Relatório Técnico

Resumo

O objetivo do projeto é prever os dados de geração de energia mensal nacional para os anos de 2019 e 2020. Para realizar esta tarefa, iremos obter dados horários de geração de energia disponibilizados pela ONS. Em seguida, treinaremos três modelos: regressão polinomial, SARIMA e XGBoostRegressor. Compararemos os resultados utilizando as métricas de MAE (Erro Absoluto Médio) e MAPE (Erro Percentual Absoluto Médio), que nos permitem avaliar a precisão das previsões em termos de diferenças absolutas e percentuais entre os valores previstos e reais. Por fim, apresentaremos as conclusões da pesquisa e discutiremos possíveis aplicações práticas.

EDA (Análise Exploratória dos Dados)

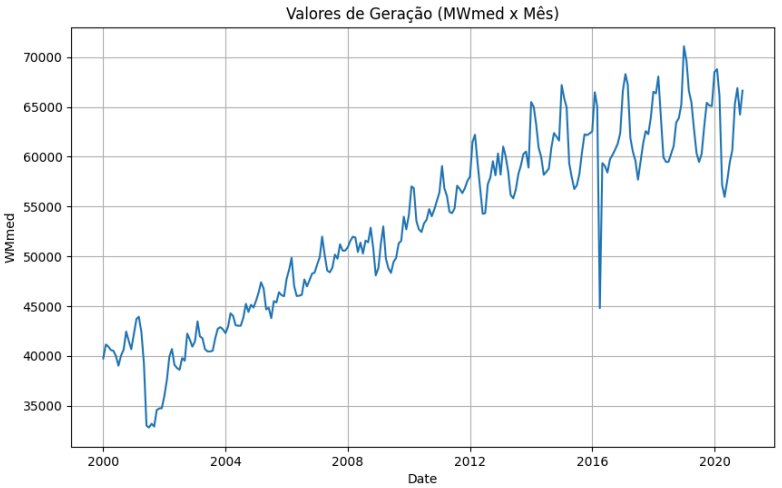
Inicialmente, realizamos o download dos dados disponíveis e visualizamos as colunas, tipos de dados e informações disponíveis. A imagem abaixo corresponde às primeiras linhas do dataframe do ano de 2000.

	din_instante	id_subsistema	nom_subsistema	id_estado	nom_estado	cod_modalidadeoperacao	nom_tipousina	nom_tipocombustivel	nom_usina	ceg	val_geracao
0	2000-01-01 00:00:00	N	NORTE	PA	PARA	TIPO I	HIDROELÉTRICA	Hidráulica	Tucuruí	UHE.PH.PA.002889-4.01	2422.5
1	2000-01-01 00:00:00	NE	NORDESTE	AL	ALAGOAS	TIPO I	HIDROELÉTRICA	Hidráulica	Xingó	UHE.PH.SE.027053-9.01	1995.6
2	2000-01-01 00:00:00	NE	NORDESTE	BA	BAHIA	Pequenas Usinas (Tipo III)	HIDROELÉTRICA	Hidráulica	Pequenas Centrais Hidroelétricas da Chesf	-	13.0
3	2000-01-01 00:00:00	NE	NORDESTE	BA	BAHIA	TIPO I	HIDROELÉTRICA	Hidráulica	Apolônio Sales	UHE.PH.AL.001510-5.01	202.5
4	2000-01-01 00:00:00	NE	NORDESTE	BA	BAHIA	TIPO I	HIDROELÉTRICA	Hidráulica	Paulo Afonso II	UHE.PH.BA.027048-2.01	269.8

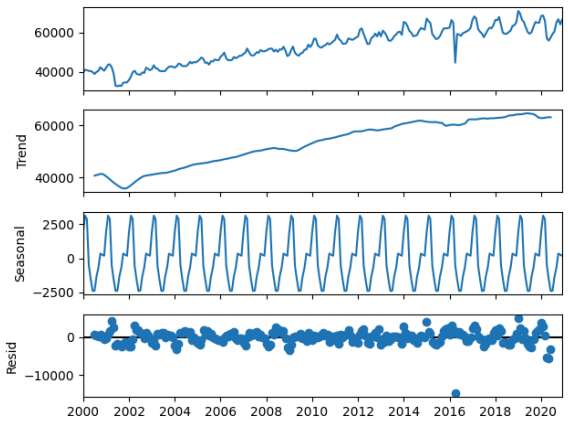
Como primeira etapa, devemos identificar qual é o nosso atributo alvo, neste caso a coluna *'val_geracao'*, e então remover as demais colunas que não interferem na previsão. Como estamos prevendo a produção geral, sem considerar as diferenças de estado, região ou tipo de usina, removeremos as colunas que não são relevantes para a previsão. Além disso, os dados serão transformados da escala de produção horária para a escala de produção mensal. Inicialmente, somamos a geração média de todas as usinas a cada hora e, por fim, calculamos a média dessa produção ao longo do mês. Posteriormente, concatenamos os dados dos anos de 2000 a 2020. As linhas iniciais do conjunto de dados gerado estão disponíveis abaixo, assim como dados estatísticos, como valores mínimos, máximos, média e quartis dos dados de geração ao longo do período, e um gráfico mostrando a geração de energia ao longo desse período.

	mes	val_geracao
0	2000-01	39748.958266
1	2000-02	41143.642385
2	2000-03	40954.942706
3	2000-04	40590.800694
4	2000-05	40521.976882
5	2000-06	39985.127500
6	2000-07	39018.919798
7	2000-08	40047.184274
8	2000-09	40640.307375
9	2000-10	42452.584185
10	2000-11	41505.840445
11	2000-12	40671.505165

	val_geracao
count	252.000000
mean	52948.963851
std	9072.680775
min	32808.255632
25%	45210.088353
50%	54002.991239
75%	60289.495970
max	71101.798364

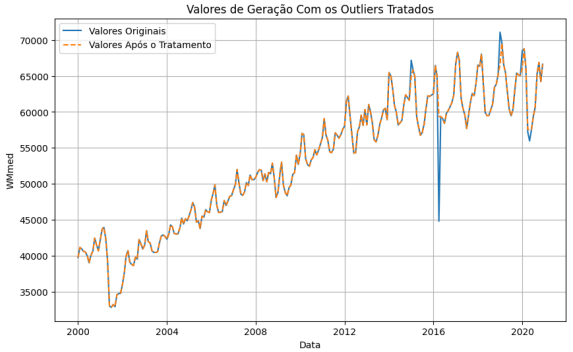


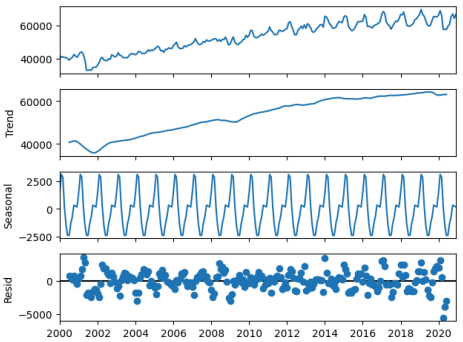
Através do gráfico, destacam-se algumas peculiaridades. Notavelmente, a presença de outliers que precisarão ser tratados, bem como a observação de uma tendência de crescimento no modelo, além de uma sazonalidade anual, demonstrada no gráfico abaixo



O gráfico abaixo evidencia os comportamentos de sazonalidade e tendência dos dados, além de ressaltar a presença de outliers. Utilizaremos o método de Tukey, aplicado aos dados do gráfico de resíduos, para identificar as amostras que distanciam-se além de 1.5 interquartis da média. Substituímos esses valores pela mediana dos valores vizinhos, considerando uma janela de 5 amostras.

Um gráfico da nova série temporal, a que nos referimos como "dados tratados", está disponível abaixo, juntamente com uma nova análise de tendência e sazonalidade.





Métricas

É uma prática recomendada que as métricas escolhidas sejam intuitivas para permitir uma compreensão do desempenho do modelo, além de serem amplamente utilizadas para facilitar a comparação entre diferentes modelos. Portanto, as métricas selecionadas para este trabalho foram o MAE (Erro Absoluto Médio) e o MAPE (Erro Percentual Absoluto Médio).

O MAE é uma métrica intuitiva, pois está na mesma ordem de grandeza dos valores observados. Por outro lado, o MAPE é a versão percentual do MAE e é ainda mais intuitivo, pois trabalha com valores normalizados, facilitando a interpretação. Ambas as métricas são amplamente utilizadas na avaliação de modelos de previsão.

Modelo

Utilizamos três modelos e comparamos seus resultados ao final. São eles: Regressão Polinomial, SARIMAX e XGBoostRegression.

Regressão Polinomial

A regressão polinomial é um modelo simples, facilmente implementado, utilizado como *baseline* neste trabalho. Ele serve como ponto de referência para comparar modelos mais avançados. Isso nos permite determinar se a complexidade adicional dos modelos mais avançados vale a pena em comparação com a melhoria nos resultados. Com base em testes empíricos, foi escolhido um polinômio de grau 4

SARIMA

O modelo SARIMA é amplamente utilizado na área de análise de séries temporais, sendo sua sigla uma abreviação para "Seasonal Autoregressive Integrated Moving Average". Essa nomenclatura reflete suas principais características:

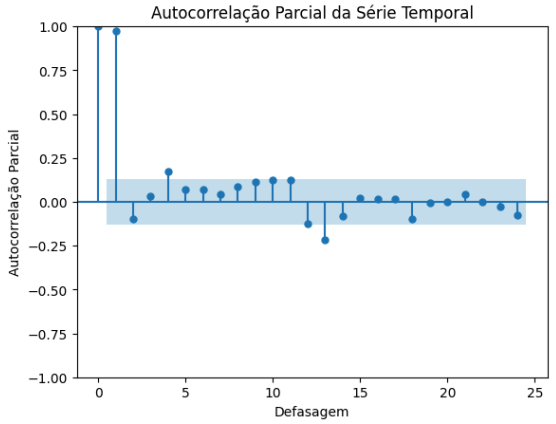
S: Indica a sazonalidade, ou seja, padrões que se repetem em intervalos regulares de tempo.
AR: Representa a parte autorregressiva do modelo, que utiliza as observações passadas para prever valores futuros.
I: Refere-se à integração, que envolve a diferenciação dos dados para tornar a série temporal estacionária.
MA: Indica a parte de média móvel, que leva em consideração os erros residuais das previsões passadas para melhorar as previsões futuras.

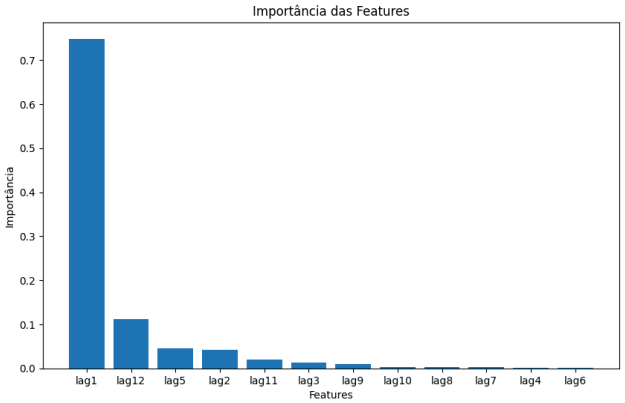
Dada a complexidade do modelo SARIMA e a presença de diversos hiperparâmetros, recorremos aos insights obtidos na análise prévia dos dados, como a identificação da sazonalidade anual e a aplicação de uma diferenciação de ordem 1 para tornar a série estacionária, e realizamos uma busca exaustiva para determinar os demais parâmetros do modelo.

XGBoostRegression

Por fim, o último modelo testado foi o XGBoostRegressor, um modelo de aprendizado de máquina baseado em árvores de decisão que demonstra alta eficácia em tarefas de regressão. Por não ter sido desenvolvido especificamente para lidar com séries temporais, os dados precisam passar por um pré-processamento. Nesse processo, são adicionadas colunas de lag-x, onde x representa o atraso dos dados. Observando o gráfico de correlação exibido abaixo, podemos notar que as relações significativas estão predominantemente nos primeiros 12 meses. Portanto, criamos colunas de lag para os 12 primeiros meses.

Ao final do treinamento, podemos observar a importância relativa atribuída a cada um dos atributos pelo modelo, conforme apresentado no gráfico abaixo:

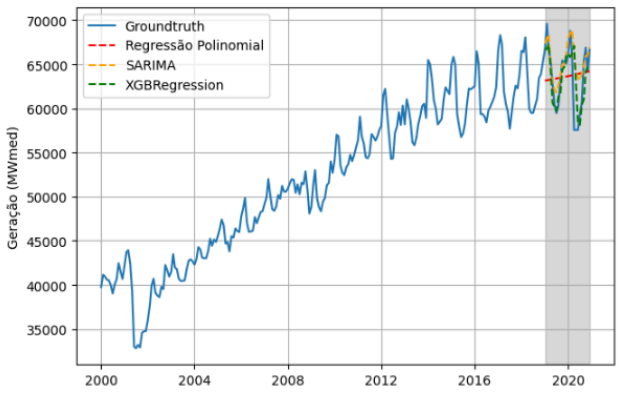


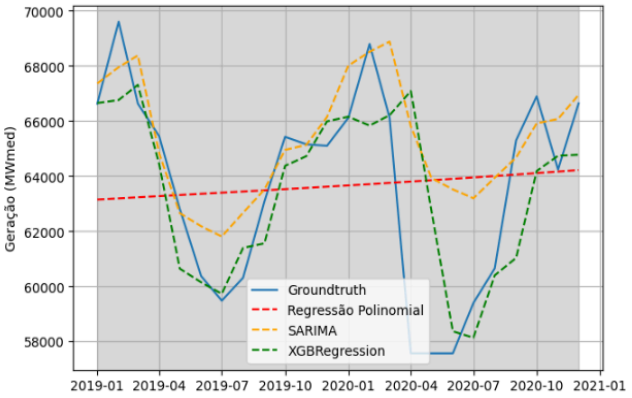


Comparações Finais

Por fim, os resultados obtidos estão listados abaixo:

Método	MAE	MAPE (%)
Regressão Polinomial	3109.10	4.97
SARIMA	2073	3.42
XGBoosterRegression	1735.57	2.79





Conclusões e Recomendações

Em última análise, o modelo XGBoostRegressor apresentou o melhor desempenho entre os testados, contudo, é importante ressaltar que o modelo SARIMA, apesar de mais simples, também obteve resultados satisfatórios.

As aplicações práticas deste trabalho resultaram em um modelo capaz de prever com boa precisão o comportamento futuro da geração de energia do país. Isso é crucial para previsões como o custo da energia, uma vez que em períodos de baixa produção, o custo tende a aumentar. Essa capacidade permite que indivíduos e empresas se preparem, considerando soluções alternativas. Além disso, é uma métrica valiosa para as instituições responsáveis no condicionamento dos sistemas de energia para as cargas previstas.

Como sugestão para pesquisas futuras, recomendo a repetição desses testes com a divisão dos dados por tipo de usina. Cada tipo de usina apresenta comportamentos específicos, como as usinas sustentáveis, que estão em constante crescimento, e todas possuem padrões cíclicos distintos. Dessa forma, com pequenas modificações nos dados, seria possível prever outras informações relevantes, como o aumento do uso de energia sustentável, os padrões cíclicos da energia solar e eólica, entre outros insights que podem ser úteis de diversas maneiras, como por exemplo no estabelecimento de metas internacionais para produção de energia limpa.