Avaliação do modelo de previsão Prophet como ferramenta para preenchimento de falha de dados em séries climáticas

Miguel de Lima[1], Ivairton Monteiro Santos[1]

[1]Instituto de Ciências Exatas e da Terra – Campus Universitário do Araguaia – UFMT – Barra do Garças/MT

miguelfljunior@gmail.com, ivairton@ufmt.br

Abstract. *The development of forecasting systems faces the challenge of dealing with inconsistent and missing climate data, resulting from various causes, such as sensor failures. Two forecasting models were compared in this study: Prophet (Facebook) and Long Short Term-Memory (LSTM). Both were applied to real daily meteorological data (temperature, precipitation, and humidity) from Brazilian stations (INMET). While Prophet stood out for its simplicity and speed in analyzing large volumes of data, the results showed that LSTM performed better at filling data gaps in some scenarios. The choice of model depends on balancing applicability and efficiency.*

Resumo. *O desenvolvimento de sistemas de previsão enfrenta o desafio de lidar com dados climáticos inconsistentes e ausentes, resultantes de diversas causas, como falhas em sensores. Dois modelos de previsão foram comparados neste estudo: o Prophet (Facebook) e o Long Short Term-Memory (LSTM). Ambos foram aplicados a dados meteorológicos diários reais (temperatura, precipitação e umidade) de estações brasileiras (INMET). Embora o Prophet tenha se destacado pela simplicidade e rapidez na análise de grandes volumes de dados, os resultados demonstraram que o LSTM apresentou melhor desempenho em preencher lacunas em alguns cenários. A escolha do modelo depende de um equilíbrio entre aplicabilidade e eficiência.*

1. Introdução

Falhas em séries de dados temporais são comuns, especialmente em contextos meteorológicos. Elas podem ocorrer por falhas de sensores ou descontinuidade de estações, impactando negativamente a qualidade das inferências estatísticas e o desenvolvimento de modelos de previsão e análise de risco [Wood *et al*., 2004].

Há diversos métodos para lidar com falhas em séries temporais e preencher lacunas, entre eles destacam-se a Interpolação, Média Móvel Ponderada, Imputação Clássica, Amostra Aleatória e Decomposição Sazonal. Mais recentemente técnicas que empregam Machine Learning (ML) têm sido empregadas para esta mesma finalidade. Uma técnica de ML promissora em sistemas comerciais, com bons resultados em séries com sazonalidade e com muitas temporadas de dados é o Prophet[1] [Taylor *et al*.,2017].

[1] Projeto Prophet. Disponível em: <https://facebook.github.io/prophet/>

Este trabalho consiste do estudo de séries de dados climáticos, com foco nos dados da temperatura compensada média, umidade relativa do ar e precipitação, que são variáveis climáticas importantes para diferentes tipos de estudos, como por exemplo no desenvolvimento de sistemas de previsão de dispersão de doenças em lavouras e análise de risco. Na série dos dados temporais, as lacunas apresentam-se de várias tamanhos e quantidades, no entanto, este trabalho lidou com o contexto de uma única ocorrência de lacuna longa (período mínimo de um mês).

Sendo assim, este trabalho propõe o desenvolvimento de um sistema de preenchimento de lacunas em séries temporais utilizando o modelo de ML Prophet e compará-lo com uma técnica reconhecida pela literatura para esta finalidade, o LSTM.

2. Metodologia

2.1 Coleta, avaliação e preparação dos dados

Os dados utilizados neste trabalho foram obtidos do Instituto Nacional de Meteorologia (INMET), da base BDMEP[2]. Foram selecionados dados diários das estações convencionais disponíveis no sistema, que possuem dados para as variáveis da (*i*) média diária da umidade relativa do ar, (*ii*) média diária de temperatura compensada média e (*iii*) precipitação total diária, entre as datas de 01/01/2001 e 31/12/2021.

Após a coleta, foi avaliada a quantidade de dados faltantes, por variável, e para cada estação meteorológica. As estações que possuíam mais de 30% dos seus registros com falha foram descartadas. Ao final da análise, foram identificadas 4 estações com série de dados mais consistente, com menos de 10 dias de dados faltando, sendo elas: Belo Horizonte/MG, Salinas/MG, Caratinga/MG e Irati/PR. Estas estações foram usadas nos testes dos modelos de preenchimento de dados faltantes implementados neste trabalho.

Após a seleção das séries, foi modelada uma base de dados em PostgreSQL, para melhor performance na manipulação destes dados. A base de dados possui duas entidades: “estacao” e “dados_diarios”, que respectivamente armazenam as informações das estações meteorológicas e a série dos dados diários registrados.

Após o carregamento dos dados no SGBD, foi feita uma exploração (análise estatística) nos dados. Foram obtidos os valores de máximo, mínimo, média e desvio padrão das três variáveis meteorológicas adotadas neste trabalho. Estes valores foram calculados para toda a série histórica e nos períodos selecionados para testes. Essas métricas nos ajudam a compreender melhor nosso conjunto de dados, tanto a média para verificar a tendência central quanto o desvio padrão para analisar o grau de variação da série temporal.

2.2 Método para preenchimento de dados baseado no Prophet

Foi empregado o modelo Prophet com sua configuração padrão, alterando apenas o uso ‘com’ ou ‘sem’ regressores (opção disponível pelo modelo) e definida a sazonalidade como anual para as séries temporais trabalhadas.

---

[2] BDMEP - Banco de Dados Meteorológico para Ensino e Pesquisa. Disponível em: <https://bdmep.inmet.gov.br/>

O Prophet é um modelo de fácil utilização. Para o processo de treino do modelo é necessário um conjunto de dados de entrada com a primeira coluna contendo a data do registro (o nome desta coluna deve ser necessariamente "ds") e uma segunda coluna com os dados que se deseja que o modelo utilize (que terá o nome "y"). Para realizar o treinamento do modelo não é necessário normalizar os dados.

O Prophet tem como possibilidade adicionar junto ao conjunto de dados de treino do modelo variáveis que podem ajudar no processo de previsão. No modelo este recurso é chamado de "regressores" e é valioso quando há correlação entre variáveis. Com essa estratégia, vale destacar que isso irá implicar na necessidade de informar essas variáveis como parâmetro no processo de previsão. Neste trabalho, vamos avaliar o modelo com e sem o uso de regressores.

2.3 Método de referência LSTM

Com o objetivo de estabelecer um modelo como referência para comparação com o Prophet, utilizamos o LSTM, uma rede neural recorrente com bons resultados para problemas de previsão de séries temporais [Hewage, *et al*., 2019] [Karevan & Suykens, 2020].

Para o desenvolvimento da rede neural com arquitetura LSTM, utilizamos a biblioteca Keras, uma API que facilita a construção, treinamento e uso de uma rede neural.

Diferente do Prophet que normaliza os dados de forma interna, na utilização da LSTM é necessário normalizar os dados previamente. Outra diferença técnica em relação ao Prophet é o método de entrada dos dados para treinamento. No Prophet é informado como entrada também a variável que se deseja prever, enquanto que no LSTM a variável alvo não compõe o conjunto de dados de entrada para treino.

Uma etapa fundamental na utilização da rede LSTM é a calibração dos *hiperparâmetros* do modelo, como número de épocas, tamanho de lote e taxa de aprendizagem. Para definir estes parâmetros foi realizado um estudo em trabalhos que utilizaram o LSTM em dados climáticos [Attri, *et al*., 2020]. Os primeiros padrões de configuração adotados foram originados destes trabalhos. Partindo disto, foram realizados testes empíricos até se chegar no melhor cenário encontrado em relação a série utilizada.

3. Testes

Sabe-se que as estações climáticas do ano determinam alterações no padrão destes dados meteorológicos, portanto o planejamento dos testes considerou os períodos em que a tendência dos dados é mais uniforme e também períodos mais desafiadores, onde ocorre a mudança dessas tendências (transição entre estações do ano).

Os testes foram pensados para lidar com uma única falha "longa" (sequência de vários dias sem registros), o que é justamente o tipo de falha que mais impacta na série climática.

Foram definidos 5 casos de testes. Quatro deles possuem ausência de dados (gerados artificialmente) em uma posição intermediária da série, de tamanho equivalente a 6 meses. O posicionamento preciso desta lacuna varia em cada caso de teste, de modo que a lacuna irá coincidir com a transição das estações do ano. Adicionalmente, um dos casos de teste tem estrutura diferente, o posicionamento da

lacuna de dados ocorre no final da série e o seu tamanho é equivalente a 20% do tamanho total. Este caso baseia-se num teste clássico na área de ML que divide seu conjunto de dados numa proporção 80x20, onde 80% dos dados são usados para treino e 20% para testes.

Assim, os casos de testes foram nomeados conforme: Caso de Teste 1–Proporção 80/20; Caso de Teste 2–Outono/Inverno; Caso de Teste 3–Primavera/Verão; Caso de Teste 4–Inverno/Primavera; e Caso de Teste 5–Verão/Outono.

Para avaliar a qualidade dos modelos serão comparados os valores preditos com os valores medidos (reais) e foram adotadas três métricas de erros: *mean absolute error* (MAE), *mean absolute percentage error* (MAPE) e *root mean squared error* (RMSE).

4. Resultados

Antes de realizar os testes para os cenários propostos foi feita uma análise exploratória dos dados que compõem as séries temporais utilizadas. Foram calculadas a média, o desvio padrão, valor mínimo e máximo para cada variável e para cada série das estações meteorológicas. Outra análise foi quanto a correlação entre as variáveis, pelo método de Pearson. Os resultados desta análise demonstraram que nenhuma das tem correlação forte, conforme pode ser observado no gráfico da Figura 1.

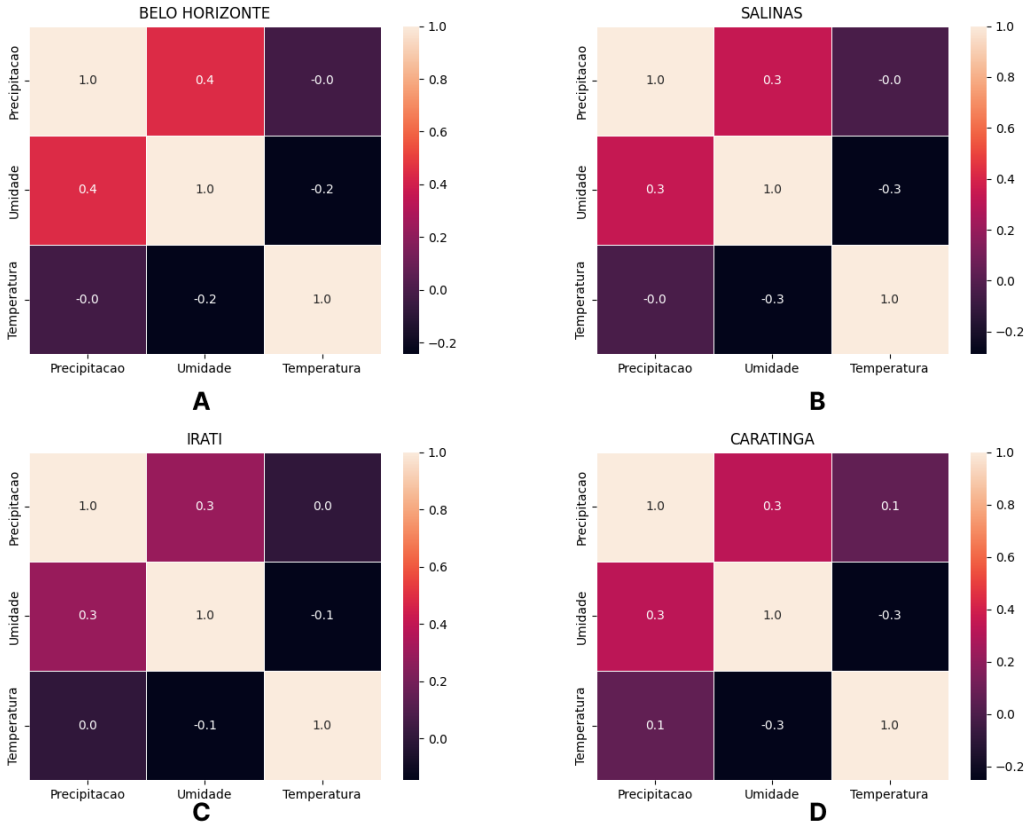

Figura 1. Comparação das médias dia-a-dia das correlações de Pearson entre as variáveis climáticas para cada estação do ano. Belo Horizonte/MG (A), Salinas/MG (B), Irati/PR (C) e Caratinga/MG (D).

Com a aplicação dos modelos, obtivemos resultados com padrões diferentes, para as diferentes variáveis analisadas. Ou seja, a depender da variável temos comportamentos distintos entre os modelos de tal forma que hora um modelo se adapta melhor, hora outro. Os cenários de testes propostos auxiliaram na busca pelo

mapeamento do comportamento dos modelos, mas mesmo assim não foi possível definir um padrão.

Na sequência são apresentados alguns dos gráficos (Figura 2, 3 e 4) que demonstram a comparação das predições dos modelos, para algumas das séries das estações meteorológicas. Nos gráficos a linha amarela representa o valor medido (real), a linha em vermelho corresponde ao Prophet sem regressores, enquanto que a linha azul representa o Prophet com regressores e a linha verde corresponde à predição pelo LSTM.

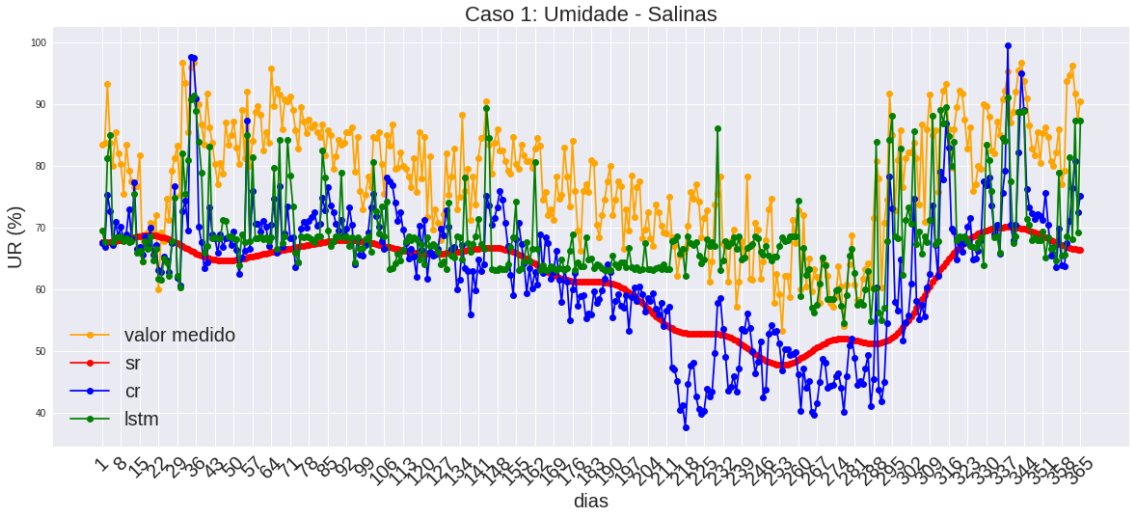

Figura 2. Comparação dos valores preditos pelos modelos e o valor real de média diária de temperatura compensada média, do ano de 2018, em Caratinga/MG.

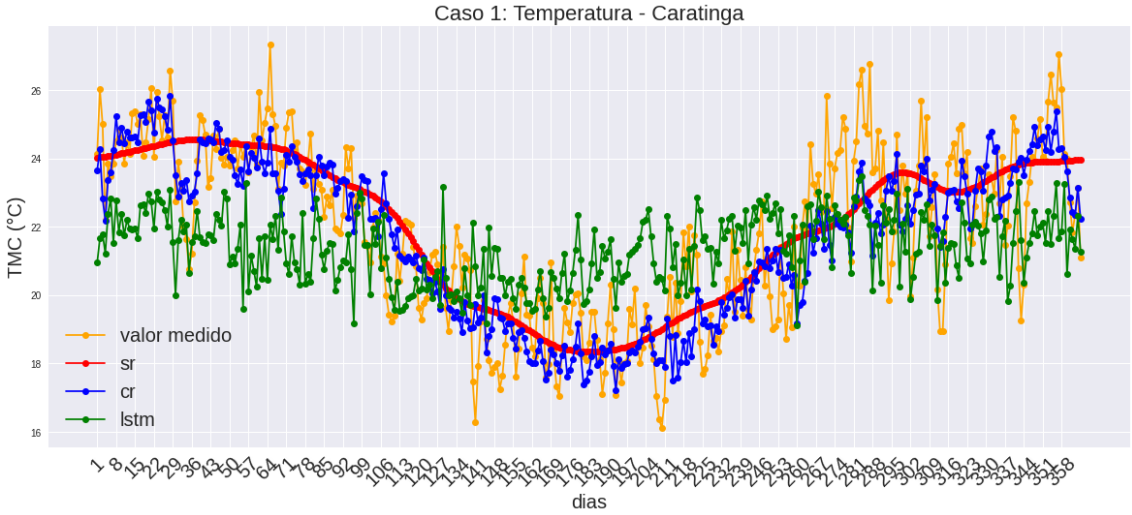

Figura 3. Comparação dos valores preditos pelos modelos e o valor real de média diária de temperatura compensada média, do ano de 2018, em Caratinga/MG.

5. Conclusões

De modo geral, avaliando um balanço entre usabilidade e resultados, o Prophet mostrou-se como uma alternativa mais simples que o LSTM, mesmo em contextos onde não há correlação entre as variáveis. No contexto da precipitação, por exemplo, os

regressores podem ter prejudicado, devido a não existência de correlação com a variável temperatura. No entanto, o Prophet mostrou bons resultados, sendo algumas vezes melhores que o LSTM.

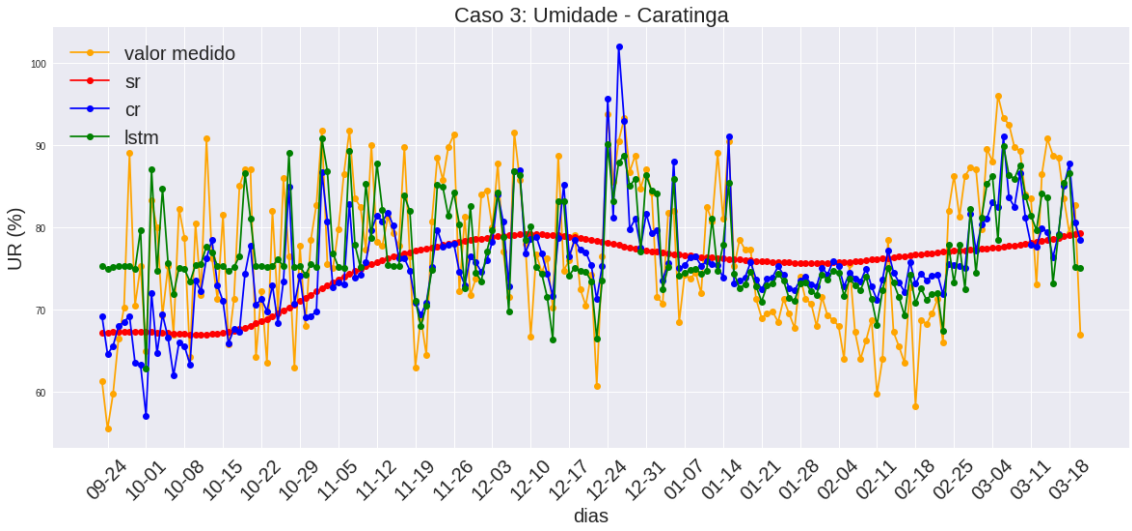

Figura 4. Comparação dos valores preditos pelos modelos e o valor real de média diária de umidade relativa do ar, estações climáticas Primavera e Verão, 23/09/2010 a 21/03/2011, em Salinas/MG.

O LSTM, desde que bem configurado e num cenário favorável, com variáveis de forte correlação, é uma alternativa poderosa, mostrou-se bem em testes com longos períodos de ausência de dados, por exemplo no contexto da umidade. Tanto o Prophet quando o LSTM se mostraram como boas opções e viáveis para lidar com o problema de dados faltantes, sendo necessário avaliar o contexto e tipo da série temporal, a praticidade de implementação e a precisão desejada.

Após os testes, avaliamos que o Prophet é uma opção válida para lidar com o problema de lacunas em séries temporais, principalmente pensando na simplicidade de implementação e uso.

Referencias

Attri, P. Sharma, Y.. Takach, K. Shah, F., “Timeseries forecasting for weather prediction”. Keras. 2020. Disponível em: < https://keras.io/examples/timeseries/timeseries_weather_forecasting/>.

Wood, A. M., White, I. R., Thompson, S.G., “Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals”. Clin Trials. 2004;1(4): 368-76. doi: 10.1191/1740774504cn032oa.

Taylor, S. J., Letham, B., “Forecasting at scale”. PeerJ Preprints 5:e3190v2. 2017. https://doi.org/10.7287/peerj.preprints.3190v2

Hewage, P., Behera, A., Trovati, M., Pereira, E., “Long-Short Term Memory for an Effective Short-Term Weather Forecasting Model Using Surface Weather Data”. In: Artificial Intelligence Applications and Innovations. AIAI 2019.

Karevan, Z., Suykens, J. A. K.. “Transductive LSTM for time-series prediction: An application to weather forecasting”. Neural Networks. (2020). doi:10.1016/j.neunet.2019.12.030