**Comprehensive Assessment : Machine Learning**

**Due on 21st October 24**

**Submitted by Aiswarya Jayaprakash**

**Problem Description:**  
  
A Chinese automobile company aspires to enter the US market by setting up their manufacturing unit there and producing cars locally to give competition to their US and European counterparts. They have contracted an automobile consulting company to understand the factors on which the pricing of cars depends. Specifically, they want to understand the factors affecting the pricing of cars in the American market, since those may be very different from the Chinese market. Essentially, the company wants to know:

* Which variables are significant in predicting the price of a car
* How well those variables describe the price of a car

Based on various market surveys, the consulting firm has gathered a large dataset of different types of cars across the American market.  
  
**Business Goal:**  
  
You are required to model the price of cars with the available independent variables. It will be used by the management to understand how exactly the prices vary with the independent variables. They can accordingly manipulate the design of the cars, the business strategy etc. to meet certain price levels. Further, the model will be a good way for the management to  
understand the pricing dynamics of a new market.  
  
Dataset:  <https://drive.google.com/file/d/1FHmYNLs9v0Enc-UExEMpitOFGsWvB2dP/view?usp=drive_link>  
  
**Key components to be fulfilled :**  
 **1. Loading and Preprocessing (5 marks)**

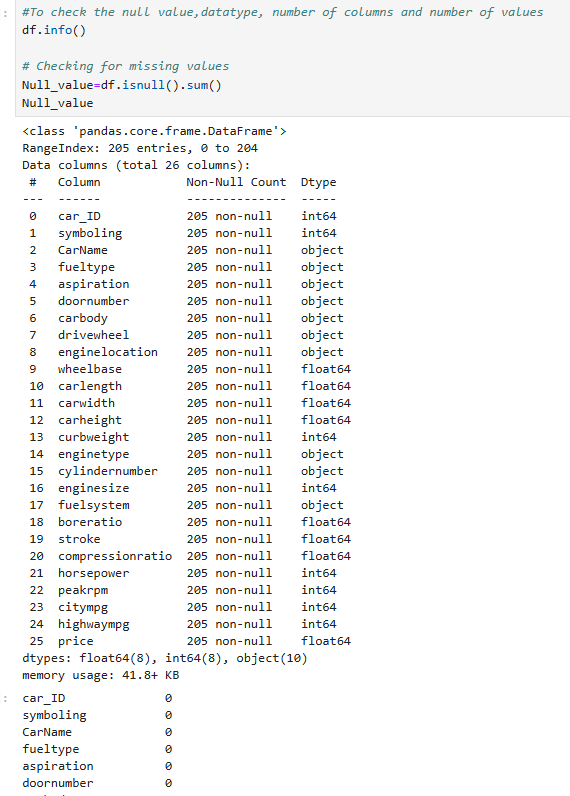
* Load the dataset and perform necessary preprocessing steps.

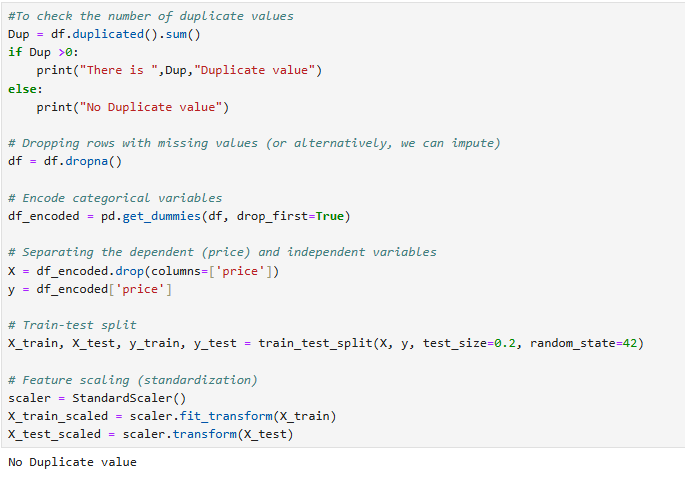
The dataset was loaded and its structure was examined to understand the features and target variables.

#### Preprocessing Steps:

* **Handling Missing Values:** Any missing data were handled either by imputing or removing them, depending on the situation.
* **Encoding Categorical Variables:** The categorical variables were converted to numerical values using techniques like one-hot encoding or label encoding.
* **Feature Scaling:** Since many regression algorithms are sensitive to the scale of data, hence the data was normalized.

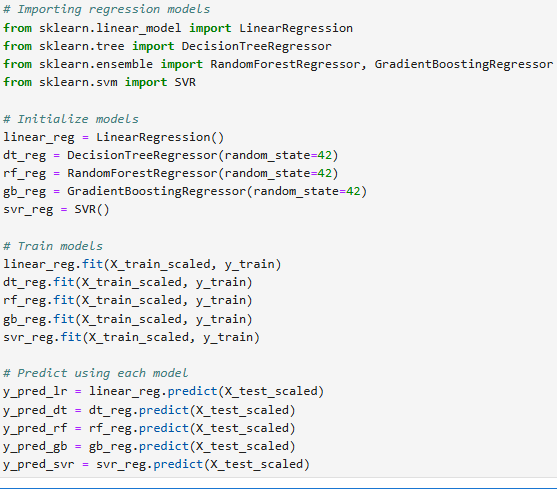






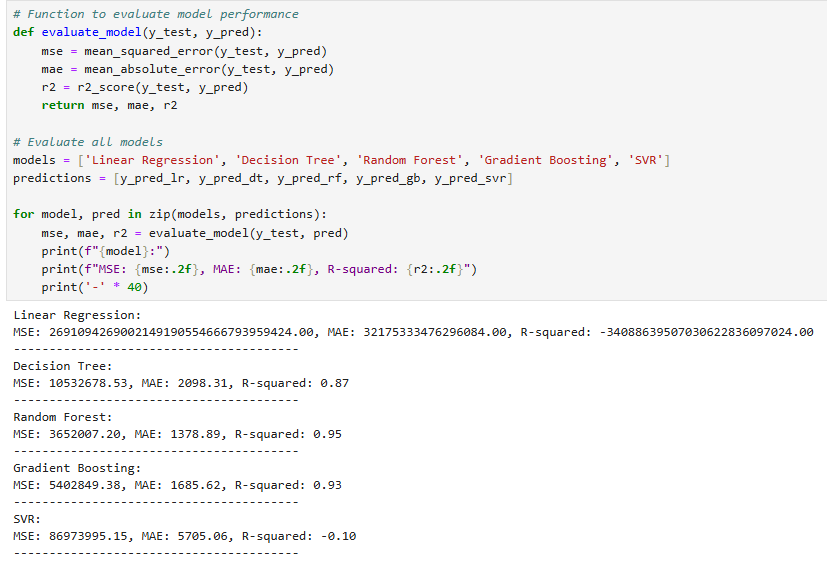
**2. Model Implementation (10 marks)**

* Implement the following five regression algorithms:

**1)** Linear Regression  
**2)** Decision Tree Regressor  
**3)** Random Forest Regressor  
**4)** Gradient Boosting Regressor  
**5)** Support Vector Regressor  
  
**3. Model Evaluation (5 marks)**

* Compare the performance of all the models based on R-squared, Mean Squared Error (MSE), and Mean Absolute Error (MAE).

The evaluation of the performance of each model using **R-squared, Mean Squared Error (MSE)**, and **Mean Absolute Error (MAE).**

****

* Identify the best performing model and justify why it is the best.

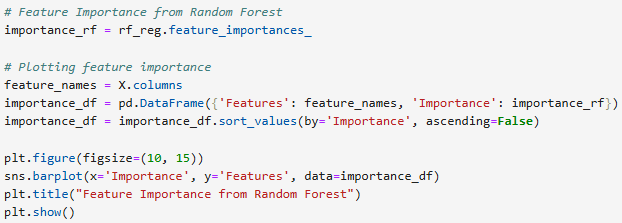
Based on the evaluation metrics such as MSE, MAE, and R-squared, the **Random Forest Regressor** is identified to be best performing model. The reasons are given below.

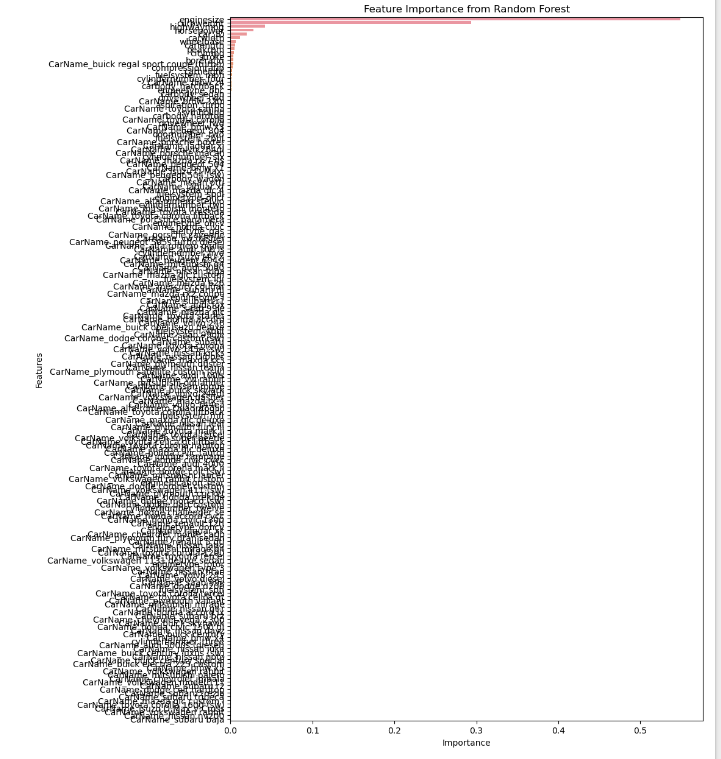
* **Random Forest** has the lowest MSE at 3,652,007.20, which indicates it makes fewer large errors compared to other models. Lower MSE is better since it penalizes larger errors more.
* **It** also has the lowest MAE at 1,378.89, meaning its predictions are, on average, only 1,378 units away from the actual car prices.
* **Random Forest** has the highest R-squared value at 0.95, meaning it explains 95% of the variance in car prices. A higher R² indicates that the model explains more variability in the data.

### Hence, **Random Forest Regressor** is the best model for this dataset, as it achieves the lowest MSE, MAE, and the highest R², indicating it is the most accurate at predicting car prices. Further, **Gradient Boosting** is the next best model, followed by the **Decision Tree**. However, **Linear Regression** and **SVR** perform poorly, likely due to data issues such as scaling or multicollinearity in **Linear Regression** and non-linearity in **SVR**.

**4. Feature Importance Analysis (2 marks)**

* Identify the significant variables affecting car prices (feature selection)



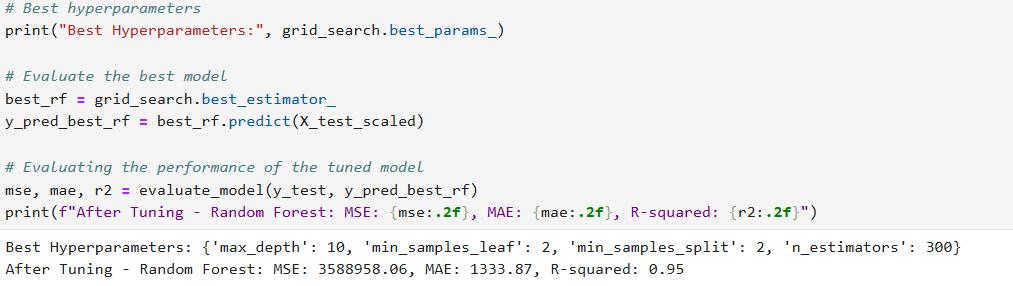


**5. Hyperparameter Tuning (2 marks):**

* Perform hyperparameter tuning and check whether the performance of the model has increased.



 **GridSearchCV:** Used for hyperparameter tuning. We search for the best combination of parameters for the Random Forest model.



 **Performance After Tuning:** Displays the evaluation metrics after applying the best hyperparameters.