Risco Relativo: Análise de Inadimplência

Cliente: Banco Super Caja
Analista de Dados: Cristiane Thiel
Links: Apresentação - Dashboard - GitHub

Com a queda recente nas taxas de juros, o banco **Super Caja** enfrentou um aumento expressivo na solicitação de empréstimos. Essa alta demanda, somada a um processo de análise manual e ineficiente, gerou gargalos na concessão de crédito e aumentou a exposição ao risco de **inadimplência**. A instituição busca automatizar o processo de análise de crédito com base em dados, aumentando a **eficiência operacional** e **reduzindo perdas financeiras**.

O objetivo central é criar um **sistema de classificação de risco**, baseado em dados históricos de clientes, que permita ao banco:

- Identificar o perfil dos clientes com maior risco de inadimplência
- Calcular o risco relativo de inadimplência entre diferentes segmentos
- Criar uma pontuação de crédito automatizada
- Apoiar a tomada de decisão sobre concessão de crédito de forma mais rápida e precisa

Ferramentas e Tecnologias: BigQuery, SQL, Python, VS Code e Looker Studio

Problema Central

O banco precisa de um **sistema confiável e escalável** que permita avaliar o **risco de inadimplência** de novos solicitantes de crédito, substituindo processos manuais e subjetivos por critérios objetivos baseados em dados históricos. A ausência de uma **ferramenta automatizada** compromete a agilidade das decisões, aumenta o risco de inadimplência e dificulta a criação de políticas de crédito mais estratégicas.

Possíveis Stakeholders

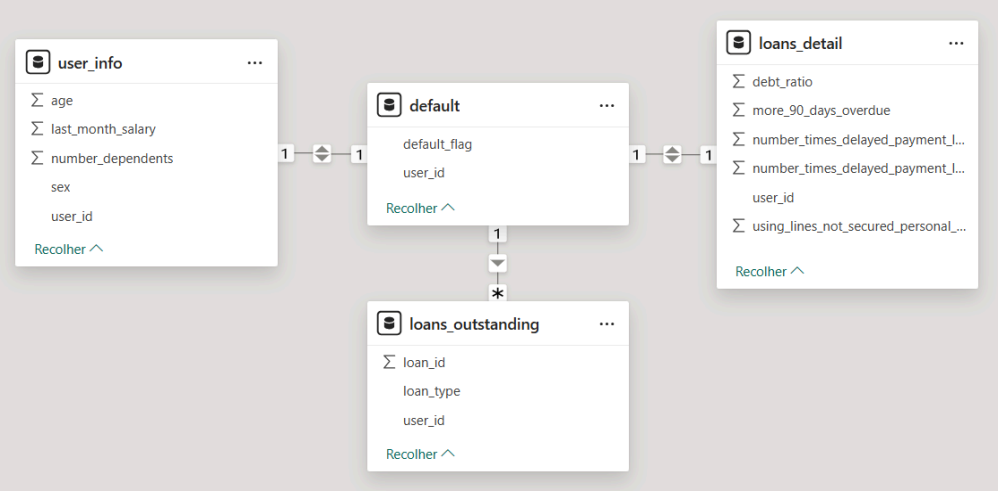
- **Executivos da área de Crédito:** interessados nos impactos estratégicos
- **Analistas de Risco e Concessão de Crédito:** foco nos critérios usados para definir inadimplência e nos segmentos analisados

- **Diretoria Financeira:** de olho nos impactos financeiros da redução da inadimplência e eficiência na análise

Processamento e Análise

Descrição das variáveis que compõem as tabelas deste conjunto de dados:

Arquivo	Variável	Descrição
user_info	user_id	Número de identificação do cliente (único para cada cliente)
	age	Idade do cliente
	sex	Gênero do cliente
	last_month_salary	Último salário mensal que o cliente informou ao banco
	number_dependents	Número de dependentes
loans_outstanding	loan_id	Número de identificação do empréstimo (único para cada empréstimo)
	user_id	Número de identificação do cliente
	loan_type	Tipo de empréstimo (real state = imóveis, others= outros)
loans_detail	user_id	Número de identificação do cliente
	more_90_days_overdue	Número de vezes que o cliente apresentou atraso superior a 90 dias
	using_lines_not_secured_personal_assets	Quanto o cliente está utilizando em relação ao seu limite de crédito, em linhas que não são garantidas por bens pessoais, como imóveis e automóveis
	number_times_delayed_payment_loan_30_59_days	Número de vezes que o cliente atrasou o pagamento de um empréstimo (entre 30 e 59 dias)
	debt_ratio	Relação entre dívidas e ativos do cliente. Taxa de endividamento = Dívidas / Patrimônio
	number_times_delayed_payment_loan_60_89_days	Número de vezes que o cliente atrasou o pagamento de um empréstimo (entre 60 e 89 dias)
default	user_id	Número de identificação do cliente
	default_flag	Classificação dos clientes inadimplentes (1 para clientes já registrados alguma vez como inadimplentes, 0 para clientes sem histórico de inadimplência)



Extração e Tratamento dos Dados

Dados carregados no BigQuery.

Identificando e Tratando Valores Nulos

Apenas a tabela user_info trouxe valores nulos em duas colunas.

Coluna last_month_salary

- Identifiquei 7199 valores nulos nesta coluna (20% da base)
- Análise cruzada com a tabela default mostrou que apenas 130 (1.8%) desses clientes com salário nulo eram inadimplentes (default_flag = 1), enquanto 7069 (98.2%) eram adimplentes
- Identifiquei também 378 registros com last_month_salary igual a 0.

Coluna number_dependents

- Identifiquei 943 valores nulos nesta coluna

Tratamentos Aplicados (na tabela final)

Coluna last_month_salary

- Por conta do baixo percentual de inadimplência entre os clientes com salário nulo e a alta quantidade de registros, decidi não excluir esses dados
- Os valores nulos serão mantidos assim
- Salários Zero: Os 378 registros com salário 0 serão mantidos como estão, pois podem representar informação relevante (sem renda declarada)

Coluna number_dependents

- Decidi tratar os valores nulos assumindo que a ausência de informação significa ausência de dependentes. Os valores nulos serão substituídos por 0
- Função COALESCE(number_dependents, 0)

Identificando e Tratando Valores Duplicados

Nenhuma linha duplicada foi encontrada em nenhuma das tabelas analisadas.

Identificando Variáveis Fora do Escopo

A variável sexo foi excluída por ser sensível. A idade foi mantida por sua relevância potencial, considerando o risco de viés.

Correlação entre Variáveis Numéricas

Correlação de **0.98**, extremamente forte e quase perfeita entre as variáveis **more_90_days_overdue** e **number_times_delayed_payment_loan_30_59_days**.

Isso sugere que essas duas variáveis estão muito interligadas e provavelmente medem informações muito semelhantes. Talvez possa eliminar uma delas para evitar redundância.

Analisar qual é mais interessante manter, que seja mais fácil de entender ou que traga mais insights para o modelo de crédito.

Correlação de **0.99** entre **more_90_days_overdue** e **number_times_delayed_payment_loan_60_89_days** é ainda mais forte, praticamente perfeita. Isso confirma que essas duas variáveis também estão extremamente relacionadas, ou seja, elas medem um comportamento muito semelhante em relação aos atrasos no pagamento. Nesse caso é mais seguro afirmar as variáveis são redundantes.

Correlação entre **number_times_delayed_payment_loan_30_59_days** e **number_times_delayed_payment_loan_60_89_days** é de **0.99**, o que indica que essas duas variáveis também são extremamente correlacionadas.

Então, todas as variáveis relacionadas aos atrasos de pagamento (30-59 dias, 60-89 dias e mais de 90 dias) são altamente correlacionadas.

Para ajudar a tomar a decisão de qual dessas duas variáveis manter para o modelo, podemos usar o **desvio padrão**.

Ao analisar os **coeficientes de desvio padrão**, podemos identificar qual das duas variáveis com correlação tem um **desvio maior**, o que a torna **mais representativa** e mais interessante para a análise, por isso excluímos a variável com menor desvio.

Variável	Desvio Padrão
more_90_days_overdue	4,1214
number_times_delayed_payment_loan_60_89_days	4,1055
number_times_delayed_payment_loan_30_59_days	4,1440

Entre as variáveis **altamente correlacionadas**, devemos manter aquela que tem o maior desvio padrão. Porque o **maior desvio padrão** indica que essa variável tem **mais variabilidade** (mais capacidade de separar, distinguir casos diferentes), e isso é melhor para o modelo. **O desvio padrão mede a dispersão dos dados.** Quanto maior o desvio padrão, mais variada a variável é (ou seja, traz mais informação diferenciada).

Dados Inconsistentes em Variáveis Categóricas

Precisa corrigir valores inconsistentes na variável loan_type, como diferentes formas de escrita de "real estate" e "other".

Realizei a análise dos valores distintos na variável loan_type para identificar as inconsistências. COUNT (DISTINCT loan_type)

Usei a função CASE WHEN para substituir "others" por "other". Usei a função LOWER() para padronizar os valores para minúsculas, garantindo consistência nos dados.

Realizei uma nova contagem dos valores distintos para confirmar que a padronização foi concluída corretamente.

Dados Discrepantes em Variáveis Numéricas (Outliers)

As variáveis numéricas para análise de outliers são: idade (age), salário (last month salary), número de dependentes (number dependents), uso de crédito (using lines not secured personal assets), taxa de endividamento (debt ratio) e atrasos no pagamento (number times delayed payment loan 30 59 days).

A análise foi realizada utilizando **funções agregadas e analíticas** no BigQuery para calcular as estatísticas essenciais, como mínimo, máximo, média, desvio padrão e variância. A mediana, porém, foi calculada separadamente com a função analítica **PERCENTILE_CONT()**, pois ela separa os 50% inferiores dos 50% superiores da distribuição de idades. Diferentemente das funções agregadas, PERCENTILE_CONT() não pode ser usada diretamente com outras funções agregadas, pois é uma função analítica que opera sobre as linhas de dados. Para garantir que a mediana seja calculada de forma única e sem interferir nas outras agregações, ela foi extraída em um **SELECT DISTINCT**, utilizando a sintaxe **OVER()**, que permite calcular o percentil contínuo de forma analítica. O uso de dois SELECTs separados foi necessário devido à limitação entre funções agregadas e analíticas no BigQuery.

Para identificar outliers, utilizei o **intervalo interquartil (IQR)**. O IQR é a diferença entre o terceiro quartil e o primeiro quartil, o que representa a **faixa de dispersão considerada normal** para a maioria central dos dados. Com base nesta medida, valores menores que Q1 - (1,5 * IQR) ou maiores que Q3 + (1,5 * IQR) são classificados como **potenciais outliers**, sendo então analisados individualmente.

Após a identificação dos outliers, foi realizada visualização gráfica por meio de histogramas e boxplots, o que auxiliou na análise da distribuição de cada variável. A decisão sobre o tratamento dos outliers foi tomada caso a caso, considerando se o valor extremo era plausível ou indicativo de erro.

Idade (age)

Mínimo	Máximo	Média	Mediana	Desvio Padrão
21	109	52.4	52	14.8

Agora, para definir os valores aceitáveis para a variável idade, o uso do intervalo interquartil (IQR) é o método estatístico mais apropriado e amplamente reconhecido, pois permite identificar valores atípicos com base na dispersão central dos dados. (o processo se repete para todas as variáveis seguintes)

Apliquei a função PERCENTILE_CONT() para calcular os quartis no BigQuery.

- ➔ Q1 (41) = 25% da base tem idade ≤ 41
- ➔ Q2 (52) = 50% da base tem idade ≤ 52 (mediana)
- ➔ Q3 (63) = 75% da base tem idade ≤ 63
- ➔ IQR = 63 - 41 = 22 = esse é o intervalo interquartil, dispersão central dos dados
- ➔ Limite superior = Q3 + (1,5 × IQR) = 63 + (1,5 × 22) = 96
- ➔ Ou seja, idades maiores que 96 podem ser consideradas outliers.

Ou seja, idades superiores a 96 anos podem ser consideradas outliers estatísticos pelo método IQR. A maior parte das pessoas tem **entre 41 e 63 anos**.Existem registros com **idades de 97 e 98 anos**. Além disso, apenas 3 registros com idades ainda mais elevadas (101, 103 e 109 anos). Considerando a baixíssima frequência desses casos mais extremos (totalizando menos de 0,1% da base com idade > 96 anos) e a possibilidade de serem reais, decidi manter todos esses registros na análise. Eles serão agrupados na faixa etária “60 ou mais”.

Salário (last_month_salary)

Mínimo	Máximo	Média	Mediana	Desvio Padrão
0	1560100	6675	5400.0	12962

- ➔ Q1 = 3400
- ➔ Q2 = 5400
- ➔ Q3 = 8300
- ➔ IQR = Q3 - Q1 = 8300 - 3400 =4900
- ➔ Inferior = Q1 - (1,5 x IQR) = 3400 - (1,5 x 4900) = - 3950 (negativo = abaixo de 1)
- ➔ Superior = Q3 + (1,5 x IQR) = 8300 + (1,5 x 4900) = 15650

A maioria das pessoas ganha entre R$3.400 (Q1) e R$ 8.300 (Q3). Valores acima de R$ 15.650, embora identificados como outliers estatísticos pelo método IQR, serão mantidos nos dados sem alterações, pois a análise preliminar sugere que podem ser rendas elevadas legítimas. A decisão sobre os salários iguais a zero (manter como 0) foi definida na seção de tratamento de nulos.

Número de Dependentes (number_dependents)

Mínimo	Máximo	Média	Mediana	Desvio Padrão
0	13	0.75	0	1.11

- ➔ Q1 = 0
- ➔ Q2 = 0
- ➔ Q3 = 1
- ➔ IQR = 1
- ➔ Inferior = -1,5 (abaixo de 1)
- ➔ Superior = 2,5

A maioria não tem dependentes. Porque a mediana é 0, e 75% da base tem no máximo 1 dependente (Q3 = 1). Com base no IQR, valores acima de 2,5 são considerados outliers estatísticos. Por isso, investiguei os 905 registros (cerca de 2,5% da base) com mais de 3 dependentes.

A distribuição mostra que registros acima de 3 dependentes são consistentes. Porque encontrei 669 registros com 4 dependentes, 171 com 5, 40 com 6, 12 com 7, 8 com 8 e 3 com 9. Apenas dois registros têm 10 e 13 dependentes, o que foge ao padrão, mas mesmo assim, optei por não remover da análise.

Assim, mantive todos os registros como estão e vou **criar uma variável auxiliar** que identifica a “quantidade de dependentes”, o que pode apoiar análises futuras.

Uso de Crédito (using_lines_not_secured_personal_assets)

Mínimo	Máximo	Média	Mediana	Desvio Padrão
0	22000	5.80	0.15	223.40

- ➔ Q1 = 0.029
- ➔ Q2 = 0.149
- ➔ Q3 = 0.548
- ➔ IQR = 0.519

- ➔ Inferior = -0.748
- ➔ Superior = 1.327

A mediana de 0,15 (15%) indica que 50% da base utiliza até 15% do crédito não garantido. Além disso, 75% da base está utilizando no máximo 54,8% (Q3). Com base no IQR, valores acima de 132,7% (ou 1,327) são considerados outliers estatísticos. Esses valores indicam uma utilização do crédito não garantido muito além do normal, sugerindo possíveis clientes endividados ou com comportamentos financeiros extremos.

O valor máximo de 22.000 é incompatível com uma fração ou porcentagem, sendo um valor anômalo e irrelevante para a análise.

Foram identificados **177 registros com valores acima de 1,327**, indicando uma utilização extremamente alta do crédito não garantido. Esses registros apresentam uma grande variação, com valores que vão de **1,327 até 22.000**. Embora a maioria dos registros elevados fique entre **2.000 e 3.000**, valores como **22.000** e **18.300** são outliers extremos e merecem investigação adicional, já que indicam uma utilização de crédito muito além do razoável.

Esses registros com valores elevados foram mantidos na análise, pois podem representar clientes com **alto risco de endividamento** ou comportamentos financeiros fora do padrão. Além disso, valores acima de 100% de utilização do crédito são anômalos no contexto dessa variável, o que sugere que podem ser **erro de dados** ou **acessos anômalos** a limites de crédito muito elevados.

Decidi não remover esses registros, porque podem ser relevantes para **segmentação futura**, especialmente ao analisar **comportamentos de clientes endividados** ou com **altas necessidades de crédito**.

Para facilitar a análise e identificar diferentes perfis de clientes, vale criar uma nova variável categórica, “faixa de uso de crédito”, segmentando os clientes com base na utilização do crédito não garantido.

Taxa de Endividamento (debt_ratio)

Mínimo	Máximo	Média	Mediana	Desvio Padrão
0	307001	351.58	0.37	2011.63

- ➔ Q1 = 0.176
- ➔ Q2 = 0.366
- ➔ Q3 = 0.873
- ➔ IQR = 0.697

- ➔ Inferior = -0.869
- ➔ Superior = 1.919

Metade das pessoas tem uma taxa de endividamento de até 37% (Mediana=0.37). A média elevada (351.58) é influenciada por valores extremos.

A variável debt_ratio representa a **relação entre dívidas e patrimônio do cliente** (Dívidas / Ativos). O valor teórico esperado está entre 0 e 1, indicando, respectivamente, ausência de dívidas e endividamento total equivalente ao patrimônio. No entanto, foram encontrados valores bem acima desse intervalo, chegando a até 307.001.

Estatisticamente, utilizando o método do IQR, valores acima de 1.919 são considerados outliers. Além disso, ao analisar o contexto de negócio, podemos entender que valores acima de 1.0 indicam um cliente cujo nível de endividamento **ultrapassa seu próprio patrimônio**. Ou seja, essa pessoa já deve mais do que possui. Apesar de estatisticamente extremos, esses casos são importantes para a análise de risco, pois representam clientes com maior chance de inadimplência.

Com base nisso, decidi não remover os outliers acima de 1 (100%), manter pode ajudar a aprofundar o entendimento do perfil de clientes altamente endividados. Valores entre 1 e 2 serão tratados como “superendividamento”.

Já os valores acima de 2 serão tratados como **valores extremos de superendividamento**. Embora ainda possam representar casos reais, serão monitorados com atenção, pois a possibilidade de erro de entrada ou distorção estatística é maior nesse grupo. Esses registros não serão removidos a princípio, mas ficarão marcados para análises específicas de comportamento de risco elevado.

Para facilitar a interpretação, vou criar uma nova variável categórica segmentando os clientes por faixas de endividamento.

Atrasos no Pagamento (number_times_delayed_payment_loan_30_59_days)

Mínimo	Máximo	Média	Mediana	Desvio Padrão
0	98	0.42	0	4.14

- ➔ Q1 = 0
- ➔ Q2 = 0
- ➔ Q3 = 0
- ➔ IQR = 0

- ➔ Inferior = 0
- ➔ Superior = 0

A maior parte da base nunca atrasou o pagamento (mediana = 0 e média < 1), o que indica que os atrasos são casos raros. Considerando isso, vou manter todos os dados como estão e criar uma variável de segmentação, para classificar os clientes conforme o número de atrasos.

Tratamento de Dados por Variável (Resumo)

user_id

Nenhum tratamento aplicado. Utilizada como identificador único do cliente.

age

Nenhum tratamento aplicado aos valores. Outliers estatísticos (idades > 96 anos) foram mantidos, pois eram poucos casos (<0,1%) e considerados plausíveis, não indicando erro de dado.

sex

Variável removida. Excluída por ser considerada sensível e para evitar potenciais vieses no modelo de risco.

last_month_salary

Os valores nulos foram mantidos, pois representam uma parcela significativa dos dados e, apesar da ausência de informação sobre renda, estão associados a baixa inadimplência (1,8%), o que indica que não são, por si só, indicadores de risco. **Os valores iguais a zero também foram mantidos**, sendo interpretados como ausência de declaração de renda, o que pode carregar significado comportamental relevante. **Os outliers também foram mantidos.**

number_dependents

Valores nulos foram substituídos por 0 (zero). Assumindo que a ausência de informação indica ausência de dependentes. Os outliers estatísticos seriam os registros com mais de 3 dependentes, mas a distribuição dos dados, mesmo para contagens mais altas, pareceu

consistente. Pela quantidade que os registros aparecem, não identifiquei como erro de digitação e mesmo o valor extremo de 13 dependentes, ainda que alto, não é inconsistente.

```
SELECT
  number_dependents,
  COUNT (number_dependents)
FROM `laboratoria-projeto-03.projeto03_risco_credito.user_info`
GROUP BY number_dependents;
```

loan_id

Nenhum tratamento aplicado. Utilizada como identificador único do empréstimo. Usada como chave primária em loans_outstanding, mas não incluída na tabela unificada final.

loan_type

Valores padronizados. Inconsistências como “others” e diferentes capitalizações de “real estate” foram corrigidas para “other” e todos os valores convertidos para minúsculas para garantir consistência. Esta variável original não está presente na tabela unificada final. Em vez disso, criei as colunas qtd_emprestimos_real_estate e qtd_emprestimos_other (tipo STRING) que contêm a contagem ou ‘não informado’ (para os 425 clientes sem dados em loans_outstanding).

using_lines_not_secured_personal_assets

A análise estatística revelou valores extremos significativamente elevados (como 22.000, 18.300 e 13.930), incompatíveis com a interpretação usual de proporção, mas que formam uma sequência coerente entre si, sugerindo não serem erros de digitação isolados. Esses valores podem indicar fenômenos reais, como **superendividamento**, erros sistemáticos de cálculo ou codificação específica no sistema. Por isso, **mantive todos os registros**, inclusive os outliers estatísticos (177 observações acima do limite definido pelo IQR).

number_times_delayed_payment_loan_30_59_days

Nenhum tratamento aplicado aos valores. Durante a análise da variável de atrasos de pagamento entre 30 e 59 dias, identifiquei valores aparentemente extremos, como 96 e 98 atrasos, que inicialmente levantaram a suspeita de possíveis erros ou outliers. Para investigar, realizei uma **análise de frequência** utilizando um **agrupamento por número de atrasos**, o que revelou que esses valores elevados apareciam de forma recorrente (por exemplo, o valor 98 ocorreu 62 vezes). Essa frequência indicou que não se tratavam de erros

pontuais, mas de registros reais de clientes com **histórico significativo de inadimplência**. Por isso, decidi **manter todos os dados originais**, considerando que esses comportamentos extremos são relevantes para a análise de risco e a compreensão mais precisa do perfil dos clientes.

```
SELECT
  number_times_delayed_payment_loan_30_59_days,
  COUNT (number_times_delayed_payment_loan_30_59_days)
FROM `laboratoria-projeto-03.projeto03_risco_credito.loans_detail`
GROUP BY number_times_delayed_payment_loan_30_59_days;
```

number_times_delayed_payment_loan_60_89_days

Fiz a mesma análise da variável anterior e pelas mesmas razões **nenhum tratamento aplicado aos valores.**

more_90_days_overdue

Fiz a mesma análise da variável anterior e pelas mesmas razões **nenhum tratamento aplicado aos valores.**

Durante a análise da variável number_times_delayed_payment_loan_30_59_days (e, por consequência, das variáveis referentes a 60-89 dias e 90+ dias de atraso), identifiquei **valores elevados como 96 e 98**, que **inicialmente pareciam ser possíveis erros ou outliers**. Para investigar, fiz uma **análise de frequência** que revelou, por exemplo, que o valor 98 aparecia 62 vezes na faixa de 30-59 dias. Ao cruzar esses dados com as demais faixas de atraso, verifiquei que **os mesmos 62 clientes** também apresentavam o valor 98 nas variáveis de 60-89 dias e 90+ dias. Essa consistência nos valores entre diferentes métricas para o mesmo grupo de clientes reforça a interpretação de que não se tratam de erros de digitação, mas sim de um padrão real de inadimplência grave e contínua. A decisão foi **manter todos os dados originais**, considerando que representam comportamentos legítimos e relevantes para a análise de risco; sua exclusão poderia distorcer a compreensão do perfil desses clientes.

debt_ratio

Inicialmente nenhum valor foi removido ou tratado, até a EDA. Outliers estatísticos e valores indicando alto endividamento (dívida > patrimônio) foram mantidos, por serem relevantes para a análise de risco.

Durante a Análise Exploratória de Dados (EDA), foi identificado que a variável Taxa de Endividamento (debt_ratio) apresentava **valores extremos e não realistas**, particularmente para o grupo de 7.577 clientes (aproximadamente 21% da base) cujo salario era “não informado”.

Investigações detalhadas revelaram que, para este segmento de clientes com salário não informado:

- A mediana da taxa de endividamento era de 1140.0 (equivalente a um DTI de 114.000%).
- Aproximadamente 92% (6.981 clientes) deste grupo possuíam uma taxa de endividamento superior a 5.0 (500%).
- Não foram encontrados clientes neste grupo com taxa de endividamento no intervalo mais típico de (0, 1), exceto por aqueles com valor exato de 0.0.

Essa análise indica que os valores de **taxa de endividamento para clientes sem informação salarial** são, muito provavelmente, fruto de cálculo original da métrica (possivelmente devido à divisão por uma renda nula ou zero) e **não representam uma medida confiável** de seu endividamento real em relação à renda.

Considerando que este grupo de “salário não informado” demonstrou uma **baixa taxa de inadimplência** (1,78%), e para preservar a integridade das análises, decidi tratar assim:

- Para clientes com salario “não informado” (ou seja, NULL ou 0 na base original), a coluna numérica taxa de endividamento será definida como NULL na tabela analítica final (tabela_unificada_auxiliar).
- Consequentemente, na variável categórica faixa_endividamento, esses clientes serão classificados como “não Informado”. Mantendo a nomeclatura da faixa salarial.

Esta abordagem garante que as **estatísticas descritivas** (como média e mediana) e as faixas de endividamento para os demais clientes (com salário informado) **não sejam distorcidas por esses valores problemáticos**, permitindo uma análise de risco mais precisa e segmentada. O grupo com “Salário Não Informado” e “DTI Não Informado” será analisado com base em suas **outras características disponíveis**.

default_flag

Nenhum tratamento aplicado. Utilizada como variável alvo (target) para classificar clientes como inadimplentes (1) ou não inadimplentes (0).

Tabela Criada

Criei a tabela tabela_unificada_auxiliar no BigQuery. Esta tabela agrega informações de todas as fontes originais (user_info, loans_detail, default, loans_outstanding) em um único local, com uma linha por cliente, totalizando 36.000 registros. Além dos dados brutos, foram aplicados tratamentos e criadas novas variáveis categóricas (faixas) para enriquecer a análise.

Dados Demográficos e de Identificação

- id_cliente (INTEGER): Identificador único do cliente, originário de user_info.user_id.
- idade (INTEGER): Idade do cliente, de user_info.age. Nenhum tratamento de nulo aplicado, pois não foram identificados nulos nesta coluna.
- salario (INTEGER): Último salário mensal informado (user_info.last_month_salary). Valores NULL e 0 originais foram mantidos para serem categorizados como “não informado” na faixa_salarial.
- numero_dependentes (INTEGER): Número de dependentes (user_info.number_dependents). Valores NULL originais foram tratados como 0 (zero) usando COALESCE, assumindo que a ausência de informação indica ausência de dependentes.

Informações de Empréstimos

- qtd_emprestimos_real_estate (STRING): Quantidade de empréstimos do tipo “Real Estate” por cliente. Calculada a partir de loans_outstanding. Para os 425 clientes sem registros em loans_outstanding, este campo é “não informado”.
- qtd_emprestimos_other (STRING): Quantidade de empréstimos do tipo “Other” por cliente. Lógica similar à anterior.

Histórico de Atrasos

- atrasos_30_59_dias (INTEGER): Nº de atrasos entre 30-59 dias.
- atrasos_60_89_dias (INTEGER): Nº de atrasos entre 60-89 dias.
- atrasos_acima_90_dias (INTEGER): Nº de atrasos superiores a 90 dias.

Comportamento de Crédito e Endividamento

- uso_linha_credito (FLOAT): Proporção de utilização de linhas de crédito não garantidas.
- taxa_endividamento (FLOAT): Taxa de Dívida (Debt Ratio). Para clientes com salario nulo ou zero, este campo numérico foi definido como NULL para evitar distorções causadas por valores de DTI não confiáveis. Para os demais, reflete o debt_ratio original.

Indicador de Risco

- indicador_inadimplencia (INTEGER): Flag de inadimplência (default.default_flag: 1 para inadimplente, 0 para adimplente).

Criação de Faixas Categóricas Derivadas

Para facilitar a segmentação e análise, as seguintes variáveis de faixa (todas do tipo STRING) foram criadas:

- faixa_etaria: '18 a 25', '26 a 35', '36 a 45', '46 a 60', '60 ou mais'
- faixa_salarial: 'até 2.000', '2.000 - 5.000', '5.000 - 10.000', '10.000 ou mais', 'não informado'
- faixa_dependentes: '0', '1', '2', '3 a 5', '6 ou mais', baseada no numero_dependentes já tratado
- faixa_uso_credito: 'Muito Baixo (até 2.9%)', 'Baixo (2.9% - 14.9%)', 'Moderado (14.9% - 54.8%)', 'Elevado (54.8% - 132.7%)', 'Extremo (> 132.7%)', 'Não Informado (Crédito)'. A categoria 'Não Informado' é atribuída se o uso_linha_credito numérico for NULL
- faixa_endividamento: Baixo (≤30%), Moderado (30-50%), Elevado (50-70%), Muito Elevado (70-100%), Superendividamento (100-200%), e Extremo Superendividamento (>200%). Valores NULL de DTI (devido a dados faltantes, como salário) são categorizados como 'Não Informado (DTI)'.
- faixa_atraso_pagamento (baseada em atrasos_30_59_dias): 'sem atraso', 'baixo', 'moderado', 'elevado', 'extremo', 'Não Informado (Atr.)'
- status_inadimplencia (baseada em indicador_inadimplencia): 'sim', 'não'

- historico_atrasos (baseada nas três colunas de atraso): 'sim','não'.
- faixa_atraso_atingida (baseada nas três colunas de atraso): 'atraso 30-59 dias', 'atraso entre 60 e 89 dias', 'atraso acima 90 dias', 'não atrasou'

Nessa mesma tabela serão criados os “quartis” para segmentação.

Análise Exploratória de Dados (EDA)

Variável	Média	Mediana	Min	Max	Desvio Padrão	Nulos
idade	52	52	21	109	14.79	0
salário	6675.05	5400	0	1.560.100	12961.77	7199
dependentes	0.73	0	0	13	1.11	0
uso linha credito	5.807040	0.149655	0	22000	223.40	0
taxa endividamento	4.793712	0.290967	0	5696	90.72	7577
atrasos 30 59 dias	0.419278	0.000000	0	98	4.14	0
atrasos 60 89 dias	0.237861	0.000000	0	98	4.10	0
acima 90 dias	0.260806	0.000000	0	98	4.12	0

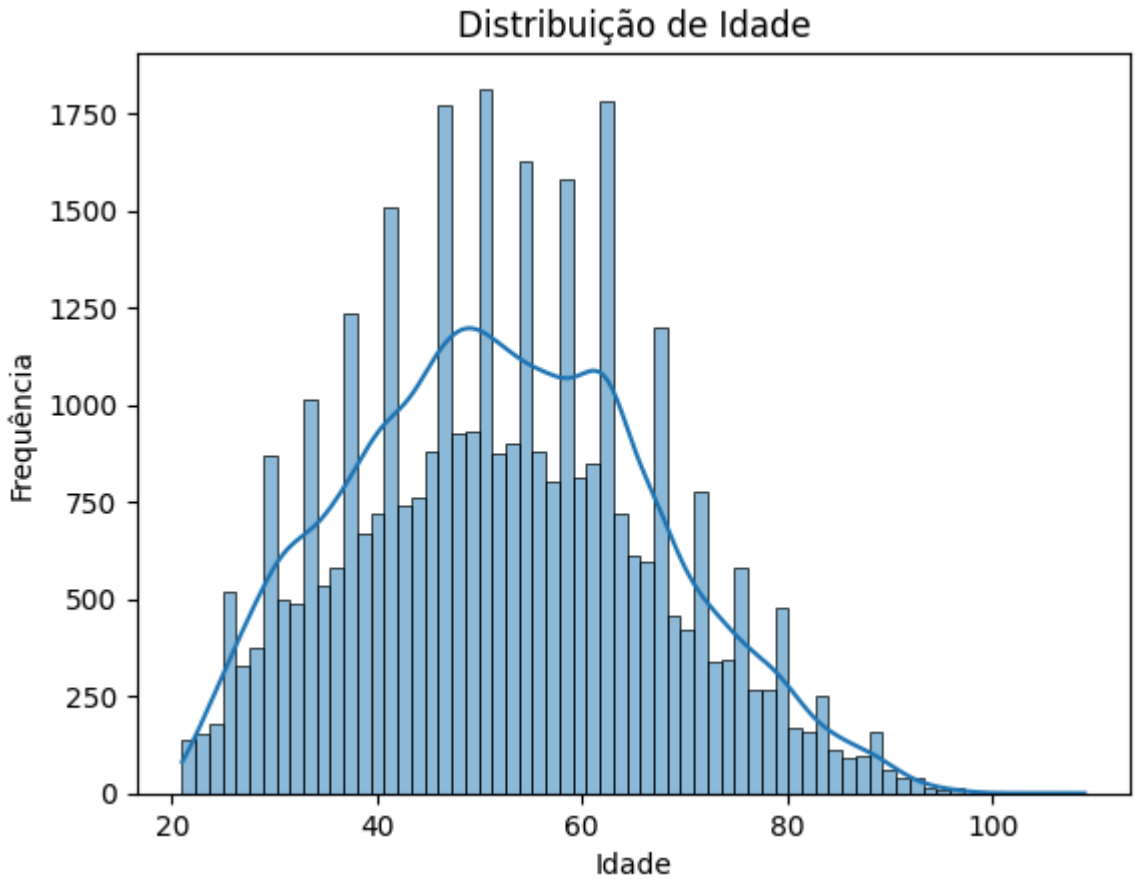
A grande diferença entre a **média e a mediana para uso de linha de crédito** (5.80 e 0.15) indica uma forte **assimetria à direita na distribuição**, com valores extremos (máximo de 22.000) elevando significativamente a média. A mediana de 0.15 (15%) é, portanto, uma medida mais representativa do comportamento típico de utilização de crédito da maioria dos clientes.

- Total de Clientes: 36.000 clientes
- 425 clientes não estão na tabela onde são informados tipos de empréstimos
- Empréstimos Imobiliários: 36.562
- Outros Empréstimos: 268.773

Nota: Estes totais de empréstimos são a soma das quantidades de empréstimos por cliente, tratando "não informado" como 0 para a soma.

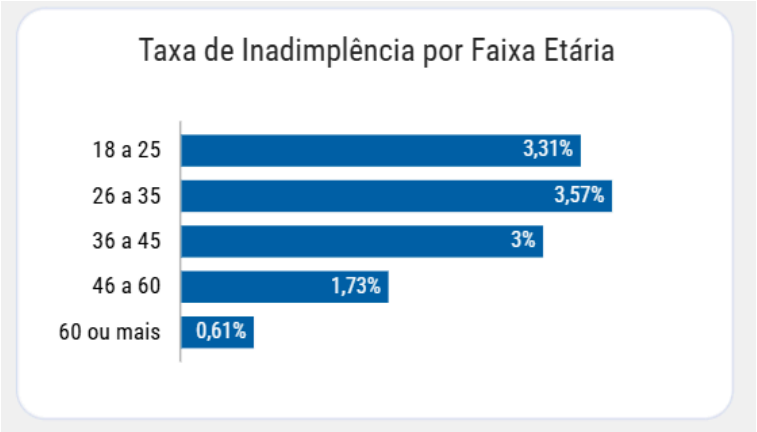
Faixa Etária

- A maior concentração de clientes está na faixa de **46 a 60 anos (36%)**, seguida de perto pela faixa de 60 ou mais anos (30%).
- As faixas mais jovens, como 18 a 25 anos (2%), possuem a menor representatividade.



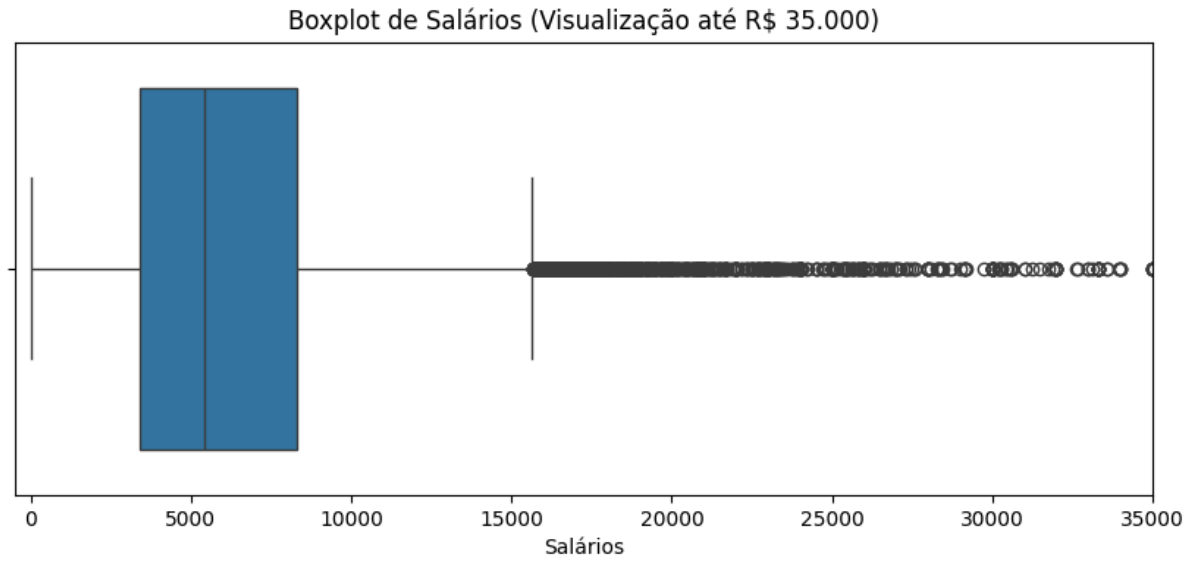
Taxa de Inadimplência por Faixa Etária

Podemos observar uma tendência clara de diminuição da inadimplência com o aumento da idade. As faixas mais jovens apresentam taxas mais elevadas (18-25 anos: 3,31%; 26-35 anos: 3,57%), enquanto a faixa de “60 ou mais anos” registra a menor taxa (0,61%). Isso sugere que a **idade pode ser um fator relevante na previsibilidade do risco**.



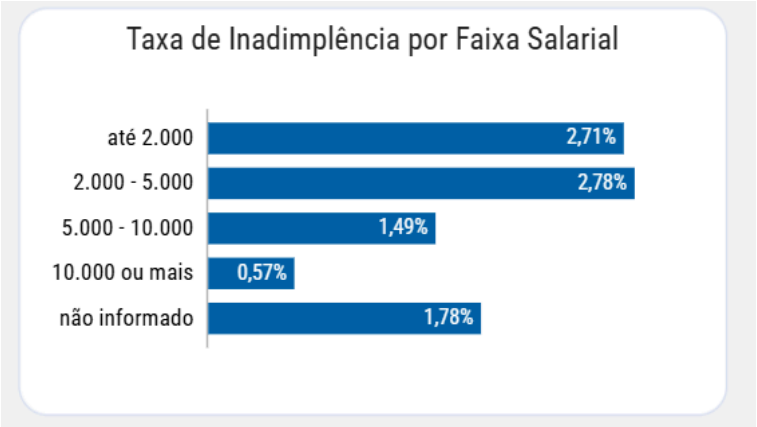
Faixa Salarial

- As faixas salariais predominantes são de R$5.000 a R$10.000 (31%) e de R$2.000 a R$5.000 (29%).
- 7.577 clientes com salário classificado como “não informado”, o que representa aproximadamente 21% da base. Não excluí valores pois perderia muita informação e não imputei valores que seriam na verdade uma suposição.
- A faixa até R$2.000 é a que tem menos representantes (7%).



Taxa de Inadimplência por Faixa Salarial

A inadimplência demonstra uma relação inversa com a renda. As taxas mais altas ocorrem nas faixas de até R$2.000 e (2,71%) de R$2.000 a R$5.000 (2,78%).



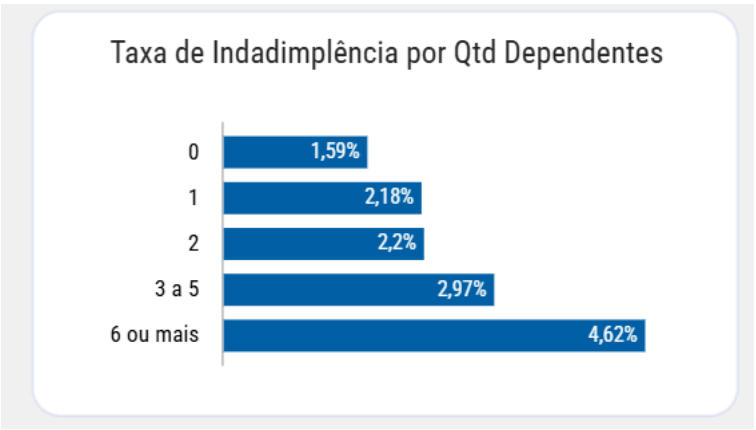
Dependentes

A maioria dos clientes informou não ter dependentes (58%). Além disso, 943 clientes simplesmente não possuem essa informação no cadastro, então considerei como ausência de dependentes. Provavelmente por conta da base ser composta por pessoas acima de 46 anos.

O número de clientes diminui à medida que o número de dependentes aumenta, com apenas 65 clientes informando “6 ou mais” dependentes.

Taxa de Inadimplência por Quantidade de Dependentes

Existe uma **correlação positiva** entre o número de dependentes e a taxa de inadimplência. Clientes sem dependentes apresentam a menor taxa (1,59%), enquanto aqueles com “6 ou mais” dependentes registram a **taxa mais elevada** (4,62%). Cada aumento no número de dependentes parece estar associado a um aumento progressivo na taxa de inadimplência.



Histórico de Atrasos

- Aproximadamente 80% dos clientes não possuem histórico de atrasos registrados.

- Histórico de Atrasos e Status de Inadimplência não comunicam a mesma coisa.
- Nenhum cliente inadimplente (default_flag = 1) sem ter registro de atraso.
- Para 29.378 clientes, o histórico de atrasos é igual ao status de inadimplência. Para 6.622 clientes, eles são diferentes.
- Estes são clientes que tiveram **algum atraso registrado**, mas não são considerados inadimplentes pela flag default_flag. Atrasos **não necessariamente classificam um cliente como inadimplente** no sistema do banco. A inadimplência (default_flag=1) pode ser acionada por **critérios mais severos** (ex: atrasos muito longos, múltiplos atrasos, um valor específico de dívida não paga, etc.).
- A taxa de inadimplência é de 9,35%. Isso significa que, entre os clientes que tiveram pelo menos um tipo de atraso, cerca de 9% deles são efetivamente classificados como inadimplentes (default_flag = 1). Os outros 91% tiveram atrasos, **mas não o suficiente** (ou do tipo certo) para serem marcados como inadimplentes.

Faixa de Atraso Atingida

Detalha o tipo de atraso para os 20% que tiveram histórico.

- Dos clientes com histórico de atraso, a faixa mais comum é “atraso entre 30 e 59 dias” (57%).
- Seguida por “atraso acima 90 dias” (27%).
- E “atraso entre 60 e 89 dias” (16%).

Uso de Crédito

Se refere ao quanto o cliente está utilizando do **limite de crédito** em linhas que **não são garantidas por bens pessoais**, como imóveis ou automóveis.

Exemplos: cartão de crédito, cheque especial, empréstimos pessoais sem garantia, crédito consignado…

Já as linhas garantidas por bens pessoais incluem: financiamento de imóveis (garantido pelo próprio imóvel), crédito com garantia de veículo (como um refinanciamento de carro).

Por que essa métrica é importante?

Créditos não garantidos representam **mais risco para o banco**. Se muitos clientes estiverem usando muito dessas linhas, isso pode indicar:

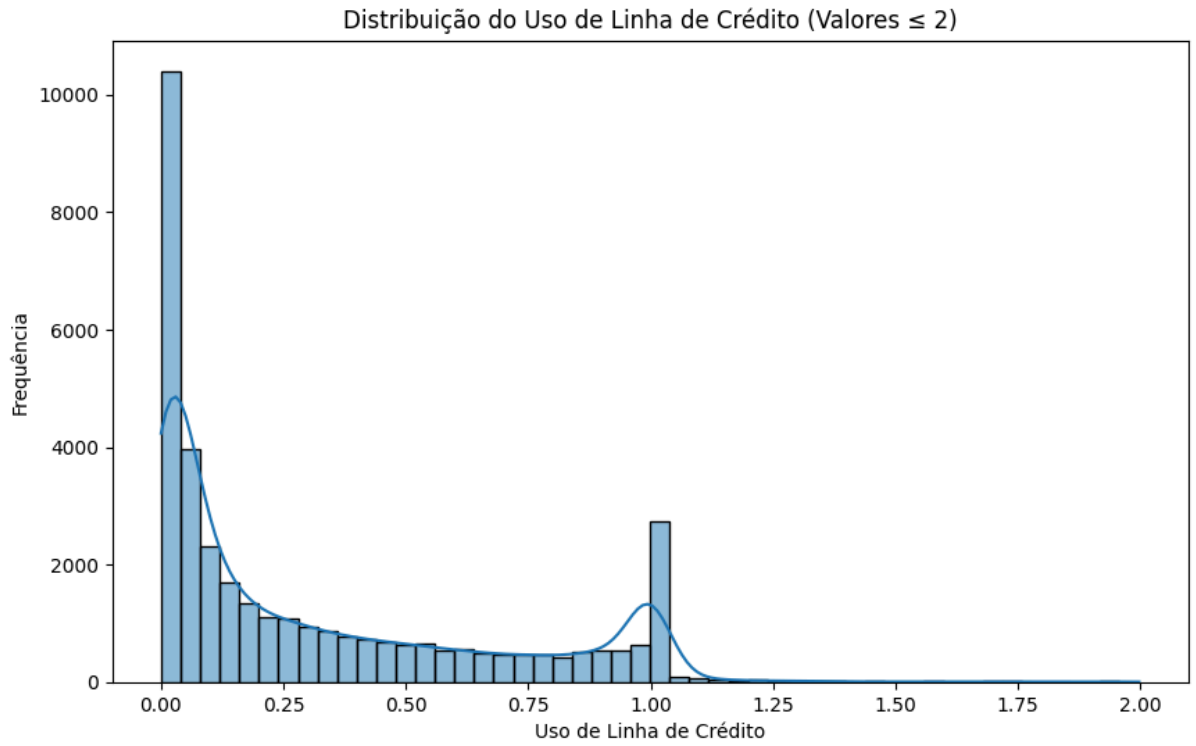
- Maior exposição ao risco de inadimplência
- Clientes mais endividados em modalidades mais caras
- Necessidade de atenção ou intervenção na gestão de crédito

A distribuição do uso de crédito é relativamente equilibrada entre as faixas:

- Baixo (2.9% - 14.9%): 9.048 clientes
- Moderado (14.9% - 54.8%): 9.024 clientes
- Muito Baixo (até 2.9%): 8.918 clientes
- Elevado (54.8% - 132.7%): 8.833 clientes

Aproximadamente 25% dos clientes em cada uma dessas faixas.

- A faixa de uso Extremo (> 132.7%) é a menos comum, com apenas 177 clientes.



Este gráfico é excelente indica:

A maioria dos clientes (que têm uso de linha de crédito ≤ 2, ou seja, ≤ 200%) está concentrada em **valores muito baixos**, especialmente perto de zero. A primeira barra, que provavelmente representa o uso de crédito próximo de 0% até uns 2,5% ou 5%, é de longe a mais alta. Se alinha com a mediana de 0.15 (15%).

Mesmo com o zoom, a distribuição é assimétrica à direita. A frequência cai rapidamente à medida que o uso de crédito aumenta.

Existe um pico na frequência em torno do valor 1.0 (uso de 100% da linha de crédito). Isso é um padrão comum e importante:

- Muitos clientes tendem a usar seus **cartões de crédito até o limite** (ou muito perto dele).
- Clientes que estão consistentemente no limite ou acima dele são geralmente considerados de maior risco.

Após o pico em 1.0, a frequência cai novamente, indicando que menos clientes (dentro deste subconjunto de uso ≤ 200%) ultrapassam significativamente seus limites.

A linha azul suave ajuda a visualizar a **forma geral da distribuição**, confirmando os picos perto de zero e em 1.0.

A **maioria dos clientes usa muito pouco de suas linhas de crédito não garantidas**, ou até o limite. Existe um grupo menor que usa uma porção intermediária. O comportamento de "estourar o limite" (valores > 1.0) é menos frequente, mas presente. Este **padrão bimodal** (picos em 0 e 1) é clássico para utilização de crédito.

Endividamento (Taxa de Dívida sobre a Renda - DTI)

Mede a **relação entre as dívidas e o patrimônio de um cliente**. Ela indica o quanto do patrimônio do cliente está **comprometido com dívidas**, ajudando a avaliar sua **saúde financeira**. Um debt ratio baixo sugere que o cliente possui uma relação saudável entre suas dívidas e ativos, enquanto um debt ratio alto indica que o cliente está mais endividado em relação ao seu patrimônio, o que pode ser um sinal de risco financeiro.

- A maior parte dos clientes, 15.050, está na faixa de endividamento Baixo (DTI ≤ 30%).
- A faixa Moderado (DTI 30%-70%) também é significativa, com 10.723 clientes.
- Um número considerável, 7.515 clientes, está em Casos Extremos (DTI > 200%).
- As faixas de "Elevado (70%-100%)" e "Superendividamento (100%-200%)" têm menos clientes.

A maior parte dos clientes tem um **nível de endividamento relativamente controlado**, o que é positivo para a saúde financeira geral da base de clientes.

Entretanto, há um número considerável de clientes em situações de **endividamento extremo e moderado**, que requerem atenção especial. Esses grupos podem representar potenciais riscos financeiros.

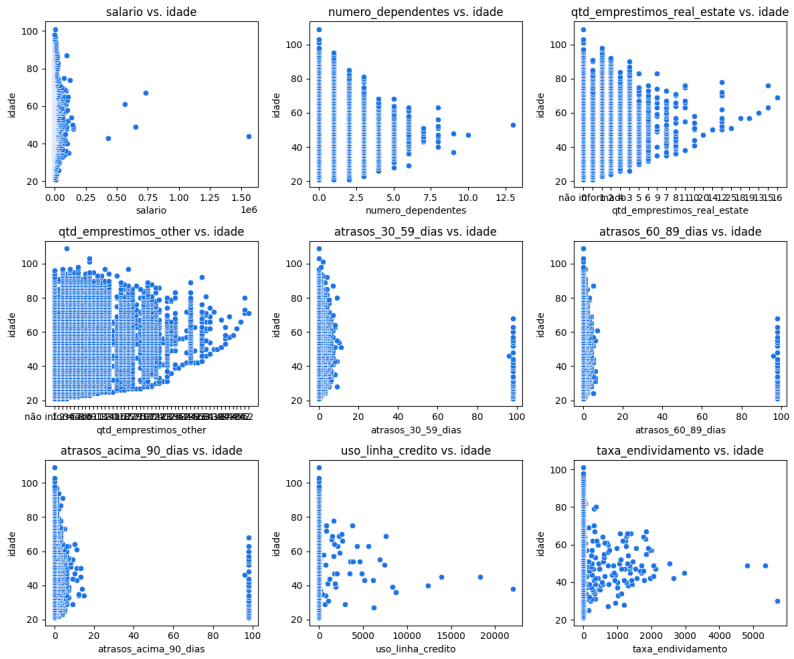
As faixas com baixo número de clientes em níveis elevados de endividamento indicam que a maioria está bem posicionada, mas os casos mais críticos devem ser monitorados com mais rigor para evitar que a situação se agrave.

Correlação entre Variáveis Numéricas

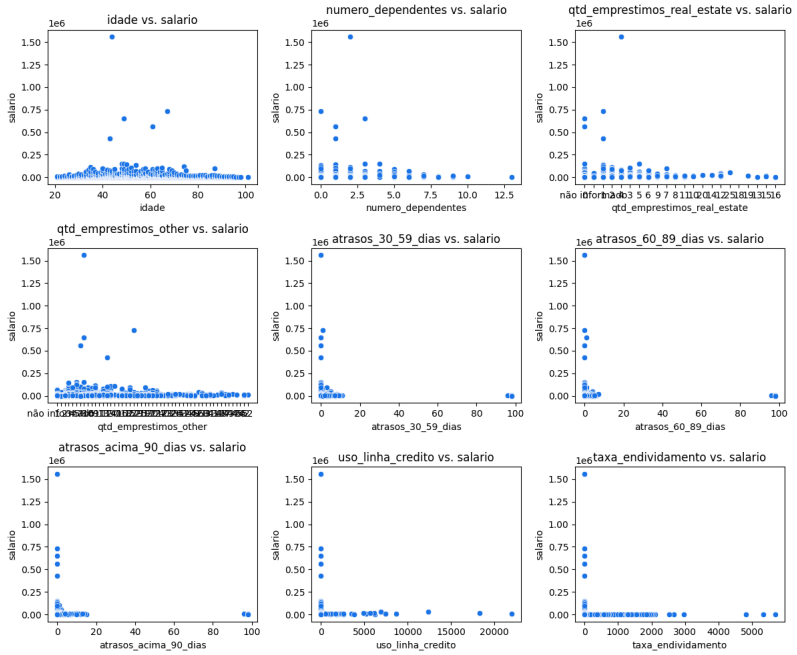
variaveis	coeficiente de correlação
idade_x_salario	0.035248879336409834
idade_x_taxa_endividamento	-0.010977792880697987
salario_x_taxa_endividamento	-0.025639658129925084
numero_dependentes_x_salario	0.0773953424857159
idade_x_uso_linha_credito	-0.0075838866718118359
salario_x_uso_linha_credito	0.007403555196104793
uso_linha_credito_x_taxa_endividamento	-0.001052109345077994

Gráficos de Dispersão entre Variáveis

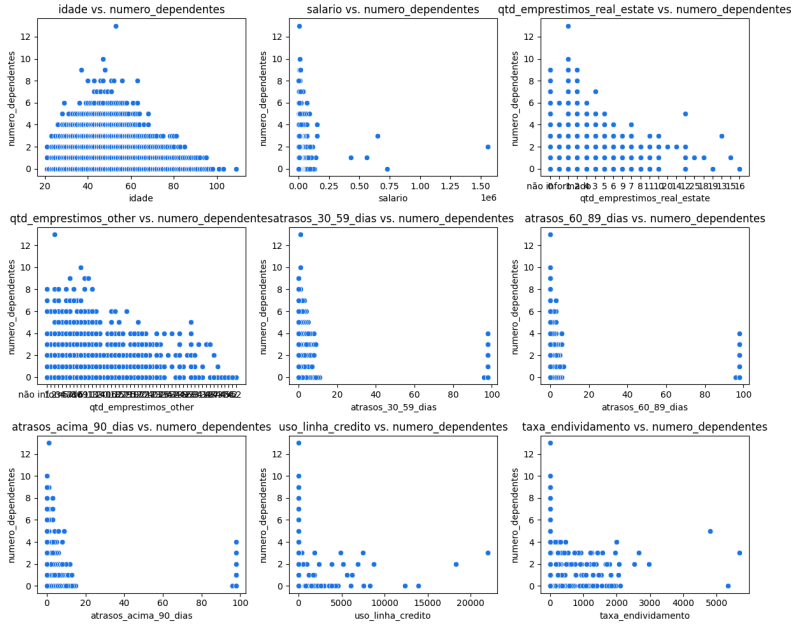
Idade



Salário



Dependentes



Tipos de Empréstimo

Tipo	Inadimplentes	Total de Clientes	Taxa de Inadimplência (%)
Real Estate	294	22.456	1,31
Other	619	35.485	1,74

Other tem **mais clientes e mais inadimplentes** em números absolutos e a taxa de inadimplência (1,74%)é ligeiramente maior do que para Real Estate (1,31%).

Lembrando que o total de clientes aqui não é o todo de clientes da base. Já que o mesmo cliente pode ter mais de um tipo de empréstimo e nem todos os clientes têm um tipo de empréstimo definido.

Análise dos Atrasos por Tipo de Empréstimo

Comparando os atrasos com base no tipo de empréstimo. As três tabelas de análise com os dados de atrasos de 30–59 dias, 60–89 dias e mais de 90 dias.

Atrasos entre 30–59 dias

A média de atrasos entre 30 e 59 dias é praticamente a mesma para os dois tipos de empréstimos: 0.2829 para real estate e 0.2826 para other. A diferença é insignificante. Mesmo com um número muito maior de empréstimos e clientes na categoria other, o comportamento médio de atraso nesse período é equivalente. Não há indicativo de que o tipo de empréstimo impacte o atraso nesse intervalo.

Tipo	Total de Clientes	Qtd de Empréstimos	Média dos Atrasos
Real Estate	22.456	36.562	0,2829
Other	35.485	268.773	0,2826

Atrasos entre 60–89 dias

A média de atrasos entre 60 e 89 dias apresenta uma diferença mais clara. Other tem uma média de 0.0632, enquanto real estate tem 0.0586. Embora a diferença ainda não seja grande, ela é mais perceptível que no período anterior. Isso indica uma leve tendência de maior atraso para empréstimos do tipo other nesse intervalo de tempo.

Tipo	Total de Clientes	Qtd de Empréstimos	Média dos Atrasos
Real Estate	22.456	36.562	0,0586
Other	35.485	268.773	0,0632

Atrasos acima de 90 dias

Após os 90 dias, a **diferença volta a ser mínima** (0.0628 vs 0.0626), o que indica que os dois tipos de empréstimos voltam a apresentar um comportamento de atraso muito semelhante.

Tipo	Total de Clientes	Qtd de Empréstimos	Média dos Atrasos
Real Estate	22.456	36.562	0,0626
Other	35.485	268.773	0,0628

Reflexão

A análise indica que **não há diferenças significativas nos atrasos entre os tipos de empréstimo nos períodos de 30–59 dias e acima de 90 dias**. A única diferença relevante ocorre **entre 60 e 89 dias**, quando os empréstimos classificados como *other* apresentam uma média de atraso **ligeiramente superior** à dos empréstimos de *real estate*.

Como os dados de atraso estão associados ao cliente, e não a cada empréstimo individual, **essa interpretação é feita por inferência**. Ainda assim, é possível supor que clientes que contratam empréstimos *other* possam ter se planejado menos financeiramente, o que contribui para uma **maior incidência de atraso nesse intervalo**. Já os que contratam empréstimos *real estate*, por envolverem valores mais altos e maior exigência de planejamento, tendem a manter maior regularidade nos pagamentos.

Na etapa inicial da análise, foi observada uma **alta correlação entre os atrasos nos três períodos** (30–59, 60–89 e 90+ dias), indicando que clientes inadimplentes em um período tendem a repetir o comportamento nos demais. A faixa de 30–59 dias foi adotada como referência principal por apresentar o **maior desvio padrão**, sugerindo maior variabilidade no comportamento de atraso nesse período.

Porém, ao segmentar os dados por tipo de empréstimo, percebemos que **o tipo de crédito só influencia significativamente os atrasos entre 60 e 89 dias**. Nos outros dois intervalos, os resultados são muito similares entre as categorias.

Interpretações:

- ★ **Até 59 dias**: comportamento de atraso é semelhante entre os dois tipos de empréstimo.
- ★ **Entre 60 e 89 dias**: clientes de empréstimos *other* demonstram maior propensão à inadimplência.
- ★ **Após 90 dias**: os atrasos voltam a se estabilizar, sugerindo reorganização financeira por parte dos clientes.

Implicações:

- ★ A faixa de **30–59 dias deve ser monitorada de perto** para análises de risco gerais, pois concentra a maior variação de comportamento.
- ★ A **segmentação por tipo de empréstimo é especialmente útil para o intervalo de 60–89 dias**, onde há maior variação no comportamento.
- ★ **Empréstimos do tipo *other*** podem demandar estratégias específicas de acompanhamento, especialmente no médio prazo.

É importante reforçar que os insights aqui apresentados são baseados em médias por cliente. Não é possível afirmar com precisão o comportamento em nível de empréstimo individual. Ainda assim, as tendências observadas permitem traçar estratégias de mitigação de risco baseadas em perfis agregados.

Quartis para Variáveis de Risco Relativo

As **faixas manuais** e os NTILES (quartis, decis, etc.) são duas formas diferentes de segmentar as variáveis numéricas, e elas servem a propósitos complementares.

As faixas manuais que eu criei (faixa_etaria, faixa_salarial, faixa_uso_credito…) foram criadas nos **limites que eu defini**, com base em regras de negócios e também com base no IQR. Eu defini os limites e os nomes das categorias com base no meu conhecimento do negócio, em padrões observados na análise das medidas centrais no processamento inicial.

Fiz isso porque não **mais intuitivas** e fáceis de entender para qualquer pessoa, pois os nomes e os limites são explícitos e muitas vezes alinhados com conceitos de negócio.

Os grupos (categorias) resultantes **podem ter tamanhos muito diferentes**. Pode haver muito mais clientes na faixa etária 46 a 60 do que na 18 a 25, por exemplo.

Essas faixas são **boas para criar dashboards** de perfil de cliente, entender a composição da base, e para comunicação geral dos dados. Para uma **análise inicial da taxa de inadimplência** para cada uma dessas faixas manuais, começando com uma boa ideia de quais segmentos são mais arriscados.

Por outro lado, a função **NTILE(*N*) OVER (ORDER BY *variavel*)** automaticamente divide os clientes em N grupos de **tamanhos o mais iguais possível**, com base na **ordenação da variável**.

NTILE(4) (quartis) cria 4 grupos, cada um contendo aproximadamente 25% dos clientes (com base nos valores não nulos da variável de ordenação). O grupo 1 tem os menores valores, o grupo 4 tem os maiores.

O objetivo é ter **grupos com número de observações (clientes) similar**. Isso é importante para **análises estatísticas**.

Os valores que definem os limites entre os NTILES são determinados pelos dados.

Os limites podem não ser tão redondos ou intuitivos quanto os das faixas manuais.

Quando temos **grupos de tamanho similar**, podemos calcular **taxas de inadimplência** para cada NTILE e comparar de forma mais analítica. É **estatisticamente mais sólido** comparar o risco do "quartil superior de renda" com o "quartil inferior de renda". É possível ver se o **risco aumenta ou diminui** consistentemente através dos NTILES.

As **variáveis transformadas em NTILES** podem ser usadas como **features em modelos de pontuação**, pois já representam uma forma de *binning* ou **categorização ordenada** que pode capturar **relações não lineares**. Os limites entre os NTILES podem ajudar a informar onde definir pontos de corte para classificar clientes em categorias de risco.

As duas maneiras de segmentar não são mutuamente exclusivas. Por isso, mantive ambas na tabela final.

As faixas manuais são boas para a EDA descritiva, para entender **o que os dados nos dizem** e para comunicar o **perfil dos clientes** de forma intuitiva no seu dashboard.

Os NTILES são uma **ferramenta mais técnica para a análise quantitativa de risco**, para entender "o quanto mais arriscado" um grupo é em relação a outro, e para preparar os dados para modelagem.

Depois de calcular o risco relativo por NTILE, é possível voltar às faixas manuais e ver se os padrões de risco se mantêm ou se há nuances.

Por exemplo, na faixa etária 60 ou mais pode abranger dois ou mais quartis de idade, e a taxa de inadimplência dentro dessa faixa manual pode ser uma média ponderada das taxas dos quartis que ela contém.

Em resumo, **as faixas manuais são como rótulos descritivos e intuitivos** para segmentação e apresentação dos perfis. Já os NTILES são como uma **ferramenta analítica para criar grupos de tamanho igual** para uma comparação de risco mais padronizada e para modelagem.

Criação de Colunas de Quartis e Decis

As seguintes colunas de NTILE foram adicionadas à tabela unificada:

NTILE(4) sobre idade

Os clientes foram divididos em quatro grupos com base em sua idade. O Quartil 1 representa os 25% de clientes mais jovens, enquanto o Quartil 4 representa os 25% de clientes mais velhos. O objetivo é avaliar como o risco de inadimplência varia entre diferentes faixas etárias de igual representatividade.

Quartil Idade	Idade Mínima	Idade Máxima	Total de Clientes
1	21	41	9.000
2	41	52	9.000
3	52	63	9.000
4	63	109	9.000

NTILE(10) sobre Salário

Os clientes foram segmentados em dez grupos (decis) com base em seu salário. Para este cálculo, salários nulos ou iguais a zero foram excluídos da ordenação para não distorcer a formação dos decis. O Decil 1 contém os 10% de clientes com os menores salários (válidos), e o Decil 10 os 10% com os maiores salários. Para uma análise mais granular da relação entre diferentes níveis de renda e o risco de inadimplência, dada a ampla variação e assimetria da variável salário.

Decil Salário	Salário Mínimo	Salário Máximo	Total Clientes
1	0	2000	2881
2	2000	3000	2880
3	3000	3800	2880
4	3800	4566	2880
5	4566	5400	2880
6	5400	6333	2880
7	6333	7500	2880
8	7500	9086	2880

9	9091	11666	2880
10	11666	1560100	2880

NTILE(4) sobre Número de Dependentes

Os clientes foram agrupados em quatro grupos com base no número de dependentes (onde nulos foram tratados como 0). Para analisar se diferentes contagens de dependentes, quando agrupadas em quartis, mostram um padrão de risco distinto. As faixas manuais para dependentes também serão utilizadas, porque podem oferecer uma interpretação mais direta para esta variável discreta.

Quartil Dependentes	Dependentes Mínimo	Dependentes Máximo	Total de Clientes
1	0	0	9000
2	0	0	9000
3	0	1	9000
4	1	13	9000

NTILE(4) sobre Uso Linha de Crédito

Segmenta os clientes em quatro grupos com base na sua taxa de utilização de linhas de crédito não garantidas. O Quartil 1 representa os clientes com o menor uso percentual de crédito, e o Quartil 4 aqueles com o maior uso. Para avaliar o impacto de diferentes níveis de utilização de crédito no risco de inadimplência. Esta é uma variável chave em modelos de crédito.

Quartil Uso Crédito	Mínimo Uso Crédito (%)	Máximo Uso Crédito (%)	Total de Clientes
1	0.000	0.029	9000
2	0.029	0.149	9000
3	0.149	0.548	9000
4	0.548	22000.000	9000

Apesar dos valores atingirem um máximo atípico de 22.000, a análise indica que estes, juntamente com outros dados sequencialmente elevados, representam **casos reais de uso desproporcional do crédito** ou **situações anômalas**, e não meros erros. Esses clientes já se enquadram no Quartil 4 de risco (que inicia em 54,8% de utilização, um patamar significativamente acima da mediana de 15% da base) e, portanto, contribuem para a identificação de **perfis inadimplentes**. Remover poderia esconder ou ignorar um risco

inerente. Além disso, esses 177 clientes (aproximadamente 0,5% da base) com comportamento extremo, estão destacados no dashboard e são uma oportunidade para investigar **potenciais fraudes ou falhas operacionais**. Por estas razões, os dados originais foram preservados para garantir a integridade da análise e futuras investigações.

NTILE(4) sobre Taxa de Endividamento

Os clientes foram divididos em quatro grupos com base na sua taxa de endividamento (DTI) numérica, onde os valores de DTI para clientes com salário não informado já foram tratados como NULL. Os clientes com taxa de endividamento = null não participam da formação dos quartis e recebem null para esta coluna. O Quartil 1 agrupa clientes com os menores DTIs válidos, e o Quartil 4 aqueles com os maiores DTIs válidos. Para analisar como o risco de inadimplência se distribui entre os diferentes níveis de endividamento calculado de forma confiável.

Quartil Taxa Endividamento	Mínimo Taxa Endividamento	Máximo Taxa Endividamento	Total Clientes
1	0.0	0.141990564	7106
2	0.142043984	0.290966689	7106
3	0.291059629	0.474236641	7106
4	0.474262869	5696.0	7105

Análise de Risco Relativo

A análise do Risco Relativo (RR) permitiu identificar padrões claros de propensão à inadimplência.

- O **uso do crédito** foi a variável com a maior capacidade para identificar o risco. Clientes no **quartil de maior uso** apresentaram uma taxa de inadimplência de 7,11% e um RR de 44,65, **risco extremamente elevado**.
- A **idade** mostrou **relação inversa com o risco**, quanto mais jovem o cliente, maior o risco (Q1 com RR = 2,46) e quanto mais velho, menor o risco (Q4 com RR = 0,23).
- O **salário**, segmentado por decis, indicou que os extremos de renda são mais estáveis: Decis 9 e 10 apresentaram os menores RRs (0,37 e 0,30), enquanto o **maior risco foi observado no Decil 4** (RR = 2,00), apontando risco em faixas de renda média-baixa.

- O **número de dependentes** teve **correlação positiva** com o risco. O Quartil 4 (maior número de dependentes) com RR = 1,48, frente a RR = 0,78 no Quartil 1 (menor número).
- A **taxa de endividamento** (DTI) também indicou **maior risco no Quartil 4** (RR = 1,52), com o **menor risco relativo no Quartil 3** (RR = 0,73), mostrando que níveis intermediários podem ter melhor desempenho de pagamento.

As tabelas apresentadas a seguir, mostram que algumas variáveis, especialmente **uso do crédito** e **idade**, têm maior capacidade de segmentação.

$$\text{Risco Relativo (RR)} = \frac{\text{Proporção de inadimplentes no grupo}}{\text{Proporção de inadimplentes no grupo de referência}}$$

RR = 1: Indica que a incidência (risco) do desfecho é a mesma no grupo exposto e no grupo não exposto. Não há diferença de risco.

RR > 1: Indica que a incidência do desfecho é maior no grupo exposto em comparação com o grupo não exposto. O fator de risco aumenta o risco.

RR < 1: Indica que a incidência do desfecho é menor no grupo exposto em comparação com o grupo não exposto. O fator de risco é protetor (ou seja, a exposição reduz o risco).

$$\text{RR} = \frac{\text{Inadimplentes no Q1 / Total Q1}}{\text{Inadimplentes nos Q2+Q3+Q4 / Total Q2+Q3+Q4}}$$

Risco Relativo por Quartil de Idade

Quartil (Idade)	Taxa de Inadimplência (%)	Risco Relativo (vs Outros Quartis)
Q1 (Mais Jovens)	3.42	**2.46**
Q2	2.28	1.29
Q3	1.34	0.65
Q4 (Mais Velhos)	0.54	0.23

O Quartil 1 (clientes mais jovens) apresenta a **maior taxa de inadimplência (3,42%)** e um **Risco Relativo de 2,46**.

Indica que este grupo tem aproximadamente **2,5 vezes mais chances de ser inadimplente** em comparação com a média dos clientes agrupados por quartis de idade.

O risco diminui consistentemente nos quartis subsequentes, com o **Quartil 4 (clientes mais velhos) demonstrando o menor risco (Taxa de Inadimplência de 0,54% e RR de 0,23)**.

Risco Relativo por Decil de Salário

Decil 1 = 10% salários mais baixos e Decil 10 = 10% salários mais altos

Decil (Salário)	Taxa de Inadimplência (%)	Risco Relativo (vs Outros Decis)
Decil 1 (Menor Renda)	1.97	1.04
Decil 2	1.61	0.84
Decil 3	2.72	**1.51**
Decil 4	3.44	**2.00**
Decil 5	2.44	**1.33**
Decil 6	2.08	**1.11**
Decil 7	1.86	0.98
Decil 8	1.47	0.76
Decil 9	0.75	0.37
Decil 10 (Maior Renda)	0.61	0.30

Decil 1 representa os 10% de clientes com os menores salários (válidos) e Decil 10 os 10% com os maiores.

A análise por decil de salário revela que os grupos de maior renda (Decis 9 e 10) apresentam as **menores taxas de inadimplência** (0,75% e 0,61%, respectivamente) e são consideravelmente **menos arriscados em relação aos demais** (RR de 0,37 e 0,30).

Existe um pico de risco no Decil 4, que possui uma taxa de inadimplência de 3,44% e um Risco Relativo de 2,00, indicando ser o **grupo mais propenso à inadimplência** quando comparado aos outros decis combinados.

Os decis intermediários apresentam taxas de inadimplência e riscos relativos variados, mas, de forma geral, as faixas de renda mais baixas e médias-baixas (excluindo o Decil 2) tendem a ter um risco maior ou similar à média dos outros decis.

A variável salário, quando segmentada em decis, demonstra ser um bom indicador para diferenciar níveis de risco, especialmente nos extremos de renda e em faixas específicas de renda média-baixa.

Risco Relativo por Quartil Dependentes

Quartil (Dependentes)	Taxa de Inadimplência (%)	Risco Relativo (vs Outros Quartis)
Q1 (Menor Nº de Dependentes)	1.57	0.78
Q2	1.59	0.79
Q3	1.92	1.02
Q4 (Maior Nº de Dependentes)	2.51	**1.48**

O Quartil 1 representa os 25% de clientes com o menor número de dependentes (provavelmente muitos com 0) e o Quartil 4 são os 25% com o maior número.

A análise por quartil de número de dependentes mostra uma tendência de aumento da inadimplência **conforme o número de dependentes aumenta**.

O Quartil 1 (menor número de dependentes) apresenta a **menor taxa de inadimplência** (1,57%) e um Risco Relativo de 0,78 em relação aos demais quartis.

O Quartil 4 (maior número de dependentes) possui a **taxa de inadimplência mais alta** (2,51%) e um Risco Relativo de 1,48, indicando ser o **grupo mais arriscado nesta segmentação**.

Confirmando as descobertas sobre as faixas manuais de dependentes, onde um **maior número de dependentes está associado a um maior risco**.

Risco Relativo por Quartil Uso do Crédito

Quartil (Uso de Crédito)	Taxa de Inadimplência (%)	Risco Relativo (vs Outros Quartis)
Q1 (Menor Uso)	0.09	0.04
Q2	0.01	0.00
Q3	0.38	0.16
Q4 (Maior Uso)	7.11	**44.65**

O Quartil 4 (clientes com maior uso de linha de crédito) apresenta a **taxa de inadimplência mais elevada**, atingindo 7,11%, e um Risco Relativo (RR) de 44,65 em comparação com os demais quartis.

Isso indica que este grupo tem aproximadamente **45 vezes mais chances de ser inadimplente**.

Por outro lado, os quartis de menor uso (Q1 e Q2) demonstram taxas de inadimplência muito baixas (0,09% e 0,01%, respectivamente) e RRs significativamente inferiores a 1 (0,04 e 0,00), indicando um risco consideravelmente menor.

Risco Relativo por Taxa de Endividamento

Quartil (Taxa de Endividamento)	Taxa de Inadimplência (%)	RR (vs Outros Quartis)
Q1 (Menor DTI)	1.70	0.87
Q2	1.86	0.97
Q3	1.48	0.73
Q4 (Maior DTI)	2.56	**1.52**

O Quartil 1 representa os 25% de clientes com o menor DTI (entre aqueles com DTI válido - porque eu tratei esse campo) e Quartil 4 os 25% com o maior DTI válido. Clientes com salário não informado (e, portanto, DTI não calculado) foram "excluídos" da formação desses quartis.

A análise por quartil da taxa de endividamento (DTI), considerando apenas clientes com DTI calculável, mostra que o Quartil 4 apresenta a **maior taxa de inadimplência** (2,56%) e um **maior Risco Relativo** (1,52) em relação aos demais quartis.

O Quartil 3 exibiu a menor taxa de inadimplência (1,48%) e o **menor Risco Relativo** (0,73), enquanto os Quartis 1 e 2 apresentaram taxas de inadimplência de 1,70% e 1,86% e Riscos Relativos de 0,87 e 0,97, respectivamente.

Isso indica que, embora um DTI muito alto (Q4) esteja associado a um risco maior, a relação não é perfeitamente linear nos quartis inferiores, com o **Quartil 3 mostrando o melhor desempenho em termos de adimplência**.

Risco Relativo por Histórico de Atraso

Histórico de Atraso	Taxa de Inadimplência (%)	RR
Não	0.00	0.00
Sim	9.35	4.93

A variável Histórico de Atraso classifica os clientes com base na **ocorrência de atrasos em pagamentos anteriores**. Clientes que registraram **pelo menos um atraso superior a 30 dias** foram classificados como "Sim", enquanto os demais foram marcados como "Não".

A análise mostra que clientes com histórico de atraso apresentam uma **taxa de inadimplência de 9,35%**, enquanto aqueles sem histórico têm taxa nula.

O risco relativo (RR) de inadimplência para o grupo com histórico de atraso é **4,93 vezes maior do que a média da base**, evidenciando **forte correlação entre atrasos passados e inadimplência futura**. Logo, é uma variável categórica com alto poder explicativo para o risco de crédito.

Segmentação por Score

Esta análise tem como objetivo verificar se a **classificação baseada no risco relativo das variáveis** é uma métrica válida para o processo de concessão de crédito. Para isso, será criada uma pontuação a partir da soma das categorias das variáveis analisadas, previamente transformadas em variáveis dummy, que assumem valores 0 ou 1 para indicar a presença ou ausência de uma característica.

Variáveis dummy são utilizadas para converter variáveis categóricas em representações booleanas, adequadas para modelos estatísticos.

Estamos dizendo ao modelo: "Este cliente pertence à categoria X? Sim/Não"

Criação das Variáveis Dummy

Para preparar os dados para a **modelagem de risco** e facilitar a **construção de um score de crédito**, foi criada a tabela **Tabela_Variaveis_Dummy** no BigQuery. Nessa mesma tabela, a variável alvo **default_flag** (que corresponde à status_inadimplencia da tabela unificada, originalmente como sim/não) foi padronizada para o formato binário (0 ou 1), **adequado para a modelagem**.

Esta tabela deriva da **Tabela Unificada Auxiliar** e transforma as categorias definidas pelas colunas de **NTILE** (quartis para idade, dependentes, uso_credito, taxa_endividamento e decis para salário), bem como a **informação de histórico de atrasos**, em um conjunto completo de variáveis dummy (0 ou 1).

Cada coluna dummy representa a presença de um cliente em uma categoria NTILE específica (ex: dummy_idade_q1, dummy_idade_q2, etc.; dummy_salario_d4, dummy_salario_d5, etc.) ou indica se o cliente possui um histórico de atrasos (dummy_historico_atrasos_sim).

A decisão de utilizar **uma única variável dummy (dummy_historico_atrasos_sim) para capturar o histórico de pagamentos**, em vez de criar dummies para faixas específicas de atraso (30-59, 60-89, >90 dias), se baseia na **análise da distribuição dos dados**. Eu observei que **a grande maioria dos clientes não apresentava atrasos nem mesmo na primeira faixa**

de 30-59 dias, conforme evidenciado pelos quartis que permaneciam em zero. Dado que os **atrasos são cumulativos**, ou seja, para atingir faixas de atraso mais severas é necessário ter passado pelas anteriores, a **baixa incidência de atrasos** na faixa inicial indicou que a granularização em múltiplas dummies **resultaria em variáveis com pouca variação e diferenciação**, não agregando valor significativo ao modelo de score.

Por isso, a **variável binária dummy_historico_atrasos_sim** foi considerada a forma mais eficiente de incorporar o **impacto do histórico de atrasos** na avaliação de risco.

Esta representação **one-hot encoding** permite que **cada segmento seja tratado como uma característica distinta**, sendo fundamental para a **aplicação da regressão logística** e para a **experimentação flexível na construção da pontuação de risco**. A tabela contém o id_cliente e todas as variáveis dummy geradas.

A validação dessa pontuação será feita com base na variável **indicador de inadimplência**, que informa se o cliente já foi considerado inadimplente, permitindo avaliar a eficácia da classificação dos perfis de risco.

- **dummy_q1_idade** ⇒ RR = 2,46 (jovens com maior risco)
- **dummy_d4_salario** ⇒ RR = 2,00 (faixa de renda com maior risco)
- **dummy_q4_dependentes** ⇒ RR = 1,48 (risco elevado)
- **dummy_q4_uso_credito** ⇒ RR = 44,65 (forte risco)
- **dummy_q4_taxa_endividamento** ⇒ RR = 1,52 (risco elevado)
- **dummy_historico_atrasos_sim** ⇒ RR = 4,93 (risco significativo)

Essas são as variáveis dummy com maior **Risco Relativo** dentro das suas respectivas variáveis categóricas, logo são as mais informativas para detectar inadimplência.

Os **pontos do score** vêm apenas das variáveis dummy que representam quartis com **alto risco relativo**. A lógica é:

Se aquele grupo (quartil) teve um RR alto, significa que pertencer a ele aumenta bastante a chance de inadimplência. Então, se um cliente está nesse grupo (listados acima), ele ganha 1 ponto.

O total de pontos mostra o acúmulo de fatores de risco para aquele cliente.

Usamos apenas as **dummy com RR alto** porque elas são as **mais relevantes** para prever inadimplência. As outras dummies seriam “ruído” e poderiam confundir mais do que ajudar.

Análise de Performance da Pontuação de Risco

Pontuação de Risco	Total de Clientes	Total Inadimplentes	Taxa de Inadimplência (%)

0	11.047	0	0,00%
1	11.109	15	0,14%
2	7.912	146	1,85%
3	4.129	270	6,54%
4	1.461	173	11,84%
5	311	72	23,15%
6	31	7	22,58%

Essa tabela ajuda a entender o comportamento da “pontuacao risco” antes de definir um ponto de corte fixo.

- ★ Ela mostra para cada valor possível do score (0 a 6), quantos clientes totais receberam aquele score.
- ★ Desses clientes, quantos realmente são inadimplentes.
- ★ Qual a taxa de inadimplência real para cada nível de score.

A coluna “taxa de inadimplência” deve aumentar à medida que a “pontuação de risco” aumenta. Você vai procurar um "salto" nessa taxa ou um nível de score onde a taxa de inadimplência se torna inaceitavelmente alta para o banco. Esse score (ou um score um pouco abaixo dele) pode se tornar seu ponto de corte para classificar alguém como "mau pagador previsto".

Matriz de Confusão

A matriz de confusão será utilizada para **validar a pontuação** construída a partir das variáveis com maior risco relativo. Ela permite comparar as previsões de inadimplência com os dados reais.

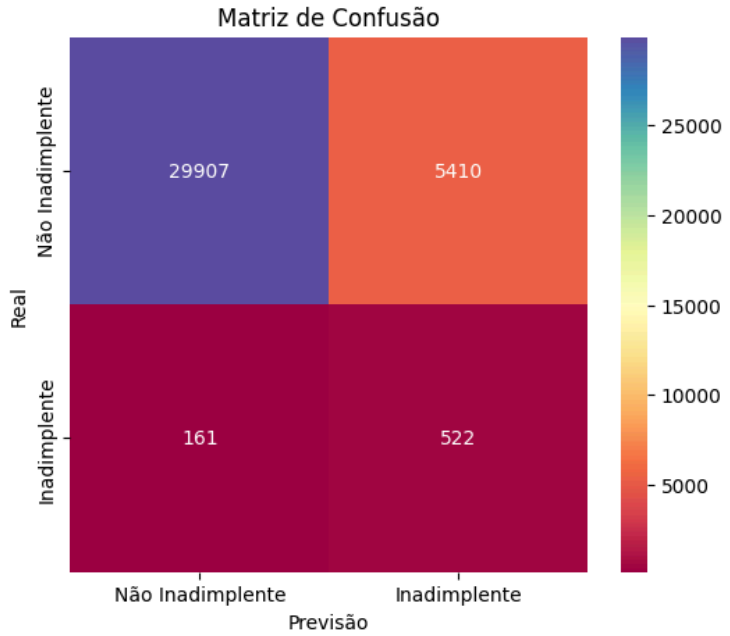
A partir da soma das variáveis dummy selecionadas, é gerada uma pontuação para cada cliente. Essa pontuação é então comparada com o indicador real de inadimplência.

A matriz de confusão apresenta quatro possíveis resultados:

- **Verdadeiro Positivo (VP):** classificado como inadimplente e é inadimplente.
- **Falso Positivo (FP):** classificado como inadimplente, mas na realidade não é.
- **Falso Negativo (FN):** classificado como inadimplente, mas é adimplente.
- **Verdadeiro Negativo (VN):** classificado como inadimplente e é adimplente.

Essas quatro classificações permitem **avaliar a qualidade da segmentação feita por score**, identificando se o modelo está conseguindo distinguir corretamente os perfis de risco.

Matriz de Confusão com Ponto de Corte 3

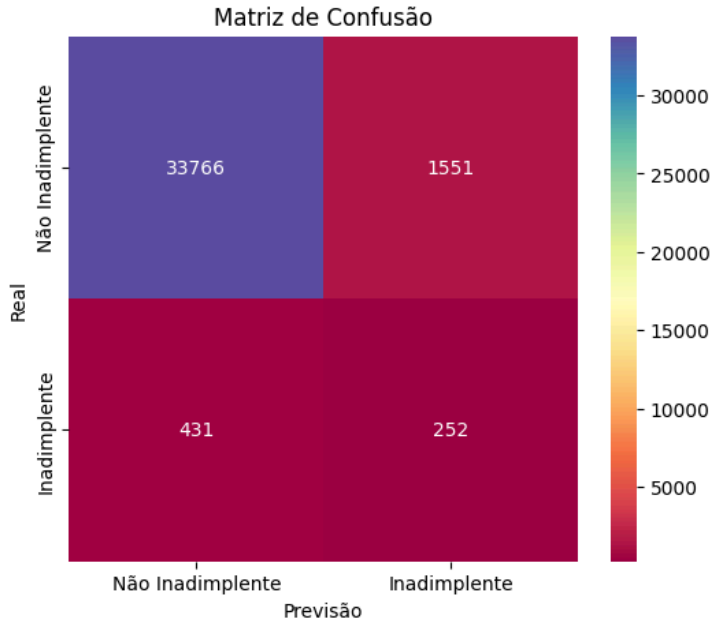


Real Inadimplência	Previsto Inadimplência	Quantidade Clientes
0 (negativo)	0 (negativo)	29.907
0 (negativo)	1 (positivo)	5.410
1 (positivo)	0 (negativo)	161
1 (positivo)	1 (positivo)	522

- Real Inadimplência = 0, Previsto Inadimplência = 0 ⇒ **29.907** (**VN** - Verdadeiro Negativo)
- Real Inadimplentes = 0, Previsto_inadimplencia = 1 ⇒ **5.410** (**FP** - Falso Positivo)
- Real Inadimplência = 1, Previsto Inadimplência = 0 ⇒ **161** (**FN** - Falso Negativo)
- Real Inadimplência = 1, Previsto Inadimplência = 1 ⇒ **522** (**VP** - Verdadeiro Positivo)
- Total de Clientes: **36.000**
- Total de Inadimplentes Reais: VP + FN = 522 + 161 = **683**
- Total de Adimplentes Reais: VN + FP = 29907 + 5410 = **35.317**

Métrica	Fórmula	Corte 3
Acurácia	(VP + VN) / Total	(522 + 29.907) / 36.000 = 0,85
Precisão	VP / (VP + FP)	522 / (522 + 5.410) = 0,088
Recall	VP / (VP + FN)	522 / (522 + 161) = 0,764
Especificidade	VN / (VN + FP)	29.907 / (29.907 + 5.410) = 0,847
F1-Score	2*(Precisão * Recall) / (Precisão + Recall)	2 * (0,088 * 0,764) / (0,088 + 0,764) = 0,158

Matriz de Confusão com Ponto de Corte 4



Real Inadimplência	Previsto Inadimplência	Quantidade Clientes
0 (negativo)	0 (negativo)	33.766
0 (negativo)	1 (positivo)	1.551
1 (positivo)	0 (negativo)	431
1 (positivo)	1 (positivo)	252

- Real Inadimplência = 0, Previsto Inadimplência = 0 ⇒ **33.766** (**VN** - Verdadeiro Negativo)
- Real Inadimplência = 0, Previsto Inadimplência = 1 ⇒ **1.551** (**FP** - Falso Positivo)
- Real Inadimplência = 1, Previsto Inadimplência = 0 ⇒ **431** (**FN** - Falso Negativo)
- Real Inadimplência = 1, Previsto Inadimplência = 1 ⇒ **252** (**VP** - Verdadeiro Positivo)
- Total De Clientes: **36.000**
- Total De Inadimplentes Reais: Vp + Fn = 252 + 431 = **683**
- Total De Adimplentes Reais: Vn + Fp = 29907 + 5410 = **35.317**

Métrica	Fórmula	Corte 4
Acurácia	(VP + VN) / Total	(252 + 33.766) / 36.000 = 0,944
Precisão	VP / (VP + FP)	252 / (252 + 1.551) = 0,1397
Recall	VP / (VP + FN)	252 / (252 + 431) = 0,369
Especificidade	VN / (VN + FP)	33.766 / (33.766 + 1.551) = 0,956
F1-Score	2 * (Precisão * Recall) / (Precisão + Recall)	2 * (0,1397 * 0,369) / (0,1397 + 0,369) = 0,202

Reflexão

Métrica	Corte 3	Corte 4
Acurácia	85%	**94,4%**
Precisão	8,8%	**13,97%**
Recall	**76,4%**	36,9%
Especificidade	84,7%	**95,6%**
F1-Score	15,8%	**20,2%**

Ponto de Corte 3 (Mais Rigoroso para Aprovar)

Nesse corte, somos **mais cautelosos** para conceder crédito. Quando o cliente atinge score 3, negamos o crédito.

Realidade	Previsão	Quantidade	O Que Significa para o Banco?
Bom Pagador	Previsto Bom	29.907	Acertou em aprovar bons clientes (VN)
Bom Pagador	Previsto Ruim	5.410	Errou, negou crédito a bons clientes (FP)
Mau Pagador	Previsto Bom	161	Aprovou quem não deveria (FN)
Mau Pagador	Previsto Ruim	522	Acertou em identificar o risco e negar crédito (VP)

Recall: Conseguiu identificar 76,4% (522 de 683) dos que realmente não pagariam. **Isso é bom para reduzir perdas.**

Precisão: Das vezes que o score negou crédito, apenas 8,8% eram realmente maus pagadores. **Ou seja, muitos alertas foram para bons clientes.**

Ponto de Corte 4 (Mais Flexível para Aprovar)

Nesse corte, somos **mais flexíveis** para conceder crédito. Quando o cliente atinge score 4, negamos o crédito.

Realidade	Previsão	Quantidade	O Que Significa para o Banco?
Bom Pagador	Previsto Bom	33.766	Acertou em aprovar mais bons clientes (VN)
Bom Pagador	Previsto Ruim	1.551	Errou menos, mas ainda negou crédito a alguns bons (FP)
Mau Pagador	Previsto Bom	431	Aprovou mais clientes que não deveriam (FN)
Mau Pagador	Previsto Ruim	252	Identificou menos riscos, mas ainda pegou alguns (VP)

Recall: Identificou apenas 36,9% dos que realmente não pagariam. **Deixou passar mais risco.**

Precisão: Das vezes que o score negou crédito, 13,97% eram realmente maus pagadores. **Melhorou um pouco, mas ainda baixo.**

Especificidade: Melhor (95,6%) em não classificar erroneamente bons clientes como ruins.

Comparando os Dois Cenários

Indicador	Corte 3	Corte 4	O Que isso Significa para o Banco?
Identificar Maus Pagadores (Recall)	**76,4%**	36,9%	O Corte 3 é muito melhor para pegar quem vai dar prejuízo.
Acertar ao Sinalizar Risco (Precisão)	8,8%	**13,97%**	O Corte 4 é um pouco mais preciso quando alerta, mas ambos ainda geram muitos "alarmes falsos".
Ajudar Bons Pagadores (Especificidade)	84,7%	**95,6%**	O Corte 4 é excelente para não negar crédito a bons clientes.
Acertos Gerais (Acurácia)	85%	**94,4%**	O Corte 4 parece melhor no geral, mas o Recall baixo é uma preocupação.

Decisão de Ponto de Corte

Se a prioridade for minimizar perdas com inadimplência, eu recomendaria o Corte 3, pois ele tem um recall alto e **captura a maior parte dos maus pagadores**, reduzindo riscos financeiros. No entanto, isso significa que alguns bons clientes terão crédito negado injustamente.

Por outro lado, se a estratégia for aumentar a base de clientes e estimular a concessão de crédito, eu recomendaria o Corte 4, que é **mais flexível e concede crédito a mais bons pagadores**. Apesar disso, ele deixa passar mais clientes inadimplentes, o que pode aumentar o risco de prejuízos.

	Corte 3	Corte 4
Proteção contra maus pagadores	Alta (76,4% Recall)	Baixa (36,9% Recall)
Clientes bons bloqueados	5.410 (FP)	1.551 (FP)
Risco financeiro	Controlado (161 FN)	Alto (431 FN)
Eficiência operacional	Alta (automatiza 85%)	Média (ainda exige análise)

Por conta do momento com queda nas taxas de juros e aumento expressivo na solicitação de empréstimos, recomendo ser mais conservador e adotar o Corte 3.

O Corte 3 reduz a inadimplência em 76,4%. Agiliza o processo (menos análises manuais). Haveria necessidade de revisão apenas para bons pagadores que fossem bloqueados. Esse cenário é melhor porque o banco terá menos prejuízos com inadimplentes, mas ainda assim os clientes de baixo risco têm oportunidade de revisão. O que torna o processo mais ágil e escalável.

Regressão Logística

Após a análise de risco relativo e a criação de um **score manual**, trabalhei no desenvolvimento de um **modelo estatístico para prever a probabilidade de inadimplência**. Este modelo é adequado para problemas de **classificação binária** (inadimplente vs. adimplente) e pode oferecer insights sobre a importância relativa de cada fator de risco.

Objetivos da Regressão Logística

1. Construir um **modelo preditivo** para classificar novos solicitantes de crédito.
2. Quantificar o **impacto de cada variável** (representada por seus respectivos NTILES/categorias) na probabilidade de inadimplência.
3. Avaliar a **performance do modelo** usando métricas padrão e comparar com a abordagem de score manual.
4. Identificar um **ponto de corte** (threshold) ótimo para a probabilidade predita, visando balancear a identificação de inadimplentes (Recall) com a minimização de falsos positivos (Precisão).

Preparação dos Dados

- **Variáveis Preditoras (X):** Foram utilizadas as variáveis dummy criadas anteriormente, representando os quartis/decis de idade, salário, número de dependentes, uso de linha de crédito, taxa de endividamento e o histórico de atrasos.
    - **Modelo 1:** Utilizando todas as 28 variáveis dummy.
    - **Modelo 2:** Utilizando apenas as 6 variáveis dummy que apresentaram o maior Risco Relativo individual.

Para uma comparação justa e para avaliar o impacto da seleção de variáveis, ambos os modelos foram avaliados utilizando o **mesmo ponto de corte** (threshold) de 0.9. Este threshold foi **identificado como ótimo para o Modelo 1** através de uma otimização baseada em custos. (mais detalhes a seguir)

- **Variável Alvo (y):** A coluna default_flag foi usada como a variável dependente (0 para adimplente, 1 para inadimplente).
- **Divisão dos Dados:** O conjunto de dados foi dividido em 80% para treinamento e 20% para teste (test_size=0.2, random_state=42) para avaliar a generalização do modelo.

Treinamento dos Modelos

- Foi instanciado um modelo **Logistic Regression** da biblioteca **scikit-learn**.
- Parâmetro max_iter=1000 foi usado para garantir a convergência do algoritmo.
- Parâmetro class_weight='balanced'. Por conta do **desbalanceamento de classes** (muito mais adimplentes do que inadimplentes), este parâmetro ajusta os pesos das classes inversamente proporcionais às suas frequências, penalizando mais os erros na classe minoritária (inadimplentes), o que é desejável em problemas de risco.

Otimização do Ponto de Corte (Threshold)

- A conversão de probabilidades (saída da Regressão Logística) em classificações binárias (inadimplente/adimplente) requer a definição de um ponto de corte (threshold). O valor padrão de 0.5 nem sempre é o ideal, especialmente quando os custos de diferentes tipos de erro são assimétricos.
- Para o Modelo 1 (Todas as Variáveis), foi implementada uma **estratégia de otimização de threshold** baseada na minimização de uma função de custo personalizada. Esta abordagem busca um ponto de corte que melhor se alinhe aos objetivos de negócio do banco.

Definição dos Custos

- **Custo de um Falso Positivo** (FP - classificar um bom pagador como inadimplente): *cust_fp = 1 unidade*.
- **Custo de um Falso Negativo** (FN - classificar um mau pagador como adimplente): *cust_fn = 5 unidades*.

Esta atribuição reflete a premissa de que **é cinco vezes mais custoso para o banco aprovar um crédito para quem se tornará inadimplente** (FN) do que negar crédito a um bom pagador (FP).

Processo de Otimização

- O modelo treinado foi usado para prever as probabilidades de inadimplência para o conjunto de teste.
- Foi realizada uma iteração através de múltiplos valores de threshold, de 0.0 a 0.99, com incrementos de 0.01.
- Para cada threshold, os clientes foram classificados como inadimplentes ou adimplentes.
- A matriz de confusão foi calculada, e o custo total foi computado usando a fórmula: Custo Total = (Número de FP * cust_fp) + (Número de FN * cust_fn).
- O threshold que resultou no menor Custo Total foi selecionado como o ***melhor_threshold***.

Comparação dos Modelos

Modelo 1 (Todas as Variáveis):

- AUC-ROC: 0.9651
- Recall: 0.8077
- Precisão: 0.2092
- F1-Score: 0.3323

Modelo 2 (Variáveis de Alto RR):

- AUC-ROC: 0.9590
- Recall: 0.7769
- Precisão: 0.1881
- F1-Score: 0.3028

O Modelo 1 demonstrou uma performance superior em todas as métricas chave (AUC-ROC, Recall, Precisão e F1-Score) quando comparado ao Modelo 2 com o mesmo threshold.

A maior AUC-ROC do Modelo 1 (0.9651 vs. 0.9590) indica uma **melhor capacidade geral de discriminação**.

Avaliação do Modelo

Modelo 1: Regressão Logística com Todas as Variáveis Dummy (Threshold Otimizado = 0.9)

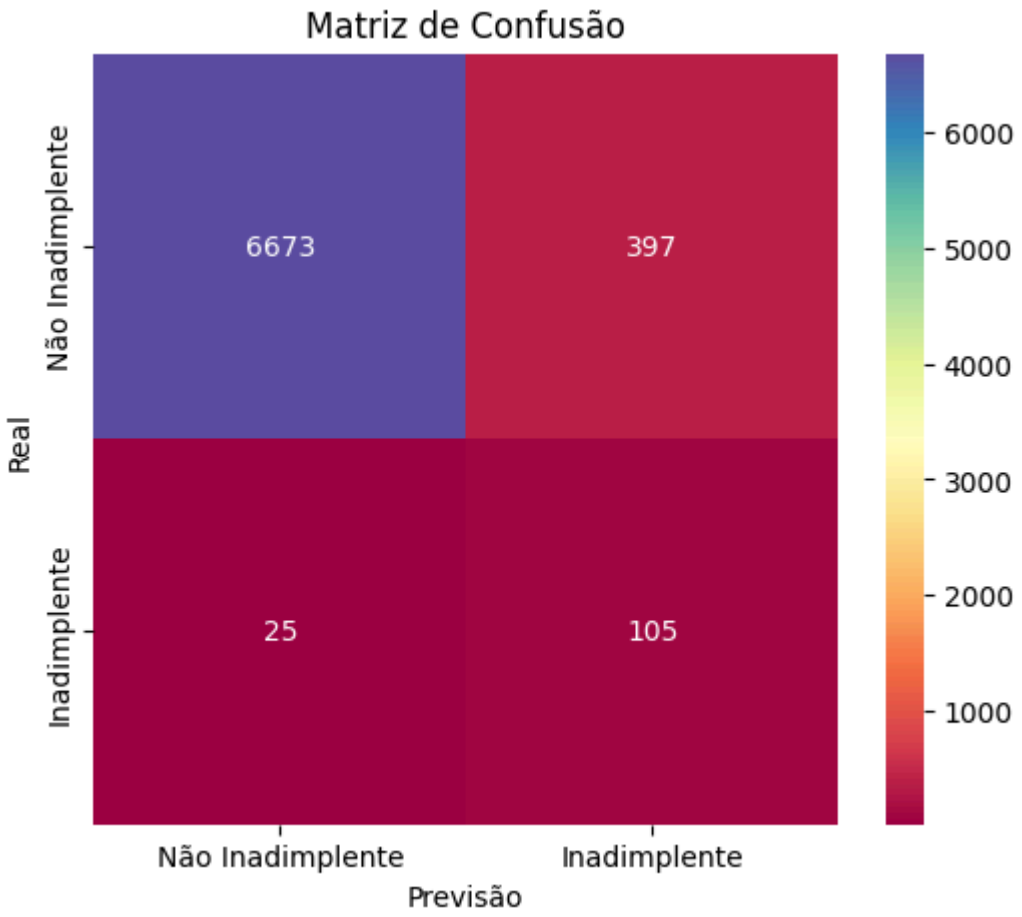
Métricas Principais (com threshold = 0.9):

- **Acurácia: 0.9414** - O modelo acertou a classificação (inadimplente ou não) em 94.14% dos casos no conjunto de teste.

- **Precisão: 0.2092** - Das vezes que o modelo previu que um cliente seria inadimplente, ele acertou em apenas 20.92% das vezes. Isso indica um número alto de “alarmes falsos”.
- **Recall (Sensibilidade): 0.8077** - O modelo conseguiu identificar corretamente 80.77% de todos os clientes que realmente eram inadimplentes. Este é um bom valor para redução de perdas.
- **F1 Score: 0.3323** - A média harmônica entre precisão e recall. O valor baixo reflete o desequilíbrio entre a precisão (baixa) e o recall (alto).
- **AUC-ROC: 0.9651** - Este valor indica uma **excelente capacidade do modelo de distinguir entre as classes** (inadimplente vs. adimplente) em diferentes pontos de corte. Um AUC tão alto é um ótimo sinal da qualidade preditiva intrínseca do modelo.

Matriz de Confusão (Teste com threshold = 0.9):

- **VN (Verdadeiro Negativo): 6673**
    - O modelo **previu corretamente** que 6673 clientes não seriam inadimplentes, e eles realmente não foram.
    - (Real = Não Inadimplente, Previsto = Não Inadimplente)
- **FP (Falso Positivo): 397**
    - O modelo **previu incorretamente** que 397 clientes seriam inadimplentes, mas eles na verdade não foram (eram bons pagadores).
    - (Real = Não Inadimplente, Previsto = Inadimplente)
- **FN (Falso Negativo): 25**
    - O modelo **previu incorretamente** que 25 clientes não seriam inadimplentes, mas eles na verdade foram (eram maus pagadores que o modelo não pegou).
    - (Real = Inadimplente, Previsto = Não Inadimplente)
- **VP (Verdadeiro Positivo): 105**
    - O modelo **previu corretamente** que 105 clientes seriam inadimplentes, e eles realmente foram.
    - (Real = Inadimplente, Previsto = Inadimplente)



- VN (Verdadeiro Negativo): 6673
- FP (Falso Positivo): 397
- FN (Falso Negativo): 25
- VP (Verdadeiro Positivo): 105

O threshold de 0.9 significa que **o modelo precisa ter uma probabilidade predita de inadimplência de 90% ou mais** para classificar um cliente como inadimplente. Isso é um critério bem rigoroso.

Ao ser tão rigoroso, o modelo consegue um Recall alto (80.77%), o que significa que ele identifica a grande maioria dos verdadeiros inadimplentes. Isso é bom porque os Falsos Negativos (não identificar um inadimplente) foram definidos como 5 vezes mais custosos.

Com FN = 25, o custo associado é 25 * 5 = 125.

Porém, essa rigorosidade para evitar aprovar maus pagadores leva a uma **precisão baixa** (20.92%). O modelo acaba classificando muitos bons pagadores como de risco (FP = 397). O custo associado é 397 * 1 = 397.

Custo Total para este Threshold (0.9): 125 (de FN) + 397 (de FP) = 522.

Esse custo (522), segundo os testes para otimizar o threshold, deve ser o menor (ou um dos menores) entre todos os thresholds testados de 0.0 a 0.99.

Coeficientes mais Notáveis

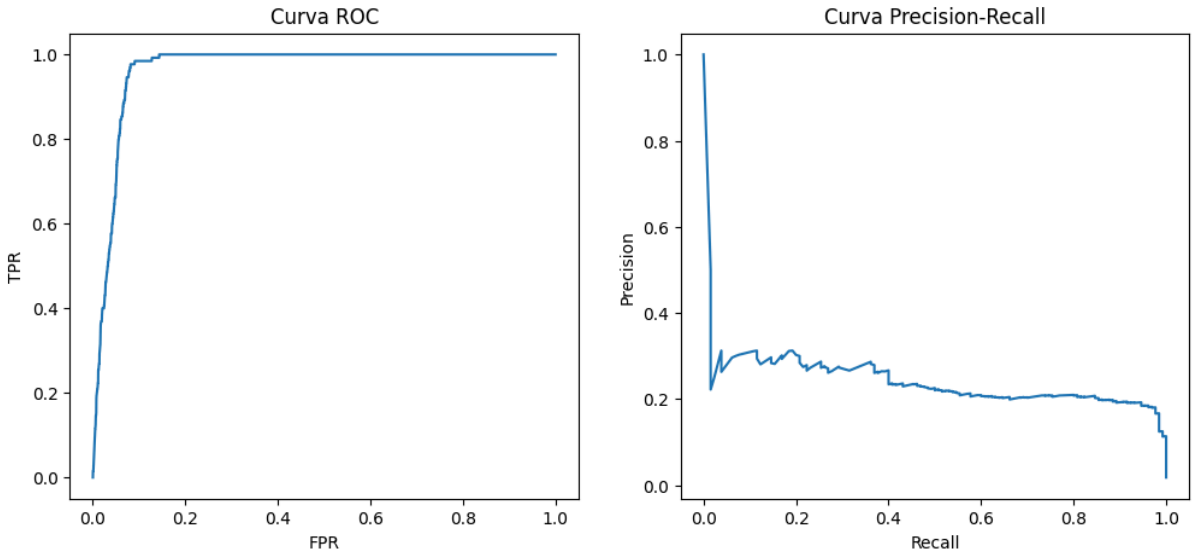
Esta parte permanece a mesma, porque os coeficientes do modelo não mudam com o threshold de decisão.

Feature	Coeficiente	Interpretação
dummy_historico_atrasos_sim	+3.488	Histórico de atraso aumenta risco
dummy_uso_credito_q4	+2.058	Uso de crédito muito alto aumenta risco
dummy_salario_d4	+0.289	Salário na categoria 4 aumenta risco (leve)
dummy_uso_credito_q2	-2.863	Uso de crédito na categoria 2 diminui risco
dummy_historico_atrasos_nao	-5.214	Sem histórico de atraso reduz muito o risco

Curvas de Performance

Estas curvas também são independentes do threshold de decisão final, porque mostram a performance em todos os thresholds.

- Curva ROC: Mostra boa capacidade de discriminação (AUC = 0.906).



O gráfico da **Curva ROC** demonstra uma performance excelente do Modelo 1. A curva ascende rapidamente em direção ao canto superior esquerdo, indicando que o modelo alcança uma **alta Taxa de Verdadeiros Positivos (Recall)** com uma **baixa Taxa de Falsos Positivos**.

A curva atinge um platô com TPR próximo de 1.0, sugerindo que o modelo é capaz de identificar quase todos os inadimplentes reais a partir de certos pontos de corte.

A AUC-ROC (Área Sob a Curva ROC) de 0.9651 confirma quantitativamente essa **capacidade de discriminação**, indicando uma probabilidade de 96,51% de o modelo ranquear corretamente um inadimplente acima de um adimplente.

A **Curva Precision-Recall** é informativa devido ao **desbalanceamento de classes** (poucos inadimplentes).

O gráfico mostra que o modelo atinge uma precisão muito alta (próxima de 1.0) quando o recall é baixo (ou seja, quando o modelo é muito restritivo e identifica poucos como inadimplentes, estes são quase sempre corretos).

Conforme o recall aumenta (na tentativa de identificar mais inadimplentes) a precisão tende a diminuir, o que é um comportamento esperado. A curva se estabiliza em torno de uma precisão de 0.2-0.3 para uma faixa significativa de recall.

O ponto de corte otimizado de 0.9, que resultou em um Recall de 80.77% e Precisão de 20.92%, representa uma **escolha estratégica nesse trade-off**, priorizando a identificação de inadimplentes (alto recall) em detrimento de uma precisão mais alta, alinhado com a função de custo definida.

A análise visual dessas curvas reforça que o Modelo 1 é eficaz. A **Curva ROC** atesta sua **forte capacidade de discriminação geral**, enquanto a **Curva Precision-Recall** oferece uma **perspectiva clara sobre o desempenho na identificação da classe minoritária** e os desafios associados à precisão em cenários de alto recall.

Das vezes que o modelo previu que um cliente seria inadimplente, ele acertou em apenas 20.92% das vezes. Isso indica um número alto de “alarmes falsos”.

A baixa precisão, apesar do bom Recall e da excelente AUC-ROC, é uma característica comum em problemas com **grande desbalanceamento de classes**, como é o caso da inadimplência neste dataset, onde apenas uma pequena porcentagem de clientes pertence à classe de interesse (inadimplentes).

Imagine um cenário onde você precisa classificar fotos entre “cachorros” e “gatos”, mas 98% das suas fotos são de cachorros e apenas 2% são de gatos.

Um modelo poderia alcançar uma **acurácia muito alta** (por exemplo, 98%) simplesmente classificando todas as fotos como “cachorro”. **Ele erraria apenas nos 2% de gatos.**

No nosso caso, é mais fácil para o modelo “acertar” ao prever que um cliente é adimplente (a classe majoritária).

Quando o modelo tenta identificar os “gatos” (os inadimplentes), ele pode ser mais cauteloso ou, ao tentar capturar o máximo possível deles (alto Recall), acabar classificando alguns “cachorros” (adimplentes) como “gatos” (inadimplentes). Estes são os Falsos Positivos.

A métrica de Precisão foca justamente nisso: das vezes que o modelo diz “isto é um gato” (ou “este cliente é inadimplente”), **quantas vezes ele realmente acerta?**

Com apenas 2% de “gatos” reais, mesmo um bom modelo pode ter dificuldade em alcançar uma precisão muito alta sem sacrificar demais a capacidade de encontrar os “gatos” (Recall).

O uso do parâmetro ***class_weight='balanced'*** no treinamento e a otimização do threshold baseada em custos (onde errar um inadimplente é mais custoso) incentivam o modelo a dar **mais atenção à classe minoritária** (inadimplentes), o que melhora o Recall, mas pode, como efeito colateral, **aumentar os Falsos Positivos** e, consequentemente, reduzir a Precisão. A Curva Precision-Recall ilustra bem esse trade-off.

Conclusões da Regressão Logística

O Modelo 1 (todas as 28 variáveis dummy) apresentou um desempenho superior ao Modelo 2 (apenas 6 variáveis de alto Risco Relativo), mesmo quando avaliados sob o mesmo ponto de corte de 0.9.

Isso sugere que, embora as variáveis de alto RR sejam individualmente importantes, o conjunto completo de variáveis NTILE captura **nuances e interações que contribuem para um poder preditivo geral maior**, como evidenciado pela AUC-ROC de 0.9651 para o Modelo 1 contra 0.9590 para o Modelo 2.

Com o threshold de 0.9, o Modelo 1 alcançou um Recall de 80.77%, **identificando a maioria dos inadimplentes**, o que está alinhado com o objetivo de minimizar perdas (dado o custo maior atribuído aos Falsos Negativos).

A Precisão de 20.92% indica que, embora eficaz em pegar inadimplentes, o modelo gera um número considerável de Falsos Positivos.

O Modelo de Regressão Logística (Modelo 1, threshold 0.9) superou o score manual com Corte 3 (Recall: 76.4%, Precisão: 8.8%) em ambas as métricas, oferecendo uma **identificação mais eficaz de inadimplentes e uma precisão significativamente melhor**, embora ainda com espaço para otimização.

Indicador	Score Manual (Corte 3)	Regressão Logística (Mod 1, thr=0.9)	Vantagem

Proteção contra Maus Pagadores (Recall)	Alta (76,4% Recall)	Muito Alta (80,77% Recall)	RL identifica proporção maior de inadimplentes
Clientes Bons Bloqueados (Falsos Positivos)	Alto (5.410 FP)	Significativamente Menor (397 FP)[1]	RL reduz drasticamente bloqueio de bons pagadores
Risco Financeiro (Inadimplentes Não Detectados) (FN)	Controlado (161 FN)	Muito Baixo (25 FN)	RL deixa passar bem menos maus pagadores
Precisão ao Sinalizar Risco	Baixa (8,8%)	Melhor, mas Baixa (20,92%)	RL é mais precisa, mas FP ainda é ponto de atenção
Eficiência Operacional (Acurácia)	Alta (85% Acurácia)	Muito Alta (94,14% Acurácia)	RL permite automatizar mais decisões corretamente
Capacidade de Discriminação Geral (AUC-ROC)	Não medida (AUC não aplicável)	Excelente (AUC-ROC: 0.9651)	RL tem capacidade superior de separar as classes

- O modelo de RL não só **captura uma porcentagem maior de inadimplentes reais** (Recall mais alto), como também **reduz drasticamente o número de maus pagadores não detectados** (FN muito menor), o que se traduz em menor risco financeiro.
- Uma das melhorias mais interessantes é a **redução no número de clientes bons bloqueados** (Falsos Positivos). Enquanto o score manual com Corte 3 bloqueava mais de 5 mil bons clientes, a RL reduz esse número para menos de 400 (no conjunto de teste), melhorando a experiência do cliente e o potencial de negócios.
- Embora a precisão do modelo de RL ainda apresente espaço para melhorias, ela é **mais que o dobro da obtida com o score manual**. A acurácia superior também indica uma **maior eficiência operacional**, com mais decisões podendo ser automatizadas corretamente.
- **A métrica AUC-ROC indica excelente capacidade de distinguir entre clientes adimplentes e inadimplentes de forma geral.**

Embora o **Score Manual com Corte 3** tenha sido uma boa primeira abordagem para segmentar o risco, o **Modelo de Regressão Logística** oferece um avanço substancial em termos de precisão, redução de risco e eficiência.

Como o Modelo de Regressão Logística Seria Usado na Prática?

Diferente do score manual que resulta em uma pontuação discreta (0-6) que podemos adicionar como uma coluna, o modelo de Regressão Logística treinado é um objeto matemático/estatístico que aprendeu os relacionamentos entre as variáveis de entrada (dummies) e a probabilidade de inadimplência.

Para usá-lo:

1. **Novos Solicitantes de Crédito:** Quando um novo cliente solicita crédito, o banco coleta as informações necessárias para gerar as mesmas 28 variáveis dummy que usou para treinar o Modelo 1 (dummy_idade_q1, dummy_salario_d4, etc.).
    - Isso significa que o sistema do banco precisaria ter a lógica para categorizar a idade do novo cliente em um dos quartis, o salário em um dos decis, e assim por diante, para então gerar os valores 0 ou 1 para cada uma das 28 dummies.
2. **Aplicando o Modelo Treinado:** Essas 28 variáveis dummy do novo solicitante são então alimentadas no modelo de Regressão Logística.
    - O modelo, com base nos coeficientes que ele aprendeu, calculará uma probabilidade de inadimplência para esse novo solicitante. Essa probabilidade será um número entre 0 e 1 (ex: 0.05, 0.30, 0.75, 0.95).
3. **Tomada de Decisão com Base no Threshold:** Essa probabilidade predita é então comparada com o ponto de corte (threshold) que o modelo encontrou (0.9 no caso).
    - **Se Probabilidade ≥ 0.9:** O cliente é classificado como de alto risco (potencial inadimplente). A política do banco pode ser negar o crédito automaticamente ou encaminhar para uma análise de risco mais aprofundada (conforme recomendação sobre Falsos Positivos).
    - **Se Probabilidade < 0.9:** O cliente é classificado como de baixo risco (potencial adimplente) e o crédito pode ser aprovado (possivelmente com limites e condições adequadas).

Links:

- Repositório no GitHub
- Dashboard no Looker Studio
- Apresentação no Google Slides
- Vídeo no Loom

Apêndice

One-Hot Encoding

One-Hot Encoding é uma técnica muito comum usada no pré-processamento de dados, especialmente para machine learning e estatística, **para converter variáveis categóricas em um formato numérico** que os algoritmos possam entender e processar.

Imagine que você tem uma variável categórica como "Cor" com três valores possíveis: "Vermelho", "Verde" e "Azul". A maioria dos algoritmos de machine learning não consegue trabalhar diretamente com esses rótulos textuais. Eles precisam de números.

O que o One-Hot Encoding faz?

Ele transforma essa única coluna categórica em múltiplas novas colunas binárias (dummy variables), onde cada nova coluna representa uma das categorias originais.

Para cada observação (linha) nos seus dados:

- Apenas uma dessas novas colunas dummy terá o valor 1 (indicando que a observação pertence àquela categoria).
- Todas as outras colunas dummy para aquela variável original terão o valor 0.

Exemplo:

Suponha que sua variável original seja cor_favorita:

id_cliente	cor_favorita
1	Vermelho
2	Verde
3	Azul
4	Vermelho

Após o One-Hot Encoding, ela se transformaria em algo assim:

id_cliente	cor_favorita_Vermelho	cor_favorita_Verde	cor_favorita_Azul
1	1	0	0
2	0	1	0
3	0	0	1
4	1	0	0

Por que "One-Hot"?

O nome vem da ideia de que, para cada linha, apenas uma das novas colunas dummy está "quente" (com o valor 1), enquanto as outras estão "frias" (com o valor 0).

Por que usar One-Hot Encoding?

- **Evita Ordenação Falsa:** Se você simplesmente atribuísse números ordinais às categorias (ex: Vermelho=1, Verde=2, Azul=3), o algoritmo poderia erroneamente assumir que existe uma relação de ordem ou magnitude entre elas (que "Azul" é "maior" que "Verde", ou que a diferença entre 1 e 2 é a mesma que entre 2 e 3). One-Hot Encoding evita isso, pois trata cada categoria como distinta e independente.
- **Compatibilidade com Algoritmos:** Muitos algoritmos (como regressão linear, regressão logística, redes neurais, máquinas de vetores de suporte) esperam entradas numéricas.
- **Interpretabilidade em Alguns Modelos:** Em modelos como regressão linear ou logística, o coeficiente associado a cada variável dummy pode ser interpretado como o impacto daquela categoria específica na variável alvo, em relação a uma categoria de referência (se uma dummy for omitida para evitar multicolinearidade).

Quando criei as colunas dummy para os NTILES na query SQL:

CASE WHEN quartil_idade_ntile = 1 THEN 1 ELSE 0 END AS dummy_idade_q1,

CASE WHEN quartil_idade_ntile = 2 THEN 1 ELSE 0 END AS dummy_idade_q2,

CASE WHEN quartil_idade_ntile = 3 THEN 1 ELSE 0 END AS dummy_idade_q3,

CASE WHEN quartil_idade_ntile = 4 THEN 1 ELSE 0 END AS dummy_idade_q4,

Estava, na prática, fazendo um One-Hot Encoding da variável quartil_idade_ntile. Se um cliente está no quartil 1, dummy_idade_q1 é 1 e as outras são 0.

A função pd.get_dummies() em Pandas é uma maneira automatizada de fazer o One-Hot Encoding.

Consideração: Multicolinearidade (e drop_first=True)

Se incluimos uma coluna dummy para todas as N categorias de uma variável categórica, essas colunas dummy serão perfeitamente multicolineares. Isso significa que uma coluna pode ser perfeitamente prevista pelas outras.

Exemplo: se cor_favorita_Vermelho = 0 e cor_favorita_Verde = 0, então cor_favorita_Azul tem que ser 1).

Isso pode ser um problema para alguns modelos estatísticos (especialmente regressões lineares).

Para evitar isso:

- A prática comum é remover uma das colunas dummy. A categoria correspondente à dummy removida se torna a categoria de referência.
- A função pd.get_dummies(..., drop_first=True) em Pandas faz isso automaticamente.

No entanto, para alguns algoritmos baseados em árvores (como Random Forest, XGBoost), a multicolinearidade não é um problema tão grande, e você pode manter todas as dummies.

Em resumo, One-Hot Encoding é simplesmente uma forma de representar informações categóricas como um conjunto de indicadores binários (0 ou 1), tornando-as adequadas para uso em modelos de aprendizado de máquina e análises estatísticas.

Organização dos Scripts SQL para Análise de Risco

Para facilitar o desenvolvimento, a depuração e o entendimento do processo de criação do score de risco e sua validação, optei por dividir a lógica SQL no BigQuery em **múltiplos scripts autônomos**. Esta abordagem modular permite que cada etapa do cálculo seja executada e verificada individualmente.

Os scripts foram nomeados sequencialmente (ex: 01_dados_para_score_base.sql, 02_score_por_cliente.sql, etc.) e cada um cumpre um objetivo específico no fluxo de análise:

- **01_dados_para_score_base.sql:** Este script é responsável por preparar o conjunto de dados inicial para o cálculo do score. Ele realiza o join entre a tabela_unificada_auxiliar (contendo as características dos clientes e NTILES) e a tabela_variavel_dummy (contendo as variáveis indicadoras binárias), selecionando as seis dummies de alto risco identificadas e o indicador_inadimplencia real. O filtro indicador_inadimplencia IS NOT NULL é aplicado para garantir a consistência nas análises de performance subsequentes.
- **02_score_por_cliente.sql:** Utilizando o resultado do script anterior (conceitualmente, através de uma CTE), este script calcula a pontuacao_risco para cada cliente, somando as seis variáveis dummy de alto risco.

- **03_previsao_por_cliente.sql:** Este script aplica um ponto de corte definido (ex: >=4) à pontuacao_risco calculada no passo anterior para gerar a previsao_inadimplencia (0 ou 1) para cada cliente.
- **04_analise_performance_score.sql:** Agrega os dados por pontuacao_risco para analisar a distribuição de clientes e a taxa de inadimplência real em cada nível de score. O resultado desta análise é fundamental para a escolha e validação do ponto de corte.
- **05_matriz_confusao.sql:** Com base na previsao_inadimplencia (utilizando o ponto de corte definido) e no indicador_inadimplencia real, este script gera as contagens para os quatro quadrantes da Matriz de Confusão (VN, FP, FN, VP).
- **06_visualizacao_amostra.sql:** Script para visualização de uma amostra de clientes, mostrando seus dados base, score e previsão, para fins de verificação e entendimento.

Embora cada script possa ser executado independentemente para fins de teste e aprendizado, em um fluxo de produção final para atualizar a tabela analítica principal, a lógica dessas CTEs será combinada em uma única query CREATE OR REPLACE TABLE para maior eficiência.

Após as análises, uma coluna com score de cada cliente foi criada na tabela unificada.

Esta abordagem de scripts separados foi adotada para fins didáticos e para facilitar a compreensão de cada componente do processo de scoring e validação.