REVISTA MIRANTE
Daqui se vê do alto!
Câmpus Anápolis de Ciências Exatas e Tecnológicas Henrique Santillo
Universidade Estadual de Goiás

Revista Mirante, Anápolis (GO), v. 18, n. 1, edição extra, p. 17-38, jun. 2025. ISSN 1981-4089

APLICANDO REGRESSÃO LOGÍSTICA NO COMBATE À FRAUDE EM EMPRESAS FINANCEIRAS

APPLYING LOGISTIC REGRESSION IN COMBATING FRAUD IN FINANCIAL COMPANIES

BRUNO REBELO PINTO
Escola Superior de Agricultura "Luiz de Queiroz" (ESALQ),
da Universidade de São Paulo (USP)
luizribeiro@unemat.br

LUIZ FERNANDO CALDEIRA RIBEIRO
Universidade do Estado de Mato Grosso (UNEMAT), Campus de Cáceres / MT
luizribeiro@unemat.br

Resumo: O custo médio de uma transação fraudulenta é 3,86 vezes o valor da operação no Brasil. Isso mostra que cada vez mais, as empresas, principalmente no ramo financeiro, precisam calibrar a segurança a fim de evitar prejuízos. Existem diversas ferramentas que mitigam a fraude, como por exemplo: fornecedores voltados em documentoscopia, biometria facial, biometria por voz etc. Nos últimos anos os dados têm ajudado a atuar fortemente nessa frente, eles ajudam a identificar padrões, entender comportamentos massivos e identificar anomalias. Os modelos matemáticos contribuem no volume expressivo de informações e favorece no processamento de grande volumetria das informações. Com a finalidade de combatermos as fraudes por meio de dados, foram propostos dois modelos logísticos: um modelo desbalanceado e outro balanceado a fim de justificar que, apenas olhando a acurácia dos modelos não justifica se de fato estão acertando na discriminação dos eventos fraudulentos. Portanto, foi importante avaliar a curva ROC para descriminar o quanto estava-se de fato classificando o evento como fraudulento, assim podendo tomar decisões com mais segurança e sugerindo a importância de usar modelos estatísticos para criar soluções inovadoras no combate à fraude financeira.
Palavras-chave: Dados; Modelos matemáticos; Soluções inovadoras; Curvas ROC

Abstract: The average cost of a fraudulent transaction in Brazil is 3.86 times the value of the original operation. This highlights the increasing need for companies—especially those in the financial sector—to calibrate their security measures in order to prevent losses. Various tools are available to mitigate fraud, such as document examination providers, facial biometrics, voice biometrics, among others. In recent years, data has played a pivotal role in this effort, helping to identify patterns, understand widespread behaviors, and detect anomalies. Mathematical models contribute to handling the vast amount of information and facilitate the processing of large data volumes. To combat fraud through data, two logistic models were proposed: one unbalanced and one balanced, in order to demonstrate that simply examining the accuracy of the models is insufficient to assess whether they are effectively discriminating fraudulent events. Therefore, evaluating the ROC curve was crucial to determine how accurately fraudulent events were being classified, enabling more informed decision-making and underscoring the importance of using statistical models to create innovative solutions for combating financial fraud.
Keywords: Data; Mathematical models; Innovative solutions; ROC curves

REVISTA MIRANTE
Daqui se vê do alto!
Câmpus Anápolis de Ciências Exatas e Tecnológicas Henrique Santillo
Universidade Estadual de Goiás
Revista Mirante, Anápolis (GO), v. 18, n. 1, edição extra, p. 17-38, jun. 2025. ISSN 1981-4089

Introdução

O papel da indústria bancária na sociedade esteve tradicionalmente vinculado a seu principal produto: O crédito. Por meio desta operação, os bancos têm como objetivo primordial a geração de lucros e retorno financeiro, além da contribuição ao desenvolvimento econômico do país. Esta é a visão tradicional vigente no Brasil e no mundo até pouco anos atrás (Silva, 2011).

A atual revolução tecnológica, impulsionada pelas tecnologias da informação e comunicação, tem provocado profundas transformações na sociedade. Antigas tradições estão sendo rompidas, enquanto surgem novas formas de produção, consumo, organização do trabalho e de relações interpessoais. Esse cenário de rápidas mudanças tem aumentado a incerteza em relação ao futuro próximo. Dentro desse contexto, o sistema financeiro ocupa uma posição de destaque, sendo o setor bancário reconhecido como um dos que mais investem em tecnologia digital em todo o mundo (Barroso, 2018).

Com uma sociedade cada vez mais conectada e tecnológica, até mesmo em tarefas simples, como realizar um pagamento na internet banking, podem gerar grande quantidade de dados. Nos processos de compras do governo essa também é uma realidade e, hoje, pode-se contar com bases de dados que não existiam em um passado. Esses dados podem ser analisados com a finalidade de se obter valor a partir dessas informações, como disponibilizar um serviço mais personalizado para o cliente, reduzindo custos operacionais ou evitando perdas com fraudes (Souza; Caitité, 2009).

A partir dos anos 80, houve um crescimento acelerado dos crimes virtuais e desde então a sociedade sofre com as fraudes ocorridas nesse meio. Visto que, acessar uma informação de qualquer pessoa é de fácil acesso. A internet cresce de maneira incontrolável e à medida que a tecnologia cria oportunidades de mercado e mais facilidades, as pessoas a usufruem para o bem ou mal (Aragão, 2015).

Os fraudadores invadem, roubam dados pessoais ou de empresas, clonam cartões e assim geram graves problemas às pessoas. Os fraudadores renascem na pele dos antigos piratas.

Revista Mirante, Anápolis (GO), v. 18, n. 1, edição extra, p. 17-38, jun. 2025. ISSN 1981-4089

O triângulo da fraude surge por meio da seguinte hipótese: Pessoas confiáveis podem se tornar fraudadores, quando acreditam que nunca serão descobertas por ter certa confiança já adquirida em público (Machado, 2015).

A transformação do mercado financeiro ampliou as oportunidades de crédito para as empresas em diversos ramos, mas também introduziu novos riscos, para os quais muitas organizações ainda não estavam preparadas. No contexto brasileiro, estima-se que essas empresas paguem, em média, cerca de quatro vezes o valor de cada ação fraudulenta sofrida (Carvalho et al., 2013).

O crescimento do e-commerce trouxe maior facilidade, agilidade e velocidade às transações comerciais e financeiras. No entanto, as empresas que atuam nesse segmento precisam adotar medidas urgentes para minimizar os impactos financeiros decorrentes de fraudes. Isso exige investimentos em tecnologias de detecção e na implementação de práticas eficazes de prevenção contra agentes mal-intencionados, muitas vezes denominados "piratas da era pós-industrial" (Delgado, 2022).

Krauter e Famá (2005), destacam a importância das melhores práticas de governança corporativa, a utilização de ferramentas e conhecimentos de outras áreas para auxiliar no processo de tomada de decisão e utilizar cada vez mais modelos matemáticos para explicar os fenômenos de estudo.

O reconhecimento de padrões visa a classificar dados baseados em conhecimento a priori (preliminar) ou informações que fazem a soma de um determinado padrão. Portanto, podem-se aplicar diversas técnicas estatísticas como importantes ferramentas no controle e, principalmente, na tomada de decisão. Neste Contexto, este trabalho tem como importância estudar técnicas estatísticas que se mostrem viáveis na detecção de transações fraudulentas ao analisar um grande volume de dados.

Metodologia

Banco de dados

Revista Mirante, Anápolis (GO), v. 18, n. 1, edição extra, p. 17-38, jun. 2025. ISSN 1981-4089

Os dados foram obtidos no site https://www.kaggle.com/ onde foi determinada uma amostra de 400.000 observações para realizar os estudos. A escolha dessa base de dados justifica-se pela sua relevância, pelo grande volume de informações disponíveis e pela qualidade dos registros, aspectos fundamentais para assegurar a robustez das análises estatísticas. A quantidade de observações trabalhada foi considerada suficiente para representar a população de interesse, além de garantir maior significância estatística e melhor desempenho preditivo nos modelos aplicados.

Para isso, foram utilizados softwares como RStudio (v. 4.1.3) e o SAS Guide (v. 7.1) para tratamento de uma base que contém seis milhões transações financeiras por meios de pagamentos. A escolha da ferramenta foi devida ao grande poder computacional desses softwares e por contar com uma gama de algoritmos já implementada pela comunidade. Assim, é possível testar técnicas de classificação, dado que ambas as linguagens de programação contam com um grande repositório de bibliotecas.

Regressão logística

Quando pretende modelar relações entre variáveis, o modelo de regressão logística se apresenta como uma das ferramentas estatísticas mais importantes nas análises de dados.
Quando a variável resposta é do tipo dicotômica e as variáveis explicativas podem ser do tipo numéricas (contínuas, discretas) e/ou categóricas (Figueira, 2006).

Os métodos estatísticos frequentemente utilizados para detecção de fraude podem ser classificados em não supervisionados ou supervisionados (Bolton; Hand, 2002).

Nos métodos supervisionados, os casos fraudulentos e legítimos, apontados em nossa base de dados, serão utilizadas para construir modelos que permitam atribuir as novas observações a ocorrência ou não do evento.

Métodos estatísticos tradicionais, como análise discriminante e árvore de decisão, têm se mostrado como ferramentas efetivas na detecção de fraude (Hand, 1981). No entanto, ferramentas mais poderosas, como redes neurais e regressão logística binaria, têm sido bastante utilizadas (Ripley, 1996).

REVISTA MIRANTE
Daqui se vê do alto!
Câmpus Anápolis de Ciências Exatas e Tecnológicas Henrique Santillo
Universidade Estadual de Goiás
Revista Mirante, Anápolis (GO), v. 18, n. 1, edição extra, p. 17-38, jun. 2025. ISSN 1981-4089

No conjunto dos métodos supervisionados, dois deles têm sido, tradicionalmente, adotados pela maioria das instituições financeiras para detectar fraudes: sistemas baseados em regras e modelos de pontuação.

Com base nessa motivação encontrada na literatura, serão aplicados os modelos propostos com o objetivo de realizar predições sobre as transações e atribuir uma pontuação que permita classificar o risco associado a cada ocorrência.

O modelo aplicado nos dados originais será confrontado com modelo que teve os dados balanceados, visto que, as transações fraudulentas representam apenas 0,13%. A ideia é verificar a acurácia e a precisão das ocorrências positivas e que são de fatos positivas, ou seja, fraude em ambas as modelagens.

É esperado que a classificação dos dados balanceados seja melhor que o saturado devido à melhora na classe minoritária. A precisão tem que ser levada em consideração a fim de poder tomar decisões com o menor impacto possível nos falsos positivos.

O modelo para identificação da fraude é dado pelo seguinte modelo saturado:

$$Fraude_i = \frac{1}{1 + e^{-\left(\beta_0 + tempo_i\beta_1 + amount_i\,\beta_2 + atg_{sld_{org_i}}\,\beta_3 + nv_{sld_{org_i}}\beta_4 + atg_{sld_{dest_i}}\,\beta_5 + nv_{sld_{dst_i}}\,\beta_6 + type_i\,\beta_7\right)}}$$

$i = 1,2, \ldots, 400.000.$

.Onde os $\beta_{1,\ldots 7}$ são parâmetros que serão estimados pelo modelo.

A variável "*type*" é do formato categórica e precisa ser feito o processo de $(n-1)$ *dummy*.

Ajustando o primeiro modelo:

$$\widehat{Fraude_i} = 1/1 + exp - \{-20.65840 + tempo_i * 0.00494 + amount_i * -0.00003 + atg\_sld\_org_i * 0.00004 + nv\_sld\_org_i * -0.000051 + atg\_sld\_dest_i * 0.000005 + nv\_sld\_dst_i * -0.000005\},$$

$$i = 1,2,3, \ldots, 400.000.$$

As variáveis *tempo, amount, atg_sld_org, nv_sld_org, atg_sld_org e nv_sld_dest* apresentaram estatisticamente significamente a um α de 5%.

Revista Mirante, Anápolis (GO), v. 18, n. 1, edição extra, p. 17-38, jun. 2025. ISSN 1981-4089

A formulação deste modelo segue os princípios da regressão logística binária, amplamente utilizada para problemas de classificação com resposta binária, conforme descrito por Hosmer, Lemeshow e Sturdivant (2013).

Função resposta

O modelo de regressão logística é um tipo de modelo linear generalizado onde as variáveis respostas $Y_i, \dots, Y_n$ são independentes e binárias com:

$$Y_i = Bernoulli(p_i),$$

sendo a família *Bernoulli* exponencial. Sabendo que a esperança de Y_i que é $E[Y_i] = p_i = p(Y_i = 1)$, a regressão logística é descrita como uma relação de p_i com x_i conforme a seguir:

$$\log(\frac{p_i}{1-p_i})_i = \alpha + \beta x_i,$$

onde α é uma constante, que representa a interceptação da reta com o eixo vertical e o β representa a inclinação em relação à variável x_i.

Interpretando a equação anterior tem-se que o lado esquerdo representa o *log* das chances de sucesso para Y_i.

O modelo assume que este logito é uma função linear da preditora x_i. A função de probabilidade Bernoulli pode ser escrita na forma de família exponencial:

$$p_i^{y_i^*}(1-p_i)^{1-y_i^*} = (1-p)e^{y_i \log(\frac{p_i}{1-p_i})}.$$

Realizando uma ligação canônica, ao utilizar o parâmetro natural $\log(\frac{p_i}{1-p_i})$ pode-se reescrever a equação da seguinte forma:

$$p_i = \frac{e^{\alpha+\beta}}{1+e^{(\alpha+\beta x_i)}}$$

Fazendo p uma função de x_i:

$$p(x) = \frac{1}{1+e^{-(\alpha+\beta x_i)}}$$

REVISTA MIRANTE
Daqui se vê do alto!
Câmpus Anápolis de Ciências Exatas e Tecnológicas Henrique Santillo
Universidade Estadual de Goiás
Revista Mirante, Anápolis (GO), v. 18, n. 1, edição extra, p. 17-38, jun. 2025. ISSN 1981-4089

Estimação dos parâmetros

No modelo em que $Y_i \sim Bernoulli(p_i)$, não existe uma conexão direta entre Y_i e o preditor linear $\alpha + \beta x_i$. O que impossibilita a utilização direta do método dos mínimos quadrados para estimação dos parâmetros. Dessa forma, o método de estimação adotado é o da máxima verossimilhança, conforme proposto por Cox (1958), considerando $F_i = p(x_i)$, a função de verossimilhança para uma amostra de tamanho n é dada por:

$$L(\alpha, \beta|y) = \prod_{i=1}^{n} p(x_i)^y (1 - p(x_i))^{1-y_i} = \prod_{i=1}^{n} {F_i}^{y_i} (1 - F_i)^{1-y_i}$$

Com o log da verossimilhança, note que:

$$logL(\alpha, \beta|y) = \sum_{i=1}^{n} [\log(1 - F_i) + y_i \log(\frac{F_i}{1 - F_i})]$$

Para testar a significância do modelo deve-se avaliar a hipótese $H_0 = \beta = 0$, pois como é observada em uma regressão linear simples, esta hipótese declara que não existe relação entre a variável preditora e a variável resposta.

A estatística do teste de Wald (1943), é dada por: $Z = \frac{\hat{\beta}}{se(\hat{\beta})}$, onde se é o erro padrão (standard error), e possui aproximadamente uma distribuição normal padrão. No caso de H_0 verdadeiro e a amostra seja grande o suficiente. Logo, H_0 pode ser rejeitada se $|Z| \geq Z_{\frac{\alpha}{2}}$, alternativamente a hipótese pode ser testada como *log* da estatística do teste da razão de verossimilhança:

$$-2 \log \lambda (y^*) = 2\left[\log\left(\hat{\alpha}, \hat{\beta}\middle|\hat{y}\right)\right] - \log(\widehat{\alpha_0}, 0|y^*),$$

onde $\widehat{\alpha_0}$ são o Estimador de Máxima Verossimilhança para α assumindo $\beta = 0$.

Com argumentos binomiais padrão, Freitas (2013) demonstra que é possível mostrar que:

$$\widehat{\alpha_0} = \sum_{i=1}^{n} \frac{y_i}{n}.$$

Revista Mirante, Anápolis (GO), v. 18, n. 1, edição extra, p. 17-38, jun. 2025. ISSN 1981-4089

Portanto, pode-se concluir que, de acordo com H_0, $-2\log\lambda$ possui uma distribuição aproximada χ_1^2 e a hipótese H_0 pode ser rejeitada no nível α se $-2\log\lambda \geq \chi_{1,\alpha}^2$.

Critério de seleção do modelo

Umas das formas de selecionar o modelo mais adequado é observar o critério de Akaike (AIC), definido por:

$$AIC = -2\log(\hat{L}) + 2K,$$

onde $\hat{L}$ é a log-verossimilhança maximizada e K é o número de parâmetros do modelo.

Segundo este critério, o melhor modelo é o que apresenta menor valor AIC. Quanto mais parâmetros, maior o AIC.

Curvas ROC

Para a avaliação do uso de duas métricas foi utilizada a técnica de ROC (Receiver Operating Characteristics Curves) (Fawcett, 2006).

A avaliação baseada em coluna única, ou seja, a taxa de positivos reais *(tp rate*) e a taxa dos falsos positivos (*fp rate*), que são definidos da seguinte forma:

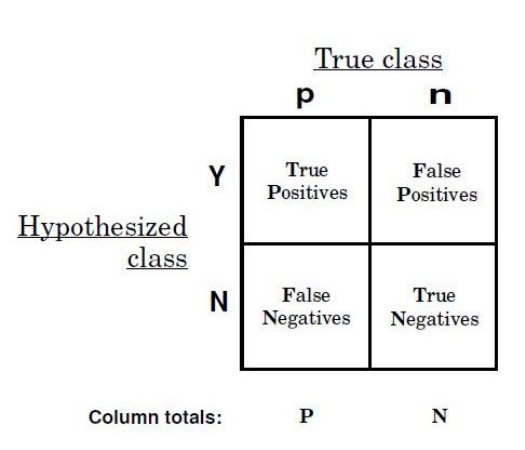
$$fp\ rate = \frac{FP}{N} \qquad tp\ rate = \frac{TP}{P}$$

$$precision = \frac{TP}{TP + FP} \qquad acuracy = \frac{TP + TN}{P + N}$$

O gráfico da curva ROC, conforme a Figura 1, foi formada pela tp rate sobre a taxa fp rate, onde cada ponto no espaço ROC corresponde ao desempenho de um único classificador em uma determinada distribuição.

Revista Mirante, Anápolis (GO), v. 18, n. 1, edição extra, p. 17-38, jun. 2025. ISSN 1981-4089

Figura 1. Tipo de transações



Fonte: https://www.scielo.br/j/ambiagua/a/PdyR9MWQPwJR5P67jWSctZg/?lang=pt

O ponto positivo em utilizar a curva ROC é a possibilidade de observar visualmente as compensações relativas entre os benefícios (que são representados pelos verdadeiros positivos) e os respectivos custos (que são os falsos positivos) de classificação.

Matriz de confusão

A matriz de confusão foi usada para medir a efetividade do modelo de classificação, ao mostrar o número de classificações corretas e as classificações preditas para cada classe em um determinado conjunto de dados.

Está métrica tem uma medição de desempenho para o problema de classificação de aprendizado de máquina em que a saída pode ser de duas ou mais classes.

Acurácia balanceada

A precisão balanceada é a média entre a sensibilidade e a especificidade, que mede a precisão média obtida nas classes minoritária e majoritária. Essa quantidade se reduz à precisão tradicional se um classificador tiver um desempenho igualmente bom em ambas as classes.

A função calcula a precisão balanceada o que evita estimativas de desempenho infladas em conjuntos de dados desequilibrados.

Revista Mirante, Anápolis (GO), v. 18, n. 1, edição extra, p. 17-38, jun. 2025. ISSN 1981-4089

Em cada amostra, a precisão é ponderada de acordo com a prevalência inversa da sua classe verdadeira. Para conjuntos de dados balanceados, a pontuação da precisão balanceada é equivalente à precisão simples. No caso binário, a precisão balanceada corresponde à média aritmética entre sensibilidade (taxa de verdadeiros positivos) e especificidade (taxa de verdadeiros negativos). Alternativamente, pode ser interpretada como a área sob a curva ROC calculada a partir de previsões binárias, em vez de pontuações contínuas.

Matematicamente, a precisão balanceada é dada por:

$$\text{Precisão Balanceada} = \ especificidade\ x\ \frac{1}{2}(sensibilidade).$$

Esse conceito e métrica têm sido amplamente discutidos em estudos sobre avaliação de classificadores, como apresentado por Brodley e Kubat (1998) e Powers (2011).

Dados desbalanceados

Uma estratégia muito utilizada na construção de amostra para o ajuste de modelos de regressão logística, na situação de dados desbalanceados é selecionar uma amostragem contendo todos os eventos presentes na base original de dados via amostragem aleatória (simples sem reposição). O número de não eventos deve ser igual ou superior aos números de eventos (Barella, 2016).

A sub amostragem tenta reduzir o viés associado a classes de dados não balanceadas. No aprendizado de máquina, sub amostragem e super amostragem são duas técnicas que lidam com desequilíbrios em um conjunto de treinamento.

Foram empregadas as técnicas de amostragens a fim de melhorar a precisão de previsão. Há um pacote no software R (v. 4.1.3) que fornece uma função chamada *ovun.sample* que permite oversampling, undersampling de uma só vez.

Resultados e discussão

Os processos financeiros de uma empresa têm diferentes nomenclaturas. Cash in e cash out são apenas nomes que se podem dar para os processos de receber e pagar contas, respectivamente. O cash in seria o dinheiro via recebimento que se pode entender também como

Revista Mirante, Anápolis (GO), v. 18, n. 1, edição extra, p. 17-38, jun. 2025. ISSN 1981-4089

depósito e o cash out é uma forma de saque. Assim como para o cash in ou cash out, existem várias formas de uma empresa no meio financeiro receber recursos e fazer pagamentos, sejam para fornecedores, funcionários, parceiros e entre outros. Por exemplo, aplicativos como delivery precisam fazer muitos pagamentos de uma vez, dependendo do tamanho do negócio ou do número de estabelecimentos cadastrados, são milhares de transferências de valores a serem feitas.

As variáveis utilizadas estão descritas na Tabela 1, disponíveis na base de dados, onde foram sinalizadas as ocorrências fraudulentas. A partir dessas informações será possível encontrar padrões, fazer os apontamentos de incidentes e realizar predições de risco à empresa.

Tabela 1. Variáveis do dataset

Variável	Descrição da variável
tempo	tempo da transação
tipo	tipo de ocorrência da transação
valor	valor transacional
cli_inic_trans	cliente que iniciou a transação
atg_sld_org	saldo inicial antes da transação
nv_sld_org	novo saldo após a transação
clie_dst	cliente que é o destinatário da transação
atg_sld_dst	destinatário do saldo inicial antes da transação
nv_sld_dst	novo destinatário do saldo após a transação.
fraude	variável target

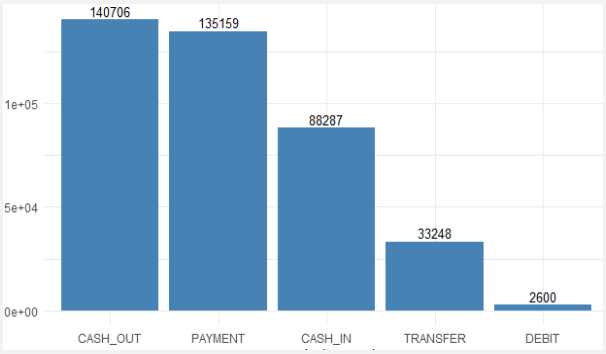
Fonte: PINTO, B. R e RIBEIRO, L. F. C.

A base de dados engloba informações importantes, como por exemplo: O tempo que a transação ocorreu, o valor envolvido e o saldo antes e depois da transação ocorrer.

Outro ponto importante é que, pode-se identificar o cliente envolvido e isso nos ajuda a apontar qual a melhor decisão a aplicar, a fim de não o impactar financeiramente ou a empresa. As transações ocorridas por tipo de transferência via pagamento, cash in, cash out, débito ou pagamento estão descritas na Figura 2. Observe que, 91% das transações realizadas giram por meio de cash in, cash out e payment.

Revista Mirante, Anápolis (GO), v. 18, n. 1, edição extra, p. 17-38, jun. 2025. ISSN 1981-4089

Figura 2. Tipo de transações



Fonte: PINTO, B. R e RIBEIRO, L. F. C.

Para Gil, Arima e Nakamura (2013) o risco está relacionado com a incerteza e com a dificuldade de controle sobre os eventos futuros. Assi (2013) afirma que a principal característica é a possibilidade de ocorrência (pode ser que ocorra ou não), podendo, por isso, ser considerado como uma ameaça para o alcance dos objetivos organizacionais.

Em um mundo de pagamentos e com a evolução constante da tecnologia as empresas estão vulneráveis a ocorrência de transações fraudulentas devido a facilidade no comércio.

Coimbra (2007) cita como principais riscos operacionais: fraudes internas e externas, práticas internas relacionadas a empregados e clientes, danos a ativos, interrupção e falhas nas atividades e sistemas, entre outros. Os riscos são inerentes a todos os segmentos do mercado, assim, todas as empresas estão suscetíveis a eles.

Assi (2013) aponta que o principal desafio das empresas é fazer com que os colaborados entendam a estratégia e a perspectiva de riscos, pois quando as pessoas têm conhecimento dos riscos referentes às suas atividades, estes podem ser identificados, auxiliando o seu gerenciamento.

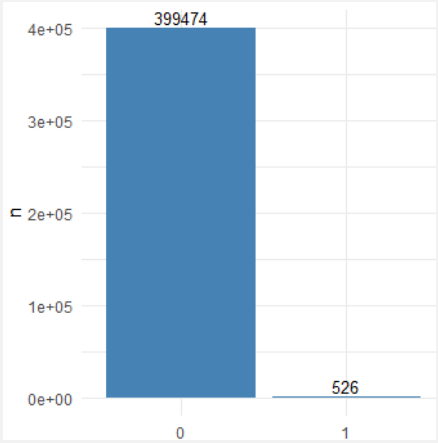
O risco, apesar de estar presente em todas as empresas, tem uma relevância maior quando se trata de instituições financeiras, já que estas atuam diretamente com o crédito e podem estar expostas a ataques de fraude. Para reduzir os prejuízos, a gestão de riscos e principalmente as ferramentas estatísticas se tornam essenciais nessa frente.

A base de dados está sinalizada com a ocorrência de um risco ou não por meio de uma marcação (flag), em que, caso igual a 1 é uma ação fraudulenta e no caso 0 é um apontamento de uma não fraude.

Revista Mirante, Anápolis (GO), v. 18, n. 1, edição extra, p. 17-38, jun. 2025. ISSN 1981-4089

Embora as transações fraudulentas correspondam a apenas 0,1315% das 400 mil operações analisadas (526 transações), é relevante observar que essas operações movimentaram aproximadamente 3,91% do valor total transacionado. Este dado evidencia que, mesmo em número reduzido, as fraudes concentram volumes financeiros significativamente altos, justificando a necessidade de se adotar modelos estatísticos robustos e técnicas de classificação mais refinadas. Assim, a aplicação de metodologias como a regressão logística torna-se crucial para minimizar impactos financeiros, especialmente diante da elevada assimetria entre frequência e impacto financeiro desses eventos (Figura 3).

Figura 3. Quantidade de fraude e não fraude nos dados



Fonte: PINTO, B. R e RIBEIRO, L. F. C.

Fazendo um estudo das informações nas bases analisadas, nada mais convencional que utilizar análises descritivas e gráficos para explorar o comportamento das variáveis no meio de pagamentos.

A Tabela 2 apresenta o sumário das variáveis, observa-se que a mediana possui valor aproximado da média apenas no tempo. Os valores em si são bem discrepantes quando atenta-se para o mínimo e o máximo de cada variável. Isso faz sentido quando se olha para o perfil de pessoas, em que, há comportamentos diferentes por conta da vida financeira de cada um, umas com mais condições e outas com menos em fazer qualquer tipo de circulação financeira.

Tabela 2. Sumário das variáveis.

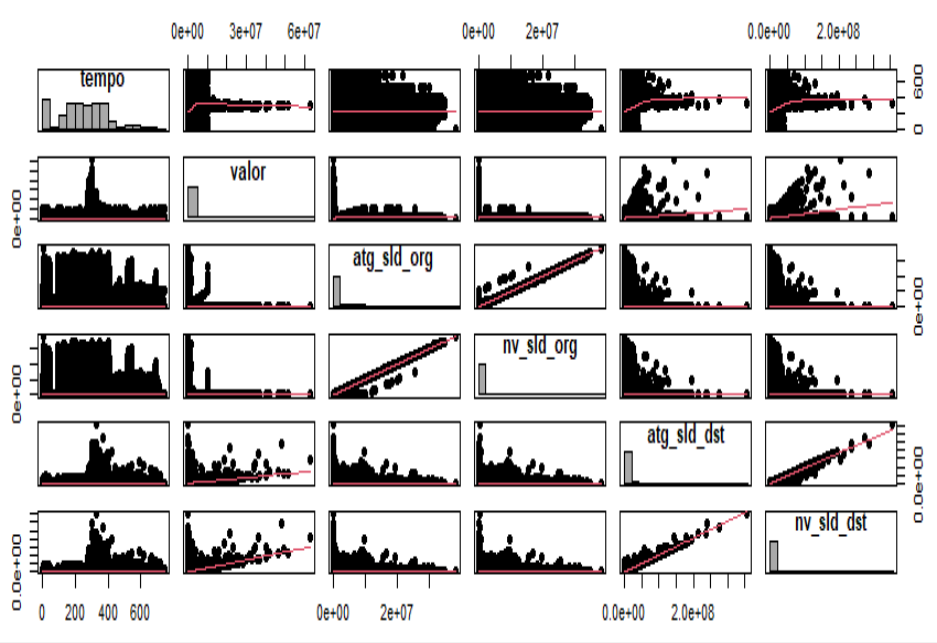
REVISTA MIRANTE
Daqui se vê do alto!
Câmpus Anápolis de Ciências Exatas e Tecnológicas Henrique Santillo
Universidade Estadual de Goiás
Revista Mirante, Anápolis (GO), v. 18, n. 1, edição extra, p. 17-38, jun. 2025. ISSN 1981-4089

Medidas	tempo	valor	atg_sld_org	nv_sld_org	atg_sld_dst	nv_sld_dst
Mínimo	1	0	0	0	0	0
1º Quartil	156	13.436	0	0	0	0
Mediana	240	75.035	14.219	0	133.600	215.872
Média	244	178.528	839.868	861.180	1.104.434	1.226.411
3º Quartil	335	208.973	107.200	144.232	944.008	1.112.261
Máximo	741	62.785.417	38.166.700	38.259.597	355.553.416	355.381.434

Fonte: PINTO, B. R e RIBEIRO, L. F. C.

O gráfico de dispersão, conforme a Figura 4, nos dá indícios de variáveis que possuem forte tipo de relação linear, como é perceptível a variável “atg_sld_org” está fortemente relacionada com “nv_sld_org” e as variáveis “atg_sld_dst” com “nv_sld_dst”.

Figura 4. Dispersão das variáveis.

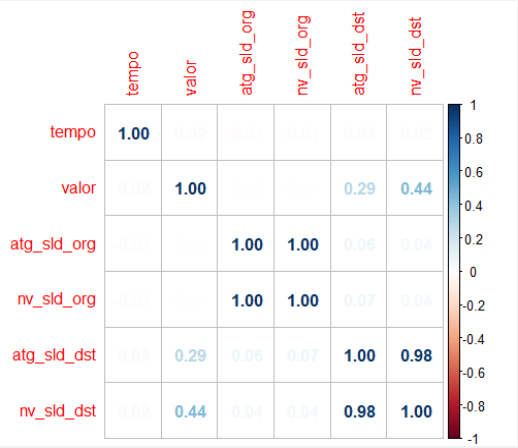


Fonte: PINTO, B. R e RIBEIRO, L. F. C.

É possível verificar, conforme a Figura 5, visualmente a matriz de correlações, uma outra métrica que nos ajuda a concluir sobre a relação linear, fica evidente a correlação muito forte das variáveis referente a saldos antigos e novos após a ocorrência da transação.

REVISTA MIRANTE
Daqui se vê do alto!
Câmpus Anápolis de Ciências Exatas e Tecnológicas Henrique Santillo
Universidade Estadual de Goiás
Revista Mirante, Anápolis (GO), v. 18, n. 1, edição extra, p. 17-38, jun. 2025. ISSN 1981-4089

Figura 5. Correlação das variáveis.



Fonte: PINTO, B. R e RIBEIRO, L. F. C.

Essa correlação se deve por conta da seguinte influência. Veja, quando o dinheiro é enviado a alguém, essa pessoa automaticamente irá ter um aumento em sua conta, enquanto a outra pessoa que enviou terá seu saldo menor em sua conta.

Matematicamente, correlações altas entre variáveis independentes causam uma instabilidade numérica ao ajustar a curva de regressão, o chamado efeito de multicolinearidade, em outras palavras, redundância (Miloca, 2009).

Existem funções no próprio software "R" que já fazem o ajuste do melhor modelo e será utilizado esse ferramental nas aplicações.

O modelo ajustado apresentou uma acurácia de 99,8%. Mas, se observarmos a matriz de confusão, conforme a Tabela 3, esse modelo classificou melhor as operações não fraudulentas.

Tabela 3. Matriz de confusão.

	fraude	não fraude
fraude	276	30
não fraude	250	399.444

Fonte: PINTO, B. R e RIBEIRO, L. F. C.

O modelo conseguiu classificar corretamente as operações fraudulentas em 52,5% das vezes. O que pode ser um problema quando for aplicar alguma política de fraude, pois esse

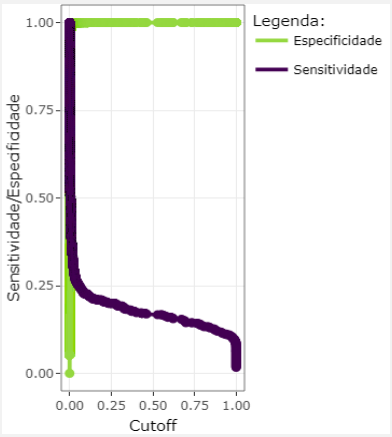
Revista Mirante, Anápolis (GO), v. 18, n. 1, edição extra, p. 17-38, jun. 2025. ISSN 1981-4089

falso positivo pode impactar negativamente os clientes bons, ou seja, as operações não fraudulentas. O que pode ocasionar um problema na regra de negócio da empresa.

O fato de as fraudes ocorrerem em menos de 1% das operações pode prejudicar a classificação modelo.

Portanto, é evidente o processo de balancear a base para que se possa melhorar a classificação do evento fraudulento. Visto que, na Figura 6, houve uma grande perca na especificidade, visualizando o gráfico da curva ROC.

Figura 6. Curva ROC - Modelo desbalanceado



Fonte: PINTO, B. R e RIBEIRO, L. F. C.

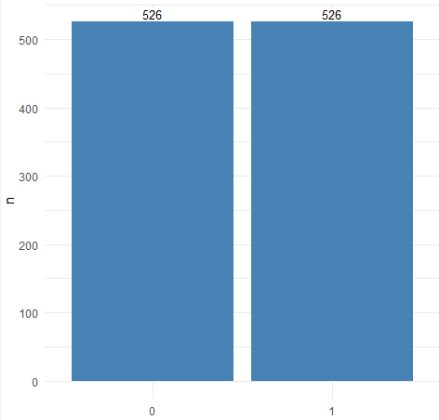
De um modo geral a curva ROC conseguiu classificar corretamente todas aquelas transações não fraudulentas. Ou ainda, se olhar por outro ângulo, caso o modelo classifique uma transação que não conteve fraude, e de fato não foi, ele teve um acerto bem significativo. No entanto, se o modelo classificar alguma transação como não fraudulenta, seria necessário fazer outra checagem para avaliar se de fato a transação não foi uma fraude, pois conforme visto na matriz de confusão há 47% de chance de uma não fraude ser fraude.

A capacidade de distinguir entre transações legítimas e fraudulentas é um problema reconhecido nesta área e o alto desbalanceamento usualmente encontrado nas classes, que pode comprometer o desempenho dos classificadores (Nicola, V., Lauretto, M., E Delgado, K. V., 2020). A fim de melhorar a classificação neste trabalho, usou-se como base este artigo para melhorar os resultados.

Revista Mirante, Anápolis (GO), v. 18, n. 1, edição extra, p. 17-38, jun. 2025. ISSN 1981-4089

Pode-se verificar, conforme a Figura 7, que as transações fraudulentas agora representam agora 50% dos eventos. Em contrapartida, no primeiro modelo a fraude estava contida em apenas 0,13% da base.

Figura 7. Base Balanceada



Fonte: PINTO, B. R e RIBEIRO, L. F. C.

Portanto, será utilizado técnicas de balanceamento (sub amostragem) nos dados para avaliar melhorias nos resultados obtidos neste primeiro modelo apresentado.

Ajustando o modelo balanceado, tem-se que:

$$\widehat{Fraude}_i = 1/1 + exp - \{-208.38 + tempo_i * 0.0034 + amount_i * 0.000008 + atg\_sld\_org_i * 0.000021 + nv\_sld\_org_i * 0.000015\}, \text{ onde } i = 1,2, \dots, 1052.$$

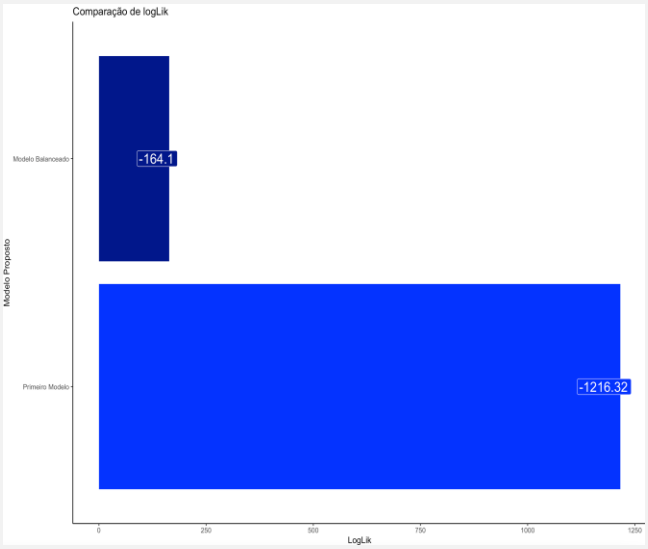
Diferente como no primeiro modelo proposto, apenas as variáveis tempo, *amount*, *atg_sld_org* e *nv_sld_org* apresentaram estatisticamente significante a um α de 5%.

O valor de probabilidade logarítmica (logLik) de um modelo de regressão é uma maneira de medir a qualidade do ajuste de um modelo. Quanto maior o valor da probabilidade de log, melhor o modelo se ajusta a um conjunto de dados (Silva, Stefani, 2019).

Na Figura 8, o modelo balanceado apresentou menor probabilidade logarítmica, considerando dessa forma, que de fato, o modelo com dados balanceados é a melhor proposta neste trabalho para a classificação de eventos fraudulentos.

Revista Mirante, Anápolis (GO), v. 18, n. 1, edição extra, p. 17-38, jun. 2025. ISSN 1981-4089

Figura 8. Comparação do LogLik nos modelos



Fonte: PINTO, B. R e RIBEIRO, L. F. C.

Aplicando um modelo balanceado com o processo de sub amostragem, pode-se perceber que já houve um ganho significante no logLik.

Observando o critério de AIC, pode-se ver que o menor valor apresentado é do modelo balanceado, conforme a Tabela 4.

Tabela 4. Critério de seleção do modelo.

Modelos	Valor critério
Primeiro modelo	2.454,6
Modelo balanceado	350,21

Fonte: PINTO, B. R e RIBEIRO, L. F. C.

Com o comportamento melhor do modelo balanceado, o próximo passo será avaliar a matriz de confusão e verificar como foi a classificação do evento fraudulento. Lembrando que, nesse ajuste o evento da fraude é 50%, enquanto as transações que não apresentaram riscos representam 50%.

Observem que a matriz de confusão, Tabela 5, do modelo balanceado classificou significativamente bem as transações fraudulentas, 94% das vezes encontrou o evento de interesse contra 52% no primeiro modelo proposto, melhorando nossos apontamentos do risco da fraude. É possível verificar também que o modelo balanceado classificou bem o evento não

Revista Mirante, Anápolis (GO), v. 18, n. 1, edição extra, p. 17-38, jun. 2025. ISSN 1981-4089

fraude, o que não gera falso positivo, conforme visto no primeiro modelo proposto nesse trabalho.

Tabela 5. Matriz de confusão.

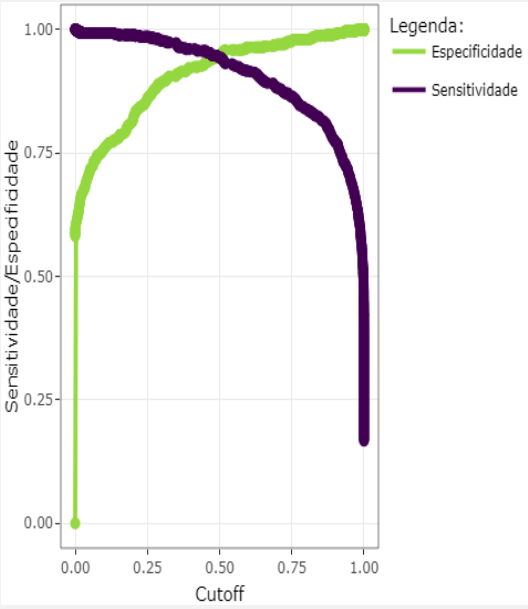
	fraude	não fraude
fraude	496	27
não fraude	30	499

Fonte: PINTO, B. R e RIBEIRO, L. F. C.

A acurácia do modelo proposto ficou em 95%, bem menor que o primeiro modelo proposto. Contudo, o interesse é classificar com uma assertividade o evento fraudulento, atendendo dessa forma a expectativa dos autores.

Na Figura 9, como a curva ROC apresentou uma melhora significativa em relação ao primeiro modelo proposto.

Figura 9. Curva ROC - Modelo Balanceado.



Fonte: PINTO, B. R e RIBEIRO, L. F. C.

O equilíbrio em encontrar o evento de fraude com eventos não fraudulentos, são de extrema rigidez no mercado financeiro.

Revista Mirante, Anápolis (GO), v. 18, n. 1, edição extra, p. 17-38, jun. 2025. ISSN 1981-4089

Conclusões

As fraudes financeiras corresponderam a apenas 0,1315 % das 400 000 transações avaliadas; todavia, concentraram 3,91 % do montante financeiro movimentado, evidenciando a assimetria entre frequência e impacto e reforçando a necessidade de empregar técnicas estatísticas específicas para sua detecção.

O modelo estimado com a base desbalanceada apresentou elevada acurácia global, mas baixa sensibilidade, identificando pouco mais da metade dos eventos fraudulentos. Já o modelo construído a partir da amostra balanceada alcançou sensibilidade de 94 %; embora tenha exibido discreta redução na acurácia total, mostrou-se muito mais eficaz na identificação de fraudes, que é o objetivo central deste estudo.

A análise comparativa das curvas ROC, dos valores de log-verossimilhança e do critério de informação de Akaike (AIC) confirma a superioridade do modelo balanceado na classificação do evento de interesse. Conclui-se, portanto, que modelos logísticos ajustados para bases desbalanceadas constituem instrumento analítico de elevado valor para mitigar fraudes no setor financeiro.

Recomenda-se, por fim, que a avaliação de modelos de detecção considere, além da acurácia global, métricas voltadas à classe minoritária, sobretudo sensibilidade e especificidade, de modo a alinhar o desempenho estatístico às exigências estratégicas das instituições financeiras.

Referências

ASSI, M. Gestão de riscos com controles internos. p. 103-130. In: ASSI, M. **Gestão de Riscos com Controles Internos**, 2ed. Saint Paul Editora, São Paulo, SP, Brasil, 2021.

ARAGÃO, D. F. D. Crimes cibernéticos na pós-modernidade: direitos fundamentais e a efetividade da investigação criminal de fraudes bancárias eletrônicas no Brasil. **Dissertação de pós-graduação em ciências sociais**. Universidade Federal do Maranhão, Maranhão, MA, Brasil, 2015.

Revista Mirante, Anápolis (GO), v. 18, n. 1, edição extra, p. 17-38, jun. 2025. ISSN 1981-4089

BARELLA, V. H. Técnicas para o problema de dados desbalanceados em classificação hierárquica. **Dissertação de doutorado**. Universidade de São Paulo, São Carlos, SP, Brasil, 2016.

ARROSO, L. C. Tecnologia Bancária: evolução recente e tendências. In: Fortaleza: Banco do Nordeste do Brasil, 2018, Fortaleza, CE, Brasil. **Anais...** p. 01-24, 2018.

BOLTON, R. J.; HAND, D. J. Statistical fraud detection: A review. In: **Statistical Science**, Edimburgo, Reino Unido, Escócia, 2002. Disponível em: https://doi.org/10.1214/ss/1042727940. Acesso em: 01 de maio de 2022.

BRODLEY, C. E.; KUBAT, M. **Learning when training data are costly: The effect of training set size on classifier performance.** In: INTERNATIONAL CONFERENCE ON MACHINE LEARNING – ICML, 15, 1998, Madison. Proceedings […]. San Francisco: Morgan Kaufmann, 1998. p. 214-222.

CARVALHO ALVES, L.; GONÇALVES, F. V.; MOIZINHO, L. C. S. O custo da fraude: uma análise de um ecommerce brasileiro. In: **Anais** do Congresso Brasileiro de Custos-ABC. São Leopoldo, RS, Brasil, 2013. Disponível em: https://anaiscbc.emnuvens.com.br/anais/article. Acesso em: 18 maio de 2022.

COIMBRA, F. Riscos operacionais: estrutura para gestão em bancos. **Tese de doutorado em Administração**. Universidade de São Paulo, São Paulo, SP, Brasil, 2007.

COX, D. R. The regression analysis of binary sequences. **Journal of the Royal Statistical Society**. Series B (Methodological), v. 20, n. 2, p. 215-242, 1958.

DELGADO, S. L. Detecção de fraude em e-commerce: aplicação para uma empresa que atua no Brasil. **Monografia** – título em bacharel em Ciências Atuariais. Osasco, São Paulo, SP, Brasil, 2022.

FAWCETT, T. An introduction to ROC analysis. In: **Pattern recognition letters**, Roma, Itália, 2006. p. 861-874.

FIGUEIRA, C. V. Modelos de regressão logística. **Dissertação** de mestre em matemática. Universidade Federal do Rio Grande do Sul, Rio Grande do Sul, RS, Brasil, 2006.

FREITAS, L. D. R. Comparação das funções de ligação logit e probit em regressão binária considerando diferentes tamanhos amostrais. **Dissertação** de mestre em Estatística aplicada e Biometria. Universidade Federal de Viçosa, Viçosa, MG, Brasil, 2013.

GIL, A. D. L.; ARIMA, C. H.; NAKAMURA, W. T. Gestão de Riscos: controle interno, risco e auditoria. p. 100-126. In: GIL, A. D. L., ARIMA, C. H., & NAKAMURA, W. T. **Gestão: controle interno, risco e auditoria**. 2ed, Saraiva, São Paulo, SP, Brasil, 2013.

Revista Mirante, Anápolis (GO), v. 18, n. 1, edição extra, p. 17-38, jun. 2025. ISSN 1981-4089

HOSMER, D. W.; LEMESHOW, S.; STURDIVANT, R. X. **Applied Logistic Regression**. 3rd ed. Hoboken: Wiley, 2013.

KRAUTER, E.; FAMÁ, R. Um estudo sobre a evolução da função financeira da empresa. In: **SEMEAD. VIII–Seminário em administração** FEA-USP, São Paulo, SP, Brasil, 2005. Anais... p. 3-10.

MACHADO, M. R. R. Investigação da ocorrência de fraudes corporativas em instituições bancárias brasileiras à luz do triângulo de fraude de Cressey. **Tese** de doutorado em Administração. Universidade de Brasília, Brasília, Brasil, 2015.

MILOCA, S. A.; CONEJO, P. D. Análise fatorial e a multicolinearidade em modelos de regressão. In: **Encontro Regional de Matemática Aplicada e Computacional**, Pato Branco, Paraná, PR, Brasil, 2009. Anais... p. 10-13.

NICOLA, V.; LAURETTO, M.; DELGADO, K. V. Avaliação empírica de classificadores e métodos de balanceamento para detecção de fraudes em transações com cartões de créditos. In: **Anais** do XVII Encontro Nacional de Inteligência Artificial e Computacional. SBC, 2020. p. 70-81.

POWERS, D. M. W. Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation. **Journal of Machine Learning Technologies**, v. 2, n. 1, p. 37-63, 2011.

RIPLEY, B. D. **Pattern recognition and neural networks**. Cambridge University Press, 1: 05-122, 2007.

SILVA, M. D. C. Crédito bancário e desenvolvimento sustentável nas instituições financeiras brasileiras. **Tese** de mestrado em Desenvolvimento Sustentável. Universidade de Brasília, Brasília, Brasil, 2011.

SILVA, S. Classificação de risco de crédito: um comparativo entre modelos de análise discriminante e regressão logística nas empresas de capital aberto brasileiras. In: **Revista de Finanças e Contabilidade da Unimep**, Piracicaba, SP, Brasil, 2019.

SOUZA, I. M. D. A.; CAITITÉ, A. M. L. A incrível história da fraude dos embriões clonados e o que ela nos diz sobre ciência, tecnologia e mídia. In: **História, Ciências, Saúde-Manguinhos**, Rio de Janeiro, RJ, Brasil, 2009. Anais... p. 02-23.

WALD, A. Tests of statistical hypotheses concerning several parameters when the number of observations is large. In: **Transactions of the American Mathematical Society**, Providence, Ilha, EUA, 1943. Anais… p. 01-57.