Análise de regressão I

27 de agosto de 2025

Prova Prática III

Nome: Ivan Cunha Vieira
Matrícula: 2023100838

Resumo

O presente trabalho tem como objetivo analisar os padrões de gastos de clientes titulares de cartão de crédito, bem como a relação entre seu comportamento de consumo e a probabilidade de inadimplência. O código desenvolvido realiza desde a preparação e limpeza dos dados até a construção de modelos estatísticos de regressão logística, de forma a identificar variáveis relevantes para a predição do risco de crédito.

A base de dados utilizada contém informações socioeconômicas, financeiras e de histórico de crédito dos clientes, permitindo investigar como fatores como idade, renda, número de dependentes, estabilidade e condição habitacional e relatórios negativos de crédito influenciam tanto os gastos quanto a inadimplência.

1 Introdução

O acesso ao crédito e a avaliação de risco de inadimplência têm sido objeto de diversos estudos na literatura, uma vez que tais análises são fundamentais para a sustentabilidade das instituições financeiras e para o consumo das famílias. Modelos estatísticos e de aprendizado de máquina têm sido amplamente empregados na detecção de padrões de inadimplência, permitindo identificar variáveis que melhor explicam o comportamento de pagamento dos clientes (Hand e Henley 1997; Lessmann et al. 2015).

Entre as abordagens tradicionais, a regressão logística permanece como uma das técnicas mais utilizadas devido à sua interpretabilidade e capacidade de estimar probabilidades de inadimplência (Hosmer, Lemeshow e Sturdivant 2013). Ao mesmo tempo, transformações de variáveis e medidas derivadas, como a razão entre gastos e renda, têm se mostrado úteis para capturar relações não lineares e efeitos de heterogeneidade entre consumidores (Anderson 2007).

No presente estudo, com base em dados de 13444 usuários de cartão de crédito, buscou-se: (i) identificar os padrões de gastos dos titulares; (ii) destacar as variáveis mais relevantes na pontuação de crédito; (iii) analisar se os gastos estão de fato associados à inadimplência; e (iv) avaliar a adequação do modelo empregado para aprovação de crédito, propondo estratégias alternativas quando necessário. O relatório foi desenvolvido em linguagem R, versão 4.3.1, com a utilização do ambiente de desenvolvimento integrado RStudio (R Core Team 2023), com ênfase na aplicação de testes estatísticos não paramétricos e regressão logística penalizada para mitigar problemas de desbalanceamento de classes.

2 Metodologia

A metodologia aplicada pode ser dividida em três etapas principais: (i) identificação e preparação dos dados, (ii) análise exploratória e estatística descritiva, e (iii) modelagem preditiva.

Preparação e identificação dos dados

- **Leitura e filtragem inicial**: A base foi importada a partir de um arquivo `.csv`. Suas colunas e variáveis foram analisadas e descritas conforme Tabela 2.1. Foram considerados apenas indivíduos que possuíam cartão de crédito aprovado (`CARDHLDR = 1`).
- **Correção e consistência**: Foram removidos registros com idades inconsistentes (menores de 18 anos) (*Credit Card Accountability Responsibility and Disclosure Act of 2009* 2009).
- **Transformação de variáveis**:
  - A idade (`AGE`) foi discretizada em faixas etárias, permitindo comparar gastos entre grupos;
  - A renda total (`INCOME`) e a renda per capita (`INCPER`) foram categorizadas em faixas, facilitando a análise de padrões de consumo;
  - Foi criada a variável `TOTAL_DRG`, que consolida o número de relatórios negativos (`MAJORDRG` + `MINORDRG`);
  - A estabilidade habitacional (`ACADMOS`) foi transformada em faixas de tempo de residência, refletindo diferentes níveis de estabilidade (*baixa, média, alta, muito alta*);
  - Variáveis categóricas (`DEFAULT`, `CARDHLDR`, `OWNRENT`, `SELFEMPL`) foram convertidas para fatores, adequando-se às técnicas de modelagem estatística.

Tabela 2.1: Descrição das colunas do banco de dados.

Coluna	**Descrição**	**Tipo**
CARDHLDR	Titular de cartão de crédito (sim ou não)	Binária
DEFAULT	Inadimplência (sim ou não)	Binária
AGE	Idade do titular em anos mais décimos de anos	Contínua
ACADMOS	Meses morando no endereço atual	Inteira
ADEPCNT	1 + número de dependentes	Discreta
MAJORDRG	Relatórios negativos graves de crédito	Discreta
MINORDRG	Relatórios negativos de crédito	Discreta
OWNRENT	Situação de moradia (proprietário ou inquilino)	Binária
INCOME	Renda mensal	Contínua
SELFEMPL	Trabalhador autônomo (sim ou não)	Binária
INCPER	Renda anual por número de dependentes	Contínua
EXP_INC	Razão entre despesas mensais renda anual	Contínua
SPENDING	Gasto mensal	Contínua
LOGSPEND	Log do gasto mensal	Contínua

Análise exploratória

A análise exploratória buscou compreender a distribuição dos gastos e sua relação com fatores demográficos, econômicos e comportamentais. Foram calculadas estatísticas descritivas (média, mediana, desvio-padrão, percentis) contemplando as seguintes dimensões:

- Gastos por faixa etária;
- Gastos por status de inadimplência;
- Gastos por número de dependentes;
- Gastos em função de relatórios negativos;
- Gastos por faixas de renda e renda per capita;
- Gastos entre autônomos e não autônomos;
- Gastos em função da estabilidade habitacional;
- Gastos por tipo de moradia.

Além disso, foi realizada a análise de correlação entre variáveis numéricas, permitindo identificar possíveis relações relevantes.

Modelagem estatística e preditiva

- **Identificação de variáveis relevantes**: Foi ajustado um modelo de regressão logística, tendo a inadimplência (`DEFAULT`) como variável resposta, para identificar os fatores mais relevantes;
- **Testes de associação**: Foram aplicados testes não paramétricos (Mann e Whitney 1947; Brunner e Munzel 2000; Massey Jr. 1951) para verificar se os padrões de gastos diferem significativamente entre inadimplentes e não inadimplentes (Tabela 2.2);
- **Modelo final de classificação**: Foi ajustada uma regressão logística penalizada (`glmnet`), com ponderação de classes para corrigir o desbalanceamento entre inadimplentes e adimplentes. O desempenho do modelo foi avaliado por meio da matriz de confusão e métricas como *precision*, *recall* e *f1-score*.

Tabela 2.2: Hipóteses dos testes estatísticos aplicados.

Teste Estatístico	**Hipóteses**
Mann-Whitney (Wilcoxon)	H_0: As distribuições dos gastos são iguais entre os grupos, H_1: As distribuições dos gastos diferem entre os grupos.
Brunner-Munzel	H_0: As medianas dos gastos são iguais entre os grupos, H_1: As medianas dos gastos diferem entre os grupos.
Kolmogorov-Smirnov	H_0: As distribuições dos gastos são idênticas entre os grupos, H_1: As distribuições dos gastos diferem entre os grupos.

3 Resultados e discussões

A Tabela 3.1 apresenta as estatísticas descritivas da variável `SPENDING` (gastos mensais) para os 10461 titulares de cartão analisados.

Tabela 3.1: Estatísticas descritivas dos gastos mensais.

Estatística	Valor
n	10461
Faltantes (NAs)	0
Média	226,9962
Mediana	139,9484
Desvio padrão	294,1705
Quantil (10%)	19,93025
Quantil (25%)	58,74917
Quantil (75%)	284,7775
Quantil (90%)	521,0816
Mínimo	0,1111111
Máximo	4810,309

A diferença significativa entre média ($226,99) e mediana ($139,95) indica uma distribuição assimétrica à direita, com alguns valores extremamente altos elevando a média. O desvio padrão de $294,17 confirma a alta variabilidade dos gastos. A amplitude entre o percentil 10 ($19,93) e 90 ($521,08) demonstra a grande dispersão dos valores.

Gastos por faixa etária

A Tabela 3.2 apresenta os gastos médios, medianos e a variância por faixa etária.

Tabela 3.2: Gastos por faixa etária.

Faixa etária	Média	Mediana	Variância	n
18-25	195,53	130,70	48950,13	2674
26-35	240,33	153,01	85874,17	4062
36-45	246,46	145,07	114933,58	2439
46-55	214,07	119,89	88529,50	931
56-65	227,08	99,50	191860,85	289
66+	143,97	61,28	44473,80	66

Observa-se que os maiores gastos médios ocorrem na faixa etária de 36-45 anos ($246,46), seguida pela faixa de 26-35 anos ($240,33). Em todas as faixas etárias, a média é superior à mediana, indicando distribuições assimétricas à direita. A faixa de 56-65 anos apresenta a maior diferença relativa entre média e mediana (128%), sugerindo a presença de *outliers* com gastos muito elevados. A faixa de 66+ anos apresenta os menores gastos médios ($143,97) e a menor mediana ($61,28), o que pode estar relacionado à redução da renda na aposentadoria e mudanças nos padrões de consumo.

Gastos por renda

As Tabelas 3.3 e 3.4 apresentam os gastos por faixa de renda e renda per capita, respectivamente.

Tabela 3.3: Gastos por faixa de renda.

Renda mensal	Média	Mediana	Variância	n
Até 2000	155,36	95,64	40091,42	3942
2000-5000	251,71	161,46	89519,87	5995
5000+	483,13	336,66	291240,06	524

Tabela 3.4: Gastos por faixa de renda per capita.

Renda anual per capita	Média	Mediana	Variância	n
Até 10000	186,04	109,13	58619,22	1621
10000-20000	210,98	133,78	68783,10	3820
20000-50000	243,79	153,70	92177,29	4602
50000-100000	345,16	191,10	273198,60	408
100000+	431,93	285,69	123391,45	10

Como esperado, existe uma correlação positiva entre renda e gastos. Clientes com renda entre $5000-$10000 apresentam gastos médios significativamente maiores ($483,13) comparados àqueles com renda até $2000 ($155,36). O mesmo padrão é observado na renda per capita. Em todos os grupos, a média supera a mediana, com a maior diferença relativa (43%) na

faixa de "Até 2000", indicando que mesmo entre a população de menor renda existem *outliers* com gastos elevados. A variância aumenta progressivamente com a renda, sugerindo maior heterogeneidade nos padrões de gastos entre indivíduos de maior poder aquisitivo.

Gastos e inadimplência

A Tabela 3.5 compara os gastos entre clientes inadimplentes e não inadimplentes.

Tabela 3.5: Gastos por status de inadimplência.

Status	Média	Mediana	Variância	n	*Label*
0	231,68	144,72	89211,98	9468	Não Inadimplente
1	182,37	109,27	58886,24	993	Inadimplente

Clientes inadimplentes apresentam gastos médios menores ($182,37) comparados aos não inadimplentes ($231,68). A diferença entre média e mediana é maior para não inadimplentes (60,2%) do que para inadimplentes (66,9%), sugerindo que entre os clientes em dia com seus pagamentos há maior dispersão de gastos, incluindo valores muito elevados. A variância também é maior para não inadimplentes, indicando maior heterogeneidade nos padrões de consumo deste grupo.

Gastos por número de dependentes

A Tabela 3.6 apresenta os gastos por número de dependentes.

Tabela 3.6: Gastos por número de dependentes.

Nº de dependentes	Média	Mediana	Variância	n
0	217,65	138,93	75183,02	5372
1	219,09	128,42	94666,03	2042
2	230,98	138,05	101349,72	1496
3	260,34	158,31	97124,53	1036
4	278,75	158,98	108059,06	393
5	263,96	171,08	90062,16	88
6	330,40	236,74	108792,43	24
7	204,34	127,71	26723,21	7
9	206,37	133,30	21193,30	3

Observa-se uma tendência geral de aumento nos gastos com o crescimento do número de dependentes, atingindo o pico em clientes com 6 dependentes ($330,40). No entanto, o pequeno tamanho amostral para famílias com muitos dependentes requer cautela na interpretação. Para a maioria das categorias, a média supera a mediana, com exceção dos grupos com 5 e 6 dependentes, onde a mediana é maior, sugerindo uma distribuição mais equilibrada ou mesmo assimétrica à esquerda para esses grupos.

Gastos e relatórios negativos

A Tabela 3.7 apresenta os gastos por número de relatórios negativos.

Tabela 3.7: Gastos por número de relatórios negativos.

Nº de relatórios	Média	Mediana	Variância	n
0	215,32	132,71	76849,84	8238
1	250,77	154,84	97594,17	1349
2	268,22	172,32	132929,88	489
3	310,57	172,69	140778,58	207
4	302,02	223,01	82893,95	94
5	475,34	313,11	281688,62	44
6	272,48	180,44	54667,84	22
7	455,77	146,54	320379,75	11
8	1101,85	587,26	2488294,59	5
9	619,68	619,68	442206,81	2

Existe uma relação positiva entre o número de relatórios negativos (TOTAL_DRG) e os gastos médios, com exceção de algumas flutuações provavelmente devido ao pequeno tamanho amostral em categorias com muitos relatórios. Clientes com 8 relatórios negativos apresentam gastos médios extremamente elevados ($1101,85), embora com pequeno número de observações (n=5). A diferença entre média e mediana é particularmente pronunciada para clientes com 7 e 8 relatórios negativos, indicando a presença de valores extremos que distorcem a média.

Gastos e estabilidade habitacional

A Tabela 3.8 apresenta os gastos por estabilidade habitacional.

Tabela 3.8: Gastos por estabilidade habitacional.

Estabilidade	Média	Mediana	Variância	n
Baixa (<1 ano)	242,26	147,13	104922,16	1874
Média (1-3 anos)	223,68	144,21	69720,43	3636
Alta (3-10 anos)	223,69	136,37	82973,13	3306
Muito Alta (>10 anos)	223,59	129,35	109767,85	1645

A estabilidade habitacional não apresenta um padrão claro em relação aos gastos. Clientes com estabilidade "Baixa"(<1 ano) apresentam gastos médios ligeiramente maiores ($242,26) comparados às outras categorias ($223,59-$223,69). A diferença entre média e mediana é consistente em todos os grupos (cerca de 64-73%), indicando distribuições similares em termos de assimetria. A variância é maior para os grupos com estabilidade "Baixa"e "Muito Alta", sugerindo maior diversidade nos padrões de gastos nestes extremos.

Gastos e tipo de emprego

A Tabela 3.9 compara os gastos entre autônomos e não autônomos.

Tabela 3.9: Gastos por tipo de emprego.

Tipo	Média	Mediana	Variância	n	*Label*
0	226,27	140,86	83404,36	9899	Não Autônomo
1	239,77	126,72	141775,98	562	Autônomo

Clientes autônomos apresentam gastos médios ligeiramente superiores ($239,77) aos não autônomos ($226,27). No entanto, a mediana é menor para autônomos ($126,72 vs $140,86), indicando uma distribuição mais assimétrica à direita para este grupo. A variância também é consideravelmente maior para autônomos, sugerindo maior heterogeneidade nos padrões de gastos entre trabalhadores autônomos.

Correlação

A Figura 3.1 apresenta a matriz de correlação entre as variáveis consideradas no estudo. De maneira geral, observa-se que a variável *DEFAULT*, que indica inadimplência, apresenta correlação baixa e negativa com a maioria das demais variáveis. Esse resultado sugere que a inadimplência não é explicada por uma única característica isolada, mas sim por um conjunto de fatores.

A variável *AGE* mostra correlação positiva moderada com o número de dependentes (*ADEPCNT*, $0,26$) e com o tempo de residência no mesmo endereço (*ACADMOS*, $0,40$), o que indica que indivíduos mais velhos tendem a acumular maior estabilidade habitacional. Por outro lado, a relação entre idade e gastos mensais é praticamente nula, em concordância com as análises por faixa etária, que demonstraram maior variabilidade interna do que associação linear direta.

No caso de *ADEPCNT*, observa-se correlação positiva com a renda total (*INCOME*, $0,32$) e negativa com a renda per capita (*INCPER*, $-0,54$). Esse resultado era esperado, já que um maior número de dependentes tende a aumentar a renda necessária para sustentar a família, mas, ao mesmo tempo, reduz os recursos disponíveis por indivíduo. A relação com os gastos é fraca, refletindo a heterogeneidade dos padrões de consumo identificada nas análises por número de dependentes. Já a variável *ACADMOS* apresenta correlações muito baixas com as demais, reforçando que esse fator, embora possa ter importância social, não se mostrou determinante nos padrões de consumo.

Os relatórios negativos de crédito, sejam graves (*MAJORDRG*) ou menores (*MINORDRG*), também não apresentam correlações expressivas com as demais variáveis, embora a análise descritiva tenha mostrado que clientes com mais registros negativos tendem a ter gastos médios mais elevados. Essa aparente contradição sugere que a relação é não linear e possivelmente influenciada por valores extremos, como indicado na tabela de gastos por

relatórios negativos.

Em relação à renda, os resultados são consistentes com as análises anteriores. A variável *INCOME* apresenta correlação positiva com a renda per capita (*INCPER*, $0,38$), com os gastos (*SPENDING*, $0,29$) e com o logaritmo dos gastos (*LOGSPEND*, $0,26$). Esses valores confirmam que indivíduos com maior renda tendem a apresentar maiores níveis de consumo. Ainda assim, a correlação com gastos não é perfeita, evidenciando a existência de fatores adicionais que modulam o comportamento de consumo. A renda per capita, por sua vez, mostra uma associação mais fraca com os gastos, sugerindo que o volume absoluto de renda explica melhor os padrões observados do que a renda ajustada ao tamanho da família.

Entre as variáveis derivadas, destaca-se a razão entre despesas e renda (*EXP_INC*), que apresenta forte correlação com os gastos mensais (*SPENDING*, $0,84$) e com seu logaritmo (*LOGSPEND*, $0,65$). Esse resultado confirma a consistência da medida, uma vez que está diretamente associada ao comportamento de consumo. Finalmente, a forte correlação entre *SPENDING* e *LOGSPEND* ($0,68$) é esperada, pois o logaritmo constitui apenas uma transformação para atenuar a influência de valores extremos, aspecto já observado na distribuição assimétrica dos gastos mensais.

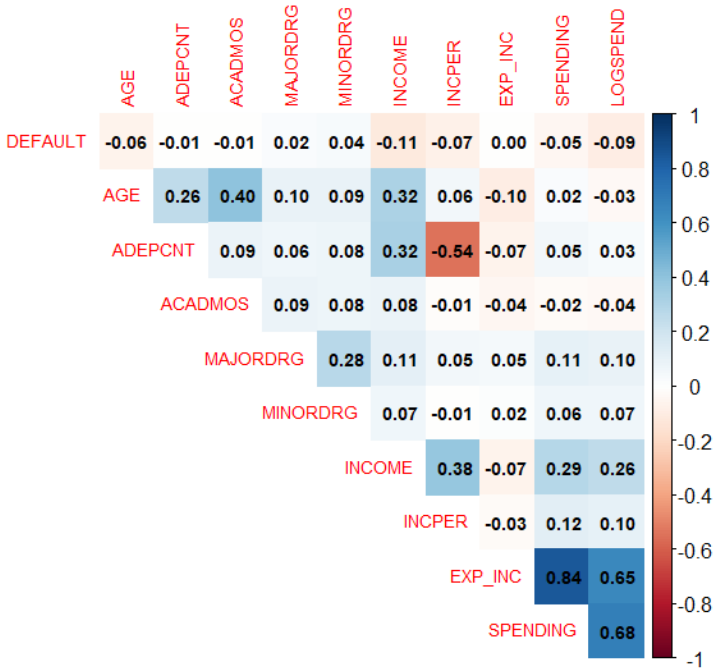

Figura 3.1: Matriz de correlação das variáveis.

De maneira geral, a análise da matriz de correlação mostra que a inadimplência não está fortemente associada a nenhuma variável isolada, o que reforça a importância do uso de modelos multivariados para compreender seu comportamento. A renda se confirma como fator central na explicação dos gastos, mais relevante do que a renda per capita, enquanto idade e número de dependentes exercem influência indireta por meio de suas relações com a renda e a estabilidade familiar. Variáveis derivadas como *EXP_INC* e *LOGSPEND* se

mostraram consistentes, permitindo capturar relações mais claras com os gastos e reduzindo distorções provocadas por distribuições assimétricas e valores extremos.

Variáveis relevantes na pontuação de crédito

$$\begin{aligned}\log\left(\frac{p}{1-p}\right) = \beta_0 &+ \beta_1 \cdot \text{AGE} + \beta_2 \cdot \text{ADEPCNT} + \beta_3 \cdot \text{ACADMOS} + \beta_4 \cdot \text{MAJORDRG} \\ &+ \beta_5 \cdot \text{MINORDRG} + \beta_6 \cdot \text{OWNRENT} + \beta_7 \cdot \text{INCOME} \\ &+ \beta_8 \cdot \text{SELFEMPL} + \beta_9 \cdot \text{INCPER} + \beta_{10} \cdot \text{EXP\_INC} \\ &+ \beta_{11} \cdot \text{SPENDING} + \beta_{12} \cdot \text{LOGSPEND}. \end{aligned} \tag{1}$$

Tabela 3.10: Variáveis mais relevantes para inadimplência.

Variável	**Importância**	**Ranque**
`LOGSPEND`	7.5715	1
`MINORDRG`	4.6299	2
`INCOME`	4.6265	3
`EXP_INC`	3.6129	4
`MAJORDRG`	3.5800	5
`AGE`	3.4157	6
`OWNRENT`	3.3562	7
`ADEPCNT`	2.1476	8
`SPENDING`	1.9970	9
`INCPER`	0.9700	10

A equação logística estimada mostra como diferentes características dos clientes influenciam a probabilidade de inadimplência. A Tabela 3.10 apresenta o ranqueamento das variáveis mais importantes segundo o modelo ajustado, permitindo identificar quais fatores exercem maior impacto na pontuação de crédito.

Os resultados indicam que a variável mais relevante é o `LOGSPEND`, transformação logarítmica dos gastos mensais. Esse achado confirma a importância do consumo no processo de avaliação de risco, ao mesmo tempo em que o uso do logaritmo reduz a influência de valores extremos, permitindo capturar melhor a relação entre gastos e inadimplência. Logo em seguida aparecem os relatórios negativos menores (`MINORDRG`), que representam histórico de crédito com ocorrências de menor gravidade, mas que ainda assim se revelam altamente preditivos para o risco. Em terceiro lugar está a `INCOME`, reforçando a relevância da renda como determinante da capacidade de pagamento.

Também se destacam variáveis derivadas, como a razão entre despesas e renda (`EXP_INC`) e o número de relatórios negativos graves (`MAJORDRG`), ambos associados a maior probabilidade de inadimplência. A idade (`AGE`) e a situação habitacional (`OWNRENT`) aparecem como fatores que refletem efeitos do tempo de vida sobre a estabilidade financeira e possíveis diferenças

entre proprietários e inquilinos em termos de comprometimento com o crédito.

Variáveis como o número de dependentes (`ADEPCNT`) e os gastos mensais (`SPENDING`) também se mostraram relevantes, ainda que em menor grau. O resultado para `SPENDING` sugere que o logaritmo de gastos captura melhor as diferenças entre clientes, sendo mais informativo do que os valores absolutos. Por fim, a renda per capita (`INCPER`) apresentou baixa importância relativa, possivelmente porque seu efeito já está parcialmente refletido pela renda total e pelo número de dependentes.

De maneira geral, a análise revela que variáveis ligadas a consumo, renda e histórico de crédito são os principais determinantes da pontuação de crédito. O modelo confirma que não apenas o volume de renda, mas também o padrão de gastos e a proporção destes em relação à renda, são cruciais para explicar a inadimplência. Além disso, o histórico de relatórios negativos, mesmo quando de menor gravidade, exerce impacto significativo, evidenciando que comportamentos passados de crédito são fortes indicadores do risco futuro.

Relação entre inadimplência e gastos

Com o objetivo de verificar se os gastos diferem de maneira significativa entre clientes inadimplentes e não inadimplentes, foram aplicados testes estatísticos não paramétricos, cujos resultados são apresentados na Tabela 3.11. Em todos os casos, os valores de p encontrados foram muito inferiores a 0,01, indicando evidências estatísticas robustas de que existem diferenças entre os grupos analisados.

O teste de Mann-Whitney (Wilcoxon) revelou uma estatística de aproximadamente 5,38 milhões com $p < 0,001$, rejeitando a hipótese nula de igualdade das distribuições e sugerindo que os padrões de gastos dos inadimplentes diferem significativamente dos não inadimplentes. O valor da estatística obtida é alto porque a base de dados apresenta um volume considerável de observações, ao mesmo tempo que possui classes de inadimplentes e não inadimplentes muito desbalanceadas.

O teste de Brunner-Munzel apresentou valor estatístico negativo (-7,42) e igualmente significativo, confirmando a diferença entre as distribuições de gastos. Por fim, o teste de Kolmogorov-Smirnov indicou uma distância de 0,1214 entre as distribuições acumuladas, também com significância estatística elevada, reforçando que a forma da distribuição de gastos varia conforme o status de inadimplência.

Esses resultados, em conjunto, apontam de forma consistente que gastos e inadimplência estão relacionados, ainda que não de forma linear simples, como sugerido pela análise de correlação. A evidência estatística confirma que inadimplentes tendem a apresentar padrões de gastos distintos, em concordância com os achados da análise descritiva e da modelagem, nos quais variáveis associadas ao consumo desempenharam papel relevante na previsão do risco de crédito.

Tabela 3.11: Resultados dos testes estatísticos para comparação de gastos entre inadimplentes e não inadimplentes.

Teste Estatístico	**p-valor**	
Mann-Whitney (Wilcoxon)	5375978,0000	0,0000
Brunner-Munzel	-7,4169	0,0000
Kolmogorov-Smirnov	0,1214	0,0000

Modelo de regressão logística penalizada

Para estimar a probabilidade de inadimplência, foi ajustado um modelo de regressão logística penalizada do tipo *elastic net*, que combina os termos de penalização Lasso (L_1) e Ridge (L_2). Esse modelo é adequado em situações com grande número de variáveis explicativas ou quando existe correlação entre elas, pois realiza simultaneamente seleção de variáveis e regularização. Além disso, foram utilizados pesos de classe inversamente proporcionais às frequências das classes, a fim de lidar com o desbalanceamento entre clientes adimplentes e inadimplentes. A formulação geral do Modelo (2) é dada por:

$$\min_{\beta_0,\boldsymbol{\beta}} \left[-\frac{1}{N} \sum_{i=1}^{N} \big(y_i \log(p_i) + (1-y_i) \log(1-p_i) \big) \; + \; \lambda \left(\alpha \, \|\boldsymbol{\beta}\|_1 + \frac{1-\alpha}{2} \, \|\boldsymbol{\beta}\|_2^2 \right) \right], \qquad (2)$$

onde $p_i = \dfrac{1}{1+\exp\big(-(\beta_0 + \mathbf{x}_i^\top \boldsymbol{\beta})\big)}$ é a probabilidade estimada de inadimplência para a observação i. Essa especificação explicita o compromisso entre ajuste e parcimônia imposto pela penalização *elastic net*, reduzindo variância do estimador e a influência de multicolinearidade.

Os resultados da avaliação estão sumarizados na matriz de confusão (Tabela 3.12), que revelou um desempenho modesto do classificador, com maior acerto em identificar adimplentes do que inadimplentes. Em termos de métricas derivadas, destacam-se os valores de precisão (*precision*), sensibilidade (*recall*) e a média harmônica entre eles (*f1-score*), apresentados na Tabela 3.13. Observa-se que, apesar do ajuste com pesos, o modelo ainda apresenta baixa precisão para a classe inadimplente, reflexo do forte desbalanceamento de classes. Por outro lado, a sensibilidade foi relativamente mais elevada, indicando que o modelo conseguiu capturar uma proporção razoável dos inadimplentes, embora com muitos falsos positivos.

Tabela 3.12: Matriz de confusão para a classe positiva (inadimplente).

	Classe Real	
Classe Predita	Adimplente (0)	Inadimplente (1)
Adimplente (0)	5492	340
Inadimplente (1)	3976	653

Tabela 3.13: Métricas de desempenho para a classe positiva (inadimplente).

Métrica	**Valor**
Acurácia	0,5874
Precisão (Precision)	0,1411
Sensibilidade (Recall)	0,6576
F1-Score	0,2323

Em síntese, o modelo de regressão logística penalizada com pesos de classes apresentou desempenho limitado, mas conseguiu evidenciar a relação entre variáveis socioeconômicas e comportamentais com a inadimplência. Sua principal contribuição se dá pelo fato de possuir poucos resultados falsos negativos, que são os tipos de resultados mais perigosos para um banco quando há o intuito de aumentar o lucro, pois esse é o tipo de classe mais perigosa. Por esse ponto de vista, é aceitável dizer que o modelo cumpriu parcialmente seu propósito.

4 Informações Computacionais

A fim de explicitar o ambiente de execução computacional utilizado no presente relatório, as Tabelas 4.1 e 4.2 apresentam as especificações técnicas do sistema utilizado para a realização das simulações e análises estatísticas deste estudo e detalham os pacotes estatísticos utilizados no desenvolvimento da pesquisa, respectivamente. Essas informações são fundamentais para garantir a reprodutibilidade e consistência dos resultados, além de permitir a avaliação da capacidade computacional empregada nos experimentos.

Tabela 4.1: Especificações técnicas do sistema utilizado.

Componente	**Especificação**
Arquitetura	64 bits
Sistema Operacional	Windows 11 Pro, versão 23H2
Processador	11th Gen Intel(R) Core(TM) i3-1115G4 @ 3,00GHz 3,00 GHz
Memória RAM	8,00 GB

Tabela 4.2: Pacotes utilizados no estudo e respectivas funções mais utilizadas.

Pacote	**Versão**	**Funções Principais**
`tidyverse`	2.0.0	`filter()`, `mutate()`, `group_by()`, `summarise()`, `%>%` (pipe operator)
`corrplot`	0.95	`corrplot()`, correlation matrix visualization
`caret`	7.0.1	`confusionMatrix()`, model evaluation metrics
`lawstat`	3.6	`brunner.munzel.test()`, non-parametric tests
`pROC`	1.18.5	`roc()`, `auc()`, ROC curve analysis
`ROSE`	0.0.4	oversampling/undersampling techniques
`glmnet`	4.1.10	`cv.glmnet()`, `glmnet()`, regularized regression
`gamlss`	5.4.22	flexible distributional regression models
`ggplot2`	3.5.2	`ggplot()`, `geom_histogram()`, `geom_bar()`, `geom_boxplot()`
`stats (base R)`	base R	`wilcox.test()`, `ks.test()`, `cut()`, `quantile()`, `sd()`, `var()`

5 Conclusões

Em primeiro lugar, quanto aos padrões de gastos, verificou-se que os clientes apresentam distribuição assimétrica, com médias bastante influenciadas por valores extremos. Os maiores gastos médios foram observados entre indivíduos de 26 a 45 anos e em faixas de renda mais elevadas, com forte heterogeneidade em todos os grupos. Também foi identificada uma tendência de aumento de gastos com o número de dependentes, ainda que esse efeito seja limitado pelo tamanho da amostra em categorias específicas.

No que diz respeito às variáveis mais relevantes na pontuação de crédito, a análise por regressão logística destacou o `LOGSPEND`, os relatórios negativos (`MINORDRG` e `MAJORDRG`), a renda total (`INCOME`) e a razão entre despesas e renda (`EXP_INC`) como determinantes centrais. Esses resultados confirmam que tanto o comportamento de consumo quanto o histórico de crédito e a renda são cruciais para explicar a inadimplência.

Quanto à associação entre inadimplência e gastos, os testes estatísticos aplicados (Mann-Whitney, Brunner-Munzel e Kolmogorov-Smirnov) rejeitaram as hipóteses de igualdade entre as distribuições, indicando diferenças significativas entre inadimplentes e não inadimplentes. Embora a correlação linear simples seja fraca, a evidência mostra que os padrões de consumo estão relacionados à inadimplência, especialmente quando analisados em conjunto com outras variáveis.

Por fim, sobre a adequação do modelo de aprovação de crédito, verificou-se que a regressão logística penalizada com pesos de classe apresentou boa sensibilidade para identificar inadimplentes, mas baixa precisão. Isso significa que o modelo consegue reduzir o risco de aprovar inadimplentes, mas ainda gera um número elevado de falsos positivos. Assim, embora útil, o modelo não é suficiente para maximizar os lucros da empresa.

De maneira geral, esse estudo demonstrou que há uma relação clara entre renda, gastos, histórico de crédito e inadimplência, e que modelos estatísticos podem capturar tais padrões de maneira eficaz, ainda que demandem aprimoramentos metodológicos para aplicação prática.

Bibliografia

Anderson, Raymond (2007). *The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation*. Oxford University Press.

Brunner, Edgar e Ulrich Munzel (2000). *Nonparametric Analysis of Factorial Designs*. John Wiley & Sons.

Credit Card Accountability Responsibility and Disclosure Act of 2009 (2009). URL: `https://www.congress.gov/bill/111th-congress/house-bill/627`.

Hand, David J. e William E. Henley (1997). *Statistical Methods in Credit Scoring*. Oxford University Press.

Hosmer, David W., Stanley Lemeshow e Rodney X. Sturdivant (2013). *Applied Logistic Regression*. 3rd. Wiley.

Lessmann, Stefan et al. (2015). "Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research". Em: *European Journal of Operational Research* 247.1, pp. 124–136.

Mann, Henry B. e Donald R. Whitney (1947). "On a test of whether one of two random variables is stochastically larger than the other". Em: *Annals of Mathematical Statistics* 18.1, pp. 50–60.

Massey Jr., Frank J. (1951). "The Kolmogorov-Smirnov Test for Goodness of Fit". Em: *Journal of the American Statistical Association* 46.253, pp. 68–78.

R Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: `https://www.R-project.org/`.