Universidade Federal do Paraná
Setor de Ciências Exatas
Departamento de Estatística
Programa de Especialização em *Data Science* e *Big Data*

Jordy Roney Withoeft

Além do Número de Finalizações: Criação e Aplicação de um Modelo de Estimação de Gols Esperados (*xG*)

Curitiba
2020

Jordy Roney Withoeft

Além do Número de Finalizações: Criação e Aplicação de um Modelo de Estimação de Gols Esperados (*xG*)

Monografia apresentada ao Programa de Especialização em Data Science e Big Data da Universidade Federal do Paraná como requisito parcial para a obtenção do grau de especialista.

Orientador: Paulo Justiniano Ribeiro Junior

Curitiba
2020

Especialização em Data Science e Big Data
Universidade Federal do Paraná
dsbd.leg.ufpr.br

Além do Número de Finalizações: Criação e Aplicação de um Modelo de Estimação de Gols Esperados (*xG*)

Jordy Roney Withoeft[1]
Paulo Justiniano Ribeiro Junior[2]

Resumo

O futebol tem cada vez mais se tornado um esporte onde avaliações e decisões são feitas baseadas em informações geradas por dados coletados em treinos e partidas. Neste contexto, a métrica de gols esperados (*xG*) tem ganhado bastante destaque, por ser rica em contexto e avaliar a qualidade das chances criadas em uma partida de futebol de forma *data-driven*. O propósito deste projeto está em estudar o processo de criação de um modelo de estimação de gols esperados, através da realização de etapas de extração de informações de uma base pública de eventos ocorridos em diversas partidas, criação de *features* que quantifiquem aspectos relacionados à finalização e a jogada que a proporcionou, exploração dessas variáveis em diferentes modelos e aplicação do melhor modelo em diferentes análises. Neste processo, foi constatado que distância e ângulo da finalização são aspectos fundamentais, mas que outras características também contribuem para o refinamento dos resultados. Os modelos de regressão logística e XGBoost foram os que tiveram melhores performances nos testes. Por mais que ainda haja espaço de melhora, os resultados obtidos neste projeto foram satisfatórios e evidenciaram que há muito potencial para utilização da estatística de gols esperados na avaliação de performance de times e jogadores.
Palavras-chave: futebol, finalização, gols esperados, chances de gol.

Abstract

The football has become even more a sport where evaluations and decisions are made based on information generated by data collected in practice sessions and matches. In this context, the expected goals measure (xG) has gained prominence, because it is a measure rich in context and that evaluates the quality of the chances created in a football match in a data-driven way. The purpose of this project is to study the process to create a expected goal model, through the execution of steps of data extraction from a public dataset of football events from several matches, the creation of features to quantify aspects related to shots and how they happened, the exploration of those variables in different models and the application of the best model in different analysis. In this process, it was found that shot distance and angle are fundamental aspects, but other features have also contributed to better results, and that models that used logistic regression and XGBoost had the best performances. As much as there is opportunities to improve, the results obtained in this project were considered very satisfactory and showed that there is huge potential to use the expected goals measure to performance evaluation of football clubs and players.
***Keywrods**: football, shot, expected goals, goal chances.*

1 Introdução

O futebol é um esporte com uma característica muito particular, quando comparado com outros esportes coletivos de grande popularidade como o basquete, beisebol, hóquei, críquete e futebol americano: as pontuações (gols) não ocorrem com muita frequência. Neste contexto, sempre houve um forte apelo para a análise de partidas e performance de times baseadas nos números de finalizações totais e que atingiram o gol. Porém, é de conhecimento geral que há diferentes tipos de finalização e que algumas possuem mais chance de se transformarem em gol que outras.

A análise da qualidade de uma chance de gol sempre foi, historicamente, realizada de forma empírica, a partir da opinião de especialistas ou simples espectadores quanto a diversos aspectos observados em cada finalização, como o local onde o chute ocorreu, qual a parte do corpo utilizada, qual o tipo de passe ou ação que originou a chance, etc. Baseados nestes pontos e na crescente geração de dados de partidas de futebol, diversas postagens e estudos foram realizados buscando quantificar as chances de gol [1, 2, 3]. O aumento do interesse nesta ideia fez com que a métrica conhecida como *Expected Goals (xG)*, traduzida literalmente como gols esperados e com origem no hóquei [4], ganhasse importância e notoriedade também no mundo do futebol.

O *xG* é uma métrica que visa diferenciar as finalizações a partir da definição da probabilidade de cada uma ser convertida em gol, baseada no contexto de que como ela foi criada, como e onde foi realizada [5, 6]. A ideia

[1]Aluno do programa de Especialização em Data Science & Big Data, jordyrw@hotmail.com.
[2]Professor do Departamento de Estatística - DEST/UFPR, paulojus@ufpr.br.

é que ela seja uma métrica de fácil interpretação e que sua aplicação permita avaliar a performance de times e jogadores de uma forma quantitativa e *data-driven*.

Neste projeto, o objetivo principal está em criar um modelo próprio de cálculo de *xG*, a partir da análise e processamento de uma grande quantidade de dados brutos de partidas de futebol e da aplicação de conceitos e técnicas estatísticas e de *Machine Learning*. A partir da criação deste modelo, este trabalho também pretende exemplificar algumas possíveis aplicações desta métrica como ferramenta para maior compreensão dos contextos que envolvem uma partida de futebol, para utilização por times de futebol, analistas profissionais, mídias convencionais/alternativas e os amantes do futebol.

2 Metodologia

Este projeto está embasado em dados gerados pela empresa *Wyscout*, uma das maiores empresas de scouting de futebol do Mundo [7], e liberados gratuitamente para utilização em projetos acadêmicos [8]. Tratam-se de dados gerados a partir da análise de partidas profissionais de futebol e do registro de todas as ações onde a bola está envolvida, através da utilização de avançados softwares que permitem a geração de informações relacionadas aos eventos que ocorrem, tais como a localização em campo do jogador com a bola, o tipo de ação executada, características da ação, etc. A base de dados é composta por eventos da temporada 2017/2018 das ligas da primeira divisão da Alemanha, Espanha, França, Inglaterra e Itália, além da Eurocopa de 2016 e da Copa do Mundo de 2018, totalizando 1941 partidas e mais de 3 milhões de eventos, distribuídos conforme Tabela 1.

Tabela 1: Tipos de eventos registrados

Tipo Evento	# Registros	% do Total
Passe	1.665.508	51,23%
Duelo	879.083	27,04%
Outros	257.240	7,91%
Bola Parada Indireta	190.406	5,86%
Interrupção	130.097	4,00%
Falta	51.049	1,57%
Chute	43.078	1,32%
Tentativa Salva	17.619	0,54%
Impedimento	8.182	0,25%
Goleiro Saindo do Gol	6.165	0,19%
Cobrança de Falta	2.209	0,07%
Pênalti	658	0,02%

Os dados estão disponíveis em 21 arquivos no formato *JSON*, contendo, além dos dados de eventos, destacados na Tabela 1, detalhes sobre as partidas, técnicos, jogadores, times, juízes, competições e dados auxiliares, sendo, portanto, uma fonte completa para as mais variadas análises. Neste contexto, a ideia é extrair todos os chutes realizados e os eventos que o precedem, com o intuito de gerar features e variáveis que serão utilizadas na criação de um modelo de *Expected Goals (xG)*. Para tanto, será utilizado Python e algumas de suas bibliotecas mais tradicionais e populares (*Pandas, Matplotlib, Scikit-learn*, etc). A Figura 1 resume o processo que foi aplicado neste projeto.

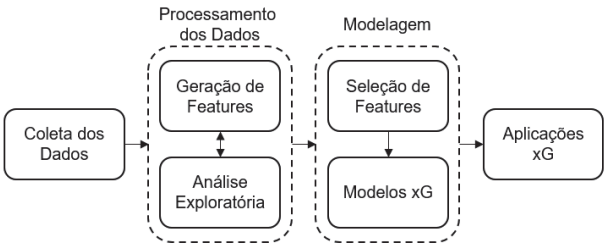

Figura 1: Processo aplicado para realização do projeto

Quanto as técnicas que serão utilizadas para gerar o modelo de *xG*, optou-se, por questões acadêmicas e de limitação de tempo, a comparação de apenas três diferentes tipos de classificadores: Regressão Logística, *Random Forest* e *XGBoost*, algoritmos já conceituados e utilizados com frequência em estudos relativos ao tema aqui discutido [9, 10]. Importante reforçar que, mesmo que sejam utilizados classificadores na fase de modelagem, o ponto principal não está na classificação em si em gol (1) ou não (0), mas na probabilidade de uma determinada finalização terminar em sucesso (gol), número este que permanecerá entre 0 e 1 e formará o valor considerado como *xG*. Devido a isso que métricas usuais de avaliação de problemas de classificação, tais como acurácia, precisão e recall, não serão consideradas na avaliação do modelo, uma vez que não há interesse na classificação em si, e sim em buscar um modelo que minimize a taxa de erro da probabilidade em relação a variável resposta. Portanto, a métrica avaliativa aplicada em diferentes etapas do projeto será o *Brier Score*, por ter seu uso recomendado para cenários como o proposto aqui [11], mas serão também apresentados os resultados obtidos com RMSE e ROC AUC, com o intuito principal de ser utilizado para comparação com outros modelos.

3 Etapas do Projeto

Nos próximos parágrafos serão apresentadas as atividades realizadas para que os dados brutos de eventos de partidas de futebol fossem convertidos em *xG* e a aplicação desta métrica em contextos reais para avaliação do seu potencial como ferramenta para maior compreensão do jogo.

3.1 Coleta dos Dados

Para simplificar as futuras etapas do projeto, foi realizada a carga dos arquivos *JSON* em um banco de dados tabular e relacional *SQLITE*. A opção por essa *engine* está no fato de ser altamente confiável, rápida e estar presente na grande maioria dos computadores, sem necessidade de configurações adicionais, o que torna simples e direto o processo de recriar a base para reprodução ou complemento das análises realizadas. A tabela de eventos, foco principal deste projeto, em seu estado puro, possui as seguintes informações: posição inicial e final de cada ação em coordenadas espaciais (x e y); código dos jogadores e times envolvidos; código da partida sendo disputada; informações temporais do momento em que o evento ocorreu; e *tags* que descrevem características da ação realizada (ex.: chute realizado com perna direita, passe incompleto, etc).

3.2 Processamento dos Dados

A partir dos dados brutos descritos anteriormente, são criadas todas as *features* que serão utilizadas na fase de modelagem e análise exploratória e visual. Esta é uma fase crítica e onde é necessário despender uma boa quantidade de esforço e tempo, pois a qualidade da informação gerada aqui está diretamente relacionada a boa parte do sucesso das etapas subsequentes, além de ser o momento em que aspectos importantes de uma partida de futebol são convertidos em dados que auxiliarão na compreensão do jogo e suas nuances. Um dos pontos principais de um modelo *xG* está relacionado a exata posição onde acontece a finalização [3, 12, 13]. Na Figura 2 é possível ter uma visão clara do padrão existente nas tentativas de chute a gol: zonas centrais do campo, com maior quantidade localizada dentro da grande área, com redução visível conforme há o distanciamento do gol.

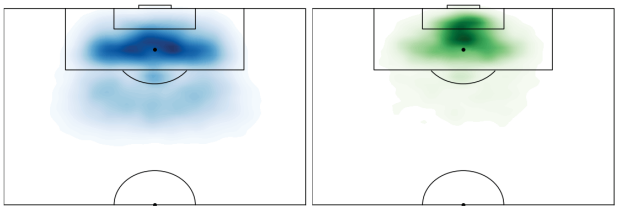

Figura 2: Distribuição dos chutes realizados (esquerda) e convertidos (direita)

A base original fornece esses dados em percentuais, onde o comprimento e a largura possuem valores de 0 a 100. Para interpretação é necessário aplicar a seguinte lógica: a coordenada x indica a posição percentual do jogador em relação ao gol de seu adversário, o que na prática significa que um jogador na grande área do oponente, em fase de ataque, está em uma posição próxima a 100%; já a coordenada y informa o quão próximo ou distante o atleta está da linha da direita do campo, o que se torna mais fácil de compreender se projetarmos que um jogador que atua pela faixa esquerda do campo, um lateral esquerdo por exemplo, terá uma boa parte de suas ações com bola reportadas com o valor de y próximo a 100% [8]. Para facilitar a interpretação e visualização desta informação, os valores foram convertidos de percentual para metros. Para tanto, foi utilizada uma dimensão de campo (105m x 68m) dentro dos limites definidos pela *IFAB* (*International Football Association Board*) como parte da regra do esporte [14] e em linha com metragens populares entre alguns dos mais famosos campos de futebol do Mundo [15]. Após a conversão supracitada, foram geradas, através da aplicação do teorema de Pitágoras, que permite o cálculo da distância entre dois pontos, três diferentes medidas de distância relacionadas aos eventos que antecedem uma finalização: distância entre o ponto inicial e final de uma ação específica; distância entre o ponto inicial e o local onde o chute foi realizado; e distância da ação em relação ao gol. Para os chutes apenas a última medida possui relevância e pode ser vista na Figura 3.

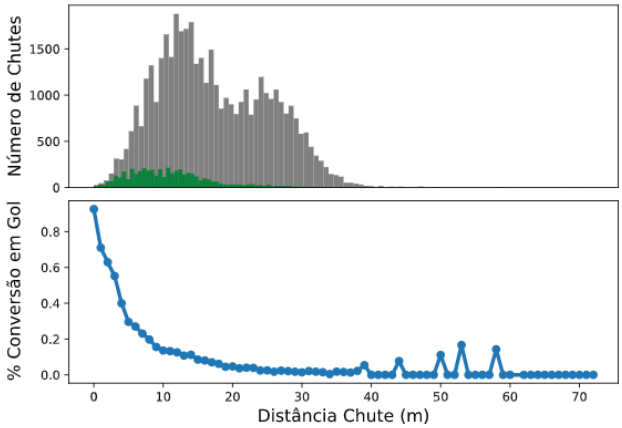

Figura 3: Distribuição dos chutes em relação à distância (esqueda) e a taxa de conversão (direita)

Além disso, a posição da finalização também pode ser avaliada pela ótica do ângulo da bola em relação ao gol (ver Figura 4), uma vez que um chute à 10 metros de frente para o gol e outro da mesma distância, mas de uma posição lateral, possuem dinâmicas diferentes. Para realização deste cálculo foi aplicada a lei dos cossenos, através da ideia de que a posição da bola e das traves geram um triângulo, onde o ângulo do canto próximo a bola é a informação utilizada como valor para esta variável.

Outro ponto importante para avaliação de uma chance de gol está relacionado a parte do corpo utilizada para realização da finalização [6, 9, 13]. Os dados em questão permitem gerar duas variáveis relacionadas a isso: se a finalização foi feita utilizando o pé ou a cabeça/outra parte do corpo menos usual; e, caso seja a primeira opção, se foi com o pé tido como melhor/preferido do jogador envolvido. É possível verificar isso na Figura 5.

O contexto da jogada que gera uma finalização também possui aspectos que influenciam na qualidade e no resultado da ação final [6, 13, 16], conforme é possível ver na Figura 6. Trabalhar os dados para que gerem informações neste sentido que sejam úteis é complexo, pois envolve analisar uma gama de eventos defensivos

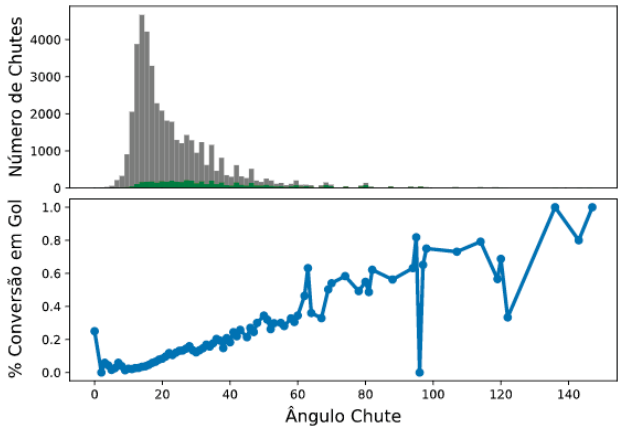

Figura 4: Distribuição dos chutes em relação ao ângulo (esquerda) e a taxa de conversão (direita)

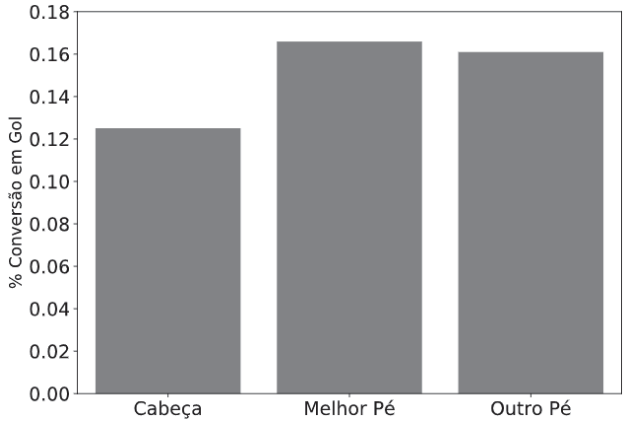

Figura 5: Taxa de conversão para diferentes partes do corpo (finalizações a menos de 20m do gol)

e ofensivos que ocorrem antes do chute em si. Um dos primeiros pontos que se destacam numa análise como essas é a ideia de que o futebol é composto por dois padrões de jogadas: as que se iniciam através de uma bola parada e as que são geradas através da bola em movimento. Para emular esta situação, foi criada uma variável qualitativa de classificação do tipo de jogada, que possui os seguintes possíveis valores: falta/lateral, escanteio, contra-ataque, rebote de chute, jogada com grande posse de bola e jogada normal.

Nesta mesma linha da análise da construção da jogada, foi verificado que as últimas duas ações antes da finalização (ir além disso não apresentou ganhos consideráveis) podem possuir características com relativa influência na probabilidade de conversão de um chute em gol. Foram gerados a partir dessa observação as seguintes variáveis: tipo de ação, local em que as ações ocorreram e, tendo sido passes, se foram classificados como *through pass* (passes nas costas dos defensores ou no espaço deixado por eles). No caso de ocorrência do último, constatou-se taxa de conversão de 20% contra 10% em situações regulares.

Um resumo das principais features geradas nesta etapa pode ser encontrado na Tabela 2.

Importante ressaltar que algumas dessas variáveis criadas, apesar de serem majoritariamente focadas no ataque, tentam também emular comportamentos defensivos, uma vez que a base de dados não possui detalhes da pressão realizada na bola, do posicionamento defensivo ou do goleiro, etc. Um exemplo disso são os *through pass*, citados anteriormente, que representam jogadas onde as ações finais foram passes com potencial maior que o normal de "quebrar" o posicionamento defensivo e a concentração dos marcadores, aumentando as chances

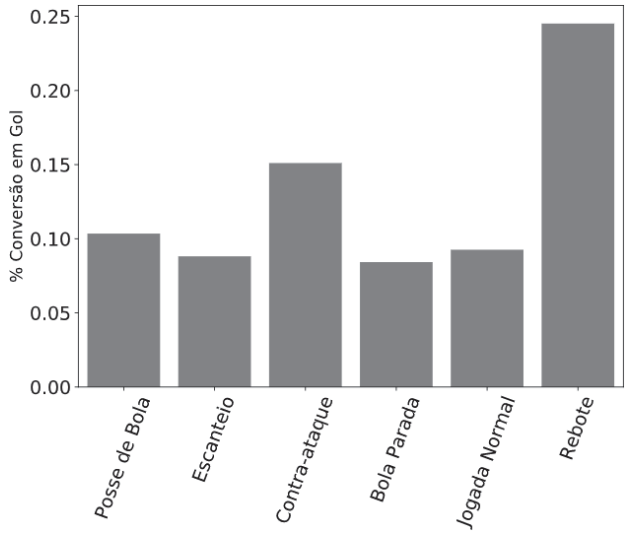

Figura 6: Taxa de conversão por tipo de jogada

Tabela 2: Resumo das features geradas (*UAC = Última Ação Chave; **PAC = Penúltima Ação Chave)

Feature	Alternativas	Tipo
Distância Chute	NA	Contínua
Ângulo Chute	NA	Contínua
Cabeça	NA	Binária
Melhor Pé	NA	Binária
Tipo de Jogada	Bola Parada Indireta; Escanteio; Contra-ataque; Posse de Bola; Jogada Regular; Rebote	Binária
Goleiro Saindo do Gol	NA	Binária
Posse Perdida em Situação Perigosa	NA	Binária
Drible/Aceleração	NA	Binária
Bola Perdida	NA	Binária
UAC*: Distância Para o Gol	NA	Contínua
UAC*: Tipo de Ação	Cruzamento; Passe Cabeça; Passe Simples; Passe Alto; Passe Inteligente; Ação Defensiva; Escanteio; Falta; Chute; Outro	Binária
UAC*: Passe Chave	NA	Binária
PAC**: Zona do Campo	14, 17, 13 ou 15, 16 ou 18	Binária
PAC**: Tipo de Ação	Cruzamento; Passe Cabeça; Passe Simples; Passe Alto; Passe Inteligente; Ação Defensiva; Escanteio; Falta; Chute; Outro	Binária
PAC**: Passe Chave	NA	Binária

de gol.

3.3 Modelagem

Pelas características dos valores obtidos para distância e ângulo do chute e distância do último passe em relação ao gol, foi decidido realizar algumas transformações nos números para averiguar qual deles iria gerar os melhores resultados quando realizado os treinos e testes dos diferentes modelos. Desta forma, estas variáveis foram transformadas das seguintes formas: transformação logarítmica; normalização dos valores entre o intervalo 0 e 1; e representação de valores em faixas não lineares, de modo a agrupar em grupos maiores os conjuntos de dados com poucas amostras. Especificamente para o modelo de Regressão Linear foram ainda testados termos quadráticos e produtos entre variáveis, com o intuito de captar possíveis interações, mas a inclusão de tais termos não resultou em melhoria no ajuste. Por fim, importante ressaltar que as variações de uma mesma variável não foram utilizadas juntas em nenhum dos testes realizados. Para checagem de quais conjuntos de *features* contribuem mais no processo de predição, foram utilizados dois métodos distintos de seleção de *features* disponíveis no pacote *Scikit-learn*: *SelectKBest* e *Recursive Feature Elimination*. Em suas próprias documentações, o primeiro é definido como um método que seleciona as *k features* com maior score no teste estatístico univariado definido pelo usuário [17], enquanto que o segundo aplica um estimador para recursivamente eliminar *features* até que seja alcançado o conjunto de variáveis que apresenta o melhor score [18]. Uma série de possibilidades de *features* foram geradas na etapa de seleção e aplicadas nos algoritmos de classificação definidos (Regressão Logística, *Random Forest* e *XGBoost*). Os melhores valores de *Brier Score* obtidos nos testes preliminares foram executados novamente com diferentes hiperparâmetros para otimização dos resultados. Os valores obtidos para as melhores combinações de *features* e hiperparâmetros podem ser visualizados na Tabela 3. Para comparação, foi realizada a avaliação de um modelo "ingênuo", onde todas as finalizações receberam a mesma probabilidade (10,43%), que é o percentual médio de gols considerando toda a base.

Tabela 3: Resultados obtidos na fase de modelagem

Método	# Features	Brier Score	RMSE	ROC AUC
Ingênuo	NA	0,0915	0,3025	0,5000
Regressão Logística	34	0,0767	0,2769	0,8011
Random Forest	41	0,0773	0,2781	0,7982
XGBoost	22	0,0768	0,2771	0,7997

Os resultados indicam que os modelos gerados são todos superiores ao teste "ingênuo" e aos testes com modelos apenas utilizando distância e ângulo. Os valores obtidos nas diferentes métricas de avaliação exibidas estão em linha com resultados obtidos em outros artigos sobre o tema [19, 20, 21]. Os melhores modelos de Regressão Logística e *XGBoost* obtiveram resultados muito próximos, o que é indicativo de que ambos poderiam ser aplicados para análise. Pela maior simplicidade e interpretabilidade, o modelo de Regressão Logística foi o escolhido. Na Figura 7 é possível ver uma representação gráfica da métrica *xG* aplicada e como as *features* adicionadas que não são relacionadas a localização adicionam um nível de "caos" a métrica, apesar de ainda existir um padrão claro relacionado a proximidade com o gol (distância + ângulo).

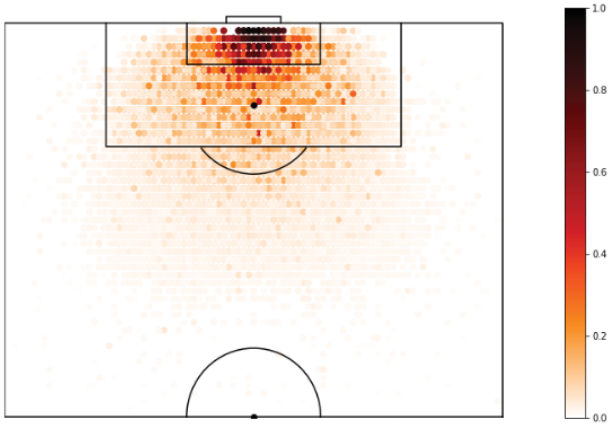

Figura 7: Mapa gráfico do local onde as finalizações aconteceram e *xG* de cada uma

Quanto as *features* do melhor modelo, foram utilizadas 34 variáveis do total de 41 disponíveis, que estão resumidas na Tabela 4.

Tabela 4: Resumo das features utilizadas pelo melhor modelo (*UAC = Última Ação Chave; **PAC = Penúltima Ação Chave)

Variável	Valores	Tipo
Distância Chute	Log(dist em metros)	Contínua
Ângulo Chute	Log(âng em graus)	Contínua
É Cabeça?	Sim ou Não	Binária
É Melhor Pé?	Sim ou Não	Binária
Tipo de Jogada	Bola Parada; Escanteio; Contra-ataque; Jogada Regular; Rebote	Binária
Posse Perdida em Situação Perigosa	Sim ou Não	Binária
Drible/Aceleração	Sim ou Não	Binária
Bola Perdida	Sim ou Não	Binária
UAC*: Distância Para o Gol	Faixas(dist em metros)	Discreta
UAC*: Tipo de Ação	Cruzamento; Passe Cabeça; Passe Simples; Passe Inteligente; Escanteio; Chute; Outro	Binária
UAC*: Passe Chave	Sim ou Não	Binária
PAC**: Zona do Campo	14, 17, 13 ou 15, 16 ou 18	Binária
PAC**: Tipo de Ação	Cruzamento; Passe Cabeça; Passe Simples; Passe Alto; Passe Inteligente; Ação Defensiva; Chute; Outro	Binária
PAC**: Passe Chave	Sim ou Não	Binária

O processo descrito nos parágrafos anteriores foi também aplicado para todas as finalizações oriundas de cobranças de faltas diretas, porém em um cenário muito

mais simples, onde apenas distância e ângulo foram considerados. O modelo de Regressão Logística também obteve melhores resultados para estas finalizações. Já para os pênaltis, por todas as finalizações serem da mesma posição e condições similares, não considerando aqui possíveis fatores psicológicos (ex.: pressão pelo momento da partida) ou físicos (ex.: cansaço devido ao tempo já transcorrido de partida), optou-se por aplicar para todas as cobranças um único valor de *xG*, baseado na taxa média de conversão de finalizações deste tipo na base de dados considerada: 72,5%. Todas essas probabilidades foram incorporadas a base principal e analisados na etapa subsequente.

3.4 Aplicação do Modelo

Uma vez que o modelo é criado, a sua aplicação prática consiste em associar a probabilidade de gol gerada pelo modelo para cada finalização realizada. Apesar de ser possível analisar de forma individual, registro a registro, o ganho analítico está na soma destas probabilidades e na consequente estimação de quantos gols eram esperados. Neste contexto, a aplicação mais direta está em totalizar os gols realizados pelos times analisados e compará-los com a quantidade esperada a partir das chances criadas. Através desta análise, é possível ter um indicativo se um determinado time está criando um número considerado de chances e se o grau de eficiência na conversão das finalizações está abaixo ou acima do valor esperado. Na Figura 8 é possível observar esta relação e como há um padrão, de certa forma já esperado, de que times que geram uma maior quantidade de chances são também os times que mais a convertem e são, consequentemente, os que terminam em melhores posições.

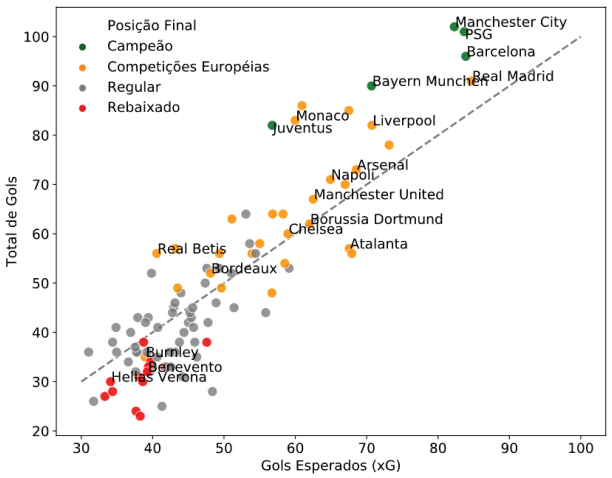

Figura 8: Gráfico de total de gols feitos x total de gols esperados (*xG*)

Da mesma forma em que é possível avaliar o sistema ofensivo dos times, a métrica de gols esperados também permite analisar sistemas defensivos, a partir da aplicação da mesma lógica, porém através de uma perspectiva diferente: analisar o nível de periculosidade das finalizações que o time permite que sejam realizadas e qual o grau de eficiência na contenção destas chances. A Figura 9 mostra como foram as performances dos clubes contidos na base de dados para esta visão.

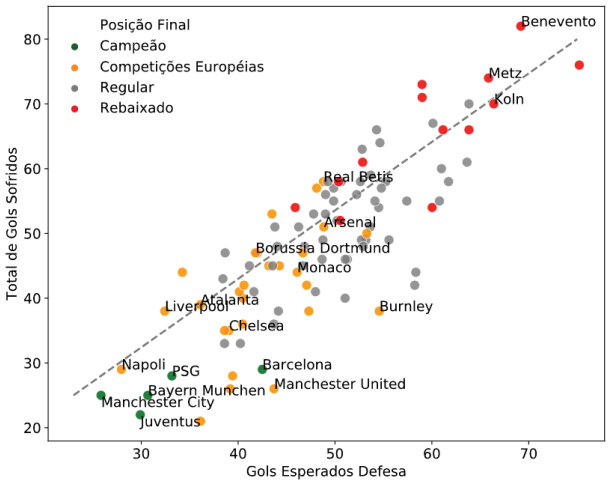

Figura 9: Gráfico de total de gols sofridos x total de gols esperados (*xG* Defesa)

A melhor forma de avaliar a performance geral de um clube de futebol e de seus adversários, está em cruzar as duas visões apresentadas nas Figuras 8 e 9, buscando com isso observar quais clubes conseguem alcançar o equilíbrio necessário entre as duas fases principais do jogo: ataque e defesa. A Figura 10 é um exemplo disso e detalha o cruzamento entre o valor médio por partida de gols esperados defensivos e ofensivos e permite observar alguns padrões interessantes, por mais que apenas uma temporada tenho sido considerada: todos os clubes rebaixados tiveram médias de *xG* ofensivo menor que o defensivo e apenas três clubes nesta mesma condição se classificaram para competições europeias.

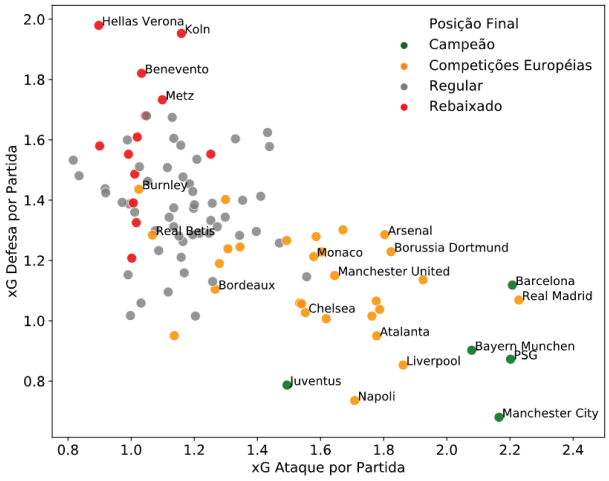

Figura 10: Gráfico da média *xG* Defesa por partida x *xG* Ataque por partida

É interessante observar nos gráficos anteriores, especialmente no ilustrado na Figura 10, a performance dos

times italianos Juventus e Napoli. Olhando apenas para os clubes da liga italiana, estes times destoaram dos demais quanto ao nível de equilíbrio de suas performances defensivas e ofensivas, porém, por mais que o Napoli tenha tido melhores médias por partida, a Juventus acabou se sagrando campeã. Isto ocorreu devido a excelente defesa e um ataque assustadoramente eficiente na transformação de chances em gols, tendo a maior diferença entre gols esperados e gols realizados da base de dados analisada: cerca de 25 gols a mais. Destrinchando os números ofensivos da Juventus na temporada (Tabela 5), é possível observar que a equipe teve um dos ataques que mais fizeram gols na Itália e na base de dados, mas teve desempenho apenas mediano quanto a criação de chances por partida, média de xG por chute e criação/aproveitamento de grandes chances ($xG > 0{,}5$). A questão é a excelente performance em boas chances ($xG <= 0{,}5$ e $xG > 0{,}2$), mas, especialmente, nas chances regulares ou não tão boas ($xG < 0{,}2$), onde o time teve rendimento muito melhor que os demais na liga italiana. Isso ocorre devido a excelente performance em jogadas normais, onde o time marcou 7 gols de chutes de fora da área, e em jogadas de bola parada, onde o clube conseguiu marcar incríveis 26 gols, maior número da base de dados.

Tabela 5: Detalhamento do ataque da Juventus no campeonato italiano na temporada 2017/2018

Estatística	Total	Posição BD	Posição Itália
Gols	82	10º	2º
Chutes	530	19º	8º
xG por partida	1,49	25º	8º
xG por chute	0,107	59º	10º
xG vs Gols ($xG > 0{,}5$)	-0,6	66º	13º
xG vs Gols ($xG > 0{,}2$)	5,3	11º	3º
xG vs Gols ($xG > 0{,}1$)	10,1	1º	1º
xG vs Gols ($xG < 0{,}1$)	10,4	2º	1º
xG vs Gols (jogada regular)	13,4	6º	2º
xG vs Gols (bola parada)	12,0	1º	1º

Ainda olhando para as performances de Juventus e Napoli, a métrica xG também pode ser utilizada para avaliação da performance dos times durante as rodadas de um campeonato. Na Figura 11, onde a área preenchida representa a diferença entre xG do ataque e defesa (verde = ataque > defesa; Vermelho = ataque < defesa) e a linha preta a diferença entre gols feitos e sofridos, é possível observar como a Juventus foi um time que oscilou bastante e teve algumas performances conservadoras quanto as chances criadas e oferecidas, com 8 partidas com xG ofensivo acima de 2 gols, 9 partidas onde permitiu que o adversário criasse mais chances que ela (perdeu apenas 1 dessas partidas) e 9 partidas onde a diferenças de xG's foi menor do que 0,5 (também perdeu apenas 1), porém, a sua alta eficiência ofensiva, brevemente detalhada anteriormente, lhe garantiu diversas vitórias nas situações relatadas. Já o Napoli apresentou maior consistência e bom nível por cerca de 2/3 do campeonato, mas visivelmente o time diminuiu a sua eficiência a partir da 27ª rodada, onde nas últimas 11 rodadas perdeu apenas um confronto na "disputa entre xG's", mas viu seus números oscilaram bastante e acabou contabilizando 2 derrotas e 3 empates, o que permitiu que a Juventus assumisse a liderança e conquistasse o título.

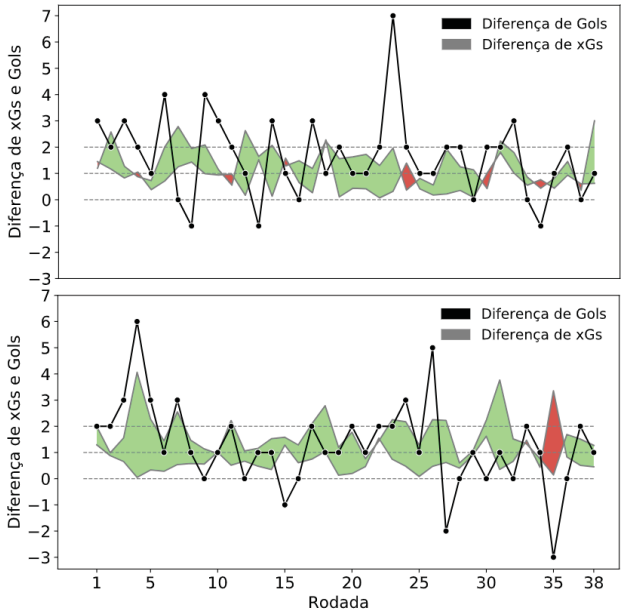

Figura 11: Gráfico da diferença de gols e xG's de Juventus e Napoli no transcorrer do campeonato italiano da temporada 2017/2018

É interessante observar com mais detalhe a partida do Napoli na 27ª rodada, que foi um marco da queda da performance do time na temporada. Nesta rodada o Napoli perdeu para a Roma por 4 a 2, mas quando olhamos na Figura 12, que mostra a evolução do xG de cada time durante toda a partida, é possível observar que a Roma teve um xG acumulado abaixo de 0,5 e mesmo assim conseguiu converter chances de baixa probabilidade em 4 gols, enquanto que o Napoli teve xG acumulado próximo a 2,5 e os transformou em 2 gols, próximo ao esperado, mas abaixo do total de seu adversário.

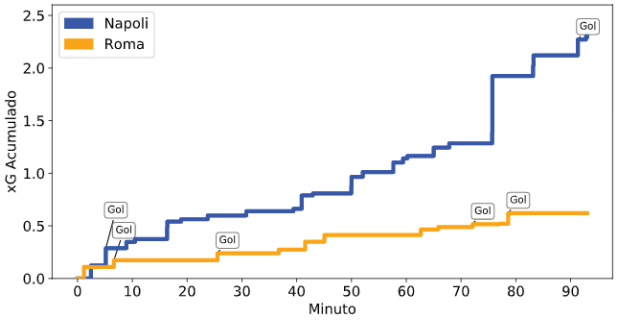

Figura 12: Gráfico de evolução do xG na partida entre Napoli (2) e Roma (4)

A métrica de gols esperados pode também ser apli-

cada no contexto individual, utilizando a mesma lógica citada anteriormente, mas totalizando as probabilidades das chances criadas na ótica dos jogadores que as finalizaram. Na Tabela 6 é possível verificar os 10 jogadores que mais gols fizeram na temporada 2017/2018.

Tabela 6: Top 10 jogadores que mais fizeram gols na temporada 2017/2018

Jogador	Gols	*xG*	Gols vs *xG*
L. Messi	34	25,10	8,90
M. Salah	32	20,89	11,11
R. Lewandowski	29	25,15	3,85
C. Immobile	29	19,66	9,34
H. Kane	29	25,98	3,02
M. Icardi	29	19,39	9,61
E. Cavani	28	22,58	5,42
C. Ronaldo	26	25,96	0,04
L. Suarez	25	23,44	1,58
P. Aubameyang	23	20,12	2,88

Analisando as performances individuais, é interessante constatar os incríveis números obtidos pelo jogador Mohamed Salah, do Liverpool, que marcou 32 gols, cerca de 11 gols a mais do que as chances obtidas indicavam. Esta performance impactante acabou influenciando na escolha de Salah como o 3ª melhor jogador do Mundo no ano de 2018. As suas finalizações estão destrinchadas na Tabela 7.

Tabela 7: Detalhes das finalizações de Salah pelo Liverpool na temporada 2017/2018

Estatística	Total	Posição BD
Gols	32	2º
Chutes	142	6º
Chutes p/ 1 Gol	4.44	21º
xG	20,89	7º
xG por chute	0,15	9º
xG vs Gols (*xG* > 0,5)	0,15	> 50º
xG vs Gols (*xG* > 0,2)	6,81	1º
xG vs Gols (*xG* > 0,1)	-0,19	> 50º
xG vs Gols (*xG* < 0,1)	4,34	7º

Necessário também destacar que Salah foi o 5º em número de finalizações em situação de boas chances ($xG <= 0,5$ e $xG > 0,2$) e com uma boa taxa de conversão (15/27 55%), o que indica que ele atua em um time que lhe possibilita boas chances durante a temporada e que ele é um atacante que costuma aproveitá-las. Por não ser um atacante propriamente de área, Salah teve apenas 4 grandes chances ($xG > 0,5$), mas foi o 3º jogador da base que mais fez gols em chances de baixa probabilidade ($xG < 0,1$), com 8 gols em 66 chutes (12%). O mapa das finalizações de Salah, apresentado na Figura 13, dá uma boa ideia de como os números descritos ficam dispostos dentro de um campo de futebol.

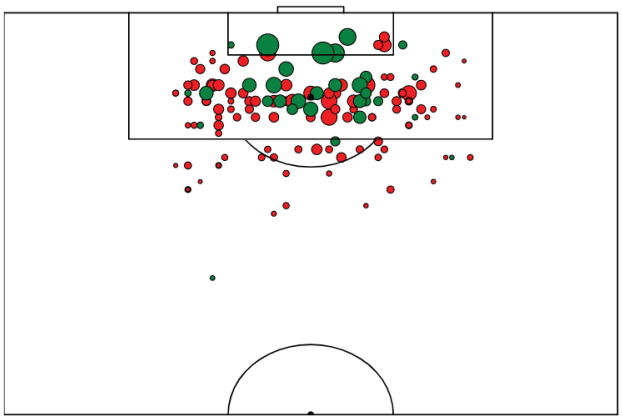

Figura 13: Mapa das finalizações de Salah sem considerar pênaltis

Além da aplicação para análise direta das finalizações, que é muito focada nos atacantes, existem adaptações que utilizam a métrica *xG* para avaliar performances por outras perspectivas, que não serão tratadas em maior detalhe neste artigo, como por exemplo na análise de: jogadores responsáveis por dar assistências (ex.: meio-campistas); goleiros e a contenção das finalizações que vão em direção ao gol; e todos os jogadores presentes em uma partida quanto a sua contribuição na construção de jogadas, sem olhar unicamente para o último passe.

4 Conclusão

Este trabalho descreveu as etapas da implementação de um modelo capaz de estimar a probabilidade de um chute se tornar gol, a partir de uma série de *features* relacionadas a finalização e aos eventos que a antecedem. Ficou claro durante o processo que distância e ângulo são características fundamentais para este tipo de análise. Outros aspectos relacionados ao contexto da jogada também possuem alguma influência, auxiliam a refinar o resultado e são importantes pontos à serem considerados na análise pós aplicação do modelo.

A métrica de gols esperados (*xG*) tem se tornado bastante popular no mundo do futebol, especialmente por carregar consigo mais detalhes sobre as finalizações feitas em uma partida, por um time ou por um jogador de futebol, quando comparadas com a utilização apenas de métricas de medida diretas, como, por exemplo, o número de finalizações. Os números gerados não devem ser vistos como definitivos e sua utilização deve se somar a outras métricas de avaliação e, especialmente, à observação visual do jogo. A combinação de formas de avaliação é recomendável, uma vez que o futebol é um esporte complexo, repleto de nuances que ainda não são totalmente captadas e convertidas em números.

Há muitas possibilidades para extender o modelo criado e as análises apresentadas. Por exemplo, uma sugestão para próximos trabalhos está em utilizar uma base de dados que apresente informações relativas aos demais

jogadores próximos às jogadas, uma vez que é sabido que aspectos de ocupação de espaços, concentração de jogadores em frente à bola e posicionamento do goleiro também são fatores que influenciam a qualidade de uma chance.

Referências

[1] GREEN, S. Assessing the performance of Premier League goalscorers. **OptaPro**, 2012. Disponível em: https://opta.kota.co.uk/news-analysis/assessing-the-performance-of-premier-league-goalscorers/. Acessado em: 22 de Ago. de 2020.

[2] RATHKE, A. An examination of expected goals and shot efficiency in soccer. Journal of Human Sport and Exercise, vol. 12, núm. 2, Março, 2017, p. 514-529. Disponível em: https://www.redalyc.org/pdf/3010/301052437005.pdf. Acessado em: 10 de Ago. de 2020.

[3] LUCEY, P.; BIALKOWSKI, A.; MONFORT, P.; MATTHEWS, I. "Quality vs Quantity": Improved Shot Prediction in Soccer using Strategic Features from Spatiotemporal Data. 9th Annual MIT Sloan Sports Analytics Conference, 2015. Disponível em: http://www.sloansportsconference.com/wp-content/uploads/2015/02/SSAC15-RP-Finalist-Quality-vs-Quantity.pdf. Acessado em: 10 de Ago. de 2020.

[4] RYDER, A. Shot Quality: a methodology for the study of the quality of a hockey team's shots allowed. **Hockey Analytics**, 2004. Disponível em: http://hockeyanalytics.com/Research_files/Shot_Quality.pdf. Acessado em: 10 de Ago. de 2020.

[5] TRAINOR, C. Goal Expectation and Efficiency. **StatsBomb**, 2013. Disponível em: https://statsbomb.com/2013/08/goal-expectation-and-efficiency/. Acessado em: 18 de Ago. de 2020.

[6] BRECHOT, M.; FLEPP, R. Dealing With Randomness in Match Outcomes: How to Rethink Performance Evaluation in European Club Football Using Expected Goals. Journal of Sports Economics, vol. 21, núm. 4, p. 514-529, Janeiro, 2020. Disponível em: https://journals.sagepub.com/doi/abs/10.1177/1527002519897962. Acessado em: 18 de Ago. de 2020.

[7] BAILEY, G. What is Wyscout? **SkySports**, 2014. Disponível em: https://www.skysports.com/football/news/11662/9066763. Acessado em: 18 de Ago. de 2020.

[8] PAPPALARDO, L.; CINTIA, P.; ROSSI, A. et al. A public data set of spatio-temporal match events in soccer competitions. Sci Data 6, núm. 236, Outubro, 2019. Disponível em: https://www.nature.com/articles/s41597-019-0247-7. Acessado em: 20 de Maio de 2020.

[9] EGGELS, H.; VAN ELK, R.; PECHENIZKIY, M. Explaining soccer match outcomes with goal scoring opportunities predictive analytics. CEUR Workshop Proceedings, vol. 1842, 3rd Workshop on Machine Learning and Data Mining for Sports Analytics (MLSA 2016), Setembro, 2016. Disponível em: http://ceur-ws.org/Vol-1842/paper_07.pdf. Acessado em: 18 de Ago. de 2020.

[10] BRANSEN, L.; VAN HAAREN, J.;VAN DE VELDEN, M. Measuring soccer players' contributions to chance creation by valuing their passes, Journal of Quantitative Analysis in Sports, vol. 15, núm. 2, p. 97-116, Fevereiro, 2019. Disponível em: https://www.degruyter.com/view/journals/jqas/15/2/article-p97.xml. Acessado em: 18 de Ago. de 2020.

[11] RUFIBACH, K. Use of Brier score to assess binary predictions. Journal of Clinical Epidemiology, vol. 63, núm. 8, p. 938-939, Agosto, 2010. Disponível em: https://www.jclinepi.com/article/S0895-4356(09)00363-1/fulltext. Acessado em: 18 de Ago. de 2020.

[12] POLLARD, R.; ENSUM, J.; TAYLOR, S. Estimating the probability of a shot resulting in a goal: The effects of distance, angle and space. International Journal of Soccer and Science, vol. 2, núm. 1, Janeiro, 2004. Disponível em: https://www.docplayer.es/17986422-Estimating-the-probability-of-a-shot-resulting-in-a-goal-the-effects-of-distance-angle-and-space.html. Acessado em: 18 de Ago. de 2020.

[13] xG stats explained: the science behind Sportec Solutions' Expected goals model. **Bundesliga**, 2019. Disponível em: https://www.bundesliga.com/en/bundesliga/news/expected-goals-xg-model-what-is-it-and-why-is-it-useful-sportec-solutions-3177. Acessado em: 22 de Ago. de 2020.

[14] Laws of the game 2020/2021: The Field of play. **IFAB**, 2020. Disponível em: https://www.theifab.com/laws/chapter/21/section/18/. Acessado em: 22 de Ago. de 2020.

[15] Are All Football Pitches The Same Size? **Football Stadiums**, 2020. Disponível em: https://www.football-stadiums.co.uk/articles/are-all-football-pitches-the-same-size/. Acessado em: 22 de Ago. de 2020.

[16] FAIRCHILD, A.; PELECHRINIS, K.; KOKKODIS, M. Spatial analysis of shots in MLS:A model for expected goals and fractal. Journal of Sports Analytics 4, vol. 4. Disponível

em: https://content.iospress.com/download/journal-of-sports-analytics/jsa207?id=journal-of-sports-analytics%2Fjsa207. Acessado em: 22 de Ago. de 2020.

[17] RFE. **Scikit Learn**, 2020. Disponível em: https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFE.html. Acessado em: 25 de Ago. de 2020.

[18] Univariate feature selection. **Scikit Learn**, 2020. Disponível em: https://scikit-learn.org/stable/modules/feature_selection.html#univariate-feature-selection. Acessado em: 25 de Ago. de 2020.

[19] DAVIS, J.; ROBBERECHTS, P. How the data availability affects the ability to learn good xG models. **DTAI Sports Analytics Lab**. Disponível em: https://dtai.cs.kuleuven.be/sports/blog/how-data-availability-affects-the-ability-to-learn-good-xg-models#fn-1. Acessado em: 30 de Ago. de 2020.

[20] GELADE, G. Assessing Expected Goals Models. **Business Analytic**, 2017. Disponível em: https://business-analytic.co.uk/blog/evaluating-expected-goals-models/. Acessado em: 30 de Ago. de 2020.

[21] DECROSS, T.; DZYUVA, V.; VAN HAAREN, J.; DAVIS, J. Predicting Soccer Highlights from Spatio-temporal Match Event Streams.Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17), p. 1302-1308, Fevereiro, 2017. Disponível em: https://dl.acm.org/doi/10.5555/3298239.3298430. Acessado em: 30 de Ago. de 2020.