METCS521- Final Project Submission
House Price Predictor

Professor: Alan Burstein
Team Members: Lakshitaa Sharma, Rajat Chawla, Nishka Sharesh Doddamani

Introduction:

One of Real Estate's star factors is House price forecasting. This factor derives useful knowledge from historical data of property markets. To analyse this, Machine learning techniques are applied to discover useful models.

As students, we always look for houses because of different opportunities available at various locations. Hence, we decided to work on the House Price Predictor because we agreed this was very relatable for us. We thought this would be a very helpful tool for new residents looking to buy a house. Using the data set provided, we predicted the price of houses based on information such as but not limited to House condition, sale price, number of rooms, size and area, etc. Our submission includes a total of the following files:

1. Download the data Script
2. Analysis Script
3. README that has been updated weekly as per the deliverable dates.

All these files guide the user through various steps about downloading the dataset, providing a correlation to find ideal numeric and categorical variables, Data partitioning, K-means clustering, Random Forest Classification, Multiple linear Regression models by changing choice of variables and making new features which will be used for predicting Sales Prices.

Downloading the Dataset:

To download the dataset, the user needs to run the download_dataset.ipynb script. The script writes a csv file to your working directory that will contain all the housing data. This csv file will be read into a dataframe by **project_analysis.ipynb** to be used for exploration and prediction of housing prices. In **project_analysis**.ipynb, each cell has to be run sequentially to reproduce results.

Extensions:

1. Exploratory Data Analysis and Correlations

Our initial steps include reading the dataset and checking Datatypes of columns using pandas, and NA Values in the dataset. Following this, we have used a describe function that checks Range and Statistics of data, column-wise. We sort the data by Sale Price to check if it contains outliers. Removal of NaN values from Sale Price has then been executed. As for data analysis and predictions, if 40% of our data is NA, we remove it because of overfitting risk. In our dataset, we had 50% values as NA for Sales Price.

We used a 'unique' function to check the existence of various levels in our categorical variables in order to label encode it later. Segregation of categorical variables on the basis of Nominal and ordinal

variables was done followed by discrete and continuous numeric data. We have changed the NaN values to ‘None’ for all nominal and ordinal columns; 0 for all the NAn values in numeric data.

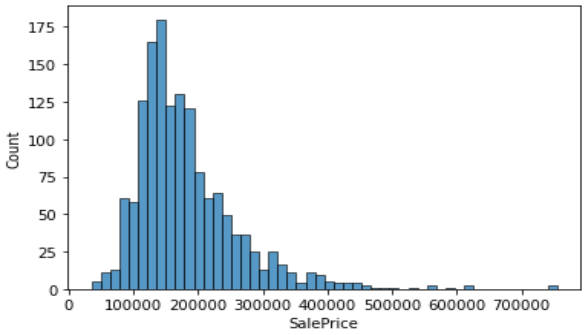


Figure 1: Histogram for Sales Price VS Count

For our first visualisation[Fig 1], we created a histogram that shows us the distribution of sale price (X-axis) against count of our data set (Y-axis) from the graph we can observe that it is right skewed with the lowest sale price lesser than $ 50000 and highest sale price more than $750,000. We also calculated our mean selling price to be $180,000 approximately and standard deviation to be $80,000. We also notice that most of the sale prices lie between 120,000 to 180,000.

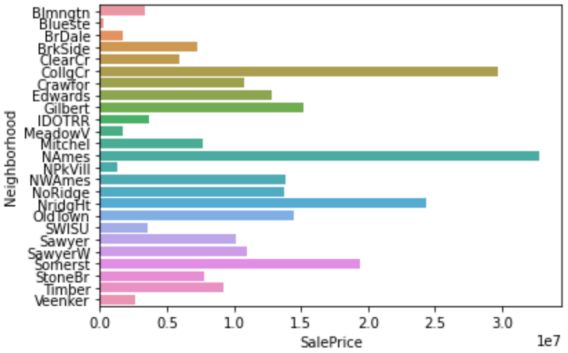


Figure 2 :Bar Plot for comparing Neighbourhoods

In our second visualisation, we wanted to see the column features difference and hence, we grouped the data based on different neighbourhood and plotted a bar graph for SalesPrice against Neighbourhood. We noticed that Names (32815593.0) had the highest total sale price whereas Blueste (275000.0) had the lowest. We also found out the top 5 highest and least expensive neighbourhoods.

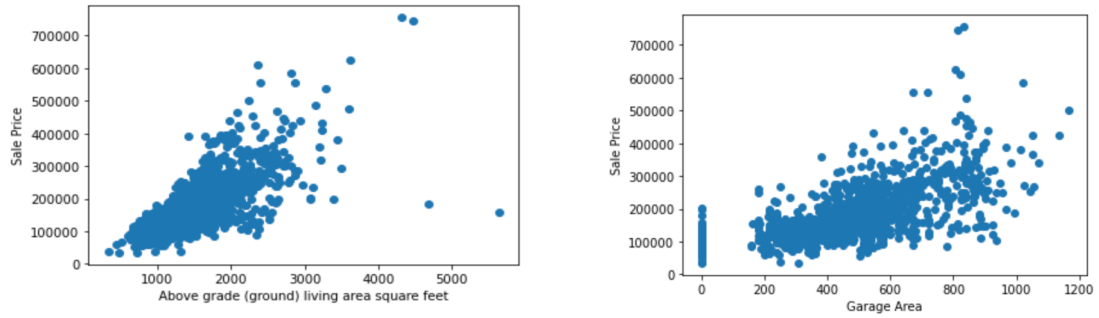


Fig.3 Scatter Plot

In our fifth scatter plot visualisation , we can perceive that as the living area increases, the sale price tends to increase too. Therefore, concluding that this variable does bring an impact. We can conclude

that there are various properties which have 0's implying to not having a garage. We also discovered that there are some outliers in the garage column. These outliers were removed as they would impact our regression model.

Numerical Data-

We looked at the correlations for numeric variables to check the ideal variables that impact our sale prices the most. We kept the variables that showed a correlation higher than 0.5 with the Sales Price

```
10 Highly correlated features with SalePrice:
OverallQual     0.790982
GrLivArea       0.708624
GarageCars      0.640409
GarageArea      0.623431
TotalBsmtSF     0.613581
1stFlrSF        0.605852
FullBath        0.560664
TotRmsAbvGrd    0.533723
YearBuilt       0.522897
YearRemodAdd    0.507101
Name: SalePrice, dtype: float64
```

Categorical Data-

We built various visualisations to check the levels of our variables and its impact on sales prices which will give us a clearer understanding of variable importance and build domain knowledge. For example- sale conditions had 5 levels from which partial had the highest impact on sale price whereas others had approximately similar impact. We converted the partial level as 1 and the rest of the levels as 0.

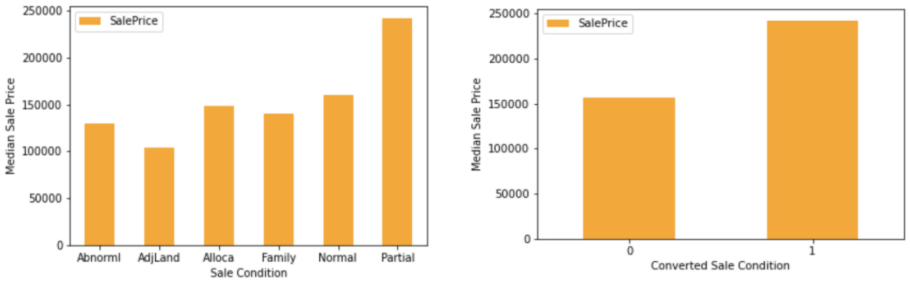


Fig.4 Bar Graph

Label Encoding-

After studying the levels in variables and its impact on sale prices, we needed to convert our categorical data to numeric for running our linear regression model. We used a label encoder from sk.learn to convert levels of variables into numeric in the sequence of higher proportion in categorical columns. For example, the neighbourhood which had the highest sales price was given the most value (based on number of neighbourhoods) and the one with lowest was given 0.

Model Preparation

Prior to building our linear regression model, we looked at the correlation matrix between all the variables to see if there exists multiple variables which are highly correlated with one another. We did this step to eliminate one of the variable from the family of highly correlated pairs of variables to build a relevant and reliable model.

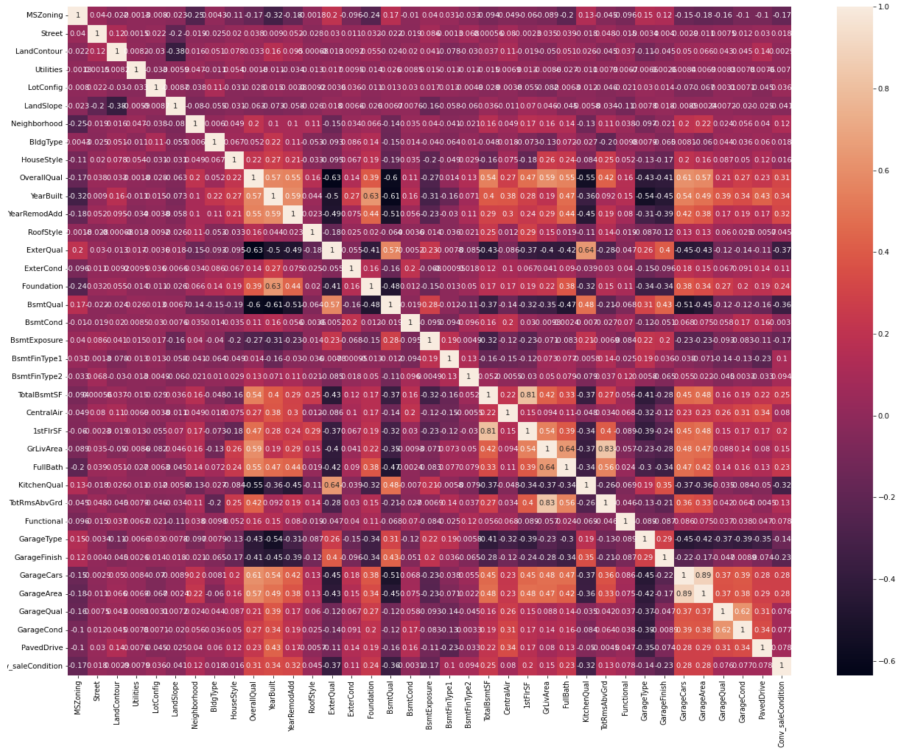


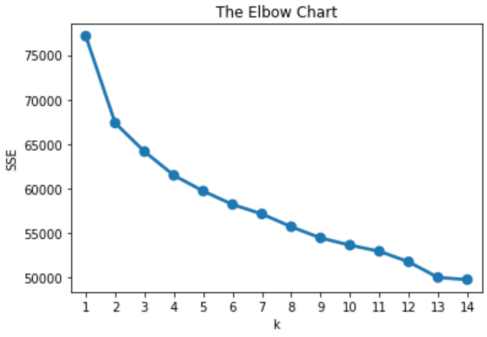
Figure 5 : Correlation Heatmap

Data Partition

We split the dataset into train and test by a ratio of 60:40 using sklearn.Splitting the data is an essential factor because we build our model based on the training set so that we can measure our model's performance based on our test (validation) data. This helps to check if our model is overfitting or meeting the satisfactory error rate.

3. K-means Clustering

We built a clustering model to divide our data into segments/groups on the basis of similarities between variables in respective clusters.



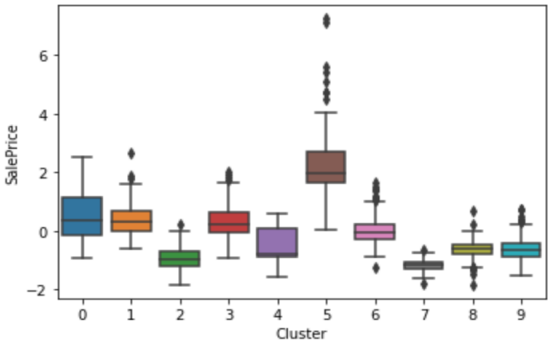


Figure 6.1: Clustering K VS SSE

Figure 6.2: Box Plot Cluster VS Sales Price

In our seventh visualisation(Fig 6.1), we created an elbow chart that helped us determine the optimal k-value in the clustering and how many clusters would be optimal for our dataset. We also built boxplots of clusters based on sale prices (Fig.6.2)and found out that cluster 5 has the highest median sales prices whereas cluster 2 and 7 had lowest median sale prices.

4. **Random Forest Classification**

We chose a random forest model classification to study the feature importance in our data. This model classified our variables into coefficient values in the order of decreasing order. This coefficient shows the impact of variables on sales prices by splitting the data into tree splits making decision nodes.

5. Multiple Linear Regression Model

We made four linear regression models.

Model1- We used all the numeric and categorical variables that were strongly correlated with sales price.

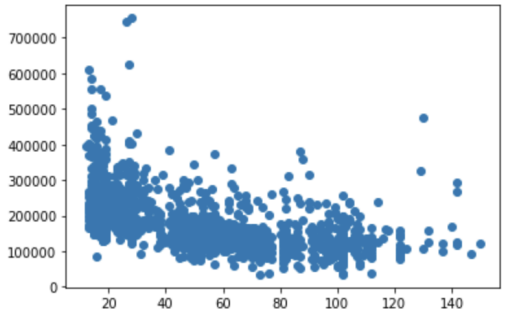
Model2- We used variables that showed higher coefficient from random forest model

Model3- we used only 4 variables that had highest coefficient value in random forest model

Model4- we made new features such as-

- Age- subtracting current year by year built
- Age_renovation-subtracting current year by year build
- Total basement score- combining all the basement quality, condition, finished quality
- Total garage score- combining all garage quality, condition
- Total exterior score- combining all exterior condition, quality

Our model 2 performed the best with an R squared value of 0.80 and Root mean square error of 35436.58. Also, we were able to generate the regression equations by getting the intercepts and coefficients. Below picture shows age versus sale price.



Conclusion:

In a nutshell, apart from the correlation technique, we came up with the most important feature selection with the help of a random forest model that gave us the variables that are the most relevant in determining sales prices. Building different regression models helped us to derive the difference between accuracies and predictions. In order to accomplish making classification variables more useful for a regression analysis, we must clean up the data and deal with categorical variables thoroughly including keeping outliers and multicollinearity in mind.

We think that this project has really helped us understand the way Python is a useful tool to analyse data. We recognized different means of explorations and machine learning algorithms. We now know that Regression and feature engineering was the most complex and challenging thing according to us but we were able to work it through. We understood that there is more to coding than just writing clean code.