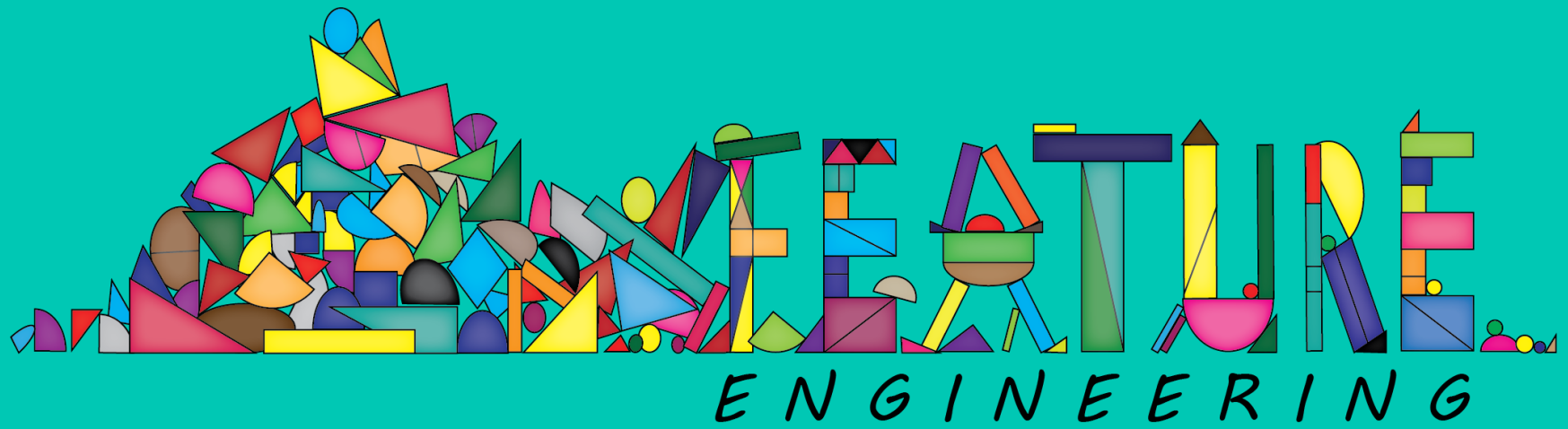
Problems in variables

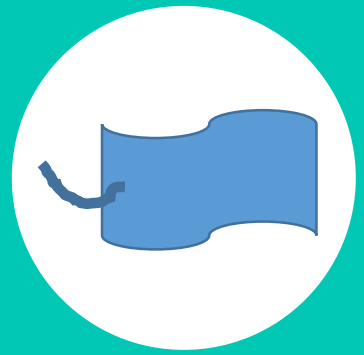


Problems with variables

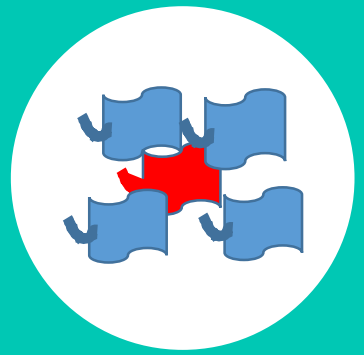
- Problems found in data
- Impact on machine learning models

Problems in variables

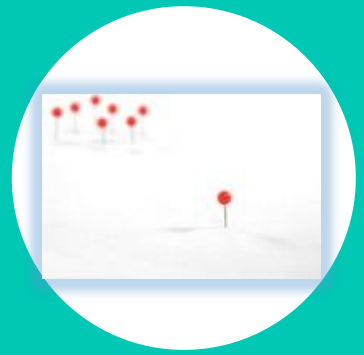
Missing data
Missing values within a variable



Labels
Cardinality



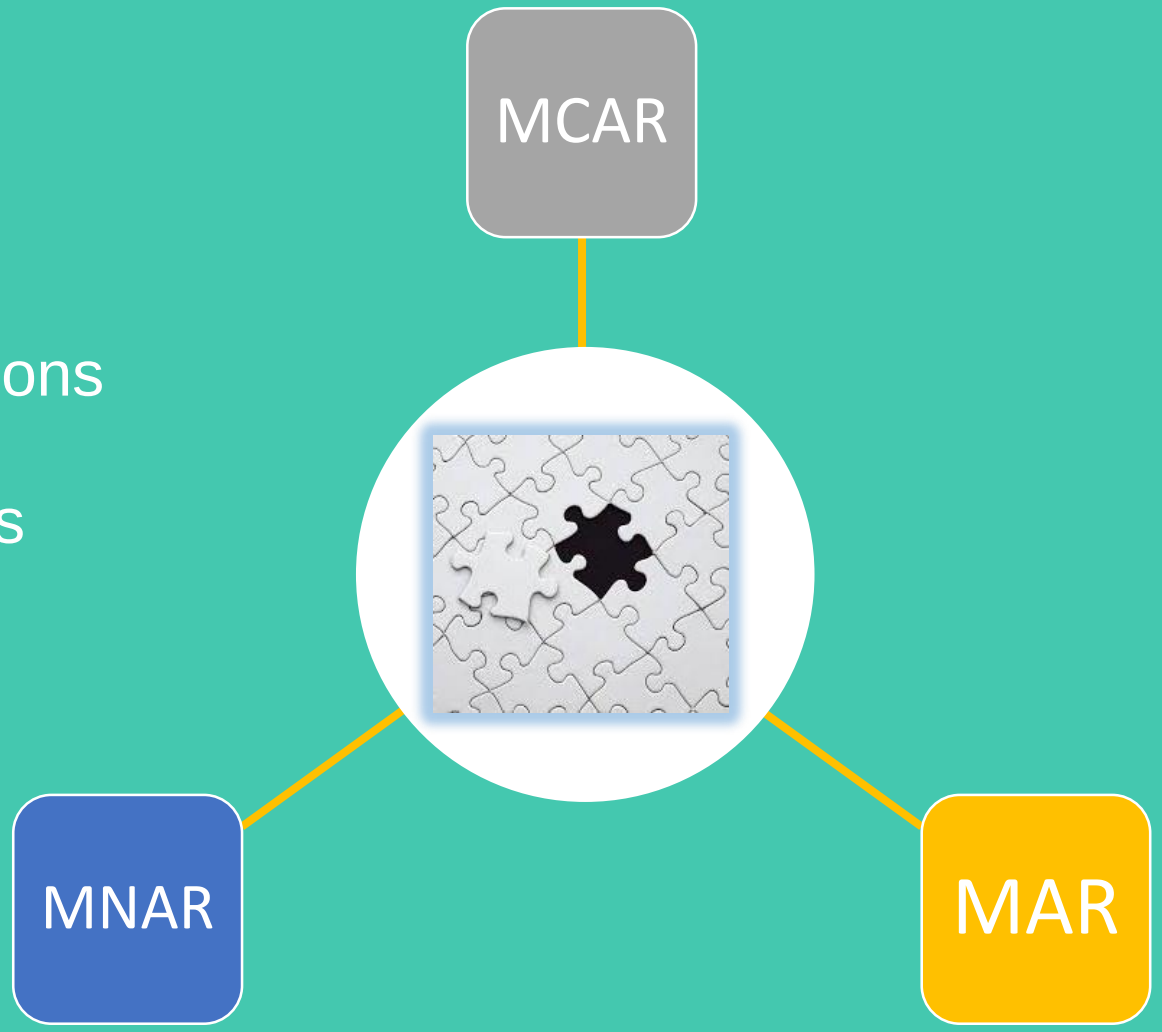
Labels
Infrequent categories



Outliers
Unusual or unexpected values

Missing data

- Missing values for certain observations
- Affects all machine learning models
  - Scikit-learn



Mechanisms of missing data

Missing data completely at random MCAR

- the probability of being missing is the same for all the observations
- there is absolutely no relationship between the data missing and any other values, observed or missing, within the dataset
- disregarding those cases would not bias the inferences made

Missing data at random MAR

- the probability an observation being missing depends only on available information

Gender	Weight
Male	60 kg
Male	NA
Male	NA
Male	77 kg
Male	80 kg
Male	62 kg
Female	NA
Female	NA
Female	60 kg
Female	55 kg
Female	NA
Female	58 kg

2 NA / 6 men = 33%

3 NA / 6 women = 50%

Missing data not at random MNAR

- there is a mechanism or a reason why missing values are introduced in the dataset.

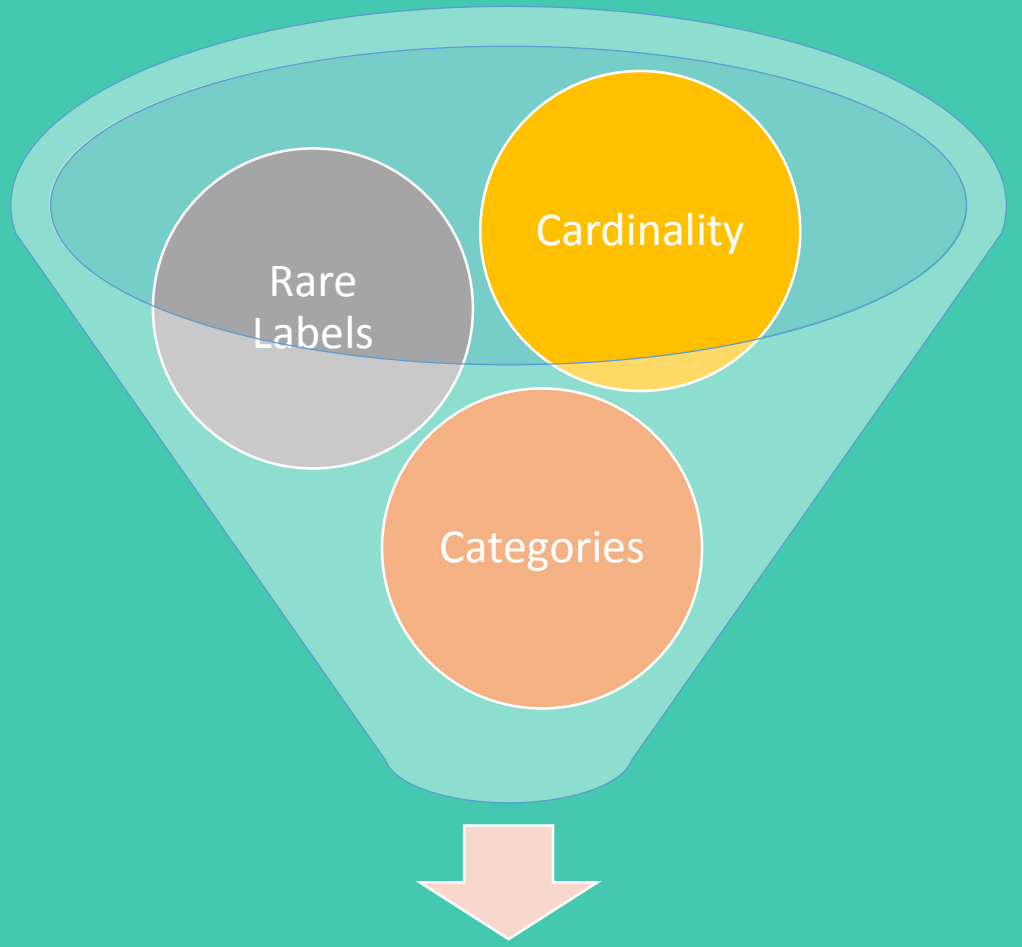
Target = depression	No of clinic visits	No sports classes weekly
Yes	1	NA
Yes	NA	NA
Yes	NA	0
Yes	4	2
Yes	NA	1
Yes	3	NA
No	0	0
No	NA	5
No	1	2
No	1	1
No	2	1
No	NA	2

More NA overall for depressed patients

Less NA for non-depressed patients

Labels

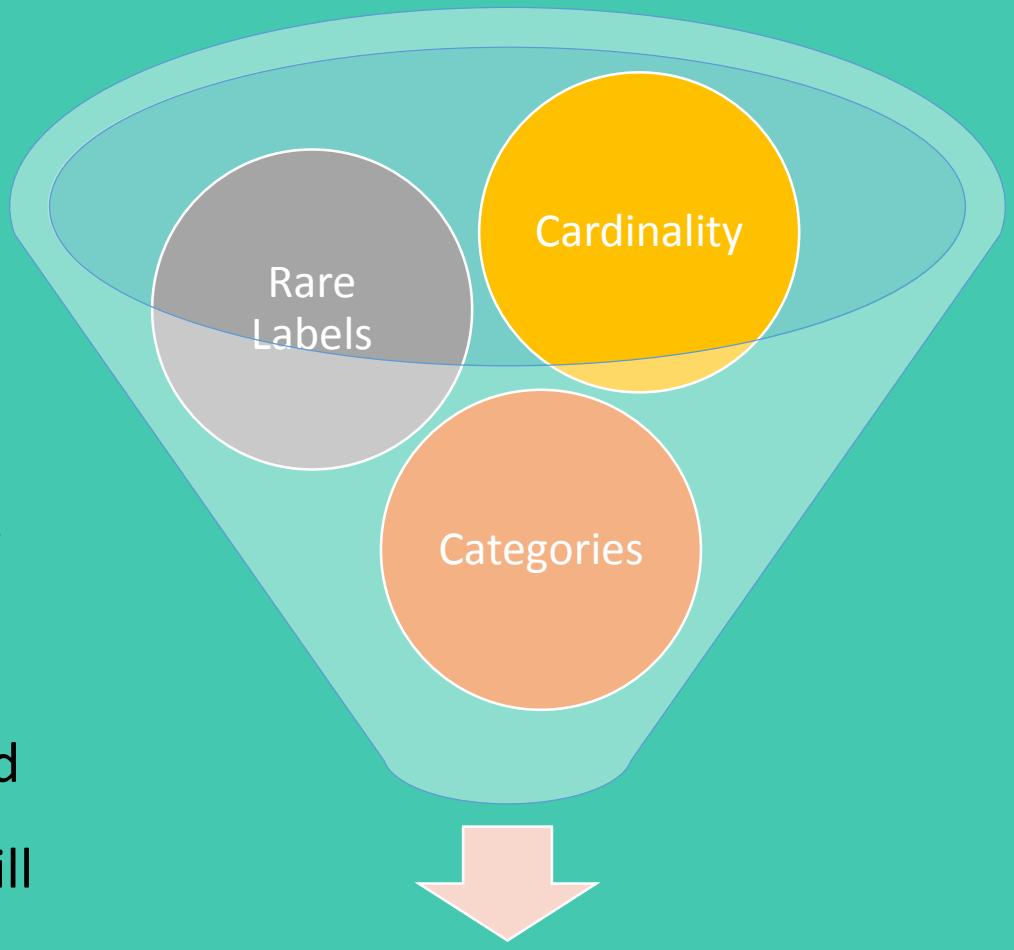
- Cardinality: high number of labels
- Rare Labels: infrequent categories
- Categories: strings
    - Scikit-learn



Tree based methods

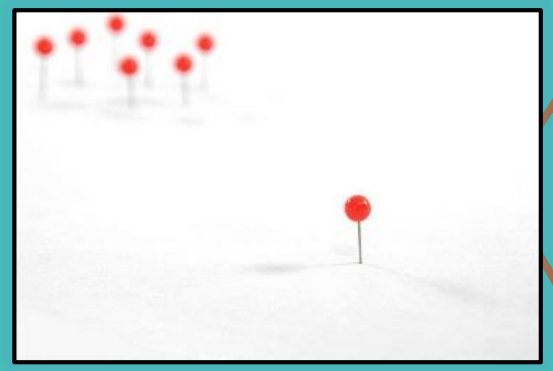
Labels

- Cardinality: Variables with too many labels tend to dominate over those with only a few labels, particularly in **Tree based** algorithms
- Rare Labels: Rare labels may be present in training set, but not in test set, causing over-fitting to the train set
- Rare Labels: Rare labels may appear in the test set, and not in the train set. Thus, the machine learning model will not know how to evaluate it for scoring.

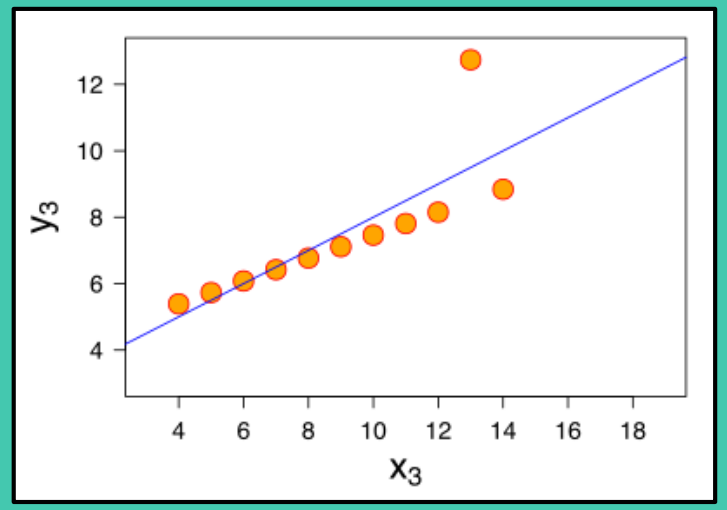


Tree based methods

Outliers



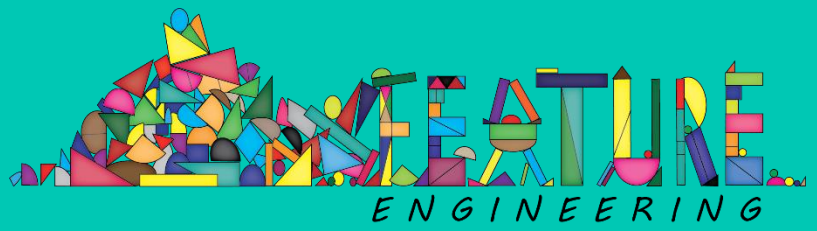
Linear models



Adaboost

Tremendous weights

Bad generalisation



How can we address these variable problems?

Problems with variables

- Practical examples of missing data
- Practical examples of how outliers, highly cardinal variables and rare labels affect ML algorithms performance
- Table with comparison of different machine learning models
- Additional reading resources