

Modelos y Simulación de Sistemas 1
Proyecto - Entrega #1

Andrés Darío Peña
Andrés Ramirez Aristizabal
Leider Caicedo Palacios

Docente: Raúl Ramos Pollán

Universidad de Antioquia
Facultad de ingeniería

Septiembre 2023

1. Describas el problema predictivo a resolver.

Predecir los resultados de un torneo de baloncesto universitario masculino en Estados Unidos en el año 2017. Utilizaremos los resultados recopilados de torneos pasados para poder construir y probar diferentes modelos de predicción. Después pronosticamos los resultados de todos los enfrentamientos posibles en el torneo de 2017.

2. Dataset que utilizaremos.

El dataset que vamos a utilizar es el de Machine Learning mania, un dataset basado en la competición de basketball NCAA del 2017. (link a continuación) [https://www.kaggle.com/competitions/march-machine-learning-mania-2017/data?select=Regular Season Detailed Results.csv](https://www.kaggle.com/competitions/march-machine-learning-mania-2017/data?select=Regular+Season+Detailed+Results.csv), la cual posee 34 columnas con 76.600 instancias.

El archivo que contiene los datos se llama: ***RegularSeasonDetailedResults.csv***

Este archivo contiene los resultados del juego por juego para 32 temporadas de datos históricos, de 1985 a 2015. Cada año, incluye todos los juegos jugados desde el día 0 hasta el 132.

"season" - Este es el año de la entrada asociada en seasons.csv (el año en que ocurre el torneo final)

"daynum" - Este número entero siempre varía de 0 a 132 y te dice en qué día se jugó el juego.

"wteam" - identifica el ID del equipo que ganó el juego.

"wscore" - identifica la cantidad de puntos anotados por el equipo ganador.

"lteam" - identifica el ID del equipo que perdió el juego.

"lscore" - identifica la cantidad de puntos anotados por el equipo perdedor.

"numot" - esto indica el número de períodos de tiempo extra en el juego, un número entero 0 o superior.

También las estadísticas totales a nivel de equipo para cada juego.

wfgm - goles de campo realizados

wfga - intentos de goles de campo

wfgm3 - tres punteros hechos

wfga3 - intento de tres punteros

wftm - tiros libres hechos

wfta - intento de tiros libres

wor - rebotes ofensivos

wdr - rebotes defensivos

wast - asiste

wto - pérdidas de balón

wstl - roba
wblk - bloques
wpf - faltas personales

3. Las métricas de desempeño requeridas (de machine learning y de negocio)

Se usará la métrica de desempeño LogLoss:

$$\text{LogLoss} = -\frac{1}{n} \sum_{i=1}^n [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)],$$

La cual nos permitirá medir una probabilidad de que un caso ocurra o no, en el caso de este dataset se medirá la probabilidad de victoria de un equipo con respecto a otro, con las miras de encontrar el equipo ganador, donde menores valores de LogLoss nos mostrará con más precisión que el equipo elegido tiene una mejor probabilidad de ganar que el equipo contrario. Como métrica de negocio, se podrá usar cuando se habla de temas como ganancia de dinero en términos de apuestas a un equipo ganador o la medición del desempeño general entre los equipos enlistados en el torneo para la planeación de estrategias futuras.

4. Un primer criterio sobre cuál sería el desempeño deseable en producción.

Con este modelo esperamos predecir con un alto nivel de precisión el resultado de los partidos del torneo teniendo en cuenta resultados obtenidos en anteriores torneos. Esta información puede ser muy útil para realizar apuestas que permitan obtener una alto nivel de ganancias.

Para que al apostador le resulte eficiente y le genere ganancias el uso del dataset el margen de error de este debe ser menor a 0.5.