Análise de Dados de Inadimplência

Introdução a mineração de dados

Maria Nilza Ramos
Teliana dos Santos

Universidade Federal do Amazonas
Departamento de Estatística

07 de julho de 2025

Sumário

Introdução:

Banco de dados

Para a análise, foram utilizados os dados do arquivo Credit.csv, disponível na plataforma Kaggle.
O arquivo .csv contém dados onde a adimplência (0) e a inadimplência (1) são indicadas na coluna 'default'. **O objetivo é identificar os fatores que mais influenciam o perfil do público inadimplente**.

- O banco contêm 15 colunas e 7081 linhas

Bibliotecas Utilizadas - AED

pandas: manipulação, filtragem e agregamento de dados tabulares.

numpy: operações numéricas vetorizadas (média, soma, arredondamento, seleção condicional).

matplotlib: gráficos básicos (histogramas, pizza, dispersão).

seaborn: visualizações estatísticas e gráficos de dispersão por classe.

Variáveis

Variável	**Tipo**
default	Qualitativa (Categórica Binária)
idade	Quantitativa
sexo	Qualitativa (Categórica Binária)
dependentes	Quantitativa (Discreta)
escolaridade	Qualitativa (Categórica Ordinal)
estado_civil	Qualitativa (Categórica)
salario_anual	Qualitativa (Categórica Ordinal)
tipo_cartao	Qualitativa (Categórica Nominal)
meses_de_relacionamento	Quantitativa (Discreta)
qtd_produtos	Quantitativa (Discreta)
iteracoes_12m	Quantitativa (Discreta)
meses_inativo_12m	Quantitativa (Discreta)
limite_credito	Quantitativa (Contínua)
valor_transacoes_12m	Quantitativa (Contínua)
qtd_transacoes_12m	Quantitativa (Discreta)

Head dos Dados

As tabelas a seguir mostra as duas primeiras linhas do DataFrame, fornecendo uma visão geral da estrutura dos dados e dos tipos de valores em cada coluna.

Tabela 1:

default	idade	sexo	dependentes	escolaridade	estado civil	salario_anual	tipo_cartao
0	76	M	3	ensino medio	casado	60K-80K	blue
0	49	F	5	mestrado	solteiro	menos que $40K	blue

Tabela 2

meses de relacionamento	qtd produtos	iteracoes_12m	meses inativo_12m	limite credito	tran
39	5	3	1	12.691,51	
44	6	2	1	8.256,96	

Análise exploratória de dados

Variáveis qualitativas

A Tabela a seguir apresenta as estatísticas descritivas para as variáveis identificadas como tipo 'object' no conjunto de dados. Isso inclui o número de valores únicos, o valor mais frequente (moda) e sua frequência.

Tabela: Estatísticas Descritivas para Variáveis Categóricas/Objeto

Variável	**count**	**unique**	**top**	**freq**
sexo	7081	2	M	3706
escolaridade	7081	5	mestrado	2591
estado_civil	7081	3	casado	3564
salario_anual	7081	5	menos que $40K	2792
tipo_cartao	7081	4	blue	6598
limite_credito	7081	6509	1.438,33	10
valor_transacoes_12m	7081	7044	4.141,61	2

Análise exploratória de dados

Variáveis quatitativas

A tabela a seguir apresenta as estatísticas descritivas para as variáveis quantitativas do conjunto de dados, incluindo contagem, média, desvio padrão, valores mínimo e máximo, e os quartis (25%, 50% e 75%).

Tabela: Estatísticas Descritivas de Variáveis Numéricas

Variável	count	mean	std	min	25%	50%	75%	max
default	7081.0	0.157181	0.363997	0.0	0.0	0.0	0.0	1.0
idade	7081.0	46.34769	8.041225	26.0	41.0	46.0	52.0	73.0
dependentes	7081.0	2.337805	1.291649	0.0	1.0	2.0	3.0	5.0
meses_de_relacionamento	7081.0	35.98135	8.002609	13.0	31.0	36.0	40.0	56.0
qtd_produtos	7081.0	3.819376	1.544444	1.0	3.0	4.0	5.0	6.0
iteracoes_12m	7081.0	2.454456	1.104917	0.0	2.0	2.0	3.0	6.0
meses_inativo_12m	7081.0	2.342607	0.995104	0.0	2.0	2.0	3.0	6.0
qtd_transacoes_12m	7081.0	64.50331	23.809330	10.0	44.0	67.0	80.0	134.0

Mineraçao de dados

Uma função lambda foi criada com o objetivo de limpar o formato dos dados, realizando a substituição de

- Pontos por espaço vazio
- Vírgulas por pontos.

```
limpa = lambda valor: float(valor.replace(".",
    "").replace(",", "."))
```

foi aplicado a função nas colunas. 'valor transacoes 12m' e 'limite credito'

```
df['valor_transacoes_12m'] = df['valor_
   transacoes_12m'].apply(limpa)

df['limite_credito'] = df['limite_credito'].
   apply(limpa)
```

Mineraçao de dados

O conjunto de dados não apresenta valores ausentes (*NA*).

```
df.isna().any()
```

```
default                     False
idade                       False
sexo                        False
dependentes                 False
escolaridade                False
estado_civil                False
salario_anual               False
tipo_cartao                 False
meses_de_relacionamento     False
qtd_produtos                False
iteracoes_12m               False
meses_inativo_12m           False
limite_credito              False
valor_transacoes_12m        False
qtd_transacoes_12m          False
dtype: bool
```

Segmentação do banco de dados

Os dados foram separados entre clientes adimplentes e inadimplentes para permitir uma análise comparativa e independente das características de cada grupo.

```
df_adimplente = df[df['default'] == 0]
df_inadimplente = df[df['default'] == 1]
```

Ordinalização das Variáveis

O código faz conversão de duas colunas categóricas ordinais para um formato numérico, mantendo a hierarquia original de seus valores.

```
#Na variavel salario_anual
salario_mapping = {
    'menos que $40K': 0,
    '$40K - $60K': 1,
    '$60K - $80K': 2,
    '$80K - $120K': 3,
    '$120K +': 4
}

df['salario_anual_ord'] = df['salario_anual'].
   map(salario_mapping)
```

Mineraçao de dados

```
#Na varivel escolaridade
escolaridade_mapping = {
    'sem educacao formal': 1,
    'ensino medio': 2,
    'graduacao': 3,
    'mestrado': 4,
    'doutorado': 5
}
df['escolaridade_ord'] = df['escolaridade'].
   map(escolaridade_mapping)
```

Análise Exploratória da Variável Estado Civil

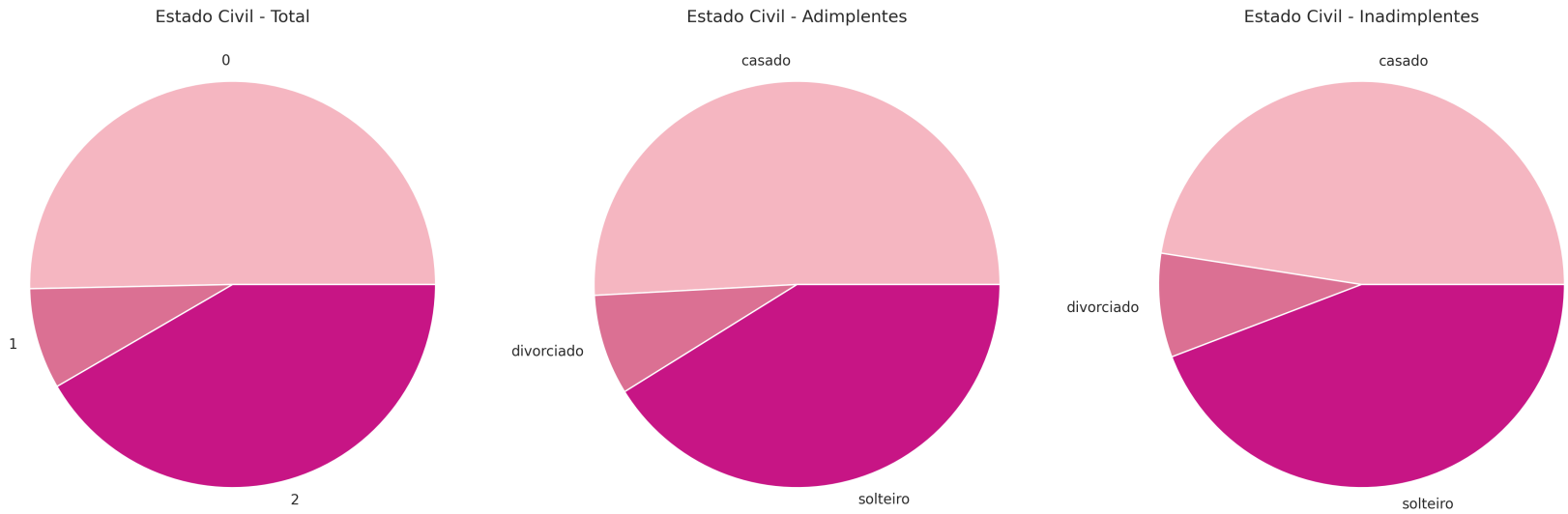


Gráfico Setorial do Estado Civil dos Clientes

Análise Exploratória da Variável Salário Anual

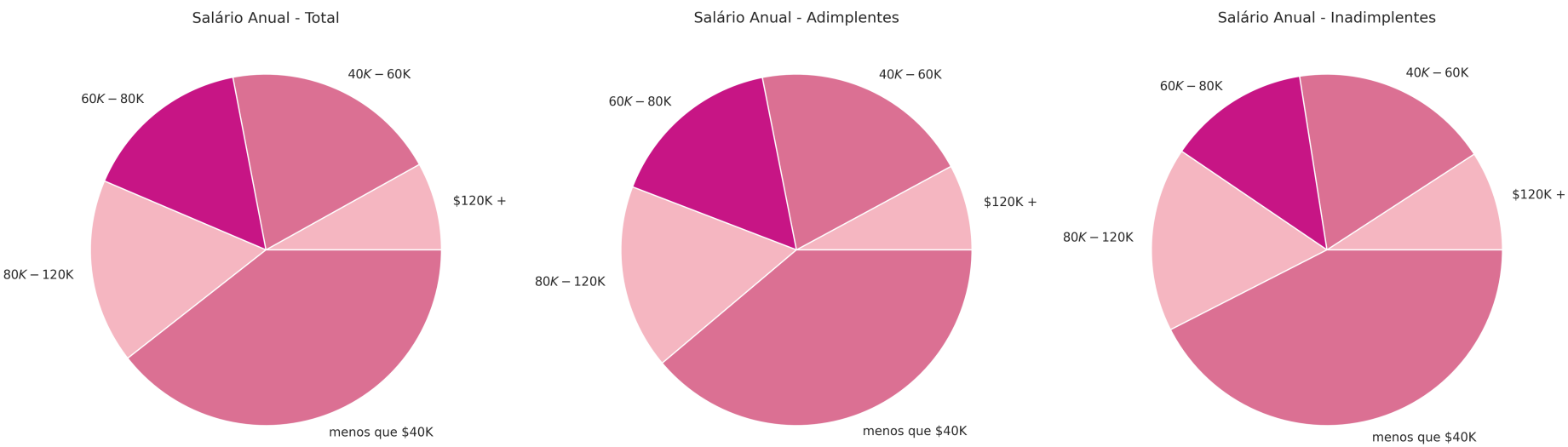


Gráfico Setorial dos Salários Anuais

Análise Exploratória da Variável Sexo

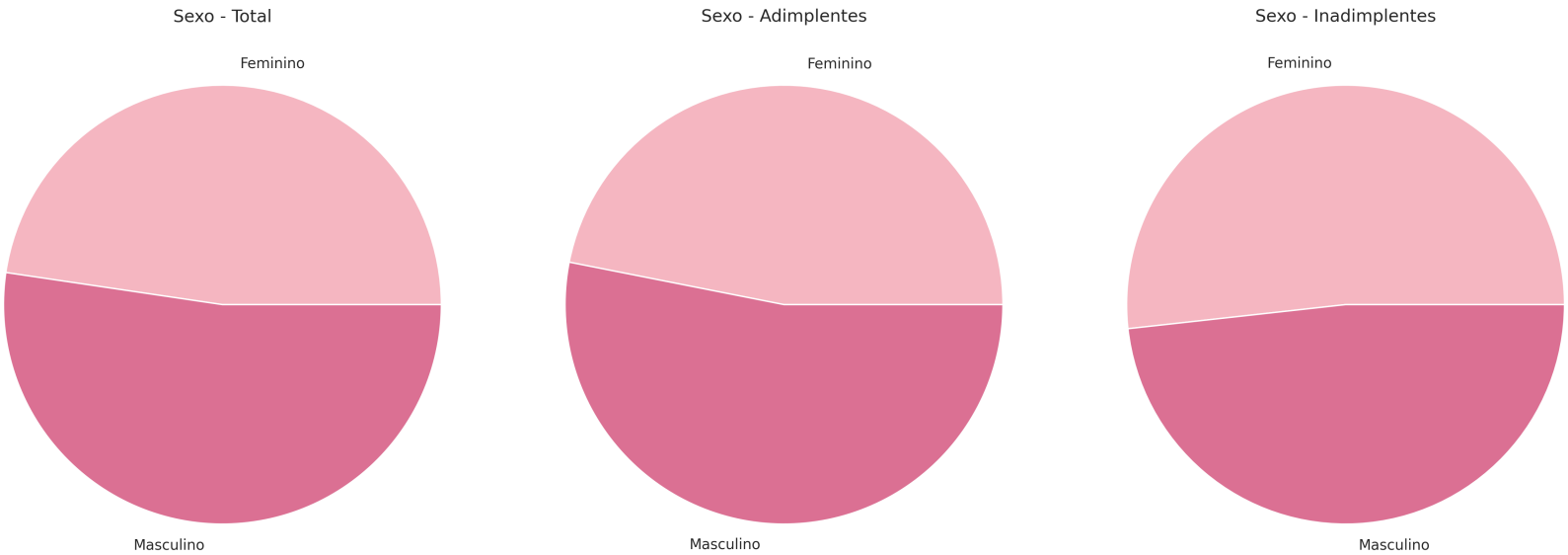
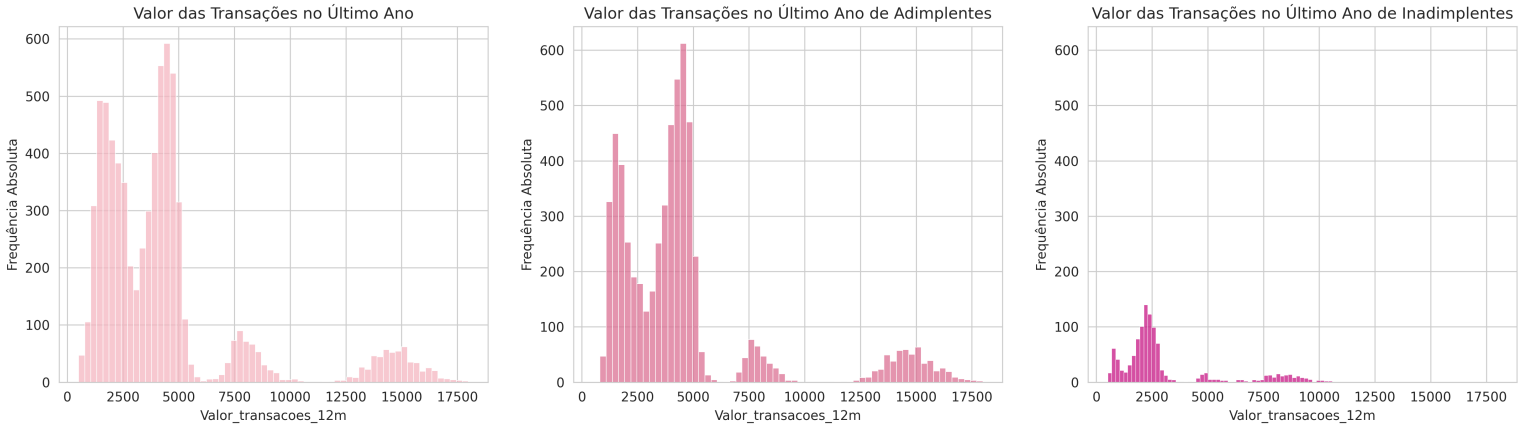


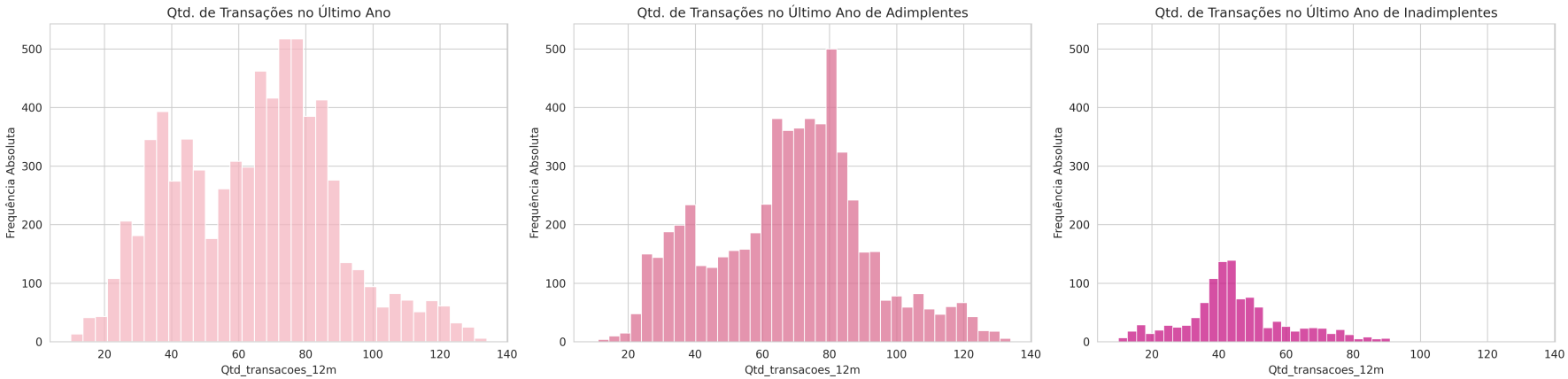
Gráfico Setorial do Sexo

Valor das Transações nos Últimos 12 Meses



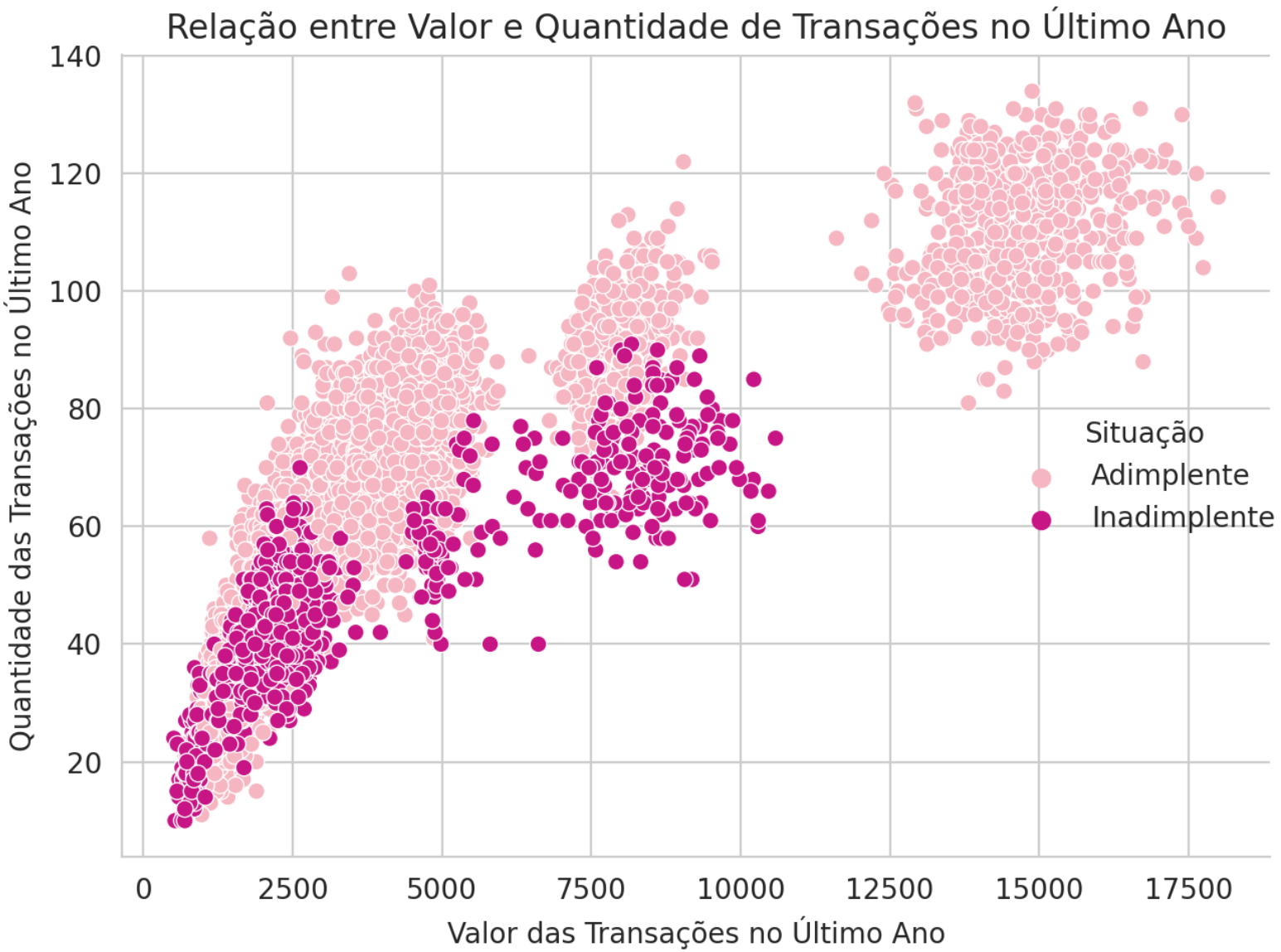
Distribuição dos Valores de Transações nos Últimos 12 Meses

Quantidade de Transações nos Últimos 12 Meses



Distribuição da Quantidade de Transações nos Últimos 12 Meses

Relação entre Valor e Quantidade de Transações



Padrões Observados nas Transações

- O histograma da quantidade de transações anuais mostra picos distintos:
  - Inadimplentes: concentrações entre **20-40** e **60-80** transações.
  - Acima de **90 transações**, praticamente só há clientes adimplentes.
- O histograma do valor das transações revela que valores superiores a **R$11.000** são associados apenas a adimplência.
- O gráfico de dispersão indica uma concentração de inadimplentes em faixas de menor volume e menor frequência de transações.

Esses padrões sugerem que a movimentação financeira anual se relaciona diretamente ao comportamento de pagamento dos clientes.

Bibliotecas Utilizadas na Implementação de Modelos

pandas: manipulação e agregação de dados.

numpy: operações numéricas básicas.

matplotlib & seaborn: visualização dos dados e gráficos.

scikit-learn: treinamento de modelos e cálculo das métricas:

- Acurácia
- Precisão
- Recall
- F1-Score
- Matriz de Confusão

xgboost: implementação do modelo XGBoost.

Visão Geral das Abordagens

Abordagem 1

Apenas variáveis de transações recentes:

- Valor das Transações
- Quantidade de Transações

Abordagem 2

Variáveis contínuas de perfil do cliente

Abordagem 3

Todas as variáveis disponíveis (inclusas as nominais)

Abordagem 1

Considerou somente variáveis relacionadas às transações financeiras mais recentes, com foco em medir a atividade transacional dos clientes.

- **Valor das Transações nos Últimos 12 Meses**
- **Quantidade de Transações nos Últimos 12 Meses**

Objetivo: avaliar se apenas o comportamento recente de gastos é suficiente para predizer inadimplência.

Métricas

Abordagem 1:

Modelo	Acurácia	Precisão	Recall	F1-Score
Logistic Regression	75.51%	35.28%	82.04%	49.34%
Random Forest	89.41%	64.00%	62.14%	63.05%
k-NN	89.27%	68.24%	49.03%	57.06%
XGBoost	86.10%	51.50%	75.24%	61.14%

Abordagem 2

Incluiu variáveis contínuas que descrevem o perfil do cliente e seu relacionamento financeiro com a instituição:

- Idade
- Número de Dependentes
- Meses de Relacionamento
- Quantidade de Produtos Contratados
- Meses de Inatividade nos Últimos 12 Meses
- Limite de Crédito
- Valor das Transações nos Últimos 12 Meses
- Quantidade de Transações nos Últimos 12 Meses

Objetivo: incorporar informações adicionais de perfil financeiro ao modelo.

Métricas

Abordagem 2:

Modelo	Acurácia	Precisão	Recall	F1-Score
Logistic Regression	75.72%	34.80%	76.70%	47.88%
Random Forest	93.93%	85.29%	70.39%	77.13%
k-NN	87.65%	60.40%	43.69%	50.70%
XGBoost	94.14%	77.09%	84.95%	80.83%

Abordagem 3

Considerou todas as variáveis disponíveis, incluindo aspectos demográficos e informações sobre o tipo de cartão utilizado:

- Todas as variáveis da Abordagem 2
- Estado Civil
- Sexo
- Tipo de Cartão

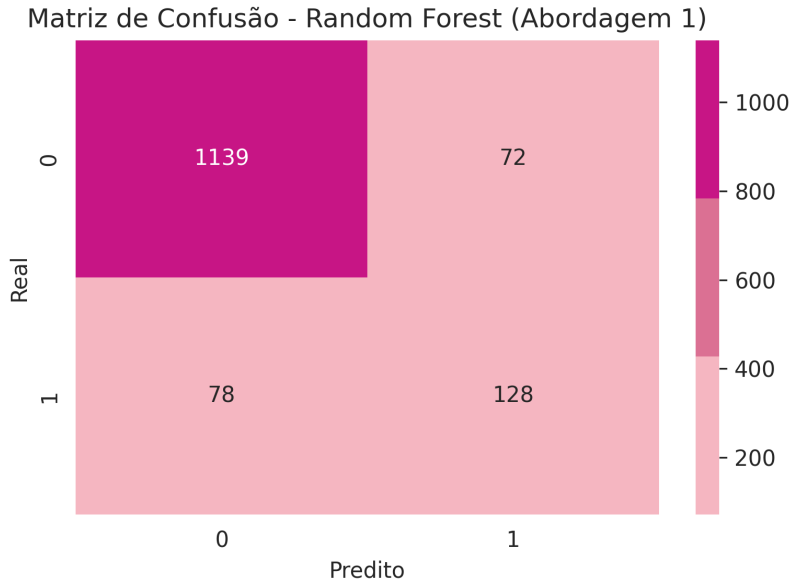
Objetivo: avaliar o ganho preditivo ao incluir dados demográficos e categóricos.

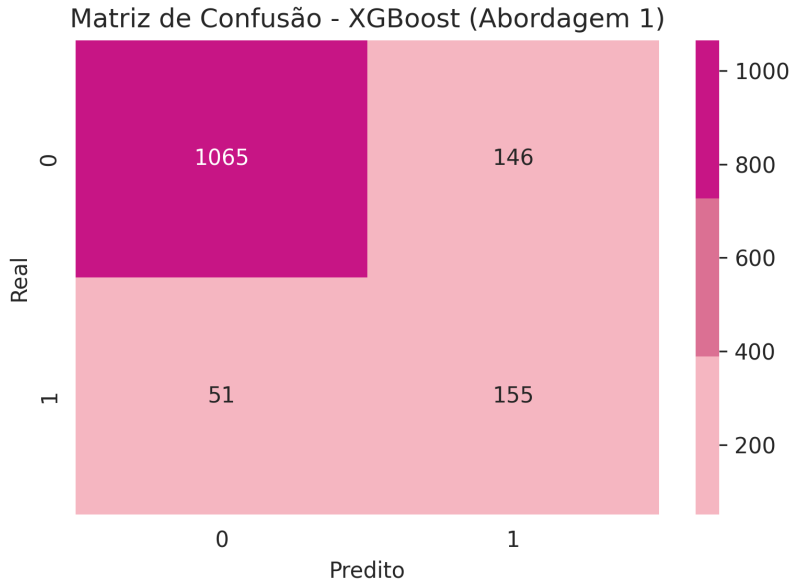
Métricas

Abordagem 3:

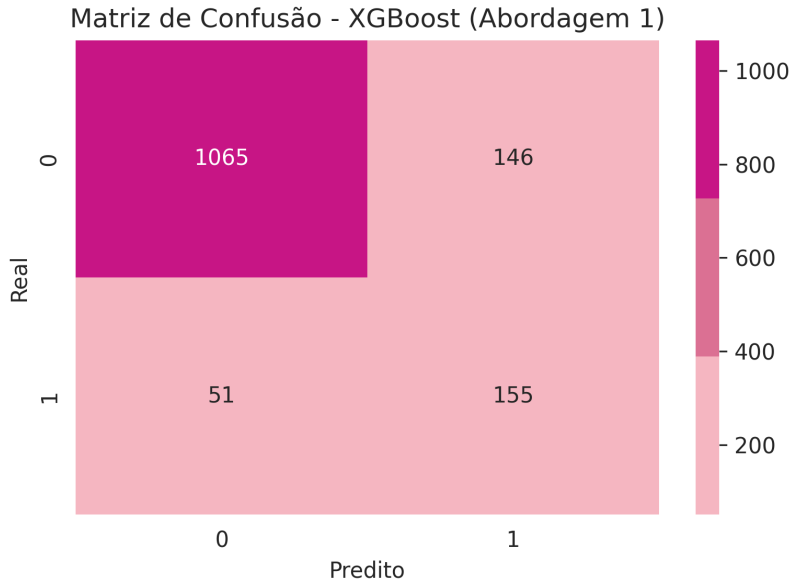
Modelo	Acurácia	Precisão	Recall	F1-Score
Logistic Regression	77.42%	36.74%	76.70%	49.69%
Random Forest	93.72%	87.74%	66.02%	75.35%
k-NN	87.65%	60.40%	43.69%	50.70%
XGBoost	94.21%	77.19%	85.44%	81.11%

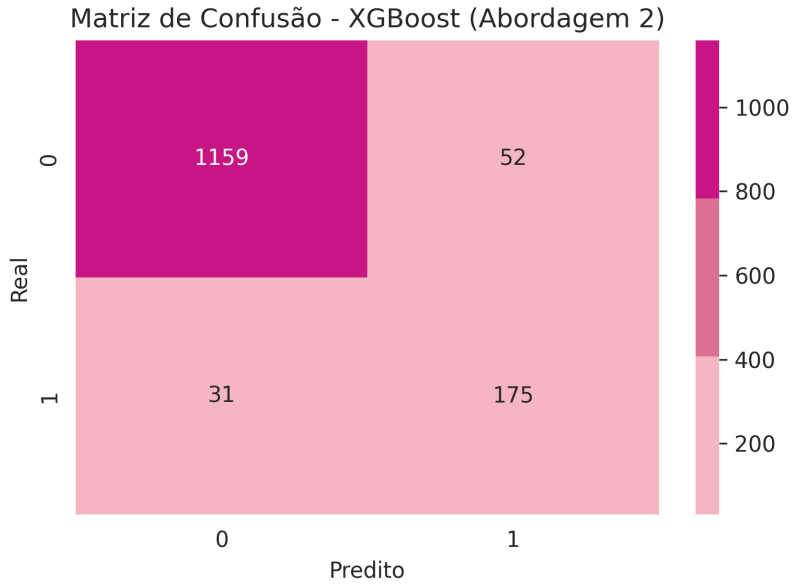
Matrizes de Confusão - Comparação na Abordagem 1



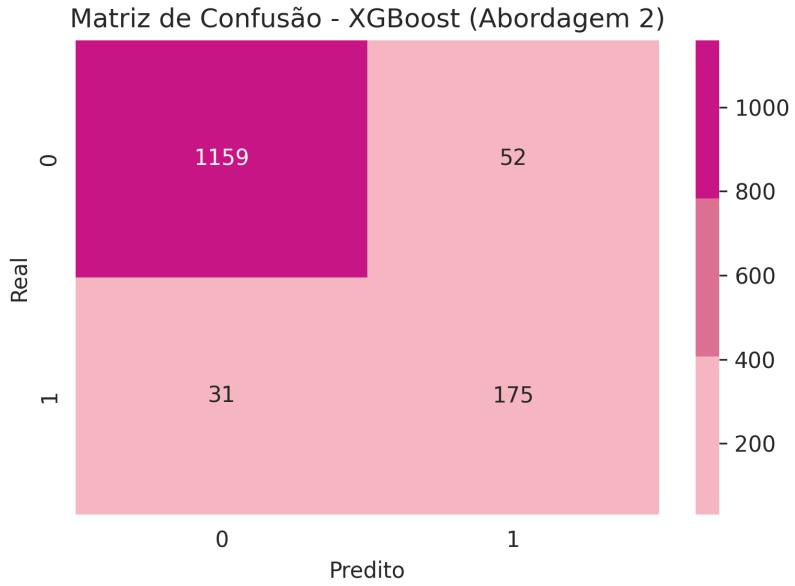


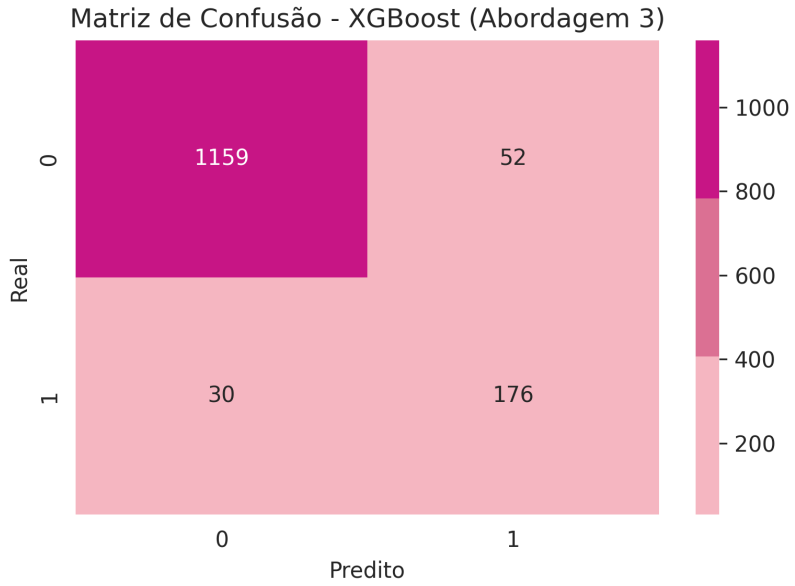
Matrizes de Confusão - Comparação dos Modelos XGBoost





Matrizes de Confusão - Comparação dos Modelos XGBoost





Conclusão

- A primeira abordagem apesar de ter métricas maiores de F1 e acurácia modelada pelo Random Forest, indicou que o modelo XGBoost oferece menos risco de deixar passar inadimplentes quando se olha o recall, e as demais abordagens apresentaram desempenho satisfatório na previsão da inadimplência.
- A Abordagem 3, que incluiu todas as variáveis disponíveis, obteve ligeira melhora no F1-Score em relação à Abordagem 2.
- Entretanto, a análise exploratória revelou que variáveis adicionais como estado civil, sexo e tipo de cartão não apresentavam diferenças expressivas entre os grupos de adimplentes e inadimplentes.
- O modelo com melhor desempenho foi o **XGBoost**, com F1-Score de 81.11%.

Conclusão principal: A inclusão de variáveis pouco discriminativas não trouxe ganho substancial, reforçando a importância da seleção criteriosa de atributos.

Referências

PEREZ, André Marcos. *Dataset de Crédito*. EBAC – Escola Britânica de Artes Criativas e Tecnologia. Disponível em: `https://raw.githubusercontent.com/andre-marcos-perez/ebac-course-utils/develop/dataset/credito.csv`. Acesso em: 05 jul. 2025.

OBRIGADA!