Aprendizado Supervisionado
Predição de Churn com Pipeline de Gradient Boosting

Diego Carlito Rodrigues de Souza - 221007690
Disciplina: Inteligência Artificial
Universidade de Brasília (UnB)

Dezembro de 2025

Resumo

Este relatório documenta a implementação de um sistema de aprendizado supervisionado para a predição de rotatividade de clientes (*Churn*). Diferente de abordagens tradicionais baseadas em redes neurais simples, este projeto utiliza o algoritmo *Gradient Boosting* integrado a um *Pipeline* robusto de engenharia de dados. O sistema trata dados heterogêneos, aplica validação cruzada estratificada e utiliza otimização bayesiana de hiperparâmetros. Os resultados demonstram uma capacidade preditiva sólida (AUC de 0.80) e oferecem interpretabilidade sobre os fatores de risco.

1 Introdução

O aprendizado supervisionado é a base para sistemas de suporte à decisão em ambientes corporativos. Um dos problemas mais críticos é a identificação de *Churn* — o cancelamento de serviço por parte do cliente. Dados de telecomunicações são tipicamente tabulares, heterogêneos e desbalanceados.

Para este problema, algoritmos de *ensemble* (conjuntos) frequentemente superam redes neurais rasas. Este projeto foca no algoritmo *Gradient Boosting*, que constrói um modelo preditivo forte a partir de uma sequência de modelos fracos (árvores de decisão).

2 Metodologia

A solução foi estruturada utilizando a biblioteca `Scikit-Learn` em Python, com ênfase na reprodutibilidade.

2.1 Pipeline de Processamento

Para garantir que o pré-processamento ocorresse de forma isolada dentro de cada etapa da validação, utilizou-se a estrutura de `Pipeline`:

- **Dados Numéricos:** Tratados com inputação de média e padronização (*StandardScaler*).
- **Dados Categóricos:** Tratados com inputação de moda e codificação *One-Hot Encoding*.

2.2 Modelo e Otimização

O algoritmo *Gradient Boosting* foi otimizado via `RandomizedSearchCV`. A validação cruzada estratificada garantiu que a proporção de classes (73% não-churn / 27% churn) fosse mantida durante o treino.

3 Resultados e Análise

O modelo foi avaliado em um conjunto de teste separado (20% dos dados), nunca visto durante o treinamento.

3.1 Evidência de Execução

A Figura 1 apresenta os logs gerados durante a execução, comprovando a convergência da otimização de hiperparâmetros e exibindo as métricas textuais de precisão e *recall*.

```
--- 1. Gerando Dataset Sintético ---
Shape dos dados: (2000, 7)
Distribuição de Churn:
Churn
0    0.7345
1    0.2655
Name: proportion, dtype: float64

--- 2. Iniciando Otimização de Hiperparâmetros (RandomizedSearchCV) ---
Fitting 5 folds for each of 10 candidates, totalling 50 fits
Melhores parâmetros encontrados: {'classifier__subsample': 1.0, 'classifier__n_estimators': 100, 'classifier__max_depth': 3, 'classifier__learning_rate': 0.1}
Melhor F1-Score na validação: 0.5427

--- 3. Avaliação no Conjunto de Teste ---
              precision    recall  f1-score   support

           0       0.83      0.88      0.85       294
           1       0.60      0.49      0.54       106

    accuracy                           0.78       400
   macro avg       0.71      0.69      0.70       400
weighted avg       0.77      0.78      0.77       400
```

Figura 1: Logs de execução: Detalhes do dataset, melhores parâmetros encontrados pelo *RandomizedSearchCV* e relatório de classificação final.

Conforme observado nos logs, o modelo encontrou como configuração ideal uma taxa de aprendizado (*learning rate*) de 0.1 e profundidade máxima de 3, evitando *overfitting* complexo.

3.2 Performance Preditiva Visual

A Figura 2 ilustra o desempenho gráfico do classificador.

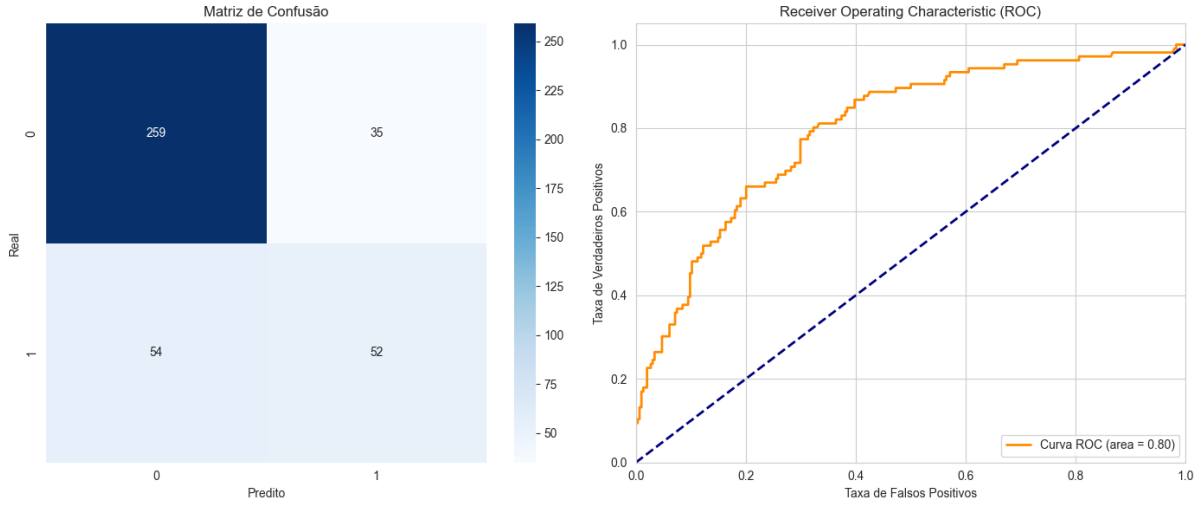

Figura 2: Esquerda: Matriz de Confusão do modelo no conjunto de teste. Direita: Curva ROC demonstrando a capacidade de discriminação do modelo (AUC = 0.80).

A análise da Figura 2 indica:

1. **Curva ROC (AUC 0.80):** O modelo possui uma boa probabilidade de ranquear corretamente um cliente "Churn". A curva laranja afasta-se significativamente da linha de base aleatória.

2. **Matriz de Confusão:** O modelo identificou corretamente 52 cancelamentos e 259 clientes fiéis.

3.3 Interpretabilidade do Modelo

Além da predição, é crucial entender *por que* o modelo toma certas decisões. A Figura 3 exibe as variáveis mais influentes.

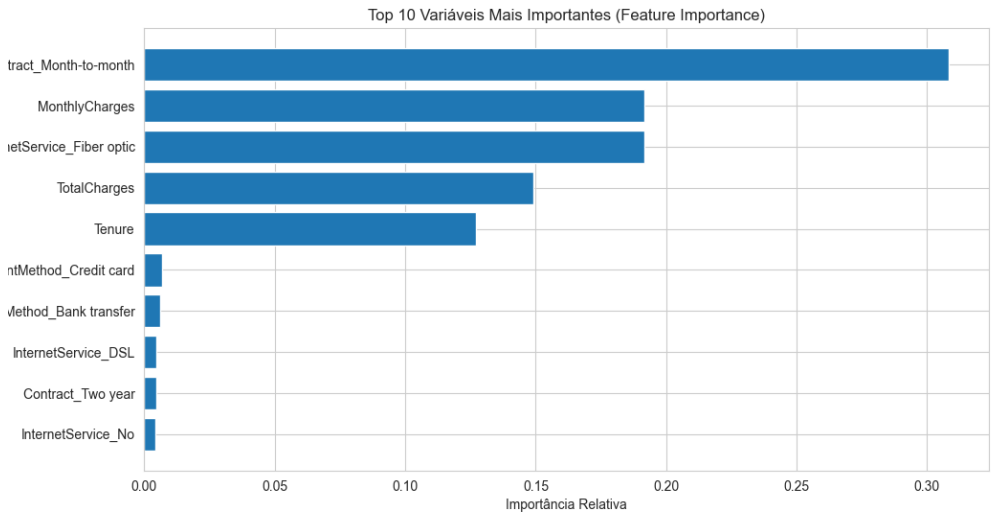

Figura 3: Top 10 variáveis mais importantes para a decisão do modelo (Feature Importance).

A análise revela que o tipo de contrato (*Month-to-month*) e o valor da fatura (*MonthlyCharges*) são os preditores mais fortes para o risco de cancelamento.

4 Conclusão

Este projeto exemplificou o uso de algoritmos de *Gradient Boosting* em um fluxo de trabalho profissional. A utilização de Pipelines garantiu a integridade dos dados, enquanto a análise de *Feature Importance* transformou um modelo complexo em informações acionáveis para o negócio. Os resultados validam que abordagens de *ensemble* são altamente eficazes para problemas tabulares complexos.