

**House Price Prediction**

Submitted by:

Pallavi Bajpai

**ACKNOWLEDGMENT**

I have taken efforts in this project. However, it would not have been possible without the kind support and help of many individuals and refrences. I would like to extend my sincere thanks to all of them.

I am highly indebted to my institution i.e. Data Trained and the mentors and chat support team for their guidance and constant supervision as well as for providing necessary information regarding the project & also for their support in completing the project.

I would like to express my gratitude towards my parents & member of Flip and work for their kind co-operation and encouragement which help me in completion of this project.

My thanks and appreciations also go to Anurag Lal whose git reference acted as a major helping hand in developing the project.

**INTRODUCTION**

* Business Problem Framing

A US-based housing company named Surprise Housing has decided to enter the Australian market. The company uses data analytics to purchase houses at a price below their actual values and flip them at a higher price. For the same purpose, the company has collected a data set from the sale of houses in Australia. The data is provided in the CSV file below.

The company is looking at prospective properties to buy houses to enter the market. You are required to build a model using Machine Learning in order to predict the actual value of the prospective properties and decide whether to invest in them or not. For this company wants to know:

* Which variables are important to predict the price of variable?
* How do these variables describe the price of the house?

Business Goal- You are required to model the price of houses with the available independent variables. 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.

* Conceptual Background of the Domain Problem

A good understanding of programming concepts along with some mathematic basic concepts like statistics , probability are very helpful. Thorough understanding of machine learning and the different models is also very important to solve this problem.

* Review of Literature

Considerable about of online research is done in order to understand the problem and requirement of ML in solving this problem.

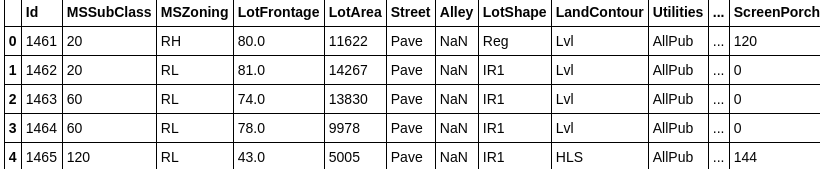
* Motivation for the Problem Undertaken

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 comes as a very important tool to 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. Predictive modelling, Market mix modelling, recommendation systems are some of the machine learning techniques used for achieving the business goals for housing companies. Our problem is related to one such housing company.

**Analytical Problem Framing**

* Data Sources and their formats

The dataset consists of both numerical and categorical variables. There is a total of 79 explanatory variables describing every aspect of the residential home. Overview of the dataset:



Selected features are : ‘BsmtFinType1’, ’BsmtFinType2', ’BsmtCond’, ‘BsmtQual’, ’Electrical’, ’MasVnrArea’, ’MasVnrType’, ‘GarageCond’, ’GarageFinish’, ’GarageQual’, ’GarageType’, ’GarageYrBlt’, ’MiscFeature’, ’Fence’, ’PoolQC’, ’PoolArea’, ’Alley’, ’FireplaceQu’ . (**Note:** This is just one of the possible approach. The selected features may vary).

* Data Preprocessing Done

Fair amount of data cleaning process was also involved. This includes checking for space and null values and then dealing with them. Also our available data has 81 columns out of which ‘BsmtFinType1’, ’BsmtFinType2', ’BsmtCond’, ‘BsmtQual’, ’Electrical’, ’MasVnrArea’, ’MasVnrType’, ‘GarageCond’, ’GarageFinish’, ’GarageQual’, ’GarageType’, ’GarageYrBlt’, ’MiscFeature’, ’Fence’, ’PoolQC’, ’PoolArea’, ’Alley’, ’FireplaceQu’ were the considered features while solving the problem. Outliers were also detected and removed.

* State the set of assumptions (if any) related to the problem under consideration

Selected features are : ‘BsmtFinType1’, ’BsmtFinType2', ’BsmtCond’, ‘BsmtQual’, ’Electrical’, ’MasVnrArea’, ’MasVnrType’, ‘GarageCond’, ’GarageFinish’, ’GarageQual’, ’GarageType’, ’GarageYrBlt’, ’MiscFeature’, ’Fence’, ’PoolQC’, ’PoolArea’, ’Alley’, ’FireplaceQu’ . (Note: This is just one of the possible approach. The selected features may vary).

* Hardware and Software Requirements and Tools Used

Python code was written in Jupiter notebook. Below are the libraries needed in the process.

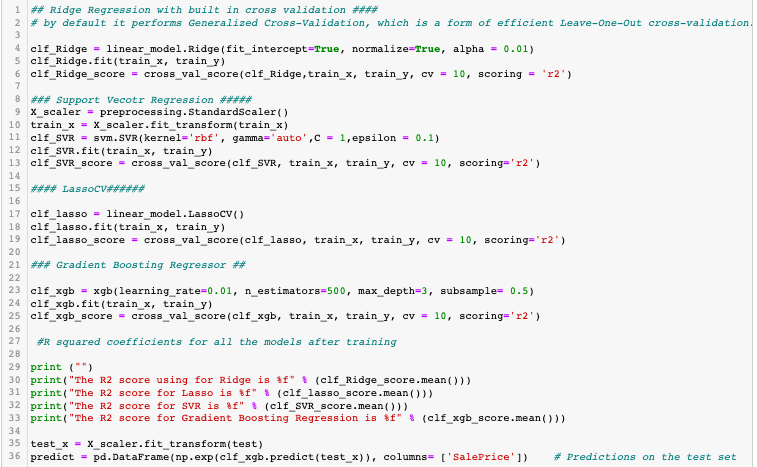
Pandas, numpy, matplotlib.pyplot, sklearn.preprocessing, sklearn.model\_selection.cross\_val\_score, sklearn.linear\_model, sklearn.linear.svm, sklearn.ensemble.GradientBoostingRegressor, sklearn.metrics.confusion\_matrix,accuracy\_score, sklearn.metrics.f1\_score, scipy.stats, seaborn

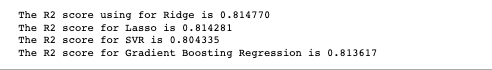
**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)
* Correlation heatmap was drawn to find the correlation between features and accordingly we selected few out of all features.
* Outliers were detected with the help of scatterplot and boxplot and were removed.
* Skewness was checked using histogram.
* Testing of Identified Approaches (Algorithms)

Below algorithms were used for training and testing of data

* Ridge
* Lasso
* SVR
* Gradient Boosting Regression
* Run and Evaluate selected models



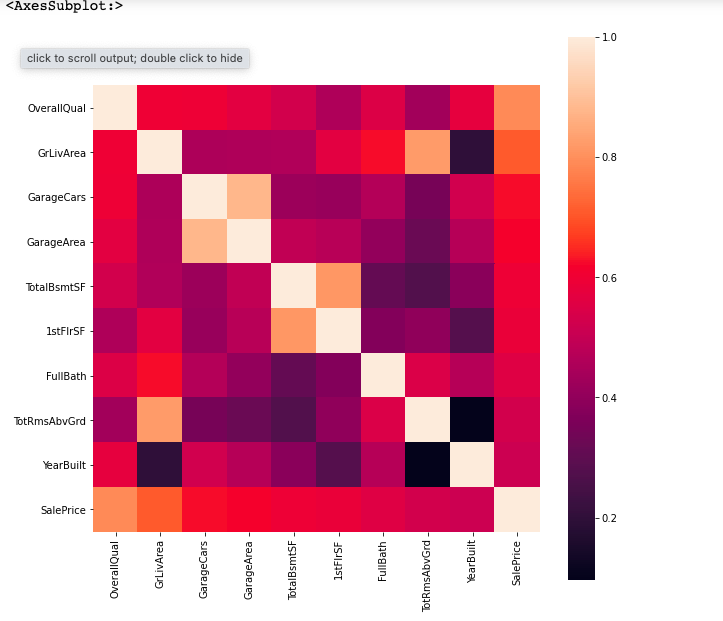


* Key Metrics for success in solving problem under consideration

R2 score was the key metric used to finalize the model.

* Visualizations

Correlation heatmap:

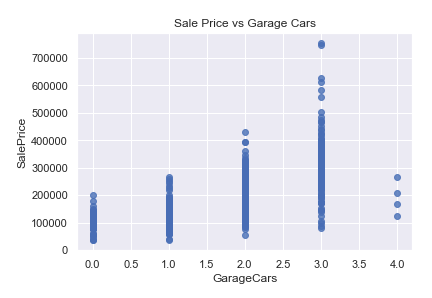


Here GarageArea, TotalRmsAbvGrd, 1stFlrSF, KitchenQual, ExternalQual were dropped because they correlated less with sales price compare to the other variables they correlated with

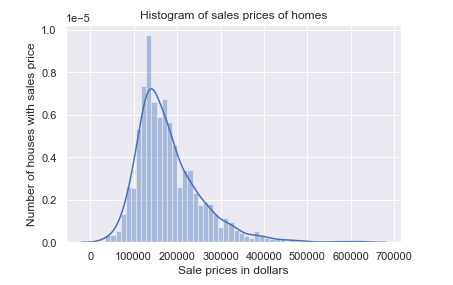
scatter plots for the continous variables and boxplots for the categorical variables to detect outliers

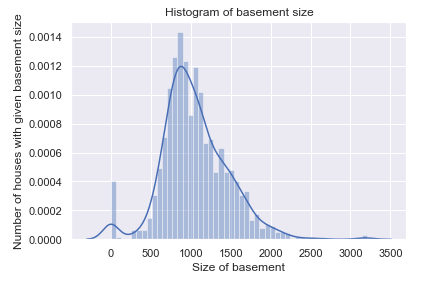




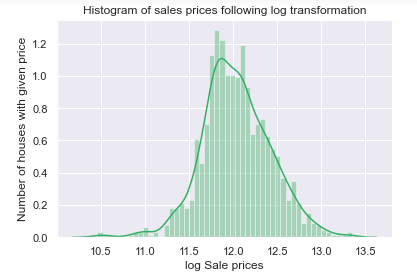


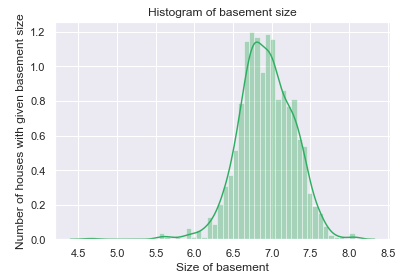
Four outliers in the graphs above will be dropped. These corresponding to ground living area > 4500 and sales price between 7000 and 8000 in Year Built graph.The isolated TotalBsmtSF point > 6000 is also included as part of these outliers which will be removed.

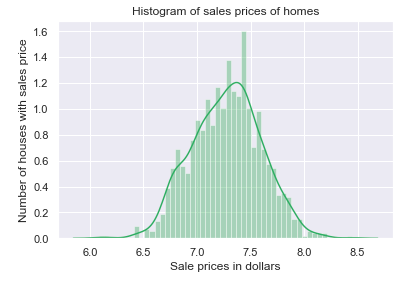












The distrbutions are now centred and more normally distributed and less skewed.

* Interpretation of the Results
* Here GarageArea, TotalRmsAbvGrd, 1stFlrSF, KitchenQual, ExternalQual were dropped because they correlated less with sales price compare to the other variables they correlated with
* Four outliers in the graphs above will be dropped. These corresponding to ground living area > 4500 and sales price between 7000 and 8000 in Year Built graph.The isolated TotalBsmtSF point > 6000 is also included as part of these outliers which will be removed.
* The distrbutions are now centred and more normally distributed and less skewed.

**CONCLUSION**

* Key Findings and Conclusions of the Study

R2 score was highest in Lasso and Gradient Boosting Regression followed by Ridge and SVR.

* Learning Outcomes of the Study in respect of Data Science
* Correlation heatmap was drawn to find the correlation between features and accordingly we selected few out of all features.
* Outliers were detected with the help of scatterplot and boxplot and were removed.
* Skewness was checked using histogram.

We have finalised Gradient boosting regressor algorithm as it works great in this case.

* Limitations of this work and Scope for Future Work

BsmtFinType1’, ’BsmtFinType2', ’BsmtCond’, ‘BsmtQual’, ’Electrical’, ’MasVnrArea’, ’MasVnrType’, ‘GarageCond’, ’GarageFinish’, ’GarageQual’, ’GarageType’, ’GarageYrBlt’, ’MiscFeature’, ’Fence’, ’PoolQC’, ’PoolArea’, ’Alley’, ’FireplaceQu’ . This is just one of the possible approach. The selected features may vary.