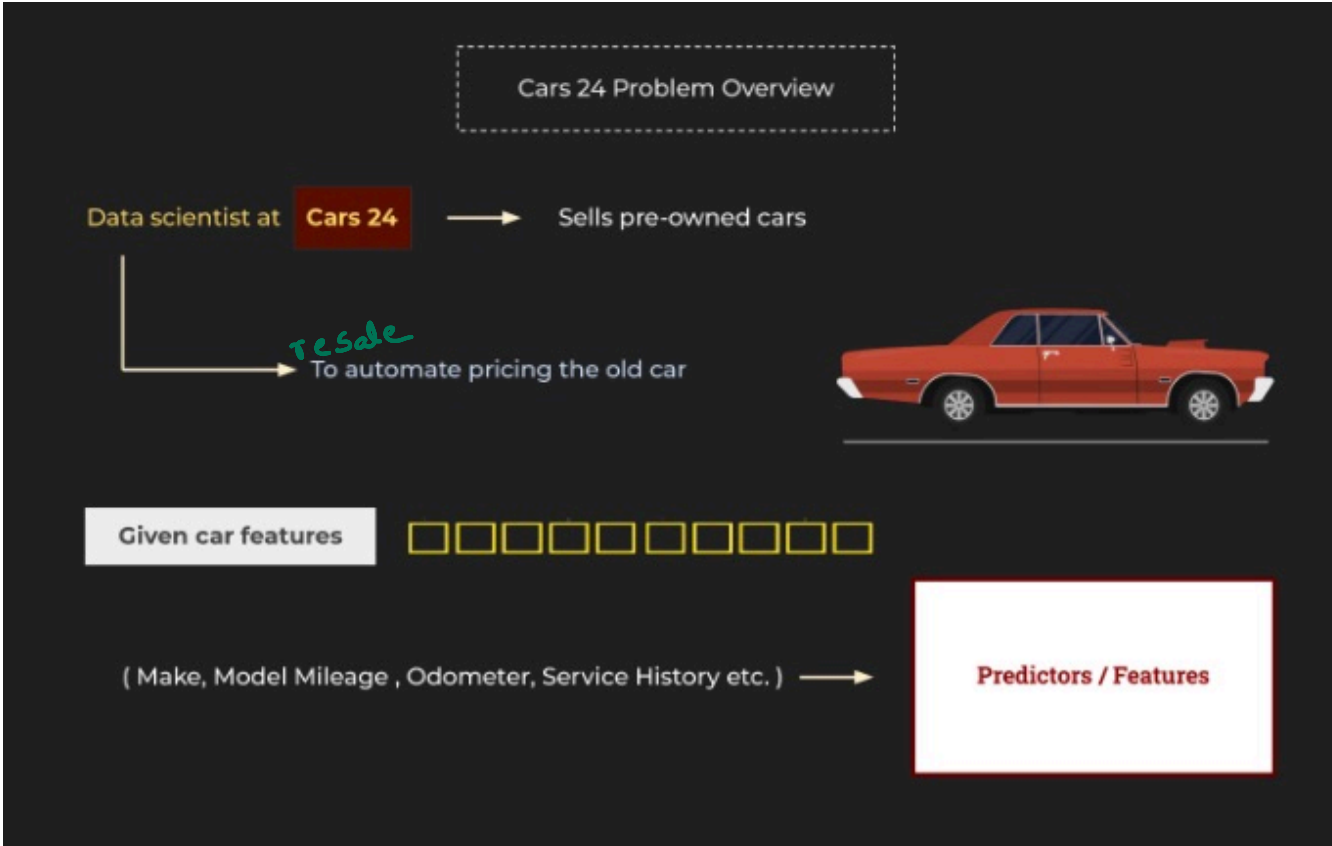
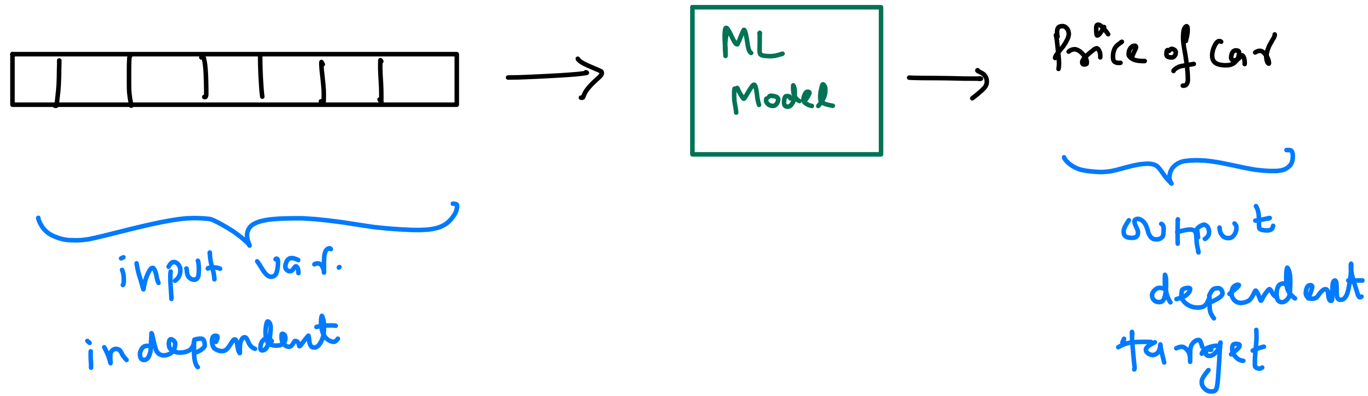
Linear Regression

Agenda

1. Linear Regression
2. Cars24 Case
3. Intuition lin. reg
4. Maths → Algebraic
5. Sklearn → Code

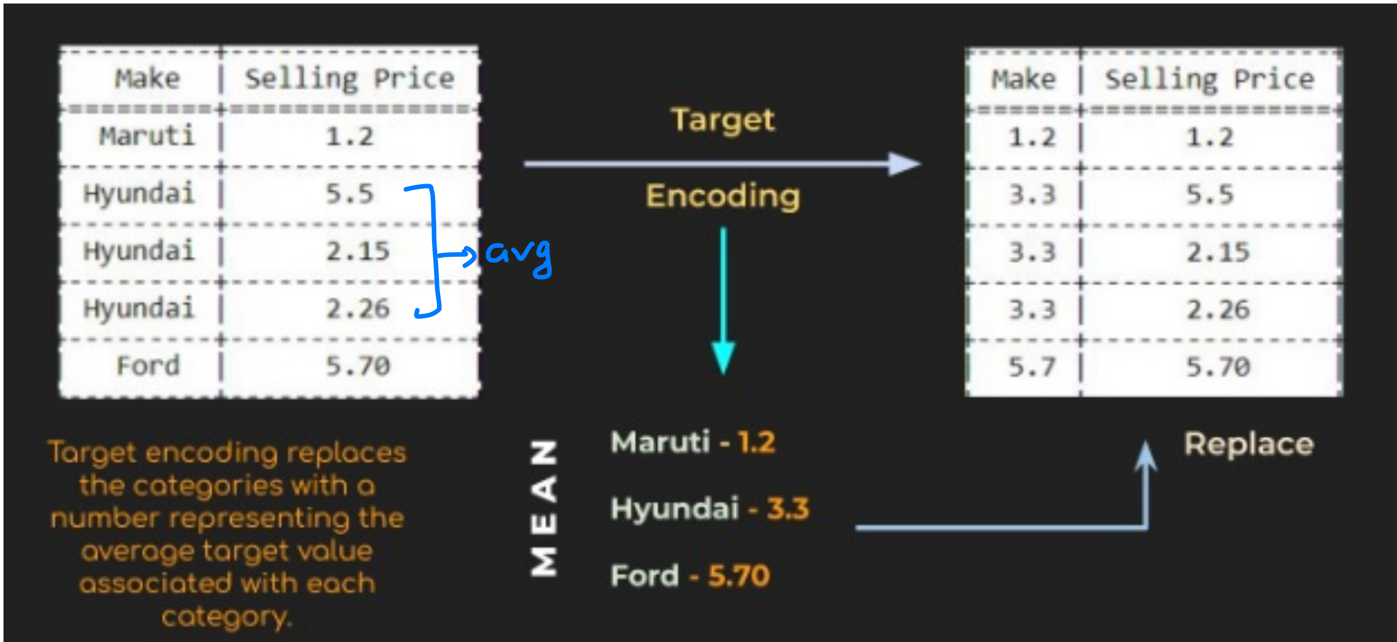

Cars 24 Problem Overview
Data scientist at Cars 24 → Sells pre-owned cars
resale
To automate pricing the old car
Given car features
(Make, Model Mileage , Odometer, Service History etc.) →
Predictors / Features



OHE $\rightarrow$ 3300 New Cols.

"Curse of Dimn"

"Target Encoding"



Make	Selling Price
Maruti	1.2
Hyundai	5.5
Hyundai	2.15
Hyundai	2.26
Ford	5.70

Make	Selling Price
1.2	1.2
3.3	5.5
3.3	2.15
3.3	2.26
5.7	5.70

Min Max Scaling (Normalization) / Standardisation

$$X_{Scaled} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Mix - Max Scaler

Km_Driven
10,000
5000
2000

Scaling $\longrightarrow$

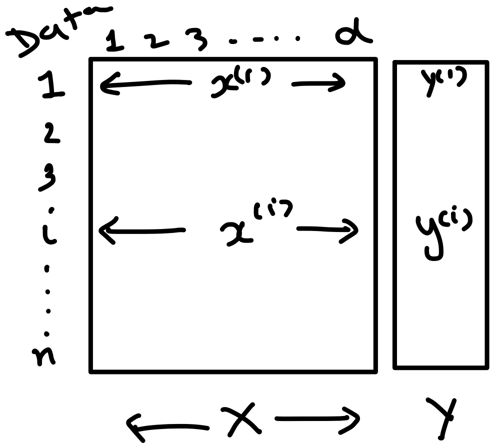
Km_Driven
1
0.375
0

$\rightarrow$ [0, 1]

$X_{min} = 2000$

$X_{max} = 10000$

Goal of ML



$x^{(i)} \longrightarrow$ ML Model Algo $\longrightarrow \hat{y}^{(i)}$ [Predicted]

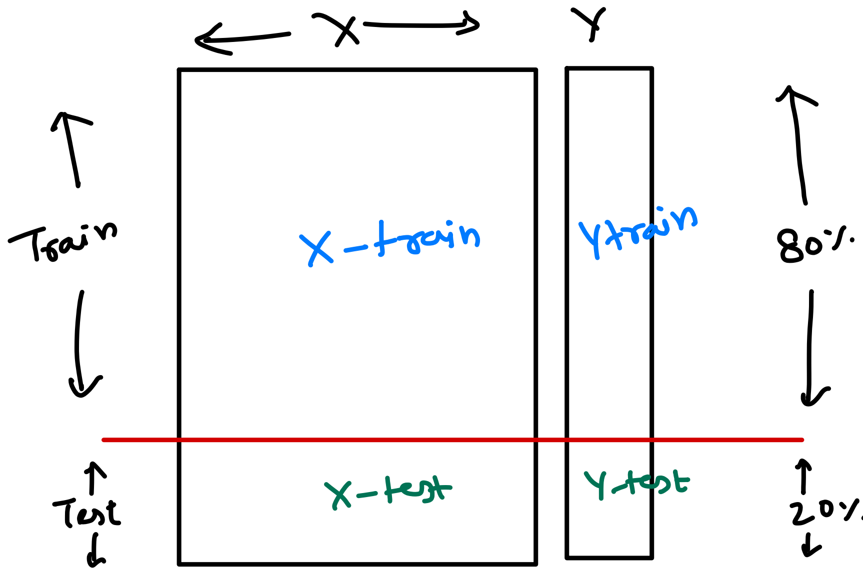
ideally,

$y^{(i)} \approx \hat{y}^{(i)}$

original Predicted

$x_{new} \longrightarrow$ ML Algo $\longrightarrow \hat{y}_{new}$

Train | Test



ML Train

↓

(X-train , y-train)

Evaluate

↓

(x-test , y-test)

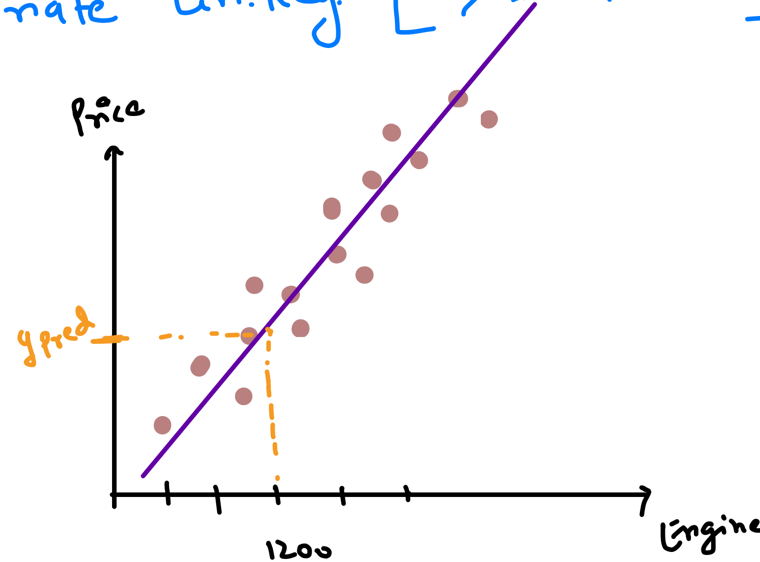
Intuition L.R

→ Univariate Lin. Reg. [1 input var.]

→ Multivariate Lin. Reg. [> 1 input var]

input → output

engine → Price



New Car (1200 cc)

↓

input this in line eq.

↓

get Price

$$x^{(i)} \xrightarrow[\text{line eq}]{f(\,)} \hat{y}^{(i)}$$

St. line →

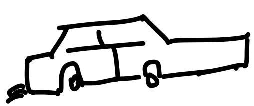
$$y = mx + c$$

$$\boxed{y = \underbrace{w_1}_{\text{weight}}\, \underbrace{x}_{\text{Engine}} + \underbrace{w_0}_{\text{intercept}}}$$

For. eg. →

$$\hat{y} = \overset{w_1}{200} \cdot x + \overset{w_0}{50000}$$

(1000cc)
newcar



$$\text{Price} = 200 \times 1000 + 50000$$

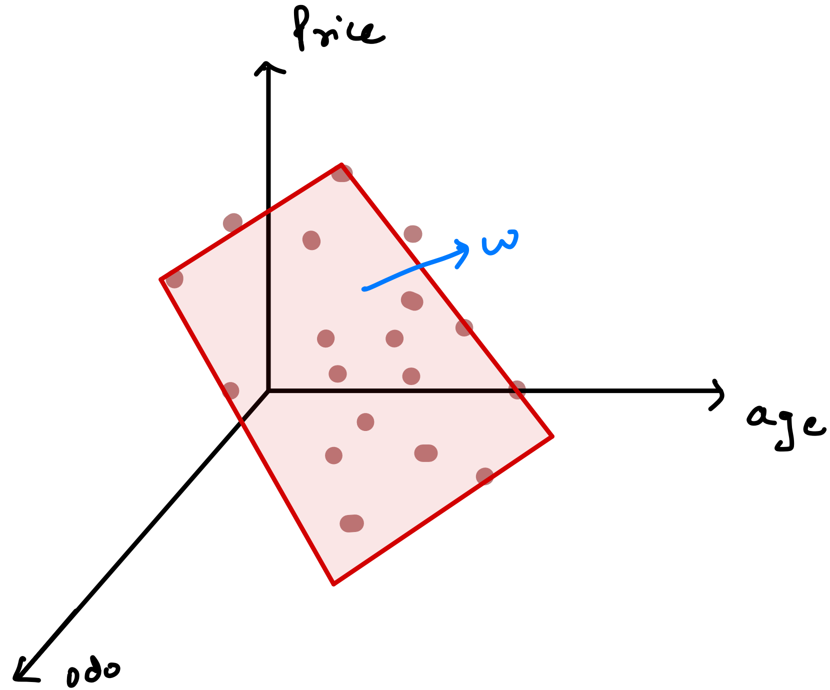
$$= 2.5\ L$$

2-input features.

age , odometer

x_1 x_2

age	odo	y



$$\boxed{y = \underbrace{w_1}_{\text{weight}}\overbrace{x_1}^{\text{age}} + \underbrace{w_2}_{\text{weight}}\overbrace{x_2}^{\text{odo}} + \underbrace{w_0}_{\text{intercept}}}$$

$$y = -10000x_1 - 10.x_2 + 5000$$

d-features

$$y = w_1 x_1 + w_2 x_2 + \dots + w_d x_d + w_0$$

$$\boxed{y = w^T x + w_0}$$

$$w = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{bmatrix} \quad x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix}$$

1 feature $\longrightarrow$ line 2D

2 features $\longrightarrow$ plane 3D

$\vdots$

d features $\longrightarrow$ hyperplane "d+1"

Evaluation Metrics

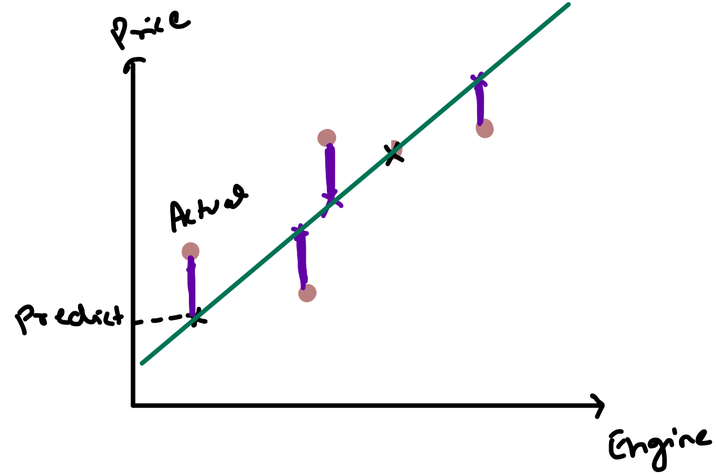
$x^{(1)} \rightarrow y^{(1)} - \hat{y}^{(1)} \rightarrow e^{(1)}$

$x^{(2)} \rightarrow y^{(2)} - \hat{y}^{(2)} \rightarrow e^{(2)}$

$x^{(3)} \rightarrow y^{(3)} - \hat{y}^{(3)} \rightarrow e^{(3)}$

$x^{(4)} \rightarrow y^{(4)} - \hat{y}^{(4)} \rightarrow e^{(4)}$

$x^{(5)} \rightarrow y^{(5)} - \hat{y}^{(5)} \rightarrow e^{(5)}$



$$\text{Total error} = \sum_{i=1}^{n} e^{(i)}$$

$$\text{Total error} = e_1 + e_2 + e_3 + e_4 + e_5$$

$$\text{Total error} \Rightarrow 5 + (-4) + (+3) + (0) + (-1)$$
$$\Rightarrow 3$$

$|5| \to 5$

$|-4| \to +4$

$|3| \to 3$

$|-1| \to +1$

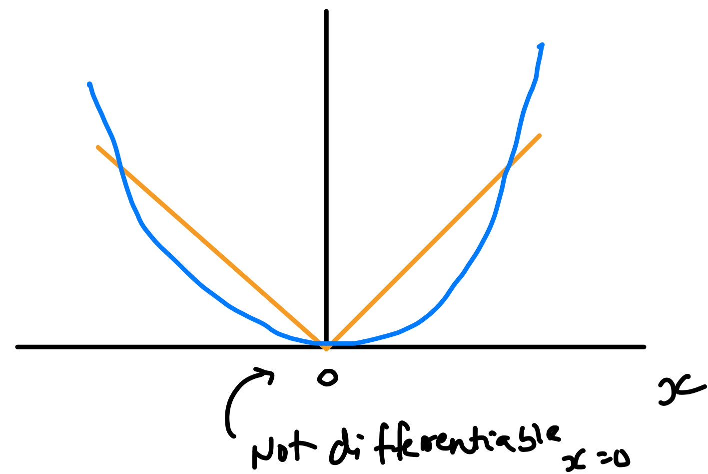
$$\text{Error} = \sum_{i=1}^{n} |e_i| \Rightarrow \sum_{i=1}^{n} \left| y^{(i)} - \hat{y}^{(i)} \right|$$

mean Absolute Error

$$\boxed{MAE = \frac{1}{n} \sum_{i=1}^{n} \left| y^{(i)} - \hat{y}^{(i)} \right|}$$

mean squared error

$$\boxed{MSE = \frac{1}{n} \sum_{i=1}^{n} \left(y^{(i)} - \hat{y}^{(i)} \right)^2}$$



MSE

M_1 9.62 ✓ Better Model

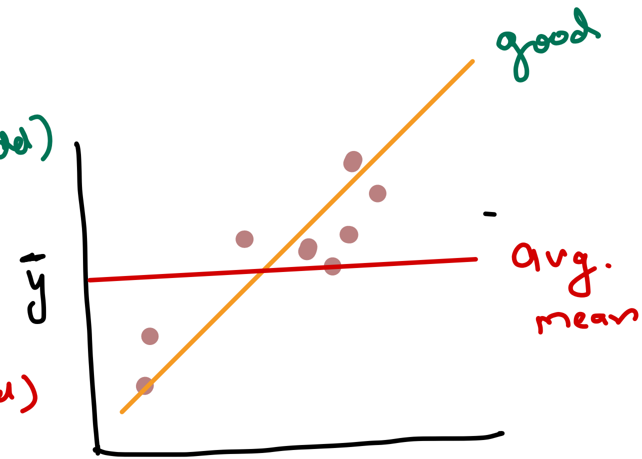
M_2 15.41

R2_Score / R_squared / Coeff. of Determination

~~accuracy~~ → Performance

$$\text{R2\_Score} = 1 - \frac{SS_{res}\ (\text{mse of good model})}{SS_{total}\ (\text{mse of avg. model})}$$

mean



$$R_2\text{-Score} \Rightarrow 1 - \frac{\cancel{\frac{1}{n}} \sum_{i=1}^{n} \left(y^{(i)} - \hat{y}^{(i)}\right)^2}{\cancel{\frac{1}{n}} \sum_{i=1}^{n} \left(y^{(i)} - \bar{y}\right)^2}$$

$\boxed{R_2\text{-Score}}$

$R_2 = 1$

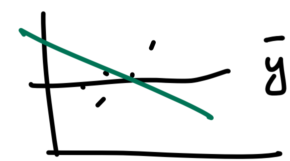
$R_2 = 0$

Prediction = avg

avg model

$R_2 = -ve$

worse than avg model



$$\text{R2-Score} = 1 - \frac{SS_{res}}{SS_{total}}$$

Generally

$$[0, 1]$$

Bad Best

Quiz time!

Quiz Ended!

In a multiple linear regression with five features, the coefficient of determination R2 is found to be 0.85. What does this value indicate about the model's performance?

33 users have participated

A	The model explains 85% of the variation in the target variable	27%
B	The model's predictions are 85% accurate	24%
C	The model has an 85% probability of making correct predictions	24%
D	The model is 85% confident in its predictions	24%

https://colab.research.google.com/drive/1ajDuNR_-K_9mozthz5BJuYBZYgcXcsiZ?usp=sharing