

House Price Prediction

Submitted by:

Aamir Rahman

**ACKNOWLEDGMENT**

I would like to express my special thanks of gratitude to FlipRobo, who gave me the golden opportunity to do this wonderful project of House Prediction Price.

Secondly, I would also like to thank DataTrained teachers and mentors who teach me all the basic and professional concept for building the project.

**INTRODUCTION**

* Business Problem Framing

A benefit to this study is that we can have two clients at the same time. However, in this case, we can have both clients with no conflict of interest.

1. Client House buyer: This client wants to find their next dream home with a reasonable price tag. They have their locations of interest ready. Now, they want to know if the house price matches the house value. With this study, they can understand which feature (ex. Number of Bathrooms, location, etc.) influence the final price of the house. If all matches, they can ensure that they are getting a fair price.
2. Client House Seller: Think of the average house-flipper. This client wants to take advantages of the feature that influence the house price most. They typically want to buy a house at a low price and invest on the features that will give the highest return. For example, buying a house at a good location but small square footage. The client will invest on making rooms at a small cost to get a large return.

* Conceptual Background of the Domain Problem

To understand this problem the developer should have a basic knowledge related to house selling trends that has been carried out from years. Like what are the most important thing that a buyer look at the time of buying house like locations, number of bedrooms, square footage, etc. Sometimes, some buyer focus only one thing that is most important to them. Some buyer prefer location over anything, some prefer quality, some prefer square footage. So before solving this problem the developer should be aware from these things.

* Review of Literature

This is a comprehensive summary of the research done on the topic. The review should enumerate, describe, summarize, evaluate and clarify the research done.

* Motivation for the Problem Undertaken

Desire to face the challenge in solving the unsolved problems. And also the objective behind taking this project is to improve my skill in statistical as well as in analytical knowledge in machine learning and artificial intelligence.

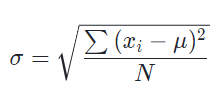
**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

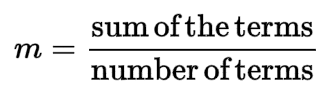
There are various mathematical modelling is carried out in this project. For example:

* finding the mu and sigma for our target variable using the scipy built-in function scipy.stats.norm. And making it normally distributed.

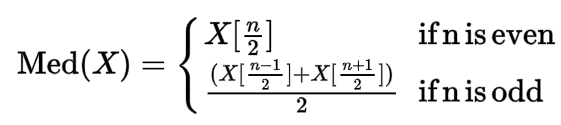
The formula for finding sigma:



The formula for finding mu:



* Treating the Null values using median and mode:





* Data Sources and their formats

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.

The given data is given in form of numerical and categorical format.

* Data Preprocessing Done

In the given data there are some ambiguities present. In order to develop a machine learning model, data pre-processing is the most important step to build an effective machine learning model.

These are some data pre-processing which is done on our data:

* **Null Values:** This is the first thing that we should check in our data. Null values can be treated in two ways either replace it or remove it. If a column contains more than 75% null values then we should better remove it. And we replace the null value by using mean, median or mode of that column.
* **Outliers:** An outlier is any data point that is distinctly different from rest of your data points. When you are looking at a variable that is relatively normally distributed, you can think of outliers as anything that falls 3or more standard deviations from its mean.

There are three causes of outliers – data entry/ an experiment measurement errors, sampling problems, and natural variations.

* **Fixing skewed features:** We fix the target variable to be more normal so that our models will be more accurate when making predictions. We use the numpy function log1p which applies log(1+x) to all elements of the column.
* **Converting label column into numerical column:** Machine understand numbers, not text. We need to convert each text category to numbers in order for the machine to process them using mathematical equations. We use label encoding in our data pre-processing to convert the label data into numerical data.

Label Encoding**:** It is popular encoding technique for handling categorical variables. In this technique, each label is assigned a unique integer based on alphabetical ordering.

* Data Inputs- Logic- Output Relationships

In machine learning model the relation between input and output can be defined in 3 ways, either the relation is positive, negative or no relation. The type of relation between input and output can be find by using the correlation value. We can build the correlation matrix by using heatmap, if the value comes near to 1 then there is positive relationship, if the value comes near to -1 then there is negative relationship, and if the value comes near to 0 then there is no relationship.

* Hardware and Software Requirements and Tools Used

List of tools used in the project:

1. Scientific computing libraries: Pandas, Numpy, Scipy.
2. Visualization libraries: Matplotlib, Seaborn.
3. Algorithmic libraries: Scikit learn, Stats model.
4. Programming language: Python.
5. IDE used: Jupyter notebook.

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

The dataset contains 81 columns. My first approach after data pre-processing is to make the feature selection. The advantage of feature selection is:

1. It enables the machine learning algorithm to train faster.
2. It reduces the complexity of a model and makes it easier to interpret.
3. It improves the accuracy of a model if the right subset is chosen.
4. It reduces overfitting.

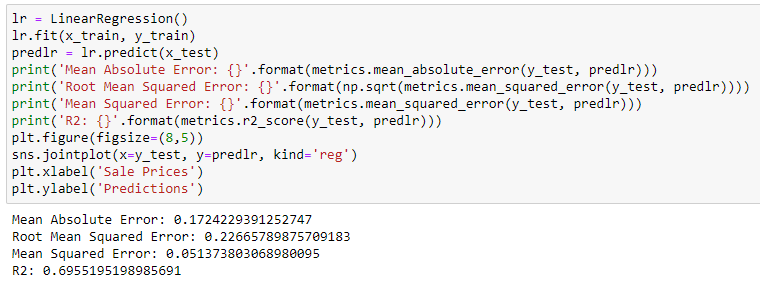
I use the mutual\_info\_classif algorithm for feature selection and print all the feature names with their importance values. The value comes between 1 to 5, so I set the threshold to 1, if any feature which is less than 1 is not much important.

* Testing of Identified Approaches (Algorithms)

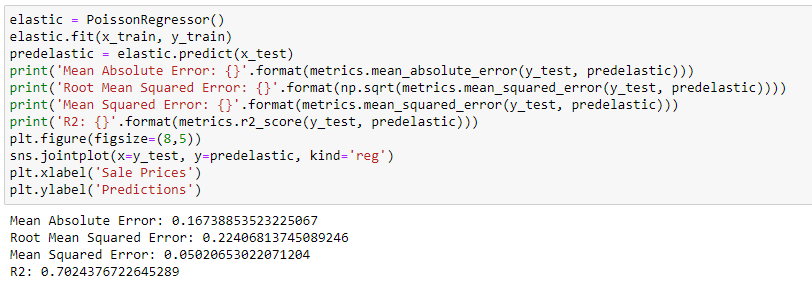
1. **Linear Regression:** Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an exploratory variable, and the other is considered to be a dependent variable.
2. **Poisson Regression:** Poisson regression is a generalized linear model form of regression analysis used to model count data and contingency tables. Poisson regression assumes the response variable Y has a Poisson distribution, and assumes the logarithm of its expected value can be modelled by a linear combination of unknown parameters.
3. **Support Vector Regression:** The support vector regression uses the same principles as the support vector machine for classification, with only a few minor differences. The main idea s to minimize error, individualizing the hyper plane which maximises the margin, keeping in mind that part of error is tolerated.
4. **Ridge:** Ridge regression is a model tuning method that is used to analyze any data that suffers from multicollinearity. This methods performs L2 regularization. When the issue of multicollinearity occurs. Least-squares are unbiased, and variance are large, this result in predicted values to be far away from the actual values.
5. **Random Forest Regression:** It is a supervised learning algorithm that uses ensemble learning method for regression. Ensemble learning is a technique that combines predictions from multiple machine learning algorithm to make a more accurate prediction than a single model.
6. **AdaBoost Regression:** An adaboost regression is a meta-estimator that begins by fitting a regressor on the original and then fits additional copies of the regressor on the same dataset but where the weights of instances are adjusted according to the error of the current prediction.

* Run and Evaluate selected models

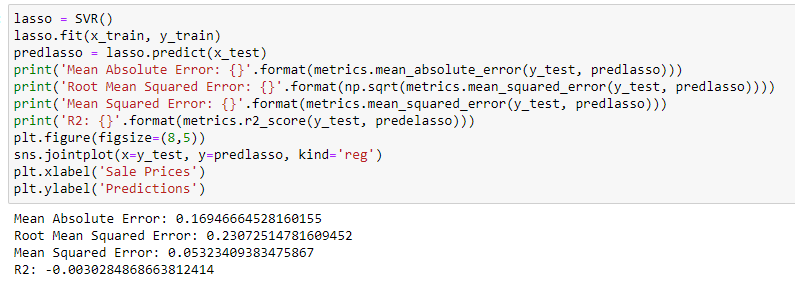
1. **Linear Regression:**



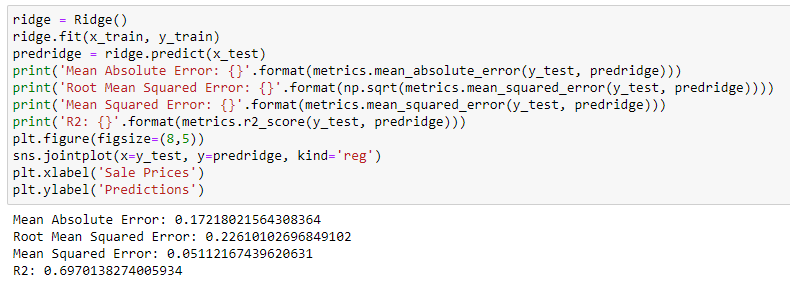
1. **Poisson Regression:**



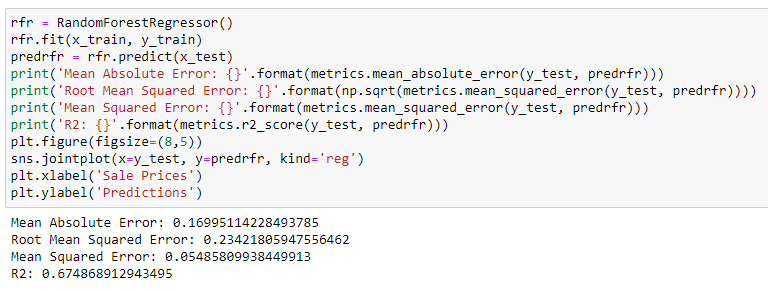
1. **Support vector regression:**

****

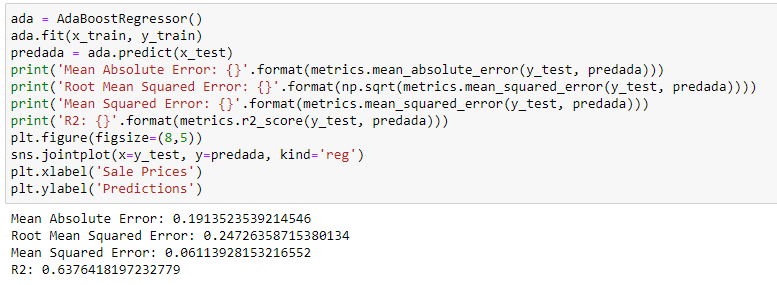
1. **Ridge regression:**



1. **Random forest regression:**



1. **AdaBoost regression:**



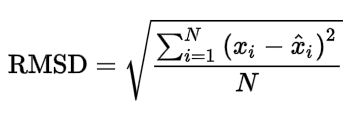
* Key Metrics for success in solving problem under consideration

This project is regression problem. These are the metrics used in project:-

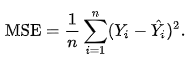
1. **Mean Absolute Error:** It is a measure of error between paired observations expressing the same phenomenon. Example of Y versus X include comparisons of predicted versus observed, subsequent time versus initial time, and one technique of measurement versus an alternative technique of measurement. MAE is calculated as:



1. **Root mean squared error:** It is frequently used measure of the differences between values (sample or population values) predicted by a model or an estimator and the values observed. RMSE is calculated as:

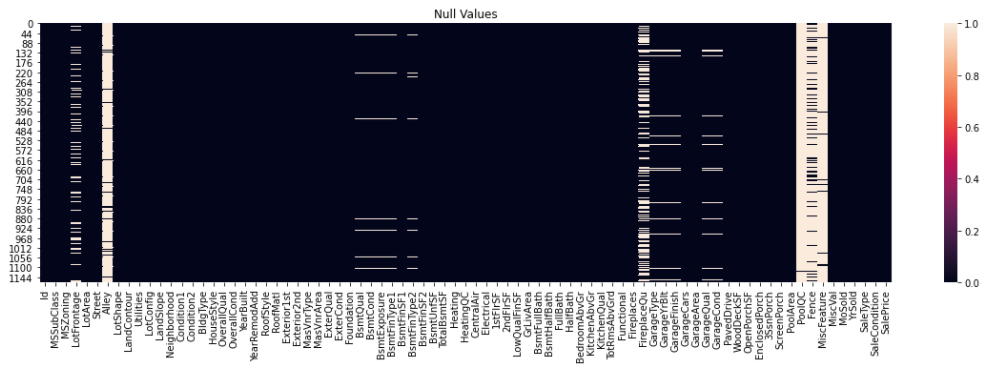


1. **Mean Squared Error:** It measures the average of the squares of the error, that is, the average squared difference between the estimated values and the actual value.MSE is a risk function, corresponding to the expected value of the squared error loss.

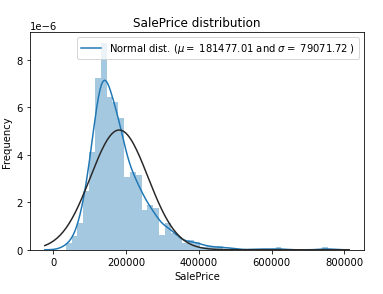
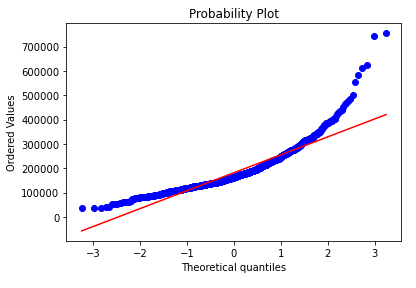


1. **R2 Score:** The main purpose of this metrics is either in prediction of future outcomes or the testing of hypotheses, on the basis of other related information. It provides a measure of how well observed outcomes are replicated by the model, based on the proportion of total variation of outcomes explained in the model.

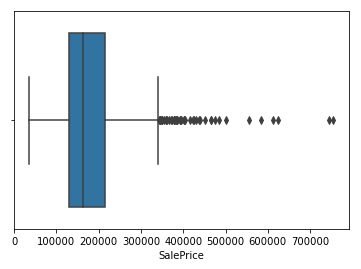
* Visualizations



From the above heatmap we can see that there are certain columns which present more than 50% of null values.

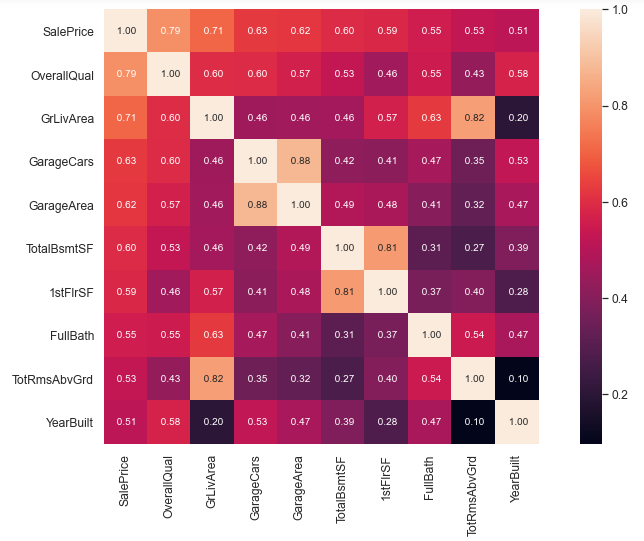
The two graphs mentioned above are density distribution plot (left) and probability distribution plot (right). Looking at the kurtosis score, we can see that there is very nice peak. However, looking at the skewness score, we can see that the sale prices deviate from the normal distribution. We want our data to be as normal as possible.



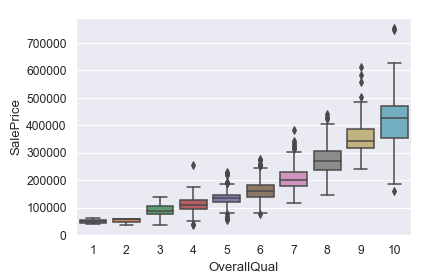
The boxplot is shown in the above picture which shows that there are certain

entries in which the sale price is very much high that the plot is showing them

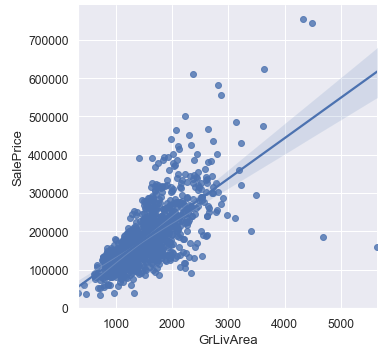
as an outlier.



The heatmap shown above shows the top 10 correlated values.



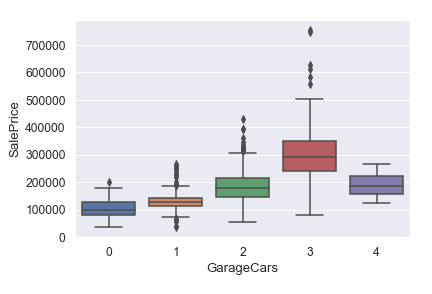
The above boxplot show a positive linear relationship with the target variable.



The above scatterplot shows the relationship between above grade (ground)

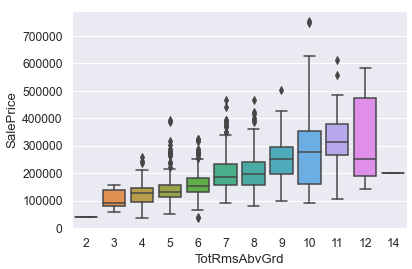
living area square feet and Sale price. We can see that there are two points on

the bottom right which is behaving opposite to our imagination.



In the above boxplot we can see that the price of the 4-car garage capacity is

less as compare to 3-car garage capacity.



We can conclude from the above boxplot that houses with more than 11 rooms

has the same price compared to 10 rooms.

* Interpretation of the Results

In the given project, data pre-processing, data visualization and modelling is done.

1. We can conclude from the data pre-processing that there are many null values present in the dataset, if that null value is present with some actual true data then our model can perform even more accurate. The data contains more labelled data due to which we perform the label encoding on independent variable, but more often we use one hot encoding for the independent variable. The data could be more precise.
2. In the visualization phase firstly we saw that our target variable is not normally distributed. From the heatmap we saw that there are multicollinearity is present in our data. In GrLiveArea column we spot out the noise present which is behaving just opposite to our imagination. In Garage cars the price of the 4-car garage capacity is less as compared to 3-car garage capacity.
3. In modelling phase, we train our data with Linear Regression, Poisson Regression, Support vector regression, Ridge regression, Random forest regression, Adaboost regression. And perform a cross validation on these model. Out of these model support vector regression performs worst while Ridge regression performs very well. Then we perform the hyper parameter tuning of the best model.

**CONCLUSION**

* Key Findings and Conclusions of the Study

This study employs machine learning techniques, to develop a price prediction model for housing problems. It uses a rather large publicly available dataset of real estate transaction. The regression model performances of the model are compared with one another. The empirical result shows that the Ridge algorithm provides superior performances for all metrics under study, the coefficient of determination of r2 score, and the computational time.

* Limitations of this work and Scope for Future Work

The Study can be enlarged in a subsequent research by increasing the dataset size so potentially uncovered details and feature of the dataset and of this study can be addressed. An increased dataset would potentially be good enough for employing deep neural networks, which can assure that more in-depth analysis on the housing price prediction can be performed. Then, the enlarged housing price prediction problem can be tackled as a classification problem.