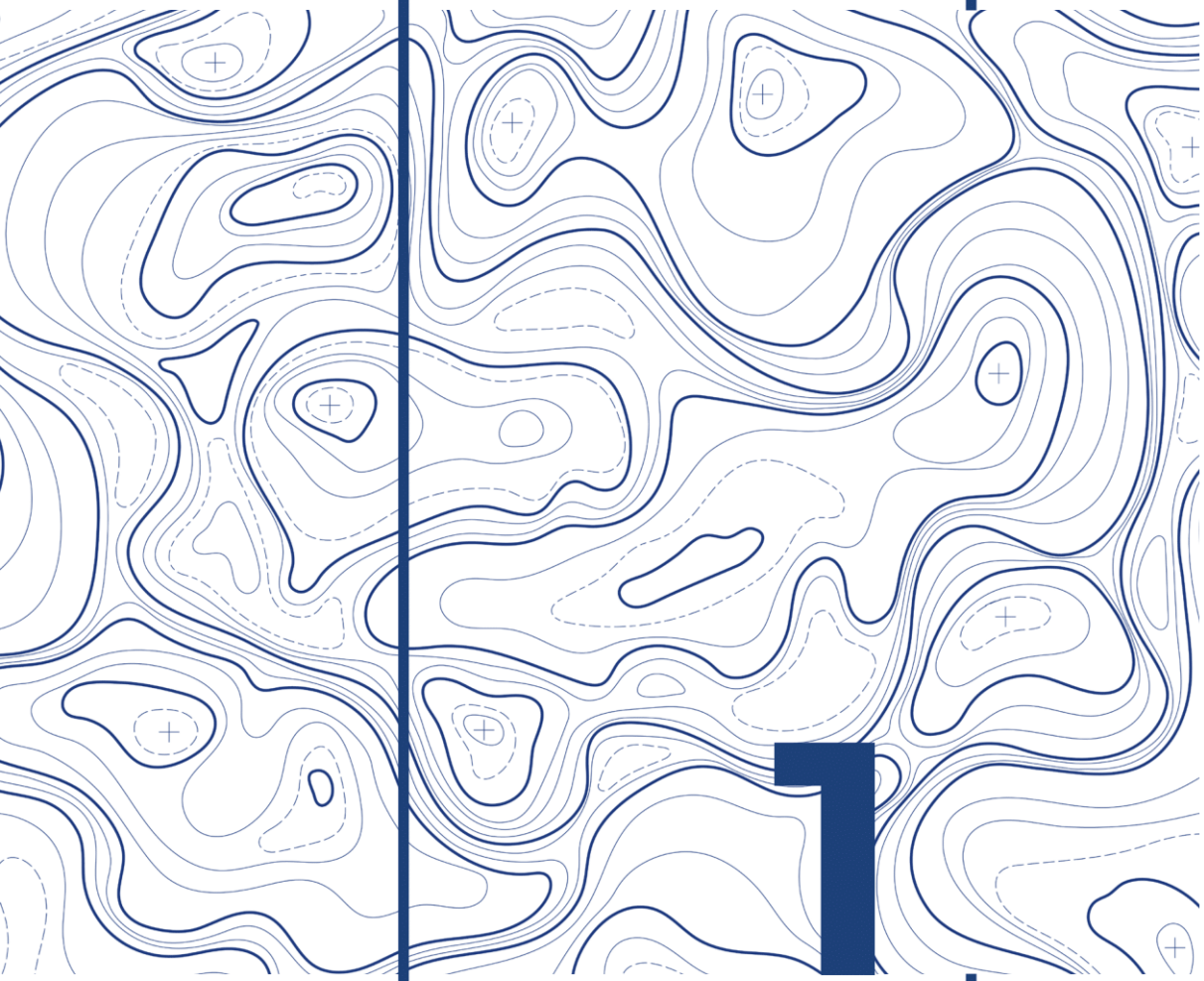
Proyecto de Datos I

Preparación de datos

Óscar Marín Esteban

Carlos Mantilla Mateos

Álvaro Enol Alonso Ortega



1

EXTRACCIÓN DE LOS DATOS

Vamos a usar diferentes fuentes de datos, entre ellas las siguientes:

Fuente 1: Página de Whoscored (Link), de esta página hemos sacado todas las características generales de los jugadores.

Librerías o herramientas:

Las librerías utilizadas son selenium para hacer todo el tema relacionado con la extracción de datos, pandas para el manejo del dataframe, HTML para el manejo en si de diferentes elementos de la web y time para esperar el tiempo suficiente antes de acceder a un elemento.

Pseudocódigo, Aspectos y Problemas:

En cuanto a la captura, lo que hacemos es, primero que nada, inicializar el navegador, una vez inicializado, tuvimos un problema con las cookies, porque al iniciarlo nos salía un aviso de cookies, por lo que lo que hicimos fue una función que pasa las cookies nada más iniciar la web, que pique en el botón correspondiente. Una vez pasadas las cookies pasamos a leer la tabla leemos la primera página de el primer año seleccionado, tenemos dos bucles, uno por los años y otro por el número de páginas que hay, en este segundo, para a partir de la primera página, teníamos un problema al cambiar la página, ya que salía un icono privacy que no se podía quitar, por lo que optamos por la solución de bajar la página un poco para que desapareciera con (driver.execute_script("window.scrollBy(0, 750)")), a partir de aquí ya habríamos hecho todo lo necesario para tener todos los datos en nuestras manos.

Fuente 2: Página de Transfermarket, de aquí hemos sacado principalmente, aunque se pondrá detalladamente más adelante, la posición exacta, la edad y el precio.

Librerías o herramientas:

En este caso, hemos utilizado una API, ya que al sacarlo normal no nos dejó, por lo que tras investigar un poco descubrimos esta API.

Pseudocódigo, Aspectos y Problemas:

A partir de aquí, todo fue bastante sencillo, simplemente usar la API, e ir añadiendo lo leído a un dataframe, que luego descargaremos, lo realmente difícil fue encontrarla y aplicarla al principio.

2

TRANSFORMACIÓN DE LOS DATOS

Variables:

Leyenda de utilidad (la utilidad la hemos tratado para todas las variables, considerando TODOS los modelos):

SI

NO

QUIZÁS

WhoScored:

Todas las variables aquí que ponen número son numéricas, todas las variables que ponen tasa son tasas y todas las que ponen % son de porcentaje.

Apps: Son las apariciones de un jugador en la temporada completa, entre paréntesis las veces que han empezado de suplentes.

Mins: Número de minutos jugados

Goals: Número de Goles en la temporada

Assists: Número de asistencias en la temporada

Yel: Número de tarjetas amarillas

Red: Número de tarjetas rojas

SpG: Número de tiros por partido

PS%: Porcentaje de acierto de pase

AerialsWon: Duelos aéreos ganados por partido

MotM: Número de veces que ha sido jugador del partido

Tackles: Número de segadas por partido

Inter: Número de intercepciones por partido

Fouls: Número de faltas por partido

Offsides_won: Tasa de fueras de juego provocados por partido

Clear: Número de despejes por partido

Drb_deffensive: Número de veces que le han regateado su defensa por partido

Blocks: Tasa de bloqueos de tiro por partido

OwnG: Número de goles en propia

KeyP: Número de pases clave por partido

Drb_offensive: Número de veces que han regateado por partido

Fouled: Tasa de faltas recibidas por partido

Offsides: Tasa de fueras de juego por partido

Disp: Tasa de pérdidas de balón

UnsTch: Tasa de malos controles por partido

AvgP: Tasa media de pases por partido

Crosses: Tasa de centros por partido

LongB: Tasa de pases largos por partido

Transfermarket:

Age, Id son enteras, Height es numérica decimal, y el resto son strings

Id: Id del jugador
Position: Posición en el campo
DateofBirth: Año de nacimiento
Age: Edad del Jugador
Nationality: nacionalidad
CurrentClub: Club Actual
Height: Altura del jugador
Foot: Pie dominante
JoinedOn: Fecha de incorporación al equipo actual
SignedFrom: El equipo en el que estaba antes de que lo ficharan
MarketValue: Valor de mercado del jugador
Status: Jugador lesionado o no

Últimas añadidas, explicado en el punto 1.4:
OffTarget: Tiros a puerta que van fuera.
OnPost: Tiros a puerta que dan al palo.
OnTarget: Tiros a puerta que entran.
Blocked: Tiros a puerta bloqueados.
OpenPlay: Goles a juego abierto.
Counter: Goles en contrataque.
SetPiece: Goles desde zonas determinadas (corneres y fueras).
PenaltyScored: Goles de penalti.
OwnG: Goles en propia puerta.
AccLB: Salvadas en área de la portería.
InAccLB: Salvadas en área de penalti.
AccSP: Salvadas fuera de esas áreas.
LongB: Pases precisos de lejos.
InAccSP: Pases fallidos de lejos.
Short: Pases precisos de cerca.
InAccLB: Pases fallidos de cerca.
KeyP: Pases clave.
Titularidades: Número de veces que es titular
Suplencias: Número de veces que es suplente

En cuanto a las transformaciones, hemos hecho varias, que bajo nuestro punto de vista van a aportar muchas mejoras a la hora de la calidad del modelo cogido, la primera es la variable del equipo en el que están, una cualitativa, en la posición en la que quedó el equipo en la liga ese año, de tal manera que para cada año tendremos nuevas.

En la columna de Apps, es decir apariciones, hemos hecho una separación entre titularidades y suplencias, ya que venían junta, así que no se podían analizar y eran importantes, ahora estas 2 variables se podrán utilizar sin ningún inconveniente.

1.2 Limpieza de los datos

En nuestro caso para el tema del proceso de limpieza no ha sido el principal problema, ya que en todas las filas en las que no hay datos tras el análisis detallado hemos visto que se pueden sustituir por 0, lo comprobamos porque había una opción en los datos de whoScored que limitaba la extracción a mínimo de partidos, hemos preferido dejarlos todos para tenerlos en cuenta, y simplemente se sustituyen por 0 los valores.

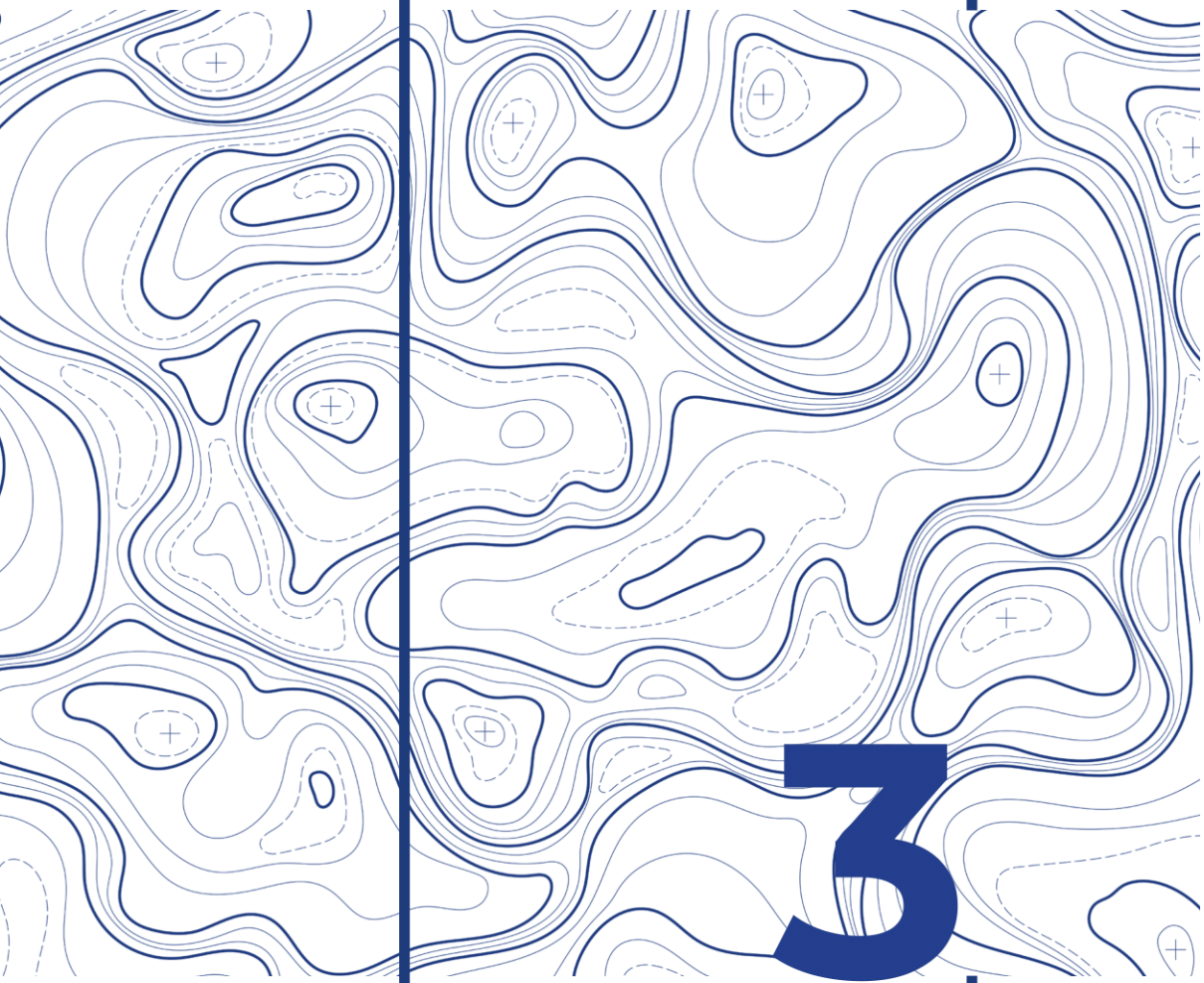
1.3 Integración de los datos

Posiblemente la parte más problemática de esta segunda entrega, ya que al tener que juntar datos de 2 webs distintas, había diferencias significativas en como estaban escritos los nombres, por lo que hemos perdido valores, unos 500, es relativamente una gran cantidad, pero al final tenemos bastantes datos, al tener bastantes años, además para cerrar este punto, muchos de ellos son los canteranos de los que no tenemos casi partidos, ya que también en transfermarket faltaban algunos de ellos, por lo que la pérdida no ha sido tan tan significativa finalmente. Este trabajo lo hemos hecho con Unicode y con fuzzywuzzy, dos librerías que hemos investigado que hacen el trabajo correctamente para los nombres que son un poco diferentes ya sea por tildes o por alguna letra mal escrita, cogiendo los datos que coincidan en más del 80% del nombre.

También, tuvimos que quitar todas las columnas que al juntar los datos se repetían, y creaba nuevas variables como apps_x con apps_y, a pesar de ser la misma variable.

1.4 Creación de nuevas variables

Para este punto no hemos tenido que crear ninguna variable adicional, aunque sí que hemos sacado valores adicionales con los que no contábamos en la entrega anterior para sacarlos, pero que hemos descubierto y que resultan útiles, por lo que se pueden incluir en nuevas variables a pesar de que no las hemos creado nosotros, pero sí que son adicionales, entre las que se incluyen valores muy específicos interesantes que nos darán un punto de vista más específico.



3

CARGA DE LOS DATOS

La separación en sí, está aún por definir, aunque sí que es verdad que en la anterior entrega dijimos que lo íbamos a hacer en 12 modelos, resulta que no tenemos tantos datos como para hacer esto, incluyendo el inconveniente de la pérdida de algunas variables por juntarlas de diferentes sitios web, quedando muy pobre cada modelo si lo hiciésemos de esta manera.

Habíamos pensado también en poner un número a cada posición o solamente incluir 4 modelos, aunque esto lo veremos más adelante dependiendo de la eficacia del modelo para predecir según los diferentes modelos que hemos explicado, así que en próximas entregas lo definiremos finalmente valorando los resultados.

En cualquier caso, tenemos los datos cargados correctamente siguiendo el formato dicho a la espera solamente de esta cuestión, pero que actualmente no es urgencia, ya que estamos pendientes del entrenamiento del modelo.

4

EXPLORACIÓN DE LOS DATOS

Pasando ahora al análisis exploratorio de las diferentes variables, tras haber estudiado los resultados podemos llegar a diferentes conclusiones:

En general no hay muchos outliers, aunque los que salen como outliers realmente no lo son, pero por ejemplo en la variable ownG, como no es algo normal que haya goles en propia, realmente cuando los hay pues lo marca como si fuese outlier, pero realmente no lo es, sino que es un dato atípico.

En cuanto a forma de las gráficas de cada columna, muchas de ellas tienen la misma forma de muy altas cuando hay pocas unidades de la columna, y pocas cuando hay muchas, lo cual parece que tiene sentido, ya que no es normal tener 15 tarjetas amarillas por ejemplo, también se podrá aplicar en la mayoría de las clases. Sin embargo, hay excepciones, como porcentaje de acierto de pase que casi parece que sigue normalidad según el dibujo, tasa media de pases, edad, altura, que es la única que realmente con el test cumple, titularidades y las posiciones de los equipos.

Solo la variable height sigue una distribución normal, aunque estas últimas columnas explicadas en el párrafo anterior se quedan bastante cerca de serlo, es necesario tener en cuenta que es bastante difícil que sean normales, y el intentar lograr que lo sean nos podrían dar una vista perjudicada de los datos, puesto que en la mayoría de los casos solo muy pocos logran tener muchísimos goles, o minutos por poner dos ejemplos, por lo que intentar una transformación sobre estas variables desorientaría totalmente nuestras predicciones de los valores reales.

Otro punto a ver son las correlaciones, aunque tendríamos que considerar si son significativas o no, SpG con goals, keyP con assists y las posiciones de los equipos en los diferentes años, todas ellas significativas, por lo que podemos decir que hay una correlación alta con certeza entre estas variables.

Por último, las qqplot nos confirman otra vez el punto 1 y 2, tras ver la forma de las colas y la línea que siguen en una distribución.

Un último párrafo dedicado a entender la variable respuesta, esta tiene una forma muy parecida a variables como goles, ya que siguiendo los puntos explicados anteriormente tener un precio muy alto solo lo logran unos pocos, así como también encontramos muchos valores atípicos, siguiendo también este principio.

La variable respuesta tiene una correlación considerable con minutos, goles, asistencias, número de tiros, jugador del partido, pases clave, offdrb o media de pases, casualmente coinciden con muchas de las variables que consideramos importantes, estas variables aportarán un peso alto al rendimiento del modelo.