Documentação - Análise de Crédito - POD Bank

Squad 09

- Product Manager (PM): Rafael Salomão
- Visual Analyst (VA): Pedro Melo
- Líder Técnico: Diego Brum
- Desenvolvedores: Vinicius Ranieri, Diego Lucena

1. Entendimento do Negócio

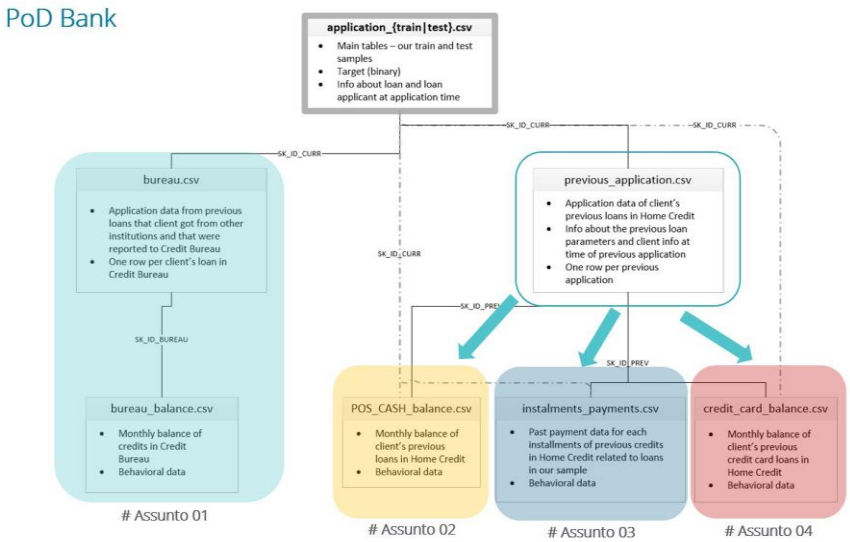
O desafio central é antever a capacidade de reembolso de empréstimos pelos clientes da PoD Bank, visando ampliar o acesso ao crédito para aqueles com histórico limitado ou inexistente.

A empresa busca construir um modelo preditivo que, ao utilizar uma gama de fontes de dados, incluindo registros bancários e informações socioeconômicas, possa avaliar a probabilidade de inadimplência. Essa iniciativa é crucial para minimizar os riscos financeiros, mantendo um equilíbrio entre responsabilidade financeira e inclusão social.

Kaggle Competition Link: [PoD Academy - Análise de Crédito]

2. Entendimento dos Dados

Para enfrentar esse desafio, a PoD Bank reuniu dados disponíveis e os disponibilizou em formato CSV para treinamento e avaliação de modelos.



Os principais conjuntos de dados incluem:

- application_train.csv/application_test.csv: Principais dados de treino/teste sobre solicitações de empréstimo, com identificação única (SK_ID_CURR) e variável TARGET indicando a inadimplência (0: pago, 1: não pago).
- previous_application.csv: Informações sobre aplicações de empréstimos anteriores dos clientes na PoD Bank.
- installments_payments.csv: Detalhes do histórico de pagamentos de empréstimos anteriores.
- bureau.csv: Dados de crédito de outras instituições financeiras.
- POS_CASH_balance.csv: Histórico de pagamentos de POS (Point of Sale) ou empréstimos em dinheiro.
- bureau_balance.csv: Informações mensais sobre créditos anteriores do cliente em outras instituições financeiras.
- credit_card_balance.csv: Informações mensais sobre saldos de cartões de crédito do cliente na PoD Bank.

Para os metadados, consulte o arquivo HomeCredit_columns_description.csv.

3. Preparação dos Dados

Feature Engineering:

Na etapa de Feature Engineering, criamos variáveis preditivas a partir das tabelas mencionadas, enriquecendo o conjunto de dados. Incrementamos variáveis em lotes para avaliação progressiva, adaptando dinamicamente as predições às nuances dos dados.

Entendimento do Público e Safras:

Compreendemos o público e a ausência de informações sobre safras. Sugerimos acompanhar tipos de créditos e utilizar Out-of-Sample na validação cruzada para mitigar viéses temporais, reforçando a robustez do modelo.

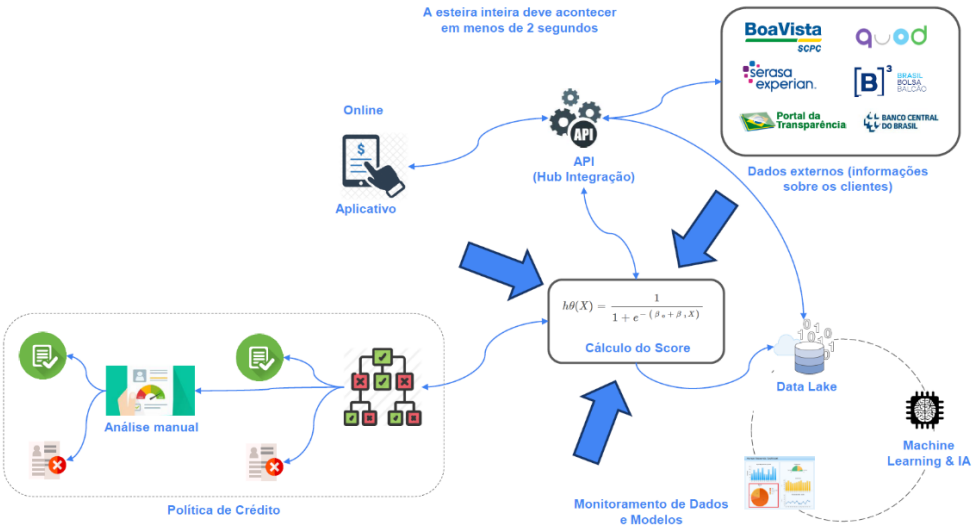
Tratamentos:

Excluímos variáveis com mais de 70% de nulos e substituímos nulos pela média. Utilizamos one-hot encoding para baixa cardinalidade, label encoding para alta

cardinalidade e selecionamos variáveis com feature importance (XGBoost). Na regressão logística, categorizamos variáveis contínuas.

4. Desenho da Arquitetura

Em resumo, desenvolvemos cinco modelos com equações distintas para encontrar o melhor cenário. A arquitetura envolveu a ingestão contínua e enriquecimento dos dados, garantindo modelos sempre atualizados e adaptáveis. Manipulamos variáveis explicativas estruturadamente, realizamos amostragem representativa e avaliamos modelos considerando KS, GINI e AUC.



5. Desenvolvimento dos Modelos

Modelos Desenvolvidos:

Criamos cinco modelos (LightGBM, XGBoost, Random Forest, Regressão Logística, Árvore de Decisão), treinando-os com grid search para otimização.

Treinamento e Otimização:

O treinamento incluiu otimização de hiperparâmetros, assegurando ajuste ideal para cada modelo.

Regressão Logística Específica:

Na regressão logística, além do treinamento convencional, avaliamos a linearidade das variáveis e categorizamos variáveis contínuas.

6. Avaliação dos Modelos

Durante o processo de avaliação dos modelos, foram considerados diversos critérios para garantir sua eficácia e aplicabilidade prática. A capacidade de implantação em produção foi um aspecto central, assegurando que o modelo pudesse ser integrado eficientemente no ambiente operacional da PoD Bank.

Além disso, métricas fundamentais, como o KS, o Gini e a ordenação do Score em 10 faixas na base de teste foram cuidadosamente analisadas. Essas métricas fornecem insights sobre a capacidade preditiva, a discriminação entre bons e maus pagadores, e a consistência da classificação em diferentes faixas de risco.

7. Desenvolvimento dos Indicadores de Negócio

Durante o desenvolvimento, a equipe analisou indicadores-chave de negócio para avaliar o impacto do modelo proposto. A Taxa de Inadimplência, que calcula a porcentagem de clientes inadimplentes, fornece uma medida crítica da eficácia do modelo na prevenção de riscos. O Valor Médio de Empréstimo oferece insights sobre o perfil econômico dos clientes atendidos, enquanto a Taxa de Aprovação de Crédito é um indicador crucial da aceitação e eficácia do modelo na aprovação de solicitações de crédito.

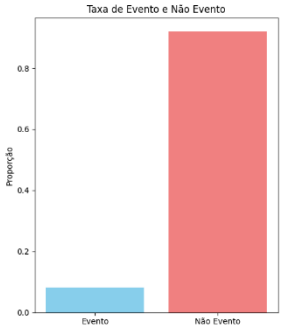
Esses indicadores não apenas servem como métricas de desempenho, mas também orientam estratégias futuras. A análise desses dados permite ajustes nas políticas de crédito, aprimoramento contínuo dos modelos e uma compreensão mais profunda das dinâmicas do mercado. O monitoramento constante desses indicadores é vital para garantir a adaptação eficaz às mudanças nas condições econômicas e comportamentais dos clientes.

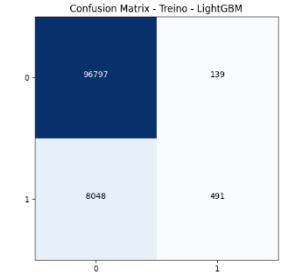
8. Resultados

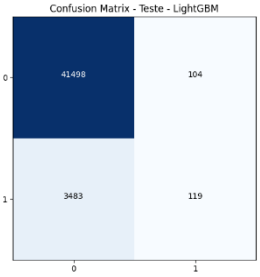
O modelo escolhido, LightGBM, se destaca pela ordenação eficaz das taxas de score, eficiência na identificação de clientes propensos a inadimplência e bom desempenho na métrica AUC (0.76).

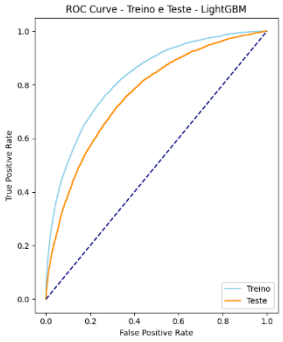
A consistência nas métricas sugere que o modelo não tem overfitting, garantindo robustez e generalização. O Modelo LightGBM permite uma taxa de aprovação

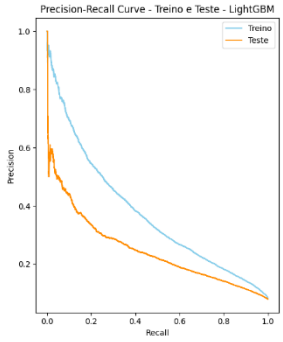
eficiente de 80% para 10% de apetite de risco, capturando R$33.2 bi entre os R$70 bi disponíveis.

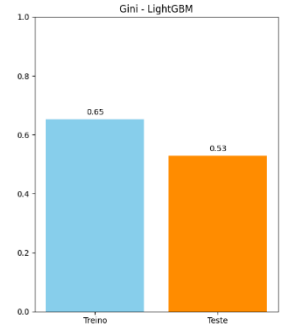


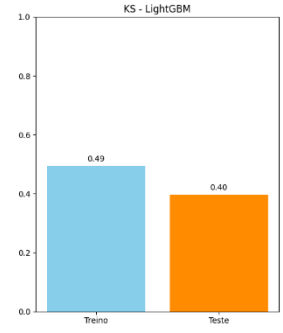


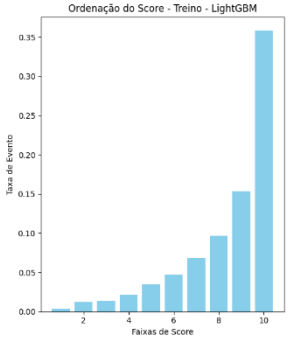


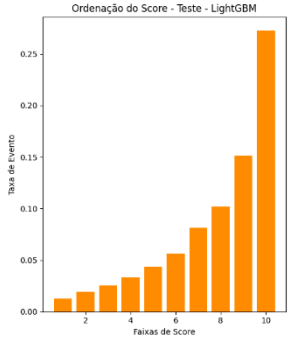












9. Implantação

Ao longo do desenvolvimento, foram salvos arquivos .pkl e .py para a implantação em produção, contendo informações para tratamento de nulos, categorização de variáveis, fórmulas de cálculo de escore e regras de ETL. A definição do apetite de risco cabe ao time de negócio.

10. Próximos Passos

Os próximos passos envolvem a confirmação do modelo e apetite de risco com o time de negócios, a definição de estratégia para incorporar o modelo no fluxo operacional e o incremento na coleta de informações de safra e tipos de crédito para o desenvolvimento futuro de modelos customizados.