Caso prático: prevendo o churn de SaaS

ROCKETSEAT

Plataforma completa de aprendizado contínuo em programação.

#BoostingPeople

rocketseat.com.br

Cenário

> Você é analista de dados em uma empresa SaaS que oferece planos mensais de assinatura. A diretoria está preocupada com o aumento da taxa de cancelamento (churn) e quer prever quais clientes estão mais propensos a sair e entender o porquê.

🎯 Objetivo

Prever o churn de clientes e gerar insights acionáveis sobre os principais fatores que influenciam essa saída.

 Contexto do Negócio

A empresa oferece três tipos de planos: Basic, Standard e Premium. Os clientes acessam a plataforma via navegador e interagem com funcionalidades, suporte e e-mails. A meta da área de dados é prever o risco de churn (cancelamento) com base no comportamento e no perfil do cliente.

📊 Variáveis disponíveis

- id_cliente, codigo_cliente, constante: identificadores & coluna constante (irrelevantes para modelagem).
- idade (18–69 anos)
- tempo_de_contrato_meses (1–59 meses)
- frequencia_uso_mensal (0–29 acessos/mês)
- ultimo_login_dias (0–59 dias desde último acesso)
- suporte_chamados_abertos (número de chamados — Poisson média ≈ 2)
- uso_funcionalidades_premium (proporção [0,1] de uso de recursos pagos)
- interacao_emails_marketing (0–9 interações)
- plano (Basic | Standard | Premium)
- pagamento_em_dia (1 = sim / 0 = não)
- avaliacao_satisfacao (nota 1.0–5.0, simétrica em torno de 3.5)
- regiao (Sul | Sudeste | Centro-Oeste | Norte | Nordeste)
- tempo_total_login_horas (média ≈ 50 h/mês)
- navegador_usado (Chrome | Firefox | Safari | Edge)
- hora_ultimo_acesso (0–23)
- churn (0 = cliente ativo / 1 = cliente cancelou)

📊 EDA inicial

1. Cerca de 80 % dos clientes permaneceram ativos (churn = 0) e 20 % cancelaram (churn = 1).

2. Há um leve desequilíbrio de classes (4:1). É importante lembrar disso ao criar modelos (ex.: balancear amostras ou usar métricas além de acurácia).

3. As variáveis mostram amplitude razoável (nenhum viés extremo), mas veremos que “pagamento_em_dia” pode ser importante para churn. A distribuição exponencial de “tempo_total_login_horas” indica que há muitos usuários de “baixa atividade” e poucos de “alta atividade”.

📊 EDA Detalhada

1. Nenhum missing → parte do pré-processamento não exige imputação
2. Principais preditores univariados de churn:
   a. Não pagamento em dia (inadimplência)
   b. Maior número de dias sem login
   c. Baixa frequência de uso mensal
   d. Baixo uso de funcionalidades premium
   e. Avaliação de satisfação baixa
   f. Plano mais simples (Basic) churna mais
3. Variáveis removíveis: id_cliente e constante.
4. Multicolinearidade leve entre “tempo_total_login_horas” e “frequencia_uso_mensal” → avaliar manter apenas a mais forte.
5. Categorias com pouco impacto (ex.: navegador) podem ser combinadas ou excluídas se não melhorarem o modelo.

Próximos passos

1. Pré-processamento: Remover colunas irrelevantes, codificar (One-Hot / LabelEncoder), verificar correlação.
2. Feature Engineering: criar uso_total, suporte_freq, bin_..., etc.
3. Feature Selection: usar VIF e SelectKBest / RFE para escolher as variáveis mais relevantes.
4. Treinamento comparativo:
   a. Regressão Logística
   b. Random Forest
   c. XGBoost
   d. SVM (ou outro alg. de sua escolha)
5. Avaliar com validação cruzada (AUC, F1, Precision/Recall).
6. XAI: aplicar SHAP e/ou LIME para explicar as previsões de churn e extrair insights acionáveis (por ex., “inadimplência + 30 dias sem login → 90 % de chance de churn”).

Conclusão

- **Vamos para a prática?**