MET CS 521 Summer 2022
Final Project Writeup : House Price Predictor
Team Members: Rohit V Satyendra, Serena Jacob, Vijaya Palaniappan

Introduction:

The overarching goal of this project is to create a service to predict housing prices and get a feel for modern data science using python. We are using Jupyter Notebook for this task as it allows us to continue to develop code and visualize results for ease in rerunning and testing various analyses. We are also utilizing GitHub in order to collaborate with a group and display the efficiency of pushing, pulling, and merging changes with others. In order to solve our problem, we first needed to download the data-set for analysis. In our case, it is the House Prediction Data which we loaded into python using pandas. We then worked with the data using different techniques- which we will further elaborate on- to understand and interpret the data set as well as display key aspects of this data for reader-friendly visualization. We utilized machine learning such as classification, clustering, and linear regression to perform our analysis of the data. Finally, we customized our project using extension tasks to further customize our project :-

- Research making a network request for data
- Pandas, using dataframe, plotting
- Sklearn and linear regression model

Methods and Techniques:

We have used multiple packages and modules to explore the dataset. The list are such:-

1. Pandas were used to read the data in the CSV file with the read_csv function.
2. Numpy is a library for working with arrays.
3. Matplotlib and %matploplib inline were used to be able to add plots to Jupyter notebook.
4. Sklearn were used to build a simple regression model.
5. Seaborn were used for the techniques to visualize distribution

Extensions:

We chose three extensions to customize our project. We used quantitative methods to intelligently pick features, we compared different regression models, and we also created a simple UI that lets users enter information and get an estimate of their housing price. In order to intelligently pick features that would best work for the estimation, we used brute force method to check the r2 metric and absolute

- Use quantitative methods to intelligently pick features
    - Implemented function to assign numerical values to nominal categorical values in order to add those categories to the graphs.
- Compare different regression models https://devopedia.org/types-of-regression
    - Linear regression, decision tree regression, log regression
    - Regression uncovers useful relationships, how predictors are correlated to the response variable

- Regression makes no claim that predictors influence or cause the outcome
- With one predictor- univariate regression
- With multiple predictors- multivariate regression

Initially we imported the modules that we needed but decided to drop nan attributes with nan values because these details would not add anything to our prediction values.

Feature engineering shows that we could assign numerical values to the descriptions of each attribute so that we could use them for our further analysis such as heatmap, linear regression, r2 score, mse value, mean absolute error, explained variance etc. We could make anything in the attributes list usable because initially if it was a string as an input the linear regression model would not be able to do anything with it, therefore we used different types of modules to visualise the distributions.

Since we wanted to use neighborhood as an input, we made a code to be able to use the average prices of each neighborhood as a variable. The code block with the keys was just brute force checking how different inputs would affect the linear regression model since we wanted the best one.

Using linear regression in our graphs had different metrics to have the data of what worked well. Linear regression model was good because of the straight line presentation.

The number of parameters vs r2 score graph were used to check the different keys from earlier datas and we wanted to see how changing the input parameters and increasing them generally increases the r2 score and various metrics because there is more data to base the calculation off.

Then we have the distribution plot of the absolute error of ground truth vs prediction which is our accuracy for the house prediction price. The n is the UI based on the metrics we initially decided were the best based off of metrics so that last key is the attributes we are adding for the user to put there preferences as and we can generate the prediction off of those attributes and the amount of whatever they want.

Everything is input as a number except for neighborhood where they have to select a neighborhood they want to live in. Scatter plots were used to show the different averages house price of each neighborhood to show that the neighborhoods had varying average prices. This made us want to use that more because people generally pick a place to live based on the area and so wanted to include that as a parameter.

Conclusion:

We have used Exploratory Analysis to explore and extract the data from the House Price Prediction that are helpful and we have taken into account certain attributes to compare with saleprice. As mentioned earlier, there certain variables that did not any values to our prediction power.

In conclusion, we have used different selection model such as Linear Regression, Decision Tree Regression, Random Forest Regression, Navie bales Regression, SVM Regression, KNN classification and Logistic Regression to asses the correctness of the dataset given to us. We have concluded that from all the above models, Decision tree is the best model.

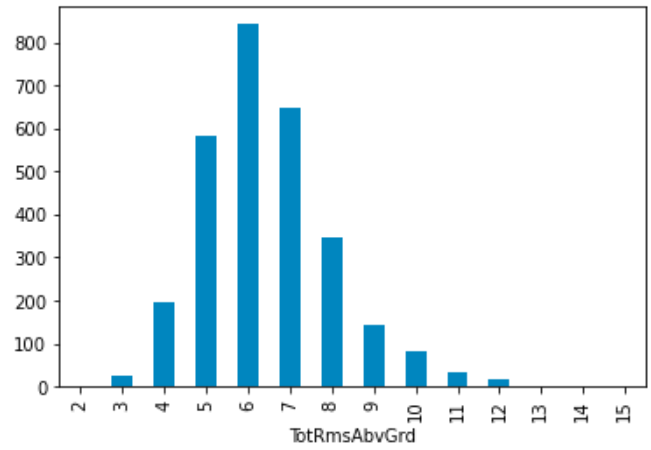

Figure A : Bar chart of TotRmsAbvGrd comparing to Neighbourhood

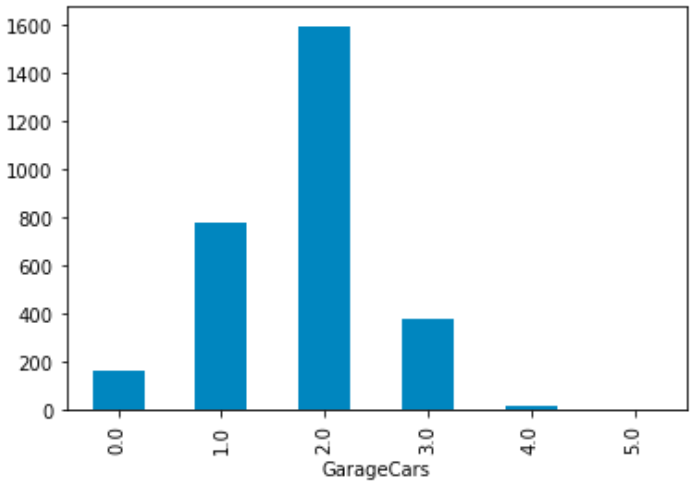

Figure B : Bar Chart showing the comparison between GarageCars and Neighbourhood

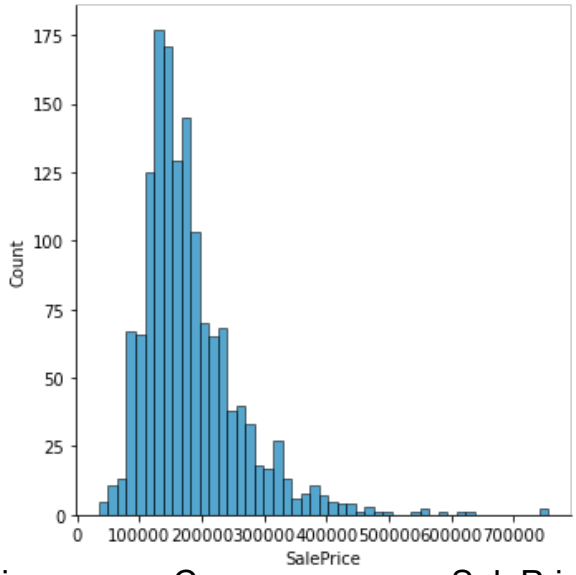

Figure C : SalePrice distribution (bell-shaped)- the distribution is concentrated between 100000 to 200000.

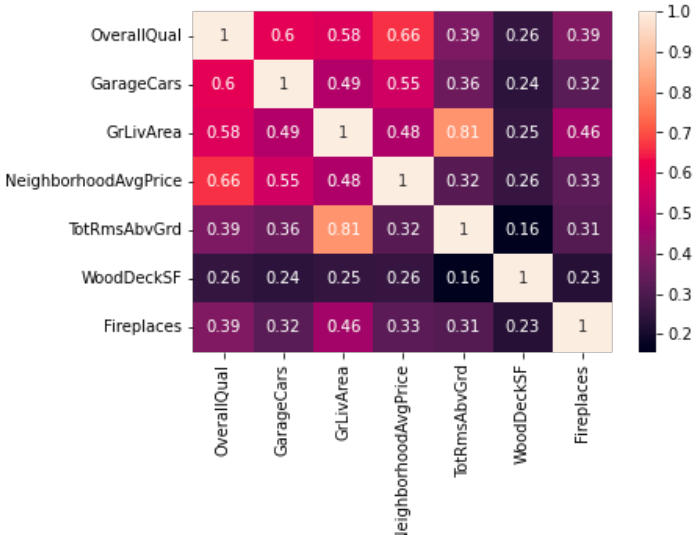

Figure D : Heatmap of selected parameters (all the attributes are below the moderate heat map level) - shows how well are the parameters are related to each other.

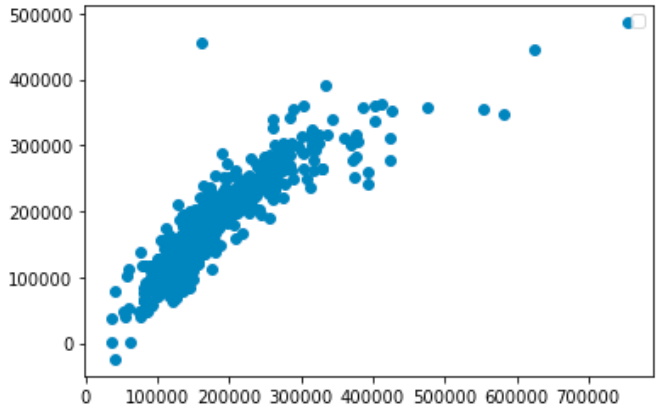

Figure E : Linear Regression and comparison of different input parameters to various metrics

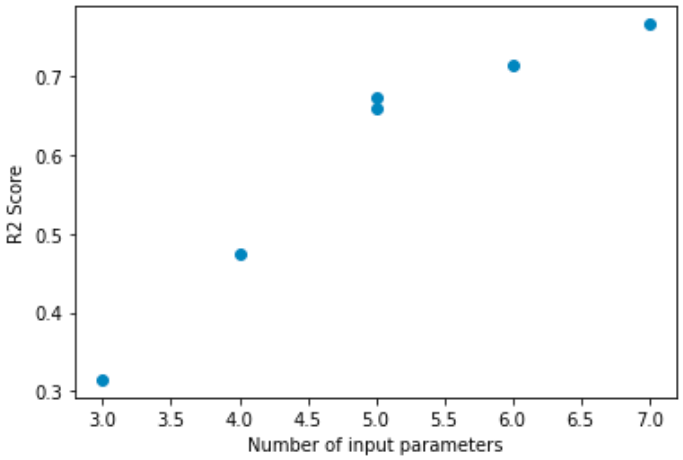

Figure F : Scatter plot shows increment of input numbers increases the r2 score

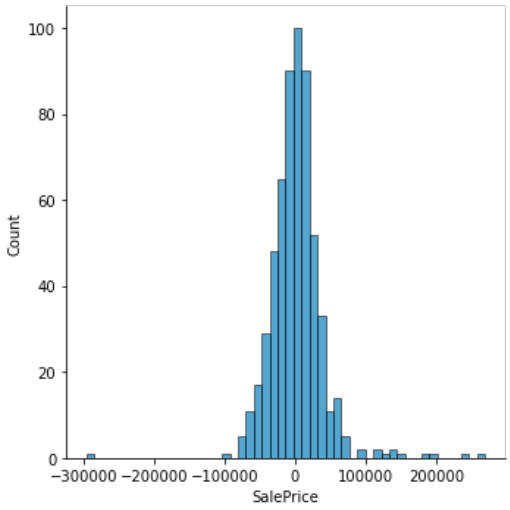

Figure G : Distibution Plot of absolute error of the ground truth versus the predictions

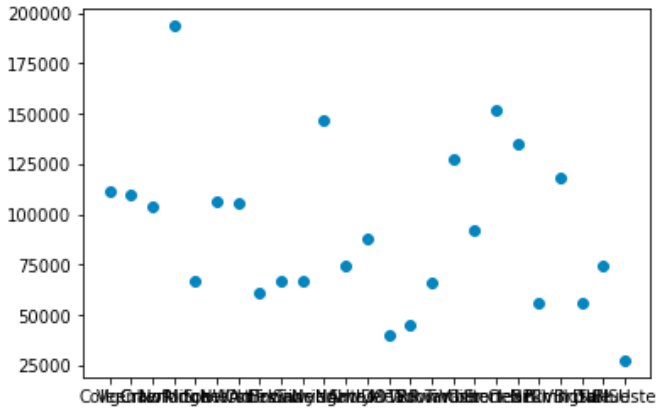

Figure H : Scatter plot for average house prices in each neighbourhood

Takeaways:

1. To help people who plans to invest in buying a new house so that they can compare the price range with the facilities available in the household.
2. House price predictions are beneficial for property investors to know the trend of the housing prices