Data Cleaning

① Why Data cleaning is important

a) See Instances are proper

b) Labels: O/p variables being predicted (If Someone labled it wrong in training it affects o/p)

c) Algo: The algo you use affects and gives results based on available data

d) Features: Proper feature data is inp for good prediction / classify

e) Model: Model itself assumes this data is for real world (not sure)

Garbage-in → garbage-out

Main Problems

→ Lack of Data
→ Too much data
→ Bad data

② How Can Data be Messy?

a) Duplicate data → adds weightn

b) Inconsistent text & types

c) Missing data

d) Outliers → few values dominate over other values by large amount

e) Getting data from Multiple Sources like DB, cloud etc and when talking together abt them might create mismatch

↳ Its Suggestable to have a look to the data & filter duplicates

Remove → You Can clean it quick but sometimes you might lose good info and make it biased

Impute → Replace value with mean (or) median

Mask → Create a category for missing data

To find the use plots (Histogram, Density plot, Box plot

Statistic- (~~Box plots~~ Inter Quartile Range, Standard deviation

Residuals (standardized, scaled standardized

Scanned with CamScanner

Data cleaning is very important part before training & analyzing your data. With a wrong data, you get wrong insights and can't fulfill your task although your model is pretty good & robust. Garbage in → Garbage out

~~How to~~ what to check for clean lines

1. Validity
2. Accuracy
3. Completeness
4. Consistency
5. Uniformity

How to check

1. Inspect data
   - ↳ Using Data Profiling & visualization find if data is proper (or) not: Outliers are there (or) not Data is skewed (or) not. Data is missing (or) not
2. clean data
   - ↳ Next Important part is cleaning. If for example data is missing. you can either del the row (or) fill the value using Statistical methods. In general If data is not skewed "mean" is best to fill. ~~For~~ For skewed data "median" is preferred.

   Remove outliers. Sometimes outliers also might contain good information. Need to analyze & take decision

Scanned with CamScanner

③ Verify

↳ Verify If data is cleaned or not after ya clean it

④ Report

↳ find the reason for getting bad data. If ya Can find the Source of Problem. You Can prevent next observations to be noisy. you Can even Condenstand better about t Data.

Scanned with CamScanner