**Previsão do desempenho dos clubes da Premier League utilizando modelos lineares generalizados**

**RESUMO**

A proposta deste artigo é prever a pontuação que os times vão obter no final de um campeonato de pontos corridos, com base nas características quantitativas (scouts) de cada equipe da temporada anterior. O campeonato escolhido foi a Premier League, visando prever a temporada atual (2019/20) com base nos resultados do campeonato passado (2018/19). Deste modo, o objetivo principal é apresentar um bom modelo a partir dos dados coletados, que possa prever o resultado final da PL 19/20 com o menor erro possível, além de gerar insights a partir de todos os processos analíticos, desde a exploração dos dados, passando pelas etapas de modelagem, chegando até as predições e interpretações do modelo. As métricas adotadas para que o modelo seja adequado e consequentemente as previsões sejam eficientes serão: a ocorrência de um bom ajuste nos dados de modelagem (PL 18/19), mensurado pelo coeficiente de determinação ajustado, e um ótimo desempenho na previsão da (PL 19/20), calculado pela raiz do quadrado médio do erro (RMSE) e pelo erro médio absoluto percentual (MAPE).

Palavras-Chave. Modelagem. Futebol. Previsão. Regressão.

**ABSTRACT**

The purpose of this article is to predict the score that teams will obtain at the end of a running points championship, based on the quantitative characteristics (scouts) of each team from the previous season. The league chosen was the Premier League, aiming to predict the current season (2019/20) based on the results of the past championship (2018/19). Thus, the main objective is to present a good model from the collected data, which can predict the final result of PL 19/20 with the least possible error, in addition to generating insights from all analytical processes, from the exploration of data, passing through the modeling steps, reaching the predictions and interpretations of the model. The metrics adopted for the model to be adequate and consequently the forecasts are efficient will be: the occurrence of a good fit in the modeling data (PL 18/19), measured by the adjusted determination coefficient, and an excellent performance in the forecast of (PL 19/20), calculated by the root of the mean square of the error (RMSE) and by the percentage absolute mean error (MAPE).

Key words. Modeling. Soccer. Forecast. Regression.

**INTRODUÇÃO**

Dentre as principais ligas do futebol do mundo a Premier League é sem dúvidas um dos grandes destaques, se não o maior. Em uma análise feita pelo Globo Esporte em 2018, comparando as 5 maiores ligas europeias (Inglesa, Espanhola ; Italiana, Francesa e Alemã) podemos perceber essa dominância. Segundo a pesquisa, a Premier League é melhor em relação as outras ligas nos seguintes critérios: faturamento dos clubes, chegando a totalizar cerca de 4,4 bilhões de euros dentre os 30 clubes que mais faturam no mundo; valor de mercado dos jogadores, obtendo cerca de 6,4 bilhões de euros; salários, dispondo de mais igualdade entre os atletas; variedade, contabilizando mais clubes diferentes que foram campeões nas últimas 5 temporadas; equilíbrio, mensurando a menor diferença média do número de pontos entre o primeiro e o quarto colocado. Além de ficar em segundo lugar na quantidade de mais jogadores estrelas, apresentando 45 atletas premiados pela FIFA, Guardian (jornal diário britânico) e FourFourTwo (revista especializada em futebol do Reino Unido), e média de público, ficando atrás apenas do campeonato alemão . (retirar detalhes de cada critério ?)

Entretanto, podemos afirmar que além de ser uma das melhores ligas da Europa, a Premier League é também a melhor do mundo. Segundo a IFFHS (Federação Internacional de História e Estatísticas do Futebol) em ranking feito em 2020 com base no desempenho dos clubes e das ligas no ano anterior, temos que o campeonato inglês lidera com 1287 pontos, seguido dos campeonatos brasileiro, espanhol e italiano com 1165, 1141 e 931 pontos, respectivamente. Logo, há uma larga vantagem entre a primeira e a quarta melhor liga, comprovando a supremacia da liga inglesa em relação aos demais campeonatos do mundo e desbancando campeonatos disputados e tradicionais como o brasileiro e o espanhol.

De outro modo, com a chegada da revolução digital acompanhada do Big Data, das mais diversas tecnologias e do fácil acesso às informações, por consequência os dados se tornam cada vez mais acessíveis e tão diversificados quanto se possa imaginar. (Descrever sobre a importância das estatísticas para o futebol no contexto de análise de desempenho). Desta forma, podemos obter diversas informações sobre os campeonatos, times, atletas e entre outros segmentos que definem o futebol atualmente nas áreas de desempenho, marketing e faturamento. Transformar essas informações em conhecimento é fundamental e uma das áreas que mais cresce é a de análise de desempenho.

Com a evolução do futebol mundial os jogos e campeonatos são decididos por detalhes minuciosos, por conta do progresso das equipes que propõe disputas mais acirradas. E isso tudo também se deve a revolução digital e no quanto podemos nos beneficiar dela, tornando as informações dos times adversários mais acessíveis, otimizando as análises dos clubes dentro do campo, mensurando as performances coletivas e individuais dos atletas e das estratégias propostas, e fora do campo, desde a melhora do condicionamento físico dos jogadores até a identificação de padrões de determinados campeonatos.

Desta forma, a ideia de poder prever resultados de uma liga com base nas estatísticas dos clubes participantes é bastante útil, pois além de poder projetar o desempenho final será possível também inferir quais são as variáveis atuais que tem mais significância para a eficiência de um time neste campeonato. É claro que essas variáveis poderão ser diferentes para cada campeonato em diferentes épocas, sempre de acordo com conjunto de dados coletado. Outros fatores como as escolhas feitas durante o processo de modelagem, as características de jogo dos clubes da liga e entre outras situações também definirão essas variáveis. Logo, o papel desse artigo é definir um bom conjunto de características que definam o rendimento dos times no contexto atual da Premier League.

**MATERIAIS E MÉTODOS**

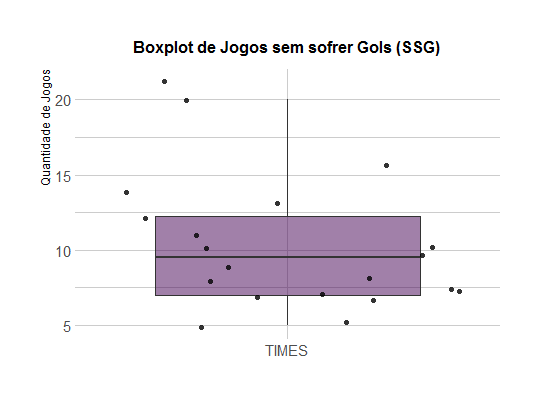
As variáveis que vão compor o dataset inicial foram coletadas no site oficial da Premier League, que são as seguintes: B (Bloqueios), CC (Cruzamentos Certos), D (Desarmes), DF (Defesas do Goleiro), DRC (Dribles Certos), F (Finalizações), FC (Finalizações Certas), GM (Gols Marcados), GS (Gols Sofridos), MC (Cortes), MI (Interceptações), P (Passes), PB (Posse de Bola), PC (Passes Certos), PL (Passes Longos), PP (Perca de Posse), PR (Passes Regressores), PTS (Pontos), SSG (Jogos sem sofrer Gols) ,T (Toques na Bola) e TIME (Nome do Time). Essas variáveis apresentam fundamentos futebolísticos com potencial para explicar a pontuação dos times ao final da competição. As variáveis restantes do site foram descartadas, pois no contexto do futebol não tem influência no desempenho de uma equipe tais como: cartões amarelos, cartões vermelhos, impedimentos, escanteios e entre outras.

É importante ressaltar que durante o processo de análise e modelagem foram criadas novas variáveis a partir das previamente selecionadas, em busca de mensurar a eficiência de aspectos importantes ofensivos, defensivos e de transições ofensivas e defensivas. Essas variáveis produzidas visam compreender os dados com mais efetividade e atingir um melhor desempenho na etapa de modelagem, e elas são: BICD (Volume das Ações Defensivas: soma de B, MI, MD e C), DFXGS (Taxa de Defesas do Goleiro por Gol Sofrido), PPFC (Passes Progressores: utilizando P, PC, PR), PXGM (Taxa de Passes por Gol Marcado), PXPR (Taxa de Passes por Passe Regressor), TXFC (Taxa de Toques na Bola por Finalização Certa) e TXGM (Taxa de Toques na Bola por Gol Marcado).

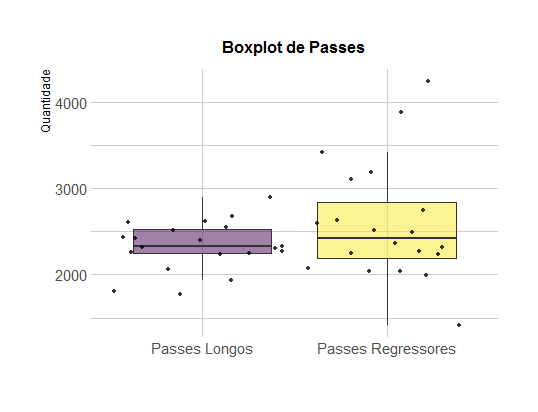
Deste modo, será realizada uma análise exploratória dos dados e em seguida entraremos na etapa de modelagem em busca de determinar bons modelos que possam explicar a pontuação dos times em função dos scouts coletados. Desta forma, poderemos analisar a avaliação, viabilidade e performance de cada modelo e escolher o ideal para realizarmos as predições e interpretações.

**RESULTADOS**

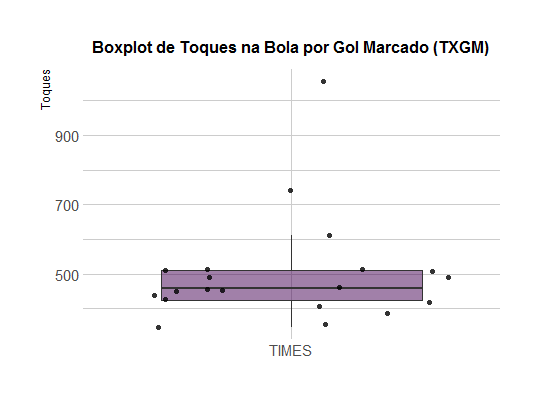
Realizando uma análise dos outliers, foram identificadas ocorrências em 5 times: Manchester City (GM,FC,PR,PL,F), Liverpool (SSG, DFXGS), Chelsea (PL), Burnley (B), Huddersfield (TXGM,PXGM,FXGM) e Fulham (PXGM,TXGM). Estes outliers foram avaliados novamente no banco de dados e de fato não são resultantes de erros de mensuração ou de coleta dos dados, logo há uma explicação natural sobre eles. Veja alguns dos boxplots a seguir:



Em relação aos outliers referentes às variáveis GM, FC, F, SSG e DFXGS, evidencia-se o alto nível que o Manchester City e o Liverpool apresentaram em comparação com os demais times nos aspectos ofensivos e defensivos, respectivamente.

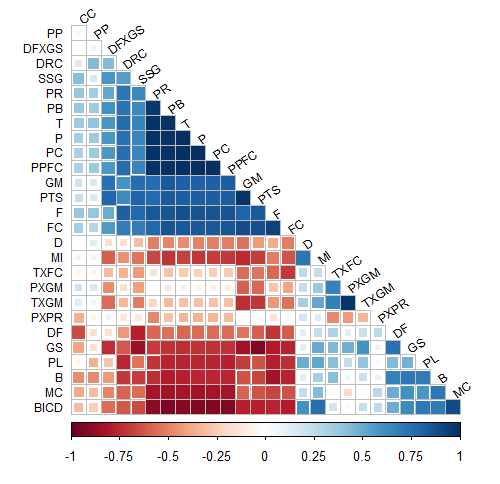


Nas variáveis PR e PL temos o reflexo do estilo de jogo do Manchester City que preserva bastante a posse de bola resultando na ocorrência de mais passes regressores para a reconstrução das jogadas, além de utilizar bem menos passes longos em comparação com as demais equipes, aderindo mais aos passes curtos. A equipe do Chelsea também possui essa característica de poucos passes longos.



Nos aspectos de B, TXGM, PXGM e FXGM, todos representam a inefetividade das equipes do Burnley, Huddersfield e Fulham, onde a primeira demonstra mais frequência nas fases defensivas e assim realiza mais bloqueios que os clubes restantes, e os dois restantes mostram a ineficiência do time com a posse de bola em relação a criação de oportunidades e conclusões de gols. Desta forma a decisão será manter essas observações, pois além de serem poucas, elas aconteceram de forma natural e podem ajudar a captar comportamentos de equipes semelhantes em dados futuros.

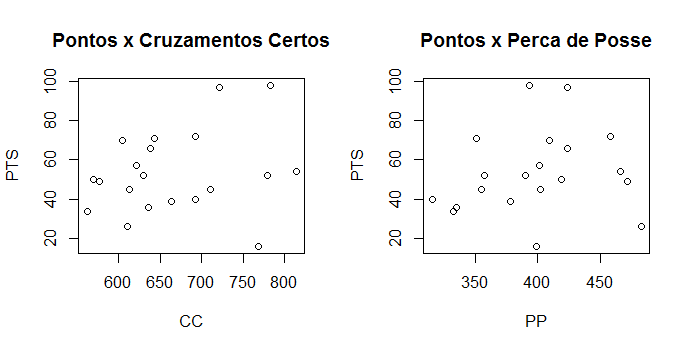
Realizando uma análise de correlação entre as variáveis, obtemos:

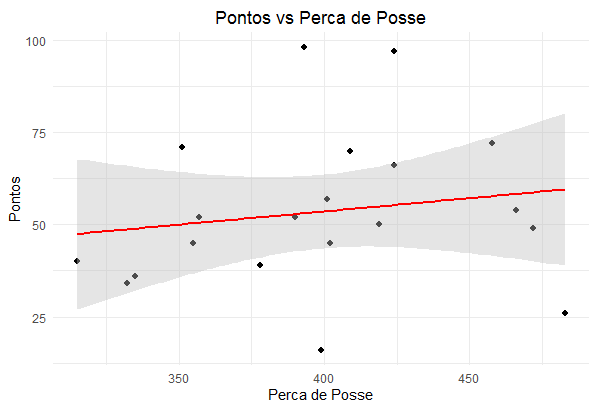


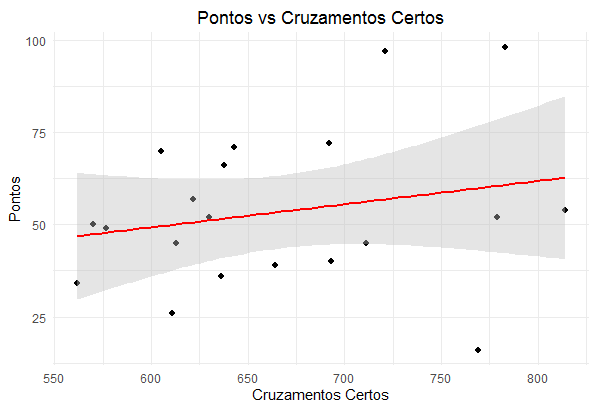
Percebemos que apenas as variáveis dependentes PXPR, CC e PP possuem fraca correlação com a independente, sendo insignificativas para a etapa de modelagem, deste modo removeremos elas posteriormente. Por outro lado, há muitos indícios de fortes correlações entre variáveis dependentes, tornando necessária uma análise mais detalhada sobre elas, a fim de eliminar variáveis semelhantes que terão a mesma significância no futuro modelo preditivo. A seguir especificaremos a análise utilizando as correlações em conjunto com os gráficos de dispersão.

Fraca correlação entre a variável resposta e a preditora:

IDEIA DE MUDAR O FUNDO DO GRÁFICO PARA VERDE E COLOCAR A LINHA DE REGRESSÃO BRANCA PRA FICAR PARECIDO COM UM CAMPO DE FUTEBOL!

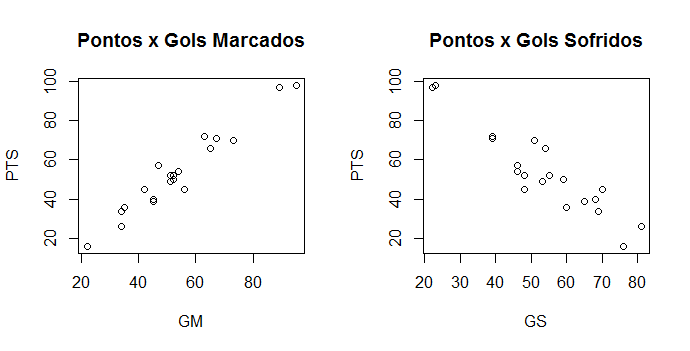


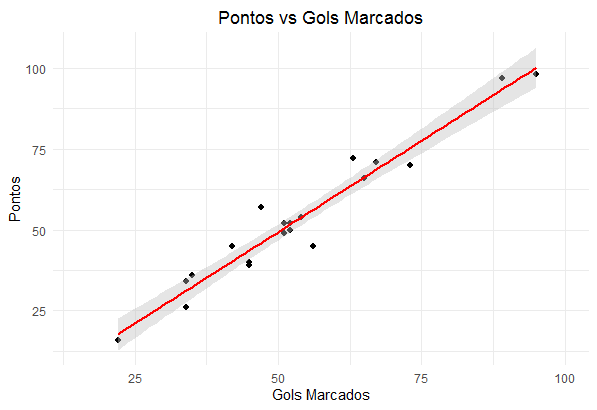


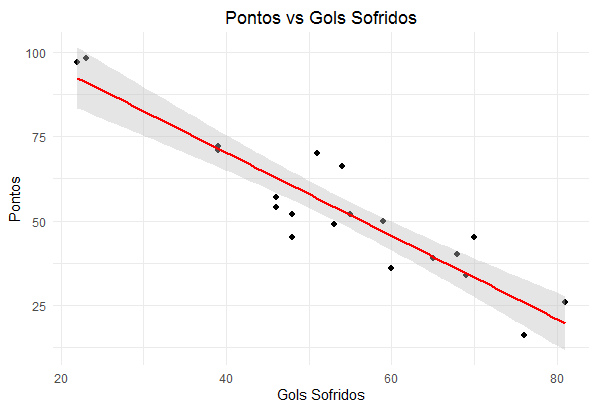


Analisando os gráficos é fácil perceber que não há nenhuma relação entre o número de cruzamentos certos e perca de posse com o total de pontos alcançado pelos times durante o campeonato. E de fato, futebolisticamente falando, a determinação desses fundamentos é muito pouca ou quase nula no aumento da performance de uma equipe, uma vez que a perca de posse é uma condição natural que ocorre alternadamente entre as equipes no jogo. A quantidade assertiva de cruzamentos por si só também não influencia na pontuação dos clubes, de modo que o cruzamento possa acabar não gerando nenhum tipo de finalização e muito menos um gol. Logo, variáveis desse tipo serão removidas das etapas de modelagem, pois terão muito pouca ou nula contribuição na explicação da quantidade de pontos obtidos no campeonato.

Forte correlação entre a variável resposta e a preditora:

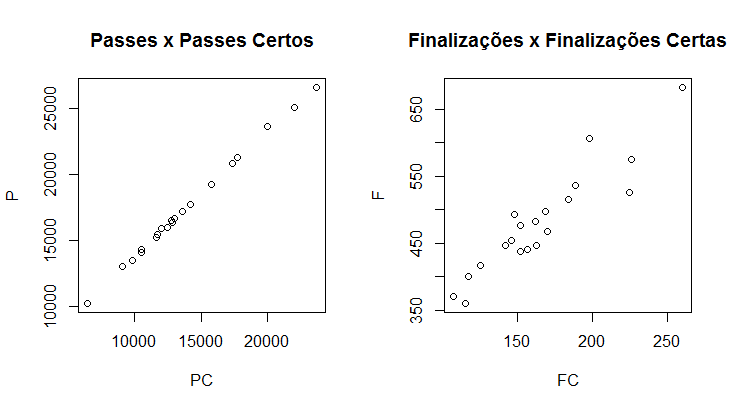


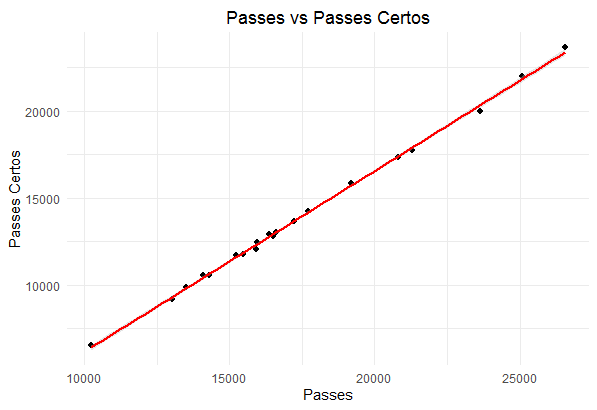


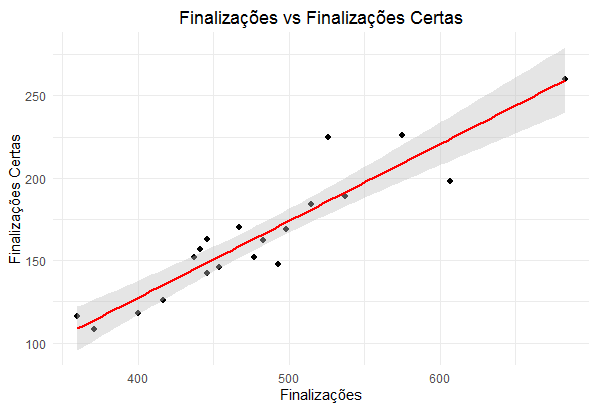


É intuitivo que o número de Gols Marcados e Sofridos teria grande impacto no desempenho dos times, uma vez que o Gol é o maior objetivo de uma equipe em um jogo de futebol. Graficamente temos mais uma prova definitiva desse fato, onde há uma alta correlação entre a pontuação e a quantidade de gols, tornando ambas variáveis fundamentais para as predições futuras, de forma que conseguem captar com uma boa precisão a eficiência das equipes.

Forte correlação entre variáveis preditoras:





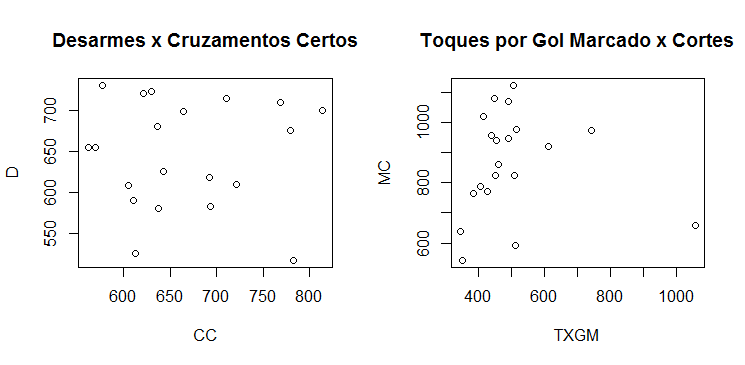


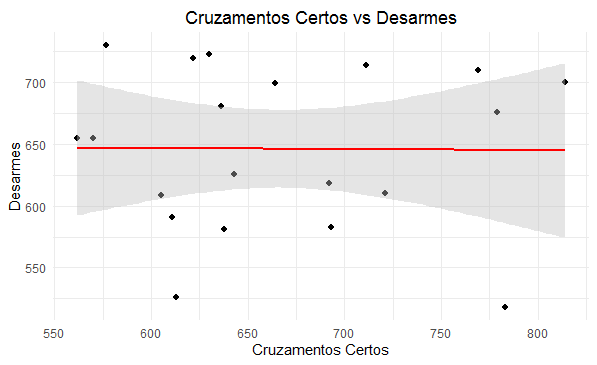
Em um modelo preditivo onde há duas variáveis altamente correlacionadas resulta em ineficiência nas predições futuras, uma vez que ambas adicionam semelhantes contribuições para o modelo, sejam significativas ou não. Dois exemplos claros são os relacionamentos entre Passes x Passes Certos e Finalizações x Finalizações Certas.

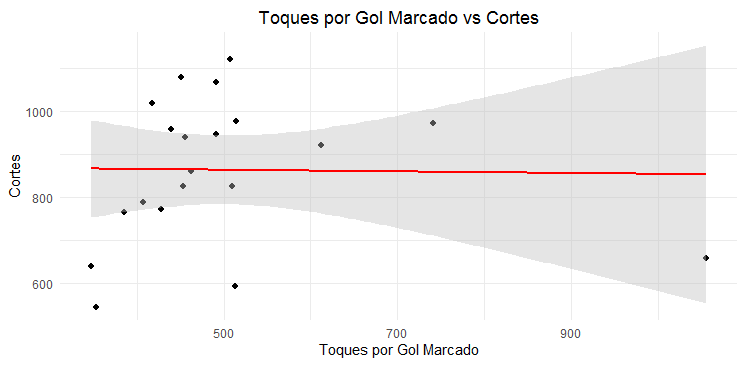
De maneira óbvia esses relacionamentos possuem uma correlação forte, pois se tratam de fundamentos iguais, onde a única diferença é que um mostra a frequência absoluta desse fundamento e o outro representa a assertividade do mesmo. Desta forma, um pode ser substituído pelo outro, e como critério de escolha optamos pelo que mostra a eficiência do time, que são os casos dos Passes e Finalizações praticados de forma correta.

Entretanto, podemos ir mais além nessa escolha, onde podemos substituir a frequência Finalizações Corretas (FC) pelo número total de vezes que um time precisa Finalizar Corretamente para ter um Gol Marcado, que é a variável FCXGM. Logo, esse tipo de raciocínio pode e deve ser feito para todas as outras as variáveis com finalidade de selecionar as melhores variáveis e obter um modelo mais performático.

Correlações Fracas entre preditoras:





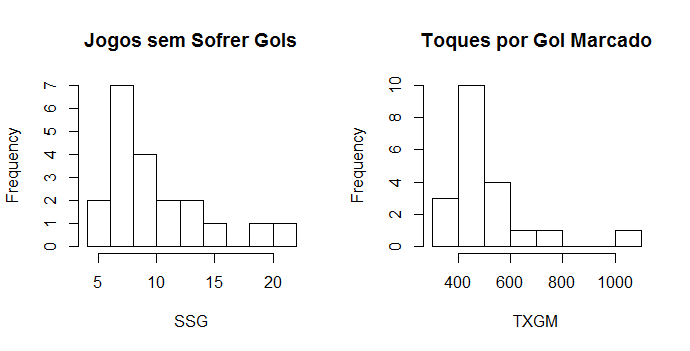


Como podemos visualizar ambas as variáveis não tem qualquer relação entre si, de forma conceitual podemos inferir que o número de Desarmes e Cruzamentos Certos ou a quantidade de Toques por Gol Marcados e Cortes, não possuem interferência umas nas outras no futebol. Assim sendo, a correlação entre as variáveis é baixa, o que é muito bom para o modelo preditivo, uma vez que variáveis dependentes altamente correlacionadas são prejudiciais para a generalização.

Prosseguindo com o processo com uma análise da simetria das variáveis, temos:

As variáveis DRC, MI, PP e PXPR apresentaram simetria, já BICD,D,DF,GS,MC e PL apresentaram assimetria moderada negativa, já B, CC, F, FC, GM, P, PB, PC, PPFC, PR, PTS, T e TXFC apresentaram assimetria moderada positiva e as restantes DFXGS, PXGM, SSG e TXGM apresentaram forte assimetria positiva. Não houve nenhuma forte assimetria negativa.

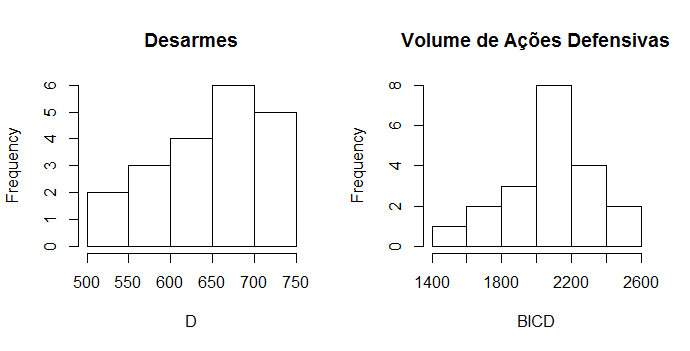
Em fortes assimetrias positivas, a média é influenciada pelos valores menores.



A quantidade de clubes com alto é SSG é muito pequena, portanto a concentração de observações está na parte inferior. A mesma lógica pode ser empregada para TXGM, pois são poucos os clubes que possuem alta eficiência e objetividade no aspecto ofensivo em busca de gols. Logo, há uma quantidade maior de clubes que necessitam de mais toques na bola para balançar a rede.

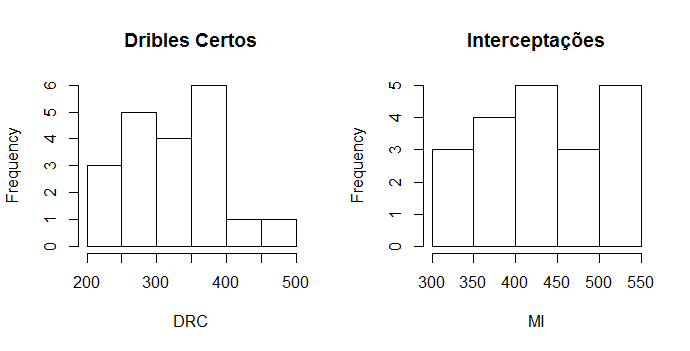
Não houve forte assimetria negativa em nenhuma das variáveis. Concluindo que para esse conjunto de dados são raros valores menores que deslocam a média.

Em assimetrias moderadas negativas, a média continua sendo influenciada pelos valores maiores.



Algumas equipes estão condicionadas a ter que desarmar mais que outras por conta de estilos de jogo defensivos e assim a distribuição é influenciada por essas observações. Analogicamente, o volume de ações defensivas segue o mesmo comportamento.

Em simetrias podemos inferir com certeza que a média daquela característica é correspondente a de todas equipes em geral, pois os dados possuem distribuição normal. Além disso, não é necessário realizar nenhum tipo de transformação para aumentar a performance dentro do modelo.



Percebemos que as distribuições de Dribles Certos e de Interpretações são normais. Fato esperado no caso das Interceptações, pois é um fundamento que todos os times praticam constantemente para retomar a posse ou quebrar uma jogada.

Concluir que para a possível melhora da performance do modelo, poderão ser requisitadas transformações nas variáveis que possuírem forte assimetria.

#É necessário ou possibilitaria mais resultados significativos tentar transformar as variáveis que possuem assimetria moderada?

Seleção das variáveis & Escolha e Motivações do modelo completo inicial

Como as variáveis CC, PP e PXPR tiveram uma fraca correlação com a preditora, assim serão descartadas. Entre as variáveis que possuem forte correlação com a independente, é evidente que GM e GS mostram grande importância na explicação da pontuação dos times, tanto pelos gráficos de dispersão e coeficientes de correlação quanto pela natureza do futebol.

Porém temos algumas que apresentam conceitos futebolísticos relacionados e/ou superficiais em relação às demais, que consequentemente causa correlação e dispensabilidade entre elas. Desta forma selecionaremos as variáveis mais significativas no contexto do futebol e levando em conta a correlação. Descartando variáveis com valores com insights superficiais e priorizando as que retratam eficiência. E alguns exemplos dessas escolhas são: FC em vez de F; PPFC e PR em vez de P, PC e PB; TXGM em vez de T. Outras variáveis inseridas serão BICD\* (em vez de ...), DFXGS e SSG.

Algumas variáveis que possuem correlação moderada com a variável resposta também podem ter potenciais para estimar a performance dos clubes, como DF, TXFC e DRC. Portanto, as 12 variáveis dependentes que irão compor o modelo inicial são GM, FC, GS, PR, DF, SSG, DRC, TXGM, DFXGS, TXFC, BICD e PPFC. Sendo que as restantes serão descartadas.

Logo, podemos estimar o primeiro Modelo de Regressão Linear Múltipla (MRLM) em função de todas as variáveis selecionadas (tive que tirar B, MI, MC e D , pois afetou em BICD). Obtendo os seguintes resultados:

Call:

lm(formula = PTS ~ GM + FC + GS + PR + DF + SSG + DRC + TXGM +

DFXGS + TXFC + BICD + PPFC, data = sel)

Residuals:

Min 1Q Median 3Q Max

-3.3967 -0.6075 0.2734 0.9072 2.9042

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.299e+02 5.033e+01 2.581 0.0364 \*

GM 2.568e-01 2.167e-01 1.185 0.2747

FC -8.935e-02 1.462e-01 -0.611 0.5604

GS -4.056e-01 2.256e-01 -1.798 0.1153

PR 2.691e-03 6.195e-03 0.434 0.6771

DF -9.472e-02 1.348e-01 -0.703 0.5049

SSG -5.121e-01 4.446e-01 -1.152 0.2872

DRC -1.329e-02 1.818e-02 -0.731 0.4883

TXGM -2.923e-02 1.527e-02 -1.914 0.0972 .

DFXGS 6.607e+00 5.652e+00 1.169 0.2806

TXFC -1.350e-01 1.504e-01 -0.898 0.3992

BICD -1.475e-02 7.730e-03 -1.909 0.0980 .

PPFC 9.315e-04 1.604e-03 0.581 0.5796

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.407 on 7 degrees of freedom

Multiple R-squared: 0.9952, Adjusted R-squared: 0.9869

F-statistic: 120.1 on 12 and 7 DF, p-value: 6.592e-07

Percebemos pelo teste t que nenhuma das variáveis foi significativa a 5%. Outra conclusão é a partir do coeficiente de determinação ajustado que infere que o modelo consegue captar cerca de 99% dos dados. Porém, é uma conclusão equivocada, pois podem haver variáveis não significativas para o modelo, além de problemas de multicolinearidade em função do alto número de variáveis preditoras. Resultando em um modelo com parâmetros inflacionados por conta da alta variabilidade, que por consequência é incapaz de generalizar e prever com eficiência novas amostras.

Deste modo, verificaremos a multicolinearidade do modelo pelo critério de VIF, temos:

GM FC GS PR DF SSG

50.314129 108.423476 41.447806 58.186938 35.106451 13.263651

DRC TXGM DFXGS TXFC BICD PPFC

4.889378 18.970666 24.740903 31.932396 14.101840 113.407263

Segundo a literatura, o valor máximo de corte aceito para a não ocorrência de multicolinearidade é de VIF<10. Assim, percebemos que muitas variáveis estão inflacionando e prejudicando o modelo. Logo, iremos adotar critérios que selecionem modelos que incluam apenas variáveis significativas. E em seguida avaliar a multicolinearidade e os demais pressupostos destes modelos selecionados.

Primeiramente, utilizaremos o procedimento de todas as regressões possíveis / algoritmo exaustivo, selecionando os dois melhores modelos para cada quantidade de variáveis selecionadas (1,2,3,...8,...,15).

2 subsets of each size up to 8

Selection Algorithm: exhaustive

GM FC GS PR DF SSG DRC TXGM DFXGS TXFC BICD PPFC

1 ( 1 ) "\*" " " " " " " " " " " " " " " " " " " " " " "

1 ( 2 ) " " " " "\*" " " " " " " " " " " " " " " " " " "

2 ( 1 ) "\*" " " "\*" " " " " " " " " " " " " " " " " " "

2 ( 2 ) "\*" " " " " " " " " "\*" " " " " " " " " " " " "

3 ( 1 ) "\*" " " "\*" " " "\*" " " " " " " " " " " " " " "

3 ( 2 ) "\*" " " "\*" " " " " " " " " " " "\*" " " " " " "

4 ( 1 ) "\*" " " "\*" " " " " " " " " "\*" " " " " "\*" " "

4 ( 2 ) "\*" " " "\*" " " "\*" " " " " " " " " " " "\*" " "

5 ( 1 ) "\*" " " "\*" " " " " "\*" " " "\*" " " " " "\*" " "

5 ( 2 ) "\*" " " "\*" " " "\*" " " " " "\*" " " " " "\*" " "

6 ( 1 ) "\*" " " "\*" " " " " "\*" " " "\*" "\*" " " "\*" " "

6 ( 2 ) "\*" " " "\*" " " " " " " " " "\*" "\*" " " "\*" "\*"

7 ( 1 ) "\*" " " "\*" " " " " "\*" " " "\*" "\*" " " "\*" "\*"

7 ( 2 ) "\*" "\*" "\*" " " " " "\*" " " "\*" "\*" " " "\*" " "

8 ( 1 ) "\*" " " "\*" " " " " " " "\*" "\*" "\*" "\*" "\*" "\*"

8 ( 2 ) "\*" "\*" "\*" " " " " " " "\*" "\*" "\*" " " "\*" "\*"

Adotaremos o critério de menor Cp da Mallows para identificar o melhor modelo dentre esses os selecionados (pág 294). Com base nesse critério o modelo escolhido tem as seguintes variáveis: GM, GS, SSG, TXGM e BICD. Assim, avaliaremos o modelo posteriormente, pois agora selecionaremos mais dois novos modelos por meio de dois métodos mais robustos, os algoritmos de forward e backward. Esses algoritmos consistem em (...).

Aplicando o método de forward, com base no teste F, temos:

Single term additions

Model:

PTS ~ GM + GS

Df Sum of Sq RSS AIC F value Pr(>F)

<none> 119.200 41.701

FC 1 1.1828 118.017 43.502 0.1604 0.69413

PR 1 2.1947 117.005 43.330 0.3001 0.59137

DF 1 19.7277 99.472 40.083 3.1732 0.09385 .

SSG 1 6.6816 112.518 42.548 0.9501 0.34420

DRC 1 1.9838 117.216 43.366 0.2708 0.60993

TXGM 1 0.6897 118.510 43.585 0.0931 0.76418

DFXGS 1 14.0510 105.149 41.193 2.1381 0.16305

TXFC 1 0.0080 119.192 43.700 0.0011 0.97431

BICD 1 9.5983 109.601 42.022 1.4012 0.25381

PPFC 1 5.4184 113.781 42.771 0.7619 0.39563

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Portanto, ao nível de 5% de significância utilizando o critério do teste F o modelo selecionado contêm as variáveis GM e GS.

Aplicando o método de backward ao nível de 5% de significância, com base no teste F temos:

Single term deletions

Model:

PTS ~ GM + GS + TXGM + BICD

Df Sum of Sq RSS AIC F value Pr(>F)

<none> 80.92 37.953

GM 1 83.46 164.37 50.128 15.4709 0.001328 \*\*

GS 1 376.53 457.44 70.598 69.7998 5.027e-07 \*\*\*

TXGM 1 28.69 109.60 42.022 5.3177 0.035805 \*

BICD 1 37.59 118.51 43.585 6.9692 0.018561 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Logo, ao nível de 5% de significância com base no teste F, o modelo selecionado contêm as variáveis GM, GS, TXGM e BICD.

Deste modo, temos 4 modelos a serem avaliados, com as respectivas variáveis:

Modelo 1 (escolhido com base no modelo com menor cp de mallows dentre os modelos selecionados a partir do método exaustivo/todas as regressões possíveis): GM, GS, SSG, TXGM e BICD.

Modelo 2 (selecionado pelo critério de forward ao nível de 5%): GM e GS.

Modelo 3 (selecionado pelo critério de backward ao nível de 5%): GM, GS, TXGM e BICD.

Analisando os parâmetros de cada um desses modelos, temos:

Modelo 1

Call:

lm(formula = PTS ~ GM + GS + SSG + TXGM + BICD, data = sel)

Coefficients:

(Intercept) GM GS SSG TXGM

105.98709 0.46756 -0.59369 -0.39133 -0.02012

BICD

-0.01531

No modelo 1 temos que o coeficiente de SSG tem interpretação inadequada, pois é no futebol é totalmente equivocado dizer que quanto mais Jogos Sem Sofrer Gol um time ficar, menos pontos ele terá. Uma solução para esse problema é retirar SSG do modelo. Um fato curioso é que ao realizar essa alteração o modelo fica igual ao modelo 3.

Modelo 1 (adaptado):

...

Modelo 2:

Call:

lm(formula = PTS ~ GM + GS, data = sel)

Coefficients:

(Intercept) GM GS

38.7568 0.7698 -0.4957

O modelo 2 não apresenta nenhum problema de interpretação.

Modelo 3:

Call:

lm(formula = PTS ~ GM + GS + TXGM + BICD, data = sel)

Coefficients:

(Intercept) GM GS TXGM BICD

92.61945 0.48091 -0.51042 -0.01887 -0.01361

O modelo 3 é igual ao modelo 1 (adaptado) que também não tem empecilhos de interpretação.

Agora vamos avaliar a multicolinearidade dos modelos remanescentes, utilizando o critério de VIF e considerando valores menores que 10:

Modelo 2

GM GS

3.233398 3.233398

Não há problemas de multicolinearidade.

Modelo 3

GM GS TXGM BICD

17.195185 3.262286 5.849906 6.734504

Para que seja mantida a variável GM que é bastante importante para o modelo e não ocorra multicolinearidade, deve ser retirada a variável BICD ou TXGM. Escolhemos por retirar a variável BICD pois é a mais inflacionada das duas, assim temos:

Modelo 3 adaptado

GM GS BICD

4.846100 3.235180 2.525382

Por fim, testaremos os pressupostos de normalidade, homogeneidade e independência para os 2 modelos restantes.

Modelo 2:

Shapiro-Wilk normality test

data: residuals(m2)

W = 0.90685, p-value = 0.05553

studentized Breusch-Pagan test

data: m2

BP = 2.1906, df = 2, p-value = 0.3344

Durbin-Watson test

data: m2

DW = 1.839, p-value = 0.6323

alternative hypothesis: true autocorrelation is not 0

Portanto, ao nível de 5% de significância há normalidade, homogeneidade e independência.

Modelo 3:

Shapiro-Wilk normality test

data: residuals(m3)

W = 0.9604, p-value = 0.5518

studentized Breusch-Pagan test

data: m3

BP = 3.8274, df = 3, p-value = 0.2807

Durbin-Watson test

data: m3

DW = 1.6715, p-value = 0.3825

alternative hypothesis: true autocorrelation is not 0

Portanto, ao nível de 5% de significância há normalidade, homogeneidade e independência. (concluir pra cada teste de forma separada)

Agora vamos comparar o Root Mean Squared Error (RMSE) dos dois modelos para inferir qual é o mais eficiente:

Modelo 2 : RMSE=3.65429e-14 vs Modelo 3: 3.177644e-14

O modelo 3 apresenta um RMSE menor em relação ao modelo 2, mas essa diferença não é tão significativa. Porém, comparar o desempenho dos modelos utilizando os dados que foram usados na modelagem não é um bom indicador para previsões futuras. Desta forma, foram coletados dados da temporada anterior a essa (Premier League 17/18) para mensurar a performance dos modelos. Assim, obteremos um melhor parâmetro para poder inferir a edição futura (Premier League 19/20). Segue abaixo a comparação das predições:

Modelo 2: MSE=4.38558 vs Modelo 3: 4.413387

Como podemos observar o modelo 2 apresenta o menor erro para essas novas observações. Portanto, é o modelo mais adequado para a previsão do próximo campeonato. Porém, será que é possível diminuir o erro do modelo 2 adicionando uma ou duas variáveis que possa contribuir significativamente para o aumento da performance ?

Tomaremos como base as variáveis que apareceram nos modelos anteriores e analisaremos se o RMSE decresce. Caso essa condição seja satisfeita, avaliaremos a multicolinearidade, verificaremos as interpretações dos parâmetros e testaremos todos os pressupostos desse novo modelo. Desta forma, se o modelo atender esses requisitos vai defini-lo como o modelo final, que por sua vez terá mais eficiência na predição da Premier League 19/20.

Após as adições das variáveis (DFXGS) e (BICD) houve um bom decréscimo do erro além desse novo modelo satisfazer todos os pressupostos. Assim, temos:

Modelo 13

Call:

lm(formula = PTS ~ GM + GS + DFXGS + BICD, data = sel)

Coefficients:

(Intercept) GM GS DFXGS BICD

44.46917 0.68427 -0.46419 2.87644 -0.00441

Avaliando Multicolinearidade:

GM GS DFXGS BICD

5.371126 3.586451 2.480658 2.528993

Avaliando os pressupostos:

Shapiro-Wilk normality test

data: residuals(m13)

W = 0.92635, p-value = 0.1314

studentized Breusch-Pagan test

data: m13

BP = 5.8573, df = 4, p-value = 0.2101

Durbin-Watson test

data: m13

DW = 1.6994, p-value = 0.4498

alternative hypothesis: true autocorrelation is not 0

Logo, ao nível de 5% de significância o modelo satisfaz os pressupostos de normalidade, homogeneidade e independência. Outros pontos positivos é a adequação da multicolinearidade e interpretação dos parâmetros. Portanto, esse será o nosso modelo preditivo final para a previsão da atual temporada da Premier League.

Call:

lm(formula = PTS ~ GM + GS + DFXGS + BICD, data = sel)

Residuals:

Min 1Q Median 3Q Max

-3.6269 -1.1067 -0.4538 0.3898 5.1458

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 44.469172 11.867061 3.747 0.00194 \*\*

GM 0.684269 0.073902 9.259 1.36e-07 \*\*\*

GS -0.464187 0.069277 -6.700 7.10e-06 \*\*\*

DFXGS 2.876445 1.867766 1.540 0.14438

BICD -0.004410 0.003417 -1.291 0.21634

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.512 on 15 degrees of freedom

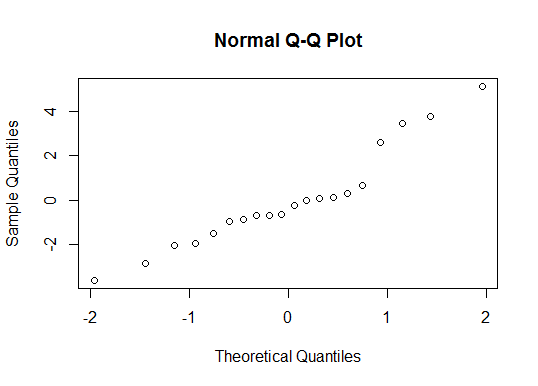
Multiple R-squared: 0.9887, Adjusted R-squared: 0.9857

F-statistic: 328.5 on 4 and 15 DF, p-value: 2.086e-14

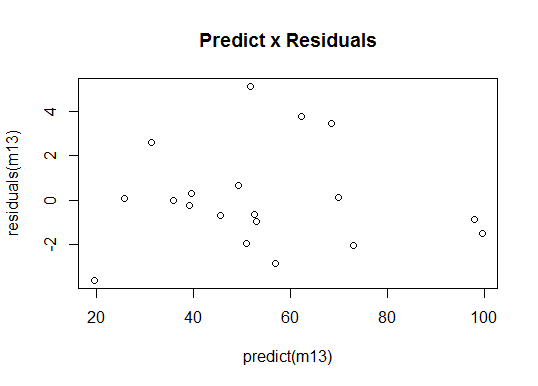
PTS=44.46917 + 0.68427\*GM -0.46419\*GS + 2.87644\*DFXGS -0.00441\*BICD

PTS=44.47 + 0.68\*GM -0.46\*GS + 2.88\*DFXGS -0.0044\*BICD

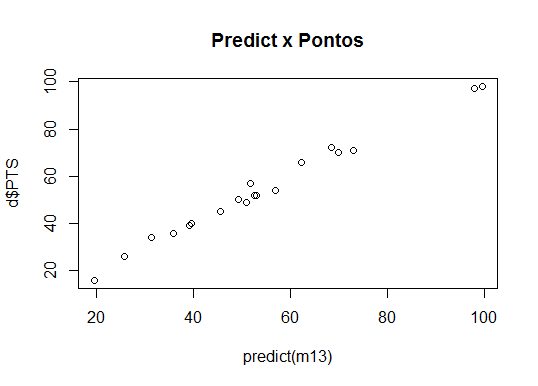
Gráficos de Resíduos:



...



...



...

Mean Squared Error: 2.175288

Interpretações a cerca do modelo:

O modelo consegue captar cerca de 98.6% dos dados quando introduzida as variáveis preditoras GM, GS, DFXGS e BICD na explicação da variável resposta PTS.

Sobre os parâmetros:

Para cada Gol Marcado (GM) é esperado um acréscimo médio de aproximadamente 0.68 pontos. Para cada Gol Sofrido (GS) é esperado um decréscimo médio de cerca de 0.46 pontos. Para cada Defesa por Gol Sofrido (DFXGS) é esperado um aumento médio em torno de 2.88 pontos. Para cada aumento unitário no volume defensivo (Bloqueios, Cortes, Interceptações e Desarmes – BICD) é esperado uma diminuição média de aproximadamente 0.0044 pontos. A seguir temos o intervalo de confiança com 5% de significância dos parâmetros:

2.5 % 97.5 %

(Intercept) 19.17513080 69.763213284

GM 0.52675150 0.841787402

GS -0.61184721 -0.316527727

DFXGS -1.10460522 6.857494500

BICD -0.01169208 0.002872377

Predições Treino

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Premier League 2018/2019 | | | | |
| Campeonato | Time | Pontos Reais | Time | Predição |
| UCL | Manchester City | 98 | Manchester City | 100 |
| UCL | Liverpool | 97 | Liverpool | 98 |
| UCL | Chelsea | 72 | Tottenham Hotspur | 73 |
| UCL | Tottenham Hotspur | 71 | Arsenal | 70 |
| UEL | Arsenal | 70 | Chelsea | 69 |
| UEL | Manchester United | 66 | Manchester United | 62 |
| UEL | Wolverhampton Wanderers | 57 | Everton | 57 |
| NA | Everton | 54 | Leicester City | 53 |
| NA | Leicester City | 52 | West Ham United | 53 |
| NA | West Ham United | 52 | Wolverhampton Wanderers | 52 |
| NA | Watford | 50 | Crystal Palace | 51 |
| NA | Crystal Palace | 49 | Watford | 49 |
| NA | AFC Bournemouth | 45 | Newcastle United | 46 |
| NA | Newcastle United | 45 | AFC Bournemouth | 46 |
| NA | Burnley | 40 | Burnley | 40 |
| NA | Southampton | 39 | Southampton | 39 |
| NA | Brightonand Hove Albion | 36 | Brightonand Hove Albion | 36 |
| REB | Cardiff City | 34 | Cardiff City | 31 |
| REB | Fulham | 26 | Fulham | 26 |
| REB | Huddersfield Town | 16 | Huddersfield Town | 20 |

Intervalo de Confiança das Predições

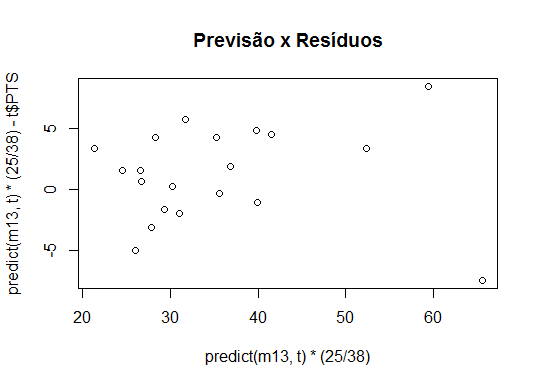
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Premier League 2018/2019 - Intervalo de Confiança : 95% | | | | |
| Campeonato | Time | Predições | | |
| Mínima | Esperada | Máxima |
| UCL | Manchester City | 94 | 100 | 105 |
| UCL | Liverpool | 93 | 98 | 103 |
| UCL | Tottenham Hotspur | 69 | 73 | 77 |
| UCL | Arsenal | 66 | 70 | 73 |
| UEL | Chelsea | 64 | 69 | 73 |
| UEL | Manchester United | 59 | 62 | 65 |
| UEL | Everton | 54 | 57 | 59 |
| NA | Leicester City | 50 | 53 | 56 |
| NA | West Ham United | 49 | 53 | 56 |
| NA | Wolverhampton Wanderers | 49 | 52 | 55 |
| NA | Crystal Palace | 49 | 51 | 53 |
| NA | Watford | 47 | 49 | 51 |
| NA | Newcastle United | 41 | 46 | 50 |
| NA | AFC Bournemouth | 41 | 46 | 50 |
| NA | Burnley | 37 | 40 | 42 |
| NA | Southampton | 35 | 39 | 43 |
| NA | Brightonand Hove Albion | 33 | 36 | 39 |
| REB | Cardiff City | 29 | 31 | 34 |
| REB | Fulham | 22 | 26 | 30 |
| REB | Huddersfield Town | 14 | 20 | 25 |

É possível perceber que o modelo apresentou um ótimo ajuste para os dados de treino, agora resta analisarmos a eficiência dele em dados futuros.

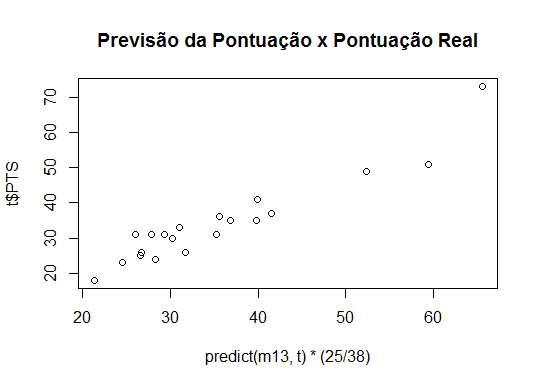
**DISCUSSÃO**

Portanto, com o modelo já definido iremos prever a pontuação dos times da Premier League 2019/2020 para a atual rodada. A seguir apresentamos alguns gráficos sobre essa previsão:

Previsão x Resíduos



Previsão x Pontos Atuais



Mean Squared Error: 3.955448

Tabela de Comparação da Previsão com a Pontuação Real da 25º Rodada

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Premier League 2019/2020 | | | | |
| Campeonato | Time | Pontos Reais R25 | Time | Predição 25R |
| UCL | Liverpool | 73 | Liverpool | 66 |
| UCL | Manchester City | 51 | Manchester City | 59 |
| UCL | Leicester City | 49 | Leicester City | 52 |
| UCL | Chelsea | 41 | Tottenham Hotspur | 41 |
| UEL | Tottenham Hotspur | 37 | Chelsea | 40 |
| UEL | Sheffield United | 36 | Manchester United | 40 |
| UEL | Manchester United | 35 | Wolverhampton Wanderers | 37 |
| NA | Wolverhampton Wanderers | 35 | Sheffield United | 36 |
| NA | Everton | 33 | Arsenal | 35 |
| NA | Arsenal | 31 | Brighton and Hove Albion | 32 |
| NA | Burnley | 31 | Everton | 31 |
| NA | Newcastle United | 31 | Crystal Palace | 30 |
| NA | Southampton | 31 | Burnley | 29 |
| NA | Crystal Palace | 30 | West Ham United | 28 |
| NA | AFC Bournemouth | 26 | Newcastle United | 28 |
| NA | Brighton and Hove Albion | 26 | AFC Bournemouth | 27 |
| NA | Aston Villa | 25 | Aston Villa | 27 |
| REB | West Ham United | 24 | Southampton | 26 |
| REB | Watford | 23 | Watford | 25 |
| REB | Norwich City | 18 | Norwich City | 21 |

Como podemos perceber intuitivamente, conseguimos uma ótima aproximação da previsão do campeonato em comparação com a pontuação atual dos times, apesar de alguns erros maiores (considerando uma margem de erro de 5 pontos na estimação) nas campanhas do Liverpool, Manchester City e Brighton and Hove Albion. Porém, as campanhas do Manchester City e principalmente do Liverpool até o momento, são altamente fora da curva, identificando por sua vez uma anormalidade ou exceção.

Outros erros no modelo são ‘normais’ uma vez que ainda não é possível estimar algo com 100% de eficácia, mas é viável de acordo com o contexto da questão, chegar tão perto quanto se deseja. Ressaltamos que esse não é um modelo único e ideal, mas foi um dos que apresentou melhor performance e atendeu os resquisitos, abrindo possibilidades para a melhoria dessa solução. Portanto, esta estimação continua sendo excelente e mostramos a partir de algumas métricas de análise criadas com base na tabela.

A predição conseguiu acertar 70% dos campeonatos que os times iriam se classificar e indo mais além podemos dividir os 20 times em 4 grupos de 5 e mensurar o percentual de acerto de cada time contido nesses grupos de acordo com a classificação real, obtendo uma taxa de acerto de 75%.Podemos mensurar quem brigaria na parte de cima e de baixo da tabela, dividindo os times em 2 grupos de 10, e assim obtemos uma eficiência de 90% em prever esse contexto.

Outra análise estimativa que podemos fazer é considerar que as equipes manterão as médias de seus desempenhos atuais até o fim do campeonato, e assim podemos prever o resultado final da Premier League 2019/2020.

Tabela de Comparação da Previsão para o final do campeonato com a Pontuação Real da 25º Rodada

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Premier League 2019/2020 | | | | |
| Campeonato | Time | Pontos Reais R25 | Time | Predição 38R |
| UCL | Liverpool | 73 | Liverpool | 100 |
| UCL | Manchester City | 51 | Manchester City | 90 |
| UCL | Leicester City | 49 | Leicester City | 80 |
| UCL | Chelsea | 41 | Tottenham Hotspur | 63 |
| UEL | Tottenham Hotspur | 37 | Chelsea | 61 |
| UEL | Sheffield United | 36 | Manchester United | 61 |
| UEL | Manchester United | 35 | Wolverhampton Wanderers | 56 |
| NA | Wolverhampton Wanderers | 35 | Sheffield United | 54 |
| NA | Everton | 33 | Arsenal | 54 |
| NA | Arsenal | 31 | Brighton and Hove Albion | 48 |
| NA | Burnley | 31 | Everton | 47 |
| NA | Newcastle United | 31 | Crystal Palace | 46 |
| NA | Southampton | 31 | Burnley | 45 |
| NA | Crystal Palace | 30 | West Ham United | 43 |
| NA | AFC Bournemouth | 26 | Newcastle United | 42 |
| NA | Brighton and Hove Albion | 26 | AFC Bournemouth | 40 |
| NA | Aston Villa | 25 | Aston Villa | 40 |
| REB | West Ham United | 24 | Southampton | 39 |
| REB | Watford | 23 | Watford | 37 |
| REB | Norwich City | 18 | Norwich City | 32 |

Intervalos de Confiança para Previsão de Teste

##É ADEQUADO E POSSÍVEL ESTIMAR PARA OS DADOS DE TESTE ???##

##Construir um intervalo de predição

...

\*\*\*Mensurar o desempenho de possíveis times do Brasil na Premier League

Comentar resultados parciais

\*\*\*Comentar se esse modelo serviria para outras ligas e até testar se tiver mais páginas.

Comentar resultados parciais

Hora do Meu Show! Interpretar! Inferir! Gerar Insights!

Estimando a pontuação de dois times fictícios, um que possui a média e o outro possui a mediana dos scouts, respectivamente:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Premier League 2019/2020 | | | | |
| Campeonato | Time | 38R | Time | 25R |
| NA | Mean | 54 | Mean | 35 |
| NA | Median | 50 | Median | 33 |

**CONCLUSÃO**

Portanto, com base nas análises feitas, no modelo preditivo final e nos resultados obtidos, concluímos que obtemos uma ótima previsão do final da Premier League 2019/2020, considerando é claro a manutenção dos desempenhos médios dos clubes, pois uma vez que algum aumenta ou diminui sua performance a classificação também sofre alterações.

Analisando de forma mais específica o modelo, temos uma evidência da intuitiva importância dos Gols Marcados e Sofridos para que um time tenha ótimo rendimento. Outro fator a se destacar é a Taxa de Defesas por Gols Sofridos, que torna ainda mais relevante o papel do goleiro no sucesso de uma equipe. É claro que essa não é uma variável influenciada apenas pelo desempenho do goleiro, ele tem um papel central, mas devemos considerar todo o contexto defensivo da equipe. Desta forma, vamos ao sentido da última variável do modelo, o BICD, que mensura o volume das ações defensivas da equipe (Bloqueios, Interceptações, Cortes e Desarmes) ao longo do campeonato.

A interpretação do BICD nos diz que quanto mais atitudes defensivas um time realiza, mais ele é propenso a ter uma baixa performance. Desta forma, todas essa variáveis em conjunto produzem o modelo que reflete grande parte do estilo da Premier League atualmente, de forma que defina os clubes de alta e baixa performance. Logo, uma medida a se adotar pelos clubes para aumentar a eficiência no campeonato seria contratar jogadores, treinadores, promoção de atletas da base e entre outros profissionais que tenham bons aproveitamentos nas variáveis ressaltadas.

Logo, a partir de todos esses insights temos que o Liverpool se sagraria atual campeão, Manchester City, Leicester City e Tottenham conseguiriam classificação para UEFA Champions League, além do Chelsea, Manchester United e Wolverhampton conseguiriam vaga para a UEFA Europa League. Entretanto, em uma situação não tão confortável, Southampton, Watford e Norwich seriam despromovidos para a segunda divisão inglesa e os outros clubes do campeonato não teriam classificação em nenhum outro campeonato do continente europeu.

**REFERÊNCIAS**

1. Site Oficial da Premier League. Disponível em [https://www.premierleague.com/stats. Acesso em 12/02/2020](https://www.premierleague.com/stats.%20Acesso%20em%2012/02/2020).

2. Site do Globoesporte. Disponível em <https://globoesporte.globo.com/futebol/futebol-internacional/noticia/ranking-ge-ingles-e-a-melhor-liga-da-europa-frances-fica-em-ultimo-no-top-5.ghtml>. Acesso em 01/03/2020.

3. Site da IFFHS. Disponível em <https://iffhs.de/posts/58>. Acesso em 01/03/2020.