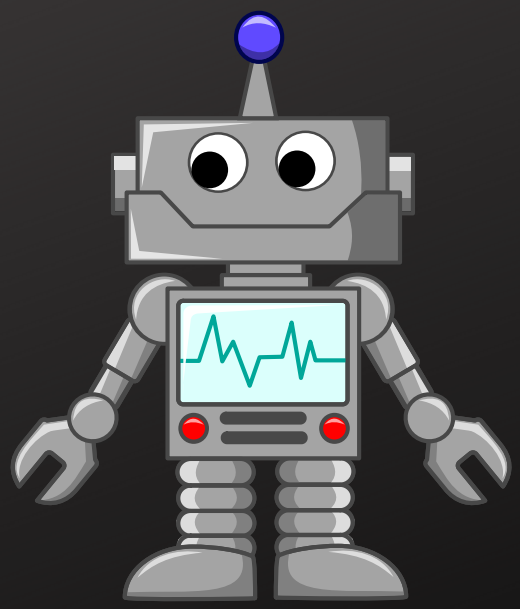
MACHINE LEARNING ENGINEERING

ML : The ability to Learn,Predict & Find Ways to Improve the Performance

Types of Machine Learning

Supervised

Unsupervised



Reinforecement

@Biranchi

MACHINE LEARNING WORK FLOW

USERS

Data

Database

>10k

No

Data Gathering

Yes

Load Data to System

Do EDA

Explore & visualize the data properly

Clean the Data
Establish a Benchmark & Reproducebility

Feature Engineering

Grouping, Class deriving, Feature splitting (dependent & independent)

USERS

Imputation(missing data), Handling outliers, Balancing the dataset, Binning, Encoding, transforming(feature scalling), Feature selection

Human Interaction & Supervision

Training set | Validation | Test

Model Buiilding

Performance Mapping, Robustness, Fairness, Constraints applying to generalise the model

Hyperparameter -Tuning (croos-validation and other techniques)

Overfitting or underfitting

Model Score

Model Validation & Diagnostics

Review & Documentation

Good

Bias testing & Modification

Deployment

AWS,AZURE, GCP,HEROKU

@Biranchi

BEST PRACTICES TO MAKE A MODEL

STAGE 1- ALL ABOUT DATA

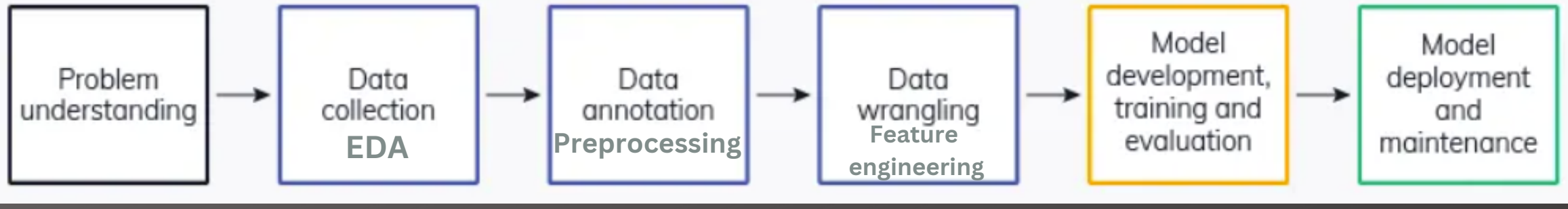
Process	Challenges to Overcome	Best Practice
Data collection (web scraping, from Database, User input & more)	• Uncleaned Data(need to clean) • Finding relevant data • Curse of Dimensionality • Baised Data • Incomplete data, Sparsity • Data Quality issue & many more	• Enrich the Data • Using data-related Key Performance Indicators (KPIs) to understand the Data • Dimension-reduction techniques • Data Modification • Data representation
Untidy Data	• Matching proper Rows and column • Multiple variables in same column	Restructure the data to be fit & tidy by using cast techniques
Missing Data	• Information Loss • Bias	• Central tendency Imputation • Tree-Based Modeling techniques • Apply best missing technique accoring to dataset(more than 20+ ways there)
Outliers	• Undue influence on Squared loss function • Unknown categorical data	• Winsorizing • Capping • Robust techniques and 10+ effective techniques availabe
EDA	• Unstable parameter estimation • Non-determinism on colinearlty	• Use Profiling process • Use statical approcah and graphical representation

BEST PRACTICES TO MAKE A MODEL

STAGE 2- TRAING & TESTING (MODEL BUILDING)

Process	Challenges to Overcome	Best Practice
Overfitting & Underfitting	• High variance and Low Bias (Not a Generalize model)	• Regularization • Noise reduction • Cross validation • Removing features • Ensembling & many more techniques
Hyperparameter Tuning	• Combinatorial explosion of hyperparameters in convetional algorithms	• Parameter estimation • GridsearchCv • Randomsearchcv & many more
Ensemble models	• Single models that fail to provide adequate • High-variance and low-bias models that fail to generalize well	• Powerful ensemble model like (bagging & boosting) • Custom combinations of predictions
Model Interpretation & Validation	• Imperfections in the Algorithm When Data Grows • Nonrepresentative training data.	• Variable selection by regularization (e.g. L1) • Surrogate Models • Partial dependency plots & variable dependency measures
Model deployment & Decay	• Trained modellogic must be transferred from a development environment to an operational computing system to assit decission making process	• Database, web scoring

MACHINE LEARNING LIFE CYCLE



70% 30%

On an Average a Data scientist spends Time