Universidade Federal do Paraná

Nayara Korisztek

O Papel da Regressão Logística na Detecção de Fraudes em transações online de cartão crédito

Curitiba
2023

Nayara Korisztek

O Papel da Regressão Logística na Detecção de Fraudes em transações online de cartão crédito

Trabalho de Conclusão de Curso apresentado à disciplina Laboratório B do Curso de Graduação em Estatística da Universidade Federal do Paraná, como exigência parcial para obtenção do grau de Bacharel em Estatística.

Orientador: Prof. Fernando Lucambio Pérez

Curitiba
2023

Agradecimentos

Inicialmente, expresso minha gratidão a meu Pai, Deus. Foi Ele quem plantou a semente em meu coração, e sem Sua orientação, a jornada teria sido significativamente mais árdua e impossível de ser finalizada. Agradeço por toda a força, coragem e determinação que Ele me concedeu. Deus foi meu alicerce e rocha, e sou grata por tê-Lo ao meu lado nesta caminhada.

Agradeço ao meu marido, Allyson, enviado pelo Senhor para me proporcionar conforto, força e tranquilidade. Ele esteve presente em cada passo, em cada conquista e, igualmente, em cada derrota. Afinal, enfrentamos juntos diversos desafios até alcançarmos essa linha de chegada. Em muitos momentos, foi graças ao seu apoio que permaneci firme no propósito, focada no alvo e relembrando tudo que fiz para chegar até aqui. Sem ele, nada disso seria possível; esta vitória também é dele.

Agradeço à minha mãe, Wilma, por sempre me incentivar nas provas com suas palavras de fé: "Você é capaz, você vai conseguir!". Estendo meus agradecimentos à família pelo apoio e compreensão diante de minha ausência em eventos e em suas vidas, nos quais não pude estar tão presente ao longo desses anos.

Aos amigos da Betel, minha segunda casa, meu sincero agradecimento pelas orações, consolos e conversas que fortaleceram minha fé e esperança. O apoio emocional e espiritual de vocês foi crucial nesse período.

Aos amigos que a vida me deu antes e durante esse período. Agradeço pelas orações, por me acolherem, apoiarem, e acima de tudo, por tornarem tudo mais leve em muitos momentos.

Aos colegas de profissão, agradeço por compartilharem suas vivências e conhecimento profissional comigo. Isso contribuiu significativamente para meu crescimento profissional e o desenvolvimento do tema abordado.

Aos meus colegas do curso, por todo conhecimento e desafios enfrentados juntos, espero que todos continuem trilhando seus caminhos com muito empenho e determinação, assim como tivemos até aqui. Um agradecimento especial ao colega Fabio Pereira por conceder-me a base para o estudo. Minha gratidão pela gentileza e empatia.

Ao meu estimado orientador, Professor Lucambio, expresso meu agradecimento pela parceria e orientação exemplar. Você me guiou quando muitas vezes não enxerguei o caminho. Sem seu suporte e incentivo, certamente este trabalho não teria evoluído da maneira que evoluiu, e não teria alcançado o patamar que atingiu. Gratidão.

Por último, mas não menos importante, à Professora Nivea Machado, agradeço por aceitar fazer parte da banca avaliadora. É uma grande honra tê-la comigo neste momento.

Nayara Korisztek

"Porque vivemos por fé, e não pelo que vemos".
(Bíblia Sagrada NVI, 2 Coríntios 5:7)

Resumo

Neste trabalho, é abordado o aumento significativo das transações com cartão de crédito, impulsionado pelo avanço tecnológico e econômico, resultando em oportunidades crescentes para fraudes e consideráveis perdas financeiras. No ambiente do comércio eletrônico (e-commerce), transações sem a presença física do portador do cartão são percebidas como mais arriscadas, e esta complexidade será explorada ao longo da pesquisa.

A Estatística é enfatizada como uma ferramenta crucial na identificação e prevenção de fraudes, com destaque para o papel da regressão logística na análise do comportamento transacional. O objetivo é explorar a fraude em transações online com cartões de crédito, identificando padrões de comportamento que possam esclarecer a ocorrência de fraudes e, assim, contribuir com insights para o desenvolvimento de regras mitigatórias no futuro.

A principal proposta do estudo é a comparação da eficácia de modelos estatísticos na previsão de transações fraudulentas, utilizando dados reais. Além da abordagem previamente mencionada da regressão logística, o foco é avaliar como esses modelos se destacam na identificação precisa de transações fraudulentas, promovendo uma compreensão mais abrangente das estratégias de detecção de fraudes.

Foi identificado que a seleção do modelo adequado deve considerar diversos elementos, incluindo o leque de variáveis disponíveis, a eficiência computacional e a capacidade de resposta em tempo real.

Palavras chaves: Cartão de Crédito, Fraudes, Comércio Eletrônico, Regressão Logística, Modelos Estatísticos, Detecção de Fraudes.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

CBK	Chargeback
CNP	Cartão Não Presente
CP	Cartão Presente
PCA	Principal Component Analysis
VP	Verdadeiro Positivo
VN	Verdadeiros Negativos

Sumário

1 Introdução

1.1 Contextualização do Tema

Com o avanço tecnológico e econômico, que facilitaram o processo de comunicação e aumento do poder de compra, transações com cartão de crédito tornaram-se o principal meio de pagamento no varejo nacional e internacional (BOLTON; HAND, 2002). Neste aspecto, o aumento do número de transações com cartão de crédito é crucial para a geração de mais oportunidades para fraudadores produzirem novas formas de fraudes, o que resulta em grandes perdas para o sistema financeiro (CHAN et al., 1999). Evidentemente que em 2023, o cartão de crédito continua sendo um meio de pagamento muito importante, mas hoje, o mercado financeiro possui outros novos meios de pagamentos, como por exemplo o PIX, que está em alta e que também já possui o módulo de parcelamento assim como o com o cartão de crédito.

No comércio eletrônico (e-commerce), a transação é realizada pelo módulo CNP (Cartão Não Presente), estas, são consideradas transações mais arriscadas do que compras CP(Cartão Presente), ou seja, transações presenciais em terminais. Esse fato ocorre devido a "facilidade"de poder se passar pelo portador quando não se está presencialmente.

Quando a fraude ocorre, esse prejuízo não afeta apenas indivíduos, mas também instituições financeiras, empresas, bem como, a economia em geral. Quanto mais perda financeira por fraude as instituições financeiras e empresas assumem para si, mais altas serão as suas taxas de juros, seguros, serviços, etc, assim garantindo o equilíbrio financeiro entre lucros e prejuízo relacionados a essas perdas.

No contexto da identificação e prevenção de fraudes, a Estatística desempenha um papel fundamental ao fornecer métodos e técnicas que auxiliam na obtenção de informações cruciais. A análise das transações de cartão de crédito, por meio da Estatística, contribui para mitigar perdas. Um problema encontrado na detecção de fraude é a estrutura dos bancos de dados. Em geral, dentre todas as transações realizadas, a porcentagem que são fraudes é bem pequena. Para este tipo de conjuntos de dados, (CRAMER, 2004) sugere o uso da regressão logística limitada para esse tipo de comportamento. Segundo (HOSMER et al., 1997), a regressão logística busca explicar a relação entre uma variável resposta dicotômica dependente e um conjunto de variáveis explicativas independentes (qualitativas ou quantitativas).

Neste trabalho, iremos explorar alguns aspectos do mundo da fraude relacionada a cartões de crédito, utilizando a metodologia da regressão logística, do qual tem como principal relacionar variáveis independentes a uma variável dependente, neste caso, a classificação de uma transação como fraude ou não. Por meio disto, espera-se ter insumos para a identificação do modus operandi que possa explicar as fraudes ocorridas no cartão de crédito, assim possibilitando em um futuro a criação de possíveis regras mitigatórias que visam diminuir a ocorrência de fraudes.

1.2 Organização do Trabalho

Após essa breve introdução no Capítulo 1, a estruturação deste trabalho é dividida em quatro partes distintas. O Capítulo 2 aborda uma fundamentação histórica do cenário de fraude no Brasil, além de fornecer um resumo conciso dos métodos e técnicas estatísticas que serão aplicados ao longo deste estudo. No Capítulo 3, são apresentados os detalhes relacionados aos dados obtidos, as manipulações necessárias e técnicas empregadas. Os ajustes dos modelos e os resultados obtidos são discutidos e expostos no Capítulo 4. Por fim, o Capítulo 5 incorpora a discussão sobre os objetivos propostos, suas realizações, e sugere ideias para futuras pesquisas.

2 Fundamentação Teórica

2.1 O que é Fraude?

O termo fraude, “latu sensu”, significa: qualquer ato ardiloso, enganoso, de má-fé, com o intuito de lesar ou ludibriar outrem, ou de não cumprir determinado dever (Houassis, 2023). Uma outra maneira de resumir o que é fraude seria: ***Uma forma de enganar e manipular informações envolvendo práticas criminosas para obter uma vantagem de forma ilegal de terceiros***.

Os praticantes de fraudes, conhecidos como "golpistas"ou "fraudadores"existem desde os primórdios da economia, há milhares de anos, desde o antigo Egito. Devido ao avanço da tecnologia, novos sistemas e aparelhos evoluem a cada dia, transformando-se em poderosas ferramentas nas mãos desses golpistas. Embora existam várias soluções antifraude em vigor, os casos de fraude continuam a crescer e as estratégias utilizadas por esses fraudadores estão se tornando cada vez mais diversificadas e sofisticadas.

2.2 Tipos de Fraude

Os fraudadores estão constantemente adotando novas estratégias para realizar golpes. No entanto, em grande parte das situações, as fraudes mantêm um padrão e um modus operandi já conhecido, o que permite a sua caracterização e identificação. Portanto, nesta seção, serão abordados diversos tipos de fraudes, juntamente com suas definições.

2.2.1 Phishing

Como sugere o próprio nome "pescaria", o phishing é uma fraude amplamente reconhecida pelo uso da técnica de Engenharia Social. Geralmente, ela se inicia com o envio de um e-mail fraudulento para a vítima, solicitando informações pessoais, como nome de usuário, senhas e detalhes do cartão.

O golpista, muitas vezes, se passa por uma instituição confiável, como um banco ou uma loja conhecida. O e-mail costuma alegar mudanças no sistema e solicitar as informações mencionadas. Ao clicar no link fornecido no e-mail, a vítima é redirecionada para um site falso, projetado para parecer legítimo.

O phishing proporciona aos criminosos acesso indevido a contas bancárias ou outros serviços e pode ser utilizado para roubo de identidade. A astúcia desse método torna essencial que os usuários estejam atentos e adotem práticas de segurança rigorosas ao lidar com comunicações eletrônicas suspeitas.

2.2.2 Autofraude

Em geral isso ocorre quando o titular do cartão realiza uma compra e, posteriormente, contesta-a, alegando não reconhecer a transação devido ao suposto comprometimento do seu cartão. Em seguida, solicita um reembolso, mas não retorna o produto adquirido, criando então o que chamamos de processo de CBK (chargeback).

Este tipo de fraude geralmente não é cometido por criminosos ou fraudadores, mas sim por consumidores que estão plenamente cientes de suas ações.

2.2.3 Fraude Amigável

Nesse tipo de fraude, a pessoa que utilizou os dados do cartão ou conta bancária é alguém próximo ao titular, como um familiar ou amigo. A vítima não tem conhecimento de que alguém próximo tenha usado seus dados e/ou cartão para realizar a transação, o que a leva a contestar a compra.

Geralmente, quando o processo de CBK é iniciado, durante a investigação é identificado o vínculo do suposto fraudador com a vítima, e na grande maioria das vezes, a vítima assume o prejuízo, pois neste caso, a mesma deixou seu cartão e/ou dados expostos.

2.2.4 Fraude Application

Pode ocorrer por meio de invasão de conta ou de um vazamento de dados do cartão e/ou informações cadastrais.

Popularmente conhecida como “Fraude Limpa”, onde fraudador faz uma personificação da vítima, ou seja, possui os dados corretos da mesma. Por conta disso, esse tipo de fraude ainda mais difícil de identificar.

Esse tipo de fraude em especial, tem grande possibilidade de ser o tipo característico do estudo deste trabalho, visto que objetivo principal é analisar a ocorrência de fraude em transações online de cartão de crédito em um aplicativo de delivery. Neste caso, a fraude pode ser oriunda tanto de vazamento de dados pessoais ou dados do cartão, conforme caracteriza a Fraude Application. Posteriormente neste trabalho, serão realizadas análises mais cautelosas afim de entender o comportamento das variáveis, do qual pode fornecer melhores insumos para compreender se de fato as ocorrências se enquadram neste tipo de fraude ou não.

2.3 Evolução da Fraude no Brasil

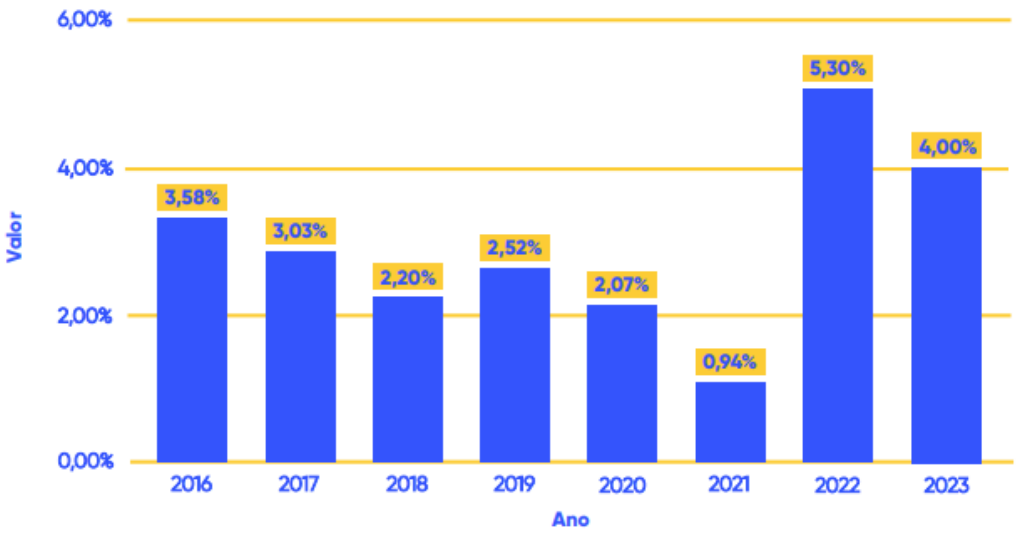
Atualmente, no mercado de aplicações e soluções antifraude, existem diversos tipos de ferramentas que contribuem no valor evitado de fraude para a economia, instituições financeira, varejo, etc., mas isso não quer dizer que necessariamente o número de tentativas

de fraude diminua, pelo contrário, mesmo com ferramentas que auxiliam a suavizar esse prejuízo, historicamente é possível ver o crescimento das tentativas de fraude, bem como, as fraudes efetivas.

Segundo (Serasa Experian, 2023), em abril de 2023, os brasileiros sofreram uma tentativa de golpe a cada 11 segundos. Esse indicador nos faz pensar na seguinte reflexão: *"O que podemos fazer hoje em nossas vidas com 11 segundos?* Alguns podem alegar que não há tempo para muitas ações nesse breve intervalo, enquanto outros afirmariam que poderiam concluir algumas tarefas. O estudo do conduzido pelo (Serasa Experian, 2023) revela um indicador muito significativo e preocupante. Além disso, os dados, que são do Indicador de Tentativas de Fraude da Serasa Experian, mostram que o total foi 232.478 investidas de criminosos contra consumidores e empresas a fim de fraudá-los ou roubar suas identidades. Apesar de alarmantes, os números sofreram uma queda de 14,8% relativos a março e um recuo de 22,2% comparados a abril de 2022 (Serasa Experian, 2023).

O Censo de Fraudes de 2023 conduzido pela Konduto, uma empresa líder em soluções antifraude, oferece uma visão do comportamento das fraudes e-commerce (compras on-line) ao longo do tempo. Apesar de não serem dados provenientes de uma entidade oficial do governo ou relacionado, sua significativa relevância é respaldada pela sólida atuação e reputação da empresa no mercado. Este Censo demonstra o comportamento fraudulento no primeiro semestre de 2023 no âmbito do comércio eletrônico brasileiro.

Figura 1 – Tentativas de fraude ao longo dos anos



Fonte: Censo da fraude 2023 - Konduto

Podemos ver uma grande oscilação ao longo dos anos detectada na Figura 1, essa discrepância pode ocorrer por vários motivos, pode estar relacionado a melhoria nas

ferramenta antifraude, por aquecimento do mercado, por novos tipos de modus operandi, entre outros motivos.

A fraude em si não é algo que identificamos só de olhar uma única variável. Para que tal detecção seja realizada com sucesso, todo o cenário deve ser levado em conta: desde o mercado financeiro, até os dados de geolocalização do usuário. Atrelado a isso, podemos meditar sobre a citação da Konduto: "Uma informação gera um insight, mas uma informação com uma análise completa gera um bom combate. É por isso que os números nos dizem mais sobre uma oscilação do mercado - que passou por pandemia, por mudanças de comportamento e até de tecnologias - do que sobre uma tendência ou outra em si". (Konduto, 2023)

O aumento da fraude muitas vezes é diretamente proporcional com o aquecimento do mercado, ou seja, quanto mais se vende, mais fraude acontece. Diante disso, esse fator também deve ser levado em conta. Podemos ver o reflexo na Figura 2:

Figura 2 – Proporção de Pedido vs Fraude

Mês	% de pedidos	% de fraudes
Janeiro	17,39%	18,51%
Fevereiro	14,75%	15,76%
Março	16,58%	17,21%
Abril	17,51%	17,34%
Maio	17,80%	17,03%
Junho	15,97%	14,15%

Fonte: Censo da fraude 2023 - Konduto

Quando falamos em analisar a proporção de pedido vs proporção de fraude, deve-se considerar o tipo de segmentação de mercado que está em estudo, ou seja, se é uma instituição financeira, indústria, varejo, etc.

No varejo por exemplo, é comum ter um aumento bem expressivo de vendas no mês de dezembro devido ao Natal, assim como, no mês de maio e agosto, respectivamente onde ocorre a data do dia das mães e dia dos pais. Logo, para avaliar corretamente tais proporções, além de analisar todos os fatores que podem justificar o aumento da fraude, deve-se atentar ao tipo de segmentação de mercado , se possui um aumento sazonal de vendas em algum período do ano, pois isso pode impactar diretamente sua proporção de fraude.

Analisando o cenário do primeiro semestre de 2023 apresentado no Censo da Fraude (Konduto 2023), podemos notar que janeiro apresenta o maior percentual de fraude

do ano, porém nos demais meses do semestre do ano, tanto pedidos quanto fraude tiveram proporções bem parecidas, isso pode ser reflexo do que acontece no mercado econômico.

Janeiro geralmente é o mês que muitas pessoas tiram férias, e é um mês que historicamente é muito cobiçado pelos fraudadores, assim como, o aquecimento do mercado financeiro, varejo, etc sempre ocorre nessa época do ano.

Portanto, se uma empresa sofre um aumento significativo de fraude nessa época do ano, esse fator sazonalidade deve ser levado em conta dependendo do tipo de segmentação de mercado , para que interpretações não sejam feitas de forma errônea. Segundo (Konduto, 2023): "18,5% de todas as fraudes do primeiro semestre aconteceram em janeiro. Isso não significa que 18,5% dos pedidos de janeiro foram fraudes".

Voltando um pouco o olhar para fraude por região geográfica, podemos observar o seguinte cenário:

Figura 3 – Proporção de Pedido e Fraude por Região

Regiões	Transações	Fraudes
Norte	2,81%	4,89%
Nordeste	12,43%	13,00%
Centro-Oeste	6,63%	5,97%
Sudeste	63,82%	65,94%
Sul	14,31%	10,21%

Fonte: Censo da fraude 2023 - Konduto

Pode-se observar que na maioria das regiões possui uma proporção de transação relativamente próxima a de fraude, até mesmo o Sudeste, que possui o maior percentual de fraude, mas como já explicado anteriormente, a proporção de pedidos/transação também impacta diretamente essa correlação. Levando em conta que o Sudeste é a região com maior índice em ambos percentuais, pode-se considerar que essa relação diretamente proporcional pode estar ocorrendo.

Importante também lembrar que, no Sudeste está localizada a cidade de São Paulo, uma das cidades com maior movimentação do país, se não a maior, no que diz respeito a giro econômico.

Outro ponto importante de analisar é por onde está ocorrendo a fraude?. É importante entender qual tipo de dispositivo, e muitas vezes entender até qual o tipo de sistema operacional está sendo usado no ato da fraude. É possível ter uma noção deste cenário de dispositivo por meio da Figura 4 e 5 a seguir:

Figura 4 – Proporção de Pedido por Tipo de Dispositivo

Dispositivo	2022	2023
Desktop	27,93%	25,06%
Mobile	72,07%	74,94%

Fonte: Censo da fraude 2023 - Konduto

Atualmente, o celular (Mobile) é um dos bens mais utilizado pela sociedade (se não o maior), o que justifica ter a maior proporção de pedidos (Figura 4). Por meio dele é possível marcar compromissos na agenda, registrar momentos, conversar em redes sociais, fazer ligações, efetuar pagamentos, realizar compras, etc.

Devido a essa gigantesca facilidade que o Mobile nos proporciona, podemos estar em qualquer lugar e a qualquer instante, gerando um grande obstáculo para a prevenção à fraude quanto a geolocalização. Criar medidas mitigatórias relacionadas ao tema torna-se um grande desafio.

Figura 5 – Proporção de Fraude por Tipo de Dispositivo

Dispositivo	2022	2023
Desktop	27,20%	26,17%
Mobile	72,80%	73,83%

Fonte: Censo da fraude 2023 - Konduto

Diante dessa facilidade que aparelho traz para a sociedade, é comum e esperado que o maior índice de fraude seja o dispositivo do tipo Mobile, conforme demonstrado na Figura 5.

Além do Mobile ser o dispositivo com mais uso hoje na atualidade, ao contrário do desktop, ele também é mais suscetível para furto e roubo, o que acaba se tornando uma grande oportunidade para os fraudadores. Além disso, os fraudadores costumam usar manobras que auxiliam a mascarar o real local da compra. Quando isso ocorre, o que pode-se ser analisado em contra partida é o comportamento do cliente, se é uma região/local do qual o mesmo realiza compras, se aquele local faz parte do seu perfil.

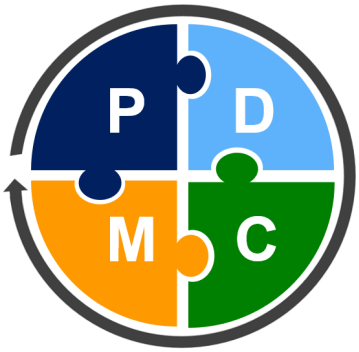
Portanto, a evolução da fraude no país pode ser analisada por diversas frentes, desde região até o tipo de dispositivo, mas o mais importante é sempre lembrar que a fraude é uma coisa mutável, hoje ela pode ser de um jeito, amanhã já pode ter um comportamento totalmente diferente, o que torna esse meio sempre desafiador e cheio de motivações para estudos e análises de como combater a fraude.

2.4 Pilares da Gestão e Combate a Fraude

Nas seções anteriores, exploramos conceitos cruciais relacionados à fraude, examinando seu comportamento ao longo dos anos no Brasil. Contudo, é igualmente essencial compreender as medidas preventivas contra a fraude e identificar os principais pilares estratégicos para garantir uma gestão eficaz. Nessa seção, aprofundaremos ainda mais nesse tema:

Podemos sintetizar a estratégia de gestão e combate a fraude em quatro pilares essenciais: Prevenção, Detecção, Correção e Melhoria Contínua.

Figura 6 – Os 4 Pilares



Fonte: A autora (2023)

Prevenção: Tem como objetivo principal a redução de chance de ocorrência de fraudes por meio de estratégias. Engloba desde a criação de políticas e processos, até a contratação e implementação de sistemas e tecnologias antifraude. É importante ter essas medidas mitigatórias alinhadas e em sinergia, para que a prevenção tenha um bom desempenho.

Detecção: Para que esse pilar realize o seu propósito, é necessário fazer uso de ferramentas e processos eficientes e ágeis, que permitam atuação em tempo real ou o mais brevemente possível. Por meio dessas medidas, será possível ter uma possibilidade maior de mitigação das fraudes que passaram pelo pilar anterior. Tendo em vista que como a fraude é mutável, nem sempre o pilar de Prevenção conseguirá mitigar a fraude, por isso é de suma importante ter o pilar de Detecção bem estruturado.

Correção: A fraude é uma coisa inevitável muitas vezes, por mais robustas que sejam as ferramentas antifraude, por mais eficazes que sejam os pilares de prevenção e detecção. A fraude é como um organismo vivo que está em constante mudança. Diante disso, faz-se necessário remediar a fraude após a sua detecção e realizar os ajustes necessários para futuras ocorrências.

Melhoria Contínua: Consiste na melhoria contínua dos processos, políticas, parametrização nas ferramentas, regras, etc.

Apesar de a Melhoria Contínua frequentemente integrar-se aos demais pilares anteriores mencionados, optamos por criar esse pilar para enfatizar a relevância e o papel da melhoria contínua, que pode ser compreendida como um processo de aprimoramento da qualidade.

Em termos gerais, a gestão de combate à fraude frequentemente concentra-se em resolver problemas imediatos, ou seja, em apagar incêndios que ocorrem no momento. No entanto, ao considerar a Melhoria Contínua como um pilar tão crucial quanto os anteriores, torna-se possível identificar falhas e fragilidades nos demais pilares antes que sua ineficiência se manifeste. Essa abordagem não só contribui para mitigar a fraude de maneira mais eficaz, mas também permite uma correção proativa e contínua dos processos, fortalecendo a integridade do sistema de combate à fraude.

Considerando os princípios fundamentais dos pilares de gestão e combate à fraude, é relevante destacar que de acordo com o propósito deste trabalho estaremos mais direcionados ao segundo pilar, o de "Detecção". O foco principal reside na criação de um modelo capaz de detectar transações fraudulentas em transações online do cartão de crédito

2.5 Regressão Logística

Segundo (HOSMER et al., 1997), a regressão logística busca explicar a relação entre uma variável resposta dicotômica e um conjunto de variáveis explicativas independentes (qualitativas ou quantitativas). É considerado um algoritmo de modelagem preditiva, utilizado quando a variável Y é categórica binária, ou seja, pode assumir somente dois valores, 0 ou 1.

A Regressão Logística é uma extensão da Regressão Linear. Na Regressão Linear Y sempre é uma variável contínua, e caso seja categórica, não pode-se fazer uso da mesma. Diante disso, deve-se fazer uso da Regressão Logística quando tiver esse tipo de variável.

A expressão mais usada para o modelo de Regressão Logística é:

$$\pi_i = \frac{\exp(\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip})}$$

É possível observar que $exp(\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip})$ é sempre positivo, do qual o numerador da equação acima é menor que o denominador, o que leva a $0<\pi<1$.

Um modelo de regressão logística também pode ser escrito como:

$$\log\left(\frac{\pi}{1-\pi_i}\right) = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}$$

A Regressão Logística tem como intuito determinar uma equação matemática que pode ser usada para prever o evento de interesse, a probabilidade do evento 1 ocorrer. É importante ressaltar que se Y possuir mais que duas classes, não será possível fazer uso da regressão logística, pois não atende o pressuposto da resposta ser dicotômica.

O modelo de Regressão Logística por meio da função logística, também conhecida como função sigmoidal, pode calcular a probabilidade de que um determinado evento ocorra, dado um conjunto de valores das variáveis independentes. A função logística desempenha um papel fundamental ao converter a combinação linear das variáveis independentes em um valor compreendido entre 0 e 1, o que representa uma probabilidade estimada. A seguir, um limite de decisão é empregado para classificar as observações em uma das duas categorias possíveis. O modelo é ajustado aos dados por meio da otimização da verossimilhança, resultando na estimação dos coeficientes que melhor se adaptam aos dados observados. Esses coeficientes podem ser interpretados como as mudanças logarítmicas nas probabilidades de sucesso, associadas a um aumento unitário nas variáveis independentes.

2.5.1 Vantagens do Modelo de Regressão Logística

O modelo de Regressão Logística é altamente utilizado em diversas áreas e um dos motivos é o grande número de vantagens e facilidades de interpretação e análise desse modelo. Diante disso, serão expostas abaixo algumas dessas vantagens:

- **Previsão de probabilidade:** Devido a regressão modelar diretamente a probabilidade de um evento ocorrer, ela é muito vantajosa para previsão de probabilidades em vez de apenas classificação.
- **Coeficientes de fácil interpretação:** Os coeficientes na regressão logística são relativamente fáceis de interpretar. pois representam a mudança nas chances (odds) de um evento ocorrer para cada unidade de alteração nas variáveis independentes.
- **Estudo Caso Controle:** O modelo é adequado para Estudos de Caso e Controle, pois pode-se avaliar a associação entre exposições e doenças.
- **Modelagem Flexível:** Pode acomodar variáveis independentes categóricas, tornando-a flexível para modelar dados com diferentes tipos de preditores.
- **Código Simples:** Diferente de alguns outros modelos, o modelo de Regressão Logística possui funções de fácil entendimento, o que torna o código do modelo de simples execução, o que também auxilia muito na aplicação do modelo em uma análise.

2.5.2 Desvantagens do Modelo de Regressão Logística

Embora o modelo logístico seja amplamente utilizado na atualidade e possua muitas vantagens, como qualquer modelo estatístico ele possui algumas desvantagens e limitações. Abaixo são listadas algumas das desvantagens mais comuns do modelo de regressão logística:

- **Amostras Suficientemente Grandes**: O modelo de regressão logística requer um tamanho amostral suficientemente grande para produzir estimativas confiáveis. Quando utilizado em conjuntos de dados pequenos, pode retornar resultados não confiáveis, do qual precisarão de reajuste.
- **Ineficaz com desequilíbrio de classes**: Quando a variável dependente binária possui classes desequilibradas, ou seja, uma classe é muito mais comum do que a outra, o modelo de regressão logística pode ter dificuldade em fornecer previsões precisas, e a métrica de acurácia pode ser enganosa. Veremos no decorrer do estudo que manobras foram feitas para tratar essa desvantagem em especial.
- **Sensibilidade a outlier:** O modelo de regressão logística é sensível a outliers (valores atípicos). Valores extremos podem distorcer as estimativas dos coeficientes e afetar a interpretação dos resultados, trazendo prejuízos para a análise.
- **Limitação à resposta binária**: Conforme já mencionado, o modelo de regressão logística é especificamente utilizado onde a variável resposta é binária. Se a variável resposta se trata de uma variável dependente contínua ou multiclasse, pode ser necessário considerar outros tipos de modelos, pois o mesmo não pode ser utilizado nesse cenário.

2.5.3 Áreas de aplicação

A Regressão logística desde a sua criação até os tempos atuais é um modelo amplamente utilizado em diversas áreas, tais como:

- **Ciências sociais:** Pesquisa sobre os fatores que influenciam o voto em uma determinada eleição.
- **Marketing:** Estudo para prever a probabilidade de um cliente comprar um produto com base em informações comportamentais ou até mesmo demográficas.
- **Finanças:** Avaliar o risco de inadimplência ao conceder empréstimos pessoais.
- **Medicina:** Prever a probabilidade de um paciente desenvolver uma determinada doença.

Conforme citado no item 2.4, o modelo de Regressão Logística é uma ferramenta é utilizada amplamente em diversas área e é essencial e flexível na análise de dados. Com o progresso contínuo das técnicas de aprendizado de máquina, espera-se que o modelo de regressão logística continue a se aprimorar, tornando-se ainda mais robusto.

Portanto, por meio deste modelo estatístico, podemos responder a perguntas críticas, como qual é a probabilidade de um paciente desenvolver uma doença". ou qual é a probabilidade de uma transação ser fraude?. A segunda pergunta em questão, é o que vamos buscar responder ao longo deste trabalho.

3 Materiais e Métodos

3.1 Material

3.1.1 Origem dos Dados

Para o desenvolvimento das análises propostas neste trabalho, utilizam-se dados reais provenientes de uma empresa de delivery, especializa em vendas online por meio de um aplicativo

3.1.2 Banco de Dados

Os dados transacionais da base referem-se a transações financeiras ocorridas em 4 cidades, sendo elas: Recife, São Paulo, Fortaleza e Teresina. Essas operações ocorreram durante os cinco primeiros meses de 2022. Nesse estudo em questão, foi analisado apenas a cidade de Recife.

No conjunto de dados além de conter transações genuínas (legítimas), também possui tentativas de fraude, sendo a variável "fraud". A variável fraude identifica as transações fraudulentas, a variável resposta do modelo de regressão.

Composta por 31 variáveis, que estão separadas da seguinte maneira:

- **Dados do usuário:** Possui 6 variáveis que descrevem dados básicos do usuário, tais como: de localização, sistema operacional do dispositivo, etc.
- **Características da Compra:** Contém 7 variáveis referentes a características específicas das compras, como: valor em reais da transação, status da transação, etc.
- **Perfil do usuário:** Possui 18 variáveis que estão relacionadas ao comportamento dos usuários no momento da compra, como: quantidade de pedidos nos últimos 90 dias, se cerveja é um item do pedido, valor do pedido aprovado no último dia, etc.

A base em questão é muito desbalanceada, do qual é um comportamento esperado de um estudo que envolve transações legítima vs fraudulenta.

Possui em sua totalidade 24923 transações registradas, sendo 24688 transações legítimas e 235 são marcadas como fraudulentas, onde "1"o que representa a fraude, e ”0” o que representa a não fraude, conforme demonstrado pela saída do R a seguir:

```
1 > table (dados$fraud)
2     0     1
3 24688   235
```

Listing 3.1 – Proporção de fraude e transações legítimas.

É importante ressaltar, que o volume de fraudes pode ser superior ou inferior ao registrado no conjunto, pois as marcações de fraude podem muitas vezes serem marcadas de forma errônea, ou até mesmo existir um número maior de fraude, da qual não foi computado devido a inexistência da abertura de CBK (chargeback) junto a instituição emissora do cartão. Abaixo, a Tabela 1 contém o dicionário de dados com suas respectivas descrições:

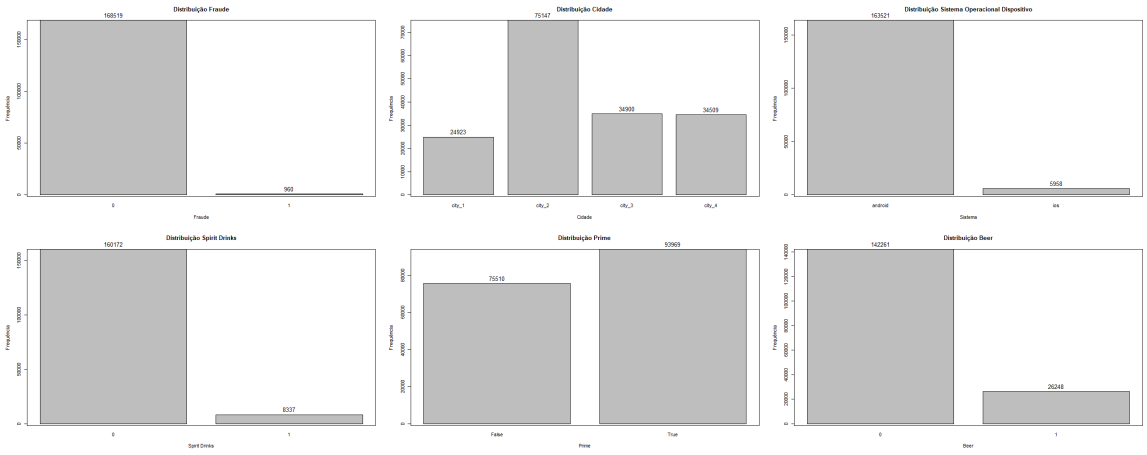
Tabela 1 – Descrição das variáveis

Variável	Descrição
account_age	tempo em dias desde a criação da conta do usuário na plataforma
amount	valor em reais da transação
approved_1d	quantidades de pedidos aprovados no último dia do usuário
approved_7d	quantidades de pedidos aprovados nos últimos 7 dias do usuário
approved_90d	quantidades de pedidos aprovados nos últimos 90 dias do usuário
approved_amount_1d	valor dos pedidos aprovados no último dia do usuário
approved_amount_7d	valor dos pedidos aprovados nos últimos 7 dias do usuário
approved_amount_90d	valor dos pedidos aprovados nos últimos 90 dias do usuário
attemp_1d	quantidades de tentativas de pedidos no último dia do usuário
attemp_7d	quantidades de tentativas de pedidos nos últimos 7 dias do usuário
attempt_amount_1d	valor das tentativas de pedidos no último dia do usuário
attempt_amount_7d	valor das tentativas de pedidos nos últimos 7 dias do usuário
beer	se cerveja é um item do pedido
cards_1d	quantidade de cartões diferentes utilizados pelo cliente no último dia
cards_7d	quantidade de cartões diferentes utilizados pelo cliente nos últimos 7 dias
cards_90d	quantidade de cartões diferentes utilizados pelo cliente nos últimos 90 dias
cities	cidade da compra
discount	valor de desconto aplicado pelo cliente na hora da transação
discount_rate	percentual do desconto aplicado pelo cliente em relação ao valor da transação
fraud	se o pedido resultou em fraude ou não
latitude	latitude do endereço de entrega
longitude	longitude do endereço de entrega
order_attempt	número da tentativa da transação
order_h_duration	tempo em horas entre o começo da compra e o fim
order_sec_duration	tempo em segundos entre o começo da compra e o fim
plataform	sistema operacional do dispositivo
prime	classificação do cliente na plataforma
retail	seguimento do estabelecimento de venda
spirit_drinks	se bebidas destiladas estão presentes no carrinho de compra
status	status da transação
timezone	data e hora da transação

Fonte: A autora (2023).

Por meio da estatística descritiva e análise exploratória é possível ver as distribuições das variáveis. Diante disso, foram escolhidas aleatoriamente algumas variáveis para termos noção do comportamento e distribuição das mesmas:

Figura 7 – Distribuição das variáveis.

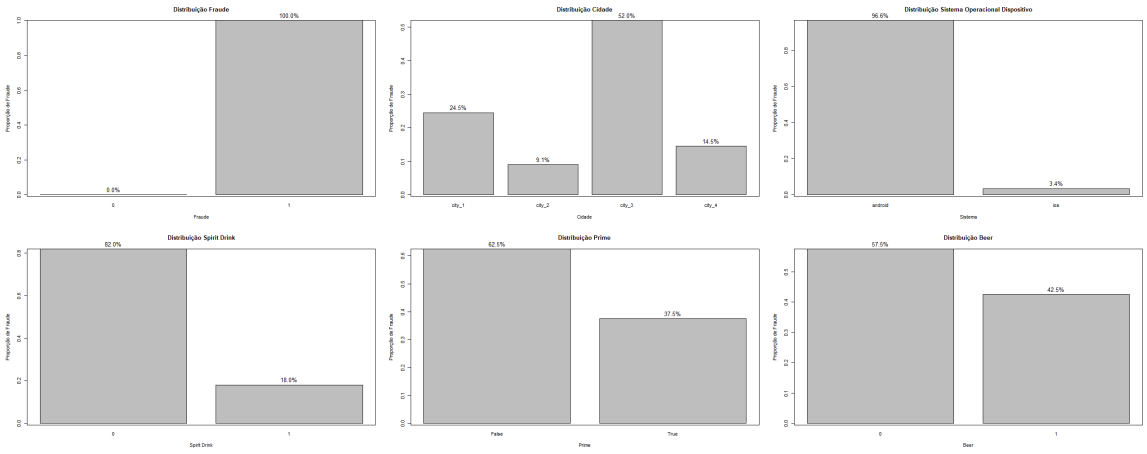


Fonte: A autora (2023).

A Figura 6 nos mostra claramente o desbalanceamento entre transações legítimas e transações fraudulentas, que é um comportamento esperado.

Além disso, é possível fazer outras observações, como o sistema operacional por exemplo, a classe de Android é bem superior a classe iOs. Neste caso, o desiquilibrio de classe pode estar relacionado ao poder de compra dos usuários, visto que os dispositivos com sistema operacional iOs possui um valor de mercado superior aos dispositivos com sistema operacional Android. Abaixo veremos esses mesmos comportamentos, porém agora com a visão de proporção de cada classe:

Figura 8 – Proporção das variáveis.



Fonte: A autora (2023).

Além da distribuição e proporção de cada classe da variável, abaixo será demonstrado a proporção de fraude das respectivas variáveis:

```
prop_city_fraud <- prop.table(tab_cont_city_fraud, margin = 1)
> prop_city_fraud
                      0            1
  city_1 0.1456699650 0.0013866025
  city_2 0.4428867293 0.0005133379
  city_3 0.2029809003 0.0029443176
  city_4 0.2027979868 0.0008201606

prop_sistema_fraud <- prop.table(tab_cont_sistema_fraud, margin = 1)
> prop_sistema_fraud
                      0           1
  android 0.994331003 0.005668997
  ios     0.994461229 0.005538771

prop_spirit_fraud <- prop.table(tab_cont_spirit_fraud, margin = 1)
> prop_spirit_fraud
                0           1
  0 0.995086532 0.004913468
  1 0.979249130 0.020750870

prop_prime_fraud <- prop.table(tab_cont_prime_fraud, margin = 1)
> prop_prime_fraud
                     0           1
  False 0.992054033 0.007945967
  True  0.996168949 0.003831051

prop_beer_fraud <- prop.table(tab_cont_beer_fraud, margin = 1)
> prop_beer_fraud
                0           1
  0 0.996119808 0.003880192
  1 0.984455959 0.015544041
```

Listing 3.2 – Proporção de fraude das variáveis selecionadas.

Nota-se que na variável “cities” há uma diferença significativa entre as classes, no que diz respeito a proporção de fraude.

É possível observar que a cidade 3 que representa Fortaleza, tem a maior proporção de fraude, mas, por critério de escolha da autora, foi decidido apresentar uma análise somente com a cidade de Recife, da qual possui a proporção de 29% de fraude.

Embora a visualização gráfica auxilie a ter um norte de como as variáveis se comportam, mesmo assim se faz necessário uma investigação mais detalhada e para identificar

possíveis modus operandi da fraude, ou seja, fatores e comportamentos relacionados à ocorrência de fraude.

3.1.3 Recursos Computacionais

Utiliza-se o *software* R, versão 3.1.1 (R CORE TEAM, 2021) para realizar a aplicação de regressão logística.

3.2 Métodos

3.2.1 Criação de Cenários

Visando analisar de forma segregada os tipos de variáveis, foram criados 2 tipos de cenários para este estudo. O intuito é que por meio dessa separação possam ser identificado quais variáveis são mais significativas em relação a resposta, transação fraudulenta. Na criação dos cenários foi levado em conta que as variáveis possuem características específicas, atráves disso que foi realizada a segregação.

Tabela 2 – Cenários

Cenário 1	**Cenário 2**
amount	approved_1d
account_age	approved_7d
discount	approved_90d
fraud	approved_amount_1d
latitude	approved_amount_7d
longitude	approved_amount_90d
plataform	attemp_1d
prime	attemp_7d
retail	attempt_amount_1d
status	attempt_amount_7d
	beer
	cards_1d
	cards_7d
	cards_90d
	fraud
	order_attempt
	order_sec_duration
	spirit_drinks

Fonte: A autora (2023).

Conforme demonstrado na Tabela 2, no Cenário 1 foram utilizadas 10 variáveis, todas relacionadas a informações básicas coletadas no momento da transação, algumas

referentes a informações cadastrais do usuário, bem como, informações da transação. Enquanto no cenário 2, foram selecionadas 18 variáveis, estas relacionadas ao comportamento do usuário, como tempo de duração do pedido, se o pedido possui ou não cerveja ou destilados.

Vale lembrar, que a variável resposta "fraud"foi inserida em todos os cenários.

Essa segregação foi realizada afim entender qual tipo de variável está mais correlacionada com a existência de fraude, com a pretensão de buscar uma explicação se a fraude está mais relacionada ao comportamento do usuário ou às suas informações cadastrais e transacionais.

3.2.2 Tratamento de Valores Ausentes

Após a criação dos cenários, foi realizado por meio da função *summary* uma primeira análise descritiva das variáveis, que incluiu informações estatísticas, como média, mínimo, máximo e quartis. Além disso, foram identificados valores ausentes (NA) em algumas variáveis, que foram tratados de acordo com a escolha da autora.

No Cenário 1, foram realizados os seguintes tratamentos:

- Nas variáveis "*latitude*"e "*longitude*", por se tratar de informações de localização do usuário, para os casos de dados ausentes, os mesmo foram substituídos por zeros.
- Nas variáveis "*amount*"e "*discount*", por se tratarem de valores monetários, em caso de NA, os mesmo foram substituído pela média da variável.

Em contrapartida, no Cenário 2, todas as variáveis que continham NA foram substituída por zeros. Vale ressaltar que do total de 18 variáveis, 15 continham NA's.

Esse volume grande de NA's é consequência do tipo de variáveis inseridas nesse grupo, do qual pode ser considerado um comportamento normal, pois nem todos os usuários teriam cerveja em seu pedido variável "*beer*", ou nem todos os usuários teriam a quantidades de pedidos aprovados nos últimos 90 dias do usuário variável "*approved_90d*", como pode ocorrer esse tipo de situação, é compreensível o grande volume de NA's presente nessas variáveis.

3.2.3 Redução de dados desiquilibrados

Conforme citado anteriormente no item 4.1.2, a base possui um grande desiquilíbrio entre as classes de fraude e não fraude, diante disso, é necessário realizar uma manobra para que essa desigualdade diminua, afim de obter um treinamento eficaz dos modelos de detecção de fraudes. Uma conduta usual de resolver esse tipo de cenário, é a criação de subamostras balanceadas.

Mediante a problemática do desbalanceamento dos dados, foram geradas subamostras balanceadas a partir do conjunto de dados original, com o objetivo de igualar o número de casos de fraude aos casos de não fraude.

As subamostras foram subsequentemente divididas em conjuntos de treinamento e teste, com a equalização do desequilíbrio de classes aplicada apenas aos dados de treinamento. Onde 70% dos casos de fraude foram incorporados ao conjunto de treinamento, enquanto os 30% restantes foram alocados ao conjunto de teste. Essa abordagem estratégica possibilita uma melhor avaliação do desempenho dos modelos.

Por meio dessa estratégia, é possível minimizar o viés que é causado pelo desbaleançamento do conjunto de dados, possibilitando que o modelo de detecção de fraude seja avaliado de forma mais eficaz, pois esta equilibrando a quantidade de fraude com a de não fraude.

Ajustado o balanceamento da base, foi iniciado o processo de repetição aleatória para criação das subamostras, tanto para teste, quanto para treino. Nesta criação, 10.000 subamostras foram geradas para o treinamento, cada uma delas acompanhada por conjuntos de teste complementares, totalizando também 10.000.

3.3 Acurácia, Especificidade e Sensibilidade

A seguir exploraremos como as métricas Acurácia, Especificidade e Sensibilidade são aplicadas e como são úteis no contexto de Regressão Logística. Antes de entrar no detalhamento do assunto, é válido abordar o conceito de Verdadeiro Positivo e Verdadeiro Negativo para que dê insumo para o entendimento do cálculo de cada métrica:

- **Verdadeiro Positivo ou VP**: São os casos em que o modelo previu corretamente a classe positiva (geralmente a classe minoritária, como fraude por exemplo) corretamente.
- **Verdadeiros Negativos ou VN**: São os casos em que o modelo previu corretamente a classe negativa (geralmente a classe majoritária, como não fraude) corretamente.

3.3.1 Acurária

- **Importância**: Mede a proporção de previsões corretas em relação ao número total de previsões. É uma métrica geral que avalia o desempenho global do modelo.
- **Utilidade**: Muito útil quando as classes são balanceadas.
- **Limitações**: Quando há um desbalanceamento entre as classes, a acurácia pode ser enganosa. O modelo pode alcançar uma alta acurácia simplesmente prevendo a classe majoritária na maioria das vezes, enquanto a classe minoritária é frequentemente classifica de forma indevida

- **Como Calcular**: (Verdadeiros Positivos + Verdadeiros Negativos) / (Total de Observações)

3.3.2 Especificidade

- **Importância**: Mede a proporção de verdadeiros negativos em relação ao total de observações que eram realmente negativass. No contexto deste estudo, ela auxilia na identificação da proporção de verdadeiros negativos (transações legítimas corretamente identificadas como legítimas).
- **Utilidade**: Ela é importante quando o custo de classificar incorretamente negativos verdadeiros é alto, como diagnósticos na área da saúde por exemplo.
- **Como Calcular**: (Verdadeiros Negativos) / (Verdadeiros Negativos + Falsos Positivos).

3.3.3 Sensibilidade

- **Importância**: Mede a capacidade do modelo de identificar corretamente as instâncias positivas. No contexto deste estudo, ela auxilia a identificar proporção de VP (fraudes corretamente identificadas)
- **Utilidade**: Em problemas de detecção de eventos raros, como fraude por exemplo, a sensibilidade ajuda a avaliar o desempenho do modelo de identificar esses eventos.
- **Como Calcular**: Verdadeiros Positivos / (Verdadeiros Positivos + Falsos Negativos)

Em resumo, ambas as métricas desempenham papéis importantes na avaliação do modelo, mas a ênfase pode variar dependendo dos objetivos específicos do problema.

Em geral, é recomendado avaliar todas essas métricas juntas para obter uma compreensão completa do desempenho do modelo.

4 Resultados e Discussões

4.1 Ajuste dos Modelos

Por meio da função "*glm*"presente no software R, foi ajustado um modelo de regressão logística a cada uma das 10.000 subamostras em cada um dos dois diferentes cenários mencionados, resultando em 20.000 modelos ajustados.

Como descrito na seção 3.2.1, criaram-se cenários para avaliar o desempenho e eficiência dos modelos em relação às variáveis específicas de cada cenário. Nesse contexto, a velocidade de convergência dos modelos foi analisada.

No cenário 1, teve-se o custo computacional de 1,36 segundos por modelo ajustado, totalizando 1 minuto e 42 segundos para ajustar os 10.000 modelos. Foi observado também que para a convergência dos 10.000 modelos foram necessárias 18 iterações.

Enquanto no Cenário 2, teve o custo computacional de 2,58 segundos por modelo ajustado, totalizando 3 minutos e 59 segundos para ajustar os 10.000 modelos. Para a convergência dos 10.000 modelos foram necessárias 25 iterações.

4.1.1 Selecionando as variáveis significativas

Após realizar o ajuste dos 10.000 modelos em cada cenário, o próximo passo envolve a identificação das variáveis com maior relevância, ou seja, mais significativas. Para isso, foi utilizado a soma dos p-valores de cada variável em todas as iterações, resultando em uma métrica que demonstra a importância geral das variáveis de forma abrangente.

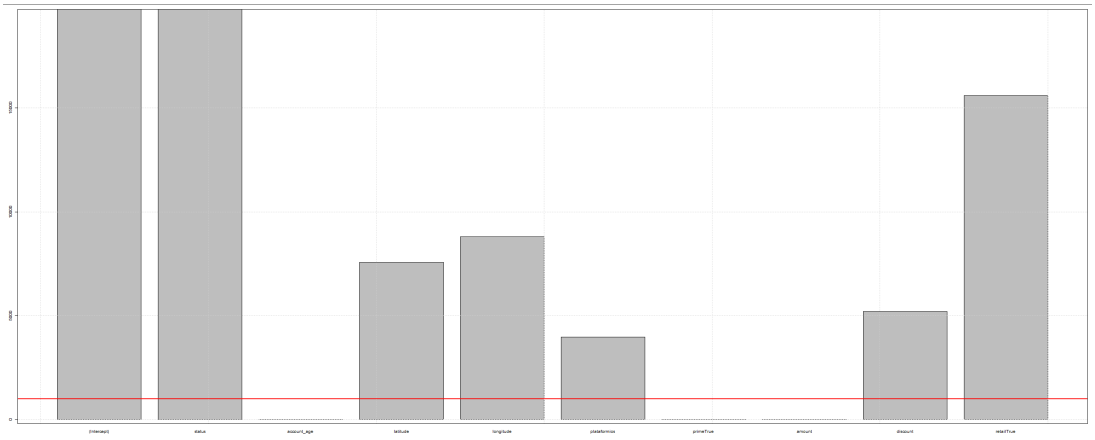
Essa seleção se faz necessária, pois por meio dela é possível identificar as variáveis mais pertinentes na detecção de transações fraudulentas, aprimorando a eficiência computacional e a compreensão dos modelos, evitando então, a inclusão de variáveis pouco informativas que não irão agregar no estudo.

No primeiro cenário, foi possível reduzir o número de variáveis de 10 para 3.

```
1 > summary(modelo_final1)
2 > significant_variables_mod1
3 account_age    primeTrue        amount
```

Listing 4.1 – Variáveis significativas Cenário 1.

Figura 9 – Variáveis significativas Cenário 1.



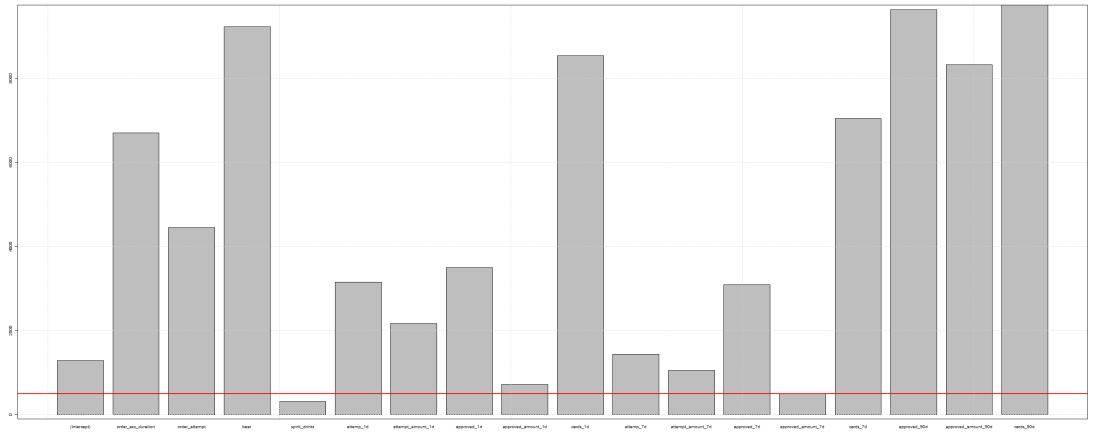
Fonte: A autora (2023).

No segundo cenário, a redução foi de 18 para 2 variáveis.

```
> significant_variables_mod2
      spirit_drinks approved_amount_7d
```

Listing 4.2 – Variáveis significativas.

Figura 10 – Variáveis significativas no Cenário 2.



Fonte: A autora (2023).

4.1.2 Ajustando os Modelos com as variáveis significativas selecionadas

O passo seguinte após a seleção de variáveis significativas, é realizar o ajuste do modelo com as mesmas.

Além do ajuste, os resultados do modelo devem ser analisados afim de entender se de fato essas variáveis podem esclarecer melhor o comportamento da resposta.

4.1.2.1 Cenário 1

Das 10 variáveis utilizadas no Cenário 1, foi realizada a seleção de 3 variáveis: *accountage*, *amount* e *prime*.

Foi feito uso da função *summary* para termos uma visão geral e interpretação estatística do modelo. Esta função nos da informações para analisar a significância das variáveis selecionadas e avaliar o ajuste global do modelo aos dados.

Abaixo serão demonstrados os resultados da saída desta função aplicada para o modelo com as variáveis significativas do Cenário 1:

```
> summary(modelo_final1)
Call:
glm(formula = fraud ~ ., family = binomial(link = "logit"), data = Cenario1
    _significant_mod1)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.3314407  0.2679896   1.237    0.216
account_age -0.0039937  0.0005839  -6.839 7.96e-12 ***
primeTrue   -1.6841657  0.3965610  -4.247 2.17e-05 ***
amount       0.0084436  0.0011691   7.222 5.12e-13 ***
---
Signif. codes: 0    ***      0.001    **      0.01    *    0.05    .    0.1
           1
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 460.25  on 331  degrees of freedom
Residual deviance: 235.77  on 328  degrees of freedom
AIC: 243.77
Number of Fisher Scoring iterations: 7
```

Listing 4.3 – Ajuste modelo final1 com as variáveis significativas selecionadas.

O coeficiente para o intercepto é de 0.331, mas o p-valor é 0.216, indicando que não é estatisticamente significativo a um nível de significância usual de 5%.

Na variável *amount*, o intercepto é 0.0084436, e p-valor de 0.000000000000512. Isso indica que há uma forte evidência estatística de que, mantendo as outras variáveis constantes, o aumento no valor da transação está positivamente associado à ocorrência de fraude. Em outras palavras, à medida que o valor da transação aumenta, a probabilidade de fraude também aumenta.

Temos um interpretação um pouco diferente para a variável *prime*, que tem intercepto de -1.6841657, e p-valor de 0.0000217. Ou seja, mantendo as outras variáveis constantes, a presença do status prime (True) está negativamente associada à ocorrência de fraude. Em comparação com contas prime (False), ou seja, contas prime têm menor probabilidade de fraude.

Na variável *accountage* com intercepto de -0.0039937, e um p-valor também muito próximo de zero, de 0.00000000000796. Pode-se notar que há uma forte evidência estatística de que, mantendo as outras variáveis constantes, o aumento no tempo da conta usuário está negativamente associado à ocorrência de fraude. Em termos práticos, à medida que a aumenta o tempo de conta do usuário, a probabilidade de fraude diminui.

Além de olhar para os resultados das variáveis significativas, é necessário também atentar-se para o valor do AIC, do qual é uma ferramenta útil na escolha de modelos, pois ajuda a evitar a seleção de modelos excessivamente complexos que podem se ajustar muito bem aos dados de treinamento, mas podem ter um desempenho inferior na previsão de novos dados. Ele auxiliar a promover a escolha de modelos que são eficazes na explicação dos dados observados sem serem excessivamente complexos.

No cenário 1 termos um AIC de 243.77, um valor relativamente baixo, que sugere que o modelo teve um ajuste relativamente bom.

Em resumo, as variáveis *accountage*, *prime* e *amount* parecem ser estatisticamente significativas para prever a ocorrência de fraude, com *accountage* e *amount* associados positivamente à fraude, enquanto $prime(True)$ está associado negativamente. O modelo que incorpora as variáveis significantes do Cenário 1 demonstra um ajuste eficaz aos dados.

Além da análise das medidas acima descritas, foi avaliado também o ganho computacional de um modelo somente com as variáveis significativas, e o ganho computacional foi de 25% na convergência dos 10.000 modelos, saindo de 1 min e 42 segundos para apenas 36 segundos.

4.1.2.2 Cenário 2

Assim como observado no Cenário 1, foram implementadas as mesmas medidas e procedimentos no Cenário 2. Este processo incluiu seleção das variáveis significativas. Posteriormente a essa etapa, das 18 variáveis inicialmente consideradas neste cenário, apenas duas, $approved\_amount\_7d$ e $spirit\_drinks$, foram identificadas como significativas, conforme representado de maneira ilustrada na Figura 10.

A análise dessas variáveis selecionadas foi realizada por meio da função *summary*, da qual resultaram informações detalhadas sobre cada uma delas. Esse procedimento permitiu a obtenção de ideias relevantes para o entendimento do impacto dessas variáveis no contexto do Cenário 2.

```
> summary(modelo_final2)

Call:
glm(formula = fraud ~ ., family = binomial(link = "logit"), data = Cenario2
    _significant_mod2)

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)        -0.8283911  0.1485779  -5.575 2.47e-08 ***
spirit_drinks       1.3568435  0.3526765   3.847 0.000119 ***
approved_amount_7d  0.0005593  0.0001555   3.598 0.000321 ***
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1
            1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 460.25  on 331  degrees of freedom
Residual deviance: 350.08  on 329  degrees of freedom
AIC: 356.08

Number of Fisher Scoring iterations: 8
```

Listing 4.4 – Ajuste modelo final2 com as variáveis significativas selecionadas

Olhando o resultado da saída, podemos observar que o valor do intercepto é -0.8283. Isso sugere que, quando todas as outras variáveis são zero, há uma tendência negativa na ocorrência de fraude.

Olhando com foco nos resultados da variável *spiritdrinks*, a estimativa do coeficiente é 1.3568, com um p-valor muito baixo 0.000119. Isso indica que um aumento na quantidade de spirit-drinks está associado a um aumento significativo nas chances de a transação ser fraudulenta. Pedidos contendo bebida destilada, têm uma probabilidade consideravelmente maior da transação ser fraudulenta.

Já no caso da variável *approved_amount_7d*, a estimativa do coeficiente é 0.0005593, e o p-valor é relativamente alto 0.000321. Isso sugere que, estatisticamente, o valor dos pedidos aprovados nos últimos 7 dias é um preditor significativo para a ocorrência de fraude no contexto deste modelo. Em termos mais simples, o valor dos pedidos aprovados nos últimos 7 dias parece ter um impacto estatisticamente significativo na probabilidade de fraude.

O AIC é 356.08, indica que este modelo é razoavelmente bom em termos de ajuste e complexidade.

Portanto, mesmo a variável *spiritdrink* e *approved_amount_7d* tenha um impacto significativo na detecção de fraude, a primeira impressão que temos é que o modelo

não performa bem para o fim esperado, detecção de fraude, resultando possivelmente em futuros ajustes, porém isso ainda será confirmado com as demais análises.

Além da análise das medidas acima descritas, foi avaliado também o ganho computacional do modelo somente com as variáveis significativas, e o ganho computacional foi de 11% na convergência dos 10.000 modelos, saindo de 3 minutos e 59 segundos para apenas 41,90 segundos.

4.1.3 Métricas do desempenho dos Modelos

Conforme explicado na seção 3.3, métricas de Acurácia, Especificidade e Sensibilidade desempenham papéis importantes na avaliação do modelo, portanto, abaixo será apresentado o resultado dessas métricas em relação a cada modelo de cada Cenário:

```
> print(metrics_table)
      Metrica          Valor
1 Sensibilidade    0.8734940
2 Especificidade   0.8132530
3 Acuracia         0.8433735
```

Listing 4.5 – Métricas de desempenho do modelo-final1.

```
> print(metrics_table2)
   Metrica           Valor
1 Sensibilidade    0.5783133
2 Especificidade   0.8975904
3 Acuracia         0.7379518
```

Listing 4.6 – Métricas de desempenho do modelo-final2.

Assim como foi realizado o estudo com cenários distintos, é importante que haja uma comparação do desempenho entre eles.

A análise das variáveis significativas, valor do intercepto e p-valor, são observações importantes que sempre devem ser feitas, pois seus valores nos dão norte do quanto cada variável está influenciando a variável resposta, assim, entendendo o quanto essas variáveis independentes são diretamente responsáveis pela ocorrência ou não de fraude, que é o objetivo deste estudo.

Embora tenha sido realizada essa análise das variáveis significativas de cada cenário, é importante analisar também outras métricas, como acurácia, especificidade e sensibilidade. Essas métricas nos dão uma avaliação e compreensão mais abrangente do desempenho do modelo, e não apenas das variáveis indepentes.

Será realizada essa comparação de forma segregada nas seções a seguir.

4.1.3.1 Cenário 1

Neste cenário, temos a acurácia do modelo com valor de 0.8433735, indicando que ele classifica corretamente aproximadamente 84,3% de todas as transações (fraudulentas e não fraudulentas).

A sensibilidade do modelo, nos retorna 0.8734940, ou seja, significa que o modelo identifica corretamente 87,3% das transações fraudulentas. Em outras palavras, quando uma transação é realmente fraudulenta, o modelo tem uma alta taxa de detecção, que é bom ponto, pois detectar quando uma transação realmente é fraude é um dos grandes desafios.

Já a especificidade do modelo, com valor de 0.8132530, indica que o modelo classifica corretamente 81,3% das transações não fraudulentas. Em termos simples, quando uma transação é legítima, o modelo tem uma boa taxa de identificação. Este é outro resultado extremamente importante, pois quando queremos mitigar a fraude, o intuito é que as transações legítimas não sejam impactadas pelas estratégias e medidas de prevenção e combate a fraude. Ter um indicador identificando 81,3% que estas transações legítimas não serão impactadas, transmite uma certa segurança em prosseguir nessa direção para prevenir e combater a fraude.

Portanto, levando em conta que a acurácia é relativamente alta, sugerindo um bom desempenho geral do modelo, que a sensibilidade alta indica que o modelo é eficaz na identificação de transações fraudulentas, e a especificidade também é razoavelmente alta, indicando uma boa capacidade de distinguir transações legítimas das fraudulentas, pode-se inferir que o modelo parece ter um desempenho aceitável para a detecção de fraude.

4.1.3.2 Cenário 2

Neste Cenário, temos a acurácia do modelo com valor de 0.7379518, indicando que ele classifica corretamente aproximadamente 73% da proporção total de previsões corretas feitas pelo modelo, considerando tanto as transações fraudulentas quanto as não fraudulentas.

Já na sensibilidade, temos o valor de 0.5783133, que indica que o modelo identificou corretamente cerca de 57% das transações fraudulentas.

Quanto a especifidade, com retorno de 0.8975904, sugere que o modelo identificou corretamente cerca de 89% das transações não fraudulentas.

O modelo tem uma sensibilidade de 57%, o que significa que está identificando aproximadamente apenas metade das transações fraudulentas. Isso é um ponto de atenção e indica que há espaço para melhorias na identificação de transações fraudulentas neste modelo. Dependendo das necessidades específicas, pode ser interessante ajustar o modelo para otimizar essa sensibilidade, a fim de alcançar um equilíbrio adequado em relação aos objetivos e às consequências de falsos positivos e falsos negativos no contexto da detecção

de fraudes.

Por fim, é importante ressaltar que é sempre útil considerar essas métricas em conjunto para ter uma compreensão abrangente do desempenho do modelo, especialmente em cenários de detecção de fraudes, nos quais a classe de interesse (fraude) pode ser substancialmente menor em comparação com a classe de transações não fraudulentas.

5 Considerações Finais

De maneira geral, este trabalho foi essencial para empregar os conhecimentos adquiridos durante o curso de Estatística. Ele proporcionou uma experiência significativa, permitindo uma exploração mais profunda e abrangente das técnicas estudadas.

O propósito deste estudo foi aplicar o modelo de Regressão Logística em um contexto real de detecção de fraudes em transações online de cartão de crédito, afim de analisar e entender os fatores que levam a explicação da variável resposta, a ocorrência da fraude. Para que o desenvolvimento das análises se aproximasse da realidade, o grupo de variáveis foi dividido, criando os dois Cenários com informações bem específicas em cada um, conforme demonstrado no decorrer deste trabalho.

Considerando os resultados dos dois cenários, observa-se que o modelo no Cenário 1 apresenta desempenho mais sólido na detecção de fraudes em comparação ao Cenário 2. No Cenário 1, a acurácia (84,3%) indica uma boa capacidade de classificação global, enquanto a sensibilidade alta (87,3%) destaca a eficácia na identificação de transações fraudulentas. Além disso, a especificidade robusta (81,3%) sugere que o modelo também é eficiente na distinção de transações não fraudulentas.

No entanto, no Cenário 2, apesar de uma acurácia razoável (73%), a sensibilidade mais baixa (57%) indica uma limitação na identificação de transações fraudulentas, sugerindo espaço para melhorias nesse aspecto. A alta especificidade (89%) destaca a capacidade de identificar corretamente transações não fraudulentas, mas a sensibilidade reduzida é um fator crítico. No entanto, vale ressaltar que se faz necessária uma abordagem cautelosa ao interpretar tais resultados e ter em mente que o comportamento das variáveis pode impactar diretamente o desempenho do modelo. No Cenário 2 por exemplo, o número de NA's era expressivo, talvez uma outra abordagem de tratamento resultaria em um outro desempenho.

Outro fator importante que deve ser levado em conta é análise do AIC, do qual é a medida que leva em consideração o ajuste do modelo e o número de parâmetros. Modelos com menor AIC são geralmente preferíveis, neste caso, seria o Cenário 1 com 243.77, contra 356.08 do Cenário 2.

Embora o modelo do Cenário 1 demonstre um desempenho mais consistente na detecção de fraudes, o Cenário 2 aponta para a necessidade de refinamento, especialmente na sensibilidade, para otimizar a identificação de transações fraudulentas. A escolha entre os dois modelos dependerá das prioridades e requisitos específicos do contexto em que serão aplicados.

Uma possível abordagem para estudos futuros seria a aplicação da regressão logística exata. Essa técnica refere-se à inferência condicional em dados binomiais modelados por uma regressão logística, sendo robusta independentemente do tamanho reduzido ou

do desequilíbrio no conjunto de dados. À medida que a amostra aumenta ou os dados se tornam mais equilibrados e menos esparsos, a solução derivada usando a abordagem convencional para grandes conjuntos de dados coincidirá com aquela obtida pela abordagem exata.

Portanto, seria benéfico aplicar a mesma abordagem empregada no estudo a uma base de dados mais robusta, que contenha informações adicionais, como endereço, possibilitando uma análise das regiões de risco, e-mail, entre outros. A seleção do modelo adequado deve considerar diversos elementos, incluindo o leque de variáveis disponíveis, a eficiência computacional e a capacidade de resposta em tempo real. Tendo em vista esses aspectos, torna-se crucial adaptar a estratégia de modelagem às particularidades e requisitos específicos do cenário em questão. Essa abordagem mais precisa alinhada aos objetivos da organização, é essencial para garantir resultados eficazes e relevantes.

Referências Bibliográficas

BOLTON, R. J.; HAND, D. J. Statistical fraud detection: A review. *Statistical science*, Institute of Mathematical Statistics, v. 17, n. 3, p. 235–255, 2002.

CHAN, P. K. et al. Distributed data mining in credit card fraud detection. *IEEE Intelligent Systems and Their Applications*, IEEE, v. 14, n. 6, p. 67–74, 1999.

CRAMER, J. Scoring bank loans that may go wrong: a case study. *Statistica Neerlandica*, Wiley Online Library, v. 58, n. 3, p. 365–380, 2004.

HOSMER, D. W. et al. A comparison of goodness-of-fit tests for the logistic regression model. *Statistics in medicine*, Wiley Online Library, v. 16, n. 9, p. 965–980, 1997.

Houassis. Siginificado de fraude. 2023. Acessado em 11 de Outubro de 2023. Disponível em: <https://houaiss.uol.com.br/corporativo/apps/uol_www/v6-1/html/index.php#1>.

Konduto. Censo da fraude 2023. 2023. Acessado em 01 de Novembro de 2023. Disponível em: <https://content.konduto.com/censo-da-fraude-2023>.

R CORE TEAM. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2021. Disponível em: <www.R-project.org/>.

Serasa Experian. Brasil sofreu uma tentativa de fraude a cada 11 segundos em abril. 2023. Acessado em 02 de Novembro de 2023. Disponível em: <https://www.serasaexperian.com.br/sala-de-imprensa/prevencao-a-fraude/brasil-sofreu-uma-tentativa-de-fraude-a-cada-11-segundos-em-abril-mostra-serasa-experian/>.