Agenda

- Cars24 Data
- Data Notation
- Goal of ML
- Linear Regression Intuition
- SkLearn Implentation
- Loss function and Evaluation Metric

Cars24 Dataset

- Goal : To build a ML Model that can predict price of used cars accurately

Target → Sell_Price

Features : 17

Ordinal Encoding:

OHE: X

Target Encoding

↓

replace each of the category with mean of that category

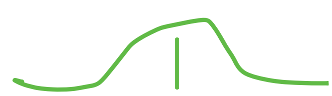
↓

(Target)

Feature Scaling

① Min-max Scaler: [-1 , 1]

② Standard Scaler:



mean = 0
std = 1

① Feature Scaling make model training Faster

② Weights are more Interpretable

[10 20 -5]

[-1, 1]

max_val = 20
min_val = -5

$$x_i = \frac{x_i - min}{max - min}$$

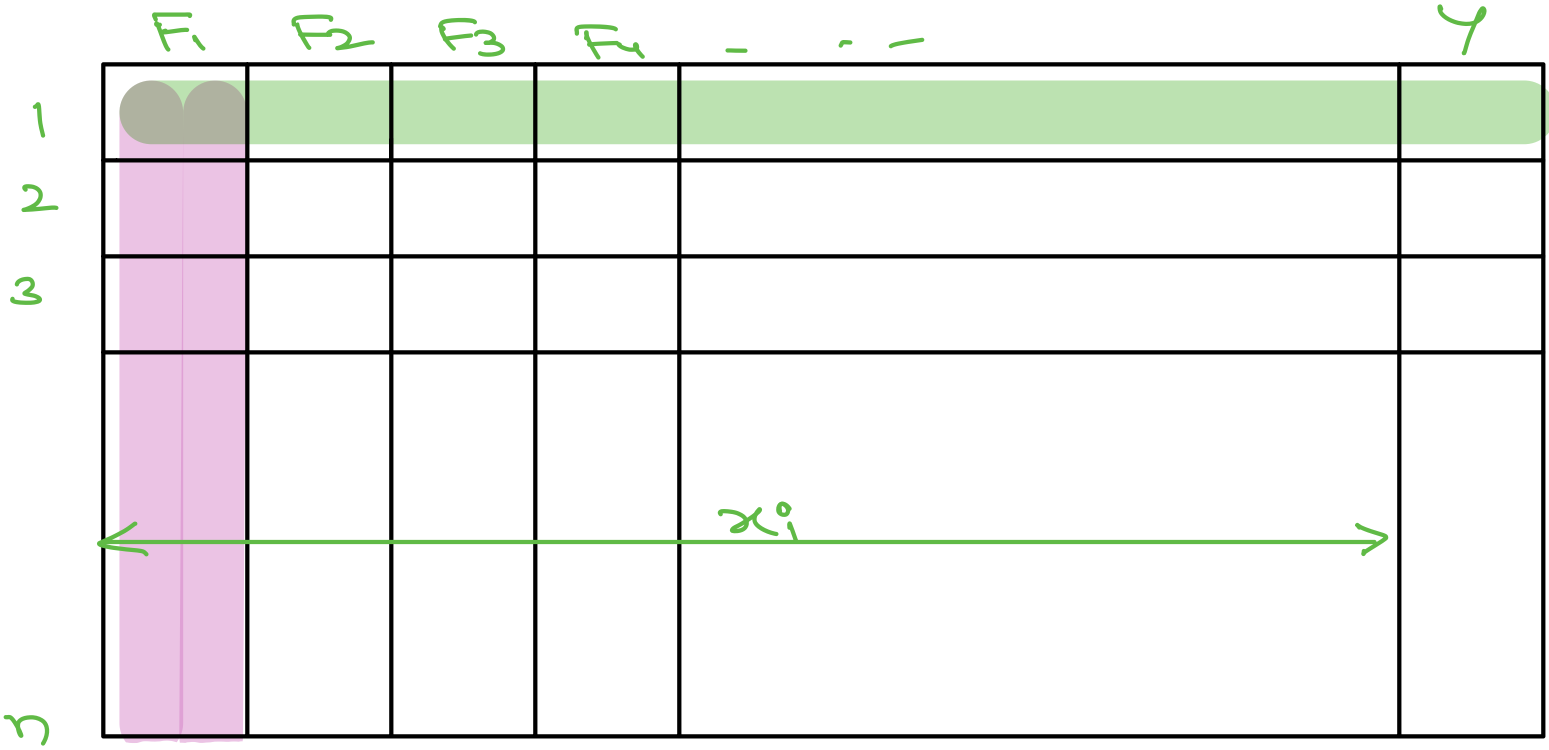
⇒ Target_Var : Selling_price

[a, b]

$\mathbb{R}$

⇒ Regression

Data Annotation

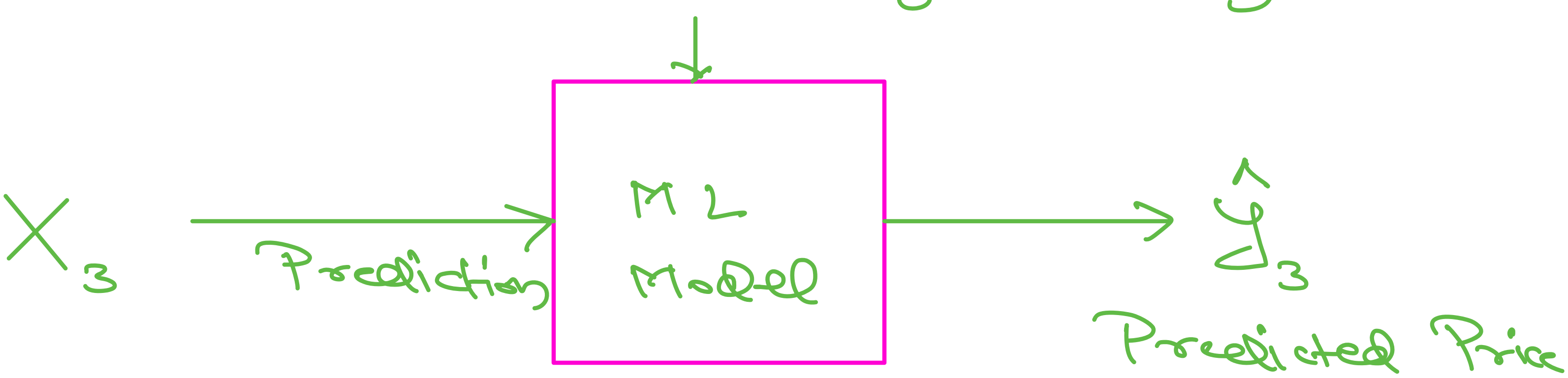


$x_3 \in [\quad]$

y_i = Scalar, y_3 = 1000 dollars

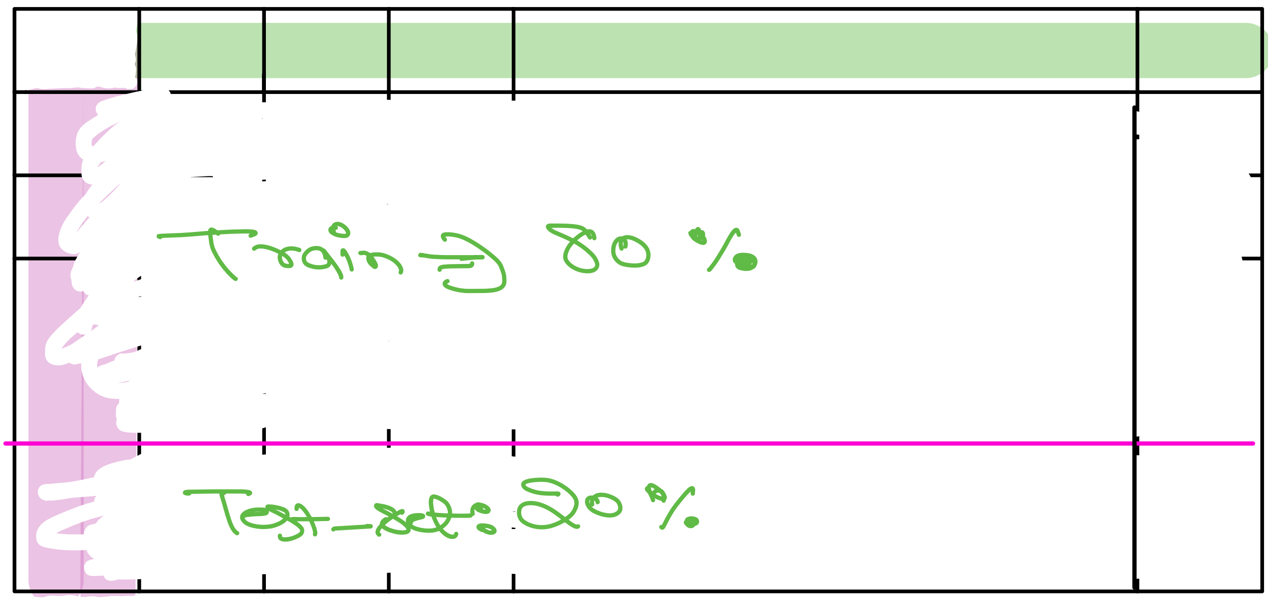
$$\boxed{X_{n \times d}, \; Y / y}$$

Training (X, Y)

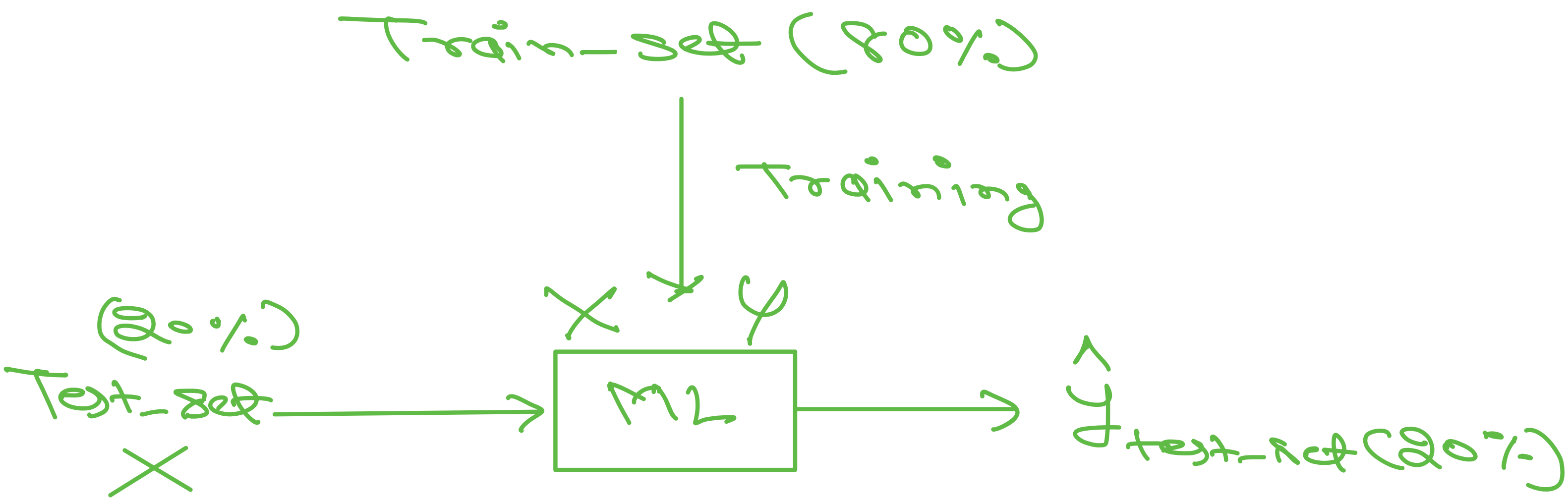


$\boxed{\hat{y} \;\; vs \;\; y}$

Predicted vs Actual



$$1 \begin{cases} 0.8 \rightarrow 80\% \\ 0.2 \rightarrow 20\% \end{cases}$$

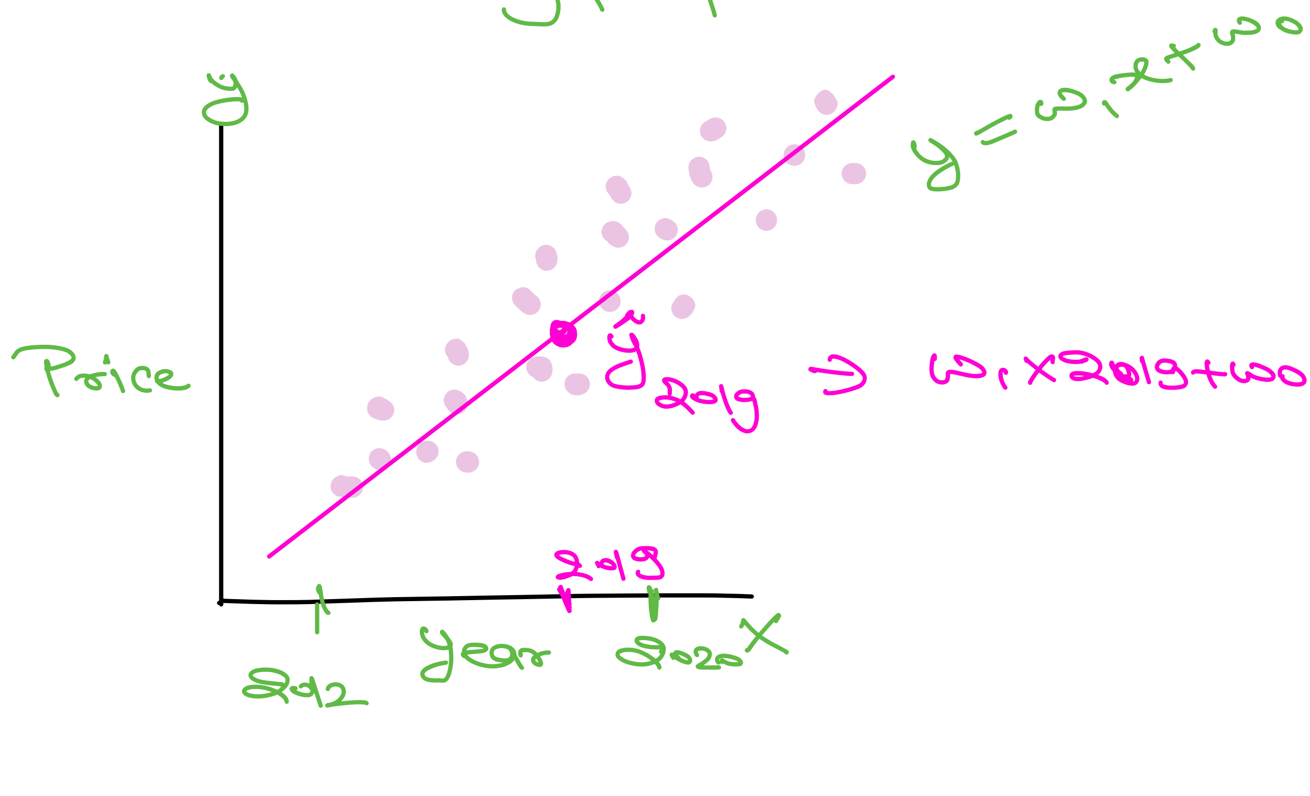


$y_{test\_data} - \hat{y}_{test\_set}$

Linear Regression Intuition

① Linear Model

Linear Hyperplane



$$y = mx + c$$

$$\downarrow$$

$$\boxed{y = w_1x + w_0}$$

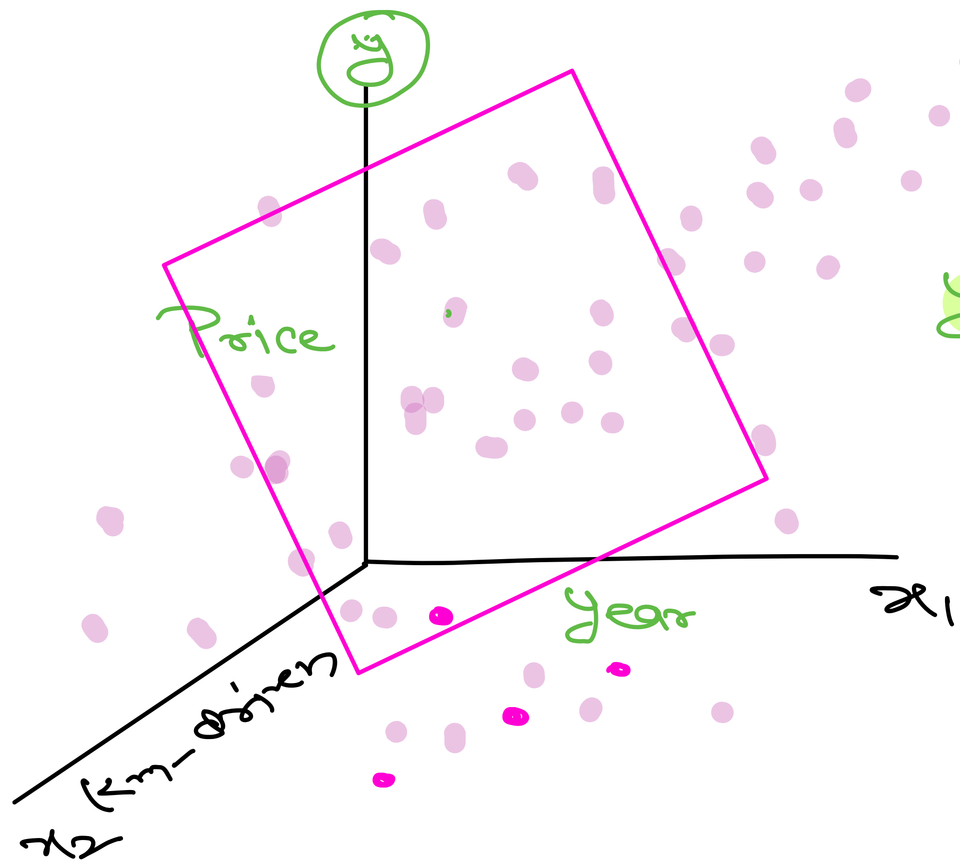
$w_1 = 2$
$w_0 \Rightarrow 5$ } Training

$y_i = 2x + 5$

x 2012 $\Rightarrow$ $\hat{y} \Rightarrow 2 \times 2012 + 5 \Rightarrow 4029$ dollars

$y = 3500$ $\Rightarrow$

Diff $\Rightarrow 4029 - 3500$
$\Rightarrow 529$



$y = w_1 x_1 + w_2 x_2 + w_0$

w_1 } Training ⇒ 2
w_2 } 3
w_0 -2

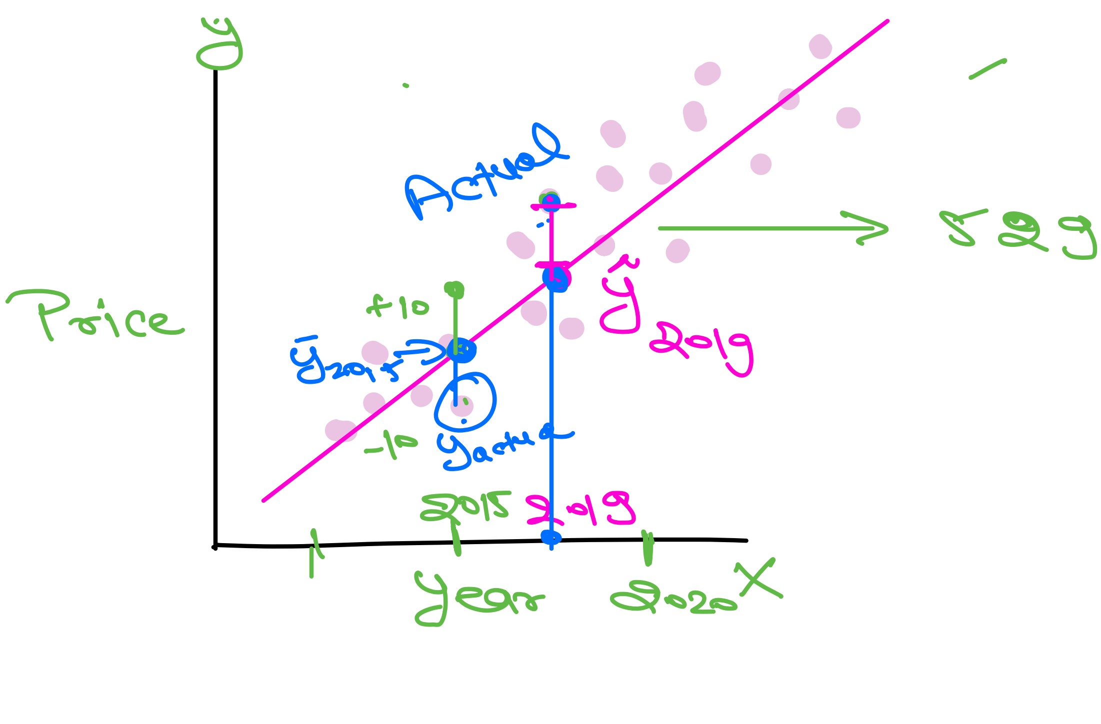
$x_1 \Rightarrow 2012$
$x_2 \Rightarrow 1000$

$y = 2 \times 2012 + 3 \times 1000 - 2$

2 features 1 target

How many parameters ⇒ 3

Dimension of Hyperplane ⇒ 2d

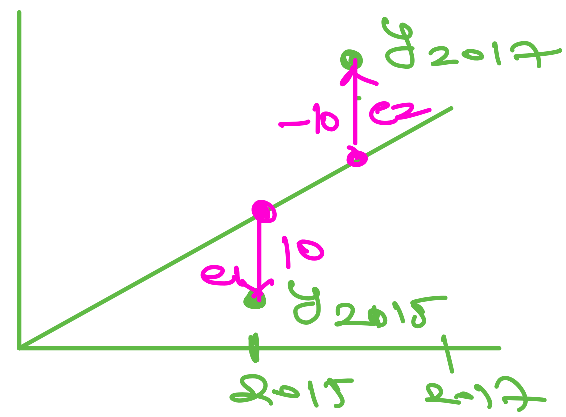


Distance of Each of the prediction to their actual Values

$$\Rightarrow \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ \vdots \end{bmatrix} \quad 20\% \text{ test\_set}$$

$$e_i \Rightarrow \hat{y}_i - y_i$$

⇒ Sum ⇒ $e_1 + e_2 + e_3 - - - -$



$e_1 + e_2 \Rightarrow 0$

② Sum of absolute

$|e_1| + |e_2| - - - -$

$|e_1| + |e_2| = 20$

For a Good Model

min (Sum of absolute Error)

↓

MAE = Mean Absolute Error

$$\min \boxed{\frac{1}{n} \sum_{i=0}^{n} |e_i|}$$

③ Sum of errors Square ✓

Total Error $= e_1^2 + e_2^2 \ldots$

↓

$$\boxed{MSE = \frac{1}{n} \sum_{i=0}^{n} e_i^2}$$

MAE	MSE

Fill this H.W

① In Regression, Loss function itself can be used as Evaluation metric

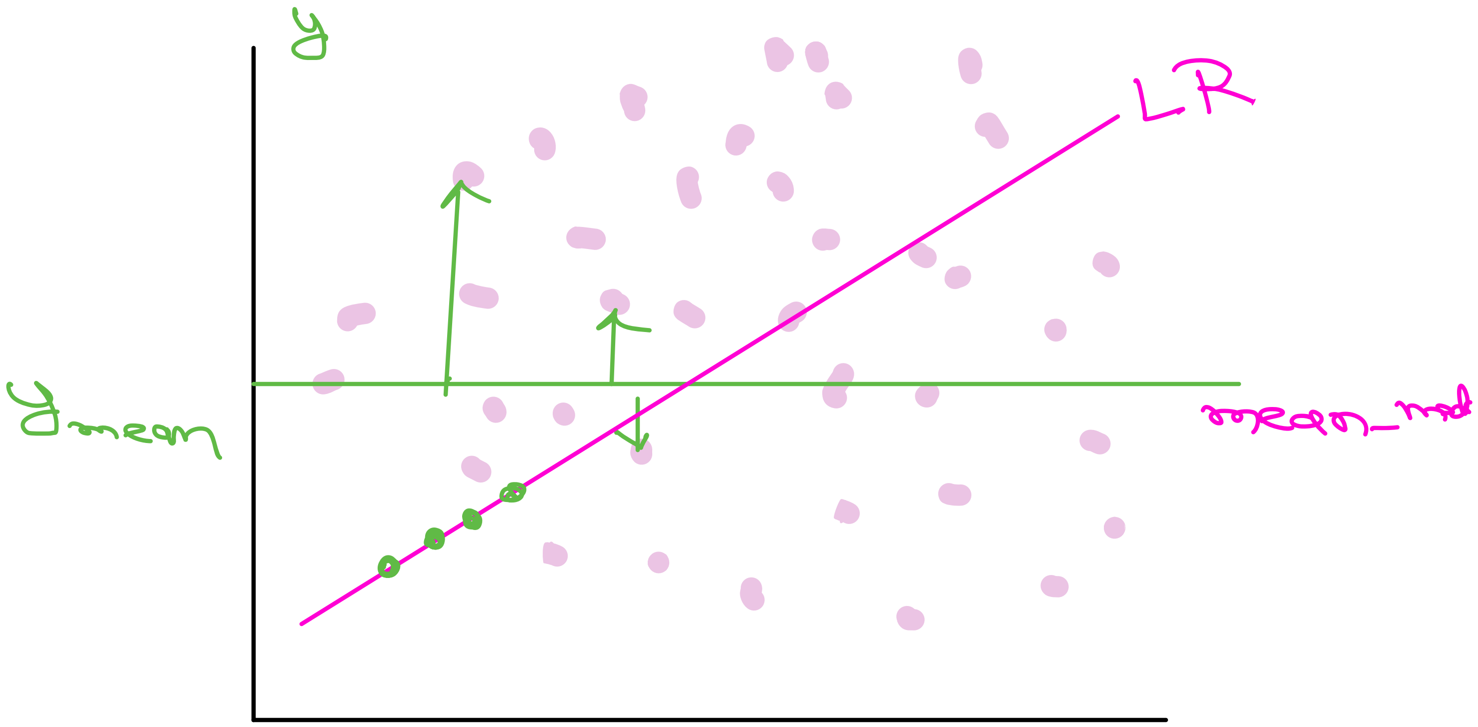
MSE and MAE so minimum

m_1 ⇒ $\boxed{2000}$

MSE

m_2 ⇒ $\boxed{1500}$

m_2 is better than m_1



To compare any model with mean_model as Baseline

Error

$$SS_{Total} \text{ (Mean)} = \sum_{i=0}^{n} (y_i - y_{mean})^2$$

$$SS_{res} \text{ (LR)} = \sum_{i=0}^{n} (y_i - \hat{y}_i)^2$$

R^2
(r2-score)

$$1 - \frac{SS_{Reg}}{SS_{mean}}$$

How well Linear Regression Model is doing compared to mean Model

Todo: Calculate R2-score

Case 1) LR has same Error as Mean

Case 2) LR has 0 Error

Case 3) LR has ∞ Error

① Ordinal Encoding

Education_status

B.Tech ⇒ 3

12^{th} ⇒ 2

10^{th} ⇒ 1

② OHE

Make		Maruti	BMW	Ford
Maruti		1	0	0
BMW	⇒	0	1	0
⇒ Ford		0	0	1

③ Target Encoding

Make

Maruti	ymean of Maruti
BMW	ymean of BMW
Ford	ymean of Ford