Previsão sobre séries temporais do mercado de energia usando Aprendizado de Máquina

Alessandra Cristina Nery Lima

Centro de Ciências Computacionais – Universidade Federal do Rio Grande (FURG)
Programa de Pós-Graduação em Ciência de Dados – Rio Grande – RS – Brasil

alessandra.nery@furg.br

Abstract. *The Brazilian electricity market is composed of a national interconnected system categorized by four submarkets. The Preço da Liquidação das Diferenças (PLD) is a strategic variable for the management decision of companies operating in regulated and free contract environments. In this sense, the present study aims to predict, considering a modeling of machine learning, positive or negative differences in PLD in view of the recent change in the calculation methodology from the weekly frequency base to an hourly basis. The results obtained from the study indicated that the machine learning model used predicts the PLDs with satisfactory performance in the short term.*

Resumo. *O mercado de energia elétrica brasileiro é composto por um sistema interligado nacional categorizado por quatro submercados. O Preço de Liquidação das Diferenças (PLD) é uma variável estratégica para decisão de gestão das empresas que atuam em ambientes regulados e de livre contratação. Neste sentido, o presente estudo avalia a qualidade da predição, via Aprendizado de Máquina (AM), das diferenças positivas ou negativas no PLD, tendo em vista a recente alteração da metodologia de cálculo que abandona a base de frequência semanal para uma base horária. Os resultados obtidos indicaram que o modelo de Aprendizado de Máquina utilizado prevê os PLDs com performance satisfatória no curto prazo.*

1. INTRODUÇÃO

É importante estudar e analisar a trajetória do Preço de Liquidação das Diferenças (PLD)[1] no mercado de energia elétrica para informar estrategicamente os agentes geradores, comercializadores e consumidores de energia elétrica que atuam no ambiente de livre contratação.

A partir de janeiro de 2021, com o advento da mudança na medição no PLD, passando de base semanal para base horária, surge um novo contexto que fez intensificar a necessidade dos agentes em controlar e planejar os preços.

[1] PLD é o valor do preço da energia calculado pela Câmara de Comercialização de Energia Elétrica (CCEE) diariamente para cada hora do dia seguinte, disponível em https://www.ccee.org.br/precos/conceitos-precos#:~:text=O%20Pre%C3%A7o%20de%20Liquida%C3%A7%C3%A3o%20das,apura%C3%A7%C3%A3o%20e%20para%20cada%20submercado. Acesso em 07/04/2022.

Em um cenário caracterizado por alta sazonalidade e imprecisão do PLD, a utilização de algoritmos de Aprendizado de Máquina (AM)[2] pode trazer soluções que permitam a realização de planejamento financeiro e estratégico aos agentes.

Desta forma, pretende-se com este estudo obter uma previsão confiável do preço (PLD) para uma perspectiva futura com intuito de auxiliar e beneficiar os agentes de mercado, proporcionando otimização e redução de custos, gerando contratações estratégicas, atingindo assim o melhor preço, volume, prazo e forma de pagamento.

1.1 Objetivos

Cabe ressaltar que, entre outros, o principal fator que direcionou a escolha da presente temática foi a busca por novas experiências, aprendizados, novos projetos e oportunidades. Ademais, estudar o comportamento dos preços no mercado de curto prazo de energia elétrica brasileiro é uma oportunidade de aplicar e unir conhecimentos advindos das Ciências Econômicas à Ciência de Dados.

O objetivo geral do estudo busca desenvolver, por meio da técnica de AM, a previsão da trajetória comportamental do PLD no mercado de curto prazo de energia elétrica.

Os objetivos específicos consistem em:

- Fazer uma revisão bibliográfica levantar conceitos e técnicas de AM;
- Adquirir e caracterizar os dados do PLD;
- Efetuar a preparação, treinamento e validação de um modelo preditivo;
- Comparar dados previstos com dados reais e traçar uma análise da eficácia desta abordagem

1.2 Organização do texto

O texto está estruturado em seções que contemplam a parte introdutória, a fundamentação teórica, a metodologia, o desenvolvimento, a conclusão e as referências bibliográficas.

2.FUNDAMENTAÇÃO TEÓRICA

Neste capítulo são apresentadas informações que podem auxiliar no embasamento das ideias discutidas no trabalho. Alguns tópicos se propõem a contextualizar sobre o setor de energia elétrica brasileiro, as técnicas de AM e as séries temporais.

2.1 Mercado de energia elétrica

A comercialização de energia elétrica no mercado brasileiro é composta por dois ambientes de contratação: livre e regulado. No ambiente de contratação livre os agentes (Comercializadores, Geradores, Consumidores e Distribuidores) estão expostos a fortes oscilações de preços do mercado de curto prazo (MCP).

De acordo com Leite et al. (2013), a partir de 2005, a oferta e demanda de energia elétrica ficaram bem próximas e houve uma maior volatilidade e imprevisibilidade do

[2] AM é uma forma de Inteligência Artificial em que um algoritmo computacional constrói, a partir de dados, modelos de aprendizado para a resolução de problemas (MEHTA, 2017). Seu conceito será aboardado na seção 2.2.

PLD, descrevendo assim um cenário de intensas incertezas e riscos no MCP (ou mercado *spot*).

No entanto, há que se ressaltar que os preços, no mercado de curto prazo brasileiro, não são dependentes somente da relação entre oferta e demanda, mas sim de cálculos matemáticos que incluem fatores complexos como capacidade de geração, hidrologia e pluviometria (Clímaco, 2010).

Assim sendo, a precificação da energia é decorrente principalmente da disponibilidade de água nos reservatórios e do nível de precipitação pluviométrico. De acordo com Leite et al. (2013), a volatilidade dos preços está relacionada, principalmente, com a dinâmica das afluências.

O MCP é definido basicamente pela liquidação das diferenças de balanço energético. É onde são contabilizadas as diferenças entre a energia contratada e o volume que foi realmente gerado ou consumido.

O preço de referência de compra e venda no MCP é determinado pelo PLD. A Câmara de Comercialização de Energia Elétrica (CCEE)[3] é a entidade responsável pela contabilização e pela liquidação financeira no MCP.

O cálculo do preço é apurado com base em informações previstas, anteriores à operação real do sistema, considerando-se os valores de disponibilidades declaradas de geração e o consumo previsto de cada submercado (CCEE, 2019).

O PLD é o valor de energia calculado em R$/MWh, sendo o balizador das negociações no Ambiente de Contratação Livre (ACL)[4]. É possível dizer que o PLD serve para equilibrar os custos entre oferta e demanda de energia no país.

Por exemplo: caso um agente contrate mais energia do que consome em um determinado mês, essa sobra será liquidada ao chamado PLD. Contudo, essa liquidação é facultativa ao consumidor, visto que existe a possibilidade de venda desse excedente de energia para um agente comercializador.

O MCP é o garantidor da liquidez das negociações de energia elétrica e também o responsável por balancear as operações, de tal forma que não falte nem sobre energia para o consumidor.

É importante ressaltar que em algumas vendas podem ocorrer o *spread* positivo (quando o valor do PLD está mais alto do que o valor contratado), o que viabiliza maior rapidez na negociação e possibilidade de lucros na operação. É nesse sentido que obter uma predição fidedigna com a realidade pode trazer aos agentes do mercado a possibilidade de encontrar oportunidades no MCP.

Uma análise aproximada das variações do PLD no MCP pode auxiliar o momento da negociação, trazendo lucros maiores nas chamadas operações estruturadas, operações "não comuns" de compra, venda ou troca de energia.

---

[3] CCEE informa os PLDs em base horária. Diponível em https://www.ccee.org.br/web/guest/precos/painel-precos.

[4] A CCEE estabelece que para participar do ALC as empresas de geração, distribuição e comercialização precisam estar assoociadas à CCEE e possuirem demanda a partir de 0,5 MW.

2.2 Aprendizado de Máquina

AM é um método em que um algoritmo computacional constrói, a partir de dados, modelos de aprendizado para a resolução de problemas (Mehta, 2017). O conceito surgiu em 1959 com Arthur Samuel definindo AM como campo de estudo que dava aos computadores a capacidade de aprender sem ser explicitamente programado.

De acordo com Mitchell (1997), um programa de computador aprende com experiência E em relação a uma tarefa T e a alguma medida de desempenho P, quando o seu desempenho em fazer T, conforme medido por P, melhorar com a experiência E.

Dessa forma, temos três principais elementos de AM que são: a tarefa (T), a experiência (E) e a medida de desempenho (P).

De acordo com Gerón (2019), podemos dizer sinteticamente que AM é útil para:

- Solucionar problemas que demandam muita afinação manual ou longas listas de regras;
- Problemas complexos em que não há uma boa solução usando uma abordagem tradicional;
- Ambientes flutuantes: um sistema de AM pode adaptar-se a novos dados e;
- Obter *insights* sobre problemas complexos e grandes quantidades de dados.

Os sistemas de AM são classificados de acordo com a abordagem e o tipo de supervisão que recebem durante o treinamento dos dados, podendo possuir uma abordagem de aprendizado supervisionado, de aprendizado não supervisionado, de aprendizado semi supervisionado ou de aprendizado por reforço.

No aprendizado supervisionado os dados de treinamento (chamados de *features*) incluem soluções desejadas, chamadas de rótulos (ou *labels*). Entre as tarefas típicas que compõem este tipo de aprendizado temos a classificação e a regressão linear.

A classificação é um processo de aprendizagem que atribui uma classe a cada nova instância, já a regressão linear faz previsão de um valor numérico alvo com base em valores chamados preditores.

No aprendizado não supervisionado os dados de treinamento não são rotulados e o sistema tenta aprender sem o apoio humano. Exemplos de tarefas que regem um aprendizado não supervisionado são: agrupamento (*clustering*), algoritmos de visualização, detecção de anomalias, entre outras.

Com relação à etapa de validação e testes, uma boa opção é dividir os dados em dois conjuntos: o conjunto de treinamento e de testes (Gerón, 2019). Com isso, como o próprio nome diz, o modelo é treinado com conjunto de treino e testado com outro conjunto (conjunto de testes).

Gerón (2019) aponta ainda que com os testes temos as taxas de erro em casos novos, chamada de erro de generalização (ou erro fora da amostra). Trata-se de métricas para avaliar a qualidade de previsão do modelo, caracterizadas como funções de utilidade, que medem o quão bom é o modelo e funções de custo, que apontam o nível de deficiência do modelo em análise.

Em geral, o processo de AM considera as seguintes etapas: análise dos dados e seleção de *features*; escolha da abordagem de AM; definição do modelo e de seus parâmetros; treino com os dados de treinamento; avaliação com os dados de teste e aplicação do modelo para fazer previsões sobre novos casos.

Em síntese, o presente trabalho está baseado no processo acima descrito de tal forma que, por meio da utilização de dados é feito o aprendizado das informações criando modelos de decisões que são utilizados como ferramentas de predição. Ou seja, é fazer uso da aprendizagem de máquina para aprender através da experiência (Géron, 2019).

2.3 Séries temporais

Uma série temporal consiste num conjunto de registros observados num intervalo de tempo regular. No campo da Ciência de Dados, as séries temporais têm encontrado representatividade ao propor soluções de problemas como previsão de demanda, de vendas, de valores mensais do IPC-A, de valores de fechamentos diários do IBOVESPA[5], de valor de Bitcoin[6], entre outros.

O objetivo da análise de séries temporais é buscar identificar padrões não aleatórios, ou seja, há uma suposição sobre a existência de um sistema causal mais ou menos constante, ordenado no tempo, que influenciou o conjunto de dados registrados no passado e que pode vir a influenciar também no futuro.

Assim sendo, por meio da identificação de padrões não aleatórios numa série temporal de uma determinada variável (preço, vendas, temperatura etc), aliada à observação comportamental de sua trajetória num período regular, pode-se viabilizar uma leitura preditiva orientando a tomada de decisão.

O modelo clássico das séries temporais trata de quatro componentes padrão: tendência, variações cíclicas ou ciclos, variações sazonais ou sazonalidade e variações irregulares (Souza,1989).

Segundo Morettin e Toloi (2006) a tendência representa o comportamento da variável no longo prazo. Pode ser causada por um crescimento demográfico, por uma mudança do perfil de consumo ou por qualquer outra variável de interesse no longo prazo.

Variações cíclicas ou ciclos são caracterizados por flutuações nos valores das variáveis com duração superior a um ano e que se repetem com certa periodicidade, como, por exemplo, os períodos de recessão ou crescimento econômico ou fenômenos climáticos como El Niño (que se repete com periodicidade superior a um ano).

As variações sazonais ou sazonalidades consistem em flutuações com duração inferior a um ano e que se repetem todos os anos. A sazonalidade pode estar atrelada, por exemplo, às estações do ano (período de seca ou chuva), festas comemorativas (natal, dia das mães), feriados ou até mesmo por exigências legais (entrega da Declaração do Imposto de Renda).

---

[5] https://www.b3.com.br/pt_br/

[6] Bitcoin é uma rede de pagamento inovadora e um novo tipo de dinheiro. Disponível em https://bitcoin.org/pt_BR/. Acessado em 08/04/2022.

Por fim, as variações irregulares apontam flutuações inexplicáveis, resultantes de fatos fortuitos e inesperados, catástrofes naturais, atentados terroristas, pandemia (COVID 19), decisões intempestivas do governo etc.

Em resumo, a decomposição clássica é útil tanto para planejamento como para previsão (Silver, 2000). Ademais, vale dizer que a decomposição da série temporal é uma ferramenta útil, à medida que permite realizar previsões e auxilia na tomada de decisão acerca do método de previsão mais adequado às características dos dados disponíveis (Souza; Samohyl; Meurer, 2004).

2.4 Prophet

A abordagem de AM adotada para este trabalho foi o Prophet, considrado um modelo de previsão de séries temporais automático, de código aberto lançado pela equipe Core Data Science do Facebook, disponível nas linguagens de programação Python e R (Taylor; Lethham, 2018).

De acordo com Lyla (2019), o Prophet é considerado uma ferramenta precisa e rápida e se utuliza de um baixo poder computacional, é robusto para dados ausentes e alterações de tendências lidando bem com outliers, é inteiramente automatizado, ou seja, desempenha uma boa previsão sem muito esforço manual.

O Prophet modela o compotamento da série temporal com combinação de três componentes: tendência, sazonalidade e feriados. A tendência da série pode assumir um comportamento de crescimento saturado ou de um modelo linear por partes, e é acrescido pelos valores correspondentes à sazonalidade da série e aos feriados (Taylor, Letham, 2018).

Segundo Satrio et al. (2020), o Prophet é um modelo de fácil utilização, pois seus parâmetros são simples e autimaticamente otimizados, fazendo com que o modelo seja flexível e útil em diversas aplicações.

2.5 Métricas de Erro e Validação Cruzada

A técnica de validação cruzada para medir o desempenho do modelo de previsão para dados históricos, consiste em selecionar pontos de corte (cutoff points) no histórico temporal e, para cada um deles, adaptar o modelo utilizando apenas dados até esse ponto de corte de tal forma que seja possível comparar valores previstos com valores reais.

Os dados são separados em conjunto de treinos e de testes, de modo que o primeiro comtempla a maior parte dos dados e é utilizado para elaboração do modelo, para elaboração de parâmetros e próprio fit do modelo, enquanto que o segundo é utilizado para mensuração da previsão (Zhang, 2007).

Nesse sentido, buscando avaliar a qualidade da previsão e verificar se o modelo apresenta boa assertividade foi realizada além da validação cruzada, a medição do erro verificado entre o conjunto o conjunto de teste (período jan/2022) e a previsão de 744 horas (31 dias).

As principais métricas utilizadas para avaliar o modelo utilizado neste trabalho foram:

- O Erro Absoluto Médio, do inglês *Mean Absolut Error* (MAE) que consiste em calcular o residual de cada ponto, no qual os valores residuais positivos e negativos não se acumulam. Após este agrupamento, calculamos a média desses residuais (Rabelo,2019). Representado pela equação 1:

$$MAE = \frac{1}{n}\sum_{t=1}^{n} |Yreal - Yprevisto| \quad (1)$$

onde:
n: número de dados utilizados no processo de estimativas,
Y real: valor real da amostra,
Y predito: o valor estimado na predição.

- O Root Mean Squared Error (RMSE) que baseia-se no quadrado da média das diferenças entre as predições e as observações reais , Pode-se assemelhar esta fórmula com a distância Euclidiana entre o vetor de 31 valores reais e o vetor de valores pressupostos, mediados pela quantidade de número de pontos ’n’ (Silva, 2019), como segue na equação 2:

$$RMSE = \sqrt{\sum_{t=1}^{n} \frac{(Yreal - Yprevisto)^2}{n}} \quad (2)$$

onde:
n: número de dados utilizados no processo de estimativas,
Y real: valor real da amostra,
Y predito: o valor estimado na predição

2.6 Trabalhos relacionados

A literatura vem explorando análises do PLD em diferentes perspectivas. Olivi et al. (2018) e Monlevade (2018) aplicam redes neurais artifciais para desenvover modelos preditivos do PLD no mercado de energia elétrica. Os autores demonstram que modelos de regressão obtiveram os melhores percentuais de acerto de tendências do PLD por considerarem variáveis de entrada relacionadas à operação do Sistema Interligado Nacional (Lagasse, 2020). O trabalho de Bianchi (2020) evidencia avanços que a mudança de medição para base horária do PLD relacionando a carga de energia fotovoltaica e redução de preços.

Trabalhos anteriores observaram que o algoritmo Prophet (adotado neste trabalho) obteve bom desempenho em séries temporais univariadas (Taylor; Lethham, 2018). Em modelos clássicos de séries temporais de temperatura e precipitação, especialmente quando aplicado a decomposição sazonal de séries temporais, o Prophet obteve resultados competitivos (Papacharalampous et al., 2018). Há também um estudo sobre previsão de valor de Bitcoin ao longo do tempo em que o Prophet resultou em melhor desempenho que o Arima (Sama et al., 2019).

3.METODOLOGIA

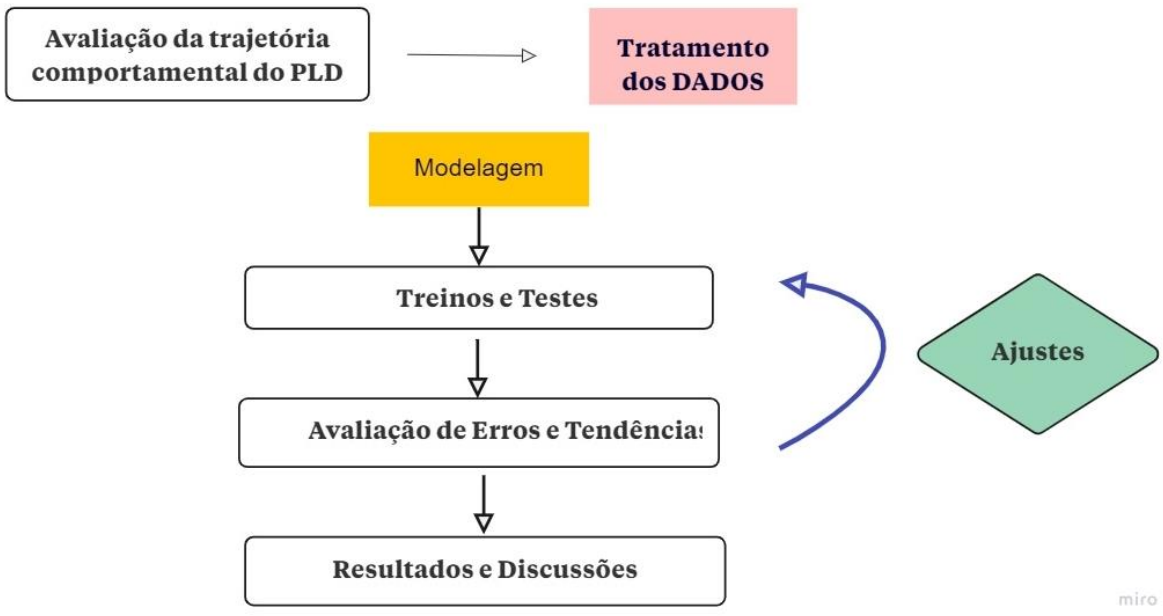
Nesta seção são apresentadas as principais etapas do processo que compõem o trabalho, as ferramentas, a caracterização dos dados e as análises preliminares.

3.1 Descrição do processo

Conforme demonstrado na Figura 1, a implementação do estudo proposto se deu primeiramente por meio da avaliação das variáveis que traçam a linha temporal dos PLDs, tendo como um primeiro passo o tratamento e a reorganização do conjunto de dados. Após este passo, foi realizada a entrada dos dados no modelo de AM para que fossem realizados os treinamentos e os testes sendo o ajustado até que se obtivesse um resultado com melhor desempenho e performance.

Foram utilizados, por meio da modelagem de AM, experimentos que promoveram ajustes nos critérios (como períodos da base de dados, quantidade de horas previstas e horizontes temporais de previsão) gerando assim avaliações de erros e tendências dos treinos e testes até a consolidação dos resultados e das discussões finais desse trabalho.

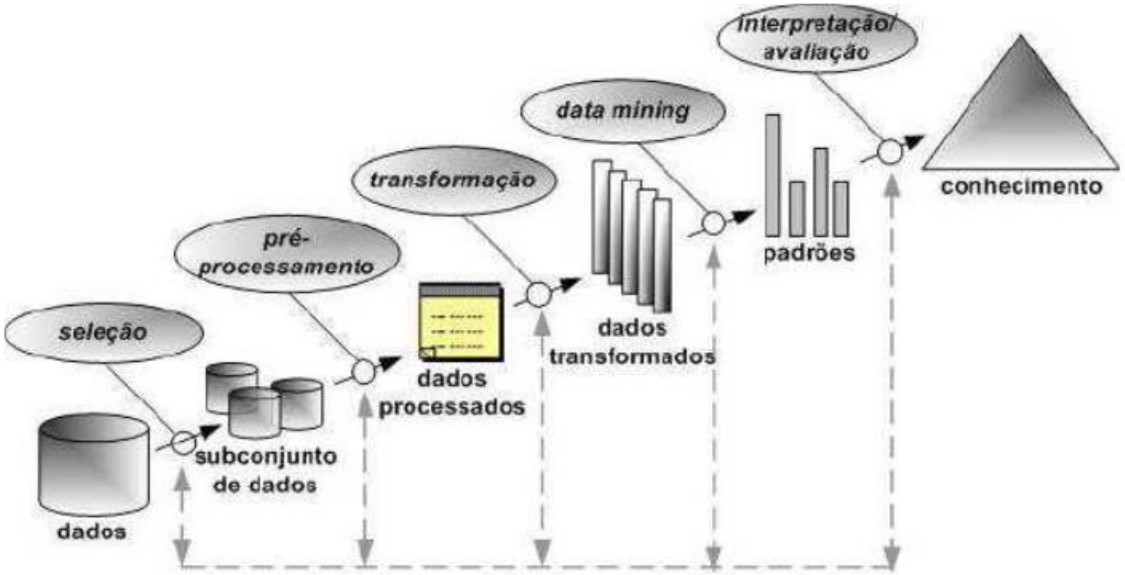
Figura 1 – Diagrama de processo



Fonte: elaborado pela autora

A metodologia desenvolvida nesse estudo está intimamente correlacionada com o processo de Descoberta de Conhecimento em Base de Dados, caracterizado pelo termo KDD (*Knowledge Discovery in Database*), em que Fayyad, Piatetsky-Shapiro e Smith (1996) explicitam várias etapas iniciando-se pela seleção dos dados a analisar, o pré – processamento dos mesmos, a transformação em dados tratados, a identificação de padrões resultando na interpretação do conhecimento.

Figura 2 – As etapas do processo de KDD



Fonte: Fayyad et al. (1996)

É importante mencionar que o trabalho se estruturou com a seguinte sistemática: construir dataset - base de séries temporais; avaliar o dataset e direcionar modelagem de AM; realizar treinos, previsões e testes; avaliar métodos e índices de acurácia, sensibilidade e validação cruzada; apresentar resultados e fazer inferências.

3.2 Ferramentas

As principais ferramentas utilizadas no desenvolvimento do trabalho foram: Jupyter Notebook, ambiente Colaboratory, Linguagem Python de programação e suas bibliotecas (Pandas) e o pacote Prophet desenvolvido pela empresa Facebook.

O Jupyter Notebook é uma aplicação *web* que permite a utilização de código interativo e textos explicativos (em *markdown*). É possível fazer uso de diversas linguagens de programação como Python e R.

O Jupyter Notebook pode ser utilizado por meio do Google Colaboratory, que é um ambiente em nuvem gratuito hospedado pelo próprio Google para incentivar a pesquisa no campo da Ciência de Dados. O ambiente colaborativo permite o compartilhamento do trabalho viabilizando sua modificação e leitura.

A linguagem de programação Python é considerada de alto nível, dinâmica, interpretada, modular, multiplataforma e orientada a objetos. É tida ainda como uma

linguagem de sintaxe relativamente simples e de fácil compreensão. Um de seus grandes atrativos é possuir um grande número de bibliotecas nativas e de terceiros.

A abordagem de AM utilizada foi o Prophet. Trata-se de um modelo bastante flexível de regressão não-paramétrica aditiva, que simplifica a preparação dos dados e a geração de previsões, ao mesmo tempo mantendo uma forma de uso muito semelhante a outros algoritmos de AM disponíveis na biblioteca de Scikit-Learn.

3.3 Caracterização dos dados

A predição do PLD busca solucionar uma tarefa de regressão simples baseada em séries temporais. A origem dos dados resultam da comercialização de energia elétrica – contratos de compra e venda que são registrados pela Câmara de Comercialização de Energia Elétrica (CCEE), por meio de programas de modelagem estatística.

Dentre os modelos utilizados pela CCEE temos o DECOMP[7], que é utilizado para determinar o despacho de geração que minimiza o custo total operacional num período de curto prazo (até 12 meses), e também o NEWAVE[8], que consiste num modelo de planejamento de operação de sistemas hidrotérmicos interligados de longo e médio prazo (5 anos).

Dessa forma, os dados reais que compõem o dataset do estudo foram coletados no site oficial da CCEE[9] e contemplam a informação da série temporal do PLD em base horária subdividida em 4 submercados (Sudeste/Centro-Oeste, Sul, Norte e Nordeste).

O período definido para o estudo delimita uma série temporal de 17/04/2018 a 31/01/2022. Com o objetivo de avaliar os impactos que novo PLD horário causaria no mercado, a CCEE criou o PLD "sombra" em 2018, tendo sua adoção formal somente a partir de janeiro de 2021.

3.4 Análises preliminares

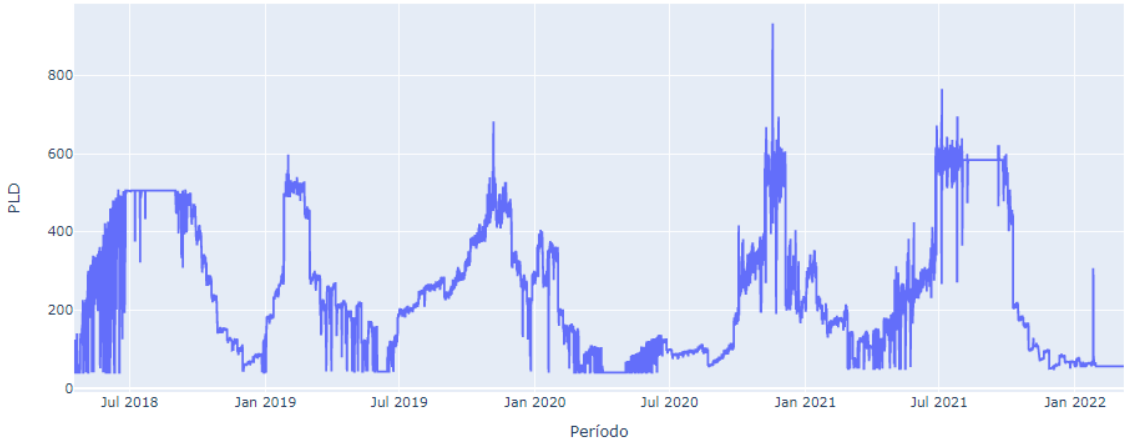
O Gráfico 1 apresenta a PLD em base horária do submercado Sudeste/Centro-Oeste calculado e divulgado pela CCEE diariamente, considerando o período de 17/04/2018 a 31/01/ 2022.

[7] DECOMP- é o modelo computacional usado no planejamento da operação de sistemas hidrotérmicos de curto prazo - com horizonte utilizado oficialmente de 2 meses - e discretização semanal para o primeiro mês. Disponível em https://www.ccee.org.br/pt/web/guest/precos/conceitos-precos.

[8] NEWAVE – é o modelo computacional para o planejamento da operação de sistemas hidrotérmicos de médio prazo (até 5 anos) - consegue determinar a estratégia de geração hidráulica e térmica em cada estágio que minimiza o valor esperado do custo de operação para todo o período de planejamento. Disponível em https://www.ccee.org.br/pt/web/guest/precos/conceitos-precos.

[9] https://www.ccee.org.br/web/guest/precos/painel-precos

Gráfico 1 – PLD em base horária submercado Sudeste/CO



Fonte: elaborado pela autora

Não é possível dizer que há uma tendência crescente ou decrescente. Há uma sucessão irregular de "picos e vales" no valor do PLD, começando em cerca de R$ 40,00 em abril de 2018, atingindo o maior pico, de R$933,00 em novembro de 2020, e terminando em R$ 62,77 no final de janeiro de 2022.

Observam-se flutuações sazonais que podem estar relacionadas a estações do ano, que se repetem anualmente (com maior ou menor intensidade). Tais padrões serão incorporados a um modelo de AM, possibilitando fazer previsões que auxiliarão na tomada de decisões.

4. DESENVOLVIMENTO

Esta seção do trabalho apresenta como se deu a organização dos dados, os treinamentos, os testes, os componentes da série temporal e como o modelo se comportou com as previsões.

4.1 Preparação dos dados

Inicialmente, os dados foram reorganizados, conforme demonstrado na Figura 2, estabelecendo a separação do *dataset* em quatro séries de dados uma para cada submercado. Assim, cada série passou a ser representada por um grande *dataframe* com um tempo (*timestep*) para cada linha e um único atributo (o preço naquele instante).

Figura 2 – Série univariadas – por submercado

```
series = pld_horarioCCEE.set_index('Submercado')
s_sudeste= series.loc[['SUDESTE']]
s_sul= series.loc[['SUL']]
s_nordeste= series.loc[['NORDESTE']]
s_norte= series.loc[['NORTE']]
```

```
# mantém todas menos a última linha
s_sudeste= s_sudeste.iloc[0:24]
s_sul = s_sul.iloc[0:24]
s_nordeste = s_nordeste.iloc[0:24]
s_norte = s_norte.iloc[0:24]
```

Fonte: elaborado pela autora

Após a fase de separação, cortes e consolidação de dados que resultaram em séries univariadas por região de submercado, tratou-se de desenvolver um layout do *dataframe* que se enquadrasse na abordagem de AM escolhida para o trabalho, o Prophet.

A entrada de dados para o Prophet é sempre um *dataframe* com duas colunas: `ds` e y. Sendo que a coluna `ds` (*datestamp*) deve ser formatada idealmente YYYY-MM-DD[10] para uma data ou YYYY-MM-DD HH:MM:SS[11] para um *timestamp*.

Neste sentido, foi criada uma nova coluna combinando dia e hora conforme consta na Figura 3.

Figura 3 – Reorganização dos dados - *timestamp*

```
c_sudeste['ds'] = pd.to_datetime(c_sudeste['Dia'].map(lambda d: str(d.date())) + ' ' + c_sudeste['Hora'].map(str) + ':00')
c_sudeste.head(24)
```

	Hora	Dia	y	ds
0	0	2018-04-17	40.16	2018-04-17 00:00:00
1	1	2018-04-17	40.16	2018-04-17 01:00:00
2	2	2018-04-17	40.16	2018-04-17 02:00:00
3	3	2018-04-17	40.16	2018-04-17 03:00:00
4	4	2018-04-17	40.16	2018-04-17 04:00:00
5	5	2018-04-17	40.16	2018-04-17 05:00:00

Fonte: elaborado pela autora

Ainda buscando estabelecer a mesma leitura do Prophet, foi definido '`ds`' como índice (é um *timestamp*) e removida as colunas 'Dia' e 'Hora' (Figura 4)

Figura 4 – Definição índice *timestamp:* '`ds`'

```
c_sudeste.drop(columns=['Dia', 'Hora'], inplace=True)
c_sudeste.head(4)
```

	y	ds
0	40.16	2018-04-17 00:00:00
1	40.16	2018-04-17 01:00:00
2	40.16	2018-04-17 02:00:00
3	40.16	2018-04-17 03:00:00

Fonte: elaborado pela autora

Antes de prosseguir para as fases de treinamento, previsões e testes, o conjunto de dados foi separado usando um critério arbitrário em dados de treino (cerca de 95% dos dias

[10] Ano (YYYY), Mes (MM) e Dia (DD)

[11] Ano (YYYY), Mes (MM) , Dia (DD), Hora(HH), Minutos (MM) e Segundos(SS)

da série temporal) e de teste (cerca de 5% dos dias da série temporal), conforme consta na Figura 5.

Figura 5 – Conjunto de treino e teste

```
treino= c_sudeste[c_sudeste.ds<'2021-11-01']
treino
```

```
teste= c_sudeste[c_sudeste.ds>='2021-11-01']
teste
```

Fonte: elaborado pela autora

4.2 Treinamento

Conforme mencionado, a abordagem de AM definida para treinar os dados é a biblioteca Prophet do Facebook. O Prophet segue a API (Aplicattion Programing Interface) modelo sklearn.

O modelo foi ajustado e instanciado ao novo objeto Prophet (m_sudeste). Em seguida, foi chamado o método de ajuste e informado o *dataframe* histórico da série temporal submercado Sudeste- CO - conjunto de dados de treino - conforme Figura 6.

Figura 6 – Conjunto de treino e teste

```
m_sudeste = Prophet()
m_sudeste.fit(treino)
```

Fonte: elaborado pela autora

O Prophet constrói um modelo procurando encontrar uma melhor linha que pode ser representada como a soma dos seguintes componentes:

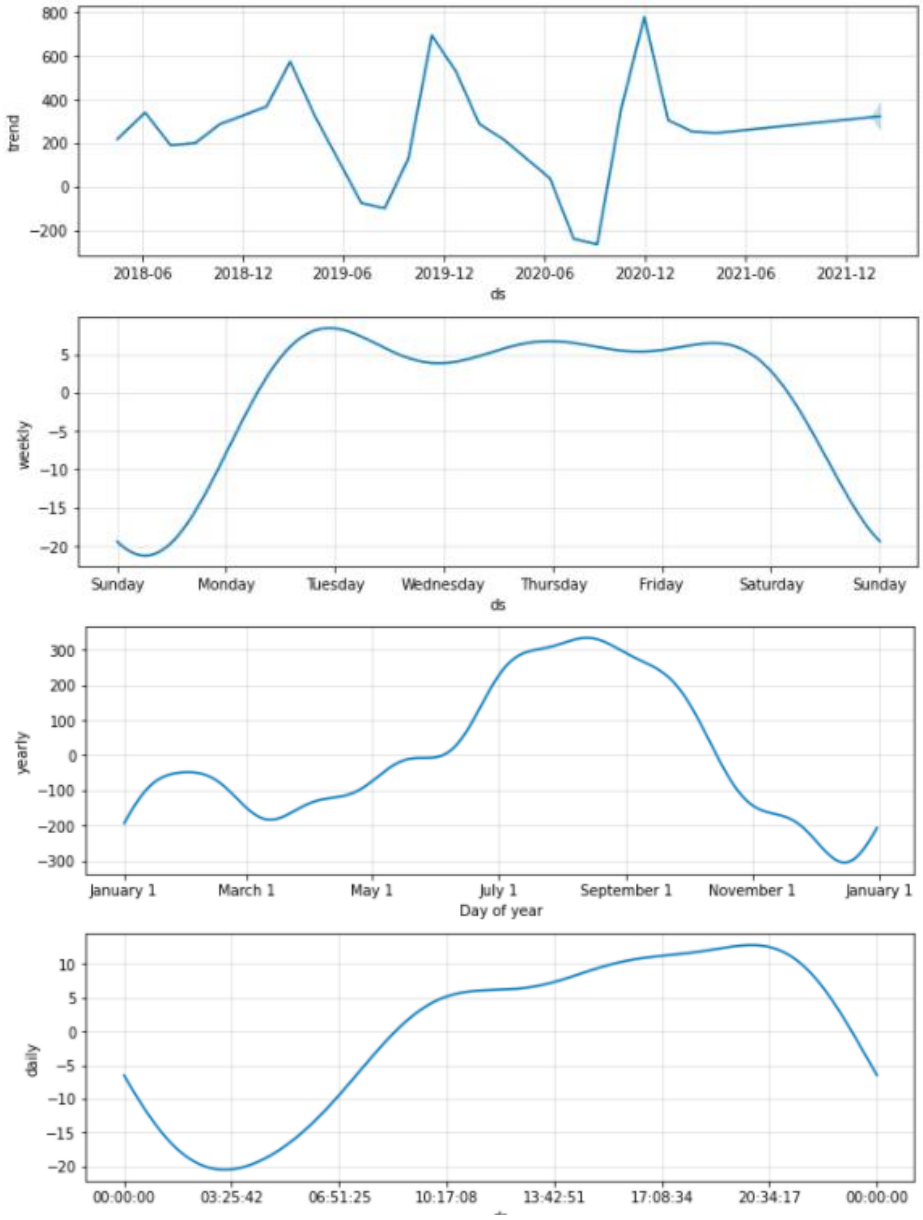
$y(t) = g(t) + s(t) + h(t) + \epsilon_t$

Tais componentes representam: tendência geral de crescimento g(t), sazonalidade anual s(t), sazonalidade semanal s(t) e efeitos de feriados h(t).

Para visualizar os componentes de previsão utilizamos o método `Prophet.plot_components`.

Assim, conforme demonstrado no Gráfico 2, é possível visualizar a tendência, a sazonalidade anual, a sazonalidade semanal e a sazonalidade diária da série temporal.

Gráfico 2 – Componentes de Previsão do PLD horário no Sudeste/CO



Fonte: elaborado pela autora

Os componentes gerados a partir do modelo Prophet para o conjunto de dados revelaram os seguintes pontos:

(1) Tendência: não há uma tendência definida na série cronológica analisada.

(2) Sazonalidade semanal: verifica-se que há mudanças periódicas dos valores aos finais de semana que é quando o nivel de consume sofre redução.

(3) Sazonalidade anual: há mudanças periódicas dos valores anuais que sofrem picos de julho a novembro. O que pode estar atrelado ao fato de termos de maio a novembro o período seco, com poucos chuvas e baixa nos reservatórios.

(4) Sazonalidade diária: há queda nos valores às 3h25 e valores crescentes alcançando o pico às 20h34. Também conhecido horário de ponta, há um período do dia,

geralmente entre 18h e 21h, em que se verifica o maior consumo de energia elétrica pela população.

O Gráfico 3 apresenta uma figura interativa da previsão do PLD e seus componentes informando variações constantes no preço e ausência de tendência.

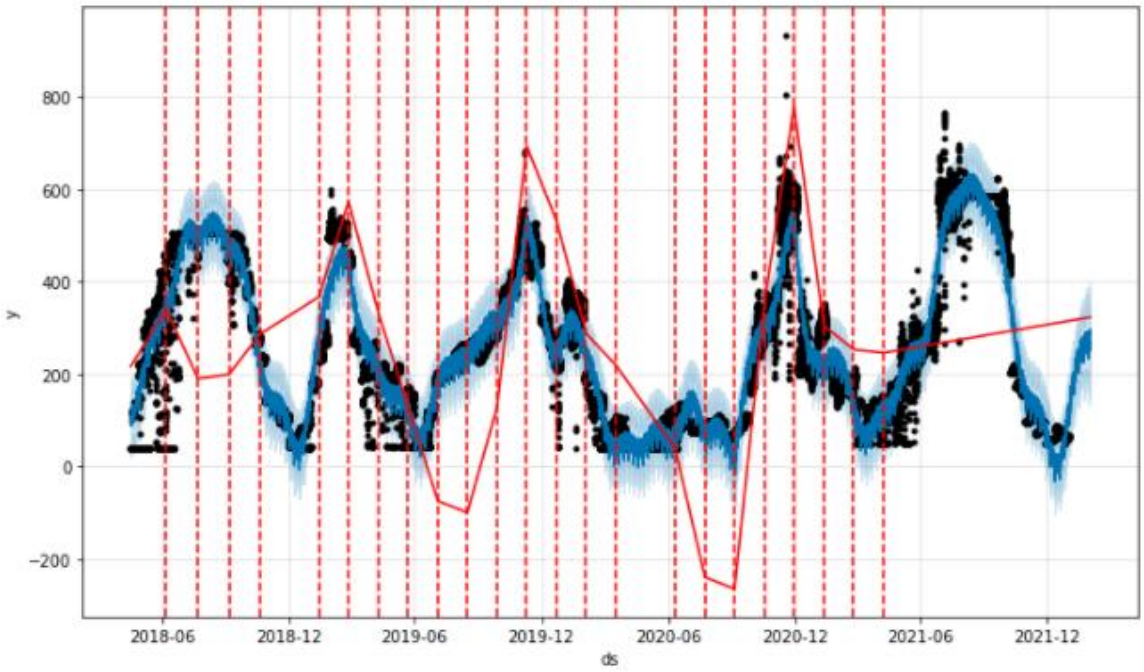
Gráfico 3 – Componentes de Previsão do PLD horário no Sudeste/CO

Fonte: elaborado pela autora

Com o propósito de reafirmar esta característica sazonal demonstrada pelos preços, vale ressaltar a existência de inúmeros *change points* que representam variações abruptas nos dados da série temporal, conforme Gráficos 4 e 5.

Os pontos de mudança são apontados no modelo como “n_changepoints”. Há um número de *changepoints* colocados automaticamente, ou seja, um padrão de 25 que é suficiente para capturar as mudanças de tendênia em uma série temporal típica.

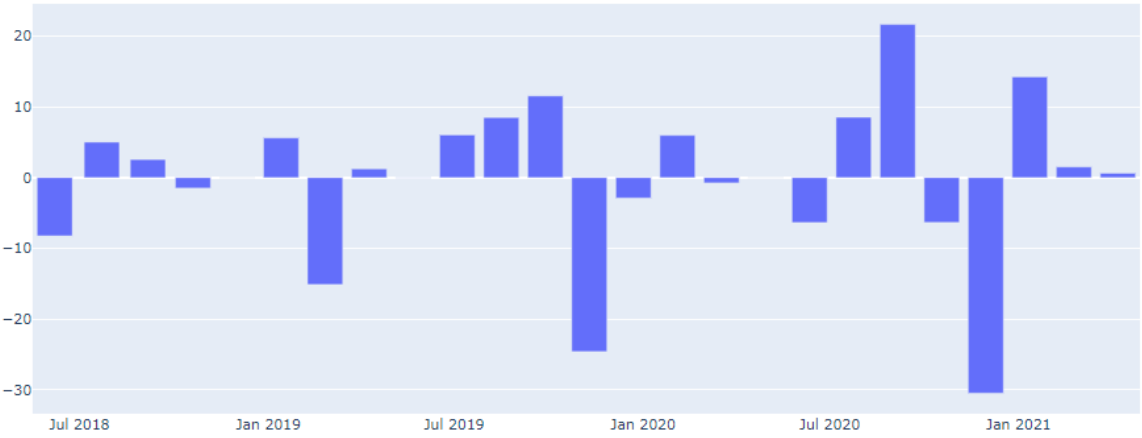
Gráfico 4 – Pontos de Mudança na Previsão do PLD horário no Sudeste/CO



Fonte: elaborado pela autora

Para tornar mais nítido, é possível visualizar por meio do Gráfico 5 todos os pontos de mudança na previsão:

Gráfico 5 – Todos pontos de mudança na previsão do PLD horário no Sudeste/CO



Fonte: elaborado pela autora

Segundo os boletins semanais divulgados pela CCEE, as intensas variações do PLD estão atreladas, entre outros fatores, à previsão de afluências no Sistema Interligado Nacional – SIN, que corresponde à estimativa do volume de água que deverá chegar aos reservatórios.

O modelo Prophet inclui ainda a funcionalidade de validação cruzada para medir o erro de previsão para dados históricos.

Isto é feito selecionando pontos de corte (*cutoff points*) no histórico temporal e, para cada um deles, o modelo é adaptado utilizando-se de dados até esse ponto de corte. Dessa forma, é possível comparar os valores previstos com os valores reais.

Tabela 1 – Validação cruzada série PLD no submercado Sudeste/CO

	ds	yhat	yhat_lower	yhat_upper	y	cutoff
0	2019-05-27 00:00:00:00	78,184399	17,856456	140,780740	42,35	2019-05-26 23:00:00:00
1	2019-05-27 01:00:00:00	68,556922	7,732578	134,140638	42,35	2019-05-26 23:00:00:00
2	2019-05-27 02:00:00:00	63,323538	8,371856	123,695759	42,35	2019-05-26 23:00:00:00
3	2019-05-27 03:00:00:00	63,670573	- 0,008064	125,574832	42,35	2019-05-26 23:00:00:00
4	2019-05-27 04:00:00:00	68,370788	7,533528	131,916490	42,35	2019-05-26 23:00:00:00

Fonte: elaborado pela autora

O resultado do *cross_validation* representado pelo *dataframe* acima indica valores verdadeiros(y) e os valores de previsão fora de amostra (`yhat`), em cada data de previsão simulada e para cada data de corte. Em particular, foi feita uma previsão para cada ponto observado entre o corte e o horizonte de corte (100 dias). Este *dataframe* pode então ser usado para calcular as medidas de erro de `yhat` vs y.

Buscando alcançar métricas de desempenho, o utilitário `performance_metrics` calculou algumas estatísticas úteis do desempenho de previsão (`yhat, yhat`_lower, e `yhat`_upper em comparação com y) em função da distância do corte (a que distância no futuro estava a previsão).

As estatísticas calculadas na Tabela 2 são: erro quadrático médio (MSE); raiz do erro quadrático médio (RMSE); erro absoluto médio (MAE); erro percentual absoluto médio (MAPE); erro percentual absoluto médio (MDAPE) e cobertura das estimativas `yhat`_lower e `yhat`_upper.

Tabela 2 –Métricas de desempenho série PLD no submercado Sudeste/CO

	horizon	mse	rmse	mae	mape	madpe	coverage
0	10 days 00:00:00	6.578,20	81,11	52,81	0,37	0,16	0,73
1	10 days 01:00:00	6.625,17	81,40	53,01	0,37	0,16	0,73
2	10 days 02:00:00	6.671,64	81,68	53,19	0,37	0,16	0,73
3	10 days 03:00:00	6.718,55	81,97	53,39	0,37	0,16	0,73
4	10 days 04:00:00	6.765,80	82,25	53,58	0,38	0,16	0,72

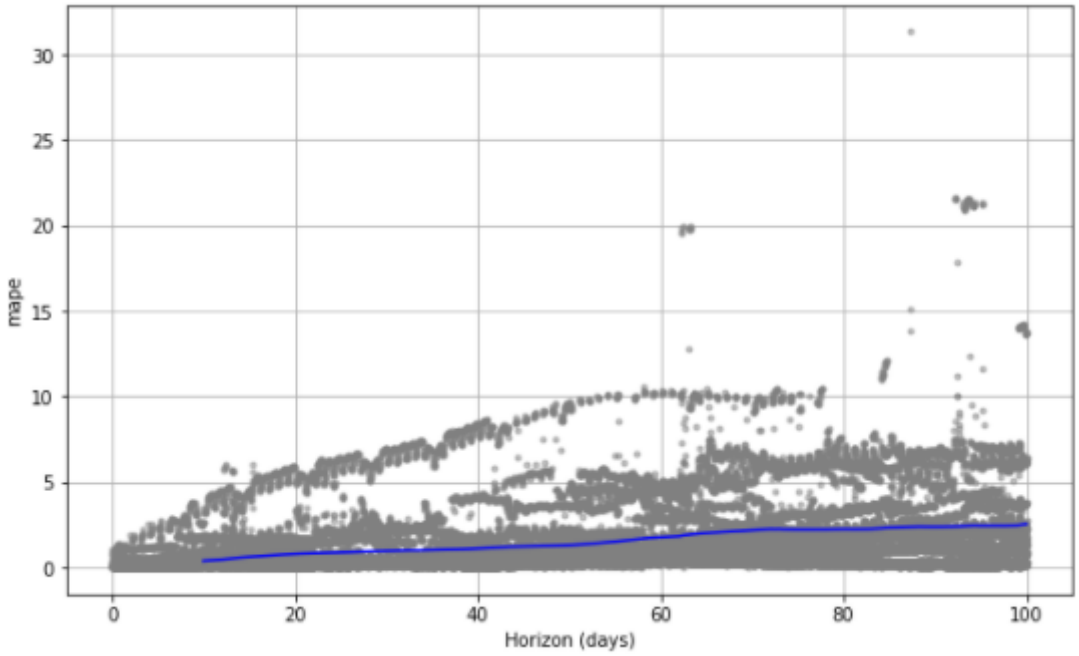
Fonte: elaborado pela autora

Métricas de desempenho de validação cruzada podem ser visualizadas com `plot_cross_validation_metric`, aqui mostrado para MAPE.

Os pontos mostram o erro percentual absoluto para cada previsão no *dataframe* da validação cruzada. A linha azul no Gráfico 6 mostra o MAPE, onde a média é tomada sobre uma janela rolante dos pontos.

Observamos para esta previsão erros em torno de 5% típicos para previsões de um mês no futuro, e que os erros aumentam até cerca de 11% para previsões que são maiores que um ano.

Gráfico 6 – Visualização métricas de desempenho MAPE cross validation



Fonte: elaborado pela autora

4.3 Previsões

Foram realizadas as previsões em um *dataframe* com uma coluna '`ds`' contendo as datas e horas para as quais uma previsão deve ser feita. Com isso, foi possível obter um *dataframe* adequado, que se estende para o futuro, mostrando um número específico de dias e horas, a partir de um método auxiliar, qual seja: `Prophet.make_future_dataframe`. Por padrão, o *dataframe* incluiu as datas e horas do histórico e, dessa forma, viu-se o modelo se ajustando bem.

Foi realizada uma previsão de 744 horas o que correponde a um período de 31 dias, limitando-se ao submercado Sudeste/CO, conforme demonstrado na Figura 7.

Figura 7 – Previsão de 744 horas (31 dias)– submercado Sudeste/CO

	ds
0	2018-04-17 00:00:00
1	2018-04-17 01:00:00
2	2018-04-17 02:00:00
3	2018-04-17 03:00:00
4	2018-04-17 04:00:00
...	...
32923	2022-01-31 19:00:00

Fonte: elaborado pela autora

Por meio do método `predict,` foi atribuída a cada linha no futuro um valor predito nomeado por '`yhat`', significando que, ao ler as datas históricas, o modelo forneceu um ajuste na amostra.

Assim sendo, o objeto de previsão *(forecast*) resultou em novos *dataframes*, em que mateve-se a coluna `ds` (datestamp no formato YYYY-MM-DD HH:MM:SS) e incluiu-se a coluna '`yhat`' com a previsão, bem como colunas para componentes e intervalos de incerteza (' `yhat`_lower' e '`yhat`_upper'), conforme Figura 8.

Figura 8 – *Dataframe* forescast – submercado Sudeste/CO

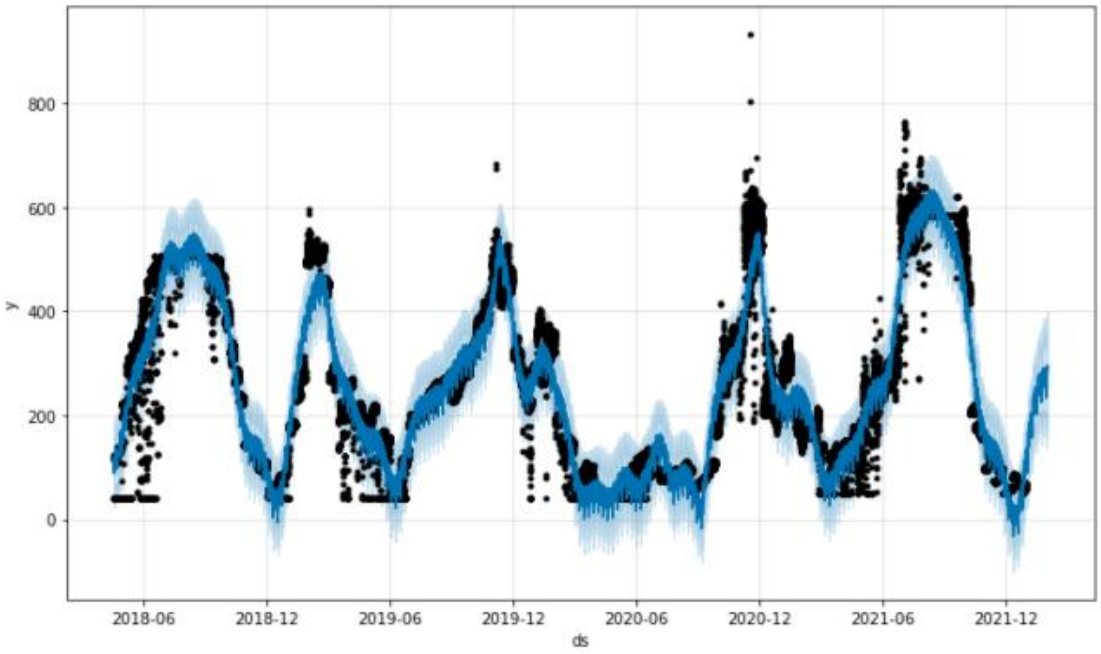
	ds	yhat	yhat_lower	yhat_upper
20	2018-04-17 20:00:00	121.126361	56.847194	189.655412
21	2018-04-17 21:00:00	120.224822	56.025850	184.060267
22	2018-04-17 22:00:00	116.616998	52.958113	181.979630
23	2018-04-17 23:00:00	110.185301	48.656667	171.572875
24	2018-04-18 00:00:00	102.337366	33.781912	165.728542
...	...	...	...	...
32923	2022-01-31 19:00:00	295.516912	198.982959	392.522663
32924	2022-01-31 20:00:00	296.217502	203.995563	391.880411
32925	2022-01-31 21:00:00	295.375231	199.026368	395.453013
32926	2022-01-31 22:00:00	291.733005	194.923581	387.425185
32927	2022-01-31 23:00:00	285.177467	195.802799	383.479893

Fonte: elaborado pela autora

A representação gráfica do *forecast* gerada por meio do método `Prophet.plot,` considerando o *dataframe* de previsão, pode ser visualizada conforme Gráficos 7, 8 e 9. Estas imagens apresentam a série temporal do PLD real nos pontos pretos e os pontos azuis

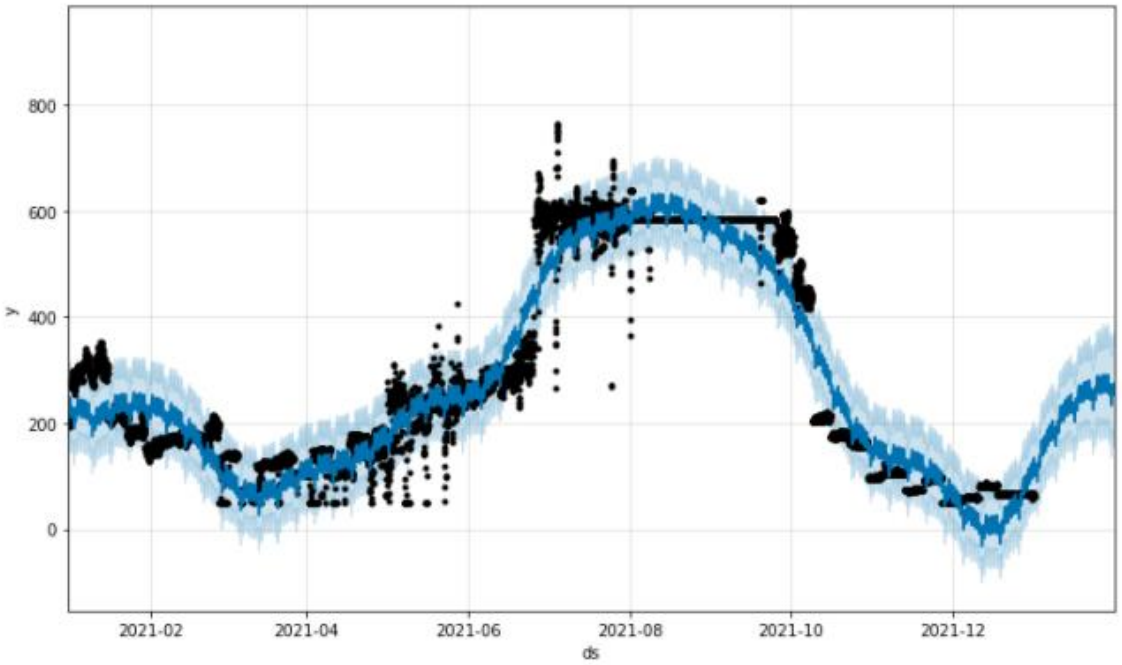
representam a previsão do modelo (*forecast).* As faixas em azul claro representam os limites inferior e superior do PLD.

Gráfico 7 – Previsão 744 horas a frente PLD horário no Sudeste/CO



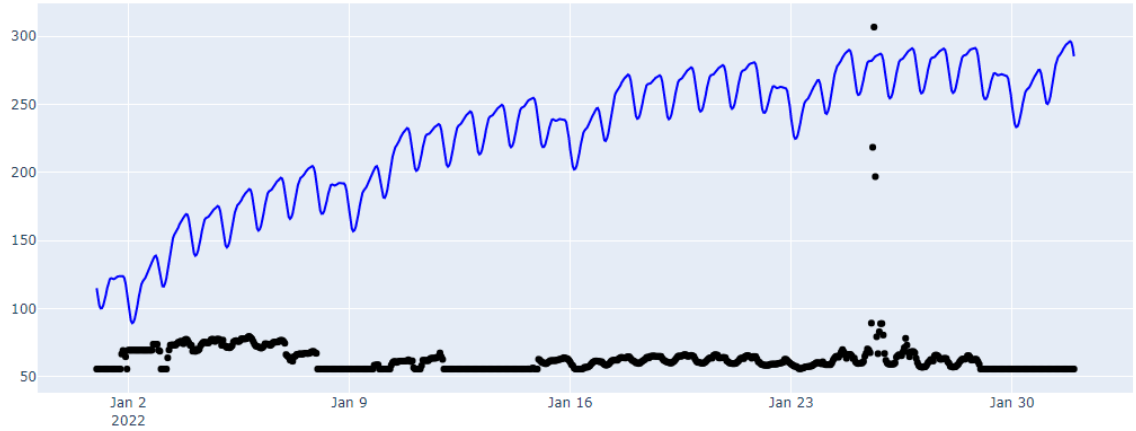
Fonte: elaborado pela autora

Gráfico 8 – Previsão final 744 horas a frente PLD horário no Sudeste/CO



Fonte: elaborado pela autora

Gráfico 9 – Comparativo dados de teste e previsão PLD horário no Sudeste/CO



Fonte: elaborado pela autora

As imagens indicam que o modelo se ajusta relativamente bem à série original, demonstrando um aumento de faixas limítrofes para um horizonte de 744 horas. É importante mencionar, que antes de apresentar as análises acima mencionadas, foram feitos exercícios e experiências alterando o critério de dados de treino e de teste, previsão de horas (`método predict`) e horizonte temporal da validação cruzada.

A Tabela 3 informa o cálculo do erro entre os dados de teste (série real período de 01 a 31/01/2022) e as previsões (744hs = 31 dias a frente dos dados de treino), demonstrando que os valores ficaram próximos dos PLDs mais recentes, tendo como escala a mesma unidade dos preços (R$/KWh). Isto significa que os dados previstos apresentam os erros MAE = 164,22 e RMSE = 172,23 para mais ou para menos comparados aos valores reais.

Tabela 3 – Cálculo do erro dados de teste e previsão PLD horário no Sudeste/CO

MAE	RMSE
164,22	172,23

Fonte: elaborado pela autora

5. CONCLUSÃO

O estudo da trajetória comportamental do PLD e a aplicação do modelo de AM apresentaram resultados satisfatórios para períodos de curto prazo (30 dias). À medida que ocorre o aumento do horizonte temporal da previsão (mais horas), maior é a medida do erro e pior o desempenho do modelo.

Conforme observado na seção anterior, mais especificamente na Figura 8, verificou-se que o modelo previu o PLD próximo do real e dentro das margens de limites máximo e mínimo para a validação cruzada.

Após realizar o c*ross-validation* do modelo de previsão, nota-se que quanto maior o número de dias previstos, maior será a dispersão dos dados, aumentando o erro para quase 10%, quando considerado períodos acima de 100 dias.

O PLD é uma variável que se altera em função de vários fatores, tais como a energia natural afluente, o volume de produção das usinas hidrelétricas, as condições climáticas (quanto mais chuva, maior o volume de água nas usinas), a demanda de energia pelos consumidores, o preço do combustível, a disponibilidade de geração e transmissão de energia, entre outros.

Apesar de altamente volátil e impactado por diversos fatores, o estudo em tese demonstrou ser possível a utilização da técnica de AM adotada, bem como de análise de dados para orientar a previsão do PLD em base horária.

No entanto, fica clara a necessidade de realizar análises mais amplas, correlacionando outras variáveis e outros modelos que sejam capazes de trazer mais assertividade e segurança para os tomadores de decisões.

No que tange a trabalhos futuros, o propósito de encontrar soluções e explicações para que o modelo funcione de forma a atingir uma previsão mais precisa do PLD em horizontes temporais mais longos, seria é viável buscar compreender e explorar as variáveis externas que são utilizadas nos modelos DECOMP e NEWAVE notadamente correlacionadas ao PLD.

Nesse sentido, ou seja, por meio do reconhecimento das varíáveis que impactam de maneira significativa o preço da energia, seria interessante lançar mão de funcionalidades do Prophet, como a a função `Prophet.add_regressors,` que captam e modelam fatores externos que possuem efeito sobre o valor alvo.

Além disso, a paralelização é um outro método a ser explorado. É possível fazer combinações de parâmetros paralelizando o loop. O modelo Prophet tem ainda uma série de parâmetros de entrada que podem ser considerados para aperfeiçoamento.

Concluímos que a problemática levantada para este estudo pode ser solucionada de maneira eficaz no curto prazo, apontando o Prophet como uma ferramenta poderosa para fazer previsões e validações dentro de curto horizonte temporal.

REFERÊNCIAS

Alencar, Victor., Pessamilio, Lucas., Rooke, Felipe., Bernardino, Heder. E Vieira, Alex (2020) “Predicão de Séries Temporais de Demanda em Modelos de Compartilhamento de Veículos para Modelos Uni e Multi Variaveis”

Bianchi, Matheus Gabriel (2020) “Impacto da geração solar fotovoltaica no preço de liquidação das diferenças em base horária”

Bouzada, Marco (2012) “Aprendendo Decomposição Clássica: Tutorial para um Método de Análise de Séries Temporais”

CCEE, Câmera de Comercialização de Energia Elétrica. Preços médios. Disponível em: https://www.ccee.org.br/

Clímaco, F. (2010) “Gestão de consumidores livres de energia elétrica”. São Paulo: Universidade de São Paulo

Garcia, Vinícius (2021) “Séries Temporais para Predição de Finanças no contexto de Criptomoedas”

Gerón, Aurélien. (2019) Hands-on machine learning with Scikit-Learn, Keras, and Tensor Flow: Concepts, tools, and techniques to build intelligent systems.

Giron, Raphael e Pedrini, Helio (2019) “Gerenciador de modelos para séries temporais univariadas”

Lagasse, William (2020) “Previsão do comportamento do preço de liquidação das diferenças (PLD) com ferramentas estatísticas”.

Leite, A.L.S. et AL (2013) “Causas da volatilidade do preço spot de eletricidade no Brasil”. Ensaios FEE p 647-668. Porto Alegre, 2013.

Lyla, Y (2019) “Um Início Rápido da Previsão de Séries Temporais com um Exemplo Prático usando o FB Prophet”.

Makridakis, S., & Wheelwright, S. (1982) “The handbook of forecasting: a manager’s guide”. New York: John Wiley & Sons, Inc.

Mehta, R.(2017) Big Data Analytics with Java. Birmingham: Packt Publishing Ltd.

Mitchell, Thomas M. (1997), Machine Learning, McGraw-Hill, Inc., New York, NY, USA

Monlevade, João. (2018) “ Previsão do Preço de Liquidação das diferenças por meio de redes neurais artificiais”

Morretin, P. A. e Toloi C.M.C (2006) “Análise de séries temporais”. São Paulo: Edgard Blucher.

Olivi, Leonardo Rocha, e Luís Henrique Lopes Lima (2018) “Redes neurais artificiais aplicadas à predição do Preço de Liquidação das Diferenças no mercado de energia elétrica”.

Papacharalampous, G., Tyralis, H., and Koutsoyiannis, D. (2018) “Predictability of monthly temperature and precipitation using automatic time series forecasting methods”. Acta Geophysica, 66(4):807–831.

Ramallo, Priscila, e Eduarda A Antoniolli (2020) “Análise Estatística do Histórico de Valores do Preço da Liquidação das Diferenças (PLD) no Mercado Livre de Energia”,

Reis, Marcelo (2021) “Cap 4 Análise de Séries Temporais” INE 7001 Análise de Séries Temporais.

Samal, K., Babu, K. S., Das, S. K., and Acharaya, A. (2019) ” Time series based air pollution forecasting using sarima and prophet model”. In Proceedings of the 2019 International Conference on Information Technology and Computer Communications.

Samuel, A. L (1959) “Some Studies in Machine Learning Using the Game of Checkers”. IBM Journal of Research and Development, v. 3, n. 3, p. 210–229, 1959. Disponível em: https://ieeexplore.ieee.org/document/5392560/authors .

Satrio, C. B. A. et al. (2020) “Time series analysis and forecasting of coronavirus disease in Indonesia using ARIMA model and PROPHET “. 5th International Conference on Computer Science and Computational Intelligence.

Silver, M. (2000) “Estatística para administração”. São Paulo: Atlas.

Souza, G., Samohyl, R., & Meurer, R. (2004) “Previsão do consumo de energia elétrica do setor industrial em Santa Catarina” – um estudo comparativo entre diferentes métodos de previsão através de suas discrepâncias. Anais do Simpósio Brasileiro de Pesquisa Operacional, São João Del Rey, MG, Brasil, 36.

Souza, R. (1989) “Modelos estruturais para previsão de séries temporais: abordagens clássica e bayesiana” Anais do Colóquio Brasileiro de Matemática, Rio de Janeiro, RJ, Brasil.

Taylor, S. J. and Letham, B. (2018) “Forecasting at scale”. The American Statistician.

Zhang, G. P. (2007) “Avoiding Pitfalls in Neural Network Research” IEEE Transactions on systems, mas, and cybernetics – part c: Applications ande Reviews.