RELATÓRIO FINAL

TEMA: *Previsão de Inadimplência*

INTEGRANTES: *Gabriel da Cunha Teixeira, Micael Ferrari, Matheus Werneck, Pedro Henrique Marazo, Thiago Malaquias*

LINK GITHUB: https://github.com/Micaelferrari/aprendizado-supervisionado

LINK NOTION (Relatório): https://www.notion.so/RELAT-RIO-FINAL-21564dc8807880b69b95c6bb87b6cee8?source=copy_link

Introdução ao problema

O projeto tem como objetivo prever a inadimplência de clientes com base em características financeiras e comportamentais. No entanto a variável de interesse foi o "default payment next month", que indica se um cliente irá se tornar inadimplente ou não no próximo mês. Trata-se, portanto, de um problema de classificação binária em um contexto de crédito e risco financeiro.

A motivação da modelagem, se deve ao fato do possível auxilio a instituições financeiras na tomada de decisão sobre a disponibilização de crédito, com o intuito de prever comportamentos futuros baseados nos dados históricos.

Descrição do dataset

O conjunto de dados é composto por 30.000 registros de clientes, com 23 variáveis explicativas e uma variável-alvo (default payment next month) (uma variável binária, com desbalanceamento: aproximadamente 78% dos clientes são adimplentes e 22% inadimplentes). As variáveis abrangem aspectos sociodemográficos (gênero, idade, estado civil, escolaridade), limites de crédito, histórico de pagamento mensal, faturas e pagamentos realizados entre abril e setembro de 2005.

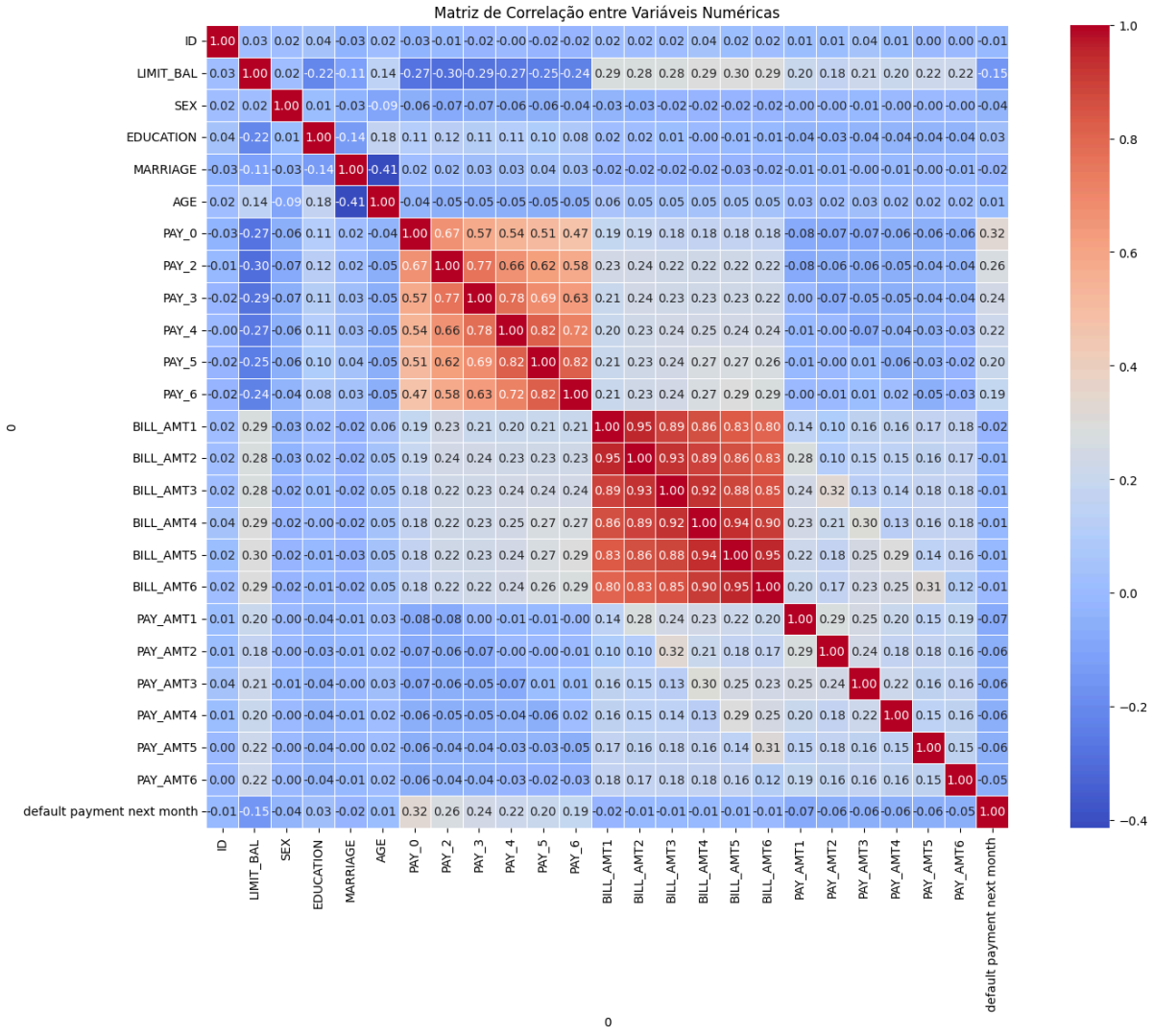
Principais colunas:

- **X1**: Valor do crédito concedido.

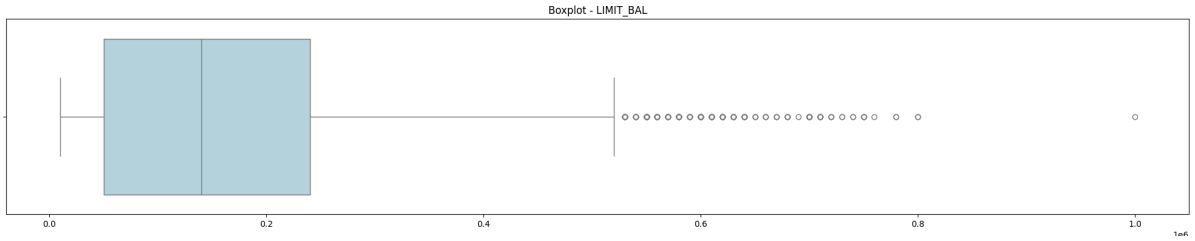
- **X2-X5**: Perfil do cliente (gênero **(X2)**, escolaridade **(X3),** estado civil **(X4)** e idade **(X5)**).
- **X6-X11**: Histórico de pagamento (atrasos mensais entre abril e setembro/2005).
- **X12-X17**: Valor das faturas mensais.
- **X18-X23**: Valor dos pagamentos efetuados (setembro - abril).

EDA e preparação dos dados

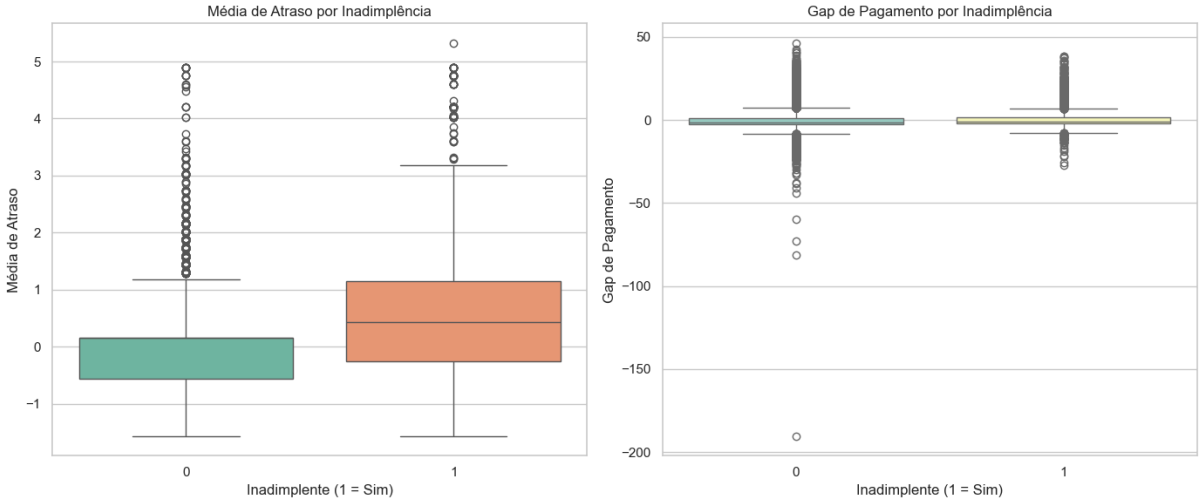
Análise Exploratória (EDA)



- A matriz de correlação revelou forte correlação entre os valores de fatura e pagamentos ao longo dos meses, evidenciando a estabilidade do comportamento financeiro mensal.
- A idade e o limite de crédito apresentaram correlações fracas com a inadimplência.



- A presença de outliers foi confirmada em variáveis como limite de crédito e valor de fatura. Embora identificados, optou-se por manter os outliers, dada a natureza dos dados financeiros reais.



- Foram criadas duas novas variáveis derivadas:
  - **Média de Atraso**: média dos valores de atraso (X6 a X11).
  - **Gap de Pagamento**: diferença entre fatura e valor pago (X12–X17 menos X18–X23).

Tratamento e Engenharia de Dados

- Após a análise dos outliers, eles foram mantidos, mesmo apresentando comportamentos financeiros extremos, eles eram reais, o que poderia interferir na distribuição natural dos dados.
- Tratamento de valores nulos estavam ausentes no dataset.
- Criação de Novas Features (Média de Atraso - Gap de Pagamento).
- As variáveis numéricas foram padronizadas para otimizar o desempenho da RNA.
- A base foi balanceada com **SMOTE**, equilibrando a proporção entre inadimplentes e adimplentes para evitar viés nos modelos preditivos.
- Codificação de Variáveis Categóricas, como “SEX”, “EDUCATION”, “MARRIAGE”.

Descrição dos modelos implementados

Random Forest

Foi treinado um modelo de Random Forest com validação em conjunto de teste. A escolha se deu pela robustez da técnica frente a variáveis correlacionadas e sua boa performance em classificação. O modelo também possibilitou análise da importância das variáveis, destacando o histórico de pagamento e o gap de pagamento como principais preditores.

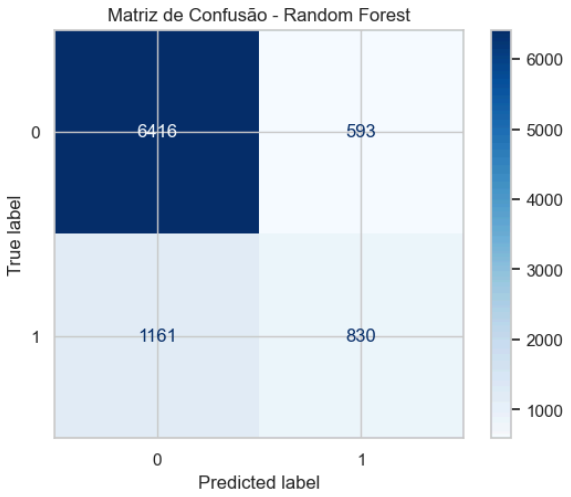
Rede Neural Artificial (RNA)

A RNA foi construída com múltiplas camadas ocultas e função de ativação ReLU, com saída sigmoide para classificação binária. Houve otimização por meio de regularização e escolha de número adequado de épocas. A normalização dos dados e o balanceamento com SMOTE foram fundamentais para sua performance.

Resultados e comparação entre modelos

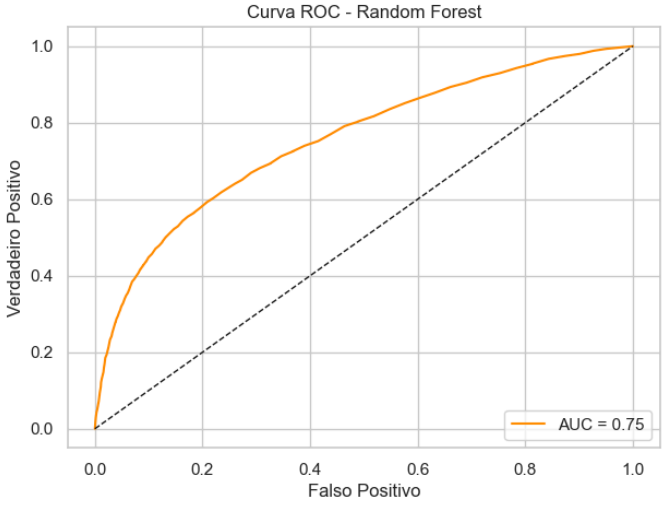
Random Forest

O desempenho do modelo Random Forest foi avaliado com base em diferentes critérios:



Matriz de Confusão:

A matriz mostra que o modelo foi capaz de identificar corretamente 6.416 clientes adimplentes e 830 inadimplentes. No entanto, ainda houve 1.161 inadimplentes classificados incorretamente como adimplentes, evidenciando um certo desafio em capturar corretamente a classe positiva.

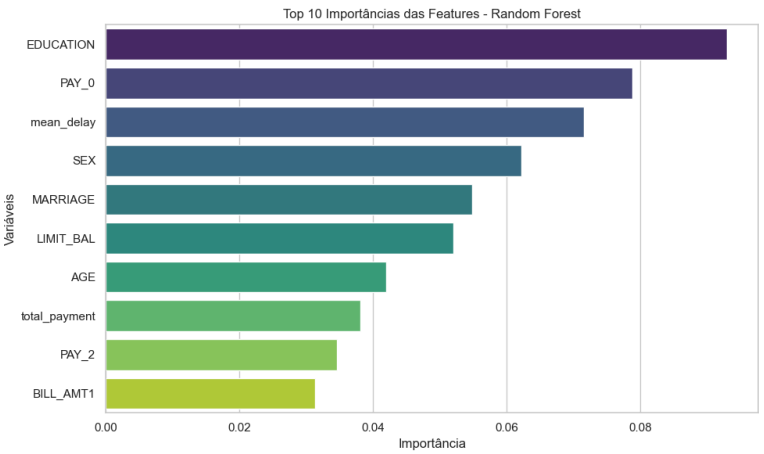


Curva ROC e AUC:

A

Área sob a Curva ROC (AUC) de 0.75 indica que o modelo possui boa capacidade discriminativa, sendo capaz de separar, em média, 75% das vezes

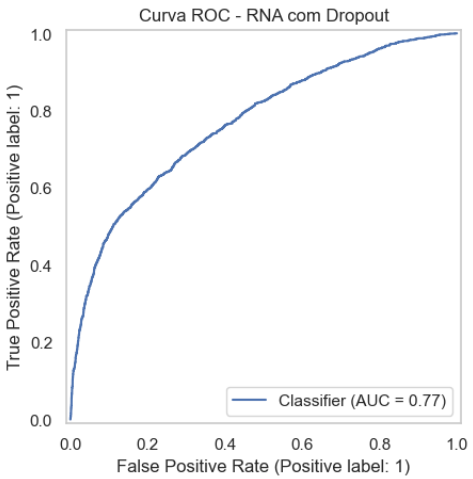
um inadimplente de um adimplente. Isso reforça sua utilidade prática em cenários de risco de crédito.



Importância das Variáveis:
A análise das variáveis mais relevantes destaca que fatores como nível educacional, histórico de pagamentos, sexo e limite de crédito têm grande influência na previsão da inadimplência. Isso sugere que o modelo capturou bem os padrões subjacentes ao comportamento de crédito.

Rede Neural Artificial (RNA)

A RNA apresentou um desempenho competitivo, com algumas diferenças importantes:

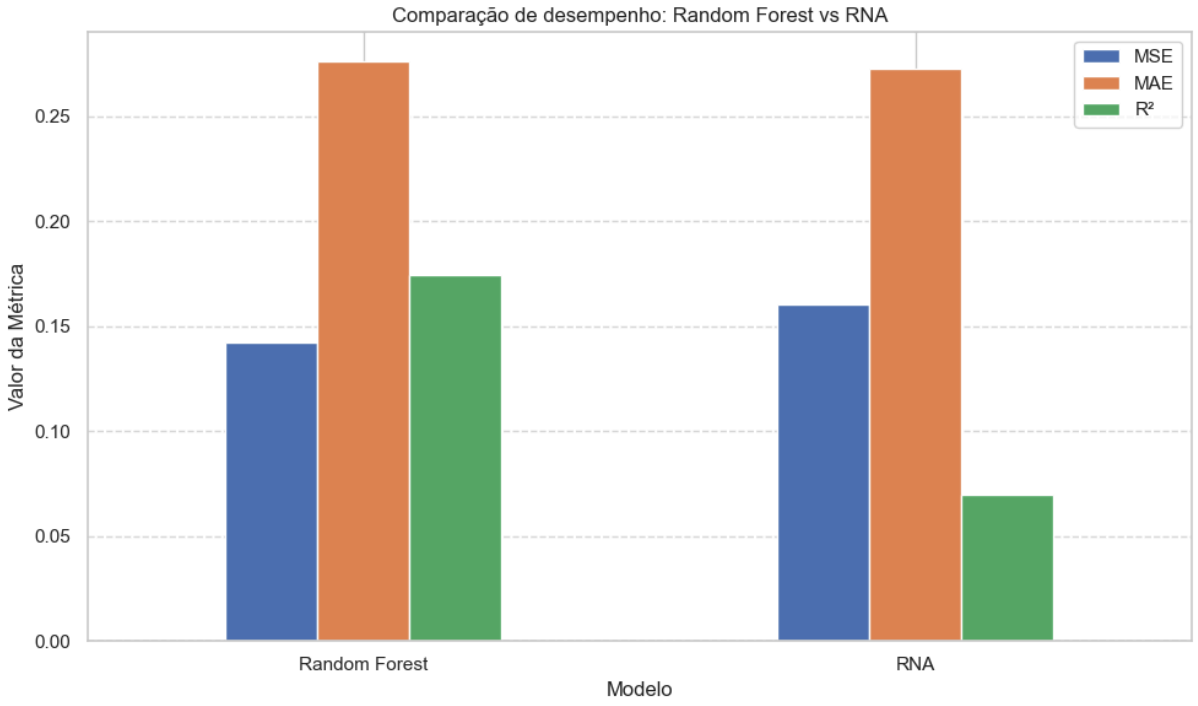


Curva ROC e AUC:
O modelo alcançou uma
AUC de 0.77, levemente superior ao Random Forest, o que indica melhor capacidade de separação entre as classes. Esse ganho pode ser atribuído à habilidade da rede em capturar relações não lineares e interações mais complexas entre as variáveis.

Capacidade de Generalização:

A adoção do Dropout ajudou a conter o overfitting, tornando a RNA mais generalizável, mesmo com maior complexidade em sua arquitetura.

Comparação:



- A RNA obteve **melhor desempenho geral**, principalmente em métricas voltadas à classe positiva (inadimplente).
- A **Random Forest** se destacou em **facilidade de interpretação**, fundamental para áreas que demandam justificativas técnicas, como bancos e crédito.

- A principal dificuldade da RNA foi o **tempo de ajuste de hiperparâmetros** e sensibilidade à normalização.
- Em ambos os modelos, as variáveis mais relevantes foram aquelas ligadas ao comportamento de pagamento recente e à diferença entre fatura e pagamento validando a importância dessas variáveis na identificação de risco.

Conclusões finais com aprendizados do grupo

O projeto mostrou que o histórico de pagamento foi o fator mais relevante na previsão da inadimplência, enquanto idade e limite de crédito tiveram baixa correlação. A Random Forest apresentou melhor desempenho e interpretabilidade em comparação à Rede Neural, que, apesar de mais complexa, demonstrou bom potencial de generalização, exigindo maior cuidado no ajuste e pré-processamento.

Insights:

- Atrasos consecutivos nos pagamentos anteriores são fortes indicativos de inadimplência futura.
- Ao contrário do que se pensa, idade não é um bom indicativo.
- O balanceamento da base com SMOTE foi essencial para mitigar o viés do modelo em prever apenas a classe majoritária (clientes adimplentes).