Project Housing

Submitted By:

Aman Saxena

**ACKNOWLEDGMENT**

* <https://scikit-learn.org/stable/> - For the libraries used in the project.
* Rest project is done by myself only.

**INTRODUCTION**

* Business Problem Framing

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.In this project we have to build the model that will predict the Sale price of the house using the independent variables so that the company can accordingly manipulate the strategy of the firm and concentrate on areas that will yield high returns.

* Conceptual Background of the Domain Problem

For more understanding we can simply correlate it with the real estate business in which we buy lots of property at the lower price and then sale it with the large profit.

* Motivation for the Problem Undertaken

In this project we have to build the model that will predict the Sale price of the house using the independent variables so that the company can accordingly manipulate the strategy of the firm and concentrate on areas that will yield high returns.

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

First of all we load the dataset then we see the shape of the dataset i.e 1168 rows and 80 columns and we notice that there are columns which has more than 70% of the empty values that means these columns are not usefull for us ['Alley','MiscFeature','Fence','PoolQC'] so we remove these columns from the dataset. Then we notice that most of the columns have Nan so we dealt those columns by the mean or mode of that column. After that we check for the outliers using the boxplot we found lots of outliers in the columns but those were the categorical columns in which the outliers were present and some outliers are the possible values so we did not remove the outliers from the dataset.

We separately visualize the features and for that we make two different dataframes for the categorical and continuous features . For the categorical we have used the count plot so that we can observe how many unique categories were present in the columns and the number of values lies under those categories.For continuous features we have used the distribution plot and we observe that there is lot of skewness present in the data set which we have dealt with the power transform to remove the skewness.

Since the categorical values were present so we have to encode them to make the prediction for that we have used ordinal encoder.

* Data Sources and their formats

We load the dtatset and check their datatypes so we observe that some columns are of float , some are of int and some are of object data types. Then we notice that most of the columns have Nan so we dealt those columns by the mean or mode of that column.

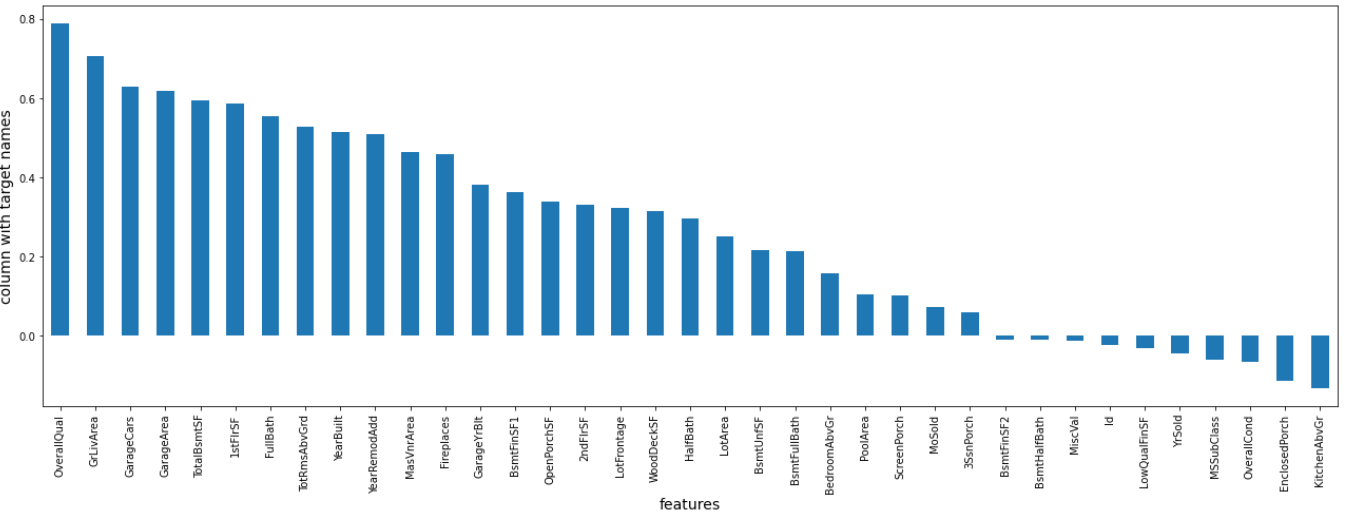
* Data Preprocessing Done

First of all we load the dataset then we see the shape of the dataset i.e 1168 rows and 80 columns and we notice that there are columns which has more than 70% of the empty values that means these columns are not usefull for us ['Alley','MiscFeature','Fence','PoolQC'] so we remove these columns from the dataset. Then we notice that most of the columns have Nan so we dealt those columns by the mean or mode of that column. After that we check for the outliers using the boxplot we found lots of outliers in the columns but those were the categorical columns in which the outliers were present and some outliers are the possible values so we did not remove the outliers from the dataset.

We separately visualize the features and for that we make two different dataframes for the categorical and continuous features . For the categorical we have used the count plot so that we can observe how many unique categories were present in the columns and the number of values lies under those categories.For continuous features we have used the distribution plot and we observe that there is lot of skewness present in the data set which we have dealt with the power transform to remove the skewness.

Since the categorical values were present so we have to encode them to make the prediction for that we have used ordinal encoder.

* Data Inputs- Logic- Output Relationships



In the above picture we can clearly see that what features are positively related with the target variable and what features are negatively related with the target variable.

* Hardware and Software Requirements and Tools Used
* import pandas as pd
* import numpy as np
* import matplotlib.pyplot as plt
* import seaborn as sns
* from sklearn.linear\_model import LinearRegression
* from sklearn.preprocessing import StandardScaler
* import warnings
* warnings.filterwarnings('ignore')
* from sklearn.preprocessing import OrdinalEncoder
* from sklearn.preprocessing import power\_transform
* from statsmodels.stats.outliers\_influence import variance\_inflation\_factor
* from sklearn.metrics import r2\_score
* from sklearn.model\_selection import train\_test\_split,GridSearchCV
* from sklearn.linear\_model import Lasso,Ridge

**Model/s Development and Evaluation**

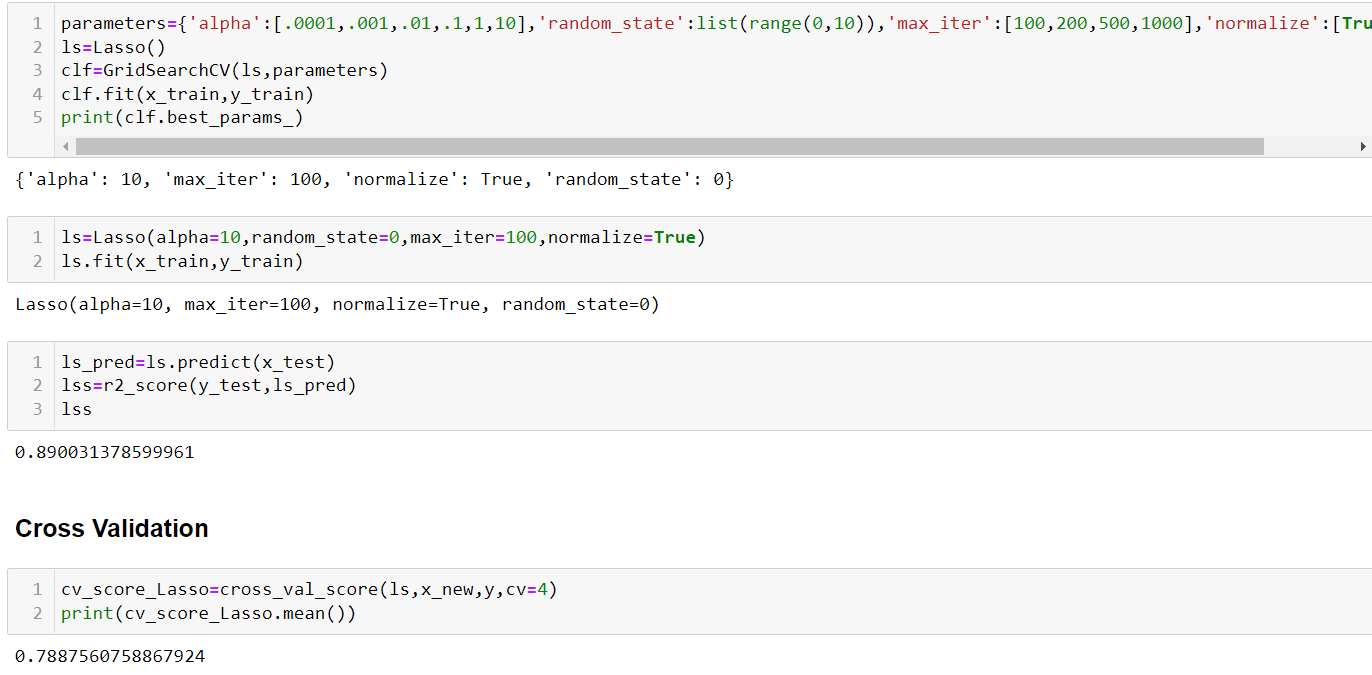
* Identification of possible problem-solving approaches (methods)

Since the problem is of regression and we have to predict the continuous output so the approaches which we can use are Linear Regression, Lasso Regression, Ridge Regression, Random forest Regressor but the approaches which we have used in our project are Linear Regression, Lasso Regression, Ridge Regression.

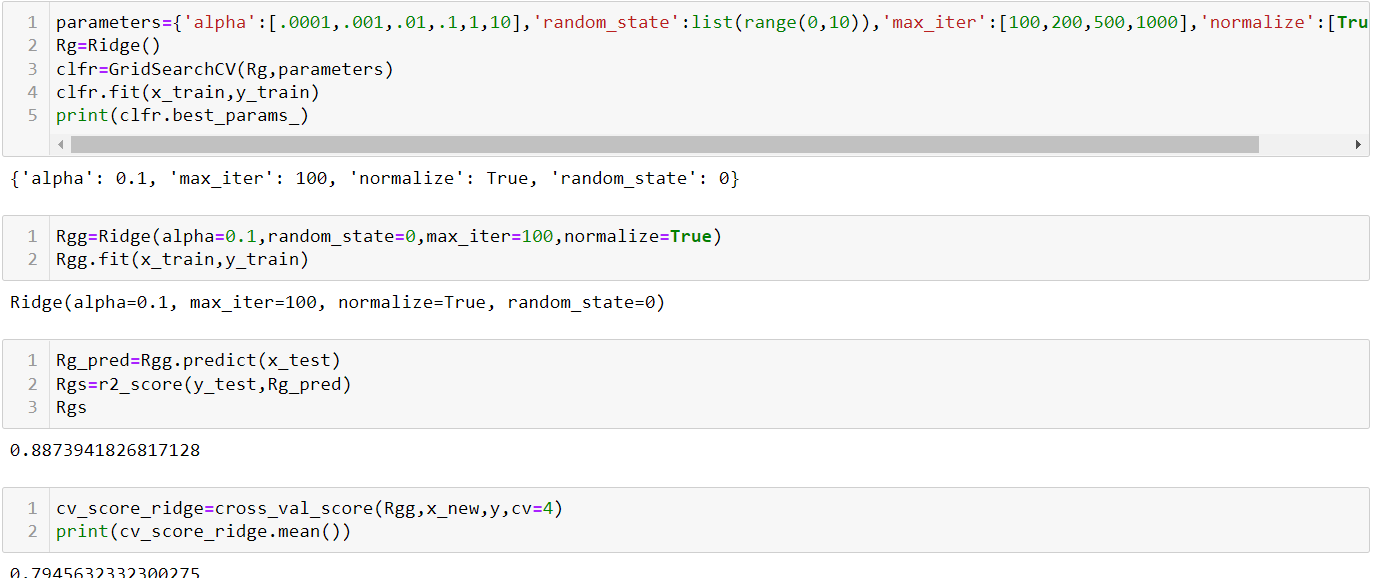
* Testing of Identified Approaches (Algorithms)
* Lasso Regression
* Ridge Regression
* Linear Regression
* Run and Evaluate selected models
* Linear Regression



* Lasso Regression



* Ridge Regression



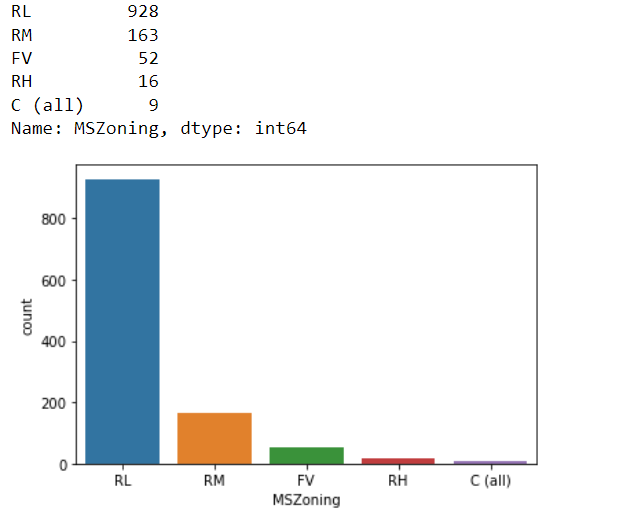
* Key Metrics for success in solving problem under consideration

R2 Score 🡪 for calculating the accuracy

Cross Val Score 🡪 For cross validation

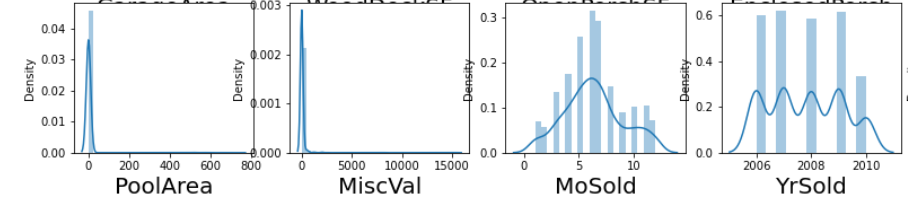
* Visualizations
* Count plot 🡪 For categorical columns
* Dist plot 🡪 For continuous data

**Count plot**



From the above plot we can clearly see that there were 5 categories in MSZoning column and the values lies under these categories are also present.

**Dist plot**

****

From the above distribution plot we can clearly see that the first two columns are rightly skewed which we dealt with power transform function.

**CONCLUSION**

* Key Findings and Conclusions of the Study

First of all we load the dataset then we see the shape of the dataset i.e 1168 rows and 80 columns and we notice that there are columns which has more than 70% of the empty values that means these columns are not usefull for us ['Alley','MiscFeature','Fence','PoolQC'] so we remove these columns from the dataset. Then we notice that most of the columns have Nan so we dealt those columns by the mean or mode of that column. After that we check for the outliers using the boxplot we found lots of outliers in the columns but those were the categorical columns in which the outliers were present and some outliers are the possible values so we did not remove the outliers from the dataset.

We separately visualize the features and for that we make two different dataframes for the categorical and continuous features . For the categorical we have used the count plot so that we can observe how many unique categories were present in the columns and the number of values lies under those categories.For continuous features we have used the distribution plot and we observe that there is lot of skewness present in the data set which we have dealt with the power transform to remove the skewness.

Since the categorical values were present so we have to encode them to make the prediction for that we have used ordinal encoder.