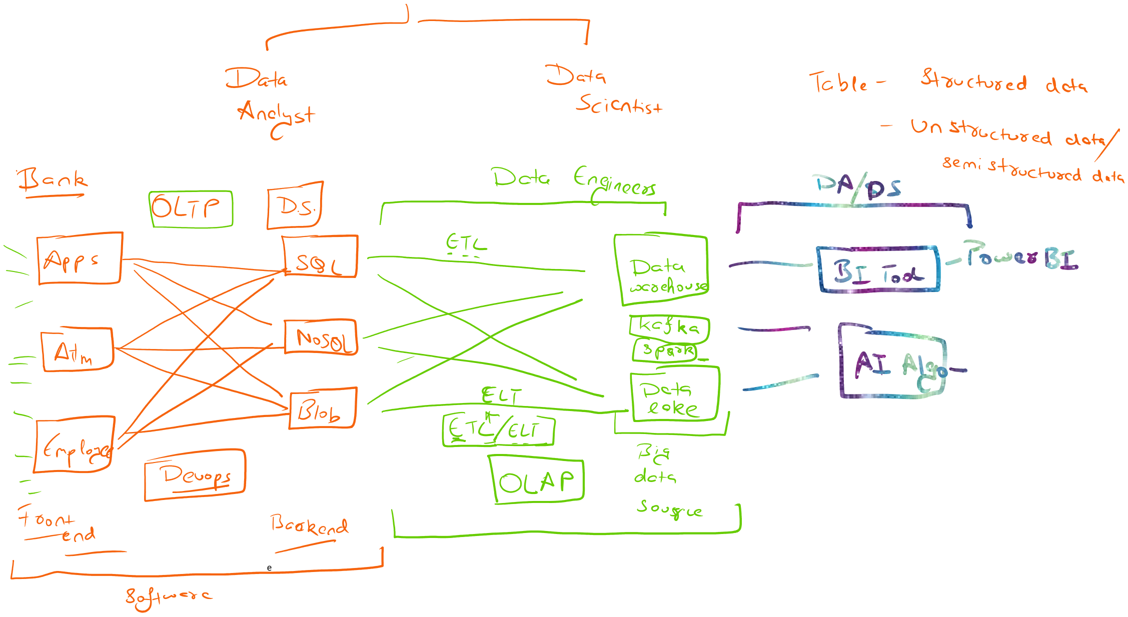
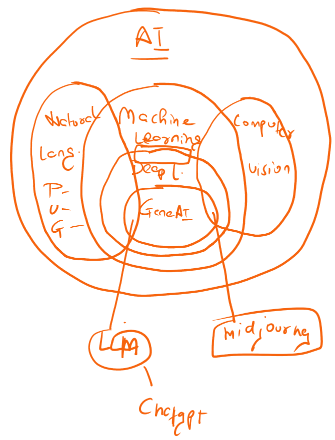
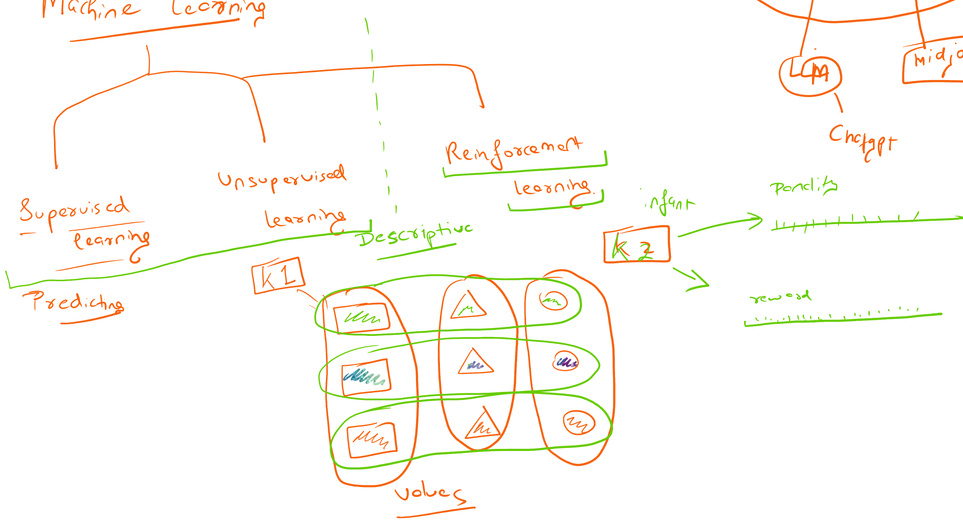
Data Science



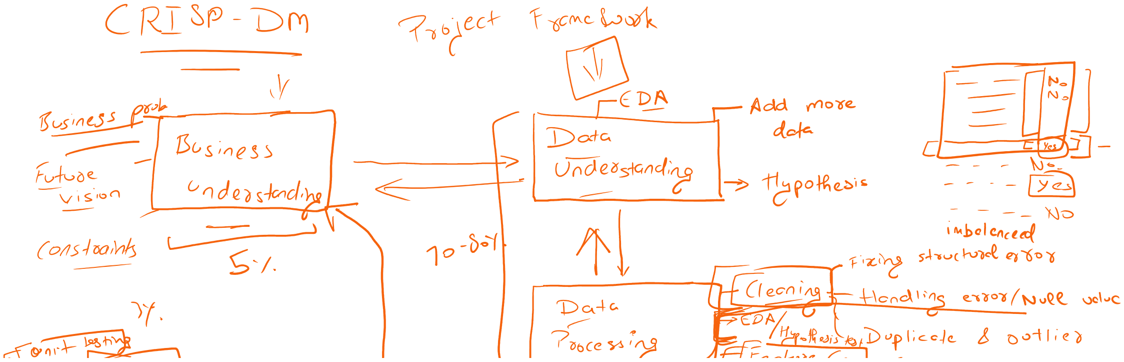
Artificial Intelligence (AI)



Machine Learning



CRISP-DM



5%.

7%.

- unit testing
- closed
- open
- prod.

deployment

16 month

2%.

10%

Data Processing
- Cleaning – Handling error/Null value
- EDA / Hypothesis – Duplicate & outlier
- Feature Eng. – Feature transformation – encoding, scaling
  - Feature selection
  - Feature Sampling
  - Demension Reduction

Modelling
- spliting data
- Selection of model
- Hyperparameter tuning

Evaluation
- Parameter
- Hypothesis testing

Rice	Rice
F1	F2

✓ F1 F2 – no change

✓ F1 > F2 – Hypothesis

35 – 5000...

5.9

Age	salary	Height
20	10k	5.5
30	20k	5.7
40	100k	6
50	300k	7.1
35	80k	5.9

Project Framework

Business Understanding
- Business prob.
- Future/vision
- constraints

Data Analysis

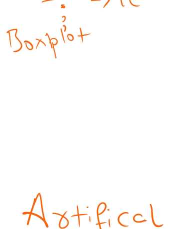
Data understanding
- Data collect.
- Explore Data Analysis (EDA)
- Add more data
- Hypothesis

Data Processing
- Data integration
- Cleaning
  - structural error
  - Handling null/error
  - duplicate/outlier

Data Analytics/Visualiz
- Univariate
- Bivariate
- Multivariate
- Visualisation

Storytelling

IQR



Boxplot

UQ => 75% 95

LQ – 25%

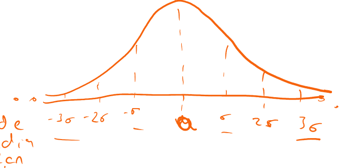
IQR = UQ – LQ

UL = UQ + 1.5 × IQR

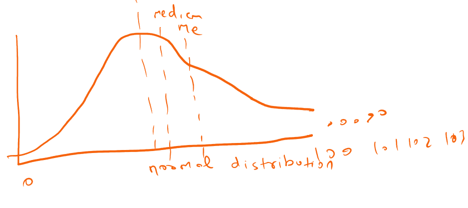
LL = LQ – 1.5 × IQR

Z-score

z-score = $\frac{x - \bar{x}}{\sigma}$



mode
median
mean



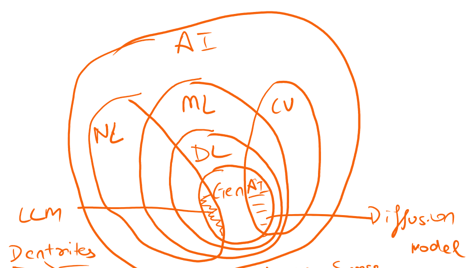
Artifical Intelligence (AI)

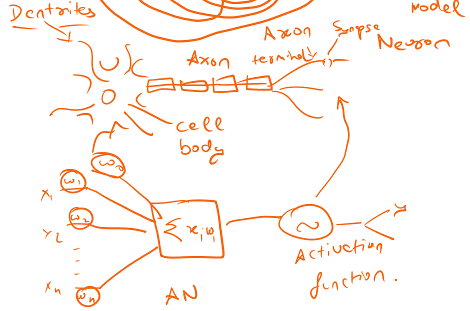
Artifical Intelligence (AI)

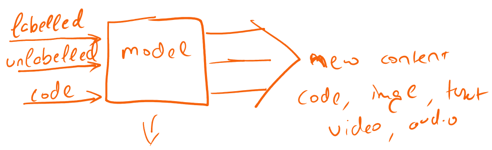
Machine learning - understanding the data

Deep learning - Processing of huge data

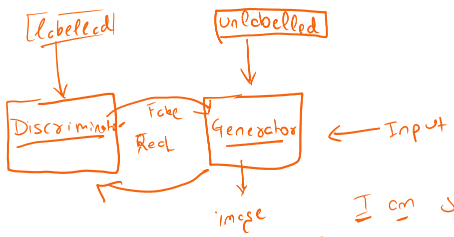
Generative AI - Generating new content







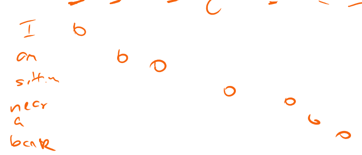
GAN



Large Language Model

Attention model

I am sitting near a bank



Large language models

transformers

Image creating model

Diffusion model

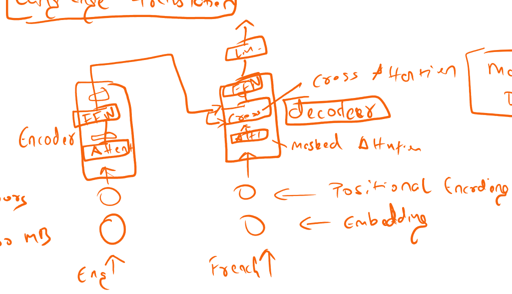
Language translation

1 layer of transform

50,00

4-5 hours

50-60 MB



Chatgpt-2 → Chatgpt 3 → Chatgpt 4

2018 → 2018 | 2020 | 2022

small 117M Parameter

12 layers

175B Parameter

96 layers

578 GB

Roadmap of AI Eng.

Management Data → Python

→ Maths: Stats, Prob, Linear Al. (Top)

Pipeline → Machine learning: Supervised, Unsupervi, Reinforcem

Deployment

→ Deep learning - Neuron

FFN, CNN, RNN

GAN, VAE

→ Attention

→ LLM

→ Diffusion

DS

CRISP-DM

Pandas
Numpy
matplotlib

null / error /
header deletion | duplicate

DS

CRISP-DM

Pandas
Numpy
matplotlib

soft skill

Business prob.
Future goal
constraints

Business understanding → Data collection, EDA, Add more data, Hypothesis → Data understanding → Data processing

Header, datatype, null/error, duplicate, outlier
Structural error
Cleaning
Hypothesis testing + DA

Feature engineering
ML Lib

Modelling
Splitting of data
Selection model
Hyperparameter tuning

Evaluation

Deployment
production | open test | close testing | unit testing

Handling null values

→ removing null values

categorical → mode
numeric → mean / median

mean - if there are not outlier
median - if there are outlier

bfill/ffill

→ interpolate

2
ffill 3
□
5 bfill
6

2
4
6 interpolate
8
10

→ Cleaning of Null values

→ removing null values

→ filling null values

Quantitative
continous | discrete

Qualitative
ordinal | categorical

→ mean / median | mode

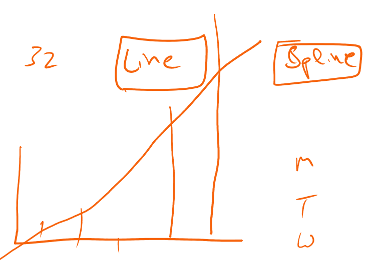
→ Constant

→ bfill / ffill

→ interpolate (Line)

→ Supervised learning model

32



C1
32
34
36
38
40

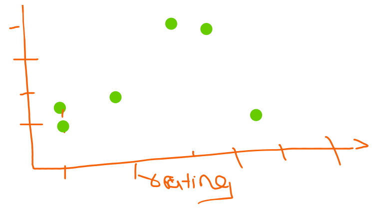
M
T
W
T
F
M
T
W
T
F
M
T

R | P
-1 | 25
2 | 30
-1 | 31
3 | 62

Price

2	30
-1	31
3	62
4	40
5	25
6	60

Price



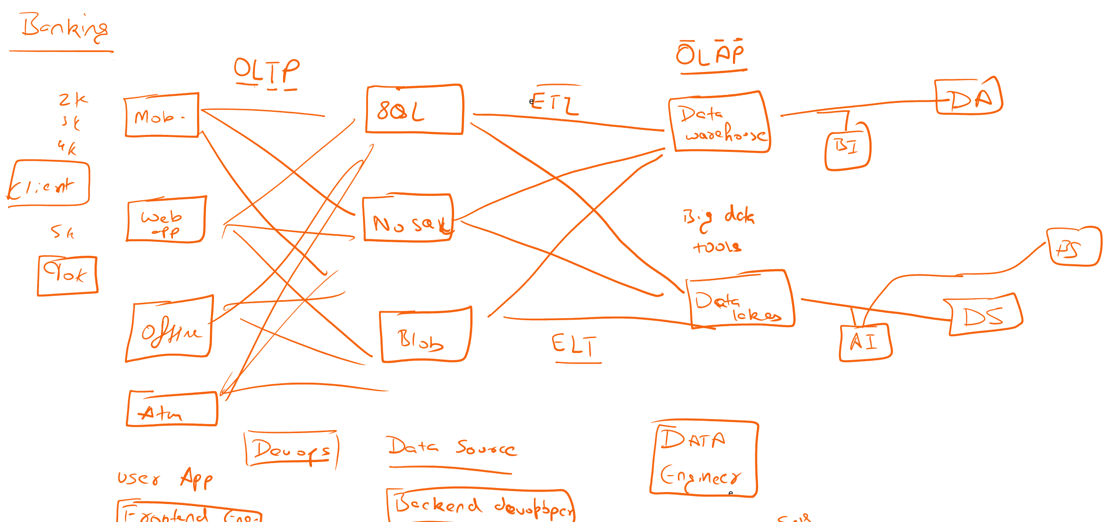
Rating

Data Science → Data Engineer.

Data Analyst

Data Scientist

Banking



Self

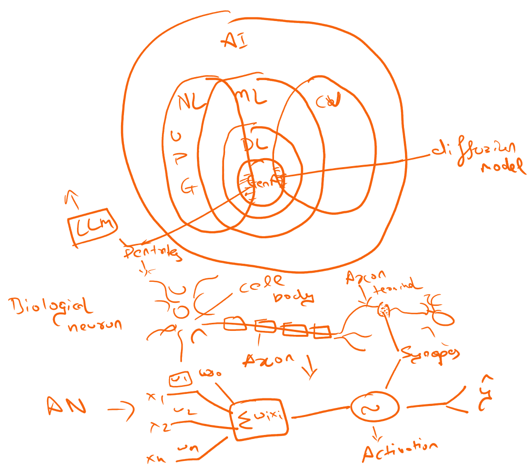
AI → 1950's → 1970 → 1990's → 2000

ML → machine learning → understanding

DL → Deep learning → Processing

Gen AI → Generative AI → Generate

NL → Natural language → language



DE

→ SQL

→ Python

→ Spark

DS

→ Python

→ Cloud

→ ML

DA

→ Python

→ SQL

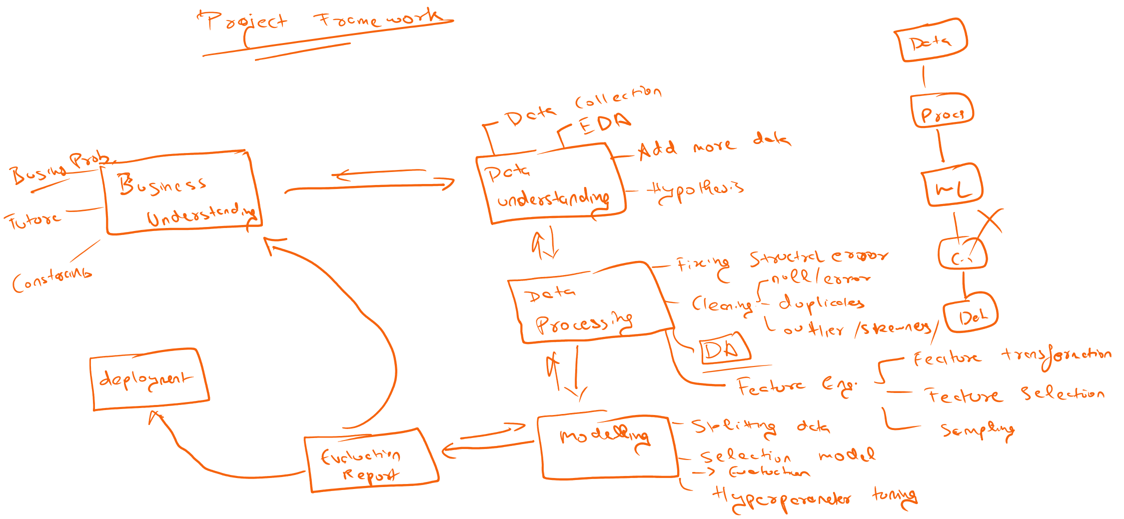
→ BI

- -> Python
- -> Spark
- -> Hadoop
- -> Azure, GCP, AWS

- -> Cloud
- -> ML
- -> DP
- -> DL
- -> GenAI
- -> Devops

- -> SQL
- -> BI
- -> Excel
- -> Storytelling

Project Framework



Today's topic

Prereq. -> • Machine learning

-> • Fully connected neural network

Deep learning -> FFN

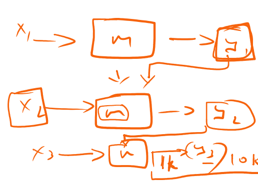
Limitation
- High dimension data = (1Lk) — Image, Audio, Text, Video
  - CNN
- Series data -> RNN

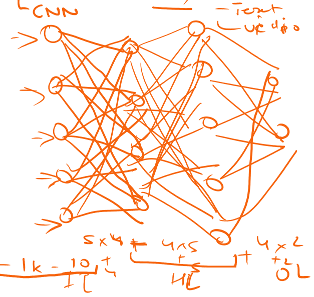
CNN - model

Convolutional Neural Network

RNN - model

Recurrent Neural Network





1k (x,y) 10k - 1k - 10

5x4 + 4x5 + 4x2

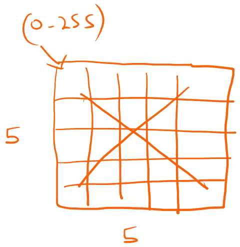
IL HL OL

2\ + 25 + 10

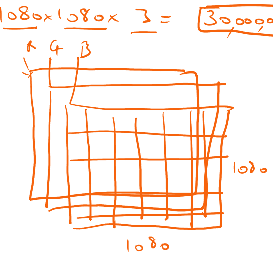
59

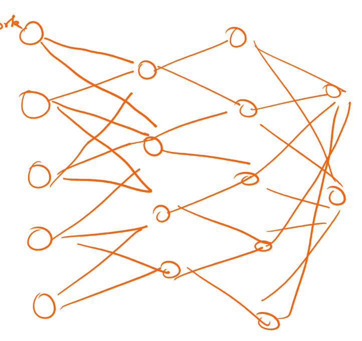
Convolutional Neural Network -> local neural network

(0-255)



1080x1080x3 = 3000000

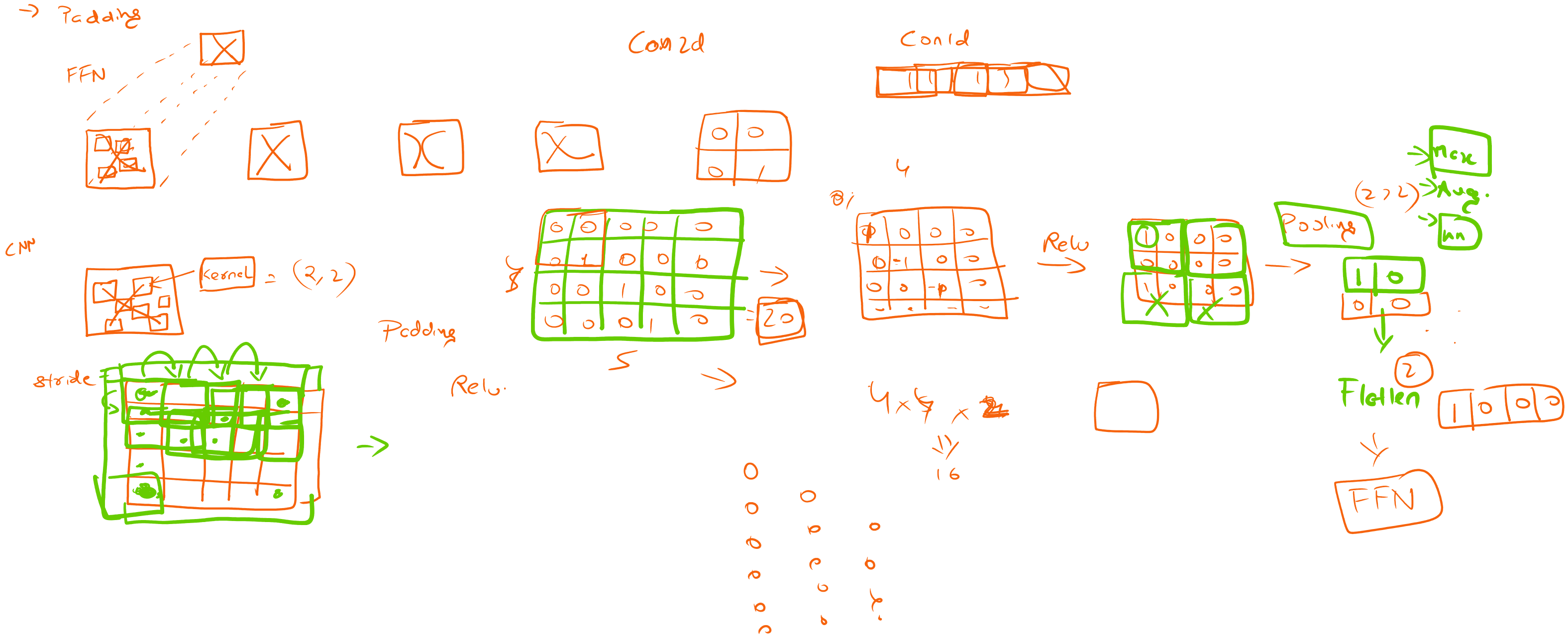




Conv -> Activation (Relu) -> Pooling

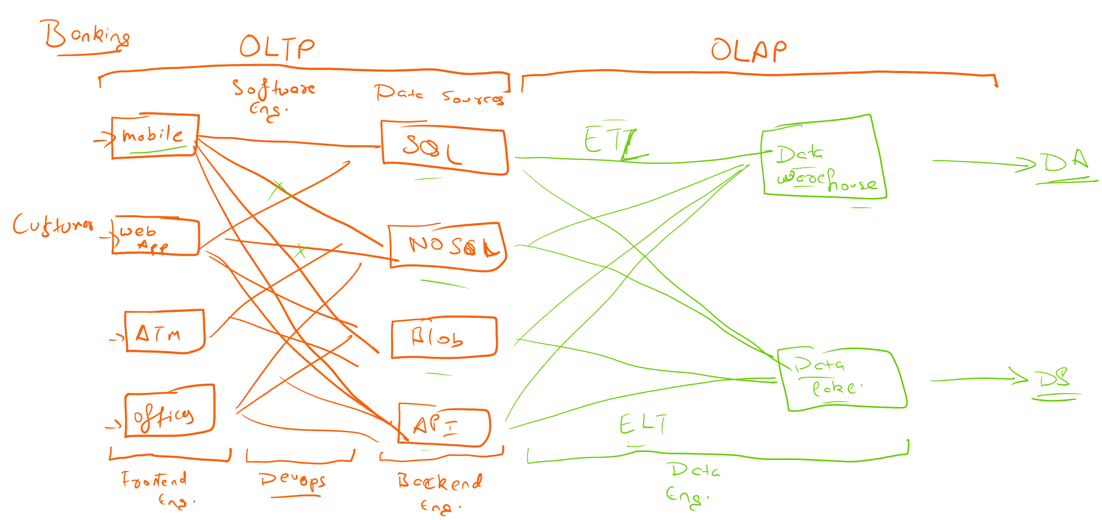
- -> Kernels
- -> strides
- -> Padding

-> strides
-> Padding



Data Science

- Data Analyst
  - visualization & BI Tools
- Data Engineer
  - Manageing data
- Data Scientist
  - AI



Data Science

AI -> Artificial Intelligence

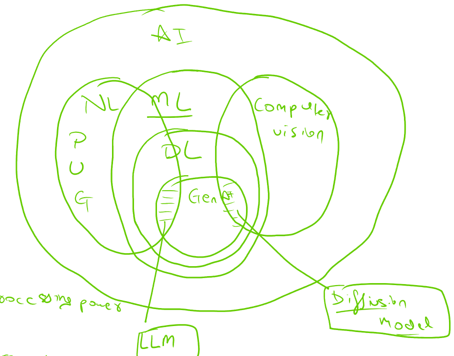
ML -> Machine learning -> understand

DL -> Deep learning -> Advance processing power

GenAI -> Generative AI -> Generate Content

NL -> Natural language

CV -> Computer vision



Large language model
GPT, Bert, LenAI

Generative AI

-> Attension model

transformer

Diffusion model

Prompt Eng.

labelled -> unlabelled -> code -> Model -> new content
Text, image, code

Project Framework -> CRISP_DM

Business Prob.
Future
contraints

Business understanding -> Data understanding

Data collection
EDA
— Add more data
— hypothesis

Data processing
— Structure error — Header — datatypes
-> Cleaning — null/error/duplicates/ outlier
-> EDA / visualization
-> Feature Engineering — Feature transformation -> Feature Selection — Feature balancing — Feature generation

Modelling
— Splitting of data
— selection model
— hyperparameter

Modelling -> Evaluation -> deployment
Evaluation -> Business understanding

Feature transformation -> Scaling

age | Salary | Height(ft)
0-100 | 100-1cr | 0-8
50 | 50,00,000 | 4

Scaling — Single
(0-1)

↓

Encoding

label Encoding

a	0	0
b	1	2
c	2	1
a	0	
c	2	
b	1	
a	0	
a	0	

one-hot Encoding

	M	F	O
M	1	0	0
F	0	1	0
O	0	0	1
M	1	0	0
F	0	1	0
F	0	1	0
M	1	0	0
O	0	0	1

String index

8+

LE | OE
M | 0
F | 1
M | 0
F | 1
4 | 0
O | 2

Scaling

-> Standardization = $z = \frac{x - \bar{x}}{}$

age | Salary | Height
50 | 5,00,000 | 4

→ ∘ Standardization = $z = \frac{x - \bar{x}}{\sigma}$

age	Salary	Right
50	5,00,000	4
0	0	0

→ ∘ Normalisation = $z = \frac{x_{max} - x}{x_{max} - x_{min}}$

[0 − 1]

7, 5, 2, 1, 3

55 − 1, − 2, = 1, = 2, = 0

N: $\frac{7-7}{7-1} = 0$, $\frac{2}{6}$, $\frac{5}{6}$, = 1, $\frac{4}{6}$

→ Feature balancing

majority → $\boxed{<50k}$ = $\boxed{1000}$ 10 > 500 = (000 (000

minority → $\boxed{>50k}$ $\boxed{100}$ ✗ 500 < 500 ≠ 1000

→ $\boxed{100000}$ 500 → 700

$\boxed{100}$

Undersampling

Random undersampler

ENN

Oversampling

Random oversampler

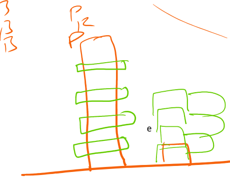
SMOTE

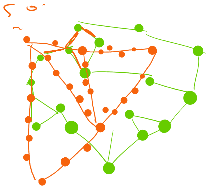
42, 44, 54, 58

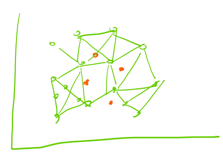
P, 13, 13, 13

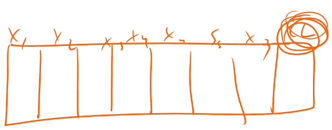
Privat, P12, P

> 50k, > x2, > 50









I.G (x_1) = $\boxed{\cdot 4}$

IG (x_2) = ·2

= $\boxed{\cdot 6}$

= $\boxed{\cdot 01}$ ✗

= ·45

= $\boxed{\cdot 1}$

Root Feature Selection

EnId	Name	Date	Age	Workclass	Married	Income

Pred

→ Correlation

→ Mutual Info gain → Decision tree ML

→ Chisq'

→ t test

13+

Feature generation

→ ∘ Calculated colums

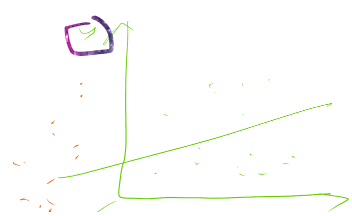
Mineral	Vitam	Carb	Fat	Prot	Amino Acid

nutrit; Energy gain; food; Body build

— unsupervised learning

→ PCA

LDA



Artificial Intelligence

Machine Learning :

- Supervised learning (predictive model)
  - 187 – classification
  - Regression
- unsupervised learning
  - Clustering
  - Association
  - dominsion reduction

Fond.
- → Naive Bayes – Prob.
- → Decision tree – Stats
- → Logistic Regression – log.
- → KNN – Coordinate.
- → SVM – "

⇒ Ensemble Technique
- Random forest
- Xgboost
- Ada boost

Type of Day	wind	Temp	Humi	y/N
S	H	H	L	y
R	L	L	H	y
C	L	L	H	N
S	H	L	L	N

Independent vari. | dependent variable

features / columns attribute | target label class

Naive Bayes – Prob. → Bayes theorm

$$\frac{\text{Occurence of Event}}{\text{Total Event}}$$

HH – 1, HT, TH = 2/4, TT = 1/4

P(y) = 4/6   P(N) = 2/6

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

Day	Temp	Hum	wind	y/N
S	H	H	H	y
S	H	L	L	N
S	N	L	H	y
R	N	H	L	y
R	N	H	H	y
R	H	L	L	N

S N H L = ? y/N

Day	y	N
S	2/4	1/2
R	2/4	1/2

Temp	y	N
H	1/4	2/2
N	3/4	0

Humid	y	N
H	3/4	0
L	1/4	2/2

wind	y	N
H	3/4	0
L	1/4	1

P(y | S N H L) = P(S|y) · P(N|y) · P(H|y) · P(L|y) · P(y)

~~P(S) · P(N) · P(H) · P(L)~~

P(N | S N H L) = (2/4) · (3/4) · (3/4) · (2/4) · (4/6)

→ 9/64 = y

P(N | S N H L) = 0

→ Laplace Smoothening : 1

Zero case senerio.

$$P(y)\ P(y \mid \underbrace{S N H W}_{B}) = \frac{P(B|A) \cdot P(A)}{\cancel{P(B)}} = 0.005$$

$$P(N)\ P(N \mid B) = \frac{P(B|N) \cdot P(N)}{\cancel{P(B)}} = 0.0016$$

Normalisation : $\frac{P(y)}{P(y) + P(N)}$ $\frac{(N)}{P(y) + P(N)}$

Normalisation : $\frac{P(y)}{P(y)+P(N)}$ $\frac{(N)}{P(y)+P(N)}$

$\frac{.005}{.0066} = \boxed{.755}$ $\frac{.0016}{.0066} = \boxed{.242}$

- → Maths
- → Assumptions
- → Limitation
- → use case
- → Types
- → when to Apply
  - → whenever you limited data
  - → fast & Simple
  - → Binary data, Multiclass classification
  - →

Naive

Assumptions:

- All feature should independent in nature
- All feature should contribute equally in pred.

Limitation

- Assumption
- Zero case senerio
- Inability to handle missing values

Use case

- Text Data
- Medical Daignose
- Recommedation

Types of Naive Bayes :-

- Gaussian Naive Bayes: – Normally distributed data & most of feature are continous in nature
- Multinomial Naive Bayes :- when most features are discrete / freq. based.
- Bernoulli NB – Binary data
- Complement NB – Imbalanced data
- Categorical NB – categorical data
- out of core NB Big data

KNN – K-Nearest Neighbours

Step-1 – Plot all data point

Step-2 $\boxed{K = 3}$ – $\boxed{K = 5}$ $\boxed{K = 1}$ $= k = \boxed{4} = 3 = 5$

Step-3 calculate dist

- Euclidean distance $\sqrt{(x_1-x_2)^2+(y_1-y_2)^2+(z_1-z_2)^2}$

$A = \sqrt{(x_1-x_2)^2+(y_1-y_2)^2}$ $= (a_1 - a)^2$ $\sqrt{\sum (x_i - y_i)^2}$

$A = \sqrt{(1-3)^2 + (2-2)^2} = 2$

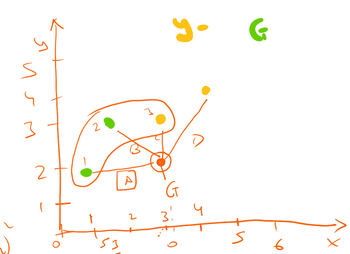
$B = \sqrt{(2-3)^2 + (3-2)^2} = 1.414$

$C = \sqrt{\ }$

D =

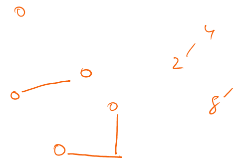
= 2.83

G

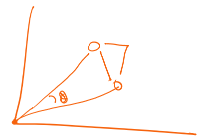


a	b	target
1	2	0
2	3	0
3	3	2
5	4	1
3	2	? = 0 = 1

Distance Metrics:



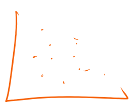
- Euclidean dist : Continous data
- Manhattan dist : $\sum |x_i - y_i|$ Categorical data, outliers
- Minkowski dist : $\sum (|x_i - y_i|^p)^{1/p}$ $p = 2 =$
- Cosine similarity :



$\boxed{\text{lazy learner}}$

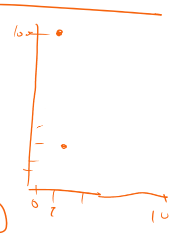
Assumptions:

- KNN assumes that similar data point are close to each.
- Normalization



df.shape

$\boxed{(x \; \boxed{y})}$



Limitation:

- Computationally complex
- $\boxed{\text{Curse of Dimensionality}}$
- Choice of k value

100 → 800 → complexity higher, computationally expensive

100 → 60 → $\boxed{\text{lossing on information}}$

$\boxed{\text{Use Cases}}$

When to Apply:

- non-linear
- Small - medium data
- donot need fast model

- Recommendation
- Anomaly detection
- Customer segmentation

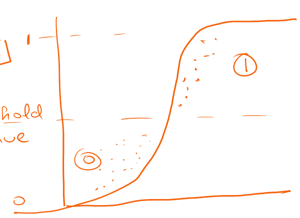
Logistic Regression:

Binary Classification

Sigmoid - $\sigma(z) = \dfrac{1}{1+e^{-z}} = [0,1]$ $\left(z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \ldots\right.$

$\boxed{\text{Gradient Descent}}$

$\boxed{.05 <} \; \sigma(z) = \boxed{1}$, $.5 > \sigma = \boxed{0}$



I	CB	P
50k	650	$\boxed{0}$ =
60k	700	1
70k	800	1
80k	720	1
$\boxed{65}$k	$\boxed{675}$	= ? = 1

.8	1.1	.5
β_0 ,	β_1 ,	β_2
-2	+0.03	+0.01

$z = -2 + \boxed{.03} \times \boxed{65} + \boxed{0.00} \times \boxed{675} = \boxed{.375}$

$\sigma = \dfrac{1}{1+e^{\boxed{-z}}} \sim \boxed{.592} = \boxed{1}$

Assumptions:

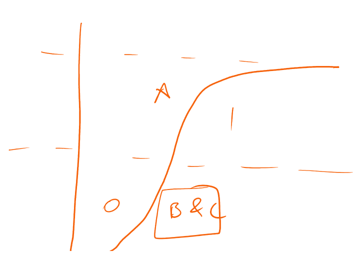
- Linear data
- Independence
- Large datasets

Limitations

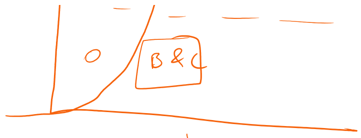
- Linearity
- Sensitive to outliers
- Scaling, selection

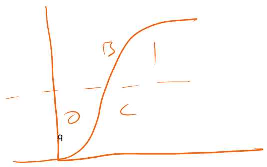
When to Apply:

- Binary classification
- Linearity data
- large dataset Available

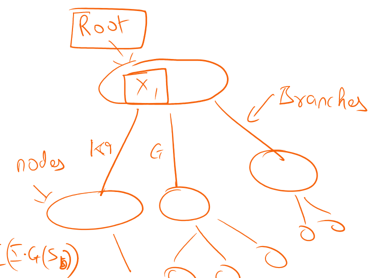


A
B
C
A
B





A
B
C
C
D
D



Decision Tree model :-

→ Entropy : $H(S) = -p_1\left(\log_2(p_1)\right) - p_2\log_2(p_2)$

Information Gain = $H(S) - \sum \frac{S_v}{S}\left(H(S_v)\right)$ $E(A) = \sum\left(I.G(S_v)\right)$

y = 2
N = 2

	Y	N
S	$\boxed{1}$	1
R	1	1

	y	N	T	W
H	1	1		
L	1	1		

ToD	T	H	W	Y/N
S	H	L	H	Y
S	L	H	H	N
R	H	H	L	Y
R	L	H	H	N

S L L H = ?

$T(S) = \left[-\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right] = \boxed{1} \times \frac{2}{4} = .5$

$T(R) = \left[-\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right) = \boxed{1} \times \frac{2}{4} = .5$

1

y = 5
N = 5

$I.G = -\frac{y}{y+N}\log_2\left(\frac{y}{y+N}\right) - \frac{N}{y+N}\log_2\left(\frac{N}{y+N}\right)$

$= -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\left(\frac{1}{2}\right) = \boxed{1}$ $1 - \left(\frac{1}{3}\right)^2 - \left(\frac{2}{3}\right)^2$

	y	N
S	1	2
C	3	0
R	1	3

$\boxed{E(S)} = -\frac{1}{3}\log_2\left(\frac{1}{3}\right) - \frac{2}{3}\log_2\left(\frac{2}{3}\right) = a$

⇒ $I.G = (a) \times \boxed{\frac{3}{10}} = .2757$ $1 - \left(\frac{3}{3}\right)^2 - \left(\frac{0}{3}\right)^2$

$I.G(C) = \boxed{\left[-\frac{3}{3}\log_2\left(\frac{3}{3}\right) - \frac{0}{3}\log_2\left(\frac{0}{3}\right)\right]} \boxed{\times \frac{3}{10}} =$

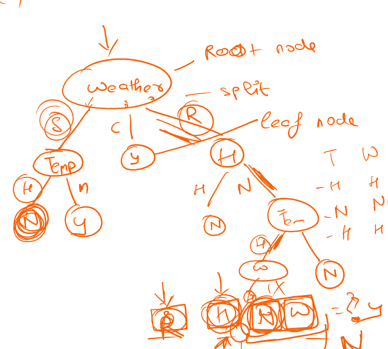
$I.G(R) = \left[-\frac{1}{4}\log_2\left(\frac{1}{4}\right) - \frac{3}{4}\log\left(\frac{3}{4}\right)\right] \times \frac{4}{10} = 0.0754$

$E(T) = .117$

ToD = $\boxed{.89}$
Temp = .51
Hum = .37
wind = .21

IG $1 - .117 = \boxed{0.89} = $ ToD

Binary , Multi



Temp	Hum	Win	Play
H	H	W	N
M	N	S	y
H	H	S	N

	y	N
Temp		
H	0	2
M	1	0

H	W	N
H	S	N

Day	Weather	Temperature	Humidity	Wind	Play?
1	Sunny	Hot	High	Weak	No
2	Cloudy	Hot	High	Weak	Yes
3	Sunny	Mild	Normal	Strong	Yes
4	Cloudy	Mild	High	Strong	Yes

Temp
Maybe

2	Cloudy	Hot	High	Weak	Yes
3	Sunny	Mild	Normal	Strong	Yes
4	Cloudy	Mild	High	Strong	Yes
5	Rainy	Mild	High	Strong	No
6	Rainy	Cool	Normal	Strong	No
7	Rainy	Mild	High	Weak	Yes
8	Sunny	Hot	High	Strong	No
9	Cloudy	Hot	Normal	Weak	Yes
10	Rainy	Mild	High	Strong	No

Maybe

Temp		
H	0	2
M	1	0

| H | S | N |

$I(H) = \left[-\frac{2}{2}\log\left(\frac{2}{2}\right) - \frac{0}{2}\log\left(\frac{0}{2}\right)\right] \times \frac{2}{3} = 0$

$I(M) = \left[-\frac{1}{1}\log\left(\frac{1}{1}\right) - \frac{0}{1}\log\left(\frac{0}{1}\right)\right] \times 1 = 0$

$I.G = 1 - \left[-\frac{1}{3}\log\left(\frac{1}{3}\right) - \frac{2}{3}\log\frac{2}{3}\right]$

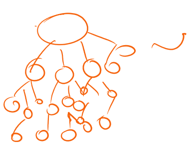
Gini - $1 - \sum P_i^2$ = Multiclass classification

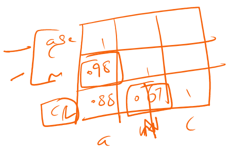
Entropy — Binary class

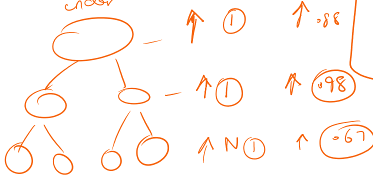
Assumptions
- Independent feature
- Hierarchical model - its assume data can be splited in hierarchy.

→ Limitation:
- Overfitting:

overfit



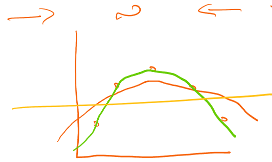




Overfit | good fit | underfit

model [complex

data [baises, variance



Limitation
- Overfitting
- Biases
- Instability

Use Cases
- Marketing campaign
- Credit Scoring
- Customers segmentation

When to Apply:
- Missing values, outlier
- interpretable model
- mix continuous & categorical

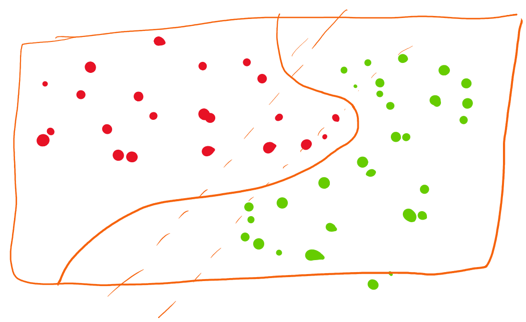
SVM - Support Vector Machine

$y = mx + c$ $\qquad \min \frac{1}{2}||w||^2$

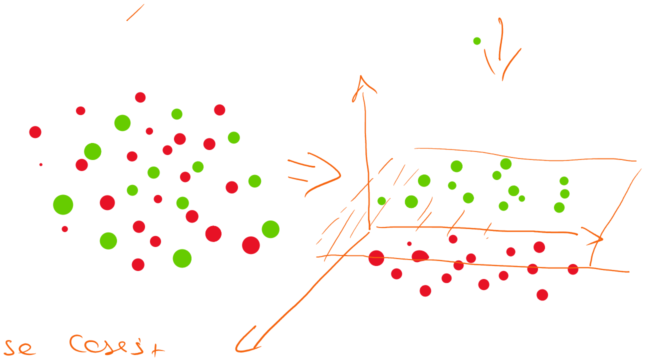
hyperplane = $n-1$

cne

ctor



- Linear Kernel
- Polynomial kernel
- Radial Kernel



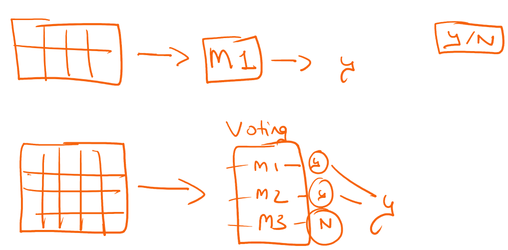
- Assumptions :-
  - Data linearly seperable
  - Classes are well seperated
  - Sensitive to outliers.
- Limitation
  - High computation & memory usage
- Use cases:
  - Text Classification
  - Image Classification
- When to Apply :
  - High dimension
  - Less instance
  - noise free data.

Ensemble Models

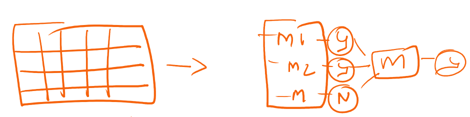
- Voting
- Stacking
- Bagging
- Boosting

Boosting - Xgboost

- Voting -

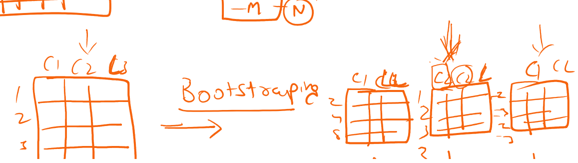


- Stacking

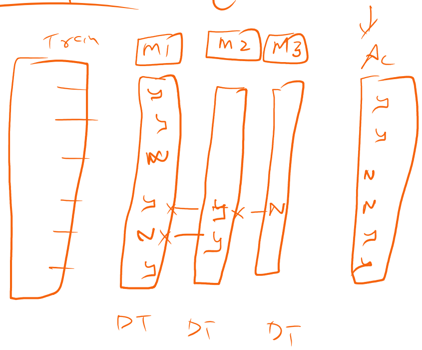


- Bagging

Random forest



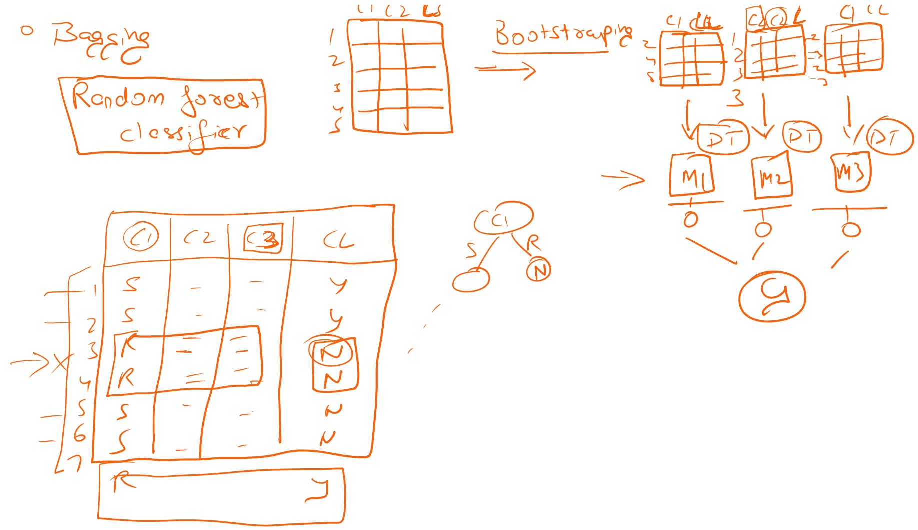
Boosting — Xgboost



Bagging

$\boxed{\text{Random forest classifier}}$

Bootstrapping

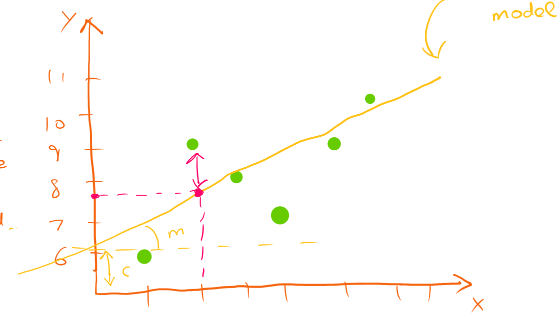


Regression Models

Linear Regression

Simple Linear regression — $\hat{y} = mx + c$

Significance of m & c



Multi-Linear Regression model — $\hat{y} = \overset{.9}{m_1}x_1 + \overset{2}{m_2}x_2 + \overset{.05}{m_3}x_3 \; ---- \; m_nx_n + m_0/b/c$

m = importance of feature

m_0 = helper value to find best fit line

P1 — 65 ? $\boxed{7.8}$

P	Area	Loci	Ame	Price
80 la	1000	M	P	1cr
2.2cr	1000	A	m	2cr
80	1000	M	H	Sold

→ $\boxed{20 \text{ la}}$

→ + 200

→ 30

$y = mx + c$

Error — $(y_{act} - \hat{y})$

Loss function

L1 error — $\boxed{\text{individual}}$

M $\boxed{A}$ E $= \frac{1}{n} \sum |y_{act} - \hat{y}|$

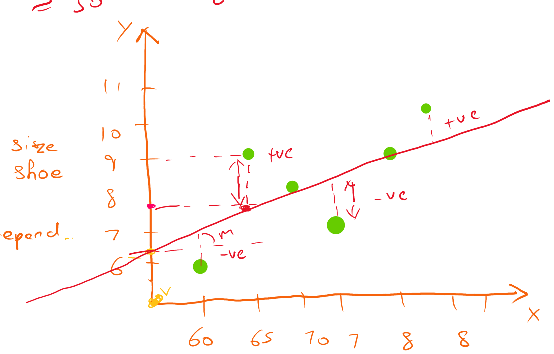
M $\boxed{S}$ E

L2 error =

Cost function

$\boxed{\text{Entire data}}$

(Gradient descent)



$$J(m,b) = \frac{1}{2n} \sum_{1}^{n} (y_i - \hat{y}_i)^2$$

$$= \frac{1}{2n} \sum_{1}^{n} \left(y_i - (mx_i + b)\right)^2$$

P − 65 = 7.8

m b c