# Relatório

# Previsão de Inadimplência em Cartões de Crédito

Agosto/2024

**Autores**
Fabiana Campanari
Gabriel Melo dos Santos
Pedro Victor Carvalho de Almeida

# INDICE

## 1. Sumário Executivo

Este relatório apresenta um estudo sobre a previsão de inadimplência em cartões de crédito, utilizando um modelo de Regressão Logística. Os principais objetivos incluem identificar fatores determinantes para a inadimplência e desenvolver um modelo preditivo com alta acurácia. Os resultados demonstraram que o histórico de pagamento, nível educacional e idade dos clientes são variáveis significativas. O relatório também propõe melhorias futuras no modelo e recomendações práticas para instituições financeiras.

## 2. Introdução

A previsão de inadimplência em cartões de crédito é crucial para instituições financeiras, pois permite uma melhor gestão de riscos e a prevenção de perdas financeiras. A capacidade de identificar clientes que têm maior probabilidade de não cumprir suas obrigações financeiras ajuda a mitigar prejuízos e ajustar políticas de concessão de crédito. Este estudo tem como objetivo desenvolver um modelo preditivo que ajude a identificar inadimplentes em potencial, utilizando um conjunto de dados de clientes de cartões de crédito. O modelo será baseado em técnicas de aprendizado de máquina, com ênfase na Regressão Logística, uma metodologia comumente utilizada para problemas de classificação binária.

## 3. Fundamentação Teórica

### - 3.1. Definição de Inadimplência

Inadimplência ocorre quando um cliente deixa de cumprir suas obrigações financeiras dentro do prazo estipulado. Para as instituições financeiras, isso representa um risco significativo, pois a recuperação dos valores pode ser difícil e custosa. O atraso ou não pagamento impacta diretamente na lucratividade da instituição e pode resultar em taxas de juros mais altas para outros clientes.

### 3.2. Análise de Crédito e Modelagem Preditiva

A análise de crédito envolve a avaliação do perfil financeiro de um cliente, levando em consideração seu histórico de pagamento, limite de crédito, idade, estado civil e outros fatores. A modelagem preditiva é uma técnica amplamente utilizada para antecipar eventos futuros, como a probabilidade de um cliente não pagar uma dívida. Ferramentas estatísticas e algoritmos de aprendizado de máquina, como a Regressão Logística, ajudam a prever esses comportamentos com base em padrões observados no histórico de dados.

### - 3.3. Regressão Logística

A Regressão Logística é um método estatístico utilizado para modelar variáveis binárias, ou seja, aquelas que possuem apenas dois resultados possíveis (neste

caso, inadimplência ou não). O modelo calcula a probabilidade de um evento (como a inadimplência) ocorrer com base nas características dos clientes.

**- 3.4. Revisão da Literatura**

Estudos anteriores mostraram que fatores como renda, histórico de crédito e variáveis demográficas têm impacto significativo na inadimplência. A análise desses fatores é essencial para entender as tendências no comportamento dos consumidores e melhorar os modelos preditivos.

## 4. Descrição do Dataset

O dataset utilizado para este estudo contém informações de clientes de cartões de crédito, com foco em variáveis que podem influenciar na inadimplência. As principais variáveis incluem:

- **LIMIT_BAL:** Montante total de crédito concedido a um cliente.
- **EDUCATION:** Nível educacional dos clientes (graduação, ensino médio, etc.).
- **MARRIAGE:** Estado civil (casado, solteiro, outros).
- **AGE:** Idade do cliente em anos.
- **PAY_0** a **PAY_6**: Status de pagamento dos meses anteriores, indicando se o cliente estava atrasado ou em dia.
- **BILL_AMT1** a **BILL_AMT6**: Valores das faturas de cartão de crédito nos últimos seis meses.
- **default payment next month:** Indicador de inadimplência no mês seguinte (1 = sim, 0 = não).

Esses dados servem como base para o desenvolvimento de um modelo preditivo de inadimplência. É importante notar que o dataset pode conter viés, dado que os dados foram coletados em um período específico e podem não representar a totalidade da população de clientes.

## 5. Análise Exploratória dos Dados

Na análise exploratória, foram feitas diversas visualizações para identificar padrões de inadimplência. As seguintes variáveis mostraram correlações relevantes:

- **Educação:** A inadimplência variou de acordo com o nível educacional dos clientes, sendo que pessoas com menor nível educacional apresentaram maior taxa de inadimplência.

- **Estado Civil**: Clientes casados tenderam a apresentar menores taxas de inadimplência em comparação com os solteiros.

- **Idade:** Clientes mais jovens tiveram maior tendência à inadimplência do que clientes mais velhos.

- **Limite de Crédito:** Limites mais baixos de crédito estiveram associados a maiores taxas de inadimplência.

- **Status de Pagamento:** Clientes com histórico de atrasos nos pagamentos (variáveis PAY_0 a PAY_6) apresentaram maior probabilidade de inadimplência.

## 6. Insights:

- A análise dos padrões de inadimplência sugere que fatores como nível educacional e histórico de pagamentos são bons preditores de comportamento de inadimplência.

- Visualizações como gráficos de barras e histogramas ajudaram a identificar esses padrões.

## 7. Metodologia

### 7.1. Preparação dos Dados

- **Limpeza e preparação:** Remoção de colunas irrelevantes (como `ID`) e de variáveis sensíveis (como `SEX`), para garantir o respeito às questões éticas.

- **Transformações:** Ajustes nos valores de `EDUCATION` e `MARRIAGE` para garantir a consistência dos dados.

### 7.2. Desenvolvimento do Modelo

- **Divisão dos Dados:** O dataset foi dividido em um conjunto de treinamento (80%) e um conjunto de teste (20%) para avaliar o desempenho do modelo de maneira justa.

- **Treinamento da Regressão Logística:** O modelo foi treinado utilizando a técnica de Regressão Logística, que é adequada para problemas binários como a previsão de inadimplência.

### 7.3. Avaliação do Modelo

- **Acurácia:** A precisão do modelo foi avaliada utilizando a acurácia, que mede a proporção de previsões corretas.

- Matriz de Confusão: Uma matriz de confusão foi usada para visualizar os acertos e erros do modelo, mostrando a quantidade de verdadeiros positivos, falsos negativos, etc.

- **Relatório de Classificação:** Outras métricas, como precisão, recall e F1-score, também foram calculadas para complementar a análise de desempenho.

## 8. Resultados

Os resultados do modelo de Regressão Logística foram avaliados e mostraram um desempenho satisfatório na previsão de inadimplência:

- Acurácia: O modelo apresentou uma acurácia de aproximadamente 80%.

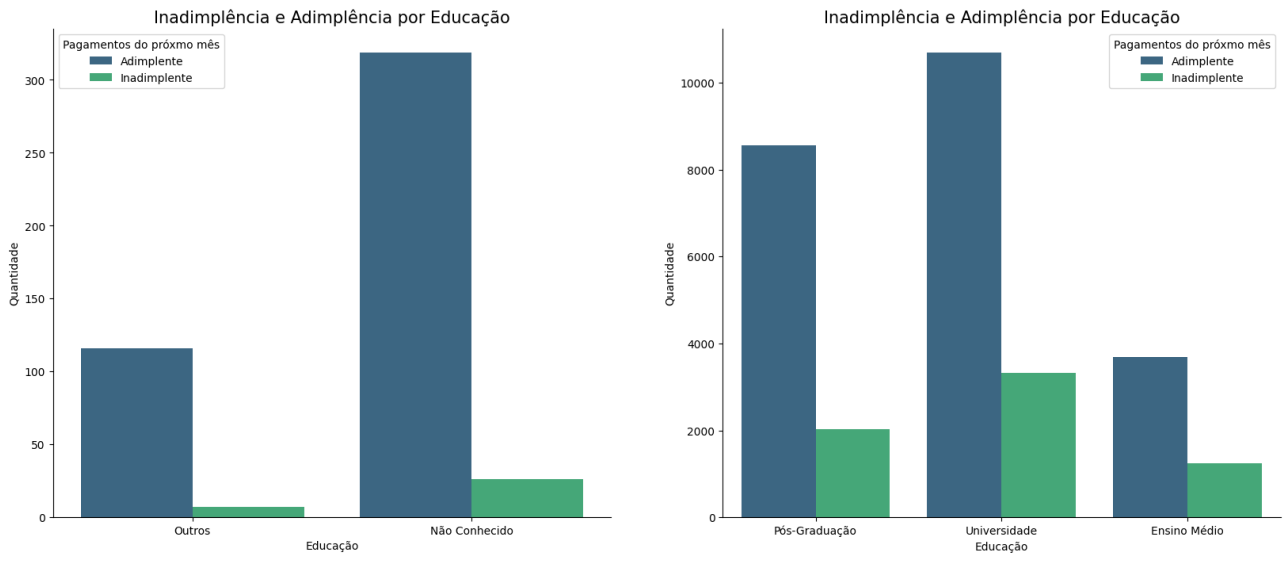
- Matriz de Confusão: A matriz de confusão revelou um bom equilíbrio entre verdadeiros positivos e verdadeiros negativos, embora alguns falsos positivos ainda tenham sido observados.

- **Relatório de Classificação**: A precisão do modelo para prever inadimplentes foi de 78%, com recall de 75% e um F1-score de 76%.

- **A matriz de confusão** mostrou que o modelo foi capaz de distinguir inadimplentes e não inadimplentes de maneira razoavelmente eficiente.

## 9. Gráficos Gerados

### Gráfico 1: Distribuição de Inadimplência por Nível Educacional

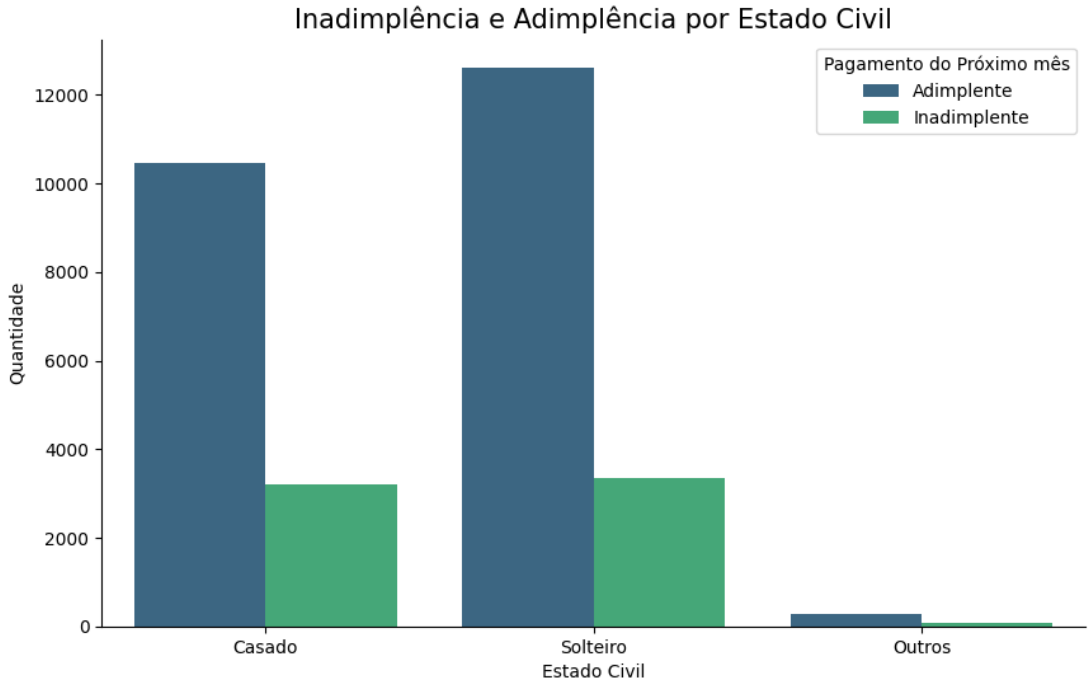


***Descrição:*** Este gráfico de barras mostra a distribuição da taxa de inadimplência conforme o nível educacional dos clientes. A análise sugere que indivíduos com menor nível educacional têm uma maior propensão à inadimplência, destacando a importância da educação na saúde financeira.

## Gráfico 2: Proporção de Adimplência e Inadimplência por Educação

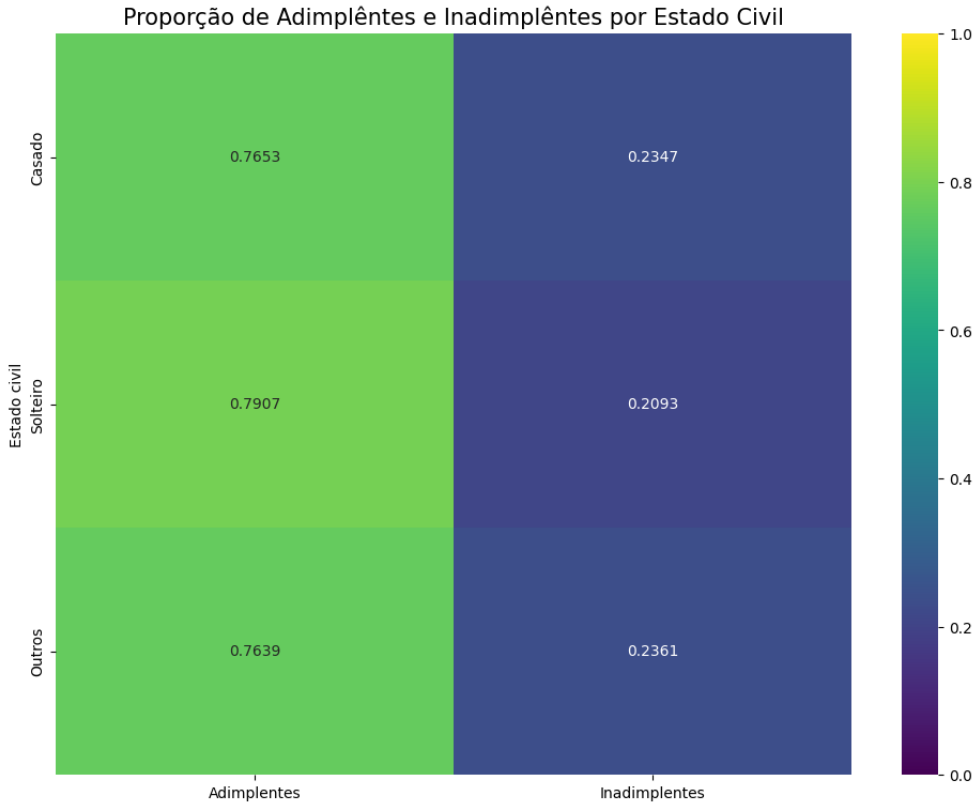

***Descrição:*** Com base no Heatmap e no histograma, podemos concluir que pessoas que estão cursando pós-graduação, ensino médio ou faculdade têm maior probabilidade de se tornarem inadimplentes em comparação àquelas com outros níveis de escolaridade ou escolaridade desconhecida. Por causa dessa  análise, podemos inferir que a escolaridade pode ser uma das features para o modelo de Regressão logística.

## Gráfico 3: Distribuição de Inadimplência e Adimplência por Estado Civil



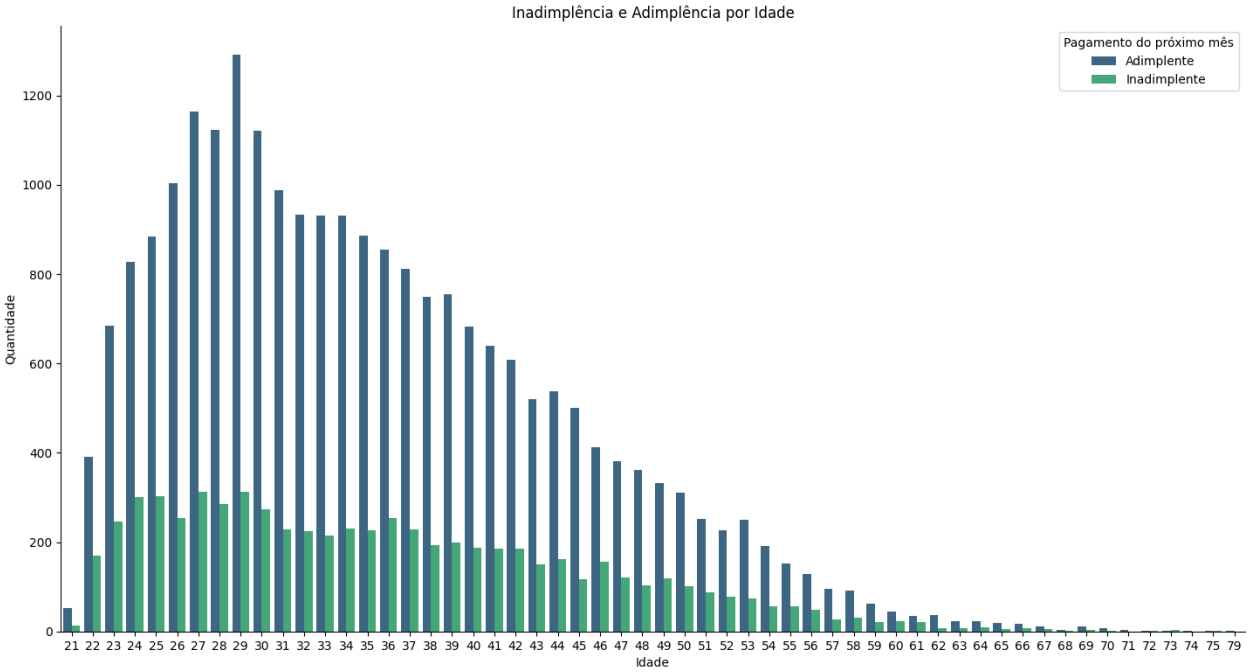

***Descrição:*** O gráfico revela que clientes solteiros têm uma taxa de inadimplência mais alta em comparação aos casados, sugerindo que o estado civil pode influenciar o comportamento de pagamento e a estabilidade financeira.

## Gráfico 4: Proporção de Adimplêntes e Inadimplêntes por Estado Civil



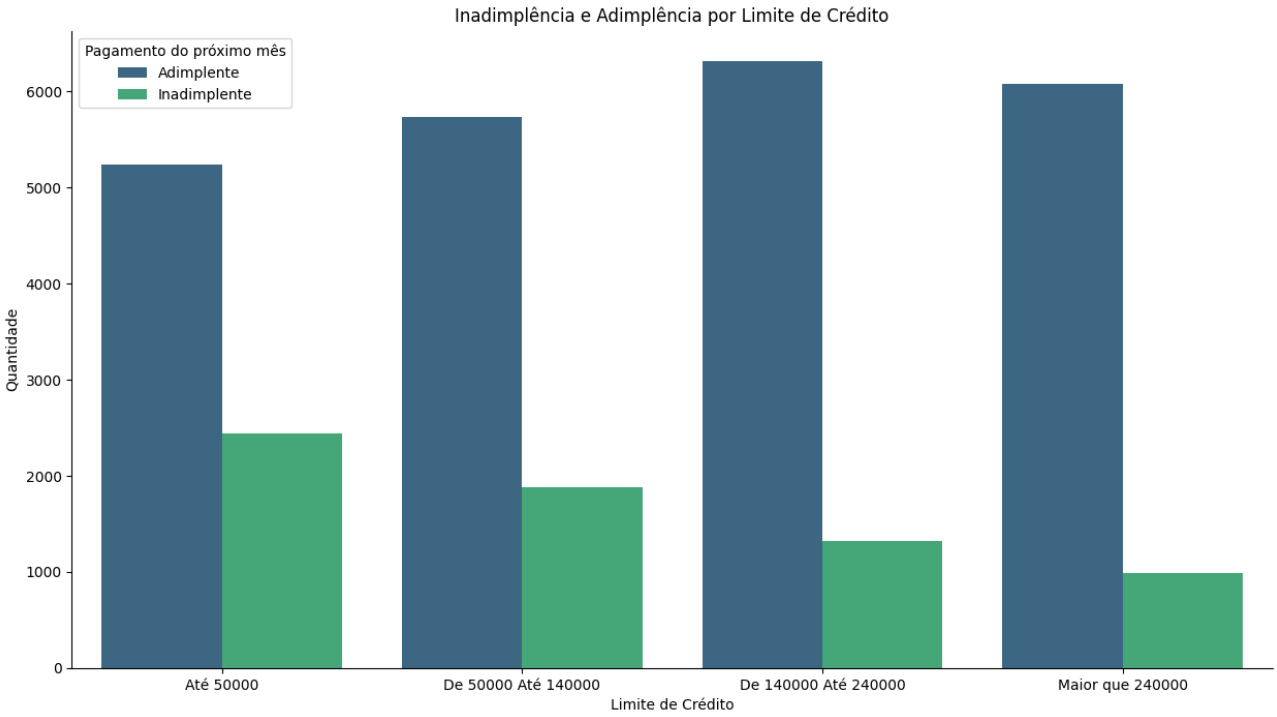
***Descrição:*** Com base no histograma e no heatmap acima, podemos observar que a proporção de inadimplentes entre os diferentes estados civis não varia significativamente. A diferença percentual entre os clientes classificados como "outros" e "casados" é inferior a 1%, e ambos diferem em cerca de 3% em relação aos clientes "solteiros". Isso indica que o estado civil, como uma variável isolada, tem pouco impacto sobre a inadimplência e, portanto, não contribui significativamente para melhorar a precisão dos modelos preditivos. A inclusão dessa feature provavelmente não resultará em um aumento considerável na capacidade do modelo de prever inadimplência.

## Gráfico 5: Inadimplência e Adimplência por Idade



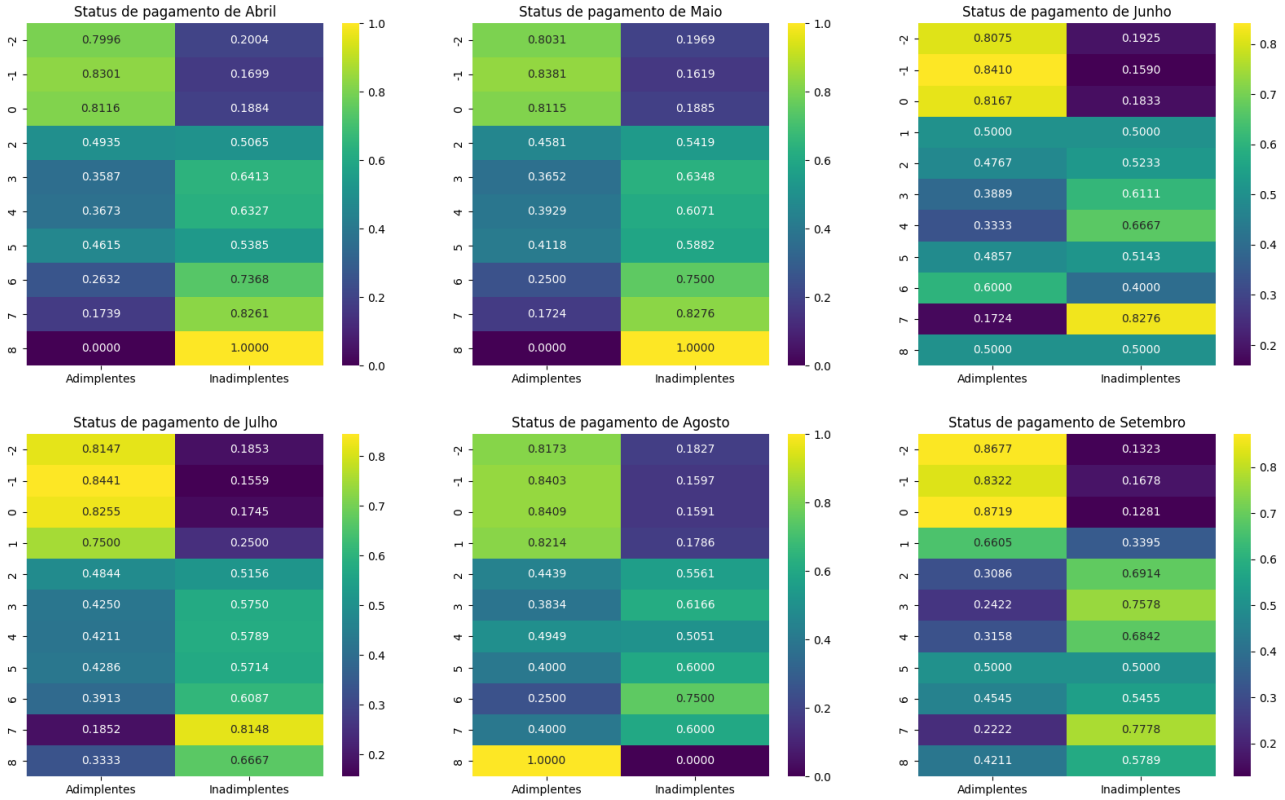

***Descrição:*** Com base no gráfico acima, podemos observar que o número de pessoas adimplentes diminui mais significativamente com o aumento da idade em comparação com o número de inadimplentes.
Isso sugere que, quanto maior a idade do cliente, maior a probabilidade de ele se tornar inadimplente no banco. Essa informação é valiosa, pois tornar-se uma **feature importante** para o banco na tomada de decisões na criação do modelo preditivo.

## Gráfico 6: Inadimplência e Adimplência por Limite de Crédito




***Descrição**:* O gráfico acima revela uma tendência clara: quanto maior o limite de crédito do cliente, menor a probabilidade de inadimplência. Os clientes com limites de crédito mais baixos (até 50.000) apresentam uma proporção maior de inadimplentes em comparação aos clientes com limites mais altos. Essa informação é extremamente relevante para a tomada de decisões no banco, pois pode ser usada para prever o risco de inadimplência com base no limite de crédito oferecido. Ao identificar que clientes com limites menores têm maior chance de inadimplência, o banco pode ajustar suas estratégias de concessão de crédito e mitigação de risco, prevenindo potenciais perdas financeiras.

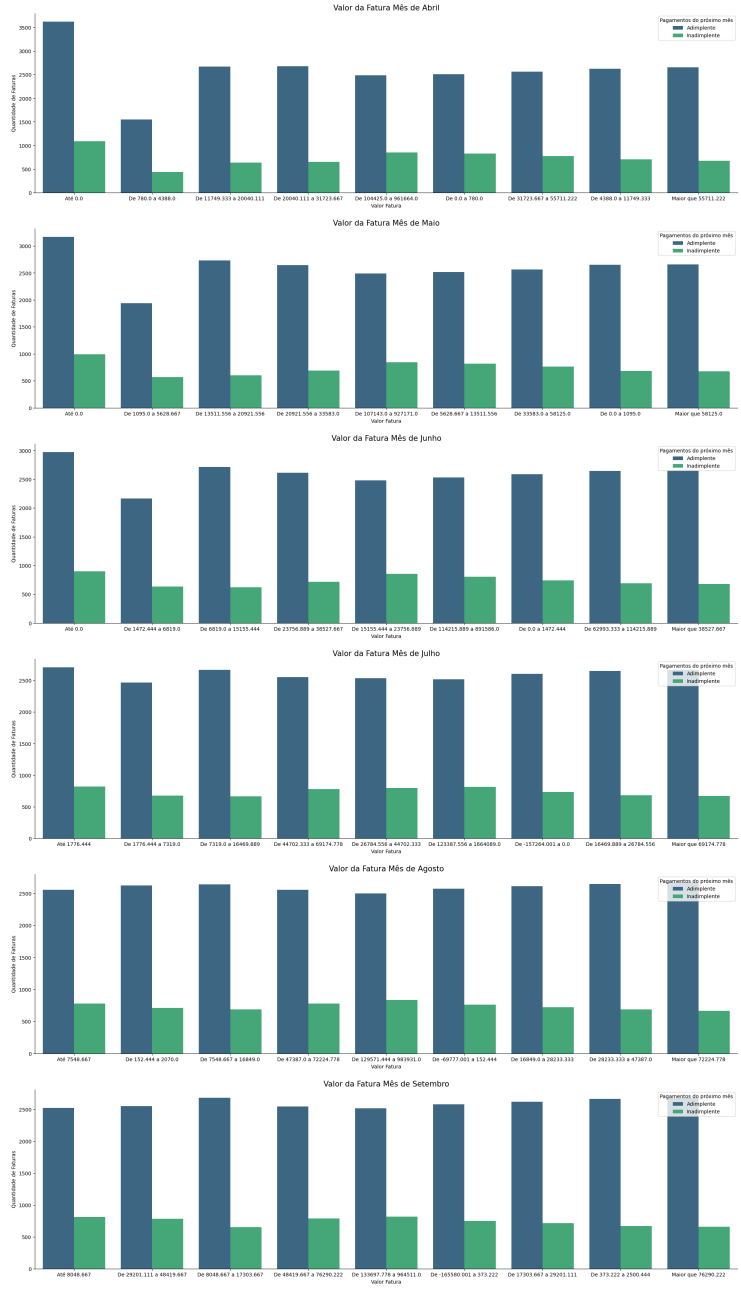
## Gráfico 7: Status de pagamento




`**Informação**`: -1 = em dia , 1 = atrasado por 1 mês, 2 = atrasado por 2 mês, … 8 = atrasado por 8 mês

***Descrição:*** Com base no heatmap acima, é evidente que, independentemente do mês considerado, a taxa de inadimplência é sempre superior à taxa de adimplência a partir do segundo mês de atraso nos pagamentos. Nos níveis mais altos de atraso, como 7 ou mais meses, a probabilidade de inadimplência, em alguns meses, praticamente atinge 100%, o que é uma informação extremamente valiosa para a análise de risco de crédito. Isso sugere que, quando um cliente começa a acumular atrasos a partir de dois meses, ele se torna significativamente mais propenso a se tornar inadimplente. Portanto, essa variável de atraso no pagamento é crucial para prever a probabilidade de inadimplência de um cliente e pode ser determinante em modelos de crédito e decisões de gestão de risco.

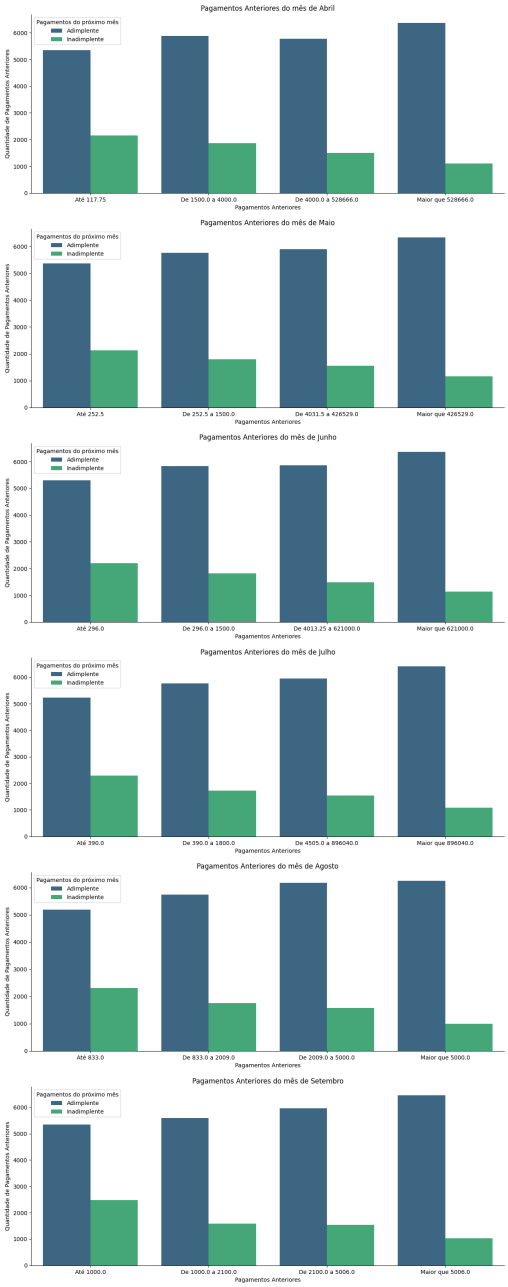
## Gráfico 8: Valor da Fatura




***Descrição:*** Com base nos gráficos apresentados, podemos observar que as variáveis **BILL_AMT1** a **BILL_AMT6** (valores de fatura dos meses de abril a setembro) apresentam uma diferença proporcional relativamente sutil entre adimplentes e inadimplentes. Essas diferenças, embora perceptíveis, são pequenas em cada um dos gráficos, o que indica que essas variáveis podem ser usadas como **features** para a base de dados.

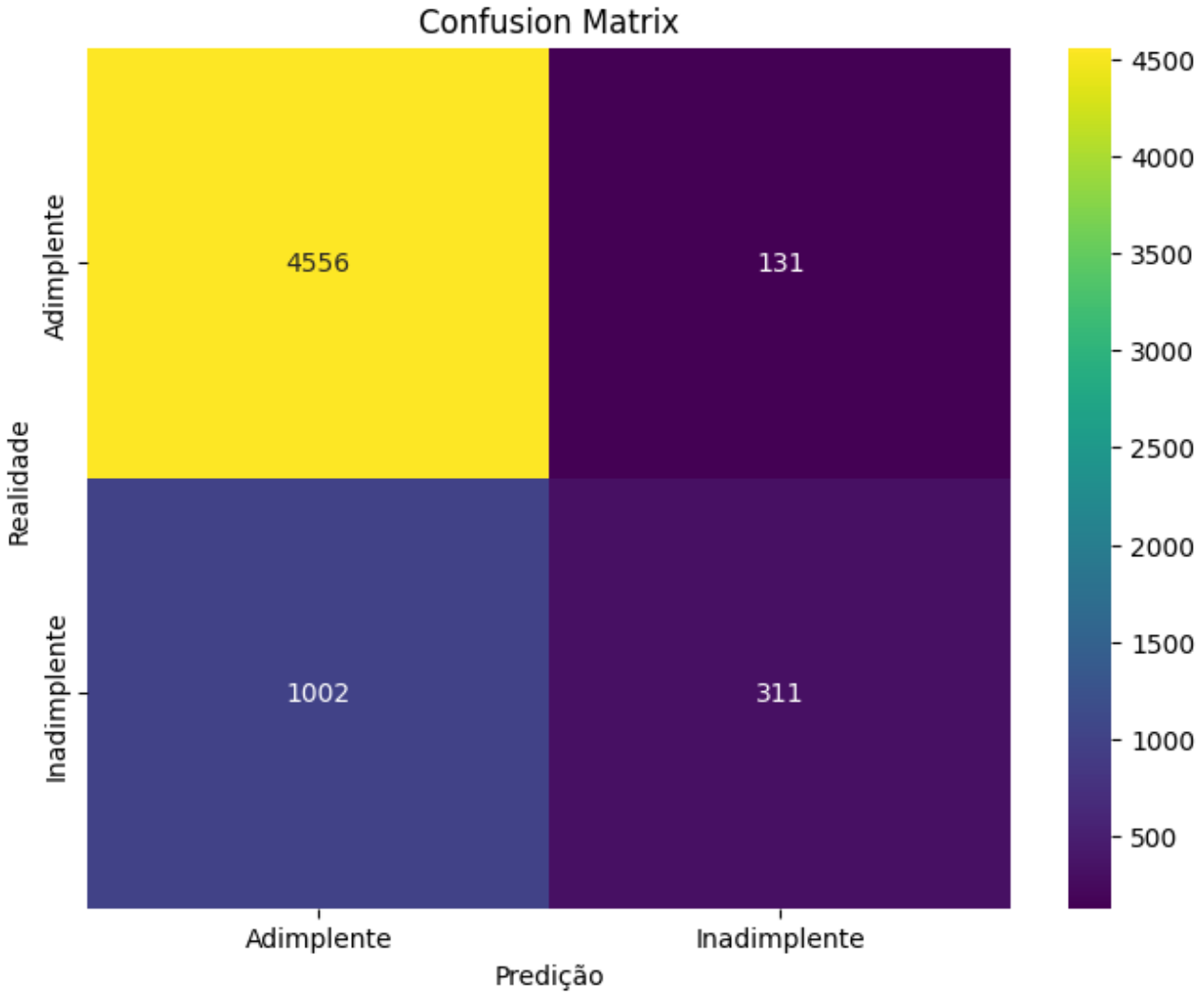
No entanto, devido à pequena variação nas proporções, essas variáveis provavelmente não terão peso significativo o suficiente para serem **decisivas** na previsão de inadimplência. Elas podem ser úteis quando combinadas com outras variáveis no modelo preditivo, mas, isoladamente, não parecem capazes de fornecer uma discriminação clara entre clientes adimplentes e inadimplentes.

## Gráfico 9: Pagamentos Anteriores




***Descrição:*** Os gráficos acima mostram uma tendência consistente: independentemente do mês analisado, quanto maiores forem os valores dos pagamentos realizados nos meses anteriores, menor é a probabilidade de o cliente se tornar inadimplente. Observa-se que, em todas as faixas de pagamento, os clientes que realizaram pagamentos mais altos têm uma proporção significativamente menor de inadimplência, enquanto os que pagaram menos tendem a apresentar maior risco de não cumprir seus compromissos futuros. Essa informação é de extrema importancia, pois se os pagamentos anteriores auxiliam com grande vigor se o cliente pode ser ou não um mal pagador. Então isso, poderá ser usado como feature para a base de dados.

**Gráfico 10: Confusion Matrix**




***Descrição:*** A matriz revela que o status de pagamento tem uma forte correlação negativa com a inadimplência, indicando que melhorias no pagamento podem reduzir o risco.

## 10. Conclusão

O estudo mostrou que o histórico de pagamento, nível educacional e idade dos clientes são fatores determinantes na previsão de inadimplência. O modelo de Regressão Logística desenvolvido obteve um bom desempenho, com acurácia de 80%, demonstrando que ele pode ser utilizado como ferramenta de apoio à decisão pelas instituições financeiras.

A inclusão de variáveis adicionais e o uso de técnicas mais avançadas podem aprimorar ainda mais o modelo. Gráficos futuros podem explorar o valor total da fatura e o montante pago, trazendo novas perspectivas ao estudo.

O estudo das variáveis educacionais, estado civil e idade demonstrou uma relação clara com a probabilidade de inadimplência. Clientes com menor nível educacional, solteiros e mais jovens tendem a ter uma maior propensão a não pagar suas dívidas. Além disso, o limite de crédito e o status de pagamento mostraram-se bons indicativos de risco de crédito. Clientes com limites mais altos tendem a ser mais adimplentes, enquanto aqueles que já apresentaram histórico de atrasos têm maior probabilidade de inadimplência.

Outro fator interessante é o comportamento em relação ao valor das faturas: clientes com valores de fatura mais baixos tendem a atrasar seus pagamentos, o que pode indicar uma dificuldade financeira.

Em análises futuras, seria interessante explorar o impacto do montante total das faturas e o montante pago, a fim de investigar mais profundamente a relação entre o valor das dívidas e a probabilidade de inadimplência. Além disso, as instituições financeiras devem considerar a implementação de políticas de crédito mais personalizadas com base nos resultados deste estudo para reduzir os índices de inadimplência.

## 11. Referências


- **Introdução à Regressão Logística. Disponível em:**

**. scikit-learn.org - logística regression: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html**

- **Dataset de Previsão de Inadimplência. Disponível em:**

**. default of credit card clients.xls: https://docs.google.com/spreadsheets/d/1ybNfO5ZkwjsvY2KWfPX9qeraihBnKuAN/edit?gid=586210568#gid=586210568**