Primer trabajo práctico

Computación Científica Actuarial - Docente: Rodrigo Del Rosso - Colaboradores: Santiago Silva - Joaquín Auza

Fecha de entrega: 01/11/2019

Consideraciones generales

El objetivo de este primer trabajo práctico es simular una situación de la vida real en la que se solicita analizar un conjunto de datos y posteriormente entregar un reporte exponiendo los resultados. **Es en base a esa idea que se solicita, aparte de entregar el script con los comandos en R, un muy breve informe describiendo las variables del dataset, comentando toda información que considere relevante acerca de las mismas. Y lo mismo para el modelo de regresión que se solicita al final.**

- No olviden que tienen el grupo para hacer todas las consultas que consideren necesarias tanto entre ustedes como a nosotros.
- Pueden usar cualquier paquete o diseñar cualquier función adicional que consideren necesario, siempre indicando el uso de los mismos.
- La fecha de entrega es inclusive. Tienen hasta las 23:59 de ese día para entregar el trabajo.

Descripcción del dataset:

UFC Fight Data

El dataset contiene información acerca de todos los encuentros disputados en la UFC desde 1993 hasta 2019, donde cada encuentro es una fila distinta. Cuenta con un total de 5144 registros y 145 variables. Algunas de esas variables son inherentes a cada participante (datos del participante de la esquina roja o azul), mientras que otros son datos referentes a la pelea en sí (fecha, referee, ganador del encuentro, duración del mismo, etc.) A continuación se muestran las primeras diez columnas de las primeras diez filas:

R_fighter	B_fighter	Referee	date	location	Winner
Henry Cejudo	Marlon Moraes	Marc Goddard	2019-06-08	Chicago, Illinois, USA	Red
Valentina Shevchenko	Jessica Eye	Robert Madrigal	2019-06-08	Chicago, Illinois, USA	Red
Tony Ferguson	Donald Cerrone	Dan Miragliotta	2019-06-08	Chicago, Illinois, USA	Red
Jimmie Rivera	Petr Yan	Kevin MacDonald	2019-06-08	Chicago, Illinois, USA	Blue
Tai Tuivasa	Blagoy Ivanov	Dan Miragliotta	2019-06-08	Chicago, Illinois, USA	Blue
Tatiana Suarez	Nina Ansaroff	Robert Madrigal	2019-06-08	Chicago, Illinois, USA	Red
Aljamain Sterling	Pedro Munhoz	Marc Goddard	2019-06-08	Chicago, Illinois, USA	Red
Karolina Kowalkiewicz	Alexa Grasso	Kevin MacDonald	2019-06-08	Chicago, Illinois, USA	Blue
Ricardo Lamas	Calvin Kattar	Dan Miragliotta	2019-06-08	Chicago, Illinois, USA	Blue
Yan Xiaonan	Angela Hill	Robert Madrigal	2019-06-08	Chicago, Illinois, USA	Red

El dataset posee datos faltantes o NA, el tratamiento de los mismos deberá sex explicitado en el informe.

Consignas

1. Importar el dataset, guardarlo en un objeto bidimensional (puede ser un data.frame, data.table, tibble, etc.)
2. Dividir el dataset original en dos datasets distintos, uno con toda la información referente al participante de la esquina roja y la información en común del encuentro y otro con la información referente al participante de la esquina azul y la información en común del encuentro.
3. Para los dos datasets obtenidos en el item anterior

a) Reemplazar el prefijo que indica el color de la esquina en los nombres de las columnas (R_ o B_) por un " (un campo vacío). Por ejemplo, el nombre de la columna "R_fighter" tiene que pasar a llamarse "fighter". Ambos datasets tendrian que tener nombres de columnas identicos al finalizar.
b) Crear una variable (o recodear la variable "Winner") que indique si el participante ganó la pelea (1) o no (0). Por ejemplo si se está trabajando con los datos del participante de la esquina azul y la variable Winner toma el valor de Red, entonces la nueva variable tendria que tomar el valor 0.
c) Crear una variable que haga referencia al color de la esquina de cada dataset.

4. Unir las filas de ambos datasets en uno solo.
5. Reordenar las columnas del dataset obtenido en el punto anterior de forma tal que la primer columna sea la que se calculó en el ítem 3.b (la cual indica si el participante ganó o no la pelea), manteniendo el orden de las demás.
6. Obtener estadísticas descriptivas básicas para todas las variables continuas: media, desvío, varianza, número de observaciones, máximo y mínimo, cuartiles, etc.
7. Para cada variable numérica graficar el histograma de la misma a efectos de poder visualizar la distribución de la misma. Utilizar por default 10 intervalos, aunque se puede varíar el número de los mismos si se considerase necesario.
8. Graficar el número de encuentros por año, para cada una de las categorías de peso (weight_class).
9. Crear una lista de data.frames (u otro tipo de array de datos) donde cada elemento de la lista sea un subset del los datos el cual contenga la info relacionada a cada una de las distintas categorías de peso. Elegir una de las categorías de peso y crear un nuevo dataset el cual solo contenga los datos pertenecientes a dicha categoría. Estos datos van a ser la base a partir de la cual se va a trabajar en los siguientes puntos.
10. Graficar la distribución, separando los casos que ganaron de los que perdieron (puede ser en 2 gráficos separados o dentro del mismo gráfico utilizando colores distintos, o de cualquier forma en la que se pueda discriminar los casos que ganaron de los que no) de un mínimo de 4 las siguientes variables :

longest_win_streak	losses
total_rounds_fought	total_title_bouts
win_by_Decision_Majority	win_by_Decision_Split
win_by_Decision_Unanimous	win_by_KO/TKO
win_by_Submission	win_by_TKO_Doctor_Stoppage
wins	Stance
Height_cms	Reach_cms
Weight_lbs	age

11. Discretizar las variables countinuas del punto anterior, el criterio para definir los intervalos es libre.
12. Crear un nuevo dataset el cual va a estar compuesto por la variable que indica si se gano o no el encuentro y las variables del punto anterior.
13. Transformar las variables del dataset del punto anterior, exepto la que indica si se ganó o perdió, en variables dummy (también conocido como one-hot-encoding) en el que para cada nivel de la variable se genera una columna la cual indica fila por fila si la variable toma un valor perteneciente a esa subcategoría o nivel.
14. Con estos nuevos datos (previamente dividiéndolos en una población de entrenamiento y una poblción de validación), estimar la probabilidad de ganar el encuentro. Se sugiere utilizar una regresión logística,

pero se puede utilizar otro tipo de modelos siempre y cuando se comente el motivo detrás de su elección. Aclaración: el número de variables regresoras a utlizar es de libre criterio, y si se desease utilizar variables que no se encuentren dentro de las listadas, se puede hacer.

15. Analizar y comentar sobre los resultados obtenidos en el punto 14.