Proyecto final del Bootcamp de Ciencia de Datos de Código Facilito

Guillermo Sangabriel Cuéllar

Creación de un modelo de Goles Esperados

Durante el Bootcamp se mostró el uso de la ciencia de datos en distintas disciplinas, para este proyecto decidí aplicar lo aprendido en el fútbol.

Aunque existen muchas cuestiones a analizar dentro de un partido de fútbol como las trayectorias de los jugadores o la geometría del control de juego por parte de los equipos, en este caso se analizarán los goles esperados. Los goles esperados expresan la probabilidad de que un tiro se convierta en gol.

El primer reto fue conseguir los datos para la elaboración de este proyecto, estos se consiguieron gracias a los datos abiertos de statsbomb. Una vez que se encuentra disponible el conjunto de datos, la primera pregunta que se tuvo que resolver fue: ¿De toda la información cuál es la relevante para elaborar nuestro modelo? Esta pregunta se pudo resolver gracias a la misma documentación de statsbomb. Luego se compararon la frecuencia de tiros con la frecuencia de los goles y, como se puede ver en el notebook correspondiente, los goles son eventos escasos comparados con la cantidad de tiros. El dataset es muy interesante y con él se pueden hacer cosas como graficar los tiros en un campo de juego. Gracias a visualizaciones como la anterior pude encontrar que, tanto los tiros como los goles ocurren con mayor frecuencia cerca de la portería, sin embargo, existe un límite de cercanía ya que no es sencillo rematar dentro de las primeras 5 yardas a partir de la línea de gol. Un hallazgo muy curioso es que en todo el dataset hay tres tiros directos a partir del tiro de esquina y todos terminaron en gol, aunque este hecho no se incluyó en los notebooks debido a que no encajaba con la narración que se llevaba.

De acuerdo a diversas fuentes, la distancia y el ángulo de tiro son features relevantes para el modelo que intentamos implementar y, aunque estas no estaban disponibles en los datos originales fue posible deducirlas usando los conceptos de distancia entre dos puntos y la ley de cosenos (sí, se tuvieron que desempolvar los conocimientos de trigonometría).

Seguía preguntarse cómo afectaban las distintas features que teníamos disponibles a las proporciones entre los tiros y los goles, estos resultados se pueden encontrar en el notebook de limpieza y análisis de datos.

Para determinar las probabilidades buscadas se implementa la regresión logística con una grilla de posibles parámetros para que se determinara el mejor modelo. Sin embargo, los resultados no fueron muy alentadores cuando se analizó el desempeño. Después, se uso el Random Forest Classifier, aunque su desempeño no varió significativamentes respecto a la regresión logística.

Aquí les dejo los links al repositorio, mi correo personal y mi cuenta de twitter para cualquier comentario o aclaración.

- repositorio
- correo: omemgsc@gmail.com
- twitter: @omemgsc

Últimas consideraciones

A pesar de que los modelos de machine learning que se implementaron no dieron resultados satisfactorios, es posible que su desempeño mejore si se eliminan algunas features que no aportan mucha información, esto lo intentaré en un futuro cercano.

Por último me gustaría indicar que el proyecto se encuentra dividido en tres notebooks, uno para la descarga de los datos, otro para la transformación y el análisis de los datos y por último un notebook para la implementación de los modelos de ML.