Apostila de Modelagem Preditiva e Ciência de Dados para Iniciantes

1. Regressão Logística e Função Sigmoide

A regressão logística é um modelo estatístico usado para prever probabilidades de uma variável binária (ex: "vai pagar" ou "não vai pagar").
Ela utiliza uma função chamada sigmoide para transformar qualquer valor real (positivo ou negativo) em uma probabilidade entre 0 e 1.

Exemplo:
Se um cliente tem score 0.8, isso pode ser interpretado como 80% de chance de pagar o empréstimo.

Função Sigmoide:
sigmoid(x) = 1 / (1 + e^-x)

2. Overfitting, Validação Cruzada, Treino/Teste

Overfitting ocorre quando um modelo aprende demais os dados de treino, incluindo ruídos, e falha em generalizar para novos dados.

Solução: separar dados em treino e teste, e usar validação cruzada.

Validação cruzada (k-fold) divide os dados em 'k' partes, treinando com 'k-1' e testando com a parte restante, rotativamente.

3. Imputação de Valores Ausentes

Imputar valores ausentes significa preencher dados faltantes.

Exemplos:
- Média ou mediana (para variáveis numéricas)

- Moda (para variáveis categóricas)

Importante: a imputação feita no treino deve ser repetida exatamente igual na produção.

4. Seleção de Variáveis

Usar todas as variáveis pode aumentar o risco de overfitting. Selecionar apenas as mais relevantes melhora a performance.

Técnicas comuns:
- Ganho de informação
- Importância de variáveis em modelos de árvore

5. Balanceamento de Classes

Quando uma classe ocorre muito mais do que outra (ex: 90% bons pagadores, 10% maus), o modelo pode se tornar tendencioso.

Técnicas para balancear:
- Oversampling (duplicar exemplos da classe minoritária)
- Undersampling (reduzir exemplos da classe majoritária)

6. Construção de Targets

Target é o que o modelo tenta prever. Em crédito, costuma ser "pagou ou não".

Cuidado: se você construir o target só com clientes aprovados, o modelo terá viés (viés de aprovação). Inclua clientes negados quando possível.

7. Cross-validation, GroupKFold e Vazamento

Cross-validation ajuda a testar o modelo com diferentes divisões de dados.

GroupKFold é uma variação onde grupos (ex: mesmo cliente com vários registros) não aparecem em treino e teste ao mesmo tempo, evitando vazamento.

Vazamento ocorre quando o modelo tem acesso a dados que não deveria durante o treino.

8. Representatividade dos Dados

Os dados usados para treinar o modelo devem representar a população real.

Se você treinar só com um tipo de cliente (ex: só homens), o modelo pode não funcionar bem para outros grupos.

9. Testes A/B e Boas Práticas

Testes A/B comparam dois grupos (A = controle, B = variação).

Boas práticas:
- Distribuir usuários aleatoriamente
- Duração suficiente para capturar sazonalidades
- Não mudar o teste no meio

10. Viés de Aprovação e População Modelada

Viés de aprovação acontece quando o modelo só vê dados de clientes que foram aprovados no passado.

Isso impede que ele aprenda a identificar bons pagadores entre os reprovados.

Solução: usar informações de todos os clientes, mesmo os que não foram aprovados.