



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

Previsão de Resultados da NBA na temporada 2018-19

Gustavo Pompeu da Silva

Orientador: Eduardo Monteiro de Castro Gomes

Brasília

2019

Gustavo Pompeu da Silva

Previsão de Resultados da NBA na temporada 2018-19

Relatório apresentado à disciplina de Trabalho de Conclusão de Curso II de graduação em Estatística, Departamento de Estatística, Instituto de Exatas, Universidade de Brasília, como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Orientador: Eduardo Monteiro de Castro Gomes

Brasília

2019

Gustavo Pompeu da Silva

Previsão de Resultados da NBA na temporada 2018-19/ Gustavo Pompeu da Silva. – Brasília, 2019-

43 p. : il. (algumas color.) ; 30 cm.

Orientador: Eduardo Monteiro de Castro Gomes

Relatório Final – Universidade de Brasília

Instituto de Ciências Exatas

Departamento de Estatística

Trabalho de Conclusão de Curso de Graduação, 2019.

1. NBA. 2. Previsão. 3. Resultados. 4. Regressão. 5. R. 6. Estatística.

Gustavo Pompeu da Silva

Previsão de Resultados da NBA na temporada 2018-19

Relatório apresentado à disciplina de Trabalho de Conclusão de Curso II de graduação em Estatística, Departamento de Estatística, Instituto de Exatas, Universidade de Brasília, como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Eduardo Monteiro de Castro Gomes
Orientador

Leandro Tavares Correia
Membro da Banca

Donald Matthew Pianto
Membro da Banca

Brasília
2019

Resumo

Este texto apresenta algumas notas de aula de TCC 1 com o formato da monografia que deve ser apresentada para conclusão do curso de Bacharelado em Estatística na Universidade de Brasília. O objetivo é apenas padronizar a apresentação do Trabalho de Conclusão de Curso, utilizando normas da ABNT e o pacote \LaTeX .

Palavras-chave: \LaTeX , abntex, pesquisa, monografia, slides, poster.

Abstract

This is the english abstract.

Keywords: L^AT_EX, abntex, research, pesquisa, monograph, slides, poster.monografia, slides, poster.

Lista de ilustrações

Figura 1 – Classificação Final da NBA separado por conferência Leste (esquerda)	
e Oeste (direita)	31

Lista de tabelas

Tabela 1 – Média de pontos sofridos por equipe	32
Tabela 2 – Média de pontos sofridos por equipe	33
Tabela 3 – Porcentagem de Acerto das previsões dos jogos da temporada 2018/19 para cada método	34

Lista de abreviaturas e siglas

NBA	<i>National Basketball Association</i>
SVM	<i>Support Vector Machine</i> (Máquina de Vetores de Suporte)
NA	Não aplicável

Sumário

1	INTRODUÇÃO	17
2	REVISÃO DE LITERATURA	19
2.1	Regressão Linear	19
2.2	Regressão Logística	20
2.3	Regressão de Probit	21
2.4	Máquina de Vetores de Suporte (SVM)	21
2.5	Análise de Discriminante Linear	21
2.6	Árvores de Regressão e Classificação	22
2.7	<i>Random Forest</i>	23
3	METODOLOGIA	25
3.1	Criação das Bases de Dados	25
3.1.1	Lidando com valores faltantes	29
4	RESULTADOS	31
4.1	Resultados Reais da Temporada 2018/19	31
4.1.1	Temporada Regular	31
4.1.2	<i>Playoffs</i>	33
4.2	Previsões	33
4.2.1	Previsões das casas de aposta	34
5	CONCLUSÃO	37
	REFERÊNCIAS	39
	APÊNDICES	41
	APÊNDICE A – CÓDIGOS EM R	43

1 Introdução

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), divididos em 2 conferências, uma do Leste e uma do Oeste. 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, atualmente, em uma temporada regular, cada time joga 82 vezes, ou seja, são 1230 jogos por temporada, isso nos permite ter muitas observações para trabalhar. Os 8 melhores times de cada conferência se classificam para os *playoffs* para disputar o título de campeão da NBA.

O objetivo geral desse trabalho é ajustar modelos utilizando diversas técnicas estatísticas para obter previsões para os resultados dos jogos da temporada de 2018-19 da NBA e compará-las para chegar em uma conclusão sobre qual técnica funcionou melhor para esse problema em específico, julgando principalmente pela acurácia das previsões.

2 Revisão de Literatura

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

- 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*

2.1 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:

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

Onde Y é a variável resposta (dependente), $\beta_j, j = 0, 1, \dots, p$ são constantes, denominados coeficientes de regressão, $X_j, j = 1, \dots, p$ 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 $X_1 = X_2 = \dots = X_p = 0$. 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, $\epsilon_i \stackrel{iid}{\sim} N(0, \sigma^2)$;

- 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)

2.2 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:

$$\text{logit}(\pi) = \log\left(\frac{\pi}{1 - \pi}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p \quad (2.1)$$

Onde $\pi = P(Y = 1)$. Baseado em 2.1, chegamos em:

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

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)

2.3 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:

$$\text{probit}(\pi) = \Phi^{-1}(\pi) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p \quad (2.2)$$

Onde $\pi = P(Y = 1)$. Baseado em 2.2, chegamos em:

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

. 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)

2.4 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.

2.5 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 (\mathbf{S}_p^2):

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

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, $\bar{\mathbf{x}}_1$ e $\bar{\mathbf{x}}_2$ correspondem às médias das variáveis independentes para cada população, e \mathbf{x}_{1j} e \mathbf{x}_{2j} são referentes à cada observação j de cada população.

Então, para uma nova observação \mathbf{x}_0 , temos: $\hat{y}_0 = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' \mathbf{S}_p^{-1} \mathbf{x}_0$ e $\hat{m} = \frac{1}{2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' \mathbf{S}_p^{-1} (\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)$, 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)

2.6 Árvores de Regressão e Classificação

As árvores de regressão e classificação são métodos estatísticos não-paramétricos utilizados baseado na teoria de árvores de decisão, a diferença é que em vez de decisões, os nós terminais da árvore são resultados numéricos (caso a variável resposta seja quantitativa), ou classes (caso a variável resposta seja qualitativa). No primeiro caso, chamamos o método de Árvore de Regressão, e no segundo de Árvore de Classificação.

Seu processo de construção é automático, e a ideia geral é particionar o espaço recursivamente em sub-regiões.

São alternativas não-paramétricas à regressão linear e regressão logística, logo, não necessitam de pressupostos para serem aplicadas.

O pacote usado no R é o *tree*. (RIPLEY, 2019)

2.7 Random Forest

O *random forest* é um método baseado nas Árvores de Classificação, a diferença é que ele separa a base de dados em vários subgrupos de observações aleatoriamente, e para cada grupo constrói uma árvore diferente, disso vem o nome "floresta aleatória", e no final faz uma média dos resultados das árvores.

Uma vantagem desse método é a prevenção de *overfitting*, mas em compensação é um método muito mais lento de ser computado, pois constrói muito mais árvores.

O pacote usado no R é o *randomForest*. (LIAW; WIENER, 2002)

3 Metodologia

A linguagem R (R Core Team, 2018) será utilizada para toda a implementação computacional necessária para o trabalho.

3.1 Criação das Bases de Dados

Para a obtenção dos dados necessários, será utilizada uma técnica de *web scraping*, em que informações são extraídas de alguma página da internet. Com o auxílio do pacote *rvest* (WICKHAM, 2016), serão extraídas informações dos jogos das temporadas de 2000/01 até 2018/19 da NBA do site Basketball Reference (BASKETBALL...), um dos maiores sites com dados numéricos sobre a NBA e basquete em geral. Foi escolhido começar dos anos 2000, pois 18 temporadas teriam muitas observações para realizar as modelagens para as previsões da temporada 2018/19.

O número de jogos por temporada varia por alguns motivos. Nas temporadas de 2000/01 a 2004/05, a NBA era composta por apenas 29 times, mas cada time já jogava 82 jogos, isso resultava em 1189 jogos por temporada regular. Nas demais temporadas utilizadas, 30 times faziam parte da liga, resultando em 1230 jogos por temporada. A única exceção foi a temporada de 2011/12, quando aconteceu um *lockout*, quando os donos das equipes se recusaram a deixar os jogos acontecerem, pois o contrato da NBA com os times acabou antes do início da temporada, e a NBA demorou para chegar em um acordo com os donos dos times para assinarem um novo contrato. Um novo acordo foi estabelecido depois de vários meses de negociação, e a temporada começou em 25 de dezembro de 2011, com quase 2 meses de atraso. Isso diminuiu o número de jogos realizados por cada equipe de 82 para 66 jogos, que resultou em um total de apenas 990 jogos na temporada regular.

A extração dos dados foi realizada da seguinte maneira: na página da internet onde estão os dados, foi utilizada uma extensão do *Google Chrome*, o *SelectorGadget* (SELECTOR...), que permite selecionar as partes do site desejadas, e através disso, com funções do pacote *rvest*, são obtidas essas partes em HTML no R, e transformadas para texto, assim, toda a informação desejada é colocada no R.

As informações obtidas de cada um dos jogos realizados das temporadas citadas são: data do jogo, nome do time visitante, pontos marcados pelo time visitante, nome do time mandante, pontos marcados pelo time mandante, se houve prorrogação no jogo, e o público presente no ginásio.

A partir do que foi obtido, podemos criar uma base de dados com muitas variáveis derivadas dessas informações, e então criar modelos para realizar as previsões, utilizando

as diversas técnicas estatísticas citadas anteriormente.

É importante ressaltar que a temporada 2018/19 estava em andamento durante a realização desse trabalho, tendo durado de Outubro de 2018 até Junho de 2019, e as informações relacionadas aos jogos dessa temporada foram sendo extraídas gradualmente conforme os jogos foram sendo realizados.

A base de dados inicial, criada a partir das informações extraídas da internet, contém 2 linhas para cada jogo realizado, cada linha tendo informações referentes à um dos times envolvidos na partida, e contém as seguintes variáveis:

Variáveis de identificação das informações do jogo:

- *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 time jogou em casa ou não.
- *Attend*: Público presente no ginásio nesse jogo. (Tem a informação apenas se o time jogou em casa)
- *OT*: Indica se ocorreu prorrogação no jogo.

Variáveis indicadoras do resultado do jogo:

- *Win*: Se o time venceu esse jogo ou não (qualitativa, dicotômica).
- *Result*: Saldo de pontos, ou seja, os pontos marcados pelo time menos os pontos marcados pelo seu adversário (quantitativa).

Variáveis de informação sobre o jogo que podem ser identificadas antes da realização da partida:

- *weekday*: Dia da semana em que o jogo foi/será realizado. Essa variável pode ser utilizada nas modelagens, pois sabemos o dia da semana que o jogo ocorrerá mesmo antes do jogo acontecer.
- *Travel*: Variável que indica se o time teve/terá que viajar da partida anterior para essa ou não. Por exemplo, se o jogo anterior foi fora de casa, o time sempre tem que viajar, ou pra voltar pra casa, ou pra ir para outra cidade fora de casa. O time só não viaja quando joga 2 jogos seguidos em casa.

Variáveis referentes à toda informação do time desde o início da temporada até antes do jogo:

- *Games_T*: Total de jogos do time até agora na temporada.
- *Games_H*: Jogos em casa do time até agora na temporada.
- *Games_A*: Jogos fora de 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.
- *Wins_A*: Vitórias fora de casa do time até agora na temporada.
- *Loss_T*: Total de derrotas do time até agora na temporada.
- *Loss_H*: Derrotas em casa do time até agora na temporada.
- *Loss_A*: Derrotas fora de casa do time até agora na temporada.
- *Streak_T*: Número indicando a sequência de vitórias (positivo) ou derrotas (negativo) do time, considerando jogos em casa e fora de casa. Exemplos: os 5 últimos jogos do time foram vitórias, e o antes desses 5 foi derrota, logo a variável vale +5. O último jogo do time foi derrota, e o penúltimo vitória, então a variável vale -1.
- *Streak_H*: Número indicando a sequência de vitórias (positivo) ou derrotas (negativo) do time, considerando apenas jogos em casa.
- *Streak_A*: Número indicando a sequência de vitórias (positivo) ou derrotas (negativo) do time, considerando apenas jogos fora de casa.
- *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 sofridos do 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 sofridos do time em jogos fora de casa até o momento na temporada.

- *Mean_Pts_A_T*: Média de pontos sofridos do time em todos os jogos até o momento na temporada.
- *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 na temporada. Divide-se o total de vitórias de todos os adversários do time pelo total de jogos de todos os adversários do time.
- *mean_attend*: Média de público do time nos jogos em casa, até o momento na temporada.

Variáveis referentes aos últimos 3, 5, 7 ou 10 jogos do time 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, 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, 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, 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 sofridos do time nos últimos X jogos fora de casa, 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 sofridos do time nos últimos X jogos em casa, 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 sofridos do time nos últimos X jogos, onde $X = 3, 5, 7, 10$.
- *Win_Last_X_A*, *Win_Last_X_H*, *Win_Last_X_T*: Número de vitórias do time nos últimos X jogos fora de casa, em casa, e total, respectivamente, onde $X = 3, 5, 7, 10$.

Variáveis referentes apenas ao último jogo realizado pelo time:

- *OT_Last*: Indica se houve prorrogação no jogo anterior do time.
- *Days_LG*: Quantos dias atrás foi o último jogo do time.

Isso resulta em 1 banco de dados para cada temporada, com 2 linhas para cada jogo realizado na temporada, e 125 variáveis. Porém, algumas variáveis de informação não podem ser utilizadas nas modelagens.

Vamos utilizar *Win* (dicotômica, qualitativa) ou *result* (quantitativa) como variáveis dependentes para os modelos, a escolha dessa variável irá depender da técnica utilizada.

Como o objetivo é realizar uma previsão para cada jogo, as duas linhas de cada jogo serão combinadas em uma só, ou seja, uma primeira parte do banco de dados final terá apenas variáveis referentes ao time visitante de cada jogo, e a segunda parte apenas variáveis referentes ao time mandante. Como existe essa separação clara entre as variáveis, é fácil remover as que ficariam duplicadas (como o dia da semana do jogo e as variáveis dependentes), e as que não teriam propósito, (como a média de público do time visitante, que não é aplicável, e a variável que indica se o time viajou do último jogo para o atual para o time visitante, pois ela sempre será *TRUE*). Além disso, também serão retiradas as variáveis referentes à jogos fora de casa para os times mandantes e as referentes à jogos em casa para os times visitantes, pois foi julgado que elas não contribuiriam.

Isso resulta em uma base final por temporada, com 1 linha por jogo, e 151 variáveis, sendo 2 delas as variáveis dependentes. As variáveis dependentes mantidas foram as referentes ao time visitante, ou seja, a variável *Win* virou *Win_Vis* e tem valor *TRUE* quando o time visitante vence, e *FALSE* caso contrário, e a variável *result* virou *result_Vis* e é positiva quando o time visitante vence, e negativa caso contrário. De forma similar, para todas as variáveis restantes na base de dados, foi adicionado a extensão *_Vis* no nome das que são referentes ao time visitante, e a extensão *_Home* no nome das que são referentes ao time mandante. A única variável que não é referente a nenhum dos dois, é a do dia da semana do jogo (*weekday*), e o nome dela foi mantido.

3.1.1 Lidando com valores faltantes

Como existem muitas variáveis que dependem de resultados anteriores, existirão vários valores faltantes no começo de cada temporada, que nesse trabalho serão referenciados como NA (não aplicável). Por exemplo, se o time só realizou 6 jogos na temporada, não é possível obter um valor para a média de pontos marcados nos últimos 7 jogos, ou se é o primeiro jogo do time na temporada, não há como obter nenhuma informação além do dia da semana do jogo.

Na base de dados da temporada 2018/19, que é a que será feita as previsões, das 1230 observações da temporada regular, apenas 875 possuem informação completa, ou seja, nenhum NA em nenhuma variável. Nas outras 355 observações, existe pelo menos um NA em alguma variável.

No R, quando é feita uma modelagem, as observações que possuem algum valor

NA são completamente ignoradas, e as previsões de observações com algum valor NA são retornadas como NA também. Para tornar possível fazer as previsões de todos os jogos da temporada 2018/19, e não apenas daqueles que tem informação completa, é possível identificar os padrões diferentes de NA's nas linhas para a base dessa temporada transformando a base de dados toda em uma matriz de 0's e 1's, sendo 0 quando a observação tem informação, e 1 quando ela é NA. Assim, cada linha da base se torna um vetor de 0's e 1's, e o padrão de cada linha pode ser identificado colando esses 0's e 1's usando a função *paste*. Feito isso, foram identificados 61 padrões diferentes para a base da temporada 2018/19. Então, para cada padrão, são identificadas as linhas que possuem aquele padrão, e as colunas das variáveis que são NA nesse padrão são retiradas. Por fim, em uma base que contém as observações das temporadas anteriores, retiramos essas mesmas colunas, e depois são retiradas as linhas em que ainda existe algum NA nas colunas que sobraram. Com isso, é possível fazer modelos com essa base das temporadas anteriores, e utilizá-los para fazer previsões para todos os 1230 jogos da temporada regular de 2018/19.

4 Resultados

4.1 Resultados Reais da Temporada 2018/19

Para efeito de comparação com as previsões a serem feitas, será listado aqui algumas estatísticas importantes da temporada.

4.1.1 Temporada Regular































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

Figura 1 – Classificação Final da NBA separado por conferência Leste (esquerda) e Oeste (direita)

Na figura 1, é observado o número de vitórias (W) e derrotas (L) de cada time na temporada 2018/19. Os 8 primeiros times de cada conferência se classificaram para os *playoffs*.

Tabela 1 – Média de pontos sofridos por equipe

Fonte: (NBA..., b)

Time	Média de Pontos Sofridos
Indiana Pacers	104.7
Miami Heat	105.9
Memphis Grizzlies	106.1
Utah Jazz	106.5
Orlando Magic	106.6
Denver Nuggets	106.7
Detroit Pistons	107.3
Boston Celtics	108
Toronto Raptors	108.4
Houston Rockets	109.1
Milwaukee Bucks	109.3
San Antonio Spurs	110
Dallas Mavericks	110.1
Portland Trail Blazers	110.5
Oklahoma City Thunder	111.1
Golden State Warriors	111.2
Charlotte Hornets	111.8
Brooklyn Nets	112.3
Philadelphia 76ers	112.5
Chicago Bulls	113.4
Los Angeles Lakers	113.5
New York Knicks	113.8
Minnesota Timberwolves	114
Cleveland Cavaliers	114.1
LA Clippers	114.3
Sacramento Kings	115.3
New Orleans Pelicans	116.8
Phoenix Suns	116.8
Washington Wizards	116.9
Atlanta Hawks	119.4

Na tabela 1, é possível ver que o Indiana Pacers teve a melhor defesa da temporada no quesito pontos, e o Atlanta Hawks a pior.

Tabela 2 – Média de pontos sofridos por equipe

Fonte: (NBA..., b)

Time	Média de Pontos Marcados
Milwaukee Bucks	118.1
Golden State Warriors	117.7
New Orleans Pelicans	115.4
Philadelphia 76ers	115.2
LA Clippers	115.1
Portland Trail Blazers	114.7
Oklahoma City Thunder	114.5
Toronto Raptors	114.4
Sacramento Kings	114.2
Washington Wizards	114
Houston Rockets	113.9
Atlanta Hawks	113.3
Minnesota Timberwolves	112.5
Boston Celtics	112.4
Brooklyn Nets	112.2
Los Angeles Lakers	111.8
Utah Jazz	111.7
San Antonio Spurs	111.7
Charlotte Hornets	110.7
Denver Nuggets	110.7
Dallas Mavericks	108.9
Indiana Pacers	108
Phoenix Suns	107.5
Orlando Magic	107.3
Detroit Pistons	107
Miami Heat	105.7
Chicago Bulls	104.9
New York Knicks	104.6
Cleveland Cavaliers	104.5
Memphis Grizzlies	103.5

Da tabela 2, percebe-se que o Milwaukee Bucks foi o time que mais marcou pontos na temporada, e o Memphis Grizzlies o que menos fez.

4.1.2 Playoffs

A final foi disputada por Toronto Raptors e Golden State Warriors.

4.2 Previsões

Utilizando as bases de dados definidas na Metodologia, será possível aplicar os métodos estatísticos propostos na Revisão de Literatura para fazer modelos e obter

previsões para os 1230 jogos da temporada regular de 2018/19.

De início, foram utilizadas todas as bases disponíveis (da temporada 2000/01 até 2017/18) para o ajuste dos modelos. Diferentemente do usual na estatística, a base não foi dividida em treinamento e validação, pois os jogos são uma sequência histórica no tempo, e não faria sentido utilizar jogos do “futuro” para prever jogos que aconteceram antes.

Tabela 3 – Porcentagem de Acerto das previsões dos jogos da temporada 2018/19 para cada método

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

Na tabela 3, temos a porcentagem de acerto das previsões para cada método utilizado. O melhor resultado foi obtido com a Regressão de Probit, com 0.6723577 de acurácia. Isso foi obtido utilizando as temporadas de 2000/01 até 2017/18 para as modelagens.

Alguns métodos são mais eficientes computacionalmente, em termos de tempo decorrido para a execução da implementação, e para esses métodos, é possível realizar as previsões modificando as temporadas a serem utilizadas na modelagem. Esses métodos são: regressão linear, regressão logística, regressão de probit, análise de discriminante linear, regressão em árvore e classificação em árvore.

...

Inserir gráficos ou tabelas com a porcentagem de acerto com menos temporadas no modelo.

...

4.2.1 Previsões das casas de aposta

A “linha” é um artifício que as casas de aposta usam para equilibrar a aposta em um time em cada partida, a casa considera o time A favorito pra ganhar por X pontos, então coloca uma linha, que para os apostadores apostarem no time A e ganharem, o time

A precisa vencer o jogo por essa diferença de X pontos. Um exemplo: uma casa de aposta diz que o Golden State Warriors é o favorito em um jogo 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, ou seja, quem apostar nos Warriors precisa que o time vença o jogo por uma diferença de 7 pontos ou mais para ganhar a aposta, e quem apostar no Trail Blazers precisa que o time perca por no máximo 6 pontos de diferença, ou vença o jogo, para ganhar a aposta. É comum o uso de meio décimo na linha para evitar empates, mas não é obrigatório, por exemplo, caso uma linha seja -7 e o time vença por exatos 7 pontos de diferença, normalmente a aposta é ressarcida ao apostador e nem ele, nem a casa ganha a aposta.

Utilizando-se da mesma técnica de *web scraping* do pacote *rvest* (WICKHAM, 2016) que foi usada para obter os dados necessários para fazer as previsões, foi possível extrair do site da ESPN Americana (ESPN...,) a “linha” de aposta de cada jogo da temporada. O site da ESPN faz uma média da “linha” de várias casas de aposta diferentes logo antes do início de cada partida, e deixa na página de cada jogo. Como a “linha” diz qual time era considerado o favorito para vencer o jogo pelas casas de aposta, podemos obter a porcentagem de acerto média das casas de aposta para os jogos da temporada.

Na página de 4 dos 1230 jogos, o valor da “linha” não estava disponível, e em outros 16 dos 1230 jogos, a “linha” era *even*, ou seja, esses jogos foram julgados tão equilibrados, que em média não foram apontados times favoritos.

Excluindo esses 20 jogos citados acima, a porcentagem de acerto calculada para as casas de aposta foi de 0.67272, em 1210 jogos.

5 Conclusão

aaa

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>>. Nenhuma citação no texto.
- BASKETBALL Reference. <<https://www.basketball-reference.com/>>. Acessado em: 28/09/2018. Citado na página 25.
- CARVALHO, J. et al. *ANÁLISE DE PROBIT APLICADA A BIOENSAIOS COM INSETOS*. [S.l.: s.n.], 2017. ISBN 978-85-64937-08-6. Citado na página 21.
- ESPN NBA Scores. <<http://www.espn.com/nba/scoreboard>>. Acessado em: 16/05/2019. Citado na página 35.
- GROTHENDIECK, G. *sqldf: Manipulate R Data Frames Using SQL*. [S.l.], 2017. R package version 0.4-11. Disponível em: <<https://CRAN.R-project.org/package=sqldf>>. Nenhuma citação no texto.
- 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>>. Citado na página 22.
- 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>>. Nenhuma citação no texto.
- 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/>>. Citado na página 23.
- 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>>. Citado na página 21.
- 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>>. Nenhuma citação no texto.
- NBA Logos. <http://www.sportslogos.net/teams/list_by_league/6/National_Basketball_Association/NBA/logos/>. Acessado em: 13/10/2018. Nenhuma citação no texto.
- NBA Stats. <<https://stats.nba.com/>>. Acessado em: 28/09/2018. Citado 2 vezes nas páginas 32 e 33.
- R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2018. Disponível em: <<https://www.R-project.org/>>. Citado 3 vezes nas páginas 20, 21 e 25.

RIPLEY, B. *tree: Classification and Regression Trees*. [S.l.], 2019. R package version 1.0-40. Disponível em: <<https://CRAN.R-project.org/package=tree>>. Citado na página 22.

SCHUMAKER, R. P.; SOLIEMAN, O. K.; CHEN, H. *Sports Data Mining*. 1st. ed. [S.l.]: Springer Publishing Company, Incorporated, 2010. ISBN 144196729X, 9781441967299. Citado na página 17.

SELECTOR Gadget. <<https://selectorgadget.com/>>. Acessado em: 25/05/2019. Citado na página 25.

UUDMAE, J. Predicting nba game outcomes. Acessado em: 28/09/2018. Disponível em: <<http://cs229.stanford.edu/proj2017/final-reports/5231214.pdf>>. Nenhuma citação no texto.

VAPNIK, V. N. *The Nature of Statistical Learning Theory*. Berlin, Heidelberg: Springer-Verlag, 1995. ISBN 0-387-94559-8. Citado na página 21.

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>>. Citado na página 22.

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>>. Citado 2 vezes nas páginas 25 e 35.

WICKHAM, H. *stringr: Simple, Consistent Wrappers for Common String Operations*. [S.l.], 2019. R package version 1.4.0. Disponível em: <<https://CRAN.R-project.org/package=stringr>>. Nenhuma citação no texto.

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>>. Nenhuma citação no texto.

Apêndices

APÊNDICE A – Códigos em R

Quisque facilisis auctor sapien. Pellentesque gravida hendrerit lectus. Mauris rutrum sodales sapien. Fusce hendrerit sem vel lorem. Integer pellentesque massa vel augue. Integer elit tortor, feugiat quis, sagittis et, ornare non, lacus. Vestibulum posuere pellentesque eros. Quisque venenatis ipsum dictum nulla. Aliquam quis quam non metus eleifend interdum. Nam eget sapien ac mauris malesuada adipiscing. Etiam eleifend neque sed quam. Nulla facilisi. Proin a ligula. Sed id dui eu nibh egestas tincidunt. Suspendisse arcu.