

Relatório Técnico: Detecção de Fraudes em Cartões de Crédito

Breno Moreira

21 de novembro de 2025

1 Introdução e Objetivo

Este relatório descreve o processo de pré-processamento e modelagem realizado na base de dados *Credit Card Fraud Detection*. O objetivo principal foi tratar os dados para permitir a criação de um modelo preditivo capaz de identificar transações fraudulentas.

A base de dados apresenta desafios significativos, sendo o principal deles o extremo desbalanceamento entre as classes (fraudes representam apenas 0,17% do total). Além disso, as variáveis preditoras (V1 a V28) são resultado de uma transformação PCA (Análise de Componentes Principais), restando apenas *Time* e *Amount* em suas escalas originais.

2 Etapas de Pré-processamento

2.1 1. Tratamento de Dados Iniciais

A verificação inicial não apontou valores ausentes (nulos), dispensando técnicas de imputação. No entanto, foi identificada a presença de redundância nos dados.

- **Ação:** Foram removidas **1.081 linhas duplicadas** para evitar que o modelo enviesasse o aprendizado com exemplos repetidos.

2.2 2. Análise Exploratória e Desbalanceamento

A visualização da variável alvo (*Class*) confirmou o desbalanceamento severo. Se um modelo fosse treinado sem tratamento, ele tenderia a classificar todas as transações como normais, atingindo alta acurácia, mas falhando no objetivo de detectar fraudes (baixo Recall).

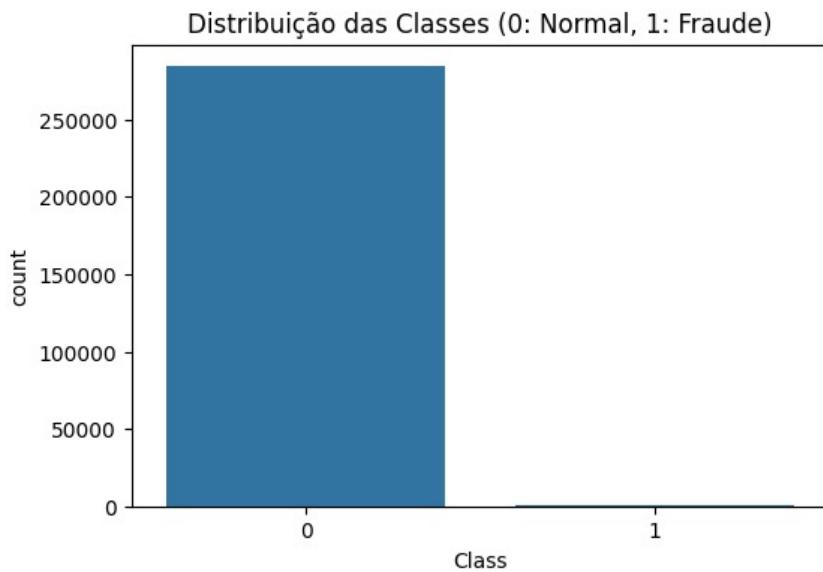


Figura 1: Distribuição das Classes. A classe 1 (Fraude) é quase invisível.

2.3 3. Normalização e Tratamento de Outliers

Como as variáveis V1-V28 já estavam padronizadas pelo PCA, o foco foi nas variáveis *Time* e *Amount*.

- **Método:** Utilizou-se o **RobustScaler**.
- **Justificativa:** Diferente do *StandardScaler* (que usa a média), o *RobustScaler* utiliza a mediana e o intervalo interquartil. Isso é crucial neste dataset, pois valores de transações fraudulentas podem ser extremos (outliers), e o uso da média distorceria a escala.

2.4 4. Análise de Correlação

A matriz de correlação (Heatmap) evidenciou que não existe multicolinearidade entre as variáveis V1-V28 (correlação próxima a zero), o que confirma a ortogonalidade do PCA.

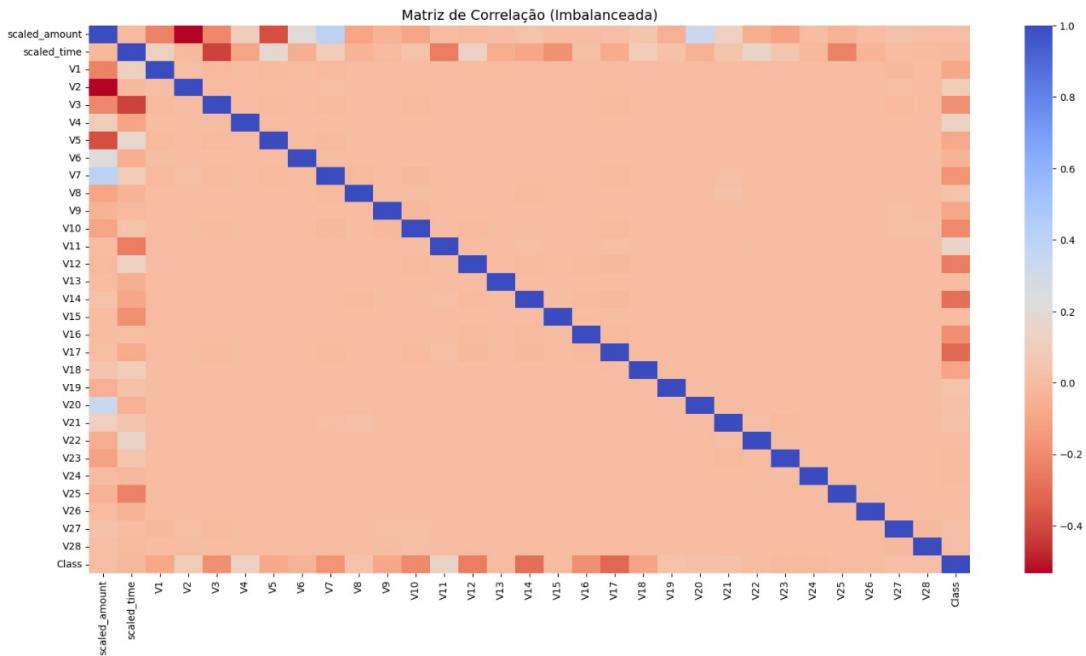


Figura 2: Matriz de Correlação das variáveis.

3 Estratégia de Modelagem e Balanceamento

Um ponto crítico do trabalho foi o momento do balanceamento.

- **Divisão Treino/Teste:** Realizada de forma *estratificada* (mantendo a proporção de 0.17% de fraudes em ambos os conjuntos).
- **SMOTE (Synthetic Minority Over-sampling Technique):** Aplicado exclusivamente nos dados de treino.

Importante: O balanceamento não foi aplicado nos dados de teste para garantir uma avaliação realista do modelo em um cenário do mundo real.

4 Resultados e Comparaçāo (Antes x Depois)

Para validar o impacto do pré-processamento, comparamos um modelo de Regressão Logística treinado com dados originais versus dados tratados com SMOTE.

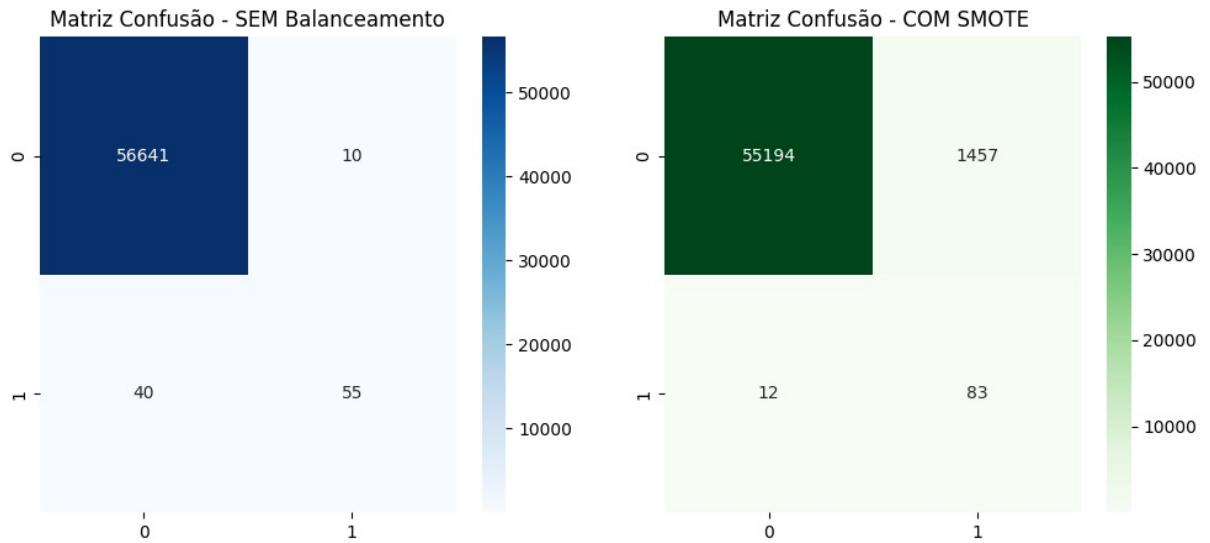


Figura 3: Matriz de Confusão: Sem Balanceamento (Esq.) vs Com SMOTE (Dir.)

4.1 Análise dos Resultados

- **Antes (Sem SMOTE):** O modelo obteve um *Recall* de apenas 0.58 para a classe de fraude. Isso significa que ele deixou passar 42% das fraudes (Falsos Negativos altos).
- **Depois (Com SMOTE):** O *Recall* subiu para **0.87**. O modelo tornou-se capaz de identificar a grande maioria das ações fraudulentas.
- **Trade-off:** Houve um aumento nos Falsos Positivos (alarmes falsos), o que é um custo aceitável em segurança bancária em troca de bloquear as fraudes reais.

5 Link para o Código

O código completo desenvolvido pode ser acessado no link abaixo:

<https://github.com;brnSalamandra/etapas-de-pr--processamento>