation between independent and dependent features. Also, helped me with feature importance and to check for multi collinearity issues. Detected outliers/skewness with the help of boxplot and distribution plot. I got to know the count of a particular category for each feature by using count plot and most importantly with predicted target value distribution as well as scatter plot helped me to select the best model.

Pre-processing: Basically, before building the model the dataset should be cleaned and scaled by performing few steps. As I mentioned above in the pre-processing steps where all the important features are present in the dataset and ready for model building.

Model Creation: Now, after performing the train test split, I have x_train, x_test, y_train & y_test, which are required to build Machine learning

models. I have built multiple regression models to get the best R2 score, MSE, RMSE & MAE out of all the models.

CONCLUSION:

Key Findings and Conclusions of the Study

I observed all the encoded dataset information by plotting various graphs and visualised further insights.

Learning Outcomes of the Study in respect of Data Science

The above study helps one to understand the business of real estate. How the price is changing across the properties. With the Study we can tell how multiple real estate amenities like swimming pool, garage, pavement and lawn size of Lot Area, and type of Building raise decides the cost. With the help of the above analysis, one can sketch the needs of a property buyer and according to need we can project the price of the property.

Limitations of this work and Scope for Future Work

During this project I have faced a problem of low amount of data. Many columns are with same entries in more than 80% of rows which lead to reduction in our model performance. One more issue is there are large number of missing values presents in this data set, so we have to fill those missing values in correct manner. We can still improve our model accuracy with some feature engineering and by doing some extensive hyperparameter tuning on it.