FUNDAÇÃO GETÚLIO VARGAS
ESCOLA DE ADMINISTRAÇÃO DE EMPRESAS DE SÃO PAULO

Programa Institucional de Bolsas de Iniciação Científica (PIBIC)

Título: Cálculo do valor atribuído pelo público à performance de jogadores de futebol

Modelo preditivo dos melhores jogadores de acordo com votação popular em cada rodada do Campeonato Brasileiro de 2019

CAIO MENESES STABEL

LILIAN SOARES PEREIRA DE CARVALHO

São Paulo – SP

2020

Cálculo do valor atribuído pelo público à performance de jogadores de futebol

Modelo preditivo dos melhores jogadores de acordo com votação popular em cada rodada do Campeonato Brasileiro de 2019

Desde a ascensão da filosofia de Moneyball no baseball dos Estados Unidos no início dos anos 2000, o uso de estatística aplicada à esportes vem ganhando espaço em diversas modalidades. Especialmente do ponto de vista gerencial, a adoção de métodos quantitativos na gestão de equipes esportivas, sobretudo com foco na avaliação de performance de atletas, revolucionou a forma de competir. No futebol, sobretudo no Brasil, esse movimento ainda está no início e há muito a se descobrir. O objetivo desse estudo é a criação de uma nova métrica, capaz de medir o valor percebido da atuação de um jogador para os expectadores. Para tal, foram desenvolvidos modelos de random forest, regressão logística e *support vector machine* para prever a probabilidade de cada jogador ser eleito na votação popular online para a "Seleção da Rodada", que premia os melhores atletas de cada posição que atuaram nos jogos da rodada em questão. Um *ensemble* dos modelos desenvolvidos foi utilizado para unificar os resultados em uma única métrica, que representa um *score* para a atuação do jogador em termos de valor percebido pelo público. Através do uso dessa métrica, clubes de futebol podem otimizar suas receitas através de melhores investimentos em contratações e oportunidades de ganhos incrementais em receitas de bilheteria, bem como obter uma melhor compreensão do valor gerado aos seus torcedores, permitindo uma gestão voltada à satisfação de seu consumidor final.

Palavras-chaves

Esportes, Futebol, Data Science, Valor Percebido, Moneyball

1. Introdução

No meio esportivo, a avaliação de performance de atletas está constantemente envolvida com questões de cunho subjetivo como aspectos físicos, emocionais e traços de personalidade, bem como elaborações menos precisas, vagamente classificadas como talento ou habilidade. No entanto, a compreensão dos componentes do talento e da efetividade de um atleta reside em aspectos, comumente ignorados pelas pessoas do meio. As variáveis qualitativas percebidas como indicativos de performance por torcedores e muitos profissionais do esporte costumam ser resultados desses dados que não são observáveis a olho nu (LEHMANN e SCHULZE, 2008). Dessa forma, em muitos esportes observa-se uma supervalorização de determinados atributos de atletas sem a real compreensão de suas origens, gerando distorções em avaliações e valores de mercado (SILVER, 2013).

Historicamente, o beisebol foi o esporte que primeiro ofereceu as condições para o desenvolvimento e aperfeiçoamento de modelos estatísticos de avaliação de performance (LEWIS, 2015). Desde os artigos de Bill James publicados no final da década de 1970 - nos quais o autor buscava compreender de maneira objetiva a origem do valor atribuído aos jogadores - a ciência começou a ganhar espaço entre jornalistas e torcedores (JAMES, 2014). A grande barreira a ser vencida era a penetração do conhecimento na tomada de decisão de times profissionais, os quais propagavam um grande criticismo à aplicação dos novos conhecimentos, encarando-os como uma ameaça ao próprio esporte como conheciam. Somente no começo dos anos 2000 que essa barreira foi rompida, com a direção do time Oakland Athletics estruturando sua estratégia na exploração de ineficiências de mercado geradas pelas falhas de avaliação que seus concorrentes apresentavam, obtendo resultados notáveis ao alcançar a liderança de sua liga em vitórias com uma das menores folhas salariais da época (LEWIS, 2015).

O beisebol possui diversas particularidades que o tornaram mais propício para seu papel de pioneiro na aplicação de estatísticas na análise de jogadores, times e partidas. Trata-se de um esporte onde os eventos seguem uma sequência definida e ordenada, além de haver registros desde o século XIX (SILVER, 2013). No futebol, por outro lado, os padrões observados não possuem a mesma ordem o que traz maiores dificuldades na obtenção de dados. Os principais avanços no campo se dão com o ascensão de novas tecnologias que passam a permitir o registro de eventos a partir de análise automatizada de imagens e considerando informações espaciais e

de localização, a exemplo do que houve no beisebol, onde tais tecnologias revolucionaram os estudos de ações defensivas (LEWIS, 2015)

Os primeiros estudos sobre o futebol datam do início do século 20, porém só no final dos anos 80 que começou a se observar um crescimento da área (LINK, 2018). Na época, a maioria das pesquisas buscavam analisar diferentes estilos de futebol e compreender qual seria mais o efetivo. Nesse contexto surgem diversos conceitos que se tornaram enraizados na cultura de algumas escolas de futebol, sobretudo o futebol inglês, tendo influência em alguns treinadores até hoje. Os principais exemplos são os estilos de posse de bola e de futebol direto, os quais foram temas de diversos estudos como "Football chance: tactics and strategy" (BATE, 1988) e "Comparison of patterns of play of successful and unsuccessful teams in the 1986 World Cup for soccer" (HUGHES, 1988), entre outros.

Desde o início dos anos 2000, a quantidade e variedade de estudos no campo de analytics aplicada à futebol vem crescendo em parte devido ao sucesso de Moneyball que elevou o campo de estatísticas nos esportes (LINK, 2018). Existe uma grande gama de aplicações do conhecimento e consequentemente de temas de pesquisa, como por exemplo, modelos de previsão de resultados (sendo muitos criados com foco na busca por ineficiências no mercado de apostas) (DIXON e COLES, 1997) , modelos de análise de treinos (para otimizar resultados e prevenir lesões) (LITTLE, 2009), modelos de performance individual de jogadores (seja para encontrar oportunidades de contratação, como parte da avaliação cotidiana da performance em um time) (HUGHES et al, 2012) ou modelos de otimização de receitas (que do ponto de vista da gestão visam encontrar oportunidades de incrementar receitas baseado no contexto do futebol) (BRANDES, FRANCK E NÜESCH, 2006)

Atualmente, grandes empresas como Microsoft e IBM utilizam esportes, entre eles o futebol como forma de promoção de suas ferramentas de análise de dados (HUSTON, 2017; BOYLES, 2016) Empresas como SAP, Wyscout e Opta lançaram no mercado produtos dedicados a insights em futebol para clubes e profissionais da área. As principais tendências atuais no campo de análise de performance se concentram na observação de eventos levando em conta o posicionamento espacial de onde ocorreram, tipo de informação que teve sua popularização através de mapas de calor (LINK, 2018)

Ao mesmo tempo que times possuem em suas mãos as ferramentas e informações para um uso sistemático de dados em suas decisões, ainda se observa que na maioria dos clubes essa inteligência não é levada em consideração, sendo muitas vezes substituída por opiniões de determinados indivíduos como técnicos e dirigentes. Tal situação é perceptível por exemplo quando treinadores de nível mundial como Josep Guardiola fazem comentários como "[Rodri]

Não tem tatuagens, brincos. Tem o cabelo certo para a posição" (ESPN, 2019), implicando que tatuagens, brincos e cabelo são causas de um mau desempenho em jogo. Ao mesmo tempo, alguns times são capazes de se aproveitar da vantagem competitiva que o conhecimento dos dados os pode trazer, como é o caso do Liverpool, que se sagrou campeão Europeu e bateu recordes em número de partidas consecutivas sem derrotas. O time pertence ao mesmo grupo que controla o time de beisebol Boston Red Sox, o qual foi um dos pioneiros nos tempos do Moneyball e parece estar aplicando sua experiência no time de futebol, a exemplo da nomeação de um presidente sem histórico no futebol profissional, mas que anteriormente ocupava uma vaga de analista de dados na equipe (WILLIAMS, 2020).

O presente artigo se propõe a contribuir com a fundamentação das bases teóricas para o estabelecimento de uma cultura de dados no contexto do futebol. Mais especificamente, a base do modelo que se popularizou com Moneyball depende de conhecimento teórico sobre a diferença entre percepção de valor de atletas e a contribuição efetiva em resultados de partidas. Assim, o problema será abordado inicialmente com o objetivo específico de modelagem da percepção de valor de um atleta (LEWIS, 2013)

O entendimento da origem do valor de um atleta permitiria à gestão de um clube de futebol maximizar ganhos vinculados à marketing, publicidade e aumento de receitas de bilheteria. Além disso, é possível obter lucros ao identificar atletas com potencial de aumento de valor percebido, mas com baixo valor de mercado (BRANDES, FRANCK e NÜESCH, 2006). Portanto, o objetivo deste estudo é criar uma métrica capaz de mensurar o valor percebido pelo público da performance de um jogador de futebol. Para isso serão aplicadas técnicas de machine learning como random forests, regressões logísticas e *suport vector machines* gerando um modelo preditivo capaz de mensurar a probabilidade de um jogador ser eleito pelo público para compor a seleção dos melhores atletas da rodada da competição. A probabilidade obtida pode ser utilizada como um índice que representa como a atuação do atleta foi recebida pelos torcedores, o que no contexto do futebol pode ser interpretado como uma medida de satisfação de seu consumidor final. Uma vez obtidas medidas objetivas do valor percebido de seus atletas, um clube pode utilizar a informação de forma a maximizar receitas, identificar oportunidades de compra de atletas e implementar soluções de aumento da satisfação de clientes. Finalmente, a aplicação de estatísticas a nível individual de atletas em futebol ainda é um campo muito pouco explorado, portanto este estudo contribui também para o crescimento do conhecimento do campo em questão.

2. Teoria

A análise da percepção de valor de um atleta tem ligação direta com a própria concepção de talento e seus impactos econômicos. A literatura científica nesse campo teve "*The Economics of Superstars"* como seu artigo seminal, publicado por Sherwin Rosen (ROSEN, 1981). Nele, o autor percebe que a noção de talento gera alterações no mecanismo de oferta e demanda, portanto, propõe um modelo econômico que retrata o efeito de *Superstars* na economia. Para Rosen (1981), o estabelecimento de tal fenômeno depende de duas pré-condições de um determinado mercado, a saber, uma oferta pequena e uma curva de distribuição de lucros desigual, onde poucos fornecedores detêm a maior parte das receitas. Nesse contexto, os fornecedores de destaque que são capazes de controlar a maior parte dos rendimentos são as denominadas "estrelas" (ROSEN, 1981).

No cenário proposto por Rosen (1981), o mercado de entretenimento é tomado como um dos exemplos onde ocorre uma influência de "estrelas" na distribuição de recursos. Dessa forma, dado o mercado de música, alguns cantores são capazes de acumular muito mais receita do que centenas de outros disponíveis no mercado. No modelo proposto pelo autor, o estabelecimento de uma estrela se dá pela capacidade de fornecimento de um serviço de qualidade superior às demais alternativas. Assim, os produtos de qualidade inferior atuariam como substitutos imperfeitos das "estrelas", fazendo com que os consumidores prefiram obter menores quantidades da opção mais satisfatória do que maiores quantidades dos substitutos. Ao mesmo tempo, é observado que para o funcionamento do modelo, o bem consumido deve permitir alta distribuição sem impactos significativos nos custos. Portanto, com o aumento de tecnologia, sobretudo no campo da comunicação, as "estrelas" tendem a se tornar cada vez mais lucrativas, devido á facilidade de acesso e aumento de demanda (ROSEN, 1981).

Em contraponto à solução proposta por Rosen (1981), Moshe Adler, em seu artigo "Stardom and Talent" (ADLER, 1985), propõe uma explicação alternativa aos fatores que permitem o estabelecimento de "estrelas". Enquanto Rosen (1981) atribuía o oferecimento de um serviço de qualidade superior ao domínio de um mercado, Adler (1985) propõe que é possível a ascensão de "estrelas" mesmo que o oferecido por elas não seja diferente de alternativas disponíveis. Para tal, o autor defende que em tais mercados, exista um consumo baseado em conhecimento. Assim, os produtos oferecidos exigem do consumidor dedicação para se informar a respeito do setor, visto que quanto mais conhecimento tiver, maior será a satisfação obtida. Logo, se estabelece para o consumidor um contexto em que a satisfação que

um fornecedor pode oferecer está condicionada a capacidade de seu cliente ouvir à seu respeito.

Tomando o entretenimento novamente como exemplo, Adler (1985) sugere que um cantor se torna uma “estrela” por ter mais pessoas falando à seu respeito e portanto, os consumidores de música terão mais conhecimento a respeito desses artistas do que de outros. Portanto, a ascensão de “estrelas” ocorre uma vez que existe uma cobertura mais extensiva a seu respeito, permitindo que os consumidores tenham mais informações e consequentemente maior satisfação ao consumi-los, enquanto também são capazes de otimizar o tempo dedicado à pesquisa, sem prejuízos para a satisfação do produto.

Egon Franck e Stephan Nüesch trazem o debate a respeito da ascensão de “estrelas” e seu impacto ao cenário do futebol em seu artigo “Talent and/or Popularity: What does it takes to be a Superstar?” (FRANCK e NÜESCH, 2012). Partindo das ideias propostas por Rosen (1981) e Adler (1985), os autores tentam compreender o papel das performances em campo e da popularidade de um atleta nos seus salários e valor de mercado.

Ao propor o cenário do futebol para os modelos econômicos de Rosen (1981) e Adler (1985), Franck e Nüesch (2007) estipulam como diferencial a premissa de que no contexto do esporte, o desempenho pode ser medido mais precisamente, permitindo uma abordagem quantitativa empírica da questão. Portanto, a análise se deu a partir de diferentes modelos criados a partir de dados do campeonato alemão de futebol. Nesse caso, a definição de “estrela” foi feita a partir dos jogadores que se encontravam acima do 95 percentil de salários na liga e foram coletados dados de eventos em campo (como gols, assistências, etc) para representar a performance, um índice compilando as citações de um determinado atleta na mídia para representar sua popularidade e estimativas de valor de mercado e salários. A partir desses dados, os autores buscaram quantificar o talento de um jogador como a capacidade de aumentar a probabilidade de vitória de seu time e realizar a análise dos resultados através da comparação dessa probabilidade com o custo do jogador, que representa seu valor. Com os modelos realizados, foi possível entre outros, determinar que ao mesmo tempo a performance em campo dos jogadores e a popularidade extra-campo são influenciam os salários e valor de mercado dos atletas. Ao mesmo tempo porém, foi observado que existe um aumento do retorno marginal pago pela performance de atletas a medida que seu valor de mercado aumenta, ou seja, dada a definição de “estrela” adotada no artigo, quanto mais próximo de ser uma “estrela”, maior a remuneração obtida por ação realizada durante o jogo.

Em contrapartida ao artigo de Franck e Nüesch (2012), outro estudo, também realizado na Alemanha trouxe resultados opostos. Erick Lehmann e Günter Shulze possuíam em seu artigo "What does it takes to be a star – The role of Performance and Media for German Soccer Players" (LEHMANN e SCHULZE, 2007) um objetivo similar aos seus conterrâneos, porém com algumas mudanças metodológicas. Os autores também optaram por utilizar o valor de mercado dos jogadores como variável dependente, porém alteraram a abordagem utilizada no cálculo das variáveis independentes. Enquanto Franck e Nüesch (2012) utilizavam 20 variáveis para definir a componente do talento relacionada à performance, Lehmann e Shulze (2007) utilizavam apenas as variáveis tidas como mais importantes para cada posição, analisando cada uma individualmente. Além disso, Franck e Nüesch (2012) consideraram a popularidade como o resíduo de uma regressão que utilizava a performance em campo para modelar a quantidade de artigos a respeito do atleta publicados em jornais, enquanto Lehmann e Shulze (2007) utilizavam a própria repercussão da imprensa como índice de popularidade. Os resultados do segundo modelo, corroboram a visão de que o mercado do futebol na Alemanha não se comporta com o "efeito superstar" de Rosen (1981) nem de Adler (1985), visto que apesar de performance e popularidade serem fatores componentes dos salários dos jogadores, ambos os fatores possuem diminuição de retornos a medida que o valor de mercado do atleta aumenta e são sujeitos a outras variáveis como idade, por exemplo. Entretanto, o estudo observou que a nível de equipe, os times são capazes de se beneficiar da popularidade de seus jogadores para incremento de suas receitas.

Muitos dos artigos publicados que aplicam estatística e analytics a futebol foram realizados tomando como objeto de estudo o futebol alemão, muito devido ao fato de ter sido a primeira liga a coletar e disponibilizar dados de partidas (LINK, 2018). No entanto, a cultura e contexto de diferentes países poderia trazer resultados distintos. Claudio Lucifora e Rob Simmons (LUCIFORA e SIMMONS, 2003), realizaram a análise do "efeito superstar" no contexto do futebol italiano. Em seu artigo, no entanto, adotaram uma metodologia diferente de outras observadas ao partir da definição inicial de "estrelas" como os jogadores de melhor performance em determinadas métricas das partidas. Além disso, baseado nas variáveis disponíveis, houve maior ênfase na análise de jogadores de ataque. O resultado final desse estudo foi a comprovação do modelo de Rosen (1981) no futebol italiano, apesar das ressalvas quanto à metodologia utilizada. Em 2007, também foi realizado um estudo no contexto do futebol espanhol por Pedro Garcia-del-Barro e Francesc Pujol, (GARCIA-DEL-BARRO e PUJOL, 2007), no qual foi adotada uma abordagem de modelagem do valor de um jogador

em função de suas menções por jornalistas e pesquisas na internet, mantendo outras variáveis como controle. O resultado obtido foi de que os jogadores recebiam um premium além de sua performance em campo devido ao seu nível de popularidade, corroborando um modelo de "efeito superstar" nos moldes de Adler (1985).

Apesar de não existir consenso a respeito da presença e forma de "efeito superstar" no futebol, Leif Brandes, se juntou a Franck e Nüesch para analisar o efeito que a presença de "estrelas" exerce na atração de torcedores e, consequentemente, receita para os times de futebol (BRANDES, FRANCK e NÜESCH, 2006). Assim como em outros estudos de Franck e Nüesch (2012), atletas no 95° percentil de valor de mercado foram definidos como "estrelas". Além disso, é introduzido o conceito de "herói local", título atribuído aos jogadores de maior destaque em equipes que não possuem "estrelas". Os modelos formulados no artigo visavam traçar a relação entre a receita obtida em uma partida e a as principais métricas de performance dos heróis locais e "estrelas" envolvidas. O resultado obtido apontou diferenças significativas no impacto dos jogadores de diferentes status. Ambos os grupos foram capazes de impactar positivamente as receitas, porém de formas distintas. Os heróis locais demonstraram maior ligação entre sua performance em campo e capacidade de atração de público, culminando em maiores ganhos marginais a medida que a performance das métricas analisadas melhorava. Por outro lado, nos jogadores com status de "estrela", sua performance se mostrou pouco significativa para justificar o aumento de ingressos vendidos. Também foram analisadas diferenças em relação à influência do local da partida para a receita incremental gerada pelos jogadores, através da criação de modelos específicos para jogos domésticos e fora de casa. Os diferentes modelos acentuaram a influência de "estrelas" na atração de público, mesmo quando seu time não era o mandante da partida, enquanto os heróis locais tinham seu impacto mais restrito à partidas locais. Portanto, o estudo de Brandes, Franck e Nüesch (2006) conclui que a diferença fundamental entre heróis locais e "estrelas" se dá uma vez que heróis locais são capazes de gerar incrementos nas receitas de bilheteria de seus clubes através de sua performance em campo, apesar de possuir uma zona de influência reduzida, o que diminui sua capacidade de gerar lucros incrementais em partidas fora de casa. Por outro lado, as "estrelas" atraem interesse do público principalmente devido a sua popularidade, o que aumenta sua influência a nível nacional ou até internacional, possibilitando aumentos de receitas mesmo em jogos fora de casa e independentemente de sua performance em campo.

Dado os resultados obtidos em sua pesquisa, Brandes, Franck e Nüesch (2006), levantam em seu artigo uma discussão sobre a relação entre heróis locais e “estrelas” como possíveis justificativas de suas descobertas. Os autores consideram que os heróis locais dependem de sua performance para se destacar e atingir esse status. Porém, ao atingir a condição de herói local, o atleta tende a obter maior atenção e ganhar popularidade. A progressão de carreira dos jogadores, comumente implica na transição de herói local para “estrela” uma vez que a alta performance em clubes de menor expressão é a principal porta de entrada para alcançar transferências para clubes maiores, onde são capazes de receber maior atenção e destaque e possivelmente alcançar a condição de “estrelas”.

A contribuição esperada do presente artigo consiste na criação de uma métrica de valor percebido a partir de uma análise da performance dos atletas. Tal abordagem remete principalmente ao artigo de Franck e Nüesch (2012), porém com foco maior na componente de performance ligada ao desempenho em campo e tomando como alvo a percepção do público em detrimento do valor de mercado do atleta. Considerando as implicações do estudo de Brandes, Franck e Nüesch (2007), a métrica desenvolvida pode ser utilizada por clubes na avaliação de atletas, como forma de identificar jogadores que são capazes de fornecer performances que se destacam aos olhos do público. Dessa forma, times podem encontrar formas de maximizar seus lucros em receita de partidas a partir da constituição de seu elenco. Além disso, os gestores de clubes (principalmente os que possuem maior atenção midiática) seriam capazes de realizar contratações de atletas com performance valorizada em times mais modestos, os quais teriam maior probabilidade de alcançar o status de “estrela” uma vez submetidos a publicidade adequada, aumentando seu valor de mercado e gerando lucros na venda de atletas, bem como trazendo receitas de bilheteria incrementais.

3. Métodos

A análise realizada possui natureza quantitativa exploratória, visando o desenvolvimento de uma métrica objetiva de valor percebido da performance de jogadores de futebol. O estudo se restringe estritamente ao uso de registros de eventos ocorridos ao longo das partidas, visto que estes representam diretamente sua contribuição como atleta para os resultados das partidas. Além disso, a escolha por uma abordagem exploratória de pesquisa foi utilizada

principalmente pela pouca compreensão que existe hoje entre os resultados de uma partida de futebol e o papel individual dos atletas envolvidos (HAIR, 2003).

A metodologia utilizada na análise foi o desenvolvimento de um modelo preditivo baseado em técnicas de machine learning (*Random forest*, regressão logística e *support vector machine*) para estabelecer uma pontuação para o desempenho de cada jogador após uma partida. A escolha dos três modelos levou em conta a natureza diferente entre eles, visando uma melhor combinação dos três em uma etapa seguinte. O algoritmo de *random forest* consiste em combinar centenas ou até milhares de árvores de decisão, sendo cada uma gerada com um número limitado de variáveis, selecionadas aleatoriamente dentre os dados fornecidos. Apesar de existirem outros modelos baseados em árvores, alguns inclusive mais avançados do que a *random forest*, como XGBoost, por exemplo, dado o risco elevado de *overfitting* que a amostra possui, visto seu desbalanceamento da variável alvo e a pouca quantidade de observações, utilizar um modelo de menor complexidade promove a redução do risco de sobre ajuste. O modelo de *support vector machine,* é um modelo que atua através da separação das observações criando hiperplanos, os quais se colocam entre as observações de cada classe, maximizando a distância entre o plano e as observações mais próximas. O SVM é um algoritmo robusto em relação à presença de *outliers* e também ao desbalanceamento de amostra, pode ser linear ou não linear, dependendo de suas configurações, sendo que no presente estudo, foi adotada a metodologia linear. O último modelo utilizado foi uma regressão logística, que assim como o SVM e ao contrário da *random forest*, possui natureza linear. A regressão logística é um dos modelos de classificação mais tradicionais e conhecidos. Seu funcionamento é similar ao de uma regressão linear simples, porém os dados passam por uma transformação para formar uma função sigmoidal como resultados fornecidos, os quais podem ser interpretados como probabilidades de um evento ocorrer.

Os dados utilizados no estudo foram comprados do site Rotowire (2019), que é provedor de estatísticas para jogadores de *fantasy games*, bem como de grandes veículos de imprensa como ESPN e Fox Sports. Os dados utilizados são todos apresentam todas as partidas do Campeonato Brasileiro de 2019 e contém informações a nível individual dos jogadores (N=10.633) de um total de 90 variáveis (Tabela 1). Os dados foram limpos para a exclusão de variáveis com alta correlação entre si, ou pouco discriminatórias, devido à ausência de variabilidade entre as observações. Também foram desprezadas no treinamento dos modelos jogadores que atuaram como substitutos nas partidas, resultando em uma amostra final de N=8.358 jogadores.

Tabela 1: Variáveis utilizadas

Acrônimo	Definição
IBS	Shots Inside Box
IBSOG	Shots On Goal Inside Box
IBG	Goals Inside Box
OBS	Shots Outside Box
OBSOG	Shots On Goal Outside Box
OBG	Goals Outside Box
APW	Assist Penalties Won
TOFF	Total offside
BCM	Big chance missed
BCS	Big chance scored
ATTDR	Attempted dribbles
PK	Penalty Kicks Taken
PKG	Penalty Kick Goals
FS	Fouls Suffered
DSP	Dispossessed
TBOX	Touches In Box
CRNW	Corners Won
AW	Aerials Won
INT	Interceptions
BLK	Blocks
TKLW	Tackles Won
BR	Ball Recoveries
DW	Duels Won
CL	Clearances
LMT	Last man tackle
Y	Yellow Cards
YR	Yellow/Red Cards
R	Red Cards
FC	Fouls Committed
EG	Errors Lead To Goal
ES	Errors Lead To Shot
OWN	Own Goals
PKC	Penalty Kicks Conceded
SA	Secondary Assists
ACRO	Accurate Crosses - Open Play
BCC	Big Chances Created
AOP	Assists From Open Play
CCOP	Chances Created From Open Play
CCSP	Chances Created From Set Play
P	Passes
ALB	Accurate long balls
ATB	Accurate throught balls
TOUCH	Touches

GC	Goals Conceded
Chance Quality Index	Big Chances Creates/Total Chances Created
Pass Precision	Accurate Passes/Total Passes
Cross Precision	Accurate Crosses/Total Crosses
Shot Precision	Shots on Goal/Total Shots
Dribble Precision	Successful Dribbles/Attemped Dribbles
Tackle Precision	Tackles Won/Tackles
IBSV*	Save of shot made inside the box
OBSV*	Save of shot made outside the box
AKS*	Accurate keeper sweeper
PKSV*	Penalty Kick Save
PUNCH*	Punches to clear ball

*Variáveis exclusivas de goleiros

Os modelos desenvolvidos tiveram como objetivo prever a probabilidade de um jogador ser eleito para integrar a "Seleção da Rodada", que consiste em uma votação popular realizada nas redes sociais da CBF, na qual após cada rodada, torcedores escolhem dentre todos os jogadores que atuaram na rodada, os melhores em cada posição que comporiam o time ideal da rodada em questão. Dessa forma, o alvo do modelo se trata de uma variável binária que indica simplesmente se o jogador foi eleito ou não na rodada em questão.

Considerando que a "Seleção da Rodada" é sempre limitada a 11 jogadores, onde normalmente são selecionados no máximo à 2 atletas de cada posição, existe um grande desbalanceamento das observações da variável alvo na base de dados (Tabela 2). Para lidar com o problema, foram criadas observações artificiais (N=1000) utilizando a técnica de *smoothed bootstrap re-sampling* (MENARDI e TORELLI, 2014).

Tabela 2: Tamanho e distribuição da base

	Seleção	Total	
GOL	37	736	5%
LAT	68	1503	5%
ZAG	69	1495	5%
VOL	43	921	5%
MC	39	914	4%
ML	60	1475	4%
MEI	21	425	5%
ATA	43	889	5%
Total	**380**	**8358**	**5%**

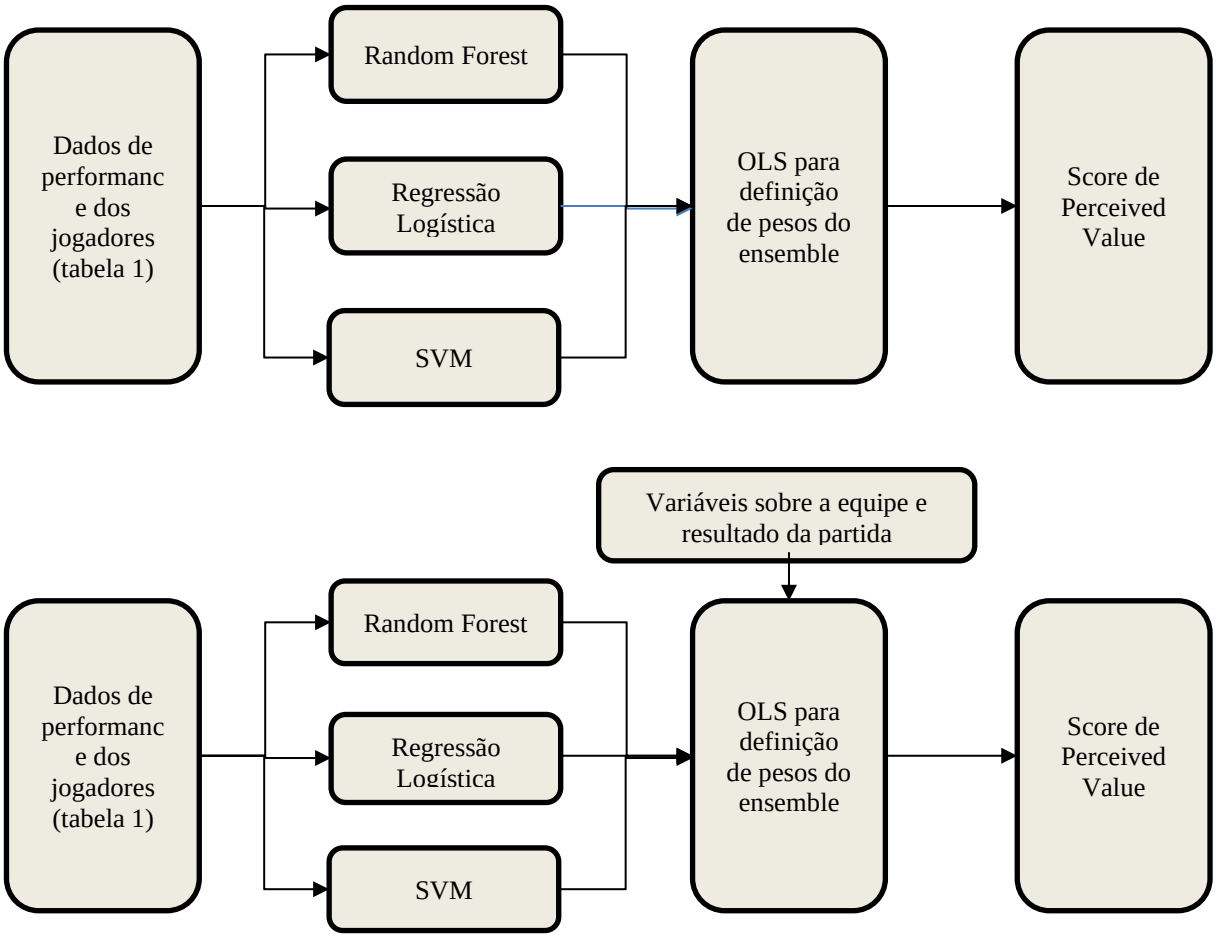
Para alcançar maior especificidade nos resultados, foram gerados modelos específicos para cada uma das principais posições dos jogadores, a saber: goleiros, zagueiros, laterais, volantes, meias centrais, meias laterais, meias ofensivos e atacantes. Para cada uma das posições, uma base artificial balanceada foi gerada a partir de cerca de 70% das observações reais, enquanto o restante foi reservado para testes dos modelos (Tabela 3). As amostras artificiais então, foram utilizadas para treinar modelos de Random Forest, Regressão Logística e *Support Vector Machine* para cada uma das posições.

Tabela 3: Tamanho e distribuição da amostra de teste

	Seleção	Total	
GOL	8	236	3%
LAT	23	503	5%
ZAG	28	495	6%
VOL	6	121	5%
MC	12	214	6%
ML	19	475	4%
MEI	9	125	7%
ATA	16	189	8%
Total	**121**	**2358**	**5%**

A partir dos resultados obtidos nos modelos, cada jogador agora tinha 3 índices que representavam a probabilidade do atleta ser eleito para a "Seleção da Rodada" baseado em métricas de performance da partida disputada na rodada em questão. Para alcançar uma métrica única, foi realizado um *ensemble* entre os modelos, utilizando uma regressão linear simples sem intercepto com a variável da "Seleção" como alvo, para otimizar os pesos atribuídos a cada uma das métricas. Nesse ponto, para méritos de comparação, também foi realizado um teste no qual foram adicionadas novas variáveis ao modelo, com o intuito de adicionar informações externas à performance individual do jogador. Assim, o resultado final obtido foram dois *scores* para cada jogador, ambos indicando a chance do atleta ser considerado pelo público o melhor de sua posição na rodada. Um dos modelos porém traz o índice obtido considerando apenas feitos individuais realizados na partida, enquanto a outra também considera: o time do jogador (através de variável *dummy*), o resultado da partida (em termos de vitória, derrota ou empate), o local do jogo (considera se o atleta pertencia à equipe mandante ou visitante) e o saldo de gols do encontro (diferença entre gols marcados e gols sofridos pela sua equipe).

Figura 1: Resumo dos modelos criados



Foram desenvolvidas versões dos modelos descritos pela para cada um dos seguintes grupos: Goleiros, Zagueiros, Laterais, Volantes, Meias Centrais, Meias Laterais, Meias Ofensivos e Atacantes

Uma vez desenvolvida a métrica de valor percebido, o desempenho do modelo foi aferido submetendo a base de teste, que até então não fora utilizada, ao modelo e atribuindo os *scores* aos jogadores. A precisão foi a principal métrica observada para a avaliação de performance do modelo, dado que ela permite medir a capacidade do modelo de discriminar performances passíveis de serem consideradas as melhores e ao mesmo tempo não é distorcida pelo desbalanceamento da base, como seria a acurácia, por exemplo. A precisão do modelo foi aferida sob dois pontos de vista. Para o primeiro modo de avaliação, os índices de valor percebido foram normalizados pelas posições, visando torná-los comparáveis entre si. Em seguida, as observações foram ordenadas de forma decrescente baseada nos *scores* e foi

admitido que as 121 primeiras seriam consideradas pertencentes à classe alvo, visto que esse é o número total de observações dessa classe na amostra de teste. O segundo modo de avaliação analisou cada posição separadamente, porém de forma similar, ou seja, em cada posição, os jogadores foram ordenados de acordo com seus *scores* de valor percebido e os primeiros foram tidos como pertencentes à classe desejada, sendo esse número definido pela quantidade de indivíduos escolhidos para a "Seleção da Rodada" naquela posição presentes na base.

4. Resultados

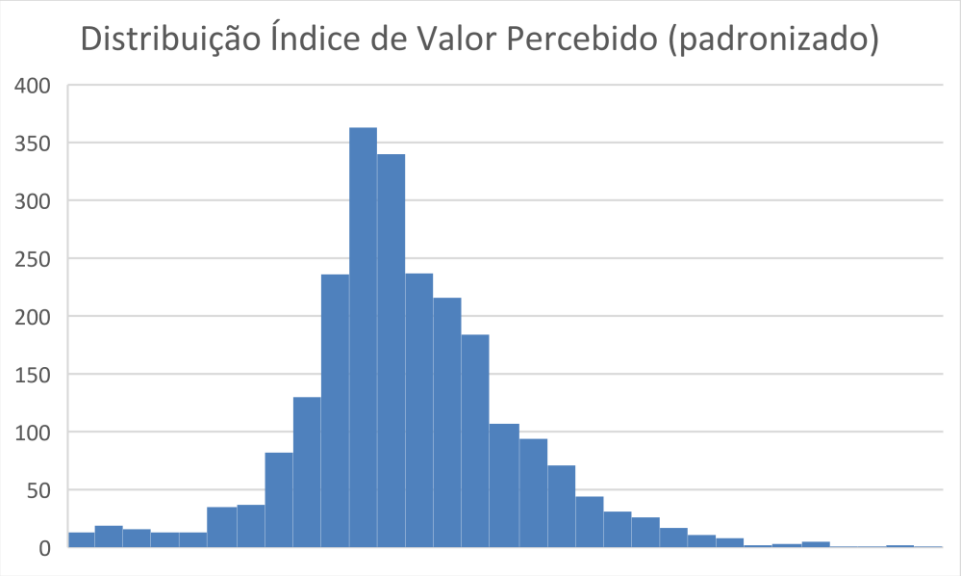
O Índice de Valor Percebido

O objetivo final do modelo é alcançar um score que represente o quanto a performance do jogador foi considerada positiva aos olhos de torcedores. O índice obtido varia de forma contínua, porém, dado a combinação entre os modelos ter sido feita de forma aditiva através de uma OLS, não existem limites positivos ou negativos que o índice pode atingir. Quanto maior a pontuação obtida por um jogador, melhor terá sido sua performance aos olhos dos espectadores. Como foram treinados modelos separadamente para cada posição do futebol, não necessariamente as pontuações obtidas são imediatamente comparáveis entre atletas de posições diferentes. Para tornar os índices mais universais, foi realizada uma normalização utilizando a técnica de z-score. Nas tabelas abaixo, estão as principais métricas descritivas do índice de valor percebido, tanto na versão original, quanto em sua versão padronizada.

Tabela 5: Métricas descritivas do Índice de Valor Percebido

IVP	
Max	2,165132
Min	-0,29078
Median	1,053894
Average	1,063407

IVP - Padronizado	
Max	4,83794
Min	-3,13748
Median	-0,11531
Average	0



Desempenho do modelo baseado apenas na performance do jogador

A tabela 5 traz a precisão do modelo baseado exclusivamente nos eventos ocorridos durante a partida avaliada. Nela, observa-se o número total de observações selecionadas para compor a "Seleção da Rodada" presentes na base de teste em cada uma das posições, bem como quantos desses teriam sido indicados de acordo com o modelo. O uso de uma amostra específica reservada para testes reduz significativamente a possibilidade de *overfitting*, entretanto, o número reduzido de observações da classe alvo, sobretudo em alguns grupos, como os volantes, pode contribuir para uma estimativa de precisão ligeiramente menor do que seria observado em uma população maior.

Tabela 5: Modelo baseado apenas na performance individual do jogador

	Total Pertencente à Classe Observada	Corretamente Classificados	Precisão
GOL	8	3	38%
LAT	23	8	35%
ZAG	28	7	25%
VOL	6	0	0%
MC	12	5	42%
ML	19	4	21%
MEI	9	3	33%
ATA	16	9	56%
Total	**121**	**39**	**32%**

Caso a precisão fosse aferida sem analisar as posições separadamente, o resultado não seria muito diferente, alcançaria 33% de precisão ao classificar como membros da seleção apenas 40 atletas entre os 121 com maior *score* de valor percebido. Apesar de o desempenho do modelo se mostrar ruim, ele ainda representa uma melhoria significativa quando comparado aos 5% de probabilidade que qualquer jogador tem de ser eleito para a "Seleção da Rodada".

Em um cenário real em que o modelo fosse usado para prever a "Seleção da Rodada", é possível que se obtivessem resultados superiores de precisão causados pela existência de maiores restrições para classificar um jogador como pertencente a Seleção ou não. A estrutura utilizada na formulação da Seleção leva em conta a organização tática de uma equipe. Assim, existem limitações quanto ao número de jogadores de uma posição que podem ser escolhidos, o que não foi levado em conta na análise do modelo, onde simplesmente foram criados pontos de corte para a classificação dos atletas. Dessa forma, enquanto no cenário real apenas dois zagueiros serão selecionados, por exemplo, é possível que o modelo considere 3 atletas acima do ponto de corte, o que resultara em uma classificação errada, que não ocorreria se o modelo fosse aplicado no contexto real. De modo semelhante, podem haver situações nas quais um atleta não alcançou pontuação suficiente para estar entre os classificados em sua posição, porém, na rodada em questão ele ainda assim é o que possui o maior *score*, o que faria com que fosse selecionado em um cenário de avaliação mais realista.

Análise da performance superior às demais observada na classificação de atacantes

Em meio as dificuldades encontradas pelo modelo para classificar corretamente os indivíduos, é notável o desempenho superior aos demais alcançado pelo modelo destinado aos atacantes. Uma das grandes vantagens de trabalhar com atletas dessa posição, é que eles possuem variáveis de performance bem definidas, facilmente observáveis e mais memoráveis. Em última instancia, a função primária de um atacante é marcar gols, o que por sua vez, são os eventos de uma partida que mais chamam a atenção do público e impactam diretamente no resultado final da partida. Dessa forma, a associação entre desempenho em campo e qualidade da atuação é muito mais direta do que em outras posições. Por outro lado, atletas de outras posições tendem a ser muito mais difíceis de ser avaliados de forma objetiva, visto que normalmente não é claro para espectadores quais ações individuais caracterizam uma boa

performance nesses casos, o que abre espaço para subjetividade e aumento da importância de fatores externos à performance em si.

Em muitos dos artigos publicados a respeito de análises estatísticas do futebol, esse fenômeno pode estar presente. Sobretudo na literatura mais antiga, onde não existiam ferramentas para registros tão extensos e detalhados dos eventos ocorridos ao longo de uma partida. Por isso, é muito comum que em casos nos quais o autor deseja utilizar variáveis que indicam performance, as métricas escolhidas estejam restritas àquelas facilmente observáveis, como por exemplo gols e assistências. Logo, alguns estudos podem estar sujeitos a um viés pela seleção das variáveis utilizadas e nesse caso, potencialmente pode se observar maior aplicabilidade dos resultados a atacantes do que a outras posições.

Desempenho do modelo baseado na performance do jogador e em fatores externos

A Tabela 6 mostra os resultados do segundo modelo desenvolvido. A amostra de teste utilizada é a mesma para todos os modelos, assim as características dela serão as mesmas. Dessa forma, mais uma vez existe o risco de uma estimativa da precisão ligeiramente abaixo da precisão real do modelo. Além disso, este modelo também está sujeito a possíveis avaliações incorretas pela inviabilidade da aplicação de uma metodologia de aferição dos resultados mais realista que considere o contexto tático envolvido na estrutura da "Seleção da Rodada".

Tabela 6: Modelo baseado em performance do jogador e adicionando informações de time e resultado da partida

	Total Pertencente à Classe Observada	Corretamente Classificados	Precisão
GOL	8	5	63%
LAT	23	10	43%
ZAG	28	14	50%
VOL	6	2	33%
MC	12	6	50%
ML	19	8	42%
MEI	9	6	67%
ATA	16	11	69%
Total	**121**	**62**	51%

Ao incluir algumas variáveis simples, mas que não refletem diretamente a atuação do jogador, a precisão do modelo melhorou drasticamente. Desprezando a divisão por posições, o desempenho real do modelo é ligeiramente pior, atingindo 48% de precisão ao acertar a classificação de 59 dos 121 indivíduos.

Dentre as novas variáveis adicionadas, duas em especial tiveram maior impacto na atribuição dos *scores*: o time ao qual o jogador pertence e o resultado da partida (em termos de vitória, derrota ou empate). Essas duas variáveis em particular ressaltam a importância do desempenho coletivo na avaliação individual, uma vez que o resultado da partida é consequência de trabalho em conjunto entre os atletas de um time. No caso da variável que representa a equipe, inicialmente poderia ser esperado um viés ligado à torcida ou influência que cada time é capaz de exercer, ainda mais pelo fado de se tratar de uma votação popular na qual os eleitores via de regra possuem um time favorito e não têm obrigação nenhuma de votar com imparcialidade. Além disso, também é esperado que jogadores de partidas menos de menor destaque (seja pela relevância dos times envolvidos ou pelo resultado final), mesmo que demonstrem desempenho individual notável não será capaz de gerar interesse o suficiente para levar os torcedores a indica-lo para a premiação. Entretanto, o que foi visto é que em geral, os coeficientes atribuídos pelo modelo aos diferentes times se mostraram muito próximos à ordem de classificação das equipes na tabela do campeonato, indicando maior influência da performance coletiva do que da parcialidade da votação.

Análise da diferença de performance entre os modelos

Ao ver os resultados dos modelos, fica claro que ao utilizar dados de fora do escopo da performance individual dos atletas ocorre uma melhoria significativa. Isso denota que existe uma certa dificuldade das pessoas em avaliar a performance de jogadores de futebol se limitando ao que de fato realizaram individualmente em uma partida. Considerando a metodologia utilizada, as variáveis externas à performance atuam basicamente atribuindo "bônus" ao *score* que o jogador obteve por sua atuação de acordo com o outro modelo, o que permite que as diferenças entre os resultados sejam interpretadas resumidamente como o efeito de vieses sobre a performance real dos atletas.

Partindo do pressuposto de que as pessoas tendem avaliar a performance de atletas em partidas de futebol de forma precisa e imparcial, os melhores jogadores deveriam ser

definidos estritamente baseados nos eventos ocorridos em campo e portanto, seria de se esperar que o registro dos eventos ocorridos fosse capaz de obter um poder preditivo significativo sobre quais seriam os melhores jogadores em campo. Os resultados obtidos, no entanto, mostram que existem falhas na compreensão do que é uma boa performance, visto que resultados coletivos de performance são atribuídos como méritos individuais aos jogadores.

A eficiência dos modelos para medição do valor percebido da performance de um jogador

A variável alvo utilizada no desenvolvimento dos modelos foi escolhida por atuar como indicador do valor atribuído à performance de um jogador, no entanto o real objetivo do estudo é mais amplo do que a previsão da "Seleção da Rodada". O intuito de utilizar o resultado final de uma votação popular como variável dependente é fruto do pressuposto de que se o modelo é capaz de prever as atuações mais bem avaliadas de todas, quanto maior a probabilidade calculada de um atleta estar entre os melhores, melhor terá sido a percepção que os torcedores tiveram sobre sua atuação.

Baseado nesse objetivo, faz mais sentido a preferência do modelo baseado em performance, uma vez que este reflete melhor quais são os atributos de performance mais relacionados à indicação para melhor jogador em sua posição. Apesar disso, foi necessário acrescentar ao modelo variáveis que não são diretamente relacionadas a performance individual para que o modelo alcançasse bons níveis de precisão. Com o desempenho final obtido, o modelo indicou ser bom o suficiente para permitir a interpretação desejada do *score* desenvolvido.

Apesar de o segundo modelo apresentar uma performance maior, no que diz respeito à forma como a performance real é computada, ambos os modelos são idênticos, visto que o acréscimo das novas variáveis ocorre em uma etapa posterior ao desenvolvimento dos modelos de performance, o que o torna como um "peso" extra adicionado aos valores do primeiro modelo. Dessa forma, o papel que a performance real dos jogadores exerceu sob a percepção de valor que os torcedores atribuíram ao atleta na partida se manteve inalterado. Portanto, o primeiro modelo pode ser considerado eficaz para medir a percepção das pessoas a respeito da performance real dos jogadores, mesmo que esta não se traduza automaticamente

como a probabilidade de integrar a “Seleção da Rodada”, visto que isso é influenciado por outros fatores.

5. Conclusão

A partir dos modelos desenvolvidos, foi possível criar o índice de valor percebido para a performance de jogadores de futebol. Essa nova métrica é especialmente valiosa do ponto de vista de gestão de uma equipe, uma vez que permite otimizar o valor gerado pelo time em campo para o consumidor final, no caso o torcedor. É importante destacar que o valor percebido em si não necessariamente implica em melhor performance, ou seja, essa métrica não indica o quanto um jogador contribuiu para o resultado do time. Por isso, os principais ganhos que o índice possibilita estão ligados ao mercado de transferências e geração de receitas.

O mercado de transferências do futebol já foi amplamente estudado em diversos estudos (FRICK, 2007) e um clube capaz de compreender as origens do valor de um atleta passa a estar em condições de tirar vantagem de algumas particularidades da indústria. Na verdade, a descoberta da influência de fatores à nível coletivo na percepção de atletas já permite por si só grandes avanços no que diz respeito à contratação de jogadores. Uma vez que um indivíduo começa a demonstrar performances dignas de alto valor percebido, mas que estão sendo menosprezadas por estarem sendo feitas em times de menor expressão, esse jogador pode representar um ótimo investimento para um clube de maior expressão. Nesse cenário, uma vez integrando um time de maior visibilidade e um melhor desempenho coletivo, o valor percebido do atleta tende a aumentar e com ele seu valor de mercado.

O comportamento observado dos modelos de valor percebido dos jogadores de futebol se mostrou alinhado com muitos dos artigos publicados sobre o assunto. O conceito de superstar, amplamente analisado por diversas abordagens, se mostra refletido no fato do valor percebido sofrer influência ao mesmo tempo da performance do jogador e de fatores externos, algo similar ao observado no efeito da popularidade sobre o valor de mercado de um jogador.

Ainda considerando o diálogo entre o índice de valor percebido e a literatura a respeito do uso de estatística em futebol, um dos maiores benefícios que um clube pode obter ao utilizar essa métrica ocorre quando ela se torna complementar ao conhecimento do estudo a respeito de heróis locais e *superstars* (BRANDES, FRANCK e NÜESCH, 2006). No artigo, são analisados principalmente os impactos dessas duas categorias de jogadores nas receitas de

bilheteria dos clubes. Nele, foi observado que os *superstars* são capazes de gerar mais receitas pelo simples fato de entrar em campo, devido ao seu status. A análise objetiva do valor que um jogador é capaz de trazer para os torcedores que o assistem possibilita maior eficiência na "criação" desse fenômeno e consequentemente, resulta em maiores receitas para o time. Os autores do artigo sugerem como uma hipótese para a formação de *superstars* a ascensão de heróis locais para times de maior visibilidade. Os heróis locais são atletas que se destacam em performances por times de menor expressão, o que os permite gerar receitas incrementais em jogos domésticos, porém os ganhos gerados por eles dependem mais da manutenção de um nível elevado de performance do que *superstars*, os quais geram receitas incrementais em qualquer partida e sem depender muito da qualidade de sua atuação. O uso do índice de valor percebido permite aos clubes aumentar suas receitas de bilheteria, uma vez que auxilia na identificação e formação de heróis locais, ou a contratação de *superstars* em potencial.

Apesar das particularidades existentes no meio esportivo, os times de futebol não deixam de ser empresas, cujo principal produto é o entretenimento. Sob essa ótica, o índice de valor percebido é uma métrica que modela o valor gerado ao cliente do produto que está sendo oferecido. No futebol brasileiro, é comum que os clubes enfrentem pressões da torcida, as vezes até de forma agressiva, o que se mostra prejudicial às operações do clube. Com maior controle e monitoramento do valor que está sendo gerado, é possível prevenir ou lidar de forma mais eficaz com essas situações. Outros usos de um conhecimento mais profundo sobre o valor gerado aos torcedores estão no auxílio ao desenvolvimento de ações de marketing mais populares e até na otimização de preços cobrados por ingressos ou produtos. Além disso, os times também se beneficiam ao aumentar o número de torcedores, porém como em toda indústria existe competição por novos clientes. Um clube capaz de gerar valor aos torcedores de forma mais eficiente tende a expandir sua base de torcedores, o que por sua vez no longo prazo traz maiores receitas.

Uma ressalva importante a ser considerada é que um atleta apresentar uma performance valorizada pelos torcedores não necessariamente implica que ele contribuiu significativamente para vencer a partida. Da mesma forma, podem existir performances que não são valorizadas, mas que na verdade contribuem muito para uma vitória do time. Na verdade, a compreensão do valor percebido e do valor gerado, bem como a relação entre as duas métricas, complementariam uma à outra de forma a gerar grandes oportunidades para os clubes, ao ponto de potencialmente revolucionar a indústria do futebol brasileiro. O manuseio correto de ambas as métricas pode permitir diversas estratégias distintas, entre elas, a formação de um

time que une jogadores baratos, porém com alta contribuição para vitórias até então desapercebida e jogadores com potencial de gerar valor para os torcedores. A mistura dos dois times no elenco se complementariam para otimizar os resultados em campo e fora dele, uma vez que com baixo custo seria formado um time vencedor e que por alcançar bons resultados maximizaria o valor percebido de seus atletas e com isso otimizaria ganhos financeiros. Portanto, seria de grande valia futuros estudos para a formulação de métricas de performance atreladas a capacidade de um jogador gerar vitórias à equipe.

6. Referências

ADLER, Moshe. Stardom and Talent. **The American Economic Review** , Vol . 75 , No . 1 (Mar ., 1985), pp . 208-212 Stable URL : http://www.jstor.org/stable/181271. *[S. l.]*, v. 75, n. 1, p. 208–212, 2016.

BATE, Richard. Football Chance: tactics and strategy**. Londres: e & Fn Spon.**, 1988

BOYLES, Ryan. Team up and connect: an IoT soccer project for good causes. Publicado em 11 de agosto de 2016. Disponível em: <https://www.ibm.com/blogs/internet-of-things/team-up-connect/>. Acesso em: 30 jan. 2020.

BRANDES, Leif; FRANCK, Egon; NÜESCH, Stephan; BRANDES. Local Heroes and Superstars – An Empirical Analysis of Star Attraction in German Soccer. **Institute for Strategy and Business Economics University of Zurich Working Paper Series**, n. 46, 2006.

CONSTANTINOU, A. C., FENTON, N. E., & NEIL, M(2012). pi-football: A Bayesian network model for forecasting Association Football match outcomes. **Knowledge-Based Systems**, 36, 322-339.

DIXON, Mark; COLES, Stuart. Modelling Association Football Scores and Inefficiences in the Football Betting Market. **Applied Statistics**, vol.46, n.2, pp.265-280, 1997.

ESPN.COM.BR. Manchester City: Preconceito? Guardiola exalta seu novo volante: 'Não tem tatuagens, brincos. Tem o cabelo certo para a posição'. 10 de agosto de 2019. Disponível em: <https://www.espn.com.br/futebol/artigo/_/id/5938185/manchester-city-preconceito-guardiola-exalta-seu-novo-volante-nao-tem-tatuagens-brincos-tem-o-cabelo-certo-para-a-posicao>. Acesso em: 30 jan. 2020.

FRANCK, Egon; NÜESCH, Stephan. Talent and/or popularity: What does it take to be a superstar? **Economic Inquiry**, *[S. l.]*, v. 50, n. 1, p. 202–216, 2012. DOI: 10.1111/j.1465-7295.2010.00360.x.

FRICK, Bernd. The Football Players' Labor Market: Empirical evidence from the major European leagues. **Scottish Journal of Political Economy** *[S. l.]*, v. 54, n. 3, p. 422–446, 2007.

GARCIA-DEL-BARRO, Pedro; PUJOL, Fransesc. Hidden Monopsy Rents in Winner‐take‐all Markets: Sport and Economic Contribution of Spanish Soccer Players. **Managerial and Decision Economics**, 28: 57‐70, 2007

HAIR, Joseph. **Essentials of Business Research**. *[S. l.]*: Wiley, 2003. 440 p.

HUGHES, Michael; CAUDRELIER, Tim; JAMES, Nic; REDWOOD-BROWN, Athalie; DONNELLY, Ian; KIRKBRIDE, Anthony; DUSCHESNE, Christophe. Moneyball and soccer - An analysis of the key performance indicators of elite male soccer players by position. **Journal of Human Sport and Exercise**, *[S. l.]*, v. 7, n. SPECIALISSUE.2, p. 402–412, 2012. DOI: 10.4100/jhse.2012.72.06.

HUGHES, Mike; ROBERTSON, K.; NICHOLSON, A.. Comparison of patterns of play of successful and unsuccessful teams in the 1986 World Cup for soccer. **Londres: e & Fn Spon**., 1988.

HUSTON, Lainie. New Garage project brings predictive analytics to sports performance data. Publicado em 27 de junho de 2017. Disponível em: <https://www.microsoft.com/en-us/garage/blog/2017/06/new-garage-project-brings-predictive-analytics-sports-performance-data/>. Acesso em: 30 jan. 2020.

JAMES, Bill. The Bill James Guide to Baseball Managers: From 1870 to Today. Nova Iorque: **Diversion Books**, 2014. 352 p.

KUMAR, G. Machine Learning for Soccer Analytics. **KU Leuven, MSc thesis**, *[S. l.]*, n. SEPTEMBER 2013, p. 1–2, 2013. DOI: 10.13140/RG.2.1.4628.3761. Disponível em: http://www.researchgate.net/publication/257048220_Machine_Learning_for_Soccer_Analytics/file/9c96052441dfabfc87.pdf%5Cnpapers3://publication/uuid/B50B6BF5-B409-4FE1-A592-2DBFF6B2DB1D.

LEWIS, Michael. Moneyball: O homem que mudou o jogo. Rio de Janeiro: Intrínseca, 2015. 336 p. Tradução de: Ana Beatriz Rodrigues e Claudio Figueiredo.

LEHMANN, Erik E.; SCHULZE, Günther G. What Does it Take to be a Star? – The Role of Performance and the Media for German Soccer Players. **Applied Economics Quarterly**, [S. l.], v. 54, n. 1, p. 59–70, 2008. DOI: 10.3790/aeq.54.1.59.

LINK, Daniel. **Data Analytics in Professional Soccer**. [s.l: s.n.]. DOI: 10.1007/978-3-658-21177-6.

LITTLE, Thomas. Optimizing the Use of Soccer Drills for Physiological Development. **Strength and Conditioning Journal**, v. 31, n. 3, p. 67-74, 2009.

doi: 10.1519/SSC.0b013e3181a5910d

LUCIFORA, Claudio; SIMMONS, Rob. Superstar Effects in Sport: Evidence From Italian Soccer. **Journal of Sports Economics**, *[S. l.]*, v. 4, n. 1, p. 35–55, 2003. DOI: 10.1177/1527002502239657.

MENARDI, Giovanna; TORELLI, Nicola. Training and assessing classification rules with imbalanced data. **Data Mining and Knowledge Discovery**, 2014 [s.l: s.n.]. v. 28 DOI: 10.1007/s10618-012-0295-5.

ROSEN, Sherwin. The Economics of Superstars. **The American Economic Review**, [s.l], v. 71, n. 5, p. 845-858, dez. 1981.

ROTO SPORTS, INC. Rotowire, 2019. Disponível em <https://www.rotowire.com/soccer/stats.php>. Acesso em: 02 de out. de 2019

RUE, H., & SALVESEN, O(2000). **Prediction and retrospective analysis of soccer matches in a league**. Journal of the Royal Statistical Society: Series D (The Statistician), 49(3), 399-418.

SILVER, Nate. **O Sinal e o Ruído: Por que tantas previsões falham e outras não**. Rio de Janeiro: Intrínseca, 2013. Tradução de: Ana Beatriz Rodrigues e Claudio Figueiredo.

WILLIAMS, Josh. **Liverpool are using incredible data science during matches, and effects are extraordinary. Publicado em 27 de janeiro de** 2020. Disponível em: <https://www.liverpool.com/liverpool-fc-news/features/liverpool-transfer-news-jurgen-klopp-

17569689?utm_source=whatsapp.com&utm_medium=social&utm_campaign=sharebar >. Acesso em: 30 jan. 2020.

7. **Anexos**