DATA PREPROCESSING

Topics:

- removing & imputing missing values from the dataset
- getting categorical data into shape for ML algorithms
- selecting relevant features for model construction

Dealing with missing data

→ there are a variety of reasons we may be missing data:
  - error in the collection process
  - measurements may not be applicable
  - fields couldve been left blank

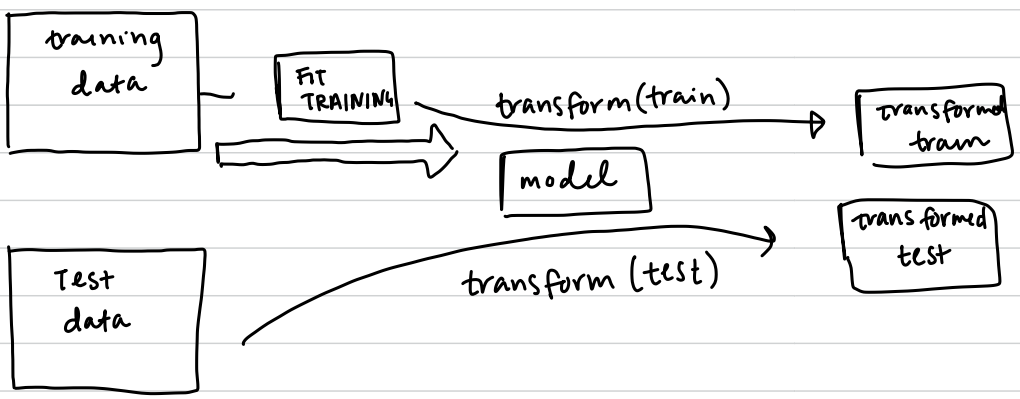
→ empty data can be represented as NaN or NULL types

→ We can remove columns/rows that contain NULLs or NaNs, but we must be sure to do this carefully to not remove important training instances or features.

imputing

- use interpolation techniques to estimate the missing values.
- the most common is mean imputation, where we simply replace the missing value w/ the mean of the entire column.
  → median
  → most-frequent
  ] options!

* Within the imputing API, we observed fit/transform methods, in fit, we learn the parameters from the data & the transform method uses those parameters to transform the data



* this is called the transformer API where the estimator API takes in both X_train & Y_train into the fit method.

HANDLING CATEGORICAL DATA

Ordinal : categories that can be sorted or ordered

Nominal : Dont imply any order

→ for ordinal features, we must map them to labels

Encoding class labels:

→ though many classifiers convert labels to integers internally, it is always good to provide class labels as integer arrays to avoid technical glitches.

Encoding nominal features:

→ label encoding nominal features using label encoding is a bad idea because it allows our model to assume untrue relationships about our data.
→ we should one hot encode this data, however this introduces multi-collinearity
→ we can remove redundant columns by blanket removing the first column of our one-hot encoded data

ADDITIONAL SCHEMES:

- Binary encoding: requires fewer feature columns ($\log_2[K]$ vs. K-1)
  → #s are converted to binary representations then the binary # position forms a new feature column.
- count or frequency encoding: replaces the label of each category by the # of times or frequency it occurs in the training data.