Universidade de Brasília
Departamento de Estatística

Interpretação de redes neurais

Davi Guerra Alves

Projeto apresentado para o Departamento de Estatística da Universidade de Brasília como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Brasília
2023

Davi Guerra Alves

Interpretação de redes neurais

Orientador(a): Thais Carvalho Valadares Rodrigues

Projeto apresentado para o Departamento de Estatística da Universidade de Brasília como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Brasília
2022

Sumário

1 Resultados 4

1.1 Análise descritiva 4

1.1.1 Condição do empréstimo 4

1.1.2 Relação entre as covariáveis e a variável resposta 5

1.2 Regressão logística 10

1.3 Rede neural 12

1.4 Interpretação da rede neural 14

1.5 Benchmark entre regressão logística e redes neurais 19

1.5.1 Complexidade da arquitetura 19

1.5.2 Resultado dos modelos 20

1.5.3 Tempo de execução 21

2 Conclusão 22

3 Anexo 23

1 Resultados

1.1 Análise descritiva

1.1.1 Condição do empréstimo

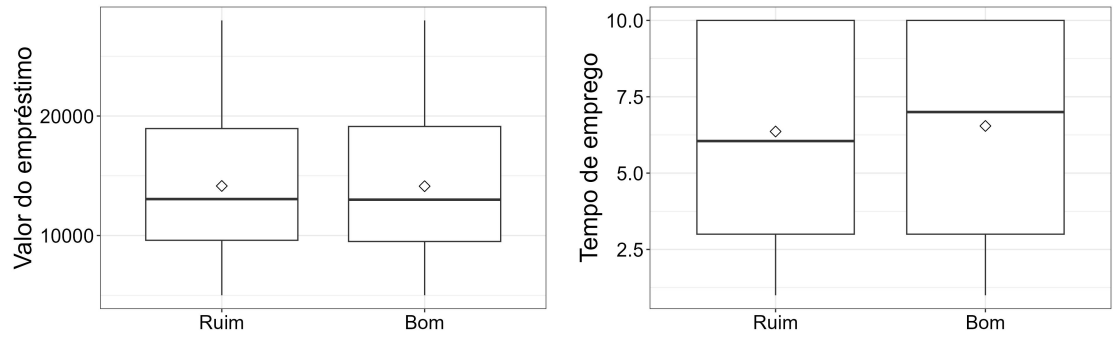
A variável "Condição do empréstimo"é a variável reposta desse estudo, como foi definido anteriormente. Com isso temos o seguinte comportamento dessa variável:

Condição do empréstimo	**Número de observações**	**Frequência relativa**
Empréstimo bom	819950	92,4%
Empréstimo ruim	67429	7,59%

Tabela 1: Número de observação em cada categoria da variável resposta

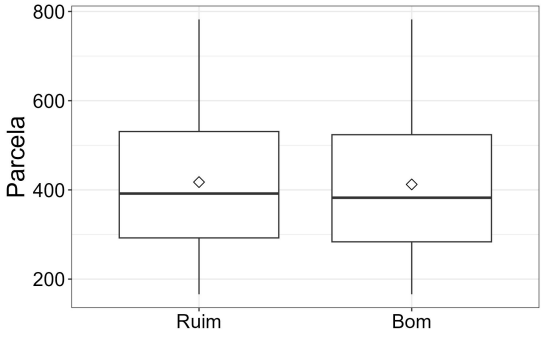
A Tabela 1 mostra a distribuição da variável "Condição do empréstimo". Uma variável composta majoritariamente por observações do tipo "Empréstimo bom", onde a mesma está presente em mais de 90% das observações na base de dados, mostrando que a cada 12 empréstimos rotulados como "bons", existe 1 rotulado como "ruim".

1.1.2 Relação entre as covariáveis e a variável resposta

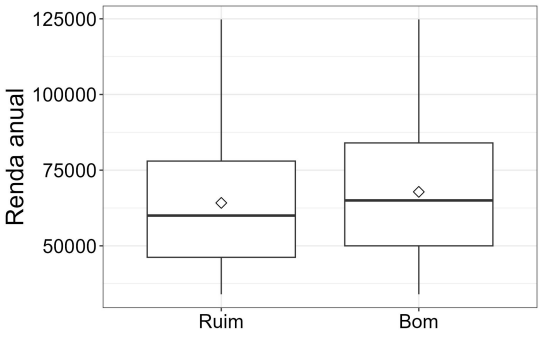


(a) Valor do empréstimo

(b) Tempo de emprego (em anos)



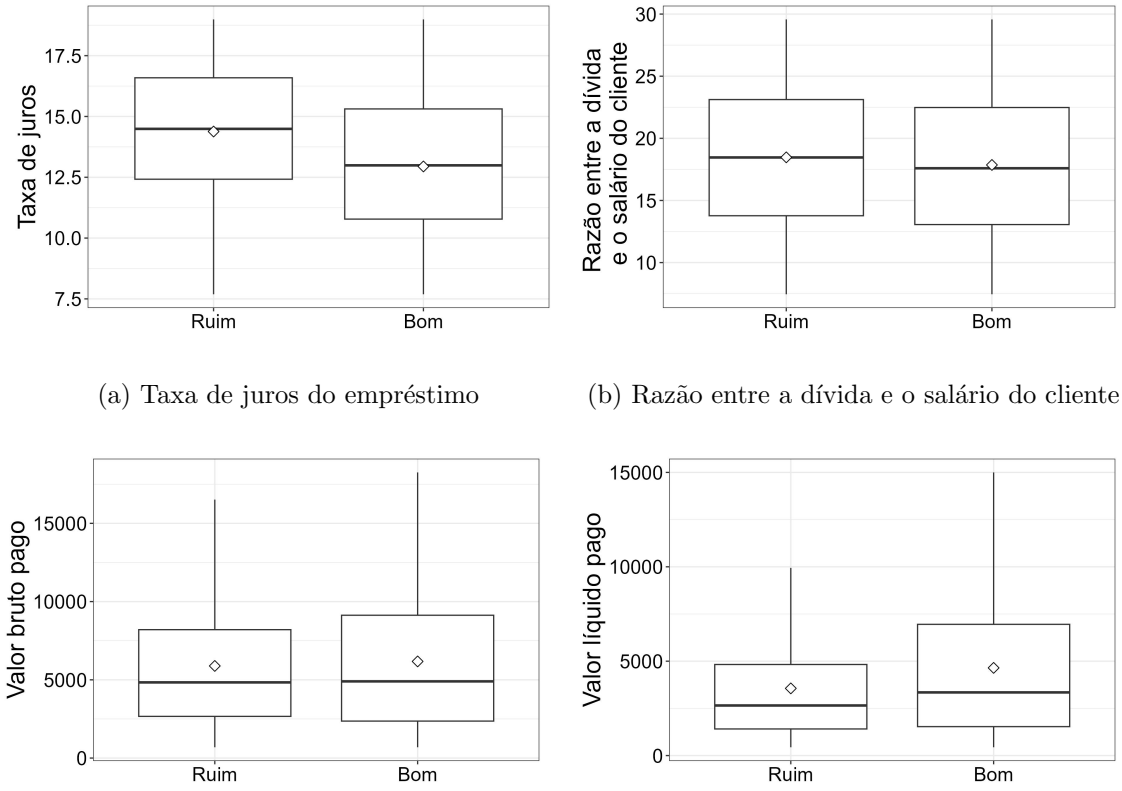
(c) Valor da parcela do emprésitmo



(d) Renda anual (em dólar)

Figura 1: Variáveis explicativas em relação à condição do empréstimo

O comportamento da variável resposta nas Figuras **??** e **??** demonstrou semelhanças, onde, em ambos os casos, não foi evidenciada uma clara diferença entre o valor do empréstimo e o valor da parcela em relação às categorias da variável resposta. A Figura **??** também apresenta um comportamento semelhante entre as classes "Empréstimo ruim"e "Empréstimo bom", mas com um detalhe: a mediana do tempo de trabalho dos clientes rotulados como "Empréstimo ruim"foi inferior em comparação ao outro caso. Por fim, a Figura **??** indica que clientes com uma renda anual elevada tendem a ser categorizados como "Empréstimo bom".



(a) Taxa de juros do empréstimo

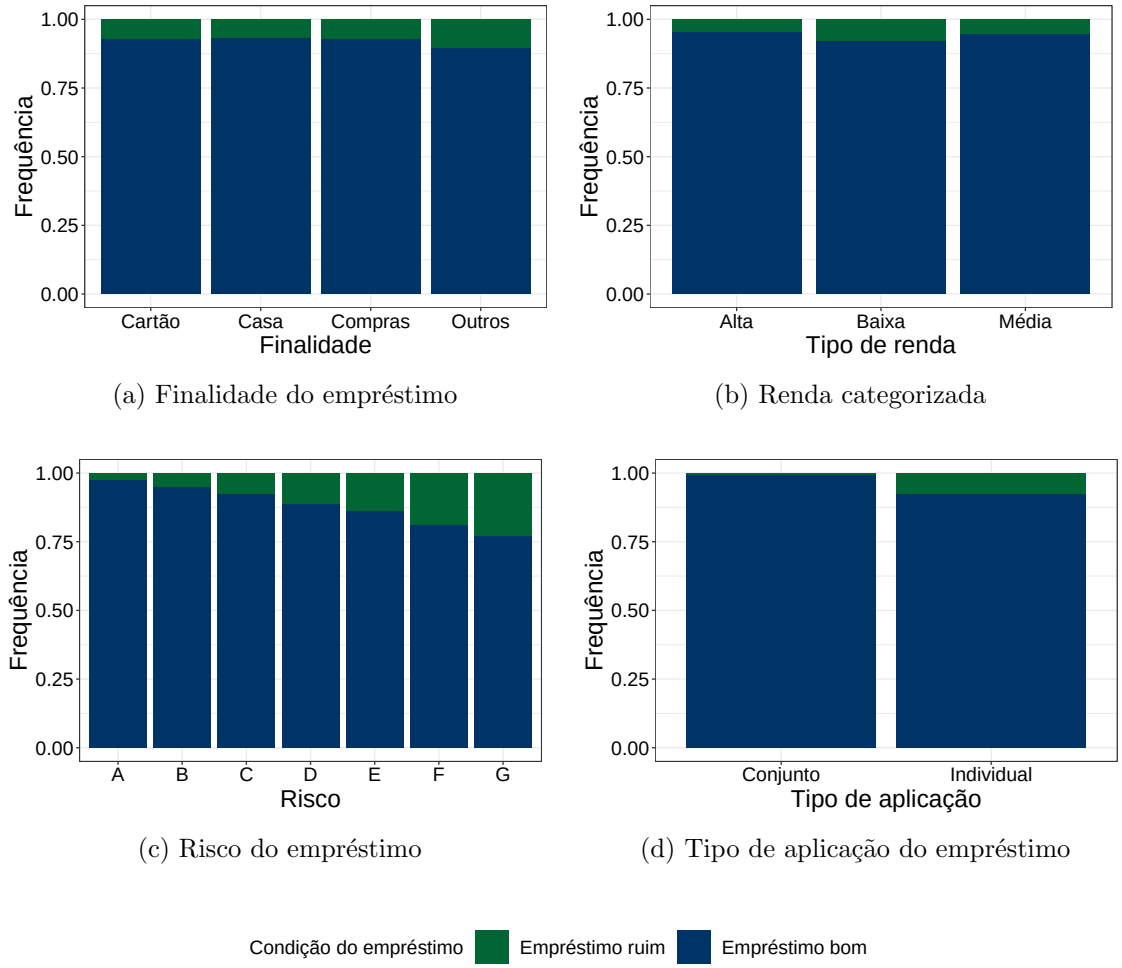
(b) Razão entre a dívida e o salário do cliente

(c) Valor bruto do emprétimo pago

(d) Valor líquido do emprétimo pago

Figura 2: Variáveis explicativas em relação à condição do empréstimo

A Figura **??** evidencia uma relação significativa entre taxas de juros elevadas e empréstimos considerados ruins. A Figura **??** complementa a informação fornecida pela Figura **??**, indicando que clientes com renda mais elevada tendem a cumprir adequadamente com seus pagamentos. As Figuras **??** e **??** seguem padrões semelhantes, sugerindo que clientes que quitaram a maior parte do empréstimo são frequentemente rotulados como bons pagadores.



(a) Finalidade do empréstimo

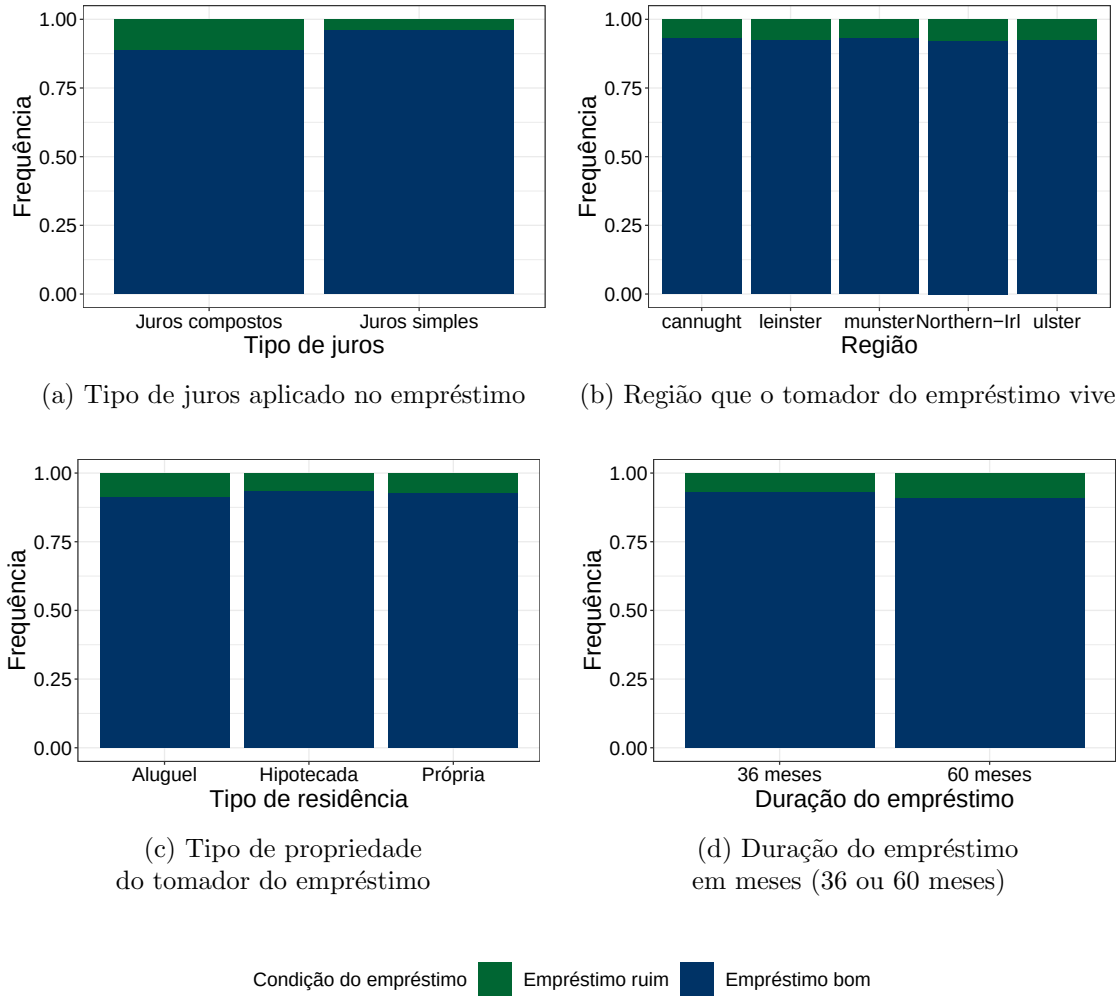
(b) Renda categorizada

(c) Risco do empréstimo

(d) Tipo de aplicação do empréstimo

Figura 3: Variáveis explicativas em relação à condição do empréstimo

A Figura **??** ilustra que as categorias da variável "Finalidade"seguem a proporção natural da condição do empréstimo, conforme indicado na Tabela de Condição do Empréstimo. Na Figura **??**, as categorias "Alta"e "Média"exibem proporções menores de empréstimos ruins em comparação com a categoria "Baixa", que apresenta uma proporção de quase 10% de empréstimos ruins. A Figura **??** revela um padrão de "cascata", indicando que à medida que o risco do empréstimo aumenta, a proporção de empréstimos ruins nas últimas categorias também aumenta, sendo a categoria G a mais afetada, com quase 25% de empréstimos classificados como ruins. Na Figura **??**, a categoria "Empréstimo conjunto"não registrou observações de empréstimos ruins, concentrando a maioria desses empréstimos na categoria "Empréstimo individual".



(a) Tipo de juros aplicado no empréstimo

(b) Região que o tomador do empréstimo vive

(c) Tipo de propriedade do tomador do empréstimo

(d) Duração do empréstimo em meses (36 ou 60 meses)

Figura 4: Variáveis explicativas em relação à condição do empréstimo

Na Figura 4, o gráfico **??** evidencia que empréstimos obtidos sob juros compostos possuem uma proporção mais elevada de rotulações ruins em comparação com empréstimos sob juros simples. As Figuras **??** e **??** destacam uma proporção natural refletida pela distribuição das categorias da variável resposta, conforme apresentado na Tabela 1. Já a Figura **??** revela uma proporção mais significativa de empréstimos ruins quando estes tendem a demorar mais para serem pagos.

Covariáveis	**Coeficiente de correlação**
Tempo de trabalho	-0.02
Renda anual	-0.03
Valor do empréstimo	0.00
Taxa de juros	0.18
DTI	0.01
Valor bruto pago	-0.04
Valor liquido pago	-0.10
Parcela	-0.01
Duração do empréstimo	0.01

Tabela 2: Valores do coeficiente de Pearson entre as covariáveis e a variável resposta

A partir da análise da Tabela 2, nota-se que as correlações entre as variáveis explicativas e a variável resposta são de baixa magnitude. Os coeficientes calculados indicam uma relação linear fraca ou inexistente entre essas variáveis. Esses resultados sugerem que outros fatores ou relações não lineares podem estar desempenhando um papel mais significativo na explicação da variabilidade na variável resposta.

Covariáveis	**Coeficiente de contingência**
Tipo de residência	0.04
Tipo de aplicação	0.01
Finalidade	0.03
Tipo de juros	0.14
Risco	0.15
Região	0.01
Prazo	0.04
Renda	0.04

Tabela 3: Valores do coeficiente de contingência entre as covariáveis e a variável resposta

Ao analisar a Tabela 3, nota-se que a maioria dos coeficientes de contingência entre as covariáveis e a variável resposta são próximos de zero. Destaca-se que a variável "Risco"exibe o maior valor de associação, atingindo 0.15. Entretanto, é importante ressaltar que esse valor ainda é relativamente baixo. Os coeficientes sugerem, em geral, uma falta de associação significativa entre as covariáveis mencionadas e a variável resposta.

1.2 Regressão logística

Falar do modelo utilizado, a normalização dos dados, os resultados métricas de avaliação e interpretação dos coeficientes

Covariáveis	**Coeficientes**	**Erro padrão**
Valor líquido pago	-4.733	0.034
Valor bruto pago	3.321	0.028
Tipo de aplicação	-1.848	0.106
Taxa de juros	1.412	1.412
Valor do empréstimo	-1.406	0.039
Risco	-1.049	0.014
Tipo de juros	-0.462	-0.462
Prazo	-0.203	0.028
Renda anual	-0.195	0.010
DTI	-0.151	0.011
Renda categorizada	-0.111	0.015
Tempo de trabalho	-0.061	0.009
Região	0.038	0.003
Duração do empréstimo	0.033	0.009
Tipo de residência	0.019	0.003
Finalidade	0.019	0.002
Parcela	0.005	0.000

Tabela 4: Estimativa dos coeficientes do modelo logístico e o erro padrão associado

Ao analisar os resultados apresentados na Tabela 4, fica evidente que as variáveis "Valor líquido pago"e "Valor bruto pago"exercem uma influência significativa no valor final de $P(Y = 1)$. Essas duas variáveis estão diretamente associadas à quantia do empréstimo que o cliente já quitou, indicando sua relevância na predição do resultado. Ao calcular a Razão de chances dessas duas variáveis, temos que:

- "Valor líquido pago": apresenta um RC de 0.008, o que sugere que, mantendo todas as outras variáveis constantes, a chance de o empréstimo ser classificado como bom é 125 vezes maior do que ser classificado como ruim.
- "Valor bruto pago": exibe um RC de 27.68, indicando que, ao manter todas as outras variáveis constantes, a chance de o empréstimo ser classificado como ruim é 27 vezes maior do que ser classificado como bom.

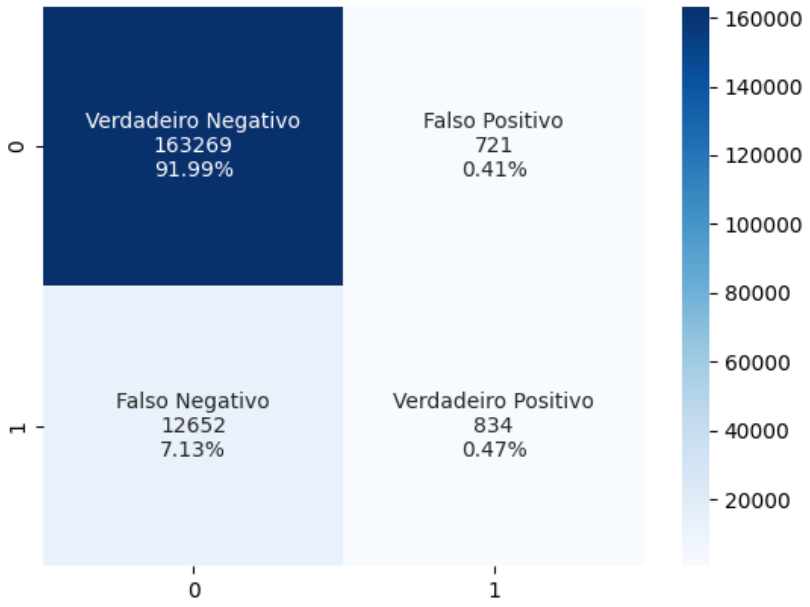


Figura 5: Matrix de confusão do modelo logístico

Visualizando os dados da Figura 5 é possível analisar os resultados do modelo logístico no conjunto de teste. O conjunto de teste apresenta uma distribuição da variável resposta de com mais de 92% dos casos como um empréstimo bom, e o restante como o empréstimo ruim.

	Precisão	**Recall**	**F1-Score**	**Tamanho da amostra**
0	0.928	0.995	0.961	163990
1	0.536	0.062	0.111	13486
Média macro	0.732	0.528	0.535	177476
Média ponderada	0.898	0.925	0.896	177476
Acurácia				0.924649

Tabela 5: Report do modelo logístico

Com base nos dados apresentados na Tabela 5 e na Figura 5, observamos que o modelo exibe uma acurácia elevada. Ele é capaz de fazer previsões precisas na maioria dos casos, alcançando uma taxa de 92,46% de classificações corretas no conjunto de teste. No entanto, é crucial destacar que essa elevada acurácia é influenciada pela proporção significativa de casos onde o empréstimo é rotulado como "bom", presente em mais de 92% dos dados de teste. Como resultado, o modelo tende a classificar uma parte considerável dos dados como "0", refletindo a influência dessa distribuição desigual na estimação dos parâmetros do modelo logístico.

Ao examinarmos a precisão do modelo, observamos uma taxa de acerto de 73% nas previsões em comparação com as rótulos reais do conjunto de teste. É importante ressaltar a notável precisão na categoria "Empréstimo bom", atingindo quase 92%. No entanto, vale destacar que esse valor elevado está correlacionado ao desequilíbrio nos dados, onde a classe "Empréstimo bom"é predominante.

Ao avaliar o recall do modelo logístico, observamos, em média, valores mais baixos em comparação com a precisão. O recall médio é de 52,8%, indicando que, ao analisar as porcentagens das rótulos reais, o modelo conseguiu acertar um pouco mais da metade delas. Esse desempenho é atribuído ao alto número de falsos negativos no modelo, visto que, ao considerar o total de "Empréstimos ruins"(13.486), o modelo acertou apenas 834 desses casos.

O F1-score acaba refletindo a real situação do modelo, pois ele balanceia os bons resultados apresentados pela precisão com os resultados ruins do recall. O F1-score médio apresentado foi de 52,57%.

A avaliação global do modelo logístico revela um viés significativo, amplificado pelo desequilíbrio nos dados. Embora o modelo tenha alcançado uma taxa geral de acerto de 92%, sua incapacidade de distinguir adequadamente entre "Empréstimos bons"e "Empréstimos ruins"é evidente. Este desempenho inferior sugere limitações na capacidade do modelo de generalizar e discriminar efetivamente entre as categorias, indicando a necessidade de refinamentos ou considerações adicionais para melhorar sua robustez.

1.3 Rede neural

Dado o modelo de rede neural escolhido(ver metodologia), o mesmo teve os seguintes resultados nos dados de validação nos dados de teste:

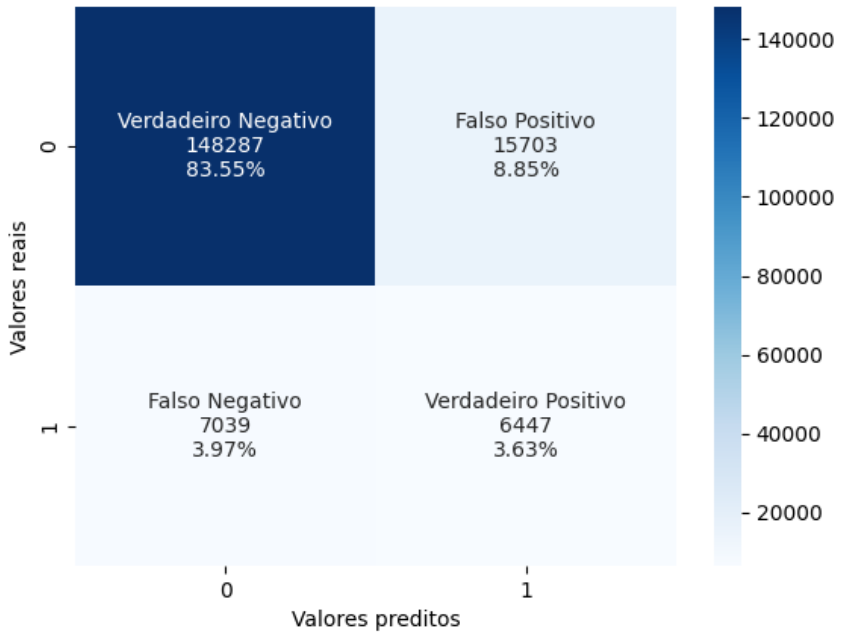


Figura 6: Matrix de confusão da rede neural

A Figura 6 estima que aproximadamente 12% da amostra de teste estimada pelo modelo consiste em empréstimos ruins, uma estimativa ligeiramente distante da distribuição real apresentada na Tabela 1. No entanto, esse resultado sugere que o modelo conseguiu generalizar seus resultados, quando comparado com o modelo logístico. Por outro lado, o número de empréstimos classificados como "0"diminuiu, representando mais de 84% da amostra de teste. Esse ajuste parece ter aumentado o número de Falsos Positivos, provavelmente devido à tentativa do modelo em variabilizar os resultados.

	Precisão	**Recall**	**F1-Score**	**Tamanho da amostra**
0	0.954	0.904	0.929	163990
1	0.291	0.478	0.362	13486
Média macro	0.622	0.691	0.645	177476
Média ponderada	0.904	0.872	0.886	177476
Acurácia				0.872

Tabela 6: Report da rede neural

Os resultados apresentados na Tabela 6 indicam um desempenho satisfatório do modelo, especialmente em termos de acurácia e métricas avaliadas para a classe "0". Contudo, ao comparar esses resultados com as métricas da classe "1", percebe-se que o modelo ainda é influenciado pelo elevado número de observações na classe "0". Entretanto, uma análise mais detalhada do Recall da classe "1"revela que o modelo conseguiu reduzir

significativamente o número de Falsos Negativos, identificando corretamente quase metade dos empréstimos considerados ruins na base original. Essa melhoria no Recall da classe "1"resultou em uma precisão inferior para essa classe, refletindo que menos de 30% das previsões do modelo foram corretas nesse contexto, como evidenciado na Tabela 6.

Em termos gerais, as decisões relacionadas à arquitetura do modelo, seus hiperparâmetros, estratégia de treinamento e outros fatores contribuíram para que o modelo de redes neurais realizasse previsões de alta qualidade, não se limitando apenas ao desbalanceamento dos dados. Os resultados obtidos pelo modelo de redes neurais foram satisfatórios devido à eficácia da arquitetura da rede. Portanto, buscar aprimorar ainda mais essa arquitetura pode ser uma abordagem promissora na busca por resultados ainda melhores.

1.4 Interpretação da rede neural

Após definir o modelo de rede neural e examinar seus resultados, esta seção aborda a interpretação do modelo. Para isso, foram construídos gráficos com base nos resultados do SHAP, destacando as variáveis de maior importância no resultado final. Utilizando uma amostra de 80 observações, o valor de SHAP foi calculado para cada observação, permitindo uma análise tanto individual quanto conjunta dessa amostra.

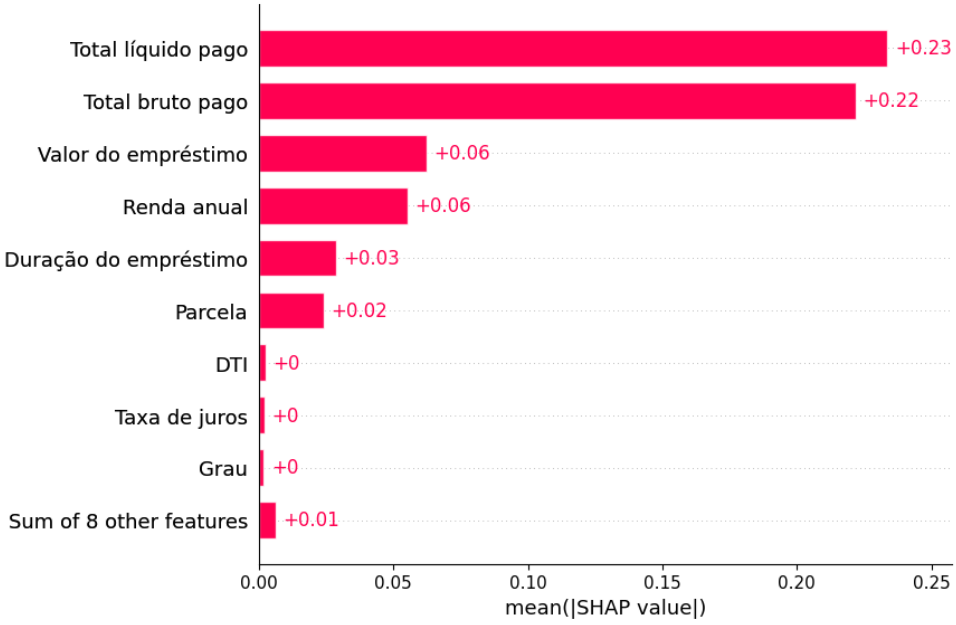


Figura 7: Média absoluta dos valores de shap

O Gráfico 7 exibe a média absoluta dos valores de SHAP para cada variável nas 80 observações consideradas. Essa representação oferece insights sobre quais variáveis o modelo considerou mais relevantes durante as predições, destacando a magnitude da

contribuição de cada variável. Vale notar que, por se tratar de valores absolutos, o gráfico não proporciona informações sobre se a contribuição de cada variável é positiva ou negativa. Entretanto, os próximos gráficos irão elucidar essa questão ao detalhar a contribuição específica de cada variável.

Ao analisar cada variável no gráfico, destaca-se que "Total líquido pago"e "Total bruto pago"são as que mais contribuem para o resultado final do modelo. O gráfico enfatiza a relevância de nove variáveis, omitindo o restante devido à sua baixa contribuição. Ao observar as variáveis omitidas, percebe-se que, somadas, suas contribuições aproximam-se de 0.01, evidenciando a sua baixa influência no resultado final do modelo de rede neural.

Os próximos graficos proporcionam a visualização dos valores SHAP para cada variável em observações individuais. O gráfico também revela os valores específicos utilizados em cada variável. A apresentação de quatro figuras distintas tem como objetivo compreender como o valor SHAP é atribuído em quatro cenários distintos que surgem em problemas de classificação binária.

Os 2 primeiros gráficos mostram casos onde o modelo acertou suas predições, tanto para casos de "Empréstimos ruins"quanto para "Empréstimos bons".

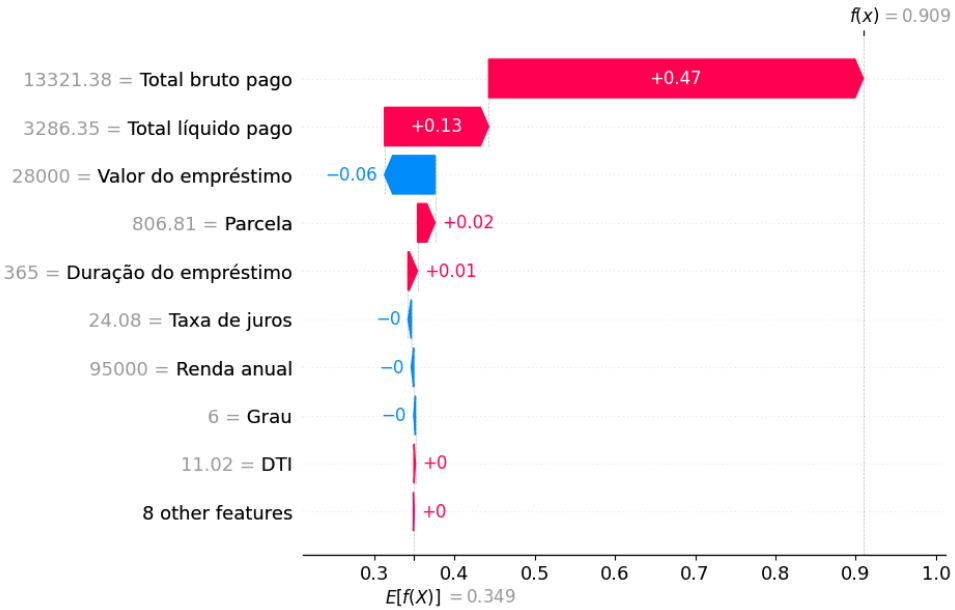


Figura 8: Valor de SHAP para uma observação do tipo Verdadeiro Positivo

Na Figura 8, cenário em que o modelo acertou a classificação de um empréstimo como ruim, observa-se que as variáveis que mais contribuíram foram as mesmas observadas na Figura 7, comportamento esse que vai prevalecer nos demais casos. Ao analisar os valores dessas duas variáveis, nota-se que, para esse cliente específico, ainda resta um montante significativo do empréstimo a ser pago, totalizando quase 90% do valor líquido

pendente e quase 50% do valor bruto pendente. Uma diferença muito grande entre o valor do empréstimo e o que falta a ser pago pode ser um dos fatores que está ocasionando a rotulação de observações como "Empréstimos ruins".

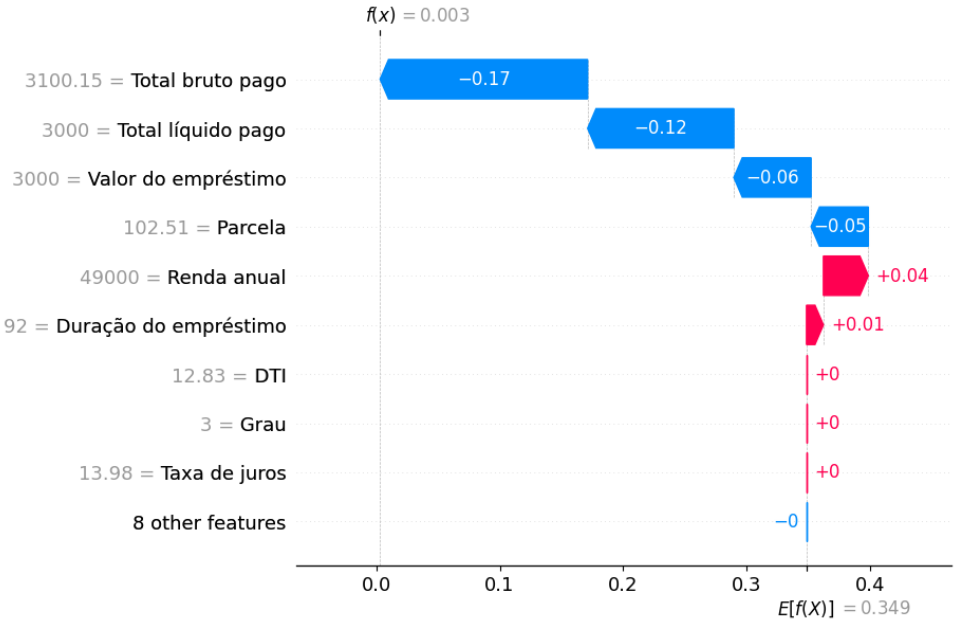


Figura 9: Valor de SHAP para uma observação do tipo Verdadeiro Negativo

Ao analisar a Figura 9, que representa o cenário em que o modelo acertou a rotulação de um empréstimo bom, nota-se um valor final bem próximo de 0. Indícios de que o modelo teve mais confiança ao realizar essa predição. Ao analisar os valores de cada variável, é possível perceber que o cliente já está finalizando ou finalizou o empréstimo, dado que as variáveis de pagamento do empréstimo chegaram no valor real do empréstimo.

Analisando as Figuras 8 e 9, é possível observar a flexibilidade do modelo em lidar com diferentes valores da mesma variável. Quando o modelo encontra valores que indicam um empréstimo ruim, atribui valores positivos para a contribuição das variáveis, aproximando-as de 1. Da mesma forma, para empréstimos bons, o modelo adiciona uma contribuição negativa, direcionando o resultado para 0. Isso demonstra como o modelo responde de forma dinâmica às variações nos valores das variáveis.

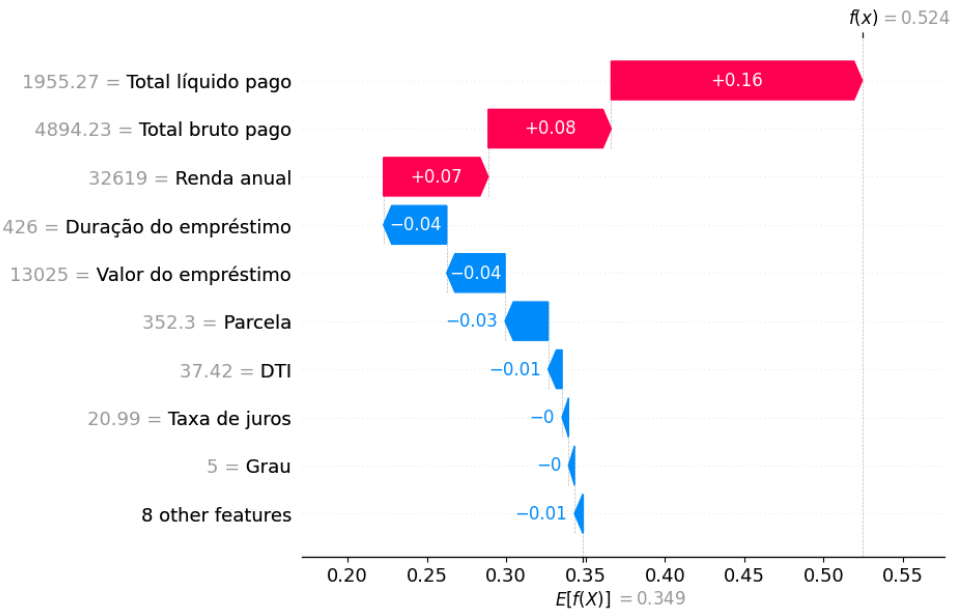


Figura 10: Valor de SHAP para uma observação do tipo Falso Positivo

Ao analisar um exemplo em que o modelo erroneamente classifica um "Empréstimo ruim", conforme ilustrado na Figura 10, destacam-se as características já discutidas nas interpretações anteriores. O resultado do modelo foi muito próximo do limiar de decisão, que é 0,5, indicando que o modelo teve alguma indecisão ao realizar a predição desse caso.

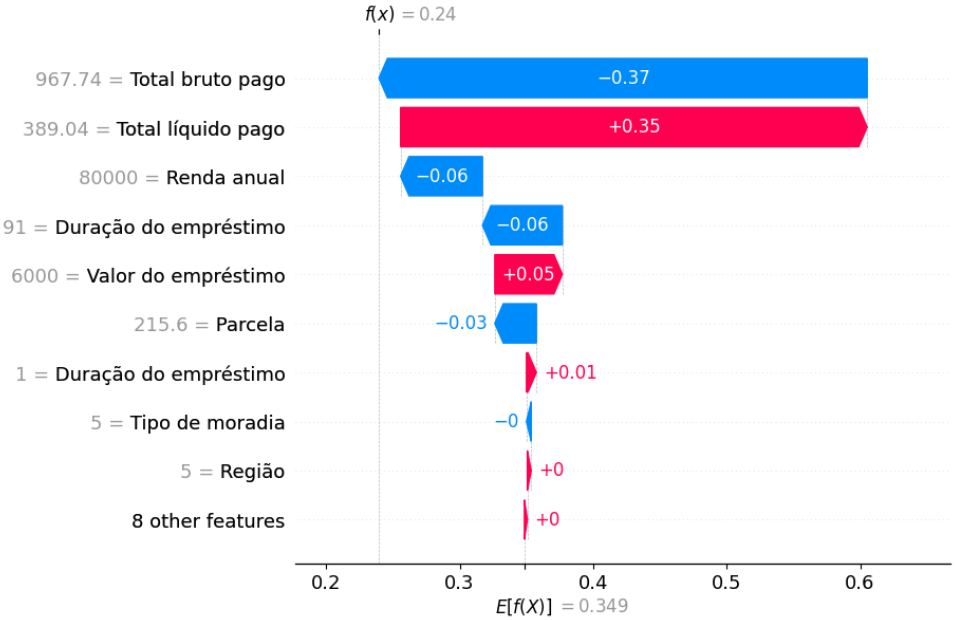


Figura 11: Valor de SHAP para uma observação do tipo Falso Negativo

Observando a Figura 11, que retrata um cenário em que o modelo classifica de forma equivocada um "Empréstimo bom", é possível perceber um comportamento de divergência entre as variáveis que mais contribuem com o resultado do modelo. Esse

padrão de divergência não foi observado nos casos anteriores, e pode ter influenciado o modelo a realizar uma predição incorreta.

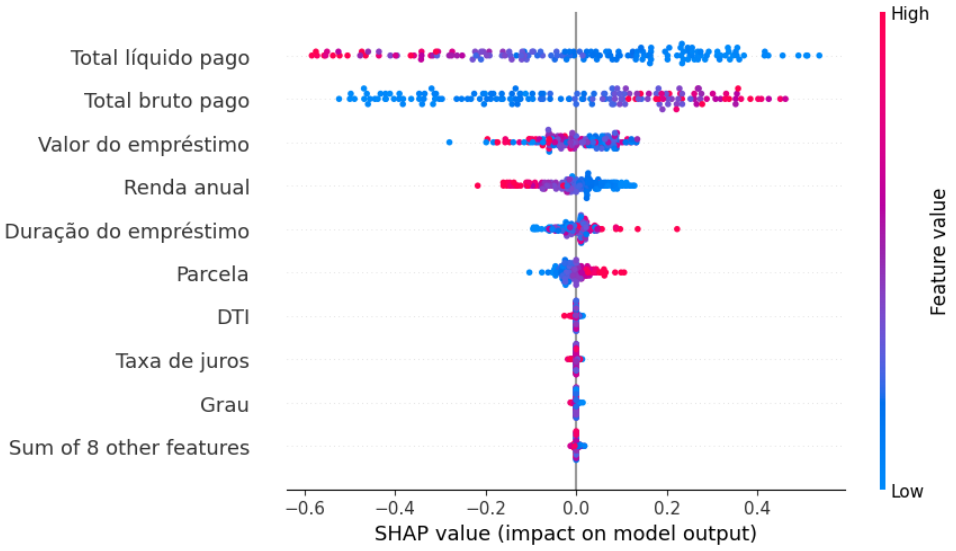


Figura 12: Valores de shap para as 80 observações utilizadas

A Figura 12 apresenta o comportamento do valor de SHAP para cada variável entre as 80 observações utilizadas no experimento. No gráfico, o eixo x refere-se à distribuição do valor de SHAP, enquanto o eixo y representa cada variável. O eixo das cores ilustra a distribuição dos valores da variável, sendo que cores mais quentes indicam valores maiores e cores mais frias indicam valores menores.

Ao analisar as variáveis que mais influenciam o resultado do modelo, observa-se uma distribuição mais dispersa no eixo x. Ao focar nas duas variáveis de maior contribuição, nota-se um comportamento heterogêneo dos valores em comparação com os valores de SHAP. Para a primeira variável, elevados valores tendem a impactar negativamente no resultado do modelo, enquanto a segunda variável apresenta o comportamento oposto, com valores mais baixos contribuindo negativamente no resultado do modelo.

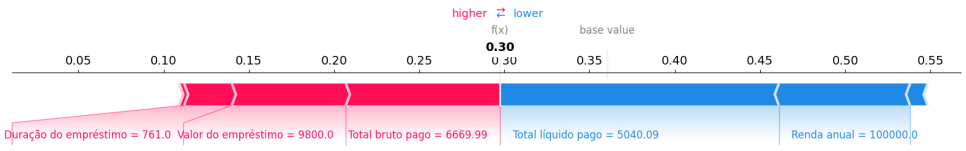


Figura 13: Gráfico de força em uma observação

A Figura 13 ilustra a "força"de contribuição de cada variável no modelo. Este gráfico proporciona uma visão clara das variáveis que tiveram impacto positivo e negativo no resultado final. Cada barra representa a intensidade da contribuição de uma variável

específica, sendo que aquelas com maior influência concentram-se no centro, enquanto as de menor influência ficam nas extremidades. A figura é centrada no valor final predito pelo modelo, que, neste caso, é observado como 0.3, indicando um empréstimo classificado como bom.

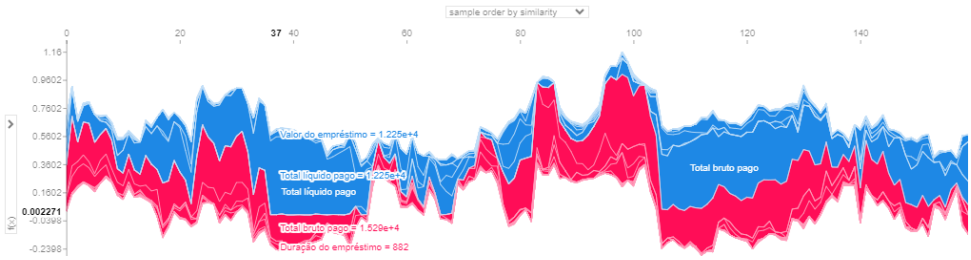


Figura 14: Gráfico de força para múltiplas observações

A Figura 14 é uma generalização da Figura 13, considerando as 80 observações. Este gráfico é um recorte obtido de uma visualização dinâmica gerada pelo pacote SHAP. Devido à natureza dinâmica do gráfico original, ao transformá-lo em uma imagem estática, parte de sua capacidade de representação é limitada. No entanto, é possível observar alguns padrões nesse recorte. Por exemplo, cores mais quentes indicam contribuições positivas quando o modelo faz predições próximas de 1, com tons de vermelho representando adições das variáveis ao resultado. Da mesma forma, cores mais frias indicam contribuições negativas quando o modelo faz predições próximas de 0.

1.5 Benchmark entre regressão logística e redes neurais

Ao obter interpretações do modelo de redes neurais, abre-se a possibilidade de realizar comparações entre os modelos de regressão logística e redes neurais no âmbito interpretativo. Até recentemente, essa capacidade de interpretação estava exclusivamente associada ao modelo logístico em comparação com o modelo de redes neurais. Essa análise comparativa se soma às comparações já realizadas entre os modelos, englobando aspectos como arquitetura, tempo de treinamento, tempo de predição e os resultados gerados por ambos os modelos. Os resultados abaixo evidenciam essas comparações.

1.5.1 Complexidade da arquitetura

Modelo	Número de parâmetros
Regressão Logística	18
Rede Neural	151233

Tabela 7: Número de parâmetros nos modelos

Ao examinar a Tabela 7, destaca-se a significativa disparidade na complexidade entre os modelos. Enquanto o modelo logístico possui apenas um parâmetro para cada covariável, a rede neural apresenta mais de 8400 parâmetros para cada parâmetro do modelo logístico, sendo esses distribuídos nos neurônios e camadas da rede.

1.5.2 Resultado dos modelos

Métricas	**Regressão logística**	**Rede neural**
Falsos Positivos	721	15703
Falsos Negativos	12652	7039

Tabela 8: Comparação dos resultados de Falsos Positivos e Falsos Negativos

Métricas	**Regressão logística**	**Rede neural**
Precisão (Classe 0)	0.928081	0.954
Precisão (Classe 1)	0.536334	0.291
Recall (Classe 0)	0.995603	0.904
Recall (Classe 1)	0.0618419	0.478
F1-Score (Classe 0)	0.960657	0.929
F1-Score (Classe 1)	0.110897	0.362
Acurácia	0.924649	0.872

Tabela 9: Comparação dos resultados da Regressão logística e Rede neural. Em verde, o modelo que obteve o melhor resultado na respectiva métrica.

Covariáveis	Média	Desvio Padrão
Total líquido pago	0.24363	0.14333
Total bruto pago	0.23619	0.13483
Valor do empréstimo	0.06298	0.03885
Renda anual	0.05132	0.04181
Duração do empréstimo	0.03105	0.03172
Parcela	0.02586	0.0215
DTI	0.00242	0.00422
Taxa de juros	0.00232	0.00355
Grau	0.00169	0.00249
Duração do empréstimo	0.00166	0.00224
Tipo de moradia	0.00117	0.00187
Finalidade	0.00094	0.00183
Total de juros	0.00077	0.00166
Região	0.00071	0.00132
Prazo	0.00068	0.00124
Renda categorizada	0.00058	0.00143
Tipo da aplicação	0	0

1.5.3 Tempo de execução

Estatísticas	**Regressão logística**	**Rede neural**
Mínimo	0.0	52.067995
Quartil 25	0.0	53.327155
Média	0.3335619	59.446688
Mediana	0.0	56.855202
Quartil 75	0.88143349	66.278052
Máximo	1.50370598	92.03124
Variância	0.28385463	59.008917
Desvio padrão	0.5327801	7.681726

Tabela 10: Tempo de predição (em ms) de cada modelo, em uma amostra com 50 observações.

Observações	**Tempo de execução**
1	317s
80	7.07hrs

Tabela 11: Tempo de execução para realizar a interpretação das variáveis

2 Conclusão

Referências

3 Anexo