Relatório de Modelagem Preditiva de Churn – Telecom X

Autor: Mateu Xauan Gonçalves de Paula

Função: Analista de Machine Learning Júnior

# 1. Resumo Executivo

A Telecom X busca reduzir a evasão de clientes (churn) por meio da antecipação de riscos de cancelamento. Neste projeto, foi desenvolvido um pipeline de machine learning para prever quais clientes têm maior probabilidade de cancelar seus serviços, permitindo que a empresa atue preventivamente.

Foram aplicadas etapas de pré-processamento, análise de variáveis, balanceamento de dados, treinamento de modelos e interpretação de resultados. Três modelos principais foram avaliados: Árvore de Decisão, K-Nearest Neighbors (KNN) e Regressão Logística.

O modelo de Árvore de Decisão com profundidade reduzida (max\_depth=3) se destacou pelo alto recall (86%) na detecção de clientes que cancelariam, sendo mais indicado quando o objetivo é maximizar a identificação de cancelamentos. A Regressão Logística apresentou o melhor equilíbrio entre precisão e recall, além de permitir interpretação clara dos fatores de risco.

# 2. Introdução

Este trabalho corresponde à segunda etapa do desafio “Telecom X – Prevendo Churn”, com foco na construção de modelos preditivos para detecção de clientes em risco de evasão.

Objetivos:

- Preparar e pré-processar os dados para modelagem.

- Analisar correlações e selecionar variáveis relevantes.

- Treinar diferentes modelos preditivos e avaliar seu desempenho.

- Interpretar resultados e extrair insights estratégicos.

- Recomendar ações de retenção baseadas nas descobertas.

# 3. Metodologia

Pré-processamento de Dados:

- Leitura de base previamente tratada (Challenge 1).

- Padronização (Z-score) para variáveis numéricas.

- One-Hot Encoding para variáveis categóricas, evitando multicolinearidade.

- Análise de desbalanceamento (74% clientes ativos vs. 26% cancelamentos).

- SMOTE aplicado somente ao conjunto de treino.

Análise de Correlação:

- Correlação de Spearman com a variável alvo cancelou.

- Fatores de retenção mais fortes: meses\_como\_cliente\_zscore (-0.41), contrato\_two\_year (-0.40), cobranca\_total\_zscore (-0.26).

- Fatores de risco mais fortes: tipo\_internet\_fiber\_optic (0.34), tem\_seguranca\_online (0.29), tem\_fatura\_digital (0.28).

Divisão de Dados:

- Treino: 70% | Teste: 30% (estratificado).

- Aplicação do SMOTE apenas no treino.

# 4. Modelos Avaliados e Resultados

Árvore de Decisão:

- Sem ajuste: Overfitting (alta acurácia no treino, baixa no teste).

- Com max\_depth=12: Acurácia 73%, Recall (cancelou) 67%, Precisão (cancelou) 48%.

- Com max\_depth=3: Acurácia 64%, Recall (cancelou) 86%, Precisão (cancelou) 41%.

K-Nearest Neighbors (KNN):

- Acurácia 70,8%, Recall (cancelou) 72%, Precisão (cancelou) 46%.

Regressão Logística:

- Acurácia 76%, Recall (cancelou) 73%, Precisão (cancelou) 53%.

# 5. Interpretação e Insights

Fatores de Retenção:

- Contratos longos, maior tempo como cliente e maior quantidade de serviços reduzem o risco de cancelamento.

Fatores de Risco:

- Internet fibra óptica, serviços adicionais como segurança online e streaming aumentam o risco de churn.

Estratégia de Modelagem:

- Se a prioridade é não perder clientes em risco, priorizar modelos com alto recall (ex.: Árvore max\_depth=3).

- Se o objetivo é equilibrar custos de falsas previsões, usar modelos mais balanceados como a Regressão Logística.

# 6. Conclusões e Recomendações Estratégicas

Modelo sugerido:

- Árvore de Decisão max\_depth=3 quando recall for prioridade.

- Regressão Logística quando for necessário equilíbrio e interpretação clara.

Ações de Retenção:

- Incentivar contratos de longo prazo.

- Monitorar clientes com muitos serviços adicionais.

- Criar programas de fidelização para novos clientes.

- Reforçar o valor dos pacotes completos para clientes atuais.