**First Milestone – Machine Learning**

**Supervisor**

**Dr /Dina Khattab**

**Team Members**

|  |  |  |  |
| --- | --- | --- | --- |
| No | Name | Department | Id |
| 1 | محمد خالد محمد علي هندي | Scientific Computing | 2017170334 |
| 2 | محمد فتحي معروف عثمان | Scientific Computing | 2018170337 |
| 3 | محمد اسامه محمود علي | Scientific Computing | 2018170310 |
| 4 | محمد قنديل علي عبد العزيز | Scientific Computing | 2018170339 |
| 5 | محمود خلاف احمد | Scientific Computing | 2017170394 |

**Preprocessing and Regression**

* **Preprocessing:**

Data preprocessing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model.

* **Feature Encoder:**

Making feature encoding to **string** values columns to convert it to **int** values. Fill null values in **“segment”** column with mode value for column after making encoding for same column, Fill null values in

**“Volume (cm3)”** columnand **“drive\_unit”** with mean values after making encoding for the same two columns.

* **Feature Clean:**

1. Convert **“fuel\_type”** string into lower because it has 4 values which expresses 2 values
2. Split **“car-info”** into three columns **“model”,** **“car”**, **“year”** and remove brackets of list from column and circle brackets
3. Then drop **“car-info”** after split it to three columns **“model”,** **“car”**, **“year”**
4. Make a list of all columns in the dataframe remove **“price (USD)”** and **“year”** from the list and create the new dataframe with two columns, convert **“year”** to **int**

* **Feature Scaling:**

Feature scaling is a method used to normalize the range of independent variables or features of data. It makes the flow of “**gradient descent”**smooth and helps algorithms quickly reach the minima of the cost function, we used **Normalization** for feature scaling, **Normalization** is a scaling technique in which values are shifted and rescaled so that they end up ranging between **“0”** and **“1”.** It is also known as “**Min-Max scaling”.**

* **Dealing with null Values:**

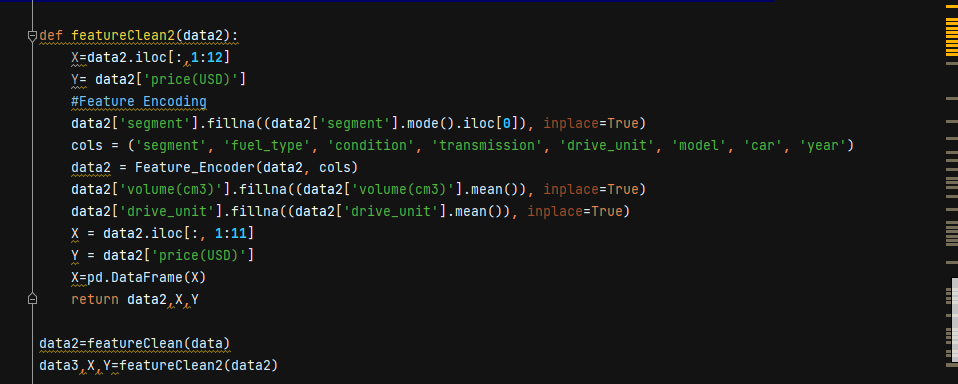
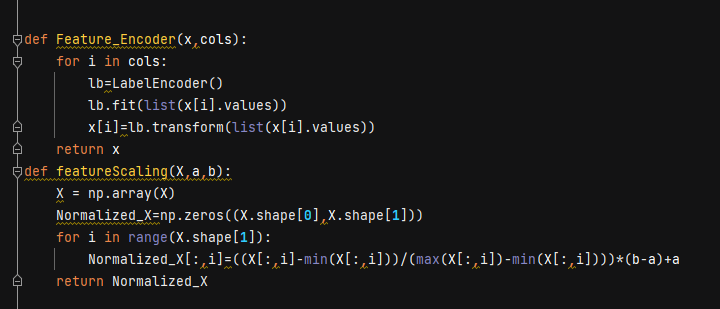
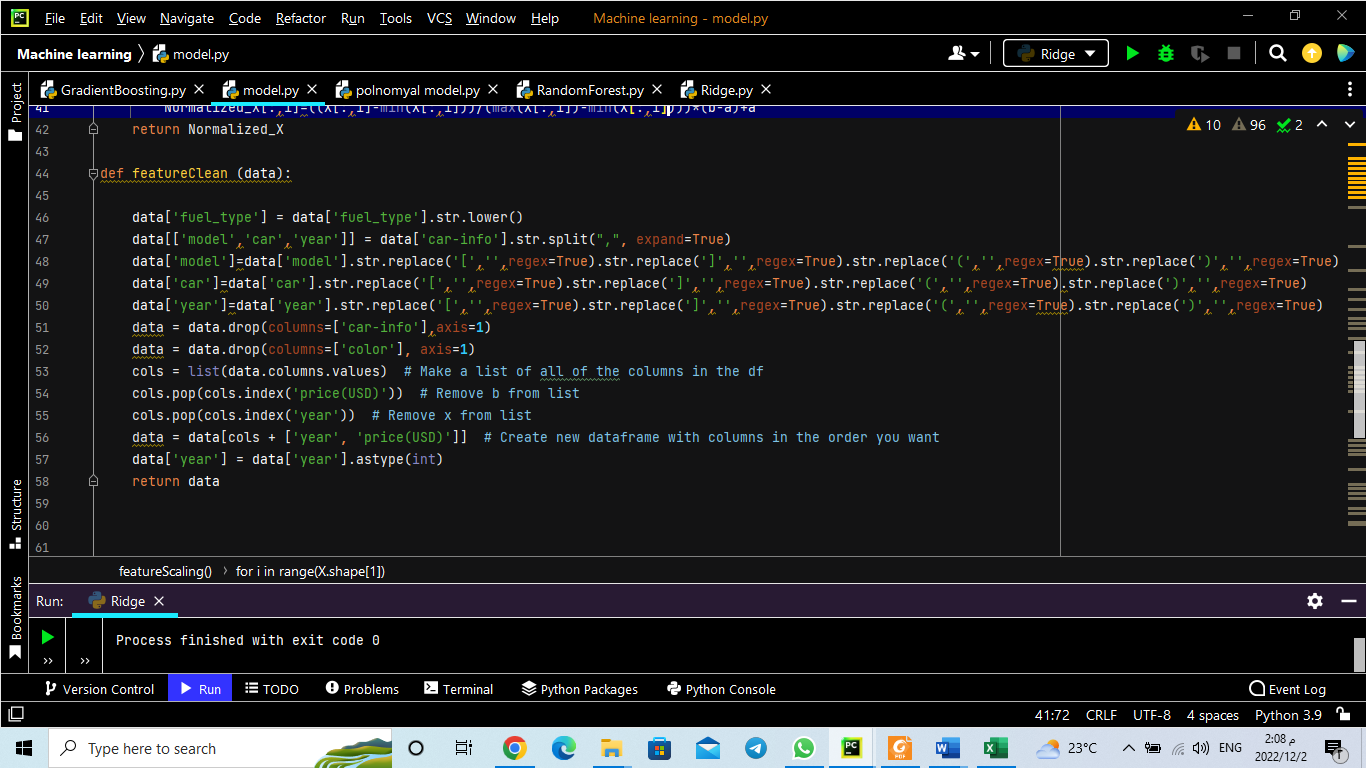
1. The **“segment”** column has **(4216)** null values which is approximately **9%** from the total number of rows **(44994),** we solve the null problem in **“segment”** column with fill null values with the mode value of column after encoding.
2. The **“volume(cm3)”** column has **(35)** null values which approximately **0.07%** from the total number of rows **(44994),** we solve the null problem in **“volume(cm3)”** column with fill null values with mean value of column after encoding.
3. The **“drive\_unit”** column has **(1543)** null values which approximately **3.5%** from the total number of rows **(44994),** we solve the null problem in **“drive\_unit”** column with fill null values with mean value of column after encoding.

* **Drop Column:**

1. Drop column **“car-info”** because we split

**“car-info”** into three columns **“model”,** **“car”**, **“year”.**

1. Drop column **“color”** because it improves Accuracy.



**Feature selection:**

By using Correlation, we are getting  
the top 50% correlation features with the target “**price (USD)**”,

We used **“Heatmap”** to visualize the correlation between the features and the target column, so that we can decide which features will we select for our model.

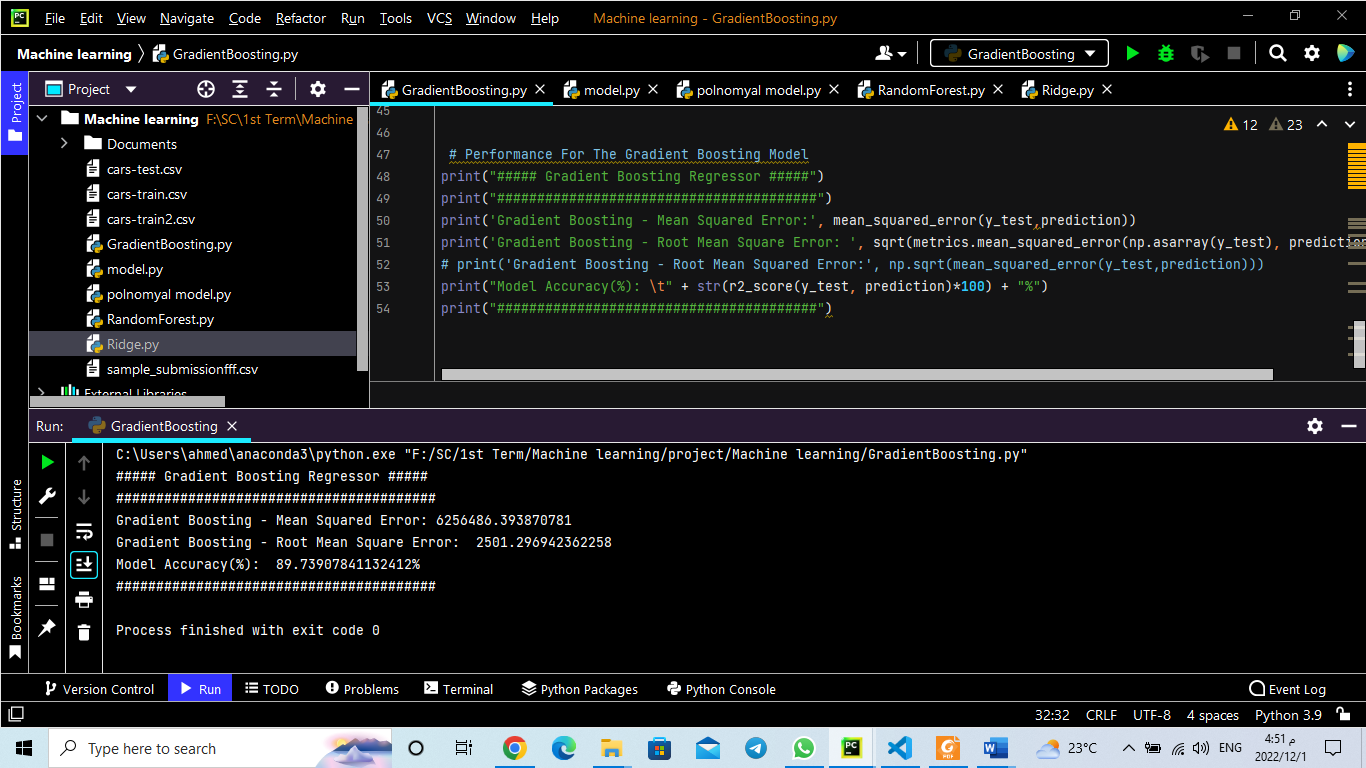
**“Models Used in Milestone”**

1. Gradient Boosting Model
2. Polynomial Regression Model
3. Random Forest Model
4. Ridge Regression

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Mean Squared Error | Root Mean Squared Error | Accuracy |
| Gradient Boosting | 6256486.393870784 | 2501.2969423622585 | 89.73907841132412% |
| Random Forest | 10693694.70046628 | 3270.1215115751097 | 84.90695250570157% |
| Polynomial Regression | 14056221.603053506 | 3749.1627869503754 | 74.55963534483972% |
| Ridge Regression | 28791277.474643026 | 5365.7504111394355 | 52.0967356349478% |

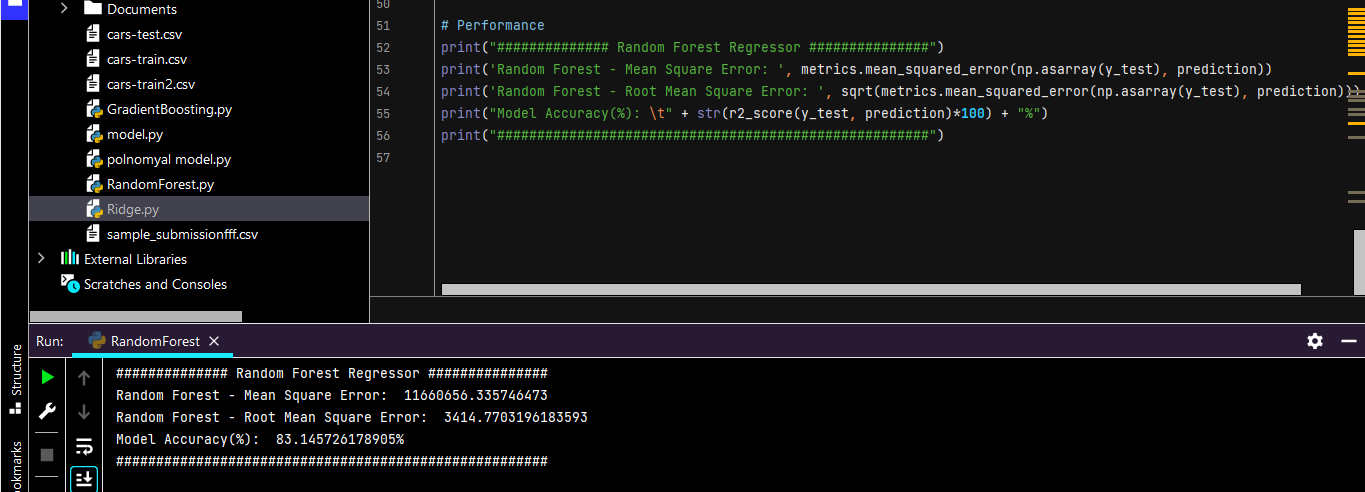
1. **Gradient Boosting**

Builds an additive model in a forward stage-wise fashion; it allows for the optimization of arbitrary differentiable loss functions. In each stage, a regression tree is fit on the negative gradient of the given loss function.



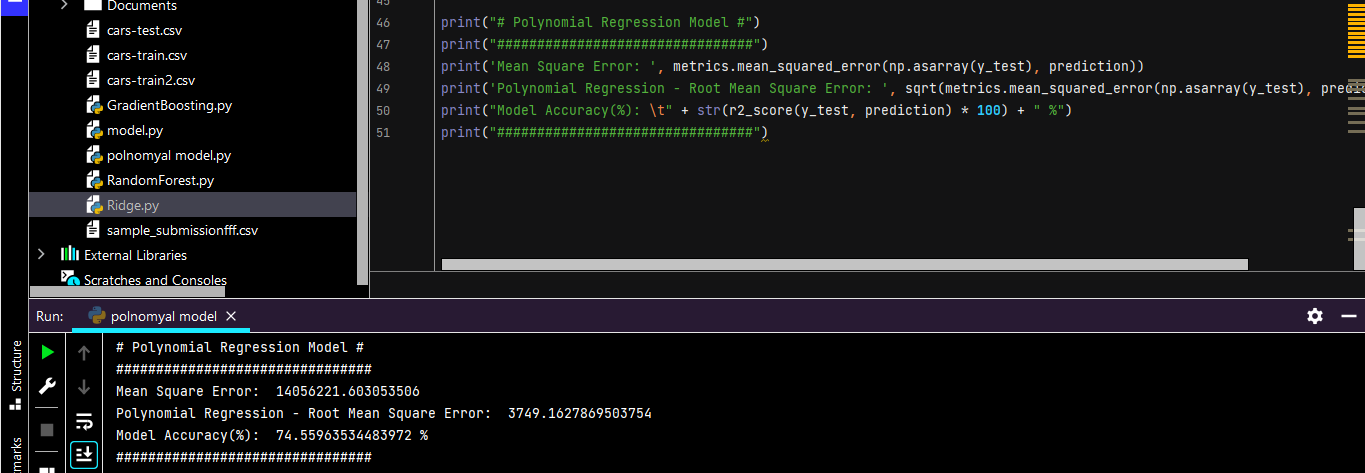
1. **Random Forest Regression**

An ensemble learning algorithm based on decision tree learners. The estimator fits multiple decision trees on randomly extracted subsets from the dataset and averages their prediction



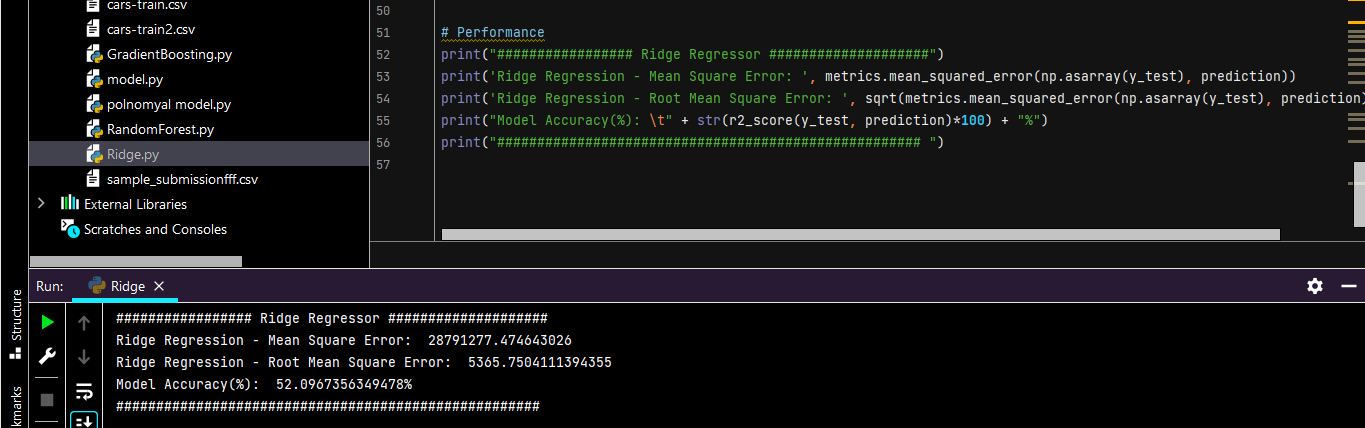
1. **Polynomial Regresssion**

form of Linear regression known as a special case of Multiple linear regression which estimates the relationship as an nth degree polynomial.



1. **Ridge Regression**

A model tuning method used to analyze any data that suffers from multicollinearity. This method performs L2 regularization. When the issue of multicollinearity occurs, least-squares are unbiased, and variances are large, this results in predicted values being far away from the actual values.



**Training and Testing size:**

* Training set: is 70% from the total of data size
* Testing set: is 30% from the total of data size
* Validation set: didn’t used validation in models
* Random\_state = 42

**Conclusion:**

In most practices, the best metrics to optimize have been obtained by the extreme gradient boosting and random forest in the **“Cars Price Prediction”**. So, it is always related to the dataset and how their features relate with each other, and there is no static way to obtain the best metric.

**The highest Models Accuracy is:**

1. Gradient Boosting: 89.73907841132412%
2. Random Forest: 84.90695250570157%
3. Polynomial Regression: 74.55963534483972%
4. Ridge Regression: 52.0967356349478%