Taller de exploración

Maestría en Ciencia de Datos

Profesores:

Norha M. Villegas, PhD (nvillega@icesi.edu.co)

Diego A. Bohórquez, MSc (dabohorquez@icesi.edu.co)

Universidad Icesi

Objetivos del taller

1. Aplicar los pasos del ciclo de la analítica que corresponden con las actividades de pre-procesamiento y análisis exploratorio, con la ayuda de R.

2. Reconocer la utilidad de la analítica de datos en cualquier campo de estudio (para este caso, en el deporte).

3. Presentar el proceso de pre-procesamiento, el análisis exploratorio y sus resultados (variables explicativas vs. variable objetivo).

Esquema de evaluación

- Análisis e integración de distintas bases de datos abiertas (sugeridas y adicionales)
- Documentación del proceso seguido para la limpieza, pre-procesamiento análisis de los datos en R (markdown)
- Aplicación de análisis univariado y bivariado
- Presentación del proceso realizado (viernes 27 de septiembre, 10 minutos por grupo)

Taller práctico

Contexto

El análisis de video y la intuición de los entrenadores, asistentes técnicos y preparadores físicos ha sido por muchos años la técnica empleada para analizar y entender el fútbol. Sin embargo, hoy en día quien esté empleando únicamente esta herramienta, está en la "prehistoria".

El vertiginoso desarrollo tecnológico de las últimas dos décadas, ha permitido que en el fútbol se logre recolectar gran cantidad de datos, para ser procesados por diferentes agentes pertenecientes a este deporte (equipos, representantes de jugadores, medios de comunicación, casas de apuestas, investigadores, etc.). En el tema de estrategia, se ha llegado a comentar que la analítica de datos constituye el jugador número 12.

Interesados, pueden leer un reporte del New York Times sobre el Liverpool:

https://www.nytimes.com/2019/05/22/magazine/soccer-data-liverpool.html

Contexto

Haremos de cuenta que su grupo de trabajo fue contratado por una firma inversionista para construir un modelo que permita pronosticar los goles que marcará un equipo en un partido de la Liga inglesa.

Una herramienta tradicional para pronosticar la cantidad de goles es el modelo de regresión de Poisson. Dicho modelo requiere de 2 supuestos:

1. Los goles marcados por un equipo siguen la distribución de Poisson
2. Los goles marcados por un equipo en un partido son independientes de los goles marcados por su rival.

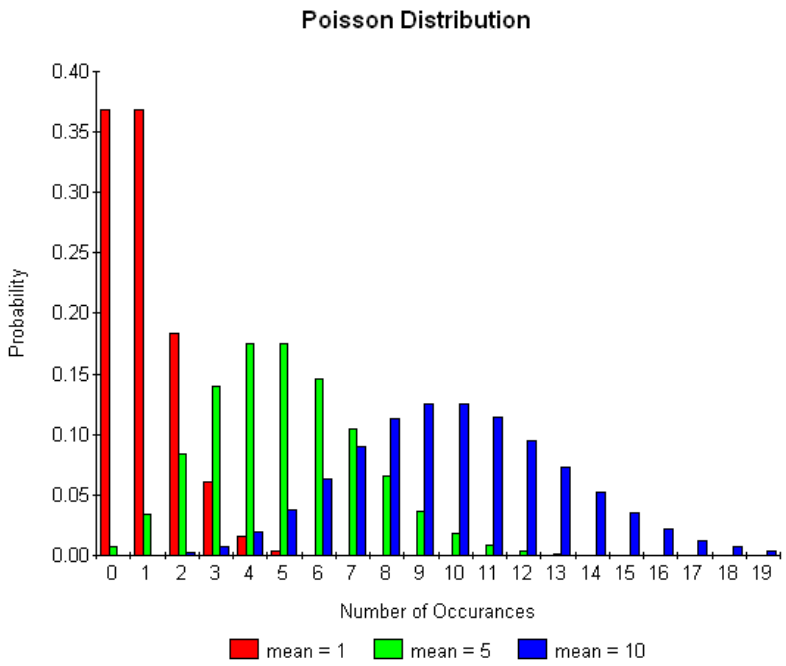
Contexto

Suponga que los goles que i le anota a j son g_{ij}, que siguen una distribución de Poisson, donde la media es λ_{ij}. Formalmente, el modelo parte de:

$$E\left[g_{ij}\middle|X\right] = \lambda_{ij} = e^{X\beta}$$

Sobre la distribución de Poisson…

- Describe el número de veces que se presenta un evento durante un intervalo.
- La media es igual que la varianza.
- Cada intervalo es independiente.
- El sesgo positivo se reduce a medida que la media aumenta.



Contexto

¿Cuál es la lógica del modelo?

- Existen variables propias que aumentan la probabilidad de anotar más goles.

- Existen variables de mi rival que disminuyen la probabilidad de anotar más goles.

¿Qué insumos requiere el modelo?:

- Variable objetivo: goles marcados por un equipo en un partido

Contexto

- **Variables explicativas:**

Según la base que encuentra en "English Premier League (football)" (en https://datahub.io/sports-data), se espera que sean incluidas en el estudio variables tales como:

- Jugar de local o visitante
- Las cuotas de las casas de apuestas Bet 365 y William Hill para el partido.
- El promedio de remates por partido en la temporada anterior del equipo y de su rival.*
- El promedio de remates por partido en la temporada actual del equipo y de su rival.
- El número de puntos promedio** por partido en la temporada anterior del equipo y de su rival.*
- El número de puntos promedio** por partido en la temporada actual del equipo y de su rival.
- El promedio de goles por partido en la temporada anterior del equipo y de su rival.*
- El promedio de goles por partido en la temporada actual del equipo y de su rival.

*Para los recién ascendidos, debe ser el dato promedio de los tres equipos que descendieron en dicha temporada.

** Asignando 3 puntos a una victoria, 1 punto a un empate y 0 puntos a una derrota.

Contexto

La muestra a utilizar será la base de datos de la temporada 2018-2019 que se encuentra en el link entregado. Esto implica que los datos que debe buscar de la temporada anterior son los de la 2017-2018.

El diccionario de variables lo encuentra disponible en Moodle.

Adicionalmente, debe agregar al menos tres variables de otras bases de datos (pueden ser bases de datos que ya existan o que usted construya).

Contexto

Objetivo:

Aplicar las técnicas de limpieza, pre-procesamiento y exploración de datos que permitan identificar las variables que deben incluirse en un modelo basado en la regresión de Poisson.

Preguntas de análisis

1. ¿Es razonable considerar que la distribución de goles en dicha temporada sigue una distribución Poisson? (Utilice gráficos y recuerde la característica de dicha distribución en la que la media es igual a la varianza)

2. ¿Cuáles son las características generales de dicha temporada?: promedio de gol, remates, remates a puerta, tiros de esquina, faltas, tarjetas amarillas y rojas POR PARTIDO. Discrimine por local y visitante estas estadísticas.

3. ¿Qué variables de las mencionadas en las diapositivas anteriores (las que van al modelo de Poisson) parecen estar correlacionadas con la cantidad de goles que un equipo anota en un partido? (utilice gráficos y pruebas estadísticas)

Revisión de literatura

Hacer revisión de literatura siempre sirve a la hora de hacer este tipo de ejercicios de análisis.

El paper seminal en este campo es el de Maher (1982): Modelling Association Football Scores. A partir de dicho estudio han surgido cientos de trabajos (incluso con metodologías diferentes a la de Poisson).