Universidad Francisco Marroquín
Data Wrangling
Luis Javier Samayoa

Laboratorio #8 - Missing Data and Feature Engineering

Parte 1: (70%)
La tabla "titanic_MD.csv" contiene missing values en varias columnas. Utilizando R o Python, realice lo siguiente:

1. **Reporte detallado de missing data para todas las columnas. (5%)**
   En la base de datos proporcionada, titulada como titanic_MD encontramos un total de 183 individuos categorizados y analizados en 12 diferentes variables. De dichas variables 6 contienen missing values.
   Sex: La variable genero contiene 51 missing values de 183, lo que representa que un 28% de la información de esa variable no se registro. Esta variable representaba sus valores faltantes con el símbolo "?". Es una variable categorica
   Age: la variable edad cuenta con 25 espacios vacíos lo que es un 14% de la data en esa columna. Es una variable numérica.
   SibSp: Esta columna representa la cantidad de hermano o pareja que tenía abordo y contiene 3 missing values, lo que significa el 1.63% de la información perdida para el estudio de la variable. Es una variable numérica.
   Parch: En esta columna se describe la cantidad de hijos o padres que tenia abordo, a la cual le faltan 12 observaciones por lo que falta el 6.55% de la data correspondiente. Es una variable numérica.
   Fare: Existe un faltante de 8 observaciones en esta columna que representa la tarifa con la que abordo, lo que representa 4.37% de la información requerida. Es una variable numérica.
   Embarked: Al igual que la variable parch tiene 12 observaciones faltantes que representan el 6.55%. es una variable de formato carácter y categorica.

2. **<u>Para cada columna</u> especificar qué tipo de modelo se utilizará (solo el nombre y el porqué) y qué valores se le darán a todos los missing values. (Ej. Imputación sectorizada por la moda, bins, y cualquier otro método visto anteriormente). (10%)**
   Sex: Standarizacion
   Age: imputacion sectorizada por media
   SibSp: Imputacion sectorizada por moda
   Parch: Imputación sectorizada por moda
   Fare: Modelo de predicción lineal
   Embarked: Modelo de predicción lineal
3. **Reporte de qué filas están completas (5%)**
   PassengerId: es una variable de carácter numérico descriptivo, la cual representa el código de pasajero en el barco, todos los valores son únicos. Variable Categórica.

Survived: Se representa con un valor 1 y 0; en el cual 1 es sobrevivió y el valor 0 es representativo de murió. Variable categórica.
Pclass: Esta variable se representa con números para el nivel de clase que tenia dentro del barco. 1 respresenta primera clase, 2 representa clase media y el 3 una clase económica. Variable Categórica.
Name: variable de caracteres que describe el nombre de los individuos.
Ticket: variable de combinación entre caracteres y números para representar el código de ticket.
Cabin: El número de cabina en el que se hospedaba, combinación de carácter y numero.

4. **Utilizar los siguientes métodos para cada columna que contiene missing values: (50%)**
    a. Imputación general (media, moda y mediana)
    b. Modelo de regresión lineal
    c. Outliers: Uno de los dos métodos vistos en clase (Standard deviation approach o Percentile approach)
5. **Al comparar los métodos del inciso 4 contra "titanic.csv", ¿Qué método (para cada columna) se acerca más a la realidad y por qué? (20%)**
   Sex: El mejor método a utilizar es el de imputación de moda ya que es mas exacto con variables categóricas.
   Age:
   SibSp: El modelo de desviación estándar que permite ver que tan dispersos son los valores y definir mejor la cantidad de parientes.
   Parch: El modelo de desviación estándar al igual que para SibSp porque permite ver que tan dispersos son los valores y definir mejor la cantidad de parientes.
   Fare: Regresión lineal, comparando los valores de tarifa junto a la correlación que existe con la variable con otra de alta correlación permite que se ajuste de la mejor manera la predicción del modelo.
   Embarked: Imputación de moda es el modelo mas confiable para esta predicción, fundamentando que es una variable categórica.

6. **Conclusiones (10%)**
   Es fundamental entender las variables de la base de dato, que significan los números, letras y símbolos. Adicionalmente de preparar la data para empezar a trabajar con ella, es fundamental analizar las estadísticas que tienen que se tienen desde un inicio así luego se puede encontrar el modelo adecuado para cambiar los valores faltantes o para eliminar dichas observaciones incompletas. Es fundamental entender como trabajar las variables categóricas y las variables numéricas para la elección del modelo, así como también la correlación que pueden existir entre las diferentes variables.

Parte 2: (30%)
Utilizando la misma tabla de "titanic_MD.csv" en R o en Python realice lo siguiente:

1. Luego del pre-procesamiento de la data con Missing Values, normalice las columnas numéricas por los métodos: **(50%)**
    a. Standarization
    b. MinMaxScaling
    c. MaxAbsScaler
2. Compare los estadísticos que considere más importantes para su conclusión y compare contra la data completa de “titanic.csv” (deberán de normalizar también). **(50%)**

El laboratorio deberá de ser entregado por medio de MiU a más tardar el Domingo, 21 de Noviembre a las 11:59pm. No estaremos aceptando entregas tarde ni por correo electrónico. La entrega será el link al documento en GitHub, en formato markdown o PDF, estén trabajando en R o en Python.