How to handle missing data

Chris Luciuk
Insight Data Science
February 20, 2018

Why Might Data be Missing?

Case Study: You have a 3 question survey to determine the demographics of a population. You ask for individuals’ *race*, *age*, and *income*. When you tabulate the results you find that some respondents did not answer every question. Why?

Why Might Data be Missing?

Case Study: You have a 3 question survey to determine the demographics of a population. You ask for individuals' *race*, *age*, and *income*. When you tabulate the results you find that some respondents did not answer every question. Why?

- Completely at random - probability of missing is the same for all rows

<u>Example:</u> the pages of the survey were sticky and so respondents randomly couldn't see one question.

Why Might Data be Missing?

Case Study: You have a 3 question survey to determine the demographics of a population. You ask for individuals' *race*, *age*, and *income*. When you tabulate the results you find that some respondents did not answer every question. Why?

- Completely at random
- At random - probability of missing depends on other recorded values

Example: age and race are highly predictive of whether a respondent will share their income (i.e., old white guys don't write down their income 80% of the time)

Why Might Data be Missing?

Case Study: You have a 3 question survey to determine the demographics of a population. You ask for individuals' *race*, *age*, and *income*. When you tabulate the results you find that some respondents did not answer every question. Why?

- Completely at random
- At random
- Depends on unobserved predictors - probability of missing value depends on something not recorded

<u>Example:</u> Grumpy people did not answer questions about their age (but you did not assess how grumpy people were in your survey)

Why Might Data be Missing?

Case Study: You have a 3 question survey to determine the demographics of a population. You ask for individuals' *race*, *age*, and *income*. When you tabulate the results you find that some respondents did not answer every question. Why?

- Completely at random
- At random
- Depends on unobserved predictors
- Depends on missing value - probability of missing depends on missing value

<u>Example:</u> individuals do not report income if over $100,000

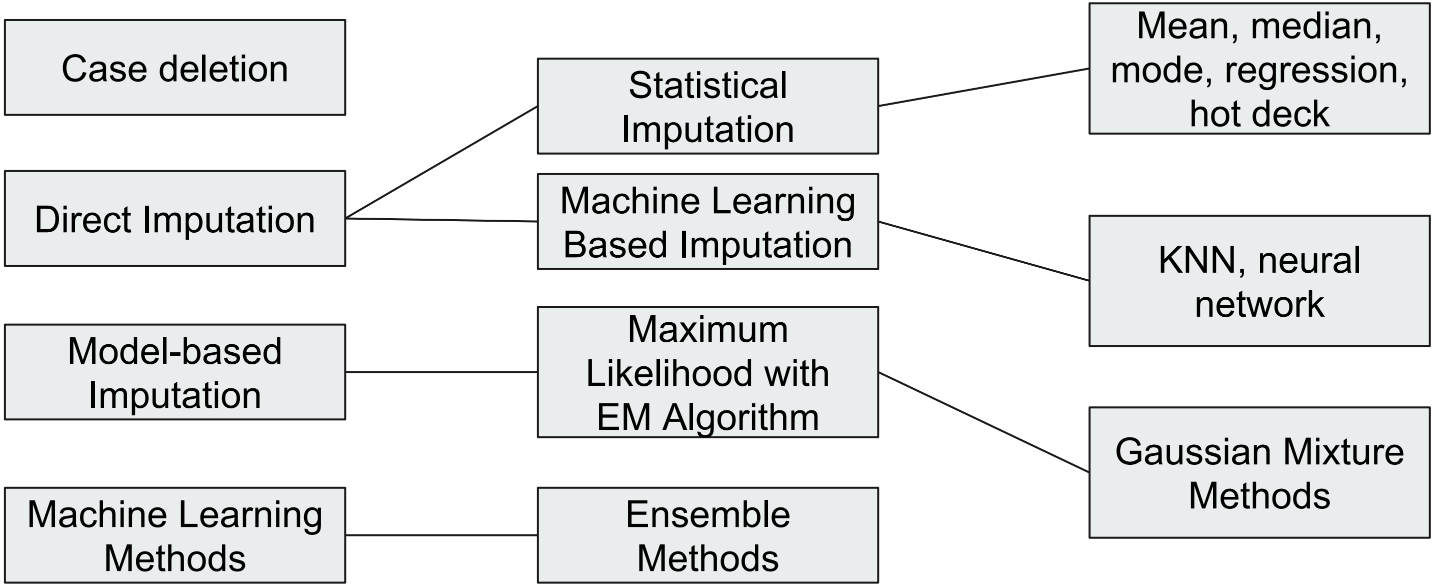
Visualize and Understand Missing Data

It’s often instructive to look at summary statistics to identify data types, missing data, etc.

Why is the data missing?

Is the data missing at random?

How to handle missing data?



Statistical Imputation - Mean, Median, Mode

Use present data to calculate the mean, mode, or median and fill missing values using it

Can use sklearn.preprocessing.Imputer

Statistical Imputation - Hot Deck

Fill missing values using a randomly selected similar record.

Last Observation Carried Forward (LOCF) - create sorted data set and use preceding value to fill in missing values

Statistical Imputation - Regression

Predict observed values of one feature using another feature. Then use prediction to fill in missing values.

Be cautious of underestimation of error

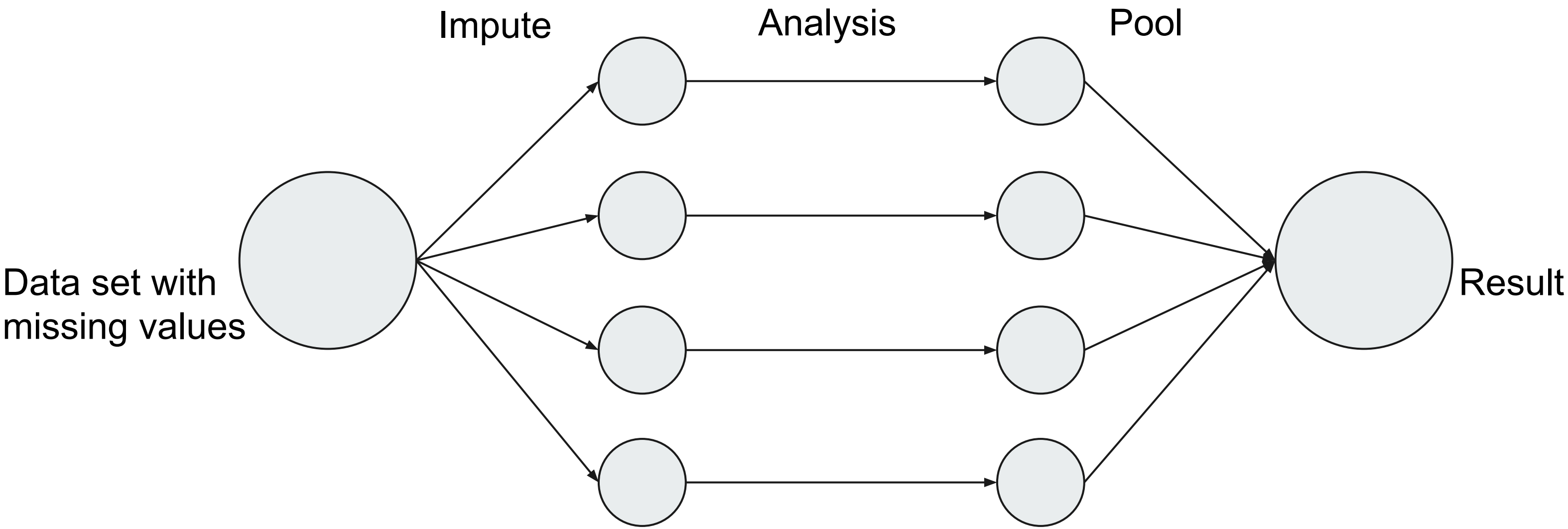
Machine Learning Imputation - kNN

Use machine learning models to impute missing values.

kNN classifier can be used to predict class label for missing values

Multiple Imputation

Repeat imputation and analysis multiple times and pool results.


Impute
Analysis
Pool
Data set with
missing values
Result