Desafio de dados: Aignosi

...

Quality prediction in a mining process

Sumário

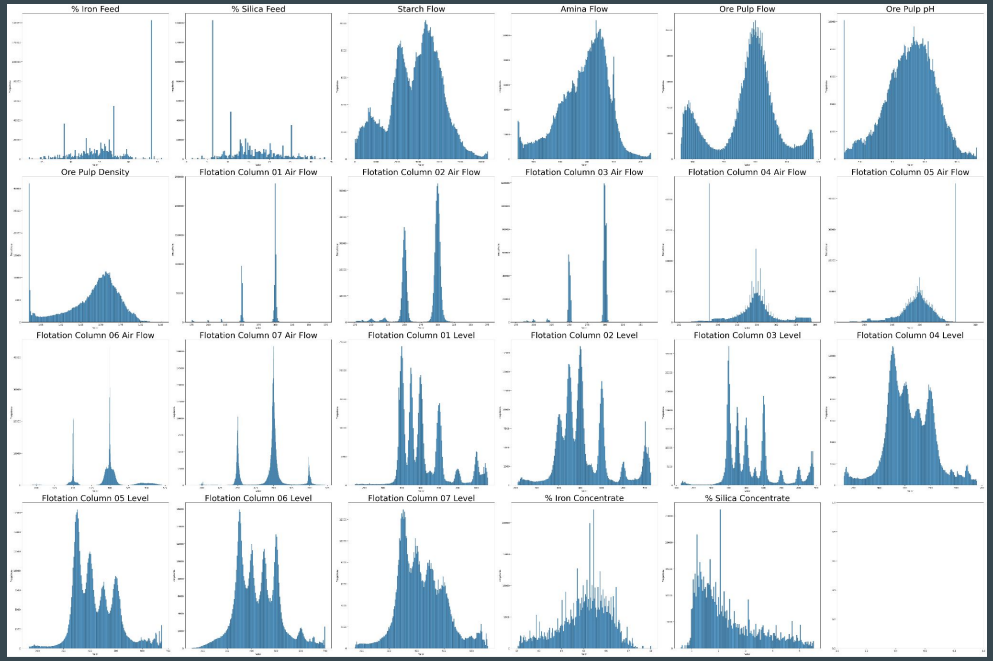
- Riqueza de dados
- Distribuição das variáveis
- Ruídos
- Tendência e sazonalidade
- Estacionariedade
- Correlação entre features
- Feature importance com XGBoost
- Melhorias

Riqueza de dados

- Potencial inexplorado
  - A empresa possui um vasto oceano de dados.
  - No dataset explorado, de forma bruta, há quase 1 milhão de entradas para 23 features e 1 target.
  - Diversos sensores que aquisitam sinais com resolução considerável, na escala de horas ou segundos, que fornecem dados robustos e confiáveis, com poucos outliers.
- Abundância de oportunidades
  - Diversos modelos aplicáveis ao processo minerador, considerando as particularidades dos dados

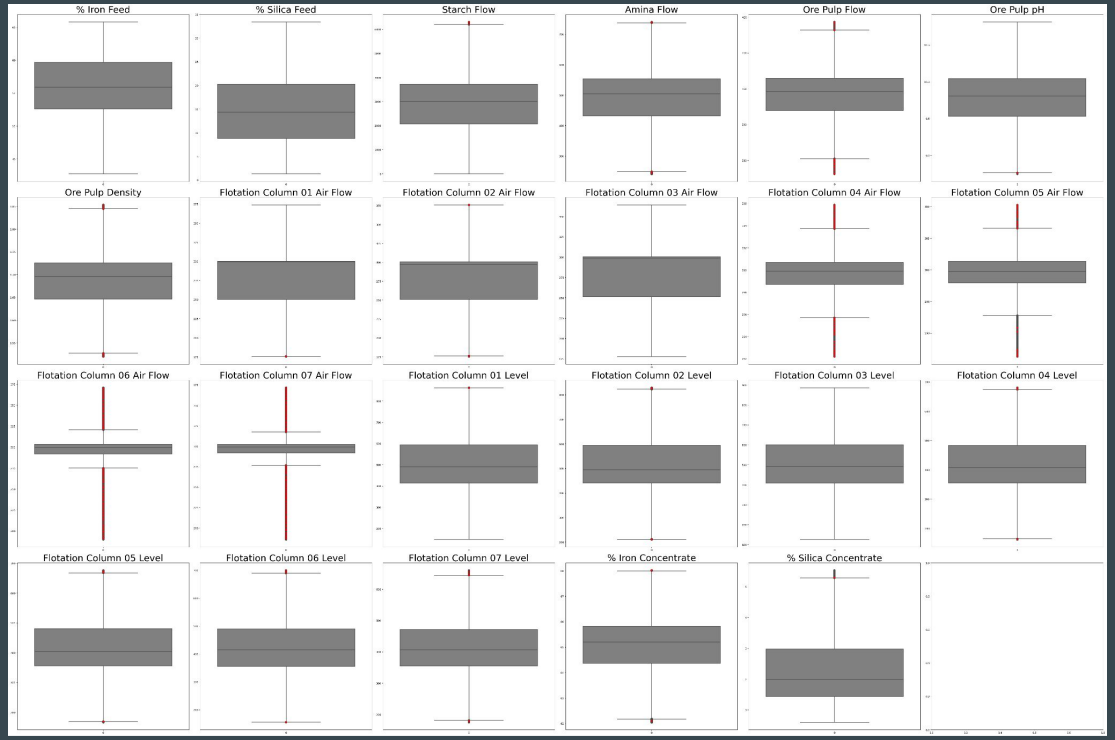
Distribuição das variáveis

- Presença de distribuição normal
  - Maior aplicabilidade em modelos
- Simétricos e bem definidos
  - Detecção de outliers e ruídos
  - Modelos mais robustos
- Ausência de outliers
  - Garbage in, garbage out



Ruídos

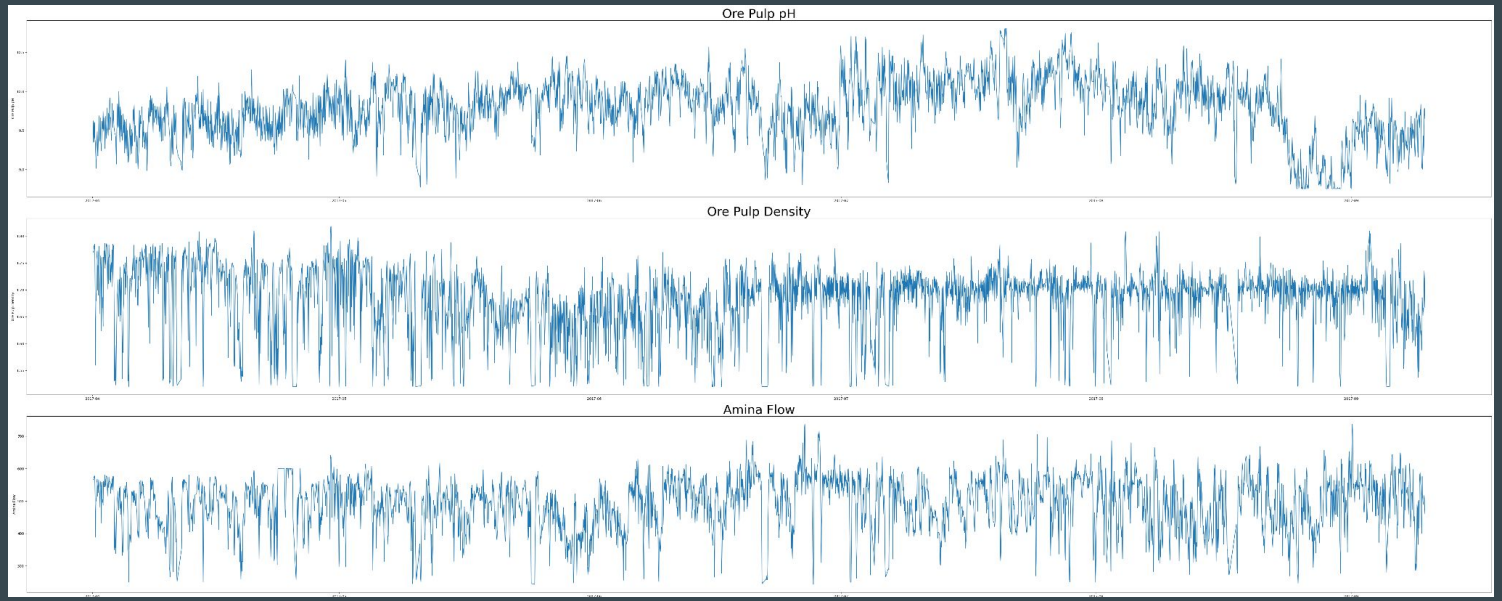
- Poucos outliers
  - Menor custo de remoção
  - Confiabilidade dos modelos
- Detecção de padrões
  - Outras abordagens
  - Aprendizado de máquina
  - Deep learning



Tendência e sazonalidade

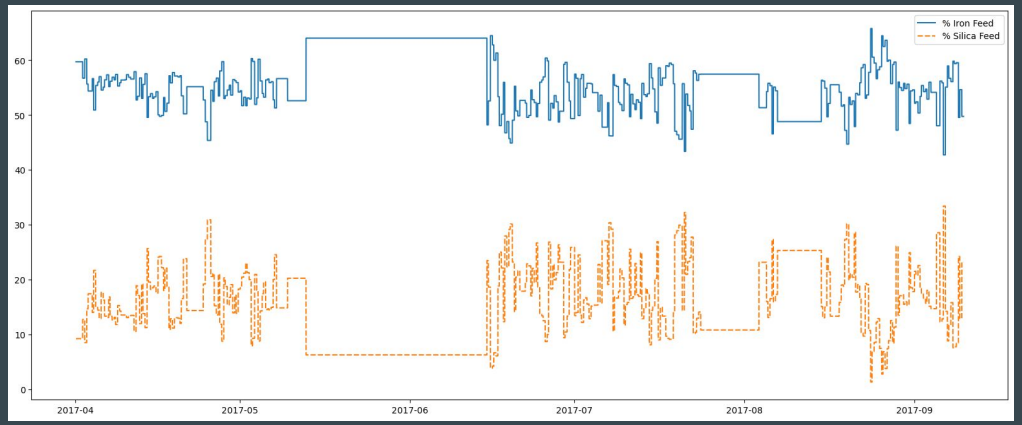
- Visualmente, os dados não apresentam tendência nem sazonalidade definidas
- Estacionariedade
  - Facilidade de predição: modelos podem assumir constância de média e variância
  - Modelos mais simples
  - Menor custo
- Teste Augmented Dickey-Fuller (ADF)
  - Todas as features são estacionárias
  - Necessários mais testes para confirmar

Estacionariedade: exemplos



Correlação entre features

- Algumas features possuem alta correlação entre si, beirando a redundância
- Eliminação de features
  - Redução da complexidade do modelo
  - Redução de custos em sensores (diminuição de frequência de aquisição)
  - Aplicação extra ciência de dados: validação de sensores e redundância (mais segurança)
- % Iron Feed x % Silica Feed (-0.97)

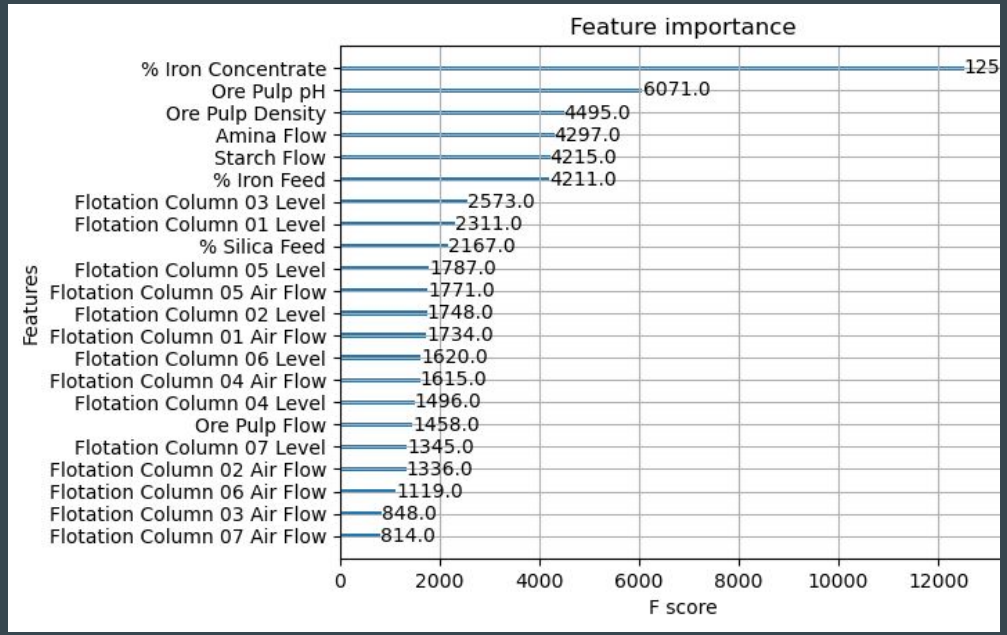


Feature importance com XGBoost

- Alto desempenho, mesmo sem colinearidade clara entre features e targets
- Robusto contra overfitting
  - Alta dimensionalidade de dados -> complexidade -> overfitting
- Interpretabilidade: Feature importance

XGBoost

- Previsão da última semana
- 1000 estimadores
- Todas as features
- MSE: 0.45
- RMSE: 0.67/ Média: 2.31
- Erro: ~29%
- Feature importance
  - Outras variáveis além do %IronConc.



Possíveis melhorias

- Explorar parâmetros do XGBoost
- Treinar outros modelos, como redes neurais e comparar os resultados
- Realizar validações e análises comparativas entre modelos para seleção de features
- Utilizando outros lags das variáveis como novas features
- Retirada de variáveis redundantes

Encerramento

- Agradecimentos
- Perguntas
- Dúvidas