CDLM DATA SCIENCE, UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA, CORSO DI HIGH DIMENSIONAL DATA ANALYSIS, ANNO ACCADEMICO 2022-2023

Feature Engineering and Selection: A Practical Approach for Predictive Models

Handling Missing Data

Progetto a cura di: Emanuela Elli (892901), Alessandro Fasani (837301), Federica Madon (825628)

Contents

1 Introduzione

La gestione dei dati mancanti riveste un ruolo fondamentale per una corretta costruzione dei modelli predittivi, per questo è fondamentale nelle fasi di pre-processing, in quanto alcune tecniche possono risultare fallimentari in presenza di missing values (`NULL`, `NaN`, o altri valori di default usati per temperare l'assenza di dati). La presenza di valori mancanti in un dataset non è un evento raro. Queste assenze di dati possono avvenire per diverse cause:

- Merging tra due datasets: quando viene eseguito un *outer join*, l'assenza di uno o più chiavi nei datasets porteranno ad ottenere uno o più valori mancanti nelle righe del nuovo dataset creato.
- Eventi randomici che non hanno permesso la misurazione del fenomeno: si pensi per esempio allo smarrimento di alcuni campioni in analisi di laboratorio.
- Misurazioni di scarsa qualità o fallite del tutto: questo può avvenire per vari fattori, soprattutto esogeni, che causano delle misurazioni con bias, come delle riprese video fuori fuoco.

Le cause di questi *missing values* possono essere identificate a seconda della strumentazione e della documentazione prodotta per il fenomeno oggetto di studi, come ad esempio note di riferimento, appunti, dati di log od altri registri specialistici.

Chiarite le cause che possono portare ad avere dei valori mancanti tra le misurazioni, un altro aspetto da considerare riguarda la natura di questi valori assenti. Comprendere la natura e la forma è altresì necessario quando non si hanno ben chiare le cause che hanno portato alla presenza di *missing values*. Distinguiamo 3 differenti meccanismi e forme in cui si possono presentare:

- **Structural deficiency**: queste presenze di missing values si presentano in modo strutturato e possono essere ricollegate all'omissione di un predittore (sia numerico che categorico) dai dati durante la raccolta. Ad esempio, in caso di variabili categoriche, una classe che indichi propriamente "assenza di..." può non essere raccolta, per praticità.
- **Random Occurencies**: le assenze randomiche di dati si presentano in due modi differenti, a seconda di quanto questa componente stocastica di casualità sia tale.
  - *Missing completely at random* (MCAR): quando la probabilità di osservare un valore mancante è indipendente dalle osservazioni, sia di quelle presenti nel *sample* sia di quelle della popolazione. Questo naturalmente è lo scenario migliore, in quanto non necessita di particolari approfondimenti circa le cause dei *missing values*.
  - *Missing at random* (MAR): quando la probabilità di osservare un valore mancante non è indipendente dalle osservazioni, sia di quelle presenti nel *sample* sia di quelle della popolazione. Ad esempio, nei test clinici che prevedono diversi trials, gli uomini sono statisticamente più inclini ad abbandonare preventivamente, portando ad uno squilibrio di dati mancanti tra uomini e donne. Nella pratica, comprendere se un dato sia mancante con una probabilità diversa tra dati osservati e non osservati,

non è impresa facile. Per questa ragione si tende ad applicare i vari metodi di gestione dei *missing data* senza fare distinzioni tra MCAR e MAR.

- **Not missing at random** (NMAR): si tratta di un'assenza di dati non dovuta ad eventi randomici, ma a delle cause specifiche relative ad un soggetto del *sample*, a sottopopolazione o ad alcune specifiche classi di questo. Questa natura di *missing values* è di fatto la più complessa e delicata da gestire.

2 Comprendere la natura e la gravità delle informazioni mancanti

Visualizzare forma e distribuzione dei dati mancanti, nonchè usare misure di sintesi del fenomeno dei *missing data*, è fondamentale per delineare eventuali approcci alla loro gestione. Per dataset di dimensioni medio-piccole (es. 100 osservazioni e 100 variabili) le visualizzazioni possono essere agevolmente plottate per comprendere la severità del fenomeno. Quando le dimensioni del dataset aumentano notevolmente, possono essere introdotte tecniche di riduzione della dimensionalità o utilizzate misure di sintesi altrettanto utili.

Per le visualizzazioni verranno utilizzati datasets "*Chicago train ridership*" e "*Scat data*". Quest'ultimo, contiene 110 osservazioni circa test di laboratorio e morfologici su campioni di escrementi di animali trovati in natura. L'obbiettivo della raccolta è focalizzato sul trovare una corrispondenza tra le misure relative ai predittori e la specie che ha prodotto l'escremento (preventivamente classificato tramite il genotipo nel DNA). Di quelle 110 osservazioni, 19 presentano uno o più valori mancanti.

- **Heatmap**: sono un ottimo modo per visualizzare la natura delle informazioni mancanti quando i datasets sono di dimensioni ridotte. Si procede distinguendo binariamente tutti i valori del dataset a seconda che siano presenti valori `NA` o meno. La funzione `heatmap()` del pacchetto `ComplexHeatmap` riorganizza righe e colonne gerarchicamente e per similarità, tramite funzioni di distanza. In **Figura 1** si può notare come 3 predittori in particolare siano più frequentemente mancanti tra le osservazioni del sample. Due di questi, `tapper` e `tapper index` (`TI`), sono mancanti congiuntamente.

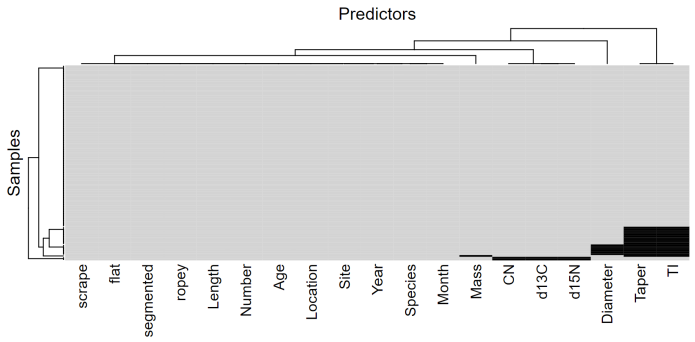

Figura 1: Heatmap con visualizzazione gerarchica delle osservazioni e dei predittori con più valori mancanti.

- **Co-occurence plot**: questo grafico si focalizza sui predittori mancanti, evidenziando quali di questi e

quante volte mancano congiuntamente tra le osservazioni del campione. In **Figura 2** è facilmente visualizzabile come `tapper` e `TI` siano i predittori mancanti nella maggior parte delle osservazioni (17 osservazioni totali).

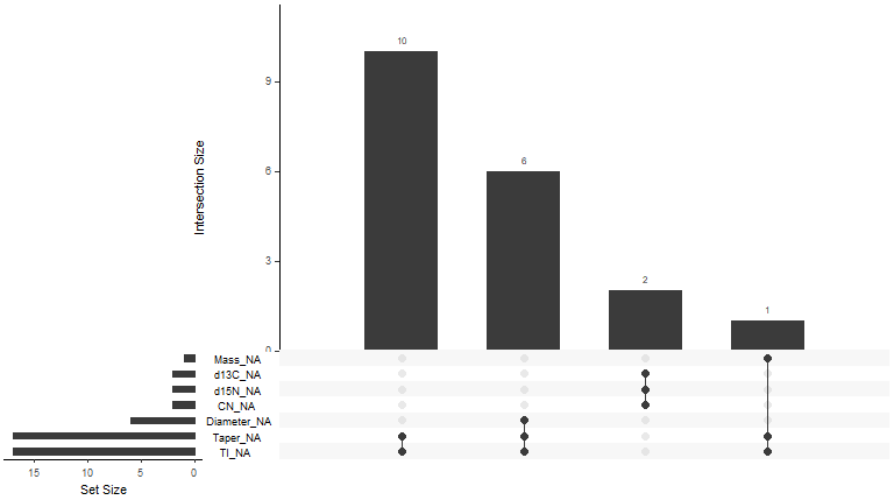

Figura 2: Co-occurence plot relativo ai predittori.

- **Scatterplot**: gli scatterplot possono essere un ottimo modo per visualizzare le relazioni tra i dati e i valori mancanti. In particolare, può essere analizzata l'interazione dei valori mancanti tra due variabili numeriche in presenza di una terza variabile categorica (binaria in questo caso). Come riportato in **Figura 3**, valori mancanti di un predittore possono essere evidenziati in modo differente sull'asse dell'altro predittore e viceversa. In questo scatterplot riportato è possibile notare come massa e diametro abbiano un qualche tipo di associazione crescente. La terza variabile in questione, dicotomica con i colori rosso e verde, fa riferimento alla presenza o meno di un campione di escrementi piatto e senza una forma ben definita. Data l'intrinseca difficoltà nel misurare il diametro di questa tipologia di campioni, possiamo considerare l'assenza di misure circa il diametro di questi, dei *missing values* sistematici, chiaramente riconducibili al valore poi assunto dalla variabile `flat`.

Quando il numero di osservazioni e/o predittori non è più contenuto in poche centinaia, le precedenti visualizzazioni risultano di difficile comprensione e non permettono di individuare eventuali pattern presenti nei dati. Per i successivi esempi di visualizzazioni verrà utilizzato il dataset "*Chicago train ridership*" che riporta il numero di passeggeri (in migliaia) presenti in 5.733 giorni (osservazioni) e lungo 137 stazioni (predittori).

- **PCA**: la *Principal Component Analysis* è una tecnica di riduzione della dimensionalità che può risultare molto utile per visualizzare nel complesso il dataset e i predittori con valori mancanti. Lo scopo della PCA è individuare le direzioni con i massimi valori di variabilità; per questa ragione, come prima fase di implementazione di questa tecnica di visualizzazione, vengono imputati a 0 tutti i valori presenti e ad 1 tutti i *missing value*, assicurandosi un peso maggiore (e proporzionale) a quelle osservazioni e predittori

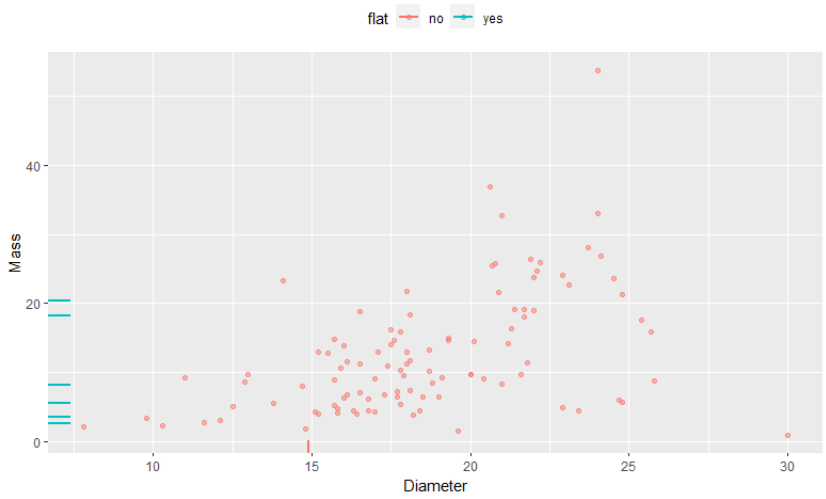

Figura 3: Scatterplot di due predittori numerici, con un terzo predittore binario indicato dai colori.

con più *missing values*. Come risultato si otterranno delle rappresentazioni di *samples* con pochi valori mancanti vicino all'origine e viceversa si allontaneranno dall'origine all'aumentare dei valori mancanti. La **Figura 4** mostra come tra le osservazioni siano presenti 8 pattern differenti con cui queste presentano dati mancanti. Il grado di *missingness* è individuabile sia dalla distanza dall'origine (0, 0) sia dalla percentuale di *missing values* con cui questi pattern si presentano.

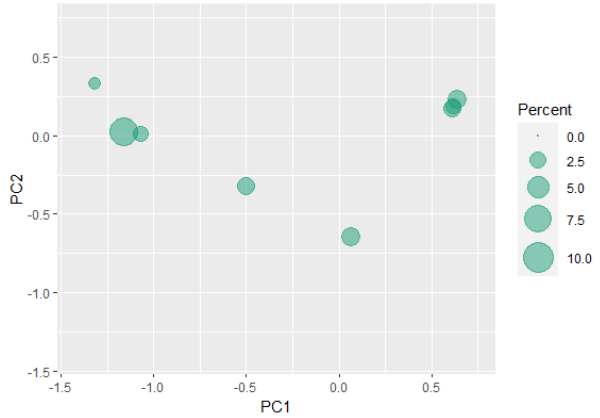

Figura 4: Rappresentazione delle prime due componenti della PCA con focus sulle osservazioni. Sono stati individuati in totale 8 pattern di *missingness*.

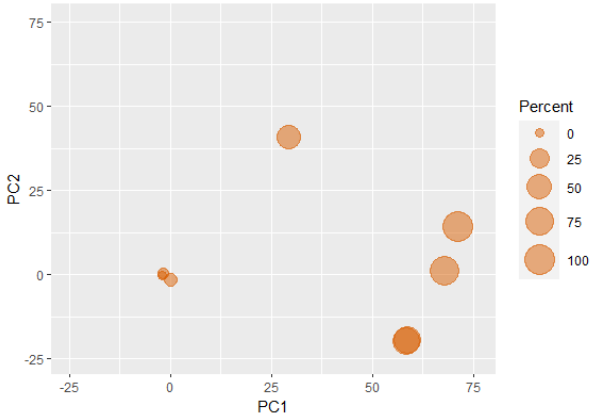

Figura 5: Rappresentazione delle prime due componenti della PCA con focus sui predittori. Sono stati individuati in totale 8 pattern di *missingness*.

Allo stesso modo può essere compreso il grado di severità dei valori mancanti, focalizzando l'attenzione sui predittori, utilizzando la trasposta della matrice (predittori su righe ed osservazioni per colonna). La **Figura 5** mostra come tra i predittori (le stazioni) siano presenti altrettanti 8 pattern differenti con cui

queste presentano dati mancanti.

- **Missing data patterns**: per avere un ulteriore comprensione di come la PCA individua dei pattern, può essere utile visualizzare una rappresentazione dei *missing values* delle stazioni (ordinate con algoritmi di clustering per distanza) e dei giorni (ordinati cronologicamente), come mostrato in **Figura 6**.

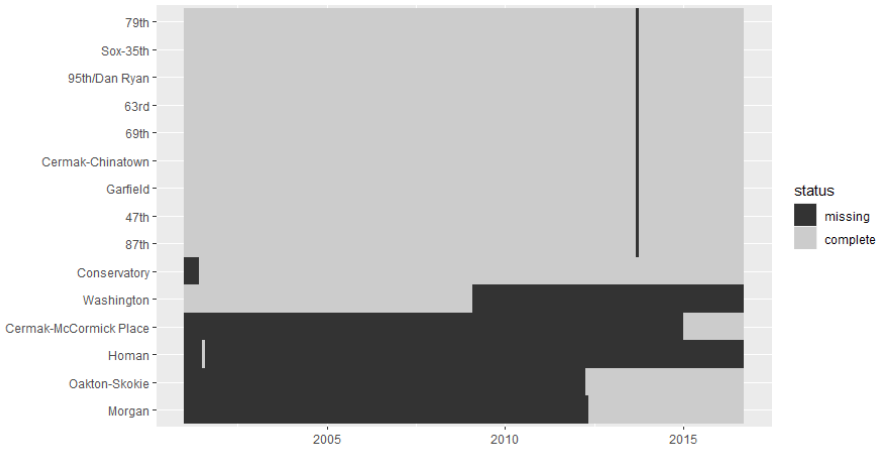

Figura 6: Missing data patterns, righe e colonne ordinate.

Le misure di sintesi sono un altro valido strumento per comprendere la gravità dei *missing data* tra osservazioni e predittori, in particolare quando le visualizzazioni risultano caotiche e di difficile interpretabilità a causa dell'elevata dimensionalità del dataset o del numero di osservazioni. Tra le più semplici si annoverano le frequenze relative % di *missing values* nei predittori e le frequenze relative % di *missing values* nelle singole osservazioni (raggruppate per range o per medesima percentuale di *missing values*) come riportato nella **Tabella 1** e **Tabella 2**.

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

Table 1: Percentuale di dati mancanti per i predittori.

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

Table 2: Percentuale di dati mancanti per le osservazioni.

3 Modelli "resistenti" ai valori mancanti

Molti tra i modelli predittivi popolari, come le *support vector machines*, le *glmnet* e le *neural networks*, non sono in grado di tollerare alcuna quantità di valori mancanti. Ciò nonostante, esistono altrettanti modelli predittivi in grado di gestire internamente i dati incompleti. Ad esempio esistono alcune implementazioni di modelli basati su alberi decisionali come la metodologia CART [1]. Essa utilizza l'idea delle *surrogate splits*, ciò comporta che alla creazione di un albero di decisione vengano catalogate un insieme di suddivisioni chiamate *splits*, utilizzando predittori alternativi rispetto al predittore corrente in grado di approssimare la logica di suddivisione originaria se il valore del predittore è mancante.

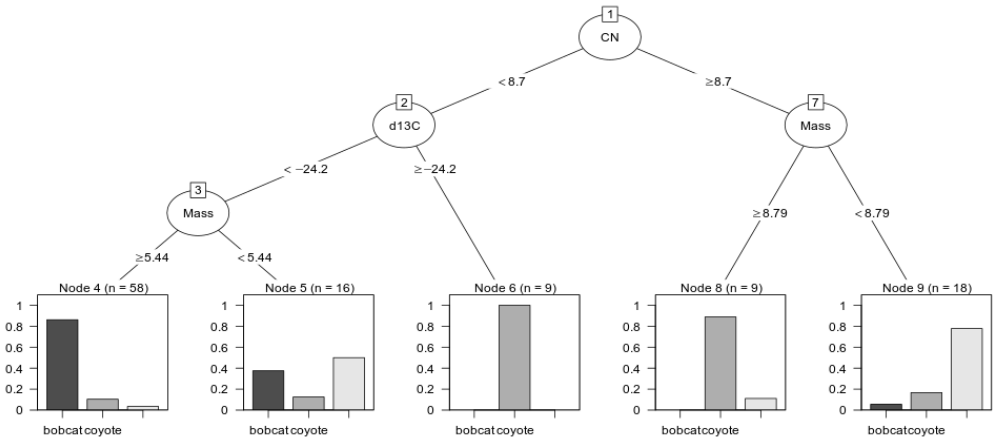

Figura 7: Albero di partizionamento ricorsivo per i dati relativi agli escrementi di animali. Tutti i tre predittori all'interno di questo albero contengono valori mancanti ma possono essere utilizzati in egual modo grazie ai predittori surrogati (*surrogate splits*).

La **Figura 7** mostra il modello di partizionamento ricorsivo per i dati sugli escrementi di animali. Tutti e tre i predittori selezionati dall'albero contengono valori mancanti, come illustrato nella **Figura 2**. La ripartizione iniziale si basa sul rapporto carbonio/azoto (`CN < 8,7`): quando un campione ha un valore mancante per `CN`, il modello CART utilizza una suddivisione alternativa basata sull'indicatore che denotava se gli escrementi fossero di forma piatta o meno. Scendendo ulteriormente nell'albero, i predittori surrogati per `d13C` e `Mass` sono rispettivamente `Mass` e `d13C`. Ciò è possibile poiché questi predittori non mancano contemporaneamente.

La metodologia C5.0 [5] [4] adotta invece un approccio diverso: in base alla distribuzione del predittore con dati mancanti, i conteggi frazionari vengono utilizzati nelle suddivisioni successive. Ad esempio, la prima suddivisione di questo modello per i dati *scat* è `d13C > 24,55`. Quando questa affermazione è vera, tutti i 13 campioni di training set sono "coyote", tuttavia esiste un solo valore mancante per questo predittore. I conteggi di ciascuna specie nei nodi successivi sono quindi frazionari a causa dell'adeguamento per il numero di valori mancanti per la variabile divisa. Questo consente al modello di tenere un resoconto corrente di dove i valori mancanti potrebbero essere arrivati nel partizionamento.

Un altro metodo che può tollerare dati mancanti è *Naive Bayes*. Questo metodo 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.

4 Cancellazione dei dati

Qualora fosse desiderabile utilizzare modelli che siano intolleranti ai dati mancanti, i valori mancanti devono essere eliminati dai dati. L'approccio più semplice per gestire i valori mancanti consiste nel rimuovere interi predittori e/o campioni che contengono valori mancanti. Tuttavia, è necessario considerare attentamente una serie di aspetti relativi ai dati prima di adottare questo approccio. Ad esempio, i valori mancanti potrebbero essere eliminati rimuovendo tutti i predittori che contengono almeno un valore mancante; allo stesso modo, i valori mancanti potrebbero essere eliminati rimuovendo tutti i campioni che contengono valori mancanti. È necessario considerare però che nessuno di questi approcci sarà appropriato per tutti i dati.

Pertanto per alcuni set di dati, può essere vero che determinati predittori siano molto più problematici di altri, oppure può essere vero che campioni specifici hanno costantemente valori mancanti tra i predittori, per questo motivo rimuovendo tali predittori il problema dei dati mancanti può essere così risolto.

Un'altra considerazione importante è il valore intrinseco dei campioni rispetto ai predittori. Quando è difficile ottenere *sample* o quando i dati contengono un piccolo numero di questi, allora non è desiderabile rimuovere tali campioni dai dati. In generale, i campioni sono più critici dei predittori e una priorità più alta dovrebbe essere posta nel conservare il maggior numero di osservazioni possibili. Per questo motivo una strategia iniziale consisterebbe nell'identificare e rimuovere prima i predittori che hanno una quantità sufficientemente elevata di dati mancanti. Naturalmente, i predittori considerati importanti o che permettono di fare una previsione della variabile risultante, non dovrebbero essere rimossi. Una volta rimossi i predittori problematici, l'attenzione può concentrarsi sui campioni che superano una soglia di mancanza dei dati.

Oltre all'eliminazione dei dati, la preoccupazione principale a causa della rimozione di campioni all'interno del training set è l'influenzabilità del modello, dovuta all'eliminazione di dati, che mette in relazione i predittori con il risultato. Ad esempio, quando i dati mancanti non sono randomici, l'eliminazione di questi all'interno dell'analisi porterebbe erroneamente a sovrastimare o sottostimare la relazione reale presente, giungendo così a delle conclusioni sbagliate.

5 Codifica dei valori mancanti

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. In altri casi, invece, i valori mancanti potrebbero essere semplicemente codificati come "mancanti" o "sconosciuti". Ad esempio, all'interno dello studio

di Kuhn e Johnson [4] viene utilizzato un set di dati in cui l'obiettivo era prevedere l'accettazione o il rifiuto di proposte di sovvenzione. Uno dei predittori categorici era lo sponsor della sovvenzione che assumeva valori quali "sovvenzioni competitive australiane", "centro di ricerca cooperativa", "industria", ecc. In totale, erano presenti circa il 10% di domande di sovvenzione aventi un valore sponsor vuoto. Per consentire l'utilizzo nella modellazione delle applicazioni che avevano uno sponsor vuoto, tali valori sono stati codificati come "sconosciuti". Per molti dei modelli studiati, l'indicatore di uno sponsor sconosciuto è stato uno dei più importanti predittori del successo della sovvenzione, ciò significa che era molto più probabile che una sovvenzione venisse finanziata con successo se il predittore dello sponsor era sconosciuto. Infatti, nel set di formazione il tasso di successo della sovvenzione associato a uno sponsor sconosciuto è stato dell'82,2% rispetto al 42,1% di uno sponsor noto.

Sfortunatamente, è impossibile stabilire se codificare i valori mancanti è una strategia vincente per tutti i dataset. Chiaramente, il meccanismo che ha portato all'identificazione dell'etichetta di sponsor mancante come fortemente associata all'accettazione della sovvenzione è stato davvero importante. Tuttavia, sarebbe problematico accettare questa tipologia di analisi come definitiva ed implicare una simile relazione di causa ed effetto. Un principio guida che può essere utilizzato per determinare se la mancanza di codifica sia una metodologia opportuna, è ad esempio pensare a come i risultati verrebbero interpretati se quell'informazione diventasse importante per il modello.

6 Metodi di imputazione

Un altro approccio alla gestione dei valori mancanti consiste nell'imputarli. L'imputazione utilizza informazioni e relazioni tra i predittori non mancanti per fornire una stima per riempire il valore assente.

Storicamente, i metodi statistici per i dati mancanti si sono occupati dell'impatto sui modelli inferenziali, nei quali le caratteristiche e la qualità della strategia di imputazione si sono concentrate sulle statistiche di test prodotte dal modello. L'obiettivo di queste tecniche è garantire che le distribuzioni statistiche siano trattabili e di qualità sufficiente per supportare i successivi test di ipotesi. L'approccio principale in questo scenario consiste nell'utilizzare più imputazioni. Esistono diverse differenze tra i modelli inferenziali e predittivi che influiscono su questo processo:

- In molti modelli predittivi, non esiste alcuna nozione di ipotesi distributive (o sono spesso intrattabili).
- Molti modelli predittivi sono costosi dal punto di vista computazionale. L'imputazione ripetuta aggraverebbe notevolmente il tempo di calcolo e le spese generali, per questo è consigliabile eseguire questo processo all'interno della fase di ricampionamento.
- Poiché i modelli predittivi sono giudicati in base alla loro capacità di prevedere con precisione campioni ancora da vedere, piuttosto che sull'adeguatezza statistica, è fondamentale che i valori imputati siano il più vicino possibile ai loro valori reali (non osservati).
- L'obiettivo generale dei modelli inferenziali è comprendere a fondo le relazioni tra il predittore e la risposta

per i dati disponibili. Al contrario, l'obiettivo dei modelli predittivi è comprendere le relazioni tra i predittori e la risposta che sono generalizzabili a campioni ancora da vedere. I metodi di imputazione multipli non mantengono il generatore di imputazione dopo che i dati mancanti sono stati stimati, il che rappresenta una sfida per l'applicazione di queste tecniche a nuovi campioni.

Alcune altre caratteristiche importanti che un metodo di imputazione predittivo dovrebbe avere sono:

- All'interno di un campione di dati, potrebbero mancare anche altre variabili. Per questo motivo, un metodo di imputazione dovrebbe essere tollerante nei confronti di altri dati mancanti.
- L'imputazione crea un modello incorporato all'interno di un altro modello. Esiste un'equazione di previsione associata ad ogni predittore nel set di addestramento che potrebbe contenere dati mancanti. È auspicabile che il metodo di imputazione sia veloce e abbia un'equazione di previsione compatta.
- Molti set di dati contengono spesso predittori sia numerici che qualitativi. Piuttosto che generare variabili fittizie per predittori qualitativi, un utile metodo di imputazione sarebbe in grado di utilizzare predittori di vario tipo come input.
- Il modello per la previsione dei valori mancanti dovrebbe essere relativamente (numericamente) stabile e non essere eccessivamente influenzato dagli *outliers*.

L'imputazione pone la domanda: quanti dati mancanti sono troppi da imputare? Una comune *best practice* è considerare come *threshold* il 20% di dati mancanti all'interno di una colonna. Naturalmente, questo dipende dalla situazione e dai modelli di valori mancanti nel set di allenamento. È anche importante considerare che l'imputazione è probabilmente il primo passo in qualsiasi sequenza di pre-elaborazione (precedente anche a qualunque altro passaggio che coinvolge la stima dei parametri).

6.1 K-Nearest Neighbors

Quando il training set è di dimensioni ridotte o moderate, *K-nearest neighbors* può essere un metodo rapido ed efficace per imputare valori mancanti. Questa procedura identifica un campione con uno o più valori mancanti [2] [6]. Quindi identifica i K campioni più simili nei dati di addestramento che sono completi. La somiglianza dei campioni per questo metodo è definita da una metrica di distanza. Quando tutti i predittori sono numerici, la distanza euclidea standard è comunemente usata come metrica di somiglianza. Dopo aver calcolato le distanze, i K campioni più vicini al campione con il valore mancante sono identificati e viene calcolato il valore medio del predittore di interesse. Questo valore viene quindi utilizzato per sostituire il valore mancante del campione.

Nel caso in cui i predittori siano sia numerici che categoriali viene utilizzata la distanza di Gower [3]. Questa metrica utilizza una metrica di distanza specializzata separata sia per i predittori qualitativi che quantitativi. 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, la distanza tra i campioni i e j è definita come

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

dove R_x è l'intervallo del predittore. La misura della distanza viene calcolata per ciascun predittore e la distanza media viene utilizzata come distanza complessiva. Una volta che i K vicini vengono trovati, i loro valori vengono utilizzati per imputare i dati mancanti. La moda viene utilizzata per imputare predittori qualitativi e la media o mediana viene utilizzata per imputare predittori quantitativi. K può essere un parametro di *tuning*, ma i valori intorno a 5-10 sono un valore predefinito ragionevole. Per i dati *scat* degli animali, la prima immagine della **Figura 8** mostra gli stessi dati della **Figura 3** ma con i valori mancanti compilati utilizzando 5 vicini in base alla distanza di Gower. I nuovi valori cadono per lo più intorno alla periferia di queste due dimensioni, ma sono all'interno della gamma dei campioni con dati completi.

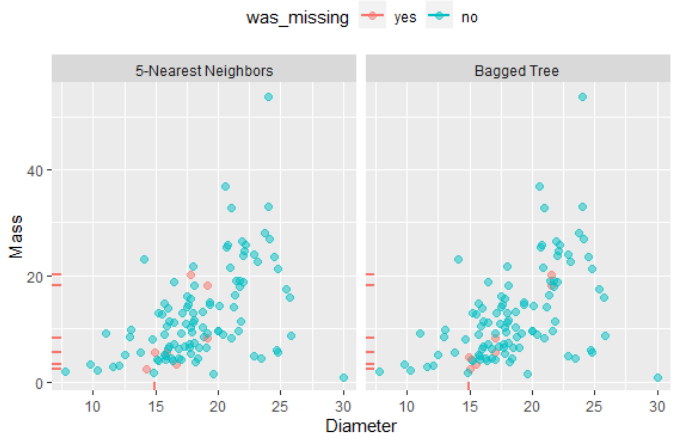

Figura 8: Un confronto dei *K-Nearest Neighbors* e le tecniche di imputazione dei *bagged tree* per i dati di *scat* animale. I trattini sui due assi mostrano dove si sono verificati i valori di quella variabile quando mancava il predittore sull'altro asse.

6.2 Trees

I modelli basati su alberi sono una scelta ragionevole per una tecnica di imputazione poiché un albero può essere costruito in presenza di altri dati mancanti. Inoltre, gli alberi hanno generalmente una buona precisione e non estrapolano valori oltre i limiti dei dati di addestramento.

Un singolo albero è noto per produrre risultati che hanno una bassa distorsione ma un'alta varianza. Gli insiemi di alberi, tuttavia, forniscono un'alternativa a bassa varianza. Le *Random Forests* sono una di queste tecniche. Tuttavia, ci sono un paio di notevoli svantaggi quando si utilizza questa tecnica in un ambiente di modellazione predittiva:

1. la selezione casuale di predittori in ogni divisione richiede un gran numero di alberi (da 500 a 2000) per ottenere un modello stabile e affidabile;

2. ognuno di questi alberi non è potato e il modello risultante di solito ha una grande impronta computazionale.

Una buona alternativa che ha un ingombro computazionale più piccolo è un *bagged tree* che è costruito in modo simile a una foresta casuale. La differenza principale è che in un modello di questo tipo, tutti i predittori vengono valutati ad ogni divisione in ogni albero. Le prestazioni di un *bagged tree*, utilizzando 25-50 alberi, sono confrontabili con le prestazioni di un modello che utilizza le *random forests*. Il secondo grafico della **Figura 8** illustra i valori imputati per i dati *scat* utilizzando un insieme di 50 alberi.

6.3 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.

Per esempio, analizzando i dati sull'utenza dei treni di Chicago si può notare come il ritardo di 14 giorni nell'utenza all'interno di una fermata è altamente correlato con l'utenza del giorno corrente. La **Figura 9** mostra la relazione tra questi predittori per la fermata di Clark/Lake. La maggior parte dei dati mostra una relazione lineare tra questi predittori, con una manciata di giorni che hanno valori lontani dalla tendenza generale. Ovviamente includere le vacanze come predittore nel modello robusto contribuirebbe a migliorare l'imputazione.

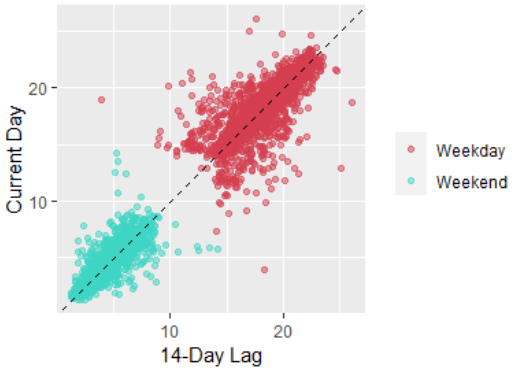

Figura 9: Ritardo di due settimane nell'utenza giornaliera rispetto all'utenza giornaliera per la stazione Clark/Lake con le festività statunitensi comuni escluse e colorate per parte della settimana.

Il concetto di imputazione lineare può essere esteso a dati ad alta dimensionalità.

7 Casi speciali

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").

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

la censura di sinistra. Ad esempio, nelle misurazioni di laboratorio lo strumento di misura potrebbe non quantificare in modo affidabile valori inferiori ad una soglia X. Quando un predittore ha valori inferiori alla soglia, tali valori sono solitamente riportati come "$< X$". Quando questi dati devono essere inclusi in un modello predittivo, è necessario capire come affrontare la gestione di valori censurati. Una pratica ampiamente accettata consiste nell'utilizzare il valore limite inferiore di X come risultato. Tuttavia i valori censurati influiscono sulle metriche che misurano la variabilità. In particolare, la variabilità sarà sottostimata.

Per mitigare il problema della variabilità, i valori censurati a sinistra possono essere imputati utilizzando valori uniformi casuali compresi tra zero e X, oppure utilizzare altri schemi di assegnazione di valori casuali che rappresentano meglio la distribuzione (se nota). Sebbene l'imputazione in questo modo aggiunga rumore casuale ai dati, è probabile che sia preferibile ai potenziali problemi di *overfitting* che possono verificarsi assegnando un valore di X ai dati.

8 Conclusioni

I valori mancanti sono occorrenze comuni nei dati. Sfortunatamente, la maggior parte delle tecniche di modellazione predittiva non è in grado di gestire i valori mancanti. Pertanto, questo problema deve essere risolto prima della modellazione. Uno dei modi migliori per comprendere la quantità e la natura dei valori mancanti è attraverso una visualizzazione appropriata. Una volta nota la gravità dei valori mancanti, è necessario prendere una decisione su come trattare questi valori, cioè se procedere ad eliminarli oppure imputarli.

References

[1] J. Friedman R. Olshen Breiman, L. and C. Stone. *Classification and Regression Trees.* New York: Chapman; Hall., 1984.

[2] Bianca NI Eskelson, Hailemariam Temesgen, Valerie Lemay, Tara M Barrett, Nicholas L Crookston, and Andrew T Hudak. The roles of nearest neighbor methods in imputing missing data in forest inventory and monitoring databases. *Scandinavian Journal of Forest Research*, 24(3):235–246, 2009.

[3] John C Gower. A general coefficient of similarity and some of its properties. *Biometrics*, pages 857–871, 1971.

[4] M Kuhn and K Johnson. *Applied Predictive Modeling.* Springer., 2013.

[5] R. Quinlan. *C4.5: Programs for Machine Learning.* Morgan Kaufmann Publishers., 1993.

[6] Gerhard Tutz and Shahla Ramzan. Improved methods for the imputation of missing data by nearest neighbor methods. *Computational Statistics & Data Analysis*, 90:84–99, 2015.