UNIVERSIDADE FEDERAL DE CAMPINA GRANDE
CENTRO DE ENGENHARIA ELÉTRICA E INFORMÁTICA
CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

Pedro Antônio Barboza Ribeiro

APLICANDO TÉCNICAS DE APRENDIZADO DE MÁQUINA NA PREVISÃO EM TEMPO REAL DE RESULTADOS DE PARTIDAS DE FUTEBOL: ROCKET E MULTIROCKET

CAMPINA GRANDE - PB

2024

Pedro Antônio Barboza Ribeiro

APLICANDO TÉCNICAS DE APRENDIZADO DE MÁQUINA NA PREVISÃO EM TEMPO REAL DE RESULTADOS DE PARTIDAS DE FUTEBOL: ROCKET E MULTIROCKET

Trabalho de Conclusão Curso apresentado ao Curso Bacharelado em Ciência da Computação do Centro de Engenharia Elétrica e Informática da Universidade Federal de Campina Grande, como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.

Orientador: Leandro Balby Marinho

CAMPINA GRANDE - PB

2024

Pedro Antônio Barboza Ribeiro

APLICANDO TÉCNICAS DE APRENDIZADO DE MÁQUINA NA PREVISÃO EM TEMPO REAL DE RESULTADOS DE PARTIDAS DE FUTEBOL

Trabalho de Conclusão Curso apresentado ao Curso Bacharelado em Ciência da Computação do Centro de Engenharia Elétrica e Informática da Universidade Federal de Campina Grande, como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.

BANCA EXAMINADORA:

Leandro Balby Marinho

Orientador – UASC/CEEI/UFCG

Melina Mongiovi Cunha Lima Sabino

Examinador – UASC/CEEI/UFCG

Francisco Vilar Brasileiro

Professor da Disciplina TCC – UASC/CEEI/UFCG

CAMPINA GRANDE - PB

RESUMO

Dada a grande popularidade do futebol e do mercado de apostas associado, uma melhoria na precisão das previsões tem implicações significativas, tanto do ponto de vista técnico quanto econômico. Tendo isso em mente, o projeto proposto visa explorar, desenvolver e avaliar diversos algoritmos de aprendizado de máquina, incluindo diferentes arquiteturas de redes neurais, para a complexa tarefa de prever resultados de partidas de futebol em tempo real. Ao utilizar uma gama de variáveis estatísticas (como cartões, chutes a gol, faltas, ataques perigosos, escanteios e gols) em uma série temporal que representa o estado do jogo, o projeto contribui para o avanço do campo da análise de dados esportivos e tem o potencial de influenciar o mercado de apostas esportivas. Esse projeto, portanto, não é apenas acadêmicamente relevante, mas também tem um alto valor comercial e social, podendo influenciar a forma como estratégias de apostas são formuladas e talvez até mesmo como o jogo é jogado e analisado.

APPLYING MACHINE LEARNING TECHNIQUES ON REAL TIME FOOTBALL RESULT PREDICTION: ROCKET AND MULTIROCKET

ABSTRACT

Given the dominant popularity of football and its associated betting market, an improvement on football match predictions has important implications, both in the technical and economic sense. tanto do ponto de vista técnico quanto econômico. With that in mind, this project aims to explore and analyse many machine learning algorithms on the context of the complex task of predicting football match results in real time. Two of those methods, never before applied on this task, are ROCKET and MultiRocket. Using an array of statistical variables (such as cards, shots on target, fauls, attacks, corners and goals) in a time series that represents the progress of the game, the project contributes to the advancement of the field of sport data analysis and has the potential of influencing the sports betting market. Thus, this project is not only academically relevant, but also has a high social and commercial value, being able to impact how betting strategies are formulated, or even how the game is played and analysed.

APLICANDO TÉCNICAS DE APRENDIZADO DE MÁQUINA NA PREVISÃO EM TEMPO REAL DE RESULTADOS DE PARTIDAS DE FUTEBOL: ROCKET E MULTIROCKET

Pedro Antônio Barboza Ribeiro
Universidade Federal de Campina Grande
Campina Grande, Paraíba, Brasil
barbozaribeiropedro@gmail.com

Leandro Balby Marinho
Universidade Federal de Campina Grande
Campina Grande, Paraíba, Brasil
lbmarinho@dsc.ufcg.edu.br

RESUMO

Dada a grande popularidade do futebol e do mercado de apostas associado, uma melhoria na precisão das previsões tem implicações significativas, tanto do ponto de vista técnico quanto econômico. Tendo isso em mente, o projeto proposto visa explorar e avaliar diversos algoritmos de aprendizado de máquina para a complexa tarefa de prever resultados de partidas de futebol em tempo real. Dois destes métodos, nunca antes aplicados nesse problema, são os métodos ROCKET e MultiRocket. Ao utilizar uma gama de variáveis estatísticas (como cartões, chutes a gol, faltas, ataques perigosos, escanteios e gols) em uma série temporal que representa o estado do jogo, o projeto contribui para o avanço do campo da análise de dados esportivos e tem o potencial de influenciar o mercado de apostas esportivas. Esse projeto, portanto, não é apenas acadêmicamente relevante, mas também tem um alto valor comercial e social, podendo influenciar a forma como estratégias de apostas são formuladas e talvez até mesmo como o jogo é jogado e analisado.

Keywords

Aprendizado de máquina, futebol, classificação, séries temporais multivariadas, estatística, apostas esportivas, previsão.

1. INTRODUÇÃO

O mercado de apostas esportivas vem apresentando um rápido crescimento nos últimos anos [12], e um dos esportes mais influentes nesse meio é o futebol. Um dos aspectos que pode explicar isso é a sua forte natureza estocástica, ou seja, numa partida, vários fatores, alguns deles imprevisíveis, são fundamentais para determinar o resultado. Apesar da incerteza ser inerente ao esporte, os resultados não são puramente aleatórios, visto que fãs do esporte conseguem formular previsões acerca do rumo das partidas com certa precisão.

Contudo, dificilmente as previsões dos apostadores conseguem consistentemente gerar lucro sobre a casa de apostas [11], visto que existe uma margem de lucro a ser superada no cálculo das *odds*, não bastando apenas ser mais preciso do que um palpite sem embasamento. Por isso, a criação de um método que consiga formular previsões precisas acerca do resultado de partidas é de grande interesse para apostadores, além de potencialmente auxiliar na elaboração de táticas para melhorar o desempenho de um time nas partidas futuras.

Já foram feitos estudos acerca da aplicação de modelos estatísticos e de técnicas de Aprendizado de Máquina no contexto desse esporte, como prever performance de jogadores específicos [17], prever o resultado da partida antes de ela ocorrer [21], entre outros. Para realizar essas previsões, os métodos utilizados em geral se beneficiam de informações sobre o desempenho dos times e dos jogadores em partidas passadas. Estes estudos mostraram que é possível, até certo ponto, fazer previsões dentro do cenário de uma partida de futebol. Para atingir esses resultados, nota-se que é essencial integrar conhecimento do domínio numa Engenharia de Atributos minuciosa e bem planejada [2]. Contudo, o caráter estocástico do esporte faz com que métodos para previsão do resultado antes da partida sejam limitados por não considerarem informações acerca de eventos imprevisíveis que ocorreram durante a partida, como por exemplo uma contusão de um jogador importante ou uma expulsão injusta.

A literatura existente ainda é escassa no que tange a estudo de predições realizadas durante a partida, sendo a grande maioria dos estudos focada em fazer predições somente antes do início da partida. Os poucos estudos sobre previsão em tempo real utilizaram abordagens mais tradicionais, como Regressão Logística [22], Árvores de Decisão [3], distribuições de Bernoulli [29] e modelos Bayesianos [18]. Fazer previsões durante a partida traz inúmeros benefícios, visto que, como mencionado, uma partida pode ser afetada por inúmeros fatores, muitos deles difíceis de prever antes de seu início, e fazer previsões após o início da partida pode ajudar a mitigar o impacto desses fatores. Além disso, eventos pontuais mudam drasticamente o rumo da partida, como uma contusão, uma expulsão ou um gol, e se adaptar à ocorrência desses eventos, principalmente quando eles não foram previstos pelo modelo, pode resultar numa melhora considerável da precisão das previsões.

Para treinar os modelos, foram utilizadas estatísticas minuto a minuto de partidas passadas, para simular que o modelo está acompanhando uma partida enquanto ela ocorre. Dessa forma, tem-se duas representações de uma partida: o estado atual e a série temporal de eventos. A primeira destas consiste na soma acumulada das estatísticas selecionadas até o momento atual da partida, enquanto a segunda também destaca a posição de cada evento no tempo. Por exemplo, enquanto para um determinado

minuto o estado da partida aponta que o time da casa está com 5 finalizações, a série temporal pode indicar que destas 5, 4 foram durante os 10 primeiros minutos, possivelmente proporcionando informações relevantes para auxiliar os modelos a formularem previsões mais precisas.

Diferentemente dos estudos feitos acerca da previsão do resultado antes da partida, este trabalho aborda o desafio de identificar a relevância da ordem dos eventos ao longo de uma partida na previsão do resultado. Portanto, esse problema pode ser definido como um problema de classificação de séries temporais multivariadas, por estar classificando em uma de três categorias (vitória do time da casa, empate, vitória do time visitante) uma série temporal minuto a minuto de um conjunto de variáveis que representam o estado da partida naquele minuto. Exemplos dessas variáveis seriam: gols feitos, chutes a gol, número de cartões, entre outras. De forma geral, a classificação de séries temporais multivariadas é um campo notadamente menos explorado do que a de séries univariadas [16]. Por se tratar de um esporte que tem uma duração aproximada predefinida, esse problema trata de séries temporais de tamanho previsível, destacando este problema de muitos dos mais explorados na literatura, como a previsão do valor de ações na bolsa de valores.

Dentre as abordagens aplicadas a esse problema, o melhor desempenho até o momento foi atingido pelo modelo CNN-BiLSTM. Este modelo superou o desempenho de modelos mais tradicionais, como uma Regressão Logística, quando aplicado entre o minuto 70 e 85 da partida, intervalo no qual modelos baseados na série temporal da partida conseguiram se beneficiar de informações sobre a ordem de eventos no decorrer do jogo [5]. Estes experimentos servirão como referência de desempenho para as abordagens aqui apresentadas, que serão aplicadas sobre a mesma base de dados dos experimentos mencionados.

Com isso em mente, esse estudo visa dar um passo a mais na aplicação de técnicas de Aprendizado de Máquina em previsões feitas durante a partida por meio do uso de modelos nunca antes aplicados nessa tarefa e que compõem o estado da arte na previsão de séries temporais multivariadas. Foram selecionados os métodos ROCKET [6] e MultiRocket [26] para a realização dos novos experimentos. Logo, este trabalho tem como objetivo abordar a previsão de resultados de partidas de futebol em tempo real, até então um problema pouco explorado, por meio da aplicação de mpetodos nunca antes utilizados para esta tarefa. Comparando várias abordagens, das mais tradicionais ao estado da arte na classificação de séries temporais multivariadas, e avaliando o desempenho de cada abordagem ao decorrer do jogo.

2. METODOLOGIA

2.1 Conjunto de dados

O conjunto de dados utilizado neste trabalho foi construído em um trabalho prévio [5]. Este contém 9.416 partidas jogadas entre 01 de julho de 2017 e 31 de fevereiro de 2020, visto que a dinâmica dos jogos após este período mudou consideravelmente devido à pandemia do COVID-19 e à ausência de torcida nos estados [13], e portanto a inclusão destes jogos poderiam prejudicar o aprendizado dos modelos.

Para padronizar a duração das partidas, os jogos que duraram menos de 95 minutos foram estendidos até 95 minutos por meio da replicação dos valores contidos no último minuto real. Com o mesmo intuito de padronizar o formato da entrada, as partidas com mais de 95 minutos tiveram seus valores posteriores aos 95 minutos removidos. Dessa forma se abstraem alguns detalhes da partida, como em que minuto da sequência foi feito o intervalo entre o primeiro e o segundo tempo do jogo.

Ele é totalmente contido em um arquivo CSV de 65.13 MB, no qual cada linha da tabela representa a série temporal completa referente a um jogo.

Os eventos observados são representados como sequências de somas acumuladas das ocorrências de cada evento minuto a minuto, sendo uma sequência referente ao time mandante (S_m) e outra para o time visitante (S_v). Para transformar os eventos observados ao longo da partida em uma série temporal foi extraída a diferença entre S_m e S_v no que chamaremos de $D = S_m - S_v$. Contudo, apenas a diferença entre os eventos não representa de forma adequada o estado da partida em um determinado momento. Por exemplo, uma partida que em um determinado minuto apresenta um placar de 1 a 0 para o time mandante e outra que apresenta um placar de 3 a 2 apresentam o mesmo $D = 1$, apesar de representarem partidas muito diferentes. Para isso, também foi construída outra série temporal contendo as razões R definidas por:

Se $S_m = S_v$, então $R = 0.5$, caso contrário, $R = \frac{S_m}{S_m + S_v}$.

Os dados são estruturados de acordo com a Tabela 1.

Tabela 1: Símbolos e Notações

Notação	Description	Domínio
M	Par de times(tm: time mandante , tv: time visitante)	$M = (tm, tv)$
G	Par de gols marcados(gm:gols do mandante, gv: gols do visitante)	$G = (gm, gv) \in \aleph^2$
r	Resultado da partida	$r \in R \mid R = \{tm\ vence, empate, tv\ vence\}$
d	Duração da partida em minutos	$d \in \aleph$
S	Série temporal minuto a minuto dos seguintes eventos para cada time: gols, cartões amarelos, cartões vermelhos, chutes a gol, chutes para fora, ataques, ataques perigosos e escanteios	*S* é um conjunto de tuplas *s = (a, e, m, t)*, na qual *a* é a frequência cumulativa de um evento *e* para o time *t* em um dado minuto *m*.
B	Série temporal minuto a minuto de posse de bola de cada time	*B* é um conjunto de tuplas *b = (q, m, t),* nas quais $q \in [0, 1]$ é a razão de posse de bola para um time $t \in M$ até um dado minuto *m*.
O	Série temporal minuto a minuto de odds no mercado de apostas	*O* é um conjunto de tuplas *o = (v, i, r), nas quais* $v \in \Re_{\geq 1}^{*}$ é o valor da odd publicado em um dado minuto *i* para o resultado $r \in R$.
D	Série temporal minuto a minuto comparando os elementos $S_m \in S$ (time mandante) e $S_v \in S$ (time visitante)	$D = S_m - S_v$.
R	Série temporal minuto a minuto comparando os elementos *S* do time mandante com os do time visitante	Se $S_m = S_v$, então *R = 0.5*, caso contrário, $R = \frac{S_m}{S_m + S_v}$.

Fonte: Costa, 2022

2.2 Estratégias de treinamento

Foram replicados experimentos realizados previamente [5] com o intuito de comparar o desempenho desses experimentos com aqueles propostos neste trabalho. Os experimentos podem ser usar modelos baseados no estado da partida ou no progresso do jogo.

Os modelos baseados no estado do jogo recebem como entrada os valores de *D*, *R*, *B* e *O* para um determinado minuto. Com isso é possível treinar um modelo único que recebe o minuto do jogo a ser analisado juntamente com as métricas correspondentes àquele minuto. Essa estratégia será chamada de Modelo Único (MU). Alternativamente, também é viável realizar o treino de um modelo diferente para cada minuto do jogo, visto que estamos trabalhando

com uma série temporal de um tamanho fixo de 95. Sendo assim, como cada modelo já teria sido treinado para avaliar a partida durante um minuto específico, incluir a minutagem na entrada do modelo se torna desnecessário. Essa segunda abordagem será chamada de Múltiplos Modelos (MM). Como exemplos de modelos baseados no estado do jogo, foram reproduzidos experimentos realizados com *Gaussian Naive Bayes* (GNB) [25] [27], Regressão Logística (RLG) [21] [20] e *Gradient Boosting* (GBO) [10] [1].

Ao invés de usarmos o estado do jogo como entrada dos nossos modelos, também é possível utilizarmos da série temporal relativa ao progresso do jogo até um determinado minuto. Porém, isso gera uma inconsistência no tamanho da entrada dos modelos, visto que para cada minuto do jogo teríamos uma série temporal de um tamanho diferente. Para contornar esse problema, uma possível solução seria treinar um modelo para cada minuto do jogo, sendo assim cada modelo teria um tamanho de entrada correspondente ao minuto que deve ser analisado. Essa abordagem será referenciada como Modelos Temporais Múltiplos (MTM). Como exemplos de modelos baseados no progresso do jogo, foram reproduzidos experimentos realizados Fully Convolutional Network (FCN) [28], InceptionTime [9] e CNN-BiLSTM [5].

Diferentemente dos experimentos já feitos, em todos os experimentos aqui citados foram utilizadas informações acerca das *odds* do mercado de apostas minuto a minuto. Isto foi feito com o fim de atingir o melhor desempenho possível para cada abordagem. Essas informações podem representar aspectos do jogo que não são captados pelas demais variáveis, como uma contusão inesperada, uma substituição equivocada ou uma expulsão de um jogador.

Além dos experimentos replicados, este trabalho inclui também experimentos utilizando ROCKET [6] e MultiRocket [26], modelos que compõem o estado da arte na classificação de séries temporais [23] [16]. O ROCKET é um método de classificação de séries temporais que se utiliza de kernels convolucionais com tamanho, pesos, vieses, dilatação e padding gerados de forma aleatória. Um maior número de kernels tende a aumentar a precisão das previsões e aumentar o tempo de treino, sendo o número recomendado 10000 [6] e o número utilizado neste trabalho 25000, a fim de adquirir uma precisão maior e mais constante, visto que como o ROCKET é um método estocástico, a precisão dele tente a se alterar ao se repetir o mesmo experimento. Esses kernels então são utilizados para realizar uma convolução da série temporal recebida como entrada. Para cada mapa de *features* resultantes da convolução para um determinado kernel são extraídas duas *features*: o valor máximo e a proporção de valores positivos. As *features* extraídas após a convolução são utilizadas para treinar um classificador linear. Como nosso dataset não é muito grande, utilizaremos um classificador utilizando regressão Ridge [15] e validação cruzada. Como os pesos dos kernels não são treinados e sim gerados aleatoriamente, o ROCKET tem um tempo de treino muito rápido quando comparado com outros métodos. A alta precisão atingida pelo ROCKET vem do uso de múltiplos valores de dilatação, possibilitando a identificação de padrões de diferentes frequências e escalas, e do uso da proporção de valores positivos para sumarizar a saída dos mapas de *features*, possibilitando ao classificador pesar a prevalência de um determinado padrão dentro de uma série temporal [6].

O MultiRocket é a variação do ROCKET que exibe o melhor desempenho na maioria das bases de dados já testadas. Diferentemente do ROCKET, o MultiRocket se utiliza de um conjunto determinado de kernels convolucionais, seguindo o mesmo padrão definido no MiniRocket [7], sendo o único aspecto definido aleatoriamente destes kernels o cálculo dos viéses. São extraídas 4 *features* para cada kernel: proporção de valores positivos, média dos valores positivos, média dos indices dos valores positivos e maior sequência de valores positivos [26]. As *features* são então normalizadas e utilizadas para se treinar um classificador linear da mesma forma que ocorre no ROCKET.

Ambos os métodos foram aplicados segundo a estratégia MTM, sendo treinado um modelo de regressão para cada minuto do jogo, totalizando 190 novos modelos treinados, 95 para cada método aqui apresentado, além dos experimentos replicados. A implementação foi feita com o uso das bibliotecas sktime [14], para a geração dos kernels de acordo com a definição do ROCKET e MultiRocket, e Scikit-learn [19], para a implementação do classificador linear. Apesar de ambos os métodos aqui propostos serem estocásticos, podendo apresentar um resultado diferente a cada iteração do experimento, foram treinados e testados somente uma vez para evitar enviesar os resultados.

2.3 Treino e avaliação

Os dados foram divididos entre um conjunto de teste de 7.532 partidas, ou 80% dos dados, e um conjunto de teste de 1.884 jogos, ou 20% dos dados. Essa divisão respeitou a ordem cronológica das partidas, a fim de simular que o modelo estaria prevendo partidas futuras a partir de partidas que já ocorreram. Estes dados foram então utilizados para treinar e testar os modelos. O desempenho nos testes foram analisados a partir da métrica Ranked Probability Score (RPS) [8]. O RPS é adotado como padrão para avaliação de modelos de predição de resultados de partidas de futebol [4] e é definida por:

$$RPS = \frac{1}{1-r} \sum_{i=1}^{r} \left(\sum_{j=1}^{r} p_j - \sum_{j=1}^{r} t_j \right)^2 .$$

Onde r é a quantidade de classes possíveis, que no caso do nosso problema são 3 (de forma que o 0 representa vitória do time mandante, 1 o empate e 2 a vitória do time visitante). p_j é a probabilidade prevista para a classe j e t_j é a probabilidade real para a classe j. Por exemplo, no caso de uma partida na qual o time mandante venceu, $t_0 = 1,\ t_1 = 0\ e\ t_2 = 0$. O objetivo do treino é diminuir o valor da RPS aproximando os valores previstos dos valores reais para cada classe.

Essa métrica é adequada para a previsão de resultados de jogos de futebol, pois ela torna possível considerar que um empate é um resultado "intermediário" entre a vitória do time mandante e a vitória do time visitante. Por exemplo, se em uma determinada partida a vitória foi do time mandante, portanto $t_0 = 1,\ t_1 = 0\ e\ t_2 = 0$, e as previsões foram de $p_0 = 0.7,\ p_1 = 0.2\ e\ p_2 = 0.1$ o valor do RPS nesse caso é de 0.50. Porém, caso os valores previstos fossem $p_0 = 0.7,\ p_1 = 0.1\ e\ p_2 = 0.2$, o valor do RPS seria de 0.65. Para nós essa diferença é útil pois no caso de a classe prevista com maior probabilidade ser a de vitória do time mandante, ou

seja, a classe 0, é incoerente que a classe com a segunda maior probabilidade seja a de vitória do time visitante e não o empate.

3. RESULTADOS E DISCUSSÕES

A Tabela 2 ilustra o desempenho obtido por cada modelo utilizado de acordo com a métrica escolhida (RPS) e com o minuto do jogo no qual foi feita a previsão (em intervalos de 15 minutos). A Figura 1 e Figura 2 comparam os desempenhos da Regressão Logística, a abordagem baseada no estado do jogo com melhor desempenho, CNN-BiLSTM, o modelo baseado no progresso do jogo com melhor desempenho, e os métodos introduzidos neste trabalho, ROCKET e MultiRocket. Cada abordagem tem um RPS diferente para cada minuto do jogo, sendo a tendência esperada que ele caia quanto maior a minutagem do jogo, visto que os dados na entrada do modelo conteriam mais informações a respeito do curso daquela partida.

Pode-se observar que os métodos aqui propostos superaram todas as abordagens já estudadas a partir de algum minuto 76, no caso do ROCKET e 78 no caso do MultiRocket. Também é esperado que o RPS de cada modelo se aproxime rapidamente de 0 nos minutos finais do jogo, visto que o placar estabelecido neste momento dificilmente se alteraria nos minutos finais da partida, sendo portanto uma tarefa trivial, na grande maioria dos jogos, prever o vencedor a partir do placar corrente. Nesse sentido, o desempenho dos métodos aqui apresentados foi ainda mais destoante daqueles adquiridos nos experimentos replicados, formulando previsões muito mais assertivas nos minutos finais. Como exemplo disso, o ROCKET atingiu um RPS no minuto 90 quase 14 vezes melhor do que o GBO com MU, que foi o modelo mais preciso nessa minutagem dentre aqueles utilizados nos experimentos replicados.

Essa melhora no desempenho é um indicativo de que, nestes intervalos, mesmo sendo gerados aleatoriamente, alguns dos kernels gerados por estes métodos apresentam um conjunto de pesos mais adequados do que aqueles aprendidos por outras abordagens aqui tratadas que se utilizam de convolução, como o FCN. Isso pode ser uma consequência do tamanho limitado da base de dados utilizada. Caso a base de dados contasse com mais jogos, possivelmente modelos como o FCN e CNN-BiLSTM conseguiriam treinar pesos mais adequados que os gerados pelo ROCKET durante este intervalo do jogo.

Ao compararmos as novas abordagens aqui propostas, se observa que o ROCKET obteve um desempenho consistentemente melhor que aquele obtido pelo MultiRocket. Isso evidencia que os filtros gerados aleatoriamente pelo ROCKET são mais adequados para esta tarefa específica do que àqueles construídos segundo o padrão do MultiRocket.

Além disso, alguma inconsistência é esperada da aplicação do ROCKET e MultiRocket, porém observa-se claramente que estes performaram de forma ainda mais inconstante nos primeiros minutos do jogo, principalmente àquelas abaixo do minuto 20, demonstrando uma grande dificuldade de estabelecer um aprendizado. Isso é reforçado ao observarmos o desempenho muito abaixo desses métodos em relação ao de todas as outras abordagens durante esse intervalo. À medida que o jogo avança, os modelos se estabilizam cada vez mais e conseguem atingir uma aprendizagem mais consistente, mostrando uma tendência clara de melhora no RPS conforme o jogo progride. Tendo em vista que estes métodos nunca foram utilizados para séries temporais muito curtas, isso possivelmente se deve à dificuldade dos filtros gerados, que têm tamanho igual a 7, 9 ou 11, de extrair *features* relevantes de uma série temporal muito curta, resultando numa dificuldade de se obter algum aprendizado com as séries temporais de curta duração.

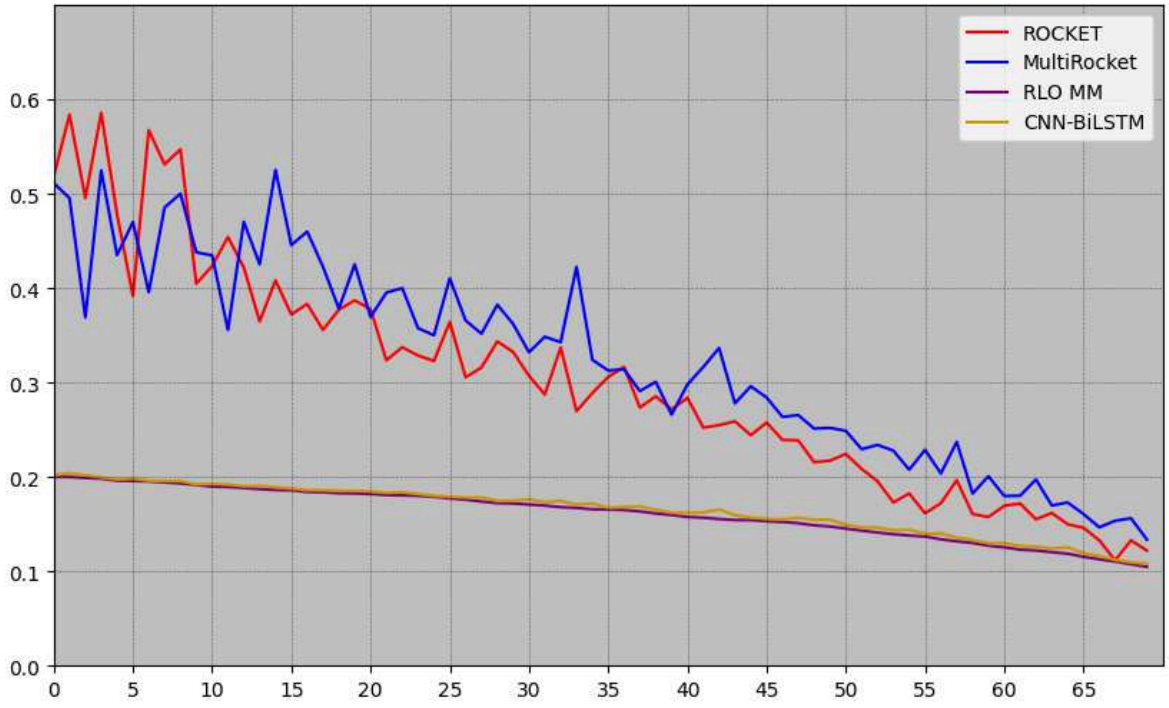
Por fim, se reforçou a hipótese já proposta [5] de que a informação acerca da ordem dos eventos por meio da formulação da representação de uma partida como uma série temporal somente é relevante para melhorar o desempenho dos métodos de previsão a partir do minuto 70. Somente um pouco após esse ponto do jogo, nos minuto 76, no caso do ROCKET e 80 no caso do MultiRocket, o desempenho destes modelos conseguem superar aquele obtido ao utilizar de RLO com MM, sendo esta uma estratégia baseada apenas no estado da partida, e não na sequência de eventos.

Tabela 2: RPS por minutagem do jogo

Modelo Utilizado	**Estratégia**	**RPS**					
		15'	**30'**	**45'**	**60'**	**75'**	**90'**
GNB	MU	0.2313	0.2154	0.1952	0.1536	0.1086	0.0283
RLG	MU	0.1861	0.1721	0.1545	0.1272	0.0924	0.0315
GBO	MU	0.1863	0.1708	0.1532	0.1260	0.0901	0.0251
GNB	MM	0.2372	0.2152	0.1926	0.1522	0.1090	0.0278
RLG	MM	0.1860	0.1709	0.1533	0.1255	0.0898	0.0252
GBO	MM	0.1875	0.1732	0.1547	0.1275	0.0913	0.0253
FCN	MTM	0.1929	0.1780	0.1633	0.1313	0.0953	0.0262
InceptionTime	MTM	0.1927	0.1781	0.1631	0.1312	0.0954	0.0259
CNN-BiLSTM	MTM	0.1894	0.1751	0.1628	0.1309	0.0932	0.0261
Rocket	**MTM**	**0.3723**	**0.3073**	**0.2577**	**0.1696**	**0.1016**	**0.0018**
MultiRocket	**MTM**	**0.4459**	**0.3323**	**0.2842**	**0.1799**	**0.1162**	**0.0061**

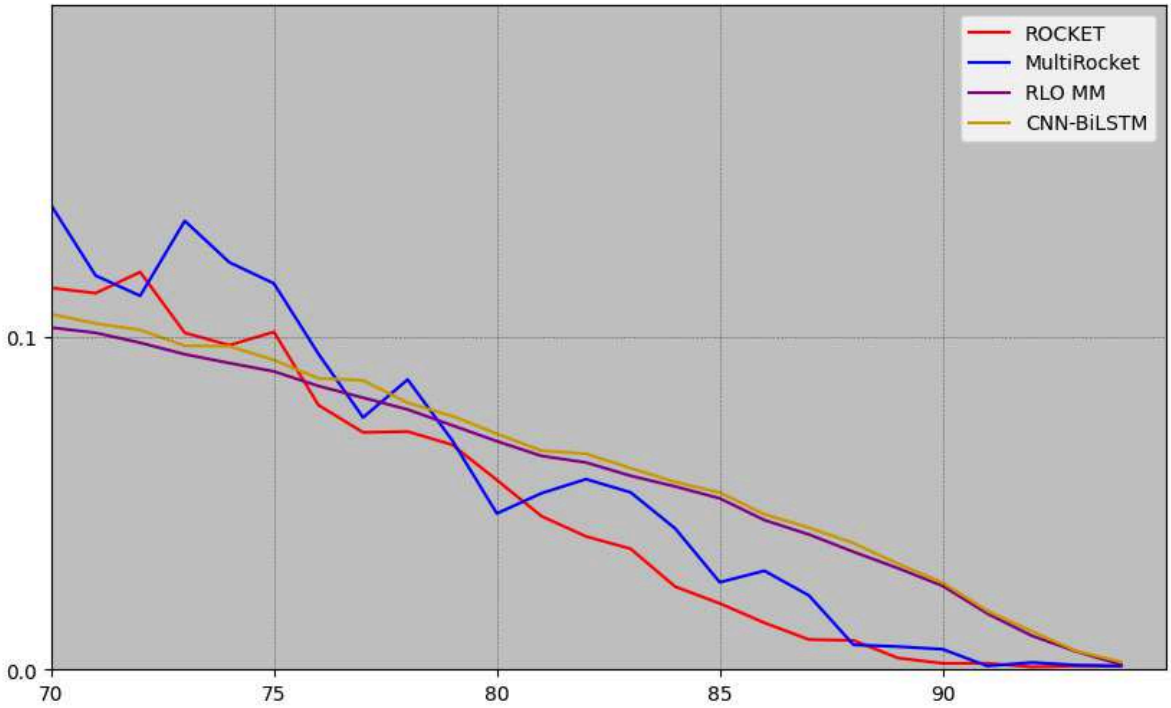
Fonte: Autor

Imagem 1: RPS do minuto 0 ao 70



Fonte: Autor

Imagem 2: RPS do minuto 70 a 95



Fonte: Autor

4. CONCLUSÃO

Esse trabalho estabelece uma nova referência do desempenho de técnicas pertencentes ao estado da arte na previsão de séries temporais multivariadas na difícil tarefa de prever resultados de jogos de futebol em tempo real. Esta nova referência pode ser utilizada por apostadores, pela mídia esportiva ou por fãs de futebol para melhor analisar e entender o funcionamento do esporte na sua atual configuração. Infelizmente, não foi possível exibir uma melhora na precisão das previsões realizadas durante a primeira metade da partida, sendo este o intervalo mais interessante para apostadores.

Mesmo nossas melhores métricas dificilmente superam de forma substancial as previsões das casas de aposta para as partidas analisadas, refletidas por meio das *odds* minuto a minuto da partida. Para atingir esse objetivo, trabalhos futuros poderiam tentar incluir informações sobre a partida que não foram utilizadas neste trabalho. Nesse sentido, poderiam ser incluídas informações sobre os times e jogadores envolvidos na partida. Isso poderia ser útil para formular previsões mais precisas ao, por exemplo, considerar uma contusão de um jogador importante ou o peso da rivalidade entre dois times.

Tendo isso em vista o RPS baixo e inconstante de ambos o ROCKET e MultiRocket durante os primeiros minutos do jogo, um trabalho futuro poderia explorar adaptar a arquitetura destes métodos para essa tarefa específica, com o intuito de adquirir resultados melhores e mais consistentes durante esse período da partida.

Ainda há muito a ser explorado neste tema, sendo possível aplicar outros modelos modernos já utilizados na classificação de séries temporais multivariadas e que não foram abordadas neste trabalho, sendo os mais notáveis os modelos TapNet [31], WEASEL+MUSE [24] e TST [30]. Também seria útil acrescentar na base de dados jogos após o retorno da torcida aos estádios em 2022, gerando mais dados para potencializar o aprendizado de quaisquer modelos empregados nessa tarefa.

5. AGRADECIMENTOS

Agradecimentos ao professor Igor Barbosa da Costa por ter disponibilizado a base de dados utilizada, até então não publicada, pela autoria do trabalho que serviu de base para este projeto e pela orientação durante a produção deste trabalho.

6. REFERÊNCIAS

[1] BABOOTA, R.; KAUR, H. . 2018. Predictive analysis and modeling football results using machine learning approach for English Premier League. International Journal of Forecasting. 35.

[2] BERRAR, D.; LOPES, P; DUBITZKY, W. 2018. Incorporating domain knowledge in machine learning for soccer outcome prediction. Machine Learning.

[3] Betsonly. Most popular football betting markets. Disponível em: http://www.betsonly.com/sport-betting/popular-football-betting-markets. Acesso em 24 de Abril de 2024.

[4] CONSTANTINUOU, A.; FENTON, N. . 2012. Solving the Problem of Inadequate Scoring Rules for Assessing Probabilistic Football Forecast Models. Journal of Quantitative Analysis in Sports. 8.

[5] COSTA, I. 2021. Modelagem e Predição de Resultados de Futebol Antes e Durante as Partidas Usando Aprendizagem de Máquina. Disponível em: http://dspace.sti.ufcg.edu.br:8080/jspui/bitstream/riufcg/20618/3/ÍGOR%20BARBOSA%20DA%20COSTA%20-%20TESE%20%28PPGCC%29%202021.pdf. Acesso em 20 de Outubro de 2023.

[6] DEMPSTER, A.; PETITJEAN, F.; WEBB, G. . 2020. ROCKET: exceptionally fast and accurate time series classification using random convolutional kernels. Data Min Knowl Disc 34, 1454–1495 (2020). Disponível em: https://doi.org/10.1007/s10618-020-00701-z. Acesso em 19 de Abril de 2024.

[7] DEMPSTER, A.; SCHIMDT, D.; WEBB, G. . 2021. MiniRocket: A Very Fast (Almost) Deterministic Transform for Time Series Classification. In Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining (KDD '21). Association for Computing Machinery, New York, NY, USA, 248–257. Disponível em: https://doi.org/10.1145/3447548.3467231. Acesso em 7 de Maio de 2024.

[8] EPSTEIN, E.; 1969. A Scoring System for Probability Forecasts of Ranked Categories. J. Appl. Meteor. Climatol, 8, 985–987. Disponível em: https://doi.org/10.1175/1520-0450(1969)008<0985:ASSFPF>2.0.CO;2. Acesso em 26 de Abril de 2024.

[9] FAWAS, H.; LUCAS, B.; FORESTIER, G. *et al*. . 2020 InceptionTime: Finding AlexNet for time series classification. Data Min Knowl Disc 34, 1936–1962. Disponível em: https://doi.org/10.1007/s10618-020-00710-y. Acesso em 26 de Abril de 2024.

[10] HUBÀCEK, O.; SOUREK, G.; ZELEZN, F. . Learning to predict soccer results from relational data with gradient boosted trees. Machine Learning, pages 1–19, 2018.

[11] KHAZAAL, Y.; CHATTON, A.; BILLIEUX, J. et al. 2012 Effects of expertise on football betting. Subst Abuse Treat Prev Policy 7, 18. Disponível em: https://doi.org/10.1186/1747-597X-7-18. Acesso em 5 de Maio de 2024

[12] LANG, S.; WILD; R.; ISENKO, A. *et al*. . 2022. Predicting the in-game status in soccer with machine learning using spatiotemporal player tracking data. Disponível em: https://www.nature.com/articles/s41598-022-19948-1. Acesso em 27 de Outubro de 2023.

[13] LINK, D.; ANZER, G. . 2022. How the COVID-19 Pandemic has Changed the Game of Soccer. Int J Sports Med.

[14] LÖNING, M, *et al*. 2019. sktime: A Unified Interface for Machine Learning with Time Series.

[15] MCDONALD, G. . 2009. Ridge regression. WIREs Comput. Stat. 1, 1, 93–100. Disponível em: https://doi.org/10.1002/wics.14. Acesso em 4 de Maio de 2024.

[16] MIDDLEHURST, M.; SCHÄFER, P.; BAGNALL, A. . 2024. Bake off redux: a review and experimental evaluation of recent time series classification algorithms. Data Min Knowl Disc. Disponível em: https://doi.org/10.1007/s10618-024-01022-1. Acesso em 18 de Abril de 2024

[17] Online Sports Betting - Worldwide. Statista. 2023. Disponível em: https://www.statista.com/outlook/dmo/eservices/online-gambling/online-sports-betting/worldwide. Acesso em 27 de Outubro de 2023.

[18] PARIATH, R.; SHAH, S.; SURVE, A.; MITTAL, J. . 2018. Player Performance Prediction in Football Game. Disponível em: https://ieeexplore.ieee.org/document/8474750. Acesso em 27 de Outubro de 2023.

[19] PEDREGOSA, F; et al. 2011. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825--2830.

[20] PRASETIO, D. *et al*. 2016. Predicting football match results with logistic regression.

[21] ROBBERECHTS, P.; HAAREN, J. V.; DAVIS, J. . 2021. A Bayesian Approach to In-Game Win Probability in Soccer. Disponível em: https://arxiv.org/pdf/1906.05029.pdf. Acesso em 27 de Outubro de 2023.

[22] RODRIGUES, F.; PINTO, A. . 2022. Prediction of football match results with Machine Learning. Disponível em: https://www.sciencedirect.com/science/article/pii/S1877050922007955. Acesso em 27 de Outubro de 2023.

[23] RUIZ, A.; FLYNN, M.; LARGE, J.; *et al*. . 2021 The great multivariate time series classification bake off: a review and experimental evaluation of recent algorithmic advances. Data Min Knowl Disc 35, 401–449. Disponível em: https://doi.org/10.1007/s10618-020-00727-3. Acesso em 18 de Abril de 2024.

[24] SCHÄFER, P; LESER, U. . 2017. Multivariate Time Series Classification with WEASEL+MUSE.

[25] TALATTINIS, K.; KYRIAKIDES, G.; KANPATAI, E.; STEPHANIDES, G. . 2019. Forecasting Soccer Outcome Using Cost-Sensitive Models Oriented to Investment Opportunities. International Journal of Computer Science in Sport. 18. 93-114.

[26] TAN, C.; DEMPSTER, A.; BERGMEIR, C. *et al*. 2022 MultiRocket: multiple pooling operators and transformations for fast and effective time series classification. Data Min Knowl Disc 36, 1623–1646. Disponível em: https://doi.org/10.1007/s10618-022-00844-1. Acesso em 19 de Abril de 2024.

[27] TAX, N.; JOUSTRA, Y. . 2015. Predicting The Dutch Football Competition Using Public Data: A Machine Learning Approach.

[28] WANG, Z.; YAN, W.; OATES, T. . 2017. Time series classification from scratch with deep neural networks: A strong baseline.

[29] YAO, W.; WANG, Y.; ZHU, M.; CAO, Y.; ZENG, D. . 2022. Goal or Miss? A Bernoulli Distribution for In-Game Outcome Prediction in Soccer.

[30] ZERVEAS, G.; JAYARAMAN, S.; PATEL, D.; BHAMIDIPATY, A.; EICKROFF, C. . 2021. A Transformer-based Framework for Multivariate Time Series Representation Learning. In Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining (KDD '21). Association for Computing Machinery, 2114–2124. Disponível em:

https://doi.org/10.1145/3447548.3467401. Acesso em 4 de Maio de 2024.

[31] ZHANG, X.; GAO, Y.; LIN, J.; LU, C. . 2020. TapNet: Multivariate Time Series Classification with Attentional Prototypical Network. Proceedings of the AAAI Conference on Artificial Intelligence, 34(04), 6845-6852. Disponível em: https://doi.org/10.1609/aaai.v34i04.6165. Acesso em 20 de Abril de 2024.

Sobre o autor:

Pedro Antônio Barboza Ribeiro é aluno do curso de Ciência da Computação na Universidade Federal de Campina Grande.