Desmistificando o Cartola FC: O Baseline

Ciro Ceissler
RA 108786
ciro.ceissler@gmail.com

Lucas de Souza e Silva
RA
lucasonline1@gmail.com

Matheus Laborão Netto
RA
mln.laborao@gmail.com

Ramon Nepomuceno
RA 192771
ramonn76@gmail.com

I. Introdução

Como baseline vamos utilizar a implementação e o tratamento dos dados disponível em [5].

II. Limpeza dos Dados

A primeira etapa realizada em [5] para implmentar o modelo de predição foi analisar os dados do fornecidos pela API do cartola e fazer as limpezas necessárias para criar amostras corretas e relevantes para o preditor. Analisando os dados previamente, alguns problemas foram detectados:

- Jogadores com todos os scouts NANs.
- Jogadores com a coluna 'ClubeID' = NAN.
- Jogadores com a coluna 'Status' = NAN.
- Jogadores com a pontuação não equivalente a soma ponderada dos scouts.

A coluna atletas.clube_id tem campos repetidos e divergentes: por exemplo, todos os Atléticos (MG, PR, e GO) são ATL. Além disso, há jogadores com siglas diferentes das equipes que eles jogam (por exemplo, Maicosuel [id: 37851]). A coluna 'athletes.atletas.scout' não é informativa. Os scouts de 2015 dos jogadores são cumulativos: ou seja, os scouts dos jogadores vão sendo somados a cada rodada. Entretanto, a pontuação não é. Isso também causa o repetimento de dados.

A. Atualização dos scouts cumulativos de 2015

Como dito anteriomente, os dados sobre os scouts de 2015 foram disponibilizados de maneira cumulativa pela , ou seja, os scouts de uma rodada são adicionados aos scouts anteriores a cada nova rodada que um jogador participa. Portanto, foi necessário tirar essa acumulação para cada jogador, de que maneira que a representação dos dados ficasse coerente.

Para isso, dada uma rodada específica, os scouts de um jogador são subtraídos do máximo dos scouts de todas as rodadas anteriores. Repare que assim há chance do scout Jogo Sem Sofrer Gols (SG) ser negativo se o jogador não sofre gols na rodada anterior e sofre na rodada atual. Quando isso acontece, esse scout é atualizado.

B. Verificação da pontuação com os scouts

Por inconsistência na base de dados fornecida pela API do cartola, alguns jogadores possuiam pontuações que não condiziam com seus scouts. Para esses casos, o jogador é removido da base de dados para evitar qualquer tipo de ruído. Ao final, mais de 4000 jogadores foram removidos.

C. Remoção das linhas duplicadas

A última operação realizada para limpar a base de dados foi apagar as linhas repetidas. A existencia de linhas repetidas deve-se ao fato de que a partir da primeira participação de um jogador no campeonato, ele aparece em todas as rodadas subsequentes, mesmo que não tenha jogado. As entradas redundantes não são necessárias ao modelo, por isso foram removidas.

III. Criação das Amostras

Uma vez a base de dados limpa, o próximo passo foi transfor Agora, vamos pegar os dados que limpamos e transformá-los em dados utilizáveis para criação dos modelos. Para isso, foi efetuar as seguintes operações:

- **Selecionar somente as colunas de interesse:** colunas como atletas.nome, atletas.foto, etc não são relevantes para criação do modelo. No entanto, colunas como o AtletaID e atletas.apelido, mesmo que não utilizadas para treinamento do modelo, são importante para avaliar o resultado e, portanto, também serão consideradas.
- **Converter todos os dados categóricos para numéricos:** as colunas Posicao, ClubeID, opponent e casa serão convertidas numéricos.

IV. treinamento do Modelo

O modelo preditor utilizado foi uma Rede Neural Artificial. Em resumo, o modelo recebe como entrada os dados de uma determinada rodada e faz uma predição das pontuações dos jogadores para a próxima rodada.

Para estimar a melhor arquitetura para rede, bem como os hiperparâmetros, foi utilizada a estratégia *GridSearch*. Nesse método, todas as combinações possíveis entre os parâmetros são testados usando uma validação cruzada com 5 folds. A combinação de parâmetros que for melhor na média dos folds, é considerada a melhor.

Optou-se também por normalizar os dados utilizando a estratégia *MinMaxScaler*, que normaliza cada atributo no intervalo [0-1], e utilizar o método de otimização *adam*. Adam é um algoritmo de otimização que pode ser usado em vez do procedimento clássico de descida de gradiente estocástico para atualizar os pesos da rede de forma iterativa com base nos dados de treinamento.

Uma vez que trata-se de um problema de regressão, a função de ativação da saída escolhida para a rede foi a função linear e a rede será treinada visando minimizar a Root Mean Squared Error (RMSE).

Configuração da Rede	fold 1	fold 2	fold 3	fold 4	fold 5
(50,50,50,50)	-19.514	-22.250	-19.641	-19.224	-17.298
(50, 100, 50)	-22.382	-19.712	-19.404	-19.025	-17.154
(50,100,100,50)	-17.114	-19.669	-22.350	-19.490	-19.038
(128)	-19.791	-19.503	-22.334	-19.013	-17.163
(128,128)	-22.282	-19.824	-20.116	-17.211	-19.424
(128, 128, 128)	-19.796	-17.251	-19.702	-19.019	-22.412

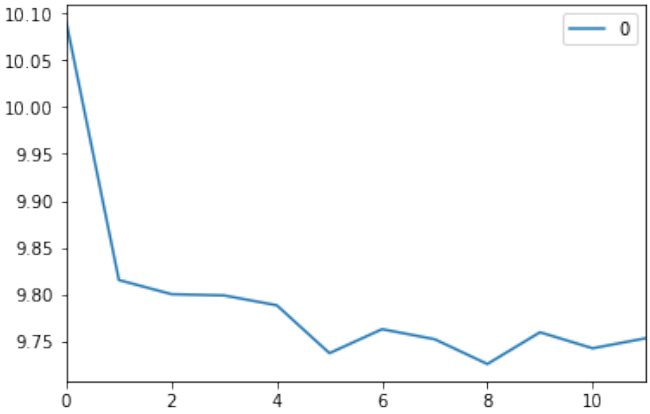

Figure 1. A boat.

V. Propostas para melhorar o Baseline.

A principio, cogitamos utilizar as seguintes estratégias para melhorar o baseline:

1) Utilização do *Random Forest*. Esse modelo consiste em um método de aprendizado conjunto para classificação, regressão e outras tarefas, que operam construindo uma multiplicidade de árvores de decisão no momento do treinamento e gerando a classe que é o modo das classes ou previsão média das árvores individuais. Essa proposta foi discutida em https://github.com/henriquepgomide/caRtola/issues/33, por esse motivo resolvemos testá-la para avaliar os resultados.
2) Acreditamos também que podemos melhorar os resultados modificando a rede para receber como entrada não só apenas os dados de uma terminada rodada, mas sim um histórico das últimas *n* rodadas. Isso deve-se ao fato de que, pela experiencia que temos assistindo futebol, o desempenho dos jogadores oscila de acordo com um determinado histórico. Faremos essa modificação na esperança que o modelo reconheça esse padrão.

References

[1] FIFA, FIFA. "FIFA/Coca-Cola World Ranking 2018."

[2] Martin, William. "Big banks like Goldman Sachs spectacularly failed to predict the World Cup winner — here's why". Business Insider (2018).

[3] GloboEsporte.com. "Com 2460.19 pontos 'Jorgito10 (O mito)' é o vencedor da liga GE AP em 2017". GloboEsporte.com.

[4] .VISCONDI, et al. "Aplicação de aprendizado de máquina para otimização da escalação de time no jogo Cartola FC."

[5] GitHub - Repositório caRtola. https://github.com/henriquepgomide/caRtola/.