Avance de trabajo de especialización: Plus Minus Ajustado para la NBA

Franco Betteo

27 July 2019

Introducción

No hay una manera única e inequívoca de comparar rendimientos de jugadores en los distintos deportes y eso da lugar a discusiones sin fin. Más díficil aún en los deportes en equipo donde hay roles diferentes y contribuciones de distinta índole. A raíz de esto se han ido desarrollando métricas que intentan resumir el aporte al equipo de manera integral para hacer comparables los jugadores. En básquetbol una de las medidas más conocidas de este tipo es el "plus-minus" (originalmente implementado en el hockey sobre hielo), que calcula la diferencia de puntos de un equipo mientras cada jugador estuvo en cancha. Es decir que valores positivos (negativos) revelan que durante un partido el equipo hizo más (menos) puntos de los que recibió mientras el jugador estuvo en cancha. Es una métrica sencilla de calcular y resume el aspecto más importante de un partido de manera general para cada jugador. Es fácil de interpretar pero no está exenta de problemas. La idea de este trabajo es aplicar un método más robusto, "Plus-Minus ajustado", basado en los aportes de Justin Jacobs[1] y Joseph Sill[2], donde el principal agregado es controlar por los otros jugadores en cancha. El objetivo es actualizar la métrica para la temporada 2017-2018 y generar un ranking de jugadores. Posteriormente comparar el ranking contra otros generados por algún medio conocido para ver similitudes y diferencias. Por último generar un ranking de equipos, basado en los rankings individuales de jugadores, y comparar contra los resultados de las rondas definitorias del certamen cuyos datos no son utilizados para calcular la métrica.

Datos

Para poder calcular la métrica "plus minus ajustado" necesitamos tener para cada momento del partido los jugadores que hay en cancha y el resultado ya que el objetivo es ver la performance del equipo en presencia y ausencia de cada jugador de la liga. A tales fines se decidió utilizar la información provista por la API de MySportsFeed[3]. En ella podemos encontrar datos a un nivel suficientemente granular. En particular, para cada partido tenemos información jugada a jugada marcada por ciertas situaciones particulares, entre ellas tiros al aro y si se convirtieron puntos o no, rebotes, faltas y sustituciones. Con la primera y la última de estas características podemos recolectar los datos necesarios para generar nuestro dataset. Al estar todas las sustituciones y tener la alineación inicial de cada equipo podemos obtener todos los segmentos del partido donde hubo distintas combinaciones de jugadores en cancha (tanto compañeros como rivales) para cada jugador. A su vez, al tener los puntos anotados podemos obtener el diferencial de puntos para cada uno de estos segmentos.

A los fines de utilizar toda esta información para entrenar el modelo de "plus minus ajustado" necesitamos armar un dataset con el formato siguiente:

- Cada observación es un segmento de un partido donde se mantuvo constante la alineación de ambos equipos.
- Las variables independientes son cada uno de los jugadores de la liga, con valor de 1 si estaban en cancha en ese segmento para el equipo local y -1 si estaban en cancha siendo del equipo visitante.
- La variable dependiente es el diferencial de puntos del equipo local. Valores positivos (negativos) es que anotó más puntos el equipo local (visitante).

[1] https://squared2020.com/2017/09/18/deep-dive-on-regularized-adjusted-plus-minus-i-introductory-example/
[2] http://www.sloansportsconference.com/wp-content/uploads/2015/09/joeSillSloanSportsPaperWithLogo.pdf
[3] https://www.mysportsfeeds.com/data-feeds/api-docs

Esta tabla esta conformada con todos los equipos de la liga y para cada partido de la temporada, de manera que para apendizar la información de cada encuentro hay que tener las variables independientes de cada equipo, es decir, todos sus jugadores. Dado que cada partido solo involucra a dos equipos, la gran mayoría de las columnas tendrán valor de 0 en cada observación, siendo el dataset una matriz dispersa (sparse matrix). Se toman valores de 1 y -1 según la localía para que el signo quede acorde a la medición del diferencial de puntos y por ende el signo de los coeficientes sea siempre positivo para aportes beneficiosos a un equipo y negativos para aportes perjudiciales.

Tanto para la consulta de la API como para el manejo y procesamiento de los datos se utilizó el lenguaje R [4]

Modelo

La metodología para calcular la métrica "plus minus ajustado" implica correr una regresión Ridge (regularización con *l2*) donde la variable dependiente es el diferencial de puntos por segmento y las variables independientes son los jugadores en cancha. La idea de fondo es calcular el aporte de cada jugador al diferencial, controlando por sus compañeros y por los advsersarios. Se intenta eliminar el factor que sobreestima el aporte de algún jugador solo por el hecho de compartir tiempo en cancha con compañeros de primer nivel, que son los que generan realmente los diferenciales positivos. De la misma manera se intenta no sobreestimar a los jugadores que anotan muchos puntos contra equipos de baja performance o que solo juegan en los minutos llamados "basura" que corresponden a los minutos finales de un partido cuando ya está todo definido y suelen haber jugadores de menor nivel. Dado que es una regresión lineal los coeficientes pueden interpretarse como un proxy del aporte neto de cada jugador, ya descontados los aportes del resto. Dada la codificación de variables postiva para el local y negativa para visitantes, todo signo positivo de un coeficiente es aporte real en puntos y negativo es tendencia a recibir más puntos que los que se convierten con el jugador en cancha. Los equipos suelen tener una plantilla de alrededor de 10 jugadores activos y los que más minutos disputan suelen ser menos aún por lo tanto es de esperar que haya una gran correlación entre los jugadores de cada equipo. Este es el principal motivo por el que se decide ir por una regresión regularizada y no una regresión lineal multivariada clásica. Ridge lo que hace es reducir la varianza de los coeficientes incluyendo una penalización *l2* que implica reducir mínimos cuadrados sumado a la diferencia al cuadrado de los coeficientes respecto de 0. Esto último ponderado por un parámetro λ a definir. A mayor λ la penalización es mayor y los coeficientes tienden a valores cercanos a 0. El procedimiento reduce la varianza de los coeficientes a costa de introducir un sesgo en su estimación, pero que de tener éxito, el tradeoff es tal que las predicciones son más certeras a pesar de no ser insesgado.

Algunos aspectos metodológicos a tener en cuenta:

- Se eliminó el primer cuartil de jugadores medido en cantidad de posesiones en toda la temporada para reducir la dimensionalidad del problema. Son jugadores con pocos minutos en cancha y perjudican más de lo que aportan.
- Se eliminaron las observaciones con menos de 5 posesiones (provisorio. Falta definir con algún criterio) ya que no se las considera representativas. Mucha varianza en diferenciales con tan poco tiempo de juego.
- Se normalizaron los datos de diferencial de puntos cada 100 posesiones para hacer comparables los segmentos de tiempo.
- Para calcular el λ óptimo se aplicó Cross Validation con 10 folds. Luego se utiliza el λ definido en el entrenamiento con todo el set de training.

Aplicando lo comentado al set de datos construido y aplicando CV para obtener λ se obtienen los primeros resultados provisorios. $\lambda = 1.369778$

Y resumiendo los primeros diez jugadores según el coeficiente obtenido.

[4] R Core Team (2013). R: A language and environment for statisticalcomputing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

coef	playerid	player.firstName	player.lastName	team.abbreviation
2.150024	9418	Joel	Embiid	PHI
2.144711	9420	Robert	Covington	PHI
2.130441	9402	Victor	Oladipo	IND
2.045717	9265	Chris	Paul	HOU
1.928693	9352	Eric	Gordon	HOU
1.827553	9218	Stephen	Curry	GSW
1.757346	9325	Giannis	Antetokounmpo	MIL
1.751562	9152	Jimmy	Butler	MIN
1.742272	9524	Otto	Porter Jr.	WAS
1.609267	9480	LaMarcus	Aldridge	SAS

Puede discutirse el orden y si falta algún jugador importante de la liga pero no es un primer ranking totalmente descabellado. Son jugadores de primer nivel y varios de ellos son "indiscutidos". Sin embargo todavía no se hizo análisis de varianza de los coeficientes para ver intervalos de confianza y qué tan concluyente es figurar más arriba en el ranking respecto a otros jugadores. Quizás es solo por la muestra elegida.

Por otra parte luego sumamos los coeficientes de cada jugador por equipo y armamos un ranking de equipos conformado por esta agregación. Nuevamente los resultados son bastante creíbles, particularmente en la parte alta de la tabla donde los equipos que más ganaron en la temporada regular figuran en los primeros puestos. Hay alguna situación excepcional como Washington que figura ante último y se clasificó a la instancia de playoffs (posterior a la temporada regular) y por lo tanto su posición en la tabla no condice con su resultado real. Ver tabla 1.

Observaciones temporales

Lo que más tiempo demandó fue conseguir los datos crudos de la API y transformarlos para llegar al dataset final, teniendo en cuenta que los datos tenían errores como más de 5 jugadores en cancha por equipo, fechas de partidos que no coincidían entre tablas por formato en distinto huso horario, etc. Además es bastante volumen de datos y los chequeos intermedios no son tan fáciles.
Por otra parte, ya superado el obstáculo de construir el dataset parece que la regresión está medianamente orientada y da resultados provisorios en sintonía con la realidad. Más allá de lo consguido hasta el momento quedan ciertas tareas pendientes:

- Analizar la variabilidad de los coeficientes para ver qué tanto nos dice un ranking de ellos.
- Utilizar la misma metodología para agregar más datos de temporadas previas y darle más robustez a la regresión.
- Coneguir los datos de la temporada 2018-2019 para utilizarla de Test set.

Table 1: Ranking de Equipos

team.abbreviation	coef_total
TOR	7.3393642
UTA	6.1522830
GSW	6.0844704
HOU	5.5569756
BOS	3.5126122
SAS	3.3015790
OKL	2.7638865
POR	2.4827007
IND	1.9754364
PHI	1.7361041
DEN	1.6316350
DET	1.3568088
MIN	1.2848449
LAC	1.2229708
NOP	1.0148622
MIA	0.8547383
CLE	0.4177646
WAS	-0.0587758
MIL	-0.7169583
ORL	-0.8992258
LAL	-1.0746973
PHX	-2.4035736
DAL	-2.5015242
CHA	-2.5132491
NYK	-2.8326085
ATL	-3.7302492
BRO	-3.9193425
MEM	-4.3131367
SAC	-5.2517649
CHI	-7.0165741