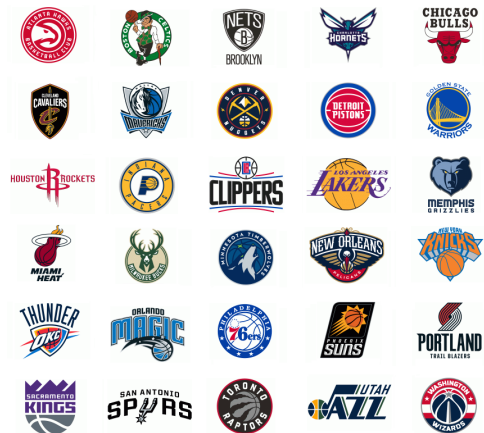


Modelos de Previsão para os Resultados da Temporada Regular de 2018/19 da NBA































Gustavo Pompeu da Silva

5 de Julho de 2019

Introdução



Introdução

Team	W	L	Team	W	L
1  Bucks	60	22	1  Warriors	57	25
2  Raptors	58	24	2  Nuggets	54	28
3  76ers	51	31	3  Trail Blazers	53	29
4  Celtics	49	33	4  Rockets	53	29
5  Pacers	48	34	5  Jazz	50	32
6  Nets	42	40	6  Thunder	49	33
7  Magic	42	40	7  Spurs	48	34
8  Pistons	41	41	8  Clippers	48	34
9  Hornets	39	43	9  Kings	39	43
10  Heat	39	43	10  Lakers	37	45
11  Wizards	32	50	11  Timberwolves	36	46
12  Hawks	29	53	12  Grizzlies	33	49
13  Bulls	22	60	13  Pelicans	33	49
14  Cavaliers	19	63	14  Mavericks	33	49
15  Knicks	17	65	15  Suns	19	63

Modelos

As técnicas estatísticas utilizadas para a obtenção das previsões dos jogos são:

- Regressão Linear;
- Regressão Logística;
- Regressão de Probit;
- Máquina de Vetores de Suporte (SVM);
- Análise de Discriminante Linear;
- Árvores de Regressão;
- Árvores de Classificação;
- *Random Forest*.

Regressão Linear

É um método estatístico que compõe uma equação para se descrever o valor esperado de uma variável Y (resposta), dado os valores de outras variáveis X (explicativas). É linear pois considera que a relação da variável resposta com as variáveis explicativas é uma função linear dependente de alguns parâmetros. A equação que determina a relação entre as variáveis é:

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

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 $P(Y_i = 0) = 1 - \pi_i$.

Matematicamente, a regressão logística estima uma função de regressão linear múltipla definida por:

$$\text{logit}(\pi) = \log\left(\frac{\pi}{1 - \pi}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p$$

$$\pi = \frac{e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p}}{1 + e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p}}$$



Regressão de Probit

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:

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

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

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



Análise de Discriminante Linear

Técnica multivariada que tem como finalidade separar observações em grupos e alocar novas observações em algum dos grupos pré-definidos. Para uma nova observação \mathbf{x}_0 , tem-se:

$$\hat{y}_0 = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' \mathbf{S}_p^{-1} \mathbf{x}_0$$

$$\hat{m} = \frac{1}{2} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' \mathbf{S}_p^{-1} (\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)$$

A regra de alocação será que a observação pertencerá à população π_1 se $\hat{y}_0 - \hat{m} \geq 0$, e pertencerá à população π_2 caso contrário.



Árvores de Regressão e Classificação

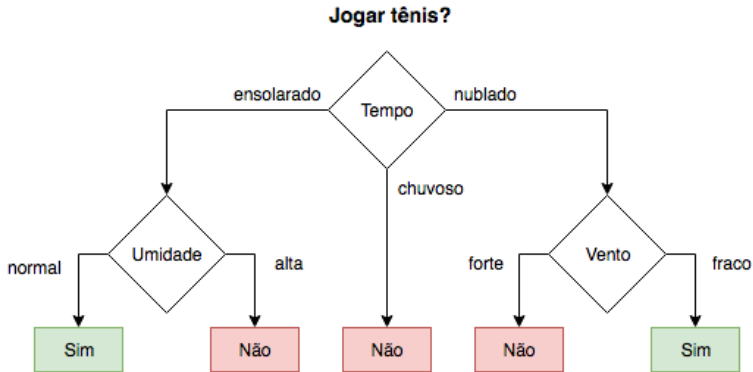


Figura: Exemplo de Árvore de Decisão para jogar tênis ou não



Random Forest

- Combinação de preditores de árvores;
- Pode ser tanto pra classificação quanto pra regressão;
- Foi utilizado apenas pra classificação.



Modelos

- Não houve preocupação em verificar os pressupostos dos modelos;
- Seleção de variáveis para regressão linear, logística e de probit;
- Método *forward*;
- Função *step*, que mede o AIC.

Web scraping

- Pacote *rvest*;
- Extensão *SelectorGadget* do Google Chrome;
- Basketball-Reference.com;
- Desde a temporada 2000/01.

Tabela: Exemplos dos dados extraídos

Data	Visitante	Pontos do Visitante	Mandante	Pontos do Mandante	Prorrogação	Público
01/12/2018	Toronto	106	Cleveland	95	-	19432
01/12/2018	Golden State	102	Detroit	111	-	20332
01/12/2018	Chicago	105	Houston	121	-	18055
01/12/2018	Boston	118	Minnesota	109	-	17663
01/12/2018	Milwaukee	134	New York	136	OT	19812
01/12/2018	Indiana	110	Sacramento	111	-	17583



Bases de Dados

Variáveis resposta, que são indicadoras do resultado do jogo:

- *Win* e *result*.

Algumas variáveis explicativas:

- *Wins_T*, *Wins_A*, *Wins_H*;
- *Mean_Pts_S_T*, *Mean_Pts_S_A*, *Mean_Pts_S_H*;
- *mean_attend*;
- *Mean_Last_X_T*, com $X = 3, 5, 7, 10$;
- *OT_Last*;
- *Days_LG*.

Para os dois times, resultando num total de 151 variáveis na base.

Valores faltantes

- Identificar padrões dos valores faltantes (NA).

Tabela: Exemplo de padrão de NA's

Variável	Valor	Vetor de 0's e 1's
<i>Wins_T_Vis</i>	1	0
<i>Loss_T_Vis</i>	1	0
<i>Mean_Last3_total_Vis</i>	NA	1
<i>Wins_T_Home</i>	2	0
<i>Loss_T_Home</i>	2	0
<i>Win_Last5_total_Home</i>	NA	1

- 61 padrões para 2018/19, implicando em 61 modelos.



Casas de Aposta

- “linha” de aposta;
- Exemplo: Golden State Warriors favorito contra o Portland Trail Blazers por 6.5 pontos, logo, a “linha” é -6.5 para os Warriors e +6.5 para o Trail Blazers;
- *web scraping* do site da ESPN;
- Em 4 jogos a “linha” não estava disponível;
- Em 16 jogos era *even* (0);
- Porcentagem de acerto 0.6727 em 1210 jogos.



Resultados

Tabela: Porcentagem de Acerto das previsões dos jogos da temporada 2018/19 para cada método utilizando dados de 2000/01 a 2017/18 na modelagem

Método	Porcentagem de Acerto
Regressão de Probit	0.6723577
Regressão Logística	0.6707317
Análise de Discriminante Linear	0.6682927
Regressão de Probit c/ Forward	0.6682927
Regressão Logística c/ Forward	0.6674797
SVM com $cost = 8$, $gamma = 10^{-4}$	0.6666667
Regressão Linear c/ Forward	0.6658537
Regressão Linear	0.6634146
SVM padrão	0.6577236
Random Forest	0.6373984
Regressão em Árvore	0.6373984
Classificação em Árvore	0.6089431



Resultados

- Evolução do esporte;
- Temporadas antigas possuem números diferentes das recentes.

Tabela: Porcentagem de acerto das previsões dos jogos da temporada 2018/19 para cada método utilizando temporadas diferentes na modelagem

Temporada de Início	Regressão Linear	Regressão Logística	Regressão de Probit	LDA	Regressão em Árvore	Classificação em Árvore
2004/2005	0.667	0.672	0.672	0.672	0.642	0.642
2005/2006	0.668	0.672	0.674	0.676	0.628	0.642
2006/2007	0.675	0.681	0.677	0.679	0.624	0.645
2007/2008	0.670	0.677	0.680	0.679	0.624	0.648



Resultados

Tabela: Porcentagem de Acerto das previsões dos jogos da temporada 2018/19 para cada método utilizando dados de 2006/07 a 2017/18 na modelagem

Método	Porcentagem de Acerto
Regressão Logística	0.6813008
Análise de Discriminante Linear	0.6788618
Regressão de Probit	0.6772358
Regressão Linear	0.6747967
SVM com $cost = 8$, $gamma = 10^{-4}$	0.6731707
Regressão Linear c/ Forward	0.6707317
Regressão Logística c/ Forward	0.6682927
Regressão de Probit c/ Forward	0.6642276
SVM padrão	0.6569106
Classificação em Árvore	0.6447154
Random Forest	0.6373984
Regressão em Árvore	0.6243902



Resultados

Tabela: Tempo de execução do código computacional para cada método

Método	Tempo (em segundos)
Regressão Linear	9.733
Classificação em Árvore	19.692
Regressão em Árvore	21.069
Regressão Logística	30.542
Regressão de Probit	33.780
Análise de Discriminante Linear	35.057
Regressão Linear c/ Forward	985.550
Regressão de Probit c/ Forward	5359.306
Regressão Logística c/ Forward	6095.090
SVM	9420.622
Random Forest	31367.020



Resultados

- Comparação das vitórias reais com as vitórias previstas para as Conferências Leste e Oeste

Time	Vitórias Reais	Vitórias Previstas
Milwaukee Bucks	60	74
Toronto Raptors	58	69
Philadelphia 76ers	51	59
Boston Celtics	49	55
Indiana Pacers	48	54
Brooklyn Nets	42	45
Orlando Magic	42	39
Detroit Pistons	41	39
Charlotte Hornets	39	36
Miami Heat	39	32
Washington Wizards	32	28
Atlanta Hawks	29	14
Chicago Bulls	22	10
Cleveland Cavaliers	19	9
New York Knicks	17	5

Time	Vitórias Reais	Vitórias Previstas
Golden State Warriors	57	65
Denver Nuggets	54	65
Portland Trail Blazers	53	63
Houston Rockets	53	60
Utah Jazz	50	59
Oklahoma City Thunder	49	54
Los Angeles Clippers	48	51
San Antonio Spurs	48	49
Sacramento Kings	39	37
Los Angeles Lakers	37	32
Minnesota Timberwolves	36	32
Dallas Mavericks	33	30
Memphis Grizzlies	33	29
New Orleans Pelicans	33	29
Phoenix Suns	19	7

Resultados

Tabela: Variáveis mais significativas no modelo

Variável	Estimativa do Parâmetro β	Erro Padrão	Z (Estatística de Teste)	p-valor
Mean_Pts_A_T_Vis	-0.34173	0.120916	-2.826	0.00471
Min_Last5home_Home	-0.18985	0.07072	-2.685	0.00726
Loss_T_Vis	-0.59691	0.227052	-2.629	0.00856
Days_LG_Vis	0.062418	0.024913	2.505	0.01223
Mean_Last3_home_opp_Home	-0.34262	0.142105	-2.411	0.01591

- Apenas para o modelo completo (875 observações)
- Parâmetros positivos indicam que a variável contribui para o aumento da probabilidade de vitória do time visitante



Resultados

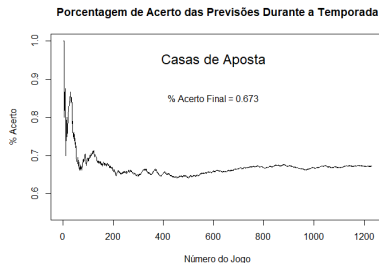
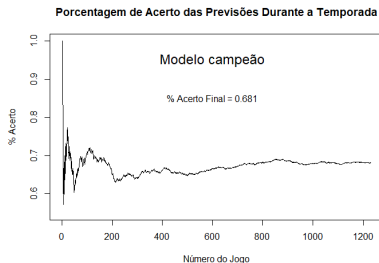
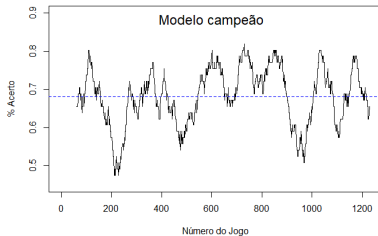


Figura: Evolução da porcentagem de acerto das previsões ao longo da temporada

Resultados

Porcentagem de Acerto das Previsões nos Últimos 61 Jogos



Porcentagem de Acerto das Previsões nos Últimos 61 Jogos

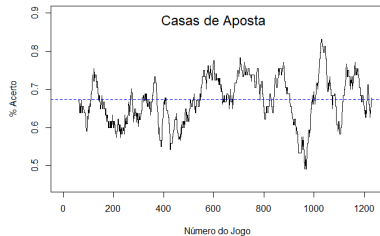


Figura: Porcentagem de acerto das previsões dos últimos 61 jogos ao longo da temporada

Resultados

Tabela: Resumo das diferenças absolutas das Previsões da Regressão Linear vs. linhas de aposta vs. resultados reais

Comparação	Mín.	1º Quartil	Mediana	Média	3º Quartil	Máx.	NA's
Regressão Linear vs. Resultados Reais	0.006	3.844	8.123	10.276	14.602	50.532	-
Linhas de aposta vs. Resultados Reais	0.000	4.000	8.000	9.927	14.000	55.000	4
Regressão Linear vs. Linhas de aposta	0.001	1.055	2.359	2.905	4.114	16.991	4



Conclusão

- Resultado melhor que das casas de aposta
- Regressão linear, logística, probit e LDA melhores tanto em acerto quanto em tempo
- Falta informações sobre jogadores, lesões, trocas, etc.



Referências I



[Basketball-reference.](#)

<https://www.basketball-reference.com/>, 2019.

Accessado em: 11/06/2019.



[Espn.](#)

<http://www.espn.com/nba/scoreboard>, 2019.

Accessado em: 16/05/2019.



[Selectorgadget.](#)

<https://selectorgadget.com/>, 2019.

Accessado em: 25/05/2019.



[Sportslogos.](#)

http://www.sportslogos.net/teams/list_by_league/6/National_Basketball_Association/NBA/logos/, 2019.

Accessado em: 13/10/2018.



[AGRESTI, A.](#)

An Introduction to Categorical Data Analysis.

Wiley Series in Probability and Statistics. Wiley, 2007.

Referências II



BREIMAN, L.

Random forests.

Accessado em: 01/06/2019.



CARVALHO, J., PRATISSOLI, D., VIANNA, U., AND MATHIAS HOLTZ, A.

ANÁLISE DE PROBIT APLICADA A BIOENSAIOS COM INSETOS.

06 2017.



JOHNSON, R. A., AND WICHERN, D. W.

Applied Multivariate Statistical Analysis.

Applied Multivariate Statistical Analysis. Pearson Prentice Hall, 2007.



KASSAMBARA, A.

Machine Learning Essentials: Practical Guide in R.

CreateSpace Independent Publishing Platform, 2018.



KUTNER, M., NACHTSHEIM, C., AND NETER, J.

Applied Linear Regression Models.

The McGraw-Hill/Irwin Series Operations and Decision Sciences. McGraw-Hill Higher Education, 2003.



MEYER, D.

Support vector machines, the interface to libsvm in package e1071.

Accessado em: 31/05/2019.

Referências III



R CORE TEAM.

R: A Language and Environment for Statistical Computing.

R Foundation for Statistical Computing, Vienna, Austria, 2018.



RIPLEY, B. D.

Pattern Recognition and Neural Networks.

Cambridge University Press, 1996.



SUÁREZ, E., PÉREZ, C. M., RIVERA, R., AND MARTÍNEZ, M. N.

Selection of Variables in a Multiple Linear Regression Model.

John Wiley & Sons, Ltd, 2017, ch. 5, pp. 77–86.



WICKHAM, H.

rvest: Easily Harvest (Scrape) Web Pages, 2016.

R package version 0.3.2.



YAN, X., AND SU, X. G.

Linear Regression Analysis: Theory and Computing.

World Scientific Publishing Co., Inc., River Edge, NJ, USA, 2009.

