Report

Machine learning :car prediction

**domain background:-**

-Car market

Our project was designed to predict the price of cars, the model on which machine learning was built is specialized in the field of predicting car prices.

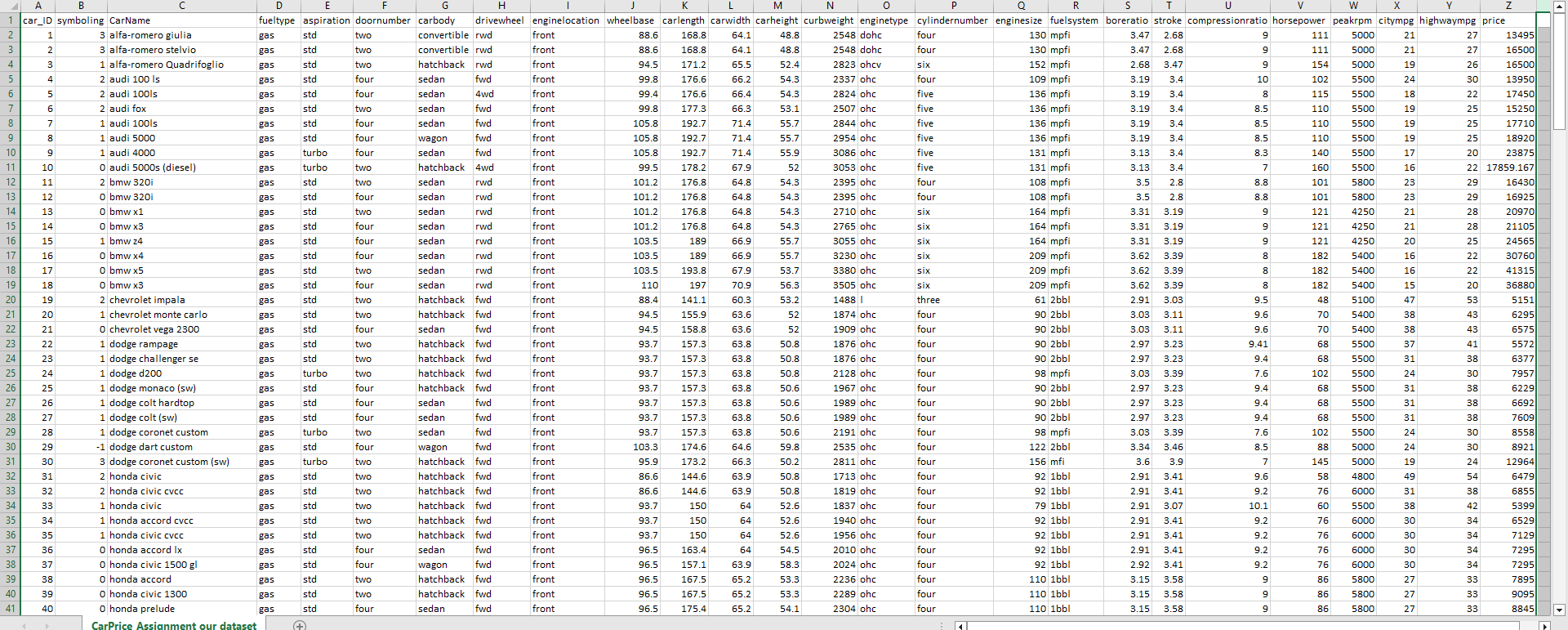
**Problem statement:-**

Let us assume, for example, that a company wanted to manufacture cars to sell them in the market, but there is a problem with pricing those cars in several respects. When they are priced based on nothing, the prices may be exaggerated or have a value less than their reasonable price.

**datasets and inputs:-**

the inputs we will used is

|  |  |  |
| --- | --- | --- |
| 1 | **Car\_ID** | Unique id of each observation (Interger) |
| 2 | **Symboling** | Its assigned insurance risk rating, A value of +3 indicates that the auto is risky, -3 that it is probably pretty safe.(Categorical) |
| 3 | **carCompany** | Name of car company (Categorical) |
| 4 | **fueltype** | Car fuel type i.e gas or diesel (Categorical) |
| 5 | **aspiration** | Aspiration used in a car (Categorical) |
| 6 | **doornumber** | Number of doors in a car (Categorical) |
| 7 | **carbody** | body of car (Categorical) |
| 8 | **drivewheel** | type of drive wheel (Categorical) |
| 9 | **enginelocation** | Location of car engine (Categorical) |
| 10 | **wheelbase** | Weelbase of car (Numeric) |
| 11 | **carlength** | Length of car (Numeric) |
| 12 | **carwidth** | Width of car (Numeric) |
| 13 | **carheight** | height of car (Numeric) |
| 14 | **curbweight** | The weight of a car without occupants or baggage. (Numeric) |
| 15 | **enginetype** | Type of engine. (Categorical) |
| 16 | **cylindernumber** | cylinder placed in the car (Categorical) |
| 17 | **enginesize** | Size of car (Numeric) |
| 18 | **fuelsystem** | Fuel system of car (Categorical) |
| 19 | **boreratio** | Boreratio of car (Numeric) |
| 20 | **stroke** | Stroke or volume inside the engine (Numeric) |
| 21 | **compressionratio** | compression ratio of car (Numeric) |
| 22 | **horsepower** | Horsepower (Numeric) |
| 23 | **peakrpm** | car peak rpm (Numeric) |
| 24 | **citympg** | Mileage in city (Numeric) |
| 25 | **highwaympg** | Mileage on highway (Numeric) |
| 26 | **price(Dependent variable)** | Price of car (Numeric) |

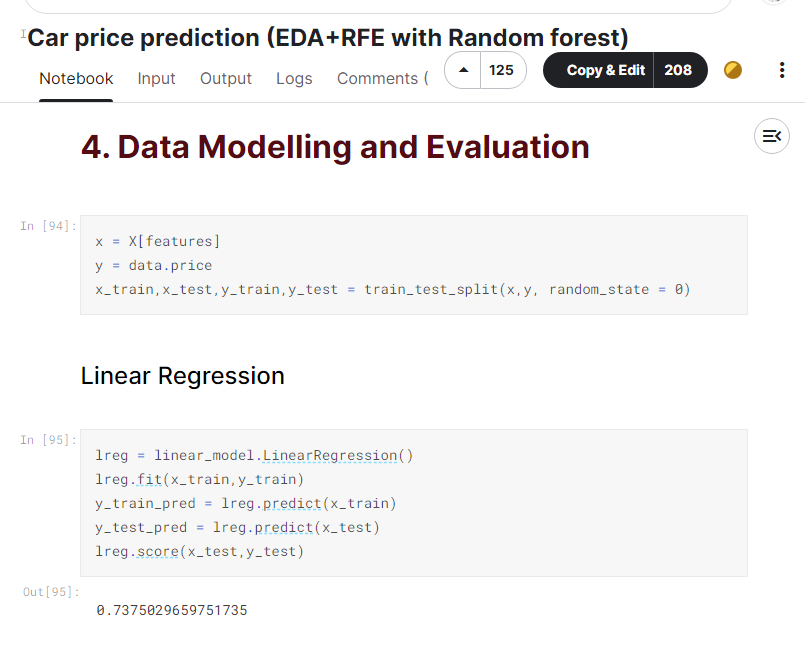
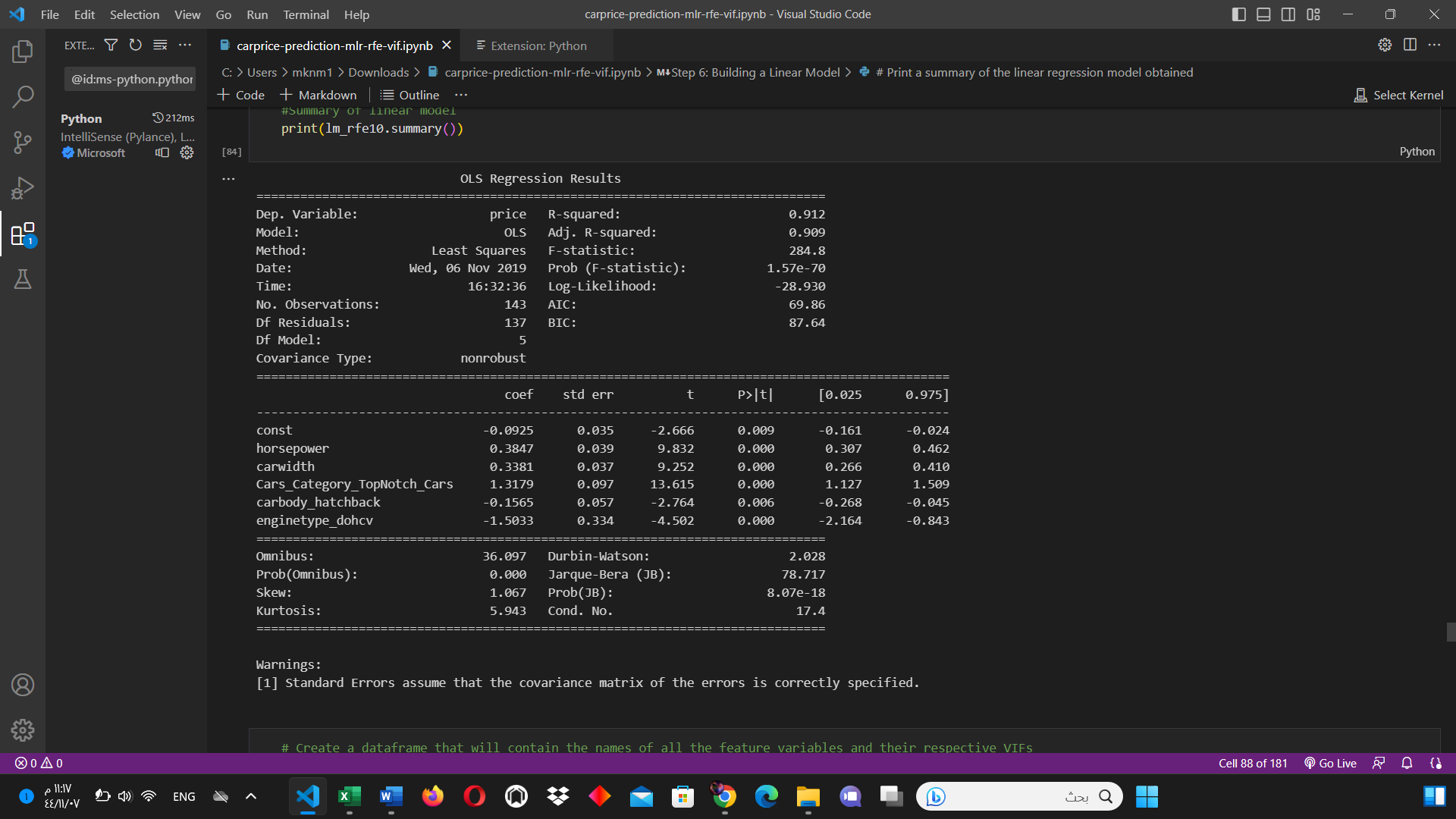
****

**solution statement:-**

Using specifications of the car, and the appropriate price for the market will be determined after machine learning algorithm from car models and their prices from other companies.

For the above problem, we will solve it using multiple linear regression or nueral network models, and these models agree with our problem, But the best model for this problem is linear regression because the number of features is not very large so no need for a nueral network.

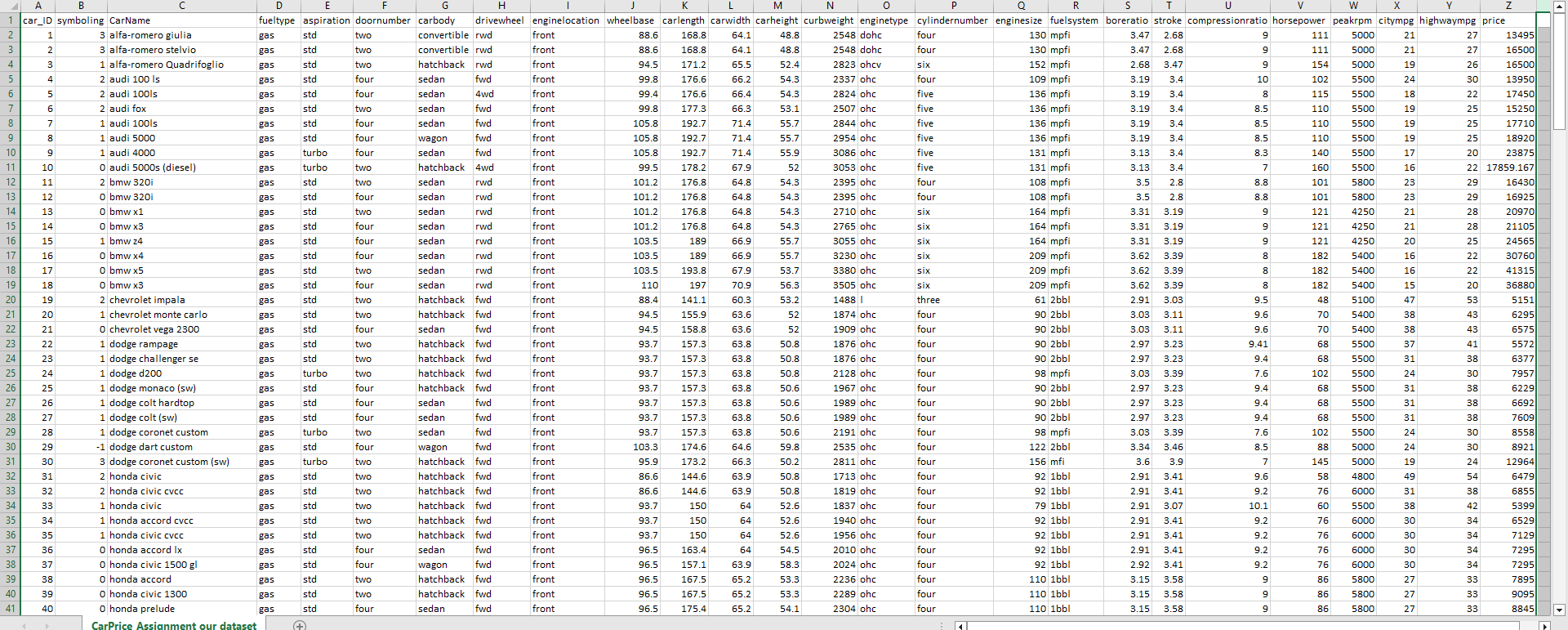
**Benchmark model**



**evaluation metrics**

the most technique used for evaluate regression model is R square as we see above there is to linear regression algorithm with different person they use r squared so we will use R squared to comparing with them after we finish our project and will use more evaluation techniques like MAE and MSE .

**project design:-**

**this figure in bottom is sample from our dataset**

1. we will clean the anonymous cell like null variable and some column that is not necessary for our project like id or any duplicate and we split the name of company from CarName
2. start with show visualizing data we use this to discover is there any skewed feature may harm our project
3. we will convert the type of string to integer to deal with it in our algorithm
4. we will see the values is not strict is there gap? , big gap between them so we will use some technique help us to scaling it like normalization or standardization
5. then we will fitting using linear regression
6. then comparing result the predicted with actual labels