Ficha técnica - Projeto 03

Análise de crédito através do risco relativo - Super Caja

Ferramentas e tecnologias

- Google BigQuery
- Google Colab
- Looker Studio
- Python
- SQL
- Notion
- Loom

Objetivo

Desenvolver e implementar um modelo automatizado de score de crédito baseado em técnicas avançadas de análise de dados, capaz de classificar os solicitantes de empréstimo do banco "Super Caja" em diferentes categorias de risco, considerando a probabilidade de inadimplência.

Equipe

Vanessa Santana do Amaral

Processamento dos dados

Conectar e importar dados

O primeiro passo foi criar um projeto chamado Risco Relativo dentro do ambiente do Google Cloud. Após isso, foi criado um dataset chamado “riscorelativo” e subido as 4 tabelas para o ambiente: “default”, “loans_detail”, “loans_outstanding” e “user_info”.

Identificar e tratar valores nulos

Nesta etapa, realizou-se a verificação de todas as variáveis presentes nas quatro tabelas, com o objetivo de identificar valores nulos. Para isso, foram utilizadas as instruções SQL **SELECT**, **FROM**, **WHERE**, **IS NULL** e **IS NOT NULL**.

Foram encontrados 7.199 valores nulos na variável **last_month_salary**, o que corresponde a aproximadamente 20% do banco de dados de usuários. Além disso, a variável **number_dependents** da tabela **user_info** apresentou 943 valores nulos.
Devido à grande quantidade de dados faltantes, optou-se por analisar a variável **last_month_salary** em relação ao comportamento dos clientes com base na variável **default_flag**, que indica se o cliente é mau pagador (1) ou bom pagador (0).

Para isso, foi utilizada a função **LEFT JOIN** para unir as tabelas, juntamente com **COALESCE**, **ORDER BY** e **CASE WHEN** para segmentar os dados. O resultado mostrou que, dentre os registros com salário nulo, 35.317 clientes são bons pagadores e 683 são maus pagadores.
Quanto aos valores nulos em **number_dependents**, foi decidido atribuir o valor zero, uma vez que essa variável não terá impacto direto na análise de crédito, servindo apenas para levantamento exploratório de hipóteses.
Para os valores nulos em **last_month_salary**, foi calculada a média (AVG) da variável com os valores existentes e, em seguida, essa média foi atribuída aos registros faltantes, garantindo um tratamento adequado para a análise.

Identificar e tratar valores duplicados e dados inconsistentes

Para identificar valores duplicados, foram utilizados os comandos SQL **COUNT** e **GROUP BY**. Nesta análise, valores duplicados foram encontrados somente na tabela **loans_outstanding**. Para isso, foram agrupados os registros pelos campos **user_id** e **loan_type**, permitindo verificar a quantidade de empréstimos por cliente.
Também foi realizada a padronização dos dados textuais da tabela **loans_outstanding**. Para isso, foram aplicadas as funções SQL **INITCAP** — que padroniza a capitalização das palavras, deixando a primeira letra maiúscula e as demais minúsculas — e **REPLACE**, utilizada para corrigir inconsistências e remover caracteres indesejados, garantindo maior uniformidade nos valores das variáveis categóricas.

Tabelas temporárias e tratamento de outliers

Para identificar outliers, foi utilizado o cálculo do **IQR (Intervalo Interquartil)**, que corresponde à diferença entre o terceiro quartil (Q3) e o primeiro quartil (Q1). Valores muito distantes de Q1 e Q3 são considerados outliers, pois se afastam significativamente da distribuição central dos dados.
No ambiente BigQuery, foram criadas tabelas temporárias por meio da

cláusula **WITH** e calculado o IQR utilizando a função **APPROX_QUANTILES**, que permite encontrar os valores dos quartis e identificar os outliers.
Utilizando uma **CTE (Common Table Expression)**, foram selecionadas as colunas **using_lines_not_secured_personal_assets** e **debt_radio** da tabela **loans_detail**. Para corrigir os outliers nessas variáveis, aplicamos a regra de que valores maiores que **1** foram ajustados para **1**, padronizando assim variáveis que deveriam estar no intervalo entre **0 e 1**.
Esse tratamento é conservador, pois não exclui linhas — apenas limita os valores máximos, considerando que o valor **1 representa 100% do patrimônio comprometido em dívidas**. Com isso, criamos uma nova tabela com dados já limpos, preparada para análises futuras ou para utilização em modelos preditivos.

Unir tabelas

Antes de realizar a união das tabelas, foram criadas versões corrigidas e limpas de cada uma delas.
Durante a correção da tabela **loans_outstanding**, foi identificado que essa tabela continha apenas **35.575 clientes**, enquanto havia **425 clientes ativos** que não estavam sendo considerados. Isso ocorreu porque as consultas anteriores estavam relacionando apenas os clientes que possuíam empréstimos registrados, deixando de fora aqueles que ainda não tinham empréstimos.
Para corrigir essa inconsistência, foi aplicado um **LEFT JOIN** entre a tabela **user_info** (que contém todos os clientes ativos) e a tabela **loans_outstanding**. Dessa forma, todos os clientes aparecem no resultado, incluindo os que ainda não possuem empréstimos, para os quais os valores relacionados a empréstimos foram preenchidos com zero.

Criação de novas variáveis

Na tabela **default**, foi criada a variável **classificacao_inadimplencia**, que classifica os clientes em duas categorias:
- **0** – bom pagador (35.317 clientes)
- **1** – mau pagador (683 clientes)

Na tabela unificada, foram agrupadas as variáveis **user_id** e **loan_type** para identificar a quantidade de empréstimos por cliente, gerando novas variáveis:
- **qtde_real_estate** – quantidade de empréstimos do tipo imobiliário
- **qtde_other** – quantidade de empréstimos de outros tipos
- **qtde_loans** – quantidade total de empréstimos por cliente

Além disso, foi criada a variável **faixa_etaria**, que classifica os clientes em categorias baseadas na idade, facilitando análises segmentadas por faixa etária.

Gerenciamento de dados fora do escopo e análise de correlação entre variáveis

Nesta etapa, o objetivo foi identificar as variáveis com potencial de correlação para serem consideradas na modelagem e análise, além de descartar aquelas que não contribuem para o estudo do risco de inadimplência.
Também foram avaliadas as novas variáveis criadas e seu poder de correlação com as demais.

Variáveis excluídas da análise

Algumas variáveis foram descartadas por não terem impacto direto na análise de risco ou por questões éticas:

- `user_id`: Identificador único dos usuários.
- `loan_id`: Identificador exclusivo dos empréstimos.
- `sex`: Variável sensível, com potencial discriminatório, optou-se por não incluí-la na análise.

Correlação entre variáveis

- `default_flag` **vs** `using_lines_not_secured_corrigida` **(0,2385)**

Correlação fraca porém significativa, indicando maior risco para usuários que utilizam crédito não garantido (cartão, cheque especial), refletindo maior exposição financeira.

- `default_flag` **vs** `more_90_days_overdue` **(0,3075)**

Correlação moderada e positiva que sugere que atrasos longos são fortemente associados ao risco de inadimplência.

- `default_flag` **vs** `age` **(-0,0782)**

Correlação negativa fraca que sugere que clientes mais jovens podem apresentar maior risco de inadimplência, possivelmente por menor estabilidade financeira.

Análise exploratória (EDA)

Na etapa de análise exploratória, o objetivo foi obter uma compreensão geral dos dados, aplicando técnicas para identificar padrões, tendências e relacionamentos entre as variáveis.
Foram criadas visualizações gráficas para analisar a distribuição de renda,

idade, dívidas, entre outras variáveis relevantes.
Também foram calculadas medidas de tendência central, como média, mediana, desvio padrão e percentis, para avaliar a dispersão e o comportamento dos dados.
Além disso, foram elaborados boxplots para as variáveis numéricas, facilitando a identificação de outliers e a visualização da distribuição.
No BigQuery, utilizamos a função NTILE para criar quartis das variáveis, a partir dos quais calculamos as medidas de tendência central e geramos gráficos de dispersão, com o objetivo de compreender melhor a distribuição dos dados.

Cálculo do risco relativo e segmentação por score

O processo de análise de risco foi conduzido em duas etapas principais: identificação dos fatores que influenciam a inadimplência e criação de um score simplificado para segmentação dos clientes.

Análise de variáveis e risco relativo

Primeiro, os dados dos clientes foram segmentados em grupos (quartis) para as variáveis financeiras e comportamentais, como idade, salário, dívidas e atrasos no pagamento. Para cada grupo, foi calculada a taxa de inadimplência, permitindo identificar quais segmentos apresentam maior ou menor risco relativo de inadimplência.
Essa análise forneceu insights importantes sobre quais características estão mais associadas ao risco de não pagamento, servindo de base para o desenvolvimento do modelo de score.

Construção do score de risco e regressão logística

A partir das variáveis mais relevantes, foram criadas variáveis binárias (dummies) que indicam alertas de risco para cada cliente. A soma desses alertas gerou um score simples, no qual os clientes com maior pontuação foram classificados como de maior risco de inadimplência.
Diversas consultas foram realizadas no BigQuery para testar diferentes pontos de corte no score (de 1 a 6), avaliando qual faixa apresentava melhor desempenho na segmentação dos clientes.

Para cada corte, foi criada uma matriz de confusão em SQL, e, posteriormente, no Google Colab, os dados foram importados e analisados em Python. Foram geradas curvas ROC específicas para cada corte e calculadas métricas como acurácia, precisão, recall e F1-score.

Com essas análises, observou-se que o ponto de corte 5 apresentou o melhor equilíbrio entre as métricas.
Para aprofundar a compreensão do comportamento dos dados e validar as hipóteses construídas, foram realizados testes complementares no Google Colab com uso da linguagem Python.

A tabela final consolidada foi utilizada para aplicar técnicas de regressão logística — um modelo estatístico que permite estimar a probabilidade de ocorrência de um evento, como a inadimplência.
Foram geradas matrizes de correlação, mapas de calor, matrizes de confusão, curvas ROC e cálculos das principais métricas de desempenho do modelo.

Resultados e conclusões

Os resultados confirmaram padrões consistentes com a lógica de comportamento de crédito:
• Histórico de atraso superior a 90 dias: clientes com esse perfil apresentaram risco relativo (RR) de 3,96, ou seja, têm quase 4 vezes mais chance de não pagar uma nova dívida.

• Uso excessivo de crédito: também mostrou forte relação com inadimplência. Clientes com esse padrão apresentaram RR = 3,95, indicando mais risco em relação aos demais.

• Idade também demonstrou ser um fator relevante: clientes mais jovens têm risco aumentado, com RR = 1,83. Já os clientes mais velhos mostraram-se significativamente mais confiáveis, com RR = 0,28, o que representa uma probabilidade até 6,5 vezes menor de inadimplência.

Esses resultados foram fundamentais para definir os alertas de risco e compor o score final utilizado na segmentação de clientes.

Com base nesses achados, algumas conclusões práticas para apoiar a tomada de decisão nas políticas de crédito:
• Rastrear o histórico de atrasos a partir de 90 dias como um dos principais alertas de risco;
• Avaliar o comportamento de uso do crédito, e não apenas a quantidade de empréstimos ativos;
• Segmentar os clientes por faixa etária, ajustando os critérios de liberação de

crédito conforme o perfil de risco;
• Valorizar perfis que, mesmo com múltiplos empréstimos, demonstram equilíbrio no uso do crédito e histórico de pagamento estável.

Desempenho do modelo de crédito escolhido

Após a construção do score, foram testados diferentes pontos de corte para classificar os clientes entre alto e baixo risco, e o escohido foi o corte 5, que apresentou o melhor equilíbrio entre os indicadores avaliados. Também foram calculadas as principais métricas de avaliação do modelo:
• Acurácia: 93,42%
• Precisão: 16,2%
• Recall (Sensibilidade): 59,15%
• F1-Score: 0,2543
Esse corte representa uma abordagem conservadora, priorizando alta acurácia e maior precisão, mesmo que com um recall moderado. Ou seja, o modelo foi eficaz em evitar falsos positivos (clientes classificados como inadimplentes indevidamente), o que é desejável em políticas de crédito mais cautelosas.