

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

Gustavo Pompeu da Silva

Orientador: Prof. Eduardo Monteiro de Castro Gomes































Departamento de Estatística  
Universidade de Brasília

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*.

# Modelos

- Não houve preocupação em verificar os pressupostos dos modelos;
- Jogos são uma sequência histórica no tempo, observações dependentes;
- 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.



# Casas de Apostas

- “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 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

**Tabela:** Variáveis mais significativas no modelo

Variável	Estimativa do Parâmetro $\beta$	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

**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
<b>Regressão Linear vs. Resultados Reais</b>	0.006	3.844	8.123	10.276	14.602	50.532	-
<b>Linhas de aposta vs. Resultados Reais</b>	0.000	4.000	8.000	9.927	14.000	55.000	4
<b>Regressão Linear vs. Linhas de aposta</b>	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.