Relatório Telecom X2 - Predição de Evasão de Clientes

Maria Clara de Freitas Pádua

1. Extração e Preparação dos Dados

O dataset utilizado foi carregado a partir do arquivo CSV dados_normalizados, originando o DataFrame df. Na etapa inicial, foram removidas colunas irrelevantes para a análise, reduzindo ruídos e simplificando o processamento.

Foi aplicada transformação de variáveis numéricas em duas etapas:

- **Normalização com MinMaxScaler:** reescalou as colunas selecionadas para o intervalo [0, 1], preservando a proporção original entre valores.
- **Padronização com StandardScaler:** aplicou-se logo em seguida, sobrescrevendo os valores normalizados, ajustando-os para média 0 e desvio padrão 1.

Adicionalmente, a coluna Genero foi convertida de texto para valores booleanos (Male → True, Female → False). Para a modelagem, a variável alvo (Y) foi definida como Cancelou, enquanto X concentrou as variáveis preditoras (excluindo Cancelou e ID_Cliente).

2. Codificação de Variáveis Categóricas (One-Hot Encoding)

Variáveis categóricas foram transformadas em variáveis binárias, adotando sempre a criação de **n−1 colunas** para evitar multicolinearidade.

- **Tipo_Internet:**
    - Internet_FiberOptic (1 se fibra ótica, 0 caso contrário)
    - Internet_DSL (1 se DSL, 0 caso contrário)
    - A categoria “No” foi a referência (0 em ambas as colunas).
- **Tipo_Contrato:**
    - Contrato_OneYear (1 se contrato anual)
    - Contrato_TwoYear (1 se contrato bienal)
    - “Month-to-month” foi a referência.
- **Metodo_Pagamento:**
    - Pagamento_MailedCheck (1 se cheque enviado)

- Pagamento_BankTransfer (1 se transferência bancária automática)
- Pagamento_CreditCard (1 se cartão automático)
- “Electronic check” foi a referência.

3. Análise da Proporção de Evasão

A distribuição mostrou forte **desbalanceamento de classes**:

- **Ativos:** 74,3%
- **Cancelados:** 25,7%

Esse desequilíbrio favorece previsões da classe majoritária, podendo comprometer a detecção de cancelamentos. Para corrigir, foi aplicado **SMOTE** (Synthetic Minority Oversampling Technique) **apenas no conjunto de treino**, equilibrando as classes em 50%-50%.

4. Divisão Treino/Teste

O conjunto balanceado foi dividido em **70% treino / 30% teste** usando train_test_split com stratify=y_res, preservando a proporção entre classes nos dois subconjuntos.

5. Análise de Correlação

Foi calculado o **coeficiente de Spearman** apenas para variáveis numéricas e booleanas. Essa métrica não paramétrica foi escolhida por lidar bem com dados sem distribuição normal e relações monotônicas.

- Correlações positivas indicam que o aumento da variável eleva a probabilidade de cancelamento.
- Correlações negativas indicam tendência de retenção.
  Os resultados foram visualizados em heatmap e gráfico de barras, auxiliando na seleção de variáveis mais relevantes.

6. Análises Exploratórias Direcionadas

Gráficos mostraram que:

- Cancelamentos concentram-se nos **primeiros meses de contrato**.

- Clientes ativos possuem contratos mais longos e variados.
- Valor mensal não apresentou relação direta clara com a evasão, mas cancelamentos são mais frequentes no início do vínculo.

7. Modelagem e Avaliação

Árvore de Decisão

- **Versão inicial:** acurácia treino 90%, teste 72,6% → indício de overfitting.
- **max_depth=3:** recall para "Cancelou" subiu para 86%, mas precisão caiu para 41%.
- Melhor para priorizar **identificação de cancelamentos** (alto recall), mesmo com aumento de falsos positivos.

KNN

- Acurácia treino 87,5%, teste 70,3% → início de overfitting.
- Recall "Cancelou": 70% (razoável), precisão baixa (45%).
- Capta boa parte dos cancelamentos, mas com alto número de falsos positivos.

Regressão Logística

- Acurácia geral 76%.
- Recall "Cancelou": 73%, precisão: 53%.
- Interpretabilidade alta via coeficientes, revelando variáveis-chave para retenção ou evasão.

8. Comparativo Geral

- **Árvore de Decisão rasa:** melhor recall para "Cancelou" (86%), indicada quando é mais importante **não perder um cliente prestes a sair**.
- **KNN:** recall razoável, mas menos robusto.
- **Regressão Logística:** maior acurácia global, equilíbrio entre métricas e boa explicabilidade.

9. Importância das Variáveis (Árvore max_depth=3)

- **Mais relevantes:** Contrato_TwoYear e Contrato_OneYear (82% da importância total).
- Seguidas por: Internet_FiberOptic e Streaming_Filmes.
- Outras variáveis foram pouco ou nada utilizadas na árvore devido à profundidade limitada.

10. Conclusões Estratégicas

- **Fatores que aumentam a evasão:** presença de serviços adicionais como segurança online, streaming de TV/filmes e múltiplas linhas.
- **Fatores que reduzem a evasão:** maior quantidade total de serviços, contratos de longo prazo e tempo de permanência como cliente.
- **Ações recomendadas:**
    1. Oferecer benefícios para contratos longos.
    2. Monitorar clientes com muitos serviços adicionais para reduzir insatisfação.
    3. Criar ações de fidelização nos primeiros meses.
    4. Reforçar o valor percebido de pacotes completos.