FLIP ROBO

Housing Price Prediction Project

Submitted by:

Priya Rajput

ACKNOWLEDGMENT

I would like to express my gratitude towards Flip-Robo for providing me this opportunity to show case my talent and also for their constant support and guidance. Also It is indeed a pleasure for me to have worked on this project.

I express my deepest thanks to **Miss Sapna Verma**, for taking part in useful decision & giving necessary advices and guidance and arranged all facilities to make my project easier. I choose this moment to acknowledge his contribution gratefully.

TABLE OF CONTENTS

INTRODUCTION

Business Problem Framing

Housing prices are an important reflection of the economy, and housing price ranges are of great interest for both buyers and sellers. In this project, house prices will be predicted given explanatory variables that cover many aspects of residential houses. The goal of this project is to create a regression model that is able to accurately estimate the price of the house given the features.

This model can be useful for potential buyers in deciding the characteristics of a house they want that best fits their budget and will be of tremendous benefit, especially to housing developers and researchers, to ascertain the most significant attributes to determine house prices and to acknowledge the best machine learning model to be used to conduct a study in this field.

Conceptual Background of the Domain Problem

1. Check whether the problem is supervised or not.

2. Check whether model is regression or classification type.

3. Perform various mathematical and statistical analysis which include description or statistical summary of the data, correlation using corr() and data visualization . Then we have used zscore to plot outliers and remove them.

4. Check whether our dataset is balanced or imbalanced. If data is imbalanced, then we apply sampling techniques to balance the dataset.

4. Building the model and check its accuracy.

5. Selecting the model, hyperparameter tuning would be done.

Literature Review

“House Price Prediction using a Machine Learning Model: A Survey of Literature” and “The impact of housing quality on house prices in eight capital cities, Australia” were reviewed and evaluated to gain insights into all the attributes that influence the price of house.

From papers we get to know location attributes and structural attributes are two prominent factors in predicting house prices. Studies suggest that there exists a close relationship between House pricing and location attributes such as distance from the closest shopping center, train station, position offering views of hills or shore, the neighborhood in which the property is situated etc.

Structural attributes of the house like lot size, lot shape, quality and condition of the house, garage capacity, rooms, Lot frontage, number of bedrooms, bathrooms, overall finishing of the house etc play a big role in influencing the house price.

Neighborhood qualities can be included in deciding house price. Factors like efficiency of public education, community social status, the socio-cultural demographics improve the worth of a property.

In above mentioned research paper, various models were built in which the house Sale Price is projected as separate and dependent variable while location, structure and various other attributes of housing properties were treated as independent variables. Therefore, the house price is set as a target or dependency variable, while other attributes are set as independent variables to determine the main variables by identifying the correlation coefficient of each attribute.

Motivation for the Problem Undertaken

The objective behind to make this project is to contribute to the world's economy. 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. It is a very large market and there are various companies working in the domain. Data science provides motivation as it can 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.

Analytical Problem Framing

Mathematical/ Analytical Modeling of the Problem

In this project we have performed various mathematical and statistical analysis. We checked description or statistical summary of the data using describe, info and unique value count.

- The dataset consist of train dataset of 1168 rows and 81 columns and test dataset of 292 rows and 80 columns.
- All of the attributes were of 'int64', 'fload64' and 'object'
- Dataset contain null values in various columns.

Data Sources and their formats

The data was provided to us by our client who is in the Housing Industry. The data was in the form of a CSV file.

There are two csv file one is train datasheet and other one is test datasheet.

Training Dataset contains 1168 entries and 81 variables,
Test Dataset contains 292 entries and 80 variables.

Data Pre-processing Done

First, we have imported the necessary libraries and dataset.

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
```

```
df=pd.read_csv(r'C:\Users\Admin\Desktop\housing folder\train.csv')
df.head()
```

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	LotConfig	LandSlope	Neighborhood	Condition1	Condi
0	127	120	RL	NaN	4928	Pave	NaN	IR1	Lvl	AllPub	Inside	Gtl	NPkVill	Norm	
1	889	20	RL	95.0	15865	Pave	NaN	IR1	Lvl	AllPub	Inside	Mod	NAmes	Norm	
2	793	60	RL	92.0	9920	Pave	NaN	IR1	Lvl	AllPub	CulDSac	Gtl	NoRidge	Norm	
3	110	20	RL	105.0	11751	Pave	NaN	IR1	Lvl	AllPub	Inside	Gtl	NWAmes	Norm	
4	422	20	RL	NaN	16635	Pave	NaN	IR1	Lvl	AllPub	FR2	Gtl	NWAmes	Norm	

```
df.info()
```

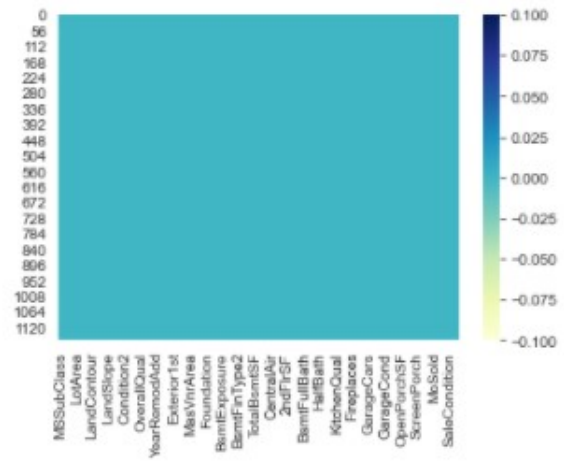
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1168 entries, 0 to 1167
Data columns (total 81 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   Id             1168 non-null   int64
 1   MSSubClass     1168 non-null   int64
 2   MSZoning       1168 non-null   object
 3   LotFrontage    954 non-null    float64
 4   LotArea        1168 non-null   int64
 5   Street         1168 non-null   object
 6   Alley          77 non-null     object
 7   LotShape       1168 non-null   object
 8   LandContour    1168 non-null   object
 9   Utilities      1168 non-null   object
 10  LotConfig      1168 non-null   object
 11  LandSlope      1168 non-null   object
 12  Neighborhood   1168 non-null   object
 13  Condition1     1168 non-null   object
 14  Condition2     1168 non-null   object
 15  BldgType       1168 non-null   object
 16  HouseStyle     1168 non-null   object
 17  OverallQual    1168 non-null   int64
 18  OverallCond    1168 non-null   int64
 19  YearBuilt      1168 non-null   int64
 20  YearRemodAdd   1168 non-null   int64
 21  RoofStyle      1168 non-null   object
 22  RoofMatl       1168 non-null   object
```

```
df.isna().sum()
```

```
Id                0
MSSubClass        0
MSZoning          0
LotFrontage     214
LotArea           0
                 ...
MoSold            0
YrSold            0
SaleType          0
SaleCondition     0
SalePrice         0
Length: 81, dtype: int64
```

```
sns.heatmap(df.isnull(), cmap='YlGnBu')
```

```
<AxesSubplot:>
```



Data Inputs- Logic- Output Relationships

EDA was performed by creating valuable insights using various visualization libraries.

Hardware and Software Requirements and Tools Used

Hardware required:
Processor: core i3
RAM: 8 GB

Software required:
Anaconda 3- language used Python 3
Microsoft Excel

MODEL DEVELOPMENT AND EVALUATION

Identification of Possible Problem-Solving Approaches (Methods)

- Converted all our categorical variables to numeric variables with the help of label encoder to checkout and dropped the columns which we felt were unnecessary.

```
df.drop(['Alley','PoolQC','Fence','MiscFeature','FireplaceQu','Id'],axis=1,inplace=True)
```

- There are missing value which we fill using mean and mode

```
df['LotFrontage']=df['LotFrontage'].fillna(df['LotFrontage'].mean())
df['BsmtQual']=df['BsmtQual'].fillna(df['BsmtQual'].mode()[0])
df['BsmtCond']=df['BsmtCond'].fillna(df['BsmtCond'].mode()[0])
df['GarageQual']=df['GarageQual'].fillna(df['GarageQual'].mode()[0])
df['GarageType']=df['GarageType'].fillna(df['GarageType'].mode()[0])
```

```
df.drop(['GarageYrBlt'],axis=1,inplace=True)
```

```
df['GarageFinish']=df['GarageFinish'].fillna(df['GarageFinish'].mode()[0])
df['GarageCond']=df['GarageCond'].fillna(df['GarageCond'].mode()[0])
```

```
df.shape
```

```
(1168, 74)
```

```
df['BsmtQual']=df['BsmtQual'].fillna(df['BsmtQual'].mode()[0])
df['BsmtCond']=df['BsmtCond'].fillna(df['BsmtCond'].mode()[0])
```

```
df['MasVnrType']=df['MasVnrType'].fillna(df['MasVnrType'].mode()[0])
df['MasVnrArea']=df['MasVnrArea'].fillna(df['MasVnrArea'].mode()[0])
df['BsmtExposure']=df['BsmtExposure'].fillna(df['BsmtExposure'].mode()[0])
df['BsmtFinType1']=df['BsmtFinType1'].fillna(df['BsmtFinType1'].mode()[0])
df['BsmtFinType2']=df['BsmtFinType2'].fillna(df['BsmtFinType2'].mode()[0])
```

- We observed skewness in data so we tried to remove the skewness through treating outliers but using zscore we loss more than 50% data which is not good for model building.

```
from scipy.stats import zscore
z=np.abs(zscore(df))
threshold=3
print(np.where(z>3))
df_new=df[(z<3).all(axis=1)]
df_new.shape
```

```
(array([   1,    1,    1, ..., 1166, 1166, 1166], dtype=int64), array([ 8, 19,

(482, 73)
```

```
df.shape
```

```
(1168, 73)
```

```
loss=(1168-482)/1182*100
loss
```

- For scaling the data, I have used Standard Scaler method

```
#seprating input and output from df_newtrain
x=df.drop(columns=["SalePrice"])
y=df[["SalePrice"]]
```

```
from sklearn.preprocessing import StandardScaler
sc=StandardScaler()
x=sc.fit_transform(x)
```

- various regression model is used and we select our model for hyper parameter tuning which have good accuracy and cross validation score.

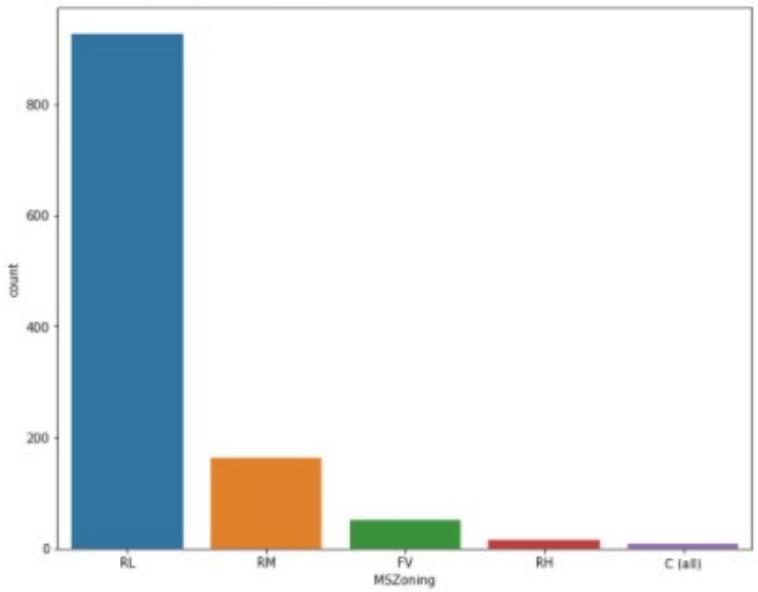
```
from sklearn.linear_model import LinearRegression
from sklearn.metrics import r2_score,mean_squared_error,mean_absolute_error
from sklearn.metrics import accuracy_score,classification_report,confusion_matrix
from sklearn.naive_bayes import GaussianNB
from sklearn.svm import SVR
from sklearn.tree import DecisionTreeRegressor
from sklearn.neighbors import KNeighborsRegressor
from sklearn.model_selection import GridSearchCV
from sklearn.model_selection import cross_val_score
from sklearn.model_selection import train_test_split
```

Visualization

Univariate Analysis

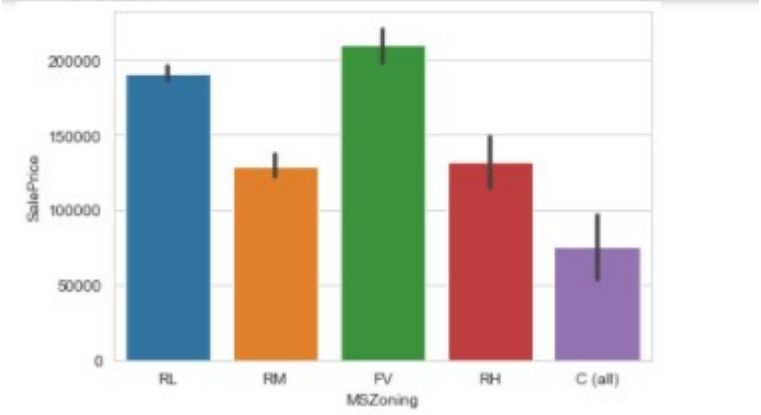
```
#for categorical columns-

object_columns = df.select_dtypes(include=['object']).columns
for column in object_columns:
    plt.figure(figsize=(10,8))
    sns.countplot(x=column,data=df)
```



Bivariate Analysis

```
for col in cat_col:
    data=df.copy()
    sns.barplot(data[col],data['SalePrice'])
    plt.xlabel(col)
    plt.ylabel('SalePrice')
    plt.show()
```

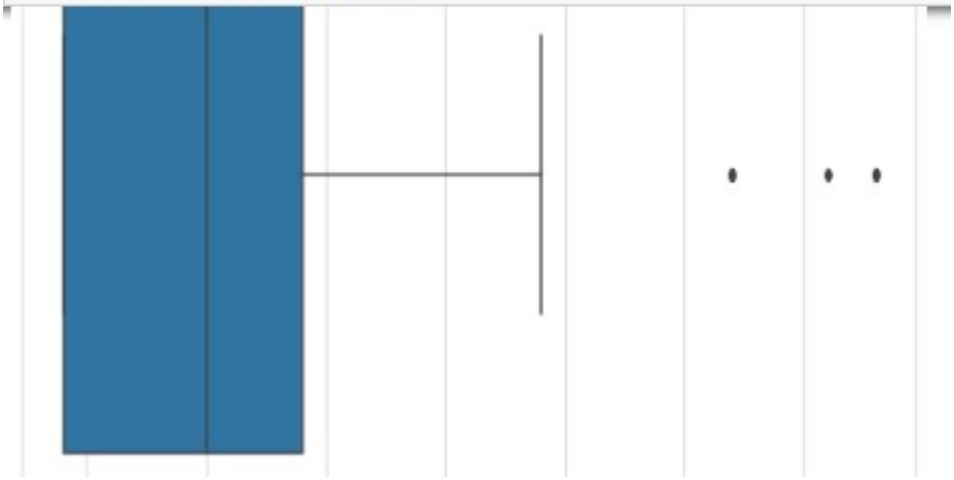


Outlier checking

```
#for checking outliers present in continous featyres-

# cont_columns = df.select_dtypes(include=(['int64']or['float64'])).columns

for column in df.columns:
    plt.figure(figsize=(10,8))
    sns.boxplot(x=column,data=df)
```



Interpretation of the Results

In the visualization part, data is correlated with each other and also check correlation with the target variable. For this work, I have used heatmap, barplots.

In the pre-processing part, data cleared in various ways. Firstly, dropped some columns which we feel are not contributing in predicting the target variable. Then we use Label Encoder to encode the object type data because a machine can only read numbers. I also replaced the null values using mean and mode.

In the modeling part, we select our model for hype parameter tuning i.e Gradient Boosting Regressor model and calculate r2score and its mean absolute error

CONCLUSION

Key Findings and Conclusions of the Study

We have to study the data very clearly so that we are able to decide which data are relevant for our findings. The techniques that I have used are heatmap, Label Encoder etc.

The conclusion of our study is we have to achieve a model with good accuracy.

Learning Outcomes of the Study in respect of Data Science

We will develop relevant programming abilities. We will demonstrate proficiency with statistical analysis of data. We will develop the ability to build and assess data-based models. We will execute statistical analyses with professional statistical software. The best algorithm for this project according to my work is Gradient Boosting Regressor because the accuracy that I have achieved is quite satisfactory than the other models.

Limitations of this work and Scope for Future Work

The results were promising for the public data due to it being rich with features and having strong correlation, whereas the local data gave a worse outcome when the same pre-processing strategy was implemented due to it being in a different shape compared with the public data in terms of the number of features and the correlation strength. Hence, the local data needs more features to be added preferably with a strong correlation with the house price.

Future scope of this work is that we can try different algorithms and approaches to achieve a good accuracy and f1-score.