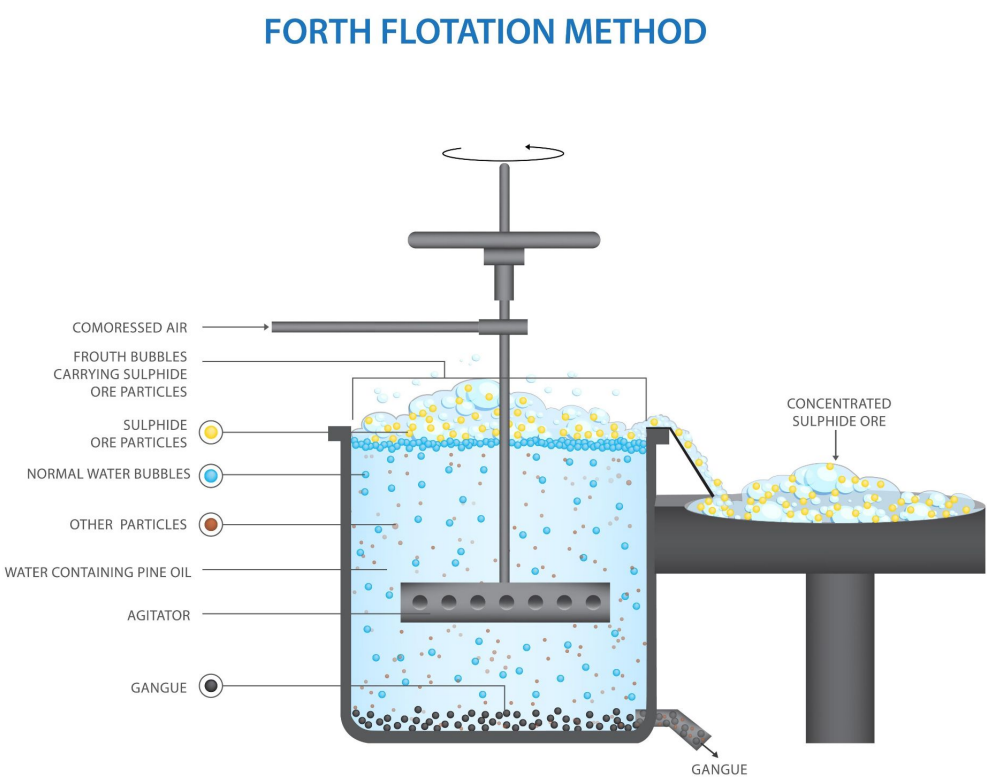
Predição de Qualidade no Processo de Flotação

Por Gabriel S. Costa

Contexto do Projeto

- **Processo industrial**: flotação de minério de ferro
- Impacto direto na qualidade do concentrado e no valor econômico
- Restrições reais:
  - pouco controle sobre coleta
  - conhecimento parcial do processo
  - dados industriais ruidosos

Coluna de Flotação



Contexto e Desafio de Negócio

O Cenário:

A flotação é a etapa final de purificação do minério. O controle preciso da Sílica determina o valor de mercado do produto.

O Problema do "Retrovisor":

- Sensores de processo: Leituras a cada 20 segundos.
- Análise de laboratório (Target): Resultado a cada 1 hora.

Hipóteses Levantadas

- **H1 (Sinal):** As variáveis de processo (vazões, níveis, pH) contêm informação suficiente para explicar a variabilidade da Sílica.
- **H2 (Inércia):** O comportamento passado imediato do processo (Lags) é um preditor mais forte do que valores instantâneos.
- **H3 (Regimes):** O processo opera em diferentes estados (ex: estabilidade vs. transição), exigindo modelos robustos a ruídos.
- **H4 (Redundância):** Existe alta colinearidade; sensores vizinhos dizem a mesma coisa.

A Física dos Dados: Ferro vs. Sílica

- **O "Espelho":** Há uma correlação negativa quase perfeita entre o teor de Ferro e o de Sílica.
- **Decisão Estratégica:** O Teor de Ferro foi **removido** das variáveis preditoras.
- **Por quê?** Usar o Ferro para prever a Sílica seria um "vazamento de dados" (Data Leakage).

Estado Inicial dos Dados

A Estrutura Bruta:

- **Volume:** ~736 mil registros cobrindo 183 dias de operação.
- **Granularidade:** Leituras de sensores a cada 20 segundos.
- **Variáveis:** 21 indicadores de processo (vazões, níveis, pH, densidade).

Qualidade e Integridade:

- **Valores Nulos:** Zero (ausência de NaNs).
- **Redundância:** Alta colinearidade detectada via VIF (Fator de Inflação de Variância), especialmente entre fluxos de ar de colunas vizinhas.
- **Sincronismo:** Identificado um desalinhamento temporal latente que exigiu a correção das janelas de observação.

Decisão de Engenharia:

- Abandono da escala de minutos para adoção da **escala horária**.
- Foco na **tendência e estabilidade** em vez de picos ruidosos.

Estado Inicial dos Dados

A Estrutura Bruta:

- **Volume:** ~736 mil registros cobrindo 183 dias de operação.
- **Granularidade:** Leituras de sensores a cada 20 segundos.
- **Variáveis:** 21 indicadores de processo (vazões, níveis, pH, densidade).

O Achado Crítico: A Ilusão dos 20 Segundos

- Embora o dataset tenha linhas a cada 20s, a variável alvo (**% Sílica**) só é atualizada de **hora em hora**.
- **Problema detectado:** Entre uma análise laboratorial e outra, os valores de Sílica eram simplesmente repetidos no banco de dados.
- **Risco:** Treinar um modelo com dados repetidos causaria um "vício" (overfitting), onde o modelo aprenderia a prever que "o futuro é igual ao agora" de forma artificial.

Qualidade e Integridade:

- **Valores Nulos:** Zero (ausência de NaNs).
- **Redundância:** Alta colinearidade detectada via VIF (Fator de Inflação de Variância), especialmente entre fluxos de ar de colunas vizinhas.
- **Sincronismo:** Identificado um desalinhamento temporal latente que exigiu a correção das janelas de observação.

Decisão de Engenharia:

- Abandono da escala de minutos para adoção da **escala horária**.
- Foco na **tendência e estabilidade** (médias horárias) em vez de picos instantâneos ruidosos.

Engenharia de Atributos

- **De 21 para 107 Atributos:** Criação de variáveis que capturam a dinâmica da planta.
- **Janelas Móveis (Rolling Statistics):** Médias e desvios padrão da última hora para capturar estabilidade.
- **Defasagens (Lags):** Inclusão do estado do processo em T-1 para respeitar o tempo de residência do minério nas colunas.

Estratégia de Validação

- **Validação Temporal (Time Series Split):** Proibido usar sorteio aleatório. O modelo foi treinado no passado e testado no "futuro" relativo.
- **Métrica Principal:** MAE (Erro Médio Absoluto).
- **Métrica de Negócio:** Ganho sobre o Baseline.

Baseline de Persistência

- **O que é?** Prever que o valor da próxima hora será igual ao valor da hora atual (estratégia "ingênua").

- **Resultado:** MAE de **1.11**.

- **Conclusão:** Se o modelo não for melhor que 1.11, ele não tem inteligência real, apenas replica o último dado.

O Modelo Lasso

Por que Lasso? Precisávamos de um modelo linear (interpretável) que fizesse seleção automática de variáveis devido à alta colinearidade (VIF alto).

Performance:

- **MAE:** 0.97 (Redução de **16%** no erro em relação ao baseline).

Veredito: O modelo prova que existe sinal útil para predição.

O que o Modelo aprendeu?

Principais Alavancas:

1. **Vazão de Amina:** Confirmado como principal reagente de controle.
2. **Densidade da Polpa:** Impacto direto no tempo de flotação.
3. **Fluxo de Ar (Colunas 1, 4 e 5):** Variáveis críticas para a formação de espuma.

Insight: O modelo de dados "concorda" com a engenharia de processos química.

Hipóteses: confirmação

H1: Existe sinal preditivo nas variáveis de processo?

(Confirmada.)

- **Evidência:** O modelo Lasso conseguiu reduzir o erro (MAE) em 16% comparado à persistência, provando que variáveis como Vazão de Amina e Fluxo de Ar contêm informação relevante sobre a Sílica futura.

H2: O comportamento passado (Lags) importa mais que o instantâneo?
(Fortemente Confirmada.)

- **Evidência:** As features de *lag* (atraso temporal) e médias móveis foram as mais selecionadas pelo modelo. Isso reflete a **inércia química** do processo: o que acontece nas colunas demora a ser refletido no concentrado.

Hipóteses: confirmação

H3: O processo opera em regimes diferentes?
(Evidências Indiretas.)

- **Evidência:** Há períodos de estabilidade intercalados com transições bruscas de qualidade. Embora não tenham sido criados modelos por regime, a necessidade de regularização (Lasso) sugere que o modelo precisa ser robusto a essas mudanças.

H4: Existe redundância excessiva entre os sensores?
(Confirmada.)

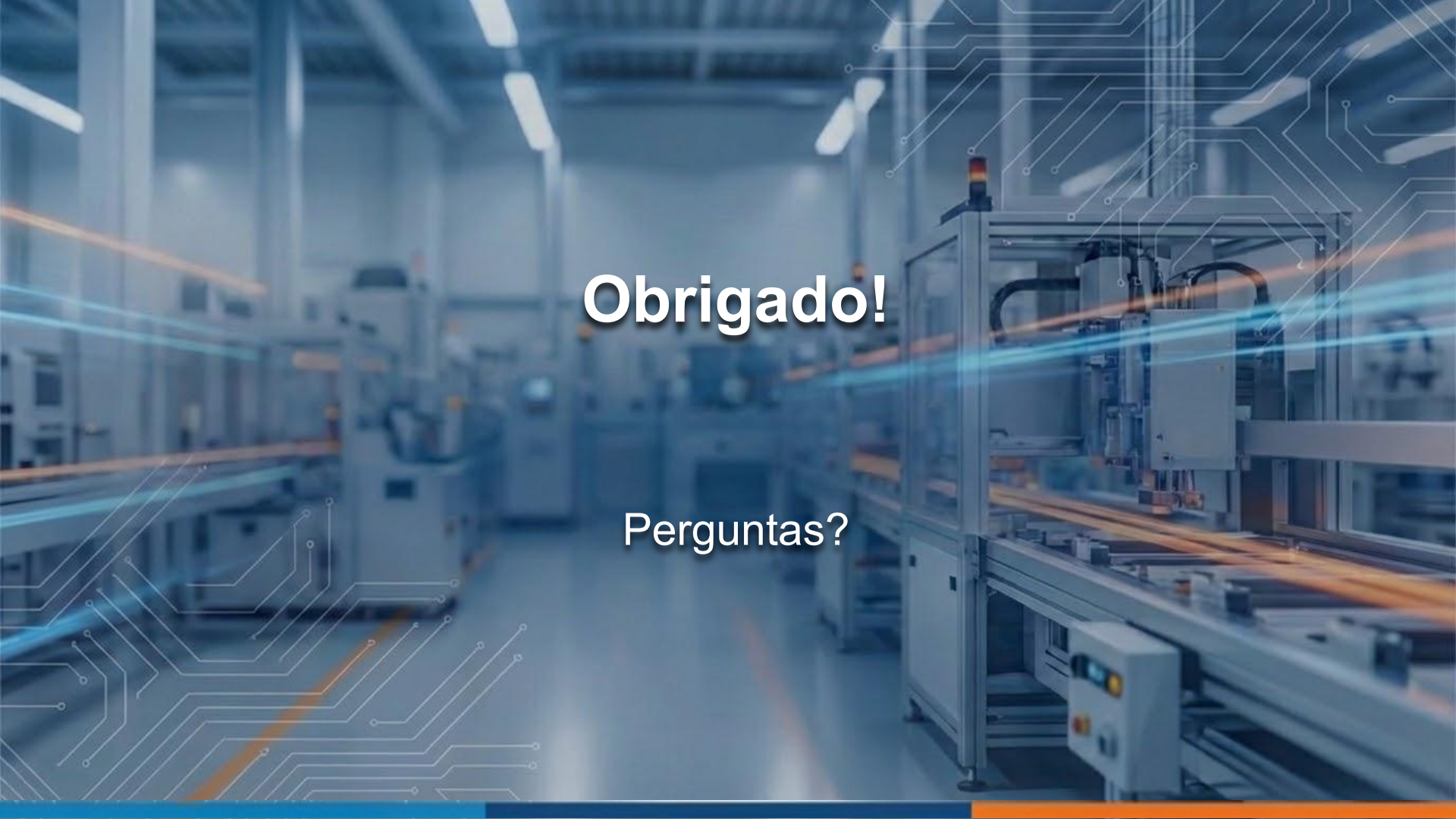
- **Evidência:** O teste de VIF (Fator de Inflação de Variância) apresentou valores altíssimos em diversas colunas (10k e além). Isso validou a estratégia de não usar todos os sensores "crus", mas sim focar em seletores de atributos ou regularização.

Conclusões Operacionais

- **Viabilidade:** É possível antecipar variações de Sílica com 1h de antecedência.
- **Valor Gerado:** Redução da variabilidade e suporte à decisão do operador, evitando que ele "persiga o erro" do passado.
- **Simplicidade:** Modelos lineares já entregam um ganho de 16%, validando o investimento em analytics.

Próximos Passos e Recomendações

- **Monitoramento de Drift:** Implementar alertas para quando o comportamento do minério mudar radicalmente.
- **Modelos Não-Lineares:** Testar algoritmos de Gradient Boosting para capturar interações complexas entre colunas.
- **Interface Homem-Máquina:** Prototipar um dashboard que mostre a "Sílica Estimada" em tempo real para os operadores.


Obrigado!
Perguntas?