Detecção de anomalias: fraudes em transações com cartão de crédito

Autor: Paulo Pinheiro
Data: 18/03/2020

Photo by Ales Nesetril & Fábio Lucas on Unsplash (Edited)

Introdução

Hoje em dia, o cartão de crédito se tornou o principal meio de pagamento pelos consumidores de todo o mundo. A facilidade com que uma compra em uma loja física ou online é efetuada e a segurança de não levar dinheiro físico na carteira é uma das principais vantagens.

Com o número crescente de pagamentos via cartão de crédito, aumenta o interesse de bandidos em querer fraudar o sistema e obter vantagens ilegais. As fraudes vão de uma simples troca de máquina em uma loja a sistemas robustos nas transações online.

É preciso um sistema de segurança rígido e detecção rápida que consiga minimizar ou mitigar essas operações fraudulentas.

Vamos construir um modelo de Machine Learning que nos possibilitará prever o quão assertivo o modelo é na detecção de transações com o cartão de crédito. Tais transações podem ou não ser fraudulentas.

Com o exemplo dos dados mostrados abaixo, podemos notar que as variáveis estão anonimizadas (por questões de segurança) e em, em sua maioria, na mesma escala de grandeza. Isso significa que não sabemos o que realmente significam. Mas isso não é impeditivo para a modelagem.

...	V21	V22	V23	V24	V25	V26	V27	V28	Amount	Class
...	-0.018307	0.277838	-0.110474	0.066928	0.128539	-0.189115	0.133558	-0.021053	149.62	0
...	-0.225775	-0.638672	0.101288	-0.339846	0.167170	0.125895	-0.008983	0.014724	2.69	0
...	0.247998	0.771679	0.909412	-0.689281	-0.327642	-0.139097	-0.055353	-0.059752	378.66	0

Tabela 1. Exemplo do conjunto de dados.

Objetivo

- Explorar alguns aspectos dos dados disponibilizados;
- Balancear os dados para termos a mesma quantidade de output na variável `[Class]` que nos diz se a transação é fraudulenta (1) ou não (0);
- Criar modelos de machine learning para classificar futuras transações;
- Comparar os modelos e escolher o melhor classificador para este conjunto de dados.

Observações iniciais sobre o conjunto de dados

- Foi observado que existem 1081 linhas duplicadas (informação redundante). Tais linhas foram removidas;
- Temos 31 variáveis e 283726 linhas após a remoção das linhas duplicadas; e
- Não existem valores faltantes;

Análise Exploratória dos Dados

1. Variável Amount

a. A variável encontra-se distorcida (skewed);
   i. Existem muitas transações com valores muito além do normal (podemos enxergar isso nos pontos fora da escala no *Gráfico 2)*; e
   ii. Temos um possível indicativo de operações fraudulentas (mas não é a única razão - temos que analisar as demais variáveis).
b. A maior transação foi de $25.691,00; e
c. A média é de $88,47 em transações.

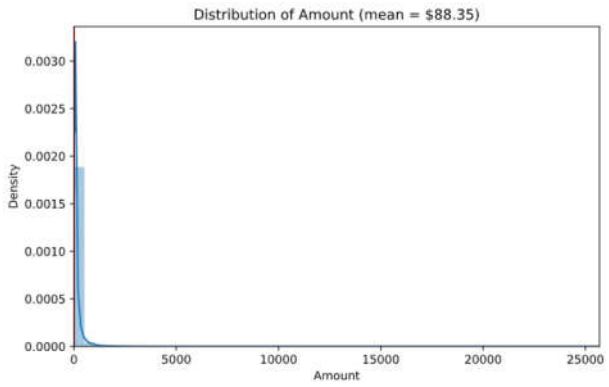

Gráfico 1. Distribuição dos valores transacionados.

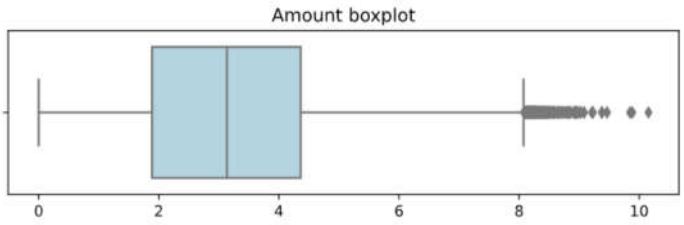

Gráfico 2. Presença de possíveis transações fraudulentas (pontos escuros)

Stats	Values
mean	88.472687
std	250.399437
min	0.000000
75%	77.510000
max	25691.160000

Tabela 2. Estatística básica da variável Amount

2. Variável Class

a. Variável alvo indicando se uma operação é fraudulenta (1) ou não (0);
b. Temos dados muito desbalanceados *(Gráfico 3)* onde 99,8% correspondem a transações normais e 0,17% a fraudulentas (o que é algo comum no dia a dia); e
c. Para uma boa modelagem, precisamos que os dados estejam balanceados *(Gráfico 4)*. Dessa forma diminuiremos o viés de respostas falsas negativas.

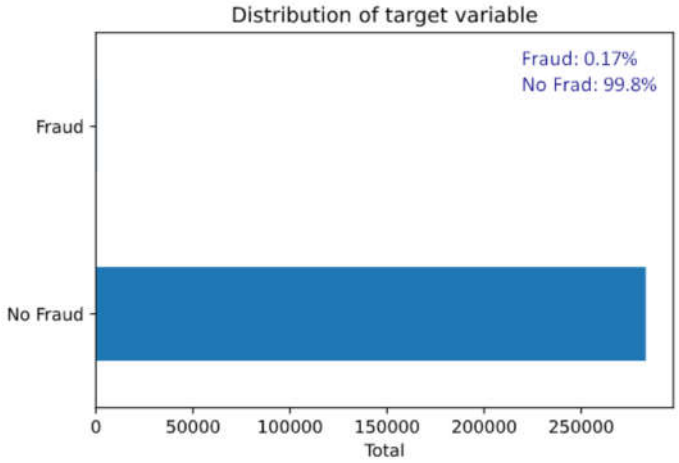

Gráfico 3. Distribuição do número de fraudes e não fraudes.

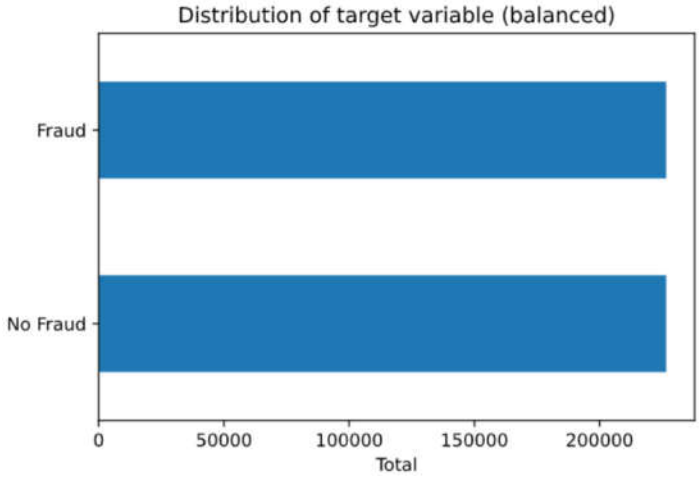

Gráfico 4. Distribuição balanceada do número de fraudes e não fraudes

3. Matriz de Correlação

Queremos saber se as variáveis têm uma forte influência em transações fraudulentas ou não. De uma amostra dos dados, temos a seguinte correlação com a variável alvo ***[Class]*** como exemplos:

- Temos V10, V12 e V14 com correlações negativas *(quanto menor é o valor, maior a probabilidade do resultado ser uma operação fraudulenta);*
- Temos V2, V4 e V11 com correlações positivas *(quanto maior é o valor, maior é a probabilidade do resultado ser uma operação fraudulenta)*.
- Nas correlações positivas, a média das transações fraudulentas são maiores que as não fraudulentas; nas negativas, o inverso.

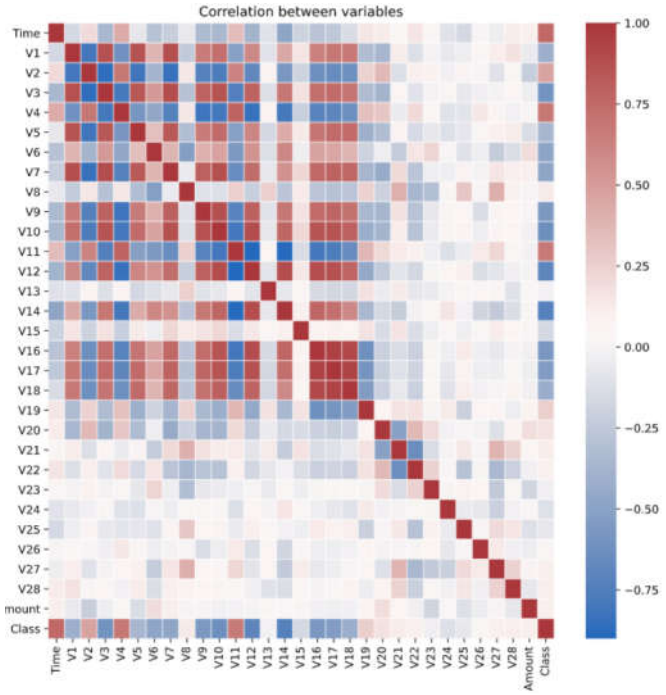

Gráfico 5. Matriz de correlação (notadamente com a variável Class)

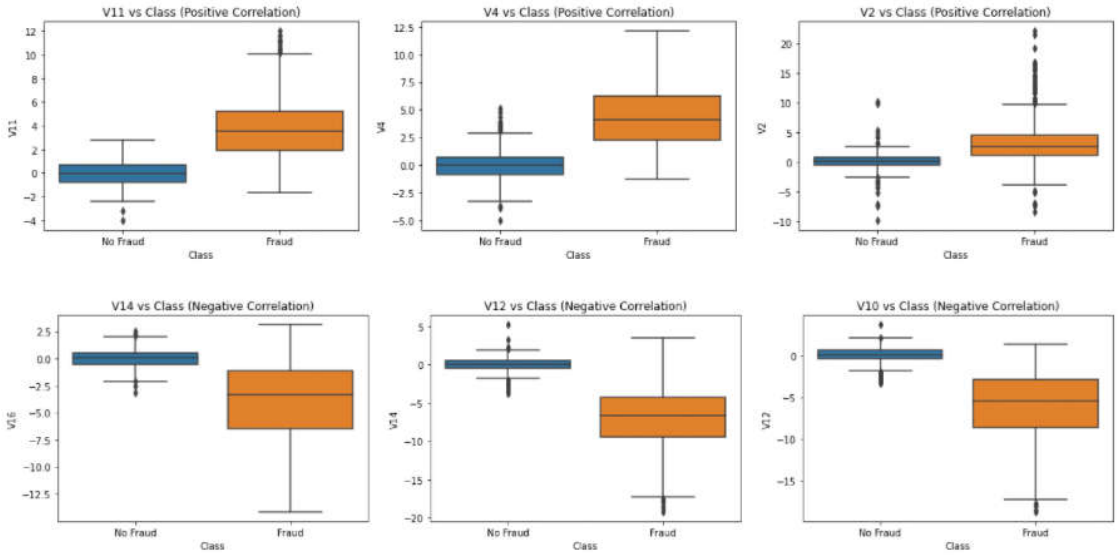

Gráfico 6. Plotagem estatística dos exemplos de correlação

4. Modelagem de Machine Learning

a. Utilizamos três modelos de machine learning: Logistic Regression, Decision Tree e Random Forest;
b. De acordo com o resultado das métricas, e com a proposta do problema de negócio, Random Forest possui a melhor performance na detecção de anomalias (fraudes) em cartões de crédito para este conjunto de dados.

	Model	AUC	Accuracy	Precision	Recall	F1_Score
0	Log_Regression	0.958	0.973	0.050	0.943	0.096
1	Decision_Tree	0.861	0.997	0.303	0.724	0.427
2	Random_Forest	0.902	0.999	0.843	0.805	0.824

Tabela 3. Métricas de avaliação dos modelos

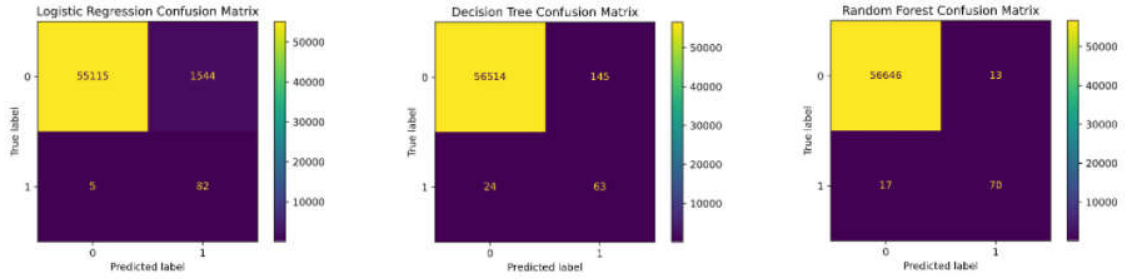

Gráfico 7. Desempenhos do modelo na assertividade.

Referências

Sigmoidal. Como lidar com dados desbalanceados.

NARKHEDE, Sarang. Understanding Confusion Matrix.

Sklearn. Documentação.