Data Science questions

① Give me a name of all the Regression Algo that you know.

1. linear Regression
2. Polynomial Regression
3. Ridge Regression
4. Lasso Regression
5. Elastic Net Regression
6. Support vector Regression.
7. Decision tree Regression.
8. Random Forest Regression
9. Gradient Boosting Regression
10. xgboost Regression
11. LightGBM Regressor
12. CatBoost Regressor
13. K-Nearest Neighbour Regressor.
14. Bayesian Ridge Regressor
15. Huber Regressor
16. Theil-Sen Regressor
17. Quantile Regressor
18. Tweedie Regressor
19. NN Regressor

③ Give me a list of All the classification Algorithm that ya know

① Logistics Regression.

② K-Nearest Neighbour

③ Support Vector machine

④ Decision tree

⑤ Random Forest

⑥ Gradient Boosting.

⑦ XG Boost

⑧ Light GBM

⑨ Cat Boost

⑩ Naive Bayes

⑪ NN Classification

⑫ CNN

⑬ RNN

⑭ LSTM

⑮ GRU

⑯ Ada Boost

⑰ Bagging classifier

⑱ Extra tree classifier

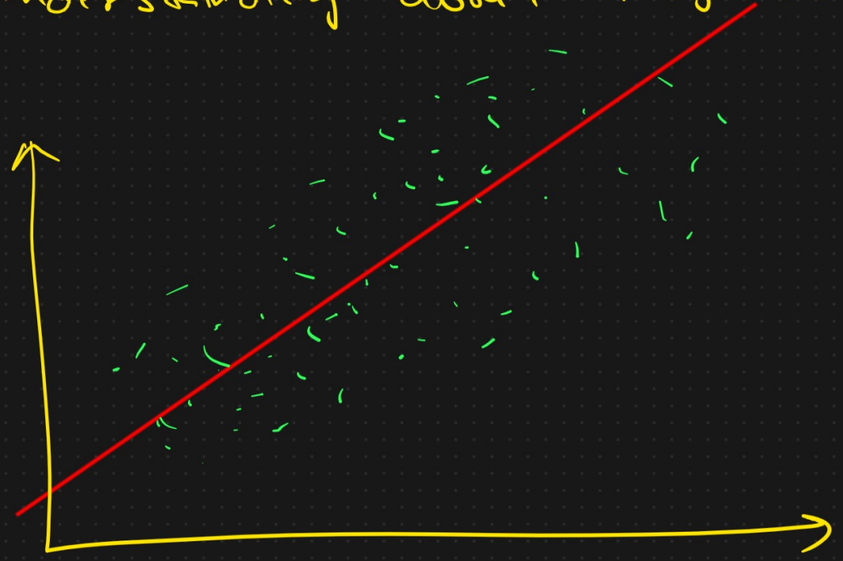
⑲ LDA

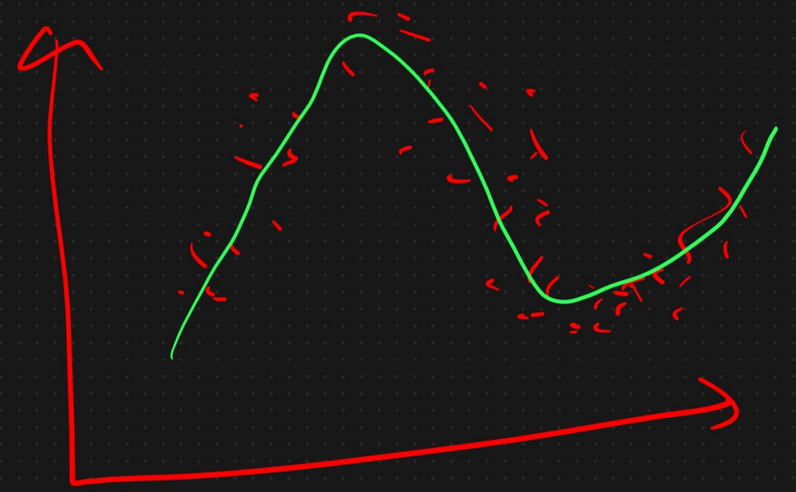
⑳ QDA

㉑ SGD

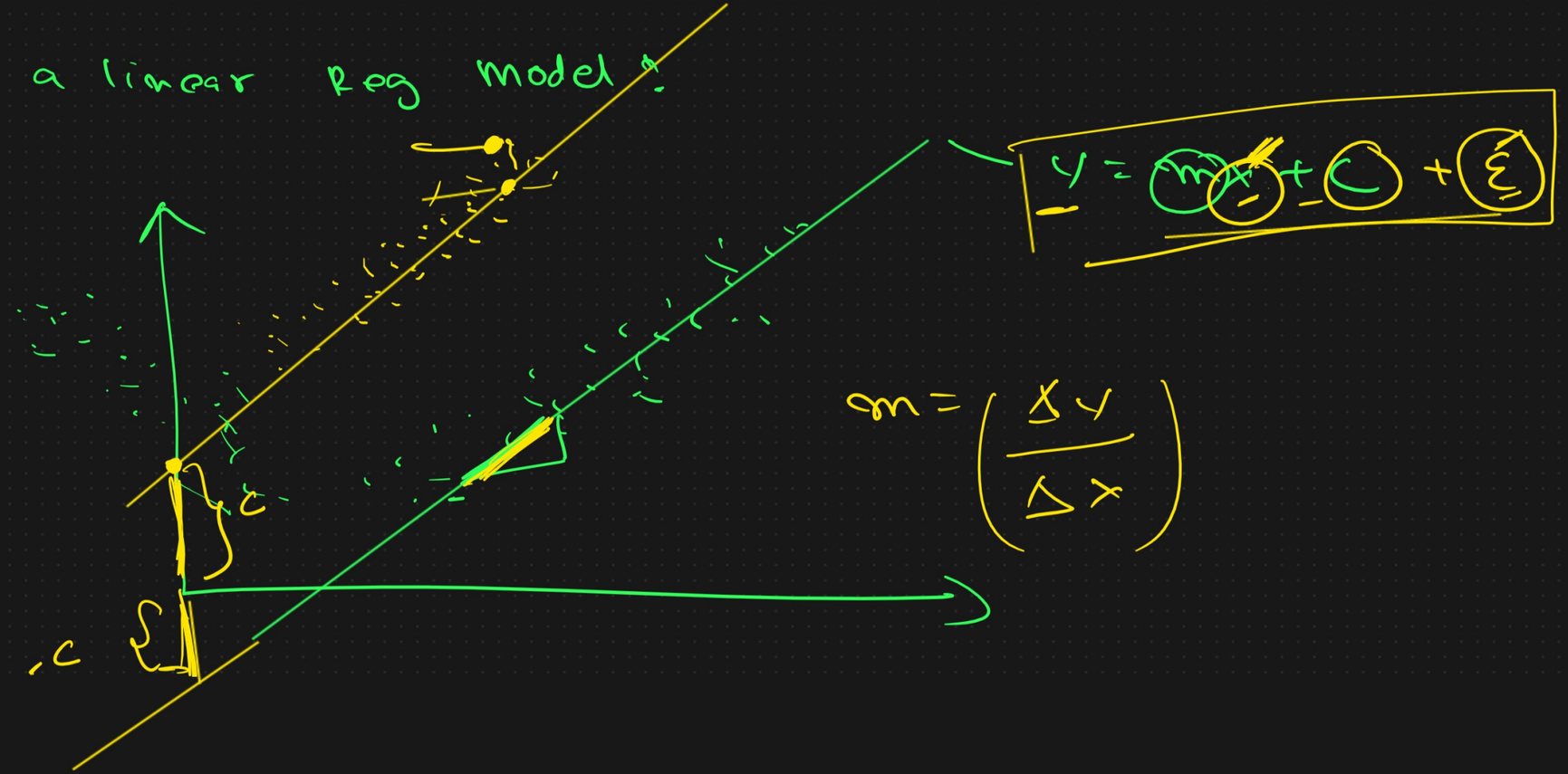
㉒ (OvR) & (OvO)

③ what is your understanding about Regression ?



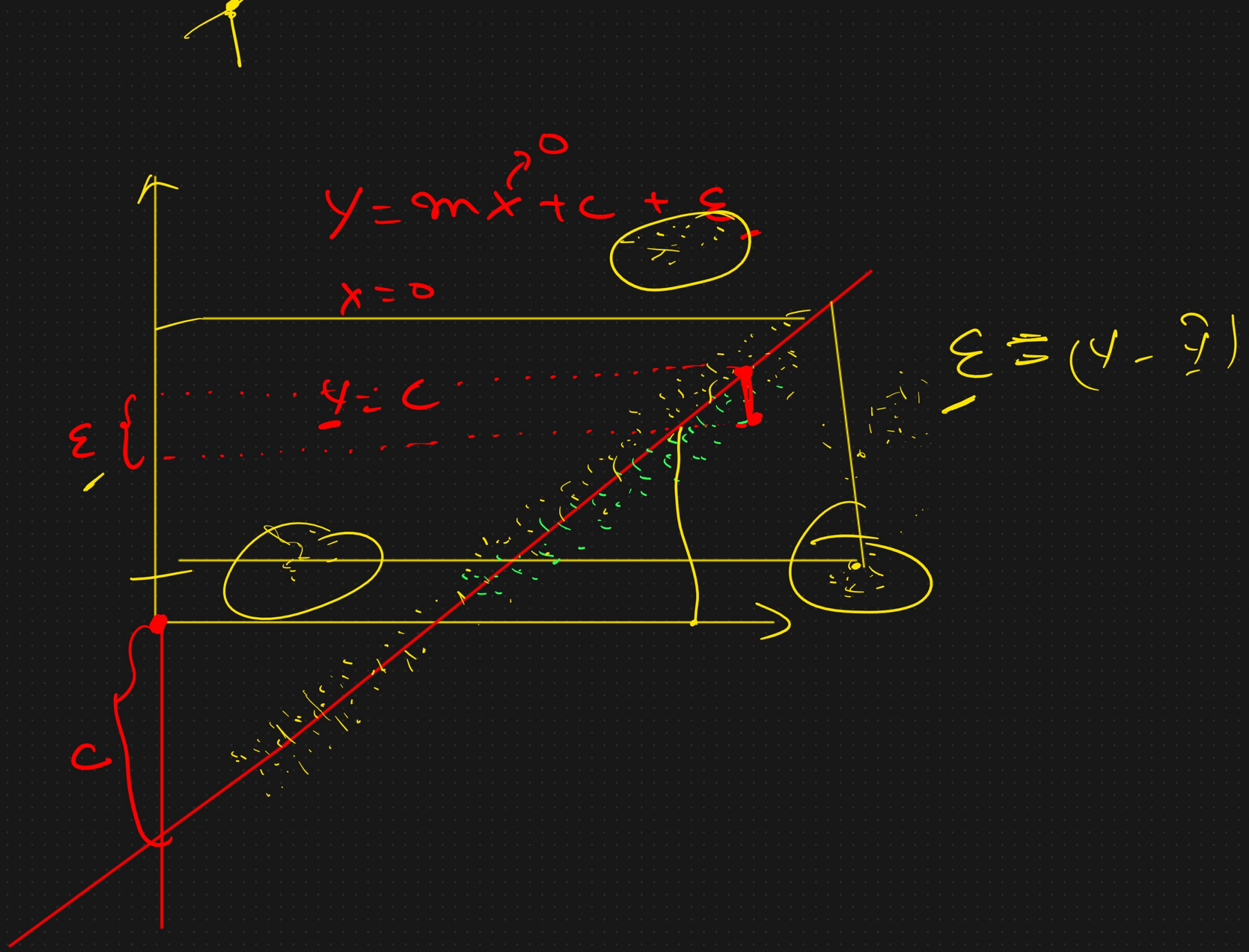


④ what is a linear Reg model ?



$$y = mx + c + \varepsilon$$

$$m = \left(\frac{\Delta y}{\Delta x}\right)$$


$y = mx^0 + c + \varepsilon$
$x = 0$
$y = c$
ε
$\varepsilon = (y - \hat{y})$
c

⑤ When you are using linear Regression what are the factors inside data you have considered.

① linearity.

② No Autocorrelation (independence of Error) →

③ Homoscedasticity. ((constant variance of Error)

④ Normality in Error →

⑤ No Multicollinearity.

⑥ Exogeneity

⑥ What do you understand by Ridge Regression? (L2)

⑦ What is overfitting

Cost function

$$J(\vec{\beta}) = \underbrace{\sum (y_i - \hat{y}_i)^2}_{\text{SSE}} + \underbrace{\left[\alpha \sum \beta^2\right]}_{L^2 \text{ Penalty term.}}$$

α → hyper parameter

Wherever we have multiple co-related features

⑧ L1-> Lasso

$$J(\beta) = \sum (y - \hat{y})^2 + \lambda \sum |\beta|$$

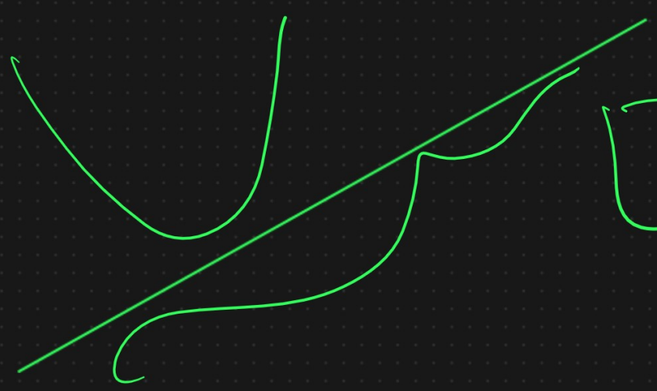
↓

L1 Pamelty term

Feature Selection.

$y = mx + c$

⑨ what do you understand by Poly Regression?



$$\left[y = \beta_0 + \beta_1 x_1 + \beta_2 x_2^2 + \beta_3 x_3^3 + \dots + \beta_n x_n^n + \varepsilon \right]$$

⑯ Lets talk about logistic Regression . . . .

$$P(Y=1|X) = \frac{1}{1+e^{-(mx+c \ldots\ldots)}}$$

↗ (Sigmoid)

e → 2.71

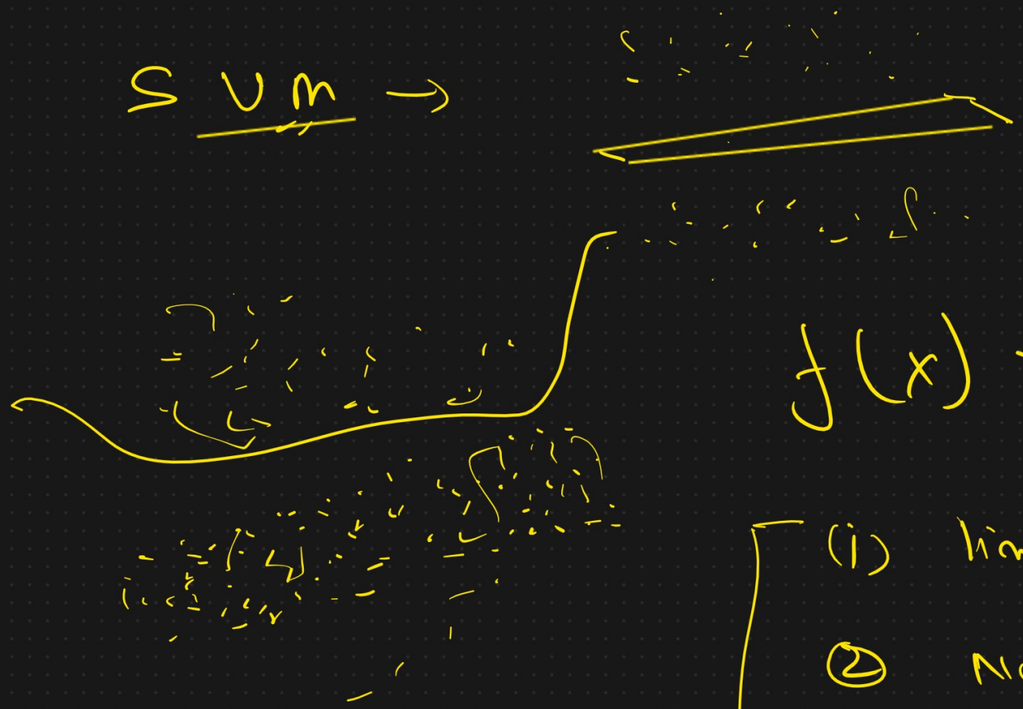
$\geq 0.5 \rightarrow 1$ ✓

$< 0.5 \rightarrow 0$ ✓

(One vs Rest) / (One vs All) ↗ Softmax

$$P(Y=k|x) = \left(\frac{e^{\beta x}}{\sum e^{\beta x}} \right)$$

① SVM →



$$f(x) = (\vec{w} \cdot \underline{x} + \vec{b})$$

- (1) linear SVM
- (2) Non linear SVM → Kernal trick (RBF, Polynomial)
- (3) SVR

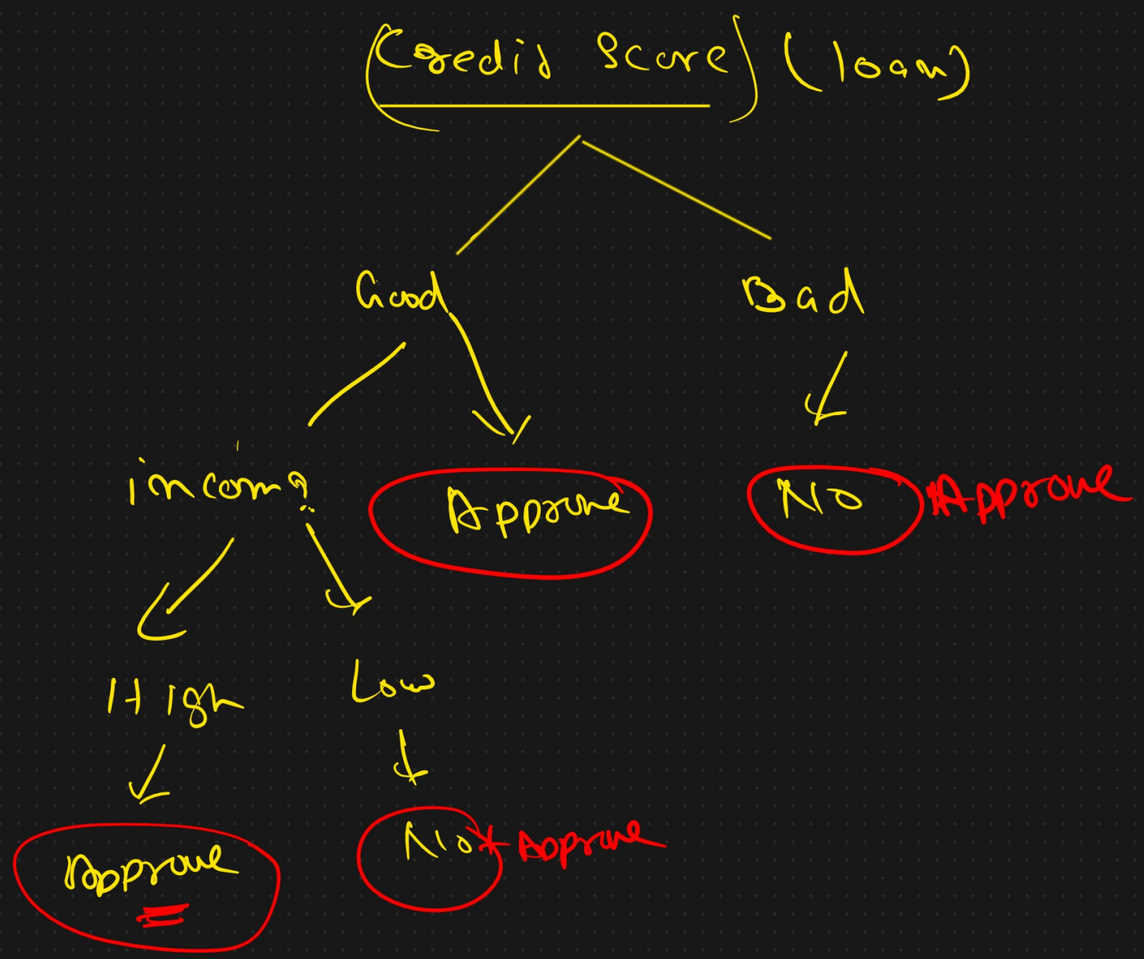
① multiple Features

② Data is Not linearly Separable.

③ Small and Huge data

① high-dimension data.

(12) Decision tree ?



$$Gini = 1 - \sum p_i^2$$

low Gini = Better Split.

$$\text{Entropy} = -\sum p_i \log_2(p_i)$$

lower Entropy = more pure Node

(13) Random Forest ?

(14) Boosting Algo ⟶ XGBoost

(15) XGBoost (Extream Gradient Boosting)

① For Boosting it usages Gradient Descent.

② Prevent overfitting

③ Parallel Computation.

④ Missing Value Automatically

(15) Stop Automatically..

(16) MAE →

$$\underline{MAE} = \frac{1}{n} \sum_{i=1}^{n} \underbrace{|y_i - \hat{y}_i|}_{\text{Absolute error}}$$

(17)

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

(18)

$$\underline{RMSE} = \sqrt{MSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$$

(19) R^2 (R-Squared)

$$R^2 = 1 - \frac{SS_{res}}{SS_{total}}$$

SS_{res} → Residual Sum of Square → $\sum (y - \hat{y})^2$

SS_{total} → Total Sum of Square = $\sum (y - \bar{y})^2$

R^2 is - ve ⟹

$$\text{Adjusted } R^2 = 1 - \frac{(1-R^2)(n-1)}{n-p-1}$$

n → No of data points

p → Number of independent variable

Confusion Matrix

Type I

Type II

A/P →	1 (T)	0 (F)
1 (T)	TP ✓	FN (Ty-2)
0 (F)	FP (Ty 1)	TN ✓

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{F1 score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$\text{Specificity} = \frac{TN}{TN + FP}$$

$$\text{False Positive Rate (FPR)} = \frac{FP}{FP + FN}$$