NOVA IMS Information Management School

MGI

Mestrado em Gestão de Informação
Master Program in Information Management

PREVISÃO DE RESULTADO DE JOGOS DA NBA COM ALGORITMOS DE MACHINE LEARNING

Lucas Mariani Lunelli

Trabalho de Projeto apresentado como requisito parcial para obtenção do grau de Mestre em Gestão da Informação

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

PREVISÃO DE RESULTADO DE JOGOS DA NBA COM ALGORITMOS DE MACHINE LEARNING

por

Lucas Mariani Lunelli

Trabalho de Projeto apresentado como requisito parcial para a obtenção do grau de Mestre em Gestão da Informação

Orientador: Mauro Castelli

Agosto 2019

AGRADECIMENTOS

À minha esposa, Isabela, pela parceria, pelo carinho e pelo amor, que tornaram a realização deste trabalho menos desgastante.

Aos meus pais, Aires e Marilice, pela educação e valores que me foram dados. Muito obrigado pelo apoio e incentivo de vocês durante toda minha vida.

Ao Professor Mauro Castelli, pelo apoio durante a realização deste trabalho e a todos professores da NOVA IMS pela qualidade do ensino.

RESUMO

A NBA é uma das maiores ligas esportivas do mundo e todos os anos movimenta mil milhões de dólares em patrocínios e apostas. Além do dinheiro envolvido, existe uma infinidade de métricas sobre jogadores e equipas, sendo uma grande oportunidade para apaixonadas por dados. Dessa forma, este projeto utilizará esses dados para desenvolver modelos de *machine learning* para prever resultados dos jogos da NBA e, a partir do resultado do modelo, construir um simples sistema apostas.

A parte experimental consiste em cinco etapas. O primeiro passo é o domínio do assunto. A seguir, é feita a coleta e análise exploratória dos dados. Na sequência, os dados são pré-processados e novas variáveis são criadas. A próxima etapa é modelagem. Com o apoio de algoritmos de *machine learning* como árvores de decisão, regressão logística, redes neuronais e *ensembles* o modelo é treinado. Após avaliação inicial, são definidas as variáveis mais importantes e que serão incorporadas ao modelo, visando sempre aumentar sua precisão. A partir do resultado do modelo é possível então criar lucrativas estratégias de apostas. Por fim, os resultados do modelo e das estratégias de apostas são avaliados através de diversas métricas e comparações com valores encontrados na literatura.

PALAVRAS-CHAVE

NBA; basquetebol; predição; aprendizado de máquina; análise de dados

ABSTRACT

NBA is one of the largest sports leagues in the world and every year it moves billions of dollars in sponsorship and betting. In addition to the money involved, many metrics about players and teams are generated, which is a great opportunity for data enthusiasts. Thus, this project will use this data to develop machine learning models to predict the results of NBA games and, from the model's results, build a simple betting system.

The experimental part consists of five steps. The first one is mastering the subject. Next, is the data collection and its exploratory analysis. After that, the data is preprocessed and new features are created. The next step is the modeling. The model is trained using machine learning algorithms such as decision trees, logistic regression, neural networks and ensembles. After first evaluation, the more important features are selected and incorporated into the model, always aiming to increase its accuracy. Using model´s results is possible now to create betting strategies. Finally, the results of the model and betting strategies are evaluated through several metrics and comparisons with the results found in the literature.

KEYWORDS

NBA; basketball; prediction; machine learning; data analysis

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABELAS

1. INTRODUÇÃO

Este projeto mescla duas áreas que estão cada vez mais conectadas: desportos e análise de dados. A explorar a crescente quantidade de dados criados diariamente, associado ao apelo que eventos desportivos têm ao redor do mundo, capazes de movimentar mil milhões de dólares em apostas, este projeto concentra-se na análise preditiva de jogos da National Basketball Association (NBA) no contexto dos mercados de apostas.

O presente trabalho é dividido da seguinte forma: o Capítulo 1 introduz o problema e apresenta os objetivos do trabalho; no Capítulo 2 é feita a revisão da literatura que começa por tratar da dinâmica dos mercados de apostas e termina com os recentes desenvolvimentos na previsão de resultados de desportos; a metodologia utilizada é relatada no Capítulo 3 onde são descritas as fontes de dados e as variáveis utilizadas, além de explicar o funcionamento dos principais modelos e como seus desempenhos são avaliados; os resultados e discussões estão no Capítulo 4 e incluem observações sobre o modelo, avaliação da precisão e os resultados das simulações de apostas; por fim, no Capítulo 5 são apresentadas as conclusões, as limitações e propostas para trabalhos futuros.

1.1. CONTEXTO

O foco deste projeto será a análise dos jogos da NBA que é a principal liga de basquetebol profissional masculino do mundo. A liga surgiu em 1946 com 11 equipas e, por meio de expansões, atualmente abriga 30 equipas dos Estados Unidos e do Canada. Desde a temporada 2004-05 as equipas estão organizadas em função de sua localização geográfica em duas conferências com três divisões de cinco equipas cada. As equipas, por sua vez, são compostas por 15 jogadores, sendo que 5 desses estão em quadra simultaneamente durante o jogo.

A temporada da NBA dura em média 8 meses e é dividida em duas partes: regular e pós-temporada. Durante a temporada regular, cada equipa joga 82 jogos, sendo metade em casa e outra metade fora, além de enfrentar todas as outras equipas no mínimo duas vezes. A temporada regular tem como principal objetivo definir os classificados para a pós-temporada. Na pós-temporada, também conhecida como *playoffs*, as oito melhores equipas de cada conferência competem pelo título em um formato de torneio. As séries são definidas em melhor de sete jogos, com a primeira equipa a vencer quatro jogos avançando para a próxima rodada, enquanto a perdedora é eliminada. O processo se repete até haver apenas uma única equipa, que será a campeã da temporada.

A NBA teve grande aumento de popularidade na década de 1980, fruto da presença de grandes estrelas como Magic Johnson, Larry Bird e principalmente Michael Jordan, que graças a sua personalidade e forma de jogar ajudou a alavancar o interesse público pela liga (Fromal, 2018). O público internacional também demonstra cada vez mais fascínio pela NBA. A temporada 2017-18 contou com 108 jogadores internacionais a representar 42 países e teve jogos transmitidos para mais de 90 países (Krasnoff, 2018).

A popularidade da NBA também é traduzida em enormes quantias. Atualmente, uma equipa vale, em média, US$ 1,65 mil milhões e os jogadores da NBA são os atletas mais bem pagos do mundo (Gaines, 2018; Badenhausen, 2018). Além disso, apenas nos casinos de Las Vegas, mais de US$ 1 mil milhão são gastos em apostas anualmente. Se for considerado o mercado de apostas ilegal e online, este valor pode chegar a quantias superiores a US$ 100 mil milhões anuais (Rybaltowski, 2018).

1.2. PROBLEMA

Avanços na área de análise de dados vem permitindo que uma variedade de problemas de diferentes áreas ganhe novas soluções. Com esse incentivo, o uso da análise de dados em competições desportivas apresenta um enorme crescimento. Constantemente são desenvolvidas novas técnicas para quantificar e melhorar o desempenho dos jogadores e das equipas. Além disso, a previsão de resultados desportivos é um tema que tem sido amplamente estudado em pesquisas onde o objetivo é construir modelos capazes de prever resultados de eventos dada alguma descrição qualitativa dos mesmos. *Machine learning* ou aprendizagem de máquina é a área que busca utilizar algoritmos para prever o futuro através do aprendizado feito com base em eventos históricos.

A NBA é um dos eventos que mais utiliza essas análises para a tomada de decisões, entretanto há ainda um amplo espaço para a criação de valor através do uso inteligente de dados. Cada temporada da NBA consiste em quase 1300 jogos e cada jogo gera uma grande quantidade de dados que podem ser divididos em três grandes grupos: pontuação, jogadas e rastreamento. Os dados de pontuação fornecem um resumo do jogo ao registar estatísticas dos jogadores e equipas. Os dados de jogadas representam uma linha de tempo do jogo, uma vez que cada evento que acontece é gravado com o tempo exato. Ainda existem os dados de rastreamento, onde todos os movimentos dos jogadores e da bola são registados usando o sistema de rastreamento de vídeo, permitindo assim a reconstrução completa do jogo. Com tanta informação disponível é possível construir modelos capazes de prever o desfecho de um jogo e também compreender fatores que levam ao sucesso de uma equipa.

Considerando a natureza de um jogo de basquetebol, torna-se interessante identificar e medir a influência que certas características têm no sucesso de uma equipa, sendo a seleção das variáveis um dos maiores desafios nesse tópico. A maneira mais simples de descrever as equipas de basquetebol de tal forma que o sucesso em um jogo possa ser previsto está relacionada com os pontos marcados e permitidos por jogo. Relacionadas a essas métricas, incluem-se os arremessos de quadra, os arremessos de três pontos, os lances livres, os rebotes ofensivos e defensivos, entre outros fatores relacionados às regras do jogo e ao comportamento e desempenho humano. O problema dessas estatísticas é que elas são números brutos, tendo assim uma interpretação limitada. Transformar essas métricas em taxas e buscar uma normalização dos dados é um dos primeiros desafios. Posteriormente, torna-se necessário encontrar outras variáveis que possam estar relacionadas com o resultado de um jogo, por exemplo, a ausência do melhor jogador da equipa impacta diretamente nas chances de vitória da mesma.

Um dos motivos da previsão de resultados desportivos ter uma grande popularidade encontra-se no fato de elas poderem fornecer informações importantes sobre como funcionam os mercados de apostas. Então, combinar modelos preditivos com apostas é um tema atraente, uma vez que permite testar a eficiência dos mercados de apostas ao criar diferentes estratégias com intuito de gerar um retorno financeiro. Portanto, por mais que seja um desafio prever os resultados desportivos, o trabalho torna-se interessante tanto pela possibilidade de lucro quanto pela popularidade da NBA. Sendo assim, este projeto visar aplicar técnicas de *machine learning* para prever o resultado de jogos da NBA e, a partir dessas previsões, criar modelos de apostas.

1.3. OBJETIVOS

O objetivo principal deste projeto é usar vários algoritmos de *machine learning* para prever resultados de jogos de basquetebol, mais especificamente da NBA. A partir do resultado do modelo, será construído um simples sistema de apostas. Combinando a saída do modelo com probabilidades de vitória de cada equipa (retiradas de sites de apostas), será possível definir uma estratégia que maximize o lucro. Cada jogo será considerado como um problema de decisão independente e terá as seguintes possibilidades: apostar na vitória de uma das equipas ou simplesmente não apostar.

A precisão de previsão é um ponto crucial para o projeto. Encontrar a melhor combinação possível de variáveis é a fórmula para obter um modelo com boa precisão. Dessa forma, se faz necessário possuir um maior entendimento sobre o problema a fim de garantir uma escolha correta das variáveis, para assim, culminar em resultados superiores aos de estudos anteriores. Também se faz necessário avaliar a precisão do modelo sob diferentes métricas e comparar os resultados com o estado-da-arte presente na literatura.

Outro objetivo deste projeto é explorar as ineficiências do mercado de apostas, ou seja, tirar proveito do mercado para obter lucro financeiro com apostas. Para verificar a viabilidade desse objetivo, simulações de apostas para verificar o quão lucrativo o modelo seria se fosse utilizado para tomar decisões de apostas são criadas. Até o momento, todos estudos mostram que apenas ganhos marginais são possíveis ao comparar modelos de previsão com às previsões produzidas pelas casas de apostas. Portanto, para obter resultados satisfatórios é preciso aliar a capacidade de estimar com precisão a probabilidade de cada resultado com um sistema de apostas que consiga encontrar as ineficiências do mercado de apostas.

Por fim, as decisões devem ser tomadas de forma rápida. Assim, o desafio de coletar dados e construir modelos preditivos sólidos no menor tempo possível possui grande importância. Uma vez que é necessário coletar dados diariamente, deve ser criado um processo prático e autônomo para extração dos dados e execução dos algoritmos.

2. REVISÃO DA LITERATURA

Antes que possa ser discutida a aplicação dos modelos preditivos na NBA com mais detalhes, é necessário colocar o assunto em contexto. Enquanto pode ser argumentado que pesquisar o tema é relevante simplesmente por causa do desporto em si, existe ainda uma vertente econômica importante. Grande parte da literatura acadêmica sobre previsão desportiva tem como principal interesse a utilização dos resultados dos modelos de previsão como ferramentas para estudar a dinâmica das apostas desportivas. Para destacar essa conexão, a revisão de literatura começa discutindo o mercado de apostas. Depois disso, passa a ser discutido a aplicação de métodos preditivos em diferentes desportos como foco especial na NBA, através da análise de diferentes estudos feitos sobre esse assunto.

2.1. CASAS DE APOSTAS E EFICIÊNCIA

O principal atrativo para usar modelos matemáticos para prever resultados desportivos é a chance de ganhos financeiros. Com uma vasta quantidade de casas de apostas disponíveis on-line, atualmente é possível apostar em praticamente todos os desportos. Além disso, enquanto antigamente as apostas se limitavam ao vencedor dos jogos, hoje os tipos de apostas são limitados apenas pela imaginação das casas de apostas e pela demanda do público. O surgimento de bolsas de apostas no início dos anos 2000 trouxe ainda mais oportunidades aos apostadores. Nesse serviço os apostadores que querem apostar em um determinado resultado são emparelhados com outros que desejam apostar o oposto. As bolsas de apostas não precisam calcular as probabilidades do evento, diminuindo custos operacionais e transferindo a incerteza da informação aos apostadores, possibilitando menores margens que são usadas para atrair um maior número de apostadores (Smith, Paton & Vaughan, 2006).

As casas de apostas usam modelos matemáticos que aliam dados estatísticos e históricos sobre as equipas, opiniões de especialistas e, principalmente, tendências de apostas para definir os preços dos eventos. A probabilidade do evento é apresentada pelo modelo e, então, convertido em preço. A partir do volume de apostas que as casas recebem, mudanças de preços podem ocorrer. No caso em que as apostas são feitas de acordo com os cálculos previsto pelo modelo, os preços permanecem estáveis. Se, por outro lado, uma quantia desproporcionalmente grande de dinheiro é colocada em dos lados, os preços são recalculados para atrair mais pessoas a colocarem dinheiro no lado com menos apostas. A prioridade das casas de apostas é garantir o lucro independentemente de qual resultado prevalecer. A lógica de receita nesta atividade é baseada em um *overround*, uma percentagem do recebimento das apostas que são mantidas pela casa de aposta em vez de distribuí-lo aos vencedores da aposta em questão. Vários sites comparam preços de uma mesma aposta em diferentes casas fazendo que margens tão baixas quanto 2% sejam encontradas. A popularidade das apostas online aumentou a concorrência e isso, por sua vez, forçou as casas de apostas a dar maior ênfase no desenvolvimento de métodos de previsão, já que não é mais possível compensar a imprecisão com *overrounds* muito grandes (Pinnacle, 2016).

Embora o conceito de apostas exista há séculos, a maior parte dos trabalhos publicados nessa área foi desenvolvido nas últimas décadas. Acadêmicos e economistas tem se esforçado para mostrar a ineficiência do mercado de apostas através do uso de estratégias lucrativas. Kyupers (2000) aponta que grande parte da pesquisa acontece devido à forma como os mercados de apostas são

constituídos: as apostas são oferecidas no mercado e informações detalhadas sobre seus preços e o resultado de seus eventos correspondentes são facilmente encontrados. Essa riqueza de informações públicas fornece um terreno fértil para a pesquisa empírica. Outros dois fatores também contribuem para a profundidade do trabalho: a popularidade do desporto em estudo e o período disponível de apostas. Como as casas de apostas de Las Vegas apostam nos desportos desde a década de 1960 é de pouca surpresa que haja uma maior quantidade de estudos relativos aos desportos nos Estados Unidos.

Ao explorar a eficiência do mercado, diversas abordagens podem ser usadas para identificar e classificar a potencial eficiência de determinado mercado. O conceito mais famoso foi criado por Fama (1970). Nessa abordagem um mercado é identificado como eficiente quando seus preços refletem totalmente todas as informações disponíveis. Em outras palavras, em mercados eficientes os ativos são sempre negociados ao "preço certo", de forma que retornos anormais não podem ser alcançados. Três níveis de eficiência são criados para quantificar a eficiência do mercado:

- Fraco, quando os preços refletem apenas preços passados;
- Semiforte, quando envolvem toda a informação disponível publicamente;
- Forte, quando todas as informações públicas e privadas são utilizadas e, portanto, não é possível alcançar, consistentemente, retornos anormais no mercado.

Vários estudos publicados recentemente testam a eficiência dos mercados de apostas desportivas. Kyupers (2000) identificou que, em geral, os fatores que promovem uma maior eficiência dos mercados de apostas são o número de participantes, o volume de apostas e a notoriedade do evento. Dado que é praticamente impossível ter acesso a todas informações disponíveis, é comum classificar as casas de apostas como um mercado semiforte. Makropoulou e Markellos (2011) argumentam ainda que, devido à extensa cobertura e regulação da imprensa, é muito improvável que os preços das apostas desportivas sejam impactados por quantidades significativas de informações privilegiadas e indisponíveis para o resto do público. Além disso, essa classificação é sustentada pelo fato de que os acadêmicos não conseguiram encontrar sinais definitivos de ineficiência. Ainda que alguns exemplos pontuais sejam relatados (Dias, 2016), os resultados falam mais em favor do mercado semiforte do que a favor da ineficiência.

2.2. MACHINE LEARNING

Desde que os computadores foram inventados, sempre houve o desejo que eles pudessem aprender da mesma forma que é comum aos seres humanos. Com o passar do tempo, começaram a surgir algoritmos capazes de realizar tarefas que precisavam de aprendizado, como fazer previsões precisas e outras aplicações feitas de forma automática, sem intervenção ou assistência humana. No campo da prospeção de dados (em inglês, *data mining*), os algoritmos de *machine learning*, aliados a outras disciplinas como estatística e ciência da computação, estão sendo usados rotineiramente para extrair conhecimentos valiosos e encontrar respostas diretamente dos bancos de dados (Mitchell, 1997).

Com a evolução da informática - armazenamento de dados mais barato, processamento distribuído, computadores mais poderosos e oportunidades analíticas disponíveis - o interesse por algoritmos de *machine learning*, que agora podem ser aplicados em enorme quantidade de dados, aumentou drasticamente (Bucheli & Thompson, 2014). Considerando a dinâmica e os benefícios desses algoritmos, o principal objetivo da pesquisa nesse campo é desenvolver algoritmos eficientes que

possam resolver um determinado problema prático. Idealmente, o mesmo algoritmo deve ser facilmente aplicado a uma ampla classe de problema de aprendizagem.

Os algoritmos devem ser capazes de analisar conjuntos de dados massivos e aprender com eles. Quando os algoritmos são treinados com dados onde o resultado desejado é conhecido, temos uma aprendizagem supervisionada, muito comum em aplicações que usam dados históricos para prever possíveis eventos futuros, como neste projeto. A resolução de um problema passa pela construção de um modelo onde várias escolhas se fazem necessárias: definição do algoritmo, forma de treinamento, função a ser aprendida e representação para essa função (Bunker & Thabtah, 2017).

Algoritmos de *machine learning* provaram ser de grande valor prático em uma variedade de aplicações em diferentes áreas. Eles são especialmente úteis para descobrir padrões automaticamente, desenvolver programas dinâmicos adaptáveis a mudanças de condições e explorar campos pouco compreendidos, onde humanos podem não ter o conhecimento necessário para desenvolver algoritmos efetivos (Mitchell, 1997). Alguns exemplos do uso de *machine learning* são:

- Segmentação de clientes e comportamento do consumidor, área que procura prever as decisões de compra com base na precificação adaptativa ou dinâmica de um produto;
- Aplicações no setor financeiro, onde aprovações de empréstimos, gestão de ativos, perfil de risco, deteção de fraudes e previsões de mercado são feitos através de técnicas *machine learning*;
- Veículos autónomos, onde visão computacional é combinada com o aprendizado profundo para trazer uma solução relativamente barata e robusta para a direção autónoma.

O contexto de resolução de problemas através de técnicas de *machine learning* traz também oportunidades para o mundo dos desportos, onde as pessoas gostam de apostar e prever quem vai ganhar um determinado evento ou jogo. Nesses momentos, os palpites geralmente são guiados pela intuição, experiência e pela capacidade de detetar padrões, mesmo que inconscientemente. Algumas dessas motivações são, justamente, as responsáveis por más decisões, como: tendência para identificar padrões inexistentes, excesso de emoção envolvida e falta de feedback, uma vez que sem saber dos erros, é impossível aprender com eles (Hayashi, 2001).

Muitos estudos e experimentos foram feitos para buscar outras formas de tomar decisões que não sejam baseadas puramente na intuição. No caso do basquetebol, por exemplo, há uma enorme quantidade de dados disponíveis. Existem estatísticas individuais e coletivas, ofensivas e defensivas, e equipas inteiras possuem uma imensidão de dados que tentam quantificar o desempenho em qualquer parte do jogo. Com o aumento da acessibilidade desses dados e a evolução da tecnologia, experimentos mais complexos foram realizados no campo da análise preditiva de resultados desportivos.

2.2.1. Previsão de resultados desportivos

Dubbs (2016) utiliza uma abordagem simples para prever os resultados dos principais desportos americanos (basquetebol, futebol americano, hóquei e beisebol). Em seu trabalho, o algoritmo utilizado é a regressão linear com o método dos mínimos quadrados sendo a técnica para encontrar o melhor ajuste para o conjunto de dados. As previsões são feitas usando apenas quatro variáveis coletadas para 30 temporadas de cada um dos despostos analisados: equipa mandante, equipa

visitante, data e resultado. Intencionalmente, nenhuma estatística das partidas foi usada, a fim de verificar o real poder preditivo destas. Os resultados são comparados com um valor ótimo teórico, definido pelo autor da seguinte forma: como se sabe a classificação de cada equipa ao final da temporada, prevê que sempre a equipa melhor classificada vença os jogos. Os resultados mostram que as estatísticas não possuem grande impactos na predição de jogos de basquetebol, uma vez que as previsões do estudo conseguiram uma alta precisão tanto em valor absoluto quanto na comparação com o valor ótimo teórico. Enquanto no futebol americano esse claramente não é o caso, uma vez que os resultados foram abaixo do esperado. Para o beisebol e o hóquei os resultados foram intermediários, mostrando que há algum potencial para melhorar as previsões ao encorpar estatísticas no conjunto de dados.

O estudo de Praet (2017) tem como objetivo descobrir se é possível prever os resultados de jogos desportivos através de sistemas de recomendação. Tênis é escolhido como o foco do trabalho devido as seguintes características: importância limitada do árbitro, do treinador e da torcida; existência de apenas dois resultados possíveis; ser um desporto individual, uma vez que estatísticas de desportos de coletivos podem estar distorcidas devido a mudança constante nos jogadores das equipas. O melhor resultado é atingido utilizando *bagging* associado a regressão logística. Entretanto, apesar do tênis reunir as características ideias para previsão dos vencedores das partidas, os resultados encontrados são similares aos valores encontrados em outros desportos. O autor atribui essa situação ao conjunto de dados limitados, com pouca informação de jogadores jovens e com uma pequena quantidade de estatísticas disponíveis. Como conclusão, é destacado que os jogadores são mais previsíveis em torneios Grand Slam e que o ranking da ATP possui grande importância no resultado das previsões.

Sillanpää e Heino (2013) focam seu trabalho na previsão dos resultados de jogos de campeonatos europeus de futebol, tendo o objetivo de avaliar como um modelo de previsão se compara com as probabilidades geradas pelo mercado de apostas. O modelo é criado com base no sistema de classificação Elo, que atribui pontuações relativas ao desempenho das equipas. Os resultados indicam que o uso de fontes de informação auxiliares, como histórico em outras competições e distâncias viajadas, não melhoram substancialmente a precisão das previsões em comparação com um modelo feito somente com o uso de variáveis baseadas na classificação Elo. Em termos de precisão e simulações de apostas, o modelo produz retornos melhores do que apostar aleatoriamente, mas é incapaz de gerar retornos positivos consistentes, mesmo que a precisão do modelo seja similar a precisão média das casas de apostas.

Bailey (2005) busca estabelecer uma abordagem estatística consistente para ajudar na previsão de resultados da liga de futebol australiano (AFL) e do críquete. Modelagem multivariada é utilizada para ponderar as contribuições de cada variável disponível e produzir equações que sirvam para determinar a probabilidade do resultado e a capacidade preditiva da abordagem escolhida. O estudo ainda incorpora o efeito de jogadores individuais para produzir melhores previsões. Usando uma estratégia de apostas, o resultado do modelo é utilizado para determinar a eficiência do mercado de apostas. O autor conclui que os mercados de apostas da AFL e do críquete são estatisticamente ineficientes ao longo de um período de análise, uma vez que a estratégia de apostas obteve retornos positivos. Após essa etapa, o estudo ainda demonstra a aplicação de modelos estatisticamente orientados para prever do desempenho individual dos jogadores.

Prever corretamente o vencedor de todas partidas do March Madness, torneio final do campeonato de basquete universitário norte-americano, é o grande objetivo de Fonseca (2018). Para isso, dados de partidas dos últimos 20 anos são coletados, processados e analisados. Neste estudo, diversos algoritmos de *machine learning* são testados, entre eles: árvores de decisão, k-vizinhos mais próximos, máquina de vetores de suporte, entre outros. As variáveis mais importantes para a previsão são a classificação da equipa, precisão nos arremessos e número de participações passadas no torneio. Com relação aos resultados, ao treinar o conjunto de dados inteiro, a precisão chega a 70%, sendo ligeiramente inferior a escolher sempre a equipa com melhor classificação. Por outro lado, ao prever apenas o torneio de 2017 a precisão chega a 85%, superando estudos anteriores e previsões de especialistas. Em ambos casos, máquina de vetores de suporte foi o algoritmo que apresentou o melhor resultado. O trabalho também se destaca pela sua extensa revisão bibliográfica.

2.2.2. Previsão de resultado na NBA

2.2.2.1. Predicting the Outcome of NBA Playoffs Based on the Maximum Entropy Principle

O estudo de Cheng et al. (2016) visa prever os jogos da NBA com auxílio do conceito de entropia. Seu primeiro passo é coletar as 14 principais estatísticas de cada equipa em mais de 10 mil jogos e converter as variáveis que possuem valores numéricos contínuos em valores discretos, uma vez que esse é um requisito do algoritmo de máxima entropia. Essa conversão é realizada com o uso do algoritmo *k-means clustering*. Para valores de k variando entre 3 e 10 é calculado o coeficiente s*ilhouette*, sendo o valor de k que maximiza esse coeficiente definido como número de cluster que terá a variável. O procedimento é repetido para todas variáveis.

O conjunto de dados discretizado é então usado para treinar o modelo de máxima entropia da NBA, denominado NBAME. Os modelos de máxima entropia são projetados para resolver problemas com dados insuficientes e aponta a melhor aproximação para a distribuição de probabilidade desconhecida, não fazendo suposições subjetivas e diminuindo o risco de fazer previsões erradas. O algoritmo tem sido amplamente usado para tarefas de processamento de linguagem natural. O modelo NBAME tem como resultado a probabilidade da vitória da equipa da casa no jogo. Como a probabilidade é um valor contínuo, o modelo faz a previsão de vitória com base em um limiar igual a 0,5. A Tabela 1 mostra os resultados obtidos pelos autores em diversas técnicas, mostrando a superioridade do NBAME em quase todos os cenários. Assumindo que a probabilidade de vitória da equipa da casa é maior, limiares de 0,6 e 0,7 também são testados. Entretanto, nesses casos, ao aumentar o nível de confiança, há uma diminuição significativa no número de jogos a ser previsto.

Ano	Naïve Bayes	Rede Neuronal	Regressão Logística	Floresta Aleatória	NBAME
2007-08	54.7%	59.3%	61.6%	64.0%	**74.4%**
2008-09	61.5%	60.4%	57.1%	60.4%	**68.2%**
2009-10	56.1%	52.4%	61.0%	64.6%	**68.3%**
2010-11	59.3%	**67.9%**	61.7%	64.2%	66.7%
2011-12	53.6%	56.0%	60.7%	58.3%	**69.0%**
2012-13	58.8%	63.5%	64.7%	**70.6%**	67.1%
2013-14	59.3%	57.1%	62.6%	62.6%	**65.2%**
2014-15	55.0%	57.5%	60.0%	56.3%	**62.5%**

Tabela 1: Comparação da precisão de previsão em função da técnica e do ano analisado.

2.2.2.2. Modelling the NBA to Make Better Predictions

Puranmalka (2013) busca modelar um jogo da NBA para prever com sucesso o desfecho dos mesmos. O principal foco do seu trabalho é a criação de novas variáveis que possam explicar o resultado do jogo. Utilizando dados de pontuação e jogadas, diversas novas variáveis são criadas gerando modelos que preveem o jogo ao nível das equipas e ao nível dos jogadores.

Nos modelos preditivos que usam estatísticas ao nível das equipas fica claro que duas variáveis criadas nesse estudo são importantes na previsão dos resultados dos jogos. São elas: relações de força entre as equipas (quando uma boa equipa nos rebotes joga com uma ruim nesse quesito, etc.) e a medida de desempenho decisivo (definida por jogos que se aproximam do fim com as equipas tendo pontuações muito próximas). Embora essas variáveis sozinhas não sejam tão significativas quanto as variáveis tradicionais, o fato de elas adicionarem valor ao modelo por si só já é relevante.

Nos modelos preditivos que usam estatísticas ao nível dos jogadores outras duas variáveis criadas para o estudo mostram-se relevantes. A primeira é a eficiência adicionado pelo contexto (métrica que mescla o tempo restante para o arremesso com a eficiência ofensiva e defensiva) e a segunda é a interação entre jogadores (como a presença de um jogador influencia a precisão de arremesso de seus companheiros de equipa, por exemplo). O fato da interação entre jogadores ser importante sugere que um determinado conjunto de jogadores não necessariamente atua de forma independente; em vez disso, o desempenho de um jogador pode ser altamente dependente de seus companheiros de equipa.

A precisão do melhor modelo criado no estudo, conseguido através do uso do algoritmo máquina de vetores de suporte (SVM), é comparada com o acerto médio das casas de apostas de Las Vegas e os resultados ano a ano são apresentados na Tabela 2. Mesmo que o modelo criado por Purumalka seja muitas vezes superior as casas de apostas, a diferença entre ambos é muita pequena para garantir um retorno financeiro consistente.

Ano	Purumalka SVM	Las Vegas Apostas
2003	**73.5%**	69.3%
2004	**73.2%**	68.5%
2005	72.5%	**73.0%**
2006	**73.0%**	69.6%
2007	70.0%	**72.2%**
2008	**72.7%**	67.0%
2009	69.5%	**71.6%**
2010	**71.2%**	67.1%
2011	70.2%	**70.4%**
2012	**67.2%**	66.7%

Tabela 2: Comparação da precisão da previsão do modelo com o acerto médio das casas de apostas.

2.2.2.3. Predicting NBA games with matrix factorization

A maioria dos trabalhos encontrados na literatura usa as estatísticas da partida associadas a algoritmos prontos como regressão logística, SVM e redes neuronais para prever os jogos da NBA. Tran (2016) apresenta uma abordagem diferente para este fim. O autor apresenta a factoração de

matrizes, técnica que ganhou visibilidade ao ser usada no problema de recomendação do Netflix, e que, em geral, pode ser aplicada a dados que são mais bem modelados como resultado da interação de pares. Como um jogo da NBA nada mais é do que uma interação entre duas equipas, a factoração de matrizes pode ser uma abordagem interessante e adequada para o problema da previsão do resultado de um jogo.

A ideia básica por trás da factoração de matrizes é encontrar duas matrizes tais que sua multiplicação devolva a matriz original. A intuição da técnica é que deve haver algumas características latentes que determinam as interações entre os pares. No caso do Netfilx, por exemplo, busca-se entender como um utilizador classifica determinado filme. Quando dois utilizadores classificam um filme de maneira semelhante, a razão para isso acontecer pode ser por ambos gostarem dos atores ou atrizes do filme ou por o filme ser de um gênero que agrade ambos. Portanto, ao descobrir essas características latentes, deve ser possível usá-los para prever a classificação de um filme antes de ele ser avaliado. Dessa forma, factoração de matrizes é uma ferramenta matemática que pode ser usada em cenários onde é preciso descobrir algo oculto sob os dados.

Após a introdução do assunto, o trabalho é dividido três partes. Primeiro, a previsão da NBA é feita com uma abordagem básica da factorização de matrizes (FM), visando descobrir a estrutura nos dados. Depois, a factoração de matrizes probabilística (FMP) é usada para incorporar o fato de que, quando duas equipas jogam entre si, as pontuações serão diferentes a cada vez. Por fim, informações complementares como a data e o local do jogo são incorporadas, combinando factorações de matrizes probabilísticas usando o processo de priorizações gaussianas (FMPG). Esse último modelo é uma generalização da factoração de matrizes probabilística que substitui as características latentes escalares por funções, cujas entradas são as informações suplementares. Os resultados obtidos para esta temporada regular 2015-16 para cada um dos modelos é apresentado na Tabela 3.

Modelo	Precisão
FM	70.9%
FMP	71.4%
FMPG	72.1%

Tabela 3: Comparação da precisão de previsão para diferentes técnicas.

2.2.2.4. The prediction of outcomes in the National Basketball Association

O objetivo do trabalho de Park (2014) é construir modelos para prever jogos da NBA antes do seu início e também durante a sua realização. O modelo pré-jogo inicial é criado empregando a classificação Elo para estimar as probabilidades de vitória. Posteriormente são incluídas outras métricas comuns em previsões da NBA como: arremessos de quadra, rebotes, roubos de bola, etc. Por fim, o melhor modelo pré-jogo é encontrado ao se utilizar uma combinação das avaliações ofensivas e defensivas (estimativa de pontos marcados/permitidos por 100 posses) nos últimos jogos de cada equipa. A probabilidade de vitória é calculada inserindo as variáveis na função logística, com os coeficientes de regressão otimizados para o menor valor qui-quadrado com a ajuda do software Risk Optimizer e variando conforme equipa e local do jogo.

O resultado do modelo foi avaliado em simulações de apostas. Ao utilizar o conjunto completo de jogos previstos, não foi possível obter lucro. Diversos cenários são avaliados e intervalos ótimos, faixa para qual o modelo fornece previsões melhores que as casas de apostas, são determinados de modo

a maximizar o lucro. Depois de obter esses valores, essas faixas são inseridas nas simulações. Apesar de o conjunto de dados estar limitado a aproximadamente 21% do total, o lucro observado fica próximo a 20%. Apesar dos lucros, a taxa de ganho nas apostas é baixa (em torno de 30%) o que faz que a estratégia seja perigosa e instável.

A previsão dos resultados enquanto o jogo acontece é investigado na sequência com uma abordagem baseada em simulações de Monte Carlo e na distribuição da probabilidade da pontuação. O tempo de jogo e a qualidade da equipa (medida pelas margens previstas por casas de apostas pré-jogo) são usados para prever a pontuação final. O modelo então analisa a percentagem de arremessos verdadeiros a cada 3 minutos para atualizar a distribuição de pontuação. Os resultados apesar de não representar fielmente o placar final apresentam um erro pequeno. Quando essas informações são incluídas em uma estratégia de apostas, o autor consegue retornos de até 10%.

2.2.2.5. Mason: Real-time NBA Matches Outcome Prediction

Os trabalhos de previsão desportiva normalmente fazem previsões antes do jogo começar. Previsões durante o jogo, ou em tempo real, ainda não foram suficientemente estudadas. Durante uma partida os dados são gerados cumulativamente tornando-se mais abrangentes e potencialmente com maior poder preditivo. Lin (2017) cria variáveis ao nível da equipa e do jogador com base nos dados em tempo real nos jogos das últimas 5 temporadas da NBA. Modelos baseados em regressão logística são criados para investigar as características e a possibilidade do uso de dados em tempo real para prever partidas da NBA. Com esse objetivo, duas hipóteses são levantadas pelo autor:

1. A precisão da previsão aumenta dinamicamente com a partida em andamento sempre que as previsões forem feitas com o mesmo conjunto de variáveis e o mesmo modelo de treinamento.
2. Resultados precisos podem ser alcançados com conjuntos de variáveis e modelos que são facilmente entendidos.

Cinco modelos são criados para avaliar essas hipóteses. O mais simples e que serve de base de comparação é denominado “diferença histórica” e faz a previsão antes do jogo começar comparando os registos do histórico entre duas equipas. O modelo “diferença presente” prediz o resultado da partida de acordo com a diferença de pontos entre as duas equipas no momento atual. “Diferença X recente” usa uma janela temporal para incluir as diferenças de pontos nos últimos X intervalos de tempo. Para incluir informações em tempo real ao nível do jogador é criado a abordagem “top K estatísticas”, onde as 18 estatísticas tradicionais presentes no conjunto de dados são selecionadas para os K jogadores líderes em cada métrica. O último modelo é uma combinação dos dois anteriores e, como esperado, é o modelo que obtém o melhor resultado. Por fim, o autor consegue demonstrar que ambas hipóteses iniciais são verdadeiras.

3. METODOLOGIA

O uso de uma abordagem inteligente e estruturada para resolver o problema da previsão de resultados no desporto é importante, pois garante que o projeto está sendo desenvolvido de forma eficaz, sem pular etapas, além de assegurar que os resultados estão corretos. Dessa forma, o presente trabalho utiliza a metodologia proposta por Bunker e Thabtah (2017), composta por seis passos:

1. Compreensão do assunto: inclui compreender o problema e as características específicas do desporto em si. É necessário entender como o desporto é jogado e quais fatores estão relacionados com o resultado dos jogos. Também é preciso haver clareza em relação ao objetivo do modelo, uma vez que cada objetivo possui diferentes fatores que devem ser levados em consideração. Objetivos comuns são: prever resultados para competir com previsões de especialistas, usar os resultados do modelo em apostas e competições online.

2. Compreensão dos dados: etapa em que os dados são obtidos e onde é definido a granularidade e o tipo de problema (classificação ou regressão). Em relação a granularidade, grande parte dos trabalhos utiliza dados no nível do jogo/equipa, porém também é possível incluir dados no nível do jogador, com estatísticas sobre os jogadores que atuaram em cada jogo. A inclusão de dados no nível do jogador traz a possibilidade de investigar se as ações ou a presença de jogadores específicos influenciam o desempenho da equipa. Em relação ao tipo de problema, geralmente a previsão do resultado de jogos é tratado como um problema de classificação com duas classes (vitória ou derrota), mas também pode ser considerado como um problema de previsão numérica, usando técnicas de regressão para prever a margem de pontos e então fazer a previsão de vitória-derrota baseada nessa margem de pontos prevista.

3. Preparação dos dados: etapa onde os dados são tratados conforme sua origem e novas variáveis (*features*) são criadas. Deve haver um tratamento diferente para variáveis "relacionadas ao jogo" e "externas". As primeiras estão relacionadas a eventos do jogo, como a quantidade de assistências ou pontos marcados no basquetebol. Já as segundas incluem eventos externos ao jogo como sequência de vitórias e dias de descanso entre jogos. Variáveis externas são conhecidas antes do jogo ser disputado, já variáveis relacionadas ao jogo são desconhecidas até que o mesmo tenha sido disputado. Dessa forma, apenas as médias anteriores das variáveis relacionadas ao jogo podem ser utilizadas para prever o resultado futuro.

4. Modelagem: momento onde os algoritmos utilizados na experimentação são selecionados. Envolve revisão da literatura para identificar algoritmos aplicados anteriormente que foram bem-sucedidos. Idealmente, devem ser criados e testados diversos modelos com diferentes combinações de variáveis e hiperparâmetros, visando encontrar um modelo otimizado.

5. Avaliação do resultado: uma das formas mais comuns de avaliar o desempenho do modelo consiste em comparar o resultado previsto com o resultado real usando uma matriz de confusão. É incomum haver um grande desequilíbrio nos valores de cada classe, embora,

dado o fenômeno da vantagem para a equipa da casa ser comumente observado, é provável que haja uma ligeira inclinação em favor da equipa que joga em casa. Em casos assim, a precisão é uma medida razoável de avaliação. Nos casos em que os dados são altamente desequilibrados, a avaliação da curva ROC pode ser mais apropriada. Também é necessário escolher o método de validação cruzada apropriado para avaliar a capacidade de generalização do modelo.

6. Implementação: a construção de um modelo robusto passa pela automatização do processo. Novos dados devem ser obtidos e pré-processados para que o modelo seja retreinado e gere previsões para os próximos jogos. Idealmente, esse processo deve ser incremental, uma vez que o conjunto de dados de treino é continuamente atualizado e, portanto, o modelo deve continuar mudando para refletir as mudanças no ambiente de aprendizado.

A Figura 1 mostra as relações entre cada etapa da metodologia utilizada para resolver o problema de previsão de resultados no desporto. É importante ressaltar que não se trata de um processo linear, uma vez que muitas iterações devem ser feitas durante todo o ciclo de desenvolvimento.

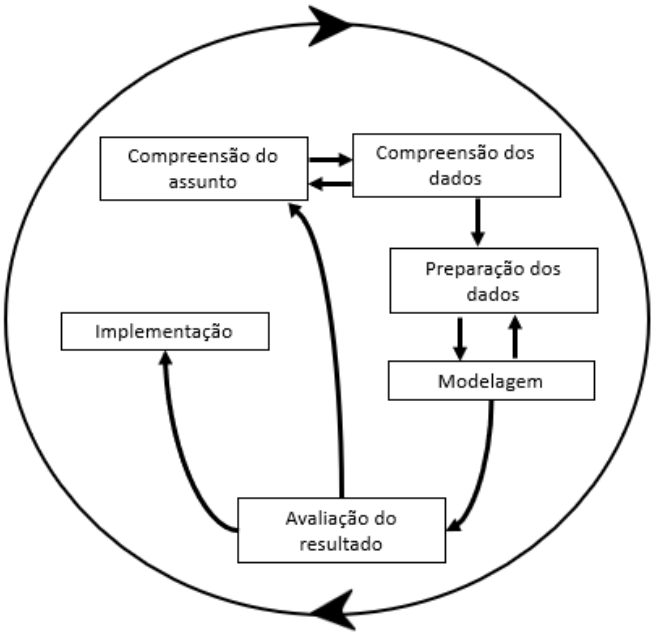

Figura 1: Metodologia para resolver problemas de previsão de resultados desportivos.

3.1. COMPREENSÃO DO ASSUNTO

Este trabalho busca criar um modelo capaz de predizer o resultado de jogos da NBA que será utilizado posteriormente na criação de um sistema de apostas. Dessa forma, o passo fundamental é entender os principais aspetos que influenciam o resultado do jogo. O trabalho de Kubatko et al. (2007) criou um ponto de partida para analisar as estatísticas do basquetebol. Diversas fórmulas foram revisadas e estatísticas de fontes não acadêmicas foram reunidas e explicadas. O artigo coloca o conceito de posses como a chave para diversas análises, dessa forma, as quatro principais formas que uma posse pode terminar (arremesso, rebote, falta ou perda de bola) se mostram determinantes para as chances de vitória de uma equipa. Oliver (2004) já havia identificado a importância desses fatores para identificar os pontos fortes e fracos e consequentemente o sucesso de uma equipa.

Dada a relevância das mesmas, as seguintes estatísticas ficaram conhecidas como os “Quatro Fatores do Sucesso no Basquetebol”:

- Percentagem de arremessos de quadra efetivos: uma equipa ideal pontua toda vez que tem a bola, no entanto, as equipas nunca são ideais. Dessa forma a percentagem de arremessos de quadra efetivos introduz uma métrica corrigida para identificar a percentagem de arremessos de quadra convertidos de uma equipa. A correção é feita para levar em conta a maior pontuação do arremesso de três pontos.

- Percentagem de rebotes ofensivos: se uma equipa não pode pontuar em cada posse, então o ideal é pegar todos os rebotes e obter uma segunda oportunidade para pontuar. Para medir a qualidade da capacidade de um rebote da equipa, é calculado o número de rebotes ofensivos dividido pelo número de rebotes disponíveis após uma tentativa perdida de arremesso de quadra.

- Taxa de lance livre: a próxima maneira de marcar pontos é através de lances livres. Atrair faltas resulta em dois efeitos positivos, uma vez que garante uma nova chance de pontuar (através de lances livres) e faz com que os oponentes fiquem mais próximos de serem expulsos do jogo. Porém, apenas sofrer faltas não é suficiente, sendo fundamental convertê-las. Sendo assim, a taxa de lance livre mede a habilidade de atrair faltas e o desempenho nos lances livres de uma equipa.

- Percentagem de bolas perdidas: o último dos quatro fatores está relacionado com o cuidado com a bola. É uma medida simples que calcula a percentagem de posses que terminam em uma perda de bola.

Oliver (2004) também definiu pesos diferentes para cada fator, sendo 40 para os aremessos, 25 para as bolas perdidas, 20 para os rebotes ofensivos e 15 para os lances livres. Porém o estudo de Kotzias (2018) mostra que a proprorção correta deve ser próxima a 43/39/10/8. Desse modo, a importância da percentagem de bolas perdidas aumentou mais de 50%, enquanto a percentagem de rebotes ofensivo e a taxa de lances livres diminuíram pela metade. As descobertas da pesquisa sustentam que os quatro fatores podem produzir modelos de projeção com um nível de precisão média de quase 94% ao projetar o número total de vitórias de uma equipa ao longa da temporada.

Jacobs (2017) também analisou os quatro fatores e concluiu que na verdade são oito fatores que influenciam o resultado de um jogo, já que devem ser analisados os mesmos quatro fatores dos oponentes para obter os fatores relacionados a qualidade defensiva da equipa. Porém, para os rebotes, o autor sugere apenas substituir os rebotes ofensivos pelos defensivos ao invés de buscar informações dos oponentes da equipa. Ainda em relação aos quatro fatores, Baghal (2012) também os classificou como indicadores das características mais gerais de uma equipa. Entretanto, seu estudo não mostrou a taxa de lance livre do oponente como um fator relevante para determinar a qualidade defensiva da equipa. A hipótese do autor é que lances livres convertidos pelo oponente não são afetados pela defesa, essa somente tem influência no número de lances livres permitidos. O estudo sugere que esta medida seja substituída pelo percentual de roubos de bola, uma estatística mais no controle da defesa. Além disso, também foi concluído que os fatores ofensivos são mais importantes para o sucesso de uma equipa que os defensivos.

Ao analisar diferenças entre estatísticas das melhores e piores equipas de basquetebol, Ibanez et al. (2008) descobriram que assistências, roubos de bola e tocos são as estatísticas mais relevantes para discriminar as equipas nesses grupos. Outro dado interessante apontado no estudo reside no fato que sucesso da equipa não depende da quantidade de posses que ela tem, mas sim de como as posses são aproveitadas.

3.2. COMPREENSÃO DOS DADOS

Os dados são a base de qualquer projeto de *machine learning*, então coletar uma quantidade suficiente de dados com alta qualidade é o primeiro desafio do projeto e uma tarefa fundamental para garantir o sucesso do mesmo. Graças a entusiastas e analistas do desposto, a NBA possui diversos repositórios com vasta quantidade de dados que vem sendo acumulados há décadas, estando a maior parte desses dados disponível online sem qualquer custo, o que facilita a criação de diversas análises.

A principal fonte de dados para este projeto é o site Basketball-Reference.com que faz parte do portal Sports-Reference.com. Basketball-Reference.com foi criado em 2004 e possui uma enorme quantidade de registos, incluindo estatísticas desde o surgimento da liga. Possui dados bem organizados, precisos e com uma fácil navegação. O site OddsPortal.com serve como fonte de dados secundária, utilizado para obter as chances de vitória das equipas em diferentes sites de apostas.

Basketball-Reference.com publica informações dos jogos da NBA em páginas da web, conforme exemplo apresentado na

Figura 2. Para coletar os dados da plataforma online de forma eficaz é necessário recorrer a um web *scraper*, *script* que extrai automaticamente a informação presente em páginas da internet. Para o presente trabalho foi utilizado um código em Python combinado com algumas bibliotecas úteis, como BeautifulSoup. Os dados coletados são armazenados em arquivos JSON e são posteriormente tratados e guardados em um formato de banco de dados.

Cleveland Cavaliers (1-0) Share & more ▾ Glossary

	Basic Box Score Stats																		
Starters	MP	FG	FGA	FG%	3P	3PA	3P%	FT	FTA	FT%	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS	+/-
LeBron James	41:12	12	19	.632	1	5	.200	4	4	1.000	1	15	16	9	0	2	4	3	29	+2
Jae Crowder	34:44	3	10	.300	1	5	.200	4	4	1.000	1	4	5	2	2	0	1	2	11	+7
Derrick Rose	31:15	5	14	.357	1	3	.333	3	4	.750	1	3	4	2	0	0	2	2	14	-7
Dwyane Wade	28:30	3	10	.300	0	1	.000	2	2	1.000	1	1	2	3	0	2	4	1	8	0
Kevin Love	28:24	4	9	.444	1	4	.250	6	7	.857	3	8	11	0	0	0	2	2	15	+1
Team Totals	**240**	**38**	**83**	**.458**	**5**	**22**	**.227**	**21**	**25**	**.840**	**9**	**41**	**50**	**19**	**3**	**4**	**17**	**25**	**102**	

	Advanced Box Score Stats														
Starters	MP	TS%	eFG%	3PAr	FTr	ORB%	DRB%	TRB%	AST%	STL%	BLK%	TOV%	USG%	ORtg	DRtg
LeBron James	41:12	.698	.658	.263	.211	2.5	35.0	19.4	43.7	0.0	4.2	16.2	26.0	126	94
Jae Crowder	34:44	.468	.350	.500	.400	3.0	11.1	7.2	8.2	2.8	0.0	7.8	15.9	100	98
Derrick Rose	31:15	.444	.393	.214	.286	3.3	9.2	6.4	10.1	0.0	0.0	11.3	24.6	88	105
Dwyane Wade	28:30	.368	.300	.100	.200	3.7	3.4	3.5	15.3	0.0	6.0	26.9	22.6	67	102
Kevin Love	28:24	.621	.500	.444	.778	11.0	27.0	19.4	0.0	0.0	0.0	14.2	21.4	117	100
Team Totals	**240**	**.543**	**.488**	**.265**	**.301**	**19.6**	**82.0**	**52.1**	**50.0**	**3.0**	**7.1**	**15.3**	**100.0**	**102.7**	**99.7**

Figura 2: Exemplo das estatísticas disponíveis em Basketeball-Reference.com.

Para este projeto, 15 anos de dados estatísticos de jogos da NBA foram coletados, desde temporada 2003-04 até a temporada 2017-18. Dessa forma, existe informações de aproximadamente 20 mil jogos que são utilizados como base para a construção do modelo. Além da informação da equipa vencedora, que é utilizado como variável dependente no processo de classificação, um amplo espectro de atributos é extraído, sendo as estatísticas descritas na Tabela 4. Todos os jogos possuem informações no nível das equipas e também no nível dos jogadores, porém, dado o escopo desde projeto, o uso das informações será limitado ao nível do jogo.

Atributo	Descrição
Minutos jogados	Quantidade de minutos jogados
Arremessos de quadra	Quantidade de arremessos de quadra convertidos
Arremessos de quadra tentados	Quantidade de arremessos de quadra tentados
Taxa de arremessos de quadra	Percentagem de arremessos de quadra convertidos por tentativa
Arremessos de 3 pontos	Quantidade de arremessos de 3 pontos convertidos
Arremessos de 3 pontos tentados	Quantidade de arremessos de 3 pontos tentados
Taxa de arremessos de 3 pontos	Percentagem de arremessos de 3 pontos convertidos por tentativa
Lances livres	Quantidade de lances livres de quadra convertidos
Lances livres tentados	Quantidade de lances livres de quadra convertidos
Taxa de lances livres	Percentagem de lances livres convertidos por tentativa
Rebotes ofensivos	Quantidade de rebotes ofensivos
Rebotes defensivos	Quantidade de rebotes defensivos
Rebotes totais	Quantidade de rebotes totais (soma dos rebotes defensivos e ofensivos)
Assistências	Quantidade de assistências
Roubos de bola	Quantidade de roubos de bola
Tocos	Quantidade de tocos
Bolas perdidas	Quantidade de bolas perdidas
Faltas pessoais	Quantidade de faltas pessoais
Pontos	Quantidade de pontos
Plus/minus	Diferença entre os pontos marcados e permitidos
Percentagem de arremessos verdadeiros	Eficiência do arremesso que leva em conta arremessos de quadra, de 3 pontos e lances livres
Percentagem de arremessos de quadra efetivos	Estatística ajustada ao fato de que arremessos de 3 pontos valem mais pontos do que arremessos de quadra
Taxa de arremessos de 3 pontos	Percentagem das tentativas de arremessos da faixa de 3 pontos
Taxa de arremessos de quadra	Percentagem das tentativas de arremessos por tentativa de lance livre
Taxa de lance livre	Percentagem das tentativas de lance livre por tentativa de arremesso de quadra
Percentagem de rebotes ofensivos	Estimativa da percentagem de rebotes ofensivos disponíveis recuperados

Atributo	Descrição
Percentagem de rebotes defensivos	Estimativa da percentagem de rebotes defensivos disponíveis recuperados
Percentagem de rebotes totais	Estimativa da percentagem de rebotes disponíveis recuperados
Percentagem de assistências	Estimativa da percentagem de arremessos feitos com assistências
Percentagem de roubos de bola	Estimativa da percentagem de posses do oponente que terminam com um roubo de bola
Percentagem de tocos	Estimativa da percentagem de arremessos de quadra do adversário bloqueados
Percentagem de bolas perdidas	Estimativa de bolas perdidas por 100 posses
Avaliação ofensiva	Estimativa de pontos marcados por 100 posses
Avaliação defensiva	Estimativa de pontos permitidos por 100 posses

Tabela 4: Métricas coletadas em Basketeball-Reference.com.

3.3. PREPARAÇÃO DOS DADOS

3.3.1. Transformação

Todo processo ETL (do inglês, *Extract Transform Load*) foi realizado no Alteryx, que é uma ferramenta com soluções para preparar, combinar, enriquecer, analisar e gerenciar dados. Possui uma interface intuitiva, permitindo uma análise de dados rápida e personalizável sem necessidade de código. O processo é construído em *workflows* que oferecem uma visão geral de todo o desenvolvimento, sendo útil para entender as relações que estão sendo contruídas. Além disso, permite a criação de fluxos de trabalho repetitivos, portanto, o mesmo processo pode ser reutilizado várias vezes.

Para guardar os dados tratados, foi escolhido o formato SQLite devido sua versatilidade, uma vez que o conteúdo do banco de dados SQLite pode ser acedido através de uma ampla variedade de programas e ainda possui um desempenho de leitura e gravação superior a ficheiros individuais. Além disso, o conteúdo pode ser acedido e atualizado usando consultas SQL concisas em vez de rotinas longas e propensas a erros, garantindo que os dados sejam atualizados de forma continua e automática.

3.3.1.1. Variáveis relacionadas ao jogo

Um problema comum em projetos de *machine learning* é a contaminação dos dados. Ao processar os dados que irão alimentar o modelo preditivo, é fundamental garantir que o passado não contenha seu próprio futuro. Quando existem dados que dependem do resultado que se pretende prever, e que, consequentemente, não estarão disponíveis no momento da coleta de dados reais, os algoritmos dão muito peso a estes. Dessa forma, uma previsão aparentemente bem-sucedida dependerá de métricas anacrônicas, tornando esses modelos ineficazes para previsões futuras reais. Para evitar a contaminação de dados, são calculadas as médias dos jogos anteriores para todas variáveis relacionadas ao jogo.

O mando de campo é definitivamente um fator central que afeta o resultado de um jogo. Cerca de 60% dos jogos na NBA são ganhos pela equipa da casa e isso pode ser explicado por várias razões: apoio dos fãs, agenda de viagens, diferença de fuso-horário, etc. Por esse motivo, é necessário ter em consideração o mando de campo no cálculo das médias das variáveis relacionadas ao jogo.

Arkes e Martinez (2011) mostraram que um maior sucesso nos últimos jogos leva a uma maior probabilidade de ganhar os próximos. Da mesma forma, jogar mal nos últimos jogos leva a uma probabilidade menor vitória. Sendo assim, além de calcular as médias das estatísticas acumuladas durante toda a temporada é necessário levar em conta a média das estatísticas nos últimos jogos para capturar o momento da equipa. Esse procedimento, no entanto, deve ser feito com cautela, uma vez que existe uma aleatoriedade alta ao ser analisado uma quantidade pequena de jogos, uma vez que a equipa pode enfrentar adversários de diferentes níveis nesse período.

Enfim, 4 tabelas com mais de 100 estatísticas divididas em ofensivas (da própria equipa) e defensivas (dos oponentes da equipa) são criadas com as seguintes informações:

- Média das estatísticas de todos os jogos da temporada atual, com exceção do jogo atual
- Média das estatísticas dos jogos da temporada atual – com exceção do jogo atual - considerando o mando de campo de cada equipa
- Média das estatísticas dos últimos 7 jogos da temporada atual
- Média das estatísticas dos últimos 7 jogos da temporada atual considerando o mando de campo de cada equipa

3.3.1.2. Variáveis externas

Como o objetivo do trabalho é prever o vencedor dos jogos da NBA, o primeiro passo é avaliar o histórico de vitórias de cada equipa. Para isso, são criadas diversas variáveis com a percentagem de vitórias, entre elas: nos últimos 5, 7 e 10 jogos; no acumulado da temporada; como visitante e como mandante. Além disso, é incorporado o total de vitórias da temporada anterior, que não se trata de uma percentagem para dar peso maior a equipas que participaram da última pós-temporada.

A força do calendário é uma variável que representa a dificuldade média dos adversários enfrentados por cada equipa, calculada através da média da percentagem de vitória dos adversários. Essa variável acaba não sendo tão importante na NBA quanto é em outros despostos, já que todas as equipas da NBA se enfrentam pelo menos duas vezes por temporada. Similar, e mais interessante que a força do calendário, é a força das vitórias que representa a dificuldade média dos adversários vencidos por cada equipa. Ainda existe a margem de vitória que calcula a margem de pontos média de todos jogos vencidos pela equipa.

Por mais que o basquetebol seja um desposto coletivo, um único jogador pode influenciar sozinho o resultado do jogo. A NBA possui uma temporada longa onde lesões, suspensões, dias de descanso e vários outros fatores são frequentes, portanto, é necessário avaliar o talento da equipa disponível em cada jogo. Para isso, é criada uma variável que mostra o valor do elenco apto em determinado jogo com base na lista de jogadores que foram eleitos para o Melhor Quinteto da NBA (*All-NBA Team*) na temporada anterior.

O cansaço de uma equipa é outro fator que impacta diretamente seu desempenho. Existem diferentes maneiras de medir isso, seja pela distância que a equipa viajou ou pelo calendário dos jogos. No presente trabalho são utilizadas as seguintes métricas:

- Distância percorrida para o jogo atual e nos últimos 5 jogos
- Intervalo de dias entre 5 jogos e intervalo de dias entre 3 jogos como mandante para a equipa da casa ou como visitante para a equipa de fora
- Indicador se a equipa disputou um jogo no dia anterior ao jogo atual

A expectativa de Pitágoras foi concebida por Bill James para estimar a provável percentagem de vitórias de uma equipa de beisebol, baseando-se na crença de que a percentagem de vitórias está relacionada às corridas marcadas e permitidas. Daryl Morey (Dewan & Zminda, 1993) adaptou a fórmula ao basquetebol na NBA, conforme Equação 1, onde PS é o número de pontos que a equipa marcou, PSA é o número de pontos que a equipa permitiu e x é um expoente determinado empiricamente considerado igual a 16,5 neste trabalho.

$$\frac{PS^x}{PS^x + PSA^x} \quad (1)$$

O sistema de classificação Elo, criado por Arpad Elo, é um método que surgiu para calcular os níveis dos jogadores de xadrez, mas que pode ser estendido para diversos desportos. Para o presente trabalho, é utilizado a versão proposta por Silver e Fischer-Baum (2015) para a NBA, que possui as seguintes regras:

- As pontuações dependem apenas do resultado de cada jogo e de onde foi disputado.
- As equipas sempre ganham pontos Elo depois de vencer jogos e perdem depois de perdê-los, sendo um sistema de soma zero.
- Vitórias contra equipas com melhores pontuações e por margens maiores rendem mais pontos.

O cálculo do Elo possui um fator K que determina a rapidez com que a pontuação reage aos novos resultados dos jogos. Esse fator deve contabilizar de forma eficiente os novos dados, mas não reagir de forma exagerada, tendo como objetivo minimizar a autocorrelação. O K ótimo para a NBA é 20 (maior que o encontrado em outros desportos), o que implica que o peso dado ao desempenho recente da equipa é relativamente alto. Isso mostra que os resultados da NBA estão sujeitos a uma aleatoriedade relativamente pequena, ou seja, uma sequência de vitórias ou derrotas não acontece ao acaso. Outros parâmetros também são definidos empiricamente: a vantagem da equipa da casa é equivalente a 100 pontos de classificação Elo e de uma temporada para outra as equipas mantêm três quartos da sua pontuação Elo, uma vez que elas costumam ser consistentes de um ano para outro.

Embora as chances postadas por casa de apostas possam não refletir verdadeiras crenças probabilísticas, elas ainda podem ser vistas como bons indicadores das probabilidades do resultado do evento, uma vez que incorporam dados externos até agora não capturados em nenhuma varável. As chances (*odds*) coletadas nesse trabalho encontram-se em formato decimal (e.g. uma equipa com chances de 3,5 dará um retorno, caso vença o jogo, de 3,5 vezes o valor apostado), mas podem ser facilmente convertidas em probabilidades, sendo apenas importante atentar ao fato de que a soma

das probabilidades de ambas as equipas ultrapassa 100%. Esse extra é a margem das casas de apostas, que existe para garantir seus lucros.

3.3.2. Normalização

A normalização dos dados é um requisito comum para muitos algoritmos de *machine learning*, uma vez que diferentes variáveis possuem diferentes escalas de avaliação. Por exemplo, um jogador pode marcar mais de 40 pontos em um jogo, mas não pode cometer mais de 6 faltas pessoais. Para eliminar o efeito de diferentes escalas, os dados são normalizados através de duas técnicas diferentes e seus resultados são comparados.

A normalização por desvio padrão assume que os dados são normalmente distribuídos dentro de cada variável e os dimensiona de forma em que a distribuição seja centralizada em torno de zero, com um desvio padrão unitário. A média e o desvio padrão são calculados para a variável e, em seguida, é realizado a normalização de acordo com a Equação 2. Este método não é indicado para os casos onde os dados não são distribuídos normalmente.

$$\frac{X-\mu}{\sigma} \quad (2)$$

A normalização min-max é provavelmente a técnica de normalização mais conhecida e consiste em reduzir o intervalo de forma que o intervalo das variáveis fique entre 0 e 1 (ou -1 a 1, se houver valores negativos). A normalização é feita de acordo com a Equação 3 e possui bons resultados para casos onde os dados não possuem distribuição normal ou quando o desvio padrão é pequeno. Como desvantagem, a normalização min-max é sensível a *outliers*.

$$\frac{X-Xmin}{Xmax-Xmin} \quad (3)$$

A normalização dos dados foi feita usando os módulos de pré-processamento presente no pacote de Python Scikit-learn, sendo a função StandardScaler utilizado para a normalização por desvio padrão e a função MinMaxScaler para a normalização min-max. A Figura 3 compara o resultado das duas normalizações para um mesmo conjunto de dados.

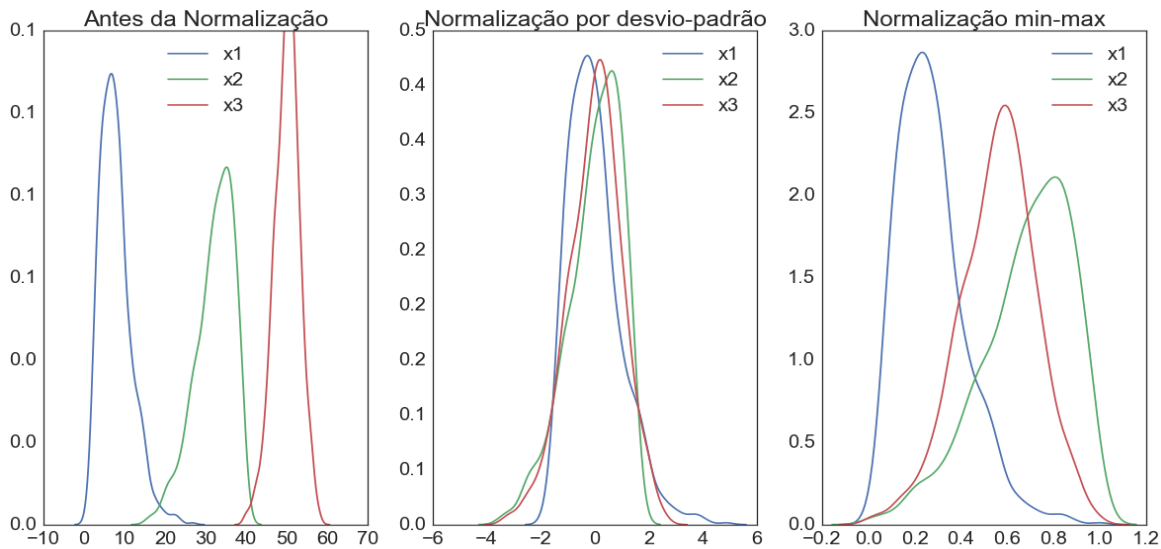

Figura 3: Comparação dos métodos de normalização.

3.4. MODELAGEM DOS DADOS

Todos os componentes da modelagem de dados são implementados na linguagem de programação Python, que ganha cada vez mais destaque no campo da ciência de dados graças a sua simplicidade e seus pacotes para computação científica que tornam a implementação sucinta e eficiente. Durante a modelagem é constantemente utilizado NumPy e Pandas que fornecem ferramentas de análise de dados e estruturas de dados de alta performance. Para a implementar os modelos de *machine learning* é utilizado principalmente o pacote Scikit-learn, no entanto, para a implementação de redes neuronais artificiais é utilizado Keras.

3.4.1. Machine learning

3.4.1.1. Regressão logística

Apesar do nome, regressão logística se trata de um algoritmo de classificação que tem como principal vantagem a alta interpretabilidade dos resultados. A função logística, definida pela Equação 4, tem papel central no algoritmo, uma vez que é aplicada na regressão linear para modelar a probabilidade de um evento ser afetado por uma ou mais variáveis.

$$\sigma(t) = \frac{1}{1+e^{-t}} \qquad (4)$$

A função logística mapeia as entradas entre -∞ e +∞ para valores entre 0 e 1, permitindo que as observações sejam interpretadas como uma probabilidade, conforme Figura 4. Dependendo do tamanho do conjunto de treino, um dos dois métodos para minimizar a perda logística e, consequentemente, definir os parâmetros do modelo é escolhido: gradiente descendente estocástico - método iterativo mais lento e adequado para grandes conjuntos de dados - ou máxima verossimilhança - aproximação numérica mais rápida e ideal para conjuntos de dados menores.

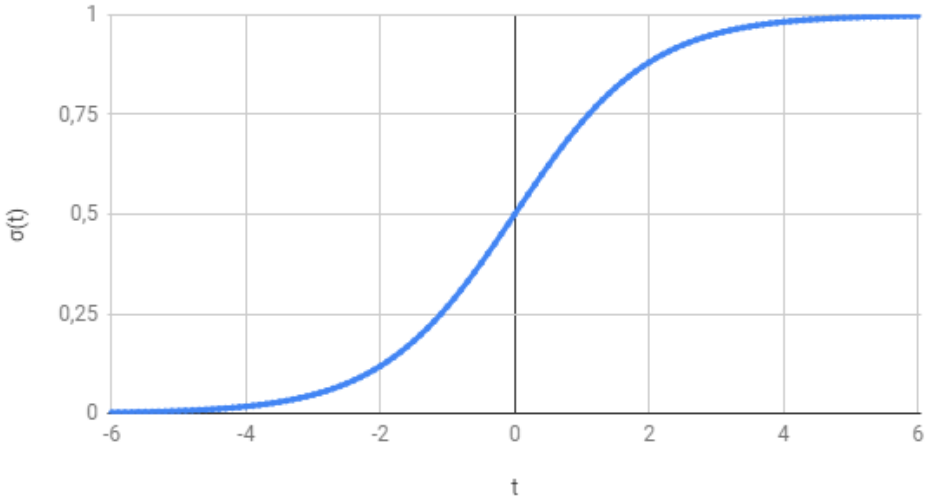

Figura 4: Função logística.

3.4.1.2. Árvore de decisão

Árvores de decisão funcionam dividindo o conjunto inicial de dados em subconjuntos mais homogéneos que por sua vez são divididos em subconjuntos ainda mais homogéneos. As árvores são compostas por nós e folhas, onde o primeiro nó apresenta a variável mais discriminadora do

conjunto de dados e cada nó subsequente representa um teste específico aplicado ao conjunto de dados com o objetivo de dividi-lo em subconjuntos menores e mais homogéneos; as folhas representam o conjunto de dados mais homogéneo que a árvore consegue produzir e que, por isso, não são submetidos a qualquer tipo de divisão. Alguns algoritmos de árvore de decisão produzem apenas árvores binárias (onde cada nó interno se ramifica para exatamente dois outros nós), enquanto outros podem produzir árvores não-binárias.

Uma das maiores complexidades inerentes à construção de uma árvore de decisão é a identificação da variável utilizada em cada nó da árvore. O objetivo é sempre criar a melhor partição, ou seja, aquela onde os nós-filhos pertençam a uma única classe dominante. Existem várias medidas propostas para a definir como deve ser feita essa partição, sendo entropia e coeficiente de Gini as duas mais conhecidas. Apesar de serem medidas diferentes, o objetivo de ambas é identificar a variável que mais discrimine o conjunto de dados.

As etapas de aprendizagem e classificação de árvores da decisão são simples e rápidas, não sendo necessário nenhum conhecimento do assunto. Além disso, a representação do conhecimento adquirido em formas de árvores é intuitiva e de fácil interpretação, conforme exposto na Figura 5. Todos esses fatores a fazem uma ótima ferramenta para análise exploratória dos dados. Outro ponto forte é a aceitação de vários tipos de variáveis (nominais, ordinais e intervalares). Como ponto fraco, árvores de decisão são mais propensas a *overfitting*, uma vez que podem produzir árvores grandes e complicadas que modelam perfeitamente todas as instâncias de treino, mas não generalizam o comportamento real. Várias técnicas podem ser utilizadas para mitigar o *overfitting*, como a poda, que remove alguns nós e folhas mais altos, e a pré-poda, que limita a profundidade da árvore.

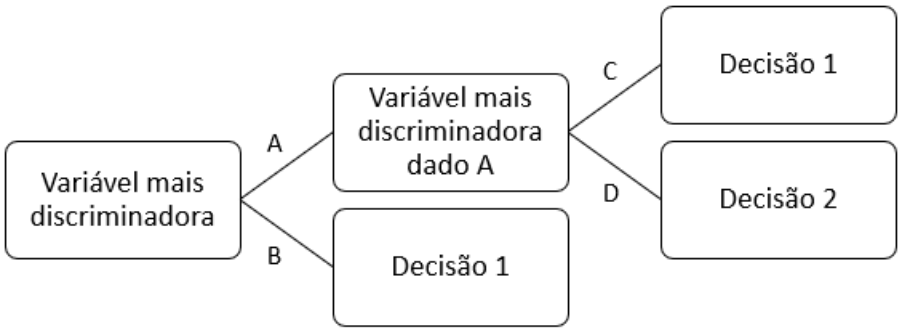

Figura 5: Exemplo de representação dos resultados de uma árvore de decisão.

3.4.1.3. k-vizinhos mais próximos

O algoritmo dos k-vizinhos mais próximos (k-NN, do inglês *k-nearest neighbors*) é um tipo de aprendizagem preguiçosa, onde a função é apenas aproximada localmente e toda a computação é adiada até o momento da classificação. O k-NN é um dos algoritmos mais simples de *machine learning* e funciona da seguinte forma: uma nova observação é classificada pela maioria dos votos de seus vizinhos, senda atribuída à classe mais comum entre os k vizinhos mais próximos, onde k é um inteiro positivo. Também podem ser atribuídos pesos às contribuições dos vizinhos, de modo que os vizinhos mais próximos contribuam mais para a decisão do que os mais distantes. A escolha ótima do valor k depende do conjunto de dados: um k menor leva a limites de decisão ruidosos que podem gerar *overfitting*, enquanto um k maior leva a um excesso de suavidade.

Um exemplo de classificação k-NN é apresentado na Figura 6. A amostra de teste (círculo verde) deve ser classificada para a primeira classe de quadrados azuis ou para a segunda classe de triângulos vermelhos. Se k = 3 (círculo de linha sólida) é atribuído à segunda classe porque há 2 triângulos e apenas 1 quadrado dentro do círculo interno. Se k = 5 (círculo de linha tracejada) é atribuído à primeira classe (3 quadrados contra 2 triângulos dentro do círculo externo).

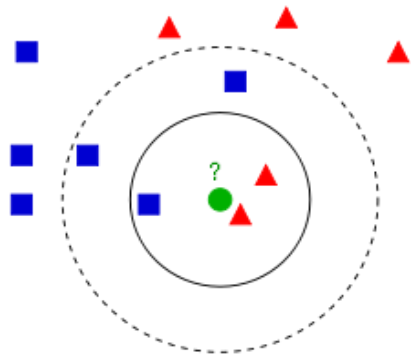

Figura 6: Exemplo de classificação k-NN.

3.4.1.4. Máquina de vetores de suporte

Máquina de vetores de suporte (SVM, do inglês *support vector machines*) é um algoritmo não-linear de aprendizagem supervisionada onde conjunto de treino é representado como pontos no espaço, mapeado de maneira que os exemplos de cada classe sejam divididos por um espaço tão amplo quanto possível. Os novos exemplos são mapeados no mesmo espaço e preditos como pertencentes a uma classe baseado em qual o lado do espaço eles são colocados. Em outras palavras, o que SVM faz é encontrar uma linha de separação, comumente chamada de hiperplano, entre as observações de cada classe. A utilização de funções *kernel* permite ao modelo ganhar flexibilidade e resolver problemas não lineares. Essas funções fazem que os dados originais sejam movidos para um novo espaço, de maior dimensionalidade, conforme mostrado na Figura 7.

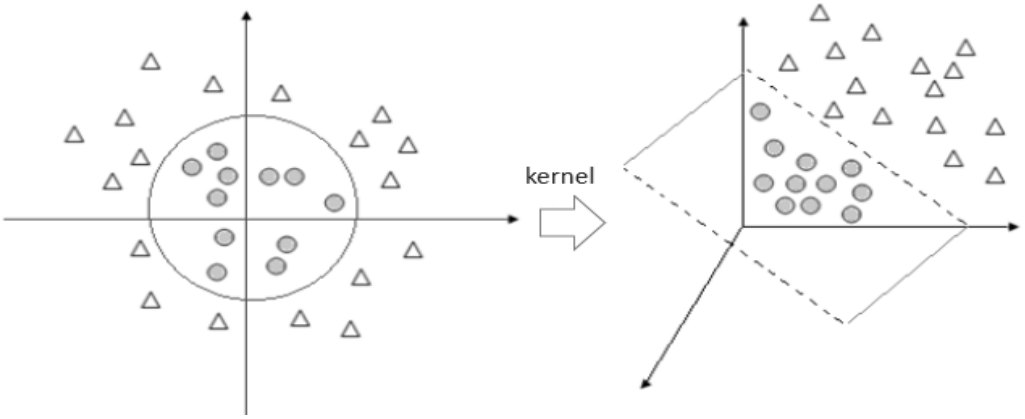

Figura 7: Exemplo de aplicação de uma função *kernel.*

As principais vantagens do SVM são a eficácia em espaços dimensionais elevados, a robustez mesmo quando existe enviesamento no conjunto de treino e a versatilidade, uma vez que diferentes funções *kernel* podem ser especificadas para problemas diferentes. Como desvantagem, o tempo de treinamento para SVMs é elevado em comparação com outros métodos e os modelos tendem a ser difíceis de configurar e interpretar.

3.4.1.5. Redes neuronais

Redes neuronais artificiais (ANN, do inglês *artificial neural networks*) são modelos computacionais inspirados pela forma como o cérebro resolve problemas. ANN são compostas de múltiplos nós, que imitam os neurônios biológicos, conectados por elos, que imitam os axônios biológicos. Assemelham-se ao cérebro humano em outros dois aspetos: o conhecimento é adquirido pela rede através de um processo de aprendizagem e as forças das conexões existentes entre os neurônios são utilizadas para guardar o conhecimento. Uma rede neuronal consiste em várias camadas, com cada neurônio de uma determinada camada sendo conectado a todos os neurônios da camada anterior. A primeira camada é a entrada, a última camada é a saída e todos as camadas intermediárias são camadas ocultas. Uma rede com três camadas é ilustrada na Figura 8.

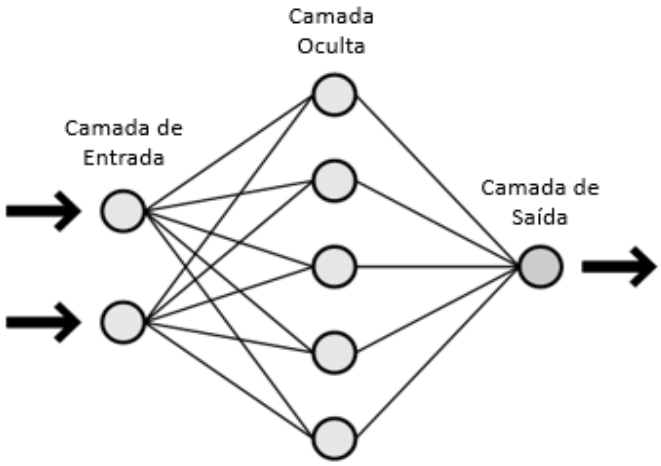

Figura 8: Rede neuronal com três camadas.

ANN funcionam da seguinte forma: cada neurónio calcula um valor a partir de suas entradas e do seu peso associado e a esse valor é aplicado uma função de ativação. O valor final é então passado como uma entrada para os neurónios da camada seguinte, que recebem a soma dos valores produzidos por todas as conexões que chegam até ele. O processo segue até chegar a camada de saída. A função de ativação não-linear permite que pequenas redes neuronais calculem problemas não triviais. A função logística, tangente ou retificadora hiperbólica são comumente usadas como funções de ativação.

ANN conseguem detetar relações complexas entre as várias variáveis de um problema, sendo que uma rede com pelo menos uma camada oculta é capaz de aproximar qualquer função contínua. No entanto, são difíceis de interpretar, sendo conhecidas por "caixa preta", uma vez que a rede treinada não dá nenhuma compreensão adicional do problema. Também são propensas ao *overfitting* e a seleção dos hiperparâmetros do modelo é um processo altamente empírico que requer uma abordagem de tentativa e erro. Apesar dos problemas, durante os últimos anos foram feitos grandes avanços no campo da aprendizagem profunda (*deep learning*), técnica que se baseia em redes neuronais de múltiplas camadas ocultas.

3.4.1.6. Ensembles

Os métodos e*nsemble* são constituídos pela combinação de dois ou mais modelos. As vantagens individuais de cada um dos modelos são destacadas e unidas de forma a melhorar a precisão de classificação do problema em estudo. Erros independentes presentes no conjunto de teste tem sua importância diminuída quando se usam esses métodos. Isso garante bom desempenho em diversos problemas, geralmente superior aos classificadores únicos de que são derivados. Porém, a performance depende da diversidade dos modelos utilizados na criação dos *ensembles*, quanto mais diferente forem os modelos melhor será a performance do método. Em geral, os algoritmos instáveis, que produzem resultados diferentes em várias execuções, são bastante adequados para *ensembles*.

Bagging, é um método de *ensemble* que usa um ponto de partida diferente para construir cada modelo do conjunto. A partir do conjunto de treino, vários subconjuntos são criados, sendo que cada observação é incluída uma única vez no subconjunto, porém pode ser incluída em vários subconjuntos. Assim, uma observação pode constar em todos os subconjuntos ou, pelo contrário, não constar em nenhum. Cada subconjunto de treino origina um modelo que será utilizado na construção do modelo final. Dessa forma, um conjunto de dados diferente é utilizado em cada iteração e cada uma dessas iterações produz um modelo individual para o conjunto. A previsão final do conjunto é feita pela maioria dos votos entre os modelos individuais. Um exemplo de *bagging* é o algoritmo de floresta aleatória que combina diversas árvores de decisão.

Boosting é uma técnica similar, uma vez que é um método de *ensemble* que usa um ponto de partida diferente na criação de cada modelo individual. Porém em *boosting*, a principal característica é a alteração da importância relativa das observações do conjunto de dados. Geralmente, cada observação do conjunto de dados tem o mesmo peso na primeira iteração, mas nas iterações seguintes os pesos das observações classificadas incorretamente aumentam e os pesos das classificadas corretamente diminuem. De uma forma geral, um classificador forte é criado através da combinação de vários classificadores mais fracos. Algoritmos diferem na forma como os pesos são recalculados e na forma como as previsões são combinadas. AdaBoost é a implementação mais popular e, como a maioria dos algoritmos de *boosting,* aplica gradiente descendente para minimizar uma função de custo convexa.

Stacking é um método diferente de *bagging* e *boosting*, no sentido de que não constrói os modelos em si. Seu funcionamento é dividido em duas etapas principais. Primeiro, diversos algoritmos são treinados usando os dados disponíveis. Segundo, um algoritmo combinador é treinado para fazer uma previsão final usando todas as previsões dos outros algoritmos como entradas adicionais. Este procedimento pode ser repetido várias vezes. Dessa forma, *stacking* gera melhores resultados do que qualquer um dos modelos treinados na primeira etapa.

3.4.1.7. Otimização dos hiperparâmetros

Os hiperparâmetros são parâmetros do modelo que o algoritmo de *machine learning* não estima. Por exemplo, para um modelo de k-NN (algoritmo dos vizinhos mais próximos) um dos hiperparâmetros é o número k de vizinhos mais próximos ao exemplo que será considerado para classificação. Os algoritmos possuem valores padrões implementados que geralmente são um bom começo, mas podem não produzir o modelo ideal. *Grid Search* é o método mais comum para selecionar os valores

de hiperparâmetros que produzem o melhor modelo. Neste projeto é utilizado o módulo GridSearchCV implementado em Python através do pacote Scikit-learn.

O funcionamento do *Grid Search* consiste em definir um conjunto de valores possíveis para cada hiperparâmetro que deve ser otimizado e treinar um modelo para cada elemento do produto cartesiano dos conjuntos. Ou seja, é uma pesquisa exaustiva que treina e avalia, usando validação cruzada, um modelo para cada possível combinação dos valores especificados de hiperparâmetro. Mesmo se tratando de um problema paralelo, uma vez que muitos modelos podem ser facilmente treinados e avaliado concorrentemente, esse método é computacionalmente dispendioso até para pequenos conjuntos de valores de hiperparâmetros.

3.4.1.8. Seleção de variáveis

A seleção de variáveis é um processo que busca reduzir a dimensionalidade do conjunto de dados sem sacrificar informações úteis, trazendo diversas vantagens na construção de modelos de *machine learning*:

- simplificação dos modelos tornando-os mais fáceis de interpretar;
- evitar a maldição da dimensionalidade - problema causado pela adição de dimensões extras que faz com que haja um aumento exponencial nos dados necessários para garantir significância estatística;
- menor tempo de treinamento;
- generalização aprimorada pela redução do *overfitting* e da variância.

A premissa central ao usar uma técnica de seleção de variáveis é que os dados contêm algumas variáveis que são redundantes ou irrelevantes e podem, portanto, ser removidos sem resultar em perda de informação. O objetivo da seleção de variáveis é, com base em um algoritmo de aprendizagem e um conjunto de dados, encontrar o melhor subconjunto de variáveis que funcione melhor (de acordo com alguns critérios) com o algoritmo de aprendizagem.

No presente trabalho, esta tarefa é utilizada usando o algoritmo de Eliminação Recursiva de Variáveis (RFE, do inglês *recursive feature elimination*), que seleciona automaticamente as variáveis que mais contribuem para o resultado e elimina aquelas que possuem pouca importância. Iterativamente, todas as variáveis são classificadas e as menos importantes para o modelo são removidas até que o número especificado de variáveis seja atingido. Dessa forma, o algoritmo busca eliminar dependências e co-linearidade que possam existir no modelo. Sua estabilidade depende do tipo de modelo utilizado para classificação e, por se tratar de um algoritmo guloso, nem sempre será encontrado o ótimo global.

O pacote Scikit-learn implementa em Python o algoritmo RFECV para a eliminação recursiva de variáveis associado com validação cruzada, utilizado neste projeto para encontrar o número ideal de variáveis para cada modelo.

3.4.2. Estratégia de apostas

Mesmo que casas de apostas sejam relativamente eficientes ainda há oportunidade de lucros através de um sistema inteligente de apostas. Esse sistema consiste em dois componentes principais: um modelo para estimar a probabilidade de ocorrer determinado evento e uma estratégia de apostas

para, combinando as probabilidades de resultados dados pelo modelo preditivo com as probabilidades da casa de apostas, determinar como será feita a aposta.

Dado a probabilidade de vitória estimada pelas casas de apostas *pC*, a probabilidade de vitória estimada pelo modelo preditivo *pM* e probabilidade real *pR*, diferentes cenários podem existir. Quando $pC < pR$, a casa de apostas subestimou a probabilidade do evento e definiu chances muito altas, criando uma oportunidade para gerar lucro. Por outro lado, quando $pC > pR$, então as probabilidades definidas estão distorcidas em favor da casa de apostas. Portanto, quando $pC < pM$ pode ser assumido que a casa de apostas colocou as probabilidades muito altas devendo ser feita a aposta no evento.

Se o objetivo fosse apenas maximizar o lucro esperado, então a solução seria trivial e todo orçamento seria apostado na opção com o maior retorno esperado de lucro entre as opções disponíveis. No entanto, as apostas são um processo contínuo e deve haver orçamento para as oportunidades seguintes. Por estas razões, para selecionar a estratégia de apostas é considerado o risco juntamento com o lucro esperado. Portanto, a maximização do lucro deve ser feita sempre minimizando o risco.

O valor apostado depende da estratégia escolhida. A estratégia de apostas combina as chances das casas de apostas com a probabilidade do modelo para definir em qual equipa e quanto será apostado em cada oportunidade. Considerando como exemplo as três opções de apostas independentes:

1. Probabilidade de uma vitória da equipa visitante de 0,1 com pagamento igual a 12
2. Probabilidade de uma vitória da equipa visitante de 0,4 com pagamento igual a 3
3. Probabilidade de uma vitória da equipa mandante de 0,8 com pagamento igual a 1,5

Diferentes apostadores farão diferentes apostas com base em suas crenças pessoais e atitude de risco. A opção de apostas 1 contém uma pequena probabilidade de ocorrência compensada por um pagamento potencial relativamente grande, enquanto opção 3 possui um menor pagamento potencial, mas tem uma probabilidade relativamente grande de acontecer. Contudo, um apostador não precisa necessariamente escolher entre as opções de apostas. O orçamento de apostas *C* pode ser dividido em diversas apostas com pesos variáveis c_i.

Cada jogo contém duas opções de apostas possíveis (vitória da equipa mandante ou vitória da equipa visitante), sendo a chance atribuída pelas casas de apostas o_i e a probabilidade resultante do modelo preditivo p_i. Os retornos ou pagamentos potenciais podem ser vistos como uma compensação pela incerteza do resultado e o risco assumido pelo apostador. O peso resultante para as opções de apostas, c_i, será utilizado para dividir o orçamento, *C*, sobre o total de apostas. Para todas as estratégias de apostas, o conjunto de opções de apostas é restrito para aqueles onde o lucro esperado é estritamente positivo.

A primeira estratégia analisada é a dos pesos iguais. Todas as opções de apostas *N* com lucro esperado positivo recebem o mesmo peso (unitário) independentes das probabilidades atribuídas e do retorno correspondente.

Outra estratégia de apostas é a dos pagamentos iguais que atribui pesos às opções de apostas de forma que o potencial retorno seja semelhante em todos os casos. Opções de apostas com uma probabilidade relativamente pequena terão pesos menores, enquanto pesos maiores serão atribuídos a opções com maior probabilidade atribuídas pelo modelo. Assim, a Equação 5 fornece os pesos:

$$c_i = \frac{1}{o_i} \qquad (5)$$

Smith e Preston (1984) propuseram uma estratégia de apostas ajustada pela variância, onde a variância das opções de apostas é levada em conta para obter os pesos ideais. O objetivo é maximizar o lucro esperado enquanto sua variância é minimizada. Os pesos ótimos são obtidos através da Equação 6:

$$c_i = \frac{1}{2o_i(1-p_i)} \qquad (6)$$

A quarta estratégia de apostas é a abordagem proposta de Kelly (1956). Nesta estratégia, o orçamento de apostas C é incorporado para obter os pesos ótimos. Os pesos seguem a Equação 7:

$$c_i = C\,\frac{p_i o_i - 1}{o_i - 1} \qquad (7)$$

Uma desvantagem de todas estas quatro estratégias de apostas é que sempre que o lucro esperado é positivo a aposta é realizada. Em uma situação hipotética onde a probabilidade atribuída pelo modelo é 0,01 e o retorno é de 101, um apostar provavelmente não fará a aposta apesar do lucro esperado ser positivo. Por este motivo, uma restrição adicional é imposta antes de aplicar as estratégias de apostas. O conjunto de opções de apostas será restrito para aqueles com $p_i > \alpha$, onde α é a probabilidade limite para a aposta ser realizada. Diversos valores para α são testados para encontrar o ponto ótimo para as apostas. Concluindo, cada uma das estratégias de apostas é aplicada em um conjunto de opções de apostas com lucro esperado positivo e $p_i > \alpha$.

3.5. AVALIAÇÃO DO RESULTADO

3.5.1. Métricas de avaliação

A avaliação do desempenho de um classificador baseia-se nas contagens de exemplos de teste corretamente e incorretamente previstos pelo mesmo. A matriz de confusão permite uma fácil visualização destes indicadores, sendo uma ferramenta útil para analisar a qualidade do classificador no reconhecimento de exemplos de diferentes classes (Han, Kamber, & Pei, 2011).

Quando um conjunto de dados tem apenas duas classes, é usualmente considerado uma como "positiva" e a outra como "negativa". Dessa forma, as entradas da matriz de confusão são:

- Verdadeiros positivos (*true positives* – TP) refere-se ao número de exemplos da classe "positiva" corretamente previstos como classe "positiva";
- Falsos positivos (*false positives* – FP) refere-se ao número de exemplos da classe "negativa" incorretamente previstos como classe "positiva";

- Verdadeiros negativos (*true negatives* – TN) representa o número de exemplos da classe "negativa" corretamente previstos como classe "negativa";

- Falsos negativos (*false negatives* - FN) representa o número de exemplos da classe "positiva" incorretamente previstos como classe "negativa".

Um classificador com boa precisão possui a maioria dos exemplos representado ao longo da diagonal da matriz de confusão, sendo as demais entradas zero ou próximas a zero. Isto é, idealmente, FP e FN são aproximadamente zero. A Figura 9 representa uma matriz de confusão para um problema de classificação binário.

	Prevista C+	**Prevista C-**
Verdadeira C+	TP	FN
Verdadeira C-	FP	TN

Figura 9: Matriz de confusão para um problema binário.

Diversas métricas podem ser derivadas da matriz de confusão. A precisão de um classificador em um determinado conjunto de teste é a percentagem de casos que são classificadas corretamente pelo classificador. A precisão é mais eficaz quando a distribuição de classes é relativamente equilibrada. A taxa de erro é simplesmente a percentagem de casos que são classificados incorretamente. Em um problema onde a principal classe de interesse é rara, ou seja, a distribuição do conjunto de dados reflete uma maioria significativa da classe negativa e uma classe minoritária positiva, outras medidas devem ser utilizadas, como a sensibilidade e a especificidade. A sensibilidade, também referida como taxa de positivos verdadeiros ou *recall*, é a proporção de exemplos positivos que são corretamente identificados, enquanto a especificidade, também conhecida como taxa de negativos verdadeiros, é a proporção de exemplos negativos que são corretamente identificados. Também é possível utilizar a exatidão, que é uma medida que fornece a percentagem de positivos corretamente previstos sobre o total de positivos previstos. Uma maneira alternativa de usar a exatidão e o *recall* é combinando ambos em uma única métrica, chamada de medida F1. A medida F1 é a média harmônica entre exatidão e ao *recall* tendo ambas igual peso. A Tabela 5 mostra as diversas métricas que podem ser utilizadas em um problema de classificação e como as mesmas podem ser calculadas.

Medida	**Fórmula**
precisão	$\frac{TP + TN}{TP + TN + FP + FN}$
taxa de erro	$\frac{FP + FN}{TP + TN + FP + FN}$
sensibilidade, taxa de positivos verdadeiros, recall	$\frac{TP}{TP + FN}$
especificidade, taxa de negativos verdadeiros	$\frac{TN}{TN + FP}$
exatidão	$\frac{TP}{TP + FP}$
medida F1	$\frac{2 * \text{exatidão} * \text{recall}}{\text{exatidão} + \text{recall}}$

Tabela 5: Medidas de avaliação de modelos.

A curva ROC (do inglês, *receiver operating characteristic*) é uma ferramenta visual útil para comparar diversos modelos de classificação. A curva ROC vêm da teoria de deteção de sinal que foi desenvolvida durante a Segunda Guerra Mundial para a análise de imagens de radar. Para um problema de duas classes, a curva ROC permite visualizar o *trade-off* entre a taxa em que o modelo pode reconhecer corretamente casos positivos versus a taxa de que identifica erroneamente casos negativos como positivos para diferentes partes do conjunto de teste. A área sob a curva ROC também é uma medida da precisão do modelo, sendo útil, uma vez que transforma a curva ROC em um único valor que representa o desempenho esperado do classificador. A Figura 10 apresenta a curva ROC para três modelos hipotéticos.

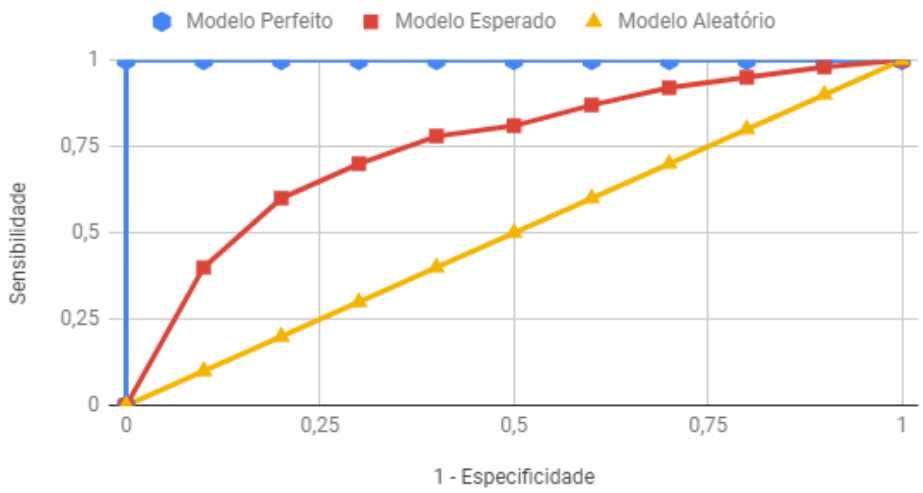

Figura 10: Curva ROC para diferentes modelos.

Além das medidas baseadas nos acertos, os classificadores também podem ser comparados utilizando aspetos adicionais:

- Velocidade: refere-se aos custos computacionais associados ao uso do classificador e na geração do resultado.
- Robustez: reflete a capacidade do classificador de fazer previsões corretas para dados com ruídos e/ou com valores ausentes. A robustez é tipicamente avaliada com uma série de conjuntos de dados sintéticos que representam graus crescentes de ruído e valores ausentes.
- Escalabilidade: relacionada à capacidade de construir o classificador de forma eficiente para qualquer volume de dados. A escalabilidade é tipicamente avaliada com uma série de conjuntos de dados de tamanho crescente.
- Interpretabilidade: refere-se ao nível de compreensão e perceção que é fornecido pelo classificador. A interpretabilidade é subjetiva e, portanto, mais difícil avaliar.

Por fim, o retorno sobre o investimento, também conhecido como ROI (do inglês, *return on investment*), mede o ganho ou a perda gerada em um investimento em relação à quantidade de dinheiro investido. Essa métrica é particularmente apropriada para modelos cujo objetivo seja ganhar tanto dinheiro quanto possível.

3.5.2. Validação cruzada

Validação cruzada é um método para avaliar a capacidade de generalização do modelo. No método de validação cruzado *k-fold* os dados são divididos em k subconjuntos de mesmo tamanho. Em cada iteração, um dos k subconjuntos é utilizado como o conjunto de testes, enquanto os demais subconjuntos formam o conjunto de treino. A medida de desempenho relatada por este método é o erro médio calculado em cada uma das iterações. Em comparação com o método *holdout* (uma parte do conjunto de dados é separado para treino e outra para teste, usualmente na proporção 70-30), essa abordagem é muito mais eficiente em termos da redução do valor da variância da estimativa, que diminui com o aumento do valor de k. Enquanto a principal vantagem é que todas as observações são utilizadas para treino e testes, a desvantagem deste método é que o algoritmo deve ser executado k vezes durante o treinamento, o que significa que é necessário um maior esforço computacional na criação do modelo.

Para o presente trabalho, será utilizado o método *k-fold* para validação cruzada com k igual a 10, devido a baixa variância proporcionada aliado a um custo computacional moderado. A Figura 11 exemplifica o funcionamento do método.

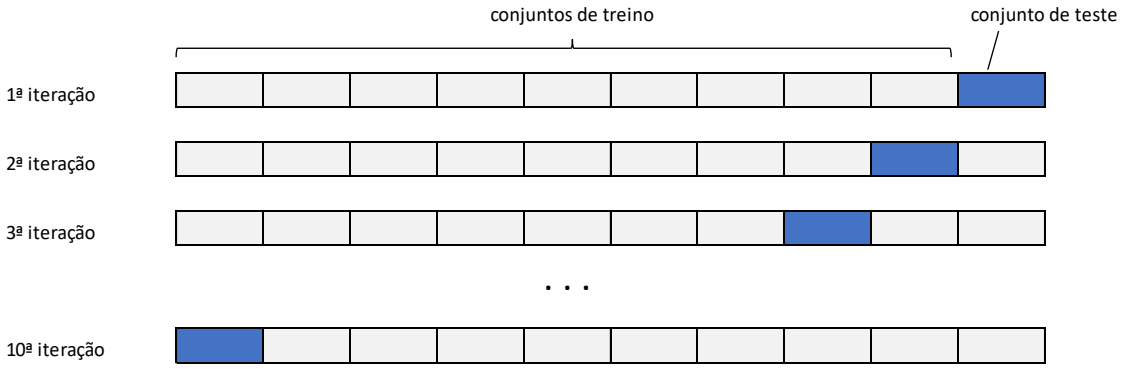

Figura 11: Diagrama para validação cruzada 10-fold.

4. RESULTADOS E DISCUSSÃO

Após definir de forma clara todas as etapas necessárias para um projeto de *machine learning* e aplicá-las com sucesso é chegada a hora de avaliar os resultados. Neste capítulo, o poder preditivo dos algoritmos será discutido e posteriormente será avaliado sua capacidade de obter retorno financeiro através de apostas.

4.1. MACHINE LEARNING

Para a construção de todos os modelos discutidos nesse capítulo foram selecionados jogos apenas das oito últimas temporadas disponíveis (de 2010-11 até 2017-18). A decisão de eliminar dados de temporadas mais antigas, mesmo que estes estivessem disponíveis, foi tomada para garantir que tendências atuais como, por exemplo, o aumento no número de arremessos de três pontos, vide Figura 12, não tenham seu impacto subestimado na previsão do modelo. Também não foram considerados jogos de pós-temporada, por possuírem algumas características diferentes dos jogos da temporada regular, uma vez que as equipas se enfrentam múltiplas vezes seguidas e o caráter eliminatório acrescenta outra dimensão a partida.

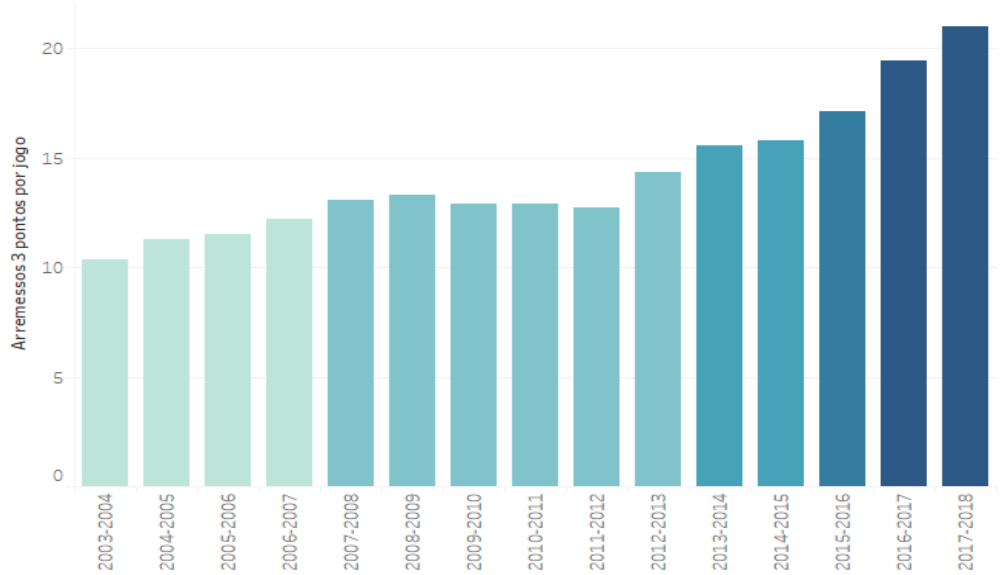

Figura 12: Evolução nos arremessos de 3 pontos ao longo dos anos.

Após o pré-processamento dos dados foram obtidas 1260 variáveis diferentes para cada partida. Esse grande número deve-se a presença de variáveis ofensivas e defensivas para cada uma das equipas e de um rácio entre as mesmas variáveis de cada equipa, criado para possibilitar uma comparação rápida entre as equipas. Com essa grande quantidade de variáveis a criação de modelos eficazes é dificultada por dois motivos: o primeiro é a maldição da dimensionalidade, uma vez que quanto mais variáveis existem no conjunto de dados mais dados são necessários para afirmar que os resultados possuem significância estatística, ou seja, não foram obtidos aleatoriamente; o segundo é relativo ao tempo de processamento, uma que o modelo demora muito mais tempo para ser computado. Para resolver esse problema foram criados diversos modelos usando subconjuntos das variáveis para buscar encontrar as que melhor predizem o resultado da partida.

Com o intuito de entender melhor o conjunto de dados, o primeiro passo foi calcular a correlação de Pearson para cada uma das variáveis. Devido a alta quantidade de variáveis fortemente correlacionadas foram eliminadas todas que possuem uma correlação entre si superior a 90%. Após essa etapa, foi calculado como cada variável se relaciona com as outras com o objetivo de perceber melhor as relações entre cada variável e, principalmente, descobrir quais são as mais relacionadas com o a vitória. A Figura 13 mostra como as 20 variáveis mais relacionadas com a vitória correlacionam-se.

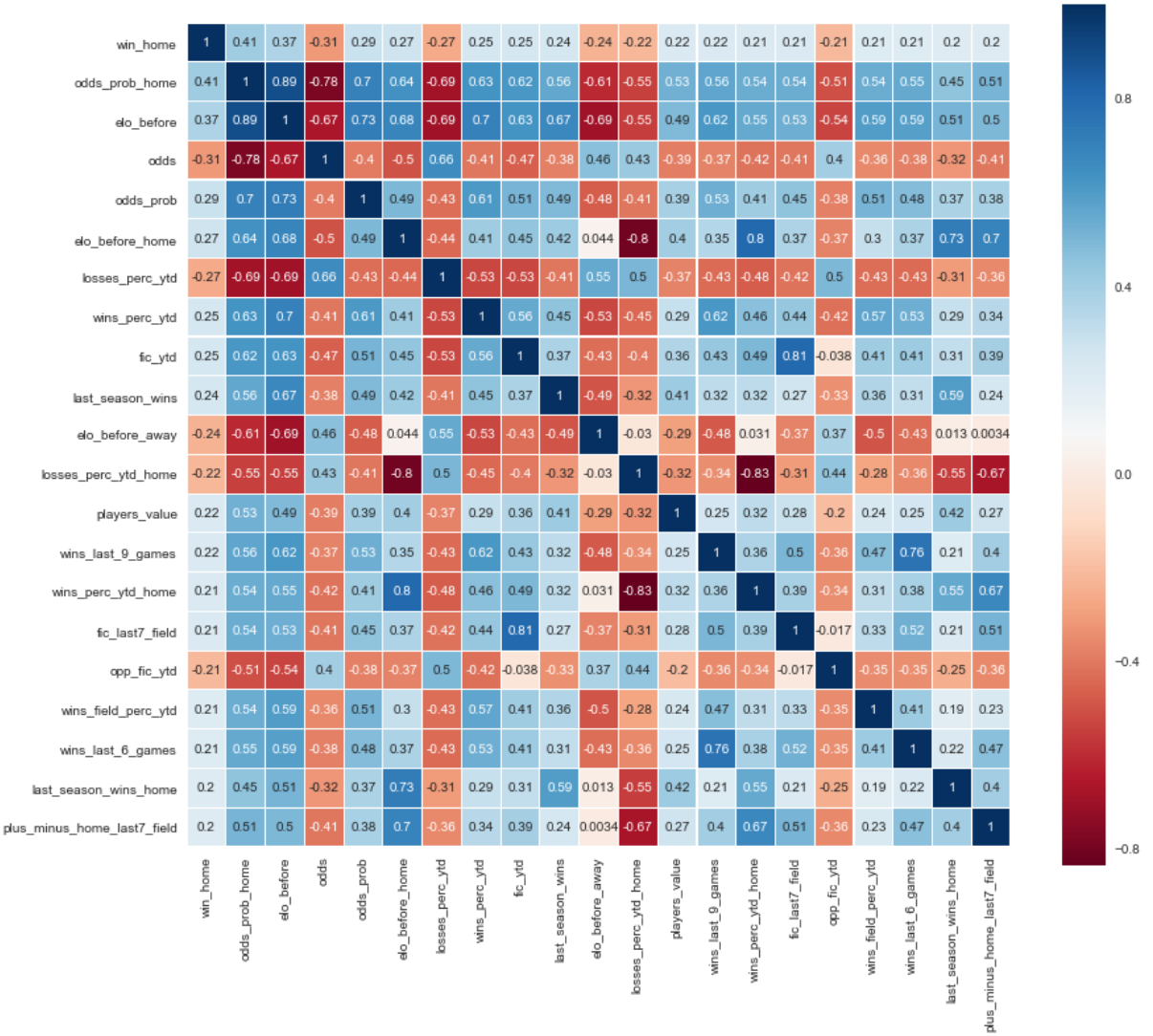

Figura 13: Correlação de Pearson entre as 20 variáveis mais relacionadas com a vitória.

Pode ser visto que as mais relacionadas com a vitória são externas, ou seja, aquelas que não dependem de estatísticas do jogo. O destaque é para as chances calculadas pelas casas de apostas, o que já é um indicativo que será bastante difícil superar seus resultados. Outras variáveis bastante correlacionadas com a vitória são as relacionadas a classificação Elo e com as percentagens de vitórias e derrotas que as equipas tem no acumulado da temporada.

Antes da criação dos modelos, foi realizada a normalização dos dados com a técnica min-max, já que não foram percebidas diferenças quando testes foram feitos outros algoritmos. Além disso,

conforme descrito no capítulo da metodologia, os modelos foram treinados usando validação cruzada com 10 subconjuntos e todos modelos criados passaram por uma otimização dos hiperparâmetros feita através de GridSearchCV.

Para o primeiro experimento foram usadas apenas as 50 variáveis mais correlacionadas com a vitória para a criação de cada modelo, com o objetivo de obter uma visão geral do desempenho de cada um dos 10 diferentes algoritmos que serão testados. A Tabela 6 apresenta os resultados do primeiro experimento para cada modelo avaliado, com modelos ordenados em função da precisão obtida com o conjunto de teste.

Modelo	Precisão de treino	Precisão de teste	3*Desvio-padrão Precisão de teste	Tempo (s)
DecisionTreeClassifier	68.8%	68.6%	2.5%	0.069
LogisticRegressionCV	69.1%	68.2%	2.1%	25.599
XGBClassifier	70.7%	68.2%	2.6%	1.886
MLPClassifier	69.2%	68.2%	2.3%	4.861
SVC	69.2%	68.2%	2.0%	3.325
AdaBoostClassifier	69.4%	68.2%	2.8%	1.512
RandomForestClassifier	68.7%	68.0%	2.4%	0.083
BaggingClassifier	76.1%	67.7%	1.8%	11.577
KNeighborsClassifier	68.4%	66.7%	2.2%	0.035
SGDClassifier	66.7%	65.9%	6.7%	0.028

Tabela 6: Resultados do experimento 1.

Os resultados deste primeiro experimento mostram que a maior parte dos algoritmos possuem resultados muito similares, com destaque para a árvore de decisão que apresentou o melhor resultado aliado a um rápido tempo de processamento. Os resultados dos conjuntos de teste e treino também estão próximos, indicando que não existe *overfit* dos dados. Um ponto relevante a se destacar é a importância dos hiperparâmetros. Quando os modelos estavam com seus hiperparâmetros padrão, os resultados foram inferiores, um exemplo é que sem a otimização dos hiperparâmetros a árvore de decisão foi o pior modelo com precisão de teste inferior a 60% e com grande indicativo de *overfit*, uma vez que a precisão do treino foi de 100%. Conclui-se que, por mais demorado que a otimização dos hiperparâmetros seja, os ganhos trazidos por ela são de grande valia para o sucesso do projeto.

Nesse primeiro experimento também foi utilizado o algoritmo de Eliminação Recursiva de Variáveis para achar o número ótimo de variáveis para cada modelo. Esse algoritmo mostra a precisão do modelo em função do número de variáveis escolhido e também permite identificar as variáveis mais importantes para o modelo. Um exemplo do resultado obtido para um classificador é mostrado na Figura 14. No entanto, os ganhos relacionados a essa técnica foram pequenos e aliado ao fato que nem todos os algoritmos possuem a capacidade de integrarem-se a essa função foi optado por não utilizar essa técnica nos experimentos seguintes.

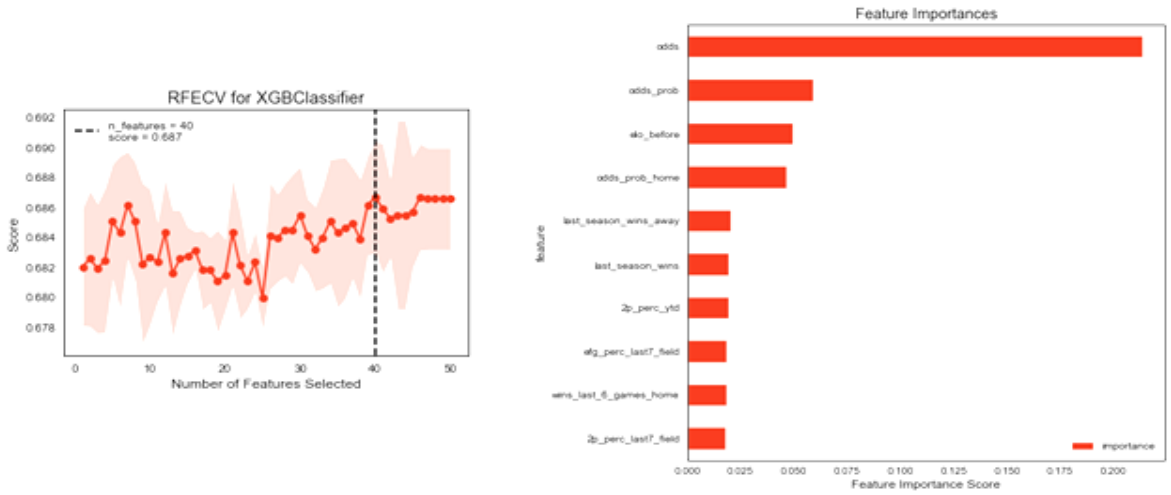

Figura 14: Seleção de variáveis feita com o algoritmo de Eliminação Recursiva de Variáveis.

Na segunda abordagem, foram utilizadas apenas 10 variáveis para criação dos modelos sendo que todas fazem parte dos "quatro fatores do sucesso". Esse experimento teve como objetivo verificar se realmente essas estatísticas conseguem prever de forma eficaz o resultado da partida. Os resultados são mostrados na Tabela 7.

Modelo	Precisão de treino	Precisão de teste	3*Desvio-padrão Precisão de teste	Tempo (s)
XGBClassifier	67.2%	64.2%	2.5%	1.046
AdaBoostClassifier	66.3%	64.2%	2.7%	6.437
BaggingClassifier	74.7%	64.2%	2.0%	5.358
RandomForestClassifier	67.7%	64.0%	2.6%	3.894
LogisticRegressionCV	64.4%	63.9%	3.0%	0.496
MLPClassifier	64.8%	63.9%	2.8%	0.813
SVC	64.4%	63.7%	2.6%	1.895
KNeighborsClassifier	98.8%	63.3%	2.7%	0.012
DecisionTreeClassifier	62.9%	62.1%	1.9%	0.035
SGDClassifier	58.9%	58.8%	14.9%	0.019

Tabela 7: Resultados do experimento 2.

Ao analisar os resultados, percebe-se que novamente os modelos apresentam resultados similares, com exceção do SGD, que mais uma vez apresentou os piores resultados para o problema em questão. Os modelos criados nessa abordagem obtiveram uma precisão de teste inferior em comparação com o primeiro experimento, o que já era esperado, uma vez que apenas 10 variáveis foram usadas, mas mostra que os "quatro fatores do sucesso" possuem um certo poder preditivo.

A terceira abordagem utilizou 24 variáveis relacionadas a partida, sendo composta por estatísticas avançadas e sobre as pontuações de cada equipa. O objetivo desta vez era encontrar algum dado relacionado ao jogo que seja capaz de prever o vencedor. Estatísticas avançadas foram escolhidas por combinarem diversas métricas tradicionais, sendo, dessa forma, mais complexas podendo

possuir algum padrão que leve a detetar o vencedor de determinada partida. Os resultados, apresentados na Tabela 8, foram bastante similares aos do segundo experimento, o que leva a conclusão de que há um limite no poder preditivo ao utilizar apenas variáveis relacionado ao jogo na construção do modelo.

Modelo	Precisão de treino	Precisão de teste	3*Desvio-padrão Precisão de teste	Tempo (s)
XGBClassifier	68.0%	64.8%	3.2%	1.265
BaggingClassifier	86.9%	64.6%	3.0%	19.663
LogisticRegressionCV	65.0%	64.6%	2.2%	1.776
KNeighborsClassifier	98.8%	64.6%	1.8%	0.029
SVC	64.9%	64.4%	2.5%	2.343
MLPClassifier	65.2%	64.2%	2.5%	2.788
RandomForestClassifier	65.4%	64.1%	2.5%	1.802
AdaBoostClassifier	64.5%	63.1%	2.7%	1.176
DecisionTreeClassifier	64.4%	63.0%	2.6%	0.079
SGDClassifier	61.5%	61.4%	13.5%	0.021

Tabela 8: Resultados do experimento 3.

O último experimento utilizando um subconjunto de variáveis foi feito para verificar se os dados externos sozinhos seriam suficientes para prever o resultado de uma partida. Sendo assim, 38 variáveis externas foram utilizadas na construção de cada modelo. Esse experimento mostrou que as variáveis externas são mais importantes que as internas, uma vez que os resultados aqui foram superiores aos apresentados nos experimentos 2 e 3, mas ainda inferiores ao do primeiro experimento. Esse comportamento já era esperado, uma vez que as variáveis mais relacionadas com a vitória eram externas. Os resultados detalhados desse experimento são apresentados na Tabela 9.

Modelo	Precisão de treino	Precisão de teste	3*Desvio-padrão Precisão de teste	Tempo (s)
SVC	67.2%	66.6%	2.0%	2.926
MLPClassifier	67.5%	66.5%	2.6%	12.411
BaggingClassifier	75.8%	66.5%	2.1%	8.926
XGBClassifier	69.6%	66.5%	1.9%	1.052
LogisticRegressionCV	67.3%	66.4%	2.7%	6.629
AdaBoostClassifier	68.0%	66.4%	1.9%	0.796
RandomForestClassifier	99.6%	65.8%	2.5%	5.564
KNeighborsClassifier	67.5%	65.6%	1.7%	0.028
DecisionTreeClassifier	68.3%	65.2%	3.2%	0.086
SGDClassifier	62.8%	62.1%	11.6%	0.013

Tabela 9: Resultados do experimento 4.

Após a realização dos 4 experimentos, todos os 40 modelos anteriores (10 de cada um dos experimentos) serviram de base para um *stacking* visando construir um modelo final que agregue todas as melhores características de cada modelo individual. O *stacking* foi criado usando *bagging* de árvores de decisão. Essa decisão foi tomada por este ser um dos algoritmos que melhor prediz probabilidades, mesmo quando nenhuma calibração é aplicada (Niculescu-Mizil & Caruana, 2005). Probabilidades reais são extremamente importantes já que o objetivo final será usar o resultado do modelo para apostas, portanto saber apenas qual equipa vai vencer é muito pouco para tomar decisões corretas.

O resultado ano-a-ano e geral do modelo é mostrado na Figura 15. Percebe-se que que os resultados não possuem grande variação ao longo dos anos, porém são ligeiramente inferiores nas últimas duas temporadas. O resultado geral é praticamente igual ao melhor modelo do primeiro experimento, o que mostra que não houve grande ganho com a criação dos outros experimentos tampouco com o uso de uma técnica de *stacking*.

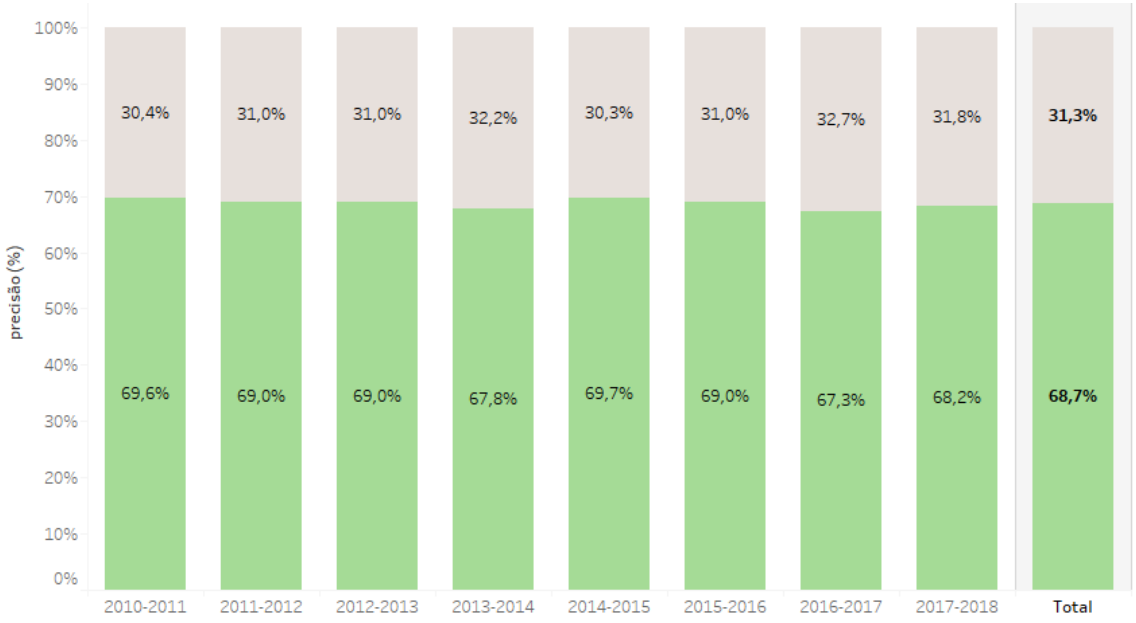

Figura 15: Precisão do modelo final ao longo das temporadas analisadas.

A Figura 16 apresenta a matriz de confusão do modelo, sendo que 1 significa vitória da equipa mandate e 0 vitoria da equipa visitante. A matriz de confusão é particularmente útil, pois a partir dela pode-se ver claramente que o modelo está superestimando o efeito da casa, uma vez que as equipas mandantes vencem aproximadamente 59% dos jogos enquanto o modelo prevê que 68% dos jogos serão vencidos pelo mandante.

	Previsto	
Real	0	1
0	20,6%	20,3%
1	11,0%	48,1%

Figura 16: Matriz de confusão.

O efeito do fator casa é mostrado com mais detalhes na Figura 17. Nos casos em que a equipa mandante vence a precisão do modelo é superior a 80%. Entretanto, quando o visitante vence a precisão cai para apenas 50%, e chega a ser inferior a 44% na temporada 2011-12.

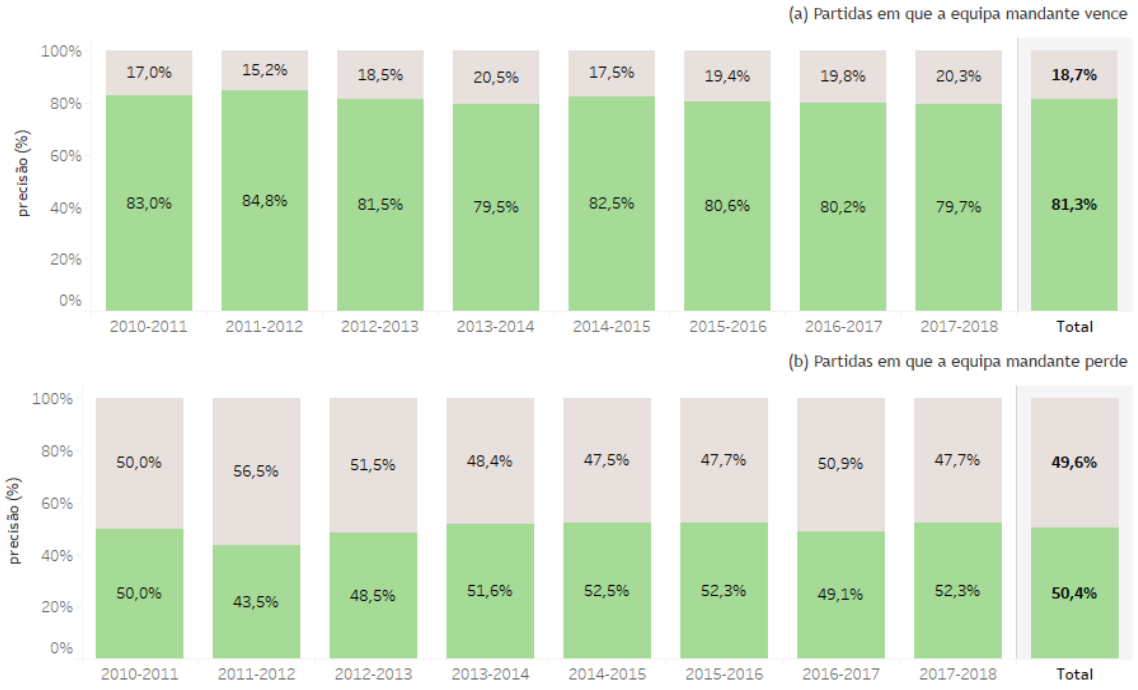

Figura 17: Precisão do modelo em função em função do vencedor.

A margem de vitória também tem uma relação com a previsão do modelo. Em partidas decididas por 10 ou mais pontos a precisão fica próxima a 77%. Entretanto, em partidas decididas por 4 pontos ou menos o modelo tem uma precisão de aproximadamente 55%. Isso é uma evidência que em jogos onde qualquer uma das equipas poderia ganhar (decididos por 4 pontos ou menos) o modelo tem dificuldade de prever corretamente o resultado, porém os resultados são previstos acertadamente em jogos fáceis (decididos por 10 pontos ou mais). Essa constatação é reforçada pelo fato que os jogos em que o modelo acerta o vencedor possuem uma margem de vitória de 12,1 pontos enquanto os que a previsão está errada possuem uma margem de 8,8 pontos. Esses dados são apresentados na Figura 18.

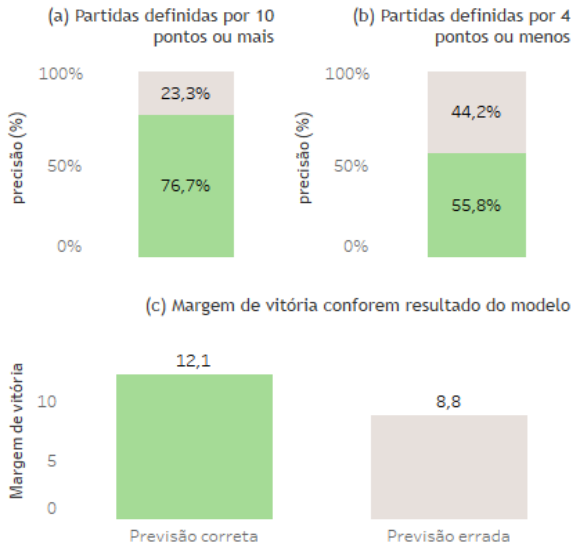

Figura 18: Relação entre a precisão do modelo e a margem de vitória.

4.2. APOSTAS

Antes de proceder com as apostas é necessário avaliar a distribuição de probabilidades atribuída pelo modelo e sua precisão. A Figura 19 mostra que a precisão do modelo cresce conforme aumenta sua certeza no resultado, com exceção das partidas onde a probabilidade prevista está entre 65 e 70%, uma vez que nessa faixa, a precisão do modelo ficou abaixo do esperado. A distribuição de probabilidades também mostra que, na maior parte dos jogos, a equipa mandante tem mais de 70% de probabilidade de vitória. Além disso, não existem partidas que uma equipa tenha mais de 80% de chance de ganhar.

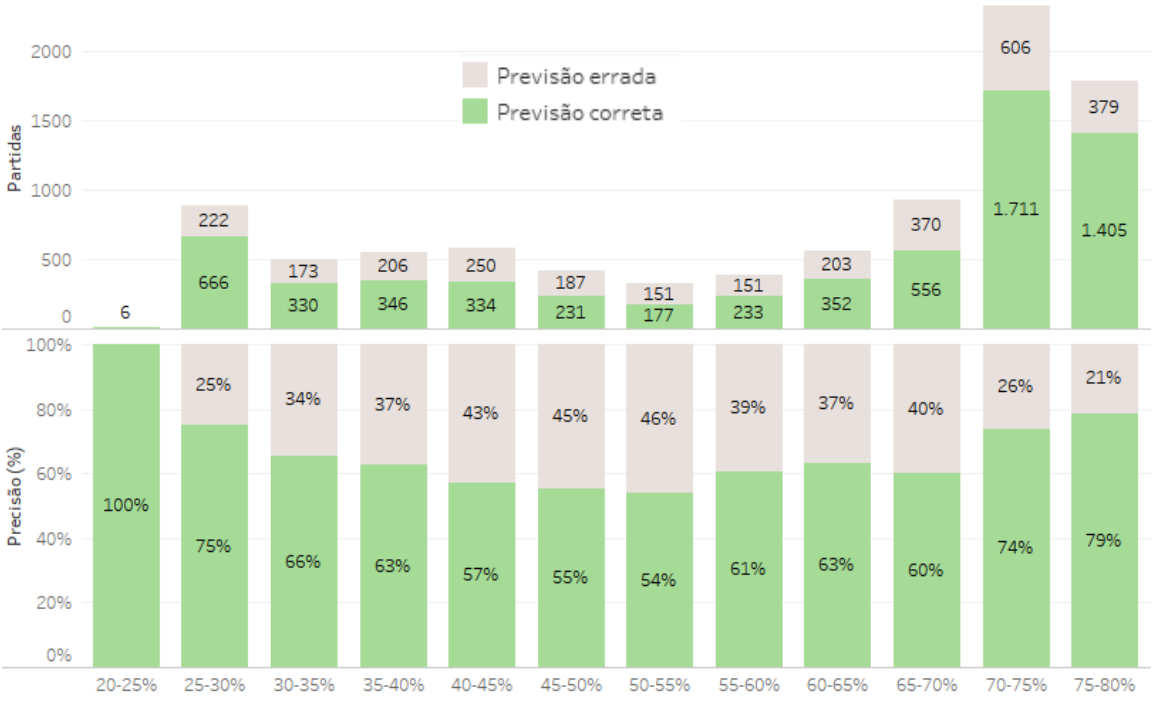

Figura 19: Precisão do modelo conforme probabilidade prevista.

Conforme detalhado na metodologia, foram testadas 4 diferentes estratégias de apostas:

- Estratégia 1: apostas iguais;
- Estratégia 2: pagamentos iguais;
- Estratégia 3: ajustada pela variância;
- Estratégia 4: critério de Kelly.

Também foram testados 4 diferentes cenários:

- Cenário 1: apostar sempre que o retorno esperado seja superior a 100%;
- Cenário 2: apostar sempre que o retorno esperado seja superior a 110%;
- Cenário 3: apostar sempre que o retorno esperado seja superior a 110% e que a probabilidade de vitória do mandante seja superior a 70% (para apostas na equipa mandante) ou que a probabilidade de vitória do visitante seja superior a 30% (para apostas na equipa visitante);
- Cenário 4: apostar sempre que o retorno esperado seja superior a 110% e que a probabilidade de vitória do visitante seja superior a 30% (para apostas na equipa visitante). Nesse cenário não é apostado na vitória da equipa mandate.

Os resultados de todas estratégias e cenários são mostrados na Figura 20. Conforme esperado, quanto menor é o número de jogos em que uma aposta é realizada, maiores foram os retornos financeiros, calculados através do ROI. Em todos os cenários o critério de Kelly apresentou os melhores resultados, chegando a obter um retorno superior a 4% em um dos cenários. Para todos os casos foram comparados os pagamentos de 3 casas de apostas e, para cada equipa/partida, foi escolhido o que apresentava o maior pagamento, de forma a maximizar o lucro potencial de cada partida.

ROI (%)	Cenário 1	Cenário 2	Cenário 3	Cenário 4
Estratégia 1	-1,7%	-0,4%	2,8%	3,8%
Estratégia 2	-1,5%	0,3%	2,4%	3,6%
Estratégia 3	-1,6%	0,5%	2,1%	3,6%
Estratégia 4	-0,4%	0,6%	2,9%	4,2%
Partidas Apostadas	92,2%	54,3%	17,5%	9,4%

Figura 20: Retorno sobre o investimento.

A Figura 21 apresenta o ROI ao longo do tempo. Percebe-se que as estratégias geralmente começam distantes uma das outras e com grandes oscilações em seus retornos, mas ao passar dos jogos estabilizam-se e acabar por terminar com resultados próximos uma das outras. O cenário 4, apesar de apresentar o melhor resultado, apresenta a maior variação. Em alguns momentos o retorno

ultrapassa os 10%, mas também possui períodos com perdas de 10%. Essa grande variação pode ser explicada pelo baixo número de jogos desde cenário (menos de 10% do total).

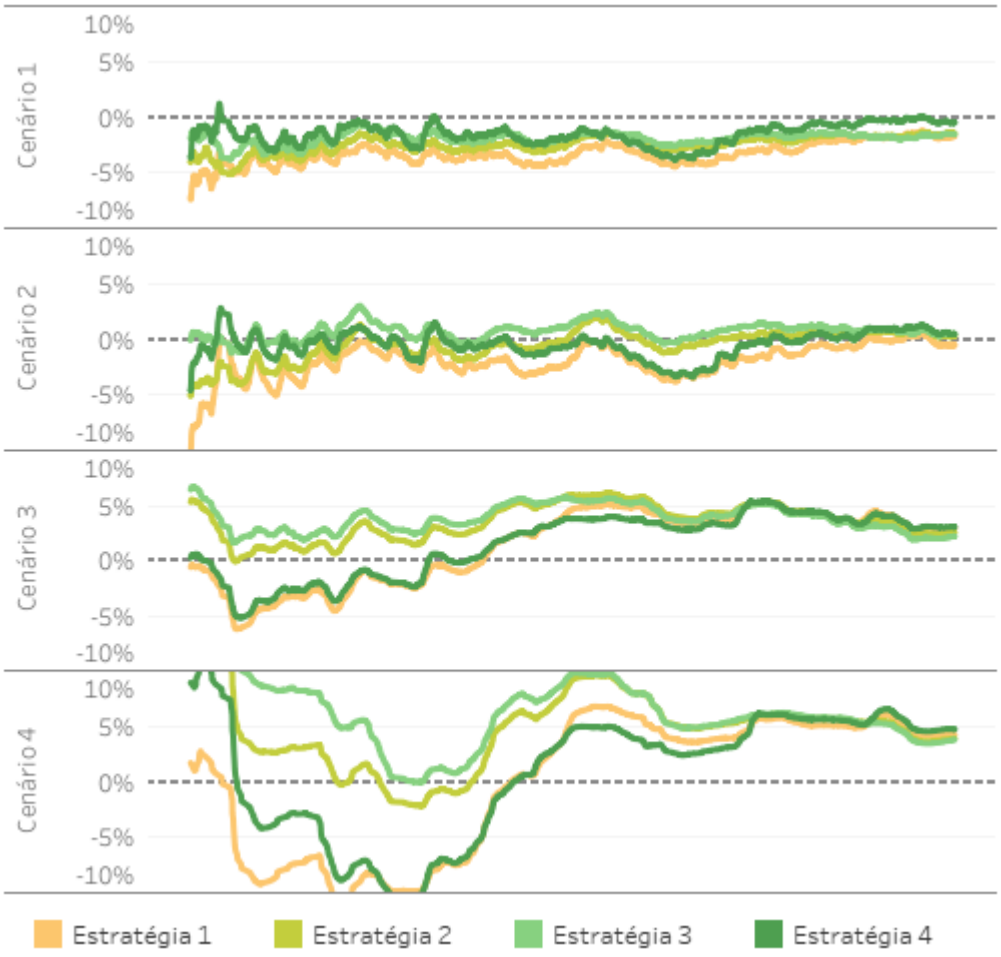

Figura 21: Evolução do retorno sobre o investimento.

Após finalizar os experimentos, pode-se concluir que vencer as casas de apostas em termos de precisão é muito difícil e, uma vez que o resultado do modelo é altamente correlacionado com os pagamentos das casas de aposta, o lucro gerado é muito baixo. No entanto, apenas o fato do modelo ser lucrativo em certos cenários mostra que existem oportunidades para explorar o mercado de apostas.

5. CONCLUSÕES

Aproveitando a crescente disponibilidade de dados sobre partidas desportivas, associada ao apelo que existe na previsão dos resultados destes eventos, este projeto explora a aplicação de técnicas de *machine learning* para prever resultados de partidas da NBA. O projeto é construído a partir da coleta de dados estatísticos, passando pela criação de novas variáveis e treinamento de modelos preditivos, com o objetivo final de explorar as ineficiências do mercado de apostas, ou seja, gerar lucro através de apostas.

Na revisão da literatura é possível ter uma visão geral sobre como funciona o mundo das apostas e também perceber as diferentes aplicações para *machine learning*. Após essa introdução, são explorados diversos estudos feitos para prever resultados desportivos. Por fim, a parte mais importante deste capítulo é sobre as previsões no basquete, uma vez que permite compreender alguns dos diferentes pontos de vista sobre os quais esse tópico pode ser abordado e também serve como *benchmark* para este projeto.

A parte experimental foi realizada em seis grandes etapas: compreensão do assunto, compreensão dos dados, preparação dos dados, modelagem, avaliação do resultado e implementação. A compreensão do assunto foi feita através de leituras e análises sobre jogos de basquete. O conjunto de dados, com estatísticas das equipas e de jogadores, foi extraído do site Basketball-Reference.com e suas principais características foram explorados com o auxílio do Alteryx, *software* que também serviu como ferramenta para preparação dos dados. Nessa etapa, diversas variáveis foram criadas buscando encontrar características que explicassem o resultado de uma partida. Depois que a qualidade do conjunto de dados foi garantida, foi possível prosseguir para a modelagem dos dados, implementada usando a biblioteca Scikit-learn de Python. Por fim, os resultados do modelo e das simulações de apostas foram avaliados através de diferentes métricas.

Mesmo que o problema de prever o vencedor de uma partida desportiva possa parecer aleatório, os resultados alcançados foram satisfatórios. O modelo final, um *stacking* do resultado de 40 modelos intermediários criados a partir de diferentes variáveis, obteve uma precisão de aproximadamente 69%. Nesse sentido, a precisão foi bastante superior a escolha ao acaso da equipa vencedora, associada a uma precisão de 50%. O modelo também foi mais eficaz que escolher a equipa da casa sempre como vencedora, precisão de 59%. O resultado também está alinhado com estudos presente na literatura, que geralmente conseguem uma precisão próxima a 70%.

Os resultados também estão de acordo com pesquisas anteriores que mostram que é difícil de superar as casas de apostas em termos de precisão. No entanto, a precisão não é uma medida direta de lucro tanto que, ao comparar várias estratégias de apostas em diferentes cenários, é possível obter um retorno sobre o investimento de até 4%. Para aumentar a lucratividade é necessário otimizar as escolhas de apostas e não apostar quando a confiança do modelo no resultado é baixa. Ser mais conservador ao fazer apostas acaba sendo uma solução viável para gerar lucro mesmo quando o resultado do modelo está correlacionado com os pagamentos das casas de apostas. Além disso, estratégias de apostas mais complexas, como o critério de Kelly, são consistentemente mais lucrativas do que estratégias mais básicas.

5.1. Limitações e Recomendações para Trabalhos Futuros

O comportamento humano, incluindo o jogo de basquete, é uma ciência inexata com muita variância. Por mais que, como em qualquer outro desporto, os resultados dos jogos geralmente tenham uma relação próxima com a qualidade dos jogadores também existem muitos fatores incertos que vão além das táticas e estatísticas, como: erros inesperados de jogadores e árbitros, lesões, cansaço e, claro, a sorte. Esses elementos imensuráveis são difíceis de quantificar de maneira que não podem ser colocados em um modelo matemático, dessa forma nenhuma previsão consegue ser totalmente precisa.

Uma maneira alternativa para obter melhores resultados é considerar um nível de granularidade diferente e encontrar maneiras de combinar dados ao nível de jogador com dados ao nível da equipa para verificar o impacto de algumas características do jogador nos resultados. Alguns exemplos de dados úteis são: idade ou senioridade, características físicas como peso, altura ou envergadura, a eficiência do jogador e comparações entre jogadores. Mesmo assim, esses novos dados dificilmente conseguirão codificar características emocionais, como liderança, motivação e destreza.

Construir novas variáveis que incluem informações adicionais sempre podem melhorar o poder preditivo, portanto, além de variáveis que representem a qualidades dos jogadores, existe a possibilidade de incluir as que representem as condições do jogo. Altitude da cidade, temperatura do dia, fuso horário, quantidade de torcedores e cobrança por resultados são mais variáveis que podem contribuir de alguma forma para o resultado do modelo.

Algumas melhorias também podem ser implementadas com objetivo de aumentar o ROI. A primeira é obter uma precisão ainda maior no modelo, que pode ser atingido com a inclusão de novas variáveis, como as já descritas nessa secção, mas também com o uso de novos algoritmos de *machine learning*, como *deep learning*. Uma segunda forma de obter melhores resultados financeiros é prever com precisão os jogos em que a casa de apostas falha, que poder ser feito aumentando a importância destes jogos no momento da criação do modelo. Por fim, uma função de perda personalizada - com o objetivo maximizar o lucro em vez de otimizar a precisão da previsão - pode aumentar os retornos financeiros.

O escopo deste projeto ficou limitado às partidas da NBA, devido a uma maior disponibilidade de dados. Ainda assim, todo código é genérico o suficiente para acomodar previsões para as partidas da WNBA, liga feminina da NBA. No entanto, como diferentes características podem ser relevantes para as mulheres, essa mudança exigirá recalibração e reavaliação dos modelos criados.

6. BIBLIOGRAFIA

Arkes, J., & Martinez, J. (2011). Finally, Evidence for a Momentum Effect in the NBA. *Journal of Quantitative Analysis in Sports*.

Baghal, T. (2012). Are the "Four Factors" Indicators of One Factor? An Application of Structural Equation Modeling Methodology to NBA Data in Prediction of Winning Percentage. *Journal of Quantitative Analysis in Sports*.

Bailey, M. J. (2005). *Predicting sporting outcomes: A statistical approach.* Doctoral dissertation, Swinburne University of Technology, Faculty of Life and Social Sciences.

Bucheli, H., & Thompson, W. (2014). Statistics and Machine Learning at Scale: New Technologies Apply Machine Learning to Big Data. *Insights From the Analytics 2014 Conference.* SAS.

Bunker, R. P., & Thabtah, F. (2017). A machine learning framework for sport result prediction. *Applied Computing and Informatics*.

Cheng, G., Zhang, Z., Kyebambe, M. N., & Kimbugwe, N. (2016). Predicting the Outcome of NBA Playoffs Based on the Maximum Entropy Principle. *Entropy*.

Dewan, J., & Zminda, D. (1993). *STATS Basketball Scoreboard.* Chicago: STATS, Inc.

Dias, A. C. (2016). *Live Betting Markets Efficiency: the NBA case.* Master's thesis, Instituto Universitário de Lisboa.

Dubbs, A. (2016). Statistics-free sports prediction. *Model Assisted Statistics and Applications* , 173-181.

Fama, E. F. (1970). Efficient capital markets: a review of theory and empirical work. *Journal of Finance*, 383-423.

Fonseca, J. G. (2018). *March madness prediction using machine learning techniques.* Master's thesis, Universidade Nova de Lisboa.

Fromal, A. (13 de Maio de 2018). *From Past to Present: The Legends Who Influenced Today's NBA*. Obtido de Bleach Report: http://bleacherreport.com/articles/2549379-from-past-to-present-the-legends-who-influenced-todays-nba

Han, J., Kamber, M., & Pei, J. (2011). *Data mining: concepts and techniques.* Waltham: Morgan Kaufmann.

Hayashi, A. M. (2001). When to trust your gut. *Harvard business review, 79*(2), 59-65.

Ibanez, S. J., Sampaio, J., Feu, S., Lorenzo, A., Gomez, M. A., & Ortega, E. (2008). Basketball game-related statistics that discriminate between teams' season-long success. *European Journal of Sport Science* , 369-372.

Jacobs, J. (5 de Setembro de 2017). *Introduction to Oliver's Four Factors*. Obtido de Squared Statistics: https://squared2020.com/2017/09/05/introduction-to-olivers-four-factors/

Kelly, J. (1956). A new interpretation of information rate. *IRE Transactions on Information Theory*, 185–189 .

Kotzias, K. (9 de Março de 2018). *The Four Factors of Basketball as a Measure of Success*. Obtido de Statathlon: https://statathlon.com/four-factors-basketball-success/

Krasnoff, L. S. (15 de Maio de 2018). *How the NBA went global*. Obtido de The Washington Post: https://www.washingtonpost.com/news/made-by-history/wp/2017/12/26/how-the-nba-went-global

Kubatko, J., Oliver, D., Pelton, K., & Rosenbaum, D. T. (2007). A Starting Point for Analyzing Basketball. *Journal of Quantitative Analysis in Sports*.

Kyupers, T. (2000). Information and efficiency: an empirical study of a fixed odds betting. *Applied economics*, 1353-1363.

Lin, R. (2017). *Mason: Real-time NBA Matches Outcome Prediction.* Master's thesis, Arizona State University.

Makropoulou, V., & Markellos, R. N. (2011). Optimal Price Setting in Fixed-Odds Betting Markets Under Information Uncertainty. *Scottish Journal of Political Economy*, 519-536.

Mitchell, T. M. (1997). *Machine Learning.* New York: McGraw-Hill.

Niculescu-Mizil, A., & Caruana, R. (2005). Predicting Good Probabilities With Supervised Learning. *Proceedings of the 22nd international conference on Machine learning* (pp. 625-632). ACM.

Oliver, D. (2004). *Basketball on Paper.* Washington, DC: Brassey's.

Park, J. (2014). *The prediction of outcomes in the National Basketball Association.* Doctoral dissertation, RMIT University, School of Mathematical and Geospatial Sciences.

Pinnacle. (10 de Junho de 2016). *How bookmakers make money*. Obtido de Pinnacle: https://www.pinnacle.com/en/betting-articles/Betting-Strategy/how-bookmakers-make-money/

Praet, R. (2017). *Techniques, Predicting Sport Results by using Recommendation.* Master's thesis, Ghent University, Department of Information Technology.

Puranmalka, K. (2013). *Modelling the NBA to Make Better Predictions.* Master's thesis, Massachusetts Institute of Technology, Master of Engineering in Computer Science and Engineering.

Rybaltowski, M. (13 de Maio de 2018). *Gaming Industry Execs Question Feasibility Of NBA's Proposed Integrity Fee*. Obtido de Forbes: https://www.forbes.com/sites/mattrybaltowski/2018/01/25/gaming-industry-execs-question-feasibility-of-nbas-proposed-integrity-fee/

Sillanpää, V., & Heino, O. (2013). *Forecasting football match results - A study on modeling principles and efficiency of fixed-odds betting markets in football.* Master's thesis, Aalto University, Department of Information and Service Economy.

Silver, N., & Fischer-Baum, R. (21 de Maio de 2015). *How We Calculate NBA Elo Ratings*. Obtido de FiveThirtyEight: https://fivethirtyeight.com/features/how-we-calculate-nba-elo-ratings/

Smith, M., Paton, D., & Vaughan, W. L. (2006). Market efficiency in person-to-person betting. *Economica*, 673-689.

Smith, R., & Preston, F. (1984). Vocabularies of motives for gambling behavior. *Sociological Perspectives*, 325–348 .

Tran, T. (2016). *Predicting NBA games with matrix factorization.* Master's thesis, Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Science.