Katherine García

20190418

Laboratorio 8

Parte 1

1. Missing data

* En el dataframe de titanic\_MD hay 6 columnas con missing values. Los cuales son; Age, SibSp, Parch, Fare, Embarked, Sex2.
* Sex2 es una columna creada para los pasajeros que tenían un ‘?’ en la especificación de género.

Text

Description automatically generated

1. Modelo a usar para los missing values

* Edad
  + Regresión Lineal
  + Ya que la edad pudo afectar en la supervivencia de la persona, se buscará hacer de manera mas exacta con este método. Se tomara en cuenta si sobrevivió, la tarifa y la clase del ticket.
* SibSp
  + Imputacion de la moda
  + Al tener unicamente dos resultados en esta columna y pocos datos faltantes se puede colocar el valor mas repetido sin sesgar demasiado la data.
* Parch
  + Imputacion de la moda
  + De igual manera, al tener unicamente dos resultados en esta columna y pocos datos faltantes se puede colocar el valor mas repetido sin sesgar demasiado la data.
* Tarifa
  + Imputación del promedio
  + Se utilizará este modelo debido a los pocos datos con los que no se cuenta y que no se desea tener tarifas que sesguen demasiado los datos, sino que se mantengan dentro de el rango.
* Embarque
  + Imputación de la moda
  + Esta variable no se considera una que sesgue de manera significativa el hecho de si la persona sobrevivió o no. Por lo que se rellenarán los valores con la moda.
* Género
  + Regresion lineal
  + Debido a que el sexo pudo afectar en la supervivencia, se tomará en cuenta la clase y si sobrevivió o no.

1. Filas completas

* Las filas que cuentan con los datos completos son PassengerId, Survived, Pclass, Name, Ticket, Cabin.
* Estas columnas se utilizarán de esta manera, sin modificaciones.

1. Resultados de los modelos

* Se creó un dataframe con las columnas que se encontraban llenas y luego se llenó con los datos ya modificados por los modelos.
* La columna edad se trabajo por medio de imputación de promedio y regresión lineal.

A picture containing text, battery, plaque

Description automatically generated

1. Comparación con los datos originales

Se realizó una comparación sobre los datos modificados y los originales. Se presenta la cantidad de datos que fueron acertados y los que no.

Text

Description automatically generated with low confidence

Las columnas que menos acertaron fueron Tarifa y Edad con ambos modelos de manejo.

Se considera que la cantidad de aciertos erroneos que los modelos generaron es debido a que estas dos casillas presentan números que varían fuertemente.

La columna de edad, con regresión lineal, a pesar de tener la misma cantidad de datos erroneos que los missing values, se mantiene al lado del promedio de los datos originales.

1. Conclusiones

* Dependiendo del tipo de variable, categorica o numerica, las predicciones con regresión lineal pueden ser más o menos acertadas.
* Por ejemplo, para predecir la edad, no se cuentan con muchas variables que tengan una correlacion alta con esta variable. Lo que provoca que la predicción no sea muy confiable.
* En otras variables como género, sibSp o embarque, la certeza de los modelos que se utilizaron fue mejor ya que los valores no eran muchos y no había mucho margen de error.
* Considero que, no existe una manera perfecta para lidiar con missing data. Depende de la percepción de la persona y su enfoque en los datos. Ya que todo esto afecta sobre qué variables son más importantes para el tipo de análisis que se esté conduciendo. Así que, el mejor modelo es el que se adapte mejor al tipo de variable e investigación que se esté haciendo.

Parte 2

1. Normalice las columnas numéricas

* Edad
* Tarifa

1. Comparación contra la data original normalizada