**Title -** OLX car sellers are finding it difficult to put a price on the cars they are selling.So we predict Car selling price using Regression

**Group number** - 20

**Participants list** - 1. Bablu 2. Shakti Singh Rathore

**Problem statement**- Regression is an important machine learning model for these kinds of problems [OLX Group](https://www.olxgroup.com/) is a global online marketplace operating in 45 countries. The OLX marketplace is a platform for buying and selling services .OLX car sellers are finding it difficult to put a price on the cars they are selling.So we predict Car selling price and we use features of the cars and their prices. For this kind of project of Price predict, we will apply the linear regression and Random forests and evaluate the result based on the training and testing a set of the data.

**Overall summary of your solution-**

1. Web Scraping-

Scrap the data from OLX site using python library bs4 in BeautifulSoup and extract feature of Second hand car and around 800 rows and 16 columns and convert into the data frame.

1. Loading the Data and getting first intuitions
   1. Identifiying missing values
   2. Plausibility check of numerical attributes
   3. Checking values of categorical attributes
2. Visualization /Plot and Histogram of "Price"
   1. Rough Data Cleansing for plausible Visualization
   2. Checking the amount of data after rough cleansing
   3. Check Distribution and Data
3. Data Spliting

Split data into traning and testing and after that using Linear Regression ,Random forests and more algorithm

1. Buidling Custom-Transformers and Preprocessing-Pipelines
   1. Calculating VIF and Noramlize data using Logrithm
   2. Data Cleansing (Filtering, Replacing NaN-Values, Frequency count and imput data )
   3. Custom-Transformers & Pipelines L1 and L2 Regularization
   4. Training and comparing Models
   5. Feature Importance (Note: I can only show numerical features. Help is appreciated if you could help me out with categorical features.)

**Detailed description and analysis of the solutions provided, including each technique used, supported by Tables / Graphs / Charts / Metrics**

**Description and Analysis of the solutions provided:**

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

%matplotlib inline

from sklearn.linear\_model import LinearRegression

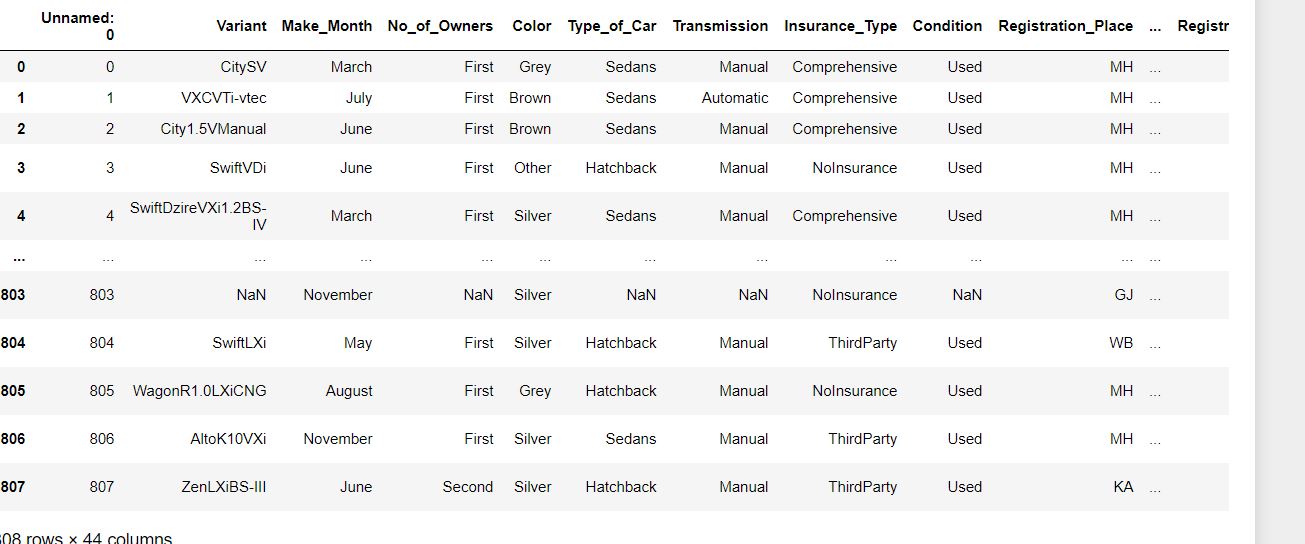
import statsmodels.formula.api as smf

from sklearn.model\_selection import train\_test\_split

from sklearn.metrics import mean\_squared\_error

import seaborn as sns

from sklearn import model\_selection



And then we read the csv file where data is stored by creating a dataframe as:

df=pd. read\_csv('forestfires.csv')

df. head ()

### **Linearity**: The relationship between the independent and dependent variables should be linear. The linearity assumption can best be tested with scatter plots.

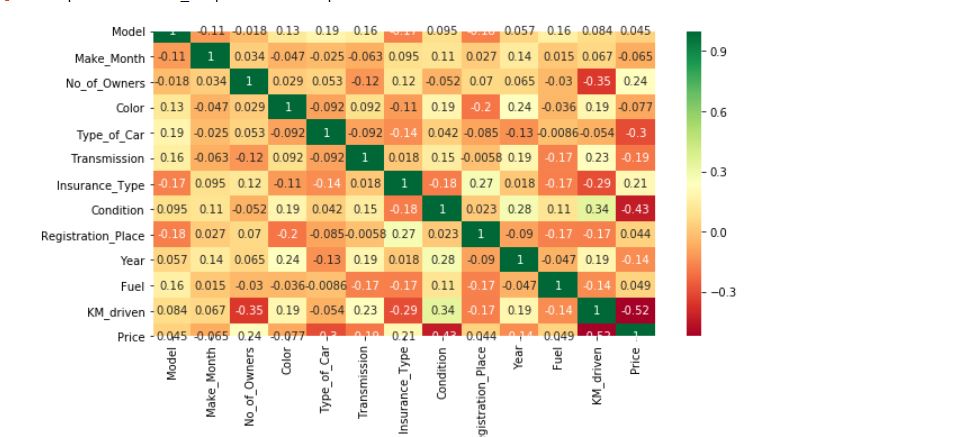
### import seaborn as sns

sns.pairplot(df)

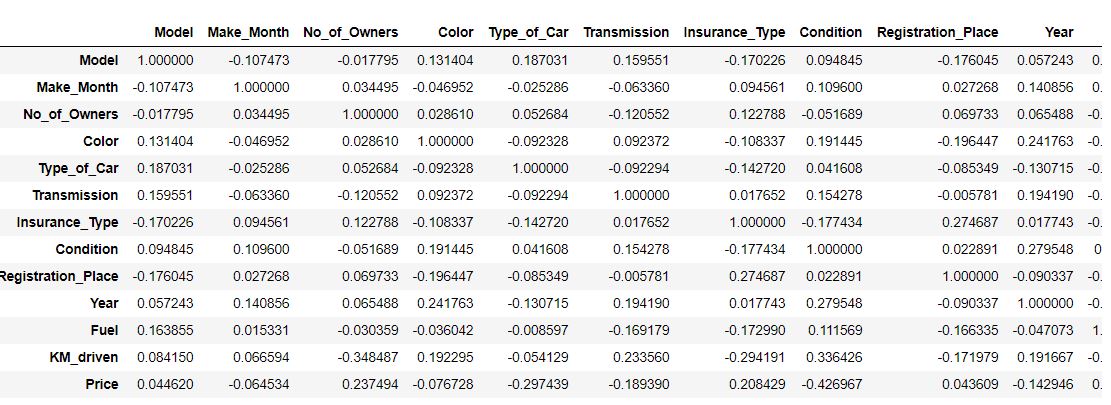
### )

#### **Multicollinearity:** Linear regression assumes that there is little or no multicollinearity in the data**.** It can be tested through VIF, Correlation matrix.

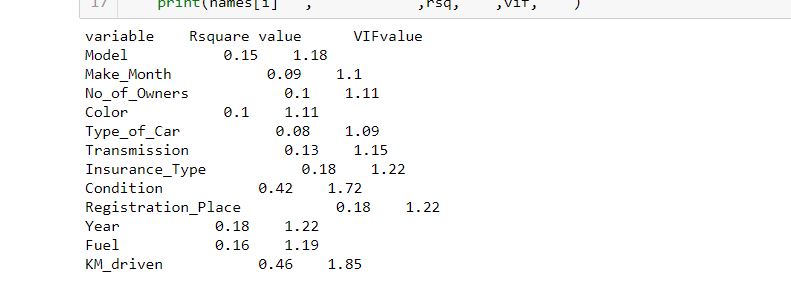
**Heat Map:**

****

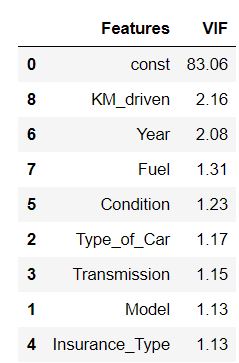
Correlation Matrix:

****

**VIF:**

****

Conclusion: With VIF > 5 there is an indication that multicollinearity may be present; with VIF > 10 there is certainly multicollinearity among the variables. We can see that there no variables that are greater than 10,if it is than centering the data the simplest way to address the problem is to remove independent variables with high VIF values



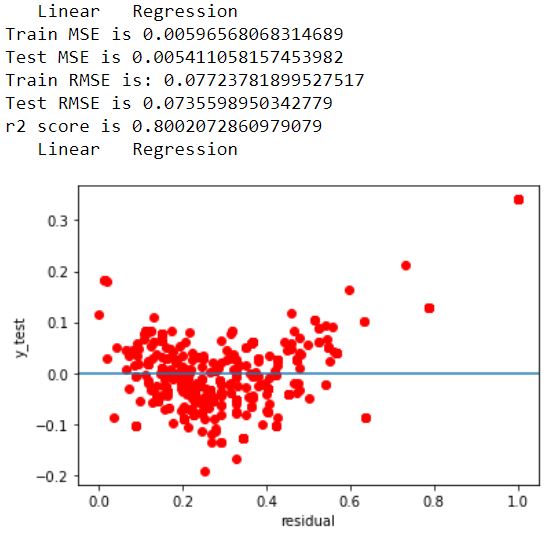
## Feature Importances-

## 

**Models-**

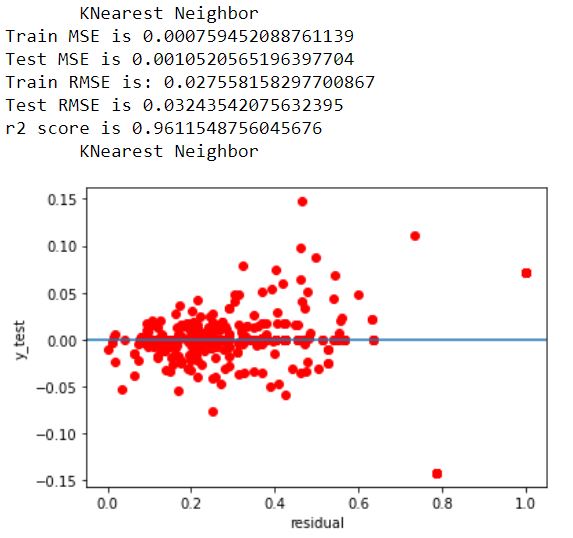
**Linear Regression :-**

* When used all the features which are encoded to fit the model and the R-squared value is 80.02%.
* There are show residual plot its show are random point.
* There are show Train RMSE ,Train MSE, and also show for test .
* As the difference between training and testing MSE and RMSE is less the model is not having overfitting

****

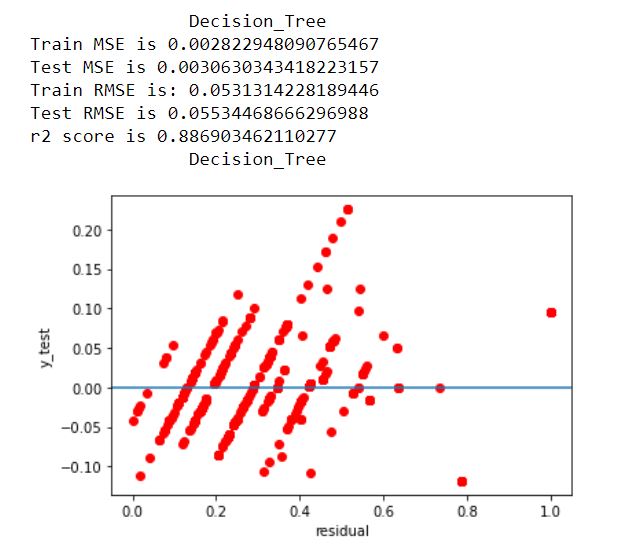
**KNearest Neighbor :-**

* When used all the features which are encoded to fit the model and the R-squared value is 96.11%.
* There are show residual plot its show are random point.
* There are show Train RMSE ,Train MSE, and also show for test .
* As the difference between training and testing MSE and RMSE is less the model is not having overfitting



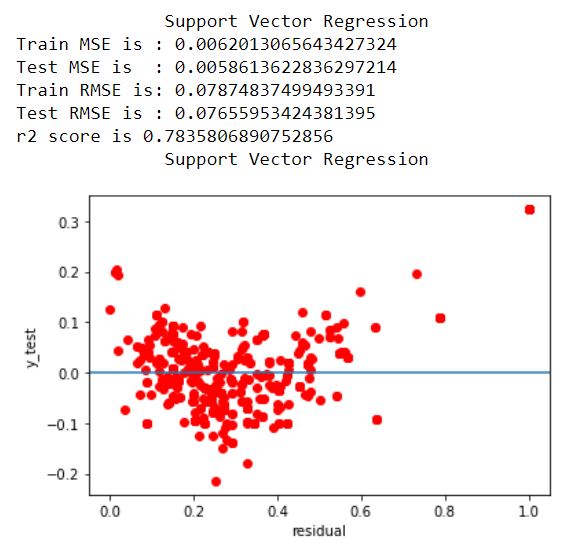
**Decision Tree :-**

* When used all the features which are encoded to fit the model and the R-squared value is 88.69%.
* There are show residual plot its show are random point.
* There are show Train RMSE ,Train MSE, and also show for test .
* As the difference between training and testing MSE and RMSE is less the model is not having overfitting

****

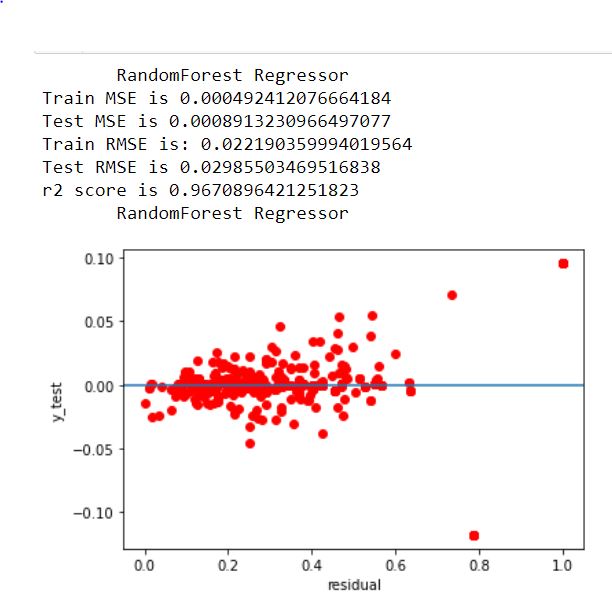
**Support Vector Regression :-**

* When used all the features which are encoded to fit the model and the R-squared value is 78.38%.
* There are show residual plot its show are random point.
* There are show Train RMSE ,Train MSE, and also show for test .
* As the difference between training and testing MSE and RMSE is less the model is not having overfit

****

**RandomForest Regressor :-**

* When used all the features which are encoded to fit the model and the R-squared value96.70%.
* There are show residual plot its show are random point.
* There are show Train RMSE ,Train MSE, and also show for test .
* As the difference between training and testing MSE and RMSE is less the model is not having overfitting

****

**Identification of the best model backed by the logic for selecting it**

## Conclusions

## I've tried to play with as much stuff as I could with this dataset in order to understand the very basic topics about:

## data interpretation and selection Random forests and Ridge and lasso given good result

## we are getting linear regression in accuracy 80 % and when we apply Regularization l1 and l2 we get more accuracy around 80 % and same as in random forests ,SVM,Decision tree gaving best accuracy 96% .

**Lessons learnt from the project**

Learn Various type of Machine learning algoritham and EDA process and Cleaning Dataset Learn about the of sklearn Library and Matplot lib libaray and learnt Web Scarping how to data scraping from any site .