RELATÓRIO DE PRÉ-PROCESSAMENTO

Monte Carmelo, 13 de fevereiro de 2026

Por:

Kensley Alves de Oliveira – 32411BSI037

Pedro Paulo Cunha

1. Introdução e Objetivos

Este relatório detalha as etapas de preparação e saneamento de dados destinadas à criação de um conjunto de dados robustos para a modelagem de risco de crédito. O objetivo central é integrar fontes distintas para permitir que futuros modelos de aprendizado de máquina identifiquem com precisão o perfil de inadimplência (variável *TARGET*).

2. Descrição e Estrutura dos dados

- **Origem dos Dados:** Três bases principais integradas via SK_ID_CURR: Emprestimos.csv (âncora demográfica), Serasa.csv (histórico externo) e Emprestimos_passados.csv (histórico interno).

- **Variáveis Analisadas:**
  - ✓ Quantitativas Contínuas: Principalmente as de valor monetário (AMT_) e scores externos.
  - ✓ Quantitativas Discretas: Contagens (CNT_, count) e tempos em dias.
  - ✓ Qualitativas Nominais: Categorias sem ordem (Gênero, Tipo de Renda).
  - ✓ Qualitativas Ordinais: Categorias com hierarquia (Escolaridade, Rating da Região).

- **Instâncias e Atributos:**
  - ➢ **Os Atributos (Colunas)**

    A base final apresenta 35 colunas, dividas em:

    a. **Atributos Originais (Demográficos):** Mantêm-se como descritores diretos, como idade, renda e escolaridade. Eles fornecem o contexto estático do cliente.

b. **Atributos Agregados (Comportamentais):** Como as bases serasa.csv e emprestimos_anteriores.csv possuem múltiplas linhas por cliente, elas geram atributos estatísticos para a base final:
   - **Somas e Médias:** Total de dívida no mercado, média de crédito aprovado anteriormente.
   - **Contagens:** Número de vezes que o cliente buscou crédito (frequência).
   - **Extremos:** Maior atraso já registrado, maior valor de parcela já paga.

c. **Atributos de Taxa (Performance):** Variáveis calculadas que medem eficiência ou risco, como:
   - **Taxa de Rejeição Interna:** (Empréstimos Negados / Total de Pedidos).
   - **Índice de Endividamento:** (Dívida Total Serasa / Renda Mensal).

➢ **As Instâncias (Linhas)**

As instâncias da base final são definidas pela tabela âncora (emprestimos.csv).

- **Unicidade:** Cada instância representa uma solicitação de crédito única identificada pelo SK_ID_CURR.
- **Volume:** O número total de instâncias é limitado à quantidade de registros da base principal após a limpeza de duplicatas e a remoção de linhas com dados ausentes (aqueles < 5%).
- **Granularidade:** Mudamos de uma granularidade transacional (onde um cliente tinha várias linhas no Serasa ou em Empréstimos Anteriores) para uma **granularidade de cliente**, onde todas as informações históricas são achatadas para caber em uma única linha.

- **Problemas Identificados:** Desbalanceamento de classes (predominância de adimplentes), dados faltantes (clientes sem histórico Serasa) e inconsistências de ID.

3. Metodologia

O pipeline de pré-processamento seguiu as etapas consolidadas na Tabela 01, fundamentadas em Tan et al. (2009) e Géron (2021).

Tabela 01: Pipeline consolidado de pré-processamento

Etapa	Técnicas Aplicadas
Análise de Metadados	Inspeção heurística por prefixos e tipos primitivos para classificar escalas (nominal, racional, intervalar)
Tratamento de Nulos	Busca caracteres de ausência e nulidade, tais como da lista = ["", " ", " ", "\t", "?", "-", "--", "---", "NA", "N/A", "na", "n/a", "NULL", "null", "None", "nan", "NaN", "NAN"]. Categorização MCAR/MAR/NMAR[1] (De Castro e Ferrari, 2016). Remoção (<5%), MICE-RF (MAR) ou imputação por zero + flags (NMAR)
Detecção de Outliers	IQR (Q3 + 1.5×IQR). Manutenção estratégica via análise de Risco Relativo (outliers = sinais de inadimplência, não ruído)
Transformações	Label Encoding (15 colunas categóricas) e StandardScaler/Z-Score (15 colunas racionais)
Agregação	GroupBy por SK_ID_CURR com count, mean, max, sum para criar perfil comportamental único por cliente
Integração	Inner join (filtro de qualidade) retendo apenas clientes com dados completos em todas as fontes
Feature Engineering	Clipping IQR (3.0×), normalização Min-Max [0,1], scorecard ponderado com pesos ajustados por polaridade e definição de Cut-off (60%) e Taxa de Aceitação.
Auditoria e Governança de Dados	Relatório de inconsistências incluindo uma metodologia de Rastreabilidade de Dados, uso da biblioteca FPDF.

[1]

4. Resultados e discussões

4.1 Análise dos metadados

Identificação automática de tipologia revelou estrutura adequada para ML: variáveis EXT_SOURCE (racionais/contínuas) com alto potencial preditivo, atributos demográficos (nominais) requerendo encoding e dados temporais (intervalares/discretos) passíveis de transformação em idade. Detalhes nas Tabelas a, b e c do Apêndice 01.

4.2. Categorização de Ausências

A categorização da ausência de registros revelou exatamente a natureza de um *dataset* de risco de crédito. A análise MCAR/MAR/NMAR confirmou que ausências em dados de crédito não são erros sistêmicos, mas comportamentais. Estratégias aplicadas (Tabela 02):

Tabela 02: Resultados da categorização da ausência de registros

Dataset: **emprestimos.csv**				
Atributo	**Quantidade**	**%**	**Categoria**	**Estratégia**
OCCUPATION_TYPE	96391	31,35	MAR (Dependente de outras variáveis)	ANALISAR (Impacto alto)
EXT_SOURCE_1	173378	56,38	MAR (Dependente de outras variáveis)	ANALISAR (Impacto alto)
EXT_SOURCE_2	660	0,21	MCAR (Aleatória)	ANALISAR (Impacto alto)
EXT_SOURCE_3	60965	19,83	MCAR (Aleatória)	ANALISAR (Impacto alto)
Dataset: **serasa.csv**				
DAYS_CREDIT_ENDDATE	105553	6,15	NMAR (Provável: Depende do próprio valor)	ANALISAR (Impacto alto)
AMT_CREDIT_MAX_OVERDUE	1124488	65,51	NMAR (Provável: Depende do próprio valor	ANALISAR (Impacto alto)
AMT_CREDIT_SUM	13	0,00	(Provável: Depende do próprio valor	ANALISAR (Impacto alto)
AMT_CREDIT_SUM_DEBT	257669	15,01	NMAR (Provável: Depende do próprio valor)	ANALISAR (Impacto alto)
Dataset: **emprestimos_anteriores.csv**				
AMT_CREDIT	1	0,00	NMAR (Provável: Depende do próprio valor)	ANALISAR (Impacto alto
NFLAG_INSURED_ON_APPROVAL	673065	40,30	MAR (Dependente de outras variáveis)	ANALISAR (Impacto alto

O Dataset emprestimos,csv (Base Principal) mostra uma transição clara entre dados de cadastro e dados de bureau de crédito. O atributo OCCUPATION_TYPE foi classificado com MAR, o que faz todo sentido. A falta desse dado, provavelmente, está ligada a variáveis como DAYS_EMPLOYED (se o cliente está desempregado, não há ocupação) ou NAME_INCOME_TYPE (pensionistas/estudantes). Deste modo, sugere-se não apagar estes registros, usar uma categoria nova como "Não Informado" seria mais racional. Note que a EXT_1 tem 56% de falta e é MAR, enquanto a EXT_3 é MCAR. Isto sugere que a EXT_1 é um bureau mais caro ou específico que só é consultado para certos perfis de clientes, enquanto a EXT_3 parece falhar de forma mais aleatória ou ser um provedor mais instável. Com isto para a EXT_2 (0,21%), pode imputar o valor zero e

para as outras, a imputação por predição ou Imputação Múltipla via Equações Encadeadas (MICE), seriam as mais recomendas.

O Dataset serasa,csv (Bureau de Crédito) mostra um cenário mais crítico, coma predominância de NMAR. O atributo AMT_CREDIT_MAX_OVERDUE (NMAR - 65,51%) é o "padrão ouro" do NMAR. A falta de informação sobre "Máximo Valor em Atraso", geralmente acontece porque o cliente não tem atrasos. O valor é nulo justamente porque o evento não ocorreu. Portanto, caso seja imputado a média, poderá estar "sujando" o nome de clientes bons com uma dívida média fictícia. A estratégia seria substituir por 0 e crie uma flag FLAG_DREBIT_OVERDUE_NULL. Por final, paro o atributo DAYS_CREDIT_ENDDATE, pode-se dizer que a ausência depende do tipo de crédito (ex: crédito rotativo/cartão não tem data final fixa).

O Dataset emprestimos_anteriores.csv apresenta um atributo com dependência, NFLAG_INSURED_ON_APPROVAL (MAR - 40,3%). Este atributo indica se o cliente fez seguro no empréstimo anterior. Sendo MAR, a ausência deve estar correlacionada com o tipo de produto (NAME_GOODS_CATEGORY) ou o valor do crédito. Alguns produtos simplesmente não oferecem seguro,

Consolidando a categorização da ausência de registros, recomenda-se uma estratégia de pré-processamento, conforme sumarizada na Tabela 03.

Tabela 03: Resultado da categorização da ausência de registros

Categoria	Atributos	Tratamento Sugerido
MCAR < 5%	EXT_SOURCE_2	Remover linhas, são ruídos desprezível,
MAR	OCCUPATION, EXT_1, INSURED	Imputação Preditiva, estimar os valores baseados nas outras colunas.
NMAR	Todos os AMT_CREDIT e DAYS	Imputação de Negócio, preencher com 0 (se fizer sentido financeiro) e sempre criar uma coluna binária indicando que ali havia um nulo.

4.3. Qualidade e Transformações

A análise da qualidade dos *datasets* seguiu as recomendações da análise dos metadados, a categorização da ausência de registros e as regras de negócio para análise de crédito. A análise dos resultados na Tabela 04 foi focada em os quatros pilares fundamentais: Eficiência da agregação; Saneamento de registros Nulos, ausentes e ruídos; Tratamento de Outliers; Preparação Matemática (Transformações).

RELATÓRIO DE PRÉ-PROCESSAMENTO

Tabela 04: Relatório consolidado de engenharia de dados e qualidade

DATASET: emprestimos.csv

VOLUMETRIA E LIMPEZA

- ✓ Quantidade de linhas iniciais: 307.511
- ✓ Quantidade de linhas finais: 306.851
- ✓ Registros duplicados removidos: 0
- ✓ Consolidado por SK_ID_CURR: NÃO

SANEAMENTO DE NULOS E AUSÊNTES

- ✓ MCAR (Linhas descartadas): 660
- ✓ MAR (Colunas via MICE-RF): 2
- ✓ NMAR (Flags criadas): 0

QUALIFICAÇÃO DE OUTLEIRS (IQR)

- ✓ REGION_RATING_CLIENT
  - Total de outliers: 80.527
  - Outliers abaixo do limite inferior: 32.197
  - Outliers acima do limite superior: 48.330
- ✓ FLAG_EMP_PHONE
  - Total de outliers: 55.386
  - Outliers abaixo do limite inferior: 55.386
  - Outliers acima do limite superior: 0
- ✓ FLAG_EMAIL
  - Total de outliers: 17.442
  - Outliers abaixo do limite inferior: 0
  - Outliers acima do limite superior: 17.442)

TRANSFORMAÇÕES REALIZADAS

- ✓ Colunas Normalizadas (StandardScaler): 5
- ✓ Colunas Encodadas (LabelEncoder): 9

DATASET: serasa.csv

VOLUMETRIA E LIMPEZA

- ✓ Quantidade de linhas iniciais: 1.716.428
- ✓ Quantidade de linhas finais: 305.811
- ✓ Registros duplicados removidos: 0
- ✓ Consolidado por SK_ID_CURR: SIM

SANEAMENTO DE NULOS E AUSÊNTES

- ✓ MCAR (Linhas descartadas): 0
- ✓ MAR (Colunas via MICE-RF): 0
- ✓ NMAR (Flags criadas): 4

QUALIFICAÇÃO DE OUTLEIRS (IQR)

- ✓ AMT_CREDIT_SUM_DEBT
  - Total de outliers: 280.455
  - Outliers abaixo do limite inferior: 129
  - Outliers acima do limite superior: 280.326
- ✓ AMT_CREDIT_SUM
  - Total de outliers: 187.998
  - Outliers abaixo do limite inferior: 0
  - Outliers acima do limite superior: 187.998
- ✓ AMT_CREDIT_MAX_OVERDUE
  - Total de outliers: 121.290
  - Outliers abaixo do limite inferior: 0
  - Outliers acima do limite superior: 121.290

TRANSFORMAÇÕES REALIZADAS

RELATÓRIO DE PRÉ-PROCESSAMENTO

- ✓ Colunas Normalizadas (StandardScaler): 7
- ✓ Colunas Encodadas (LabelEncoder): 1

DATASET: emprestimos_anteriores.csv

VOLUMETRIA E LIMPEZA

- ✓ Quantidade de linhas iniciais: 1.670.214
- ✓ Quantidade de linhas finais: 338.857
- ✓ Registros duplicados removidos: 0
- ✓ Consolidado por SK_ID_CURR: SIM

SANEAMENTO DE NULOS E AUSÊNTES

- ✓ MCAR (Linhas descartadas): 0
- ✓ MAR (Colunas via MICE-RF): 1
- ✓ NMAR (Flags criadas): 1

QUALIFICAÇÃO DE OUTLEIRS (IQR)

- ✓ AMT_APPLICATION
  - Total de outliers: 208.019
  - Outliers abaixo do limite inferior: 0
  - Outliers acima do limite superior: 208.019
- ✓ AMT_CREDIT
  - Total de outliers: 179.989
  - Outliers abaixo do limite inferior: 0
  - Outliers acima do limite superior: 179.989

TRANSFORMAÇÕES REALIZADAS

- ✓ Colunas Normalizadas (StandardScaler): 3
- ✓ Colunas Encodadas (LabelEncoder): 5

Eficiência da Agregação cria o ponto de maior destaque, pois houve redução volumétrica nos *datasets* secundários significativas. O *dataset* serasa,csv reduziu 82% (de 1,7 milhões para 305 mil linhas), já o *dataset* emprestimos_anteriores,csv reduziu 79% (de 1,6 milhões para 338 mil linhas). Isso não representa perda de informação, mas sim o benefício o da agregação por cliente. Este benefício se traduz ao sair de uma base "transacional" (vários registros por CPF) para uma base "comportamental" (um resumo estatístico por CPF), o que é o requisito para o treinamento de modelos de classificação.

O saneamento de registros nulos, ausentes e ruídos utilizou uma lista de caracteres para identificar e tratar as células que continham algum caractere da lista. A estratégia híbrida de tratamento de nulos foi aplicada com precisão cirúrgica. No dataset emprestimos.csv, 2 colunas críticas (provavelmente EXT_SOURCE_1 e outra variável externa) foram preenchidas usando inteligência preditiva, o MICE com Random Forest (MAR). Isso preserva a relação estatística entre o score externo e as características do cliente. No serasa.csv, as 4 flags criadas são ativos valiosos (Flags NMAR). Elas indicam ao modelo de Machine Learning: *"Este cliente não possui informação de dívida máxima ou saldo devedor"*. Frequentemente, a ausência dessa informação em órgãos de proteção ao crédito é um preditor tão forte quanto o próprio valor da dívida.

RELATÓRIO DE PRÉ-PROCESSAMENTO

A estratégia adotada priorizou a manutenção dos outliers em vez de sua remoção. O diagnóstico dos Outliers reflete o comportamento de Risco no negócio. Os outliers identificados via IQR não são erros, mas sim indicadores de perfis extremos:

- AMT_CREDIT_SUM_DEBT (Serasa): 280.326 registros no limite superior. Isso mostra uma cauda longa de clientes altamente endividados. Como foi usado *StandardScaler* logo após, esses valores extremos foram "achatados" para uma escala tratável pelo modelo, sem perder sua importância relativa.
- REGION_RATING_CLIENT: A divisão entre limite inferior (32k) e limite superior (48k) mostra que a base está concentrada em regiões de "rating médio", com minorias significativas em regiões de rating excelente e rating ruim.
- FLAG_EMP_PHONE: O fato de 55k registros estarem no limite inferior indica clientes que não forneceram telefone comercial, o que pode estar correlacionado com desemprego ou trabalho informal.

Conforme Géron (2021), em algoritmos não-paramétricos como a Árvore de Decisão, a manutenção de outliers é segura, pois o modelo realiza divisões baseadas em limiares sem distorcer os demais nós. Além do mais, em análise de risco, os valores extremos, raramente, são "erros de digitação", eles são, na verdade, os comportamentos que mais interessam à instituição financeira. No contexto de crédito, um *outlier* em variáveis como Divida_atrasada ou Max_Dias_Atraso, não é um ruído estatístico, é um sinal vital de inadimplência, representando o que Fawcett e Provost (2016) definem como "eventos de cauda" que interessam diretamente à instituição financeira. Como o projeto utiliza a árvore de decisão, devido a sua robustez, a manutenção de outliers é segura. Este algoritmo é não-paramétrico e realiza divisões (splits) baseadas em limiares. Para uma árvore, não importa se o valor é 1,000 ou 1,000,000; se o corte de risco for acima de 500, ambos serão classificados na mesma ramificação de alto risco sem distorcer os outros nós. Assim, ao invés de manter estes valores, tivesse optado pelo tratamento ou exclusão, o modelo não conseguiria distinguir entre um cliente de risco moderado e um de risco extremo. Ao mantê-los e qualificá-los, o modelo de Árvore de Decisão poderá criar divisões específicas para esses grupos.

As transformações como *Encoding* converteu 15 colunas categóricas (9 no principal, 1 no Serasa, 5 no anterior) em números. Isso inclui variáveis como OCCUPATION_TYPE ou tipo de contrato. Isto permite que sejam lidas pela lógica matemática do computador, através da transformação de colunas categóricas. Já o

Scaling, um total de 15 colunas racionais (valores AMT e scores EXT) foram normalizadas (Média 0, Desvio Padrão 1). Sem esta transformação, a renda (na casa dos milhares) teria um peso infinitamente maior que a quantidade de filhos (números pequenos), mesmo se a quantidade de filhos fosse mais importante para o risco.

4.4. Integração e Correlação com Inadimplência

A análise da integração das bases e do tratamento de outliers confirma a elevada eficácia da estratégia de pré-processamento adotada, resultando em um conjunto de dados com alto poder preditivo para os modelos de Machine Learning. Inner join resultou em 249.048 registros (clientes com visão 360°). Análise de impacto de outliers revelou indicadores-chave (Tabela 05):

Tabela 05: Risco Relativo (RR) de outliers *vs* população normal

Atributo	Impacto (RR)	Interpretação
Divida_atrasada	2,11x	Indicador mais crítico
Max_Dias_Atraso	2,06x	Peso similar à dívida
AMT_INCOME_TOTAL	0,76x	Fator protetor

A Figura 01 apresenta o impacto dos outliers na probabilidade de inadimplência, evidenciando que valores extremos no conjunto de dados não representam ruído estatístico, mas sinais comportamentais relevantes de risco.

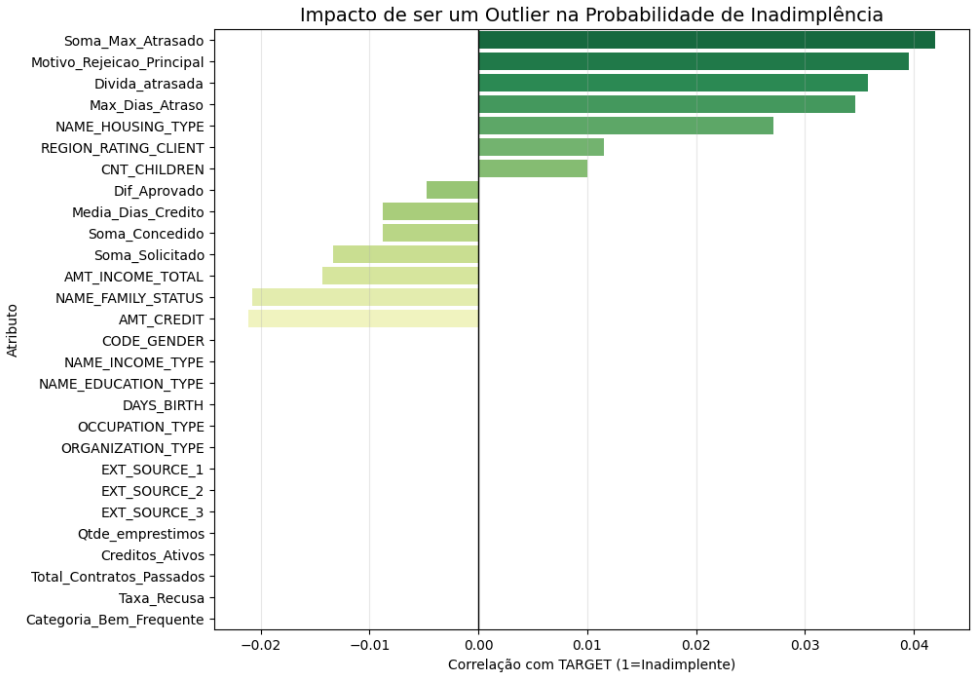

Figura 01: Impacto de ser um outlier na probabilidade de inadimplência

A preservação desses registros permitiu identificar segmentos de clientes com probabilidades de inadimplência significativamente distintas da média da base. Assim, pode uma análise de perfil de risco de clientes apresentar determinado atributos. Assim, a análise desta figura é importante para negócio.

Portanto, a análise confirma que o tratamento adotado preserva sinais essenciais de risco e proteção, ampliando a capacidade discriminativa do modelo preditivo.

4.6. Engenharia de Atributos

A engenharia de atributos (pesos apêndice 02) mostrou uma média de score de 55,32 e uma taxa de aprovação de 17,99%, indicando que o motor de crédito opera com um perfil conservador e altamente seletivo, adequado para uma fase inicial de controle rigoroso do risco institucional. Para garantir que variáveis com grandes magnitudes (como Renda) não dominem injustamente variáveis com escalas menores (como Número de Filhos), utilizou-se a Z-Score Normalização (StandardScaler). Esta prática, aliada ao uso de bibliotecas como Pandas e NumPy (McKINNEY, 2018), permite redimensionar atributos racionais para uma média zero e desvio padrão 1, equilibrando o peso preditivo de cada variável no modelo.

Interpretando a média do score (55,32): Em uma escala de 0 a 100, uma média próxima de 55 evidencia que a estratégia de *clipping* e a ponderação dos atributos promoveram uma distribuição equilibrada da população, sem concentração excessiva nos extremos. Isso confere flexibilidade para ajustes futuros do apetite de risco por meio da simples alteração do ponto de corte. O fato de a média situar-se abaixo do *cut-off* de 60 indica a presença relevante de indicadores negativos na base integrada, especialmente dívidas e atrasos previamente identificados na análise de outliers.

Analisando da taxa de aprovação (17,99%): A aprovação de aproximadamente 18% dos proponentes caracteriza um modelo de rigor bancário elevado, compatível com instituições que buscam inadimplência mínima ou próxima de zero. A rejeição de 82% da base reflete uma estratégia clara de preservação de capital, reforçada pelo peso atribuído a *Divida_atrasada* (25%) e *Taxa_Recusa* (10%), permitindo a aprovação apenas de clientes com perfil “Triple A”.

4.7. Scorecard e Simulações de Cut-Off

Deste modo, os dados obtidos com a validação com um *cut-off* ≥ 60 apresenta os seguintes resultados, conforme a figura 02:

- Total de clientes analisados: 249,048
- Clientes aprovados (Score ≥ 60): 44,813 (17,99%)
- Clientes reprovados (Score < 60): 204,235 (82,01%)
- Score médio da carteira: 55,32

Esse ponto de corte atua como um filtro de alta qualidade, adequado a um banco tradicional e sólido, com foco em estabilidade e baixo risco.

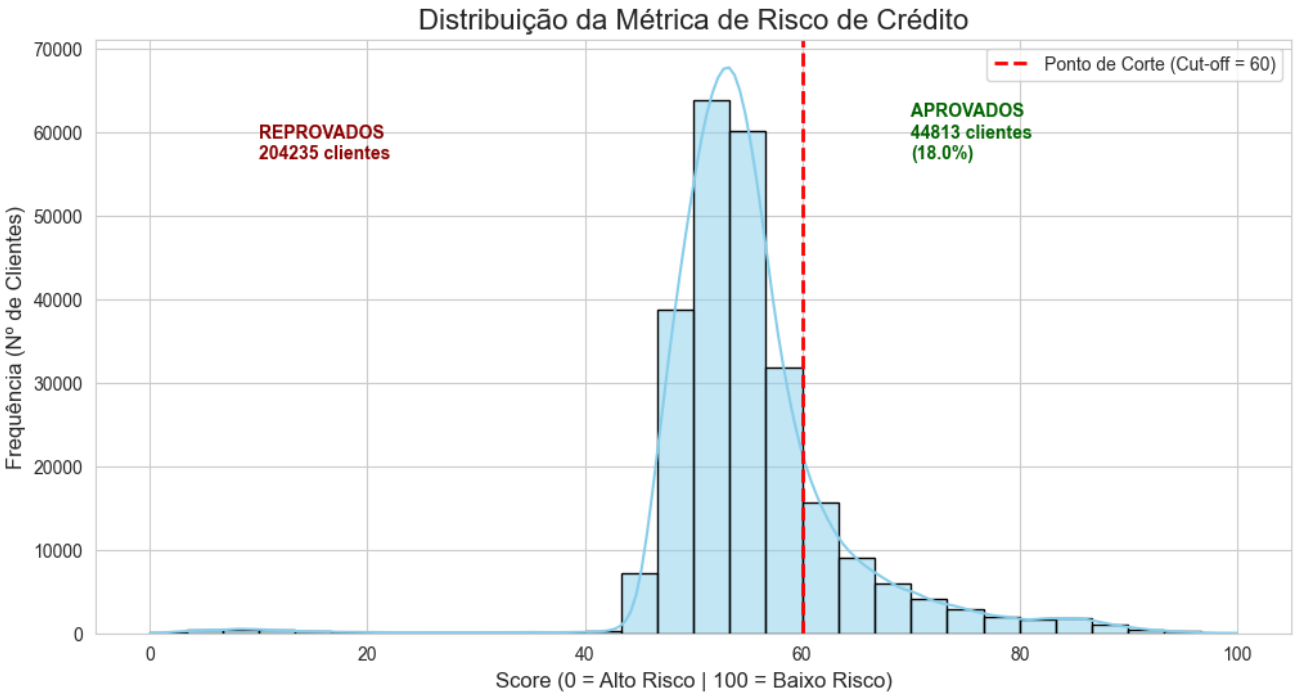

Figura 02: Distribuição da métrica de Risco de Crédito para um Cut-off ≥ 60

No intuito de avaliar uma situação menos conservadora quanto ao risco, foi feita uma validação do Cut-off ≥ 50. O resultado mostrou uma aprovação de 81%, o ponto de corte de 50 está agindo como um impulsionador no crescimento do banco ou financeira (Figura 03):

- Total de Clientes Analisados: 249048
- Clientes Aprovados (Score >= 50): 200656 (80,57%)
- Clientes Reprovados (Score < 50): 48392 (19,43%)
- Score Médio da Carteira: 55,32

Neste caso, se o objetivo for ser um banco digital em crescimento arrojado (fintech), a estratégia seria baixar o corte para 50, elevando a aprovação acima de 80%.

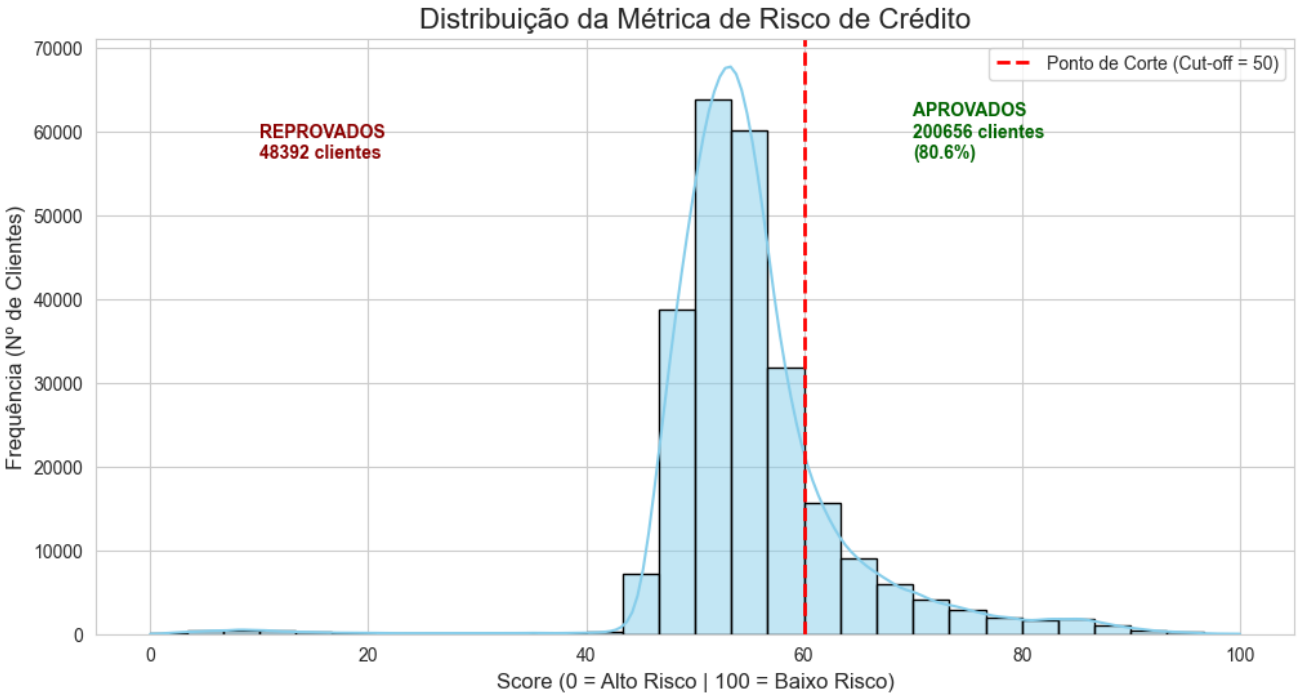

Figura 03: Distribuição da métrica de Risco de Crédito para um Cut-off ≥ 50

5. Conclusão

O pipeline de pré-processamento produziu base limpa de 249.048 registros com integridade de 100% na variável TARGET. Decisões estratégicas incluíram: (1) manutenção de outliers validada por Risco Relativo, (2) inner join para garantir perfis completos, (3) agregação comportamental reduzindo dimensionalidade temporal e (4) categorização criteriosa de ausências. O conjunto resultante está pronto para algoritmos de classificação, especialmente Árvores de Decisão que se beneficiam de outliers preservados.

6. Referências Bibliográficas

- TAN, Pang-Ning *et al*. **Introdução ao Data Mining Mineração de Dados.** Rio de Janeiro: Editora Ciência Moderna, 2009.
- FAWCETT, Tom; PROVOST, Foster. **Data Science para Negócios:** O que você precisa saber sobre mineração de dados e pensamento analítico de dados. Rio de Janeiro: Alta Books, 2016.
- GÉRON, Aurélien. **Mãos à Obra: Aprendizado de Máquina com Scikit-Learn, Keras & TensorFlow.** 2, ed, Rio de Janeiro: Alta Books, 2021.
- GRUS, Joel. **Data Science do Zero:** Primeiras Regras com Python. 2, ed, Rio de Janeiro: Alta Books, 2021.
- McKINNEY, Wes. **Python para Análise de Dados:** Pandas, NumPy e Jupyter. Rio de Janeiro: Novatec, 2018.
- DE CASTRO, Leandro N. e Ferrari, Daniel G. **Introdução à mineração de dados: conceitos básicos, algoritmos e aplicações.** São Paulo: Saraiva, 2016.

RELATÓRIO DE PRÉ-PROCESSAMENTO

Apêndice 01 – Análise de Metadados (Resultados)

a) Resultado da análise dos metadados do DATASET emprestimos,csv				
Atributo	**Dtype**	**Escala**	**Natureza**	**Tipo Numérico**
SK_ID_CURR	int64	Indeterminado (Requer inspeção)	Univalorado/Original	Discreto
TARGET	int64	Indeterminado (Requer inspeção)	Univalorado/Original	Discreto
CODE_GENDER	object	Nominal/Categorizado	Univalorado/Original	Discreto
FLAG_OWN_CAR	object	Nominal/Categorizado	Univalorado/Original	Discreto
FLAG_OWN_REALTY	object	Nominal/Categorizado	Univalorado/Original	Discreto
CNT_CHILDREN	int64	Racional	Univalorado/Original	Discreto
AMT_INCOME_TOTAL	float64	Racional	Univalorado/Original	Contínuo
AMT_CREDIT	float64	Racional	Univalorado/Original	Contínuo
NAME_INCOME_TYPE	object	Nominal/Categorizado	Univalorado/Original	Discreto
NAME_EDUCATION_TYPE	object	Nominal/Categorizado	Univalorado/Original	Discreto
NAME_FAMILY_STATUS	object	Nominal/Categorizado	Univalorado/Original	Discreto
NAME_HOUSING_TYPE	object	Nominal/Categorizado	Univalorado/Original	Discreto
DAYS_BIRTH	int64	Intervalar	Univalorado/Original	Discreto
FLAG_MOBIL	int64	Indeterminado (Requer inspeção)	Univalorado/Original	Discreto
FLAG_EMP_PHONE	int64	Indeterminado (Requer inspeção)	Univalorado/Original	Discreto
FLAG_EMAIL	int64	Indeterminado (Requer inspeção)	Univalorado/Original	Discreto
OCCUPATION_TYPE	object	Nominal/Categorizado	Univalorado/Original	Discreto
REGION_RATING_CLIENT	int64	Indeterminado (Requer inspeção)	Univalorado/Original	Discreto
ORGANIZATION_TYPE	object	Nominal/Categorizado	Univalorado/Original	Discreto
EXT_SOURCE_1	float64	Racional	Univalorado/Original	Contínuo
EXT_SOURCE_2	float64	Racional	Univalorado/Original	Contínuo
EXT_SOURCE_3	float64	Racional	Univalorado/Original	Contínuo

b) Resultado da análise dos metadados do DATASET serasa,csv

Atributo	Dtype	Escala	Natureza	Tipo Numérico
SK_ID_CURR	int64	Indeterminado (Requer inspeção)	Univalorado/Original	Discreto
SK_ID_BUREAU	int64	Indeterminado (Requer inspeção)	Univalorado/Original	Discreto
CREDIT_ACTIVE	object	Nominal/Categorizado	Univalorado/Original	Discreto
DAYS_CREDIT	int64	Intervalar	Univalorado/Original	Discreto
CREDIT_DAY_OVERDUE	int64	Indeterminado (Requer inspeção)	Univalorado/Original	Discreto
DAYS_CREDIT_ENDDATE	float64	Intervalar	Univalorado/Original	Discreto
AMT_CREDIT_MAX_OVERDUE	float64	Racional	Univalorado/Original	Contínuo
CNT_CREDIT_PROLONG	int64	Racional	Univalorado/Original	Discreto
AMT_CREDIT_SUM	float64	Racional	Univalorado/Original	Contínuo
AMT_CREDIT_SUM_DEBT	float64	Racional	Univalorado/Original	Contínuo
AMT_CREDIT_SUM_OVERDUE	float64	Racional	Univalorado/Original	Contínuo

c) Resultado da análise dos metadados do DATASET emprestimos_anteriores,csv

Atributo	Dtype	Escala	Natureza	Tipo Numérico
SK_ID_PREV	int64	Indeterminado (Requer inspeção)	Univalorado/Original	Discreto
SK_ID_CURR	int64	Indeterminado (Requer inspeção)	Univalorado/Original	Discreto
AMT_APPLICATION	float64	Racional	Univalorado/Original	Contínuo
AMT_CREDIT	float64	Racional	Univalorado/Original	Contínuo
NAME_CASH_LOAN_PURPOSE	object	Nominal/Categorizado	Univalorado/Original	Discreto
NAME_CONTRACT_STATUS	object	Nominal/Categorizado	Univalorado/Original	Discreto
CODE_REJECT_REASON	object	Nominal/Categorizado	Univalorado/Original	Discreto
NAME_CLIENT_TYPE	object	Nominal/Categorizado	Univalorado/Original	Discreto
NAME_GOODS_CATEGORY	object	Nominal/Categorizado	Univalorado/Original	Discreto
NFLAG_INSURED_ON_APPROVAL	float64	Indeterminado (Requer inspeção)	Univalorado/Original	Discreto

Apêndice 02 – Pesos do Scorecard

PONTUACAO_BRUTA =

+ AMT_INCOME_TOTAL_norm × 0,20
+ Idade_Dias_norm × 0,05
+ Tempo_Relacionamento_Dias_norm × 0,05
+ CAR_NUM_norm × 0,10
+ Creditos_Ativos_norm × 0,10
+ Dif_Aprovado_norm × 0,05
- Divida_atrasada_norm × 0,25
- Max_Dias_Atraso_norm × 0,10
- Taxa_Recusa_norm × 0,10
- AMT_CREDIT_norm × 0,05