



Juan José Fernández Moreno

Goles en la Premier League con machine learning

Análisis y modelos predictivos en el mercado de goles

Presentación del problema de negocio

En el ámbito de las apuestas deportivas, el mercado de **Over/Under 2.5 goles** es uno de los más populares entre los apostantes. Pero la mayoría de este tipo de decisiones se toman en función de:

- Intuiciones o preferencias personales
- Sensaciones recientes
- Lo que apuesta la mayoría
- Aleatoriedad

Sin embargo, los datos históricos pueden ofrecer una **base mucho más sólida** para tomar decisiones objetivas.

Objetivo

Desarrollar un sistema que, utilizando datos estadísticos previos al partido, permita predecir con fiabilidad si se superarán los 2.5 goles en un encuentro de la Premier League.

- Apostar con mayor criterio
- Minimizar el componente del azar
- Aumentar la probabilidad de acierto
- Potencialmente, **mejorar la rentabilidad**

No se busca **ganar siempre**, sino **apostar cuando hay una probabilidad real superior a la de las casas de apuestas**.



Datos y estadísticas utilizadas

Dataset: Datos históricos de partidos de la Premier League desde la temporada 1993-1994.

Datos utilizados: desde el año 2014, un total de **3120 partidos de entrenamiento y 780 de testeo**.

Estadísticas: relacionadas con el potencial ofensivo y otras interesantes.

Tratamiento de datos

TARGET: goles totales → Convertido a binario para más (1) o menos (0) de 2.5 goles.

Limpieza de features: eliminación de features irrelevantes como *FullTimeResult*, *MatchID*, *MatchWeek*, *Date*, *Time*, *HalfTimeResult*, *HalfTimeGoals* o *Referee*.

Feature engineering: generación de features más interesantes como precisión, efectividad, media de goles anotados o encajados (*en la temporada*) o rachas goleadoras previas al partido.

Distribución del target



Decisiones técnicas

Utilizar la **media de goles anotados y recibidos** por equipo, temporada y estadio (*si juega como local o visitante*) en lugar de la media acumulada en los partidos previos, con el objetivo de obtener un dato más representativo y realista del rendimiento de un equipo a lo largo de la temporada.

De la misma forma, se ha usado esta media con las nuevas features generadas, con el objetivo de obtener datos más consistentes.

Features seleccionadas

Después del análisis de correlación, las features seleccionadas para el entrenamiento han sido:

- HomeTeamShots
- AwayTeamShots
- HomeTeamShotsOnTarget
- AwayTeamShotsOnTarget
- ShotsPrecisionHomeTeam
- ShotsPrecisionAwayTeam
- GoalsTotalShotsHomeTeamRatio
- GoalsTotalShotsAwayTeamRatio
- EfectivityHomeTeam
- EfectivityAwayTeam
- GoalsPerMatchMeanHomeTeam
- GoalsPerMatchMeanAwayTeam
- GoalsReceivedPerMatchMeanHomeTeam
- GoalsReceivedPerMatchMeanAwayTeam

Features apenas relevantes incluidas en otro dataframe de entrenamiento junto a las demás:

- RachaOverHomeTeam
- RachaOverAwayTeam
- RachaUnderHomeTeam
- RachaUnderAwayTeam

MODELOS

Para el entrenamiento, se han utilizado **4 modelos para el problema de clasificación:**

- Randon Forest
- Gradient Boosting
- LightGBM
- XGBoost

Métrica elegida: f1-score de la clase 1 (más de 2.5 goles)

¿Por qué? Se busca un equilibrio entre la precisión y el recall de la clase 1, es decir, minimizar los fallos y minimizar el perder oportunidades de apuestas.

ENTRENAMIENTO

- **Entrenamiento de modelos base**, usando los dos dataframe: el dataframe con las features seleccionadas y otro añadiendo además las rachas.
- **Se guardan los resultados de cada modelo en cada dataframe** con el objetivo de comparar y elegir la mejor sinergia entre modelo-dataframe.
- **Con *optuna*, se optimizan los hiperparámetros de cada modelo** entrenándolos con el dataframe que mejor métrica han dado.
- **Se guardan los resultados de cada modelo** para comparar.



Resultados base sin rachas

Resultados base con rachas

Resultados finales optimizados

Model	Precision_0	Precision_1	Recall_0	Recall_1	F1_0	F1_1
RandomForestClassifier	0.64	0.68	0.63	0.69	0.64	0.69
GradientBoostingClassifier	0.66	0.69	0.63	0.72	0.65	0.71
LGBMClassifier	0.64	0.68	0.62	0.69	0.63	0.68
XGBRFClassifier	0.64	0.67	0.60	0.71	0.62	0.69

Model	Precision_0	Precision_1	Recall_0	Recall_1	F1_0	F1_1
RandomForestClassifier	0.65	0.70	0.65	0.69	0.65	0.70
GradientBoostingClassifier	0.64	0.68	0.62	0.71	0.63	0.69
LGBMClassifier	0.64	0.69	0.64	0.69	0.64	0.69
XGBRFClassifier	0.64	0.67	0.60	0.71	0.62	0.69

Modelo	Precision_1	Recall_1	F1_1
RandomForest	0.68	0.72	0.70
GradientBoosting	0.69	0.74	0.71
LightGBM	0.68	0.74	0.71
XGBRFClassifier	0.62	0.91	0.74

Conclusiones

- Las medias por temporada estabilizan las métricas y evitan ruido
- Las estadísticas ofensivas son buenos predictores del over
- El modelo es capaz de detectar con relativa fiabilidad partidos con más de 2.5 goles, en un mercado que muchas veces ronda el 50% de probabilidad

Mejoras

- Scraping de estadísticas actualizadas de cada equipo para realizar un sistema de predicciones automatizado basado en los modelos obtenidos
- Realizar un entrenamiento más exhaustivo de cada modelo, optimizando mejor hiperparámetros, features...