

Teste Técnico - Cientista de Dados Pleno

Previsão de Canal Aduaneiro

Informações Gerais

- **Tempo estimado:** 4-6 horas
- **Formato:** Notebook Jupyter (.ipynb) ou scripts Python + relatório em Markdown
- **Entrega:** Repositório Git com código completo e documentação

1. Contexto do Problema

Introdução

O canal aduaneiro é um sistema de classificação de risco utilizado pela Receita Federal para determinar o nível de fiscalização aplicado a uma Declaração de Importação (DI). Existem quatro tipos de canais:

- **Verde:** Sem fiscalização - liberação automática
- **Amarelo:** Fiscalização documental apenas
- **Vermelho:** Fiscalização documental e física
- **Cinza:** Fiscalização especial antifraude

Objetivo

Desenvolver um modelo de machine learning capaz de prever o canal aduaneiro que será atribuído a uma DI com base em características históricas, auxiliando importadores na previsão de custos e prazos logísticos.

2. Tarefas Solicitadas

2.1 Análise Exploratória dos Dados (EDA)

Realize uma análise exploratória completa dos dados, incluindo:

Análise Descritiva

- Informações gerais sobre o dataset (shape, tipos de dados, valores missing)
- Análise da distribuição da variável target
- Estatísticas descritivas das features categóricas e numéricas

Análise Univariada

- Distribuição de cada feature categórica
- Identificação de outliers e valores inconsistentes
- Análise de padrões temporais (sazonalidade, tendências)

Visualizações

- Heatmaps de correlação
- Análise temporal da distribuição dos canais
- Visualizações de features de risco

2.2 Pré-processamento dos Dados

Implemente um pipeline completo de pré-processamento:

Tratamento de Dados Faltantes

- Estratégia para valores missing em cada coluna
- Justificativa das decisões tomadas

Engenharia de Features

- Criação de features temporais (ano, mês, dia da semana, trimestre)
- Features de risco baseadas em histórico (NCM, país, importador)
- Features categóricas derivadas (ex: capítulo do NCM)
- Features de agregação temporal

Balanceamento de Classes

- Análise do desbalanceamento
- Estratégias para lidar com classes minoritárias
- Implementação de técnicas de over/under sampling se necessário

Divisão dos Dados

- Estratégia de divisão treino/validação
- Consideração de aspectos temporais na divisão

2.3 Modelagem

Implemente e compare pelo menos duas diferentes abordagens de modelagem a sua escolha

Métricas de Avaliação

- Acurácia global
- Precision, Recall e F1-Score por classe
- Matriz de confusão

3. Perguntas Específicas

Análise de Negócio

1. **Quais são os 5 NCMs com maior risco de canal vermelho?** Apresente uma análise detalhada.
2. **Existe sazonalidade na distribuição dos canais?** Como isso pode impactar o modelo?
3. **Qual o impacto do modo de transporte na seleção do canal?** Justifique com dados.
4. **Como o porte da empresa importadora influencia o canal selecionado?**

Modelagem

5. **Qual modelo apresentou melhor desempenho?** Justifique considerando métricas de negócio.
6. **Como você lidaria com o desbalanceamento extremo das classes?** Implemente pelo menos duas estratégias.
7. **Quais features são mais importantes para a predição?** Use técnicas de interpretabilidade.
8. **Como garantir que o modelo seja robusto a mudanças sazonais?**

Produção

9. **Como você implementaria um sistema de monitoramento do modelo em produção?**
10. **Qual estratégia de retreinamento você recomendaria?**
11. **Quais são os principais riscos de bias neste modelo?**

4. Entregáveis

Código

- Notebook Jupyter com análise completa (ou múltiplos notebooks organizados)
- Scripts Python modulares para reuso
- Testes unitários para funções críticas
- Arquivo de requirements.txt

Documentação

- README.md com instruções de uso
- Relatório executivo com principais insights
- Documentação técnica das decisões tomadas
- Análise de limitações e próximos passos

Artefatos

- Modelos treinados salvos
- Pipelines de pré-processamento
- Visualizações salvas em formato adequado
- Arquivo de configuração com hiperparâmetros

5. Dicas e Orientações

Aspectos Temporais

- Mantenha ordem cronológica na divisão dos dados
- Considere lag features se relevante
- Analise estabilidade temporal das features

Interpretabilidade

- Priorize modelos interpretáveis ou use técnicas de explicabilidade
- Documente claramente as regras de negócio inferidas
- Valide insights com conhecimento de domínio

Produção

- Pense em escalabilidade desde o início
- Considere latência vs. acurácia
- Implemente logging adequado

Boa sorte e esperamos ver sua abordagem criativa para este desafio!

Este teste foi desenvolvido para avaliar habilidades técnicas em ciência de dados. Todos os dados utilizados são simulados e não representam informações reais.