Challenges:

(1) Data Collection .:

Actual ML

Data accumulation -> Fetch data from API.

Data Gathering -> Web Scraping .

(2) Insufficient data /Labelled Data

A

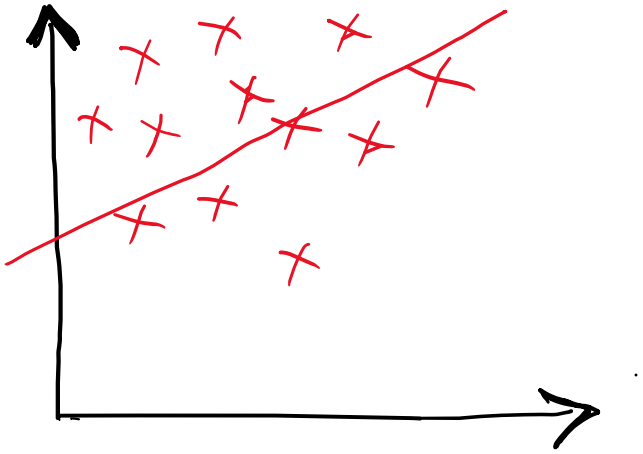
Uses algorithm - A &
having 100 rows of data.

B

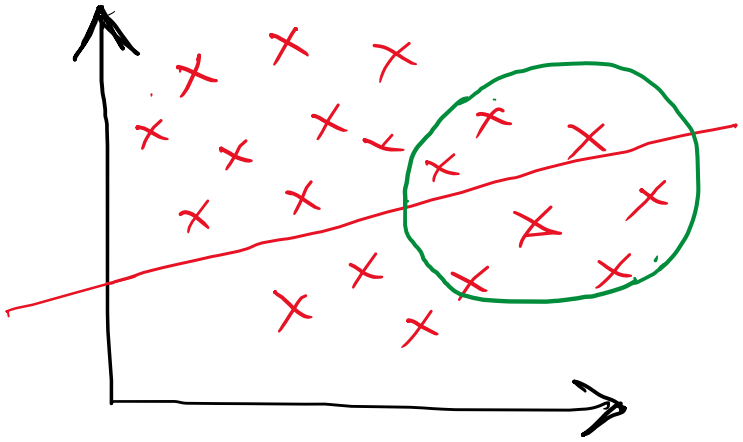
Uses algorithm - B & having
1000000 rows of data.

' B' will perform better. Although dataset 'A' is having better algorithm than algorithm B but dataset 'B' is having vast data.

(3) Non- representative data:



Model where,
Seme data is not included



Model where,
Total data is there.

(Full survey, Full data for representation i.e. valid data).

So if we are taking full data So, that will be non-representative data which won't give good result. i.e., due to poor representation of data.

So these kind of things are called "Sampling noise."

Sometime we gather full data but still not good result we get.

This kind of things are called Sampling bias.

Example: Taking survey of people just is India among all the countries that, who will win hockey match will obviously answer India. (Sampling noise).

Taking survey from people all over the countries in the World & they answered India only [Because in other countries too there are lot of Indians]. So this is called Sampling Bias.

(4) Poor Quality Data:

- Lot of outliers.
- Lot of missing values.
- Lot of abrupt values (uncertain /unexpected values).
- Different format values in Data.

(5) Irrelevant features:

If the Garbage values in data are there i.e., inappropriate values (non helpful /non-related / non-useful values) the model will give garbage output only.

Ex:

Conducting marathon competition by circulating the form to peoples by taking their data. And from that data predict who will be coming to the competition & who will not.

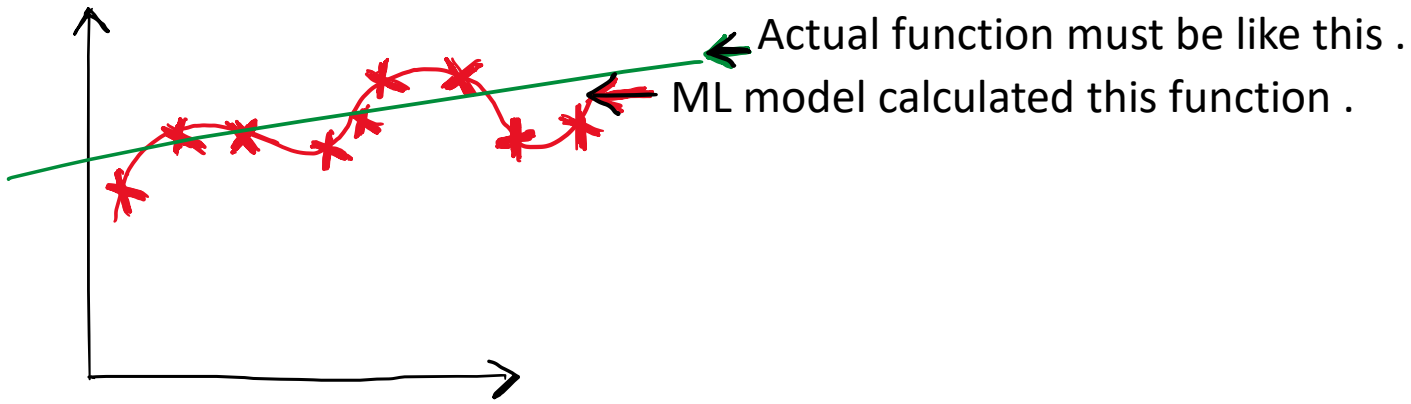
Overfitting:

ML model is trained on a dataset. This model had just learned the data and not completely understood the concept. That's why it wont give accuracy for new data.

Ex: If I went to Pune and go for Snack Center the vendor/shopkepper rated high amount for it. So by seeing this I considered that everything is costly in Pune. But that's not fair to decide & consider whole Pune as costlier just because of that one shop.

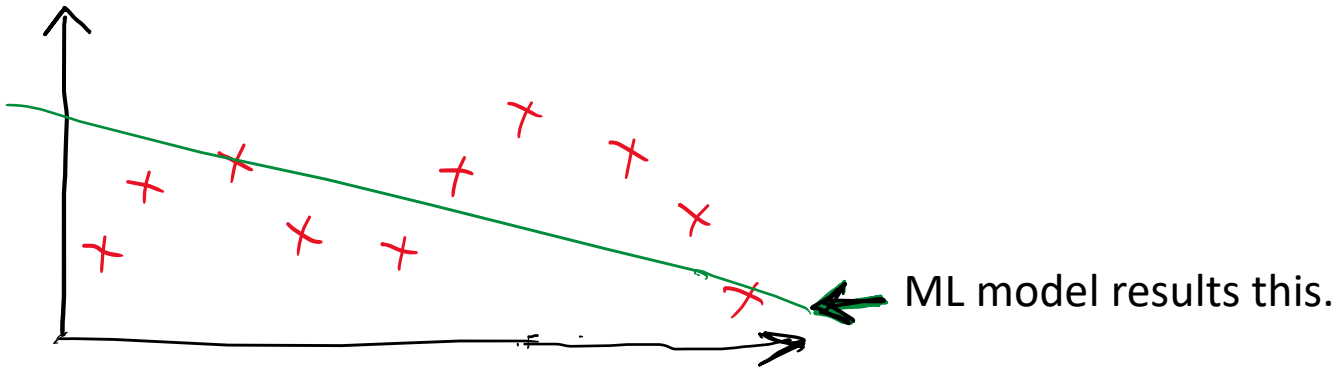
Without taking Data around we can't predict or consider that everything is like the one thing we considered/observed.

Same thing happens with the model, without testing & understanding concept of data it can't give accuracy.



Underfitting:

- Model don’t think much and just makes a small opinion.
- Opposite of Overfitting.



- Don't give good result on training nor new data.
(100%) accuracy.

Software Integration:

Integrating our software .

Offline Learning / Deployment:

Update of model constantly on server

Cost Involved:

Many multiple users uses paid softwares.

APPLICATION of ML:

B2B Application:

(i) Retail: Amazon /Big Bazaar.

(ii) Banking & Finance.

(iii) OLA

(iv) Manufacturing – Tesla.

(v) Consumer Internet – Twitter.

(vi) IMDB Movie Reviews sentiment.

Search movies

Read reviews (+ve or -ve)

ML Development Life Cycle:

SDLC (Software Development Life Cycle)

MLDLC (Machine learning Development life Cycle)

Step 1: Framing a problem. – Understanding a problem .

Step 2 : Gathering Data: csv, API, Web scraping, ETL, etc.

Step 3: Data preprocessing : removing impurities, Remove duplicates, NULL values, outliers , Scale

Step 4: Exploratory Data Analysis (EDA):

- Study relationship between Input & Output.
- Visualization
- Univariate Analysis
- Bivariate Analysis.
- Multivariate Analysis.
- Outliers detection.
- Inbalanced Dataset & Balanced Dataset.

Step 5: Feature Engineering & Selection

Feature -> means Input.

Creating new feature

- Feature selection

Step 6: Model training , Evaluation & Selection

- Try different algorithms & finally decide which to use.
- Matrix. :

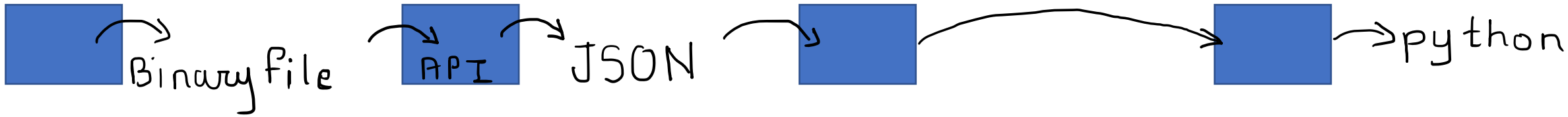
Classification – Accuracy Score.

Regression – MSE, RMSE, MAE

Clustering – Dunn Index

- Tunning the parameters for the algorithm we chosen (hyperparameter tunning)
- Ensemble Learning: Ensemble different algorithm / methods & thus create new algorithm for getting good accuracy i.e., to improve performance.

Step 7: Model Deployment



Step 8: Testing

- Beta Testing
- A/B Testing

Step 9: Optimize