Universidade de Brasília - UnB
Instituto de Ciências Exatas - IE
Departamento de Estatística - EST

Predição de Resultados da NBA na temporada 2018-19

Gustavo Pompeu da Silva

Orientador: Professor Eduardo Monteiro de Castro Gomes

Brasília

2018

Sumário

1 Introdução e Justificativa

Mineração de dados em esportes é um tópico que tem crescido rapidamente nos últimos anos. Jogadores de ligas de *fantasy* e entusiastas de esportes estão cada vez mais interessados em procurar uma vantagem nas apostas e previsões através de dados e números. Ferramentas e técnicas começaram a ser desenvolvidas para medir desempenho tanto de times quanto de atletas, e esses métodos vem chamando a atenção de grandes franquias esportivas.

Existe uma imensa quantidade de dados disponíveis sobre qualquer esporte. Esses dados podem ser de desempenho individual de jogadores ou da equipe, decisões da comissão técnica, eventos que acontecem nos jogos, entre outros. O problema não é como coletar esses dados, mas sim saber quais dados podem ser úteis e como fazer o melhor uso possível deles. Achando os meios para transformar esses dados em conhecimento, organizações esportivas tem o potencial de obter uma vantagem competitiva sobre seus oponentes. Não devemos analisar performance no sentido de marcar mais gols ou pontos do que o oponente, pois esse é o objetivo geral de qualquer esporte, o que é interessante é encontrar padrões em outras estatísticas que mostram tendências justamente para chegar às vitórias.

Data Mining envolve procedimentos para descobrir padrões escondidos e descobrir novas informações a partir de fontes de dados. A fundação científica de data mining pode ser dividida em três disciplinas: estatísticas, inteligência artificial e machine learning. *Data mining* então pode ser definido como a busca de conhecimento dentro dos dados. (SCHUMAKER; SOLIEMAN; CHEN, 2010)

A NBA (National Basketball Association) é a principal liga de basquete profissional do mundo. Atualmente, é composta por 30 times baseados em cidades da América do Norte (29 nos Estados Unidos e 1 no Canadá). É a liga onde jogam os melhores atletas de basquete do mundo, e com os maiores salários do esporte. Uma das vantagens de trabalharmos com o basquete e a NBA especificamente é a grande quantidade de dados, pois em uma temporada, cada time joga 82 jogos, ou seja, são 1230 jogos por temporada, isso nos permite ter muitas observações para trabalhar.

2 Objetivos

Utilizar diversas técnicas estatísticas para prever os resultados dos jogos da temporada de 2018-19 da NBA e chegar em uma conclusão sobre qual é a melhor técnica para esse problema em específico, através de medidas como acurácia.

2.1 Objetivo Geral

Utilizar métodos como regressão linear, regressão logística, entre outros, para prever resultados e obter probabilidades de vitória para os times em cada um dos jogos da temporada 2018-19 da NBA.

2.2 Objetivos Específicos

- Ler os dados de resultados dos jogos das temporadas anteriores automaticamente com o R;
- Utilizar os métodos acima descritos nesses dados para ajustar modelos de previsões;
- Obter previsões dos resultados dos jogos da temporada 2018-19 com esse modelos;
- Comparar os métodos e descobrir qual o melhor para essas previsões especificamente.

3 Material e Métodos

A linguagem R (R Core Team, 2018) será utilizada em todo o trabalho. Com o auxílio do pacote *rvest* (WICKHAM, 2016), serão extraídos os resultados dos jogos das temporadas anteriores direto da internet, no site Basketball Reference (BASKETBALL. . . ,), que é um dos maiores sites com dados estatísticos sobre a NBA e basquete em geral. A partir desses dados, podemos criar inúmeras variáveis, e então criar modelos e métodos de previsão. Durante a temporada de 2018-19, os dados dos jogos também irão sendo extraídos direto da internet.

A ideia é, utilizar vitória ou saldo de pontos na partida como variável resposta para os modelos, e selecionar as variáveis explicativas que são significativas. E então criar modelos com os seguintes métodos: regressão linear, regressão logística, máquina de vetores de suporte (SVM) e *random forest*, e utilizar os resultados de temporadas anteriores para ajustar esses modelos, que serão utilizados para fazer previsões para as partidas da temporada de 2018-19.

3.1 Regressão Linear

Regressão linear é uma equação para se estimar o valor esperado de uma variável Y (resposta), dados os valores de outras variáveis X (explicativas). É chamada "linear"porque se considera que a relação da resposta às variáveis explicativas é uma função linear de alguns parâmetros. Para se estimar o valor esperado, usa-se de uma equação, que determina a relação entre as variáveis:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \epsilon$$

Onde Y é a variável resposta (dependente), $\beta_j, j = 0, 1, ..., p$ são constantes, denominados coeficientes de regressão, $X_j, j = 1, ..., p$ são as variáveis explicativas (independentes) e ϵ representa o erro experimental. O parâmetro β_0 corresponde ao intercepto, e fornece a resposta média de Y quando $X_1 = X_2 = ... = X_p = 0$. Para $j \geq 1$, os parâmetros β_j indicam uma mudança na resposta média de Y a cada unidade de mudança na variável X_j, quando as demais variáveis são mantidas fixas.

As suposições necessárias para o Modelo de Regressão Linear Múltipla são:

- Os erros não devem ser correlacionados, devem seguir distribuição normal e ter média zero e variância σ^2, desconhecida. Ou seja, $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$;
- Deve existir uma relação linear entre a variável dependente e as variáveis independentes;
- Não deve haver multicolinearidade entre as variáveis independentes, ou seja, elas não podem ter correlação alta entre si.

Nesse trabalho, sabemos que não estamos cumprindo todas as suposições da regressão linear, principalmente porque as observações não são independentes umas das outras, pois os jogos são uma sequência histórica no tempo.

Para a implementação computacional é utilizada a função *lm* do pacote *stats*, que faz parte do R. (R Core Team, 2018)

3.2 Regressão Logística

A regressão logística se difere da linear essencialmente pelo fato da variável resposta ser binária, ou seja, Y tem distribuição Bernoulli $(1, \pi)$, com probabilidade de sucesso $P(Y_i = 1) = \pi_i$ e de fracasso $P(Y_i = 0) = 1 - \pi_i$.

No centro da regressão logística está a tarefa de estimar o *log odds* de um evento. Matematicamente, a regressão logística estima uma função de regressão linear múltipla definida por:

$$logit(\pi) = log(\frac{\pi}{1-\pi}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p \quad (1)$$

Onde $\pi = P(Y = 1)$. Baseado em 1, chegamos em:

$$\pi = \frac{exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p)}{1 + exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p)} = \frac{e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p}}{1 + e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p}}$$

As suposições necessárias para o Modelo de Regressão Logística são:

- A variável dependente precisa ser binária (dicotômica);
- As observações precisam ser independentes umas das outras, ou seja, as observações não devem prover de medições repetidas ou dados correspondentes;
- Não deve haver multicolinearidade entre as variáveis independentes, ou seja, elas não podem ter correlação alta entre si;
- Deve existir linearidade entre as variáveis independentes e o *log odds*;
- Regressão logística tipicamente requer uma amostra grande.

Para a implementação computacional é utilizada a função *glm* do pacote *stats*, que faz parte do R. (R Core Team, 2018)

3.3 Regressão de Probit

A Análise de Probit ou Regressão de Probit (CARVALHO et al., 2017) é outro tipo de regressão binária, parecida com a regressão logística, a diferença é a função de ligação utilizada, o *link* probit é dado por:

$$probit(\pi) = \Phi^{-1}(\pi) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p \tag{2}$$

Onde $\pi = P(Y = 1)$. Baseado em 2, chegamos em:

$$\pi = \Phi(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p)$$

. Φ é a Função de Distribuição Acumulada (f.d.a.) da distribuição Normal Padrão.

Para a implementação computacional também é usada a função *glm* do pacote *stats*, que faz parte do R. (R Core Team, 2018)

3.4 Máquina de Vetores de Suporte (SVM)

As Máquinas de Vetores de Suporte (*Support Vector Machines* - SVMs) constituem uma técnica embasada na Teoria de Aprendizado Estatístico (VAPNIK, 1995), e seu objetivo é classificar dados em grupos.

Para classificar dados em duas classes diferentes, podemos enfrentar o problema de uma maneira direta: tentamos achar um plano que separe as classes no espaço p-dimensional. Vamos chamar esse plano de hiperplano.

O SVM determina o hiperplano ótimo, e pode fazer isso para conjuntos linearmente separáveis ou não, através da utilização de funções Kernel. Para a implementação computacional, será utilizada a função *svm* do pacote *e1071* da linguagem R (MEYER et al., 2018), que possui 4 opções de função Kernel: linear, base radial (gaussiana), polinomial e sigmoidal. Cada tipo de Kernel tem vários parâmetros que podem ser ajustados. Para esse trabalho será utilizado apenas o Kernel base radial, pois foi o que apresentou melhores resultados em geral para os dados. Ele consiste em $exp(-\gamma\|\mathbf{u}-\mathbf{v}\|^2)$.

O parâmetro γ equivale a $\frac{1}{2\sigma^2}$, e pode ser especificado pelo usuário.

3.5 Análise de Discriminante Linear

A Análise de Discriminante é uma técnica multivariada que se preocupa em separar observações em grupos, e alocar novas observações em algum dos grupos pré-definidos. A Análise de Discriminante é bastante exploratória em sua natureza. Em geral, o objetivo dessa técnica é descrever algebricamente as características diferenciais das observações, nós tentamos achar "discriminantes"cujo valores numéricos são tais que as populações são separadas o melhor possível. (JOHNSON; WICHERN, 2007)

A Análise de Discriminante Linear é uma generalização da Discriminante Linear de Fisher. Para duas classes, a alocação de novas observações funciona de uma maneira muito simples. Primeiramente, é feita uma matriz de variância-covariância estimada para os dados ($\boldsymbol{S}_p{}^2$):

$$\boldsymbol{S}_p{}^2 = \frac{\sum_{j=1}^{n1}(\boldsymbol{x}_{1j}-\bar{\boldsymbol{x}}_1)^2 + \sum_{j=1}^{n2}(\boldsymbol{x}_{2j}-\bar{\boldsymbol{x}}_2)^2}{n_1+n_2-2}$$

Onde n_1 e n_2 correspondem ao número de observações da população 1 (π_1) e ao número de observações da população 2 (π_2), respectivamente, $\bar{\boldsymbol{x}}_1$ e $\bar{\boldsymbol{x}}_2$ correspondem às médias das variáveis independentes para cada população, e $\boldsymbol{x}_{1j}$ e $\boldsymbol{x}_{2j}$ são referentes à cada observação j de cada população.

Então, para uma nova observação $\boldsymbol{x}_0$, temos: $\hat{y_0} = (\bar{\boldsymbol{x}}_1-\bar{\boldsymbol{x}}_2)'\boldsymbol{S}_p{}^{-1}\boldsymbol{x}_0$ e $\hat{m} = \frac{1}{2}(\bar{\boldsymbol{x}}_1-\bar{\boldsymbol{x}}_2)'\boldsymbol{S}_p{}^{-1}(\bar{\boldsymbol{x}}_1+\bar{\boldsymbol{x}}_2)$, e a regra de alocação será que x_0 pertencerá à π_1 se $\hat{y_0}-\hat{m} \geq 0$, e x_0 pertencerá à π_2 caso contrário.

As suposições necessárias para a Análise de Discriminante Linear são:

- Normalidade multivariada dos dados;
- Matriz de variância-covariância das populações devem ser iguais.

Para a implementação em linguagem R é utilizada a função *lda* do pacote *MASS*. (VENABLES; RIPLEY, 2002)

4 Resultados

Na temporada de 2017-18, os times com mais vitórias na temporada regular foram o Houston Rockets (65), Toronto Raptors (59) e Golden State Warriors (58). Os piores times foram o Phoenix Suns com 21 vitórias, o Memphis Grizzlies com 22 e o Atlanta Hawks com 24. O time com mais pontos marcados em média durante a temporada foi o Golden State Warriors, com 113.5 pontos por jogo, e o com a menor média foi o Sacramento Kings, com 98.8 pontos por jogo. A melhor defesa, ou seja, o time que levou menos pontos em média, foi o San Antonio Spurs, sofrendo apenas 99.8 pontos por jogo, enquanto que o pior time nesse quesito foi o Phoenix Suns, sofrendo 113.3 pontos por partida. O máximo de pontos marcados por um time em um jogo foram 149, pelo Miami Heat, em um jogo de duas prorrogações contra o Denver Nuggets no dia 19/03/2018. Já o mínimo de pontos foi 69, em duas ocasiões, pelo Chicago Bulls contra o Oklahoma City Thunder em 28/10/2017 e pelo Washington Wizards contra o Utah Jazz em 04/12/2017. Já nos *playoffs*, o Golden State Warriors se deu melhor, vencendo o Cleveland Cavaliers na *NBA Finals* e conquistando o seu sexto título na história, e o terceiro nos últimos 4 anos. (NBA. . . ,)

Apenas com as informações de placar, data, público, e se o jogo foi para a prorrogação ou não, foram criadas bases de dados com mais de 200 variáveis diferentes. Elas são:

Variáveis do jogo da linha atual:

- *Team*: Nome do time
- *Opp*: Nome do time adversário
- *Pts_S*: Pontos marcados pelo time nesse jogo.
- *Pts_A*: Pontos marcados pelo time adversário nesse jogo.
- *Home*: Se o jogo foi em casa ou não.
- *Win*: Se o time venceu esse jogo ou não.
- *Attend*: Público presente no ginásio nesse jogo.
- *Result*: Saldo de pontos, ou seja, os pontos marcados pelo time que jogou em casa menos os pontos marcados pelo time que jogou fora de casa.

Variáveis do último jogo do time:

- *OT*: Se o último jogo do time foi para a prorrogação ou não.
- *Days_LG*: Quantos dias atrás foi o último jogo do time.

Variáveis da temporada do time:

- *Games_T*: Total de jogos do time até agora na temporada.
- *Games_H*: Jogos em casa do time até agora na temporada.
- *Wins_T*: Total de vitórias do time até agora na temporada.
- *Wins_H*: Vitórias em casa do time até agora na temporada.
- *Mean_Pts_S_H*, *Max_Pts_S_H*, *Min_Pts_S_H*: Média, máximo e mínimo de pontos marcados do time em jogos em casa até o momento na temporada.
- *Mean_Pts_S_A*, *Max_Pts_S_A*, *Min_Pts_S_A*: Média, máximo e mínimo de pontos marcados do time em jogos fora de casa até o momento na temporada.
- *Mean_Pts_S_T*: Média de pontos marcados do time em todos os jogos até o momento na temporada.
- *Mean_Pts_A_H*, *Max_Pts_A_H*, *Min_Pts_A_H*: Média, máximo e mínimo de pontos marcados contra o time em jogos em casa até o momento na temporada.
- *Mean_Pts_A_A*, *Max_Pts_A_A*, *Min_Pts_A_A*: Média, máximo e mínimo de pontos marcados contra o time em jogos fora de casa até o momento na temporada.
- *Mean_Pts_A_T*: Média de pontos marcados contra o time em todos os jogos até o momento na temporada.
- *Mean_Last_X_A*, *Max_Last_X_A*, *Min_Last_X_A*: Média, máximo e mínimo de pontos marcados do time nos últimos X jogos fora de casa até o momento na temporada, onde $X = 3, 5, 7, 10$.
- *Mean_Last_X_H*, *Max_Last_X_H*, *Min_Last_X_H*: Média, máximo e mínimo de pontos marcados do time nos últimos X jogos em casa até o momento na temporada, onde $X = 3, 5, 7, 10$.
- *Mean_Last_X_T*, *Max_Last_X_T*, *Min_Last_X_T*: Média, máximo e mínimo de pontos marcados do time nos últimos X jogos até o momento na temporada, onde $X = 3, 5, 7, 10$.
- *Mean_Last_X_A_Opp*, *Max_Last_X_A_Opp*, *Min_Last_X_A_Opp*: Média, máximo e mínimo de pontos marcados contra o time nos últimos X jogos fora de casa até o momento na temporada, onde $X = 3, 5, 7, 10$.
- *Mean_Last_X_H_Opp*, *Max_Last_X_H_Opp*, *Min_Last_X_H_Opp*: Média, máximo e mínimo de pontos marcados contra o time nos últimos X jogos em casa até o momento na temporada, onde $X = 3, 5, 7, 10$.

- *Mean_Last_X_T_Opp*, *Max_Last_X_T_Opp*, *Min_Last_X_T_Opp*: Média, máximo e mínimo de pontos marcados contra o time nos últimos X jogos até o momento na temporada, onde $X = 3, 5, 7, 10$.

- *Str_Sch*: A "força de calendário"do time até o momento na temporada, ou seja, a proporção de vitórias dos adversários que o time enfrentou até o momento.

Isso resulta em 101 variáveis no banco de dados, para cada time em cada jogo, mas juntando as duas linhas de cada jogo em uma só, resulta em um banco com uma linha para cada jogo e 202 variáveis. Nota-se que ainda mais variáveis podem ser criadas para o banco de dados.

Em uma análise preliminar, apenas com os resultados da temporada anterior, de 2017-18, e fazendo uma regressão linear inicial com algumas variáveis que foram selecionadas testando o nível de significância, ajustando o modelo com os primeiros 1000 jogos, e fazendo as previsões para os últimos 230, foi obtida uma acurácia de quase 68% nas previsões. Ou seja, com uma seleção de variáveis melhor e dados de mais temporadas, além de utilização de outros tipos de modelos, há uma grande chance desse resultado preliminar melhorar bastante.

Referências

AGRESTI, A. *An Introduction to Categorical Data Analysis*. Wiley, 2007. (Wiley Series in Probability and Statistics). ISBN 9780470114742. Disponível em: <https://books.google.com.br/books?id=OG9Eqwd0Fh4C>.

BASKETBALL Reference. <https://www.basketball-reference.com/>. Accessado em: 28/09/2018. 4

CARVALHO, J. et al. *ANÁLISE DE PROBIT APLICADA A BIOENSAIOS COM INSETOS*. [S.l.: s.n.], 2017. ISBN 978-85-64937-08-6. 6

JOHNSON, R. A.; WICHERN, D. W. *Applied Multivariate Statistical Analysis*. Pearson Prentice Hall, 2007. (Applied Multivariate Statistical Analysis). ISBN 9780131877153. Disponível em: <https://books.google.com.br/books?id=gFWcQgAACAAJ>. 7

KUTNER, M.; NACHTSHEIM, C.; NETER, J. *Applied Linear Regression Models*. McGraw-Hill Higher Education, 2003. (The McGraw-Hill/Irwin Series Operations and Decision Sciences). ISBN 9780072955675. Disponível em: <https://books.google.com.br/books?id=0nAMAAAACAAJ>.

LIAW, A.; WIENER, M. Classification and regression by randomforest. *R News*, v. 2, n. 3, p. 18–22, 2002. Disponível em: <https://CRAN.R-project.org/doc/Rnews/>.

MEYER, D. et al. *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*. [S.l.], 2018. R package version 1.7-0. Disponível em: <https://CRAN.R-project.org/package=e1071>. 7

MILLER, T. L. based on Fortran code by A. *leaps: Regression Subset Selection*. [S.l.], 2017. R package version 3.0. Disponível em: <https://CRAN.R-project.org/package=leaps>.

NBA Stats. <https://stats.nba.com/>. Accessado em: 28/09/2018. 8

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2018. Disponível em: <https://www.R-project.org/>. 4, 5, 6

SCHUMAKER, R. P.; SOLIEMAN, O. K.; CHEN, H. *Sports Data Mining*. 1st. ed. [S.l.]: Springer Publishing Company, Incorporated, 2010. ISBN 144196729X, 9781441967299. 3

UUDMAE, J. Predicting nba game outcomes. Accessado em: 28/09/2018. Disponível em: <http://cs229.stanford.edu/proj2017/final-reports/5231214.pdf>.

VAPNIK, V. N. *The Nature of Statistical Learning Theory*. Berlin, Heidelberg: Springer-Verlag, 1995. ISBN 0-387-94559-8. 6

VENABLES, W. N.; RIPLEY, B. D. *Modern Applied Statistics with S*. Fourth. New York: Springer, 2002. ISBN 0-387-95457-0. Disponível em: <http://www.stats.ox.ac.uk/pub/MASS4>. 7

WICKHAM, H. *rvest: Easily Harvest (Scrape) Web Pages*. [S.l.], 2016. R package version 0.3.2. Disponível em: <https://CRAN.R-project.org/package=rvest>. 4

WICKHAM, H. et al. *dplyr: A Grammar of Data Manipulation*. [S.l.], 2018. R package version 0.7.8. Disponível em: <https://CRAN.R-project.org/package=dplyr>.