

Python Project: Develop prediction models by Kate Rogatina

In this project, we will develop several models that will predict the price of the car using the variables or features. This is just an estimate but should give us an objective idea of how much the car should cost.

Some questions we want to ask in this module

- Do I know if the dealer is offering fair value for my trade-in?
- Do I know if I put a fair value on my car?

In data analytics, we often use **Model Development** to help us predict future observations from the data we have.

A model will help us understand the exact relationship between different variables and how these variables are used to predict the result.

Import libraries:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
```

Load the data and store it in dataframe df:

```
df = pd.read_csv('https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DA0101EN-SkillsNetwork/labs/Data%20files/automobileEDA.csv')
df.head()
```

| | symboling | normalized-losses | make | aspiration | num-of-doors | body-style | drive-wheels | engine-location | wheel-base | length | ... | compression-ratio | horsepower | peak-rpm | city-mpg | highway-mpg |
|---|-----------|-------------------|-------------|------------|--------------|-------------|--------------|-----------------|------------|----------|-----|-------------------|------------|----------|----------|-------------|
| 0 | 3 | 122 | alfa-romero | std | two | convertible | rwd | front | 88.6 | 0.811148 | ... | 9.0 | 111.0 | 5000.0 | 21 | 27 |
| 1 | 3 | 122 | alfa-romero | std | two | convertible | rwd | front | 88.6 | 0.811148 | ... | 9.0 | 111.0 | 5000.0 | 21 | 27 |
| 2 | 1 | 122 | alfa-romero | std | two | hatchback | rwd | front | 94.5 | 0.822681 | ... | 9.0 | 154.0 | 5000.0 | 19 | 26 |
| 3 | 2 | 164 | audi | std | four | sedan | fwd | front | 99.8 | 0.848630 | ... | 10.0 | 102.0 | 5500.0 | 24 | 30 |
| 4 | 2 | 164 | audi | std | four | sedan | 4wd | front | 99.4 | 0.848630 | ... | 8.0 | 115.0 | 5500.0 | 18 | 22 |

5 rows × 29 columns

1. Linear Regression and Multiple Linear Regression

Linear Regression

One example of a Data Model that we will be using is:

Simple Linear Regression

Simple Linear Regression is a method to help us understand the relationship between two variables:

- The predictor/independent variable (X)
- The response/dependent variable (that we want to predict)(Y)

The result of Linear Regression is a **linear function** $\hat{Y} = a + bX$

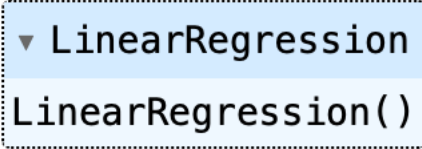
- a refers to the **intercept** of the regression line, in other words: the value of Y when X is 0
- b refers to the **slope** of the regression line, in other words: the value with which Y changes when X increases by 1 unit

Let's load the modules for linear regression:

```
from sklearn.linear_model import LinearRegression
```

Create the linear regression object:

```
In [10]: lm = LinearRegression()  
lm
```

```
Out[10]:   
▼ LinearRegression  
LinearRegression()
```

How could "highway-mpg" help us predict car price?

For this example, we want to look at how highway-mpg can help us predict car price. Using simple linear regression, we will create a linear function with "highway-mpg" as the predictor variable and the "price" as the response variable.

```
X = df[['highway-mpg']]  
Y = df['price']
```

Fit the linear model using highway-mpg:
`lm.fit(X,Y)`

```
In [12]: lm.fit(X,Y)
```

```
Out[12]:  
▼ LinearRegression  
LinearRegression()
```

```
In [ ]: |
```

```
Yhat=lm.predict(X)  
Yhat[0:5]
```

Output:

```
array([16236.50464347, 16236.50464347, 17058.23802179, 13771.3045085 ,  
       20345.17153508])
```

What is the value of the intercept (a)?

```
lm.intercept_
```

Output:

```
38423.305858157415
```

What is the value of the slope (b)?

```
lm.coef_
```

Output:

```
array([-821.73337832])
```

What is the final estimated linear model we get?

As we saw above, we should get a final linear model with the structure:

$\hat{Y} = a + b X$

Plugging in the actual values we get:

Price = 38423.31 - 821.73 x highway-mpg

We create a linear regression object called "lm1":

```
lm1 = LinearRegression()  
lm1
```

Now, we train the model using "engine-size" as the independent variable and "price" as the dependent variable:

```
X=df[['engine-size']]  
Y=df[['price']]  
lm1.fit(X,Y)
```

We find the slope and intercept of the model:

```
# Slope  
lm1.coef_
```

```
Output:  
array([[166.86001569]])
```

```
# Intercept  
lm1.intercept_
```

```
Output:  
array([-7963.33890628])
```

The equation of the predicted line. To find it, we can use x and yhat or "engine-size" or "price":

```
# using X and Y  
Yhat=-7963.34 + 166.86*X
```

```
Price=-7963.34 + 166.86*df['engine-size']
```