Relatório Técnico – Bootcamp de Dados

Objetivo:

Desenvolver um modelo de predição utilizando um conjunto de dados que detalha a geração horária de energia elétrica no Brasil, de 2000 a 2018.

Resumo:

Foram fornecidos inicialmente os dados de geração de energia elétrica no Brasil ao longo de 2000 a 2020, separados cada ano por um arquivo csv. Sendo que os anos de 2019 e 2020 foram separados para a validação do modelo.

Inicialmente, carregamos os dados e o juntamos em um único dataframe. Foram feitas as verificações do comportamento e descrições dos dados. Após isso fomos para o tratamento de dados onde lidamos com os dados faltantes e agrupamos os registros para que pudessemos fazer análises mensais. Inicialmente o total de registros eram de mais 40 milhões de linhas mas com o agrupamento se tornaram apenas por volta de 60 mil, o que poupou o processo computacional.

Depois disso fomos a análise exploratória de dados, estudamos o comportamento de nossa variável alvo(val_geracao), suas relações com as outras variáveis e por fim, a relação entre as variáveis preditivas.

Antes da modelagem com Machine Learning tivemos que preparar os dados para usarmos o algoritmo. Criamos novas variáveis e lidamos com variáveis categóricas.

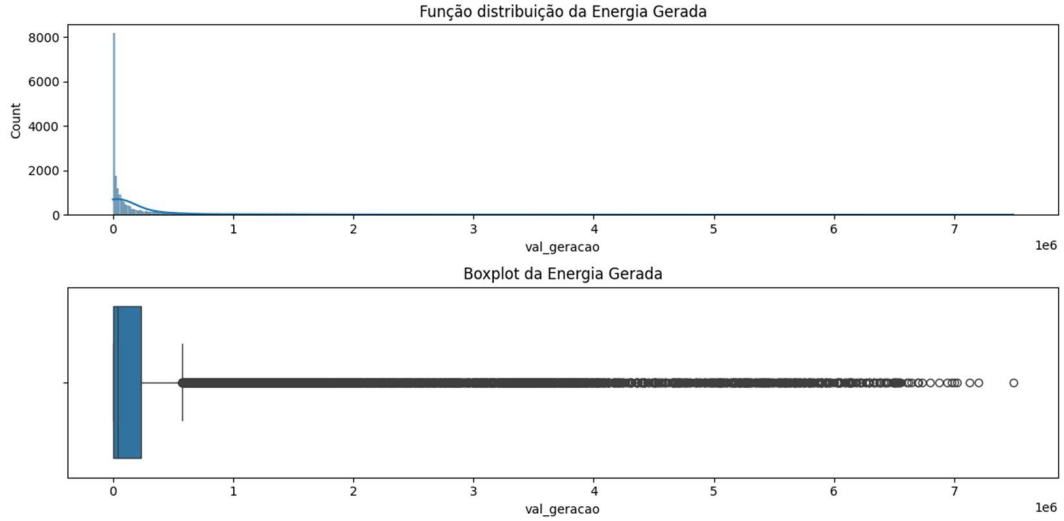
No estágio da modelagem de ML optamos por usar Regressão Linear, Random Forest e Xgboost por serem bons algoritmos e de simples execução.

Depois disso usamos as métricas MAE, MAPE e RMSE para medir nosso erro, e decidirmos qual modelo utilizar.

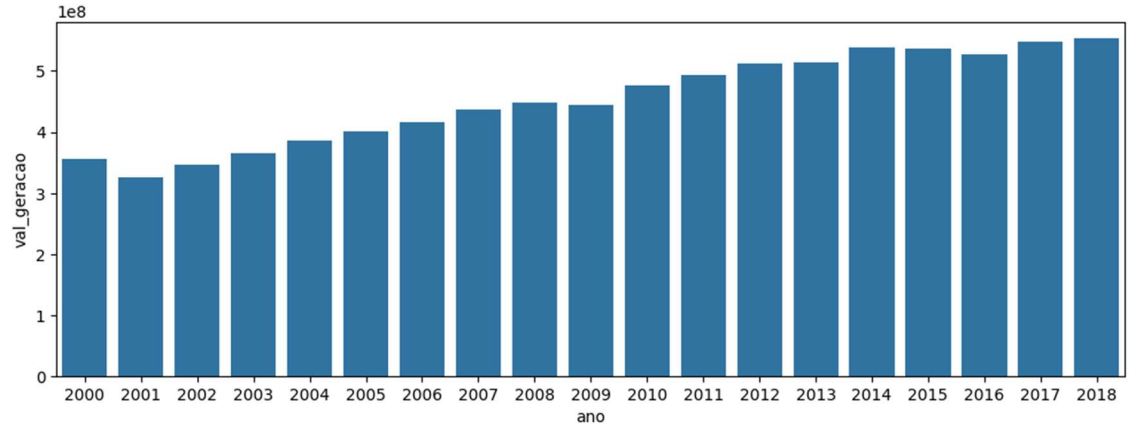
Por fim, comparamos graficamente as diferenças dos dados preditos com os valores reais.

Análise exploratória de Dados:

Pela análise da função de distribuição e pelo boxplot observamos que apresentava muitos outliers e assimetria acentuada



Pelo gráfico anual de geração de energia notamos um aumento gradativo ao longo dos anos.



Também pudemos observar que o Tipo I de modalidade operacional, subsistemas do Sudeste e usinas Hidroelétricas são os maiores geradores de energia no país.

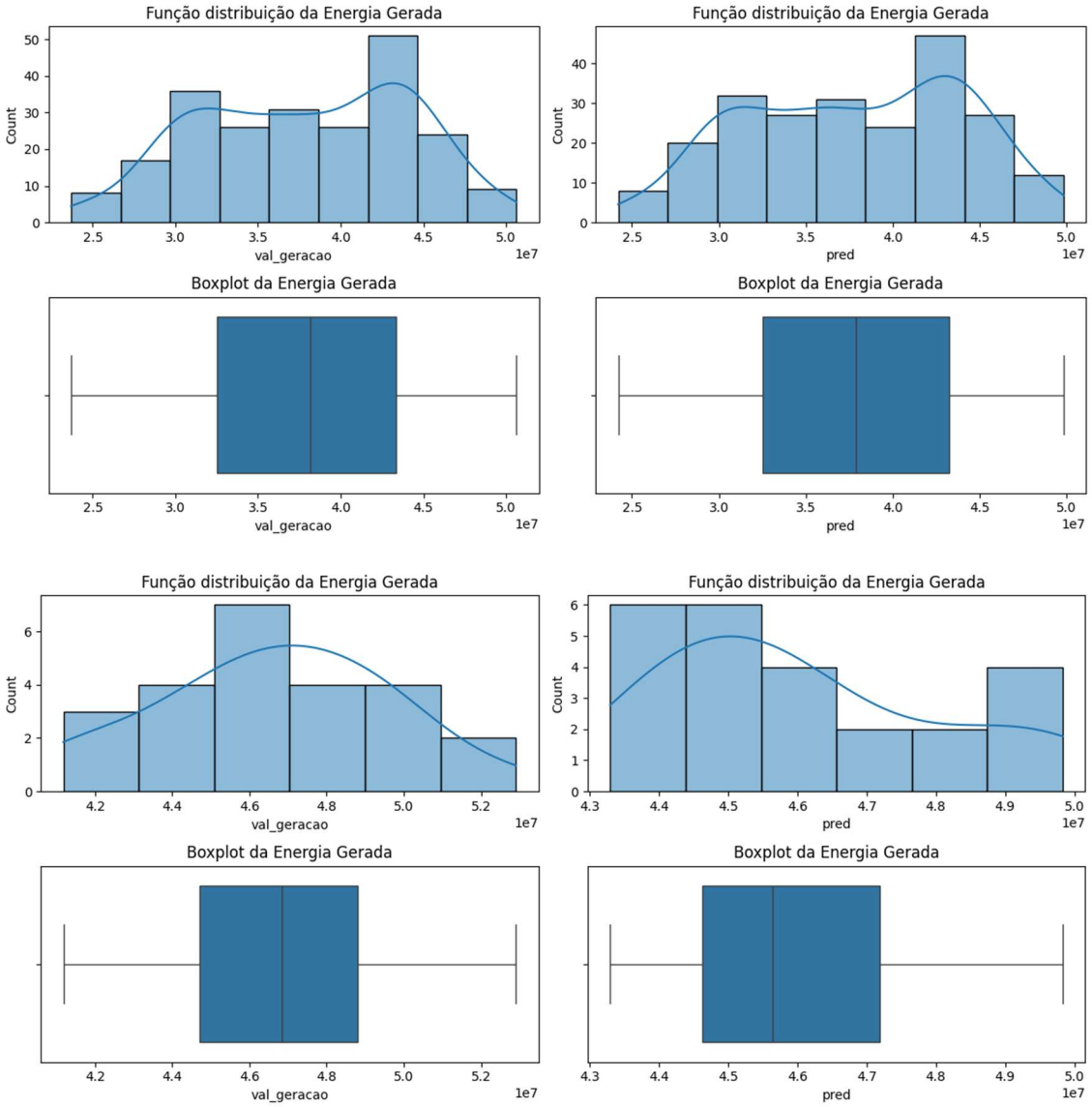
Modelos de ML:

	Model Name	MAE	MAPE	RMSE
0	random forest	1.430595e+06	0.030852	1.855479e+06
0	XGBoost Regressor	1.879300e+06	0.040161	2.322581e+06
0	Linear Regression	3.183868e+06	0.071597	4.090584e+06

A partir das métricas obtivemos a melhor opção a Random Forest.

Vale salientar que o modelo obtido com a melhor desempenho foi onde não foram consideradas nenhuma variável categórica, apenas o ano e o mês da geração de energia. Todo o dataframe foi agrupado em função do ano mês e a soma da energia gerada.

Um fato interessante é que observamos uma tendência mais uniforme na distribuição e sem outiliers quando agrupamos apenas pelo tempo.



Vemos acima a energia gerada real na esquerda e predito na direita, enquanto que a parte de cima é a base de treino e em baixo a base de validação.

Conclusão:

O trabalho pode ser considerado apenas como um esboço para uma pesquisa mais ampla, conseguimos obter informações importantes a partir da EDA e pela predição dos dados. Com certeza há muito para ser estudado ainda. No futuro podemos fazer uma análise mais aprofundada das variáveis, haviam muitos dados nulos ou 0s que poderíamos investigar. Além disso o uso de outras técnicas de Machine Learning para melhorarmos as predições.