UNIVERSIDADE FEDERAL DO PARANÁ

André Luiz Grion – GRR20159284

Bruno Henrique Abreu – GRR20159983

Maria Tereza Neves de Oliveira – GRR20159323

Predição dos resultados dos jogos do campeonato brasileiro via modelagem do número de gols, utilizando dados do Cartola

CURITIBA

Novembro de 2017

Resumo

A ideia de predizer resultados de jogos de futebol é um desafio. Número de gols variam bastante a cada partida e o time aparentemente mais forte, nem sempre é o vitorioso. É um esporte supostamente imprevisível. Modelagens estatísticas, no entanto, vêm sendo amplamente utilizadas no intuito de selecionar as variáveis mais significativas e fazer predições no esporte. Com essa motivação, o presente trabalho busca calcular a probabilidade média de gols marcados dos times mandantes e visitantes. Ampliando a análise, buscou-se também predizer qual time sairá vitorioso de determinada partida. O número de gols marcados pelos times mandantes e visitantes foram tratados como duas distribuições de Poisson independentes, com função de ligação logarítmica, no contexto de modelos lineares generalizados. Os dados foram importados da plataforma *Kaggle*, originados do Cartola F.C. A seleção de variáveis foi realizada pelo método *stepwise*, utilizando o Critério de Informação de Akaike (AIC). De modo geral, as observações foram bem ajustadas pelo modelo proposto. Dicotomizando o resultado foi possível chegar a nível de sensibilidade em torno de 0,71 e especificidade em torno 0,40.

Palavras-chave: Futebol; Predição; Cartola; Regressão dados de contagem; Modelos lineares generalizados; Distribuição Poisson.

Conteúdo

Introdução

Originado na Inglaterra, o futebol é um dos mais populares jogos coletivos do mundo. Uma das características desse esporte é que o melhor time nem sempre é o vencedor da partida ou do torneio, o que causa um clima de apreensão entre os jogadores e fãs. Na mídia, o time mandante marca mais gols do que o time visitante, isso é uma característica de "vantagem por jogar em casa" e não é um fator particular do futebol.

Com base nesses fatores, predizer o número de gols de um time torna-se um interesse de pesquisadores e estatísticos. Pensando nisso, foi proposto o modelo Poisson para descrever os dados, retirados da plataforma *kaggle*, referentes ao campeonato brasileiro de 2017.

O modelo Poisson é uma distribução de probabilidade discreta que descreve a probabilidade de um número de eventos ocorrer, em um determinado período de tempo ou espaço (por exemplo, 90 minutos). Essa distribuição possui uma taxa média de ocorrência e o número de eventos é independente do tempo, ou seja, a probabilidade de se marcar um gol não aumenta conforme o número de gols são marcados na partida. O número de gols é expresso como uma fução da taxa média de gols.

O banco de dados foi intensamente analisado e ajustado para auxiliar na escolha das variáveis mais significativas para compor o modelo. Com técnicas estatésticas específicas, chegou-se a um modelo final que obteve ajuste satisfatório.

Material e Métodos

Construção da ABT (Analytical Base Table)

Para a construção da ABT, utilizamos as bases de dados disponibilizadas na plataforma Kaggle, que estavam divididas em 34 bases. Segue abaixo breve resumo de cada base:

Tabela	Descrição
cartola_17	Informações dos atletas (9.642 observações , 37 variáveis).A base contém informações dos atletas, tais como: nome, ID do jogador(chave que identifica o jogador), time que atua, posição que atua (lateral, ataque, goleiro, meio-campo ou técnico), status do jogador (possibilidade de ser escalado na rodada), pontos na rodada, dentre outras.
cartola_2017_samples	Informações dos atletas (2.650 observações, 32 variáveis). A base contém basicamente as mesmas informações da base, cartola_17, porém sumarizada. Váriaveis com missing e observações sem ID foram descartadas.

Tabela	Descrição
cartola_2017_scouts	Informações dos atletas (2.650 observações, 38 variáveis). A base contém basicamente as mesmas informações da base, cartola_2017_samples, porém foram adicionadas as variáveis do time adversário (nome do time adversário na rodada em questão e gols tomados).
cartola_aggregated	Informações dos atletas (40.296 observações, 77 variáveis). A base contém as informações dos atletas, assim como as bases citadas acima, porém não indica de qual time é o atleta.
matches_brasileirao_2017	Informações dos jogos (300 observações, 9 variáveis). A base contém todos os jogos do campeonato brasileiro de 2017, rodada a rodada, com data, times que se enfrentaram na rodada e placar atualizado até a 35ª rodada.
tabela_times	Informações dos jogos (20 observações, 17 variáveis). A base contém a classificação do campeonato brasileiro 2017 com os 20 times, atualizada até a 26ª rodada.
teamids	Informações dos times (20 observações, 5 variáveis). A base contém o ID (identificação na base) de cada time e a posição no campeonato brasileiro 2017 até a 35ª rodada.
teamids_consolidated	Informações dos times (43 observações, 6 variáveis). A base contém informações de cada time da série A e B do campeonato brasileiro 2017, tais como: ID (identificação na base), nome do time, posição na classificação do campeonato brasileiro de 2017 até a 35ª rodada.

Tendo em vista um melhor ajuste do modelo, decidimos atualizar as bases até a 35ª rodada. Apesar das bases citadas acima conterem muitas informações úteis, não foi possível utilizá-las para a construção da ABT para realizar a modelagem. Após algumas manipulações de dados, verificou-se que quando utilizavam-se as chaves para realizar a junção das bases, tínhamos informações apenas até a 11ª rodada, impossibilitando a criação da ABT com as rodadas mais recentes. Isso pode ter acontecido, devido a problemas no carregamento das bases no GitHub.

Sendo assim, utilizamos os dados disponibilizados de cada rodada, para criar a base usada na modelagem. Segue abaixo breve resumo das 35 bases, referentes a cada rodada:

rodada	Descrição
rodada_1	772 observações, 33 variáveis
rodada_2	772 observações, 33 variáveis
rodada_3	798 observações, 33 variáveis
rodada_4	800 observações, 33 variáveis
rodada_5	805 observações, 33 variáveis
rodada_5	809 observações, 33 variáveis
rodada_7	813 observações, 33 variáveis
rodada_8	814 observações, 33 variáveis
rodada_9	816 observações, 33 variáveis
rodada_10	819 observações, 33 variáveis
rodada_11	822 observações, 33 variáveis
rodada_12	828 observações, 33 variáveis
rodada_13	830 observações, 33 variáveis
rodada_14	833 observações, 33 variáveis
rodada_15	832 observações, 33 variáveis
rodada_16	834 observações, 33 variáveis
rodada_17	819 observações, 33 variáveis
rodada_18	836 observações, 33 variáveis
rodada_19	831 observações, 33 variáveis
rodada_20	835 observações, 33 variáveis
rodada_21	836 observações, 33 variáveis
rodada_22	853 observações, 33 variáveis
rodada_23	854 observações, 33 variáveis
rodada_24	853 observações, 33 variáveis
rodada_25	840 observações, 33 variáveis
rodada_26	842 observações, 33 variáveis
rodada_27	832 observações, 33 variáveis
rodada_28	812 observações, 33 variáveis
rodada_29	820 observações, 33 variáveis
rodada_30	829 observações, 33 variáveis
rodada_31	802 observações, 33 variáveis
rodada_32	855 observações, 33 variáveis
rodada_33	844 observações, 33 variáveis
rodada_34	842 observações, 33 variáveis
rodada_35	842 observações, 33 variáveis

Todas as bases possuem informações dos atletas, tais como: nome do jogador, ID (identificação do jogador na base), apelido do jogador, preço em cada rodada, time em que atua, posição em que atua, números de jogos que já atuou no campeonato, pontos do jogador na rodada, média dos pontos até a rodada atual e variáveis que relatam a atuação do jogador. A variável status tinha os seguintes parâmetros:

Status	Descrição
Provável	O jogador tem uma possibilidade alta de jogar na rodada, ou seja, está à disposição para jogar.
Dúvida	O jogador possivelmente jogue a rodada, ou seja, está à disposição para jogar porém depende de fatores extra-campo (exemplo: recuperação de lesão).
Nulo	O jogador não irá jogar na rodada devido a problemas extra-campo (exemplo: não está inscrito no campeonato).
Suspenso	O jogador não irá jogar na rodada, devido a uma suspensão.
Contundido	O jogador não irá jogar na rodada, devido a uma contusão.

Outras variáveis referentes a jogadores também estão disponíveis, tais como:

Scouts	Descrição
FS	faltas sofridas
PE	passes errados
A	assistências para gol
FT	chutes na direção do gol
FD	chutes defendidos (apenas goleiro)
FF	chutes para fora do gol
G	gols
I	impedimentos
PP	pênalti perdido
RB	taxa de sucesso (passes certos)
FC	faltas cometidas
GC	gols marcados
CA	cartão amarelo
CV	cartão vermelho
SG	roubadas de bola (apenas zagueiros)
DD	dificuldade da defesa (apenas goleiros)
DP	pênaltis defendidos (apenas goleiros)
GS	gols sofridos (apenas goleiros)

Todos os *scouts* são atualizados a cada rodada, sendo assim, é uma visão da rodada e não acumulada em todas as rodadas anteriores.

Como a intenção também é predizer se o time da casa irá vencer a próxima rodada, tivemos que manipular os dados de tal maneira que as variáveis fizessem sentido no momento da modelagem.

O primeiro passo foi filtrar todos os jogadores que tinham **status** nas bases como **provável** e **dúvida**, pois desse modo o modelo a ser desenvolvido seria com base nos atletas que tem possibilidade

de jogar a rodada em questão.

Após o filtro, foi realizada a sumarização dos jogadores por posição em cada clube. As posições são:

Posição	Descrição
ata	jogadores de ataque
zag	jogadores de defesa (nesse caso os jogadores das laterais (ala), também foram inclusos)
mei	jogadores de meio-campo
gol	goleiros
tec	técnico do time

Essa sumarização é a média de cada variável das bases das rodadas (citadas acima). Realizamos essa manipulação pois como não se sabe qual jogador irá atuar na rodada, seria mais "justo" utilizar a média dos jogadores de cada posição como uma varíavel para cada time.

Sendo assim, seguem alguns exemplos de cada nome das variáveis disponíveis na base:

Variável	Descrição
media_preco_ata_casa	Média de preço dos jogadores de ataque para o time que joga em casa
media_preco_gol_casa	Média de preço dos jogadores que jogam no gol para o time que joga em casa
media_qt_pontos_zag_fora	Média dos pontos recebidos pelos jogadores que jogam na zaga, para o time que joga fora de casa

Após essa sumarização ficaram 5 variáves de cada posição, para cada variável, para cada rodada. Realizamos uma transposição dessas bases para unir a base **matches_brasileirao_2017**, que contém as informações dos confrontos de cada rodada.

Sendo assim, foi criada a ABT de modelagem com 350 observações (todos os confrontos até a 35ª rodada do campeonato), em que cada linha da base contém a média de cada variável (disponibilizadas nas 35 bases, por rodada) de cada posição (ataque, defesa, meio-campo, goleiro e técnico) por rodada.

Além das médias de cada variável, realizamos também a média geral do time e os mínimos e máximos para cada variável e ao todo, construímos mais de 800 variáveis para teste.

Todas as variáveis disponíveis para os times da casa também foram usadas para os times que jogaram fora de casa. O intuito de utilizar as variáveis para os times que jogam fora de casa, é verificar se alguma variável é significativa para a quantidade de gols marcados por cada time. Após a validação dos dados, foi possível então começar as análises para o desenvolvimento do modelo.

Consistência dos dados

Ainda que fosse possível utilizar técnicas de imputação de dados, para manter os dados fiéis ao fornecido pelo Cartola, optou-se pela exclusão de todas as variáveis ligadas ao *scout*, pois apresentaram pelo menos 20 observações perdidas. As variáveis ligadas ao técnico que apresentaram pelo menos 11 observações perdidas também foram desconsideradas.

Embora algumas variáveis também apresentassem número elevado de informações perdidas, como por exemplo as de ataque e as relacionadas aos goleiros, optou-se por manter essas variáveis devido a sua importância na determinação do número de gols em uma partida de futebol.

Após a verificação da consistência da base de dados, restaram 318 observações (partidas) e 150 variáveis do Cartola para a composição do modelo para predição do número de gols marcados.

Separação dos dados para ajuste e para validação

A base de dados foi dividida da seguinte forma:

- Rodadas 1 a 26 (239 partidas): Ajuste
- Rodadas 26 a 35 (79): Validação

Métodos estatísticos

Para a variável resposta, número de gols marcados, foi assumida a distribuição *Poisson* que pode ser representada como:

$$y_i|\boldsymbol{x_i} \sim Poisson(\mu_i)$$
$$g(\mu_i) = \boldsymbol{x_i'}\boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}$$

em que y_i é a i-ésima observação do número de gols com média μ_i; $\boldsymbol{x_i}$ é o vetor das p variáveis explanatórias associadas à i-ésima resposta; $\boldsymbol{\beta}$ é o vetor com $p+1$ parâmetros e g é a função de ligação canônica para a distribuição *Poisson*:

$$g(\mu_i) = log(\mu_i).$$

A seleção das variáveis foi realizada pelo método *stepwise* com eliminação bidirecional, utilizando o Critério de Informação de Akaike (AIC) para a decisão de inclusão ou exclusão.

$$AIC = 2k - 2log(\hat{L})$$

em que k é o número de parâmetros estimados no modelo e $\hat{L}$ é o máximo valor da função de verossimilhança para o modelo.

Dois modelos foram ajustados, um para o número de gols do time mandante e outro para o número de gols do time visitante. Para avaliação da capacidade preditiva, a probabilidade do placar observado foi calculado considerando independência entre os gols marcados por cada time, utilizando a função de probabilidade da distribuição Poisson com parâmetros μ_m e μ_v, os valores preditos, na escala da resposta, respectivamente, para o time mandante e o time visitante (Tabela 7).

Tabela 7: Probabilidades de placares baseados da predição da média de gols marcados com distribuições de Poisson independentes para os times mandante e visitante

		Visitante							
		0		1		...	g		
Mandante	0	$\frac{\mu_m^0 e^{-\mu_m}}{0!}$	$\frac{\mu_v^0 e^{-\mu_v}}{0!}$	$\frac{\mu_m^0 e^{-\mu_m}}{0!}$	$\frac{\mu_v^1 e^{-\mu_v}}{1!}$	...	$\frac{\mu_m^0 e^{-\mu_m}}{0!}$	$\frac{\mu_v^g e^{-\mu_v}}{g!}$
	1	$\frac{\mu_m^1 e^{-\mu_m}}{1!}$	$\frac{\mu_v^0 e^{-\mu_v}}{0!}$	$\frac{\mu_m^1 e^{-\mu_m}}{1!}$	$\frac{\mu_v^1 e^{-\mu_v}}{1!}$	...	$\frac{\mu_m^1 e^{-\mu_m}}{1!}$	$\frac{\mu_v^g e^{-\mu_v}}{g!}$
	⋮	⋮		⋮		⋱	⋮	
	g	$\frac{\mu_m^g e^{-\mu_m}}{g!}$	$\frac{\mu_v^0 e^{-\mu_v}}{0!}$	$\frac{\mu_m^g e^{-\mu_m}}{g!}$	$\frac{\mu_v^1 e^{-\mu_v}}{1!}$	...	$\frac{\mu_m^g e^{-\mu_m}}{g!}$	$\frac{\mu_v^g e^{-\mu_v}}{g!}$

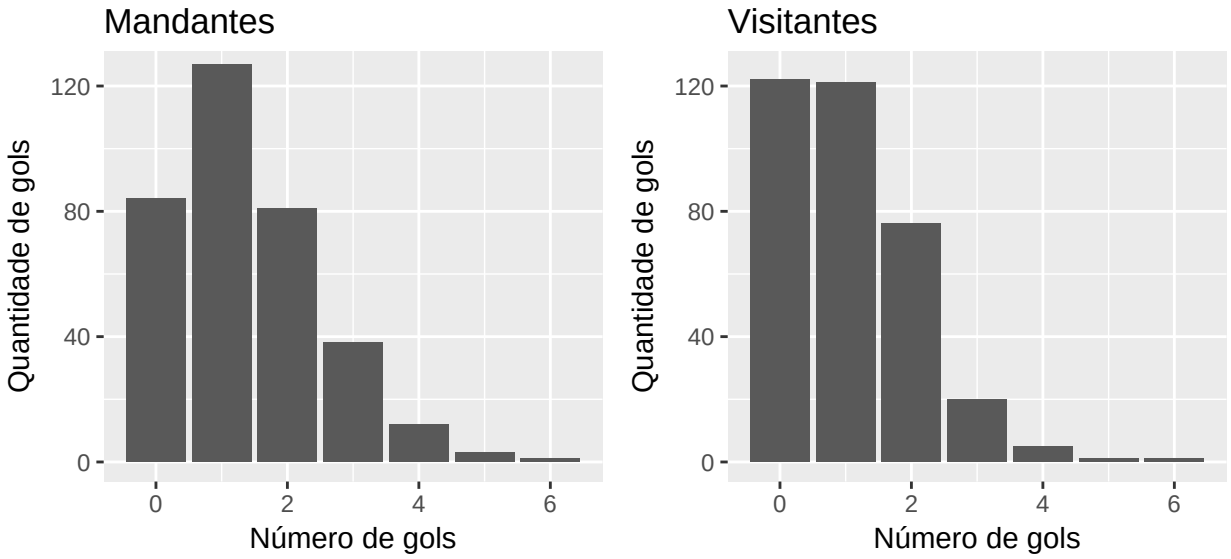
Com base na Tabela 7 é possível calcular a probabilidade de vitória do time mandante como o somatório das probabilidades abaixo da diagonal principal; a probabilidade de vitória do time visitante como o somatório das probabilidades acima da diagonal principal e, a probabilidade de empate, como o somatório das probabilidades da diagonal principal da tabela. Essas probabilidades devem ser ajustadas pelo somatório de todas as probabilidades da tabela, já que o número máximo de gols preditos (g) é finito. Neste trabalho, $g = 15$.

Resultados e discussão

Análise exploratória

Embora e a suposição de independência dos números de gols marcados entre os times mandantes e visitantes seja bem questionável, no presente estudo foi observado uma correlação de Spearman de -0,003, indicando que a suposição é válida.

Segue a distribuição dos gols marcados dos times que jogaram em casa e fora de casa:



Pode-se observar que a amplitude para gols marcados fora de casa e dentro de casa é a mesma (6 gols), porém a frequência é um pouco diferente. Para os mandantes, há uma maior frequência para

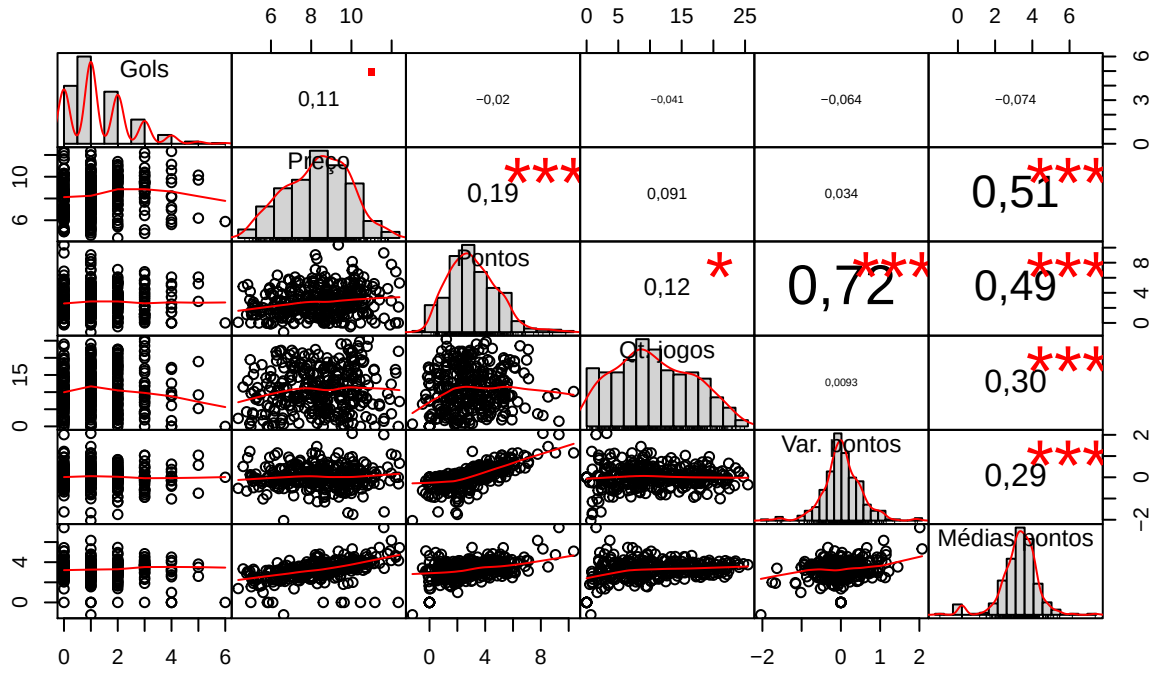


Figura 1: Associação entre o número de gols marcados e as covariáveis (médias gerais de cada posições) para o time mandante.

um gol marcado e para os visitantes, há quase um empate, entre nenhum gol marcado e um gol marcado.

Um resumo descritivo das variáveis consideradas no estudo, encontra-se na Tabela 12, em anexo.

A correlação da variável resposta (gols dos times mandantes), com as variáveis, media_pontos_geral_casa, media_qt_jogos_geral_casa, media_variacao_pontos_geral_casa e media_das_medias_pontos_geral_casa estão apresentadas na Figura 1. A mesma tabela de correlação foi utilizada para a variável resposta dos times visitantes (Figura 2).

É possível observar que não há uma forte correlação da variável resposta, com as demais variáveis. Nota-se que existe uma alta correlação entre a quantidade de pontos e variação dos pontos. Isso é esperado, dado que a variação dos pontos nas rodadas depende do número de pontos conquistados ou perdidos em cada rodada. A correlação entre pontos e médias de pontos ficou em torno de 0,5, tanto para mandantes quanto para visitantes. Isso ocorre porque a média provém da quantidade de pontos perdidos ou ganhados na rodada. E por fim, a correlação entre preço e média de pontos ficou um pouco acima de 0,5, indicando que o preço depende muito dos pontos que cada jogador recebe após cada rodada do campeonato.

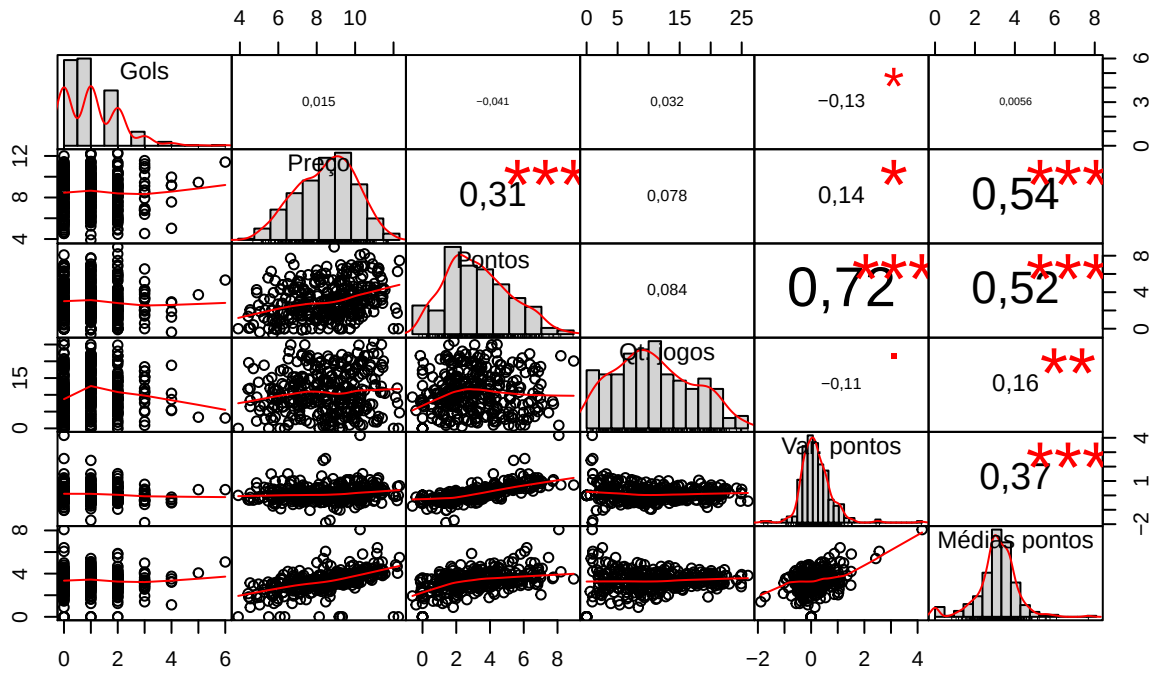


Figura 2: Associação entre o número de gols marcados e as covariáveis (médias gerais de cada posições) para o time visitante.

Ajuste de modelos

Após a seleção de variáveis, foi ajustado um modelo com função de ligação *logarítmica*, tanto para gols dos times mandantes, como para gols dos times visitantes (Tabelas 8 e 9).

Tabela 8: Coeficientes do modelo ajustado para a média de gols marcados pelo time mandante.

	Estimativa	Erro padrão	Valor Z	P-valor
(Intercept)	0,942	0,283	3,33	0,001
media_qt_jogos_mei_casa	-0,127	0,037	-3,43	0,001
max_variacao_pontos_geral_casa	-0,084	0,047	-1,78	0,074
min_pontos_mei_fora	-0,086	0,042	-2,04	0,042
max_qt_jogos_zag_casa	0,098	0,044	2,22	0,027
min_das_medias_pontos_mei_fora	0,134	0,061	2,21	0,027
min_preco_ata_fora	-0,042	0,017	-2,48	0,013
media_preco_mei_casa	0,039	0,019	2,04	0,042
media_das_medias_pontos_gol_fora	-0,063	0,022	-2,84	0,005
min_qt_jogos_ata_fora	0,044	0,021	2,06	0,040
media_preco_gol_fora	0,025	0,015	1,67	0,094
min_preco_mei_fora	-0,059	0,031	-1,89	0,059
max_qt_jogos_geral_casa	-0,070	0,038	-1,83	0,068
min_variacao_pontos_zag_fora	0,182	0,085	2,14	0,032
min_pontos_zag_fora	-0,060	0,034	-1,78	0,075
min_qt_jogos_zag_fora	0,030	0,020	1,50	0,134

Observando os coeficientes que foram significativos, verificam-se que apenas 5 variáveis pertencem ao time mandante e que o que mais determina a quantidade de gols marcados pelo time mandante, são as variáveis do time visitante. Nota-se também que o que mais influencia na queda dos gols marcados, é a média de jogos dos jogadores do meio de campo e o que mais influencia na marcação de gols, é a variação de pontos da zaga adversária (nesse caso, o mínimo dentre as rodadas).

Tabela 9: Coeficientes do modelo ajustado para a média de gols marcados pelo time visitante.

	Estimativa	Erro padrão	Valor Z	P-valor
(Intercept)	0,078	0,253	0,307	0,759
min_pontos_ata_fora	-0,056	0,022	-2,582	0,010
media_variacao_pontos_gol_fora	-0,153	0,052	-2,959	0,003
media_variacao_pontos_gol_casa	-0,111	0,048	-2,284	0,022
min_das_medias_pontos_mei_casa	0,132	0,048	2,742	0,006
min_preco_geral_casa	-0,090	0,042	-2,130	0,033
min_variacao_pontos_geral_casa	0,214	0,085	2,529	0,011
min_pontos_geral_casa	-0,085	0,041	-2,096	0,036
max_das_medias_pontos_mei_fora	0,103	0,035	2,967	0,003
max_pontos_zag_casa	0,033	0,016	2,034	0,042
max_das_medias_pontos_geral_fora	-0,048	0,030	-1,574	0,116
max_pontos_mei_fora	-0,030	0,021	-1,436	0,151
media_pontos_geral_fora	0,166	0,062	2,665	0,008
max_pontos_geral_fora	-0,037	0,024	-1,570	0,116

Com relação aos gols marcados pelo time visitante, ocorre o oposto do que ocorreu no modelo do time mandante. Apenas 5 variáveis foram significativas para os times visitantes. Nesse caso, o que mais afeta na marcação de gols, é a média da variação dos pontos do goleiro. Isso talvez possa ser explicado pelo fato de o time visitante sofrer gols e não conseguir reagir dutante a partida. A covariável que eleva o número de gols dos times visitantes é a variação dos pontos geral do time mandante (nesse caso, o mínimo dentre as rodadas).

Após o ajuste dos modelos, devemos verificar se o modelo foi bem ajustado às observações através do uso de envolope simulado e gráfico da *deviance vs valores preditos* (Figura 3).

Os gráficos demonstram que tanto o modelo de predição para os times mandantes, quanto o modelo para os times visitantes, foram bem ajustados. Não há nenhum ponto de resíduo fora dos envelopes simulados.

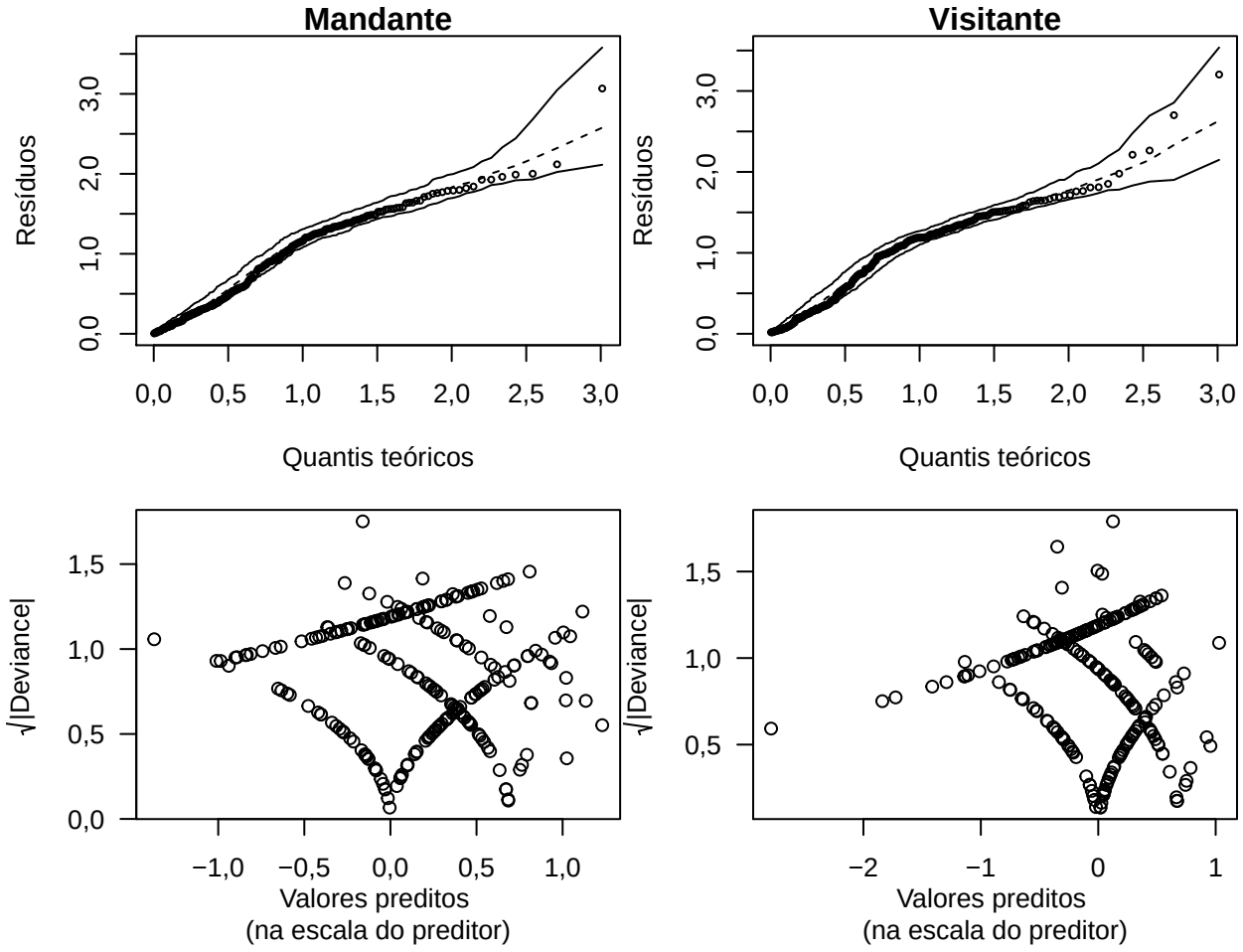


Figura 3: Resíduos com evelope simulado.

Predição

Após o ajuste e verificação do modelo, foram feitas algumas predições de resultados do campenato brasileiro.

Tabela 10: Exemplo de matriz de probabilidade conjunta do número de gols (time mandante na primeira coluna e visitante na primeira linha)

	0	1	2	3	4	5	6	7
0	0,0413	0,0427	0,0220	0,0076	0,0020	0,0004	0,0001	0,0000
1	0,0890	0,0919	0,0475	0,0163	0,0042	0,0009	0,0001	0,0000
2	0,0958	0,0990	0,0511	0,0176	0,0045	0,0009	0,0002	0,0000
3	0,0688	0,0710	0,0367	0,0126	0,0033	0,0007	0,0001	0,0000
4	0,0370	0,0382	0,0197	0,0068	0,0018	0,0004	0,0001	0,0000
5	0,0160	0,0165	0,0085	0,0029	0,0008	0,0002	0,0000	0,0000
6	0,0057	0,0059	0,0031	0,0011	0,0003	0,0001	0,0000	0,0000
7	0,0018	0,0018	0,0009	0,0003	0,0001	0,0000	0,0000	0,0000
8	0,0005	0,0005	0,0003	0,0001	0,0000	0,0000	0,0000	0,0000
9	0,0001	0,0001	0,0001	0,0000	0,0000	0,0000	0,0000	0,0000
10	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000

Como exemplo, utilizando os parâmetros da regressão foi calculada uma matriz de probabilidade conjunta (Tabela 13) para o número de gols marcados para o jogo Santos x Atlético-GO. Foi possível estimar uma probabilidade de vitória do time mandante de 0,63, fazendo a soma das probabilidades abaixo da diagonal principal; vitória do time visitante de 0,172, fazendo a soma das probabilidades acima da diagonal principal e a probabilidade de empate, fazendo a soma das probabilidades da diagonal principal 0,199. Essa matriz foi calculada para cada um dos jogos da base de validação e seus resultados estão sumarizados na tabela 13, em anexo.

Ao utilizar o desfecho da maior probabilidade como valor predito do resultado final da partida (se vitória, empate ou derrota do time mandante) observou-se 27 acertos em 79 partidas testadas, ou seja, obtém-se resultado positivo em torno de 34,18 das tentativas. Não mais que acertos casuais para o caso de 3 desfechos.

Ao dicotomizar o resultado, assumindo sucesso na vitória do time mandante e fracasso no caso de não vitória, ou seja, empate ou vitória do visitante, encontramos valores de sensibilidade e especificidade de 0,265 e 0,733.

Entretanto, utilizando as técnicas de otimização para variáveis dicotômicas, podemos encontrar o ponto em que poderemos, por exemplo, maximizar a sensibilidade e especificidade simultaneamente (Tabela 11). Onde é possível observar melhores resultados no ponto de corte em torno de 0,20.

Tabela 11: Sensibilidade e especificidade para resultado dicotomizado da predição para diferentes pontos de corte

Pontos de corte	Sensibilidade	Especificidade
0,1	0,882	0,156
0,2	0,706	0,400
0,3	0,471	0,600
0,4	0,206	0,756
0,5	0,118	0,822
0,6	0,059	0,889
0,7	0,000	0,978
0,8	0,000	0,978
0,9	0,000	0,978

Conclusão

O modelo mostra que é possível predizer o resultado de um jogo de futebol com algum nível de acurácia, porém, obter uma taxa elevada de precisão é muito mais difícil.

Inúmeras variáveis devem ser levadas em consideração. Um time por exemplo, pode estar jogando sem motivação por já estar eliminado, ou não ter mais nada a ganhar no campeonato, alterando a capacidade preditiva de qualquer modelo.

Outras distribuições e abordagens devem ser consideradas e avaliadas. Aumentar o número de variáveis e observações poderiam trazer maiores probabilidades de acerto, assim como técnicas de machine learning e inferência bayesiana estão em voga para esse tema. Outra distribuição proposta é a distribuição de Weibull, que poderia apresentar um melhor ajuste para a distribuição de gols ao longo de uma partida.

De modo geral, a regressão para dados de contagem mostra-se como um bom ponto de partida para modelagens mais sofisticadas.

Anexo

Tabela 12: Estatísticas descritivas das variáveis consideradas nos modelos

(Co)variável	Mín.	Q1	Mediana	Média	Q3	Máx.
gols_casa	0,000	1,000	1,000	1,349	2,000	6,000
gols_fora	0,000	0,000	1,000	1,066	2,000	6,000
media_preco_ata_casa	1,300	6,670	9,105	9,511	12,080	25,140
media_preco_gol_casa	0,710	5,862	9,615	9,843	12,655	24,470
media_preco_mei_casa	2,680	5,422	7,620	7,807	9,725	17,500
media_preco_zag_casa	1,350	6,055	7,365	7,567	8,975	13,770
media_pontos_ata_casa	-5,800	0,012	2,300	3,420	5,475	21,100
media_pontos_gol_casa	-8,000	-0,450	2,700	3,683	8,000	33,000
media_pontos_mei_casa	-1,230	0,633	2,160	2,484	3,903	12,220
media_pontos_zag_casa	-1,830	0,542	1,900	2,602	4,265	13,500
media_variacao_pontos_ata_casa	-5,100	-0,538	0,000	0,062	0,628	4,730
media_variacao_pontos_gol_casa	-8,290	-0,950	0,000	0,025	1,055	8,340
media_variacao_pontos_mei_casa	-3,760	-0,300	0,000	0,011	0,308	3,200
media_variacao_pontos_zag_casa	-2,160	-0,348	0,000	0,017	0,385	2,230
media_das_medias_pontos_ata_casa	-2,230	2,435	3,560	3,526	4,310	12,200
media_das_medias_pontos_gol_casa	-8,000	2,200	3,980	4,019	5,292	21,830
media_das_medias_pontos_mei_casa	-0,650	1,900	2,590	2,732	3,520	7,720
media_das_medias_pontos_zag_casa	-1,400	2,485	3,265	3,153	4,030	7,570
media_qt_jogos_ata_casa	0,000	5,330	10,000	11,271	16,000	31,000
media_qt_jogos_gol_casa	0,000	5,000	11,000	12,607	20,000	34,000
media_qt_jogos_mei_casa	0,000	5,670	9,775	10,637	15,438	28,250
media_qt_jogos_zag_casa	0,000	5,330	9,670	10,081	14,438	26,250
max_preco_ata_casa	1,300	8,228	11,200	11,952	15,005	26,120
max_preco_gol_casa	0,710	5,862	9,615	9,843	12,655	24,470
max_preco_mei_casa	3,810	7,728	11,000	11,511	14,842	23,000
max_preco_zag_casa	1,670	9,000	10,455	10,836	12,460	22,630
max_pontos_ata_casa	-5,800	0,600	3,600	5,442	8,775	29,900
max_pontos_gol_casa	-8,000	-0,450	2,700	3,683	8,000	33,000
max_pontos_mei_casa	-0,900	2,725	5,700	6,022	9,000	21,100
max_pontos_zag_casa	0,000	2,400	5,150	5,433	8,075	21,200
max_variacao_pontos_ata_casa	-5,100	-0,160	0,320	0,599	1,392	6,050
max_variacao_pontos_gol_casa	-8,290	-0,950	0,000	0,025	1,055	8,340
max_variacao_pontos_mei_casa	-2,200	0,050	0,670	0,875	1,377	5,290
max_variacao_pontos_zag_casa	-1,610	0,000	0,525	0,755	1,220	5,190
max_das_medias_pontos_ata_casa	-2,230	3,190	4,355	4,462	5,617	16,250
max_das_medias_pontos_gol_casa	-8,000	2,200	3,980	4,019	5,292	21,830
max_das_medias_pontos_mei_casa	0,000	3,172	4,105	4,355	5,435	11,350
max_das_medias_pontos_zag_casa	0,000	3,660	4,540	4,605	5,500	12,500
max_qt_jogos_ata_casa	0,000	6,000	12,000	12,767	19,000	31,000
max_qt_jogos_gol_casa	0,000	5,000	11,000	12,607	20,000	34,000
max_qt_jogos_mei_casa	0,000	7,000	13,000	14,038	21,000	33,000
max_qt_jogos_zag_casa	0,000	7,000	13,000	13,594	19,750	32,000
min_preco_ata_casa	0,770	4,178	6,455	7,291	9,640	25,140
min_preco_gol_casa	0,710	5,862	9,615	9,843	12,655	24,470
min_preco_mei_casa	0,760	2,670	4,000	4,506	5,790	14,050
min_preco_zag_casa	0,710	2,665	4,065	4,578	6,220	11,010
min_pontos_ata_casa	-6,400	-0,600	0,000	1,445	2,175	21,100
min_pontos_gol_casa	-8,000	-0,450	2,700	3,683	8,000	33,000
min_pontos_mei_casa	-4,900	-1,500	-0,100	-0,400	0,000	9,300
min_pontos_zag_casa	-5,200	-1,175	0,000	0,112	0,100	8,500
min_variacao_pontos_ata_casa	-8,140	-1,147	-0,420	-0,472	0,000	4,730
min_variacao_pontos_gol_casa	-8,290	-0,950	0,000	0,025	1,055	8,340
min_variacao_pontos_mei_casa	-6,220	-1,250	-0,730	-0,819	-0,192	1,720

(Co)variável	Mín.	Q1	Mediana	Média	Q3	Máx.
min_variacao_pontos_zag_casa	-5,810	-1,218	-0,645	-0,713	0,000	1,160
min_das_medias_pontos_ata_casa	-2,800	1,390	2,505	2,602	3,728	12,200
min_das_medias_pontos_gol_casa	-8,000	2,200	3,980	4,019	5,292	21,830
min_das_medias_pontos_mei_casa	-2,700	0,200	1,150	1,183	1,920	5,830
min_das_medias_pontos_zag_casa	-4,900	0,512	2,115	1,760	2,888	6,200
min_qt_jogos_ata_casa	0,000	4,000	8,000	9,679	14,000	31,000
min_qt_jogos_gol_casa	0,000	5,000	11,000	12,607	20,000	34,000
min_qt_jogos_mei_casa	0,000	2,250	5,000	6,950	10,000	26,000
min_qt_jogos_zag_casa	0,000	2,000	5,000	5,943	9,000	22,000
media_preco_geral_casa	4,370	7,130	8,575	8,425	9,748	12,390
media_pontos_geral_casa	-1,230	1,400	2,710	2,884	4,140	10,360
media_variacao_pontos_geral_casa	-2,040	-0,240	0,000	0,026	0,308	2,070
media_das_medias_pontos_geral_casa	-1,230	2,752	3,285	3,202	3,797	7,440
media_qt_jogos_geral_casa	0,000	6,020	10,270	10,805	15,560	25,450
max_preco_geral_casa	8,000	12,840	15,205	15,903	19,678	26,120
max_pontos_geral_casa	0,000	6,625	9,550	9,941	12,775	33,000
max_variacao_pontos_geral_casa	0,000	0,912	1,515	1,726	2,270	8,340
max_das_medias_pontos_geral_casa	0,000	5,242	6,300	6,714	7,960	21,830
max_qt_jogos_geral_casa	0,000	8,000	16,000	15,925	24,000	34,000
min_preco_geral_casa	0,710	1,943	2,695	3,020	3,950	8,660
min_pontos_geral_casa	-8,000	-2,900	-1,800	-1,947	-0,700	1,500
min_variacao_pontos_geral_casa	-8,290	-1,968	-1,390	-1,550	-0,930	0,000
min_das_medias_pontos_geral_casa	-8,000	-0,360	0,540	0,365	1,300	3,110
min_qt_jogos_geral_casa	0,000	1,000	3,000	3,679	5,000	19,000
media_preco_geral_fora	3,910	7,118	8,595	8,462	9,815	12,310
media_pontos_geral_fora	-0,600	1,645	2,900	3,071	4,375	8,960
media_variacao_pontos_geral_fora	-1,890	-0,240	0,035	0,126	0,428	4,170
media_das_medias_pontos_geral_fora	0,000	2,830	3,290	3,267	3,860	8,080
media_qt_jogos_geral_fora	0,000	5,895	10,435	10,727	15,330	26,000
max_preco_geral_fora	7,000	12,495	15,450	15,807	19,338	26,230
max_pontos_geral_fora	0,000	6,725	9,900	10,024	12,900	32,100
max_variacao_pontos_geral_fora	0,000	0,970	1,565	1,882	2,375	14,080
max_das_medias_pontos_geral_fora	0,000	5,272	6,335	6,770	7,928	22,800
max_qt_jogos_geral_fora	0,000	8,000	15,500	15,921	24,000	33,000
min_preco_geral_fora	0,710	1,770	2,835	3,084	4,053	8,690
min_pontos_geral_fora	-9,000	-2,500	-1,400	-1,619	-0,200	4,000
min_variacao_pontos_geral_fora	-6,780	-1,870	-1,355	-1,433	-0,795	0,000
min_das_medias_pontos_geral_fora	-8,300	-0,158	0,530	0,348	1,300	3,120
min_qt_jogos_geral_fora	0,000	1,000	2,000	3,642	5,000	23,000
media_preco_ata_fora	0,760	6,690	8,745	9,356	11,835	24,000
media_preco_gol_fora	0,710	6,002	9,605	9,848	12,623	24,690
media_preco_mei_fora	1,640	5,400	7,505	7,840	9,938	15,640
media_preco_zag_fora	2,250	6,070	7,550	7,648	9,220	14,330
media_pontos_ata_fora	-6,600	0,105	2,240	3,315	5,475	18,800
media_pontos_gol_fora	-9,000	0,000	2,000	3,394	7,000	23,000
media_pontos_mei_fora	-2,270	1,032	2,475	2,761	4,165	10,430
media_pontos_zag_fora	-2,270	0,830	2,200	2,932	4,873	12,600
media_variacao_pontos_ata_fora	-3,100	-0,548	0,000	0,173	0,692	7,780
media_variacao_pontos_gol_fora	-5,510	-0,650	0,000	0,094	0,790	5,290
media_variacao_pontos_mei_fora	-3,000	-0,248	0,030	0,091	0,378	5,300
media_variacao_pontos_zag_fora	-3,710	-0,338	0,005	0,133	0,560	7,380
media_das_medias_pontos_ata_fora	-3,300	2,468	3,510	3,535	4,283	13,900
media_das_medias_pontos_gol_fora	-8,300	1,940	3,910	3,884	5,137	17,000
media_das_medias_pontos_mei_fora	0,000	1,890	2,645	2,827	3,688	9,880
media_das_medias_pontos_zag_fora	-0,380	2,580	3,335	3,252	4,042	12,600
media_qt_jogos_ata_fora	0,000	5,000	10,165	11,212	16,247	33,000
media_qt_jogos_gol_fora	0,000	4,000	11,000	12,277	19,000	33,000

(Co)variável	Mín.	Q1	Mediana	Média	Q3	Máx.
media_qt_jogos_mei_fora	0,000	5,543	10,000	10,702	15,500	26,670
media_qt_jogos_zag_fora	0,000	5,000	9,000	9,974	14,250	28,500
max_preco_ata_fora	0,760	8,025	10,830	11,600	14,365	26,230
max_preco_gol_fora	0,710	6,002	9,605	9,848	12,623	24,690
max_preco_mei_fora	1,640	7,863	11,090	11,649	15,100	23,000
max_preco_zag_fora	3,960	9,000	10,650	10,798	12,293	23,110
max_pontos_ata_fora	-6,600	1,025	3,750	5,297	8,800	32,100
max_pontos_gol_fora	-9,000	0,000	2,000	3,394	7,000	23,000
max_pontos_mei_fora	0,000	3,100	5,900	6,295	8,600	22,500
max_pontos_zag_fora	0,000	2,800	5,600	6,042	8,600	22,800
max_variacao_pontos_ata_fora	-3,100	0,000	0,240	0,719	1,325	8,910
max_variacao_pontos_gol_fora	-5,510	-0,650	0,000	0,094	0,790	5,290
max_variacao_pontos_mei_fora	-0,920	0,172	0,780	0,956	1,418	9,780
max_variacao_pontos_zag_fora	-0,980	0,090	0,725	0,970	1,513	14,080
max_das_medias_pontos_ata_fora	-3,300	3,215	4,360	4,495	5,487	15,900
max_das_medias_pontos_gol_fora	-8,300	1,940	3,910	3,884	5,137	17,000
max_das_medias_pontos_mei_fora	0,000	3,240	4,175	4,534	5,707	15,900
max_das_medias_pontos_zag_fora	0,000	3,770	4,670	4,772	5,675	22,800
max_qt_jogos_ata_fora	0,000	6,000	12,000	12,777	19,000	33,000
max_qt_jogos_gol_fora	0,000	4,000	11,000	12,277	19,000	33,000
max_qt_jogos_mei_fora	0,000	7,000	13,000	13,972	21,000	32,000
max_qt_jogos_zag_fora	0,000	7,000	13,000	13,535	19,000	31,000
min_preco_ata_fora	0,760	4,220	6,375	7,314	9,707	24,000
min_preco_gol_fora	0,710	6,002	9,605	9,848	12,623	24,690
min_preco_mei_fora	0,760	2,428	4,000	4,397	5,605	13,350
min_preco_zag_fora	0,730	2,880	4,640	4,848	6,612	10,580
min_pontos_ata_fora	-6,600	-0,300	0,000	1,475	2,200	18,800
min_pontos_gol_fora	-9,000	0,000	2,000	3,394	7,000	23,000
min_pontos_mei_fora	-7,000	-1,000	0,000	-0,168	0,175	9,100
min_pontos_zag_fora	-6,000	-0,900	0,000	0,295	0,900	7,800
min_variacao_pontos_ata_fora	-5,760	-1,005	-0,310	-0,358	0,000	7,780
min_variacao_pontos_gol_fora	-5,510	-0,650	0,000	0,094	0,790	5,290
min_variacao_pontos_mei_fora	-6,010	-1,190	-0,600	-0,749	-0,150	2,600
min_variacao_pontos_zag_fora	-6,780	-1,197	-0,545	-0,671	0,000	2,560
min_das_medias_pontos_ata_fora	-3,300	1,370	2,380	2,613	3,748	13,900
min_das_medias_pontos_gol_fora	-8,300	1,940	3,910	3,884	5,137	17,000
min_das_medias_pontos_mei_fora	-2,240	0,185	1,140	1,190	1,825	9,100
min_das_medias_pontos_zag_fora	-6,000	0,815	2,095	1,852	2,945	6,100
min_qt_jogos_ata_fora	0,000	4,000	8,000	9,525	14,000	33,000
min_qt_jogos_gol_fora	0,000	4,000	11,000	12,277	19,000	33,000
min_qt_jogos_mei_fora	0,000	2,000	6,000	7,031	10,000	24,000
min_qt_jogos_zag_fora	0,000	2,000	5,000	6,142	9,000	27,000

Tabela 13: Probabilidade de vitória do time mandante, empate e vitória do time visitante para cada um dos jogos da base de validação (rodada 26 até a 35)

Rodada	Mandante	Placar	Visitante	Prob. vit. mand.	Prob. Empate	Prob. vit. visit.
27	Atlético - PR	2 x 2	Atlético - GO	0,360	0,249	0,391
27	Corinthians - SP	3 x 1	Coritiba - PR	0,200	0,381	0,418
27	Atlético - MG	1 x 0	São Paulo - SP	0,397	0,355	0,249
27	Grêmio - RS	0 x 1	Cruzeiro - MG	0,381	0,366	0,253
27	Avaí - SC	1 x 2	Vasco da Gama - RJ	0,172	0,508	0,320
27	Flamengo - RJ	1 x 1	Fluminense - RJ	0,289	0,388	0,323
27	Vitória - BA	1 x 2	Sport - PE	0,066	0,199	0,735
27	Ponte Preta - SP	1 x 1	Santos - SP	0,535	0,197	0,268
27	Palmeiras - SP	2 x 2	Bahia - BA	0,261	0,422	0,317
28	Vasco da Gama - RJ	1 x 0	Botafogo - RJ	0,302	0,362	0,336
28	Fluminense - RJ	1 x 0	Avaí - SC	0,400	0,285	0,315
28	Sport - PE	1 x 1	Atlético - MG	0,199	0,281	0,520
28	Atlético - GO	1 x 3	Palmeiras - SP	0,471	0,305	0,224
28	Chapecoense - SC	0 x 1	Flamengo - RJ	0,917	0,048	0,035
28	Coritiba - PR	0 x 1	Grêmio - RS	0,044	0,274	0,682
28	Bahia - BA	2 x 0	Corinthians - SP	0,369	0,267	0,364
28	Santos - SP	2 x 2	Vitória - BA	0,564	0,235	0,201
29	Atlético - GO	0 x 1	Vasco da Gama - RJ	0,117	0,207	0,676
29	Coritiba - PR	1 x 0	Cruzeiro - MG	0,061	0,304	0,635
29	Atlético - MG	2 x 3	Chapecoense - SC	0,247	0,342	0,411
29	Corinthians - SP	0 x 0	Grêmio - RS	0,195	0,445	0,361
29	Fluminense - RJ	3 x 1	São Paulo - SP	0,369	0,219	0,413
29	Avaí - SC	1 x 1	Botafogo - RJ	0,408	0,248	0,344
29	Palmeiras - SP	2 x 0	Ponte Preta - SP	0,309	0,379	0,312
29	Vitória - BA	2 x 3	Atlético - PR	0,042	0,243	0,716
29	Flamengo - RJ	4 x 1	Bahia - BA	0,166	0,290	0,544
29	Sport - PE	1 x 1	Santos - SP	0,622	0,232	0,146
30	Botafogo - RJ	2 x 1	Corinthians - SP	0,135	0,266	0,599
30	Cruzeiro - MG	1 x 3	Atlético - MG	0,311	0,354	0,336
30	Grêmio - RS	1 x 3	Palmeiras - SP	0,241	0,303	0,456
30	Vasco da Gama - RJ	1 x 1	Coritiba - PR	0,134	0,367	0,499
30	Bahia - BA	2 x 1	Vitória - BA	0,323	0,361	0,316
30	Santos - SP	1 x 0	Atlético - GO	0,630	0,199	0,172
30	Ponte Preta - SP	1 x 2	Avaí - SC	0,645	0,206	0,150
30	Chapecoense - SC	2 x 0	Fluminense - RJ	0,432	0,197	0,372
30	Atlético - PR	2 x 1	Sport - PE	0,424	0,301	0,275
31	São Paulo - SP	2 x 1	Santos - SP	0,270	0,321	0,409
31	Atlético - PR	0 x 0	Chapecoense - SC	0,301	0,367	0,332
31	Fluminense - RJ	1 x 1	Bahia - BA	0,362	0,251	0,387
31	Atlético - MG	0 x 0	Botafogo - RJ	0,283	0,265	0,452
31	Ponte Preta - SP	1 x 0	Corinthians - SP	0,517	0,232	0,251
31	Sport - PE	3 x 4	Coritiba - PR	0,262	0,297	0,441
31	Vitória - BA	1 x 1	Atlético - GO	0,089	0,359	0,551
31	Avaí - SC	2 x 2	Grêmio - RS	0,146	0,307	0,547
32	Santos - SP	3 x 1	Atlético - MG	0,258	0,272	0,470
32	Botafogo - RJ	1 x 2	Fluminense - RJ	0,155	0,293	0,553
32	Coritiba - PR	4 x 0	Avaí - SC	0,267	0,337	0,395
32	Corinthians - SP	3 x 2	Palmeiras - SP	0,241	0,416	0,343
32	Grêmio - RS	3 x 1	Flamengo - RJ	0,322	0,303	0,375
32	Chapecoense - SC	1 x 1	Sport - PE	0,602	0,215	0,183
32	Bahia - BA	2 x 0	Ponte Preta - SP	0,342	0,420	0,239
32	Vasco da Gama - RJ	1 x 1	Vitória - BA	0,075	0,256	0,668
32	Atlético - GO	0 x 1	São Paulo - SP	0,040	0,203	0,756

Rodada	Mandante	Placar	Visitante	Prob. vit. mand.	Prob. Empate	Prob. vit. visit.
33	Avaí - SC	1 x 2	Bahia - BA	0,049	0,160	0,791
33	Ponte Preta - SP	0 x 1	Grêmio - RS	0,106	0,333	0,561
33	Sport - PE	1 x 2	Botafogo - RJ	0,598	0,245	0,157
33	Atlético - PR	0 x 1	Corinthians - SP	0,424	0,330	0,246
33	Flamengo - RJ	2 x 0	Cruzeiro - MG	0,078	0,217	0,706
33	Santos - SP	1 x 2	Vasco da Gama - RJ	0,267	0,390	0,343
33	Vitória - BA	3 x 1	Palmeiras - SP	0,087	0,300	0,614
33	São Paulo - SP	2 x 2	Chapecoense - SC	0,681	0,180	0,139
33	Atlético - MG	3 x 2	Atlético - GO	0,376	0,249	0,376
33	Fluminense - RJ	2 x 2	Coritiba - PR	0,349	0,236	0,416
34	Botafogo - RJ	0 x 1	Atlético - PR	0,190	0,404	0,406
34	Corinthians - SP	1 x 0	Avaí - SC	0,094	0,306	0,599
34	Vasco da Gama - RJ	1 x 1	São Paulo - SP	0,181	0,247	0,572
34	Palmeiras - SP	2 x 0	Flamengo - RJ	0,264	0,353	0,383
34	Grêmio - RS	1 x 1	Vitória - BA	0,243	0,425	0,332
34	Bahia - BA	2 x 2	Atlético - MG	0,322	0,312	0,366
34	Coritiba - PR	1 x 1	Ponte Preta - SP	0,163	0,347	0,490
34	Chapecoense - SC	2 x 0	Santos - SP	0,187	0,268	0,546
35	Ponte Preta - SP	2 x 1	Atlético - PR	0,137	0,270	0,593
35	Grêmio - RS	1 x 0	São Paulo - SP	0,236	0,455	0,309
35	Vasco da Gama - RJ	1 x 1	Atlético - MG	0,248	0,389	0,363
35	Corinthians - SP	3 x 1	Fluminense - RJ	0,179	0,295	0,526
35	Palmeiras - SP	5 x 1	Sport - PE	0,557	0,262	0,181
35	Chapecoense - SC	2 x 1	Vitória - BA	0,695	0,183	0,123
35	Coritiba - PR	1 x 0	Flamengo - RJ	0,134	0,354	0,512
35	Bahia - BA	3 x 1	Santos - SP	0,210	0,331	0,459