Previsão de Churn Bancário Utilizando Machine Learning Clássico

Autores: Bruno Monteiro, Luis Filipe Colombo e Vitor Nascimento

Contexto: O Desafio do Churn no Setor Bancário

Ambiente Competitivo

O setor bancário atual é marcado por uma concorrência acirrada, impulsionada principalmente pela ascensão de fintechs e bancos digitais que oferecem serviços ágeis e desburocratizados.

Migração de Clientes

A facilidade com que os clientes podem migrar para outras instituições representa um risco constante. A baixa barreira de entrada e a oferta de melhores condições são fatores decisivos.

Alto Impacto Financeiro

A perda de clientes (churn) não significa apenas a interrupção de receitas futuras, mas também acarreta custos de aquisição de novos clientes, impactando diretamente a rentabilidade do banco.

ML como Solução

O Machine Learning surge como uma ferramenta estratégica para prever o risco de churn antecipadamente, permitindo que as instituições financeiras atuem proativamente na retenção.

Problema e Objetivos do Projeto

Problema Central

Identificar proativamente clientes com alta probabilidade de cancelar suas contas bancárias antes que o churn ocorra, minimizando perdas e otimizando estratégias de retenção.

Objetivos Específicos

- **Construir Modelos Preditivos**

  Desenvolver e implementar modelos de Machine Learning eficientes para a previsão de churn.

- **Comparar Algoritmos**

  Analisar e comparar o desempenho de dois algoritmos clássicos: Regressão Logística e Árvore de Decisão.

- **Avaliar Desempenho e Variáveis**

  Avaliar os modelos com métricas de classificação robustas e identificar as variáveis mais relevantes para a previsão de churn.

Dataset Utilizado: Churn Modeling (Kaggle)

Origem

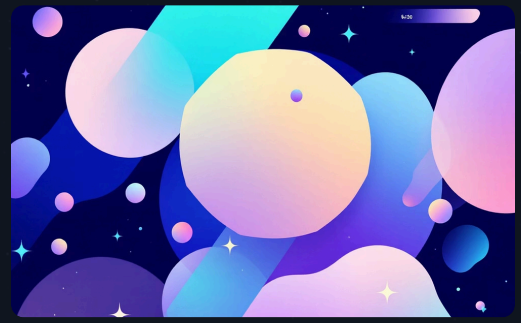
O dataset "Churn Modeling" foi obtido na plataforma Kaggle, uma das maiores comunidades de ciência de dados do mundo.

Estrutura

Composto por 10.000 registros e 14 colunas, fornecendo uma base rica para a análise preditiva.

Variável Alvo

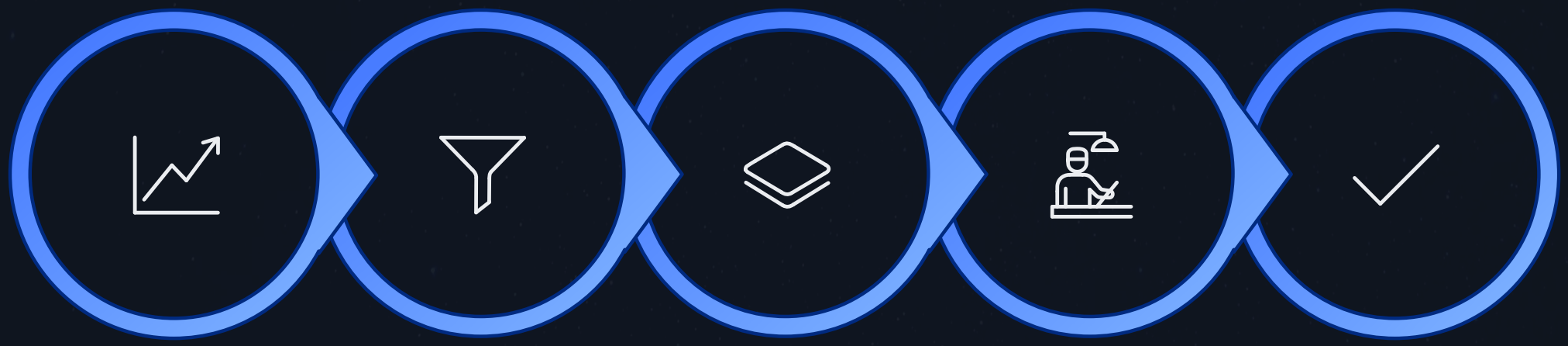
A coluna 'Exited' representa a classe alvo, indicando se o cliente cancelou (1) ou permaneceu (0) no banco.



Variáveis Principais

Inclui atributos críticos como Idade, Saldo, Score de Crédito, Geografia, Gênero e Salário Estimado, essenciais para entender o comportamento do cliente.

Metodologia: Pipeline de Machine Learning



01

Análise Exploratória de Dados (EDA)

Realização de análises gráficas e estatísticas para compreender a distribuição dos dados, identificar padrões e anomalias.

02

Pré-processamento de Dados

Etapas como remoção de colunas irrelevantes, codificação de variáveis categóricas (One-Hot Encoding) e escalonamento (StandardScaler) para preparar os dados.

03

Engenharia de Atributos

Criação de novas variáveis, como a 'BalanceSalaryRatio', para extrair mais informação e melhorar o poder preditivo dos modelos.

04

Divisão Treino/Teste

O dataset foi dividido em 70% para treinamento dos modelos e 30% para teste, garantindo uma avaliação imparcial do desempenho.

05

Treinamento e Avaliação de Modelos

Modelos de Regressão Logística e Árvore de Decisão foram treinados e avaliados utilizando métricas de classificação específicas para problemas de churn.

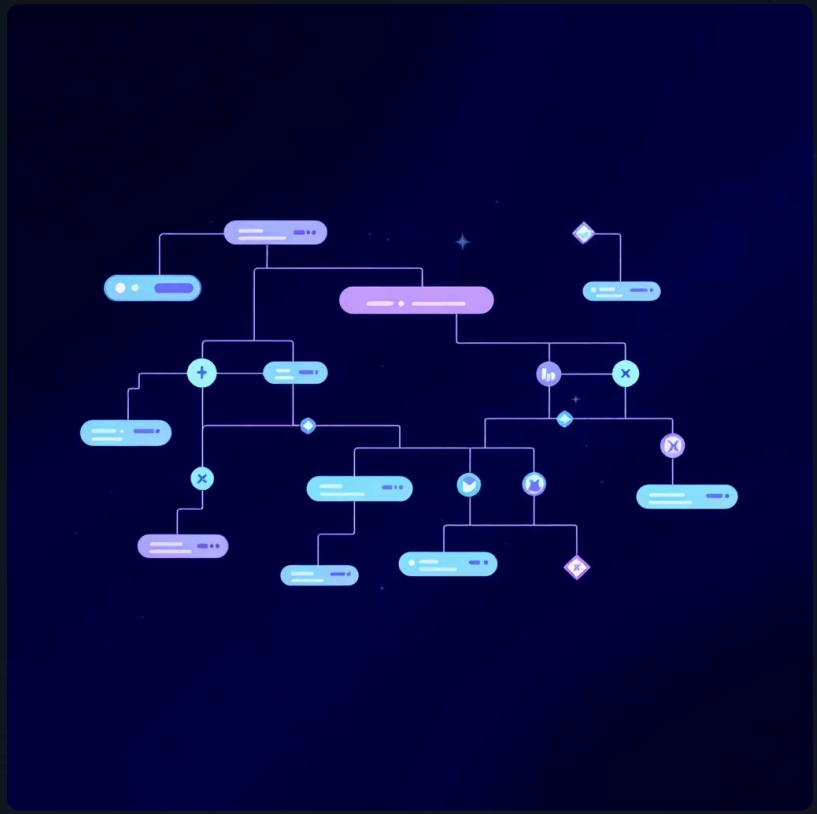
Modelo A: Regressão Logística

Visão Geral

A Regressão Logística é um algoritmo de classificação binária que estima a probabilidade de um evento ocorrer. É amplamente utilizado devido à sua simplicidade e interpretabilidade.

- **Classificação Probabilística:** Fornece a probabilidade de um cliente cancelar, não apenas uma decisão binária.
- **Interpretabilidade:** Os coeficientes do modelo indicam a força e a direção da relação entre as variáveis preditoras e a probabilidade de churn.
- **Eficiência Computacional:** Possui baixo custo computacional, sendo rápido para treinar e fazer previsões, ideal para grandes volumes de dados.
- **Desempenho Sólido:** Demonstrou uma das melhores performances no projeto, especialmente em datasets com relações lineares subjacentes.

Modelo B: Árvore de Decisão



Visão Geral

A Árvore de Decisão é um modelo de classificação que utiliza uma estrutura hierárquica de regras para tomar decisões. Ela particiona o espaço de dados em subconjuntos baseados em características das variáveis.

- **Regras Hierárquicas:** Opera através de uma série de perguntas e respostas (nós e folhas), que levam a uma classificação final.
- **Captura Relações Não Lineares:** Capaz de modelar interações complexas e padrões não lineares nos dados.
- **Fácil Interpretação:** A estrutura em árvore é intuitiva e pode ser visualizada, facilitando a compreensão das regras de classificação.
- **Risco de Overfitting:** Pode se ajustar excessivamente aos dados de treino, resultando em desempenho inferior em dados não vistos.
- **Desempenho Moderado:** Embora útil para insights, seu desempenho neste projeto foi inferior ao da Regressão Logística em termos de métricas gerais.

Resultados e Análise Comparativa

A avaliação dos modelos revelou diferenças significativas no desempenho, com a Regressão Logística se destacando em métricas cruciais para a previsão de churn.

Regressão Logística: Melhor Recall

O modelo de Regressão Logística obteve as melhores pontuações na métrica de Recall, indicando sua capacidade superior em identificar corretamente os clientes propensos ao churn.

Árvore de Decisão: Melhores Métricas Gerais

A Árvore de Decisão apresentou um desempenho superior nas demais métricas, mas inferior por quase 20% no Recall, confirmando a importância da escolha do modelo para o problema específico.

AUC-ROC da Árvore de Decisão Superior

A Curva ROC e sua Área sob a Curva (AUC-ROC) confirmaram a robustez da Árvore de Decisão, com um valor mais alto indicando melhor poder de discriminação entre as classes.

Variáveis Mais Importantes

As análises de importância das variáveis destacaram Idade, Saldo e a localização Geográfica (especialmente Alemanha) como os fatores mais influentes na previsão de churn.

Conclusão: Sucesso na Previsão de Churn

Este projeto demonstrou a eficácia do Machine Learning clássico na previsão de churn bancário, atingindo os objetivos propostos com um modelo robusto e um pipeline replicável.

1 Objetivo Alcançado

Conseguimos construir modelos preditivos eficazes para identificar clientes com alta probabilidade de churn.

2 Regressão Logística como Melhor Modelo

A Regressão Logística se destacou como o modelo mais adequado para este problema, oferecendo o melhor balanço entre desempenho e interpretabilidade.

3 Eficiência do Machine Learning

O estudo reforça a capacidade do Machine Learning de fornecer insights valiosos e ferramentas práticas para a retenção de clientes no setor bancário.

4 Pipeline Completo e Replicável

Desenvolvemos um pipeline de ML completo, desde a EDA até a avaliação, que pode ser replicado e adaptado para outros contextos de negócio.

Trabalhos Futuros e Expansões

Para aprimorar ainda mais a solução de previsão de churn, diversas direções podem ser exploradas em trabalhos futuros.

Modelos Ensemble

Explorar o uso de modelos mais avançados como Random Forest, Gradient Boosting (XGBoost, LightGBM) para capturar relações mais complexas e melhorar o desempenho preditivo.

Técnicas de Balanceamento

Aplicar métodos como SMOTE (Synthetic Minority Over-sampling Technique) para lidar com o desbalanceamento de classes e otimizar a identificação da classe minoritária (churn).

Ajuste do Limiar de Classificação

Realizar um ajuste fino do limiar de classificação para maximizar o Recall, especialmente porque o custo de um falso negativo (perder um cliente) for maior que o de um falso positivo (gastar mais recursos em um cliente que não iria cancelar).

Dashboard ou API

Desenvolver um dashboard interativo ou uma API RESTful para integrar o modelo preditivo a sistemas corporativos, facilitando o uso e a visualização dos resultados pelas equipes de negócio.