MLP Week 4 SWI(shared)

February 1, 2022

You are working as a data scientist in a big automobile company. Your 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 got some data to understand the factors on which the pricing of cars depends in the American market, since those may vary different from the indian market. 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.

1 Business Goal:

Data science team 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 management to understand the pricing dynamics of a new market.

[2]:

```
!pip install -e .
```

```
ERROR: File "setup.py" or "setup.cfg" not found. Directory cannot be
installed in editable mode: /content
```

[3]:

```
!apt-get install pandoc
```

```
Reading package lists... Done
Building dependency tree
Reading state information... Done
pandoc is already the newest version (1.19.2.4~dfsg-1build4).
pandoc set to manually installed.
0 upgraded, 0 newly installed, 0 to remove and 37 not upgraded.
```

2 Step-1: Importing Libraries

[ ]:

```
# Importing the libraries
import numpy as np
import pandas as pd
from numpy import math

from sklearn.preprocessing import MinMaxScaler
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error
from sklearn.metrics import r2_score


import seaborn as sns
import matplotlib.pyplot as plt
import warnings
warnings.filterwarnings('ignore')
```

3 Step-2: Loading the data

[ ]:

```
from google.colab import drive
drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

[ ]:

```
# Importing the dataset
dataset = pd.read_csv('/content/drive/MyDrive/Dataset/Car_price in US market.
 ↪csv')
```

4 Step-3: Data Inspection

5 Question set-1:

(i) No of data point

(ii) No of features

(iii) No of categorical features

(iv) No of numerical features

(v) No of NA values

(vi) List of all features

(vii) What about duplicate data ?

[ ]:
```
dataset.shape
```

[ ]:
```
(205, 26)
```

[ ]:
```
dataset.head(5)
```

[ ]:
```
   car_ID  symboling                   CarName  ... citympg highwaympg    price
0       1          3        alfa-romero giulia  ...      21         27  13495.0
1       2          3       alfa-romero stelvio  ...      21         27  16500.0
2       3          1  alfa-romero Quadrifoglio  ...      19         26  16500.0
3       4          2               audi 100 ls  ...      24         30  13950.0
4       5          2                audi 100ls  ...      18         22  17450.0

[5 rows x 26 columns]
```

[ ]:
```
dataset.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 205 entries, 0 to 204
Data columns (total 26 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   car_ID            205 non-null    int64
 1   symboling         205 non-null    int64
 2   CarName           205 non-null    object
 3   fueltype          205 non-null    object
 4   aspiration        205 non-null    object
 5   doornumber        205 non-null    object
 6   carbody           205 non-null    object
 7   drivewheel        205 non-null    object
 8   enginelocation    205 non-null    object
 9   wheelbase         205 non-null    float64
 10  carlength         205 non-null    float64
 11  carwidth          205 non-null    float64
 12  carheight         205 non-null    float64
 13  curbweight        205 non-null    int64
 14  enginetype        205 non-null    object
 15  cylindernumber    205 non-null    object
 16  enginesize        205 non-null    int64
 17  fuelsystem        205 non-null    object
 18  boreratio         205 non-null    float64
 19  stroke            205 non-null    float64
 20  compressionratio  205 non-null    float64
 21  horsepower        205 non-null    int64
 22  peakrpm           205 non-null    int64
 23  citympg           205 non-null    int64
 24  highwaympg        205 non-null    int64
 25  price             205 non-null    float64
dtypes: float64(8), int64(8), object(10)
```

```
memory usage: 41.8+ KB
```

[ ]:

```
dataset.describe(include='all')
```

[ ]:

```
              car_ID   symboling  ...  highwaympg         price
count     205.000000  205.000000  ...  205.000000    205.000000
unique           NaN         NaN  ...         NaN           NaN
top              NaN         NaN  ...         NaN           NaN
freq             NaN         NaN  ...         NaN           NaN
mean      103.000000    0.834146  ...   30.751220  13276.710571
std        59.322565    1.245307  ...    6.886443   7988.852332
min         1.000000   -2.000000  ...   16.000000   5118.000000
25%        52.000000    0.000000  ...   25.000000   7788.000000
50%       103.000000    1.000000  ...   30.000000  10295.000000
75%       154.000000    2.000000  ...   34.000000  16503.000000
max       205.000000    3.000000  ...   54.000000  45400.000000

[11 rows x 26 columns]
```

[ ]:

```
features=(dataset.columns)
```

[ ]:

```
len(dataset[dataset.duplicated()])
```

[ ]:

```
0
```

6 Step-5: Exploratory data analysis

Question Set 2:

1) Give list of all numeric features

2) List of all categorical features

3) Type of distribution your dependent variable follow.

4) Plot different graph (histogram,box-plot,scatter plot e.t.c) for all independent variable to get some insight.

5) some scaling is needed or not.

(5) Comment about different categorical feature.

(6) Create a function which converts string into numerical e.g- {"four": 4, "two": 2}

[ ]:

```
num_feat=dataset.describe().columns
list(num_feat)
```

[ ]:

```
['car_ID',
 'symboling',
 'wheelbase',
 'carlength',
```