Ficha Técnica: Projeto de Análise de Dados

Título do Projeto: Super Caja

Objetivo: Automatizar o processo de análise de crédito do banco Super Caja, utilizando técnicas de análise de dados para:

- Identificar perfis de clientes com maior risco de inadimplência.
- Criar uma pontuação de crédito baseada em variáveis comportamentais e socioeconômicas.
- Avaliar o risco relativo de inadimplência.
- Classificar clientes e futuros clientes em categorias de risco para subsidiar decisões de concessão de crédito.

Ferramentas e Tecnologias: Google BigQuery, Google Colab, Apresentações Google: ferramenta para criação e edição de apresentações e Google Looker Studio.

Processamento e análises:

1. **Identificar e tratar valores nulos**
    a. Foi realizada uma consulta para identificar a incidência de valores nulos na variável **last_month_salary** da tabela users, segmentada pelo status de inadimplência (default_flag). Os resultados foram os seguintes:
        - Total de clientes bons pagadores (default_flag = 0): 35.317
        - Quantidade de nulos em last_month_salary: 7.069
        - Percentual de nulos: 20,02%
        - Total de clientes maus pagadores (default_flag = 1): 683
        - Quantidade de nulos em last_month_salary: 130
        - Percentual de nulos: 19,03%
        - A proporção de valores nulos em last_month_salary é bastante semelhante entre bons e maus pagadores, não caracterizando viés relevante em relação à variável default_flag. Por isso, decidiu-se não excluir os registros nulos, e sim tratá-los substituindo os valores faltantes pela mediana de salário por grupo de inadimplência, preservando a integridade dos dados e garantindo a participação desses registros na análise de risco de crédito.
    b. A variável **number_of_dependents** apresentou 943 valores nulos, representando 2,6% dos bons pagadores e 1,5% dos maus pagadores. Analisando a distribuição, identificou-se que o valor mais frequente (moda) é 0 dependentes, representando mais de 50% dos registros. Considerando o percentual reduzido de nulos e a forte concentração neste valor, optou-se por substituir os valores ausentes pela moda.

c. Não foram encontrados nulos em outras tabelas.

2. **Identificar e tratar valores duplicados**
    a. Durante a análise de duplicidades nos dados, foi realizada a verificação utilizando as funções GROUP BY e HAVING para identificar registros duplicados nas tabelas do banco de dados. Os valores duplicados foram encontrados apenas na tabela **loans_outstanding**. Essa duplicidade foi considerada válida e mantida, pois cada usuário pode possuir múltiplos empréstimos ativos simultaneamente, o que justifica a existência de múltiplos registros com o mesmo identificador de usuário nessa tabela.
    b. Nas demais tabelas, não foram identificados registros duplicados que comprometessem a integridade dos dados.

3. **Identificar e gerenciar dados fora do escopo da análise**
    a. Foi identificada alta correlação entre as variáveis **more_90_days_overdue, number_times_delayed_payment_loan_30_59_days** e **number_times_delayed_payment_loan_60_89_days** (coeficientes acima de 0,98), indicando multicolinearidade. Para evitar redundância e melhorar a qualidade do modelo, optou-se por manter apenas a variável **number_times_delayed_payment_loan_30_59_days**, que apresentou o maior desvio padrão e, portanto, maior variabilidade e representatividade dos dados. As outras duas variáveis foram removidas da análise.

4. **Identificar e tratar dados inconsistentes em variáveis categóricas**
    a. Foi identificada inconsistência nos valores da variável **loan_type**, devido à variação de maiúsculas, minúsculas e plurais, como "OTHER", "Other", "other", "others" e "REAL ESTATE", "Real Estate", "real estate". Para garantir a padronização e evitar inconsistências na análise, optou-se por uniformizar todos os valores para maiúsculas com a função UPPER() e agrupar variações semânticas equivalentes utilizando CASE WHEN. Com isso, os valores foram consolidados em categorias únicas, como "OTHER" e "REAL ESTATE", eliminando redundâncias e facilitando o tratamento dos dados categóricos.

5. **Identificar e tratar dados discrepantes em variáveis numéricas**
    a. Foi identificado um conjunto significativo de outliers na variável **last_month_salary** da base de dados. Em especial, constatou-se a presença de valores iguais a zero, que foram considerados como outliers por não representarem salários válidos. Além disso, foi estabelecido um limite arbitrário inferior fixado em R$1518, equivalente ao salário mínimo vigente,

para sinalizar salários muito baixos, uma vez que o limite inferior estatístico (baseado no IQR) resultou em valor negativo e, portanto, não refletia uma referência realista para a variável salário. Foram encontrados 378 registros com salário zero, e 1846 registros com salário abaixo do limite arbitrário de R$1518, indicando uma concentração relevante de valores baixos que podem demandar análise ou tratamento adicional. Já na faixa superior, foram identificados 2186 registros com salários acima do limite superior calculado estatisticamente, configurando outliers altos que podem indicar possíveis inconsistências, erros de digitação ou casos especiais que requerem atenção.

b. Foi identificado um conjunto de 10 outliers na variável **age**, considerando como critério valores fora do intervalo estatístico (calculado via método do IQR). As idades identificadas como outliers variaram de 97 a 109 anos, significativamente acima da distribuição típica observada na base.
c. Foi identificada a presença de 3.230 outliers na variável **number_dependents**. Para definição desses outliers, adotou-se como limites inferior e superior -1,5 e 2,5, respectivamente. Como o número de dependentes não pode ser negativo, e considerando o comportamento típico da variável, valores acima de 3 dependentes foram classificados como outliers.
d. Foi identificada a presença de 177 outliers na variável **using_lines_not_secured_personal_assets**. Todos os outliers identificados apresentaram valores acima do limite superior, caracterizando situações em que o valor utilizado pelo cliente em linhas de crédito não garantidas por bens pessoais está desproporcionalmente elevado em relação à distribuição da base. Essa identificação é importante para mitigar o impacto de valores extremos na modelagem de risco e na análise de perfil de crédito
e. Foi identificada a presença de 7.570 outliers na variável **debt_ratio**. Para definição desses outliers, adotou-se os limites inferior e superior de -0,90 e 1,97, respectivamente. Como se trata de uma variável de razão, valores negativos não são possíveis na prática e, neste caso, todos os outliers identificados estavam acima do limite superior estabelecido. Esses casos podem indicar clientes com grau de endividamento excessivamente elevado em relação ao seu patrimônio, o que exige atenção especial na análise de risco.
f. Foi decidido não alterar os outliers.

6. Criar novas variáveis e Unir tabelas

a. Foram identificados 36562 empréstimos para imóveis e 268773 para outras categorias.
b. Foi identificado que 425 usuários não estavam presentes na base loans_outstanding.
c. Foi realizada a integração das bases users_limpo, loans_outstanding, loans_detail e default por meio da variável user_id, com o objetivo de mapear a situação de cada pessoa usuária em relação aos seus empréstimos.

Inicialmente, foi feita uma junção LEFT JOIN entre users_limpo e loans_outstanding, permitindo identificar os usuários presentes na base de clientes, mas sem registros de empréstimos ativos.

d. Para isso, foi aplicada a condição WHERE l.user_id IS NULL após o LEFT JOIN, retornando os usuários que não possuem informações na tabela loans_outstanding. Assim, foi possível quantificar os casos e analisar o perfil desses clientes.
e. Além disso, para qualificar a situação de cada usuário em relação aos empréstimos, foi criada a variável final_loan_status, com as seguintes categorias:
    - 'known': quando o usuário possui empréstimos registrados na tabela loans_outstanding.
    - 'unknown': quando o usuário não aparece em loans_outstanding, mas possui histórico nas tabelas loans_detail ou default, indicando provável relacionamento com empréstimos, mas sem detalhes sobre os contratos ativos
    - 'none': quando o usuário não possui registro em nenhuma das bases relacionadas a empréstimos.
f. Essa classificação permitiu compreender melhor os diferentes perfis e situações de clientes no banco de dados, além de identificar lacunas importantes de informação para o ajuste e aprimoramento das análises.

7. Análise Exploratória

a. Distribuição por Faixa Etária

- **41 a 60 anos**: 16.803 usuários (**42,5% do total**).
- **61 a 80 anos**: 9.689 usuários (**24,5%**).
- **25 a 40 anos**: 7.830 usuários (**19,8%**).
- Faixas extremas, como **Menor de 25 anos (1,2%)** e **Acima de 90 anos (0,3%)**, possuem representatividade marginal.

Conclusão: O perfil etário da base é predominantemente composto por indivíduos de meia-idade e idosos, o que pode influenciar diretamente o comportamento de consumo e risco de crédito.

b. Distribuição Salarial

Em relação à renda declarada (ou imputada), a maior parte da base concentra-se nas seguintes faixas:

- **R$ 5.001 a R$ 10.000**: 11.097 usuários (**28,1%**).
- **R$ 3.001 a R$ 5.000**: 7.266 usuários (**18,4%**).
- **R$ 1.501 a R$ 3.000**: 4.222 usuários (**10,7%**).
- Um ponto relevante é a alta proporção de usuários com salário **não informado** (**7.577 usuários — 19,2%**).

Conclusão: A base é composta majoritariamente por indivíduos com renda entre **R$ 3.001 e R$ 10.000**, sugerindo uma faixa de poder aquisitivo médio a alto. A ausência de informação salarial para aproximadamente **um quinto da base** pode limitar análises preditivas baseadas em renda.

c. Número de Dependentes

A distribuição de dependentes apresenta o seguinte perfil:

- **Sem dependentes:** 20.912 usuários (**53%**).
- **1 a 2 dependentes:** 10.915 usuários (**27,7%**).
- **3 a 5 dependentes:** 3.165 usuários (**8%**).
- **Mais de 6 dependentes:** 65 usuários (**0,2%**).
- **Não informado:** 943 usuários (**2,4%**).

Conclusão: A maioria significativa da base não possui dependentes ou possui até dois, o que tende a reduzir a pressão financeira sobre a renda familiar e pode impactar positivamente a capacidade de pagamento.

d. Faixa de Comprometimento de Renda (Debt Ratio)

A variável **debt_ratio** (razão entre dívidas e renda) apresenta a seguinte distribuição:

- **Entre 0,2 e 0,5:** 12.294 usuários (**31,1%**).
- **Abaixo de 0,2:** 10.109 usuários (**25,6%**).
- **Acima de 1:** 8.461 usuários (**21,4%**).

- **Entre 0,51 e 1:** 4.902 usuários (**12,4%**).

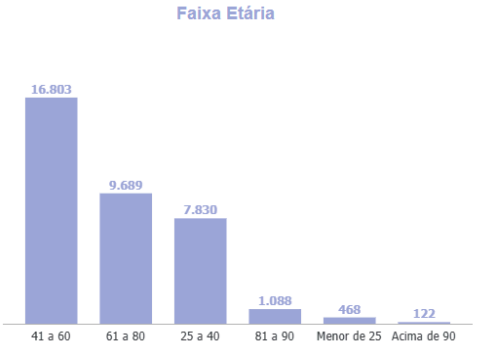
Conclusão: Embora a maior parte da base apresente comprometimento de renda moderado, há uma parcela expressiva (**21,4%**) com comprometimento superior a 100%, o que configura elevado risco de inadimplência.

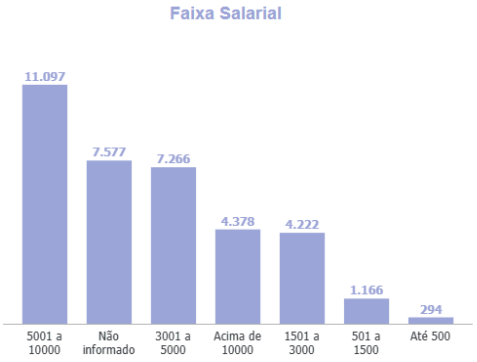
e. **Uso de Crédito Sem Garantia**

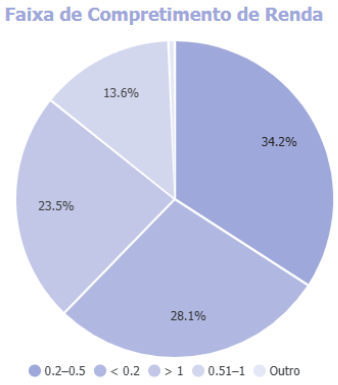
Quanto à utilização de linhas de crédito não garantidas:

- **Abaixo de 20%:** 19.712 usuários (**49,9%**).
- **Entre 20% e 50%:** 6.483 usuários (**16,4%**).
- **Acima de 80%:** 5.862 usuários (**14,8%**).
- **Entre 51% e 80%:** 3.760 usuários (**9,5%**).

Conclusão: A maioria dos usuários possui baixo nível de utilização de crédito sem garantia. No entanto, cerca de **24,3%** utilizam mais de 50% da sua capacidade de crédito, o que, combinado com alto debt ratio, pode elevar o risco de inadimplência neste segmento.







f. Através das medidas de tendência central foi observado que:

- **Inadimplentes são mais jovens** (média de 44 anos vs 52 anos).
- **Salário médio e mediano de inadimplentes é menor** (R$ 4.549 vs R$ 6.508).
- **Inadimplentes tendem a ter menos dependentes** (média de 0,73 vs 1).
- **Faixas salariais mais críticas:** de **R$ 1.501 a R$ 5.000** concentram mais inadimplentes.
- **Faixa etária de 41 a 80 anos** concentra mais inadimplentes em volume — precisa ver a proporção.
- **Pessoas sem dependentes** têm maior volume de inadimplência.
- **Quanto maior o debt ratio**, maior o volume de inadimplência — mas muitos adimplentes também mantêm ratios altos.
- **Uso alto de crédito sem garantia** está associado a maior inadimplência.

g. Através do cálculo de taxa de inadimplência por faixas foi observado que:

- Faixa Salarial
  - **1501–3000** apresenta a maior taxa de inadimplência: **3,32%**, seguida de perto por **501–1500** (3,09%) e **3001–5000** (2,48%).
  - Curiosamente, quem ganha **> 10.000** e **Até 500** tem as menores taxas (0,57% e 0,68%, respectivamente).
  - **Não informado** (19,2% da base) tem inadimplência de **1,78%**, o que sugere que a falta de dado não implica, necessariamente, maior risco.
- Faixa Etária
  - O grupo **25–40 anos** lidera a taxa de inadimplência: **3,55%**.
  - **Menor de 25** apresenta 2,14%, e **41–60**: 1,95%.
  - Idades acima de 60 anos mostram taxas muito baixas (< 1%), chegando a zero em **> 90 anos**.
- Faixa de Dependentes
  - Maior risco em **"Mais de 6" dependentes**: 4,62%.
  - Em seguida, **3–5 dependentes**: 2,97%.
  - Clientes sem dependentes apresentam 1,61%.
- Uso de Crédito Sem Garantia
  - Taxa extremamente alta em **> 80%** de utilização: **9,83%** — quase 10% de inadimplentes.
  - Faixas intermediárias (51–80%) têm 1,88%, e uso baixo (< 20%) praticamente zero (0,05%).
- Faixa de Debt Ratio
  - Maior taxa em **0.51–1**: 2,92%.
  - Ratios < 0.2 e > 1 têm taxas semelhantes (~1,8–1,9%).

- Faixa 0.2–0.5 apresenta o menor risco (1,53%).
- Um debt ratio moderadamente alto (até 100% da renda) é mais arriscado que endividamentos extremos (> 100%), talvez porque usuários já muito endividados não buscam novas linhas de crédito.

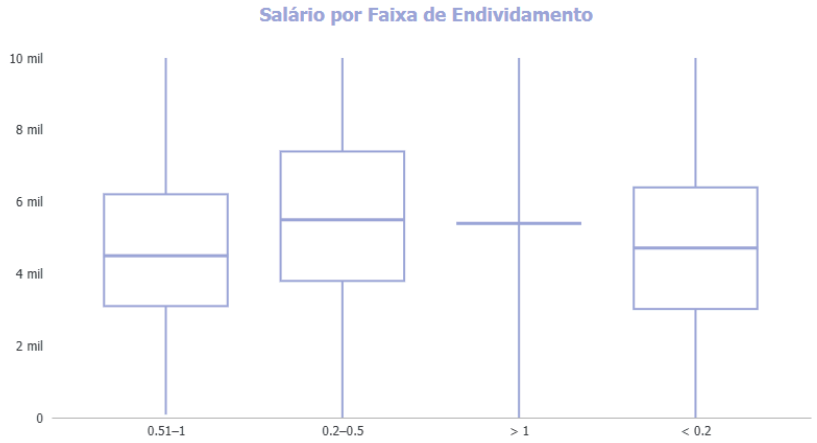
h. Insight Integrado:

- **Perfil Médio (Média/Mediana)** indica que o "cliente típico" inadimplente é mais jovem (≈ 44 anos) e ganha menos (≈ R$ 4.549) do que o típico adimplente. Isso resume onde está concentrado o centro da distribuição dos inadimplentes.
- **Taxa por Faixa**, porém, mostra que a **chance** de default é maior para quem tem **25–40 anos** (3,55%) e ganha entre **R$ 1.501–3.000** (3,32%). Ou seja, nem sempre a faixa em que o maior volume de inadimplentes está (p.ex. 41–60 anos) é a que apresenta maior risco relativo.
- A faixa etária **41–60** tem o maior número absoluto de inadimplentes (328 pessoas), mas sua taxa de default (1,95%) é bem inferior à de 25–40 anos (3,55%). Isso acontece porque 41–60 é uma faixa populacional muito grande na base — logo, concentra volume mas dilui a proporção de inadimplência. Já faixas com menos clientes podem apresentar **riscos concentrados** (alto percentual de default) ainda que em números absolutos sejam menores.
- A média de salário dos inadimplentes (≈ R$ 4.549) pode ser puxada para cima por alguns saldos mais altos, mascarando que a **faixa de renda média-baixa (R$ 1.501–3.000)** tem a maior taxa de inadimplência (3,32%).

i. Boxplot

- **Distribuição semelhante entre as faixas**
  - Em todas as faixas de debt ratio, o salário varia bastante — com mínimo próximo de 0 e máximo em 10.000.
  - A mediana está entre **4.500 e 6.500**, variando pouco entre as faixas.
  - **Faixa 0.2–0.5 tem a maior mediana:** Mediana ligeiramente acima de **6 mil**, maior do que as outras faixas.
  - **Faixa > 1:** Tem um comportamento interessante: a mediana está na casa dos **5 mil**, mas a caixa (entre 1º e 3º quartis) está mais estreita, indicando menos variação nos salários para essa faixa de endividamento alto.

- **Faixa < 0.2 e 0.51–1:** Apresentam distribuições parecidas: mediana próxima de 4.700 e grande dispersão.



8. **Técnica de Análise (Risco Relativo)**
    a. **Faixa Salarial - Quartis**

Para analisar o comportamento da inadimplência em função da remuneração, os dados de salário foram divididos em **quatro quartis**, com base na distribuição dos valores. Os limites definidos para os quartis foram:

- **Q1:** até **R$ 4.000,00**
- **Q2:** de **R$ 4.000,01 a R$ 5.400,00**
- **Q3:** de **R$ 5.400,01 a R$ 7.416,00**
- **Q4:** acima de **R$ 7.416,00**

A **taxa geral de inadimplência da base** foi de **1,89%**.

Ao calcular o **risco relativo de inadimplência** para cada quartil em relação à taxa geral, observou-se que:

- O **maior risco relativo** está no **Q1** (salários até R$ 4.000,00), com risco **2,32 vezes maior** do que a média geral.
- Em seguida, o **Q3** (entre R$ 5.400,01 e R$ 7.416,00) apresentou um risco relativo de **0,93**, próximo da média.

- Os **Q2** e **Q4** (faixas intermediária baixa e alta) mostraram os menores riscos, com valores de **0,44** e **0,43**, respectivamente.

Conclusão: há um indicativo claro de que pessoas com salários mais baixos (**até R$ 4.000,00**) apresentam maior vulnerabilidade à inadimplência, enquanto as demais faixas possuem risco reduzido ou próximo da média.

b. Faixa Etária - Quartis

Os dados de idade foram segmentados em **quatro quartis**, com os seguintes limites:

- **Q1:** até **41 anos**
- **Q2:** de **41,01 a 52 anos**
- **Q3:** de **52,01 a 63 anos**
- **Q4:** acima de **63 anos**

A **taxa geral de inadimplência da base** permanece em **1,89%**.

Principais Resultados:

- O **maior risco relativo** foi observado no **Q1** (até 41 anos), com risco **1,81 vezes maior** do que a média geral.
- Na sequência, o **Q2** (entre 41,01 e 52 anos) apresentou um risco relativo de **1,17**, ligeiramente acima da média.
- Os quartis **Q3** e **Q4** (acima de 52 anos) mostraram riscos reduzidos, com **0,70** e **0,26**, respectivamente.

Conclusão: A análise evidencia que pessoas mais jovens (**até 41 anos**) tendem a apresentar maior inadimplência no banco, enquanto o risco diminui progressivamente com o aumento da idade, sendo significativamente menor a partir dos **63 anos**. Isso sugere que a faixa etária é um fator relevante na avaliação de risco de crédito, com jovens apresentando maior vulnerabilidade financeira no contexto analisado.

c. Número de Dependentes - Quartis

Dividindo a variável **dependents_imputed** em quatro quartis com os seguintes limites:

- **Q1:** até **0 dependentes**

- **Q2:** até **0 dependentes**
  (obs: como Q1 e Q2 têm o mesmo valor, significa que boa parte das pessoas tem zero dependentes — distribuição concentrada)

- **Q3:** até **1 dependente**

- **Q4:** acima de **1 dependente**

A taxa geral de inadimplência no conjunto é de **1,89%**.

Analisando os riscos relativos:

- Pessoas no **Q4** (mais de **1 dependente**) apresentam o **maior risco relativo**, com **1,33** vezes a taxa de inadimplência geral.

- Em seguida, o **Q3** (até **1 dependente**) com risco relativo de **1,15**.

- Pessoas no **Q1** (zero dependentes) apresentam **menor risco relativo**, com **0,84**.

Conclusão: Assim como no salário e idade, esse resultado mostra que o risco de inadimplência **aumenta conforme o número de dependentes cresce**, especialmente a partir de **2 dependentes ou mais (Q4)**.
Esse comportamento faz sentido, pois mais dependentes podem significar maiores despesas e, consequentemente, maior risco financeiro.

Após a realização dos cálculos utilizando quartis, foi possível perceber que essa segmentação não trouxe classificações tão precisas quanto esperado, pois os intervalos definidos pelos quartis eram muito amplos, agrupando perfis bastante distintos dentro de uma mesma categoria.

As faixas previamente estabelecidas pela analista, por outro lado, se mostraram mais eficazes para a análise, já que consideram intervalos mais específicos e alinhados à realidade do negócio, possibilitando a identificação de comportamentos de inadimplência de forma mais clara e acionável. Dessa forma, optou-se por seguir com as faixas definidas manualmente, por trazerem insights mais relevantes e direcionados.

d. Número de Dependentes

Os dados mostram que a taxa de inadimplência varia significativamente conforme a quantidade de dependentes, quando comparada à taxa de inadimplência do grupo complementar (ou seja, todas as outras faixas juntas):

- **Faixa "Mais de 6 dependentes"** apresenta a maior taxa de inadimplência (4,62%) e o risco relativo mais elevado (2,44), indicando que o risco de inadimplência desse grupo é mais que o dobro da soma dos demais grupos.

- **Faixa "3 a 5 dependentes"** tem uma taxa de inadimplência de 2,97% e risco relativo de 1,66, sugerindo um risco 66% maior do que o restante da população considerada.

- **Faixa "1 a 2 dependentes"** apresenta inadimplência de 2,19% e risco relativo de 1,24, evidenciando um risco 24% superior ao grupo complementar.

Esse padrão indica uma tendência clara: quanto maior o número de dependentes, maior é o risco relativo de inadimplência. A análise por faixa, comparando cada grupo com o restante da amostra, oferece uma visão mais detalhada e precisa do comportamento dos diferentes perfis, permitindo ações específicas e direcionadas para mitigar riscos.

e. Faixa Etária

Os dados mostram que a taxa de inadimplência varia significativamente conforme a faixa etária, quando comparada à taxa de inadimplência do grupo complementar (ou seja, todas as outras faixas juntas):

- **Faixa "25 a 40 anos"** apresenta a maior taxa de inadimplência (3,55%) e o risco relativo mais elevado (2,47), indicando que o risco de inadimplência desse grupo é **2,5 vezes maior** do que a soma dos demais grupos.

- **Faixa "Menor de 25 anos"** tem uma taxa de inadimplência de 2,14% e risco relativo de 1,13, sugerindo um risco **13% maior** do que o restante da população considerada.

- **Faixa "41 a 60 anos"** apresenta inadimplência de 1,95% e risco relativo de 1,06, evidenciando um risco **levemente superior (6%)** em relação ao grupo complementar.

- **Faixa "61 a 80 anos"** apresenta uma taxa de inadimplência de 0,65% e risco relativo de 0,28, indicando um risco **72% menor** do que os demais grupos.

- **Faixa "81 a 90 anos"** registra inadimplência de 0,37% e risco relativo de 0,19, demonstrando um risco **81% inferior** ao restante da amostra.

- **Faixa “Acima de 90 anos”** não registrou inadimplência, com taxa de 0% e risco relativo de 0, representando o grupo de menor risco na amostra.

Esse padrão indica uma tendência clara: **o risco de inadimplência é mais elevado entre pessoas com idade entre 25 e 40 anos**, reduzindo progressivamente nas faixas etárias superiores. A análise por faixa etária, comparando cada grupo com o restante da amostra, oferece uma visão mais detalhada e precisa do comportamento dos diferentes perfis, permitindo ações específicas e direcionadas para mitigar riscos — como reforço de análise de crédito e políticas diferenciadas para faixas críticas.

f. Faixa Salarial

Os dados indicam que a inadimplência **varia significativamente de acordo com a faixa salarial**, com comportamentos distintos:

- Faixa **“1501 a 3000”** apresenta uma taxa de inadimplência de **3,32%** e risco relativo de **1,94**, indicando um risco **94% maior** do que os demais grupos.
- Faixa **“501 a 1500”** possui taxa de inadimplência de **3,09%** e risco relativo de **1,66**, representando um risco **66% superior** ao grupo complementar.
- Faixa **“3001 a 5000”** registra inadimplência de **2,48%** e risco relativo de **1,42**, apontando um risco **42% maior** em relação aos demais.
- Faixa **“Não informado”** apresenta taxa de inadimplência de **1,78%** e risco relativo de **0,92**, indicando um risco **8% menor** do que os outros grupos.
- Faixa **“5001 a 10000”** tem taxa de inadimplência de **1,49%** e risco relativo de **0,71**, ou seja, um risco **29% inferior**.
- Faixa **“Até 500”** exibe inadimplência de **0,68%** e risco relativo de **0,36**, indicando risco **64% menor**.
- Faixa **“Acima de 10000”** apresenta a menor inadimplência, com **0,57%**, e risco relativo de **0,27**, ou seja, um risco **73% inferior** ao restante da amostra.

A análise revela que o risco de inadimplência é mais elevado entre pessoas com renda mensal intermediária, especialmente nas faixas entre R$501 e R$3.000. A partir das faixas de renda acima de R$5.000, o risco reduz progressivamente, sendo significativamente menor entre os que recebem acima de R$10.000. Além disso, pessoas que não informaram a renda não apresentam aumento de risco, mantendo um comportamento de inadimplência semelhante ao da média geral. Esse padrão sugere que o risco de crédito não está concentrado nas faixas mais baixas, mas sim nos perfis de renda intermediária.

g. Crédito Sem Garantia

Os dados indicam que a inadimplência varia significativamente conforme a faixa de crédito sem garantia, com comportamentos muito distintos:

- Faixa **"> 80%"** apresenta uma taxa de inadimplência de **9,83%** e risco relativo de **27,68**, indicando um risco mais de **27 vezes maior** do que os demais grupos.
- Faixa "**51–80%**" possui taxa de inadimplência de **1,88%** e risco relativo de **0,99**, representando um **risco semelhante** ao grupo complementar.
- Faixa "**20–50%**" registra inadimplência de **0,35%** e risco relativo de **0,16**, apontando um risco **84% menor** em relação aos demais.
- Faixa "**< 20%**" tem a menor inadimplência, com **0,05%**, e risco relativo de **0,01**, ou seja, um risco **99% inferior** ao restante da amostra.

A análise revela que o risco de inadimplência aumenta drasticamente em clientes com mais de 80% do crédito sem garantia, indicando alta vulnerabilidade e necessidade de monitoramento rigoroso. Já as faixas inferiores a 80% mostram riscos significativamente menores, principalmente abaixo de 50%, onde o risco é praticamente nulo. Isso sugere que o controle da proporção de crédito sem garantia é fundamental para a gestão eficiente do risco de crédito e que estratégias diferenciadas devem ser aplicadas conforme o nível dessa exposição.

h. Debt Ratio

Os dados indicam que a inadimplência varia conforme a faixa da razão dívida/renda, apresentando diferenças claras:

- Faixa "**0.51–1**" apresenta uma taxa de inadimplência de **2,92%** e risco relativo de **1,69**, indicando um risco **69% maior** do que os demais grupos.
- Faixa "> 1" possui taxa de inadimplência de **1,88%** e risco relativo de **0,99**, representando um **risco semelhante** ao grupo complementar.
- Faixa "< 0.2" registra inadimplência de **1,84%** e risco relativo de **0,96**, apontando um risco **4% menor** em relação aos demais.
- Faixa "**0.2–0.5**" apresenta a menor inadimplência, com **1,53%**, e risco relativo de **0,73**, ou seja, um risco **27% inferior** ao restante da amostra.

A análise revela que o risco de inadimplência é mais elevado em clientes com dívida entre 51% e 100% da renda, evidenciando maior vulnerabilidade financeira nessa faixa. Já para aqueles com dívida superior à renda (> 100%), o risco não aumenta, mantendo-se próximo da média geral. Clientes com menor razão dívida/renda (< 50%) apresentam risco reduzido, especialmente na faixa de 20% a 50%, sugerindo maior capacidade de pagamento. Essas informações indicam que o controle do nível da dívida em relação à renda é importante para gestão do risco, e o monitoramento deve focar nos clientes que apresentam níveis intermediários dessa relação.

9. Aplicar segmentação por Score

a. Dummies

Foi decidido transformar em variáveis dummy apenas as categorias associadas a um risco relativo significativamente superior ao baseline (RR > 1), pois são essas que efetivamente aumentam a probabilidade de inadimplência e possuem valor discriminativo para o modelo.

No caso da variável Número de Dependentes, foi definido incluir as faixas 3 a 5 dependentes (RR=1,66) e mais de 6 dependentes (RR=2,44), por apresentarem riscos relativos superiores a 1.

Para a variável Faixa Etária, foi decidido considerar a faixa 25 a 40 anos (RR=2,47), devido à sua taxa de inadimplência expressivamente acima das demais.

Na variável Faixa Salarial, foi estabelecido incluir as faixas R$501 a R$1500 (RR=1,66) e R$1501 a R$3000 (RR=1,94), ambas com riscos relativos acima de 1 e relevantes para a análise de crédito.

Em relação ao Crédito Sem Garantia, foi decidido criar uma dummy apenas para a faixa acima de 80% (RR=27,68), uma vez que o risco relativo dessa faixa é desproporcionalmente elevado em comparação com os demais intervalos.

Por fim, para a variável Debt Ratio, foi definido incluir a faixa 0,51 a 1 (RR=1,69), que apresentou risco superior ao baseline e potencial para contribuir na classificação de risco dos clientes.

Essa seleção foi conduzida com o objetivo de manter o modelo enxuto, priorizando categorias com maior poder preditivo e eliminando variáveis com risco igual ou inferior a 1, a fim de reduzir ruído e evitar sobreajuste.

b. Corte de Score e Matriz da Confusão

Após análise detalhada dos resultados do modelo preditivo de inadimplência, foi decidido adotar o corte de score igual ou superior a 2 para a classificação dos clientes como de maior risco. Essa decisão considerou os principais indicadores de desempenho do modelo — recall, precisão, acurácia e F1 score — bem como o contexto operacional e o histórico de comportamento dos clientes.

Com o corte ≥ 2, o modelo apresenta um **recall elevado**, identificando 625 verdadeiros positivos (mau pagadores corretamente previstos), o que indica uma boa capacidade de captura dos clientes que realmente apresentam risco de inadimplência. Apesar de haver um número maior de falsos positivos (8.257), é importante destacar que cerca de 16,76% desses clientes já possuem histórico de atraso entre 30 e 59 dias, embora ainda não estejam

marcados como inadimplentes na variável default_flag. Esse dado reforça que o modelo sinaliza potenciais riscos latentes, contribuindo para uma gestão mais preventiva e eficaz.

Em termos de **precisão**, o modelo apresenta um valor menor neste ponto de corte, reflexo natural do trade-off com o recall. Porém, a prioridade neste caso foi maximizar a captura dos casos de inadimplência, evitando o custo maior de deixar riscos passarem despercebidos.

A **acurácia geral** do modelo também permanece robusta, sustentada pela alta taxa de verdadeiros negativos (27.060), mostrando que a maioria dos clientes sem risco é corretamente identificada.

O **F1 score**, métrica que pondera equilíbrio entre precisão e recall, apoia essa escolha, indicando que o modelo está bem ajustado para este cenário de análise, considerando a importância crítica da detecção precoce do risco.

Dessa forma, optou-se pelo corte ≥ 2, visando a maximização do recall e a captura de potenciais atrasos futuros, mesmo diante do aumento de falsos positivos. Essa abordagem preventiva é estratégica para o banco, permitindo agir antes que ocorram inadimplências mais severas, e garantindo uma melhor gestão do risco de crédito.

Resultados e Conclusões:

Limitações / Próximos Passos:

Links de Interesse: