

Modelos de aprendizado de máquina aplicados à detecção de fraudes bancárias em transações de cartão de crédito

Cristiano Mendieta, Gabrielly Halas, Kléber Benatti, Mateus Fernandes

23 de março de 2022

Resumo

[inserir]

1 Introdução

A digitalização dos serviços bancários possibilita ao cliente realizar transações bancárias de maneira simplificada. Assim como, esse benefício também é munido de ameaças, como as fraudes que burlam as regras de segurança dos bancos. Por essa razão, a fim de evitar possíveis transações fraudulentas, frequentemente as empresas buscam desenvolver e aperfeiçoar métodos que detectem esse tipo de ação buscando a associação automática de cada transação a uma das categorias "fraude" ou "não-fraude". Esse sistema precisa ser robusto a inconsistências nos dados, ter boa capacidade de generalização e ser eficiente para processar grandes quantidades de transações. (RAMOS, 2014, p. 1,2)

Porém, alguns desafios são enfrentados nessas situações. Existe uma grande variabilidade de ataques, os quais são aprimorados rapidamente, ou seja, a tecnologia dos atacantes pode se desenvolver mais rapidamente do que a das empresas. Isso sugere a necessidade de atualizações constantes do modelo de detecção de fraudes com tempo reduzido de implementação. (RAMOS, 2014, p. 2)

Ademais, o volume de transações financeiras realizadas diariamente é grande, sendo que as caracterizadas como fraudulentas ocorrem em quantidade bastante inferior às não-fraudulentas, isto é, as classes de transações fraudulentas e legítimas são desbalanceadas. Por isso, ao mesmo tempo que o modelo precisa tratar o problema de desbalanceamento de classes para não ter seu desempenho prejudicado, ele precisa ser eficiente para interromper ataques em andamento. (RAMOS, 2014, p. 2)

Nesse sentido, é importante que a acurácia do modelo seja alta, de modo que seja possível minimizar falsos-negativos (fraudes detectadas como não-fraudes) e falsos-positivos (não-fraudes detectadas como fraudes). Esse ponto é fundamental para evitar desgastes com clientes quando uma transação não fraudulenta é classificada como tal. (RAMOS, 2014, p. 2)

Resumidamente, os principais pontos enfrentados nesse cenário são:

1. Muitos padrões de fraude;
2. Mudança de tendência nas fraudes;
3. Automação em *real time*;
4. Classes desbalanceadas (poucas fraudes para muitas transações fidedignas);
5. Órgãos reguladores exigem interpretabilidade em alguns cenários.

Diante disso, esse trabalho implementa e analisa modelos, como: Árvores de decisão, Regressão Logística e *Random Forest*, a partir de uma base de dados simulados de transações de cartão de crédito, levando em consideração metodologias para tratar os problemas das classes desbalanceadas.

2 Análise exploratória dos dados e dataprep

A base de dados utilizada na implementação dos modelos foi extraída do *Kaggle*, plataforma que armazena e disponibiliza diversos *datasets* e permite hospedar competições de *Data Science*, tanto patrocinadas quanto focadas no aprendizado.

Foi escolhido o *dataset Credit Card Transactions Fraud Detection Dataset*, que contém dados simulados de transações de cartão de crédito geradas usando *Sparkov*, disponível por meio do link <https://www.kaggle.com/kartik2112/fraud-detection?select=fraudTrain.csv>. Ela é composta por dois arquivos no formato *comma-separated values* (.csv): *fraudTest.csv* e *fraudTrain.csv*. Para melhor compreensão dessa base foi realizada uma análise exploratória.

A base é formada 1296675 observações distribuídas em 23 variáveis, sendo 11 numéricas e 12 fatores, organizadas conforme segue:

- *index* [numérica] - identificador único de cada linha;
- *transdatetrans_time* [fator] - data e horário da transação;
- *cc_num* [numérica] - número de cartão de crédito do consumidor, contando com 983 números únicos;
- *merchant* [fator] - nome do comerciante, contando com 693 nomes únicos;
- *category* [fator] - categoria do comerciante, contando com 14 diferentes categorias;
- *amt* [numérica] - valor da transação, sendo o mínimo de 1 e máximo de 28948,90, com média de 70,35;
- *first* [fator] - primeiro nome do titular do cartão de crédito;
- *last* [fator] - sobrenome do titular do cartão de crédito;
- *gender* [fator] - sexo do titular do cartão de crédito;
- *street* [fator] - endereço do titular do cartão de crédito;
- *city* [fator] - cidade do titular do cartão de crédito, conta 894 diferentes cidades;
- *state* [fator] - estado do titular do cartão de crédito, conta com 51 diferentes estados;
- *zip* [numérica] - zip do titular do cartão de crédito;
- *lat* [numérica] - latitude da localização do titular do cartão de crédito;
- *long* [numérica] - longitude da localização do titular do cartão de crédito;
- *city_pop* [numérica] - população de titulares de cartão de crédito na cidade;
- *job* [factor] - profissão do titular do cartão de crédito, sendo 494 diferentes profissões;
- *dob* [factor] - data de nascimento do titular do cartão de crédito;
- *trans_num* [factor] - número da transação;
- *unix_time* [numérica] - UNIX time (representação da data da transação em segundos);
- *merch_lat* [numérica] - latitude da localização do comerciante;
- *merch_long* [numérica] - longitude da localização do comerciante;
- *is_fraud* [numérica] - *flag* que determina se a transação é (1) ou não (0) fraude.

Não foram identificados dados duplicados, nem valores faltantes. Para facilitar a utilização dos dados, a coluna *transdatetrans_time* foi separada em duas: uma de data (*date*) e uma de hora (*time*). Ademais, seguem observações obtidas a partir da exploração das variáveis *category*, *amt*, *gender*, *city*, *job*, *lat*, *long*, *merch_lat*, *merch_long*, *is_fraud*.

2.1 Variáveis *is_fraud* e *gender*

Em relação ao total de 1296675 observações tem-se que 7506 delas foram classificadas como fraudes, ou seja, 0.5789% do total de dados.

Do total de 1296675 transações tem-se que 45.255% foram realizadas por pessoas do sexo masculino e 54.745% por pessoas do sexo feminino, sendo que a taxa de fraudes se apresenta superior para as pessoas de sexo masculino, conforme é possível observar na tabela 1.

Sexo	Total Fraudes	Total de Observações	Taxa de fraudes
Masculino	3771	586812	0.00643
Feminino	3735	709863	0.00526

Tabela 1: Taxa de fraudes por sexo (FONTE: Autores(2022))

2.2 Variável *category*

Em relação a variável *category*, que diz respeito as 14 categorias do comerciante que constam no conjunto de dados, tem-se que o gráfico 1 aponta que a maior taxa de de fraudes ocorreu em comércios categorizados em *shopping_net* (0.0176), *misc_net* (0.0145) e *grocery_pos* (0.0141).

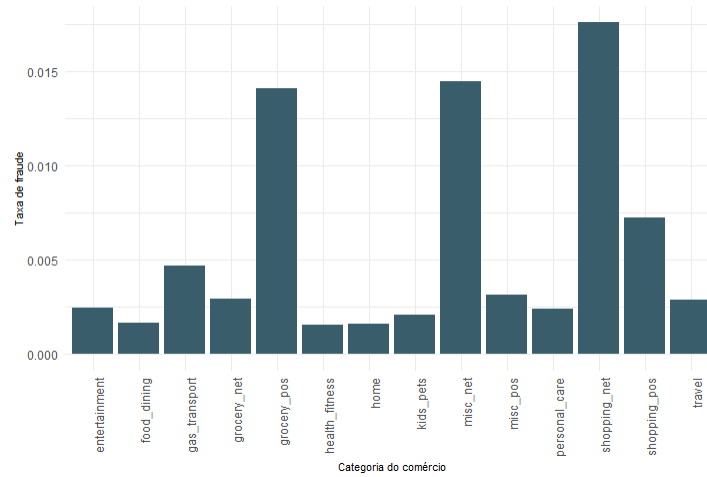


Figura 1: Taxa de fraudes por categoria de comércio (FONTE: Autores (2022))

Para auxiliar a utilização dessa variável nos modelos, as taxas de fraude relacionadas a ela cuja amplitude foi de 0.01607 foram distribuídas em 10 intervalos de amplitude 0.001607. A maior parte das taxas concentrou-se no intervalo $[0.00153, 0.00315]$ (faixa 1), conforme observa-se na tabela 2. A base foi acrescida de uma coluna *faixascategory*, a fim de classificar os dados da coluna *category* em suas respectivas faixas, de acordo com as taxas de fraude.

Faixa	Intervalo	Quantidade de taxas
1	$[0.00153, 0.00315]$	9
2	$(0.00315, 0.00475]$	1
3	$(0.00475, 0.00635]$	0
4	$(0.00635, 0.00795]$	1
5	$(0.00795, 0.00956]$	0
6	$(0.00956, 0.0112]$	0
7	$(0.0112, 0.0128]$	0
8	$(0.0128, 0.0144]$	1
9	$(0.0144, 0.016]$	1
10	$(0.016, 0.0176]$	1

Tabela 2: Faixas de taxas de fraude por categorias de comércio (FONTE: Autores (2022)).

2.3 Variável *job*

A variável *job* que traz a profissão dos titulares do cartão de crédito, aponta que das 494 diferentes profissões, as 10 com maior taxa de fraude foram *Accountant*, *chartered*; *Air traffic controller*; *Armed forces technical officer*; *Broadcast journalist*; *Careers adviser*; *Contracting civil engineer*; *Dancer*; *Engineer*, *site*; *Forest/woodland manager*; *Homeopath*.

Assim como a variável *category*, a variável *job* também teve suas taxas de fraude organizadas em 10 intervalos. Ao distribuir a amplitude do conjunto em 10 intervalos de mesmo tamanho, percebeu-se uma grande concentração dos dados nos intervalos $[0, 0.1]$ e $(0.9, 1]$. Dessa forma, os intervalos foram organizados em 0 a 0.000483, 8 faixas de igual amplitude entre 0.000483 e 0.0519, e um intervalo de 0.0519 a 1, obtendo-se as faixas conforme tabela 3. A base foi acrescida de uma coluna *faixasjob*, a fim de classificar os dados da coluna *job* em suas respectivas faixas, de acordo com as taxas de fraude.

Faixa	Intervalo	Quantidade de taxas
1	$[0, 0.000483]$	51
2	$(0.000483, 0.0069]$	214
3	$(0.0069, 0.0133]$	155
4	$(0.0133, 0.0197]$	35
5	$(0.0197, 0.0262]$	12
6	$(0.0262, 0.0326]$	4
7	$(0.0326, 0.039]$	1
8	$(0.039, 0.0454]$	2
9	$(0.0454, 0.0519]$	0
10	$(0.0519, 1]$	20

Tabela 3: Faixas de taxas de fraude por profissão (FONTE: Autores (2022)).

2.4 Variável *city*

A variável *city* que contempla 894 diferentes cidades aponta que as 10 cidades com maior taxa de fraude foram Angwin, Ashland, Beacon, Brookfield, Bruce, Buellton, Byesville, Chattanooga, Clarion e Claypool.

Da mesma forma como a variável *job*, a variável *city* também teve suas taxas de fraude organizadas em 10 intervalos. Ao distribuir a amplitude do conjunto em 10 intervalos de mesmo tamanho, também percebeu-se uma grande concentração dos dados nos intervalos $[0, 0.1]$ e $(0.9, 1]$. Dessa forma, os

intervalos foram organizados em 0 a 0.000394, 8 faixas de igual amplitude entre 0.000394 e 0.0449, e um intervalo de 0.0449 a 1, obtendo-se as faixas conforme tabela 4. A base foi acrescida de uma coluna *faixacity*, a fim de classificar os dados da coluna *city* em suas respectivas faixas, de acordo com as taxas de fraude.

Faixa	Intervalo	Quantidade de taxas
1	[0,0.000394]	193
2	(0.000394,0.00596]	248
3	(0.00596,0.0115]	230
4	(0.0115,0.0171]	69
5	(0.0171,0.0227]	55
6	(0.0227,0.0282]	31
7	(0.0282,0.0338]	8
8	(0.0338,0.0394]	0
9	(0.0394,0.0449]	2
10	(0.0449,1]	58

Tabela 4: Faixas de taxas de fraude por cidade (FONTE: Autores (2022)).

2.5 Variáveis *lat*, *long*, *merch_lat*, *merch_long*

Com o objetivo de calcular a distância geodésica entre o local de moradia do titular do cartão de crédito e o local no qual foi efetuada a transação, foram utilizadas as variáveis *lat*, *long* referente a primeira localização citada e *merch_lat*, *merch_long*, referente a segunda. Dessa forma, a base recebeu uma coluna *distGeo* que conta com essas distâncias em metros.

3 Referências

RAMOS, J. A. de P. **Árvores de Decisão Aplicadas À Detecção de Fraudes Bancárias**. Disponível em: https://repositorio.unb.br/bitstream/10482/16954/1/2014_JoseAbiliodePaivaRamos.pdf. Acesso em: 18 mar. 2022.