

Datathon 2025

Proposta de Atividade

Você foi contratado como consultor de dados por um cliente do setor financeiro. Sua missão é transformar dados brutos de operações de crédito em insights acionáveis e em uma solução preditiva robusta.

O objetivo final deste desafio é desenvolver uma solução de ponta a ponta — da análise exploratória à avaliação e interpretabilidade de um modelo de machine learning — que capacite o cliente a otimizar sua operação de crédito e a tomar decisões mais inteligentes e seguras.

Conceitos Fundamentais

Para ter sucesso neste desafio, é crucial compreender os seguintes termos do domínio de risco de crédito:

- **Adimplência (Variável Alvo):** Nosso objetivo é prever a adimplência. A variável alvo (**adimplencia**) é baseada no conceito **M6EVER60**: ela indica se um cliente, após receber um empréstimo, atrasou um pagamento por mais de 60 dias em algum momento nos 6 meses seguintes à concessão (contratação do empréstimo).
 - **Atenção:** Na base de dados, o valor **1** indica que o cliente **pagou** (adimplente) e **0** indica que **não pagou** (inadimplente).
- **Safra:** Refere-se ao período (mês/ano) em que o crédito foi concedido. A análise por safras é essencial para avaliar a evolução do comportamento da carteira de clientes e a estabilidade dos modelos ao longo do tempo.
- **Métricas de Performance (Gini, KS, AUC):** São métricas padrão para avaliar modelos de classificação em risco de crédito. Elas medem a capacidade do modelo de discriminar entre "bons" e "maus" pagadores. Valores mais altos indicam um modelo com maior poder de separação.
- **Base de Treino e Teste:** A **base de treino** é a porção dos dados usada para ensinar e ajustar os parâmetros do modelo. A **base de teste** é uma amostra separada, usada para avaliar a performance do modelo em dados não vistos.
- **Base OOT:** A base OOT (Out-of-Time) é a porção de dados com safras posteriores às da Base de Treino e Teste. Normalmente ela é utilizada para avaliar se a performance e a distribuição do modelo são estáveis em datas posteriores às de desenvolvimento dos modelos.
 - **Distribuição de Scores:** Visualize a distribuição dos escores do modelo para diferentes safras da base OOT. Se a forma da distribuição mudar drasticamente, pode indicar instabilidade.
 - **Monitoramento de KS1 e KS2:** No contexto da estabilidade, ao invés de apenas um valor de KS, é útil monitorar o KS em diferentes safras (o que você poderia chamar de KS por safra, ou KS1 para uma safra e KS2 para outra). A ideia é observar se a capacidade de

separação do modelo se mantém consistente. Se o KS variar significativamente entre safras, pode indicar problemas de estabilidade.

Dados Disponibilizados

Você receberá dois arquivos no formato `.csv`:

1. **`base_des.csv`: Sua base de desenvolvimento.** Utilize este arquivo para todas as etapas de análise, treinamento e validação dos modelos. Ela contém as seguintes colunas:
 - `id`: Identificador único do cliente.
 - `safra_concessao`: Safra em que o crédito foi concedido.
 - `vp_1, vp_2, ..., vp_333`: Variáveis preditoras anonimizadas que descrevem o comportamento histórico do cliente.
 - `faixa_idade, faixa_renda, regioao`: Variáveis demográficas.
 - `adimplencia`: A variável alvo (1 = pagou, 0 = não pagou).
 2. **`base_oot.csv`: Sua base de pontuação final (Out-of-Time).** Este arquivo possui a mesma estrutura do anterior, **exceto pela ausência da variável `adimplencia`**. Não o utilize durante o desenvolvimento. Ao final do desafio, você deverá preenchê-lo com as predições do seu modelo e entregá-lo.
-

O Desafio

Descrição

Seu cliente espera uma solução robusta e bem documentada. Organize seu trabalho em um **Jupyter Notebook** ou **RMarkdown** para o desenvolvimento técnico e em uma **apresentação** para a comunicação dos resultados. O desafio está estruturado em três macroetapas.

As tarefas a seguir representam os requisitos mínimos. **Sinta-se à vontade para explorar técnicas adicionais e trazer novas análises que agreguem valor ao projeto.** Análises e ideias extra serão consideradas na avaliação.

Leia o desafio até o final para entender todo o passo a passo e os requisitos do mesmo. Entender tudo o que é pedido pode te ajudar na organização e planejamento do seu processo.

Ferramentas Sugeridas

Você tem liberdade para usar as bibliotecas que preferir. A seguir, algumas sugestões:

Python: `numpy`, `pandas`, `scikit-learn`, `lightgbm`, `xgboost`, `catboost`, `statsmodels`, `scipy`, `plotly`, `matplotlib`, `shap`.

R: `dplyr`, `tidyr`, `data.table`, `caret`, `glmnet`, `randomForest`, `tidymodels`, `lightgbm`, `xgboost`, `catboost`, `stats`, `glm`, `ggplot2`, `plotly`, `lattice`, `shapR`.

Etapa 1: Análise Exploratória de Dados (AED)

Contexto: Antes de construir qualquer modelo, um bom consultor mergulha nos dados. Uma AED bem executada revela padrões, identifica potenciais problemas e gera insights de negócio valiosos que guiarão a modelagem.

Sua Missão: Realizar uma análise exploratória focada em compreender o perfil dos clientes e o comportamento da inadimplência. Utilize principalmente as variáveis não anonimizadas (`faixa_idade`, `faixa_renda`, `regiao`, `safra_concessao`) para fundamentar suas análises.

Fluxo de Trabalho para Análise Exploratória

Para extrair os primeiros insights da base de dados, siga o roteiro de análise abaixo. Use a base de dados `base_des.csv` para as tarefas.

Tarefa 1.1: Análise do Perfil Geral dos Clientes

Analise e visualize o perfil geral dos clientes para entender a composição da base de dados. Apresente a distribuição dos clientes pelas seguintes dimensões:

- Faixa de Idade
 - Faixa de Renda
 - Região
-

Tarefa 1.2: Evolução da Inadimplência ao Longo do Tempo

Investigue a evolução da taxa de inadimplência ao longo do tempo. Calcule a taxa de inadimplência para cada `safra_concessao` e apresente os resultados em uma visualização (como um gráfico de linhas) que seja clara e de fácil interpretação para um público não técnico.

Tarefa 1.3: Análise da Inadimplência por Perfil de Cliente

Aprofunde a análise da inadimplência cruzando-a com os diferentes perfis de clientes. Calcule e compare a taxa de inadimplência para os diferentes grupos formados pelas variáveis demográficas (`faixa_idade`, `faixa_renda`, `regiao`).

Tarefa 1.4: Formulação de Hipóteses Iniciais

Organize as informações das tarefas anteriores para formular hipóteses preliminares. Com base nas análises demográficas, liste as características que mais parecem diferenciar os clientes adimplentes dos inadimplentes. O objetivo é gerar intuição para a etapa de modelagem.

Tarefa 1.5: Divisão da Base de Dados (Train/Test Split)

Como passo final da preparação, prepare os dados para as etapas de seleção de variáveis e modelagem. Realize a divisão da sua base de dados `base_des.csv` em um conjunto de **treino** e um conjunto de **teste**. Justifique a estratégia de divisão utilizada (ex: proporção, estratificação, etc.).

Etapa 2: Análise e Seleção de Preditores

Contexto: No mercado de crédito, é comum trabalhar com centenas de variáveis. Um modelo de sucesso não depende apenas de um bom algoritmo, mas de uma seleção inteligente de variáveis que sejam estáveis, informativas e que não contenham redundâncias. Nesta etapa, seu foco será avaliar a qualidade do "book de variáveis" disponibilizado.

Sua Missão: Executar uma análise aprofundada do conjunto de variáveis preditoras (`vp_1` a `vp_333`) para definir e justificar os conjuntos de features que serão utilizados nos modelos da Etapa 3.

Tarefa 2.1: Preparação dos Dados para Análise

Antes de iniciar as análises, você precisa preparar os dados. Assim como fará na modelagem, o primeiro passo é separar sua **base de treino** (criada na Tarefa 1.5).

Tarefa 2.2: Análise e Seleção de Preditores

Nesta etapa, você tem à disposição o conjunto completo de variáveis preditoras (`vp_1` a `vp_333`), além das variáveis demográficas.

Realize as seguintes análises sobre este conjunto de dados:

- **2.2.1 - Análise de Correlação:** As variáveis preditoras disponíveis possuem alta correlação entre si? Identifique os grupos de variáveis correlacionadas, discuta os problemas que isso poderia causar no modelo e sugira como tratar essa questão.
 - **2.2.2 - Análise de Estabilidade:** As variáveis demonstram estabilidade em suas distribuições ao longo do tempo (safras)? A presença de variáveis instáveis pode ser um risco para o modelo? Utilize métricas como PSI (Índice de Estabilidade da População) para fundamentar sua análise.
 - **2.2.3 - Análise de Poder Discriminatório:** Qual o poder de discriminação de cada variável em relação à `adimplencia`? Identifique as variáveis com maior capacidade de diferenciar pagadores de não-pagadores e descreva a metodologia utilizada (ex: Information Value, Gini, KS da variável).
-

Resultado Esperado da Etapa 2

Ao concluir esta etapa, você deve ter como resultado:

- **Uma lista de variáveis finalistas:** Uma para o modelo.

- **Uma justificativa clara para a lista**, baseada nas análises de correlação, estabilidade e poder discriminatório que você acabou de realizar.

Essa lista de features selecionadas será o insumo principal para a construção dos modelos na Etapa 3.

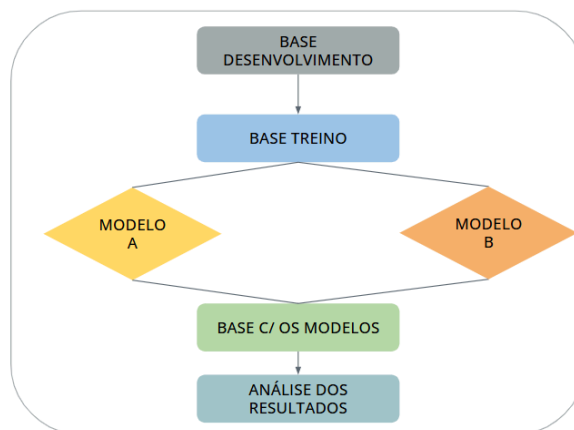
Etapa 3: Desenvolvimento do Modelo e Análise de Resultados

Contexto: Esta é a etapa mais importante do projeto. Aqui você construirá o modelo que prevê a adimplência de uma pessoa física. Um modelo preciso pode gerar grande valor financeiro, mas sua performance deve ser rigorosamente avaliada, e suas decisões, interpretáveis.

Sua Missão: Desenvolver, comparar e analisar os modelos de classificação, detalhando todo o processo de construção, validação e interpretação.

Fluxo de Trabalho para Modelagem e Avaliação

Para construir e validar sua solução, siga o fluxo de trabalho estruturado abaixo. O diagrama a seguir pode te ajudar a compreender o que será solicitado nas tarefas 3.1 e 3.2. Lembre-se: você pode adicionar novas técnicas de modelagem, além das descritas nas tarefas obrigatórias. Se o fizer, busque justificar as novas técnicas:



Tarefa 3.1: Treinamento do Modelo

O primeiro passo é construir os modelos para a sua amostra. Para isso, **considere as variáveis selecionadas na Etapa 2** e, caso seja necessário, refine essa seleção para cada algoritmo utilizado no desenvolvimento dos modelos. Seguir o passo a passo abaixo poderá ajudá-lo nessa etapa.

1. **Utilize a Base de Treino:** Separe sua base de treino (criada em 1.5).

2. **Treine o Modelo:** Para o dataframe, desenvolva **dois modelos**. O primeiro deverá utilizar obrigatoriamente a **Regressão Logística** e o segundo poderá utilizar qualquer outra técnica de **Machine Learning** que faça sentido para esse problema.

Dica 1: Realizar testes com diversos algoritmos para selecionar a melhor técnica de Machine Learning no desenvolvimento do segundo modelo é um diferencial que será considerado na avaliação, por isso, caso isso seja feito pelo grupo, deixe isso bem evidenciado na apresentação e nos códigos que serão enviados.

Dica 2: Os seguintes algoritmos de Machine Learning são bastante utilizados no mercado de crédito e costumam retornar bons resultados: XGBoost, LightGBM, CatBoost, Random Forest e Multi-Layer Perceptron (MLP)

3. **Resultado:** Ao final do passo anterior, você terá dois modelos treinados.

Tarefa 3.2: Escore e Consolidação dos Resultados na Base de Teste

Com os modelos treinados, o próximo passo é aplicá-los na base de teste e transformar as probabilidades brutas em um escore final padronizado, que varia de 1 a 999.

1. **Entendendo a Transformação do Escore:** No mercado de crédito, desejamos um escore de crédito onde **valores mais altos indicam menor risco (melhores pagadores)**. Como os modelos preveem a probabilidade de inadimplência ($P(\text{adimplencia}=1)$), você precisará convertê-la para uma escala de 1 a 999 que representa o "perfil de bom pagador".

Siga esta regra de transformação:

- **Calcule a probabilidade de ser um bom pagador:**

`prob_bom_pagador = P(adimplencia=1)`

- **Expanda a escala:** Multiplique o resultado por 1000.

`score_bruto = prob_bom_pagador * 1000`

- **Trunque para um valor inteiro:** Remova as casas decimais.

`score = trunc(score_bruto)`

- **Ajuste os Limites:** Garanta que o score final esteja entre 1 e 999.

i. Se o score for 0, ajuste para **1**.

ii. Se o score for 1000 ou maior, ajuste para **999**.

2. *Exemplo de cálculo:* Se o modelo previu uma $P(\text{adimplencia}) = 0.6253$, então:
`prob_bom_pagador = 0.6253`,

`score_bruto = 0.6253 * 1000 = 625.3`,

`score = trunc(625.3) = 625`. O score final é **625**.

3. **Utiliza a Base de Teste:** Separe sua base de teste (criada em 1.5).
 4. **Aplique os Modelos:** Escore o dataframe com os modelos correspondentes.
 5. **Organizando os Resultados:** Em um único dataframe (Base de Teste). Crie duas novas colunas para armazenar os escores consolidados:
 - **score_modelo_A:** Esta coluna deve conter o escore do modelo de **Regressão Logística**.
 - **score_modelo_B:** Esta coluna deve conter o escore do modelo de **Machine Learning**.
-

Tarefa 3.3: Análise de Performance

Com a base de teste escorada, avalie o poder preditivo dos seus modelos. Calcule as métricas de performance padrão de mercado (KS, Gini, AUC) para o **score_modelo_A** e **score_modelo_B**. Analise o desempenho ao longo do tempo (safras) por grupos demográficos.

Tarefa 3.4: Análise de Estabilidade do Modelo

Avalie a estabilidade da distribuição dos escores (**score_modelo_A** e **score_modelo_B**) ao longo das safras, utilizando as três primeiras safras da base de teste como população de referência. Métricas como PSI podem ser úteis nesta etapa.

Tarefa 3.5: Recomendação Final

Com base nas análises de performance (Tarefa 3.3) e estabilidade (Tarefa 3.4), qual dos modelos (A ou B) você recomendaria ao cliente? Apresente uma justificativa clara e fundamentada nos dados para sua escolha. Se além dos modelos obrigatórios, você incluiu modelos adicionais, considere-os também nesta avaliação.

Tarefa 3.6: Interpretabilidade do Modelo

Para o modelo que você recomendou, demonstre como suas decisões são tomadas.

- Identifique os **5 preditores mais importantes para o modelo**.
- Explique de forma clara como eles impactam a previsão (se aumentam ou diminuem a probabilidade de inadimplência).

Dica 1: Para o modelo de regressão logística pode ser realizada a análise dos coeficientes (betas).

Dica 2: Técnicas como SHAP podem ser muito úteis para modelos de boosting.

Tarefa 3.7: Escorar a Base OOT (Out-of-Time)

Como passo final, utilize os modelos desenvolvidos para gerar os escores `score_modelo_A` e `score_modelo_B` na base `base_oot.csv`. O processo de escoragem será idêntico ao que você fez na base de teste. Salve o arquivo da base com os modelos escorados, pois ele é um dos entregáveis do desafio.

Entregáveis

Ao final do desafio, você deverá submeter:

1. **Códigos com a resolução:**
 - 1.1. **Jupyter Notebook (.ipynb):** Um notebook contendo todo o código, análises, visualizações e respostas técnicas. O código deve ser limpo, bem comentado e reproduzível.
 - 1.2. **RMarkdown (.Rmd):** Se preferir, um arquivo RMarkdown para o desenvolvimento do projeto em R.
 2. **Apresentação (PDF):** Uma apresentação executiva com as principais análises, resultados e a recomendação final para o cliente.
 3. **Base de Dados Escorada:** O arquivo `base_oot.csv` preenchido com as colunas de score dos seus modelos.
-