UNIVERSIDADE FEDERAL DO PARANÁ

André Luiz Grion – GRR20159284

Bruno Henrique Abreu – GRR20159983

Maria Tereza Neves de Oliveira – GRR20159323

Predição de vitória de times mandantes no campeonato brasileiro 2017 via dados do Cartola

CURITIBA

Outubro de 2017

Resumo

Devido ao crescente sucesso de jogos tipo *fantasy* para esportes e a popularidade do jogo Cartola F.C. sobre futebol, surgiu a curiosidade de saber se as informações fornecidas por este tipo de jogo são condizentes com a realidade. Assumindo que a vitória de um time de futebol depende do bom desempenho dos seus atletas e que jogos tipo *fantasy* são baseados na performance do atleta em cada rodada, esse trabalho foi realizado com o objetivo de verificar a possibilidade de predição de vitória de um time com base nas informações de desempenho de seus atletas, fornecidas pelo Cartola F.C. Nesse contexto, o banco de dados do Cartola foi extraído da plataforma Kaggle e ajustado criteriosamente para a escolha da modelagem estatística que melhor descrevesse os dados. A variável resposta foi dicotomizada, juntando empate e derrota do time mandante em uma única classe. A seleção das variáveis foi realizada pelo método stepwise e para avaliação do poder preditivo de cada modelo testado foi utilizada a área sobre a curva ROC. O modelo com função de ligação Cauchit foi o mais adequado de acordo com o AIC, porém o modelo com função de ligação Logit foi superior para predição. A área abaixo da curva ROC do modelo Logit foi de 0,642. De modo geral, as análises confirmaram a dificuldade para realizar previsões no futebol, muitas alternativas, porém, podem ser estudadas para melhorar as estimativas. Os *fantasy games* são relativamente novos não só no Brasil, como nos Estados Unidos e as literaturas sobre o tema ainda não foram muito difundidas ou publicadas. Fica a certeza de ser um campo extremamente interessante e desafiador.

Palavras-chave: Futebol; Predição; Cartola; Regressão dados binários; Modelos lineares generalizados; Funções de ligação; Distribuição Binomial.

Conteúdo

Introdução

O Cartola FC é um game de futebol (derivado do inglês fantasy game) que permite que o jogador crie e administre seu próprio time. Os resultados oficiais de cada partida do campeonato brasileiro e o desempenho dos jogadores nos times reais se transformam em pontos para os mesmos atletas nos times virtuais. Os pontos são recebidos (ou perdidos) de acordo com as atuações dos atletas em campo após cada jogo através dos itens de scout. A pontuação geral, dada pela média dos pontos dos jogadores de cada time, determina a posição destas equipes no ranking das diferentes ligas nas quais eles estão inseridos. No início do campeonato são levados em consideração qualidades, características e o histórico profissional de cada atleta e técnico do time. A cada rodada do campeonato brasileiro (série A) as estatísticas de desempenho dos jogadores são consolidadas e transformadas em pontos, os quais são determinados por um time de profissionais especializados no game e disponibilizados no website oficial do Cartola. Por exemplo: um gol marcado vale oito pontos para o jogador, já um gol contra faz com que o atleta perca seis pontos, e assim por diante.

Dessa maneira, o objetivo desse trabalho foi desenvolver um modelo para descrever o comportamento dos dados e predizer a vitória dos times mandantes do campeonato brasileiro de 2017. O modelo considera a utilização dos dados do Cartola F.C, retirados da plataforma Kaggle e devidamente ajustados para facilitar a modelagem. Vale salientar que os dados disponibilizados originalmente na plataforma Kaggle, fazem parte de uma competição para predizer as pontuações e os preços dos jogadores.

Material e Métodos

Construção ABT

Para a construção da ABT (Analytical Base Table), utilizamos as bases de dados disponibilizadas na plataforma Kaggle, que estavam divididas em 34 bases. Cada base disponibilizava informações diferentes, como segue breve resumo abaixo:

1)*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.

2)*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.

3)*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, gols tomados).

4)*cartola_aggregated* - Informações dos atletas (40.296 observações, 77 variáveis). A base contém as informações dos atletas, assim como as bases citadas acima, porém até a 38ª rodada,porém não indica de qual time é o atleta.

5)*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 26ª rodada.

6)*tabela_times* - Informações dos jogos (20 observações, 17 variáveis). A base contém a classificação do campeonato brasileiro 2017 com os 20 times, atualizado até a 26ª rodada.

7)*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 26ª rodada.

8)*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 26ª rodada.

Apesar das bases citadas acima conter 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 as bases disponibilizadas de cada rodada para criar a base usada na modelagem. Abaixo segue breve resumo das 26 bases, referente a cada rodada:

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_6 (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)

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:

Provável - O jogador tem uma possibilidade alta de jogar na rodada, ou seja, está a disposição para jogar.

Dúvida - O jogador possivelmente jogue a rodada, ou seja, está a disposição para jogar porém depende de fatores extra-campo (recuperação de lesão).

Nulo - O jogador não irá jogar na rodada devido a problemas extra-campo (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, referente a jogadores, também estão disponíveis, tais como:

“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).

Todas as variáveis citadas acima são atualizadas após o término de cada rodada. Como a intenção 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 realizado a sumarização dos jogadores por posição em cada clube. As posições são:

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 (tecnico 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" realizar a média dos jogadores de cada posição como uma varíavel para cada time.

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 260 observações(todos os confrontos até a 26ª rodada do campeonato), em que cada linha da base contém a média de cada variável (disponibilizadas nas 26 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.

Ao todo, construímos mais de 800 variáveis para teste.

Todas as variáveis construídas para os times da casa também foram construídas para os times que jogaram fora de casa. O intuito de fazer as variáveis para os times que jogam fora de casa, é verificar se alguma variável é significativa para a vitória ou derrota do time que joga em casa. A variável resposta foi construída de tal modo que, se o time joga em casa recebe 1, se empatou ou perdeu a partida, recebe 0. Após a validação dos dados, foi possivel então começar as análises para o desenvolvimento do modelo.

Todas as variáveis ligadas ao *scout* apresentaram pelo menos 21 observações perdidas e portanto foram excluídas da base. As variáveis ligadas ao técnico apresentaram pelo menos 11 observações perdidas e também foram excluídas da base.

Restaram, portanto, após a higienização da base construída, 260 observações (jogos) e 40 variáveis do Cartola.

Base de dados para ajuste e para validacão

A base de dados foi dividida: as rodadas de 1 a 20 (200 partidas) foram utilizadas para a realização dos ajustes do modelo e as partidas referentes às rodadas 21 a 26 (60 partidas) foram utilizadas para a validação, representando essas 23,08% dos dados disponíveis.

Métodos estatísticos

Embora uma partida de futebol tenha três desfechos possíveis (vitória, empate ou derrota), a variável resposta de interesse foi dicotomizada para representar somente a probabilidade de vitória do time mandante. Para a variável resposta, então, foi atribuído o valor um no caso de vitória do time mandante e zero, caso contrário. A variável resposta apresenta, portanto, distribuição *Bernoulli* que, sendo um caso especial da distribuição *Binomial* para $m_i = 1$, no contexto de modelos lineares generalizados fica representada como:

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

em que y_i é a i-ésima variável resposta com probabilidade μ_i; $\boldsymbol{x_i}$ é o vetor com as p variáveis explanatórias associadas à i-ésima resposta e $\boldsymbol{\beta}$ é o vetor com $p+1$ parâmetros.

Foram testadas 4 funções de ligação ($g(\mu_i)$):

- Logística (*Logit*): $log(\frac{\mu_i}{1-\mu_i})$;
- *Probit*: $\Phi^{-1}(\mu_i)$;
- Complemento log-log: $log(-log(1-\mu_i))$;
- *Cauchit*: $tan[\pi(\mu_i - \frac{1}{2})]$.

A seleção das variáveis foi realizada pelo método *stepwise*, 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. O AIC também foi utilizado para comparar os modelos com diferentes funções de ligação.

Para avaliação do poder preditivo de cada modelo foi utilizada a Área sob a curva ROC (AUC).

Resultados e discussão

A proporção de vitórias do time mandante para a base completa foi 0,431 (Figura 1). Nas bases de ajuste e validação as proporções foram de 0,44 e 0,4, respectivamente.

Na Tabela 1 são apresentados os valores mínimos, máximos 1^{os} e 3^{os} quartis, média e mediana das variáveis obtidas no Cartola.

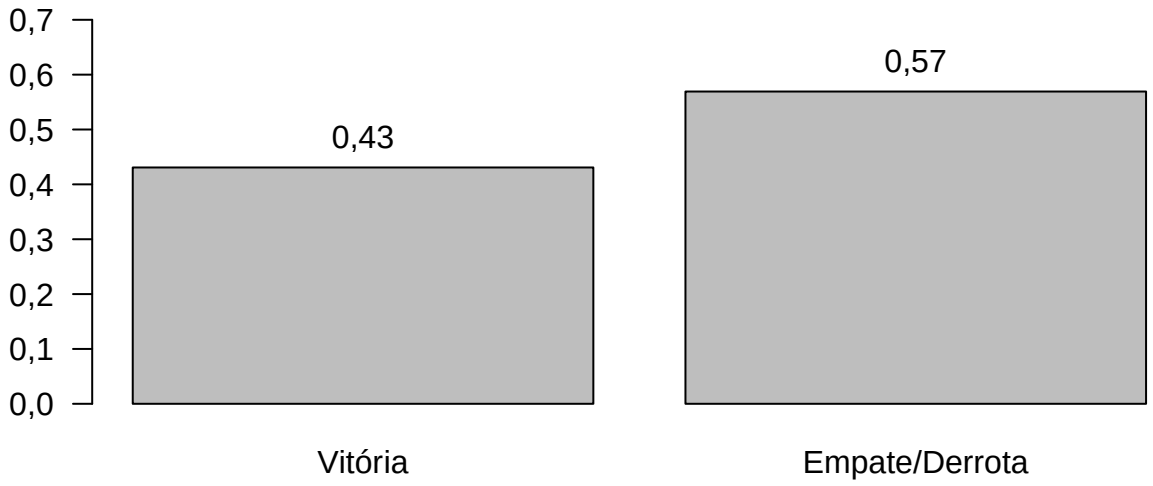

Figura 1: Proporção de resultados para o time mandante.

Tabela 1: Estatística descritiva das variáveis consideradas nos modelos

	Covariável	Mín.	Q1	Mediana	Média	Q3	Máx.
1	media_preco_ata_casa	2,250	4,038	4,765	5,117	5,985	12,830
2	media_preco_gol_casa	1,050	3,375	4,490	4,564	5,595	7,960
3	media_preco_mei_casa	3,160	4,470	5,210	5,310	6,030	8,850
4	media_preco_zag_casa	2,960	4,340	5,120	5,332	6,290	8,150
5	media_pontos_ata_casa	-1,060	0,068	0,640	0,883	1,450	5,090
6	media_pontos_gol_casa	-2,770	-0,085	0,800	0,845	1,878	5,500
7	media_pontos_mei_casa	-0,260	0,320	0,790	0,852	1,223	4,960
8	media_pontos_zag_casa	-0,670	0,268	0,760	1,011	1,708	4,290
9	media_variacao_pontos_ata_casa	-1,690	-0,192	-0,040	-0,041	0,143	1,350
10	media_variacao_pontos_gol_casa	-2,070	-0,230	0,000	-0,013	0,260	1,670
11	media_variacao_pontos_mei_casa	-1,070	-0,150	-0,020	-0,042	0,073	1,230
12	media_variacao_pontos_zag_casa	-0,690	-0,160	-0,005	-0,004	0,160	0,700
13	media_das_medias_pontos_ata_casa	-0,200	0,958	1,285	1,401	1,792	5,160
14	media_das_medias_pontos_gol_casa	-2,770	0,275	0,970	1,033	1,482	4,310
15	media_das_medias_pontos_mei_casa	-0,190	0,888	1,260	1,329	1,693	4,240
16	media_das_medias_pontos_zag_casa	-0,150	1,167	1,750	1,766	2,353	4,910
17	media_qt_jogos_ata_casa	0,000	2,232	4,105	4,343	6,133	13,130
18	media_qt_jogos_gol_casa	0,000	1,400	2,800	2,846	4,200	8,670
19	media_qt_jogos_mei_casa	0,000	2,160	4,475	4,559	6,677	11,140
20	media_qt_jogos_zag_casa	0,000	2,130	4,250	4,198	6,122	10,000
21	media_preco_ata_fora	2,260	4,060	4,775	5,166	5,973	12,550
22	media_preco_gol_fora	0,940	3,285	4,585	4,561	5,557	8,140
23	media_preco_mei_fora	3,090	4,448	5,230	5,364	6,100	8,590
24	media_preco_zag_fora	3,110	4,385	5,130	5,401	6,442	8,720

	Covariável	Mín.	Q1	Mediana	Média	Q3	Máx.
25	media_pontos_ata_fora	-1,160	0,165	0,700	1,001	1,645	4,970
26	media_pontos_gol_fora	-2,250	0,000	0,675	0,851	1,680	5,000
27	media_pontos_mei_fora	-0,760	0,370	0,890	0,939	1,303	5,590
28	media_pontos_zag_fora	-0,780	0,278	0,790	1,145	1,980	4,430
29	media_variacao_pontos_ata_fora	-1,010	-0,162	-0,005	0,015	0,152	1,830
30	media_variacao_pontos_gol_fora	-1,380	-0,170	0,000	0,026	0,192	1,320
31	media_variacao_pontos_mei_fora	-0,800	-0,110	-0,010	0,007	0,080	3,080
32	media_variacao_pontos_zag_fora	-1,400	-0,150	0,000	0,038	0,202	2,470
33	media_das_medias_pontos_ata_fora	-0,140	0,980	1,355	1,465	1,810	4,570
34	media_das_medias_pontos_gol_fora	-2,010	0,312	0,925	1,060	1,530	4,900
35	media_das_medias_pontos_mei_fora	0,000	0,930	1,295	1,395	1,750	5,590
36	media_das_medias_pontos_zag_fora	-0,080	1,220	1,790	1,857	2,425	4,830
37	media_qt_jogos_ata_fora	0,000	2,208	4,140	4,358	6,115	12,830
38	media_qt_jogos_gol_fora	0,000	1,250	2,750	2,848	4,200	8,000
39	media_qt_jogos_mei_fora	0,000	2,275	4,405	4,526	6,647	11,640
40	media_qt_jogos_zag_fora	0,000	2,000	4,155	4,175	6,000	10,890

Observa-se na Figura 2 altas correlações (acima de 0,80) entre as covariáveis relacionadas à quantidade de jogos e as diferentes posições. Isso é esperado pois a quantidade de jogos tende a aumentar com o passar das rodadas.

Correlações entre 0,50 e 0,80 ocorrem entre as médias de pontos e preços dos jogadores e, entre pontos e variação de pontos além das médias de pontos dentro de cada posição. Algumas correlações entre pontos e médias de pontos ocorrem também entre posições, demonstrando inclusive, uma coerência dos posicionamentos em campo, como as correlações entre goleiros e zagueiros e destes com o meio-campo.

As figuras a seguir descrevem as relações existentes entre as covariáveis e a probabilidade de vitória do time mandante, considerando as informações do próprio time mandante (Fig. 3) e as informações do time visitante (Fig. 4).

Para as informações do próprio time mandante, esperava-se observar relações positivas entre todas as variáveis e a probabilidade de vitória, como acontece com as variáveis relacionadas a preço ou pontos do goleiro. Entretanto, observa-se uma relação negativa entre a quantidade de jogos e a probabilidade de vitória.

Para as covariáveis relacionadas ao time visitante, esperava-se observar correlações negativas como as que ocorreram entre as quantidade de jogos e as posições e, para o preço e média de pontos para atacantes e zagueiros. Entretanto também são observadas correlações positivas, como a encontrada com a variação de pontos do goleiro.

Modelos ajustados

Na Tabela 2 observam-se os já esperados efeitos positivos para as explanatórias referentes ao time mandante em todos os modelos, como por exemplo, a média de variação de pontos dos goleiros e a média de preços dos zagueiros. A quantidade de jogos para o meio-campo, que apresentou coeficiente negativo, pode ser devido à previsibilidade apresentada pela repetição dos jogadores ou

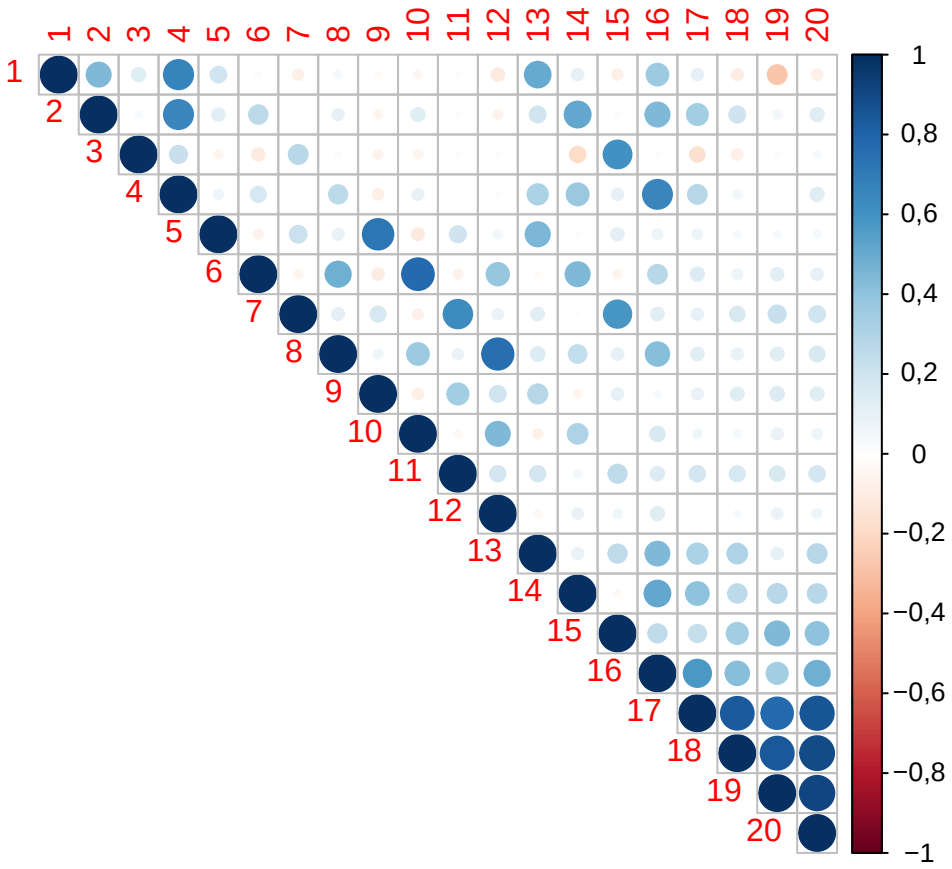

Figura 2: Correlograma das covariáveis (ver numeração na Tabela 1).

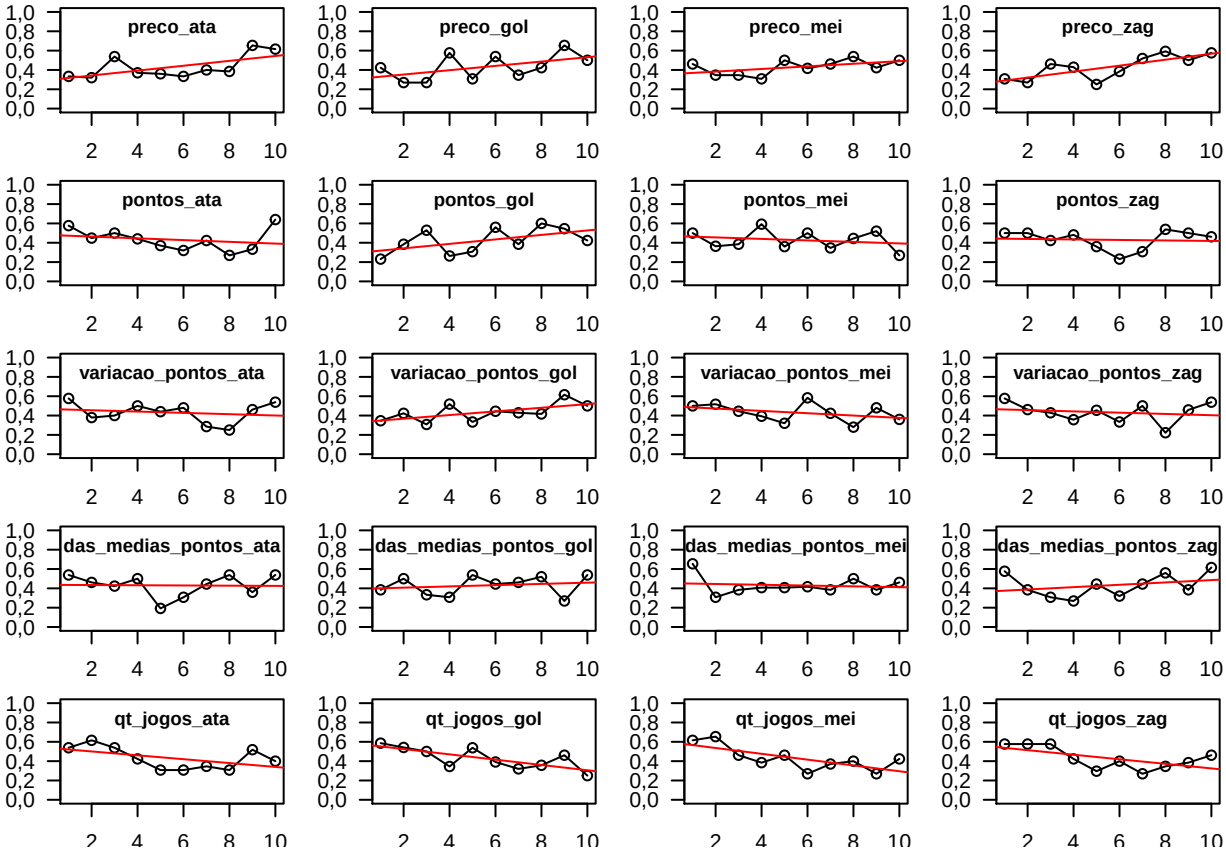

Figura 3: Relação entre cada covariável (eixo X, em decis) com a probabilidade de vitória (eixo Y) para as informações do time da casa.

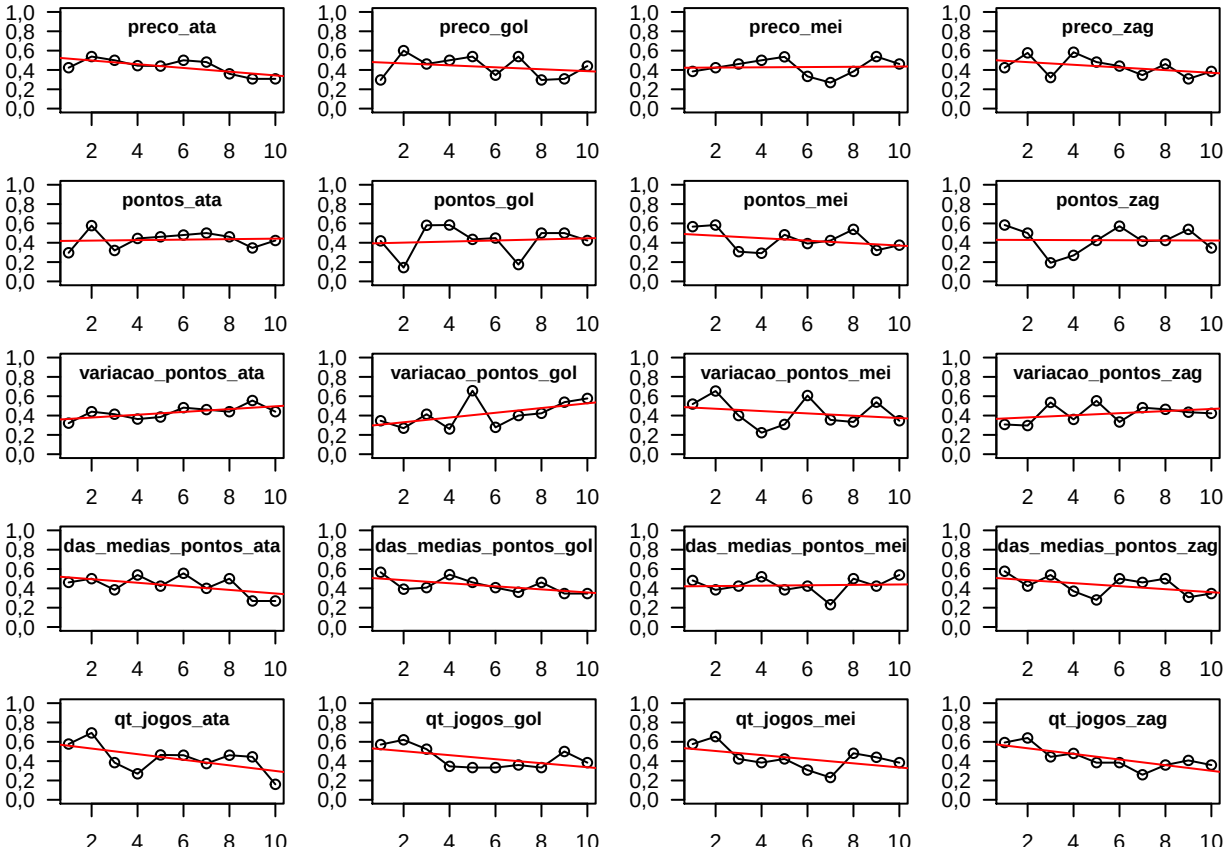

Figura 4: Relação entre cada covariável (eixo X, em decis) com a probabilidade de vitória (eixo Y) para as informações dos times visitantes.

simplesmente o cansaço do decorrer da temporada. A quantidade de jogos para os atacantes do time mandante foram significativos e apresentaram coeficientes negativos para os modelos com função de ligação complemento log-log e Cauchit.

Dois efeitos referentes aos goleiros do time visitante também foram selecionados em todos os modelos e apresentaram coeficientes com sinais opostos. Quanto maior a média de pontos do goleiro adversário, menor é a probabilidade de vitória do time mandante. Uma possível oscilação de desempenho dos goleiros pode ser observada via variação de pontos na posição. Essa variável mede a variação que foi obtida na rodada anterior e quanto maior o valor, maior a chance de vitória do time mandante. Isso demonstra um pior desempenho dos goleiros jogando fora e até um possível efeito de regressão à média.

Para os modelos com ligação Logit e Probit, também foi selecionada a média de preços dos jogadores de ataque, com efeito negativo para a probabilidade de vitória do time mandante. Finalmente, para os modelos com ligação complemento log-log e Cauchit foram selecionadas a média de pontos dos zagueiros com efeitos negativos e a quantidade de jogos dos jogadores de meio-campo, com efeito positivo para a probabilidade de vitória do time mandante. Isso confirma o exposto no coeficiente negativo para a quantidade de jogos dos jogadores de meio-campo do time da casa.

Tabela 2: Coeficientes das variáveis selecionadas para cada função de ligação

	Logit	Probit	Cloglog	Cauchit
(Intercept)	0,088	0,052	-0,878	-0,799
media_variacao_pontos_gol_casa	0,668	0,408	0,351	0,602
media_preco_zag_casa	0,314	0,189	0,226	0,415
media_qt_jogos_mei_casa	-0,225	-0,130	-0,380	-0,580
media_qt_jogos_ata_casa	-	-	-0,156	-0,257
media_variacao_pontos_gol_fora	0,923	0,569	0,910	1,256
media_das_medias_pontos_gol_fora	-0,294	-0,185	-0,233	-0,293
media_pontos_zag_fora	-	-	-0,179	-0,354
media_qt_jogos_mei_fora	-	-	0,372	0,519
media_preco_ata_fora	-0,185	-0,113	-	-

Predição

Na Figura 5 estão apresentadas as curvas ROCs de cada um dos modelos para ilustrar a capacidade preditiva de cada um. Em destaque (vermelho) sobre a curva, estão os pontos de corte ideais considerando a sensibilidade e a especificidade, igualmente importantes.

A seguir são demonstradas seis exemplos de predições que foram bem e mal sucedidas para a vitória do time mandante. Foram selecionados 3 jogos para cada caso baseados na probabilidade predita do modelo Logit, quando havia alta probabilidade de vitória e de fato a vitória ocorreu (Tabela 3) ou foi observada uma alta probabilidade de vitória e ela não ocorreu (Tabela 4).

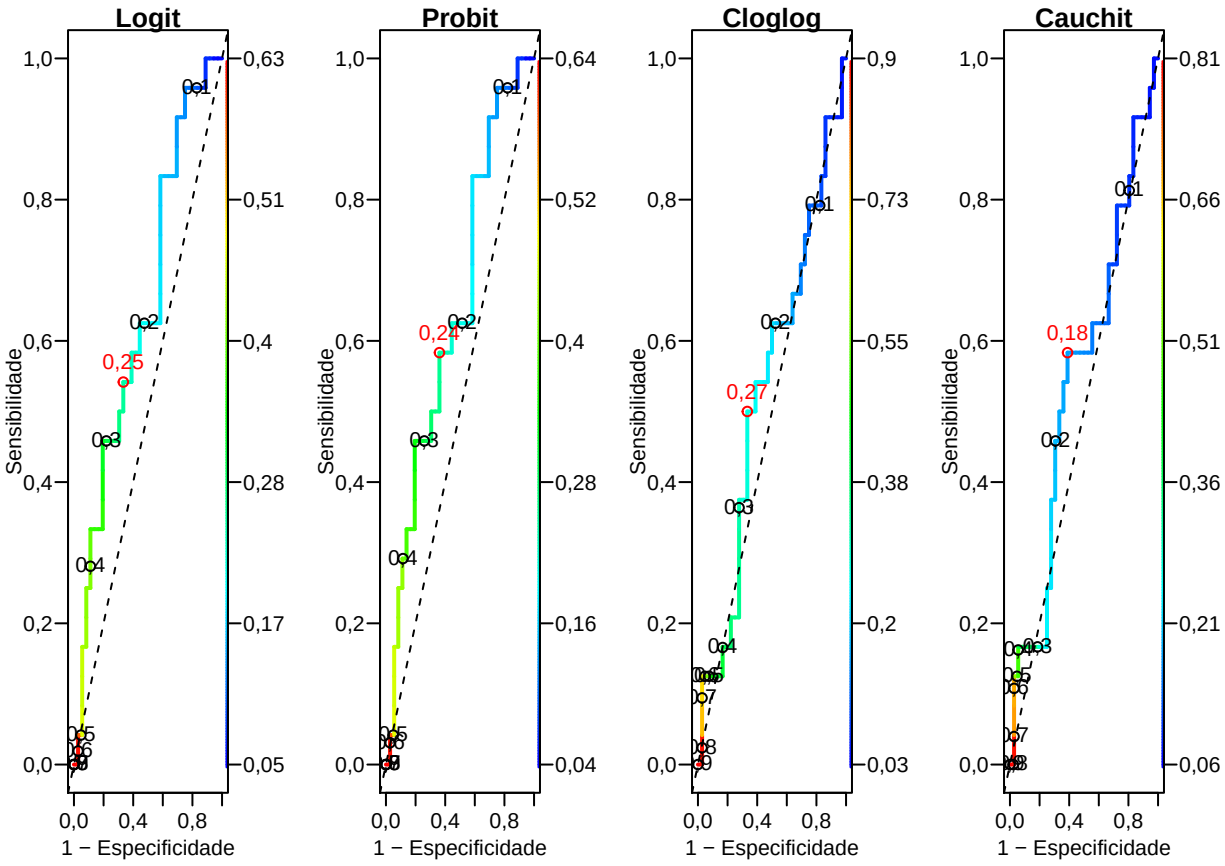

Figura 5: Curva ROC para os diferentes modelos

Tabela 3: Exemplos de predições bem-sucedidas de vitória do time mandante

Rodada	Mandante	Placar	Visitante	Logit	Probit	Cloglog	Cauchit
21	Flamengo - RJ	2 x 0	Atlético - GO	0,576	0,589	0,672	0,542
26	Grêmio - RS	1 x 0	Fluminense - RJ	0,455	0,464	0,710	0,695
22	Palmeiras - SP	4 x 2	São Paulo - SP	0,453	0,463	0,270	0,214

Na Tabela 3, nos casos de predição de vitória bem-sucedida, observa-se casos de vitórias de times que estão bem classificados no campeonato que jogaram em casa contra times que não estão muito bem posicionados no campeonato.

Tabela 4: Exemplos de predições malsucedidas de vitória do time mandante

	Rodada	Mandante	Placar	Visitante	Logit	Probit	Cloglog	Cauchit
10	24	Grêmio - RS	0 x 1	Chapecoense - SC	0,427	0,435	0,419	0,377
11	21	Palmeiras - SP	0 x 2	Chapecoense - SC	0,466	0,475	0,190	0,165
12	22	Corinthians - SP	0 x 1	Atlético - GO	0,621	0,630	0,885	0,798

Já para os exemplos de predição malsucedidas, os exemplos selecionados foram consideráveis "zebras" do campeonato, com os três times melhores classificados perdendo, em casa, para times considerados fracos. Sendo o maior erro, cujos modelos apontaram probabilidade de vitória acima de 60% para o mandante, a inesperada derrota do primeiro colocado, em casa, para o último colocado que ocorreu na rodada 22.

Resumo dos ajustes e predição

Na Tabela 5 é apresentado um resumo de medidas para identificação do melhor modelo para a predição de vitória do time mandante no campeonato brasileiro.

Tabela 5: Resumo das especificações dos modelos de predição

	G.L.	AIC	AUC	Sens.	Espec.
Logit	7	259	0,642	0,542	0,667
Probit	7	259	0,641	0,583	0,639
Cloglog	9	258	0,531	0,500	0,667
Cauchit	9	254	0,546	0,583	0,611

Embora todos os modelos tenham apresentado resultados não muito satisfatórios, é interessante observar que o modelo com função de ligação Cauchit, que foi o mais adequado segundo o AIC, apresentou desempenho inferior em relação ao modelo com função de ligação Logit para predição.

Conclusão

Esse trabalho confirma a dificuldade para realizar previsão no futebol, o que o torna um esporte muito interessante. Embora parte dos problemas de estimação possa ser devido às variáveis explanatórias escolhidas, já que o Cartola pode não representar tão bem o desempenho dos atletas, muitas coisas podem ser feitas para melhorar as estimativas.

Primeiramente, ainda com os modelos propostos, um aumento no número de informações poderia trazer benefícios. Regressões multinomiais que levariam em consideração os três desfechos possíveis numa partida de futebol também poderiam ser mais adequadas, já que agrupar empate e derrota numa mesma categoria faz com que se perca informação valiosa para contabilizar o desempenho de atletas e times.

A regressão de Poisson para modelagem do número de gols é uma abordagem mais comum nas predições de futebol e podem vir a apresentar melhores resultados para predição e verificação da confiança nos dados obtidos pelo Cartola.