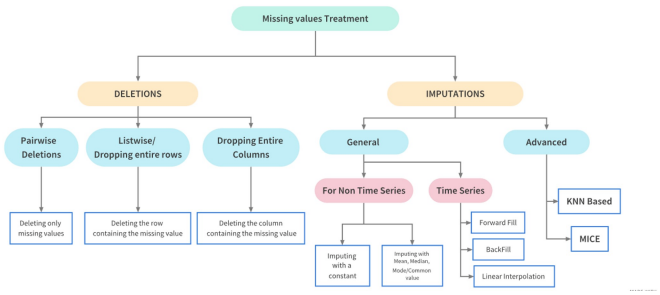
Missing values

Friday, November 29, 2024 10:32 AM



2. Мал процент недостигачи (< 5%):

- Ако е нумеричка, можеш да импутираш со медијаната (ако има аномалии) или со средната вредност (ако нема аномалии).
- Ако е категоријална (иако во примерот не е), импутирај со модалната (најчесто користена) вредност.

3. **Умерен процент недостигачи (5%-30%):**
   - Тука зависи од важноста на колоната за твојата анализа:
     - Ако колоната е клучна за твојот модел, импутирај со напредни техники:
       - Нумерички: **KNN Imputer** или регресија базирана на други атрибути.
       - Категоријални: **Честопати најкористена вредност** или со импутација базирана на соседни податоци.
     - Ако колоната не е клучна или е ирелевантна, размисли за отстранување на колоната.
   - **Пример:** Колоната `CouncilArea` (10.081001% недостигачи)

4. **Висок процент недостигачи (> 30%):**
   - Колоните со многу недостигачи вредности обично додаваат малку информации и можат да го отежнат моделот.
   - Размисли да ги отстраниш овие колони ако се ирелевантни или нивната импутација е премногу комплицирана.
   - **Примери:**
     - `BuildingArea` (47.496318%)
     - `YearBuilt` (39.580265%)

Визуелизација на Missing Values

1) **Heatmap => покажува дали групирани податоци фалат или random**
   - Ако се групирани ffill, bfill
   - Ако се случајни -> избор на импутирачки метод
2) Bar Chart со missing values

3) Matrix за зависност

- Корелација меѓу колони може да помогне така што ако често се појавуваат заедно 2 колони тогаш можеме **едната да ја пополниме врз база на другата** (KNN или линеарна регресија)
  ** Iterative Imputer ако имаат СИЛНА корелација (нумерички/категориски)
  ** KNN Imputer -> мал % missing values и слични пар соседи (нумерички) висина, тежина, возраст

  ВИСОКА КОРЕЛАЦИЈА = NOT MISSING AT RANDOM (напредни техники KNN, MICE, Iterative)
  НИСКА / НЕПОСТОЕЧКА КОРЕЛАЦИЈА = MISSING AT RANDOM / COMPLETELY (тогаш со мода, медијана)

  Ако нема корелација
  ** Mean, Median за мали количини missing values

- **Интерполација => континуирани податоци / временски серии**

- **Forward Fill/Backward Fill => кога има логичка поврзаност** и треба да се пополнат врз база на претходни или следни вредности (временски серии, цена на акции)

- **Linear Regression => СИЛНА КОРЕЛАЦИЈА**
- Се прави линеарен модел и се предава на таа колона

- Heatmap(correlationMissingValues)
- Ако има висока корелација, импутацијата на едната колона може да се направи врз база на другата со користење линеарна регресија на пример.
- Ако повеќе променливи се поврзани - MICE

NaN енкодирање

- Ако користиме KNN Imputer прво мора да направиме Encoding на категориските променливи, тогаш и NaN се енкодира и мора да го вратиме назад во NaN
- За Iterative Imputer не треба NaN енкодирање

- Ако има ? (знак) и тоа се смета за NaN и треба да се справиме со тоа
- Како гледаме содржина на колоната ? Unique() --> ќе ни врати листа од вредностите

КОГА БРИШЕМЕ А КОГА ПОПОЛНУВАМЕ:

- Бришење:

Метод	Кога да се користи?
Бришење на редици	Кога недостасувачките вредности се случајни и само во неколку редици (мал % од вкупниот сет).
Бришење на колони	Кога колоната има **многу недостасувачки вредности** и не може лесно да се пополни со импутација.

- Ако во target има missing value го тргаме редот
- Се брише цела колона ако има исти вредности
- Се брише цела колона ако не е важна за предвидување (Id, датуми)

- Пополнување: Кога колоната е важна

Техника	Кога да се користи	Примерни сценарија
MICE	Кога имате **повеќе променливи кои се поврзани** меѓу себе. Секој атрибут се импутира посебно, врз основа на другите.	Анализа на **здравствени податоци**: Висина, тежина, возраст.
Iterative Imputation	Кога сакате **точна и итеративна предикција на недостасувачки вредности** за променливи кои се поврзани.	Пополнување на **плати** врз база на искуство и возраст.
Интерполација	За временски или секвенцијални податоци, каде што вредностите зависат од **претходните и следните вредности**.	Пополнување на **температура** или **влажност** во временски податоци.

- **MICE и Iterative Imputation** се идеални за податоци каде колоните се меѓусебно поврзани.
- **Интерполација** е погодна за **временски и секвенцијални податоци.**