Análise de Risco de Crédito

com Modelagem Preditiva

Projeto

Contexto

O projeto é parte do desafio técnico proposto pela PowerofData para a posição de Cientista de Dados Júnior.

Objetivo

O objetivo é analisar uma base pública de dados relacionada a crédito ao consumidor, com foco na melhoria do processo de avaliação de risco e na redução de prejuízos. Espera-se, ao final, a construção de um modelo preditivo capaz de estimar a probabilidade de inadimplência (default).

Dados

O conjunto de dados disponível, cerca de 3GB, foi acessado, observado, e foram selecionadas 3 tabelas principais para a construção de um modelo inicial, com foco em simplicidade e interpretabilidade. O pipeline foi estruturado para permitir a inclusão de novas fontes de dados futuramente.

Metodologia

Desde os dados até a modelagem + insights

O projeto foi conduzido com foco na clareza das etapas e no uso eficiente dos dados.

A metodologia adotada seguiu o ciclo:
→ 🧹 Tratamento Dados
→ 📊 Exploração Dados
→ 🧠 Modelagem Preditiva
→ 🧾 Interpretação e Insights

Evitou-se o uso de múltiplas bases complementares inicialmente, priorizando a criação de um pipeline enxuto, confiável e extensível.

A modelagem preditiva baseou-se em variáveis interpretáveis e diretamente relacionadas ao comportamento de pagamento.

Tratamento de Dados

Raw Zone

Dados brutos extraídos da competição Kaggle

- Mais de 3GB de arquivos disponíveis
- Seleção inicial de 3 tabelas (train_base, train_static_0_0, train_static_cb_0)
- Chave primária comum: case_id

Trusted Zone

Tratamento técnico para padronização e consistência dos dados

- Remoção de colunas com mais de 95% de nulos
- Conversão de datas e variáveis para tipos apropriados
- Normalização e limpeza de dados inconsistentes
- Encoding de variáveis categóricas com baixa cardinalidade
- Aplicação de log1p em variáveis com cauda longa

Refined Zone

Base final de modelagem com variáveis selecionadas

- Seleção de variáveis mais correlacionadas com o target
- Criação de variáveis transformadas e interpretáveis
- Manutenção do target e eliminação do case_id
- Pronta para uso direto em modelagem supervisionada

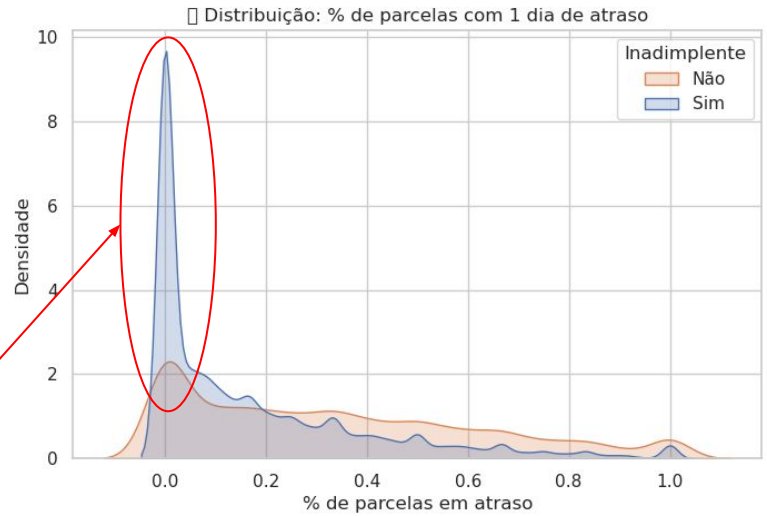
Exploração de Dados

Análise de Distribuição de Inadimplentes

- Inadimplência presente em apenas ~3% dos registros
- Amostra fortemente desbalanceada

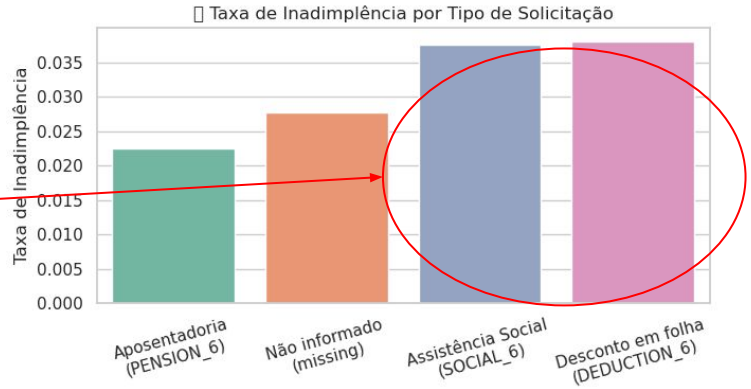
Análise Variáveis Numéricas

- Correlacionamento de variáveis com amostra inadimplente
- Exemplo: Inadimplência (y), % de Parcelas atrasadas - 1 dia(x)



Análise Variáveis Numéricas

- Correlacionamento de variáveis com amostra inadimplente
- Exemplo: Inadimplência (y), Tipo de Solicitação (x)



Modelagem Preditiva

Algoritmos utilizados

- Regressão Logística (baseline)
- LightGBM (modelo principal)

Estratégia de treino

- Validação hold-out (80/20), com estratificação
- class_weight='balanced' para lidar com desbalanceamento

Métricas de avaliação

- AUC (Área sob a curva ROC)
- Recall da classe 1 (inadimplente)

Interpretação da Modelagem

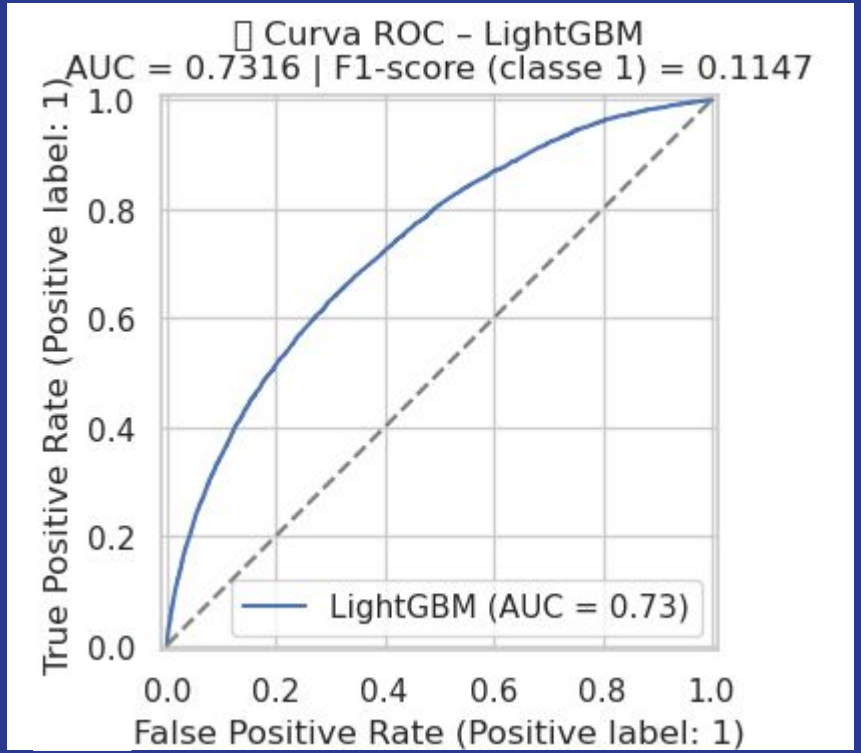
LightGBM

🟢 AUC = 0.73 → mostra com boa performance (!) se o modelo sabe diferenciar quem vai pagar de quem vai atrasar.

🔴 F1-score para inadimplentes (classe 1): ~0.12 – mostra o quanto ele acerta os casos difíceis (inadimplentes), que são minoria.

🎯 Mesmo com poucos acertos, o modelo já supera o acaso e pode ser útil.

Obs: a linha pontilhada significa uma escolha aleatória de inadimplente x adimplente. Nosso **modelo supera com a curva** acima da linha pontilhada!



Insights e Recomendações Finais

Principais Achados

- A inadimplência representa apenas 3% dos registros, exigindo estratégias específicas para tratar o desbalanceamento.
- Variáveis com maior impacto na inadimplência estão ligadas ao histórico de pagamento recente e à quantidade de parcelas em atraso.
- Modelos simples já apresentam desempenho robusto (AUC ~0.73 com LightGBM), mesmo com base enxuta e interpretável.

Recomendações de Ações

- Uso em produção: o modelo pode ser usado como suporte à decisão em concessão de crédito, atribuindo um score de risco.
- Melhoria contínua: incluir mais variáveis temporais, comportamentais e dados externos (ex.: renda estimada, região).
- Interpretação acessível: manter variáveis explicáveis é essencial para uso ético e transparente em decisões financeiras.
- Análise de segmentos: explorar subgrupos de clientes (por tipo de produto, idade ou localização) para estratégias específicas.

Próximos Passos

- Validação cruzada e tunagem de hiperparâmetros
- Teste de outros algoritmos (ex.: XGBoost, redes neurais)
- Adição de mais variáveis disponíveis, numérica e categóricas, para testes de correlação (perseguir aumento de F1-score)
- Criação de plano de monitoramento e re-treinamento periódico