Conclusión: cuándo usar qué método?

#### ¿Qué método de imputación de datos faltantes debo usar y cuándo?

Los diferentes métodos de imputación hacen diferentes suposiciones y tienen diferentes ventajas y desventajas. Visita el artículo anterior, "Descripción general de los métodos de imputación de valores faltantes", para obtener más detalles al respecto.

****Como guía te diríamos:****

1. Si hay valores faltantes en menos del 5% de las observaciones:

A - reemplazar por media / mediana o muestra aleatoria (numérica)

B - reemplazar por la categoría más frecuente o muestra aleatoria (categórica)

2 - Si hay valores faltantes en más del 5% de las observaciones:

A - hacer una por imputación la media / media y agregar una variable binaria para capturar NA (numérico)

B - agregar una etiqueta de 'Falta' o ‘Missing’ en variables categóricas (categóricas)

La clave es investigar si los datos faltantes tienen valor predictivo. Si lo hacen, es mejor tratar de diferenciarlos de las observaciones restantes. Si no lo hacen, reemplazarlos por los valores más frecuentes es una buena opción.

****Excepciones a lo anterior:****

1 - Si vos o tu organización sospechan que no faltan datos al azar y no desean atribuir la ocurrencia más común a NA

2 - Si no deseas aumentar el espacio de los datos agregando variables adicionales para capturar NA

En estos casos, reemplaza por un valor en el extremo más alejado de la distribución o un valor arbitrario.

****Nota final****  
  
La imputación de datos faltantes para competencias de datos o en entornos empresariales puede abordarse de manera diferente. En las competiciones de datos, un pequeño aumento en el rendimiento puede ser la diferencia entre el primer o segundo lugar. Por lo tanto, es posible que desees probar todos los métodos de ingeniería de variables y selecciones el que le brinde al algoritmo de aprendizaje automático el mejor rendimiento. Los diferentes métodos de imputación de datos faltantes pueden ayudar a diferentes modelos a hacer mejores predicciones.  
  
En las empresas u otras organizaciones, los científicos de datos no suelen tener tiempo para realizar estudios largos y, por lo tanto, pueden optar por simplificar la canalización de ingeniería de variables. En estos casos, es una práctica común seguir las pautas anteriores, teniendo en cuenta las excepciones, y realizar el mismo procesamiento para todas las funciones.  
  
Es posible que este procesamiento previo optimizado no conduzca a la "mejor" de las variales predictivas posibles luego de la transformación, pero facilita la entrega de modelos de aprendizaje automático en un tiempo más corto. Por lo tanto, el negocio puede comenzar a disfrutar el poder del aprendizaje automático.