Instituto de MATEMÁTICA E ESTATÍSTICA UFRGS

DEPARTAMENTO DE ESTATÍSTICA
E-mail: dest@mat.ufrgs.br

Trabalho de Conclusão de Curso

Desempenho individual no *college basketball* e sucesso profissional na *National Basketball Association*: uma abordagem de Machine Learning

Enzo Bertoldi Oestreich

16 de maio de 2022

Enzo Bertoldi Oestreich

Desempenho individual no *college basketball* e sucesso profissional na *National Basketball Association*: uma abordagem de Machine Learning

Trabalho de Conclusão apresentado à comissão de Graduação do Departamento de Estatística da Universidade Federal do Rio Grande do Sul, como parte dos requisitos para obtenção do título de Bacharel em Estatística.

Orientador: Prof. Dr. João Henrique Ferreira Flores

Porto Alegre
Maio de 2022

Enzo Bertoldi Oestreich

Desempenho individual no *college basketball* e sucesso profissional na *National Basketball Association*: uma abordagem de Machine Learning

Este Trabalho foi julgado adequado para obtenção dos créditos da disciplina Trabalho de Conclusão de Curso em Estatística e aprovado em sua forma final pela Orientador e pela Banca Examinadora.

Orientador:_____
Prof. Dr. João Henrique Ferreira Flores, UFRGS
Doutor pela Universidade Federal do Rio Grande do Sul, Porto Alegre, RS

Banca Examinadora:

Prof. Dr. João Henrique Ferreira Flores, UFRGS
Doutor pela Universidade Federal do Rio Grande do Sul, Porto Alegre, RS

Prof. Dr. Rodrigo Citton Padilha dos Reis, UFRGS
Doutor pela Universidade Federal de Minas Gerais, Belo Horizonte, MG

Porto Alegre
Maio de 2022

Resumo

Este trabalho tem como objetivo avaliar o poder preditivo de estatísticas do basquetebol colegial para o sucesso na NBA utilizando os jogadores selecionados nos *drafts* das temporadas de 2010 à 2017. O objetivo é identificar quais variáveis contribuem de forma positiva/negativa para o *Value Over Replacement Player* (VORP) de um atleta de forma a estabelecermos uma relação que nos auxilie a detectar uma nova estrela do basquetebol. Para analisar o banco de dados deste trabalho foram aplicadas técnicas de redução de dimensionalidade, como a análise de componentes principais (PCA). No intuito de avaliar a capacidade preditiva das variáveis selecionadas, foram utilizadas as técnicas de regressão linear múltipla e redes neurais artificiais. Os resultados mostram que embora não possuímos a capacidade de distinguir uma nova estrela apenas utilizando as estatísticas colegiais, as técnicas de aprendizagem de máquina se mostram como ferramenta auxiliar para melhoria do ordenamento do *draft*.

Palavras-Chave: Redes Neurais, Performance, NBA, Aprendizagem de Máquina.

Abstract

This work aims to assess the predictive power of college basketball statistics with regards to success in the NBA using players drafted between 2010 and 2017 NBA seasons. The purpose of this study is to identify which variables contribute positively/negatively to the Value Over Replacement Player (VORP) of an athlete in order to establish a relationship that helps us predict a new basketball star. In order to analyze the database of this study, dimensionality reduction techniques were applied, such as principal component analysis (PCA). To evaluate the predictive power of the selected variables, multiple linear regression and artificial neural networks were used. The results show that although we do not have the ability to distinguish a new star using only college basketball statistics, machine learning techniques are shown to be a helpful tool for improving the draft order selection.

Keywords: Neural Networks, College Performance, NBA, Machine Learning.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas

NBA *National Basketball Association*

PCA *Principal Component Analysis*

MVP *Most Valuable Player*

VORP *Value Over Replacement Player*

BPM *Box Plus/Minus*

MSE *Mean Squared Error*

RMSE *Root Mean Squared Error*

MAE *Mean Absolute Error*

1 Introdução

1.1 Comentários Iniciais

A maior inserção de novos talentos na *National Basketball Association* (NBA) ocorre através do *draft*. Um evento anual, realizado desde 1947, onde as equipes selecionam candidatos advindos do basquete universitário ou internacional. Além do acréscimo de novos jogadores aos plantéis, o *draft* possui o intuito de estabelecer a paridade na liga. De acordo com o formato atual, equipes com as piores campanhas na temporada anterior possuem maiores chances de serem as primeiras a escolher, consequentemente as melhores equipes serão as últimas a selecionarem seus jogadores NBA (2021).

Devido a esta definição do ordenamento nas escolhas, a cada temporada novas equipes se submetem ao processo de *tanking* [1] na esperança de obterem as primeiras posições do *draft* e, quem sabe, *draftar* a nova estrela da franquia.

Anualmente, os maiores meios de comunicação esportiva divulgam rankings dos melhores candidatos elegíveis para o *draft* CBS (2021). Com a ampliação da cultura orientada a dados e o alto investimento direcionado a *scouts* Alamar (2021), pressupõe-se que a primeira escolha seja destinada ao melhor jogador disponível, sendo acompanhado pelo segundo e subsequentemente até a última seleção. Tal premissa não se solidifica como verdade absoluta ao olharmos para o passado, uma vez que inúmeras equipes falharam ao prever a próxima estrela. O fracasso mais recente decorreu no draft de 2013, a equipe do *Cleveland Cavaliers* selecionou Anthony Bennett, versátil e atlético ala de força (1), como número 1 da noite. Após 4 temporadas com pontuação média abaixo dos 5 pontos, Anthony não se encontra em nenhum elenco da liga americana e até hoje é considerado a pior primeira escolha da história. Na mesma noite, o décimo quinto jogador a ser selecionado foi Giannis Antetokounmpo. Atualmente o grego é detentor de dois prêmios de *Most Valuable Player* (MVP) e nome confirmado no hall da fama do basquetebol Soliven (2021).

1.2 Problematização

Com base nesta recorrência de jogadores ficarem para trás no *draft*, gostaríamos de enteder diante tal disparidade quais fatores discriminam um atleta de alto nível performático em relação aos demais. Para avaliar o possível sucesso de um joga-

[1] Processo no qual equipes propositalmente mantém um desempenho abaixo da média ao longo da temporada para obterem mais chances de possuir a primeira escolha no *draft*

dor recém-chegado a liga, precisamos levar em conta fatores externos que possam influenciar o seu desenvolvimento Kaufman (2014). Como muitas dessas variáveis são não-mensuráveis, neste trabalho nos atemos a analisar dados pré-NBA e sua capacidade de predizer a performance de um novato no basquete profissional. Especificamente estaremos focados nos dados do basquetebol colegial visto que em torno de 75% dos atletas selecionados no *draft* advém do *college*,

1.3 Objetivo geral

Dado que a disponibilidade de dados do basquetebol colegial se assemelha com a própria NBA, neste trabalho estaremos focados em avaliar o poder preditivo das estatísticas do colegial para o sucesso na NBA.

1.4 Objetivos específicos

Estaremos interessado neste trabalho em:

- Definir o melhor modelo a ser utilizado para predição da nossa variável resposta;
- Estabelecer balizador para mensurar o sucesso de um jogador;
- Identificar as variáveis que contribuem de forma positiva para nossa variável resposta;

1.5 Hipóteses de Pesquisa

Para podermos realizar nosso estudo, tomamos como pressuposto a hipótese de existir uma relação entre dados provindos do basquetebol colegial (ou internacional) com o sucesso desenvolvido na liga profissional, tomando como base os anos do atleta sob contrato de novato.

Sob tal pressuposto esperamos poder extrair características relevantes na descoberta de novos talentos, para então realizarmos previsões relacionadas ao sucesso de um jogador de forma a melhorarmos as escolhas no *draft*. No próximo capítulo veremos alguns modelos que nos auxiliem a responder estas hipóteses de pesquisa.

2 Métodos

2.1 Introdução

Esta seção apresenta os principais conceitos abordados pela pesquisa desenvolvida. Acerca do tema que engloba as análises preditivas no contexto esportivo, elucidamos os conceitos referentes ao mesmo, subdivindo este capítulo em dois tópicos principais, sendo um deles as técnicas e métodos utilizados assim como os conceitos e definições de nossa variável de interesse.

2.2 Como avaliar o sucesso profissional de um jogador na NBA

Quando tratamos da análise de performance no basquetebol profissional nos deparamos com um impasse que se resume à definição de sucesso: quais estatísticas devem ser utilizadas para parametrizar e separar uma estrela de um atleta de baixo desempenho. Ao utilizarmos o conceito de sucesso não estamos atrelando-o a fama, dado que este pode ser monitorado através de votos do *All Star Game*, vendas de camisetas e outros fatores, mas sim àquele jogador cuja presença no plantel de sua equipe seja indispensável para a prosperidade da mesma.

No intuito de explorar tais métricas de performance, encontramos estudos que utilizam diferentes abordagens para responder a mesma pergunta, revelando uma certa subjetividade no que diz respeito ao conceito de sucesso. No trabalho de Kannan et al. (2018) temos uma visão baseada em longevidade de carreira onde, a partir de um *threshold* pré-definido pelo autor, separamos os atletas entre bem-sucedidos ou não a partir da quantidade de partidas do mesmo na liga profissional, tendo como resposta uma variável binária; já Berri et al. (2011) traz uma visão metrificada de desempenho utilizando como medida de performance a estatística *Win Shares Per 48*, a qual baseia-se no *box score*[1] para dividir entre os jogadores o crédito pelas vitórias da equipe, buscando estimá-la a partir de dados do basquetebol colegial.

Como notamos acima, a definição de um atleta bem sucedido profissionalmente varia entre autores. Partindo desse pressuposto, neste trabalho iremos utilizar como balizador de sucesso a ideia de avaliar o quão essencial um jogador é para sua equipe, ou seja, entender o impacto que o mesmo possui nas partidas. Como estamos

[1] Tabela de estatísticas contabilizadas durante uma partida

interessados na comparação/avaliação de atletas e estimação da variável resposta, iremos mensurar o sucesso através de uma métrica quantitativa denominada *Value Over Replacement Player* (VORP), a qual será abordada na próxima seção.

2.3 *Value Over Replacement Player*

Com o avanço das tecnologias de monitoramento e obtenção de dados referentes a jogadores ao longo das partidas, o chamado *play-by-play data*, surgem, cada vez mais, novas métricas avançadas para avaliação de desempenho, porém estes dados não possuem grande disponibilidade. Deste modo, decidimos utilizar uma métrica baseada no *box score* tradicional. No entanto, para falarmos de VORP, primeiramente precisamos entender o conceito de *Box Plus/Minus* (BPM), visto que o primeiro é um derivado do segundo.

Como citado anteriormente, o BPM é baseado em estatísticas presentes no *box score*, além de levar em conta a posição do atleta e a performance geral da equipe. Esta métrica foi criada por Myers (2020) para avaliar o impacto de um jogador em quadra quando comparado a média da liga, estimando o número de pontos por 100 posses de bola. Para exemplificar o uso desta estatística tomamos LeBron James na temporada de 2009-2010, com BPM +11.8, significando que sua equipe era 11.8 pontos melhor (por 100 posses) quando o mesmo se encontrava em quadra. No intuito de visualizar uma escala para esta métrica, dado que quanto maior o seu valor melhor é o desempenho do jogador, temos na Tabela (2.1) uma exemplificação criada pelo criador do BPM:

Tabela 2.1: Escala de Performance BPM

Escala	Classificação	Interpretação
+10	*all-time season*	Desempenho Histórico
+8	*MVP season*	Desempenho de *MVP*
+6	*all-NBA season*	Desempenho de *all-NBA*
+4	*all-star consideration*	Desempenho de *all-star*
+2	*good starter*	Desempenho de um bom titular
+0	*starter or 6th man*	Desempenho de titular/sexto homem
-2	*bench player*	Desempenho de reserva

Dado os conceitos apresentados na Tabela (2.1), podemos seguir para o entendimento de nossa variável resposta. O VORP é uma métrica que leva em conta o tempo em quadra de um atleta, convertendo o BPM em uma estimativa da contribuição do jogador para a equipe, tendo como referência o *replacement player*, sendo este jogador considerado um reserva com BPM no valor de -2. Para obtermos o VORP de um jogador basta seguirmos:

$$VORP = [BPM - (-2)] \times (\%PP) \times (G/82) \tag{2.1}$$

onde:
BPM = *Box Plus Minus*;
-2 = BPM de um jogador definido como *replacement player*;
$\%PP$ = número de minutos jogados dentre os minutos possíveis;
G = número de jogos disputados;

Utilizando a Equação 2.1, obtemos o número de pontos que um jogador produz a mais que um *replacement player* por 100 posses da equipe, desta forma podemos comparar os diferentes atletas de acordo com seu VORP, sendo que, ao seguir esta regra, quanto maior o VORP melhor. Para enterdermos a relação de nossa variável resposta com as covariáveis preditoras, estaremos abordando as técnicas de aprendizagem de máquina próxima seção.

2.4 Aprendizagem de Máquina

Aprendizagem de máquina é um campo da inteligência artificial baseado no uso e desenvolvimento de sistemas computacionais capazes de aprender através da experiência e repetição sem seguir instruções explícitas, utilizando algoritmos e modelos estatísticos para obter inferências de padrões nos dados (Mitchell, 1997). Esta área tem como intuito providenciar à sistemas a capacidade de solucionar tarefas de complexa execução com a menor intervenção humana possível. No intuito de replicar o comportamento humano, máquinas são alimentadas com dados e, através da repetição de tarefas, são capazes de desenvolver a habilidade de tomar decisões.

Devido a contínua disseminação da cultura orientada a dados e o aumento de visibilidade da área de *data science*, a utilização de técnicas relacionadas a inteligência artificial tem se tornado ferramenta indispensável dentro do contexto esportivo (Tichy, 2016). Por conta da vasta disponibilidade de dados e a constante necessidade de equipes se adaptarem a evoluções nos estilos de jogo, o uso de técnicas de aprendizado de máquina ganham cada vez mais força para equipes que buscam o aprimoramento da performance dentro e fora de quadra (Millington e Millington, 2015).

Dentro do contexto de aprendizagem de máquina nos deparamos com a divisão entre três tópicos: aprendizado supervisionado, não-supervisionado e semi-supervisionado, embora o último não seja tratado neste trabalho. O ponto de divergência entre os métodos citados se dá na presença ou ausência de um vetor resposta nos dados. Ambas as técnicas serão apresentadas nas seções a seguir.

2.5 Aprendizagem Não-Supervisionada

A aprendizagem não-supervisionada diz respeito ao conjunto de algoritmos utilizados para identifição de padrões dentro um *dataset* não categorizado. O termo não-supervisionado se encaixa neste contexto de forma que não possuímos uma variável resposta para monitorar nossos resultados, ou seja, o modelo trabalha sem a supervisão do valor de resposta/classe. De modo geral, as técnicas utilizadas são aplicadas como parte do processo de análise exploratória, dado que sua finalidade seja aprender sobre os dados, sem intuito de previsão (James et al., 2013).

Entre os diferentes métodos disponíveis, elucidamos a recorrência da utilização de técnicas para redução de dimensionalidade como a análise de componentes principais (PCA), dado que lidamos com grande quantidade de variáveis explicativas, como visto em em (Hoffman e Joseph, 2017). Método, este, que será melhor detalhado a seguir.

2.5.1 *Principal Component Analysis (PCA)*

A era do *Big Data* também atingiu o mundo esportivo. Com a evolução de mecanismos para coleta e rastreamento de estatísticas individuais, aliados ao desenvolvimento de novas métricas para avaliação de performance, nos deparamos com um grande volume de dados a serem analisados. Neste contexto, ressaltamos a utilização de métodos para redução de dimensionalide destes dados, sendo PCA uma técnica constantemente aplicada no cenário do basquetebol (Hoffman e Joseph, 2017).

Dado um conjunto de dados formado por uma série de variáveis correlacionadas, a aplicação de componentes principais visa a sumarização dos dados em uma representação de espaço dimensional reduzido, a qual retém a maior porção de variação possível (James et al., 2013). Tal processo é realizado através de uma transformação ortogonal dos dados, alterando as coordenadas originais.

Definindo uma série de covariáveis $X_1, X_2, ..., X_p$, os componentes principais Z_i são combinações lineares normalizadas dessas variáveis que seguem o modelo:

$$\begin{aligned} Z_1 &= \phi_{11}X_1 + \phi_{21}X_2 + \ldots + \phi_{1p}X_p, \\ Z_2 &= \phi_{21}X_1 + \phi_{22}X_2 + \ldots + \phi_{2p}X_p, \\ &\vdots \\ Z_p &= \phi_{p1}X_1 + \phi_{p2}X_2 + \ldots + \phi_{pp}X_p, \end{aligned} \tag{2.2}$$

Onde ϕ_p representa a carga de cada componente principal.

Para obtermos estes componentes devemos seguir as seguintes restrições:

- A variabilidade explicada por cada componente decresce em relação ao anterior, ou seja, $\text{Var}(Z_1) > \text{Var}(Z_2) > ... > \text{Var}(Z_p)$;
- Cada componente deve ser não-correlacionado com o anterior (ortogonal);
- A soma dos quadrados das cargas ϕ devem somar 1;

Como estamos no contexto algébrico, podemos interpretar cada vetor de cargas como a direção no espaço de variáveis onde os dados apresentam maior variabilidade. Outro resultado que podemos obter ao utilizar este método se diz respeito a variáveis latentes, visto que variáveis com altas cargas nos mesmos componentes são correlacionadas entre si, podemo supor, mesmo que de forma empírica, a presença padrões nos dados que antes não eram perceptíveis (Bishop, 2006).

Para a seleção do número de componentes a serem retidos utilizaremos dois métodos distintos descritos abaixo.

- Critério de Kayser: esta abordagem apresentada por Kayser e Guttman retém apenas aqueles componentes cujos autovalores possuem valor próximo ou acima de 1 (Floyd e Widaman, 1995);
- *Scree Plot*: esta abordagem busca encontrar um ponto de corte baseado na suavização do gráfico de variância explicada, retendo todos os componentes antes deste ponto (Dmitrienko et al., 2007);

2.6 Aprendizagem Supervisionada

Ao contrário do que foi explorado até o momento, o contexto do aprendizado supervisionado leva em consideração um vetor de saída, sendo este comumente referido como variável dependente/resposta (James et al., 2013). O intuito deste modelo de aprendizado é desenvolver uma função que melhor mapeie o conjunto de dados através da relação entre variável resposta e suas preditoras. Entre os principais resultados deste grupo de métodos está a possibilidade de predição; uma vez que conhecemos a associação entre os dados, podemos utilizá-la para predizer novas observações.

Entre os diversos métodos disponíveis, salientamos o uso de *Multiple Linear Regression* no contexto esportivo, o qual se faz presente nos trabalhos (Berri et al., 2011), (Coates et al., 2010) e (Greene, 2015). Apesar de estarmos tratando de um método simples, este é amplamente utilizado devido a sua capacidade de trabalhar com dados de alta dimensão possuindo baixo custo computacional (Krkač et al., 2020). O método será descrito na próxima seção.

2.7 *Multiple Linear Regression*

Este método é uma generalização da *Simple Linear Regression*, visto que estamos em um contexto de duas ou mais variáveis preditoras. A regressão em questão é utilizada para predizer um vetor quantitativo (Y) através do ajuste de uma função para modelar a relação entre o mesmo com as variáveis explicativas (X); tal função assume que a associação entre X e Y é linear (James et al., 2013).

Dado um conjunto de dados com p variáveis de entrada, o modelo para análise de regressão múltipla é dado por

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon, \tag{2.3}$$

Onde $\beta_1, \ldots, \beta_p$ representam os coeficientes relacionados a cada variável $X_1, \ldots, X_p$ e β_0 o intercepto da equação. Como estes valores são desconhecidos, a estimação é feita utilizando a abordagem de mínimo quadrados ordinários, no qual escolhemos $\beta_0, \beta_1, \ldots, \beta_p$ de forma a minimizar a soma do quadrado dos resíduos (James et al., 2013)

$$RSS = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip} \right)^2. \tag{2.4}$$

onde:

RSS = soma do quadrado dos resíduos;

No momento em que tratamos de uma abordagem quantitativa, conseguimos extrair informações referentes a contribuição de cada variável em relação a resposta, facilitando a interpretação dos coeficientes do modelo. Além deste resultado, podemos obter outras conclusões através da regressão, entre eles:

- Quais variáveis são úteis para explicar a variável dependente;
- Quão bem o modelo se ajusta aos dados;
- Possibilidade de prever novos valores;

Embora este método seja amplamente utilizado para predição, estamos interessados em avaliá-lo em comparação com métodos mais robustos e sofisticados. Dado nossos objetivos de avaliação, na próxima seção introduzimos as redes neurais artificiais.

2.8 Redes Neurais Artificiais

Até o momento buscamos explorar técnicas clássicas dentro do mundo da análise esportiva, portanto neste capítulo iremos nos aprofundar na abordagem de um método recente da área: as Redes Neurais Artificiais (Bunker e Thabtah, 2019). As Redes Neurais Artificiais, ou simplesmente Redes Neurais, são, como o próprio nome sugere, uma arquitetura computacional formada por um conjunto de algoritmos cuja estrutura de camadas se assemelha com o cérebro humano, assim como seu funcionamento busca replicar a atividade das sinapses entre neurônios (Haykin, 2001).

Podemos visualizar a estruturação de uma rede neural em 3 camadas principais, sendo elas:

- *Input Layer*: entrada de dados para aprendizado;
- *Hidden Layer*: camada(s) onde os cálculos são realizados e o modelo aprende as relações presentes nos dados;
- *Output Layer*: saída do modelo, variando número de nodos de acordo com a variável resposta;

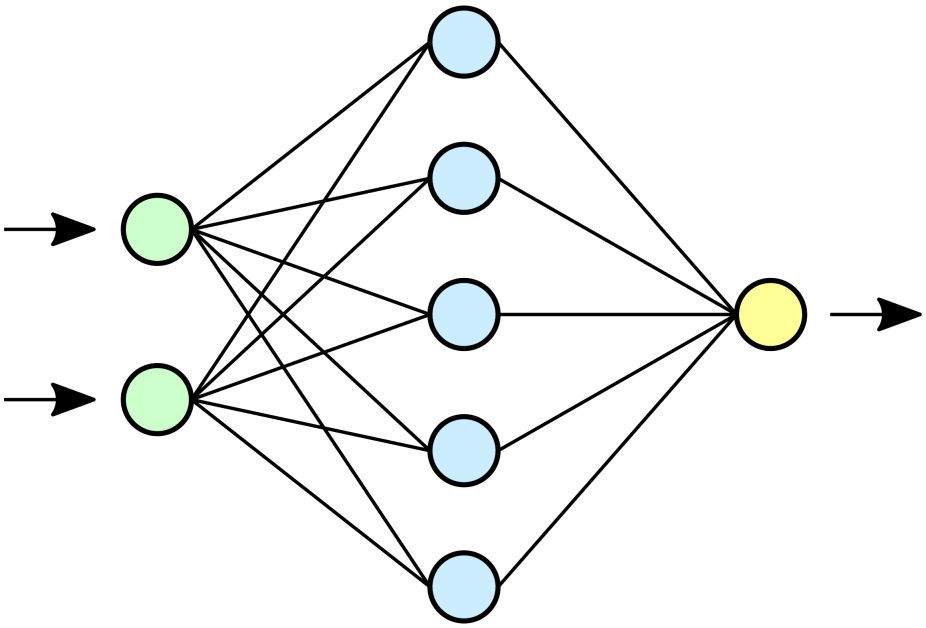


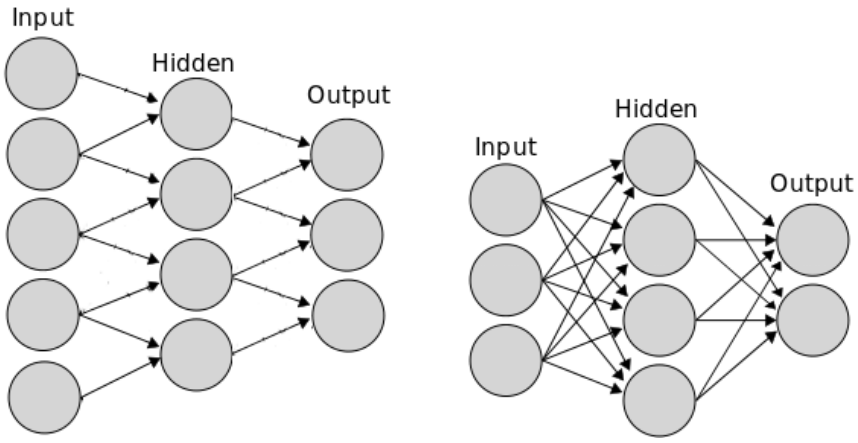
Figura 2.1: Estrutura de uma Rede Neural com 2 neurônios na camada de entrada, 5 neurônios na camada oculta e 1 neurônio na camada de saída

Dentro das redes neurais temos que um neurônio é o componente formado através da composição de peso, *bias* e *input*, onde uma função de ativação é aplicada, sendo este o valor que o neurônio irá repassar para a próxima camada até chegarmos na saída do modelo. Para que a aprendizagem por parte do neurônio aconteça, este processo é repetido inúmeras vezes em conjunto com outras funções, no intuito de buscarmos os pesos que melhor se ajustam as variáveis e explicam o banco de dados em questão (Haykin, 2001), assim como reduzir o erro de saída do modelo. Como existem diversas arquiteturas de redes, na próxima seção iremos nos aprofundar naquela utilizada para este trabalho.

2.8.1 *Multilayer Perceptrons*

A arquitetura selecionada para o estudo em questão é uma das mais clássicas e utilizadas dentre as diversas opções de redes neurais, sendo o modelo que iremos tratar na modalidade de *feed forward*, ou alimentadas adiante. Neste modelo as diferentes camadas de neurônios se postulam de maneira sequencial, onde a informação viaja em apenas uma direção, da entrada para a saída dos dados.

Dentro deste grupo, encontramos a classe dos *Multilayer Perceptrons*, objeto de interesse para este trabalho. Embora o significado de MLP se confunda diretamente com qualquer arquitetura *Feed Forward*, o mesmo tem suas particularidades. Em (Haykin, 2001), o MLP é definido como uma rede neural com uma ou mais camadas ocultas, sendo a generalização do *Single Layer Perceptron*, onde cada camada é totalmente conectada com a camada subsequente. Abaixo podemos ver a diferença explícita:



(a) Rede *Feed Forward* padrão (b) *Multilayer Perceptron*

Figura 2.2: Diferenças entre as arquiteturas *feed forward*

O propósito da rede MLP é utilizar o algoritmo *backpropagation* para correção e ajuste de pesos utilizados no processo de aprendizagem da rede neural, visando a redução do erro preditivo. Como o próprio nome sugere, este algoritmo repetidamente ajusta os pesos das conexões entre os neurônios para minimizar a diferença entre a saída do modelo e a saída verdadeira, ou seja, o algoritmo busca minimizar a função custo através do ajuste do *bias* e pesos, sendo este nível de ajuste baseado pelos gradiente da função custo, visto que o gradiente mostra qual direção determinado parâmetro deve ser ajustado para a minimização da função. A utilização do termo

propagação reversa ocorre devido a ordem qu este ajuste ocorre, sendo partir da última camada, procedendo até a camada inicial.

Com os diferentes métodos de estudos já mapeados, agora iremos analisar as métricas de desempenho utilizadas para avaliar nossos modelos de aprendizado de máquina.

2.9 Avaliação dos Modelos

Para avaliar a acertividade da predição dos nossos modelos ajustados no trabalho em questão, utilizaremos três métricas distintas que buscam mensurar o erro nestas predições. Dado as observações do banco definidas como y_i e as predições do modelo $\hat{y}_i$, temos:

- MAE: média do valor absoluto da diferença entre valores observados e preditos;

$$\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}| \tag{2.5}$$

- MSE: média do quadrado da diferença entre valores observados e preditos;

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{2.6}$$

- RSME: raiz do MSE;

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{2.7}$$

No intuito de obtermos o melhor modelo de acordo com seu poder preditivo sem entrarmos em um cenário de *overfitting*, utilizaremos ao longo deste trabalho o método de avaliação *k-fold cross-validation*. Este método divide o conjunto de treino em k partes iguais (ou o mais próximo possível) e, em cada iteração, separa uma amostra para teste enquanto as outras k-1 partes são utilizadas para treinar o modelo, obtendo assim as métricas de desempenho que se baseiam nos erros de predição citadas acima. Por fim, é calculada a média simples das métricas obtidas em cada um dos *folds*.

Com as diferentes métricas de desempenho já mapeadas, agora iremos analisar nosso banco de dados utilizado para o estudo das hipóteses deste trabalho em questão, de forma a entendermos quais variáveis estaremos tratando na construção dos modelos de aprendizado.

2.10 Banco de Dados

O horizonte de dados para este estudo foram os anos de 2010 à 2017, buscando informações de todos os jogadores *draftados* no período em questão, possuindo um $N = 480$. Para possibilitar a análise dos dados foram criados 3 bancos de dados distintos, sendo estes construídos de forma a se conectarem entre si através do nome do atleta. O primeiro banco de dados se refere as informações do *draft* (*rank*, *round*,

equipe, *college*); já as outras bases se dividem entre dados dos atletas no *college* e na NBA. Todos as informações utilizadas neste trabalho foram coletadas na base de dados pública disponível em Basketball-Reference (2022) através de um *scraper*[2] criado na linguagem Python pelo autor. A seguir podemos ver as tabelas nas quais as informações foram coletadas:

		Round 1	
Pk	Tm	Player	College
1	CLE	Andrew Wiggins	Kansas
2	MIL	Jabari Parker	Duke
3	PHI	Joel Embiid	Kansas
4	ORL	Aaron Gordon	Arizona

Figura 2.3: Tabela com dados referentes ao *draft*
Fonte: *www.basketball-reference.com*

Season	School	Conf	G	GS	MP	FG	FGA	FG%	2P	2PA	2P%	3P	3PA	3P%	FT	FTA	FT%	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS	SOS
2013-14	Arizona	Pac-12	38	38	31.2	5.0	10.1	.495	4.6	8.9	.513	0.4	1.2	.356	2.0	4.7	.422	2.7	5.3	8.0	2.0	0.9	1.0	1.4	2.4	12.4	9.04
Career	**Arizona**		**38**	**38**	**31.2**	**5.0**	**10.1**	**.495**	**4.6**	**8.9**	**.513**	**0.4**	**1.2**	**.356**	**2.0**	**4.7**	**.422**	**2.7**	**5.3**	**8.0**	**2.0**	**0.9**	**1.0**	**1.4**	**2.4**	**12.4**	**9.04**

Figura 2.4: Tabela com dados referentes ao *college*
Fonte: *www.basketball-reference.com*

[2]Método para obtenção de dados de páginas web

Em relação as variáveis observadas nas figuras 2.4 e 2.3, temos:

- Pk: número da escolhas;
- Tm: time que selecionou o jogador;
- *Player*: nome do jogador;
- *College/School*: nome da universidade que o jogador frequentou;
- *Season*: temporada em que os dados foram obtidos;
- *Conf*: conferência da universidade;

As demais variáveis serão trabalhadas com mais detalhes adiante neste trabalho.

Em relação a nossa variável resposta VORP, salientamos que a mesma é medida de forma anual, ou seja, ao longo das temporadas. Portanto para cada ano de um jogador na NBA ele terá um VORP atribuído.

Tabela 2.2: VORP de Kevin Durant ao longo de sua carreira

Season	VORP
2007-08	1.3
2008-09	3.8
2009-10	7.5
2010-11	5.3
2011-12	5.8
2012-13	8.9
2013-14	9.6
2014-15	2.8
2015-16	7.8
2016-17	5.7
2017-18	5.5
2018-19	5.1
2020-21	2.7
2021-22	4.8

Através da Tabela (2.2) podemos ver como esta métrica é coletada ao longo das temporadas em que o jogador esteve em quadra, desta forma o número de observações para cada atleta varia de acordo com sua longevidade na liga profissional. Para nossos modelos de aprendizagem de máquina iremos considerar apenas algumais faixas relevantes do VORP, as quais serão descritas, exploradas e analisadas no próximo capítulo.

3 Resultados

3.1 Introdução

Neste capítulo iremos abordar os diferentes métodos utilizados no desenvolvimento deste trabalho. O processo de obtenção dos dados ocorreu utilizando a linguagem Python (Van Rossum e Drake Jr, 1995), através da IDE *VS Code*. As transformações e subsequente análise de dados foram realizadas através do IDE *RStudio* e a linguagem R (R Core Team, 2013).

3.2 Análise Descritiva

Inicialmente foi realizado uma análise descritiva do banco de dados apresentado na seção 2.10 para entendermos as características dos atletas presentes no *draft*, assim como explorar nossa variável resposta quando avaliada em conjunto de outras variáveis.

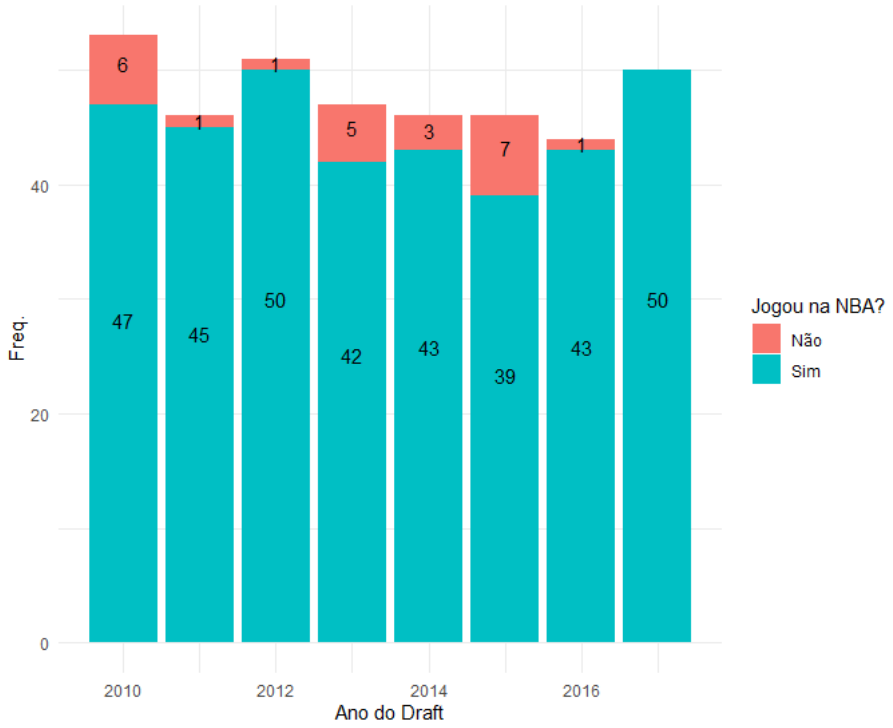


Figura 3.1: Frequência (Freq.) de atletas *draftados* com passagem pelo *college*

Dado que o *draft* é composto de 60 escolhas, podemos observar na Figura (3.1) que cerca de 75% dos atletas selecionados no processo de *draft* são advindos do *college*, sendo esta a principal fonte de novos talentos na liga. Temos um total de 480 atletas selecionados onde apenas 90 não são advindo do *college*. Além disso, podemos perceber que apesar do *draft* ser uma coroação do desempenho pré-NBA dos novatos, uma pequena parcela destes jogadores nunca põe o pé em quadra entre os profissionais. Dado que o *draft* é dividido em 2 *rounds* compostos de 30 escolhas cada, se torna interessante investigarmos em qual etapa do *draft* se encontram estes jogadores que não tiveram oportunidade na NBA.

Tabela 3.1: Número de *draftados* por *round* que jogaram (ou não) na NBA

	Jogou na NBA	
Draft Round	Sim	Não
Primeiro	204	0
Segundo	155	24

Através da Tabela (3.1) notamos que todos os atletas *draftados* entre 2010 e 2017 que não possuíram uma oportunidade de pisar nas quadras como atleta profissional da NBA foram escolhidos no segundo *round* do *draft*.

Tabela 3.2: VORP médio por *round*

Round	VORP médio
1	0.572
2	0.043

Corroborando com as afirmações acima, ao analisarmos a variável do estudo (VORP) por *rounds* aravés da Tabela (3.2), notamos que esta segue os mesmos padrões, ou seja, atletas selecionados entre as 30 primeiras escolhas possuem maiores oportunidades para se desenvolver e obter sucesso na liga profissional se comparados com atletas de segundo *round*.

Ao voltarmos nossa atenção para a variável resposta, gostaríamos de avaliar como características físicas influenciam no desempenho profissional, buscando entender se existe uma relação positiva/negativa com o sucesso de um atleta.

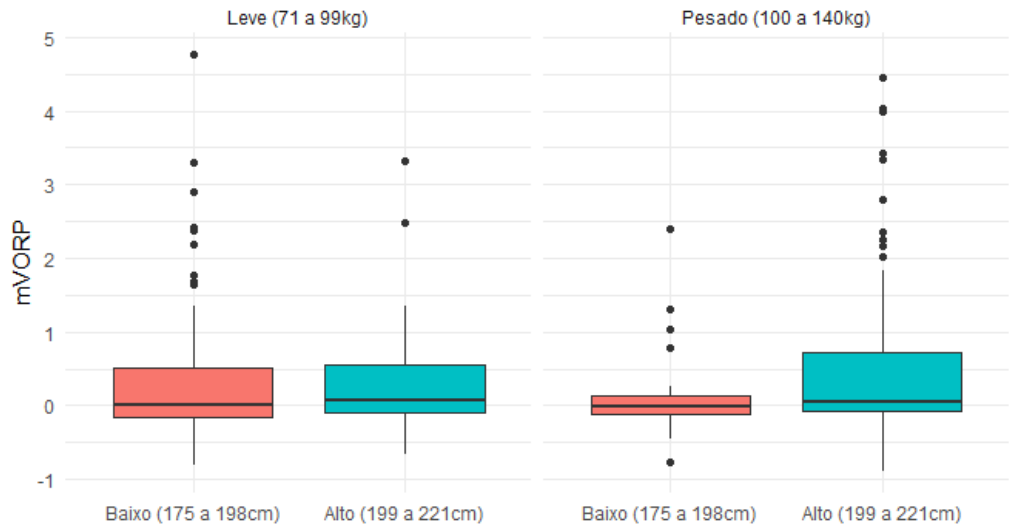


Figura 3.2: Distribuição do VORP médio (mVORP) por Altura X Peso

A partir da Figura (3.2) não conseguimos realizar inferências precisas em relação a existência de uma interação entre altura/peso e sucesso, portanto iremos avaliar como estas variáveis se comportam nos modelos de *machine learning*, mas, aparentemente, o VORP médio de um atleta não está diretamente ligado apenas as suas respectivas altura e peso.

Para fins de analisarmos a distribuição referente a variável de interesse deste trabalho, utilizamos o histograma apresentado na Figura (3.3).

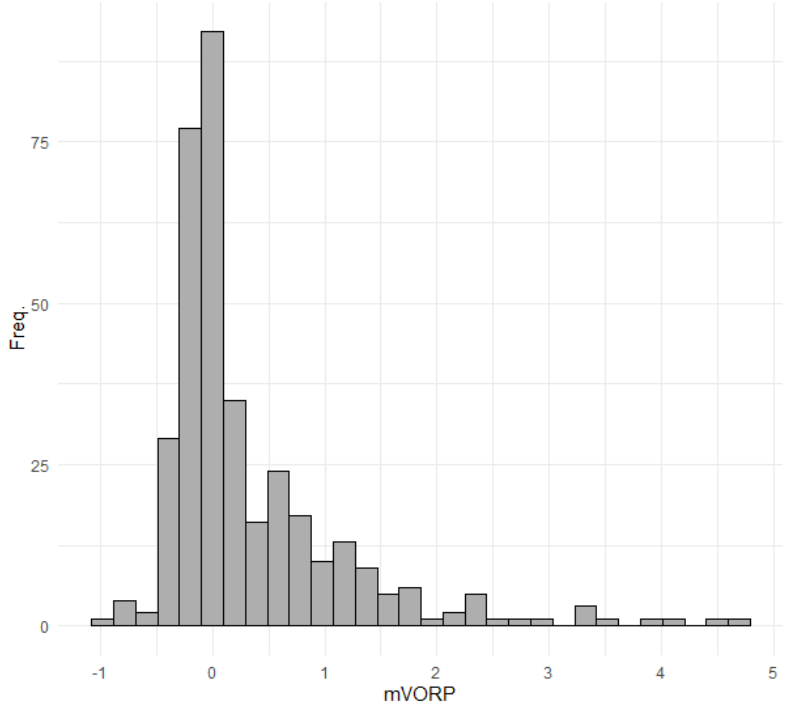


Figura 3.3: Histograma do VORP Médio

Podemos perceber através da Figura (3.3) que possuímos uma distribuição assi-

métrica a esquerda, onde grande parte dos jogadores se encontram com um VORP próximo de 0, ou seja, a grande maioria dos jogadores selecionados nos *drafts* de 2010 a 2017 se encontram no nível de *replacement player*. Tal resultado já se mostra esperado, visto que em uma liga profissional de qualquer esporte haverão mais jogadores medianos do que estrelas. Nossa análise se complementa com a Tabela (3.3), a qual nos mostra a dificuldade de um atleta na NBA se manter em alto nível ao longo de toda extensão de sua carreira, dado que o VORP máximo se encontra no valor +5.

Tabela 3.3: Estatísticas descritivas do VORP Médio

Estatísticas	VORP Médio
Mediana	0.0125
Desvio Padrão	0.8309
Mínimo	-0.9
Média	0.3435
Máximo	4.7667

As estatísticas descritivas apresentadas até agora e resumidas através da Tabela (3.3) se referem a todos os 358 jogadores com passagem pelo basquetebol colegial que jogaram pelo mesnos uma partida profissional na NBA. Em relação ao VORP, ainda não estamos considerando uma faixa de temporadas, ou seja, a média apresentada pela Tabela (3.3) se refere a todos os anos da carreira de cada jogador, portanto teremos atletas com mais e outros com menos experiência neste cálculo. A partir da próxima seção buscaremos avaliar a relação entre as covariáveis dos atletas para, após compreensão de todo o banco de dados, partirmos para as técnicas de aprendizagem de máquina.

3.3 Análise de Componentes Principais

Seguindo na linha da análise exploratória do nosso banco de dados, utilizaremos uma técnica de redução da dimensionalidade para entender o modo em que nossas variáveis estão relacionadas. Para este processo foram selecionadas 11 variáveis de interesse como mostra a Tabela (3.4), visto que estas são amplamente disponibilizadas no *box score*.

Tabela 3.4: Variáveis selecionadas para o PCA

Abreviação	Descrição
FG%	*Field Goal Percentage*
2PT%	*Two-Point Percentage*
3PT%	*Three-Point Percentage*
FT%	*Free-Throw Percentage*
RPG	*Rebounds Per Game*
APG	*Assists Per Game*
SPG	*Steals Per Game*
BPG	*Blocks Per Game*
TPG	*Turnovers Per Game*
PPG	*Points Per Game*
USG%	*Usage Percentage*

Primeiramente calculamos a correlação entre as covariáveis do banco, assim conseguimos analisá-las conjuntamente através do correlograma presente na Figura (3.4).

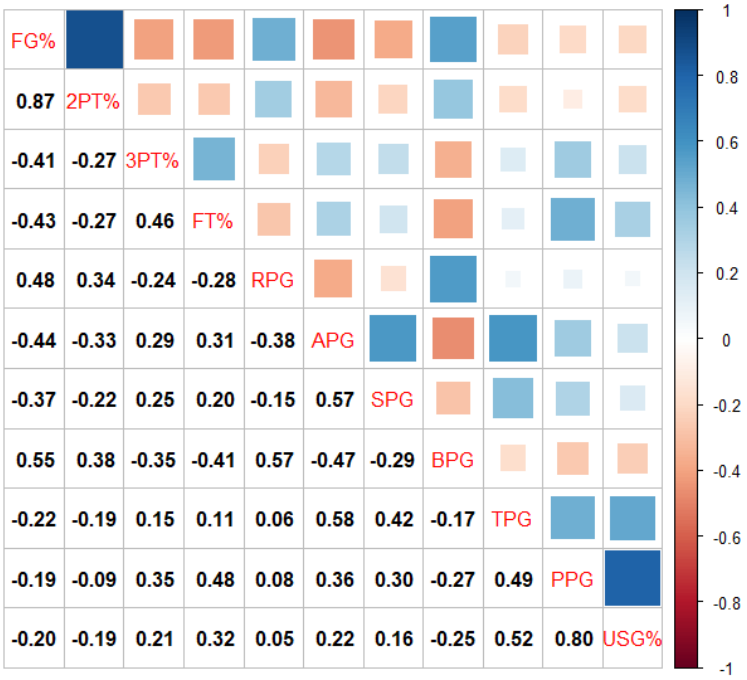
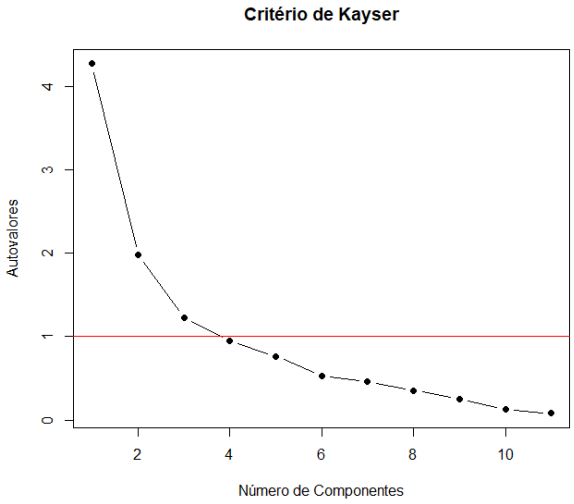


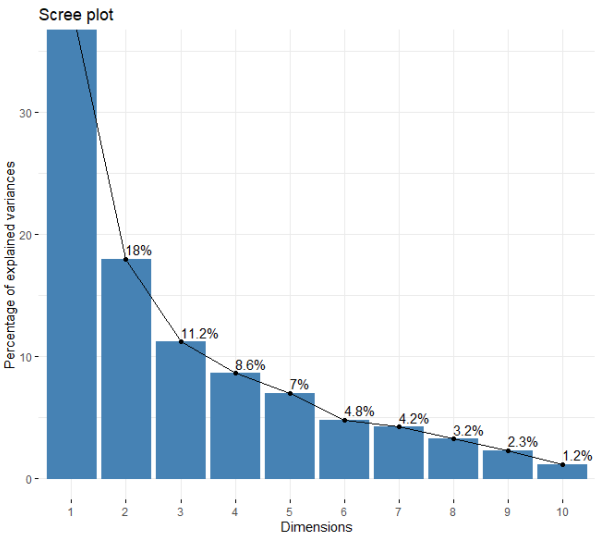
Figura 3.4: Correlograma das Variáveis

Percebemos, ao observar a Figura (3.4), no primeiro momento, as variáveis se relacionam diretamente com a função dos jogadores em quadra. Exemplificando nosso relato acima, percebemos correlações altas entre RPG e BPG, variáveis relacionadas a jogadores de defesa que permanecem na região do garrafão, jogando próximo a cesta; corroborando com as informações anteriores, temos as variáveis FG% e 2PT% positivamente correlacionadas com jogadores que se encontram perto da cesta, visto que estes possuem arremessos mais fáceis e com isso maior percentual de conversão de arremessos em pontos.

Dado que possuímos correlações significativas em nossas variáveis, cabe introduzirmos neste trabalho o método de PCA. Primeiramente devemos nos ater a escolha do número de componentes, para isso levamos em conta duas técnicas amplamente utilizadas para este método, sendo elas o *scree plot* e o critério de Kayser, o qual considera como metodologia a quantidade de autovalores que possuem valor acima de 1.



(a) *Critério de Kayser*



(b) *Scree Plot*

Figura 3.5: Critérios utilizados na escolha dos componentes

Como podemos analisar através da Figura (3.5), ambos os critérios acabam por se complementar. Dado que possuímos aproximadamente 5 autovalores bem próximos do valor 1 e que o ponto de suavização do *scree plot* ocorre na quinta dimensão, chegamos na conclusão de determinar que o número de componentes a serem escolhidos é 5.

Tabela 3.5: Variância explicada pelos componentes

Componente	% de Variância	% de Variância Acumulada
1	38.87	38.87
2	17.99	56.87
3	11.18	68.04
4	8.61	76.65
5	6.97	83.62
6	4.78	88.40
7	4.21	92.61
8	3.22	95.83
9	2.29	98.12
10	1.16	99.29
11	0.71	100.00

Através da Tabela (3.5) percebemos que a variância explicada ao retermos o número de componentes determinados está em torno de 84%, ou seja, ao reduzirmos em mais da metade a dimensionalidade do banco ainda explicamos quase toda sua variância.

Com a quantidade de componentes definida, podemos analisar a carga de cada um a fim de identificar quais variáveis que são correlacionadas com os componentes selecionados. Por mais que não nos encontramos em um contexto de análise fatorial, o formato em que as variáveis se organizaram nos permite, ao menos de forma empírica, caracterizarmos cada um dos componentes de acordo com a relação presente entre suas principais variáveis.

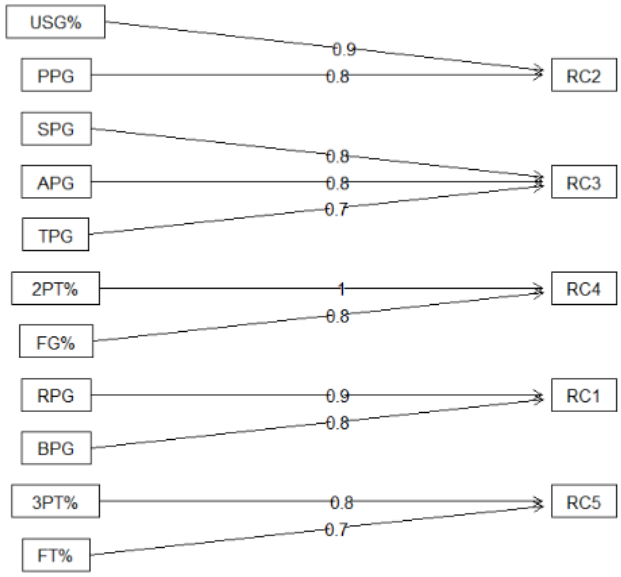


Figura 3.6: Relação entre Variáveis e Componentes

Baseado na Figura (3.6), temos os seguintes conceitos:

- Componente 1: protetores de garrafão;
- Componente 2: alto volume de jogo;
- Componente 3: controle do jogo;
- Componente 4: seleção de arremessos;
- Componente 5: arremessadores;

Utilizando estes componentes da Figura (3.6) e as variáveis selecionadas no início do capítulo, partimos para a seção de ajuste de modelos, onde estamos interessados na melhor forma de estimar nossa variável de interesse.

3.4 Ajuste de Modelos

Nas seções anteriores o foco foi centralizado na análise exploratória do banco, buscando descobrir informações e padrões dentro do mesmo. A partir deste capítulo estamos interessados na resposta de um dos objetivos principais deste trabalho em questão: qual o melhor modelo para prever o VORP de um atleta. Para prosseguirmos com nossa análise salientamos a utilização de duas variáveis resposta distintas, mas que se referem ao mesmo resultado.

Os modelos que serão tratados nos tópicos subsequentes são:

- Regressão Linear Múltipla (variáveis originais);
- Regressão Linear Múltipla (*scores* PCA);
- Regressão Neurais (variáveis originais);

- Redes Neurais (*scores* PCA);

Em relação aos modelos de redes neurais, utilizamos diferentes combinações de camadas ocultas, sendo elas:

- Uma camada (2 a 30 neurônios);
- Duas camadas (50 a 80 neurônios cada);
- Três camadas (4 a 20 neurônios cada);

3.4.1 VORP no Primeiro Ano

A primeira variável resposta a ser analisada se refere ao VORP no primeiro ano de NBA de um jogador advindo do *college*. Através desta variável buscamos avaliar quais jogadores saem do basquetebol colegial prontos para o profissional, ou seja, estamos interessados na relação das variáveis preditoras com o sucesso imediato ao ingressarem na liga.

Para treinarmos nossos diferentes modelos selecionamos apenas aqueles atletas que se encaixavam na categoria, tendo jogado basquetebol colegial e pelo menos um ano na NBA, dado que estamos interessados apenas neste primeiro contato profissional, sem considerar a longevidade de carreira na liga. Aplicando este filtro nos restaram 358 jogadores entre os anos de 2010 e 2017.

O próximo passo foi a separação do banco em conjuntos de treino e teste, sendo o treino constituído por 90% do banco e o restante para teste. Para mantermos o equilíbrio entre os conjuntos, realizamos uma amostragem estratificada por posição de cada atleta, desta forma mantivemos a mesma proporção para ambos. Tal iniciativa foi tomada devido ao conhecimento prévio das diferentes características que compõem a posição em quadra de um atleta, portanto ao estratificarmos o banco estamos levando em conta esta diversidade.

Após realizarmos a validação cruzada obtivemos que o melhor resultado para redes neurais utilizando PCA foi a combinação de 3 camadas ocultas, sendo a primeira com 4 neurônios, a segunda também com 4 neurônios e a terceira com 18 neurônios, ou seja, (4, 4, 18), enquanto para as redes com variáveis originais foram 3 camadas ocultas (4,9,11), onde ambas foram treinadas utilizando 100.000 épocas ou até atingir a convergência do algoritmo, aplicando função de ativação logística em todas as camadas ocultas e linear na camada de saída. Na Tabela (3.6) podemos ver a comparação das métricas entre os melhores modelos:

Tabela 3.6: Métricas de Desempenho

Modelo	RMSE	MSE	MAE	R2
Regressão	0.8931	0.7976	0.6073	0.1281
Regressão (PCA)	0.8866	0.7861	0.6003	0.1308
Redes Neurais	0.8896	0.7913	0.5895	0.1137
Redes Neurais (PCA)	0.8830	0.7797	0.5850	0.1374

Notamos através de Tabela (3.6) que os modelos com menores erros foram aqueles que utilizaram os *scores* do PCA, tanto no caso das redes neurais como na regressão. Podemos inferir através dos valores acima que os modelos em sua totalidade

possuem dificuldade de explicar a variável resposta, visto que o maior coeficiente de determinação se encontra abaixo de 0.15. Vale ressaltar que os valores das variáveis estão escalonados, portanto interpretamos as métricas de acordo com os desvios.

Como vimos, os melhores modelos foram aqueles que levaram em conta o PCA, desta forma podemos analisar os coeficientes da regressão para entendermos de qual forma eles afetam a nossa variável preditora.

Tabela 3.7: Coeficientes da Regressão

Coeficiente	Estimativa	P-valor	Signif.
Intercepto	-0.0053	0.9199	
RC1	0.1698	0.0016	**
RC2	-0.0677	0.1993	
RC3	0.1682	0.0017	**
RC4	0.2121	0.0001	***
RC5	0.0062	0.9050	

Embora o coeficiente de determinação tenha sido baixo, notamos através da Tabela (3.7) que possuímos significância em 3 dos 5 componentes utilizados no modelo, sendo todos eles responsáveis por aumento no VORP. Ainda analisando a Tabela (3.6), olharemos para o gráfico de observações e valores preditos do nosso melhor modelo, as redes neurais utilizando o PCA.

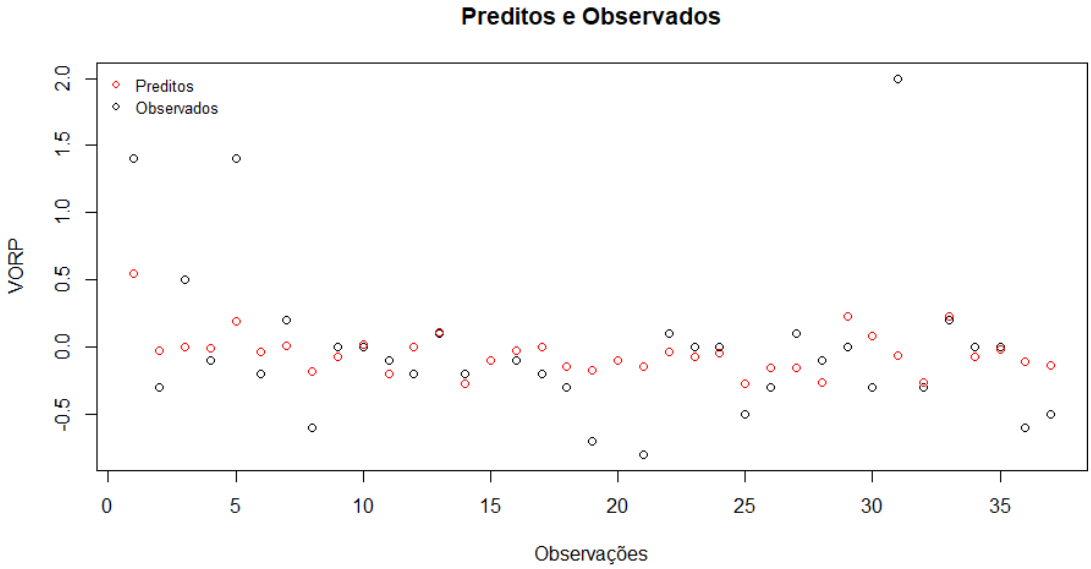


Figura 3.7: Observações e valores preditos

Podemos perceber através da Figura (3.7) que possuímos muitas observações em torno do valor 0, algo esperado dado a configuração de nossa variável resposta, sendo estes valores corretamente preditos pelo nosso modelo. Porém, salientamos que em relação a valores discrepantes o modelo de redes neurais tem dificuldade de estimá-los, dado que ele as subestima; embora estes sejam os números que mais nos interessam, manteremos este modelo para futura comparação com o *draft* da NBA.

3.4.1.1 VORP no Primeiro Ano como variável categórica

Utilizando outra abordagem para estimarmos nossa variável resposta, consideraremos o VORP em 3 faixas categóricas, sendo elas divididas em Ruim, Regular e Bom.

- Ruim: VORP < 0;
- Regular: VORP < 2;
- Bom : VORP > 2;

Após a categorização do nosso banco de dados devemos conferir como ficaram distribuídas as classes entre os nossos atletas.

Tabela 3.8: Frequência de classes do VORP no banco de dados

Classe do VORP	Contagem
Bom	6
Regular	89
Ruim	263

Podemos perceber através da Tabela (3.8) que estamos lidando com classes desbalanceadas, portando nossa solução para manter o equilíbrio entre conjuntos de treino e teste será realizar a estratificação do banco pela variável VORP categórico.

Para treinarmos nosso banco utilizamos apenas os modelos referentes a redes neurais dado que estes comportam respostas multiclasse. Após iterarmos em cada combinação de camadas especificadas no início do capítulo, os modelos com melhor desempenho foram de 3 camadas ocultas (5, 9, 4) para o PCA e 3 camadas ocultas (5, 5, 10) para as variáveis originais, onde ambas foram treinadas utilizando 100.000 épocas ou até atingir a convergência do algoritmo, aplicando função de ativação logística em todas as camadas ocultas e na camada de saída. Através da Tabela (3.9) percebemos que os modelos apresentaram performance semelhante ao compararmos a acurácia de ambos, acertando, em média, aproximadamente 75% das categorias dos jogadores.

Tabela 3.9: Métricas de Desempenho - Resposta Categórica

Modelo	Acurácia
Redes Neurais	0.7360
Redes Neurais (PCA)	0.7578

Como sabemos, mesmo estratificando o banco, que estamos lidando com classes desbalanceadas, devemos olhar para a matriz de confusão do nosso melhor modelo ao aplicá-lo no conjunto de teste.

Tabela 3.10: Matriz de Confusão

	Classe Predita		
Classe Original	Bom	Regular	Ruim
Bom	0	0	1
Regular	0	0	1
Ruim	1	9	25

Percebemos a partir da Tabela (3.10) que possuimos uma acurácia mediana, se aproximando de 70%, porém como muitos valores se concentram em apenas uma classe, nosso modelo aparenta estar se limitando a mesma resposta e não realmente entendendo a relação entre as variáveis.

3.4.2 Mediana do VORP

Nossa segunda variável resposta a ser analisada se refere a mediana do VORP nos 4 primeiros anos de NBA de um jogador advindo do *college*. O motivo de selecionarmos apenas os 4 primeiros anos no basquetebol profissional se deve ao fato desta ser a duração padrão de um contrato de novato na NBA. Com a passagem deste período de teste as equipes decidem se devem oferecer um novo contrato ao jogador ou deixá-lo livre no mercado. Através desta variável buscamos avaliar quais atletas mantém um bom desempenho neste período inicial, ou seja, estamos interessados na relação das variáveis preditoras com o sucesso no período de contrato de novato.

Assim como para a variável resposta anterior selecionamos apenas aqueles atletas que se encaixavam na categoria, tendo jogado basquetebol colegial e pelo menos 4 anos na NBA. Aplicando este filtro nos restaram 282 jogadores entre os anos de 2010 e 2017. Seguindo os mesmos passos da subseção 3.4.1, separamos nosso banco entre treino e teste, onde o primeiro contém 90% das observações e o restante ficou alocado para testarmos os modelos. Para mantermos o equilíbrio entre os conjuntos, mantivemos uma amostragem estratificada por posição visando levar em conta as disparidades que compõem as diferentes posições em quadra de um atleta.

As métricas de avaliação do modelo são obtidas utilizando o método *k-fold cross-validation*, onde definimos $k = 10$. Após realizarmos a validação cruzada obtivemos que o melhor resultado para redes neurais utilizando PCA foi a combinação de 3 camadas ocuktas (4,19,14) enquanto para as redes com variáveis originais foram 3 camadas ocultas (4,14,12), onde ambas foram treinadas utilizando 100.000 épocas ou até atingir a convergência do algoritmo, aplicando função de ativação logística em todas as camadas ocultas e linear na camada de saída. Na Tabela (3.11) podemos ver a comparação das métricas entre os melhores modelos:

Tabela 3.11: Métricas de Desempenho

Modelo	RMSE	MSE	MAE	R2
Regressão	0.9777	0.9559	0.6953	0.0965
Regressão (PCA)	0.9727	0.9461	0.6897	0.1044
Redes Neurais	0.9487	0.9000	0.6747	0.1314
Redes Neurais (PCA)	0.9632	0.9278	0.6876	0.1146

Notamos através de Tabela (3.11) que o melhor modelo de rede neural foi aquele que utilizou as variáveis originais com uma leve margem em relação às métricas do PCA, já no caso da regressão quem se sobressaiu foi o PCA. Podemos inferir através dos valores acima que os modelos em sua totalidade possuem dificuldade de explicar a variável resposta, visto que o maior coeficiente de determinação se encontra abaixo de 0.15, ainda piores se compararmos com a variável resposta do capítulo anterior. Vale ressaltar que os valores das variáveis estão escalonados, portanto interpretamos as métricas de acordo com os desvios. Para fins de comparação, olharemos para

os coeficientes de regressão do modelo que levou em conta os *scores* do PCA assim como realizado para a variável VORP no primeiro ano de NBA.

Tabela 3.12: Coeficientes da Regressão

Coeficiente	Estimativa	P-valor	Signif.
Intercepto	0.0043	0.9199	
RC1	0.1985	0.0024	**
RC2	0.0738	0.2349	
RC3	0.1482	0.0237	*
RC4	0.2083	0.0009	***
RC5	-0.0027	0.9667	

Observamos através da Tabela (3.12) que os mesmos componentes são significantes (RC1, RC3 e RC4), sendo todos eles responsáveis pelo aumento na mediana do VORP de um atleta. Dado que nosso melhor modelo se refere às redes neurais utilizando as variáveis originais do banco, olharemos para o *plot* das observações do grupo de teste em conjunto com os valores preditos.

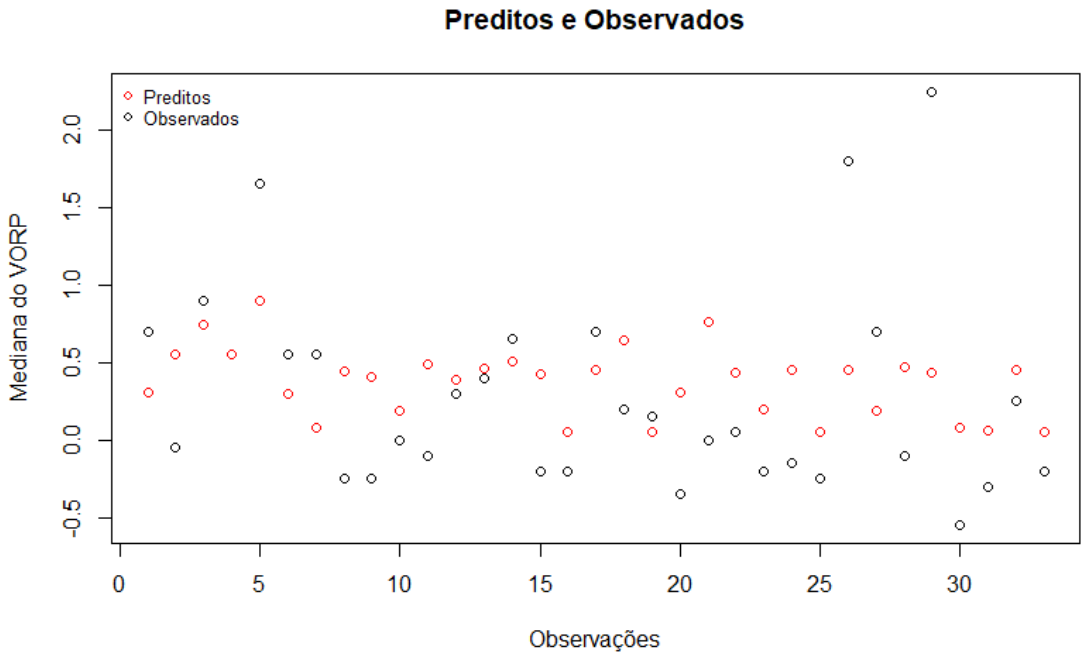


Figura 3.8: Observações e valores preditos

Podemos ver claramente através da Figura (3.8) que nosso modelo enfrenta muitas dificuldades em entender a relação entre variáveis resposta e preditoras, visto que a maioria das observações está em torno de 0, portanto o modelo acaba predizendo seus valores numa faixa que varia de -0.5 a 0.5, subestimando jogadores com o VORP alto. Embora estes sejam os números que mais nos interessam, manteremos este modelo para futura comparação com o draft da NBA.

3.4.2.1 Mediana do VORP como variável categórica

Utilizando outra abordagem para estimarmos nossa variável resposta, consideraremos o VORP em 3 faixas categóricas, sendo elas divididas em Ruim, Regular e Bom.

- Ruim: VORP < 0;
- Regular: VORP < 2;
- Bom : VORP > 2;

Após a categorização do nosso banco de dados devemos conferir como ficaram distribuídas as classes entre os nossos atletas.

Tabela 3.13: Frequência de classes do VORP no banco de dados

Classe do VORP	Contagem
Bom	15
Regular	140
Ruim	127

Podemos perceber através da Tabela (3.13) que existe um certo desbalanceamento entre as classes onde, dado que possuímos poucas observações dentro da caregoria 'Bom'. Portanto nossa solução para manter o equilíbrio entre conjuntos de treino e teste será realizar a estratificação do banco pela variável VORP categórico.

Para treinarmos nosso banco utilizamos apenas os modelos referentes a redes neurais dado que estes comportam respostas multiclasse. Após iterarmos em cada combinação de camadas especificadas no início do capítulo, os modelos com melhor desempenho foram de 3 camadas ocultas (4, 8, 5) para o PCA e 3 camadas ocultas (9, 4, 5) para as variáveis originais, onde ambas foram treinadas utilizando 100.000 épocas ou até atingir a convergência do algoritmo, aplicando função de ativação logística em todas as camadas ocultas e na camada de saída. Através da Tabela (3.9) percebemos que os modelos apresentaram performance semelhante ao compararmos a acurácia de ambos, acertando em média aproximadamente 55% das categorias dos jogadores, número bem abaixo se compararmos com a variável resposta referente ao primeiro ano do VORP.

Tabela 3.14: Métricas de Desempenho - Resposta Categórica

Modelo	Acurácia
Redes Neurais	0.5629
Redes Neurais (PCA)	0.5669

Olharemos para a matriz de confusão do nosso melhor modelo ao aplicá-lo no conjunto de teste para tentarmos entender como as previsões estão se portando.

Tabela 3.15: Matriz de Confusão

	Classe Predita		
Classe Original	Bom	Regular	Ruim
Bom	0	0	0
Regular	1	6	10
Ruim	0	8	3

Percebemos a partir da Tabela (3.15) que possuimos uma acurácia extremamente baixa, se aproximando de 30%, com isso notamos que nosso modelo aparenta estar se limitando a mesma resposta e não realmente entendendo a relação entre as variáveis, repetindo o mesmo caso que aconteceu com a variável VORP no primeiro ano, embora esta tivesse previsto mais classes 'Ruim' enquanto a mediana fixou na classe 'Regular', provavelmente por ser a de maior frequência.

Dado que avaliamos todos os modelos propostos neste trabalaho em questão, agora estamos interessados em estudarmos se nossos melhores modelos podem se sair melhor que o *draft* atual na NBA, sendo tal questionamento aprofundado na próxima seção.

3.5 Comparação entre o *draft* e os modelos

3.5.1 Introdução

Até o momento, utilizamos as métricas de desempenho para compararmos diferentes modelos entre si no intuito de selecionarmos o de melhor performance. Nesta etapa do trabalho iremos avaliar como nosso melhor modelo se comporta em relação ao próprio *draft* da NBA. Para isso ordenamos os jogadores do *draft* de 2017 de acordo com seu VORP e comparamos com o ordenamento do *draft* atual e de nossos modelos.

Tabela 3.16: Exemplo de comparação

Ordenação Modelo	Ordenação *draft*	Ordenação VORP	Jogador
5	15	1	Jogador 1
10	3	2	Jogador 2
3	7	3	Jogador 3

De acordo com o exemplo da Tabela (3.16), o erro calculado para a ordenação do modelo seria:

$$\begin{aligned} MSE &= \frac{((5-1)^2 + (10-2)^2 + (3-3)^2)}{3} = 26.7 \\ MAE &= \frac{(|(5-1)| + |(10-2)| + |(3-3)|)}{3} = 4 \end{aligned} \tag{3.1}$$

Enquanto para o *draft* atual seria:

$$\begin{aligned} MSE &= \frac{((15-1)^2 + (3-2)^2 + (7-3)^2)}{3} = 71 \\ MAE &= \frac{(|(15-1)| + |(3-2)| + |(7-3)|)}{3} = 6.3 \end{aligned} \tag{3.2}$$

De acordo com os valores presentes nas equações 3.1 e 3.2 iríamos concluir que nosso modelo se sai melhor que o *draft* atual da NBA. A metodologia até agora exemplificada será replicada nas próximas subseções com valores reais.

3.5.2 VORP no primeiro ano

Como demonstrado na subseção 3.5.1, para atingirmos os resultados esperados dividiremos o banco em treino e teste, sendo o teste composto apenas por jogadores do *draft* de 2017 enquanto o restante irá compor o treino. O intuito desta abordagem está em ordenar os jogadores *draftados* em 2017 pelo VORP no seu primeiro ano de NBA e comparar com as ordenações do modelo e do *draft* real.

Tabela 3.17: Comparação com o *draft*

Modelo	RMSE	MSE	MAE
Modelo de RN	13.51	182.56	9.96
draft	14.73	217.24	11.24

Através da Tabela (3.17) notamos que em ambas a métricas nosso modelo se sai melhor quando comparamo-os com o *draft* real da NBA, portanto por mais que tenhamos visto que ele enfrenta dificuldades em entender a relação entre covariáveis e variável resposta, isto não o impede de melhorar as predições do *draft*.

3.5.3 Mediana do VORP

Dado que possuíamos duas variáveis respostas distintas, o processo de comparação com o *draft* será realizada para ambas. Neste momento estamos interessados na mediana do VORP, portanto separamos o banco em treino e teste, sendo o teste composto apenas por jogadores do *draft* de 2017 enquanto o restante irá compor o treino. O intuito desta abordagem está em ordenar os jogadores *draftados* em 2017 pela mediana do seu VORP nos quatro primeiros anos de NBA e comparar com as ordenações do modelo e do *draft* real.

Tabela 3.18: Comparação com o *draft*

Modelo	RMSE	MSE	MAE
Modelo de RN	14.94	223.31	12.68
draft	11.99	143.85	9.21

Através da Tabela (3.17) notamos que em ambas a métricas nosso modelo se saem pior quando comparado com o *draft* real da NBA, portanto podemos perceber que a dificuldade que o modelo enfrenta em entender a relação entre covariáveis e variável resposta é refletida quando analisamos o ordenamento do *draft*, ou seja, a introdução do modelo de aprendizado de máquina não traz nenhum ganho para a escolha de jogadores no *draft*, sendo o método atual suficiente para obter melhores métricas quando avaliamos os atletas utilizando a mediana de seu VORP nos 4 anos de contrato de novato. Este resultado, porém, também pode ser explicado por outros fatores, como a limitação de tamanho do banco de dados. Mas sobre isso trataremos com mais profundidade no próximo capítulo.

4 Considerações Finais

O presente trabalho tinha como finalidade explorar e analisar o poder preditivo das estatísticas do basquetebol colegial para o sucesso na NBA, utilizando de técnicas de *machine learning*. Quanto a determinação de uma variável resposta, determinanos como balizador de sucesso o VORP de um atleta, métrica que busca sumarizar todas as contribuições em quadra através de um número, possibilitando, desta forma, compararmos o sucesso entre jogadores. Em relação a avaliação dos diferentes métodos: redes neurais artificiais e regressão linear múltipla, notamos melhores métricas de avaliação em relação as redes, embora o desempenho de ambos seja semelhante, ou seja, estamos observando resultados que nos mostram que provavelmente a utilização de redes neurais artificiais não nos oferece uma grande vantagem ao aplicá-la em um *dataset* relativamente pequeno. Tal questão se torna aparente quando analisamos os resultados do VORP como variável categórica pois, aparentemente, estamos tratando de uma relação relativamente complexa entre as variáveis e, dado a quantidade de dados coletados, os modelos não são capazes de explicá-la.

Quanto as variáveis que influenciam no aumento ou diminuição do VORP, percebemos ao utilizar a técnica de componentes principais que jogadores com características de proteger o garrafão, controlar o jogo ou selecionar arremessos de forma a ter alto aproveitamento tendem a possuir um VORP mais alto, tanto no seu primeiro ano na NBA quanto ao longo do seu contrato de novato, enquanto as características de alto volume de jogo assim como arremessadores não tem influência significatica na nossa métrica de sucesso. Nossos resultados corroboram com Berri et al. (2011) pois, embora um alto volume de jogo esteja relacionado com posições mais altas no *draft*, tal variável não se relaciona com a produção e desempenho a nível profissional. Em relação as variáveis que influenciam no aumento do VORP, podemos avaliar que por mais que elas não aumentem de forma estrondosa o VORP de um atleta, são características mais fáceis de serem mantidas na transição do *college* para o basquetebol profissional e, como o VORP se baseia em dados do *box score*, consequentemente estes atletas terão um desempenho mais alto.

Com relação a nossa variável resposta, podemos afirmar que as estatísticas do basquetebol colegial se mostram mais úteis para prever o VORP no primeiro ano de um atleta na liga, enquanto demonstra dificuldade para estimar esse valor em um período de tempo maior. Tanto na variável resposta quantitativa como categórica nossos modelos se adaptaram melhor ao tratarmos do primeiro ano de VORP, tal resultado se deve ao fato que um maior período de tempo influencia em questões de adaptação de um atleta, podendo sofrer com rotações de equipe, lesões ou até mesmo

o efeito contrário, jogadores com pior *ranking* se tornarem peças fundamentais em equipes vencedoras, sendo complicado conseguirmos estimar com precisão esta evolução/declínio do atleta. De qualquer forma, prever uma nova estrela na NBA foge do alcance dos modelos apresentados. Dentro do que foi proposto no trabalho, onde estávamos apenas interessados em estatísticas do *box score*, concluímos que apesar do nosso modelo não ser capaz de diferenciar um LeBron James ou Michael Jordan dos demais, o mesmo aparece como uma boa alternativa para estimar o potencial de um atleta ao entrar na liga profissional, dado que observamos que o mesmo superou o *draft* ao compararmos a ordenação do primeiro ano de VORP de um jogador.

Vale mencionar que este trabalho se mostra como um estudo pontual considerando uma faixa relativamente atual de jogadores selecionados no *draft*. Dado que a NBA possui mudanças em seu estilo de jogo assim como a adaptação e evolução dos atletas de acordo com estas mudanças, a realização deste estudo em outras épocas do *draft* poderia obter diferentes resultados dos obtidos neste trabalho em questão. Embora focamos no sucesso de jogadores do basquetebol profissional, este estudo, assim como os métodos do mesmo, poderiam ser estendidos para outros esportes que considerassem a universidade como principal meio de inserção de novos talentos para a liga profissional, como é o caso do futebol americano.

Referências Bibliográficas

(2013). Anthony bennett – basketball recruiting – player profiles – espn. ESPN, 2013. http://insider.espn.com/college-sports/basketball/recruiting/player/_/id/103629/anthony-bennett. Acesso em: 28 set. 2021.

(2021). 2021 nba draft prospect rankings. CBS Sports, 2021. https://www.nba.com/nba-draft-lottery-explainer. Acesso em: 28 set. 2021.

(2021). Nba draft lottery: Schedule, odds and how it works. NBA.com, 2021. https://www.nba.com/nba-draft-lottery-explainer. Acesso em: 28 set. 2021.

Alamar, B. (2021). Rockets, spurs lead the way in nba draft analytics. ESPN, 2021. https://www.espn.com/nba/story/_/id/23762871/rockets-spurs-celtics-most-analytical-draft-teams-nba. Acesso em: 28 set. 2021.

Basketball-Reference (2022). Basketball-Reference. https://www.basketball-reference.com/.

Berri, D. J., Brook, S. L., e Fenn, A. J. (2011). From college to the pros: Predicting the nba amateur player draft. *Journal of Productivity Analysis*, 35(1):25–35.

Bishop, C. M. (2006). Pattern recognition. *Machine learning*, 128(9).

Bunker, R. P. e Thabtah, F. (2019). A machine learning framework for sport result prediction. *Applied Computing and Informatics*, 15(1):27–33.

Coates, D., Oguntimein, B., et al. (2010). The length and success of nba careers: Does college production predict professional outcomes. *International Journal of Sport Finance*, 5(1):4–26.

Dmitrienko, A., Chuang-Stein, C., e D'Agostino, R. (2007). *Pharmaceutical Statistics Using SAS: A Practical Guide*. SAS Institute.

Floyd, F. J. e Widaman, K. F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological assessment*, 7(3):286.

Greene, A. C. (2015). The success of nba draft picks: Can college careers predict nba winners?

Haykin, S. (2001). *Redes neurais: princípios e prática*. Bookman Editora.

Hoffman, L. e Joseph, M. (2017). A multivariate statistical analysis of the nba. *Advanced Engineering Informatics*, 33:388–396.

James, G., Witten, D., Hastie, T., e Tibshirani, R. (2013). *An introduction to statistical learning*, volume 112. Springer.

Kannan, A., Kolovich, B., Lawrence, B., e Rafiqi, S. (2018). Predicting national basketball association success: A machine learning approach. *SMU Data Science Review*, 1(3):7.

Kaufman, S. B. (2014). What predicts nba success? Scientific American, 2014. https://blogs.scientificamerican.com/beautiful-minds/what-predicts-nba-success/. Acesso em: 28 set. 2021.

Krkač, M., Gazibara, S. B., Arbanas, Ž., Sečanj, M., e Arbanas, S. M. (2020). A comparative study of random forests and multiple linear regression in the prediction of landslide velocity. *Landslides*, 17(11):2515–2531.

Millington, B. e Millington, R. (2015). 'the datafication of everything': Toward a sociology of sport and big data. *Sociology of Sport Journal*, 32(2):140–160.

Mitchell, T. (1997). *Machine Learning*. McGraw-Hill International Editions. McGraw-Hill.

Myers, D. (2020). About box plus/minus (bpm). Basketball-Reference, 2020. https://www.basketball-reference.com/about/bpm2.html. Acesso em: 15 mar. 2022.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Soliven, E. (2021). Giannis antetokounmpo makes the case for the hall of fame in just one year. Basketball Network, 2021. https://www.basketballnetwork.net/giannis-antetokounmpo-makes-the-case-for-the-hall-of-fame-in-just-one-year/. Acesso em: 28 set. 2021.

Tichy, W. (2016). Changing the game: Dr. dave schrader on sports analytics. *Ubiquity*, 2016(May):1–10.

Van Rossum, G. e Drake Jr, F. L. (1995). *Python tutorial*. Centrum voor Wiskunde en Informatica Amsterdam, The Netherlands.