

## **Ficha Técnica do Projeto**

**Projeto: Projeto 3 - Risco Relativo**

**Analista de Dados: Bruna Derner**

**Ferramentas utilizadas: Big Query, Looker, Google Colab, Google Apresentações e Loom.**

### 1. Objetivo do Projeto

### 2. BigQuery

#### 2.1. Importação de dados

#### 2.2. Identificar e Tratar Nulos

#### 2.3. Tratar Duplicados

#### 2.4. Dados fora do escopo

#### 2.5. Dados discrepantes

##### 2.5.1. Variáveis Categóricas

##### 2.5.2. Variáveis Numéricas

#### 2.6. Criar novas variáveis

#### 2.7. Unir tabelas

#### 2.8. Quatil

### 3. Looker Studio

### 4. Análise de Risco Relativo

#### 4.1. Identificação e Separação

#### 4.2. Segmentação

##### 4.2.1. Regressão Logística

###### 4.2.1.1. Modelo com Cálculo de Risco Relativo

###### 4.2.1.2. Modelo com Variáveis Dummy

##### 4.2.2. XGBoost

##### 4.2.3. Decision Tree

##### 4.2.4. AUC ROC

## **1. Objetivo do Projeto**

O objetivo da análise é identificar o perfil de clientes com risco de inadimplência, desenvolver uma pontuação de crédito com base em dados históricos e avaliar o risco relativo de cada cliente. A partir dessa classificação, busca-se categorizar os clientes em diferentes níveis de risco de crédito, permitindo ao banco tomar decisões mais seguras na concessão de empréstimos. Essa abordagem visa reduzir a inadimplência, aumentar a eficiência operacional e fortalecer a saúde financeira da instituição por meio da automação do processo de análise de crédito.

## **2. BigQuery**

### **2.1. Importação de dados**

Inicialmente, editei o arquivo no Excel utilizando a função "Texto para colunas", com delimitador por vírgula, e salvei no formato CSV UTF-8. Em seguida, realizei a importação do arquivo para o BigQuery via upload de arquivo local. Criei um conjunto de dados

nomeado como proj03 e a tabela `user_info`. O esquema foi detectado automaticamente, sem particionamento. Nas opções avançadas, defini a marcação como v2, o delimitador como ponto e vírgula (;), aspas duplas como caractere de texto, e marquei a opção para considerar novas linhas dentro de aspas. Após essas configurações, finalizei com a criação da tabela. Realizei o mesmo processo para todas as tabelas

## 2.2. Identificar e Tratar Nulos

Foi realizada uma verificação de valores nulos na tabela `user_info`, que contém 36.000 registros. As colunas `user_id`, `age` e `sex` não apresentaram nulos. Identificaram-se 7.199 valores ausentes em `last_month_salary` e 943 em `number_dependents`.

Na tabela `loans_outstanding`, 305.335 registros, e a tabela `loans_detail`, com 36.000 registros, não foram identificados valores nulos.

Já a tabela `default`, 36.000 registros, foram identificados 35.317 valores nulos na variável `default_flag`, ou seja, apenas 683 registros possuem informação sobre inadimplência.

Para tratar os valores nulos na variável `default_flag` da tabela `default`, realizei a substituição dos nulos por 0. Essa decisão foi baseada na descrição da variável, que classifica clientes inadimplentes com valor 1 e clientes sem histórico de inadimplência com valor 0. Como apenas o valor 1 estava presente na base original, considerei que os registros ausentes representavam uma falha de importação e, portanto, deveriam ser tratados como 0 para manter a coerência com a definição da variável.

Os demais valores decidi manter até a união das tabelas em que provavelmente será tratado com “ausente”, ou média, no caso de salários.

## 2.3. Tratar Duplicados

Foi realizada uma verificação de duplicatas na tabela `user_info`, utilizando a coluna `user_id` como referência. A consulta não retornou resultados, indicando que não há registros duplicados de usuários nessa tabela. Cada `user_id` está presente apenas uma vez. O mesmo aconteceu com as tabelas `default` e `loans_detail`.

Na tabela `loans_outstanding`, foram identificados 34.510 usuários com registros duplicados, ou seja, com mais de um empréstimo ativo associado ao mesmo `user_id`. O número de ocorrências por usuário varia.

## 2.4. Dados fora do escopo

Ao realizar análise de correlação entre variáveis, identifiquei

- `more_90_days_overdue` e `number_times_delayed_payment_loan_60_89_days`: com correlação de 0.99217552634075257
- `more_90_days_overdue` e `number_times_delayed_payment_loan_30_59_days`: com correlação de 0.98291680661459857

Se eu optasse por manter as três variáveis no modelo, enfrentaria um problema de multicolinearidade. Esse fenômeno ocorre quando as variáveis independentes apresentam uma forte correlação entre si, o que compromete a estabilidade das estimativas dos coeficientes. Como resultado, torna-se difícil isolar o efeito individual de cada variável sobre

a variável dependente, o que pode levar a interpretações imprecisas e resultados não confiáveis na análise.

Sendo assim, calculei o desvio padrão das variáveis para decidir qual manter.

- `more_90_days_overdue` : 4.12136466842672
- `number_times_delayed_payment_loan_60_89_days`: 4.1055147551019706
- `number_times_delayed_payment_loan_30_59_days`: 4.144020438225871

Apesar de o desvio padrão de `number_times_delayed_payment_loan_30_59_days` ser maior, eu optei por manter apenas a variável `more_90_days_overdue` dado a sua importância para o cenário geral, visto que é uma inadimplência mais grave.

Para evitar qualquer tipo de discriminação eu irei retirar a variável “sex” da análise também.

## 2.5. Dados discrepantes

### 2.5.1. Variáveis Categóricas

Na tabela `loans_outstanding`, a coluna `loan_type` apresentava diversas variações de escrita referentes a apenas duas categorias. Para padronizar essa variável, todos os valores foram convertidos para letras minúsculas e caracteres adicionais foram removidos, garantindo uma padronização nas classificações. Nessa meta também foi realizado a alteração de variáveis em texto para numéricas e assim por diante, cada qual com sua classificação.

### 2.5.2. Variáveis Numéricas

Foi realizada uma análise exploratória da variável `last_month_salary` na tabela `user_info`, considerando apenas os registros não nulos. Os resultados revelaram uma grande dispersão nos dados: o valor mínimo foi 0, o máximo R\$1.560.100,00, a média R\$6.675,05 e a mediana R\$5.366,00. A diferença significativa entre média e mediana, além do valor máximo extremamente elevado, indica a presença de valores atípicos (outliers) que podem influenciar negativamente análises estatísticas e preditivas.

Foi preferível manter os dados como estavam para decidir no Looker Studio e depois retornar para fazer a alteração desses dados.

Após a visualização no Looker Studio pude constatar a presença de outliers e sendo assim, decidi tratar com média (R\$12441,65) os valores acima do limite superior (R\$37.000).

| Linha | quartil_salario | qtd_usuarios | salario_min | salario_max | media_salario      |
|-------|-----------------|--------------|-------------|-------------|--------------------|
| 1     | 1               | 8894         | 0.0         | 3944.0      | 2424.4677310546449 |
| 2     | 2               | 8894         | 3947.0      | 5366.0      | 4932.04564875197   |
| 3     | 3               | 8894         | 5366.0      | 7495.0      | 5963.7769282662466 |
| 4     | 4               | 8893         | 7495.0      | 1560100.0   | 12441.650399190374 |

## 2.6. Criar novas variáveis

Os valores nulos da variável `last_month_salary` foram tratados com a substituição pela mediana (R\$5.366,00), minimizando o impacto de outliers. Além disso, foi criada a variável `classificacao_dependentes`, categorizando o número de dependentes em faixas, incluindo a categoria "ausente" para valores nulos.

A partir da tabela `loans_outstanding`, foram construídas as seguintes variáveis: `total_emprestimos`, que contabiliza o número total de empréstimos por cliente;

`qtde_real_estate`, que indica a quantidade de empréstimos do tipo "real\_state"; e `qtde_other`, correspondente à quantidade de empréstimos classificados como "other". Essas variáveis permitem analisar o perfil de crédito dos usuários com base no tipo e na quantidade de financiamentos contratados.

## 2.7. Unir tabelas

Ao criar uma nova tabela que agrupava as características de empréstimos, foi identificado o total de 35575 linhas. Sendo assim, na hora da união usei JOIN INNER. A tabela `analise` foi criada com base na junção das quatro outras tabelas reunindo as variáveis `user_id`, `last_month_salary`, `faixa dependentes`, `faixa idade`, `more_90_days_overdue`, `using_lines_not_secured_personal_assets`, `debt_ratio`, `default_flag`, `total_emprestimos`, `qtde_real_estate` e `qtde_other`.

Antes de ir ao Looker Studio eu calculei a correlação de outras variáveis.

`debt_ratio` e `using_lines_not_secured_personal_assets` : 0.00081905934460664543

`default_flag` e `using_lines_not_secured_personal_assets` : 0.2805241647024424

`default_flag` e `debt_ratio`: 0.034925303326563469

`default_flag` e `more_90_days_overdue` : 0.424477864163663

`default_flag` e `last_month_salary` : -0.025026936778125137

`default_flag` e `total_emprestimos` : -0.042892348873060683

`default_flag` e `age`: -0.0695661855345519

`last_month_salary` e `total_emprestimos`: 0.096909616691683145

`last_month_salary` e `debt_ratio` : -0.044883753361757796

`qtde_other` e `qtde_real_estate` : 0.22542029340737749

## 2.8. Quartil

Foi criado quartis para as variáveis: `last_month_salary`, `debt_ratio`, `using_lines_not_secured_personal_assets`, `total_emprestimos` e `more_90_days_overdue`.

## 3. Looker Studio

No Looker Studio, foi criado um painel interativo com o objetivo de explorar o perfil dos clientes e identificar padrões associados ao risco de crédito. O dashboard permite filtrar os dados por `default_flag`, facilitando a análise comparativa entre clientes inadimplentes e adimplentes. Foram incluídas visualizações que mostram a distribuição dos usuários por faixa etária e número de dependentes, além de gráficos de dispersão que exploram relações entre variáveis como `total_emprestimos`, `last_month_salary`, `debt_ratio` e `using_lines_not_secured_personal_assets`.

Também foram inseridos gráficos de barras que evidenciam a concentração salarial — com destaque para a faixa de R\$5 mil a R\$10 mil — e a relação entre o tipo de empréstimo contratado (`qtde_real_estate` e `qtde_other`) e a idade dos clientes. Essas visualizações oferecem uma visão clara dos diferentes perfis de usuários e dos fatores que podem influenciar a inadimplência.

## 4. Análise de Risco Relativo

### 4.1. Identificação e Separação

Para realizar a análise de risco relativo, calculou-se os quartis de idade. Apresentando o seguinte resultado:

| Linha | quartil_age | qtd_usuarios | idade_min | idade_max | media_idade        |
|-------|-------------|--------------|-----------|-----------|--------------------|
| 1     | 1           | 8894         | 21        | 42        | 33.944794243310113 |
| 2     | 2           | 8894         | 42        | 52        | 47.024398470879241 |
| 3     | 3           | 8894         | 52        | 63        | 57.37901956375083  |
| 4     | 4           | 8893         | 63        | 109       | 71.790846733385791 |

Foi calculado o risco relativo de inadimplência por faixa etária (quartil\_age), comparando a taxa de cada grupo com a média geral da base. Os resultados mostraram que usuários mais jovens apresentam risco 1,74 vezes maior que a média, enquanto os mais velhos têm risco 70% menor.

| Linha | quartil_age | total | inadimplentes | taxa_quartil          | taxa_geral           | risco_relativo    |
|-------|-------------|-------|---------------|-----------------------|----------------------|-------------------|
| 1     | 1           | 8894  | 271           | 0.03046997976163706   | 0.017484188334504568 | 1.742716286206... |
| 2     | 2           | 8894  | 189           | 0.021250281088374186  | 0.017484188334504568 | 1.215399919162... |
| 3     | 3           | 8894  | 116           | 0.013042500562176748  | 0.017484188334504568 | 0.745959738745... |
| 4     | 4           | 8893  | 46            | 0.0051726076689531091 | 0.017484188334504568 | 0.295844883959... |

A partir da segmentação dos usuários em quartis de salário (quartil\_salario), foi possível identificar uma tendência clara entre renda e inadimplência. Usuários com menor faixa salarial (quartil 1), com média de R\$2.424, apresentaram um risco relativo de 1,57, o que indica que têm 57% mais chances de se tornarem inadimplentes em comparação à média da base. Por outro lado, os usuários no quartil mais alto, com média salarial acima de R\$12 mil, demonstraram um risco relativo de apenas 0,46. Esse padrão evidencia que a probabilidade de inadimplência diminui conforme aumenta a renda mensal. Veja as imagens abaixo.

| Linha | quartil_salario | qtd_usuarios | salario_min | salario_max | media_salario      |
|-------|-----------------|--------------|-------------|-------------|--------------------|
| 1     | 1               | 8894         | 0.0         | 3944.0      | 2424.4677310546435 |
| 2     | 2               | 8894         | 3947.0      | 5366.0      | 4932.045648751965  |
| 3     | 3               | 8893         | 5366.0      | 7494.0      | 5963.6047453052952 |
| 4     | 4               | 8893         | 7495.0      | 730483.0    | 12267.063083323965 |

| Linha | quartil_salario | total | inadimplentes | taxa_quartil         | taxa_geral           | risco_relativo      |
|-------|-----------------|-------|---------------|----------------------|----------------------|---------------------|
| 1     | 1               | 8894  | 245           | 0.027546660670114683 | 0.017484188334504568 | 1.5755184137288261  |
| 2     | 2               | 8894  | 176           | 0.019788621542613    | 0.017484188334504568 | 1.1318009829235651  |
| 3     | 3               | 8894  | 129           | 0.014504160107937935 | 0.017484188334504568 | 0.82955867498374924 |
| 4     | 4               | 8893  | 72            | 0.00809625548183965  | 0.017484188334504568 | 0.46306155750232403 |

De forma complementar, a variável total\_emprestimos também foi dividida em quartis e analisada em relação à inadimplência. O grupo com menor número de empréstimos (quartil 1), com média de 3,19 operações, apresentou o maior risco relativo (1,60), enquanto usuários com maior volume de empréstimos (quartil 4), com média de 15,5, apresentaram um risco de 0,51. Esse comportamento, aparentemente contraintuitivo, pode indicar que usuários com bom histórico de crédito e maior acesso ao sistema financeiro mantêm maior estabilidade no pagamento de dívidas.

| Linha | quartil_total_empres | qtd_usuarios | total_emprestimos_n | total_emprestimos_n | media_total_emprestimos |
|-------|----------------------|--------------|---------------------|---------------------|-------------------------|
| 1     | 1                    | 8894         | 1                   | 5                   | 3.1931639307398254      |
| 2     | 2                    | 8894         | 5                   | 8                   | 6.3141443669889812      |
| 3     | 3                    | 8894         | 8                   | 11                  | 9.3089723409039813      |
| 4     | 4                    | 8893         | 11                  | 57                  | 15.515911390981676      |

| Linha | quartil_total_empres | total | inadimplentes | taxa_quartil         | taxa_geral           | risco_relativo     |
|-------|----------------------|-------|---------------|----------------------|----------------------|--------------------|
| 1     | 1                    | 8894  | 250           | 0.02810883741848437  | 0.017484188334504568 | 1.6076718507437    |
| 2     | 2                    | 8894  | 161           | 0.018102091297503935 | 0.017484188334504568 | 1.0353406718789429 |
| 3     | 4                    | 8893  | 108           | 0.012144383222759474 | 0.017484188334504568 | 0.694592336253486  |
| 4     | 3                    | 8894  | 103           | 0.011580841016415561 | 0.017484188334504568 | 0.6623608025064045 |

A análise da taxa de endividamento (debt\_ratio), definida como a razão entre o total de dívidas e o patrimônio do usuário, foi conduzida por meio da variável quartil\_debt\_ratio, que segmenta os indivíduos em quatro grupos. Observa-se um crescimento expressivo da média do debt\_ratio entre os quartis, indicando uma ampla variação no nível de comprometimento financeiro entre os usuários.

| Linha | quartil_debt_ratio | qtd_usuarios | debt_ratio_min | debt_ratio_max | media_debt_ratio     |
|-------|--------------------|--------------|----------------|----------------|----------------------|
| 1     | 1                  | 8894         | 0.0            | 0.181227143    | 0.074178407682257724 |
| 2     | 2                  | 8894         | 0.181248368    | 0.369452219    | 0.2733208830163032   |
| 3     | 3                  | 8894         | 0.369609856    | 0.884247945    | 0.53851730261198572  |
| 4     | 4                  | 8893         | 0.884336405    | 9184035477.0   | 224755779.00112912   |

Na avaliação do risco relativo, os usuários com maiores níveis de endividamento (quartil 4) apresentaram risco de inadimplência 1,45 vezes superior à média da base. Esses dados evidenciam uma correlação positiva entre alto endividamento e maior probabilidade de inadimplência.

| Linha | quartil_debt_ratio | total | inadimplentes | taxa_quartil         | taxa_geral           | risco_relativo     |
|-------|--------------------|-------|---------------|----------------------|----------------------|--------------------|
| 1     | 3                  | 8894  | 204           | 0.022936811333483246 | 0.017484188334504568 | 1.3118602302068594 |
| 2     | 4                  | 8893  | 160           | 0.017991678848532554 | 0.017484188334504568 | 1.0290256833384979 |
| 3     | 2                  | 8894  | 134           | 0.015066336856307623 | 0.017484188334504568 | 0.8617121119986233 |
| 4     | 1                  | 8894  | 124           | 0.013941983359568248 | 0.017484188334504568 | 0.7974052379688753 |

A variável more\_90\_days\_overdue representa o número de vezes que um usuário teve pagamentos em atraso por mais de 90 dias. Ao segmentar os dados em quartis, observou-se que apenas o quarto grupo apresentou valores diferentes de zero. Nesse grupo, a média foi de aproximadamente 0,35 ocorrências, enquanto os demais apresentaram média nula. A análise de risco relativo evidenciou que usuários no quartil mais alto dessa variável possuem risco de inadimplência cerca de 3,68 vezes maior do que a média da base, o que sugere forte associação entre histórico de atrasos prolongados e inadimplência.

| Linha | quartil_more_90_day | qtd_usuarios | minimo | maximo | media               |
|-------|---------------------|--------------|--------|--------|---------------------|
| 1     | 1                   | 8894         | 0      | 0      | 0.0                 |
| 2     | 2                   | 8894         | 0      | 0      | 0.0                 |
| 3     | 3                   | 8894         | 0      | 0      | 0.0                 |
| 4     | 4                   | 8893         | 0      | 98     | 0.35421117733048457 |

| Linha | quartil_more_90_day | total | inadimplentes | taxa_quartil          | taxa_geral           | risco_relativo       |
|-------|---------------------|-------|---------------|-----------------------|----------------------|----------------------|
| 1     | 4                   | 8893  | 572           | 0.064320251883503876  | 0.017484188334504568 | 3.6787668179351294   |
| 2     | 3                   | 8894  | 19            | 0.0021362716438048123 | 0.017484188334504568 | 0.12218306065652122  |
| 3     | 2                   | 8894  | 16            | 0.0017989655947829998 | 0.017484188334504568 | 0.10289099844759682  |
| 4     | 1                   | 8894  | 15            | 0.0016865302451090623 | 0.017484188334504568 | 0.096460311044622007 |

Foi realizada uma análise do impacto do uso de linhas de crédito não garantidas (variável using\_lines\_not\_secured\_personal\_assets) na inadimplência. A base foi dividida em quartis e observou-se que usuários no quartil 4, com os maiores valores médios de uso dessa linha (R\$

698 milhões), apresentaram a maior taxa de inadimplência, com risco relativo de 2,71 em relação à média geral da base. Em contraste, os usuários nos quartis 2 e 3 apresentaram risco significativamente inferior, com risco relativo abaixo de 0,3, sugerindo que quanto maior o uso dessas linhas, maior o risco de inadimplência.

| Linha | quartil_using_lines_n | qtd_usuarios | minimo      | maximo       | media                 |
|-------|-----------------------|--------------|-------------|--------------|-----------------------|
| 1     | 1                     | 8894         | 0.0         | 0.028829832  | 0.0089065651233415834 |
| 2     | 2                     | 8894         | 0.028830012 | 0.144373673  | 0.073540918719024059  |
| 3     | 3                     | 8894         | 0.144389193 | 0.52937284   | 0.30827315451484161   |
| 4     | 4                     | 8893         | 0.529463329 | 7384615385.0 | 90578172.9825652      |

| Linha | quartil_using_lines_n | total | inadimplentes | taxa_quartil           | taxa_geral           | risco_relativo        |
|-------|-----------------------|-------|---------------|------------------------|----------------------|-----------------------|
| 1     | 4                     | 8893  | 584           | 0.065669627797143817   | 0.017484188334504568 | 3.7559437441855166    |
| 2     | 3                     | 8894  | 29            | 0.003260625140544187   | 0.017484188334504568 | 0.18648993468626923   |
| 3     | 1                     | 8894  | 8             | 0.00089948279739149989 | 0.017484188334504568 | 0.051445499223798409  |
| 4     | 2                     | 8894  | 1             | 0.00011243534967393749 | 0.017484188334504568 | 0.0064306874029748011 |

## 4.2. Segmentação

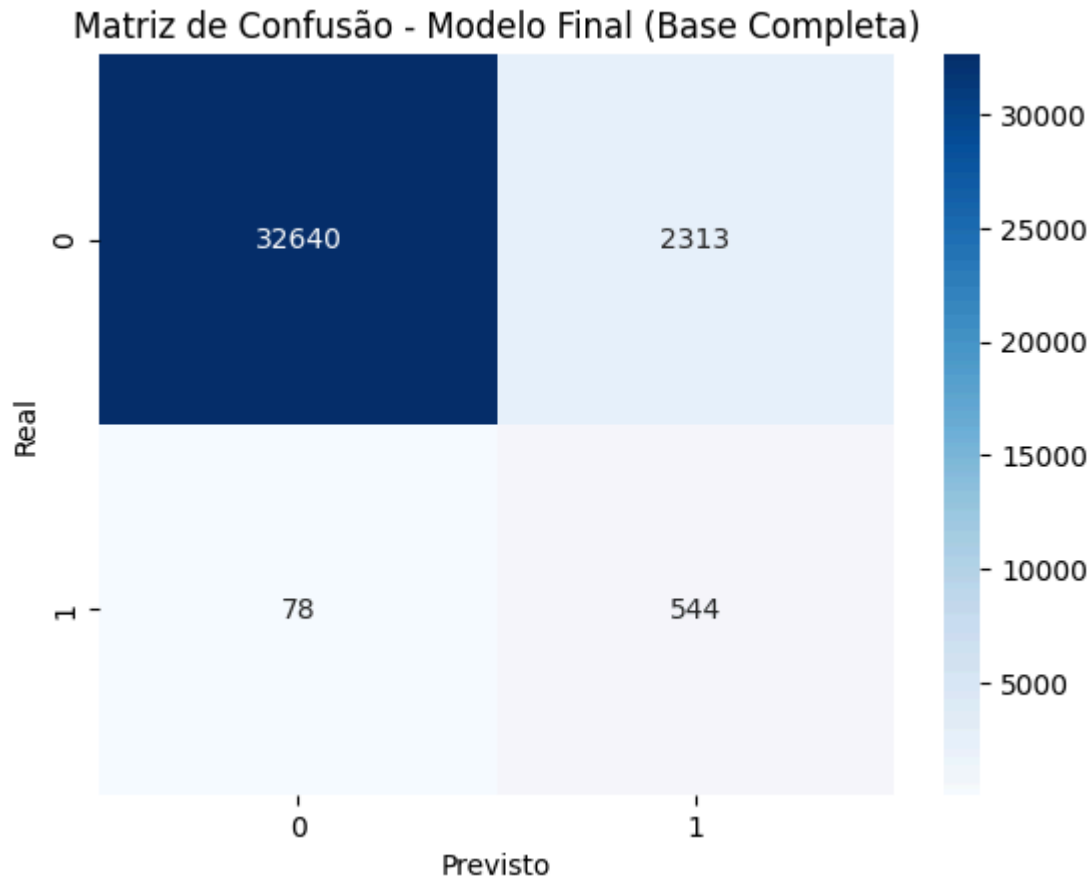
Após calcular o risco relativo para cada dimensão, criamos uma variável dummy correspondente. No caso da idade, como o segundo e o terceiro quartis apresentaram valores muito próximos, incluímos ambos; nas demais variáveis (salário, atraso superior a 90 dias e uso de linhas não garantidas) marcamos apenas o quartil com o maior risco relativo. Após vários testes, identificamos que as quatro variáveis mais explicativas eram idade, salário, atraso >90 dias e uso de linhas não garantidas, e definimos o `score_risco_total` como a soma dessas dummies. Toda essa etapa da análise foi realizada no Google Colab.

### 4.2.1. Regressão Logística

A regressão logística é um algoritmo de aprendizado supervisionado usado para tarefas de classificação binária; ele estima a probabilidade de um evento ocorrer a partir de uma combinação linear das variáveis independentes, aplicando a função logística para mapear essa combinação em um valor entre 0 e 1.

#### 4.2.1.1. Modelo com Cálculo de Risco Relativo

Neste primeiro approach, trabalhamos diretamente com as quatro métricas contínuas de risco relativas: `risco_relativo_idade`, `risco_relativo_salario`, `risco_relativo_more_90_days_overdue` e `risco_relativo_using_lines_not_secured_personal_assets`. Para simplificar a entrada no modelo, somamos os valores dessas quatro variáveis em um único score total de risco relativo. Esse score agregado foi então usado como preditor em uma regressão logística (com `class_weight='balanced'` para corrigir o desbalanceamento de classes), tendo como variável alvo `default_flag`. Dessa forma, o modelo captura o efeito combinado de idade, renda, atraso prolongado e exposição a dívidas não garantidas num único índice contínuo, facilitando tanto a interpretação dos coeficientes quanto a comparação direta entre clientes. O resultado foi o seguinte:

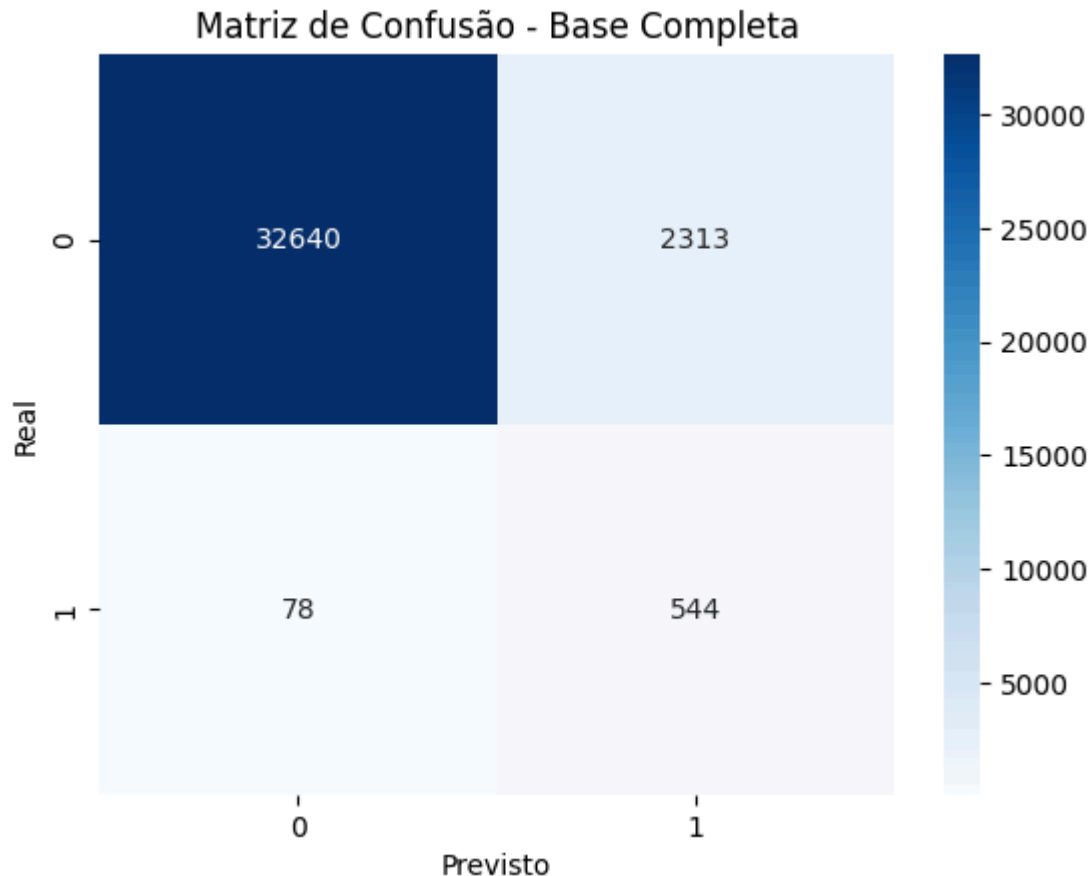


A matriz de confusão mostrou 32.640 verdadeiros negativos, 2.313 falsos positivos, 78 falsos negativos e 544 verdadeiros positivos. Com isso, a acurácia geral foi de 0,93, indicando que 93% das classificações estão corretas — valor elevado, mas influenciado pela forte predominância de clientes adimplentes na base. Para a classe 0 (adimplentes), o modelo apresentou precisão de 1,00, recall de 0,93 e F1-score de 0,96, evidenciando excelente capacidade de identificar corretamente quem paga em dia, sem quase nenhum erro de classificação. Já para a classe 1 (inadimplentes), a precisão foi de apenas 0,19, enquanto o recall atingiu 0,87, resultando em um F1-score de 0,31. Isso indica que, embora o modelo consiga identificar a maioria dos inadimplentes (alta sensibilidade), ele também comete muitos falsos alarmes — classificando muitos adimplentes como devedores.

#### 4.2.1.2. Modelo com Variáveis Dummy

No segundo approach, mantivemos as mesmas quatro dimensões de risco, mas mudamos a forma de representá-las para indicadores binários gerados via SQL. Para cada variável original, criamos uma coluna DUMMY que vale 1 quando o cliente se enquadra na condição de risco (por exemplo, atraso acima de 90 dias) e 0 caso contrário. Em seguida, somamos essas quatro dummies, obtendo um score de contagem de riscos — ou seja, quantas das quatro condições de risco cada cliente atende. Esse score discreto foi então alimentado no mesmo tipo de regressão logística, permitindo avaliar como o número de fatores de risco acumulados impacta a probabilidade de inadimplência, em contraste com a intensidade contínua do primeiro modelo.

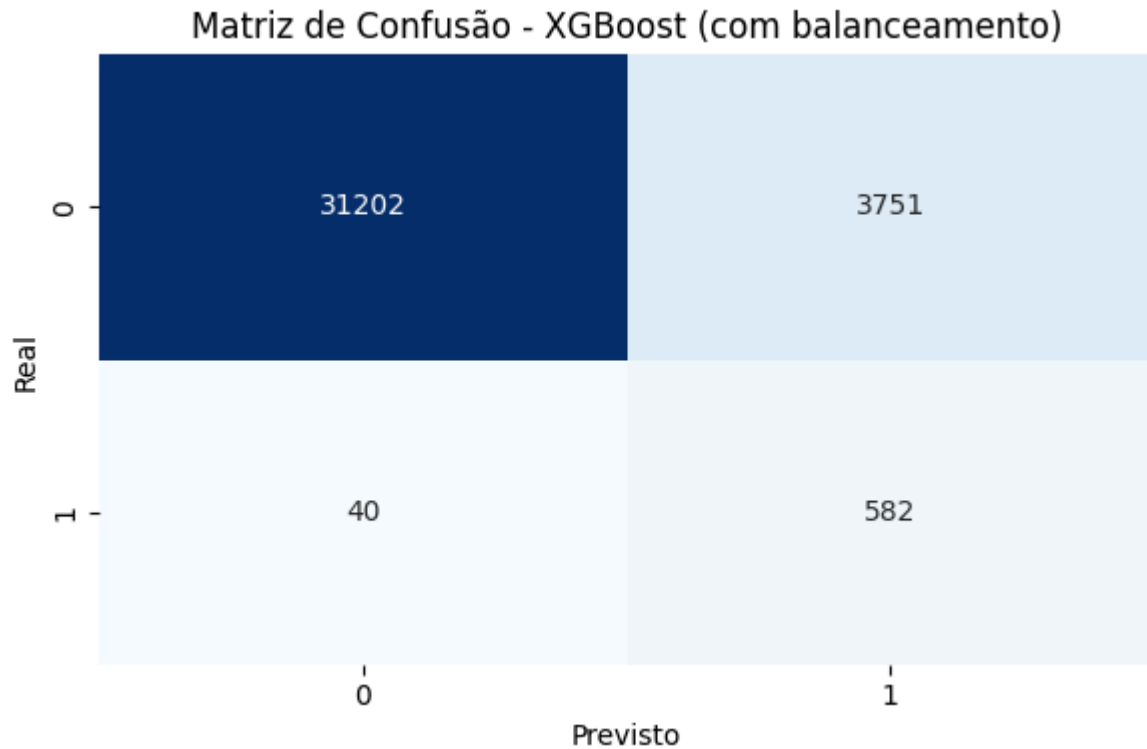




O modelo, avaliado em 35.575 observações, apresentou excelente desempenho geral, com acurácia de 93,3% e AUC-ROC de 0,95, indicando forte capacidade de distinguir entre adimplentes e inadimplentes. Entre os 34.953 adimplentes reais, 32.640 foram corretamente classificados, e 2.313 foram falsos positivos. Dos 622 inadimplentes, o modelo acertou 544 e errou 78. A classe 0 (adimplentes) teve precisão de 1,00, recall de 0,93 e F1-score de 0,96, refletindo excelente desempenho. Já a classe 1 (inadimplentes) apresentou recall alto (0,87), mas baixa precisão (0,19), o que resultou em um F1-score de 0,31. Isso mostra que o modelo identifica bem os inadimplentes, mas com muitos falsos alarmes. Apesar da alta performance global, o baixo poder de acerto nas previsões de inadimplência pode limitar sua aplicação direta em decisões críticas como concessão de crédito.

#### 4.2.2. XGBoost

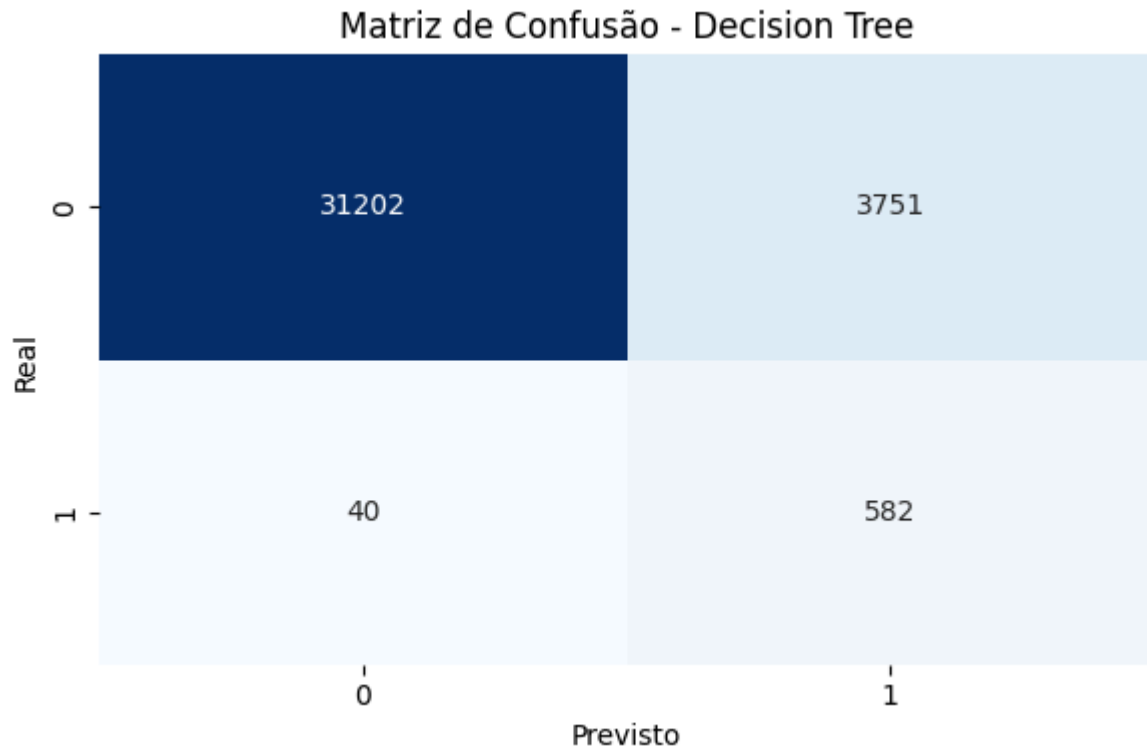
O XGBoost (eXtreme Gradient Boosting) é uma biblioteca voltada a implementar de forma escalável e eficiente o framework de Gradient Boosted Decision Trees (GBDT). Por meio de iterações sucessivas, o algoritmo ajusta novos modelos de árvore para corrigir os erros residuais dos anteriores, utilizando derivadas segunda ordem e regularização para melhorar a generalização.



O modelo XGBoost com balanceamento, avaliado em 35.575 observações, obteve acurácia de 89%. Entre os 34.953 adimplentes, classificou corretamente 31.202, com 3.751 falsos positivos. Dos 622 inadimplentes, identificou corretamente 582, com apenas 40 falsos negativos. A classe 0 (adimplentes) teve ótimo desempenho (precisão 1,00, recall 0,89, F1-score 0,94). Já a classe 1 (inadimplentes) alcançou alto recall (0,94), mas baixa precisão (0,13), com F1-score de 0,23, indicando muitos falsos positivos. Apesar da boa detecção de inadimplentes, o modelo ainda sofre com excesso de alarmes falsos, o que pode limitar sua aplicação prática.

#### **4.2.3. Decision Tree**

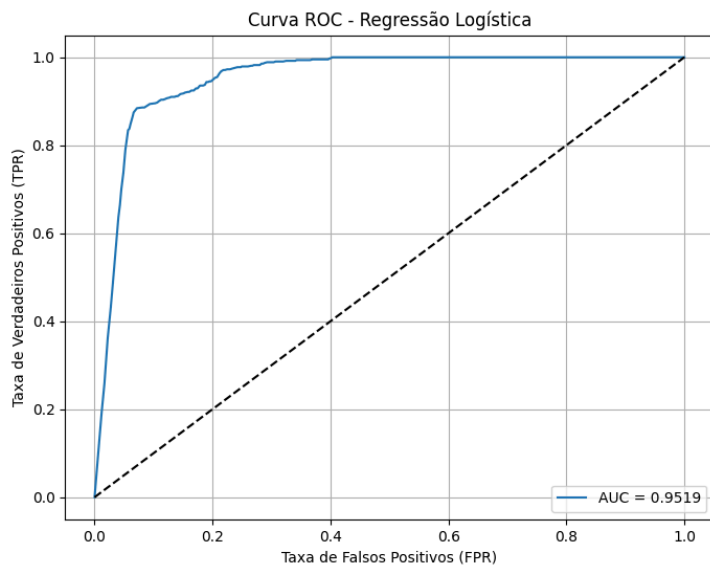
A árvore de decisão é um método não-paramétrico de aprendizado supervisionado que pode ser usado tanto para classificação quanto para regressão. Ela segmenta iterativamente os dados em subconjuntos homogêneos por meio de divisões baseadas em atributos, formando uma estrutura em “árvore” com nós de decisão e nós folha que representam as classes ou valores preditos.



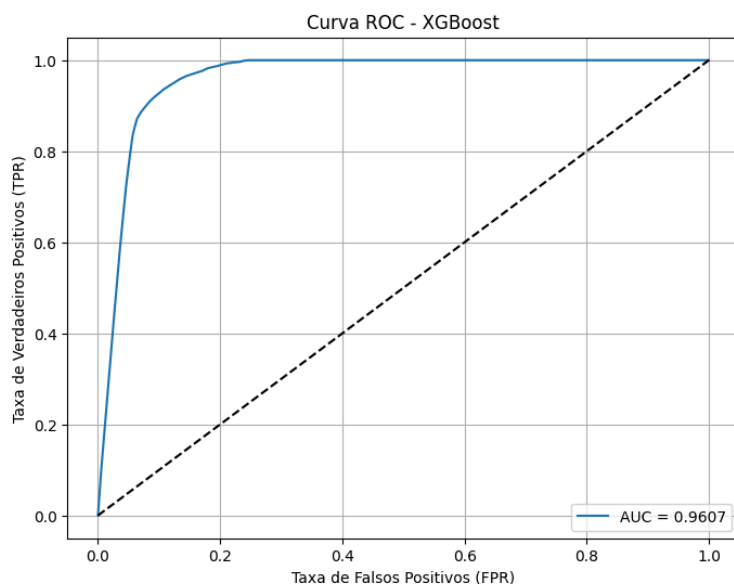
O modelo Decision Tree com balanceamento apresentou acurácia de 89% em uma base com 35.575 observações. Dos 34.953 clientes adimplentes, 31.202 foram corretamente classificados, e 3.751 foram falsos positivos. Entre os 622 inadimplentes, o modelo identificou corretamente 582, errando apenas 40 casos. A classe 0 (adimplentes) manteve desempenho elevado, com precisão de 1,00, recall de 0,89 e F1-score de 0,94. Para a classe 1 (inadimplentes), o recall também foi alto (0,94), mas a precisão ficou em 0,13, resultando em um F1-score de 0,23. Assim como no XGBoost, o modelo é eficaz em capturar inadimplentes, mas sofre com muitos falsos positivos.

#### **4.2.4. AUC ROC**

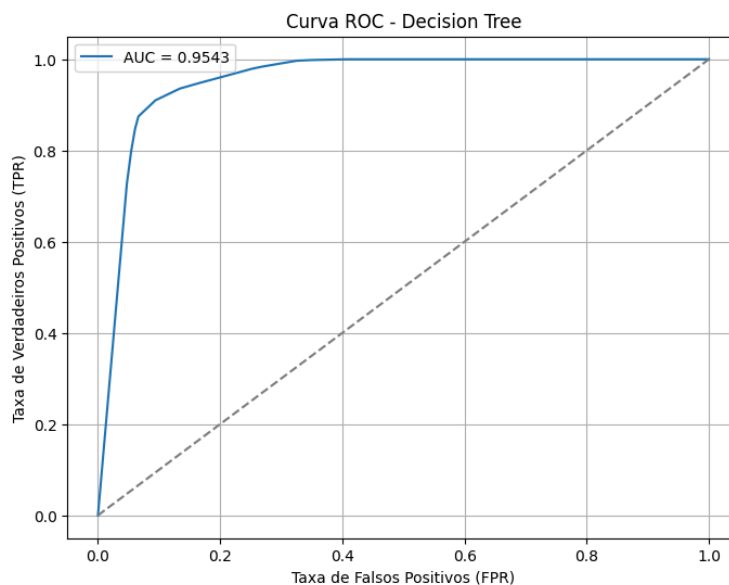
A avaliação por meio da curva ROC (Receiver Operating Characteristic) e da métrica AUC (Area Under the Curve) quantifica a habilidade do modelo em distinguir exemplos positivos de negativos em todos os possíveis limiares de decisão. A AUC corresponde à probabilidade de que, ao comparar uma instância positiva e uma negativa aleatórias, o modelo atribua uma pontuação maior à positiva; valores próximos a 1 indicam forte capacidade discriminativa, enquanto 0,5 aponta desempenho equivalente a um palpite aleatório. Para Tal foi calculado em todos os modelos a curva ROC e o que obteve melhor desempenho foi o XGBoost, como vemos nas imagens a seguir.



A curva ROC da regressão logística atinge um AUC de 0,9519, indicando excelente discriminação entre adimplentes e inadimplentes. Observa-se que, logo nos menores valores de FPR (taxa de falsos positivos), a TPR (taxa de verdadeiros positivos) já cresce rapidamente, o que significa que o modelo consegue identificar grande parte dos inadimplentes sem gerar muitos falsos positivos. A forma suave e ascendente da curva reflete a consistência da regressão logística em equilibrar sensibilidade e especificidade ao longo dos diferentes limiares de decisão.




O classificador XGBoost supera levemente os demais com um AUC de 0,9607, traduzindo-se em ainda maior capacidade de separar as duas classes antes de escolher qualquer ponto de corte. Sua curva ROC permanece bem próxima ao canto superior esquerdo do gráfico, mostrando que o XGBoost mantém TPR elevado mesmo para FPR muito baixos. Essa performance evidencia a eficácia desse algoritmo em capturar padrões não lineares e interações complexas, resultando em maior sensibilidade sem acarretar um aumento proporcional de alarmes falsos.



A árvore de decisão apresenta um AUC de 0,9543, praticamente igual ao da regressão logística, mas sua curva ROC mostra um leve achatamento nas seções intermediárias de FPR. Esse perfil indica que, embora o modelo seja capaz de identificar corretamente muitos inadimplentes logo nos primeiros limiares, ele tende a oferecer menos incrementos de TPR em troca de pequenos aumentos de FPR quando comparado ao XGBoost. Ainda assim, o desempenho global é robusto, evidenciando que a árvore de decisão captura bem as divisões de risco, mesmo sem explorar combinações de variáveis tão complexas quanto as de um ensemble.

Em última análise, reconhecemos que nenhum modelo de risco de crédito é isento de limitações nem de perda esperada — isto é, dos custos inerentes a conceder crédito a quem não pagará (falsos positivos) ou a negar crédito a quem pagaria (falsos negativos). Ainda assim, a regressão logística com o score de risco relativo destacou-se como a alternativa que melhor equilibra discriminabilidade e robustez estatística, apresentando AUC superior a 0,92, acurácia em torno de 84% e recall consistente na captação de inadimplentes.

## Anexo 1 - Links de Interesse

 bruna-derner-colab.3.ipynb

 bruna-derner-03

Dashboard:

<https://lookerstudio.google.com/reporting/acb7b497-ba2c-485b-bbcf-c640a76cacb3>

Video:

<https://www.loom.com/share/8afd71618bf04d9db5f45ffaace0d965?sid=dff82aea-2d68-4f5c-bf58-7f5276fd5e4e>