Valores Ausentes

Definición

Los valores ausentes, también conocidos como datos faltantes o missing values, son aquellos elementos de un conjunto de datos que no tienen un valor registrado.

Los valores ausentes pueden surgir por diversas razones a lo largo del proceso de recolección, almacenamiento y análisis de datos. Los más comunes son:

- **Errores de Medición o Recolección de Datos**: Los errores humanos, intencionados o no, o técnicos durante la recolección o entrada de datos pueden generar valores ausentes.
- **Problemas de Acceso**: Los datos pueden ser intencionalmente omitidos por razones de privacidad o confidencialidad, generando en lugar de estos, valores ausentes.
- **Problemas de Registro y Almacenamiento**: Los datos pueden perderse durante el almacenamiento debido a errores de base de datos o fallos de hardware o al migrar datos entre diferentes sistemas o formatos.

Tipos de Valores Ausentes

Los valores ausentes se pueden categorizar en diferentes tipos según el mecanismo que los genera. Comprender estos tipos es crucial para decidir cómo manejarlos adecuadamente en un análisis de datos. A continuación, se describen los tres principales tipos de valores ausentes:

Completamente al Azar (MCAR - Missing Completely at Random)

En este caso, la ausencia de datos no está relacionada ni con las variables observadas ni con las no observadas. Es decir, los valores ausentes son puramente aleatorios y no tienen un patrón identificable.

Ejemplo: En una encuesta, algunas respuestas se pierden debido a un error de transmisión aleatorio que no está relacionado con ninguna característica del encuestado.

Implicación: Si los datos son MCAR, los análisis realizados en el subconjunto de datos completos no estarán sesgados debido a la ausencia de datos.

Aleatorios Condicionales (MAR - Missing at Random)

Aquí, la probabilidad de que un valor esté ausente está relacionada con las variables observadas, pero no con el valor de la variable ausente en sí. Es decir, la ausencia de datos puede depender de otras características conocidas de la muestra.

Ejemplo: En un estudio de salud, es más probable que las personas mayores omitan responder preguntas sobre el uso de tecnología digital. La ausencia de datos sobre el uso de tecnología está relacionada con la edad, una variable observada.

Implicación: Si los datos son MAR, se pueden usar técnicas de imputación que tomen en cuenta las variables observadas para estimar los valores ausentes y reducir el sesgo en el análisis.

No al Azar (MNAR - Missing Not at Random)

En este caso, la ausencia de datos está relacionada con el valor de la variable ausente en sí misma. Es decir, hay un patrón sistemático en los datos faltantes que está directamente relacionado con las características no observadas.

Ejemplo: En una encuesta de ingresos, las personas con ingresos muy altos o muy bajos pueden ser menos propensas a revelar sus ingresos. La ausencia de datos sobre los ingresos depende del valor del ingreso en sí.

Implicación: Si los datos son MNAR, los métodos estándar de imputación pueden no ser efectivos y el análisis puede estar sesgado. Se requiere un enfoque más complejo, como modelos específicos que consideren el mecanismo de la ausencia.

Identificación de Tipo de Valor Ausente

MCAR

- **Imputación**

  Realizar imputación múltiple para manejar los datos faltantes pueden ser usados para detectar si los datos faltantes son MCAR al observar el impacto que la imputación tiene en el conjunto de datos. La imputación múltiple genera varios conjuntos de datos imputados y permite evaluar la variabilidad entre ellos. Si la imputación no cambia la distribución y/o las estadísticas de manera consistente en los diferentes conjuntos de datos imputados esto sugiere que los datos faltantes pueden ser MCAR.

- **Mapas de Calor**

  Los mapas de calor pueden ser usados para detectar si los datos faltantes son MCAR al identificar si pudieran existir patrones en o entre estos. Esto se hace representando de un color los valores ausentes y de otro los valores no ausentes en una matriz que represente todas las variables y sus valores de un conjunto de datos. Si los valores faltantes no muestran patrones específicos y están distribuidos aleatoriamente sugiere que los datos pueden ser MCAR. Si hay agrupaciones, bien sea por fila o por columna, o tendencias, por ejemplo, los valores faltantes de la columna 'A' tienden a ocurrir junto con valores faltantes en la columna 'B' esto sugiere que los datos faltantes no son MCAR.

- **Análisis de Relación**

  Las medidas de relación pueden ser usadas para detectar si los datos faltantes son MCAR al identificar si pudiera existir una relación entre la variable que contiene los valores faltantes a estudiar y otras variables del conjunto de datos. Esto se puede hacer creando una nueva variable binaria a partir de la variable con valores ausentes donde el 1 represente al valor faltante y el 0 a un valor no faltante y, una vez construida esta nueva variable, calcular la medida de relación oportuna, según los tipos de variables a comparar, y, si la medida de relación no es significativa, esto sugiere que los datos faltantes pueden ser MCAR.

- **Test de Little**

  El test de Little, también conocido como Little's MCAR Test, es una prueba estadística diseñada para evaluar si los datos faltantes en un conjunto de datos son Missing Completely at Random (MCAR) a través del cálculo de un estadístico chi-cuadrado basado en las diferencias entre las medias y las covarianzas de los grupos. Este test plantea la siguiente hipótesis:

- **Hipótesis nula (H0)**: Los datos están faltando completamente al azar (MCAR).
- **Hipótesis alternativa (H1)**: Los datos no están faltando completamente al azar (no MCAR).

A través de comparar el estadístico calculado con su valor crítico o a través del valor p se puede obtener un resultado.

Para más detalles sobre la prueba de Little acceder al anexo, en este mismo directorio, llamado "Anexo_Prueba-Little.pdf".

- **Resumen**

Método	Ventajas	Limitaciones
Imputación	Permite evaluar la variabilidad entre conjuntos imputados; robustez en análisis	Complejidad; requiere conocimientos avanzados; puede ser intensivo en recursos computacionales
Mapas de Calor	Visualización clara y directa; fácil identificación de patrones	Subjetividad, no proporciona medida estadística
Análisis de Relación	Detección de relaciones significativas entre valores faltantes y otras variables	equiere análisis adicional; la naturaleza discreta de las variables indicadoras puede limitar el uso de ciertas técnicas de relación
Test de Little	Rigor estadístico, interpretación clara	Complejidad de implementación, sensibilidad al tamaño de muestra, limitado a evaluación de MCAR

MAR

- **Imputación**

  Realizar imputación múltiple para manejar los datos faltantes pueden ser usados para detectar si los datos faltantes son MAR al observar el impacto que la imputación tiene en el conjunto de datos. La imputación múltiple genera varios conjuntos de datos imputados y permite evaluar la variabilidad entre ellos. Para datos MAR, se espera que la variabilidad se observe entre el conjunto de datos original y los conjuntos de datos imputados, pero no tanto entre los diferentes conjuntos de datos imputados.

- **Mapas de Calor**

  Los mapas de calor pueden ser usados para detectar si los datos faltantes son MAR al identificar si pudieran existir patrones en o entre estos. Esto se hace representando de un color los valores ausentes y de otro los valores no ausentes en una matriz que represente todas las variables de un conjunto de datos. Si hay agrupaciones por fila, o tendencias, por ejemplo, los valores faltantes de la columna 'A' tienden a ocurrir junto con valores faltantes en la columna 'B' esto sugiere que los datos faltantes pueden ser MAR.

- **Medidas de Relación**

  Las medidas de relación pueden ser usadas para detectar si los datos faltantes son MAR al identificar si existe una relación entre la variable con valores faltantes a estudiar y otras variables del conjunto de datos. Esto se puede hacer creando una nueva variable binaria a partir de la variable con valores ausentes donde el 1 represente al valor faltante y el 0 a un valor no faltante y, una vez construida esta nueva variable, calcular la medida de relación oportuna, según los tipos de variables a comparar, y, si la medida de relación es significativa, esto sugiere que los datos faltantes pueden ser MAR.

- **Modelos Estadísticos**

  Se pueden utilizar modelos estadísticos de clasificación para detectar si los datos faltantes son MAR. Al igual que en las medidas de relación, se creará una variable a partir de la variable con valores

ausentes donde el 1 represente al valor faltante y el 0 a un valor no faltante. Una vez construida esta variable se usará como variable dependiente y las demás variables del conjunto de datos como independientes. Si se puede predecir, con relativa eficacia, cuando aparecerá un valor faltante a través de las otras variables esto sugiere que los datos faltantes pueden ser MAR.

Los modelos más usados para esta tarea son:

- Regresión Logística
- Árboles de Decisión
- Bosques Aleatorios

- **Resumen**

Método	Ventajas	Limitaciones
Imputación	Permite evaluar la variabilidad entre conjuntos imputados; robustez en análisis	Complejidad; requiere conocimientos avanzados; puede ser intensivo en recursos computacionales
Mapas de Calor	Visualización clara y directa; fácil identificación de patrones	Subjetividad en la interpretación; no proporciona medidas estadísticas
Análisis de Relación	Detección de relaciones significativas entre valores faltantes y otras variables	Requiere análisis adicional; la naturaleza discreta de las variables indicadoras puede limitar el uso de ciertas técnicas de relación
Modelos Estadísticos	Rigor estadístico; permite modelar relaciones complejas; proporciona predicciones claras	Complejidad en implementación; puede ser computacionalmente intensivo; requiere supuestos específicos del modelo

MNAR

- **Imputación**

  Realizar imputación múltiple para manejar los datos faltantes pueden ser usados para detectar si los datos faltantes son MAR al observar el impacto que la imputación tiene en el conjunto de datos. La imputación múltiple genera varios conjuntos de datos imputados y permite evaluar la variabilidad entre ellos. Para datos MNAR, se espera que la variabilidad se observe tanto entre el conjunto de datos original y los conjuntos de datos imputados, como entre los diferentes conjuntos de datos imputados.

- **Comparación de Distribuciones**

  Comparar la distribución de los valores observados con la distribución esperada puede ayudar a identificar si los datos faltantes son MNAR (Missing Not At Random). Si los datos faltantes dependen del valor faltante en sí mismo, la distribución de los valores observados puede diferir significativamente de la distribución esperada de los datos completos.

  La distribución esperada se refiere a una distribución teórica que representa los datos completos si no hubiera valores faltantes. Se puede generar de varias formas dependiendo del conocimiento previo sobre los datos y las suposiciones que se pueden hacer sobre su distribución. Comúnmente, se asume que los datos siguen una distribución normal, pero también se pueden usar otras distribuciones basadas en la naturaleza de los datos. Para generar una distribución esperada, se suelen utilizar las estadísticas descriptivas de los datos observados, como la media y la desviación estándar, y se asume una forma de distribución para los datos completos.

  Para evaluar la similitud entre la distribución de los valores observados y la distribución esperada se pueden usar visualizaciones conjuntas

- **Prueba de Kolmogorov-Smirnov**

  El test de Kolmogorov-Smirnov es una prueba estadística no paramétrica que se utiliza para determinar si una muestra proviene de una distribución teórica específica, en el contexto de MNAR se usa la distribución esperada. Es una herramienta útil para comparar la distribución observada de una muestra con una distribución teórica como la normal a través del cálculo de un estadístico basado en la distancia entre la CDF de cada función. Este test plantea la siguiente hipótesis:

  - **Hipótesis Nula (H_0):** La muestra proviene de la distribución teórica especificada.
  - **Hipótesis Alternativa (H_1):** La muestra no proviene de la distribución teórica especificada.

  A través de comparar el estadístico calculado con su valor crítico o a través del valor p se puede obtener un resultado. Si se rechaza la hipótesis nula esto sugiere que los datos pueden ser MNAR.

  Para más detalles sobre la prueba de Little acceder al anexo, en este mismo directorio, llamado "Anexo_Prueba-KS.pdf".

- **Resumen**

Método	Ventajas	Limitaciones
Imputación	Permite evaluar la variabilidad entre conjuntos imputados; robustez en análisis	Complejidad; requiere conocimientos avanzados; puede ser intensivo en recursos computacionales
Comparación de Distribuciones	Visualización clara y directa; fácil identificación de patrones	Requiere conocimiento previo sobre la distribución esperada; puede no proporcionar una prueba estadística formal
Prueba de Kolmogorov-Smirnov	Proporciona una medida estadística formal; fácil de implementar con software estadístico	Sensible al tamaño de la muestra; suposiciones sobre la distribución teórica pueden no ser precisas

Métodos para Tratar Valores Ausentes

- **Eliminación de Datos Faltantes**

  - **Descripción**: Consiste en eliminar las filas o columnas que contienen valores ausentes.
  - **Cuándo Usarlo**
    - Cuando el porcentaje de datos faltantes es pequeño (MCAR).
    - Cuando la eliminación de filas o columnas no introducirá sesgo (MCAR).
  - **Ventajas**
    - Simple y fácil de implementar.
    - No introduce nuevos datos en el conjunto de datos.
  - **Desventajas**
    - Puede llevar a la pérdida de información valiosa.
    - No es adecuado si los datos faltantes no son aleatorios.
    - Puede distorsionar los resultados si el porcentaje de datos faltantes es grande.

- **Imputación Simple**

  - **Descripción**: Consiste en rellenar los valores ausentes con una única estimación, como la media, mediana, moda o un valor constante.
  - **Cuándo Usarlo**
    - Cuando el porcentaje de datos faltantes es bajo (MCAR, MAR).
    - Cuando se supone que los datos faltantes son MCAR o MAR.

    - **Ventajas**
        - Fácil de implementar.
        - Preserva el tamaño del conjunto de datos.
    - **Desventajas**
        - No refleja la incertidumbre sobre los valores faltantes.
        - Puede introducir sesgo si los datos no son MCAR o MAR.
        - Puede reducir la variabilidad en los datos.

Imputación Múltiple

- **Descripción**: Consiste en crear múltiples conjuntos de datos imputados, analizar cada uno por separado y luego combinar los resultados.
- **Cuándo Usarlo**
    - Cuando el porcentaje de datos faltantes es significativo (MAR).
    - Cuando se supone que los datos faltantes son MAR.
- **Ventajas**
    - Refleja la incertidumbre sobre los valores faltantes.
    - Proporciona estimaciones más precisas y robustas.
    - Mantiene la variabilidad en los datos.
- **Desventajas**
    - Complejo de implementar.
    - Requiere más recursos computacionales.
    - Puede ser intensivo en tiempo.
- **Imputación Basada en Regresión**
    - **Descripción**: Utiliza modelos de regresión para predecir y rellenar los valores faltantes basándose en otras variables observadas.
    - **Cuándo Usarlo**:
        - Cuando se dispone de variables predictoras que se correlacionan con los datos faltantes (MAR).
        - Cuando se supone que los datos faltantes son MAR.
    - **Ventajas**
        - Aprovecha la información de otras variables.
        - Proporciona imputaciones más precisas que la imputación simple.
    - **Desventajas**
        - Asume relaciones lineales entre las variables.
        - Puede no ser adecuado si las relaciones no son lineales.
- **Imputación Estocástica**
    - **Descripción**: Reemplazar los valores faltantes con valores generados aleatoriamente dentro de una distribución estimada de los datos.
    - **Cuándo Usarlo**
        - Cuando se quiere mantener la variabilidad de los datos (MAR).
    - **Ventajas**
        - Mantiene la variabilidad y distribución de los datos originales.
    - **Desventajas**
        - Puede introducir variabilidad adicional no deseada.
        - Complejo de implementar.
- **Modelos Basados en Árboles (Random Forests)**

- **Descripción**: Usa árboles de decisión para imputar valores faltantes basándose en el resto de los datos. Random Forests puede ser particularmente útil para capturar relaciones no lineales entre las variables.

- **Cuándo Usarlo**
  - Cuando se dispone de suficientes datos para entrenar un modelo (MNAR).
  - Cuando los datos faltantes son MNAR y se desea capturar relaciones complejas entre las variables.
- **Ventajas**
  - Puede manejar relaciones complejas y no lineales.
  - No requiere suposiciones sobre la distribución de los datos.
- **Desventajas**
  - Complejo y puede ser intensivo en recursos computacionales.
  - Puede sobreajustarse si no se maneja adecuadamente.

- **Imputación por Vecinos Cercanos (KNN Imputation)**

  - **Descripción**: Imputa valores faltantes basándose en la similitud con otros puntos de datos. KNN utiliza los valores de los vecinos más cercanos para imputar los valores faltantes.
  - **Cuándo Usarlo**
    - Cuando los datos faltantes son MNAR y se desea usar la información local de los datos (MNAR).
    - En situaciones donde se dispone de un conjunto de datos con alta densidad.
  - **Ventajas**
    - Fácil de entender e implementar.
    - Aprovecha la similitud local entre las observaciones.
  - **Desventajas**
    - Puede ser intensivo en recursos computacionales para conjuntos de datos grandes.
    - No es adecuado si hay pocos vecinos cercanos.

Resumen en Tabla

Método	Descripción	Cuándo Usarlo	Ventajas	Desventajas
Eliminación de Datos Faltantes	Eliminar filas o columnas con valores ausentes	Cuando el porcentaje de datos faltantes es pequeño (MCAR)	Simple, fácil de implementar; no introduce nuevos datos	Pérdida de información; puede distorsionar resultados si hay muchos datos faltantes
Imputación Simple	Rellenar valores ausentes con media, mediana, moda o un valor constante	Cuando el porcentaje de datos faltantes es bajo (MCAR, MAR)	Fácil de implementar; preserva el tamaño del conjunto de datos	No refleja incertidumbre; puede introducir sesgo; reduce variabilidad
Imputación Múltiple	Crear múltiples conjuntos de datos imputados	Cuando el porcentaje de datos faltantes es significativo (MAR)	Refleja incertidumbre; proporciona estimaciones más precisas; mantiene variabilidad	Complejo de implementar; requiere más recursos computacionales
Imputación Basada en Regresión	Usar modelos de regresión para predecir valores faltantes	Cuando hay variables predictoras correlacionadas con los datos faltantes (MAR)	Aprovecha información de otras variables; más precisa que imputación simple	Asume relaciones lineales; no adecuado si las relaciones no son lineales

Método	Descripción	Cuándo Usarlo	Ventajas	Desventajas
Imputación Estocástica	Reemplazar valores faltantes con valores generados aleatoriamente dentro de una distribución estimada	Cuando se quiere mantener la variabilidad de los datos (MAR)	Mantiene la variabilidad y distribución de los datos originales	Puede introducir variabilidad adicional no deseada; complejo de implementar
Modelos Basados en Árboles (Random Forests)	Usar árboles de decisión para imputar valores faltantes	Cuando hay suficientes datos y relaciones complejas entre variables (MNAR)	Maneja relaciones complejas y no lineales; no requiere suposiciones de distribución	Complejo; intensivo en recursos computacionales; riesgo de sobreajuste
Imputación por Vecinos Cercanos (KNN Imputation)	Imputar basándose en la similitud con otros datos	Cuando hay alta densidad de datos y se desea usar información local (MNAR)	Fácil de entender e implementar; aprovecha similitud local	Intensivo en recursos para grandes conjuntos de datos; no adecuado con pocos vecinos