BEDU

DATA SCIENCE

REPORTE DE PROYECTO

Bloque 2: Estadística y Programación en R

(Equipo 2)

***Gómez Barraza**, Karla Daniela*
***Guevara Moreno**, Fernando*
***Muñoz Esparza**, José Cruz*
***Rodríguez Rivera**, Gil Estéfano*
***Rosado Martínez**, Ana Elizabeth*

15 de julio de 2021

Índice

Resumen

En este reporte se presenta una descripción de los componentes principales del proyecto de este bloque del curso. El proyecto consiste en la realización de un dashboard con la información de la Primera División de la Liga Española de Futbol (obtenidos de https://www.football-data.co.uk/spainm.php). Este dashboard presenta información procesada útil para la realización de apuestas y quinielas. Además, se discuten todos los análisis preparatorios previos a la información que se presenta en el dashboard. Se mencionan las posibles áreas de oportunidad o mejora de cada método utilizado.

Los **códigos** se encuentran en el siguiente repositorio de **GitHub**: https://github.com/gilesitorr/DataScience3_Bloque2.

El **dashoboard** se encuentra en el siguiente enlace https://gilesitorr.shinyapps.io/Postwork8/.

1. Introducción

La Probabilidad y la Estadísticas son áreas del conocimiento que siempre han estado íntimamente relacionadas a las apuestas y los juegos de azar. Un ejemplo de esto es la popular anécdota de que un apostador le propuso a Blaise Pascal un par de problemas relacionados con su experiencia en las apuestas. Fue ahí, según se cuenta, cuando Pascal dio origen a la Teoría de la Probabilidad moderna. Este estrecho vínculo vive incluso hoy en día. Esto puede verse en las grandes casas de apuestas o en las quinielas sobre los resultados de los partidos de futbol que se hacen entre conocidos. La motivación del proyecto surge de un interés en el segundo ejemplo.

El objetivo de este proyecto es proporcionar de manera sencilla y condensada información útil para realizar apuestas de futbol.

En la **Sección 2** se enuncian los conceptos que son relevantes para comprender conceptual y contextualmente algunos elementos del proyecto pero que no se trataron a lo largo del curso. En la **Sección 3** se describen las partes principales en el desarrollo del proyecto:

1. El **análisis inicial** de la relación que hay entre el **desempeño** de los equipos en un partido y la **localidad** de los mismos. Es decir, si el equipo jugó de local o visitante, ¿cuántos goles es más probable que anote?.

2. El estudio de una librería (fbranks) que sirve para realizar **predicciones** sobre los marcadores de los distintos partidos haiendo analizados desempeños previos de cada equipo de la liga.

3. Las posibilidades, retos y oportunidades de utilizar la cantidad de goles totales por partido como una **serie de tiempo**.

4. La presentación de una librería (mongolite) para hacer consultas específicas sobre la base de datos de los partidos.

5. La realización, **comparación y presentación** de algunos resultados importantes obtenidos a lo largo del proyeto en un dashboard con gráficas y tablas interactivas (data tables) con el fin de que la información que se procesó sea accesible.

En el apartado de los **documentos de las librerías** se enlistan las fuentes para revisar las especificaciones de las librerías usadas que fueron más relevantes en el desarrollo del proyecto pero que no se revisaron en el curso.

2. Marco teórico

2.1. Independencia de dos variables aleatorias

Se entiende a la distribución conjunta de probabilidad como la probabilidad de que la intersección de que dos eventos aleatorios ocurra. De manera más formal, indica la distribución de probabilidad para dos -o más- variables aleatorias.

Dos **variables aleatorias** X **y** Y **son independientes** dada una distribución de probabilidad P si la distribución conjunta es igual al producto de las distribuciones marginales de cada variable. Esto es:

$$P(X = x, Y = y) = P_X(X = x)P_Y(Y = y) \tag{2.1}$$

Donde las sumas sobre todos los valores posibles de X o de Y, según sea el caso, $P_X(X = x) = \sum_Y P(X = x, Y = y)$ y $P_Y(Y = y) = \sum_X P(X = x, Y = y)$ son las probabilidades marginales. (Beichelt, 2014)

2.2. Método bootstrap

El **método bootstrap** sirve para hacer inferencias sobre algún estadístico obtenido a partir de una muestra. Consiste en formar submuestras (por medio de un remuestreo aleatorio y con reemplazo) a partir de la muestra original. Se calcula el estadístico de interés para cada submuestra. A partir de esos cálculos, se puede estimar el error estándar del estadístico. Pudiendo así obtener información como puede ser el intervalo de confianza y el intervalo de predicción de la muestra. (Yen, 2019)

Aunque se recomienda tener muestras grandes para poder obtener estadísticos confiables, es importante tener en cuenta también el tamaño de la muestra a partir de la cual se aplica el bootstrap: esto es para evitar crear muestras con demasiados datos repetidos o redundantes.

2.3. ¿Qué es un momio?

Los **momios** (**odds**, en inglés) es la proporción de la probabilidad p de que un evento ocurra contra la probabilidad de que no ocurra $1 - p$. Esto es:

$$Momio = \frac{p}{1-p} \tag{2.2}$$

(Mandeville, 2007)

3. Desarrollo del proyecto

3.1. Postworks 1-4: Análisis inicial de la relación entre los goles y la localidad de los equipos

El Postwork 4 es una recopilación de los códigos realizados en los primeros tres postworks con algunas añadiduras. Para evitar la redundancia, se discutirá únicamente el conjunto de los postworks mencionados, que es el código del Postwork 4.

Lo que hace el programa del **Postwork 4** es tomar tres bases de datos con información sobre la Liga Española de Primera División de las **temporadas 17/18**, **18/19** y **19/20**. Estas bases de datos cuentan con la información de todos los partidos jugados en las respectivas temporadas. Por ejemplo, tiene información sobre la fecha del partido, el equipo que jugó como local así como el que jugó de visitante, el marcador y algunos otros datos que no son de tanto interés para este estudio.

Los datos que son de interés en esta parte del trabajo son los **marcadores**, es decir, los goles que anotó el **equipo local** y los que metió el **equipo visitante** en cada partido, sin importar la fecha ni los equipos específicos. Las columnas en con esos datos se llaman ***FTHG*** y ***FTAG***, respectivamente. Supondremos que la probabilidad de cada marcador es independiente de la temporada que se juega (así como se ha despreciado el equipo y la fecha), por lo que, en aras de tener más datos con los cuales trabajar, se juntan las bases de datos de las tres temporadas enunciadas luego de haber escogido sólo las dos columnas citadas anteriormente (FTHG y FTAG) en cada una de las bases individuales.

Con la nueva base de datos, compuesta por dos columnas, se construye una tabla de la frecuencia de cada marcador en toda la base de datos con el comando ***table***. Esta tabla de frecuencias sirve para construir una tabla pero con probabilidades conjuntas, misma que se obtiene al dividir cada entrada entre el total de datos de la base. Por otra parte, es posible obtener las probabilidades marginales de cada variable (FTHG y FTAG) de diversas maneras; en el programa se obtuvo al agrupar la variable adecuada en la base de datos y luego dividir la frecuencia con la que aparece la variable estudiada -independientemente del valor de la otra- entre el total de datos). En el programa se pueden visualizar todas las distribuciones de probabilidad explicadas (la conjunta y las dos marginales).

De la definición de dos variables aleatorias independientes (enunciada en la **Subsección 2.1**), se puede ver que, al dividir las probabilidades marginales en ambos lados, se obtiene un uno del lado derecho de la igualdad. Este valor puede utilizarse para estudiar en qué condiciones (y porqué) las variables FTHG y FTAG son independientes.

La segunda parte del programa tiene como objetivo determinar las **condiciones de independencia** (es decir, los marcadores en los que se puede decir que los goles del local **no dependen** estadísticamente de los goles del visitante y viceversa). Esto se hace dividiendo la tabla de las probabilidades conjuntas entre el producto de sus respectivas probabilidades marginales (**Ojo**: esto se vale sólo cuando ambas probabilidades marginales son diferentes de cero: pero, en ese caso, se puede estudiar si las variables son independientes al revisar si las probabilidades conjuntas son cero; si ambos lados de la **condición de independencia, la ecuación** (2.1), son cero, necesariamente se cumple la independencia en ese caso). Esa tabla da una idea de cuán cercano está cada entrada de la tabla de la independencia de variables y, si bien, puede ser una buena estimación, no es del todo certera.

Para dar certeza a la inferencia, se necesita de una desviación estándar, sin embargo, la tabla de cocientes no cuenta con una. Esta situación se puede sortear por medio del **método bootstrap** (mencionado en la **Subsección 2.2**).

Con cada submuestra se repite el proceso de calcular las distribuciones de probabilidad conjunta y marginales para luego conseguir una tabla de coeficientes como los descritos anteriormente para evaluar la independencia de las variables. Una vez que se tiene una tabla de coeficientes para cada submuestra, se puede calcular el promedio y la desviación estándar de cada elemento de la tabla. Esta tabla de desviaciones estándar S es útil para realizar una prueba de hipótesis.

Para la prueba de hipótesis se usará la tabla de coeficientes de la muestra original M en lugar del promedio de las submuestras porque, de acuerdo a la construcción que se hizo en el método bootstrap, la muestra original funge a modo de una *-pseudo-*población y las submuestras fungen como muestras de la población. Expresado de

otro modo, es preferible usar la media poblacional en lugar de la media muestral.

La **hipótesis nula** H_0 es que las variables son independientes, es decir, que los coeficientes tienen valor 1. La prueba de hipótesis se hará a modo de tabla (donde es claro que la expresión sólo tiene sentido hablando de entrada por entrada, y no del objeto completo, que es como opera R), de la siguiente manera:

$$Z = \frac{1 - M}{S/\sqrt{N}} \tag{3.1}$$

Donde el 1 viene de la hipótesis nula y N es el número de submuestras utilizadas. Este estadístico de prueba se contrasta con una significancia bilateral del $5\,\%$, pues la negación de la hipótesis nula es que los coeficientes de M no sean iguales a 1, independientemente de si son mayores o menores.

Luego de la prueba, se obtienen los marcadores para los que las variables FTHG y FTAG son independientes. Estos marcadores pertenecen ambos al conjunto {0,1,2,3} además de algunas combinaciones de valores colindantes a esas celdas (si se organizara como una tabla).

Si se ignoran los valores colindantes (que pueden suponerse que, con una muestra más grande, fallarían la prueba) y la atención se centra en el *conjunto principal*, puede verse que es la región con resultados más frecuentes. Es decir, independientemente del talento de un equipo, es razonable que un equipo de menor categoría le meta un gol a uno de mayor clase, mientras que es posible que el segundo le meta uno o dos al primero. Por eso, en esas regiones los goles del visitante y los del local son independientes. Por el contrario, note la independencia no se cumple cuando se trata de marcadores más *exóticos*. Esto puede verse desde dos aristas, la primera es que son resultados suficientemente raros como para poder estudiarse adecuadamente; la segunda interpretación es que, por ejemplo, es improbable que un equipo bueno le meta 8 goles a otro mientras que el otro equipo le meta 1 o 2, aún siendo bueno: eso sólo es concebible -por lo general- cuando los dos equipos tienen un nivel desproporcionado, que ya no es un *estado independiente* de los equipos que jueguen y, por lo tanto, ya no es un marcador en el que las variables sean independientes. En casos más extremos, como un marcador de 8 a 6, sencillamente es improbable que dos equipos profesionales lleguen a esos marcadores, salvo que, por ejemplo, ambos porteros se lesionen y no tengan reemplazos u ocurra otro hecho extraordinario.

3.2. Postworks 5: Predicciones usando *fbranks*

El Postwork 5 presenta con un ejemplo concreto la utilidad de la librería fbranks para realizar predicciones sobre los juegos de las jornadas dado un conjunto de juegos pasados. Con esto, puede introducirse la idea de que no sólo se puede utilizar

la estadística para estudiar una muestra estática sino que también se puede usar para hacer predicciones de su estado futuro.

Lo que hace el programa del **Postwork 5** es retomar la base de datos obtenida a partir de las tres bases de datos con información sobre la Liga Española de Primera División de las **temporadas 17/18**, **18/19** y **19/20**, tomando las columnas de la fecha, el equipo local, el equipo visitante y el marcador. En este caso es importante cambiar nombre de las columnas correspondientes a *date*, *home.team*, *home.score*, *away.team* y *away.score*, pues de lo contrario no se podrá usar la librería ***fbranks***. Se guarda la base de datos en un archivo csv.

Se usará la función ***create.fbRanks.dataframes*** de fbranks para leer el archivo csv que se guardó anteriormente. Esta función crea una lista con, entre otros, los elementos **scores** y **teams**.

Antes de proceder, es necesario crear un vector con todas las fechas (sin repetir) en las que hubo partido. Para esto se usó la función ***unique***.

Se utiliza la función ***rank.teams*** utilizando el vector de fechas únicas (desde la primera hasta la penúltima fecha) y los elementos scores y teams como argumentos. Se obtiene un ranking de los equipos.

A partir del ranking, se puede utilizar la función ***predict*** (también de fbranks) para predecir los resultados de los partidos de la última fecha del vector de fechas. Se puede comparar esa predicción con los resultados reales y puede verse que este proceso puede ser útil al planear una quiniela o preparar una apuesta. Aunque las predicciones de los marcadores no eran tan buenas, parece ser que sí es un buen estimador para determinar si un equipo gana o pierde.

El método podría ser más apto para estudiar temporadas individuales. Esto se debe a que, aunque se cuenta con menos datos, es posible que un equipo destaque en una temporada y tenga un mal desempeño en otra: introduciendo inconsistencias en la información.

3.3. Postworks 6: Los goles totales por partido como serie de tiempo

El Postwork 6 es otra exploración de la evolución de un parámetro del sistema y la introducción de métodos para obtener información útil de un sistema *con mucho ruido*.

El programa del **Postwork 6** retoma el conjunto de tres jornadas utilizado en las dos subsecciones pasadas. A esta base de datos se le agrega una columna con la suma de goles por partido. Si se grafica esta columna contra la fecha, puede verse que hay tantos datos en la gráfica que la información que provee es inútil.

Una de las cosas que se puede hacer para tratar la columna de la suma de

goles es calcular el promedio de esa columna pero para cada mes. Con ese solo tratamiento, puede formarse una serie de tiempo con registros mensuales que provea de más información.

Esa serie temporal es interesante porque podría ajustarse a alguno de los modelos estocásticos que se revisaron en el curso (MA, AR, ARMA, ARIMA). Con esto, nuevamente se presenta la posibilidad de hacer predicciones sobre la evolución de ese estadístico mensual.

3.4. Postwork 7: La posibilidad de utilizar MongoDB para hacer las consultas iniciales

Algo que en este caso no fue de tanta relevancia en el gran esquema del proyecto pero de lo cual es importante dejar registro es que con una cuenta en MongoDB Atlas o con un MongoDB Server es posible hacer consultas de base de datos en RStudio con la librería ***mongolite***. Esto podría ser útil para automatizar alguna consulta que se realiza frecuentemente a través de MongoDB Compass, por ejemplo.

El código del **Postwork 7** explora un breve ejemplo de lo mismo. Se realiza una conexión a un servidor y se realiza una consulta.

3.5. Postwork 8: Dashboard con los resultados más relevantes

El Postwork 8 es la conclusión y presentación de los resultados más relevantes de todo el proyecto.

Antes de describir el código del Postwork 8, se mecionará brevemente el objetivo y el funcionamiento del código **momios.R**, que es un recurso usado para complementar el proyecto y que se obtuvo de https://github.com/beduExpert/Programacion-R-Santander-2021/blob/main/Sesion-08/Postwork/momios.R. Retomando la parte de las predicciones del Postwork 6, el código realiza predicciones sobre los goles del equipo local y el visitante. Con esas predicciones se calcula el momio de que la suma de goles por partido sea de más de 2.5 goles (momio máximo) y el momio promedio para la misma cantidad de goles totales por partido (momio promedio). Con esos valores se puede realizar una simulación de las ganancias proyectadas con cada tipo de momio, suponiendo apuestas de $1,000$ monedas con una bolsa inicial de $50,000$. Se grafican ambas simulaciones.

El programa del **Postwork 8** crea un dashboard con cuatro pestañas.

1. La primera pestaña tiene gráficas de barras para cada uno de los equipos de la liga. Se tiene la frecuencia de la cantidad de goles que anotó el equipo local o el visitante jugando contra el equipo de la gráfica correspondiente como visitante. Se estudiaron las jornadas desde $10/11$ hasta $19/20$. Como el equipo

de la gráfica correspondiente es visitante, la selección de los goles del visitante indica la frecuencia con los que anotó goles. La selección de los goles de local indica la frecuencia con la que le anotaron goles.

2. La segunda pestaña tiene dos gráficas de barras (una para el equipo visitante y otra para el local) de la cantidad de goles que anotó el equipo en las jornadas desde $17/18$ hasta $19/20$. Además, tiene un heatmap de las probabilidades conjuntas. (Esto se puede recuperar del Postwork 3 o 4).

3. La tercera pestaña tiene un data table con datos de las jornadas desde $10/11$ hasta $19/20$. El data table tiene las columnas *date*, *home.team*, *home.score*, *away.team* y *away.score*.

4. La cuarta pestaña muestra la gráfica de las simulaciones de los momios.

Las gráficas de la pestaña 1 son útiles para observar el rendimiento que tiene cada equipo cuando juega como visitante. Quizá podría resultar útil agregar otra pestaña con gráficas similares pero para analizar el desempeño de los equipos cuado juegan en casa.

Las de la pestaña 2 indica cuáles son los casos de los marcadores más frecuentes.

El data table puede servir para hacer consultas específicas, de manera alternativa a lo discutido sobre el Postwork 7.

Como se puede ver en las simulaciones de los momios, el método de apuesta de la simulación no es útil para maximizar ganancias. Con ese objetivo se puede plantear un modelo *naive*, pero que podría optimizarse. El modelo consiste en el ajuste de un modelo estocástico para la cantidad de goles de cada equipo. Esto junto con el modelo de los momios puede dar como resultado un modelo que sepa *cuándo conviene apostar*. Además, el modelo puede mejorarse calculando las probabilidades de que aparezcan los valores para los que se ha estimado que la cantidad de goles del local es independiente de los del visitante: verificando los casos en los que es válido predecir independientemente la serie temporal de la cantidad de goles de cada equipo por separado.

3.6. Retos

Como en todos los proyectos de programación, aparecieron problemas con la sintaxis. En partícular, las situaciones más retadoras en ese aspecto fueron cuando se usaron por primera vez las librerías **fbranks** y **mongolite**. Ambas situaciones se resolvieron revisando la documentación respectiva (se encuentran más abajo, en el apartado de **Documentos de las librerías**). En el caso de fbranks es importante seguir al pie de la letra el orden de utilización de las funciones que se mencionarion

en la **Subsección 3.2**. En fbranks también es fundamental que la base de datos que se guarda en un archivo csv tenga los nombres de columnas y el orden indicados (date, home.team, home.score, away.team y away.score). En el caso de mongolite es fundamental tener un cluster en MongoDB Atlas o un MongoDB Server, pues de lo contrario no se cuenta con dónde hacer la conexión desde RStudio.

Un reto conceptual importante fue entender la utilidad que tenía el método bootstrap en el Postwork 4. Esta situación se solucionó consultando diversas fuentes hasta encontrar una que explicara la información suficientemente clara (y que es la que se cita en la **Subsección 2.2**). Luego de investigar, quedó claro que era para obtener la *incertidumbre* (o el error) en cada uno de los coeficientes de la muestra. Otro reto que también surgió al implementar el código del Postwork 4 fue entender cómo realizar el análisis de lass independencias de variables. La primera aproximación utilizada fue ver si el valor 1 pertenecía al intervalo de incertidumbre de cada coeficiente obtenido con el método bootstrap. Este procedimiento daba resultados razonables pero no permitía indicar la significancia de tal observación. Por lo anterior, se optó por realizar una prueba de hpótesis, con el fin de obtener una conclusión con una justificación rigurosa.

Uno de los problemas técnicos a resolver fue el ajuste del tamaño de las ***box*** usadas en el dashboard del Postwork 8. Esto es porque la visualización del dashboard localmente en RStudio es diferente a la visualización que se obtiene en la página de **ShinyApp.io**. La única manera que se encontró fue estimar a ojo el tamaño apropiado y hacer una actualización del dashboard en la app de Shiny.

Documentos de las librerías

- FBRANKS: https://cran.r-project.org/web/packages/fbRanks/fbRanks.pdf
- MONGOLITE: https://cran.r-project.org/web/packages/mongolite/mongolite.pdf

Referencias

- BEICHELT, F. (2014). *Applied Probability and Stochastic Processes*. Second edition. CRC Press. p. 28.
- YEN, L. (2019). *An Introduction to the Bootstrap Method*. Recuperado el 10 de julio del 2021 de: https://towardsdatascience.com/an-introduction-to-the-bootstrap-method-58b
- MANDEVILLE, P. B:. (2007). *La razón de momios*. Ciencia UANL. Año X, Num. 2. Recuperado el 11 de julio del 2021 de: https://www.redalyc.org/pdf/402/

40210219.pdf.

- WALTERS, T. (2021). *How Do Odds Work in Betting?*. Recuperado el 11 de julio del 2021 de: https://www.investopedia.com/articles/investing/042115/betting-basics-fractional-decimal-american-moneyline-odds.asp.