UNIVERSIDADE FEDERAL DO PARANÁ
SETOR DE CIÊNCIAS EXATAS
DEPARTAMENTO DE ESTATÍSTICA
CURSO DE ESTATÍSTICA

Fábio Fernandes Pereira

Identificação de fraude em transações online de cartão de crédito

CURITIBA
2023

UNIVERSIDADE FEDERAL DO PARANÁ
SETOR DE CIÊNCIAS EXATAS
DEPARTAMENTO DE ESTATÍSTICA
CURSO DE ESTATÍSTICA

Fábio Fernandes Pereira

Identificação de fraude em transações online de cartão de crédito

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

Orientador: Prof. Dr. Fernando Lucambio

CURITIBA
2023

AGRADECIMENTOS

Em primeiro, gostaria de expressar meu sincero agradecimento a minha esposa e meu filho, pelo constante apoio, paciência e compreensão demonstrados ao longo de toda a minha jornada acadêmica.

Não posso deixar de mencionar meus amados pais e irmãs, cujo apoio e amor têm sido uma fonte de força e inspiração em todas as fases da minha vida. Seus ensinamentos valiosos moldaram meu caráter e me guiaram até este momento. Sou profundamente grato por tudo o que vocês fizeram e continuam fazendo por mim.

Expresso minha gratidão aos professores que passaram por minha trajetória acadêmica, pois cada um contribuiu para o meu desenvolvimento intelectual e pessoal. Em particular, gostaria de agradecer ao Dr. Lucambio, que gentilmente se dispôs a orientar e oferecer suporte durante esta etapa final da minha graduação. Seus conselhos e conhecimento foram muito importantes para a conclusão bem-sucedida deste trabalho.

Por fim, expresso meu profundo apreço a todos que fazem parte da minha vida, pois cada pessoa que cruzou meu caminho contribuiu de alguma forma para a pessoa que me tornei hoje. A todos vocês, meu mais sincero agradecimento.

"A educação é a arma mais poderosa que você pode usar para mudar o mundo."

(Nelson Mandela)

RESUMO

As fraudes em transações online feitas com cartão de crédito ocorrem quando um indivíduo utiliza dados de terceiros para realizar a transação, sem que o detentor do cartão esteja ciente dessa operação. Isso acontece porque, em compras online, não há meios de validação pessoal, como senha ou assinatura. Muitas vezes, apenas as informações descritas no cartão de crédito são suficientes para essa operação.

De acordo com a pesquisa realizada pela ACI Worldwide (2019), em 2020, foi relatado um total de 149 milhões de dólares em transações fraudulentas feitas com cartão de crédito. Esses custos são de responsabilidade dos lojistas que realizam a operação de venda. Essa responsabilidade foi definida pelas bandeiras dos cartões, que entendem que o vendedor é responsável por validar a autenticidade de seu cliente antes de realizar uma venda. No entanto, no mundo digital, não é tão simples como no mundo físico, em que o vendedor pode verificar as marcas d'água em uma cédula de dinheiro ou pedir para o cliente digitar uma senha pessoal. Portanto, é preciso usar os dados que se tem à mão e, em alguns casos, decidir se aprovará ou não uma compra em frações de segundo. Por isso, é necessário ter modelos ou algoritmos eficientes para conseguir aprovar transações de clientes reais e negar transações de fraudadores. (ACI WORLDWIDE, 2019).

Durante o pequeno intervalo de tempo que leva entre o momento em que você confirma uma compra online e essa mesma compra aparece em seu cartão de crédito, muitas validações e cruzamentos de dados estão sendo feitos, incluindo a validação de fraude. Essa validação normalmente é feita por meio de algoritmos de Machine Learning ou modelos estatísticos clássicos, que, com uma série de dados coletados do comprador, tentarão estimar se ele é um cliente real ou um fraudador.

O objetivo deste trabalho é testar modelos estatísticos na predição de transações fraudulentas. Para isso, foram utilizados dados reais e ajustados modelos de regressão logística e modelos de random forest. Esses modelos foram avaliados em termos de sua capacidade de identificar corretamente transações fraudulentas, e seus respectivos custos atrelados.

Palavras-chave: Estatística descritiva. Regressão Logística. Random Forest. Modelos de identificação de fraudes. Fraude transacional.

Sumário

AGRADECIMENTOS iii
RESUMO v
1 INTRODUÇÃO 8
2 REVISÃO DE LITERATURA 10
2.1 Regressão Logística 10
2.2 Random Forest 12
3 MATERIAL E MÉTODOS 13
3.1 Material 13
3.1.2 Recursos Computacionais 20
3.2 Métodos 21
3.2.1 Balanceamento de sub-amostragem 24
4 RESULTADOS E DISCUSSÃO 27
4.1 Modelo de Regressão Logística 27
4.1.1 Ajuste dos Modelos 27
4.1.2 Seleção das variáveis 29
4.1.3 Ajuste dos Modelos com variáveis significativas 31
4.2 Modelo de Random Forest 34
4.2.1 Ajuste dos Modelos com variáveis significativas 34
4.2.2 Ajuste dos Modelos com todas as variáveis 36
4.3 Comparação dos resultados, Regressão Logística e Random Forest 38
4.3.1 Cenário 1 38
4.3.2 Cenário 2 40
4.3.3 Cenário 3 42
4.3.4 Cenário 4 43
4.4 Comparação dos cenários no Modelo Random Forest 44
5 CONSIDERAÇÕES FINAIS 45
REFERÊNCIAS 48

1 INTRODUÇÃO

A segurança nas transações online de cartão de crédito é um assunto cada vez mais relevante em um mundo cada vez mais digital. De acordo com o relatório "Global Payment Fraud and Security Survey" da ACI Worldwide (2019), 46% dos entrevistados em todo o mundo relataram ter sofrido fraude em pagamentos eletrônicos. Com o aumento das compras realizadas pela internet, também cresce o número de fraudes cometidas por indivíduos mal-intencionados. Por isso, a identificação de fraudes em transações online de cartão de crédito é um tema crucial para garantir a segurança das transações financeiras.

A Estatística desempenha um papel fundamental ao fornecer métodos e técnicas para a detecção e prevenção desses crimes. A análise estatística de transações de cartão de crédito pode fornecer insights úteis para detectar fraudes em tempo real. Além disso, algoritmos de aprendizado de máquina têm sido cada vez mais utilizados para a detecção de fraudes em transações online de cartão de crédito. Esses algoritmos têm a capacidade de analisar grandes quantidades de dados em tempo real e identificar anomalias que podem indicar atividades fraudulentas.

Neste contexto, o presente trabalho tem como objetivo apresentar um estudo sobre a identificação de fraudes em transações online de cartão de crédito, utilizando técnicas estatísticas e algoritmos de aprendizado de máquina. Foi realizada uma comparação entre dois modelos amplamente utilizados na detecção de fraudes em transações de cartão de crédito: Random Forest e Regressão Logística. O modelo Random Forest é um algoritmo de aprendizado de máquina que cria várias árvores de decisão independentes para classificar as transações como fraudulentas ou legítimas. Já a Regressão Logística é uma técnica

estatística que analisa a relação entre variáveis independentes e uma variável dependente, neste caso, a classificação de uma transação como fraude ou não. Ao comparar esses dois modelos, foi possível avaliar a eficácia de cada um na identificação de fraudes em transações online de cartão de crédito, dentro dos cenários propostos.

2 REVISÃO DE LITERATURA

2.1 Regressão Logística

A regressão logística é um modelo estatístico utilizado para analisar a relação entre uma variável dependente binária (0 ou 1) e um conjunto de variáveis independentes. Ao longo do tempo, o modelo de regressão logística tem sido amplamente utilizado em diversas áreas, como medicina, marketing, finanças, entre outras.

Uma das primeiras referências sobre a regressão logística foi publicada por Joseph Berkson em 1944, em um artigo que discutia o uso de modelos estatísticos para avaliar o efeito de tratamentos em pacientes. Posteriormente, o modelo foi amplamente utilizado em estudos de epidemiologia, em que a variável dependente binária é frequentemente a presença ou ausência de uma doença.

Nos anos 60 e 70, com o advento dos computadores, a regressão logística tornou-se mais acessível e começou a ser utilizada em estudos de diversas áreas, incluindo psicologia, sociologia e ciência política. Naquela época, o modelo era frequentemente utilizado para estimar a probabilidade de um indivíduo pertencer a uma determinada categoria, com base em um conjunto de variáveis independentes.

Desde então, o modelo de regressão logística evoluiu e se tornou uma ferramenta ainda mais poderosa. Nos anos 90, por exemplo, a regressão logística foi amplamente utilizada em estudos de marketing para estimar a probabilidade de um consumidor comprar um determinado produto, com base em variáveis como idade, gênero, renda e preferências de consumo. Desde então, o modelo de regressão logística tem sido aplicado em diversas áreas, incluindo ciências sociais, medicina, biologia, economia, entre outras.

A regressão logística é uma técnica estatística utilizada para modelar a relação entre uma variável dependente binária (ou seja, uma variável que pode ter apenas dois valores possíveis) e uma ou mais variáveis independentes (também conhecidas como covariáveis). A técnica é aplicável quando a variável dependente representa um resultado categórico, como sucesso ou fracasso, aprovação ou reprovação, vida ou morte, entre outros.

O modelo de regressão logística é uma técnica estatística utilizada para modelar a relação entre uma variável binária dependente e um conjunto de variáveis independentes. Ele usa a função logística, também conhecida como função sigmoidal, para calcular a probabilidade de que um evento ocorra, dado um conjunto de valores das variáveis independentes. A função logística transforma a combinação linear das variáveis independentes em um valor entre 0 e 1, representando a probabilidade estimada. Em seguida, é aplicado um limiar de decisão para classificar as observações em uma das duas categorias possíveis. O modelo é ajustado aos dados através da maximização da verossimilhança, encontrando os coeficientes que melhor se ajustam aos dados observados. Esses coeficientes podem ser interpretados como as mudanças logarítmicas nas chances de sucesso associadas a um aumento unitário nas variáveis independentes. A probabilidade do um evento ocorrer na regressão logística é dado por:

$$P(Y = 1|X) = \frac{1}{1 + e^{(-X\beta)}}$$

onde P(Y=1|X) representa a probabilidade de que a variável dependente Y assume o valor 1 dado um conjunto de valores de covariáveis X, β

representa o vetor de coeficientes de regressão e e é a função exponencial.

Ao longo do tempo, foram desenvolvidas várias técnicas para melhorar o desempenho do modelo de regressão logística. Por exemplo, os métodos de regularização Lasso e Ridge podem ser utilizados para evitar a multicolinearidade (ou seja, alta correlação entre as variáveis independentes), enquanto o método de bootstrap pode ser utilizado para avaliar a incerteza dos coeficientes de regressão.

Em resumo, o modelo de regressão logística é uma ferramenta importante e versátil na análise de dados, que tem sido amplamente utilizada em diversas áreas ao longo do tempo. Com o avanço das técnicas de aprendizado de máquina e o crescente volume de dados disponíveis, espera-se que o modelo de regressão logística continue a evoluir e se tornar ainda mais poderoso.

2.2 Random Forest

O modelo de Random Forest (Floresta Aleatória) é um algoritmo de aprendizado de máquina que se baseia na construção de múltiplas árvores de decisão para estimar uma variável dependente. Cada árvore é construída a partir de uma amostra aleatória (com reposição) do conjunto de dados e das variáveis independentes, neste estudo foi definido que o valor de 100 árvores e, em cada cenário, utilizou-se a raiz quadrada no número de covariáveis para o ajuste do modelo, o que o torna menos suscetível a overfitting (sobreajuste).

O modelo de Random Forest foi proposto por Leo Breiman em 2001 como uma extensão do algoritmo de Árvore de Decisão. Desde então, o modelo de Random Forest tem sido amplamente utilizado em diversas

áreas, como ciência de dados, bioinformática, finanças, marketing, entre outras.

Uma das principais vantagens do modelo de Random Forest é sua capacidade de lidar com conjuntos de dados grandes e complexos, com muitas variáveis independentes. Além disso, o modelo é capaz de lidar com dados faltantes e valores discrepantes (outliers).

Em termos de aplicações, o modelo de Random Forest tem sido amplamente utilizado em diversas áreas, incluindo bioinformática, análise de imagem, análise financeira, reconhecimento de voz, detecção de fraudes, entre outras.

Uma das principais vantagens do modelo de Random Forest é sua capacidade de lidar com dados de alta dimensionalidade, variáveis categóricas e numéricas, além de lidar bem com dados ausentes. Além disso, o modelo é capaz de fornecer informações sobre a importância das variáveis para a classificação, o que pode ser útil na seleção de variáveis para modelos posteriores.

No geral, o modelo de Random Forest tem sido bem recebido pela comunidade de aprendizado de máquina, devido à sua facilidade de uso e desempenho.

3 MATERIAL E MÉTODOS

3.1 Material

3.1.1 Conjunto de Dados

O conjunto de dados utilizado é composto por informações reais provenientes de uma empresa de delivery, especializada em vendas online

por meio de um aplicativo. Os dados são compostos por transações financeiras, realizadas na cidade de Fortaleza nos primeiros 5 meses de 2022, representando tanto as compras realizadas pelos clientes quanto as realizadas pelos fraudadores, sendo que a variável "fraud" é utilizada para identificar as transações fraudulentas. Além desta, há outras 48 variáveis, 16 relacionadas ao comportamento do usuário no momento da compra, tais como quantidades de cartões diferentes utilizados e tempo de navegação no APP durante a compra. Há também 6 variáveis sobre a compra em si, como valor, se há itens de alto risco no pedido e percentual de desconto utilizado, e mais 5 variáveis relacionadas ao usuário, como localidade da entrega, tempo de cadastro no APP e sistema operacional do dispositivo utilizado. Foram criadas mais 21 variáveis a partir da parte local do endereço de e-mail do usuário, como a quantidade de caracteres especiais repetidos em sequência e o total de caracteres numéricos.

O conjunto de dados utilizado é desbalanceado, composto por 26.958 transações, das quais 499 estão marcadas como fraude, no entanto, é importante ressaltar que há dois fatores que precisam ser levados em conta na análise dos resultados.

O primeiro fator é que existem casos em que o detentor do cartão não entra em contato com o emissor do cartão para contestar uma transação fraudulenta, o que pode levar a casos de fraudes não marcadas no conjunto de dados. O segundo fator é que pode haver pedidos marcados como fraude, mas que na realidade não são fraudes, simplesmente porque o detentor do cartão contestou alguma outra transação com o emissor do cartão, e o emissor acabou contestando todas as transações de um período.

Quadro 1 – Descrição das variáveis disponíveis no conjunto de dados

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
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
email_count_alpha	total de letras da parte local do e-mail
email_count_alpha_max_rep	número da maior ocorrência da mesma letra da parte local do e-mail
email_count_alpha_seq_rep	quantidade de letras repetidas em sequência (maior sequência) da parte local do e-mail
email_count_num	total de números da parte local do e-mail
email_count_num_max_rep	número da maior ocorrência do mesmo número da parte local do e-mail
email_count_num_seq_rep	quantidade de números repetidos em sequência (maior sequência) da parte local do e-mail
email_count_seq_num	quantidade de números da maior sequência (maior sequência) da parte local do e-mail
email_count_seq_sp	quantidade de caracteres especiais em sequência (maior sequência) da parte local do e-mail
email_count_sp	total de caracteres especiais da parte local do e-mail
email_count_sp_max_rep	número da maior ocorrência do mesmo caractere especial da parte local do e-mail
email_count_sp_rep	quantidade de caracteres especiais repetidos em sequência (maior sequência) da parte local do e-mail
email_diff_alpha	quantidade de letras diferentes da parte local do e-mail
email_diff_num	quantidade de números diferentes da parte local do e-mail
email_diff_sp	quantidade de caracteres especiais diferentes da parte local do e-mail
email_end_num	"1" se termina com números a parte local do e-mail

email_end_sp	"1" se termina com caracteres especial a parte local do e-mail
email_ini_num	"1" se começa com números a parte local do e-mail
email_ini_sp	"1" se inicia com caractere especial a parte local do e-mail
email_qtd_seq_alpha	quantidade de sequência de letras da parte local do e-mail
email_qtd_seq_num	quantidade de sequência de números da parte local do e-mail
email_qtd_seq_sp	quantidade de sequência de caracteres especiais da parte local do e-mail
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_sec_duration	tempo em segundos entre o começo da compra e o fim
plataform	sistema operacional do dispositivo (Android/IOS)
prime	classificação do cliente na plataforma
retail	seguimento do estabelecimento de venda
spirit_drinks	se bebidas destiladas estão presentes no carrinho de compra
timezone	data e hora da transação

Fonte: Fabio Fernandes Pereira (2023).

Na Figura 1, encontra-se a distribuição de todas as transações por meio da utilização das coordenadas de longitude e latitude correspondentes. Nesse contexto, os pontos azuis representam as transações não marcadas como fraudulentas, enquanto que os pontos vermelhos representam as transações marcadas como fraudulentas. É possível observar que, de modo geral, as fraudes estão amplamente distribuídas pelo mapa, havendo apenas uma pequena concentração na região sudeste.

Quadro 1 – Descrição das variáveis disponíveis no conjunto de dados
Quadro 1 – Descrição das variáveis disponíveis no conjunto de dados

Figura 1 – Observações dispostas de acordo com as localizações geográficas.

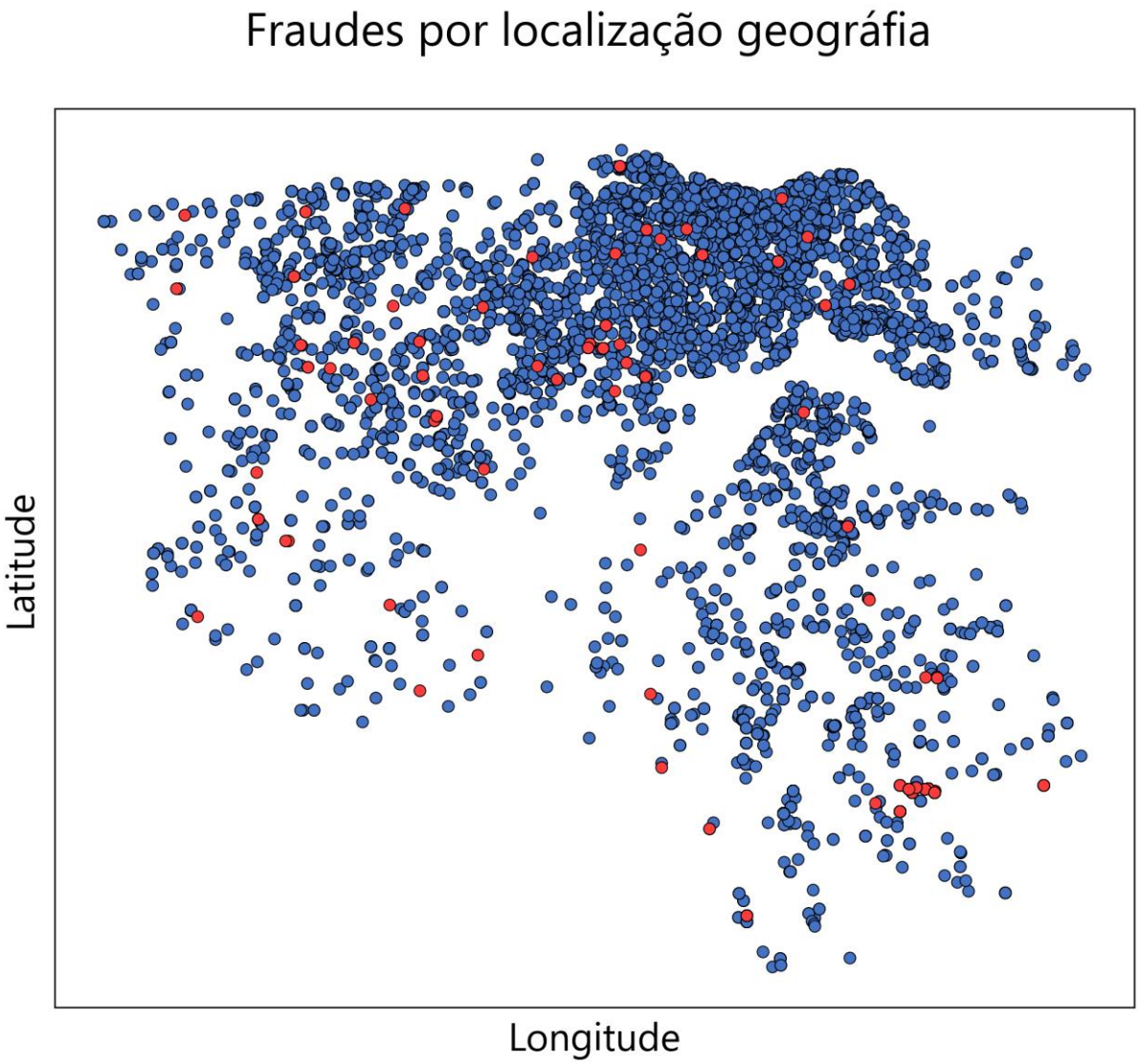

Figura 1

A Figura 2 apresenta a porcentagem geral de transações fraudulentas, bem como a porcentagem de transações fraudulentas segmentadas pelas variáveis dicotômicas.

Figura 2 – Índice de fraude.

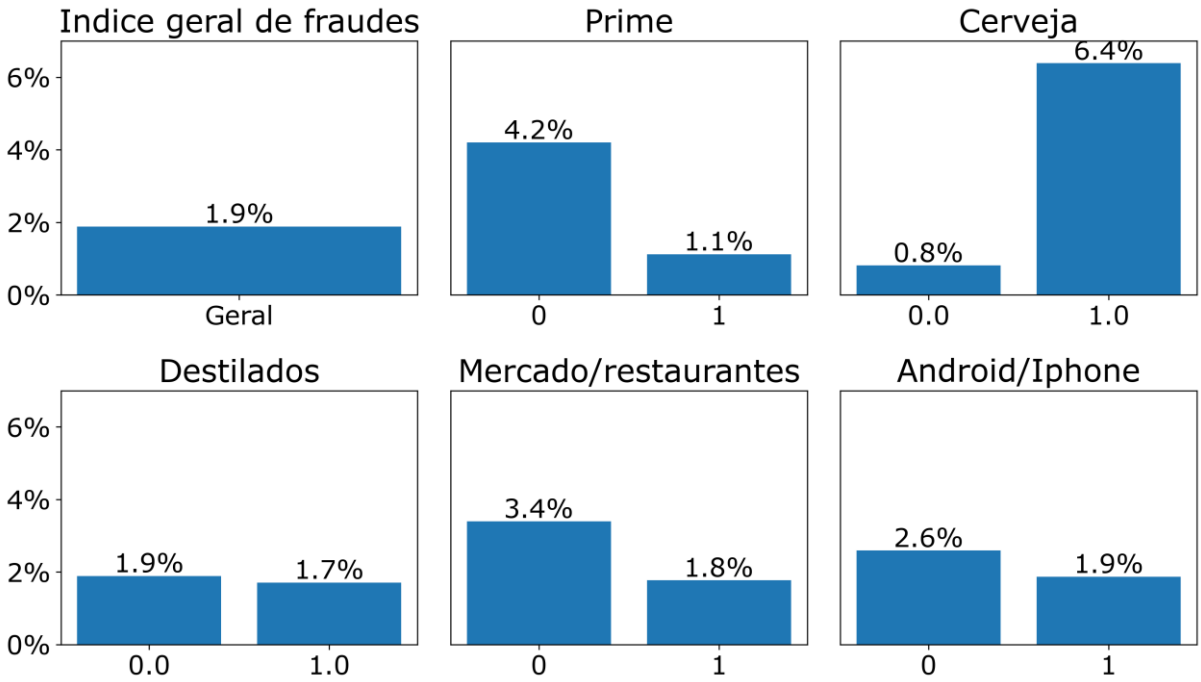

Figura 2.

A Figura 3 a seguir representa a correlação entre algumas covariáveis em relação à variável "fraude", com o objetivo de identificar qualquer concentração significativa de fraudes nas diferentes combinações dessas covariáveis. No entanto, observa-se que não há nenhuma combinação que apresente uma concentração relevante de fraudes.

Figura 3 – Relação entre as variáveis.

Figura 3.

A análise da correlação entre as covariáveis e a variável "fraud" é uma etapa crucial na investigação de possíveis padrões ou tendências relacionadas a atividades fraudulentas. Ao examinar a Figura 3, é possível visualizar os diferentes níveis de associação entre as covariáveis selecionadas e a ocorrência de fraude.

No entanto, com base na Figura 3, não é evidente que haja uma combinação específica de covariáveis que esteja fortemente correlacionada com a variável "fraude". Não há uma concentração distinta

de pontos que indique uma relação direta entre as covariáveis e a presença de fraudes.

Em suma, com base na Figura 4 apresentada, não há indícios claros de alguma combinação de covariáveis que exiba uma concentração relevante de fraudes. Essa análise sugere a necessidade de uma investigação mais abrangente e rigorosa para identificar possíveis fatores relacionados à ocorrência de fraudes.

3.1.2 Recursos Computacionais

O *software* Spyder (Python), versão 3.9.13 (main, Aug 25 2022, 23:51:50) [MSC v.1916 64 bit (AMD64)], foi utilizado para manipular os dados e ajustar os modelos aos dados descritos por meio dos pacotes: pandas, versão 1.4.4, matplotlib, versão 3.5.2, seaborn, versão 0.11.2., numpy, versão 1.23.5, sklearn, versão 1.0.2.

O software R, versão 4.2.1 (2022-06-23 ucrt), foi utilizado para obtenção das variáveis de componentes principais.

3.2 Métodos

Foi conduzida uma análise comparativa entre dois métodos amplamente difundidos para detecção de fraudes em transações de cartão de crédito: o modelo Random Forest e o modelo de Regressão Logística. Esses modelos foram aplicados em quatro diferentes cenários, diferenciando-se pelas suas covariáveis, as quais são apresentadas na Tabela 1:

Cenário 1	Cenário 2	Cenário 3
latitude	order_sec_duration	email_count_alpha
longitude	order_attempt	email_count_num
prime	account_age	email_count_sp
amount	attemp_1d	email_diff_alpha
discount_rate	attempt_amount_1d	email_diff_num
retail	approved_1d	email_diff_sp
beer	approved_amount_1d	email_count_seq_num
spirit_drinks	cards_1d	email_count_seq_sp
hour	attemp_7d	email_count_sp_rep
weekday	attempt_amount_7d	email_count_alpha_seq_rep
android	approved_7d	email_count_num_seq_rep
	approved_amount_7d	email_count_alpha_max_rep
	cards_7d	email_count_num_max_rep
	approved_90d	email_count_sp_max_rep
	approved_amount_90d	email_qtd_seq_alpha
	cards_90d	email_qtd_seq_num
		email_qtd_seq_sp
		email_ini_num
		email_end_num
		email_ini_sp
		email_end_sp

Tabela 1

A análise comparativa dos dois modelos foi realizada considerando quatro cenários distintos. Cada cenário é caracterizado por um conjunto específico de covariáveis, cujas informações são apresentadas na Tabela 1. Essas covariáveis representam os diferentes atributos e características

das transações de cartão de crédito, que são consideradas como preditores potenciais para identificar a presença de fraude.

Ao aplicar o modelo Random Forest e o modelo de Regressão Logística em cada cenário, foram avaliadas diversas métricas de desempenho, como acurácia, sensibilidade, especificidade e valor preditivo positivo. Essas métricas permitiram comparar a capacidade de detecção de fraudes dos dois modelos em cada cenário específico.

É importante ressaltar que a escolha desses quatro cenários visa abordar limitações que uma empresa possa ter em relação a captura de covariáveis para detecção de fraudes. Dessa forma, foi possível avaliar a robustez e a generalização dos modelos em diferentes contextos.

Ao final da análise comparativa, espera-se obter percepções sobre a efetividade e as limitações de cada modelo em relação à detecção de fraudes em transações de cartão de crédito.

No cenário 1, foram utilizadas 11 covariáveis, todas relacionadas a informações básicas coletadas no momento da transação. Esse cenário foi denominado “cenário básico”, pois são informações básicas e coletadas no momento da transação, sem a necessidade de percorrer o banco de dados ou desenvolver cálculos.

No cenário 2, foram utilizadas 16 covariáveis que capturam principalmente informações relacionadas ao tempo e à frequência de comportamentos do usuário. Esse cenário foi denominado "cenário comportamental", uma vez que busca mapear o comportamento do usuário por meio de características como a velocidade de compra e a alteração de cartões durante as compras em diferentes intervalos de tempo. Neste cenário é preciso percorrer o banco de dados para coletar

informações históricas do usuário. Portanto, necessita de um desenvolvimento mais elaborado para sua criação.

No cenário 3, foram utilizadas uma série de variáveis criadas a partir da parte local do endereço de e-mail do usuário, ou seja, a parte que vem antes do @. Estas funções visão identificar padrões na estrutura do email. Esse cenário foi denominado "cenário da característica do email". As variáveis consistirão em contadores de caracteres do email. Essas variáveis são especialmente úteis para detectar fraudes em que os fraudadores precisam gerar um grande número de e-mails de forma sistemática. Ao analisar as características do email, é possível extrair informações relevantes que podem indicar atividades fraudulentas. Os fraudadores muitas vezes recorrem à criação em massa de contas de email ou enviam e-mails em grandes quantidades como parte de suas estratégias para realizar fraudes em transações de cartão de crédito.

No cenário 4, foi realizada uma análise de componentes principais (PCA, do inglês Principal Component Analysis), que levará em conta a correlação entre todas as covariáveis e a variável resposta. O objetivo dessa análise é reduzir a dimensionalidade do conjunto de dados, resumindo as informações contidas em várias covariáveis em um número menor de variáveis, conhecidas como componentes principais.

A análise de componentes principais é uma técnica estatística amplamente utilizada para explorar a estrutura subjacente de um conjunto de dados multivariado. Ela busca identificar combinações lineares das covariáveis originais que capturem a maior parte da variabilidade dos dados. Essas combinações lineares, chamadas de componentes principais, são ordenadas de forma decrescente de acordo com sua importância em explicar a variância total dos dados.

Ao aplicar o PCA no cenário 4, todas as covariáveis foram consideradas em conjunto, permitindo que a análise identifique os padrões gerais e a estrutura subjacente das variáveis em relação à variável resposta, independentemente de sua quantidade original. O resultado foi a criação de novas variáveis (componentes principais) que representam o comportamento geral ou capturam a maior parte da variabilidade presente nas covariáveis originais.

3.2.1 Balanceamento de sub-amostragem

No contexto do cenário descrito, em que o conjunto de dados apresenta um desbalanceamento significativo entre as classes de fraude e não fraude, medidas devem ser tomadas para lidar com essa disparidade a fim de garantir um treinamento adequado dos modelos de detecção de fraudes. Uma abordagem comum nesses casos é a criação de sub-amostras balanceadas.

No estudo, foram criadas sub-amostras balanceadas a partir do conjunto de dados original, de modo que a quantidade de exemplos de fraude seja equiparada à quantidade de exemplos de não fraude. Estes sub-amostras foram divididas em treino e teste, em que apenas as amostras de treino foram balanceadas. Assim, 70% das fraudes foram incluídas no conjunto de treinamento, enquanto os 30% restantes foram alocados no conjunto de teste. Essa divisão estratégica permite avaliar o desempenho dos modelos em dados de teste independentes.

Ao adotar essas práticas de sub-amostragem balanceada e divisão proporcional dos dados entre treino, busca-se minimizar o viés causado pelo desbalanceamento inicial do conjunto de dados. Essas medidas contribuem para que os modelos de detecção de fraudes sejam treinados e avaliados de forma mais adequada, considerando tanto as fraudes

quanto as não fraudes de maneira equilibrada, o que pode resultar em um desempenho mais preciso e confiável na detecção de atividades fraudulentas.

Após o balanceamento do conjunto de dados, um processo de repetição aleatória foi realizado na criação das sub-amostras para treinamento e teste. Foram criadas 1000 diferentes sub-amostras para treinamento, cada uma delas acompanhada por 1000 subconjuntos complementares para teste.

Essa abordagem, conhecida como reamostragem bootstrap, envolve a geração de múltiplas amostras a partir do conjunto de dados original. Através da seleção aleatória com reposição, essa técnica permite criar variações aleatórias dos dados disponíveis, capturando a incerteza associada à amostragem.

Ao criar 1000 diferentes sub-amostras para treinamento, explorou-se a variabilidade inerente aos dados e permitiu-se que os modelos de detecção de fraudes fossem treinados em diferentes conjuntos de exemplos balanceados. Essa abordagem é útil para fornecer uma estimativa mais robusta do desempenho dos modelos.

Adicionalmente, para cada sub-amostra de treinamento, foram criados 1000 subconjuntos complementares para teste. Isso possibilitou a avaliação do desempenho dos modelos em diferentes conjuntos de dados de teste, fornecendo uma avaliação mais abrangente e estável.

Resumidamente, o conjunto de amostras de treinamento é composto por 349 casos de fraudes e 349 casos de não fraudes. Já o conjunto de amostras de teste contém todas as observações restantes, com 150 casos de fraudes e 26.110 casos de não fraudes.

Essa estratégia de repetição aleatória na criação das sub-amostras de treinamento e teste é útil para avaliar a variabilidade e a incerteza associadas ao treinamento e à avaliação dos modelos de detecção de fraudes. Ela permite obter uma visão mais robusta do desempenho dos modelos, considerando diferentes combinações de exemplos de treinamento e teste, e fornecendo uma avaliação mais confiável da capacidade dos modelos em detectar fraudes em diferentes cenários.

3.2.2 Acurácia, sensibilidade e especificidade

A acurácia é uma medida que representa a proporção de previsões corretas em relação ao total de observações. Ela fornece uma visão geral do desempenho geral do modelo, considerando tanto as classificações corretas de fraude quanto de não fraude.

A sensibilidade, também conhecida como taxa de verdadeiros positivos, mede a proporção de fraudes que foram corretamente identificadas em relação ao total de casos de fraude. Essa métrica é importante para avaliar a capacidade do modelo em detectar efetivamente as transações fraudulentas.

A especificidade, por sua vez, é a taxa de verdadeiros negativos. Ela representa a proporção de casos de não fraude que foram corretamente identificados em relação ao total de casos de não fraude. Essa métrica é útil para avaliar a capacidade do modelo em distinguir corretamente as transações legítimas das fraudulentas, entretanto, devido ao desbalanceamento dos dados, a medida de especificidade fica muita próxima da acurácia, uma vez que os dados são majoritariamente não fraude. Por este fato, os gráficos apresentarão apenas a acurácia e a sensibilidade.

4 RESULTADOS E DISCUSSÃO

4.1 Modelo de Regressão Logística

4.1.1 Ajuste dos Modelos

Nesta etapa inicial, foi ajustado um modelo de Regressão Logística para cada uma das 1000 sub-amostras em cada um dos quatro diferentes cenários mencionados, resultando em 4000 modelos ajustados. O objetivo era avaliar a velocidade de convergência dos modelos e identificar as variáveis que se mostraram estatisticamente significativas em cada cenário. O custo computacional foi de 0.674 segundos por modelo ajustado, totalizando 11 minutos e 14 segundos para ajustar os 1000 modelos.

Ao realizar o ajuste repetido dos modelos de Regressão Logística, considerou-se a reamostragem bootstrap para criar variações aleatórias dos conjuntos de dados de treinamento. Esse procedimento permitiu explorar a incerteza associada aos dados e obter uma visão mais robusta dos resultados.

Durante cada ajuste dos modelos, avaliou-se a velocidade de convergência, ou seja, a quantidade de interações necessárias para que os modelos atingissem a convergência durante o processo de estimação dos parâmetros. Essa informação é relevante para compreender a estabilidade e a eficiência dos modelos de Regressão Logística em cada cenário. A Figura 5 a seguir mostra a quantidade de interações necessárias em cada cenário, limitadas em 1000 interações.

Figura 5 – Interações necessárias para convergência dos parâmetros.

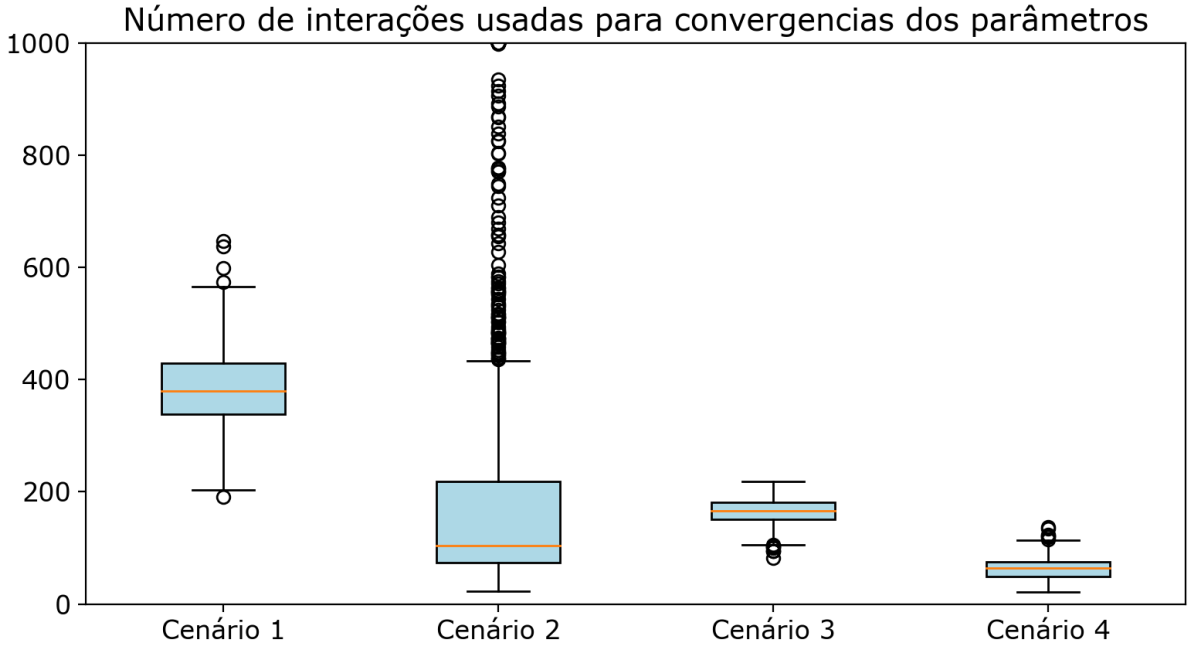

Figura 5.

Podemos Observar que o cenário 1 foi o que em média necessitou mais interações para convergir, enquanto o cenário 2, em alguns casos chegou a 1000 interações sem convergir.

Além disso, foi realizada uma análise das variáveis que se mostraram estatisticamente significativas em cada cenário. Esse processo envolveu a identificação das covariáveis que apresentaram um impacto estatisticamente significativo na variável resposta, com base nos resultados dos testes de hipóteses associados aos coeficientes estimados nos modelos de Regressão Logística.

Ao repetir esse processo 1000 vezes em cada cenário, foi possível observar a consistência dos resultados e identificar as variáveis mais relevantes em cada contexto específico. Essas informações são essenciais para compreender quais covariáveis têm um papel significativo na detecção de fraudes e podem fornecer insights valiosos para o desenvolvimento de modelos mais precisos e eficientes.

Na Figura 6, são apresentados os desempenhos de cada cenário em termos de métricas de avaliação de modelos. Essas métricas incluem a acurácia, a sensibilidade e a especificidade.

Figura 6 – Distribuição das métricas relacionadas ao desempenho dos modelos.

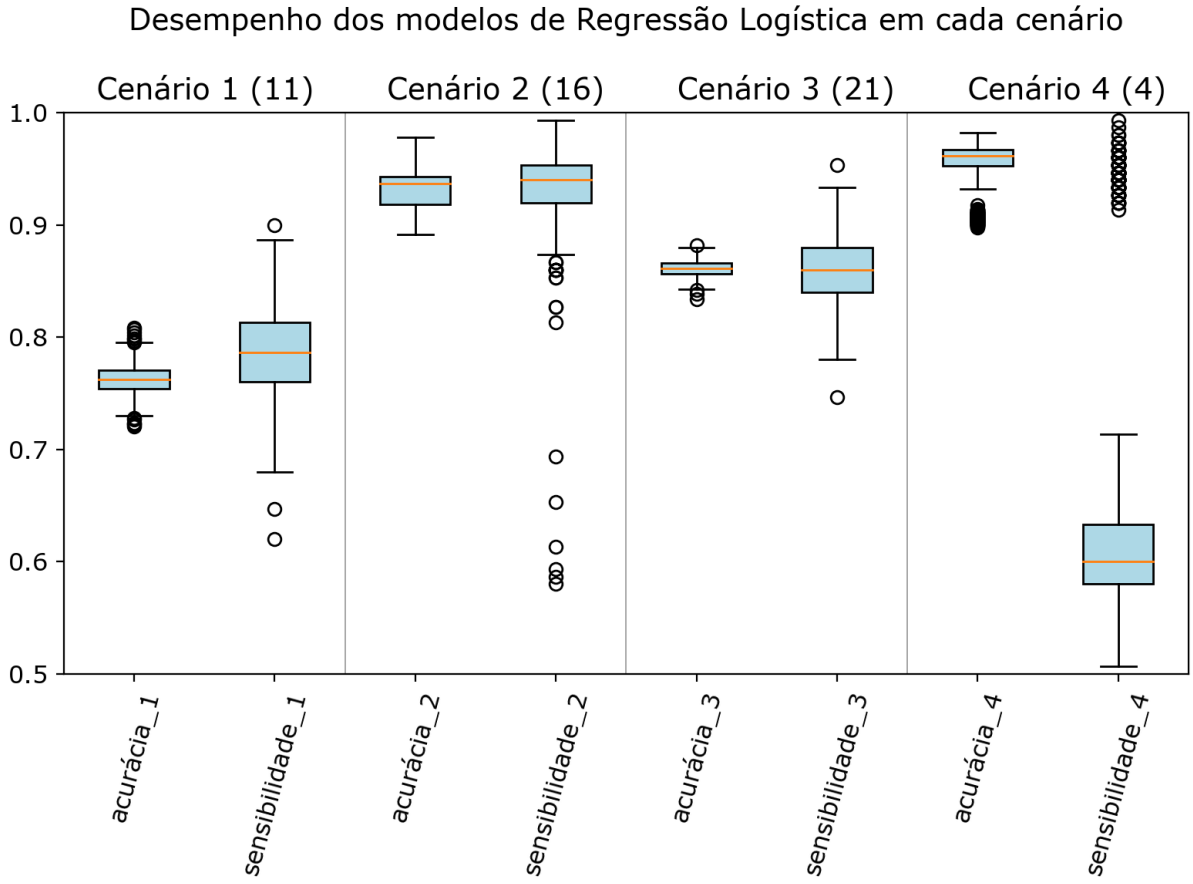

Figura 6. Entre parênteses está o número de covariáveis utilizadas no ajuste dos modelos.

4.1.2 Seleção das variáveis

Após a análise dos modelos ajustados 1000 vezes em cada cenário, o próximo passo consistiu na seleção das variáveis mais significativas. Para realizar essa seleção, somou-se o valor-p de cada covariável em cada ajuste realizado, resultando em uma métrica agregada que reflete a importância geral das variáveis.

Na Figura 7, é apresentada a representação desses valores agregados, indicando o somatório dos valores-p para cada covariável. Com auxílio do AIC (Critério de Informação de Akaike), foi estabelecido um ponto de corte de 325, de modo que as variáveis cujo somatório dos valores-p foi inferior a esse número foram consideradas significativas para os próximos ajustes dos modelos.

Essa abordagem permite identificar as variáveis que, de forma geral, apresentaram associação estatisticamente significativa com a variável resposta em cada cenário. Ao somar os valores-p, é possível avaliar a consistência da importância das variáveis ao longo das repetições dos modelos. A definição do ponto de corte de 325 permite estabelecer um critério objetivo para selecionar as variáveis mais relevantes.

Essa etapa de seleção de variáveis é fundamental para reduzir a dimensionalidade do problema e focar nas covariáveis mais relevantes na detecção de fraudes. Ao considerar apenas as variáveis significativas, é possível melhorar a eficiência computacional e interpretabilidade dos modelos, além de evitar a inclusão de variáveis pouco informativas ou redundantes.

As variáveis selecionadas foram utilizadas nas etapas seguintes do ajuste dos modelos, visando otimizar o desempenho na detecção de fraudes. No cenário 1, o número de variáveis foi reduzido de 11 para 7. No cenário 2, a redução foi de 16 para 8 variáveis. Já no cenário 3, o número de covariáveis foi reduzido de 21 para 4. Essa seleção das variáveis permitiu simplificar os modelos, tornando-os mais eficientes e rápidos e mantendo a eficiência.

Figura 7 – Soma de p-valores.

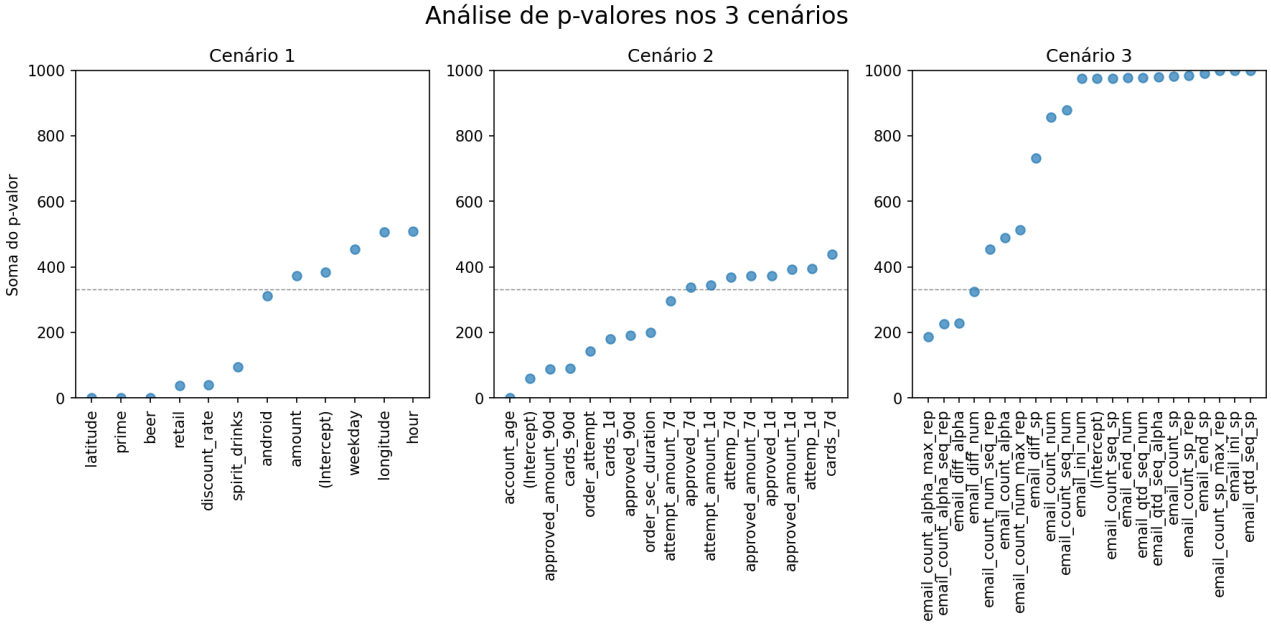

Figura 7.

4.1.3 Ajuste dos Modelos com variáveis significativas

Após identificadas as variáveis significativas, os mesmos 1000 modelos foram reajustados apenas com essas variáveis, com uma redução do tempo computacional de 81%, indo para 0.127 segundos por modelo. A quantidade de interações necessárias para ajustar os modelos também foi reduzida drasticamente como podemos verificar na Figura 8.

Figura 8 – Interações necessárias para convergência dos parâmetros.

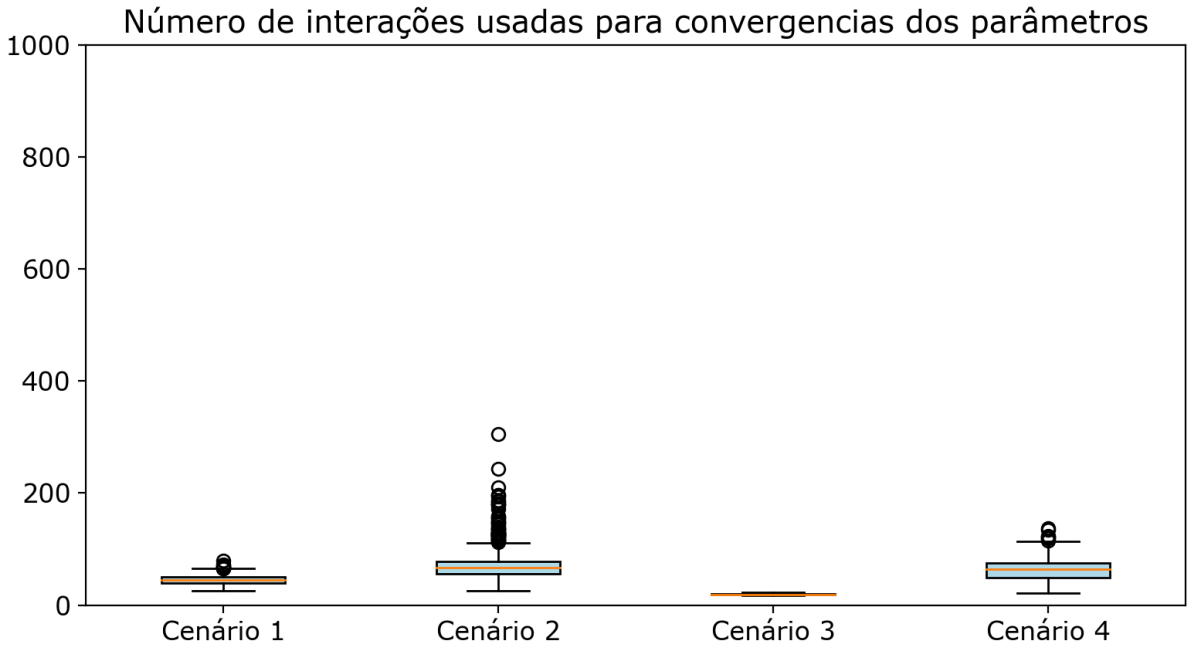

Figura 8.

Na Figura 9, é possível observar que a qualidade dos ajustes dos modelos, de forma geral, se manteve similares aos ajustes iniciais com todas as covariáveis, mesmo após a seleção das variáveis mais significativas. No entanto, algumas diferenças foram observadas entre os diferentes cenários.

Especificamente, o cenário 2, que considerou as variáveis relacionadas ao comportamento do usuário, apresentou resultados promissores em termos de acurácia, sensibilidade e especificidade. Isso indica que o comportamento do usuário desempenhou um papel relevante na detecção de fraudes neste conjunto de dados. A alta acurácia e a capacidade de identificar corretamente tanto as fraudes quanto as não fraudes destacam a eficácia desse cenário específico.

Por outro lado, o cenário das componentes principais mostrou um ajuste geralmente bom, mas com baixa sensibilidade. Isso significa que o

modelo teve dificuldade em identificar corretamente os casos de fraude, resultando em uma taxa maior de falsos negativos. Embora a acurácia possa ser alta, é importante considerar a baixa sensibilidade, pois a detecção adequada das fraudes é crucial em um contexto de detecção de fraudes em transações de cartão de crédito.

Essas observações ressaltam a importância de avaliar diferentes cenários e abordagens na detecção de fraudes. Embora certos cenários possam apresentar um bom ajuste global, é necessário considerar métricas específicas, como a sensibilidade, para garantir que o modelo esteja efetivamente identificando os casos de fraude.

Dessa forma, os resultados obtidos até o momento sugerem que o cenário 2, relacionado ao comportamento do usuário, pode ser mais relevante na detecção de fraudes nesse conjunto de dados específico. Por outro lado, o cenário das componentes principais, embora tenha tido um bom ajuste global, precisa ser aprimorado em termos de sensibilidade para melhorar a detecção de fraudes.

Essas observações fornecem informações valiosas para aprimorar os modelos de detecção de fraudes e direcionar esforços futuros na seleção de variáveis e no desenvolvimento de abordagens mais eficazes para a detecção de fraudes em transações de cartão de crédito.

Figura 9 – Distribuição das métricas relacionadas ao desempenho dos modelos.

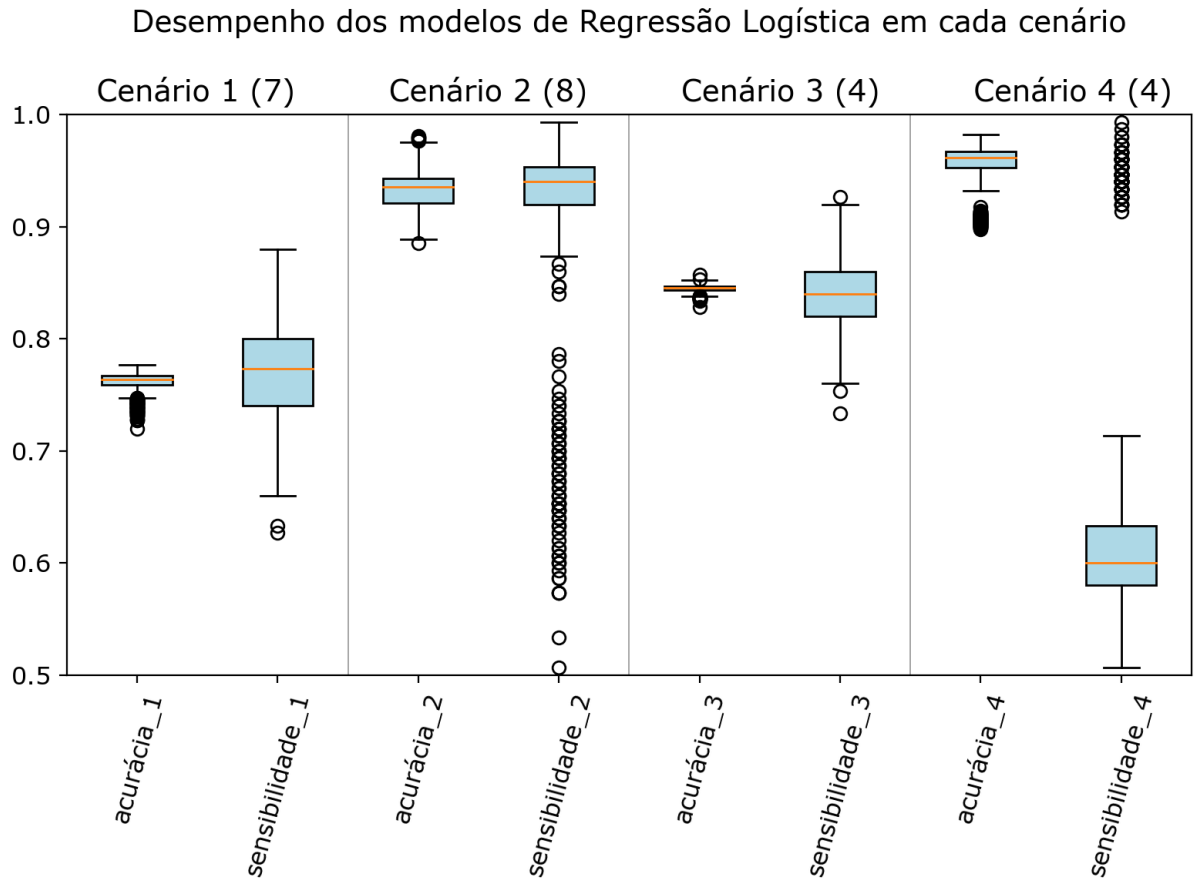

Figura 9. Entre parênteses está o número de covariáveis utilizadas no ajuste dos modelos.

4.2 Modelo de Random Forest

4.2.1 Ajuste dos Modelos com variáveis significativas

O capítulo atual apresenta o Modelo de Random Forest, uma técnica de aprendizado de máquina amplamente utilizada para análise e previsão de dados. Neste contexto, buscamos explorar a relevância das variáveis significativas identificadas no Modelo de Regressão Logística e o impacto computacional associado ao ajuste do Modelo de Random Forest utilizando exclusivamente essas variáveis selecionadas. Ao longo deste capítulo,

abordaremos aspectos relacionados à eficiência de ajuste e computacional. No presente trabalho, foi empregado o modelo de Random Forest para análise e previsão dos dados, visando à detecção de fraudes em transações online de cartão de crédito. Durante o ajuste do modelo, diversos parâmetros foram considerados.

O parâmetro 'n_estimators' determina o número de árvores no conjunto de Random Forest. Para este estudo, foi utilizado o valor de 100 árvores, buscando um equilíbrio entre desempenho e eficiência computacional, levando em consideração a complexidade do problema de detecção de fraudes.

O parâmetro ‘max_features’ que define o número de covariáveis utilizado em cada árvore. Nos modelos ajustados foi utilizado o parâmetro com o valor ‘sqrt’, que calcula a raiz quadrada do número total de covariáveis.

No que se refere ao parâmetro 'criterion', que define a medida utilizada para avaliar a qualidade das divisões nos nós das árvores, o valor padrão foi adotado. No contexto de detecção de fraudes em transações online de cartão de crédito, o critério padrão, Gini, é considerado apropriado para capturar padrões complexos e fornecer resultados confiáveis.

Em suma, o modelo de Random Forest utilizado neste estudo foi ajustado considerando os diferentes parâmetros, levando em consideração a relevância do contexto de detecção de fraudes em transações online de cartão de crédito. Essa abordagem permitiu obter resultados confiáveis e relevantes para a análise e previsão dos dados em questão.

Neste primeiro ajuste de Random Forest, consideramos apenas as variáveis significativas identificadas previamente no Modelo de Regressão Logística. Embora o custo computacional tenha sido de pouco mais de 19 vezes o tempo do modelo de Regressão Logística, com uma média de 2.482 segundos no Random Forest em comparação com 0.127 segundos na Regressão Logística, os modelos de Random Forest apresentaram um ajuste significativamente melhor aos dados.

Figura 10 – Distribuição das métricas relacionadas ao desempenho dos modelos.

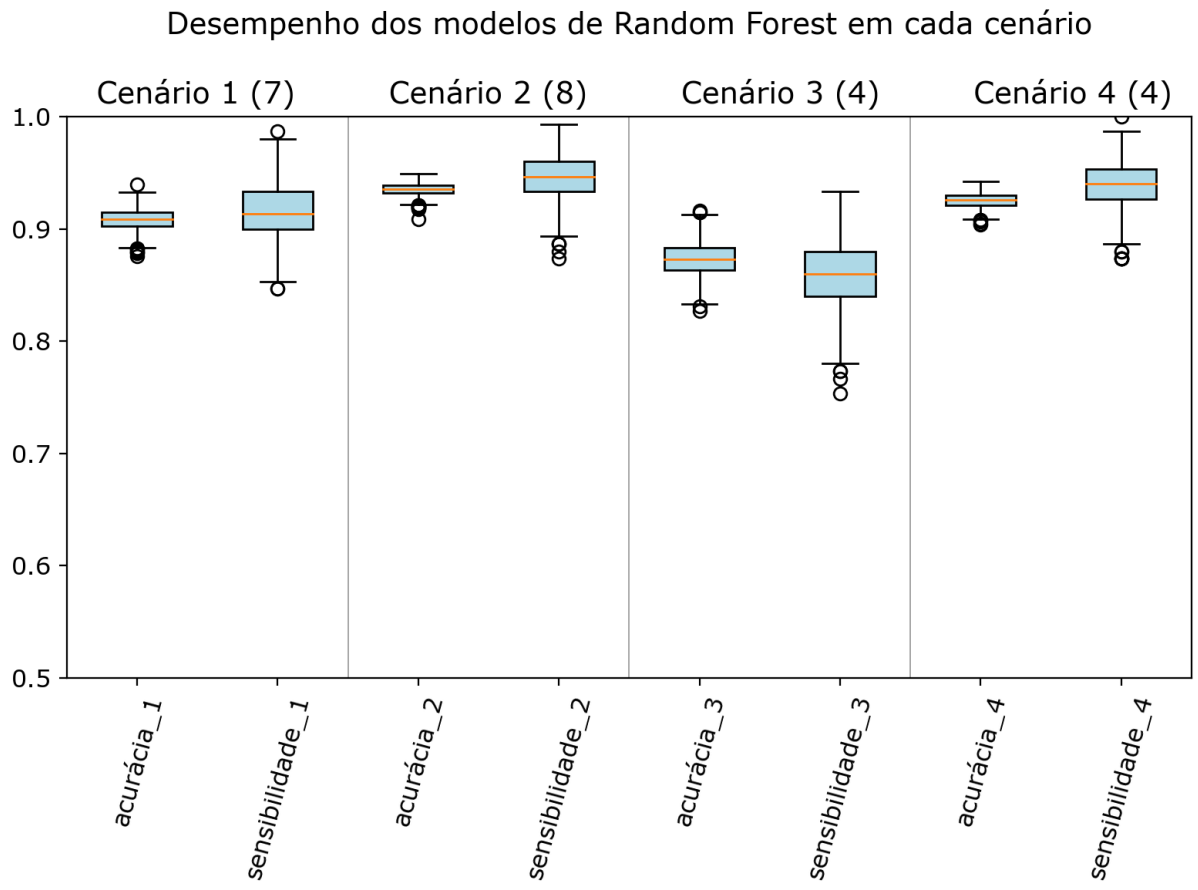

Figura 10. Entre parênteses está o número de covariáveis utilizadas no ajuste dos modelos.

4.2.2 Ajuste dos Modelos com todas as variáveis

Após ajustar os modelos de Random Forest considerando apenas as variáveis significativas, foi realizado um segundo ajuste incluindo todas as variáveis em cada cenário. Observou-se que o aumento no número de covariáveis resultou em um aumento no custo computacional, passando de 2.482 segundos para 2.705 segundos por modelo ajustado.

No que diz respeito ao ganho de informação, os cenários 1 e 3 apresentaram alguma melhora em comparação aos modelos de Random Forest com apenas as variáveis significativas, destaque para acurácia do cenário 1. Isso indica que a inclusão de todas as variáveis relevantes nesses cenários contribuiu para uma melhor adaptação do modelo aos dados. No entanto, no cenário 2, não foi observada uma alteração relevante no desempenho do modelo com o aumento das covariáveis.

Figura 11 – Distribuição das métricas relacionadas ao desempenho dos modelos.

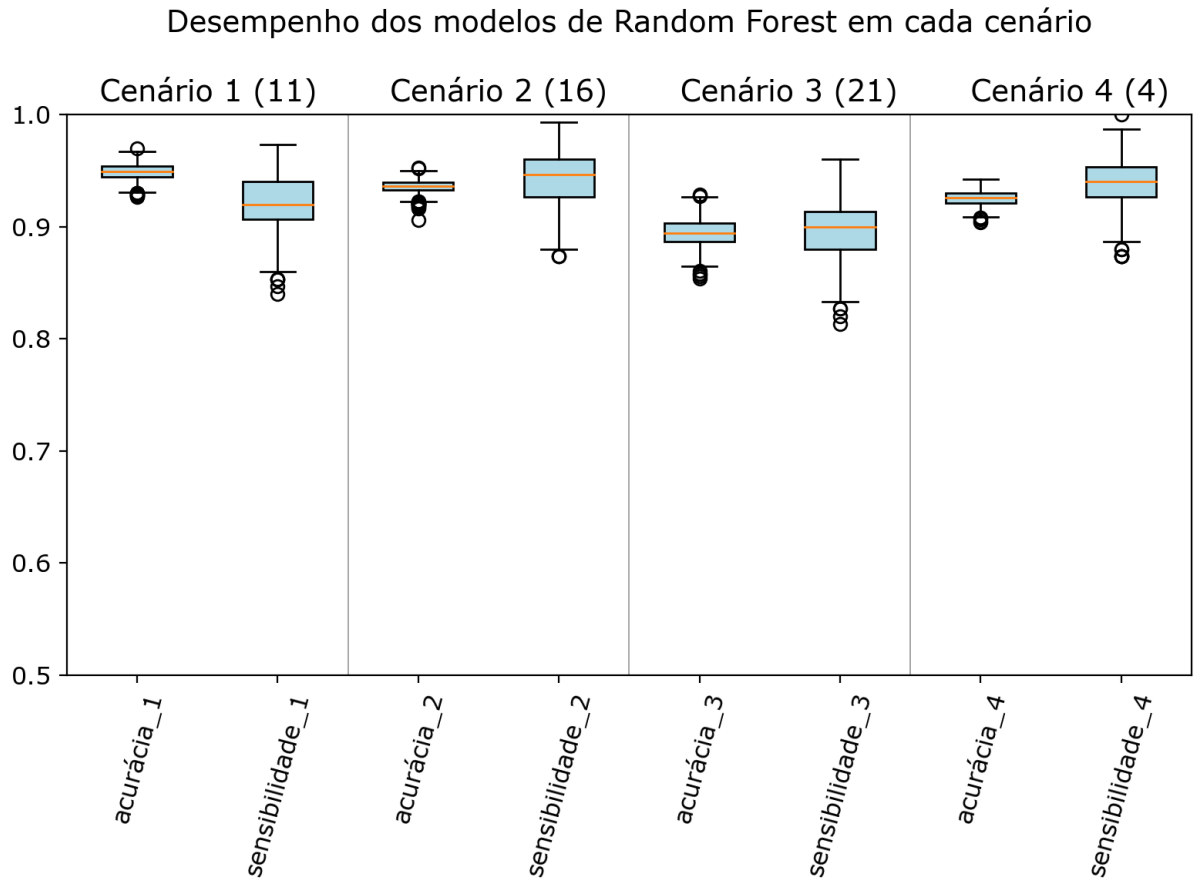

Figura 11. Entre parênteses está o número de covariáveis utilizadas no ajuste dos modelos.

4.3 Comparação dos resultados, Regressão Logística e Random Forest

No intuito de comparar o desempenho dos modelos, foram analisados os resultados em cada um dos 4 cenários.

4.3.1 Cenário 1

No cenário 1, foi possível observar uma melhora significativa no desempenho dos modelos de Random Forest em relação aos modelos de Regressão Logística, tanto em termos de acurácia quanto de sensibilidade. Essas melhorias podem ser visualizadas na Figura 11, que demonstra a comparação dos resultados obtidos pelos diferentes modelos.

Ao utilizar o modelo de Random Forest no cenário 1, observou-se um aumento substancial na acurácia em comparação com a Regressão Logística. Além disso, a sensibilidade também apresentou uma melhora considerável, indicando uma capacidade maior de detectar corretamente os casos positivos. Esses resultados destacam a eficácia do modelo de Random Forest na análise dos dados no cenário 1, fornecendo insights valiosos para a pesquisa em questão.

É importante ressaltar que a comparação do desempenho entre os modelos de Random Forest e Regressão Logística foi realizada considerando os mesmos conjuntos de variáveis e demais conFigurações relevantes. Essa abordagem permitiu uma avaliação justa e precisa do impacto do modelo de Random Forest no cenário específico.

Figura 12 – Distribuição das métricas relacionadas ao desempenho dos modelos.

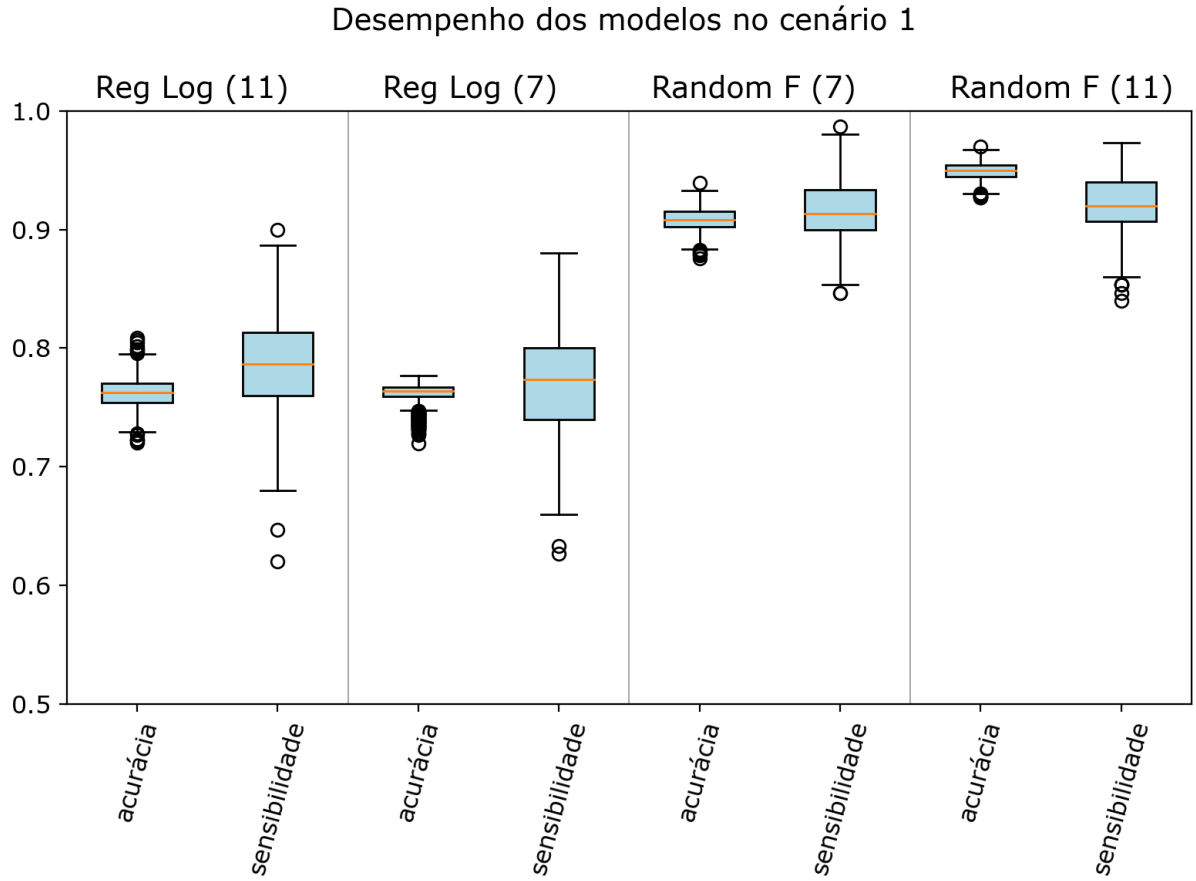

Figura 12. Entre parênteses está o número de covariáveis utilizadas no ajuste dos modelos.

4.3.2 Cenário 2

No cenário 2, ao analisar a Figura 12, observa-se que em média, os resultados dos modelos de Random Forest e Regressão Logística foram similares. No entanto, a diferença entre os modelos foi percebida em relação à consistência dos ajustes e à presença de pontos fora do limite inferior.

No que diz respeito à acurácia, os ajustes do Random Forest apresentaram uma consistência maior em relação à Regressão Logística. Isso significa que, em geral, o modelo de Random Forest manteve um

desempenho mais estável e confiável ao longo dos ajustes realizados no cenário 2.

Em relação à sensibilidade, também foi observado que o Random Forest teve um desempenho mais consistente, com menos pontos fora do limite inferior em comparação com a Regressão Logística. Isso indica que o modelo de Random Forest teve uma capacidade superior de identificar corretamente os casos de fraude, evitando ajustes com baixa capacidade de predição do evento de interesse.

Figura 13 – Distribuição das métricas relacionadas ao desempenho dos modelos.

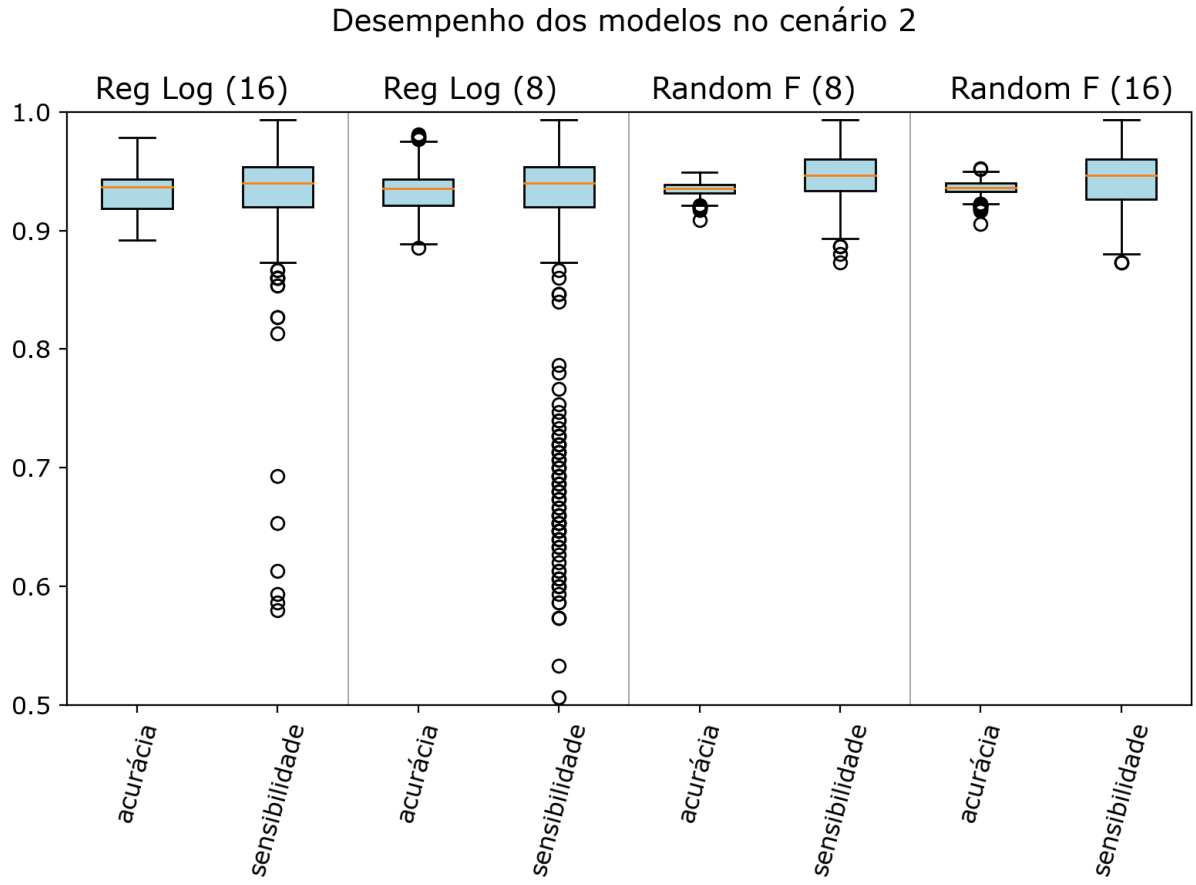

Figura 13. Entre parênteses está o número de covariáveis utilizadas no ajuste dos modelos.

4.3.3 Cenário 3

No cenário 3, ao analisar os resultados, foi observado que, em média, os modelos de Random Forest apresentaram resultados ligeiramente melhores em comparação aos modelos de Regressão Logística. No entanto, os modelos de Regressão Logística por sua vez, apresentaram uma variabilidade na acurácia muito menor que o Random Forest.

Figura 14 – Distribuição das métricas relacionadas ao desempenho dos modelos.

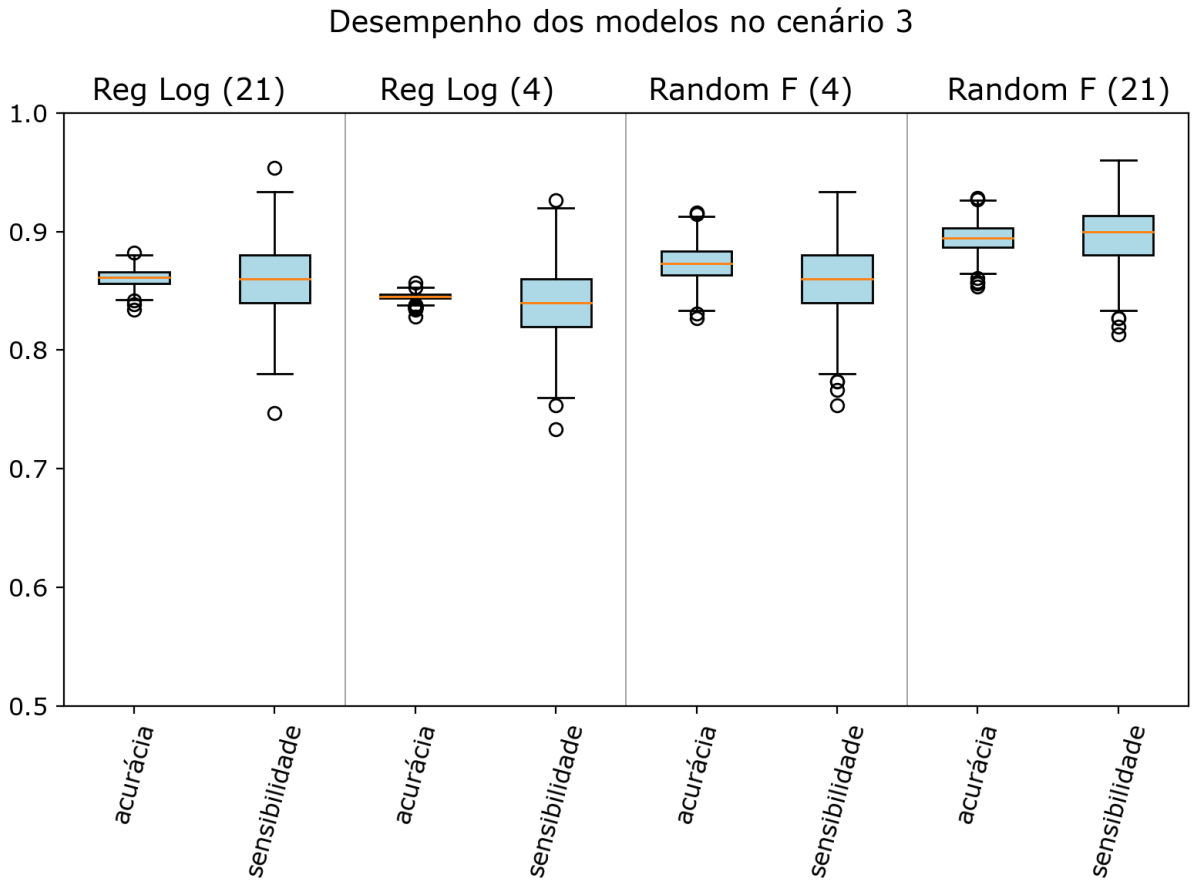

Figura 14. Entre parênteses está o número de covariáveis utilizadas no ajuste dos modelos.

4.3.4 Cenário 4

No cenário 4, os modelos de Regressão Logística mostraram uma ligeira melhora na acurácia em comparação com os modelos de Random Forest. No entanto, a sensibilidade dos modelos de Regressão Logística foi consideravelmente pior. Isso significa que os modelos de Regressão Logística tiveram dificuldade em identificar corretamente os casos positivos, o que é crucial para a detecção de fraudes em transações online de cartão de crédito. Portanto, no cenário 4, os modelos de Random Forest podem ser mais adequados, pois demonstraram uma capacidade superior de detecção dos casos positivos.

Figura 15 – Distribuição das métricas relacionadas ao desempenho dos modelos.

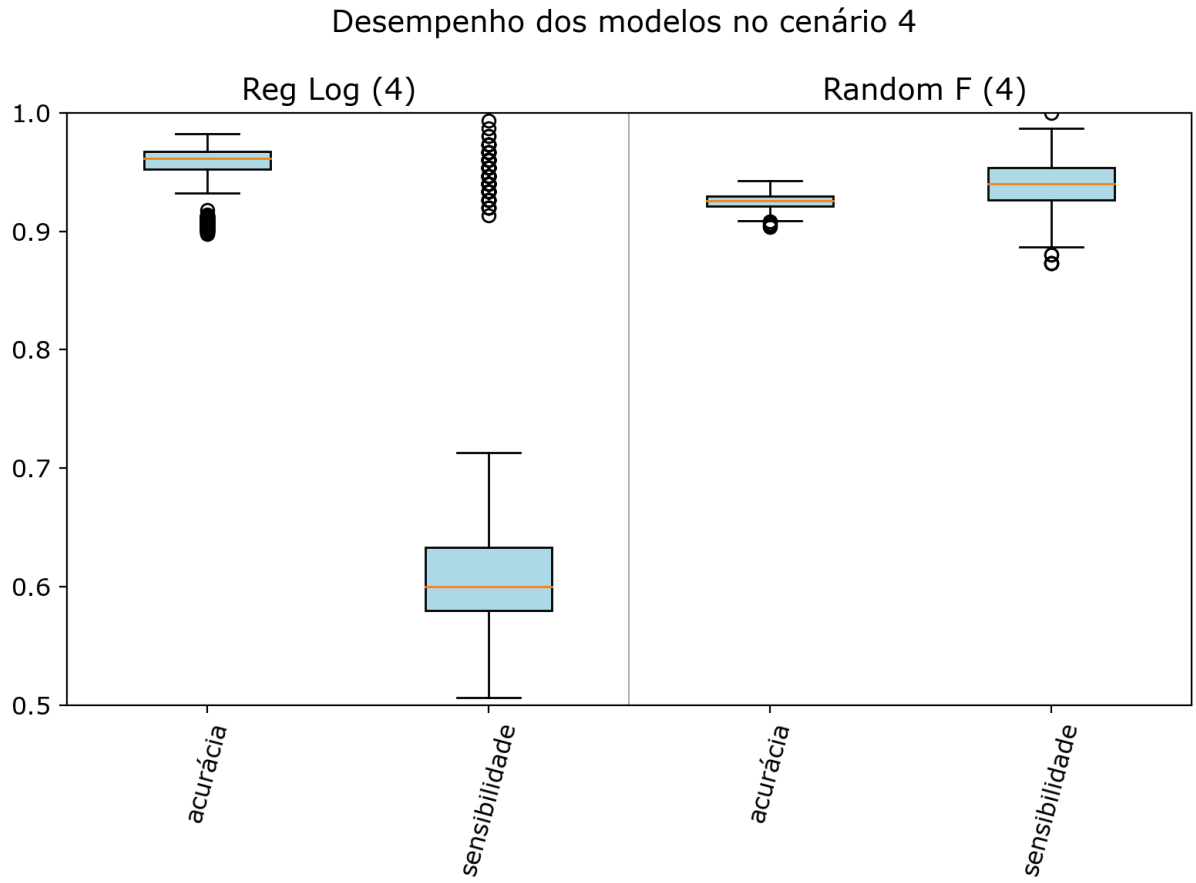

Figura 15. Entre parênteses está o número de covariáveis utilizadas no ajuste dos modelos.

4.4 Comparação dos cenários no Modelo Random Forest

Após observar um ganho de precisão nos resultados dos modelos de Random Forest ao adicionar as variáveis não significativas no modelo de Regressão Logística para cada cenário, surgiu o interesse de avaliar se um modelo que incluísse todas as variáveis de cada cenários teria um desempenho superior. Dessa forma, um modelo foi ajustado considerando todas as variáveis e os resultados foram comparados com os demais ajustes do Random Forest em cada cenário. A Figura 15 apresenta uma

visualização desses resultados, permitindo uma análise comparativa entre os modelos de Random Forest para cada conjunto de covariáveis.

Figura 16 – Distribuição das métricas relacionadas ao desempenho dos modelos.

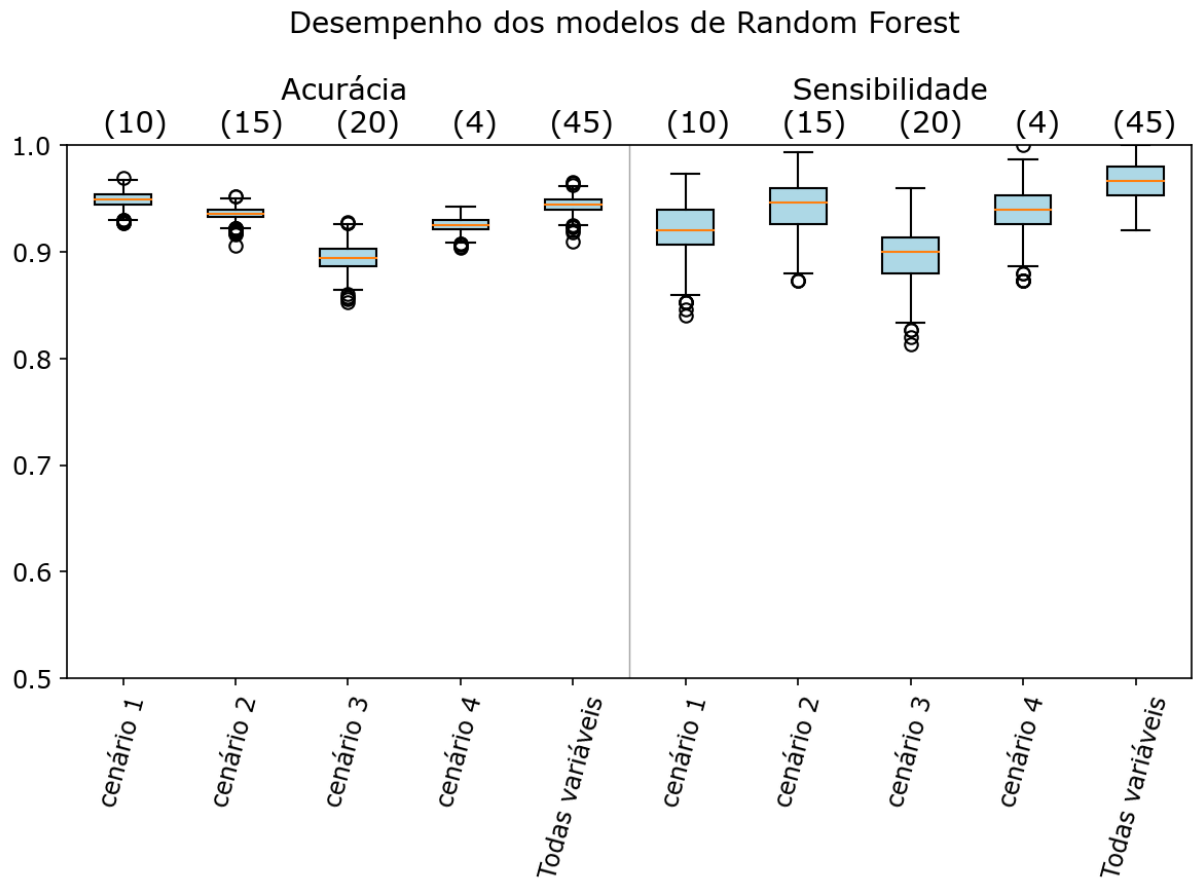

Figura 16. Entre parênteses está o número de covariáveis utilizadas no ajuste dos modelos.

5 CONSIDERAÇÕES FINAIS

Este trabalho teve como objetivo comparar os modelos de Regressão Logística e Random Forest em um cenário real de detecção de fraudes em transações online de cartão de crédito. Para tornar a análise ainda mais próxima do mundo real, o estudo foi dividido em cenários,

cada um representando grupos de covariáveis coletadas de acordo com o processo específico de obtenção das informações.

No cenário 1, foram consideradas as variáveis básicas, que consistem em informações diretas do cliente ou da compra. Essas variáveis comumente utilizadas para a análise de fraudes em transações online.

No cenário 2, incluíram-se métricas relacionadas ao comportamento do cliente, obtidas de forma interativa e com base em seu histórico. Essas métricas desempenham um papel fundamental na identificação de fraudes. No entanto, por não serem informações geralmente disponíveis nos bancos de dados das empresas, elas normalmente não são utilizadas nos ajustes iniciais, pois requerem mais tempo e recursos para serem incorporadas ao processo de prevenção.

No cenário 3, as variáveis foram selecionadas com o intuito de identificar padrões presentes nos e-mails dos clientes. Essas variáveis são fundamentais em casos de ataques de fraudes em que o fraudador utiliza algum algoritmo para criar email.

No cenário 4, foi adotada uma abordagem de criação de 4 variáveis de Componentes Principais. Essas variáveis foram geradas para sintetizar as informações contidas nas demais covariáveis, buscando capturar a essência das características dos dados.

A divisão em cenários permitiu a avaliação do desempenho dos modelos para diferentes grupos de covariáveis. De forma geral, observou-se que o modelo de Random Forest apresentou um melhor ajuste aos dados. No cenário 1, essa melhora foi especialmente significativa. Diferentemente do modelo de Regressão Logística, que manteve

resultados semelhantes ao utilizar todas as covariáveis dentro de cada cenário ou apenas as significativas, o modelo de Random Forest demonstrou um ajuste superior quando todas as variáveis foram consideradas. Isso indica que o Random Forest conseguiu capturar e aproveitar informações relevantes presentes nas covariáveis adicionais, resultando em um desempenho aprimorado na detecção de fraudes.

É importante ressaltar que o modelo de Random Forest apresenta um custo de processamento significativamente maior em comparação com a Regressão Logística. Nos cenários 2 e 3, embora tenha havido uma melhora leve nos resultados em relação à Regressão Logística, essa diferença não foi tão expressiva. Portanto, ao lidar com grandes volumes de dados e um grande número de covariáveis, além da necessidade de obter respostas rápidas, como em situações em que as decisões precisam ser tomadas em questão de frações de segundos, a Regressão Logística pode ser uma opção viável. A escolha do modelo adequado deve considerar fatores como o conjunto de variáveis disponíveis, a necessidade atual da empresa (seja reduzir fraudes ou aumentar aprovações), a eficiência computacional e a capacidade de resposta em tempo real, levando em conta as restrições e requisitos específicos do cenário em questão.

REFERÊNCIAS

1 **Global Payment Fraud and Security Survey Report**. ACI WORLDWIDE, 2019. Disponível em: https://www.aciworldwide.com/-/media/files/corporate/news/2019/global-payment-fraud-and-security-survey-report.pdf. Acesso em: 27 mar. 2023.

2 BERKSON, Joseph. **Application of the logistic function to bio-assay**. Journal of the American Statistical Association, v. 39, n. 227, p. 357-365, 1944.

3 HOSMER JR, D. W.; LEMESHOW, S.; STURDIVANT, R. X. **Applied logistic regression**. 3ª ed. Hoboken, NJ: John Wiley & Sons, 2013.

4 KLEINBAUM, D. G.; KLEIN, M. **Logistic regression: a self-learning text**. 2ª ed. New York: Springer Science & Business Media, 2010.

5 MENARD, S. **Applied logistic regression analysis**. 2. ed. Thousand Oaks, CA: Sage, 2002.

6 BREIMAN, Leo. **Bagging predictors**. v. 24, n. 2, p. 123-140, 1996.

7 BREIMAN, Leo. **Random forests**. v. 45, n. 1, p. 5-32, 2001.