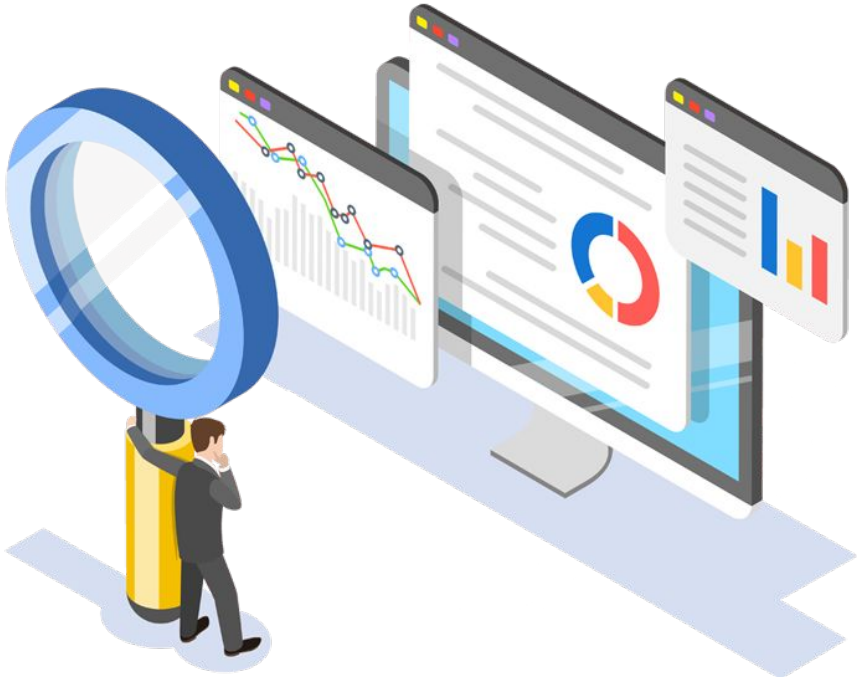
Handling Missing Data

Feature Engineering and Selection: A Practical Approach for Predictive Models



Emanuela Elli (892901)
Alessandro Fasani (837301)
Federica Madon (825628)

High Dimensional Data Analysis - 2022/2023

I missing data e le loro cause

Comprendere il fenomeno

La gestione dei dati mancanti è fondamentale per un **corretta costruzione** dei modelli predittivi. La prima fase riguarda la comprensione del fenomeno e la distribuzione di questo tra predittori ed osservazioni, mediante **visualizzazioni** e **indici di sintesi**.

Misurazioni di scarsa qualità

A causa di eventi esogeni o danni alle strumentazioni si possono ottenere alcune dei valori di alcune grandezze che risultano mancanti o con un forte bias (immagini sfuocate, sonde danneggiate).

Fase di merging

L'assenza di alcune chiavi portano ad avere valori mancanti.

df1

	A	B	C	D
0	A0	B0	C0	D0
1	A1	B1	C1	D1
2	A2	B2	C2	D2
3	A3	B3	C3	D3

df4

	B	D	F
2	B2	D2	F2
3	B3	D3	F3
6	B6	D6	F6
7	B7	D7	F7

Result

	A	B	C	D	B	D	F
0	A0	B0	C0	D0	NaN	NaN	NaN
1	A1	B1	C1	D1	NaN	NaN	NaN
2	A2	B2	C2	D2	B2	D2	F2
3	A3	B3	C3	D3	B3	D3	F3
6	NaN	NaN	NaN	NaN	B6	D6	F6
7	NaN	NaN	NaN	NaN	B7	D7	F7

Eventi randomici

Non permettono la misurazione del fenomeno, come in caso di interruzioni di corrente o smarrimento di alcuni campioni.

Natura del fenomeno dei Missing Values

3 forme di Missing Values

01 I Missing Values si presentano in modo **strutturato**; come in caso di omissione di una classe o misurazioni senza azionamento.

Structural Deficiency

02 A seconda di quanto la componente stocastica sia tale, si possono presentare in 2 modi:

Random Occurencies

2.1 La probabilità di osservare un valore mancante è **indipendente** dalle osservazioni.

MCAR

2.2 La probabilità **non** è **indipendente**. In caso di trial clinici, differenze tra uomini e donne.

MAR

03 Si tratta di un'assenza di dati dovuta a **cause specifiche**; come l'abbandono prematuro dei trials.

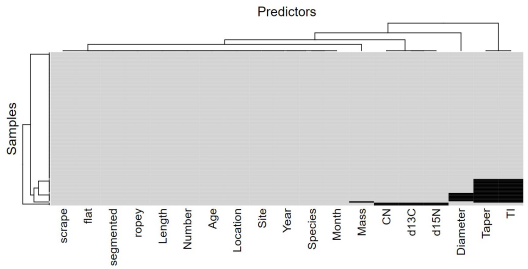
NMAR

Forma e distribuzione dei Missing Values (1 di 2)

Datasets medio-piccoli: In esame un dataset di campioni di escrementi animali, 110 osservazioni, 19 missing values.

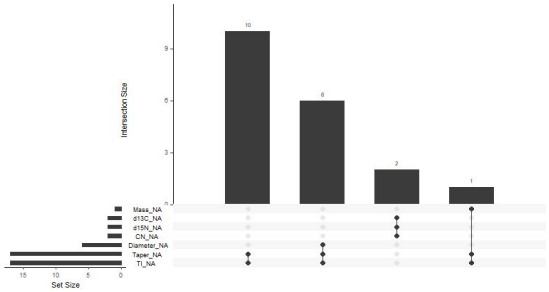
Heatmap

Molto utili quando i dataset sono di **dimensioni ridotte** (100 oss. e/o 100 predittori). Prodotto con la funzione **`heatmap()`** del pacchetto `{stats}`, nativamente **raggruppa gerarchicamente** righe e colonne.



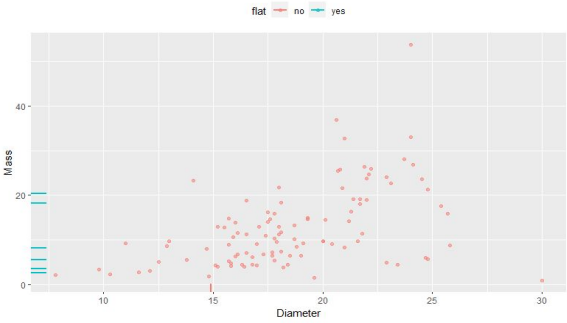
Co-occurence plot

Prodotto con i tools del pacchetto `{naniar}`, questo grafico ha un focus sui predittori, evidenziando **quali** e **quanti** di questi si presentano congiuntamente con valori ***na*** tra le osservazioni.



Scatterplot

Gli scatterplot possono essere un ottimo modo per visualizzare il **comportamento dei missing values** tra due variabili numeriche, **alla luce di una terza variabile**, in questo caso categorica binaria.

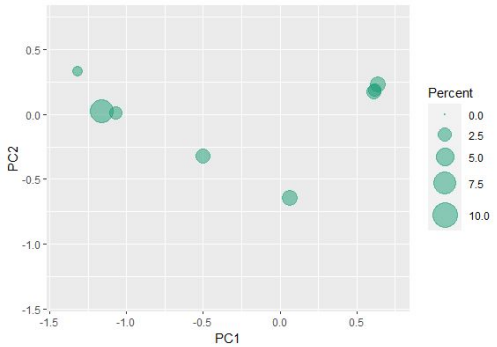


Forma e distribuzione dei Missing Values (2 di 2)

Datasets medio-grandi: In esame un dataset relativo ai passeggeri del trasporto ferroviario in 5733 giorni di misurazioni e lungo 137 stazioni.

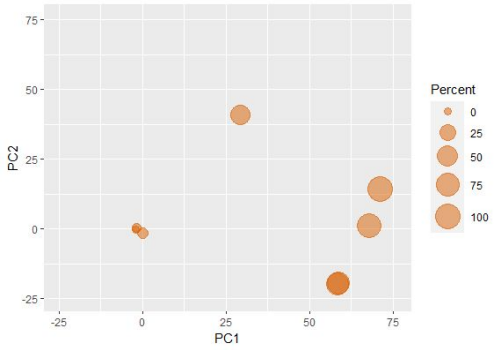
PCA su osservazioni

Lo scopo della PCA, in questo caso, è individuare le **direzioni** con i massimi valori di variabilità, nonché il numero di **patterns** con cui il fenomeno dei valori mancanti si manifesta tra le osservazioni.



PCA su predittori

La medesima visualizzazione dei patterns può essere creata in funzione dei predittori. Per entrambe le PCA sono stati creati appositamente dataset binari (0 e 1) completi.



Missing data patterns

Per avere un ulteriore comprensione di come sia utile individuare i pattern, il grafico sottostante ordina e raggruppa i predittori per intensità di missing.

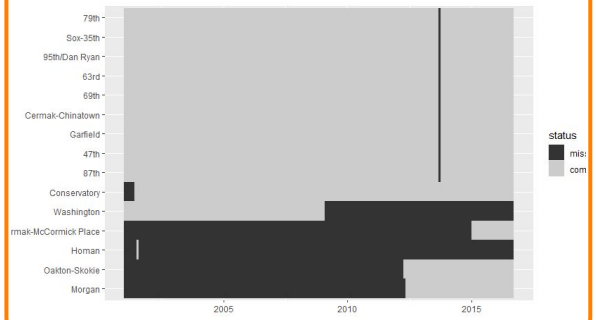


Tabelle di sintesi

Datasets molto grandi: Quando le visualizzazioni precedenti diventano caotiche e di difficile interpretabilità. Tra le più semplici sono presenti le frequenze relative percentuali.

Percentuali dati mancanti tra i predittori

Dati mancanti	Predittori
15.5%	Taper e TI
5.5%	Diameter
1.8%	d13C, d15N e CN
0.9%	Mass
0%	Tutte le altre

Percentuali dati mancanti tra le osservazioni

Dati mancanti	Id Osservazioni
15.8%	11, 13, 14, 15, 29, 60, 67, 80, e 95
10.5%	51, 68, 69, 70, 71, 72, 73, 75, 76, e 86
0%	Tutte le altre

Modelli "resistenti" ai valori mancanti

CART

- Algoritmi CART (Classification & Regression Trees) di **Breiman** 1984.
- Un albero binario che utilizza l'indice **GINI** come criterio di suddivisione.
- CART può gestire attributi sia nominali che numerici per costruire un albero decisionale.

C5.0

- Algoritmo C5.0 di Ross **Quinlan** e successore dell'algoritmo C4.5 anch'esso sviluppato da Quinlan (1994).
- Fornisce un albero binario o un albero a più rami.
- Utilizza Information_Gain (**entropia**) come criterio di suddivisione.
- La tecnica di potatura C5.0 adotta il metodo del **limite di confidenza binomiale**.

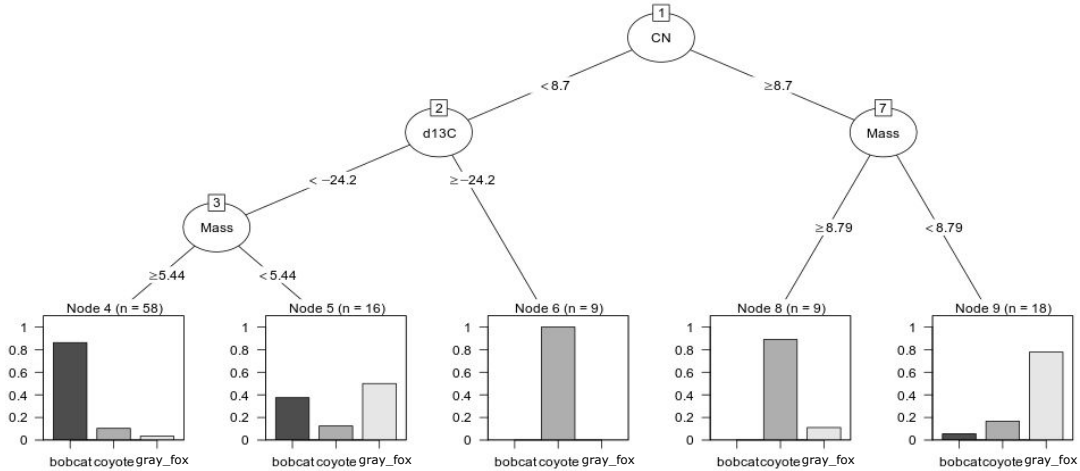
Naive Bayes

- Famiglia di classificatori probabilistici, basati sul **teorema di Bayes** della probabilità condizionata.
- Ciascuna feature nel dataset contribuisce in modo **indipendente**, e con lo stesso peso, alla determinazione della classe dell'istanza.
- L'assunzione è molto forte ma il metodo funziona bene nei casi reali.

Modelli "resistenti" ai valori mancanti

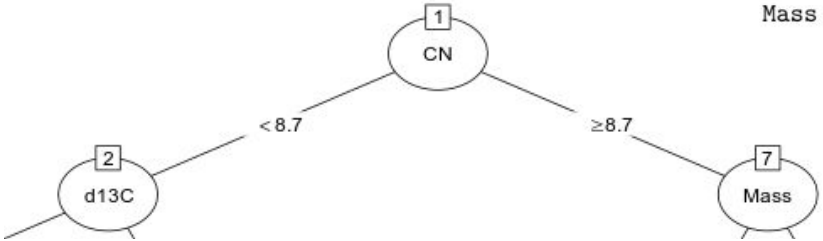
CART

- Utilizza la logica dei ***surrogate splits***.
- Per ogni divisione vengono considerate le suddivisioni alternative i cui risultati sono simili alla suddivisione originale dell'albero.
- Se il surrogate split si avvicina bene alla suddivisione originale, può essere utilizzato quando i dati del predittore non sono disponibili.
- Viene memorizzato non solo lo split migliore (chiamato ***split primario***) ma anche diversi surrogate splits per ogni divisione primaria nell'albero.



Summary(decision tree)

```
Node number 1: 110 observations,    complexity param=0.2641509
  predicted class=bobcat    expected loss=0.4818182  P(node) =1
    class counts:    57    28    25
   probabilities: 0.518 0.255 0.227
  left son=2 (83 obs) right son=3 (27 obs)
  Primary splits:
      CN        < 8.7     to the left,  improve=13.135800, (2 missing)
      d13C      < -24.195 to the left,  improve=12.812480, (2 missing)
      d15N      < 11.78   to the left,  improve=10.617790, (2 missing)
      Mass      < 8.64    to the right, improve= 8.667345, (1 missing)
      segmented < 0.5     to the right, improve= 6.139761, (0 missing)
  Surrogate splits:
      flat      < 0.5     to the left,  agree=0.806, adj=0.222, (2 split)
      Diameter < 11.85    to the right, agree=0.778, adj=0.111, (0 split)
      Month    splits as  LLLRLLLLL,    agree=0.769, adj=0.074, (0 split)
      Mass      < 3.625   to the right, agree=0.769, adj=0.074, (0 split)
```



Modelli "resistenti" ai valori mancanti

C5.0

- Utilizza i ***conteggi frazionari***, ovvero quando il valore di un attributo nell'albero non è noto, C5.0 suddivide il caso e invia una frazione a ciascun ramo sottostante.
- Consente al modello di tenere un ***resoconto corrente*** di dove i valori mancanti potrebbero essere arrivati nel partizionamento.

Naive Bayes

- ***Modella separatamente*** le distribuzioni specifiche della classe di ***ciascun predittore***.
- In caso si verifichi la presenza di missing values, il modello è in grado di utilizzare le informazioni delle distribuzioni riguardanti i dati completi evitando la cancellazione dei singoli dati mancanti.

Cancellazione dei dati

Eliminazione dei dati mancanti

- Approccio più ***semplice***.
- Potrebbe consistere nell'eliminazione di interi predittori o di campioni specifici.
- Non adatto a tutti i set di dati.

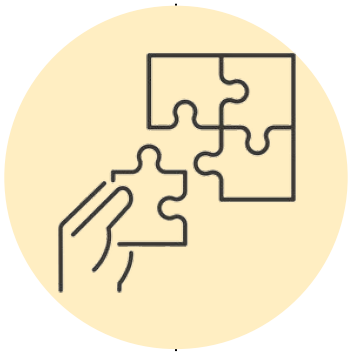
Valore intrinseco dei campioni

- Quando è difficile ottenere campioni o quando i dati contengono un piccolo numero di campioni.
- NON è desiderabile rimuovere dati mancanti.
- I campioni hanno priorità più alta di esser conservati perché più critici dei predittori.

Influenzabilità del modello

- Influenzabilità del modello a causa della rimozione di dati nella previsione dei risultati.
- Esempio studi medici.

Codifica dei valori mancanti



Predittore discreto

Nel caso in cui fosse presente un predittore di natura discreta, la mancanza strutturale dei valori può essere codificata direttamente nel predittore come se fosse una categoria naturale.

Codifica dei valori mancanti

In altri casi i valori mancanti potrebbero essere semplicemente codificati come "mancanti" o "sconosciuti".

➔ Sfortunatamente, è impossibile stabilire se codificare i valori mancanti è una strategia vincente per tutti i dataset.

Principio guida

Per determinare se la mancanza di codifica sia una metodologia opportuna, si può pensare a come i risultati verrebbero interpretati se quell'informazione diventasse importante per il modello.

Metodi di imputazione

K-Nearest Neighbors

Campioni simili

Questo algoritmo identifica un campione con uno o più valori mancanti. Quindi identifica i K campioni più simili nei dati di addestramento che sono completi.

Metrica di distanza

La somiglianza dei campioni per questo metodo è definita da una metrica di distanza: quella euclidea oppure la distanza di **Gower**.

Distanza di Gower

Per un predittore qualitativo, la distanza tra due campioni è 1 se i campioni hanno lo stesso valore e 0 altrimenti. Per un predittore quantitativo x è la seguente (con R_x intervallo del predittore):

$$d(x_i, x_j) = 1 - \frac{|x_i - x_j|}{R_x}$$

Bagged Trees

Trees

Un albero può essere costruito in presenza di dati mancanti. Un singolo albero è noto per produrre risultati che hanno bassa distorsione ma alta varianza.

Gli insiemi di alberi, tuttavia, forniscono un'alternativa a bassa varianza. Le **Random Forests** sono una di queste tecniche.

Bagged Trees

Una buona alternativa che ha un ingombro computazionale più piccolo è un **bagged tree** che è costruito in modo simile a una foresta casuale.

Prestazioni

Le prestazioni di un bagged tree, utilizzando 25-50 alberi, sono confrontabili con le prestazioni di un modello che utilizza le random forests.

K-Nearest Neighbors e Bagged Trees

In **rosa** i valori imputati.

K-Nearest Neighbors imputa i nuovi valori per lo più nell'intorno della periferia delle due dimensioni.

L'algoritmo che utilizza **Bagged Trees** ha imputato i valori utilizzando 50 alberi.

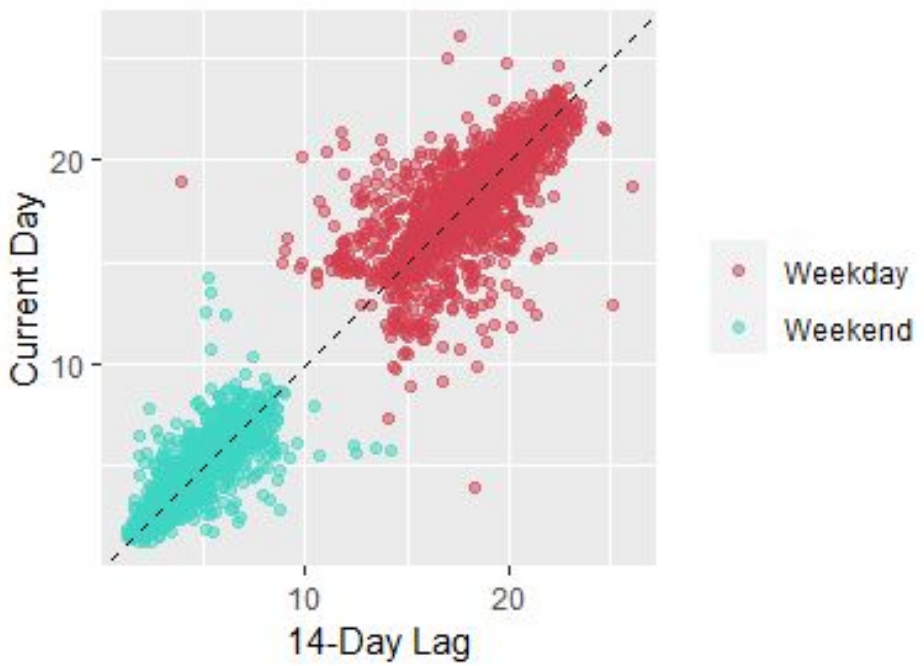
Entrambi i modelli imputano i nuovi valori all'interno della gamma dei campioni completi.

Linear Methods

Quando un predittore completo mostra una forte relazione lineare con un predittore che richiede l'imputazione, un **modello lineare semplice** può essere l'approccio migliore.

La **regressione lineare** può essere utilizzata per un predittore numerico che richiede l'imputazione.

Allo stesso modo, la **regressione logistica** è appropriata per un predittore categorico che richiede l'imputazione.



Casi speciali

Valori censurati o troncati

Ci sono situazioni in cui un punto dati non manca ma non è nemmeno completo. Questa tipologia di valori sono indicati come “**censurati**” (o in alcuni casi “**troncati**”).

Censura di sinistra

Le durate sono spesso giustamente censurate poiché il valore finale non è noto. In altri casi, può verificarsi la **censura di sinistra**.

Imputazione

Per mitigare il **problema della variabilità**, i valori censurati a sinistra possono essere imputati utilizzando valori uniformi casuali compresi tra zero e l’estremo, oppure utilizzare altri schemi di assegnazione di valori casuali che rappresentano meglio la distribuzione (se nota).

Grazie per l’attenzione!