Creación e implementación de un modelo de goles esperados

Freddy F. Tapia C.

Abril 2023

Planteamiento del proyecto

- Creación de un modelo de goles esperados.
- Se usará una base de libre acceso de Statsbomb.
- La implementación será mediante una app Shiny.

Búsqueda y datos

- Para acceder a la data se usará el paquete “StatsbombR”.
- A la fecha existen 10 competiciones disponibles.
- 4 competiciones son del fútbol femenino en diferentes categorías.

Búsqueda y datos

- La Liga cuenta con 17 temporadas.
- De cerca le sigue la Champions League que cuenta con 15 finales.
- Existen información sobre las copas mundiales del 2018 y 2022.

Búsqueda y datos

- La data posee información sobre más de 30 eventos.
- Existen funciones en el paquete StatsbombR que ayudan a la extracción y procesamiento de la data.
- Se usará el evento “Shot”.

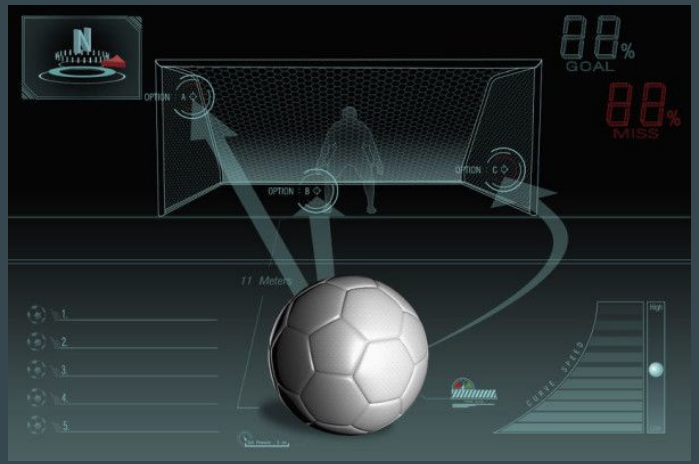
Variable objetivo

- Para definirla se usará la variable de nombre “shot.outcome.name”.
- De sus 8 categorías, no se usarán 5 de ellas.
- Se creará una variable binaria donde “1” indica que el remate fué gol.

Limpieza de datos - Filtros

- Recodificación variables “under_pressure” y “shot.first_time”.
- Eliminación de valores nulos en variables que indiquen coordenadas.
- Filtro por género, sólo se consideran competiciones masculinas.

Limpieza de datos - data final



- La data final cuenta con 7.242 observaciones y 25 variables.
- En el modelo sólo se usarán 9 de ellas.
- 5 son categóricas y 4 numéricas.

Limpieza de datos - IV

- Las variables numéricas son las más importantes, de las cuales el ángulo de tiro es la que destaca.
- La variable categórica más importante es “position.id”, está en el 8vo puesto.

Variable	IV
shot.angle	0,6357
DistToGoal	0,5589
location.x	0,3429
location.y	0,1536
shot.first_time	0,1126

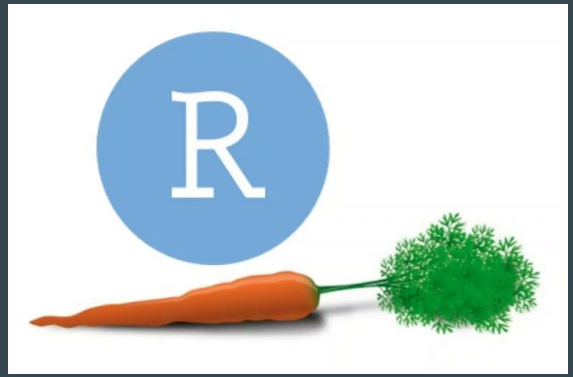
Limpieza de datos - análisis bivariado

- Se muestran las tres variables más importantes, la última columna muestra el porcentaje donde existe una mayor cantidad de goles.

CAT	Gol	No_Gol	TOTAL	var	por_gol
(28.4,168]	1331	1188	2519	shot.angle_cat	52,84
(16.3,28.4]	839	2073	2912	shot.angle_cat	28,81
[0.994,16.3]	191	1620	1811	shot.angle_cat	10,55
[0.7,11]	1032	779	1811	DistToGoal_cat	56,99
(11,16.8]	786	1433	2219	DistToGoal_cat	35,42
(16.8,80.2]	543	2669	3212	DistToGoal_cat	16,91
1	1060	1406	2466	shot.first_time	42,98
0	1301	3475	4776	shot.first_time	27,24

Modelo analítico

- El modelo a emplear es un xgboost, para ello se empleará la librería caret.
- Se realizarán dos versiones.
- La primera es usando los hiper parámetros que recomienda la función, la segunda es optimizando los mismos.



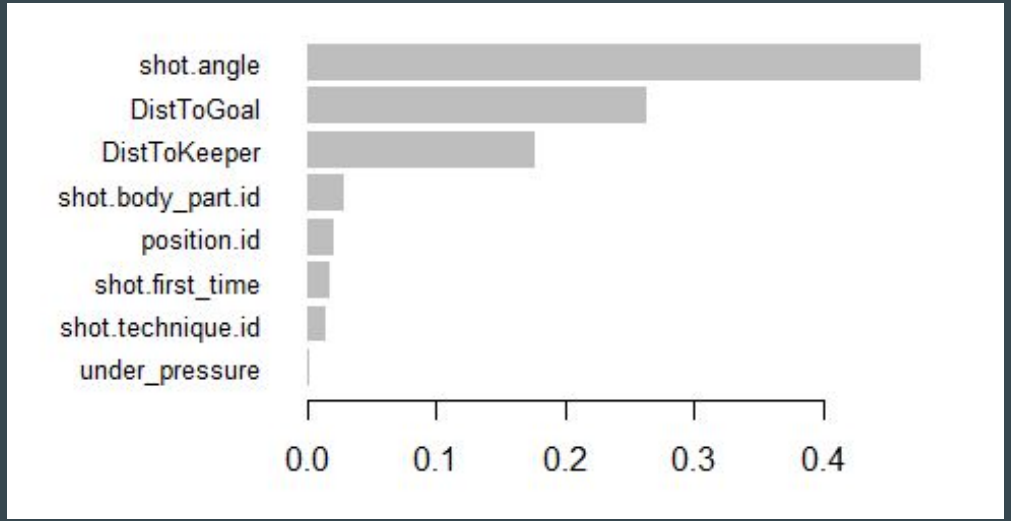
Modelo analítico - métricas

- Se usaron las métricas del Gini, Ks y AUC para elegir al mejor modelo.
- El modelo seleccionado es el inicial por ser el más conservador.

Modelo	AUC	GINI	KS	Media Score
StatsbombR	0,775	0,55	0,4098	0,1689
Inicio	0,7636	0,5271	0,4021	0,3287
Optimizado	0,7643	0,5287	0,4006	0,4059

Modelo analítico - Importancia de las variables

- Las variables numéricas son las más importantes según el modelo.
- La parte del cuerpo con la que se remata destaca dentro de las variables categóricas.



Modelo analítico - SHAP

- Los valores Shap permiten tener una mayor interpretabilidad sobre las variables.
- Valores positivos indican que la variable afecta de forma positiva al score final del modelo.

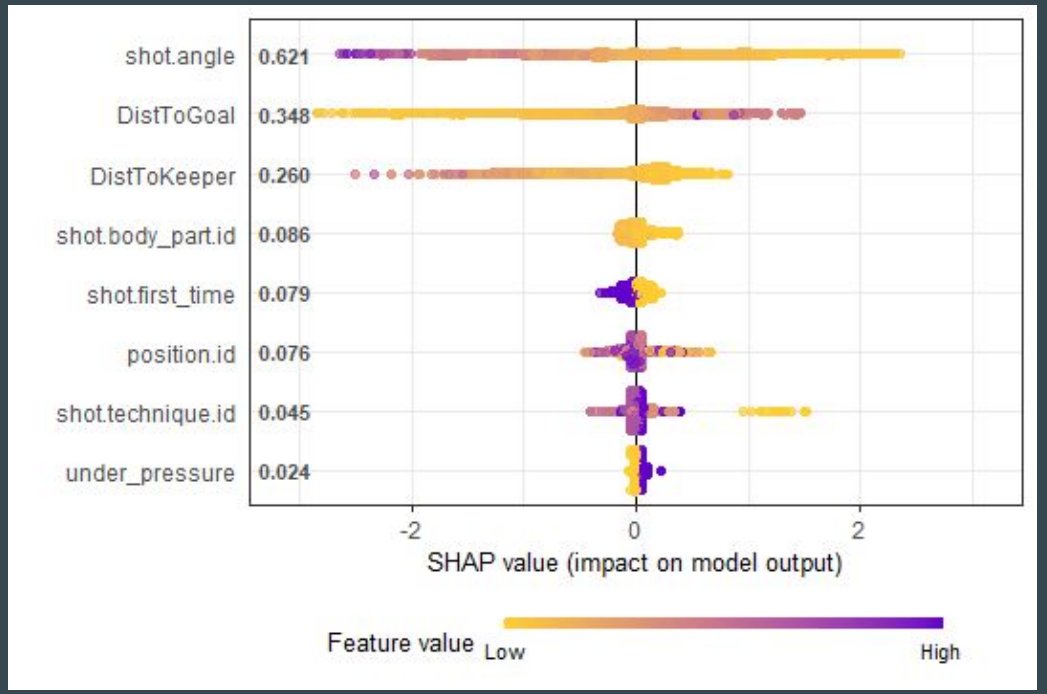
Modelo analítico - SHAP

- Un resumen de la importancia, IV y media de los valores SHAP se presenta en esta tabla:

Feature	Media SHAP	Importancia	IV	tipo
shot.angle	0,6208	0,4755	0,6358	numérico
DistToGoal	0,3478	0,2624	0,5589	numérico
DistToKeeper	0,2600	0,1772	0,0832	numérico
shot.body_part.id	0,0857	0,0286	0,0068	categórico
position.id	0,0761	0,0207	0,0644	categórico
shot.first_time	0,0785	0,0180	0,1126	categórico
shot.technique.id	0,0452	0,0155	0,0433	categórico
under_pressure	0,0235	0,0021	0,0055	categórico

Modelo analítico - SHAP

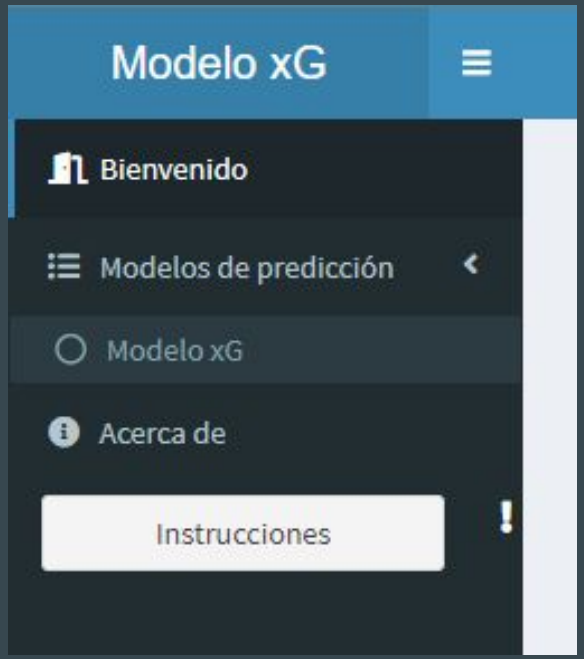
- Valores bajos de la variable “shot.angle” genera valores shap positivos.
- Esto indica que afecta positivamente al score final del modelo.



Visualización - Shiny app

- La implementación del modelo será mediante una Shiny app.
- El paquete Shiny de R permite generar aplicaciones web interactivas.
- Aquí se puede observar una galería de app gratuitas.

Visualización - secciones Shiny app



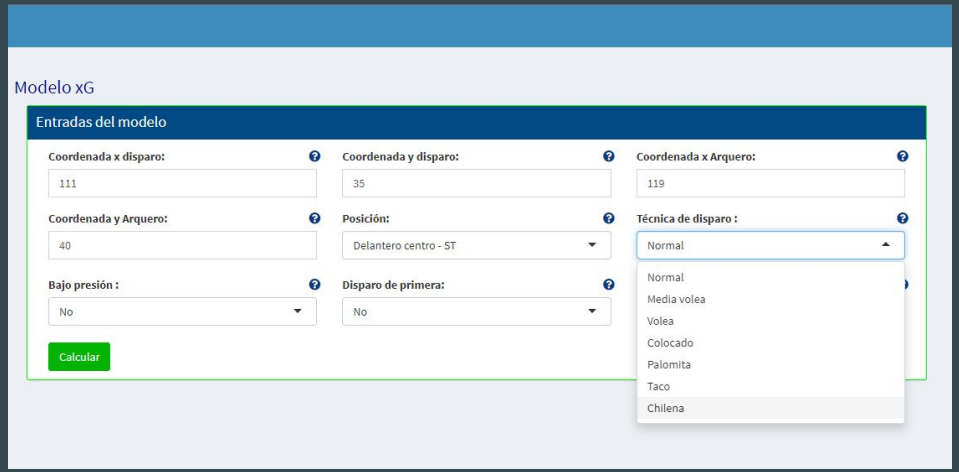
➔ La Shiny app cuenta con 3 secciones,

- Sección de bienvenida.
- Sección del modelo.
- Sección acerca de.

Visualización - sección de bienvenida

- ➔ Contiene información sobre:
  - Data usada.
  - Metodología.
  - Métricas del modelo.
  - Implementación.

Visualización - sección modelo - entradas

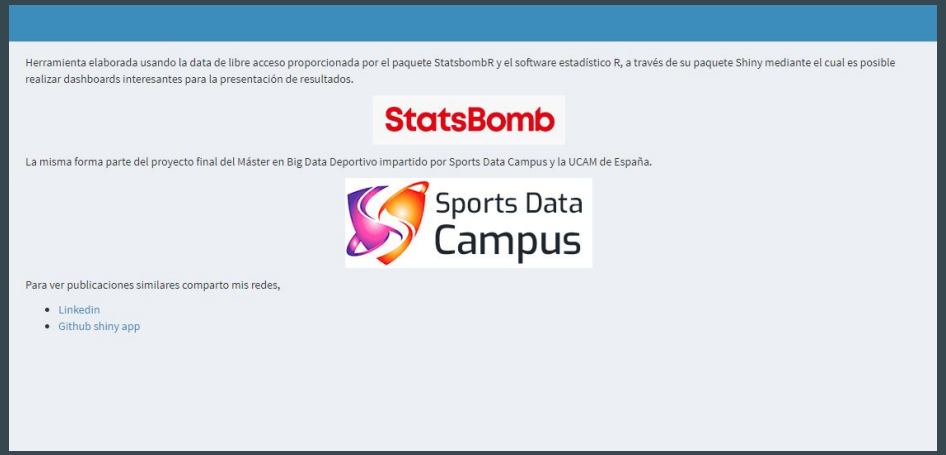


- ➔ Cuenta con las variables que el modelo espera recibir.
- ➔ Tiene una descripción de cada variable.
- ➔ La variable parte del cuerpo se condiciona por la técnica.

Visualización - sección modelo - salidas

- Cuenta con una sección de advertencias.
- Se muestran variables calculadas internamente.
- Se genera un gráfico del remate considerado.

Visualización - sección acerca de



- ➔ Cuenta con información sobre el desarrollador.
- ➔ Tiene enlace del repositorio en Github donde se encuentra la app.

Visualización - enlace Shiny app

- En el siguiente link se encuentra la Shiny app.
- Se utilizó la opción que ofrece Shiny para subir la app a la nube.
- Es una manera interactiva de probar el modelo.

Conclusiones y recomendaciones

- El modelo implementado usa pocas variables con el fin de que pueda ser usado sin la necesidad de tener mucha información.
- Mejores versiones del modelo pueden ser creadas usando la totalidad de variables disponibles en la data.
- Con la base creada es posible crear otros modelos de predicción, como por ejemplo el del xT ó expected Threat.