|  |
| --- |
| Computación Científica Actuarial |
| Trabajo Practico 1, Informe |
| Docente: Rodrigo Del Rosso – Colaboradores: Santiago Silva – Joaquín Auza |
| Alumnos: |
| 888530 Ivan Weyner |
| 889411 Cristian Soto |
| 892247 Federico von Brudersdorff |

|  |
| --- |
| 3-10-2019 |

**Descripción de variables del dataset**

El dataset original muestra cada encuentro con información referida a cada peleador e información en común con el oponente sobre el encuentro, se asigna a cada peleador un color y además se observa una columna llamada “Winner” que indica cuál fue el ganador.

Se asigna un color a cada peleador mediante R\_fighter o B\_fighter cuyas dos opciones son rojo o azul. El dataset contiene estadísticas en relación con la pelea, la mayoría de estas divididas en intentados y conectados sobre los golpes, caídas, knock downs, tiempos de pelea e información física del peleador.

El siguiente cuadro resume la información sobre algunas variables:

|  |  |  |  |
| --- | --- | --- | --- |
| **Fighter** | Nombre del peleador | **current\_lose\_streak** | Cantidad de peleas perdidas antes del evento |
| **current\_win\_streak** | Cantidad de peleas ganadas antes del evento | **Draw** | Indica si hubo empate |
| **avg\_X\_att**  **BODY**  **CLINCH**  **DISTANCE**  **GROUND**  **HEAD**  **LEG**  **SIG\_STR** | Promedio de golpes intentados a distintas partes del cuerpo | **avg\_X\_landed**  **BODY**  **CLINCH**  **DISTANCE**  **GROUND**  **HEAD**  **LEG**  **SIG\_STR** | Promedio de golpes exitosos a distintas partes del cuerpo |
| **avg\_SIG\_STR\_pct** | Promedio de golpes significativos | **avg\_SUB\_ATT** | Promedio de sumisiones intentadas |
| **avg\_TOTAL\_STR\_att** | Promedio total de intentos de golpe | **avg\_TOTAL\_STR**  **\_att** | Promedio total de golpes exitosos. |
| **avg\_**  **KD**  **PASS**  **REV** | Promedio de knockdowns | **avg\_opp\_** | Parecen ser las variables del oponente. |
| **total\_rounds\_fought** | Numero total de rounds peleados | **total\_time\_fought**  **(seconds)** | Tiempo peleado en segundos |
| **total\_title\_bouts** | Total de combates por el titulo | **win\_by\_Decision**  **\_Majority** | Numero de peleas ganadas por decisión de la mayoría del jurado (donde 2 jueces votaron a un peleador y el tercero empate) |
| **win\_by\_Decision\_Split** | Numero de peleas ganadas por decisión de la mayoría del jurado (donde 2 jueces votaron a un peleador y el tercero al otro) | **win\_by\_Decision**  **\_Una**  **Nimous** | Número de peleas ganadas por decisión unánime del jurado |
| **win\_by\_KO/TKO** | Peleas ganadas por knock out o knock out tecnico | **win\_by\_Submission** | Número de peleas ganadas por sumisión |
| **win\_by\_TKO\_Doctor\_**  **Stoppage** | Peleas ganadas por knock out técnico por decisión del médico | **Wins** | Cantidad de peleas ganadas |
| **Stance** | Forma de pararse | **Height\_cms** | Altura en centimetros |
| **Reach\_cms** | Alcance de brazos en centímetros | **Weight\_lbs** | Peso en libras |
| **Age** | Edad | **Losses** | Numero de peleas perdidas |
| **Date** | Fecha del encuentro | **Stance** | Postura |
| **Title\_bout** | Indica si la pelea es por el titulo | **Weight\_class** | Categoria de peso del peleador |
|  |  |  |  |
|  |  |  |  |

En la información en común se encuentra: el réferi del encuentro, la locación, fecha, que color ganó, la variable booleana asociada a la columna title\_bout indicando si en el encuentro hubo un campeón de categoría o no, la categoría de peso del encuentro y el número de rounds.

Tratamiento de los NA: Son removidos mediante la función na.rm: TRUE.

**Paquetes utilizados**

* data.table: Permite superar los limites de datos del data.frame que trae por defecto R.
* stringr: Permite manipular variables del tipo string, por ejemplo con expresiones regulares, altamente utilizadas en otros lenguajes de programación.
* moments: Hace posible el calculo de distintos momentos mediante funciones.
* ggplot: Mejora las visualizaciones en R y ofrece más herramientas para las mismas.
* dplyr: Hace fácil manipular el data-frames.
* stringi: También permite mejorar la manipulación de strings.
* gridExtra: Permite devolver varios gráficos en uno.
* tibble: Se utiliza para manejar el data.frame de la regresión.
* caret: Se utiliza para realizar la regresión logística.
* caTools: También es necesario para armar regresión.
* e1071: Es necesario para la función na.action.
* ggpubr,psych: Tests de correlación.

**Gráficos**

Histogramas: Se puede observar que la mayoría de los histogramas son asimétricos con un sesgo hacia la derecha, salvo ciertos parámetros que presentan una distribución normal como: numero de rounds, altura, alcance, tiempo de pelea y el promedio de golpes significativos.

Cantidad de Peleas: En la mayoría de las categorías de peso, la mayor cantidad de peleas se encuentra entre los años 2010 y 2018 y que además las peleas en las categorías de peso de mujeres han crecido con el paso de los años. Por otro lado, en la categoría Open Weight, solo se observan peleas en la década del 90, por lo que se estima que este tipo de categoría ha dejado de existir en la actualidad.

**Comentarios regresión**

Se utiliza el modelo lineal generalizado para estimar la probabilidad de ganar el encuentro, a través de la estimación de máxima verosimilitud de los parámetros del dataset UFC. A través de datos transformados en dummys, el glm utiliza los datos binarios para estimar la variable dependiente “Winner”. Se divide los datos en una población entrenamiento y otra de validación para no producir una sobreestimación (sobreajustar u “overfittting”) a través de la validación cruzada y elegimos algunas variables para realizar el glm para no haya una sobreparametrizacion del modelo.

Se analizaron los valores de varios modelos, diseñados a través del análisis de correlación y correlación parcial con la variable “Winner” (principalmente, de las variables dummy, ya que peso y alcance son intuitivamente sospechosos de correlacionarse a la altura). El modelo elegido incluye:

* Un coeficiente para las variables Height\_med, indicadora de una altura a menos de un desvío estándar de la media.
* Un coeficiente para la variable winPctge, indicadora del porcentaje de victorias sobre el total de peleas (NA para aquellos que no registraban peleas previas).
* Un coeficiente para la variable AgeDif, indicadora la diferencia en añs respecto de la edad media.
* Por último, un coeficiente para la combinación Reach\_lo, indicadora de un peleador con alcance a más de un desvío estándar de la media.
* Un intercepto, que debe interpretarse como el valor de la probabilidad de éxito de aquellos peleadores de altura muy alejada (más de un desvío estándar) de la media.

Este modelo no es el más acertado según AIC (1546,4). El más acertado incluía una variable de interacción entre winPctge y Height\_lo (AIC: 1537,6). El problema de este modelo fue su interpretación: el coeficiente de esta última variable era de -2,8, lo que implicaría que, de ser de los peleadores de menor altura, un mayor porcentaje de victoria se traduce fuertemente en una reducción de las probabilidades de éxito. Se supone que fue una anomalía dada por la muestra (recordar que se utilizó la información de una sola categoría).

El más significativo de los coeficientes resulto ser el de AgeDif (-0.04044), con un p-value inferior al 1% (0.00666). Esto implicaría que, por cada año por encima de la media (29.98528), un peleador pierde alrededor del 4% de probabilidades de éxito (o las gana, si está por debajo).

También resultó ser significativo el coeficiente de altura (0.37541), que parece indicar una clara ventaja en favor de los peleadores con una altura cercana (menos de 1 d.e.) a la media. Su p-value fue de 0.01729, lo que indica que, para la amplia mayoría de los criterios, sería significativo.

En vista de esto, es de destacar que el intercepto estimado fue de -0.35643, un valor cercano al del coeficietne de la altura, el que parece indicar que, efectivamente, los peleadores con alturas muy alejadas de la media están en gran desventaja . Si bien su p-value no es tan bajo (0.06417), sigue siendo un nivel aceptable.

Por último, los coeficientes de Reach\_lo y winPctge no alcanzaron p-values notables (0.44804 y 0.20284, respectivamente). En particular, el coeficiente de Reach\_lo podría estar sufriendo de correlación con la altura. Pero ninguna iteración del modelo que los tuviera en cuenta resultaba más satisfactoria desde un punto de vista de AIC. Sus valores (-0.12526 y 0.27845) son acordes a lo esperado, indicando una desventaja para los peleadores de alcance corto, y una ventaja creíble para los que tuvieran un mayor porcentaje de victorias (aunque hay que señalar que el valor de winPctge no tiene en cuenta la cantidad de peleas, por lo que un peleador con 1 victoria en 1 encuentro tiene el mismo valor que uno con 50 en 50).