HOUSING: PRICE PREDICTION

Submitted by: Harneet Kaur Rehsi

ACKNOWLEDGMENT

I want to thank my SME Khusboo Garg and Flip Robo Techniques to express our sincere thanks to the following people, without whom we would not have been able to complete this project.

INTRODUCTION

- **Business Problem Framing:** 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

- **Motivation for the Problem Undertaken:** This model will then be used by the management to understand how exactly the prices vary with the variables. They can accordingly manipulate the strategy of the firm and concentrate on areas that will yield high returns. Further, the model will be a good way for the management to understand the pricing dynamics of a new market.

Analytical Modeling of the Problem

In this project, the dataset contains several rows and columns containing all the necessary information.For removing NaN values or in the dataset we have used several statistical and exploratory data visualization for better understanding and model building for predictions.

- **Data Sources and their formats:** The data contains two different datasets. train dataset and test dataset. We have used a train dataset for training and building our model. We have used test dataset for testing the model.

- **Data Preprocessing Done:** Dataset contains NaN values which are very important to be removed or filled. so I have used the fill na method and filled NaN values with some meaningful data using the mean and mode method. Then we have used EDA for better visualization of data and relationship between features and the target variable.

- **Hardware and Software Requirements and Tools Used:** We have used and import several libraries like pandas , numpy , matplotlib.pyplot ,seaborn , siciklearning models etc:

- import pandas as pd

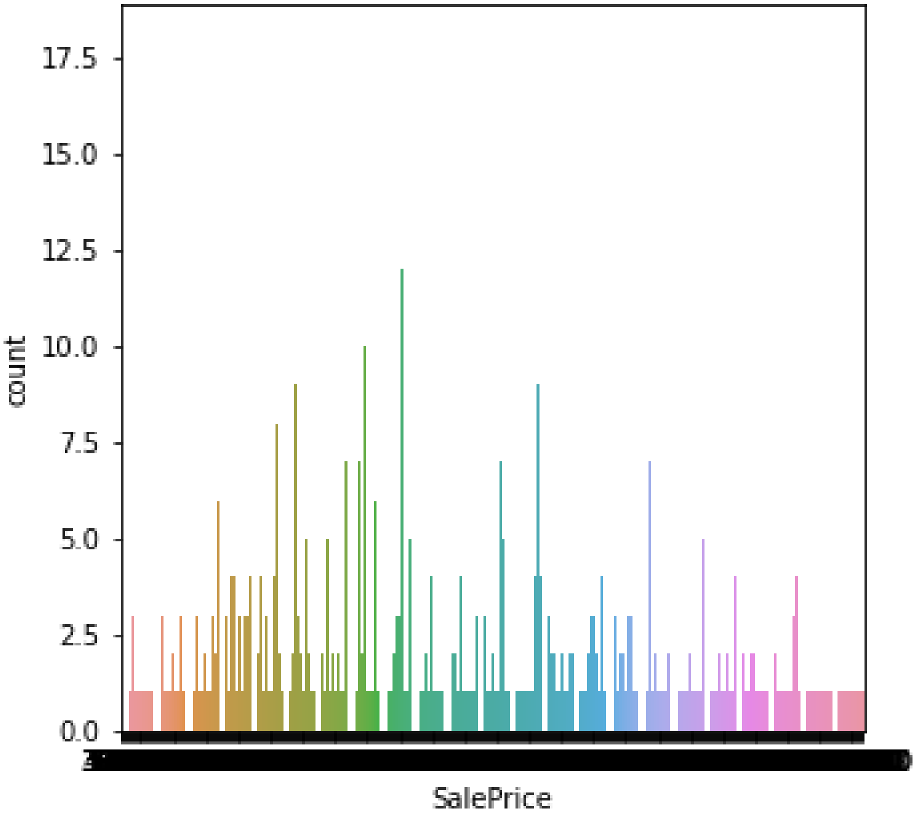
- import numpy as np
- pd.set_option('display.max_columns', 100)
- pd.set_option('display.max_rows', 100)
- import warnings
- warnings.filterwarnings('ignore')
- import matplotlib.pyplot as plt
- import seaborn as sns
- %matplotlib inline
- from sklearn.model_selection import train_test_split
- from sklearn.linear_model import LinearRegression
- from sklearn.tree import DecisionTreeRegressor
- from sklearn.linear_model import Ridge
- from sklearn.linear_model import Lasso
- from sklearn.preprocessing import LabelEncoder
- from sklearn.metrics import mean_squared_error , mean_absolute_error
- from sklearn.metrics import mean_squared_error , r2_score

Model/s Development and Evaluation

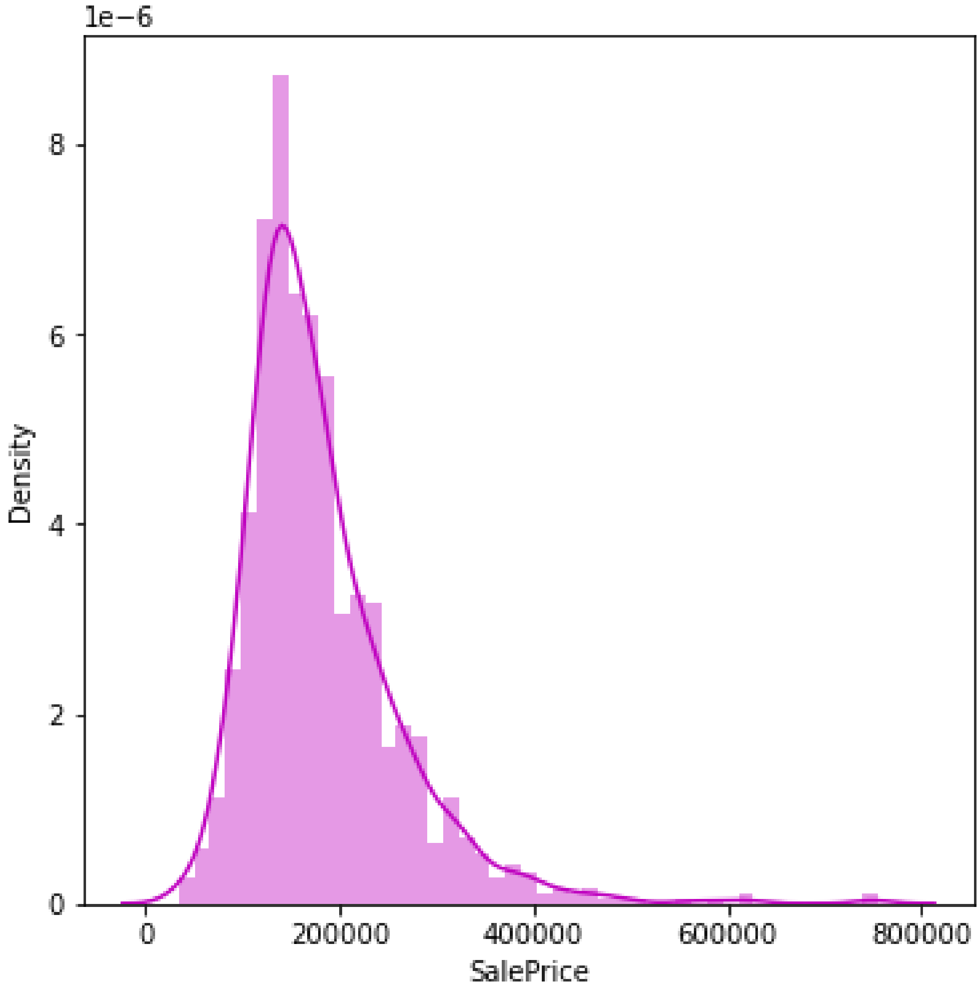
- Testing of Identified Approaches (Algorithms)
- import pandas as pd
- import numpy as np
- pd.set_option('display.max_columns', 100)
- pd.set_option('display.max_rows', 100)
- import warnings
- warnings.filterwarnings('ignore')
- import matplotlib.pyplot as plt
- import seaborn as sns
- %matplotlib inline
- from sklearn.model_selection import train_test_split
- from sklearn.linear_model import LinearRegression
- from sklearn.tree import DecisionTreeRegressor
- from sklearn.linear_model import Ridge
- from sklearn.linear_model import Lasso
- from sklearn.preprocessing import LabelEncoder

- from sklearn.metrics import mean_squared_error , mean_absolute_error
- from sklearn.metrics import mean_squared_error , r2_score

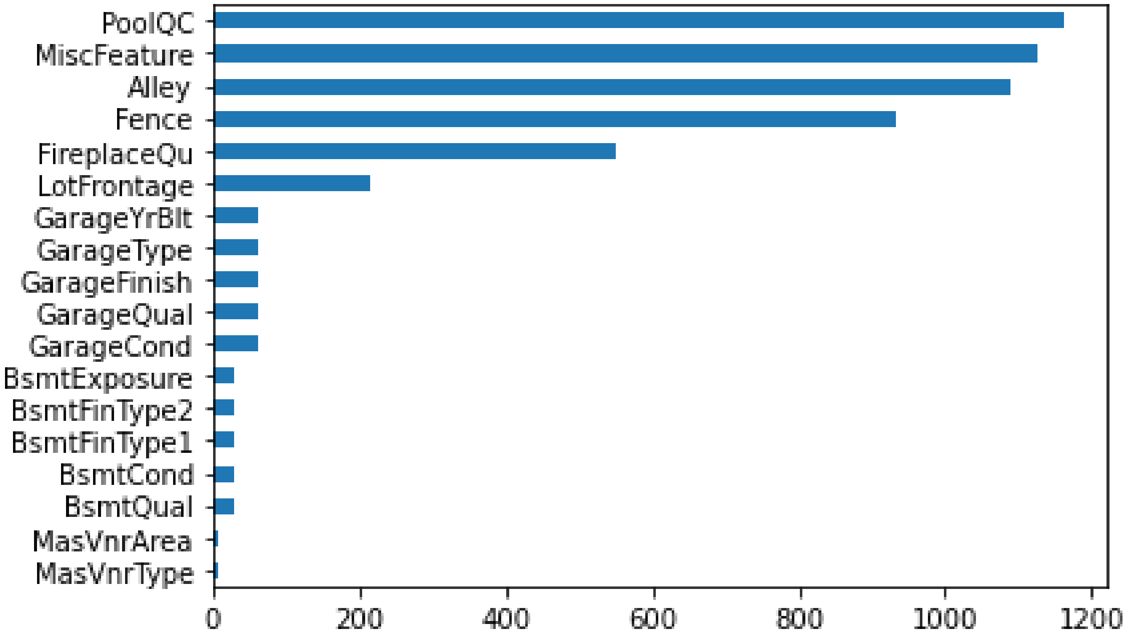
- Visualizations



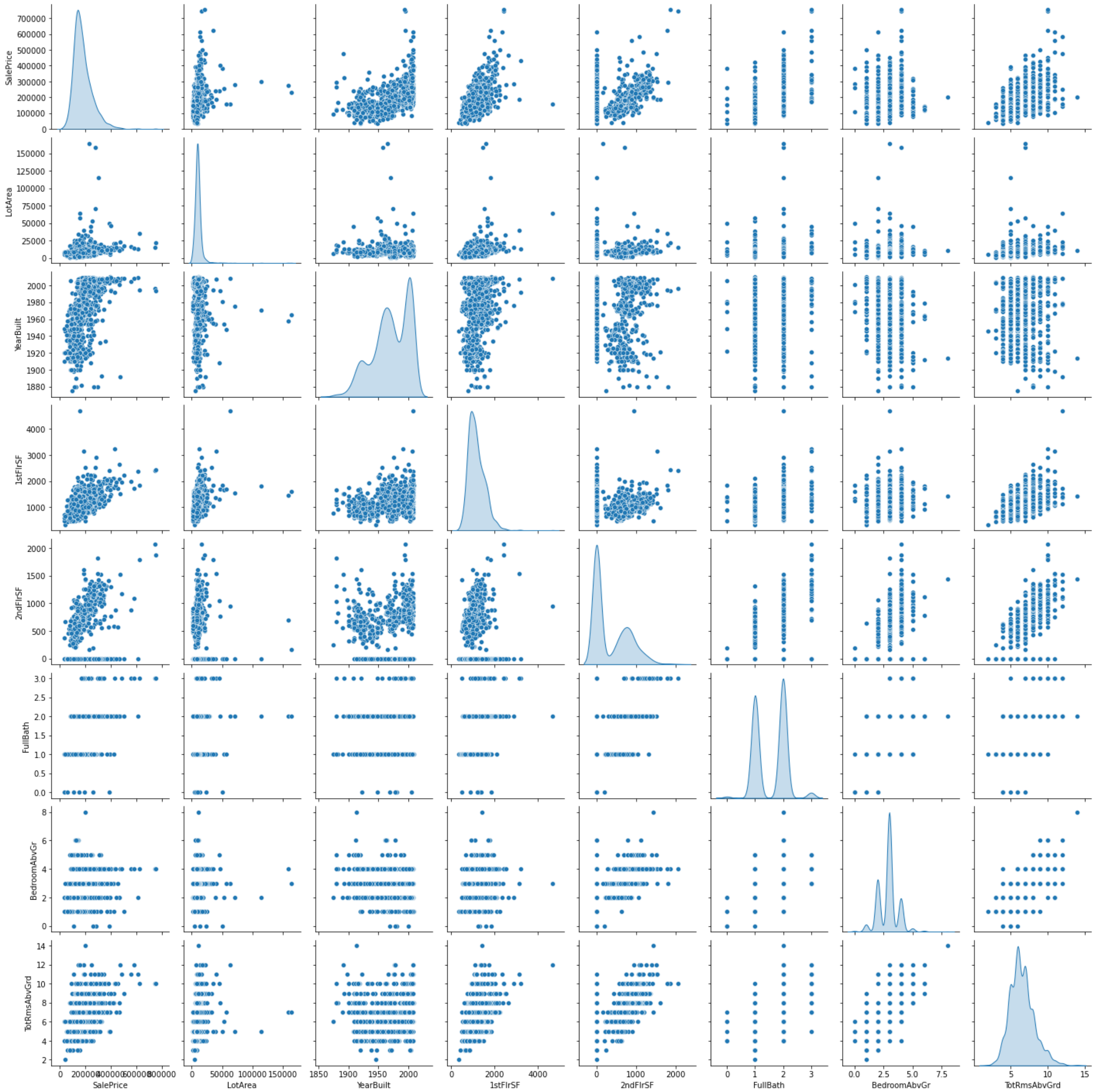
Used countplot for observing SalePrice



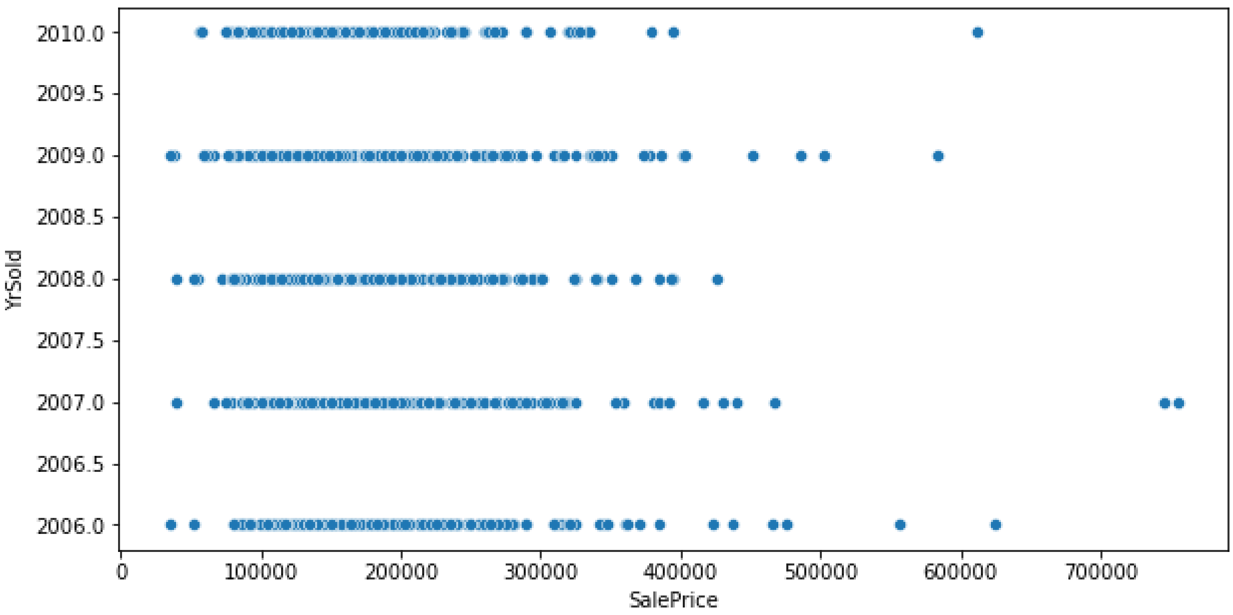
Used distplot on SalePrice to check the skewness of the data. the data is not skeweed as the curve its making is a bell shape curve.



Here it is showing the features containing null values and the features containing the most and the least null values



Used pair plot for some features ["SalePrice", "LotArea", "YearBuilt", "1stFlrSF", "2ndFlrSF", "FullBath", "BedroomAbvGr", "TotRmsAbvGrd"]



Used scatterplot and check the relationship between ‘saleprice’ and ‘yrsold’

Model Building

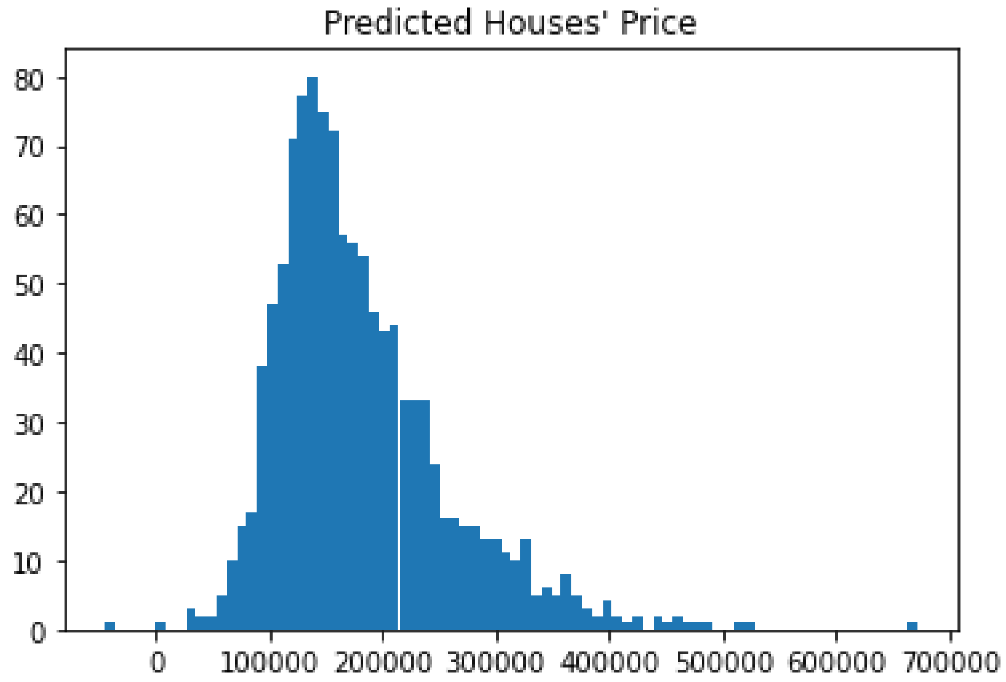
For model building first we have clean both train data and test data . we have clean data using fill na method with mean function and mode function . then we have check the null values as well as duplicate values.

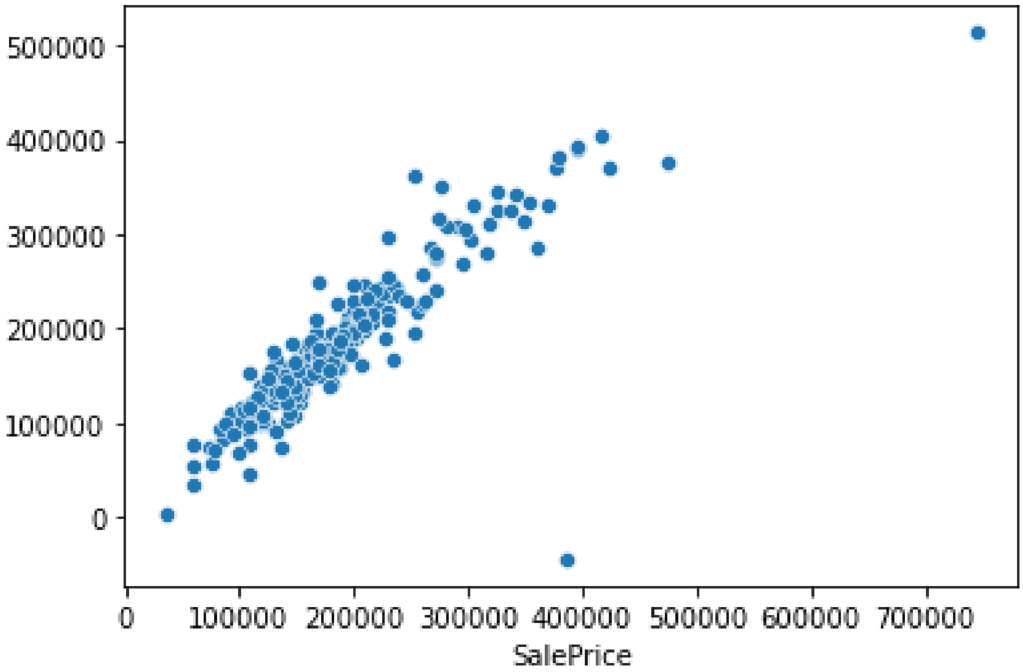
After cleaning the data we have divide features in x and target variable in y . used train_test_split for splitting the dataset .

We have used linear regression for fitting the model and for predictions we have check r2score and also checked MEAN ABSOLUTE ERROR, MEAN SQUARED ERROR, ROOT MEAN SQUARED ERROR.

Used lasso regression for checking the fitting of the model. Also used decision tree regressor for better score or prediction.

Predicted house price





Conclusion

We started this data science project with a seemingly simple question — “What is the price of this property? How can we improve our price estimations?”. After gathering data and a significant amount of time was spent cleaning the data and extracting the features needed for modeling.
We tested multiple variations of regression models, including Linear regression, Decision tree regressor, using Lasso for checking if the model is overfitted. MEAN SQUARED ERROR, MEAN ABSOLUTE ERROR, ROOT MEAN SQUARED ERROR for calculating the mean of error.
Two models that used different combinations of features were trained for this project — it is helpful to compare them side by side.