Data analysis learnings

March 18, 2024

1 Visualization

For plotting use seaborn (sns)

2 Basic analysis of data

Data with different columns ($X_1, \ldots X_n$), we should plot dimension pairs to analyze their correlation and figure out irrelvant column (low correlation or have mostly missing values) in order to drop them. After dropping the columns, we either drop the rows with na or we backfill.

3 Backfilling

It is not good to backfill with overall mean. Find a highly correlated dimension and segment the column by that dimension and backfill based on the mean of each slice. For example, we have Age and PClass in Titanic data, it is not good to backfill age with overall mean but rather with the mean of each PClass (1,2,3).

4 Outlier

outlier can cause overfitting because they can't be generalized. Training on outliers will cause the model to become wiggly hence overfitting. Sometimes you want to remove them.