Projeto: Previsão de jogos de futebol utilizando Machine Learning

Introdução

A ideia do projeto é, a partir de estatísticas de jogos de futebol, realizar a previsão de partidas futuras a partir de modelos de Machine learning.

O campeonato escolhido para ser utilizado como referência foi a Premier League, e os modelos utilizados foram: Regressão Logística, SVM, Redes Neurais e XGboost.

Base

A base foi conseguida a partir do pacote "worldfootballR" no R, onde foi possível retirar as estatísticas de cada uma das partidas do campeonato.

No pacote há a opção de escolher de qual site retirar as informações dos jogos, e, entre os possíveis, o escolhido foi o FBref. Os dados no site são apresentados como na Figura 1

Aston Villa Player Stats Glossary

Summary | Passing | Pass Types | Defensive Actions | Possession | Miscellaneous Stats

						Performance											
Player	#	Nation	Pos	Age	Min	Gls	Ast	PK	PKatt	Sh	SoT	CrdY	CrdR	Touches	Tkl	Int	Blocks
Wesley Moraes	9	BRA	FW	23-018	68	0	0	0	0	1	0	0	0	27	0	0	1
Jonathan Kodjia	26	CIV	FW	30-053	22	0	0	0	0	0	0	0	0	5	0	0	0
Jack Grealish	10	ENG	LW	24-095	90	0	0	0	1	3	0	0	0	43	0	0	1
Anwar El Ghazi	21	NED	RW	24-225	79	0	0	0	0	1	0	0	0	31	1	0	0
Trézéguet	17	EGY	RW	25-074	11	0	0	0	0	0	0	0	0	2	0	0	0
Henri Lansbury	8	ENG	LM	29-063	65	0	0	0	0	0	0	0	0	31	0	0	1
Douglas Luiz	6	BRA	LM	21-219	25	0	0	0	0	0	0	0	0	16	1	0	0
Marvelous Nakamba	11	ZIM	CM	25-329	90	0	0	0	0	1	0	0	0	45	2	2	0
John McGinn	7	SCO	RM	25-057	90	0	0	0	0	0	0	0	0	53	2	1	1
Matt Targett	18	ENG	LB	24-087	90	0	0	0	0	0	0	1	0	66	0	3	2
Kortney Hause	30	ENG	CB	24-151	90	0	0	0	0	1	0	1	0	79	1	0	3
Björn Engels	22	BEL	CB	25-090	90	0	0	0	0	0	0	0	0	54	2	4	0
Frederic Guilbert	24	FRA	RB	24-355	90	0	0	0	0	0	0	0	0	75	2	3	1
Tom Heaton	1	ENG	GK	33-243	90	0	0	0	0	0	0	0	0	39	0	0	0
14 Players					**990**	**0**	**0**	**0**	**1**	**7**	**0**	**2**	**0**	**566**	**11**	**13**	**10**

Figura 1: Exemplo das estatísticas de uma equipe no site FBref de uma partida

Escolhido o site, foram então importados os dados, que formaram um total 5 bases com informações (variáveis) em sua maioria distintas.

Exemplo: uma trazia variáveis voltadas para finalizações de jogadas, como número de chutes ao gol; outra com estatísticas defensivas, como número de desarmes; outra com a de construção de jogadas, como posse de bola. Um exemplo está na Figura 2.

	Home_Team	Away_Team	Home_Score	Away_Score	Home_Gls	Home_Ast	Home_PK	Home_PKatt	Home_Sh	Home_SoT
1	Arsenal	Leicester City	4	3	4	4	0	0	27	10
2	Watford	Liverpool	3	3	3	1	0	0	9	4
3	Crystal Palace	Huddersfield Town	0	3	0	0	0	0	14	4
4	West Bromwich Albion	Bournemouth	1	0	1	1	0	0	16	6
5	Chelsea	Burnley	2	3	2	2	0	0	19	6
6	Everton	Stoke City	1	0	1	1	0	0	9	4
7	Southampton	Swansea City	0	0	0	0	0	0	29	2
8	Brighton & Hove Albion	Manchester City	0	2	0	0	0	0	6	2
9	Newcastle United	Tottenham Hotspur	0	2	0	0	0	0	6	2
10	Manchester United	West Ham United	4	0	4	4	0	0	21	5
11	Swansea City	Manchester United	0	4	0	0	0	0	5	0
12	Bournemouth	Watford	0	2	0	0	0	0	6	2
13	Southampton	West Ham United	3	2	3	1	2	2	12	3
14	Leicester City	Brighton & Hove Albion	2	0	2	1	0	0	14	4
15	Burnley	West Bromwich Albion	0	1	0	0	0	0	20	0

Figura 2: Exemplo das variáveis na base

A partir disso, foi feita a limpeza e adequação dos dados, pois, apesar das bases não possuirem NAs, haviam dados duplicados e os nomes das variáveis não estavam no padrão desejado.

Um ponto sobre as observações é de que cada uma representava uma partida do campeonato, apresentando estatísticas dos jogos para ambos os times.

Por exemplo: para a variável posse de bola, havia a HOME_POSSESSION, indicando a posse de bola do mandante, e a AWAY_POSSESSION, indicando a posse de bola do visitante.

Após a limpeza das bases, elas foram unidas, formando uma única. E então foi construída a variável alvo "resultado", a partir das variáveis de números de gols, que indicava se a vitória havia sido do clube mandante, do visitante ou se ocorreu um empate.

Ou seja, foi construída uma variável categórica com as classes Mandante, Visitante e Empate.

Um problema encontrado foi de que não era possível realizar as previsões de jogos futuros, já que para obter as estatisticas que alimentariam as bases era necessário que os jogos já tivessem ocorrido.

Para contornar essa questão, foi decidido que cada observação traria as informações das, inicialmente, 3 partidas anteriores através de uma média.

Ou seja, para a variável Finalizações, a observação "100" traria uma média das finalizações das partidas "99","98" e "97".

Essa foi uma metodologia interessante, já que no futebol os clubes vivem diferentes momentos dentro de um campeonato: ora atravessam um bom momento e pontuam mais, ora vivem mau momento e pontuam menos.

Porém, então surgiu um outro problema, que é o de definir qual a melhor janela para realizar a média das partidas anteriores que nos levariam a melhores modelos.

Então, foram construídas bases diferentes, contendo cada uma o cálculo da média de partidas anteriores indo de duas partidas anteriores até 6.

Ainda foram adicionadas na base as variáveis que indicavam as pontuações dos times tanto a cada intervalo de jogos quanto a pontuação total até a partida jogada, além de variáveis que indicavam o valor dos jogadores do clube.

Outro ponto a ser discutido seria o de que quantos campeonatos formariam as bases de treino e teste.

Análise dos Dados

A base então possuía 97 variáveis, sendo uma delas a variável alvo "resultado".

Cada campeonato possui 20 clubes, cada um jogando duas vezes contra os outros 19, em partidas em casa e fora. Portanto, um campeonato possui 380 partidas, dois possui 760 e assim por diante.

Olhando especificamente para as classes da variável "resultado", percebeu-se que havia um número maior de jogos com vitórias de mandantes do que de visitantes e empates, como mostra a Figura 3.

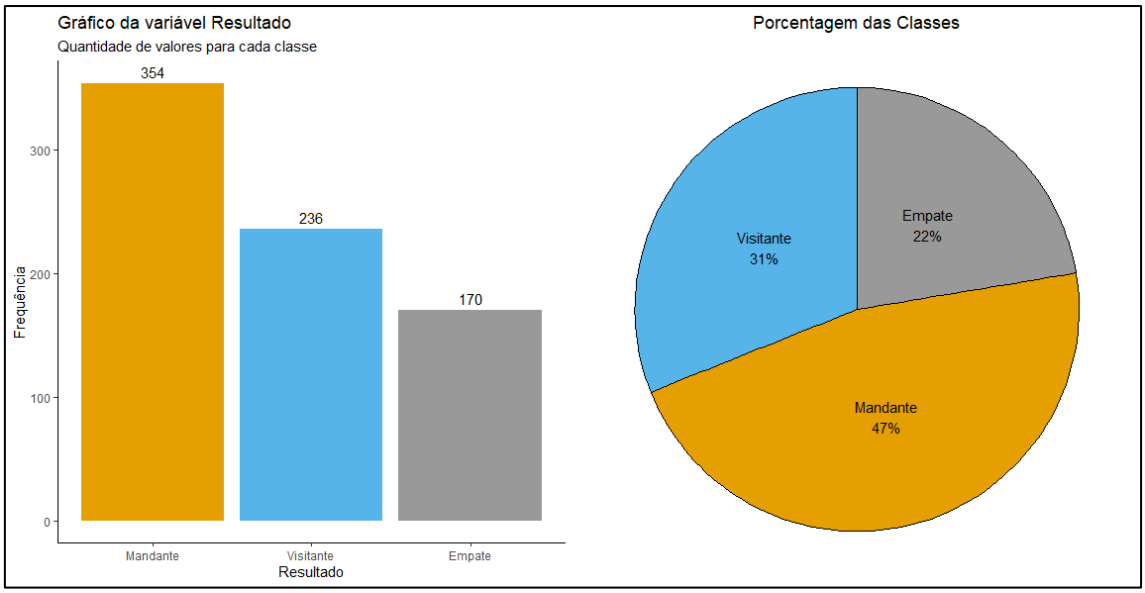

Figura 3: Gráficos para a variável alvo Resultado

E ainda foram então construídos gráficos para tentar identificar algumas possíveis relações das variáveis preditoras com a alvo "resultado".

Percebeu-se que, isoladamente, a maioria das variáveis não apresentava uma relação considerável com a resultado, como mostra a Figura 4.

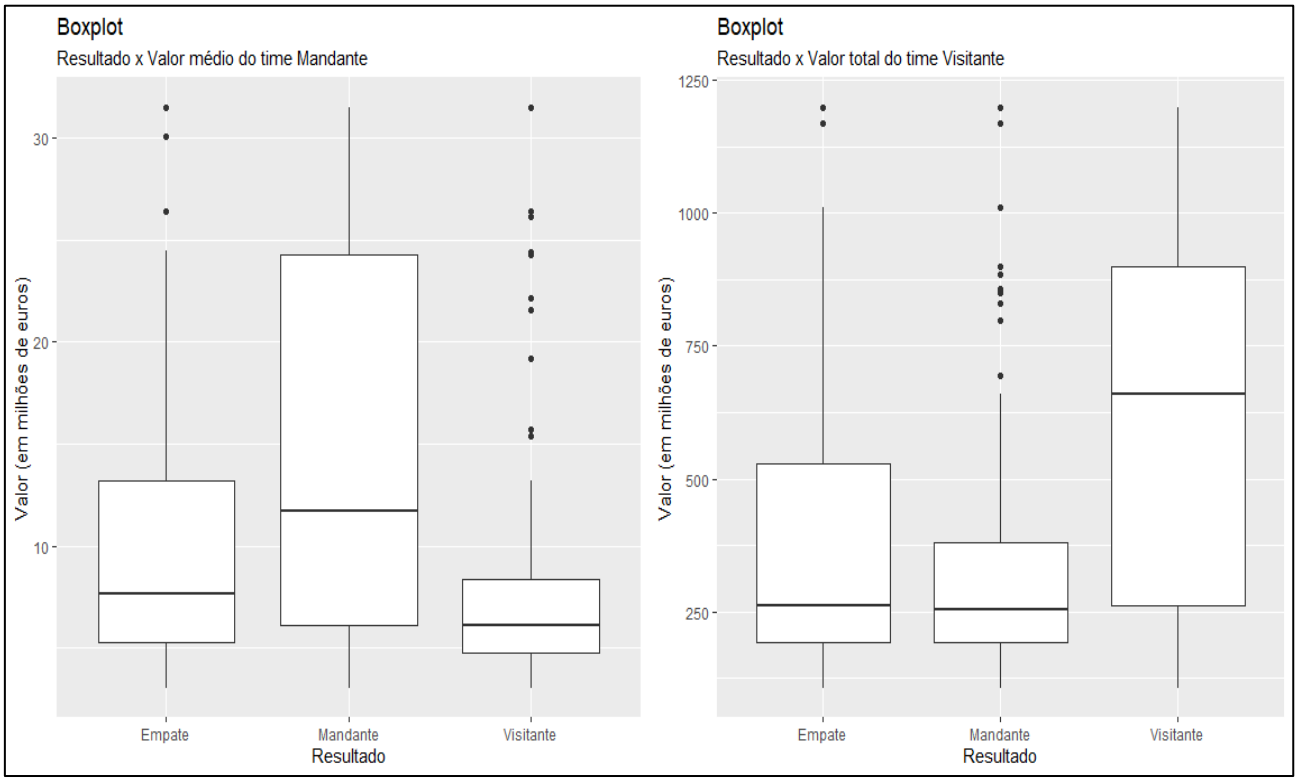

Figura 4: Gráficos de exemplos de variáveis com pouca relação com a Resultado

Porém algumas, como as de valores dos jogadores dos clubes, foi possível verificar um indício de relação, como mostra a Figura 5.

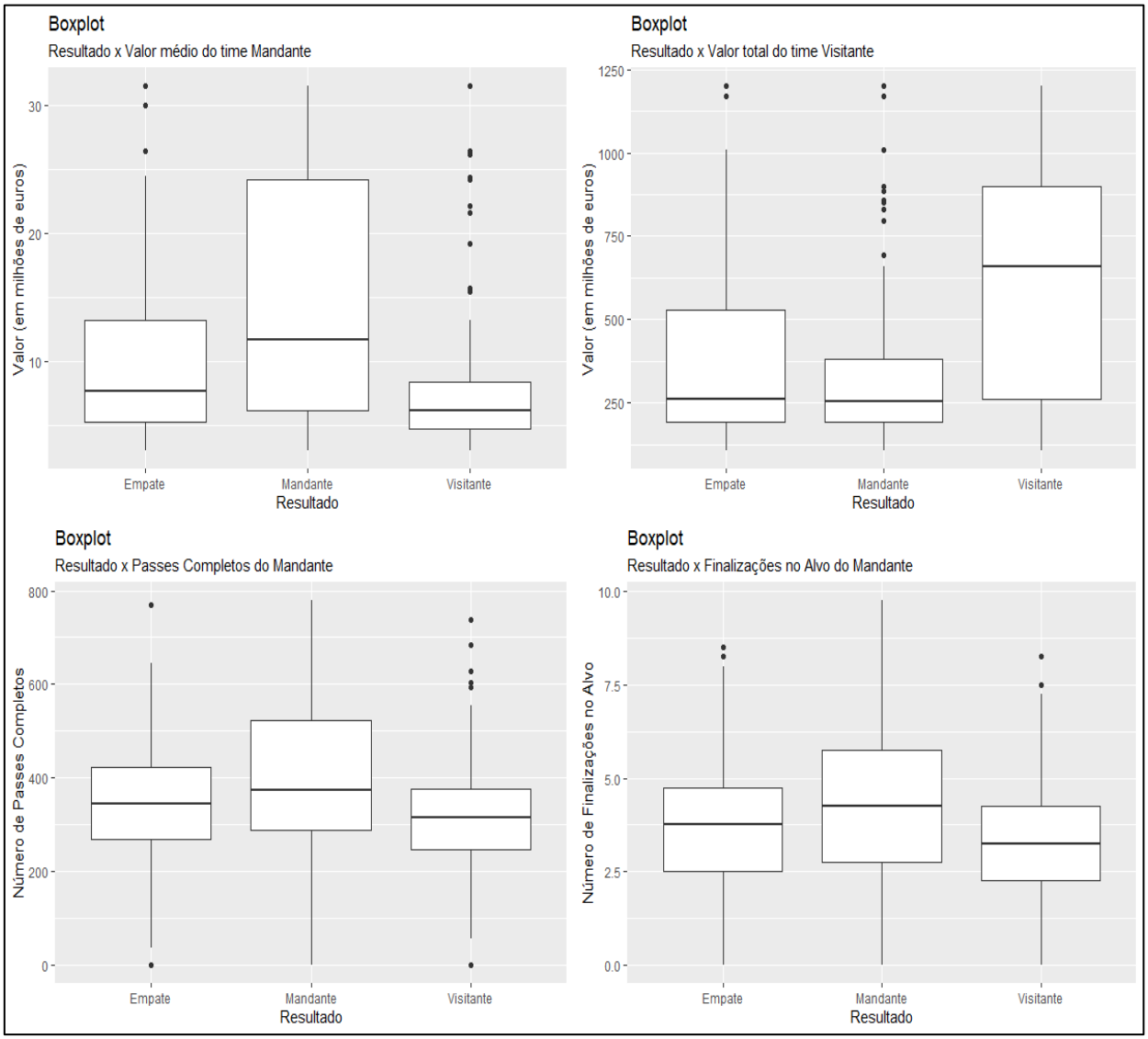

Figura 5: Gráficos com exemplos de variáveis com aparente relação com a Resultado

Pré-processamento

Inicialmente os dados das primeiras rodadas, dependendo da média utilizada, tiveram que ser retirados. Exemplo: se a média utilizada era das 4 partidas anteriores, então os 40 primeiros jogos (observações na base eram retirados).

Como os dados estavam em diferentes escalas, para que os modelos pudessem tiver informações sem privilegiar uma ou outra variável por conta da diferença de escalas entre elas, foi realizada a padronização dos dados.

Foram tentados também outros métodos, porém que geraram resultados ruins.

Um exemplo foi o rebalanceamento dos dados, já que havia uma diferença considerável entre dados de mandantes, visitantes e empates.

Outro teste foi utilizar o PCA, ambos sem sucesso.

Seleção das Variáveis

Foram realizadas diferentes abordagens para tentar selecionar as melhores variáveis para utilizar nos modelos.

Por exemplo, foi verificada a presença de multicolinearidade entre as variáveis explicativas, porém ao retirar as variáveis com grande correlação entre sí, os resultados dos modelos foram ruins. Isso indica que, apesar das variáveis apresentarem informações semelhantes, cada uma traz uma variação diferente que se mostrou importante nos modelos.

A técnica que gerou os melhores resultados foi o RFE, indicando então para a utilização nos modelos um total de 33 variáveis.

Divisão em treino e teste

A estratégia utilizada para treinar e testar os modelos foi a de deixar 1 turno inteiro (19 últimas partidas) como teste, variando apenas a base de treino.

Por conta de um problema nas bases em campeonatos anteriores a 2016 e por conta da pandemia em 2019, os campeonatos tiveram que ser limitados a esse periodo de 2016 até 2019.

E a base que gerou os melhores resultados foi a utilizando os campeonatos de 2016/2017 e 2017/2018, deixando a base de treino com o primeiro campeonato mais o primeiro turno do segundo, e a de teste com o segundo turno do último campeonato.

Resultados

Os modelos foram utilizados a partir do pacote "caret", utilizando técnicas para tentar maximizar os resultados, como a validação cruzada K-Fold e técnicas para encontrar os melhores hiperparâmetros de cada modelo.

O resultados estão na Tabela 1, e o melhor veio do modelo de Redes Neurais, utilizando uma base com média das 4 partidas anteriores, alcançando uma acurácia de 61,59%.

Tabela 1: Acurácia dos modelos

-	Regressão L.	SVM	Redes Neurais	XGBoost
Média2	55,79	57,89	61,05	58,42
Média3	56,84,3	57,89	56,84	58,42
Média4	57,89	58	61,59	60
Média5	57	57,5	57,5	58,5
Média6	57,89	57,89	57,89	58,95

Observação – Classe Empate

Os acertos dos modelos se devem muito aos acertos de vitórias de mandantes, que são superiores as das duas outras classes.

Porém o número de acertos de empates de todos os modelos foi algo em torno de zero, e portanto decidiu-se excluir as partidas que tiveram esse resultado.

Uma nova base então foi montada apenas com dados da classes Mandante e Visitante, e após realizar todo o procedimento e estimar os modelos, o melhor resultado também veio do modelo de Redes Neurais, também utilizando a média das 4 partidas anteriores, com uma acurácia de 72,61%, como mostra a Tabela 2 com os resultados.

Tabela 2: Acurácia dos modelos sem a classe Empate na base

-	Regressão L.	SVM	Redes Neurais	XGBoost
Média2	66,88	68,79	71,97	70,82
Média3	66,24	70,06	68,79	71,34
Média4	66,88	70,06	72,61	70,7
Média5	67,68	69,51	71,34	68,9
Média6	67,52	69,43	71,34	71,34

Conclusão

Os resultados estiveram dentro do que outros trabalhos envolvendo o mesmo tema obtiveram, que é algo em torno de 58 e 62% de acurácia.

Ainda verificou-se que os modelos possuem grande dificuldade de prever vitórias de visitantes e, sobretudo de empates. Então foram retiradas as observações com a classe Empate, o que resultou em uma melhora leve melhora nos acertos de vitórias de Visitantes.

E ainda, pelo menos para esse trabalho, a médias das 4 partidas anterios obteve os melhores resultados.