HOUSING: PRICE PREDICTION

Submitted by:

Snehal Kumawat

INTRODUCTION

- ## Business Problem Framing

The primary aim of this case study is to predict housing price for the given features to maximize the prediction accuracy by employing the proposed methodology.

- ## Conceptual Background of the Domain Problem

Data contains the information for various houses and names of the features. And the target column with prices of the house.

- ## Review of Literature

The reviewed literature was divided in three separate parts. First, the studies that emphasize then housing price evaluation using machine learning techniques are reviewed. The second part includes the studies focusing on hedonic-based regression, and other stochastic approaches for the price prediction problem. The third part of the literature review concentrates on the studies related to price prediction model using specifically machine learning algorithms.

- ## Motivation for the Problem Undertaken

Prediction house prices are expected to help people who plan to buy a house so they can know the price range in the future, then they can plan their finance well. In addition, house price predictions are also beneficial for property investors to know the trend of housing prices in a certain location.

- ## Data Sources and their formats

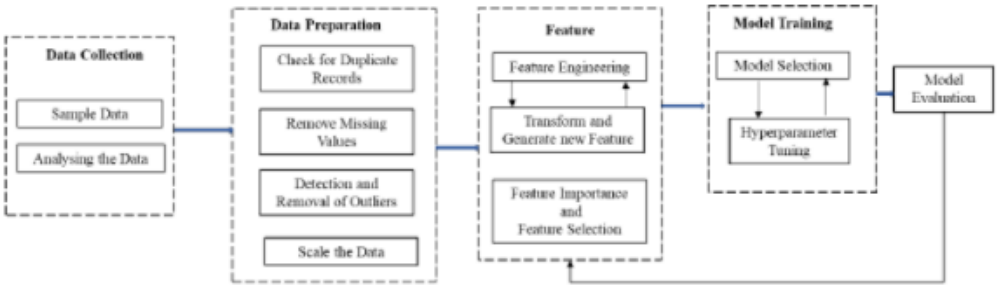
The dataset was in the form of a CSV file, so I used the read_CSV file function from the pandas module. The picture of the dataset I just have given below, you can observe that it consists of 1168 rows (records) and 81 columns (features).

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	...	PoolArea	PoolQC	Fence	MiscFeature	MiscVal
0	127	120	RL	NaN	4928	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	NaN	0
1	889	20	RL	95.0	15865	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	NaN	0
2	793	60	RL	92.0	9920	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	NaN	0
3	110	20	RL	105.0	11751	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	MnPrv	NaN	0
4	422	20	RL	NaN	16635	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	NaN	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1163	289	20	RL	NaN	9819	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	MnPrv	NaN	0
1164	554	20	RL	67.0	8777	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	MnPrv	NaN	0
1165	196	160	RL	24.0	2280	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	NaN	NaN	0
1166	31	70	C (all)	50.0	8500	Pave	Pave	Reg	Lvl	AllPub	...	0	NaN	MnPrv	NaN	0
1167	617	60	RL	NaN	7861	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	NaN	0

1168 rows × 81 columns

1. Data Preparation and Cleaning

Fig. 1 captures the research framework for the housing price prediction problem. It includes five major blocks, namely data collection, data preparation, feature processing, mode training, and model evaluation. These blocks of the diagram are explained in detail in the next subsections.



In this model, we need to feed the available independent variables to the model will predict the possible price of houses. For designing the model the machine learning method I used is multiple regression model and the tool I used for coding is jupyter notebook.

About Project

Machine Learning Regression model is developed to predict the price of houses. Data contains 1460 entries each having 81 variables. There are two datasets (test.csv, train.csv). We need to train on train.csv dataset and predict on test.csv file.

Packages used: Pandas, Numpy, Seaborn, Matplotlib, Scikit and Stats models.

2. Data Analysis

In this we need to extract data from train and test csv file. After understanding the datasets we need to clean the data if some features has missing values (such as Alley and PoolQC). Let's just remove 'Id' and the features with 30% or less NaN values.

Let's work on the categorical variables of our dataset.

Dealing with missing values Filling Categorical NaN with NA.

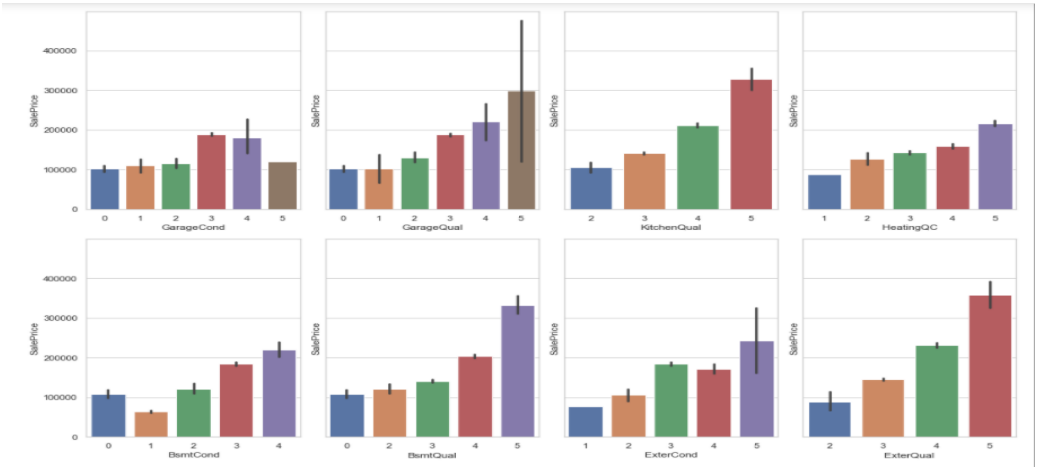
Data Visualization

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.

Let's now plot the correlation of some variables with target variable (Sale price).

Code

```
import matplotlib.pyplot as plt
ord_cols = ['ExterQual', 'ExterCond', 'BsmtQual', 'BsmtCond', \
'HeatingQC', 'KitchenQual', 'GarageQual', 'GarageCond']
f, axes = plt.subplots(2, 4, figsize=(15, 10), sharey=True)
for r in range(0, 2):
    for c in range(0, 4):
        sns.barplot(x=ord_cols.pop(), y="SalePrice", \
        data=df_tr, ax=axes[r][c])
plt.tight_layout()
plt.show()
```

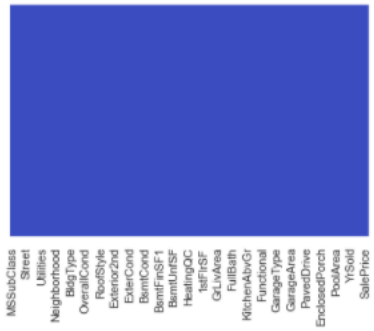


After filling all the null values then we check using heatmap is there any column is left with null values.

Checking through heat map

```
sns.heatmap(df_tr.isnull(),yticklabels=False,cbar=False,cmap='coolwarm')
```

Output

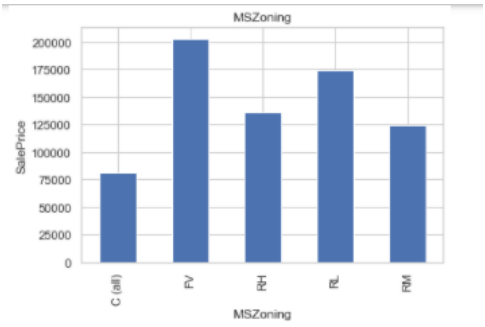


Now no null values are present hence now we do encoding of data. Let us find out the relationship between categorical variable and dependent feature Sales Price.

Code

```
for feature in categorical_features:
    data=df_tr.copy()
    data.groupby(feature)['SalePrice'].median().plot.bar()
    plt.xlabel(feature)
    plt.ylabel('SalePrice')
    plt.title(feature)
    plt.show()
```

This is the some reference of plot with respect to target column.

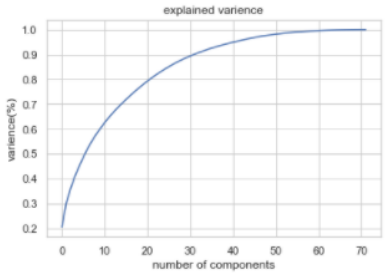


After that we checking the Outliers and then checking correlation of features with target column. And then we checking the correlation with target column. Then we scaled the data.

After the scaling data, the data contains 73 column and having muliti collinearity to remove it we need to use Principal Component Analysis (PCA).

Code

```
from sklearn.decomposition import PCA
pca=PCA()
pc=pca.fit_transform(x_t)
plt.figure()
plt.plot(np.cumsum(pca.explained_variance_ratio_))
plt.xlabel("number of components")
plt.ylabel("varience(%)")
plt.title("explained varience")
plt.show()
```



So now we selected 60 columns hence principal_x is having 60 columns and 1168 rows.

Now this train data is ready for regression.

Now we have do same EDA on test data. As the features are same only the rows are 292 only.

3. Regressor Model Building:

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables.

Splitting the Dataset

As usual for supervised machine learning problems, we need a training dataset to train our model and a test dataset to evaluate the model.

Code

```
x_train_b,x_test_b_b,y_train_b,y_test_b=train_test_split(principal_x,y,random_state=619,test_size=0.20)
```

Linear Regression

Using Linear Regression we got Training r2_score is:- 81.45 and Testing r2_score is:- 81.44

Regulariztion Methods

Lasso Regression

Firstly, we will use GridSearchCV() to search for the best model parameters in a parameter space provided by us. Then we got R2 Score: 81.54 and Cross Val Score: 77.59

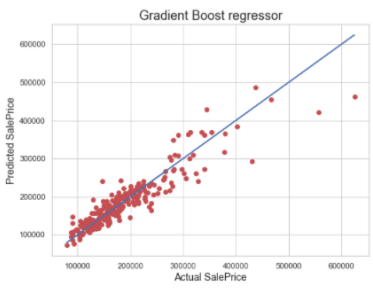
Ridge Regression

We got R2 Score: 81.54 and Cross Val Score: 77.59

GradientBoostingRegressor

We got R2 Score: 85.064 and Cross Val Score: 87.191

Real Vs Predicted



KNeighborsRegressor
We got R2 Score: 60.851 and Cross Val Score: 68.493

AdaBoostRegressor
We got R2 Score: 72.648 and Cross Val Score: 80.908

The best model is GradientBoostingRegressor.

Since the difference between the percentages score of cross validation and r2_score is optimum. After that we save the best model using pickle method.

4. Conclusion

It is seen that the most effective attribute in predicting the house price and that **Gradient Bo osting Regressor** is the most effective model for our Dataset with R2 Score of 85.064.
And here is original and predicted houses prices.

	original	predicted
0	264561	100489.040649
1	101800	138650.425753
2	204900	136927.088161
3	129000	171097.657120
4	262000	101643.849458
...	...	...
229	110000	169782.470475
230	217500	323278.349746
231	625000	137646.638671
232	198500	235655.021991
233	158500	221900.920025

We built serveral regression models to predict the price of some house given some of the house features. We eveluated and compared each model to determine the one with highest per formance.We also looked at how some models rank the features according to their importanc e. We followed the data science process starting with getting the data, then cleaning and preprocessing the data, followed by exploring the data and building models, then evaluating t he results and communicating them with visualizations.

5. Limitations of this work and Scope for Future Work

House is very important nowadays as the price of the land and price of the house increases every year. So our future generation needs a simple technique to predict the house price in future. The price of house helps the buyer to know the cost price of the house and also the right time to buy it. Here we have used various regression techniques to predict the house price.