Modelagem Preditiva de Risco de Crédito utilizando Regressão Logística

---

Análise exploratória, modelagem e aplicação prática de score

Desenvolvido por: Paulo Eduardo De Vincenzi

Projeto Pessoal de Data Science

Outubro/2025

1. CONTEXTO

O crédito desempenha um papel essencial na vida financeira das pessoas e no funcionamento da economia. Ele possibilita que consumidores realizem compras relevantes, enfrentem imprevistos com maior segurança e tenham acesso a oportunidades que, de outra forma, seriam inviáveis. Para as instituições financeiras, como bancos e administradoras de cartão de crédito, conceder crédito é igualmente importante, pois impulsiona seus negócios e amplia o acesso a serviços financeiros.

Entretanto, essa prática envolve desafios. Um dos mais relevantes é o risco de inadimplência, que ocorre quando o cliente não realiza o pagamento de suas faturas dentro do prazo acordado. Monitorar e prever esse comportamento é fundamental para equilibrar a expansão do crédito com a sustentabilidade financeira da instituição. Quando a previsão não é precisa, podem surgir consequências como:

- Perdas financeiras diretas, devido ao não recebimento do crédito concedido
- Aumento da provisão para devedores duvidosos, impactando a saúde financeira da empresa
- Reavaliação constante das políticas de crédito, muitas vezes restringindo o acesso de clientes bons ao sistema financeiro

Nesse cenário, a capacidade de prever previamente quais clientes têm maior probabilidade de se tornarem inadimplentes se torna um diferencial competitivo. Com análises robustas e baseadas em dados, é possível:

- Reduzir perdas com crédito mal concedido
- Melhorar a qualidade da carteira de clientes
- Ajustar limites de forma inteligente
- Manter clientes com bom comportamento adimplente, aumentando receita

Foi nesse contexto que este projeto foi desenvolvido: um estudo aplicado de **modelagem preditiva de score de crédito e inadimplência**, baseado em dados reais de clientes de cartão de crédito, com foco em entregar insights e uma solução funcional para apoio à decisão de concessão de crédito.

2. OBJETIVO

O objetivo principal deste projeto é desenvolver uma solução prática e assertiva para predição de risco de crédito com regressão logística, incluindo uma aplicação funcional que estima score e probabilidade de inadimplência em tempo real.

A proposta envolve três pilares:

I. **Análise Exploratória dos Dados (EDA)**
Identificar padrões, comportamentos e sinais de alerta associados à inadimplência, gerando insights relevantes para decisões de crédito.

II. **Modelagem Preditiva de Score e Probabilidade de Inadimplência**
Estimar a probabilidade real de um cliente se tornar inadimplente e traduzir essa probabilidade em um score simplificado (300 a 850), facilitando sua interpretação.

III. **Aplicação Funcional para Tomada de Decisão**
Implementar uma interface gráfica que permita ao usuário inserir dados de um cliente e obter imediatamente o score e o risco estimado. Essa aplicação mostra a viabilidade do modelo como produto final, aproximando a análise do contexto operacional do mercado financeiro.

Em termos de **impacto de negócio**, o projeto busca:

- Contribuir para redução de perdas financeiras com concessão inadequada de crédito
- Otimizar políticas de aprovação, permitindo maior inclusão de clientes com bom perfil
- Prover maior eficiência e agilidade no processo de avaliação

3. DADOS UTILIZADOS

O modelo foi treinado com a base de dados “**Default of Credit Card Clients”**, fornecida pelo UCI Machine Learning Repository. Embora a base seja antiga e proveniente de um banco de Taiwan, os resultados podem ainda ser relevantes, dependendo do contexto, mas podem não refletir com precisão dados mais recentes ou específicos do Brasil.

A base contém informações de aproximadamente **30 mil clientes de cartão de crédito**, totalizando **24 variáveis** relacionadas ao perfil do cliente e ao seu comportamento financeiro recente.

3.1. Perfil Socioeconômico e de Crédito

As variáveis a seguir ajudam a identificar características estruturais que podem influenciar a capacidade de pagamento.

- Limite de crédito disponível (LIMIT_BAL)
- Sexo (SEX), idade (AGE), escolaridade (EDUCATION) e estado civil (MARRIAGE)

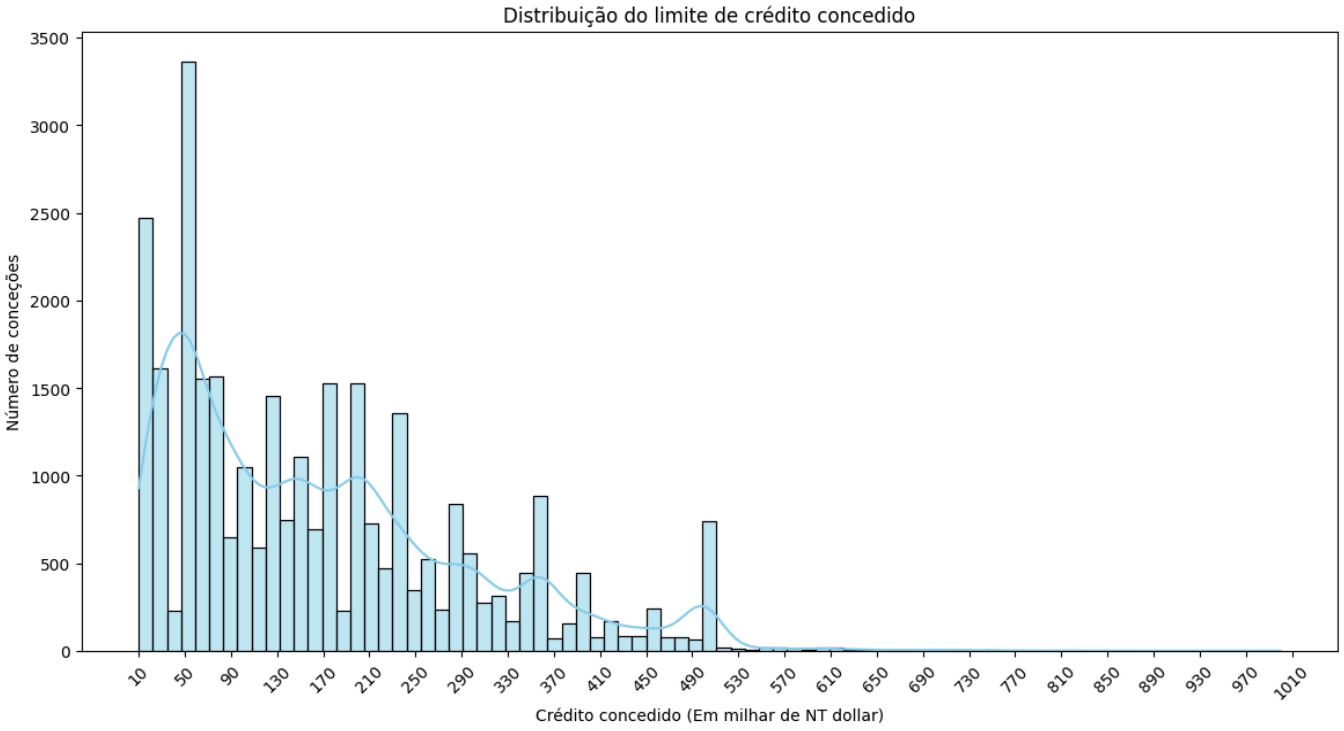

Figura 1 - Distribuição do limite de crédito concedido em milhar de NT dollar.

Fonte: Elaborado pelo autor.

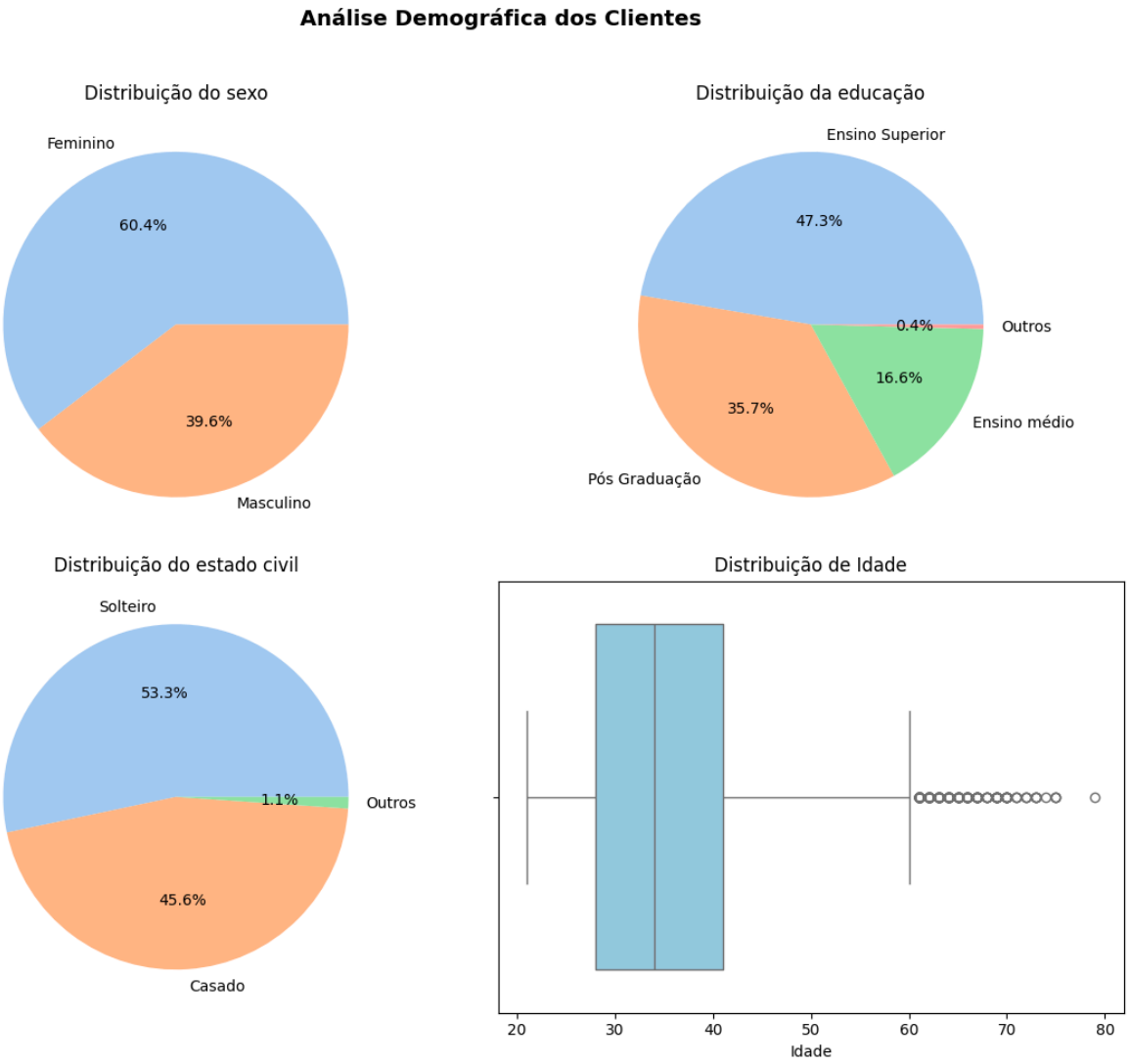

Figura 2 - Distribuições de sexo, educação, estado civil e idade.

Fonte: Elaborado pelo autor.

As Figuras 1 e 2 ajudam a visualizar quem são os clientes desta base: quanto crédito recebem e quais perfis são mais frequentes, primeiros indícios importantes para entender o comportamento de risco.

3.2. Histórico das Seis Últimas Faturas

As variáveis referentes às últimas seis faturas representam indicadores diretos de comportamento de pagamento, abrangendo:

- Quanto o cliente deveria pagar a cada mês (BILL_AMT1 a BILL_AMT6)
- Quanto o cliente pagou a cada mês (PAY_AMT1 a PAY_AMT6)
- Se houve atraso e por quanto tempo (PAY0 a PAY6)

Embora as variáveis relacionadas às últimas seis faturas sejam essenciais para a modelagem, seus gráficos individuais de distribuição apresentam pouca variação significativa entre os meses e, portanto, não agregariam informações relevantes à leitura deste relatório. Ainda assim, algumas observações importantes podem ser destacadas a partir do conjunto:

- Em média, **86% dos clientes** não apresentaram atraso (PAY) no pagamento ou não possuíam fatura no período analisado.
- Os **valores das faturas** (BILL_AMT) mantiveram-se concentrados em uma faixa semelhante ao longo dos meses, sem grandes alterações no padrão.

A principal diferença observada está no **valor efetivamente pago** (PAY_AMT). Nos meses mais antigos, mais distantes do momento de coleta, os pagamentos se concentraram em valores mínimos mais baixos. Já nas faturas mais recentes, verificou-se uma elevação desse mínimo, indicando um comportamento de pagamento crescente ao longo do tempo que pode ser visualizado no gráfico da Figura 3.

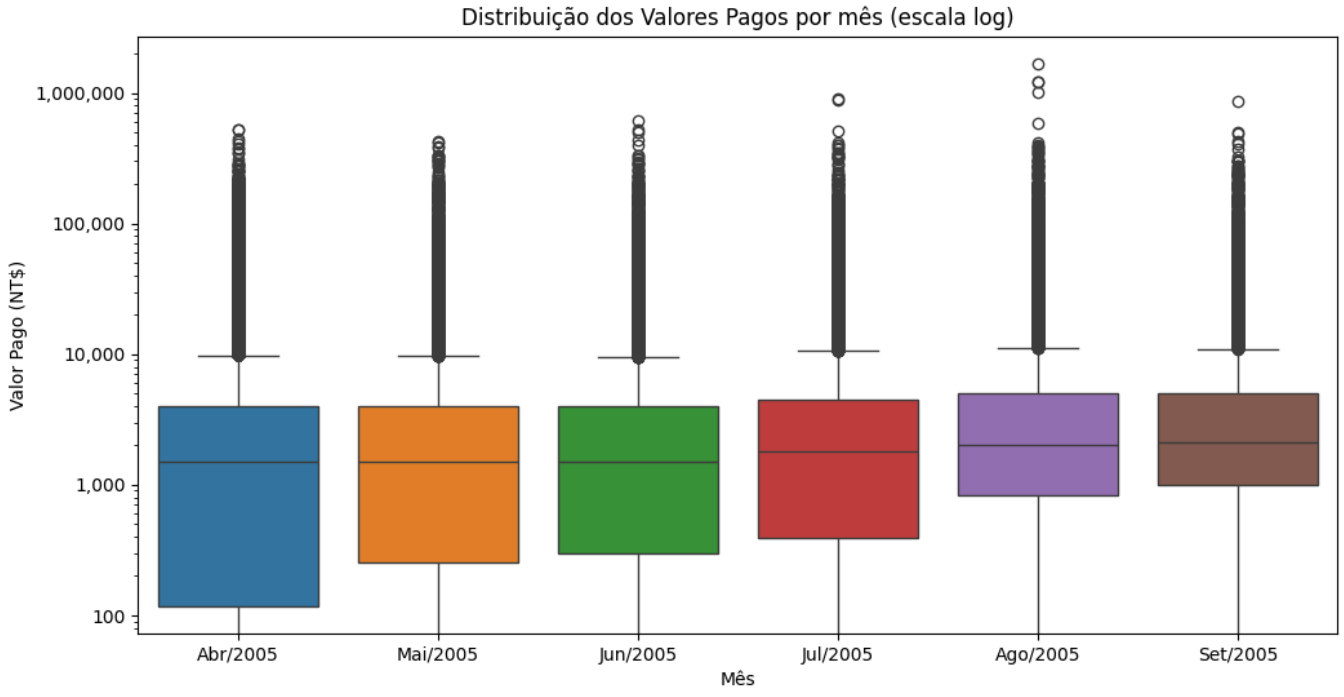

Figura 3 - Distribuição dos valores pagos a cada mês.

Fonte: Elaborado pelo autor.

4. PRINCIPAIS INSIGHTS DO EDA

O objetivo da análise exploratória foi identificar padrões que diferenciam clientes adimplentes dos inadimplentes, ajudando a direcionar a modelagem. A Figura 4 apresenta a correlação entre as variáveis numéricas e a inadimplência, destacando as que mais influenciam o risco de não pagamento. Nota-se que o **limite de crédito (LIMIT_BAL) exerce maior impacto**, seguido pelos valores pagos em cada fatura (PAY_AMT).

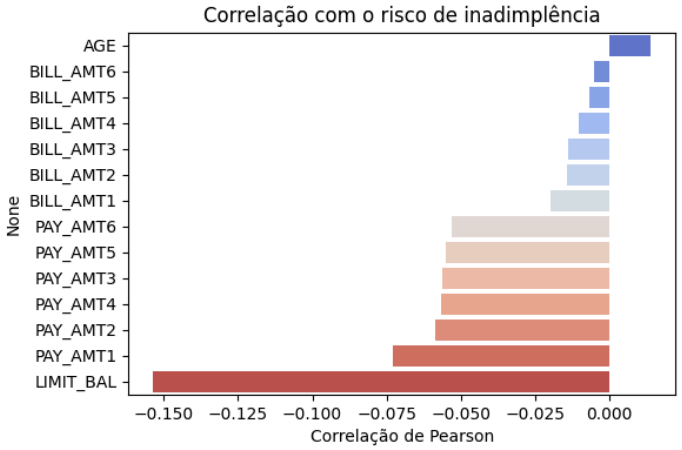

Figura 4 - Correlação entre variáveis numéricas e o risco de inadimplência.

Fonte: Elaborado pelo autor.

Observando a Figura 5, nota-se que indivíduos com menor limite de crédito apresentam maior propensão à inadimplência. Já a **idade** mostra pouca diferença entre adimplentes e inadimplentes, confirmando sua **baixa relevância** como indicador de risco.

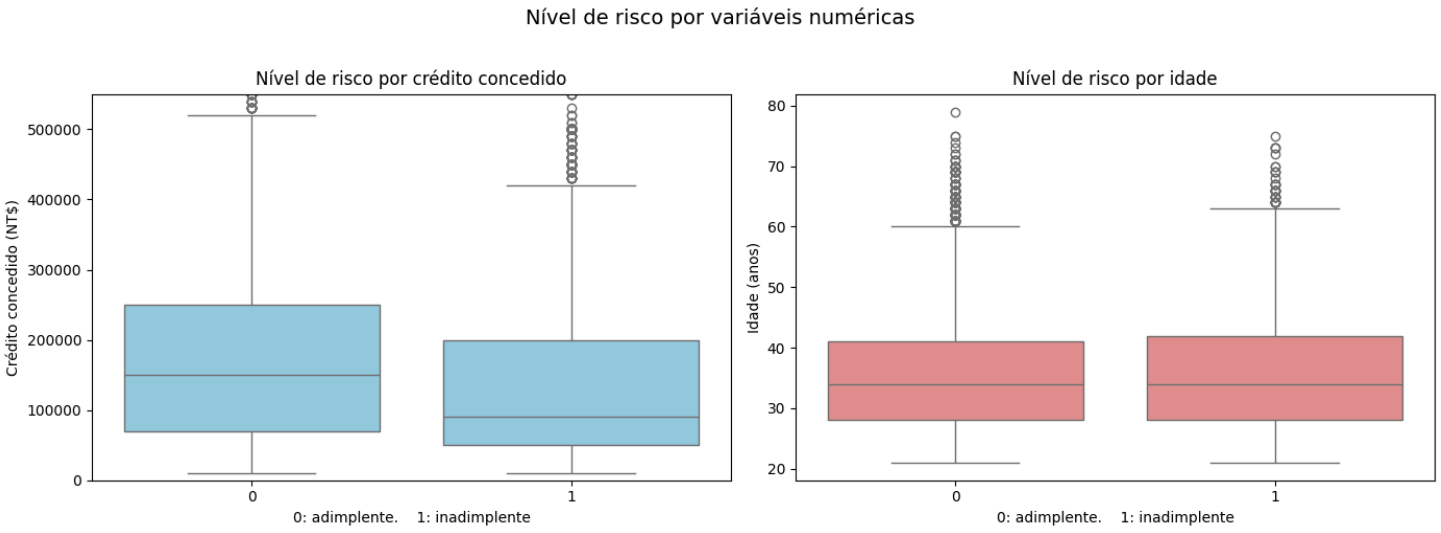

Figura 5 - Nível de risco por crédito concedido e idade.

Fonte: Elaborado pelo autor.

Além das variáveis numéricas, as variáveis categóricas também contribuem para explicar o risco de inadimplência. Conforme é exibido na Figura 6 abaixo, **a taxa de inadimplência aumenta à medida que se elevam os níveis de atraso no pagamento**.

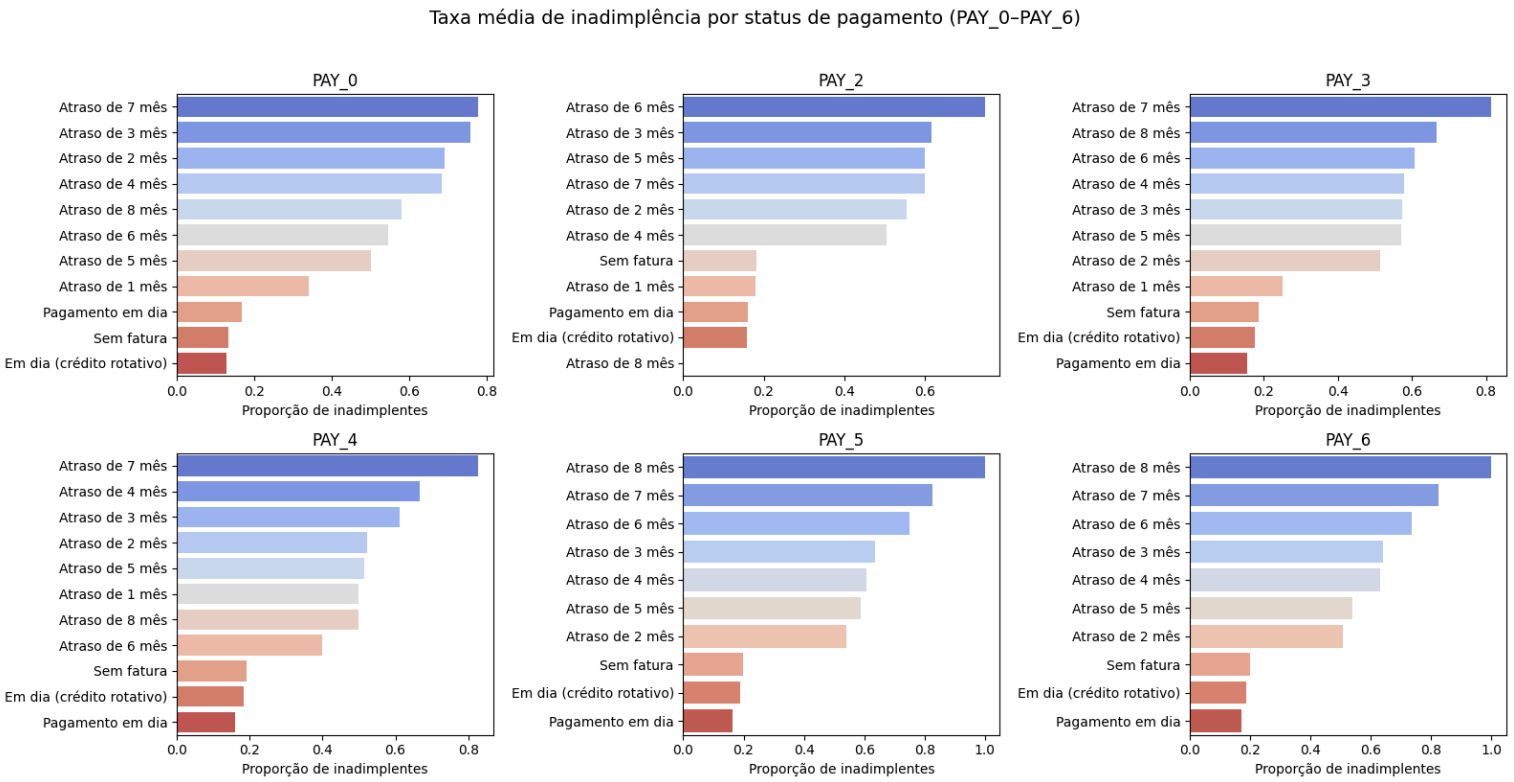

Figura 6 - Taxa média de inadimplência por atraso de pagamento.

Fonte: Elaborado pelo autor.

Já a Figura 7 a seguir evidencia que **homens apresentam maior propensão à inadimplência** em comparação às mulheres, **assim como indivíduos com menor escolaridade e pessoas casadas**, que demonstram maior risco em relação aos solteiros.

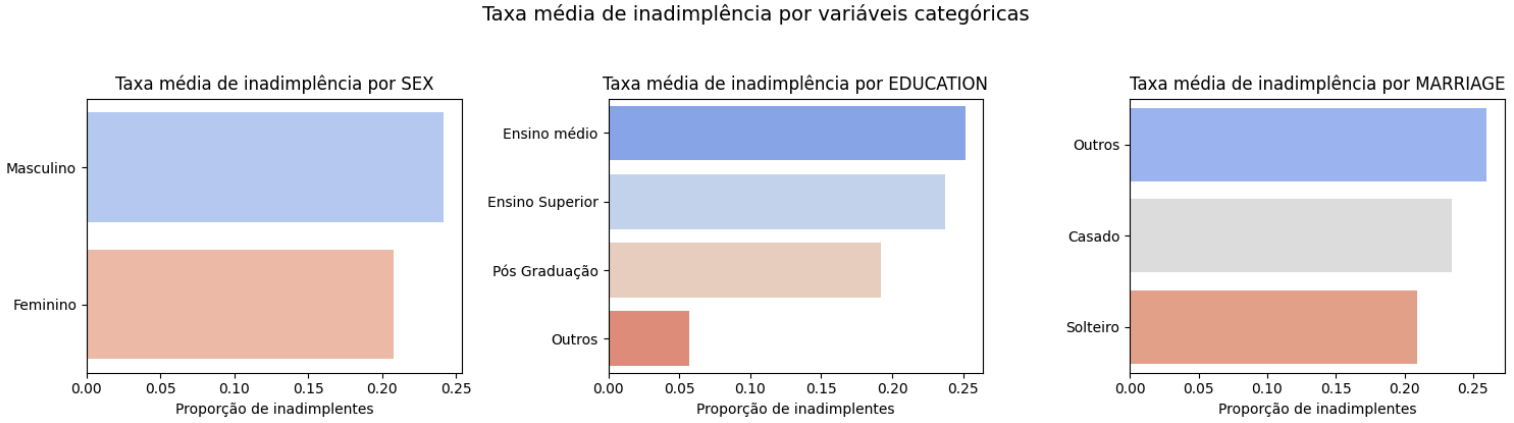

Figura 7 - Taxa média de inadimplência por sexo, nível de escolaridade e estado civil.

Fonte: Elaborado pelo autor

5. MODELAGEM PREDITIVA E RESULTADOS

A etapa de modelagem preditiva iniciou-se com o enriquecimento do conjunto de dados original. Além das 24 variáveis fornecidas pela base, **foram criadas 16 novas variáveis derivadas**, com o objetivo de capturar tendências e padrões que as variáveis originais isoladamente não evidenciam. Entre os exemplos de variáveis derivadas criadas, destacam-se:

- **PAY_to_BILL_MEAN**: proporção média entre o valor pago e o total da fatura, refletindo o comprometimento do cliente com o pagamento de suas dívidas.
- **BILL_TREND**: diferença entre a dívida mais recente e a mais antiga, indicando se há tendência de crescimento ou redução no saldo devedor ao longo dos meses.

Antes de prosseguir para o treinamento do modelo, foi realizada uma etapa de seleção de variáveis para evitar redundância e dependência excessiva entre os atributos explicativos, o que pode distorcer os coeficientes do modelo e gerar interpretações incorretas. Para isso, foi aplicado o teste de multicolinearidade por meio do Variance Inflation Factor (VIF)**.** Com base nesse critério, **foram removidas as variáveis cujo VIF apresentou valor infinito ou acima de 10**. Ao final desse processo, reduziu-se o conjunto de atributos de 40 variáveis (originais e derivadas) para um total de 24 variáveis selecionadas, garantindo um modelo mais coerente.

Além disso, **as variáveis categóricas foram convertidas para o formato *dummy***, permitindo que o modelo de regressão logística as interpretasse corretamente. Esse procedimento transforma categorias em variáveis numéricas binárias, evitando que o algoritmo atribua relações de ordem inexistentes entre os grupos. Em seguida, **as variáveis numéricas foram padronizadas**, garantindo que todas estivessem em uma mesma escala de valores. Essa etapa é essencial em modelos baseados em pesos como a regressão logística, pois evita que variáveis com magnitudes maiores exerçam influência desproporcional no treinamento.

Após o pré-processamento, **os dados foram divididos em dois conjuntos: 80% para treino do modelo e 20% para teste**. Considerando que apenas cerca de 22% dos clientes da base eram inadimplentes, **aplicou-se o parâmetro *class_weight='balanced'*** na regressão logística, a fim de compensar o desbalanceamento da variável-alvo e evitar que o modelo favorecesse excessivamente a classe majoritária (adimplentes). Durante a

etapa de predição, **definiu-se um threshold de 0,40 para classificação de inadimplência**. Esse limite reduz a quantidade de inadimplentes não identificados, porém aumenta o número de falsos positivos, classificando alguns bons pagadores como de risco. A escolha do threshold ideal depende da estratégia da instituição financeira, já que existe um equilíbrio econômico a ser considerado: cada cliente inadimplente aprovado gera prejuízo, enquanto cada cliente adimplente recusado representa perda de receita e oportunidade. Assim, a definição desse parâmetro deve considerar os custos e benefícios envolvidos no negócio, buscando maximizar o retorno esperado da carteira de crédito.

Além disso, foram analisadas as **10 variáveis com maior influência no modelo**, ilustradas na Figura 8 a seguir. Observa-se que atributos como **sexo, estado civil e nível de educação** estão entre os mais relevantes, reforçando os padrões já identificados na análise exploratória. Além disso, **as variáveis relacionadas aos atrasos de pagamento (PAY) destacaram-se como os principais indicadores de risco**, evidenciando que o histórico recente de inadimplência permanece sendo o fator mais determinante para a previsão do modelo. Essa coerência entre EDA e modelagem aumenta a confiabilidade dos resultados e reforça a interpretabilidade do modelo adotado.

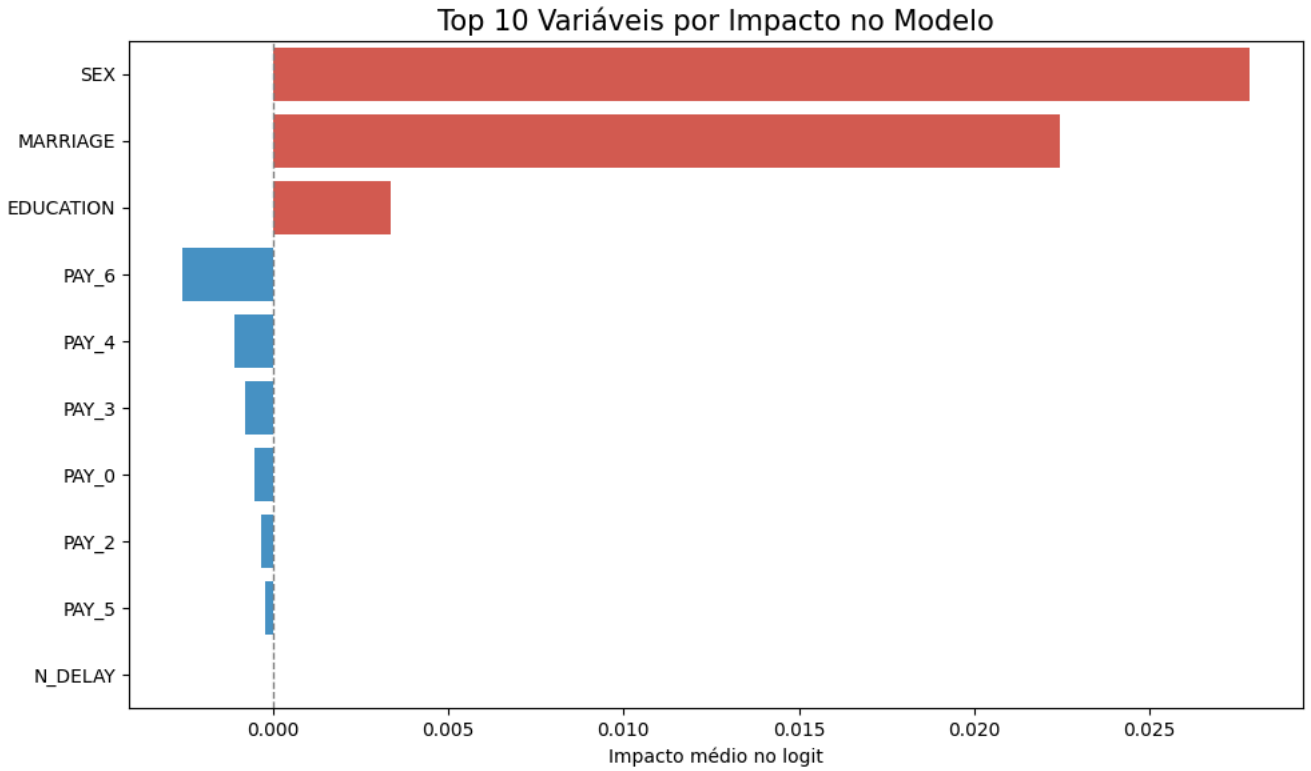

Figura 8 - As 10 variáveis mais impactantes para determinar a probabilidade de inadimplência.

Fonte: Elaborado pelo autor

Por fim, **o desempenho do modelo foi avaliado por meio da matriz de confusão**, apresentada na Figura 9 abaixo, permitindo verificar diretamente os acertos e erros na identificação de clientes adimplentes e inadimplentes. **Além disso, utilizou-se o valor da métrica AUC-ROC**, que mede a capacidade do modelo de distinguir corretamente entre as duas classes. **Após confirmada sua performance, o modelo final foi salvo** juntamente com o *scaler* aplicado às variáveis numéricas e com o conjunto e ordem das variáveis selecionadas, garantindo reprodutibilidade e permitindo sua utilização direta na aplicação funcional desenvolvida.

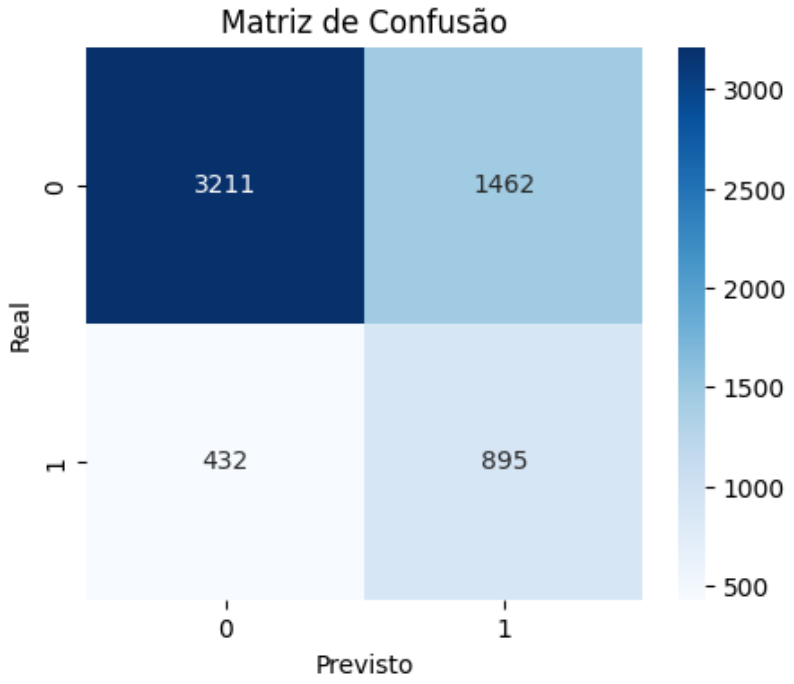

Figura 9 - Matriz de confusão do modelo de regressão logística.

Fonte: Elaborado pelo autor.

Com base nos resultados, o modelo alcançou um **AUC-ROC de 0,748**, **desempenho considerado satisfatório para o contexto de modelagem de risco de crédito**. Observando a matriz de confusão da Figura 9, nota-se que a maior parte dos acertos refere-se aos clientes não inadimplentes (3.211 casos corretamente classificados), o que é esperado dado que essa é a classe majoritária. Já entre os inadimplentes, o modelo identificou corretamente 895 casos, enquanto 432 foram classificados incorretamente como bons pagadores. Esse resultado evidencia um **modelo mais conservador, priorizando a identificação da classe negativa** (não inadimplentes), mas ainda capaz de recuperar uma parcela relevante dos clientes com risco real. Importante destacar que a proporção de inadimplentes corretamente

identificados se mantém próxima à própria taxa de inadimplência da base (aprox. 22%), o que reforça a coerência do modelo com a distribuição dos dados e indica que **ajustes adicionais no threshold podem ser realizados conforme a estratégia de risco da instituição financeira**.

6. APLICAÇÃO REAL: INTERFACE DE SCORE

Com o modelo treinado e validado, foi desenvolvida uma aplicação prática que simula o processo de análise de risco em um cenário real de concessão de crédito. **A interface**, mostrada nas Figuras 10 e 11 a seguir, **permite que o usuário insira os dados de um cliente e, de forma imediata, obtenha a probabilidade estimada de inadimplência e um score simplificado na escala de 300 a 850**, faixa comumente utilizada no mercado financeiro. Além disso, a aplicação foi compilada em um executável, permitindo sua utilização de maneira simples e direta, sem necessidade de ambiente de desenvolvimento ou instalação de bibliotecas, bastando apenas clicar no arquivo para iniciar o sistema.

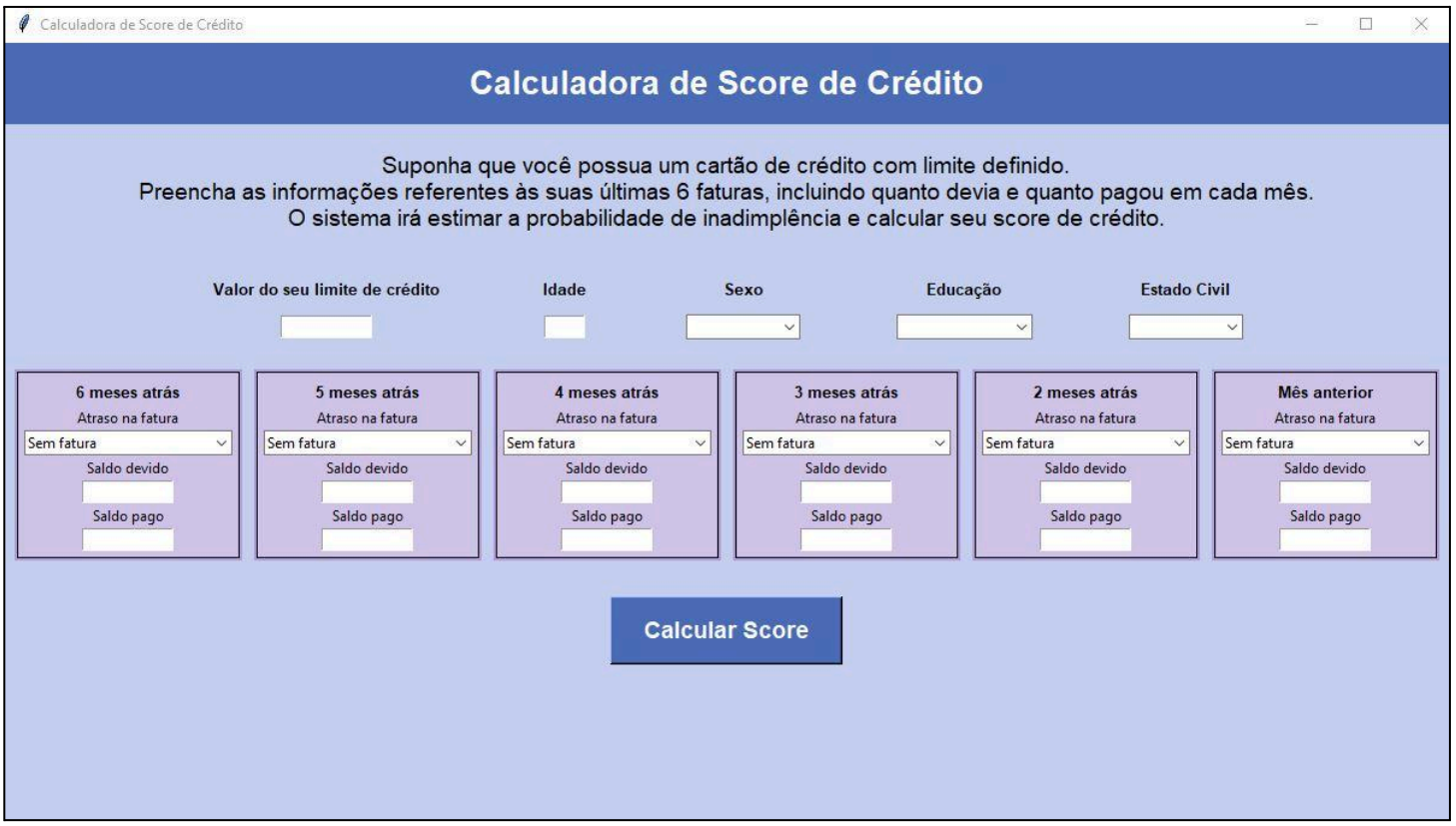

Figura 10 - Interface gráfica sem dados preenchidos.

Fonte: Elaborado pelo autor.

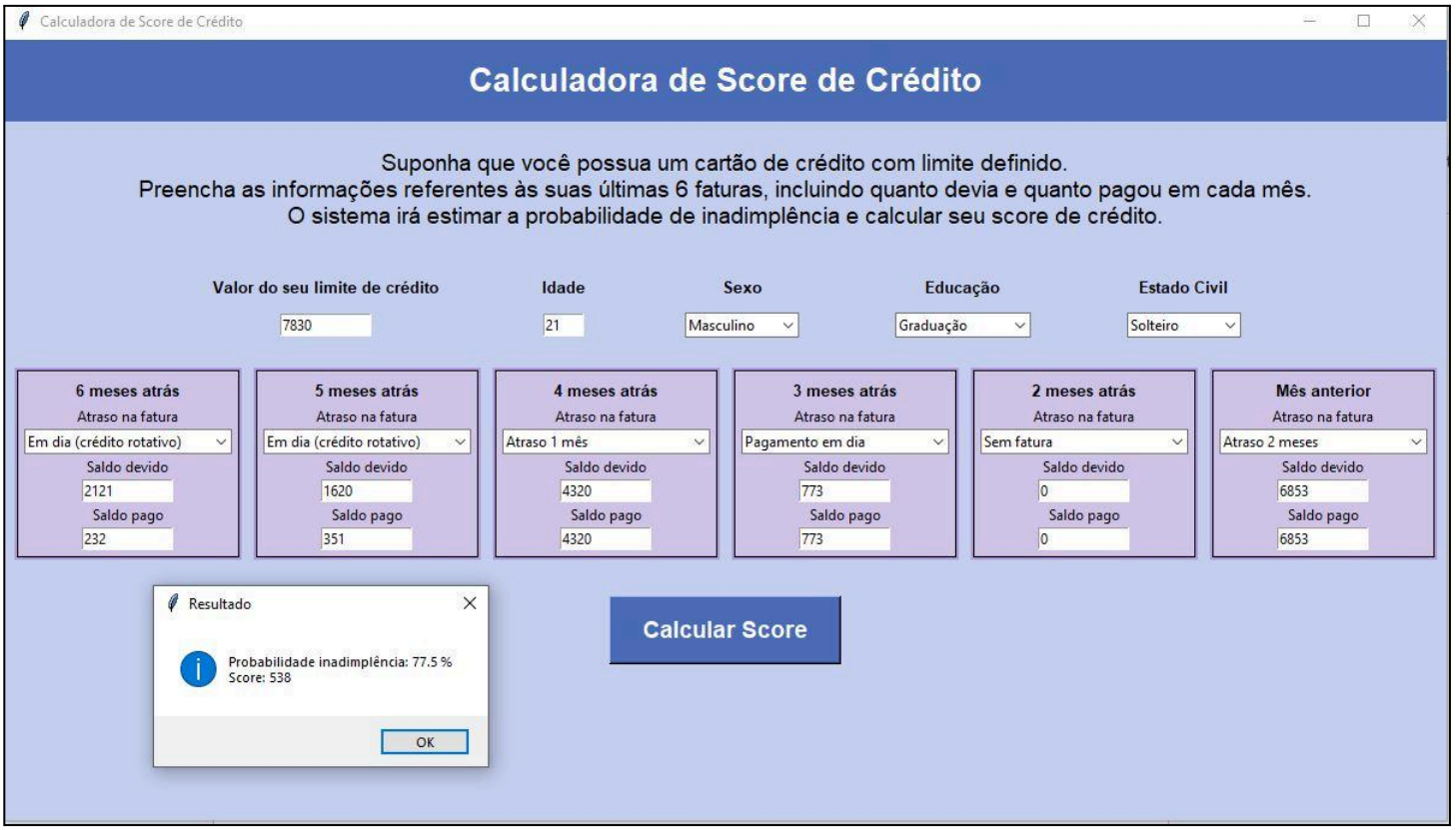

Figura 11 - Interface gráfica com dados preenchidos e apresentando o resultado.

Fonte: Elaborado pelo autor.

7. CONCLUSÕES

O presente projeto demonstrou que é possível prever o risco de inadimplência de clientes de cartão de crédito com um nível satisfatório de acurácia, utilizando regressão logística como modelo base. A partir da análise exploratória, identificou-se que o histórico recente de pagamentos é o principal indicador de risco, seguido por características socioeconômicas e pelo limite de crédito concedido, confirmando padrões reconhecidos no mercado financeiro.

O modelo desenvolvido apresentou um AUC-ROC de 0,748, desempenho compatível com benchmarks acadêmicos para este conjunto de dados, demonstrando capacidade consistente de discriminação entre clientes adimplentes e inadimplentes.

De forma geral, o trabalho atingiu seus objetivos, validando o uso de modelos estatísticos interpretáveis como ferramenta de apoio à concessão de crédito, contribuindo para evitar perdas financeiras e otimizar políticas de aprovação.

8. LIMITAÇÕES E TRABALHOS FUTUROS

Apesar da relevância dos resultados obtidos e da validação prática do modelo, **algumas limitações devem ser consideradas** antes de sua aplicação em escala real. A base de dados utilizada neste projeto, embora amplamente reconhecida na literatura, é proveniente de um **banco de Taiwan e reflete um contexto econômico, comportamental e regulatório distinto do mercado brasileiro. Além disso, trata-se de um conjunto de dados relativamente antigo**, o que pode não representar padrões mais atuais de consumo e endividamento.

Outra limitação está relacionada ao escopo das variáveis disponíveis. Informações cruciais para análise de crédito, como histórico completo de relacionamento financeiro, renda declarada, score oficial de crédito, dados de transações bancárias e informações de mercado, não foram contempladas. Como consequência, o desempenho do modelo pode ser afetado pela ausência de atributos que descrevem com maior precisão a capacidade e o comportamento de pagamento dos clientes.

Do ponto de vista estatístico, mesmo com o balanceamento da variável-alvo, **ainda há uma proporção relevante de falsos positivos e falsos negativos na classificação**, o que indica a necessidade de ajustes mais específicos ao contexto de negócio. Em aplicações reais, cada erro de classificação possui impacto financeiro distinto: aprovar um cliente inadimplente gera prejuízo direto, enquanto negar um cliente adimplente representa perda de receita e possível desgaste na experiência do cliente. Assim, **a definição do threshold ideal deve considerar métricas econômicas, e não apenas estatísticas.**

Com base nesses pontos, **algumas oportunidades de evolução** são sugeridas para trabalhos futuros:

- Testar modelos mais complexos e não lineares, como Random Forest, XGBoost e LightGBM, que podem capturar relações mais profundas entre as variáveis
- Incorporar novas fontes de dados, especialmente informações comportamentais e do sistema de crédito nacional, ampliando o poder preditivo
- Realizar análise de custo-benefício baseada em métricas de risco, como expected loss e retorno da carteira de crédito