Aplicações de Aprendizado de Máquina e Mineração de Dados (SCC0233)

Fabio Henrique Alves Cavaleti	11200550
Gabriel Akio Urakawa	11795912
Vinícius Santos Monteiro	11932463

LINKS ÚTEIS

Link para o Kaggle:
https://www.kaggle.com/competitions/predict-energy-behavior-of-prosumers/

Link para o Google Colab:
https://colab.research.google.com/drive/1CdEAulUQ9hvycc1KU9B2qmYyVsILOn89?authuser=1#scrollTo=BWOeQSCVe1AF

Link para o vídeo no YouTube:
https://www.youtube.com/watch?v=NLrM7Q5VIlA&ab_channel=GabrielUrakawa

1 - Entendimento do Negócio

Problema

Prosumidores são pessoas que são tanto consumidores quanto produtores de energia. O número de prosumidores nos Estados Unidos vem aumentando rapidamente e a energia produzida por eles começa a se tornar relevante no contexto de fornecimento de energia para a população. Entretanto, essa produção de energia não é consistente e confiável a ponto de as empresas de energia poder contar com elas em momentos de necessidade.

A tarefa é criar um modelo preditivo de energia que possa prever o consumo e a geração de energia de prosumidores, visando reduzir os custos associados ao desequilíbrio energético. Assim, o objetivo principal é melhorar o planejamento para empresas do setor, otimizando suas operações e evitando perdas.

Negócio

O impacto almejado é a redução dos custos para as empresas de energia, que enfrentam dificuldades logísticas e financeiras com os desequilíbrios de energia. Um modelo preditivo eficiente pode reduzir a necessidade de compensar esses desequilíbrios por meio da compra ou venda de energia a preços de última hora, que costumam ser mais altos. Além disso, melhora a estabilidade da rede e promove o uso de fontes produtoras de energia elétrica. Outro ponto importante levantado é que essa aplicação poderia potencialmente incentivar mais consumidores a se tornarem prosumidores, sabendo que seu comportamento energético pode ser gerenciado adequadamente, promovendo assim a produção e o uso de energia renovável.

Sucesso

O impacto do modelo seria medido diretamente pela alteração dos gastos com os desequilíbrios de energia após a implementação. Entretanto, temos que tomar cuidado com fatores naturais e provenientes dos dados como possíveis riscos para a solução.

Por se tratar de seres humanos e a natureza, fatores sociais e humanos podem atrapalhar o modelo por meio de fatores externos ao nosso dataset que seriam imprevisíveis. Além disso, um ponto de atenção é a mudança de padrões globais em uso de energia e de consumo próprio, por exemplo o encarecimento de

uma fonte de energia ou novas tecnologias que melhoram a eficiência de painéis solares.

Planejamento

Seguiremos a metodologia CRISP-DM, processo flexível e cíclico usado na indústria para mineração de dados que é dividida em 6 fases:

1. **Entendimento do Negócio**: Compreensão dos objetivos do negócio e como os dados podem ser usados para alcançar esses objetivos.
2. **Entendimento dos Dados**: Coleta e análise inicial dos dados disponíveis para entender sua qualidade, estrutura e implicações.
3. **Preparação dos Dados**: Limpeza, transformação e formatação dos dados, preparando-os para análise ou modelagem.
4. **Modelagem**: Aplicação de técnicas de mineração de dados ou algoritmos de machine learning para criar modelos preditivos.
5. **Avaliação**: Verificação dos resultados obtidos garantindo que o modelo atende aos nossos objetivos
6. **Implantação**: Implementação prática da solução, com um modelo preditivo e um relatório, gerando valor ao negócio em questão

2 - Entendimento dos Dados

Descrição dos dados

O dataset está dividido em alguns arquivos .csv. No arquivo de *train*, encontramos as informações gerais do problema. Em *gas_prices*, encontramos as informações relacionadas com o preço do gás natural usado para produção de energia. Em *client* há os dados relacionados clientes podendo ser prosumidor ou não. Em *electricity_prices* encontramos os dados relacionados com preços da eletricidade. Em *forecast_weather* vemos dados sobre as previsões do clima no momento da previsão. Por fim, temos *historical_weather* que contém os dados históricos meteorológicos.

3 - Preparação dos Dados

Uma análise prévia dos dados mostrou que a qualidade dos dados pode ser considerada ótima, com apenas dois campos com valores *NaN* no arquivo *forecast_weather.csv*. Por ser um arquivo extenso, com mais de três milhões de linhas, a abordagem utilizada foi remover as linhas para que não houvesse problemas futuros.

Após o pré-processamento dos dados, fizemos a junção dos dados para visualização e criação de gráficos para uma análise exploratória.Até o momento temos uma visualização estática que funciona no *colab* e um dashboard que funciona no jupyter, por exemplo.

Para a visualização, a nossa ideia foi variar as condições climáticas, latitudes e longitudes para podermos comparar, ao longo do tempo, possíveis relações entre condições climáticas e alterações no preço.

4 - Modelagem

Na fase de modelagem, planejamos utilizar modelos robustos de aprendizado de máquina que se destacam em problemas de previsão de séries temporais e são eficientes na captura de padrões complexos nos dados, como árvores de decisão e métodos baseados em ensemble. A escolha por esses modelos se justifica pela sua capacidade de lidar com múltiplas variáveis e por sua robustez em cenários de previsão energética, onde a variabilidade dos dados é alta devido a fatores externos, como o clima e o comportamento dos prosumidores.

Inicialmente, consideramos usar modelos como o Random Forest e o Gradient Boosting, que combinam várias árvores para melhorar a precisão preditiva. No entanto, estamos também explorando modelos mais específicos para séries

temporais e previsão energética, como o XGBoost e LightGBM, que conseguem lidar com grandes volumes de dados e são eficientes na execução.

Para lidar com o desafio de previsão simultânea de consumo e produção, estamos testando abordagens que incluem a criação de modelos separados para cada variável, bem como uma abordagem multitarefa, onde um único modelo prevê tanto o consumo quanto a geração de energia. Além disso, técnicas de otimização de hiperparâmetros estão sendo aplicadas para maximizar a performance dos modelos, incluindo o uso de ferramentas como Optuna.

5 - Implementação

Inicialmente, implementamos um pipeline com o objetivo de inferir valores distintos de consumo e produção energética. Para isso, optamos por dividir o dataset original em dois subconjuntos distintos com base na variável categórica is_consumption, separando os dados de consumo dos de produção. Essa abordagem buscava explorar as diferenças específicas entre os dois tipos de dados, permitindo treinar modelos independentes e especializados para cada categoria. Utilizamos algoritmos de aprendizado de máquina robustos e amplamente reconhecidos pela eficácia em tarefas preditivas, como Random Forest, XGBoost e LightGBM. Esses modelos foram aplicados separadamente aos subconjuntos, precedidos por uma etapa detalhada de preparação de dados que incluiu limpeza, engenharia de atributos e normalização. Além disso, utilizamos a biblioteca Optuna para realizar a otimização de hiperparâmetros, buscando minimizar métricas de erro como RMSE e MAE.

Nos testes realizados, observamos que o modelo LightGBM apresentou melhor desempenho no subconjunto de consumo, enquanto o XGBoost destacou-se na previsão de produção. A modularidade dessa abordagem permitia que o pipeline selecionasse automaticamente o modelo mais adequado, dependendo do tipo de dado, ao verificar se o dado se tratava de consumo ou produção energético. Essa separação tinha como objetivo não apenas melhorar a precisão dos resultados ao usar modelos especializados para padrões distintos, mas também facilitar a manutenção e escalabilidade do sistema.

No entanto, ao longo do processo, constatamos que a separação do dataset em dois fluxos distintos não trouxe os ganhos esperados em termos de eficiência preditiva e simplicidade operacional. Os resultados obtidos com essa abordagem foram insatisfatórios, o que nos levou a descartar essa ideia. Como alternativa, decidimos treinar o modelo utilizando consumo e produção no mesmo dataset, mantendo as informações combinadas. Essa mudança não apenas simplificou significativamente o pipeline, como também trouxe uma melhoria na performance

geral do sistema. A decisão de consolidar os dados demonstrou ser mais eficiente e eficaz, permitindo resultados mais consistentes e um processo de modelagem mais direto.

6 - Avaliação

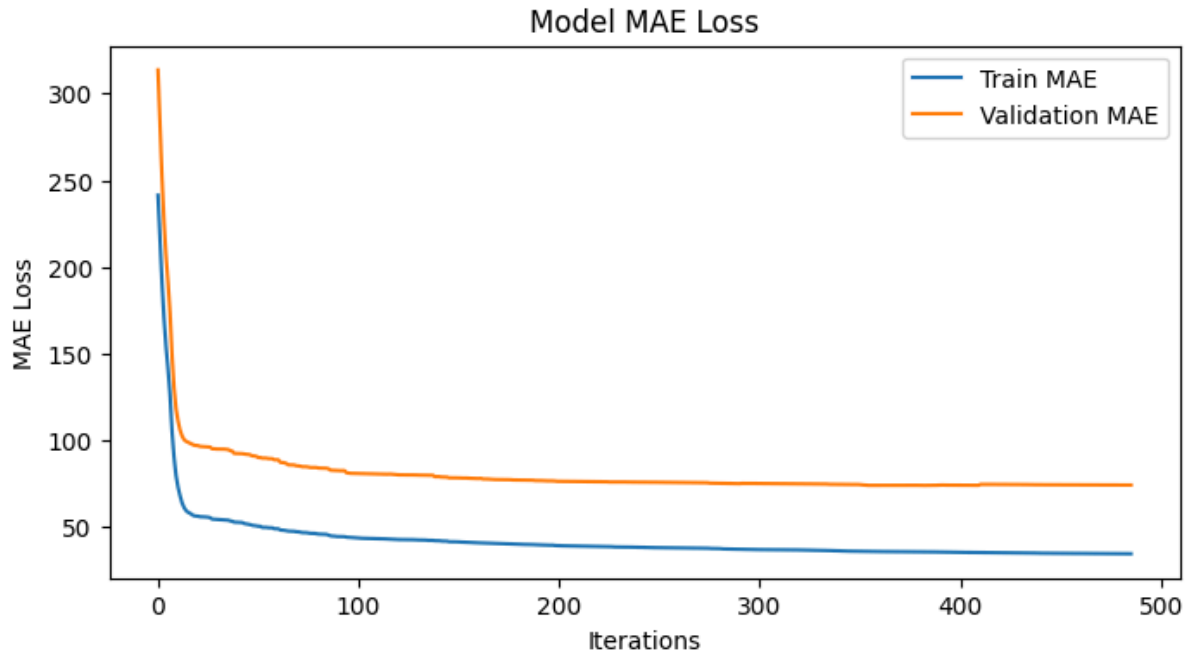

Figura 1. Validação MAE Loss do modelo

Decréscimo Inicial: Ambas as linhas (Train MAE e Validation MAE) apresentam um decréscimo significativo nas primeiras iterações. Isso indica que o modelo está aprendendo rapidamente e se ajustando aos dados de treinamento.

Estabilização: Após um certo número de iterações, ambas as linhas se estabilizam em um valor relativamente baixo. Isso sugere que o modelo convergiu, ou seja, encontrou um conjunto de parâmetros que minimizam a perda.

A distância entre as linhas Train MAE e Validation MAE pode indica um bom ajuste: Com ambas as linhas se estabilizarem em um valor baixo e relativamente próximos, isso indica que o modelo encontrou um bom ajuste aos dados, balanceando a capacidade de generalização e a capacidade de ajustar os dados de treinamento.

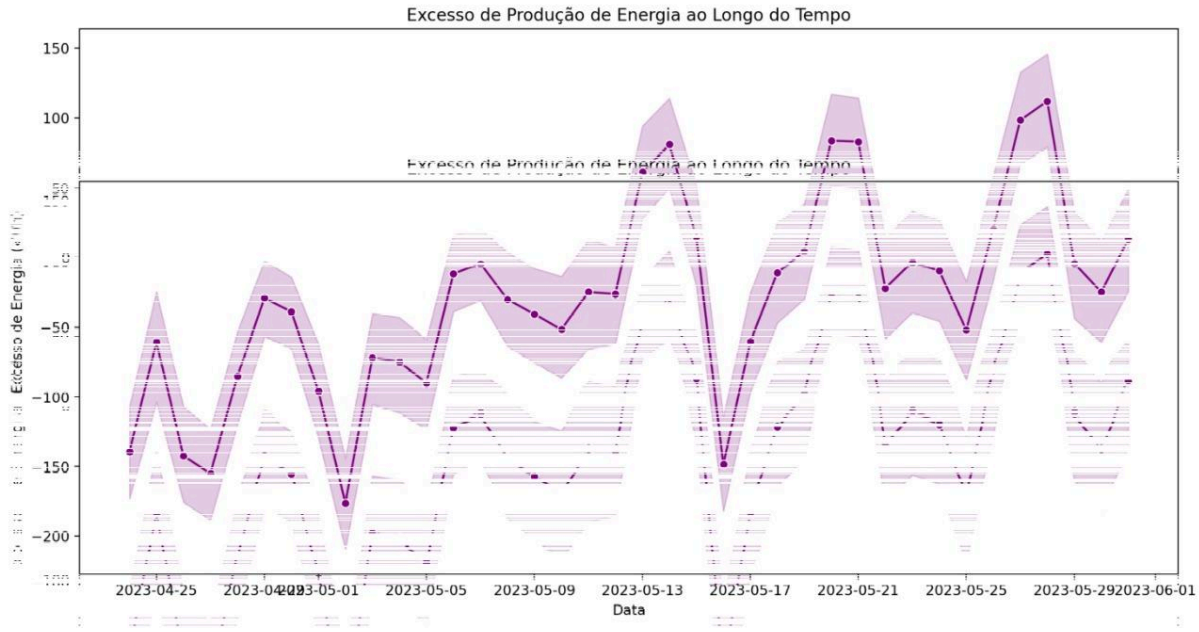

Figura 2. Excesso de produção

Variações significativas: O excesso de energia apresenta variações consideráveis ao longo do tempo, com picos positivos e negativos. Isso indica que a produção de energia, em alguns momentos, excede significativamente a demanda, enquanto em outros momentos há uma produção menor do que o necessário.

Tendência geral: A linha central, que representa a média do excesso de energia, parece oscilar em torno de um valor próximo a zero. Isso sugere que, em média, a produção de energia está equilibrada com a demanda, mas há flutuações consideráveis ao longo do tempo.

Intervalo de confiança: A área sombreada em torno da linha central representa um intervalo de confiança, indicando a incerteza associada às estimativas. Quanto maior a área sombreada, maior a incerteza sobre o valor real do excesso de energia.

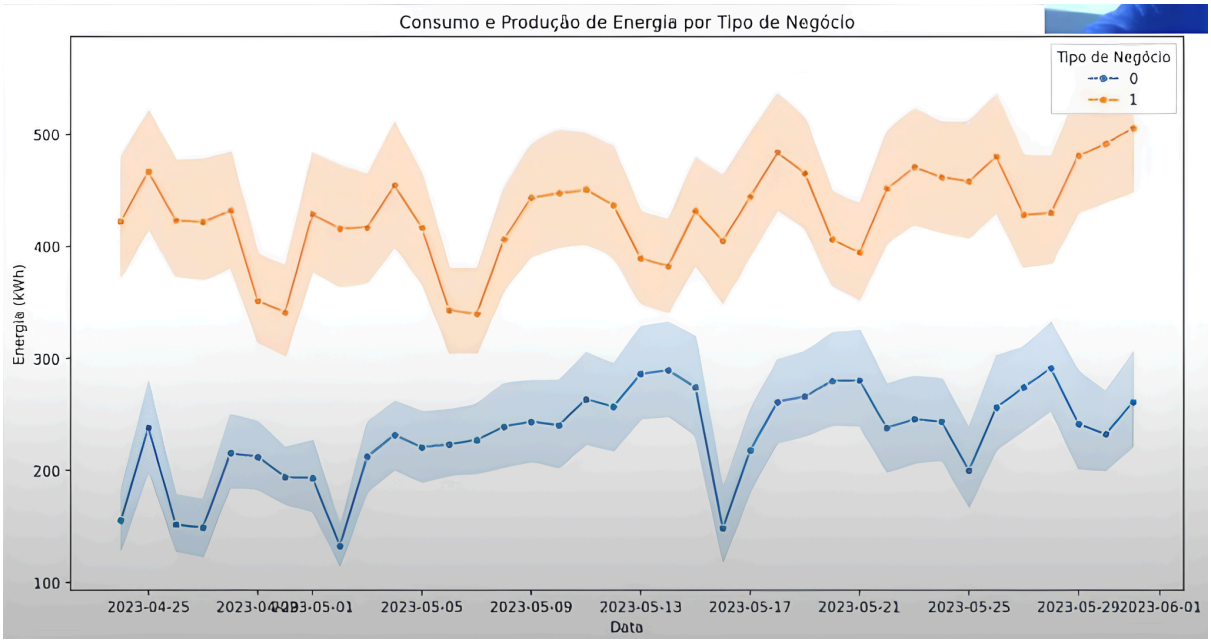

Figura 3. Consumo e Produção de energia por tipo de negócio

Diferentes padrões de consumo: As duas linhas apresentam padrões distintos, indicando que os diferentes tipos de negócios possuem perfis de consumo de energia distintos.

Variações sazonais: É possível observar flutuações no consumo ao longo do tempo, o que pode estar relacionado a fatores sazonais, como mudanças climáticas ou variações na atividade econômica.

Consumo médio: As áreas sombreadas em torno de cada linha representam um intervalo de confiança ou um desvio padrão, indicando o consumo médio e a variabilidade dos dados.