

**Housing Project**

Submitted by:

Shailendra Pratap Singh

**ACKNOWLEDGMENT**

Inside the Prediction of the dataset () I am guided by mine mentor (Sajid Choudhary), I also took some help from Kaggle, UCI Machine Learning Repository and YouTube in the completion of project.

**INTRODUCTION**

**1. 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?

**2. Conceptual Background of the Domain Problem**

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.

**3. Review of Literature:**

(i) DATA ANALYSIS  
(ii) FEATURE ENGINNERING  
(iii) FEATURE SELECTION  
(iv) MODEL BUILDING  
(v) MODEL DEPLOYMENT

**4. Motivation for the Problem Undertaken**

Houses are one of the necessary needs 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**

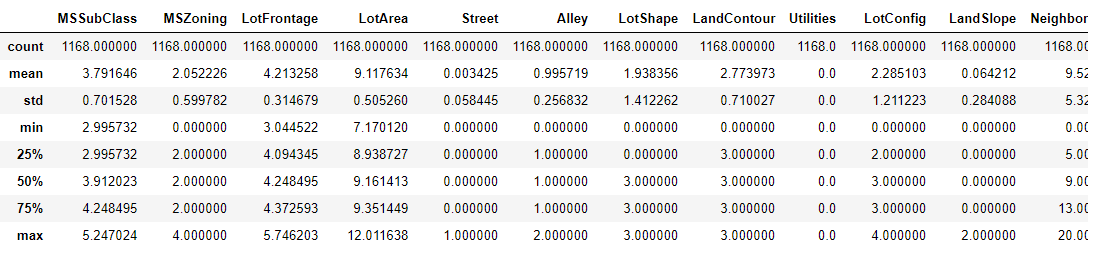
1. **Mathematical/ Analytical Modeling of the Problem**

Inside the dataset we do several Mathematical/Statistical Modeling.

(i) We check the null values inside the whole dataset and its percentage in

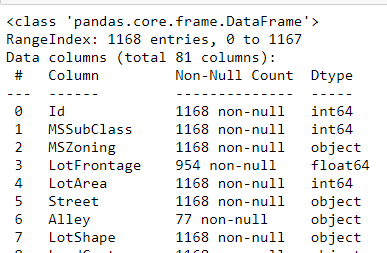
each and every feature.

(ii) Fill the null value according with suitable value like Mean, Median, Mode.  
  
(iii) We check the statistical parameter of the dataset like Mean, Median, Mode, Percentile(25th,50th,75th), Maximum Value, Minimum Value and Count.



**2. Data Sources and their formats**

We get 2 kind of data one is Train-Data and another one is Test -Data. The dataset contains all kind of value it contains numerical (int64) as well as categorical data(object) the numerical data contain temporal data(date-time), Discrete data(int64) and continuous data(float64).



3. **Data Pre-processing Done**

(i) It is the first step that is follow in data analysis. In which we find the null-value in all the feature of dataset. Fill the null values according to suitable value like mean median mode.

(ii) If null-values percentage is more than 50% in any column we drop that particular column. Drop the duplicate value inside the column.

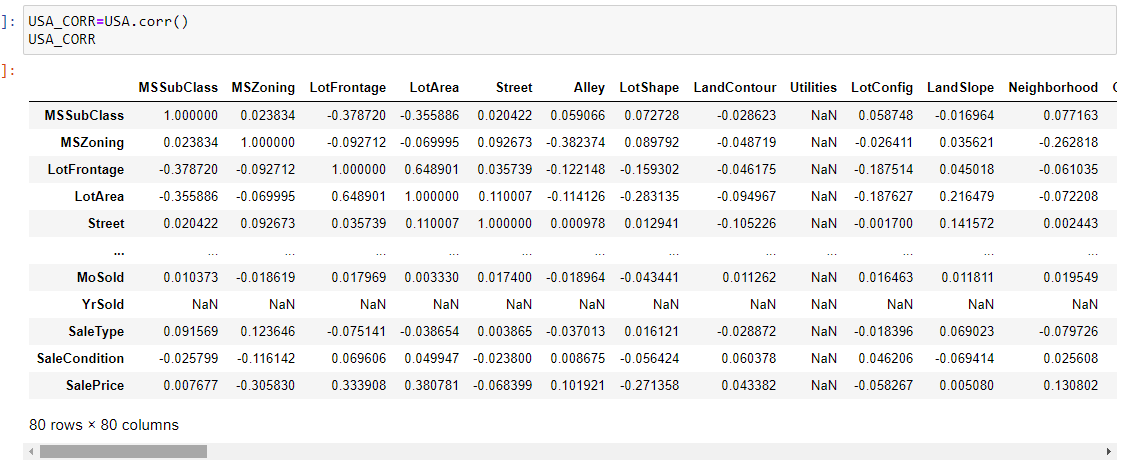
(iii) Check the skewness of the dataset and remove it by different type of transformation in my case I use Log-Transformation. For the data frame to come under the gaussian-transformation.

So by this we make the model for development.

4. **Data Inputs- Logic- Output Relationships**

Inside the dataset input-logic-output relationship is find by the help of correlation. It shows the feature inside the dataset how possible corelated to each other and dependent feature. So, from the corelation we get that some feature is positively corelated with dependent feature and some are negatively correlated so positively correlated data is important.

If we talk about the dataset than the heatmap is not possible because the number of columns is so large. But we corelated the feature by data frame.



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

No there is no assumption is required everything is based on knowledge after the full felling the missing values we same that the different type of data inside the data frame so we apply encoding technique like label encoding which convert the object data type into integer after that we use Minmaxscaler which standardize the value in between 0 and 1. Also use Pca (principal component analysis) to reduce the fear of curse of dimensionality.

[This Photo](https://en.wikipedia.org/wiki/Numerical_Python) by Unknown Author is licensed under [CC BY-SA](https://creativecommons.org/licenses/by-sa/3.0/)

6. **Hardware and Software Requirements and Tools Used**

The whole process is done on my laptop on jupitarnotebook, in the whole process we import different kind of library like NumPy, pandas, matplotlib, seaborn, we also import SciPy for statical analysis. For the machine model building we used skit learn.



[This Photo](https://scipy.in/2015) by Unknown Author is licensed under [CC BY-SA](https://creativecommons.org/licenses/by-sa/3.0/)



**MODEL/S DEVELOPMENT AND EVALUATION**

1. **Identification Of Possible Problem-Solving Approaches (Methods)**

In feature engineering process basically we standardize the data into standard format mins all the feature datatype will be the same datatype and the value is not more variate from each other.so here we apply encoding process in which we use label encoder, minmaxscaler and pca.as I mentioned earlier.

After completion the feature engineering process our dataset is ready for the   
model development so for applying the machine learning model we have   
to split the data into two parts one is independent feature which contain   
the all the feature except output the dependent feature.

**2. Testing of Identified Approaches (Algorithms)**

For the training and testing part we import train\_test\_split from sklearn which split our dataset for the training purpose and testing purpose. So in this we pass x and y feature also define the test\_size and random\_state after this aur data is randomely split into train and test data. Here inside the dataset we have to train the model from train data and predict the result by the test data. So here into the train data train test split function is applied while into test data we directly apply into the prediction.

3. **Run and Evaluate Selected Models**

For the dataset the dependent feature is in numerical datatype(int64) so here we apply regression so according to the dataset the output data is in discrete form so here we apply linear regression model

4. **Key Metrics for Success in Solving Problem Under Consideration**

For the accuracy of model we use key-metrics here inside the dataset we use Linear Regression model so here we use.

1. **Mean absolute error**

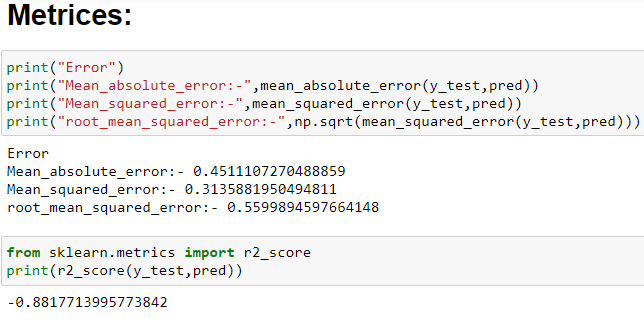
In statistics, mean absolute error (MAE) is a measure of errors between paired observations expressing the same phenomenon.

**(II) Mean squared error**

It is calculated by taking the average of the square of the difference between the original and predicted values of the data.

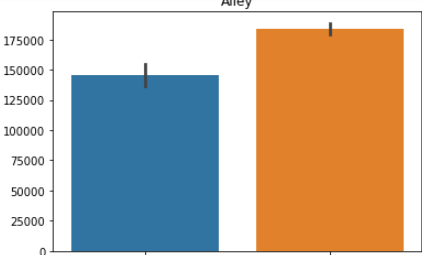
1. **Root mean squared error**

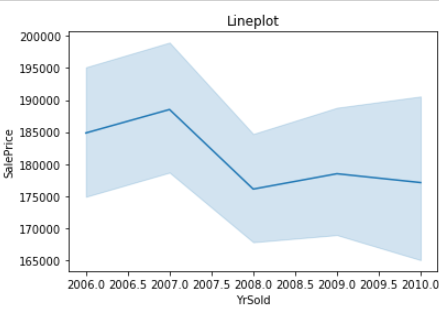
The Root Mean Square Error or**RMSE** is**a frequently applied measure of the differences between numbers (population values and samples) which is predicted by an estimator or a** mode.

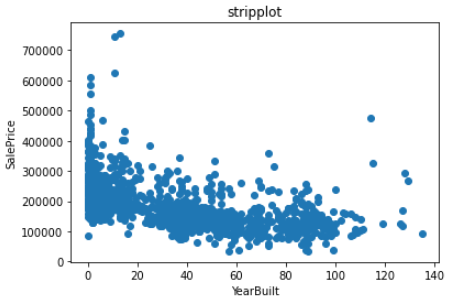
****

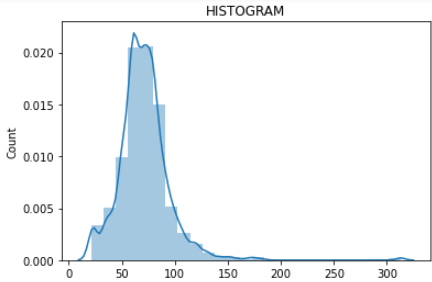
5. **Visualizations**

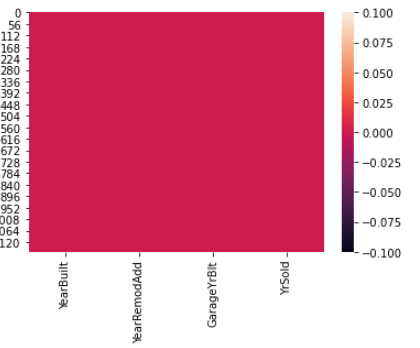
For the visualization purpose we use several method basically we use matplotlib and seaborn library for the visualization purpose we use bar plot, histogram, heatmap with annotation we also use prob plot from SciPy and dist. plot from seaborn.









****

1. **Interpretation of the Results**

In the visualization section first we have conclude the nan values so the section where value is given we replace it by 1 and there nan value is present replace it by 0 after that plot the bar graph with the sales price for check that the credibility of nan value for the out so by this we do bivariate analysis

Line plot for the comparison of temporal data year built with sales price where we analyse that when the year will passed the price of house is reduced. we also doing bivariate analysis of temporal data by the help of scatter plot.

By the help the heatmap we find the null values inside the dataset, and dist. plot and prob plot gives us the information about skewness if skew is present inside the feature we did not for proper bell shaped curve.

**CONCLUSION**

1. **Key Findings and Conclusions of the Study**

Study shows a comparison between regression problem and artificial neural network when predicting house prices in Surprise Housing in united state the result will promising for the public data. Due to it being rich with feature and having strong corelation, where a local data give a worst outcome when the same pre-processing strategy was implemented due to it being in a different shape compared with the public data in terms of number of feature and the correlation strength.

1. **Learning Outcomes of the Study in respect of Data Science**

Local data needs more feature to be added preferably with a strong correlation with the house price however LR gave the best RSME score and lasso got the best R2 score overall. The final result of study showed the lasso make s batter prediction compared to other used algorithms.

1. **Limitations of this work and Scope for Future Work**

As the model has several parameter to set, a reasonable approach would be an extensive exploration of best values as this can be tedious and error prone task one can make us of some automated search/evaluation process. Like the grid search resource available in the sci-kit learn package with the use of analytical and graphical tools, we were able to evaluate the predictive performance of various housing price model also helped identify which characteristic of housing were most strongly associated with the prise and could explain most of the price variation.