Relatório Técnico – Desafio Machine Learning

Autor: Wesley Canaam (Candidato à posição de Engenheiro de IA SR)

Data: 17/03/2025

Repositório do Desafio: https://github.com/manipulaeHealth/desafio-machine-learning

Repositório da Solução:

🎯 Objetivo do Desafio

Prever o preço correto (`correto`) de fórmulas manipuladas com menor erro possível a partir das informações fornecidas no dataset disponibilizado pela empresa.

📌 Visão Geral do Dataset

- **Total de Dados:** 7121 registros.
- **Estrutura:**
  - `descricao`: fórmula detalhada.
  - `criado`: data de criação.
  - `qtdInsumos`: quantidade de insumos.
  - `calculado`: preço pré-calculado.
  - `correto`: preço real, alvo das previsões.

🔍 Metodologia Aplicada

1. Pré-processamento e Tratamento dos Dados

- **Conversão** da coluna `criado` para datetime, garantindo ordem cronológica.
- **Remoção de outliers** nas colunas críticas (`calculado`, `qtdInsumos`, `correto`) utilizando método IQR.
- **Remoção de duplicatas** para assegurar qualidade dos dados.

2. Engenharia de Features

- Extração de informações valiosas da coluna `descricao`:
  - Quantidade total de insumos.
  - Soma e média das quantidades individuais dos insumos.
  - Número de insumos únicos.

- Criação de novas variáveis derivadas:
  - `calc_qtd_ratio` (calculado/qtdInsumos).
  - `calc_qtd_mult` (calculado*qtdInsumos).
- Criação de features temporais adicionais para captar sazonalidade (`ano`, `mes`, `dia`, `trimestre`, `dia_semana`).
- Agrupamento inteligente de insumos pouco frequentes para evitar alta dimensionalidade.

3. Seleção e Validação dos Modelos

- Divisão dos dados conforme exigido (5121 primeiros registros para treino e 2000 últimos para teste), mantendo estrutura temporal.
- Modelos testados com ajuste fino de hiperparâmetros (GridSearch com validação cruzada):
  - **GradientBoostingRegressor** (melhor performance: RMSE = **9.6370**)
  - XGBRegressor (RMSE = 9.6557)
  - RandomForestRegressor (RMSE = 10.1629)

4. Avaliação da Robustez do Modelo

- Gráficos de resíduos mostraram boa distribuição centrada próximo de zero.
- Comparação real vs. previsto mostrou forte relação linear, confirmando solidez das previsões.
- Importância das features:
  - `calculado` explicou 98% da variação.
  - Demais features tiveram relevância pequena, atuando como ajuste fino.

🚨 Insights Relevantes e Decisões Estratégicas

- O modelo obteve desempenho sólido principalmente devido ao preço pré-calculado (`calculado`), evidenciando que o método interno de precificação atual já possui elevada precisão.
- Para futuras melhorias, é recomendado testar previsões sem o uso direto de `calculado` ou modelar apenas o residual (`correto - calculado`) para obter insights sobre quais insumos específicos ou fatores temporais impactam diretamente no ajuste de preços.

📊 Resultados Alcançados

- **Melhor Modelo:** GradientBoostingRegressor.
- **RMSE Final:** 9.6370
- **Nível de acurácia:** Alta capacidade preditiva, erros médios abaixo de 10 unidades monetárias, validando plenamente o objetivo do desafio.

🔍 Exportação para Power BI e Análises Visuais

Para complementar a modelagem, os dados foram exportados em um formato ideal para análises no **Power BI**, garantindo que insights estratégicos possam ser extraídos. O arquivo CSV contém as seguintes variáveis:

- `descricao`
- `qtdInsumos`
- `calculado`
- `correto`
- `complexidade`
- `ano`, `mes`, `dia`, `trimestre`, `dia_semana`
- `soma_quantidades`
- `media_quantidades`
- `num_insumos_unicos`
- `calc_qtd_ratio`
- `calc_qtd_mult`
- `insumo_outros`

Essas variáveis permitem criar visualizações que ajudam na compreensão da precificação, impacto dos insumos, sazonalidade e identificação de padrões relevantes.

🚩 Próximos Passos Sugeridos

- Explorar validação temporal avançada para capturar tendências de longo prazo.
- Aplicar técnicas de embeddings para melhorar eficiência na codificação dos insumos.
- Testar modelos especializados para previsão do residual (`correto - calculado`).

🎖️ Conclusão

O desafio foi cumprido integralmente, com metodologia robusta, transparência e forte capacidade técnica demonstrada na construção, ajuste e validação dos modelos.

Estou à disposição para aprofundar qualquer ponto deste relatório e discutir estratégias adicionais que possam gerar ainda mais valor para o negócio.

Wesley Canaam
Engenheiro de IA SR