Un método simple para predecir descensos tras la primera jornada de liga basado en probabilidades condicionadas

Article · November 2015

CITATIONS
0

READS
389

1 author:

Jose A. Martinez
Universidad Politécnica de Cartagena
139 PUBLICATIONS **2,009** CITATIONS

SEE PROFILE

REVISTA INTERNACIONAL DE DEPORTES COLECTIVOS

UN MÉTODO SIMPLE PARA PREDECIR DESCENSOS TRAS LA PRIMERA JORNADA DE LIGA BASADO EN PROBABILIDADES CONDICIONADAS

A SIMPLE METHOD FOR PREDICTING RELEGATIONS AFTER THE FIRST MATCH OF THE SEASON BASED ON CONDITIONAL PROBABILITIES

José Antonio Martínez[1]

[1]Universidad Politécnica de Cartagena, España. E-mail: josean.martinez@upct.es.

RESUMEN

El objetivo de este estudio es aplicar un método muy sencillo, manejable y entendible por cualquier analista, que con información muy limitada resulte útil en predecir la probabilidad de descenso de un equipo de fútbol después de jugar el primer partido de la competición. A través del uso del Teorema de Bayes, y partiendo de información a priori sobre el potencial de los equipos, se puede calcular la probabilidad de descender tras la primera jornada de liga. El método se describe didácticamente y se discuten diferentes aplicaciones del mismo.

PALABRAS CLAVE: predicción, probabilidades condicionadas, Teorema de Bayes, fútbol.

ABSTRACT

The aim of this research was to apply a straightforwards and easily understandable method to predict the probability of relegations of football teams after the first match of the season. Once established the priors about the potential of teams, and after using the Bayes Theorem, the probability of relegation can be computed, even if the information is very limited. The method is described in a didactic way and different extensions of its applications are finally discussed.

KEYWORDS: prediction, conditional probabilities, Bayes Theorem, football.

Fecha de recepción: 12/01/2015 ✷ Fecha de aceptación: 05/03/2015

1. INTRODUCCIÓN

En los últimos años, diversas investigaciones han puesto énfasis en la importancia de la detección precoz de estados problemáticos en los equipos a través del análisis estadístico del marcador dentro de un mismo partido[1], o de los resultados de las primeras jornadas de la liga[2].

Así, en cuanto al desarrollo del marcador de un mismo partido de baloncesto, sería preferible comenzar ganando, ya que la dependencia del estadio temporal anterior del resultado es muy alta. Ir ganando en el primer cuarto es fundamental para obtener la victoria en el partido, si bien ello también depende, aunque en mucha menor medida, de si el equipo juega o no en casa y de la diferencia de potencial entre los contendientes. De este modo, para los equipos visitantes una ventaja de 10 o 15 puntos no garantiza la victoria al final del partido si el equipo local es más fuerte que ellos, mientras que si el visitante es el más fuerte, entonces no debe confiarse, ya que ventajas cortas del equipo local pueden ser irrecuperable.

En cuanto a la influencia de los resultados iniciales en la clasificación final, en el caso del fútbol, para los equipos con los presupuestos más bajos, los rendimientos obtenidos en los primeros partidos de la competición tienen una influencia muy destacada sobre su clasificación final. Por tanto, esos equipos deberían de tratar de comenzar lo más fuerte posible la temporada, diseñando para ello una pretemporada acorde a esas metas.

Sin embargo, no hemos encontrado ninguna investigación que lleve la detección precoz de estados problemáticos para los equipos hasta el extremo temporal de la primera jornada de competición. El conocer desde prácticamente el inicio de la liga si un equipo es más proclive al descenso dotaría a esos equipos de una información valiosa para desde ese mismo momento trataran de darle un vuelco a esa situación.

En muchas ocasiones, tras la primera jornada de liga, los analistas deportivos e incluso los propios entrenadores o jugadores no quieren realizar valoraciones sobre el

[1] MARTÍNEZ, J. A. La influencia del primer cuarto en el resultado final en baloncesto. *Revista Internacional de Medicina y Ciencias de la Actividad Física y el Deporte*. 2011, núm. 14 (56), pp. 755-769.
[2] LAGO, C. y CASÁIS, L. La influencia de los resultados iniciales en la clasificación final de los equipos de fútbol de alto nivel. *Motricidad. European Journal of Human Movement*. 2010, núm. 5 (14), pp. 107-122.

rendimiento futuro de los equipos: "*Es sólo el primer partido*". Pero lo que esta investigación muestra en el ámbito del fútbol, es que el resultado de la primera jornada de liga es un indicador de la probabilidad que tiene un equipo de descender al final de la competición. De este modo, y manjenado información extremadamente simple, se puede conocer la probabilidad de descender tras ese primer partido.

El objetivo de este estudio es aplicar un método muy sencillo, manejable y entendible por cualquier analista, que con información muy limitada resulte útil en predecir la probabilidad de descenso. A través del uso del Teorema de Bayes, y partiendo de información a priori sobre el potencial de los equipos, se puede calcular la probabilidad de descender tras la primera jornada de liga.

2. MÉTODO Y RESULTADOS

Se construyó una base de datos con los resultados de los partidos de la primera jornada de liga, tomados de www.linguasport.es. Las temporadas a considerar fueron desde la 1929-30 hasta la 2010-11. Para ello se registraron todos los partidos de las primeras jornadas de liga de esas temporadas, el margen de goles por el que un equipo perdía o ganaba (siendo 0 en caso de empate), y si el equipo en cuestión descendía o no al final de la temporada. En caso de que en uno de esos partidos ambos equipos se identificaran como clubes que iban a descender se eliminaba ese partido de la base de datos.

De este modo, se obtuvieron 606 registros, es decir, 606 equipos que participaban en el primer partido de la temporada y de los cuales descendían un 24.91% de ellos. Hay que tener en cuenta que ese porcentaje no significa que desciendan 1 de cada 4 equipos de la liga, sino que en 1 de cada 4 partidos analizados estaba involucrado un equipo que descendía.

Alternativas de modelización

Aunque parezca una idea cuestionable, el resultado del primer partido de liga está asociado a la probabilidad de descender. Una primera modelización muy sencilla nos ilustra este hecho.

Si tomamos la variable "descender" como una variable binaria (0,1), donde 0 significa permanencia y 1 significa descenso, entonces podemos implementar un simple modelo de regresión logística incluyendo el margen de goles como variable

independiente. Así, por ejemplo, si el equipo perdía 2-0 el margen sería -2, y si el equipo ganaba 3-1 el margen sería +2.

Los resultados de este modelo nos indican que la probablidad de descender frente a no descender se incrementa a medida que lo hace el margen de goles en contra, o disminuye a medida que ese margen se hace más pequeño hasta convertirse en positivo. Dicho de otro modo, cuanto más abultada sea la derrota en el primer partido, mayor es la probabilidad de descender, y cuanto más abultada sea la victoria menor es la probabilidad de descender a final de la temporada (Tabla 1).

Tabla 1. Regresión logística del margen de goles sobre la probabilidad de descenso

	Coeficiente
Margen de goles	-.46**
Constante	-1.08**
Pseudo R2: .11**	LR chi2: 72.79**

***p*<.0

Este modelo es obviamente simplista, sólo clasifica correctamente el 63.20% de los casos, con una sensitividad (porcentaje de descensos correctamente predichos) del 82.12%. Sin embargo, es un modelo que no se mejora al añadirle variables cuadráticas o cúbicas, incluso cuando se le añada una variable que tenga en cuenta si el equipo en custión ha jugado contra el Real Madrid o el Barcelona (a priori podría pensarse que una derrota frente a estos equipos no significa lo mismo que frente al resto); los valores del BIC y AIC son similares. Por tanto, el modelo más simple sería el más adecuado, si se hacen este tipo de comparaciones.

No obstante, el problema de esta clase de aproximaciones tan simplistas es la especificación correcta del modelo. Ninguno de esos modelos pasa el test de Hosmer-Lemeshow (si este test arroja un valor no significativo apoyaría la idea de que los valores observados y predichos concuerdan de manera similar en los diferentes deciles), ni el test de especificación (se usa el predictor linear y su cuadrado para estimar de nuevo el modelo. Para apoyar la buena especificación del modelo el predictor lineal tiene que ser significativo y su cuadrado no, lo que no ocurre en nuestro caso). Probablemente la causa de ese problema de especificación es que hay causas omitidas, como el potencial de los equipos, correlacionadas con el marcador en el primer partido.

Alternativas no paramétricas como los k-NN (k-vecinos próximos) podrían ser una opción, pero en este caso tenemos la consideración de que la variable independiente

tiene muy poca variabilidad (es prácticamente una variable discreta), por lo que el algoritmo de cómputo no tiene claro en muchas ocasiones qué opción manejar. Por ejemplo, si el k-NN óptimo fuera 3 y un equipo perdiera por el margen de -2 goles, entonces no sólo habría que considerar la predicción en los vecimos más próximos (-1, -2 y -3 goles), sino dentro de cada nivel la distribución de descensos. Aunque se podría buscar una solución a este problema, sin duda la complejidad asociada a este procedimiento limitaría su aplicación.

Probabilidades condicionadas

El uso de las probabilidades condicionadas, base de la estadística bayesiana, nos da una solución muy sencilla a los problemas anteriores de emplear métodos "frequentistas" cuando la información objetiva es muy escasa (sólo tenemos un resultado en la primera jornada, nada más).

De este modo, se puede construir el esquema básico de la probabilidad condicionada, basado en el Teorema de Bayes: $P(Bp|Ai)=P(Bp)\ P(Ai|Bp)/\ \Sigma P(Bp)\ P(Ai|Bp)$. Es decir, podemos conocer la probabilidad de que ocurra el evento B dado el evento A, a partir de conocer la probabilidad condicionada de A dado B y la probabilidad de B. El subíndice "p" indica las diferentes clases de esa variable en la población (mutuamente excluyentes), y el subíndice "i" los diferentes niveles de la variable dependiente. Este aparente juego de palabras se aclara bastante cuando tomamos un ejemplo concreto. En nuestro caso estamos interesados en conocer la probabilidad de que un equipo descienda tras conocer el resultado del primer partido de liga. Por tanto:

P(descender|marcador) = P(descender) P(marcador|descender) /
[(P(descender) P(marcador|descender)) + (P(No descender) P(marcador| No descender))]

Los términos de la ecuación anterior son los siguientes:

1. P(descender|marcador): Es la probabilidad de que un equipo descienda tras conocer el marcador de la primera jornada. Es la probabilidad *a posteriori*, es decir, la que queremos calcular.

2. P(descender): Es la probabilidad de descenso del equipo en cuestión. Es una probabilidad *a priori,* y tiene que ser indicada por el investigador a partir de información objetiva y/o subjetiva.

3. P(marcador|descender): Es la probabilidad de que se de ese marcador cuando un equipo desciende. También se suele identificar con la *verosimilitud*, y es simplemente una probabilidad inferida de los datos empíricos que tenemos, que tiene que ser evaluada para cada marcador. Dicho de otro modo, si hay 95 equipos que han perdido por 1 gol, y de ellos 39 descienden, la probabilidad es de .41. Esto se deriva de los datos empíricos y suele ser las probabilidades que se calculan desde la óptica frecuentista. Como veremos, la perspectiva bayesiana modifica esta probabilidad en función de las probabilidades a priori, dando la probabilidad a posteriorir, que es la que realmente nos interesa.

4. [(P(descender) P(marcador|descender)) + (P(No descender) P(marcador| No descender))]: Es simplemente una forma de ponderar el numerador de la ecuación, con el fin de que la probabilidad a posteriori esté entre 0 y 1. Para ello, hay que calcular las probabilidades en los diferentes niveles de la variable "descender". Dado que en este caso sólo tiene 2 niveles (descenso o permanencia), el cómputo es muy sencillo. Por ejemplo, para el comentado caso de que desciendan 39 de los 95 equipos que han perdido por 1 gol, entonces P (No descender) = 1- P(descender); y la P(marcador| No descender) = 1- P(marcador|descender) = 1 – .41 = .59.

Con todos estos elementos podemos realizar ya los cálculos pertinentes, aunque por motivos ilustrativos vamos a detallar en una tabla todos esos cálculos necesarios, con el fin de que quede clara la sencillez del método (Tabla 2)

Tabla 2. Tabla de probabilidades condicionadas para obtener la probabilidad de descender dado el marcador de la primera jornada

		Marcador (margen de goles)						
		+3 o más	+2	+1	0	-1	-2	-3 o más
	n	77	84	125	136	95	55	34
	descensos	7	9	11	41	39	27	17
P(A|B)	P(marcador|descender)	.09	.11	.09	.30	.41	.49	.50
	95% IC	(.04 ; .18)	.05 ; .19)	(.04 ; .15)	(.23 ; .39)	(.31 ; .52)	(.35 ; .63)	(.32 ; .68)
P(B)	P(descender)	.86	.86	.86	.86	.86	.86	.86
P(B)*P(A|B)	P(descender)*P(marcador|descender)	.08	.09	.08	.26	.35	.42	.43
P(¬B)*P(A|¬B)	P(No descender) P(marcador| No descender)]	.13	.13	.13	.10	.08	.07	.07
P(B)*P(A|B) + P(¬B)*P(A|¬B)	[P(descender) P(marcador|descender) + P(No descender) P(marcador| No descender)]	.21	.22	.20	.36	.44	.49	.50
P(B|A)	P(descender|marcador)	.38	.42	.37	.73	.81	.86	.86

Nota: P(B) ha sido determinada como .86, en consonancia con el ejemplo comentado del Real Murcia.

Para ello hemos dividido la distribución del margen de goles agrupando aquellos niveles de la variable con pocas observaciones. Como la calidad de las predicciones vendrá dada por la calidad de las verosimilitudes calculadas, es decir, de las probabilidades que nos da el análisis del histórico de partidos, no es conveniente tener niveles de la variable con muy pocos datos. Sobre cuántas observaciones sería conveniente tener en cada nivel podemos realizar diferentes simulaciones empleando el intervalo de confianza binomial[3]. Esto puede calcularse fácilmente en una hoja de cálculo o automáticamente en progamas como Stata 12.0. Por ejemplo, cuando el margen es +6, tenemos 6 observaciones en la base de datos, de las cuales una de ellas corresponde a un descenso. De este modo la P(marcador|descender) =.16, pero el intervalo de confianza al 95% es (0 ; .64). Esto nos indica que esa probabilidad es muy poco fiable. Para el caso de margen +1, hay 125 observaciones de las cuales 11 son descensos. La probabilidad es de .09, con un intervalo de confianza de (.04 ; .15). Por tanto, esta probabilidad es mucho más fiable que la anterior, ya que la amplitud del intervalo de confianza es mucho más pequeña, debido, principalmente, a la diferencia en el tamaño de las muestras.

El último paso, y seguramente el más controvertido, es asignar un valor a la probabilidad a priori, es decir a P(descender). Posiblemente la crítica más feroz al bayesianismo aplicado a la investigación es tener que confiar en ese tipo de probabilidades "subjetivas", donde es el investigador el que tiene la última palabra para indicar el valor concreto de probabilidad. No es objetivo aquí entrar en discusiones acerca de esta temática, cuyos defensores y detractores llevan discutiendo muchas décadas[4,5]. Sin embargo, en el caso que este estudio que nos ocupa no necesariamente tenemos que asignar subjetivamente esa probabilidad, sino que podemos hacerla en base a criterios más "objetivos" o a una combinación de ambos. Esta es una de las grandes ventajas de la perspectiva bayesiana, la flexibilidad.

Obviamente no todos los equipos tienen la misma probabilidad de descender a comienzos de la temporada. Y esto lo debemos de tener en cuenta a la hora indicar esa probabilidad a priori. Variables como el presupuesto, la inversión en fichajes, el número de temporadas que lleva el equipo en primera división, etc. son indicadores

[3] LEVY, P. S. y LEMESHOW, S. *Sampling of populations.* New York: John Wiley & Sons, 1999.
[4] ZYPHUR, M. J. y OSWALD, F. L. Bayesian probability and statistics in management research: A new horizon. *Journal of Management*. 2013, num. 39 (1), pp. 5-13.
[5] KRUSCHKE, J. K., AGUINIS, H. & JOO, H. The time has come: Bayesian methods for data analysis in the organizational sciences. *Organizational Research Methods.* 2012, núm. 15, pp. 722-752.

del potencial del equipo. A mayor potencial del equipo, menor es la probabilidad de descender.

Por ejemplo, en el caso del equipo Real Murcia, ha estado en Primera División 11 temporadas entre la 1940-41 y la 2007-08. Y de esas 11 temporadas ha descendido el primer año en 6 de ellas. Esto quiere decir que es un equipo débil que suele mantenerse muy poco tiempo en la máxima categoría. A la hora de establecer una probabilidad a priori, podríamos hacerla a partir del porcentaje de temporadas que ha estado fuera de Primera División en las 80 ediciones de la Liga, que es 86%, es decir, la probabilidad de descender sería de .86. Como puede verse en la Tabla 2, una vez establecidad esa probabilidad a priori podemos conocer la probabilidad de descender en función del marcador de ese primer partido. De este modo, si el Real Murcia vuelve a Primera División y gana su primer partido, la probabilidad de descender bajaría prácticamente a la mitad, con lo que sería una información muy importante para establecer valoraciones y predicciones acerca del rendimiento del equipo.

Es indudable que esa probabilidad a priori puede contener también elementos subjetivos. Por ejemplo, sobre ese .86 anterior podemos matizar algunas centésimas o incluso décimas en función de información relevante que se maneje al comienzo de la temporada (fichajes, problemas en el vestuario, opinión de expertos, etc.).

Existen muchas ventajas de emplear este enfoque, ya que nos permite manejar la información de manera más coherente. Mirando estrictamente los datos, si, por ejemplo, el Real Madrid perdiera por 2 goles en el primer partido de la temporada, diríamos que tiene un 49% de probabilidad de descender. Pero esto es totalmente irreal. Para el Real Madrid o el Barcelona podríamos establecer *priors* muy bajos, por ejemplo, de .05, ya que ellos nunca han descendido y tienen los presupuestos más grandes de la competición. Con ese .05 de probabilidad a priori, la probabilidad de descender si pierden por 2 goles de diferencia ya no es .49 sino .048, es decir, 10 veces menor.

3. DISCUSIÓN

La detección precoz es una herramienta muy importante en todas las ramas de la ciencia. El test de Apgar, por ejemplo, para los neonatos pretende detectar problemas de salud del niño a partir de información simple y no invasiva en el momento de nacer. Como hemos indicado, en las Ciencias del Deporte cada vez más se proponen

emplear las herramientas estadísticas para detectar problemas precozmente, y así que los equipos puedan tomar decisiones atendiendo a esa nueva información generada.

En esta investigación hemos visto cómo a través de un sencillo uso de las probabilidades condicionadas, y tras el análisis del histórico de datos de resultados de la primera jornada de la máxima competición del fútbol español, los equipos pueden revisar la probabilidad que a priori tienen de descender, con el sólo hecho de conocer el resultado de su primer partido.

El método aplicado aquí puede complementarse y complicarse mucho más, con el cómputo de intervalos de confianza para las probabilidades a posteriori (que en términos bayesianos se denominan "credible intervals"), empleando distribuciones para los *priors.* Pero esta forma avanzada de utilizar este método requiere de *software* muy técnico como WinBUGS, por ejemplo, que limita muchísimo la aplicación práctica a grupos de analistas muy especializados.

Sin embargo, la simplicidad de la aplicación del Teorema de Bayes permite con una hoja de cálculo actualizar las probabilidades, y que los analistas, demás profesionales e incluso los aficionados hagan revisiones de las probabilidades de descender, realizando asimismo análisis de sensibilidad variando los *priors*. Obviamente, no sólo este método es intereante para detectar estados precarios, sino también para predecir campeones, por ejemplo. Tan sólo hay que preocuparse de ir actualizando la base de datos de partidos temporada a temporada, con lo que las verosimilitudes se van convirtiendo (generalente) en más fiables a medida que crece la muestra de partidos, por lo que las probabilidades a posteriori serán también más acertadas.

Es aconsejable, asimismo, que los *priors* se establezcan combinando información cuantitativa y cualitativa. Ninguna de ellas es, obviamente, perfecta. Los presupuestos de los equipos pueden ser un buen indicador de la calidad de los mismos, pero no siempre ello es así. Por ejemplo, en la NBA, la asociación entre el número de victorias y el volumen salarial de los equipos es muy baja[6]. Por eso, esta perspectiva bayesiana permite acomodar más fácilmente esa información cualitativa, que en el caso de construcción de modelos estadísticos (como el logit realizado en este artículo) es más complejo (aunque también posible) de implementar.

[6] BERRI, D. J. y SCHMIDT, M. B. *Stumbilng on wins.* New Jersey: Pearson Education, Inc, 2010.

En definitiva, el empleo de probabilidades condicionadas permite una sencilla e intuitiva determinación de la probabilidad de descender para un equipo dado el marcador de la primera jornada de liga, un método, insistimos, que puede extenderse y aplicarse a muchos otros ámbitos del análisis deportivo. Por otro lado, esta investigación también muestra cómo el resultado del primer partido de liga sirve como indicador sobre el futuro del equipo varios meses después, al final de la temporada, en relación a si va a descender o no, lo que conlleva que, al igual que ocurre con los bebés recién nacidos, desde que la competición ve la luz se puede inferir información útil para el desarrollo de la vida del equipo tiempo después.

4. BIBLIOGRAFÍA

- BERRI, D. J. y SCHMIDT, M. B. *Stumbilng on wins.* New Jersey: Pearson Education, Inc, 2010.

- KRUSCHKE, J. K., AGUINIS, H. & JOO, H. The time has come: Bayesian methods for data analysis in the organizational sciences. *Organizational Research Methods.* 2012, núm. 15, pp. 722-752.

- LAGO, C. y CASÁIS, L. La influencia de los resultados iniciales en la clasificación final de los equipos de fútbol de alto nivel. *Motricidad. European Journal of Human Movement*. 2010, núm. 5 (14), pp. 107-122.

- LEVY, P. S. y LEMESHOW, S. *Sampling of populations.* New York: John Wiley & Sons, 1999.

- MARTÍNEZ, J. A. La influencia del primer cuarto en el resultado final en baloncesto. *Revista Internacional de Medicina y Ciencias de la Actividad Física y el Deporte*. 2011, núm. 14 (56), pp. 755-769.

- ZYPHUR, M. J. y OSWALD, F. L. Bayesian probability and statistics in management research: A new horizon. *Journal of Management*. 2013, num. 39 (1), pp. 5-13.