MACHINE LEARNING NA PREVENÇÃO DE PERDAS NO E-COMMERCE: O USO DA REGRESSÃO LOGÍSTICA PARA IDENTIFICAÇÃO DE PEDIDOS FRAUDULENTOS

MACHINE LEARNING IN E-COMMERCE LOSS PREVENTION: USING LOGISTIC REGRESSION TO IDENTIFY FRAUDULENT ORDERS

MACHINE LEARNING PARA PREVENIR PÉRDIDAS EN EL E-COMMERCE: EL USO DE LA REGRESIÓN LOGÍSTICA PARA IDENTIFICAR PEDIDOS FRAUDULENTOS

Kevin William Matos Paixão[1]
Gabrielle Maria Romeiro Lombardi[2]
Paulo Ricardo de Andrade Barroso[3]

Artigo recebido em dezembro de 2024
Artigo aceito em fevereiro de 2025

DOI: 10.26853/Refas_ISSN-2359-182X_v11n04_04

RESUMO

Nos últimos 25 anos, diversos meios de detecção e prevenção de fraude foram desenvolvidos. Nos dias atuais os modelos de aprendizado de máquina, "machine learning", são os mais avançados. Neste trabalho foi desenvolvido um modelo estatístico de machine learning cujo objetivo é identificar, por meio da técnica de regressão logística, a probabilidade de um pedido ser fraude. O modelo foi desenvolvido no software R e o dataset utilizado contém 13216 observações e 15 variáveis. Como resultado o modelo obteve 91,4% de acurácia, 87,31% de especificidade e 97,7% de sensibilidade, tal performance resultou em uma área abaixo da curva ROC de 95,1% e um Coeficiente de GINI de 90,21%. Como principais contribuições desta obra destacam-se a promoção e divulgação da utilização de técnicas de machine learning para resolução de problemas reais e cotidianos no e-commerce, e o esclarecimento e fomento do processo de desenvolvimento de modelos de regressão logística, bem como dos seus principais parâmetros de validação e indicadores de performance. Com base nos resultados já apresentados é possível afirmar que o objetivo desta pesquisa foi atingido, pois o modelo desenvolvido é altamente eficaz na prevenção de fraudes.

Palavras-chave: Regressão logística; Modelo estatístico; Fraude; e-commerce; Algoritmo.

[1] Especialista em Ciência de dados. USP-ESALQ. E-mail: paixão-k@hotmail.com. Lattes: http://lattes.cnpq.br/8596948853024607. OrcId: https://orcid.org/0000-0002-3795-9712.
[2] Doutora em Genética e Melhoramento de Plantas. Professora Orientadora - Pecege. Email: glombardi.pecege@gmail.com. Lattes: http://lattes.cnpq.br/1903604573471325. OrcId: https://orcid.org/0000-0002-0166-1260.
[3] Especialista em ciência de dados. USP-ESALQ. Email: pauloricardoandrad@gmail.com. Lattes: http://lattes.cnpq.br/5865294778495081. OrcId: https://orcid.org/0009-0004-0980-9970.

ABSTRACT

In the last 25 years, several means of detecting and preventing fraud have been developed, nowadays machine learning models are the most advanced. In this work, a statistical machine learning model was developed whose objective is to identify, using the Logistic Regression technique, the probability of an order being fraudulent. The model was developed in the R software and the data set used contains 13216 observations and 16 variables. As a result, the model obtained 91.4% accuracy, 87.31% sensitivity and 97.7% specificity, such performance proven in an area below the ROC curve of 95.1% and a GINI Coefficient of 90.21 %. The main contributions of this work include the promotion and dissemination of the use of machine learning techniques to solve real and everyday problems in e-commerce, and the clarification and promotion of the process of developing logistic regression models, as well as their main validation parameters and performance indicators. Based on the results already presented, it is possible to affirm that the objective of this research was achieved, as the model developed is highly effective in preventing fraud.

Keywords: Logistic regression; Statistical model; Fraud; e-commerce; Algorithm.

RESUMEN

Durante los últimos 25 años, se han desarrollado varios medios para detectar y prevenir el fraude. Hoy en día, los modelos de aprendizaje automático son los más avanzados. En este trabajo se desarrolló un modelo estadístico de aprendizaje automático cuyo objetivo es identificar, mediante la técnica de regresión logística, la probabilidad de que un pedido sea un fraude. El modelo fue desarrollado en el software R y el conjunto de datos utilizado contiene 13216 observaciones y 15 variables. Como resultado, el modelo obtuvo 91.4% de precisión, 87.31% de especificidad y 97.7% de sensibilidad, este desempeño resultó en un área bajo la curva ROC de 95.1% y un Coeficiente de GINI de 90.21%. Las principales aportaciones de este trabajo incluyen la promoción y difusión del uso de técnicas de aprendizaje automático para la resolución de problemas reales y cotidianos en el comercio electrónico, y la clarificación y promoción del proceso de desarrollo de modelos de regresión logística, así como sus principales parámetros de validación. e indicadores de desempeño. Con base en los resultados ya presentados, es posible afirmar que se logró el objetivo de esta investigación, ya que el modelo desarrollado es altamente efectivo para prevenir el fraude.

Palabras clave: Regresión logística; Modelo estadístico; Fraude; comercio electrónico; Algoritmo.

1 INTRODUÇÃO

O comércio eletrônico é um dos maiores avanços da era moderna (ZHANG ET AL., 2023). O comércio eletrônico ou e-commerce, como é conhecido mundialmente, refere-se a compra e venda de produtos online, e sua estrutura possui duas formas principais de negócio, sendo elas: i. o business to consumer (B2C), modelo em que uma organização negocia diretamente com o consumidor final e ii. o business to business (B2B), modelo em que uma organização faz transações comerciais com outra (YU ET AL., 2016).

No ambiente global de negócios o e-commerce impulsionou empresas em diversos países e promoveu o desenvolvimento econômico mundial (MORTEZA ET AL.,2011), modificando a economia dos Estados Unidos e desempenhando um papel fundamental no crescimento da China nas últimas três décadas (NOVIKOVA, 2020). A China vem liderando o comércio eletrônico, que está expandindo de forma acelerada, e suas transações *online* já ultrapassaram a marca de 1,57 trilhões de dólares , todo esse volume de vendas somados à

participação da Coreia do Sul e do Japão consolidam, o continente da Ásia como líder de vendas *online* com mais de 50% das transações globais (NIELSENIQ, 2022).

No Brasil, o comércio eletrônico se consolidou após a popularização da internet e o avanço das plataformas de rede durante a década de 1990. Desde então o crescimento do e-commerce tornou a comunicação entre as empresas e os consumidores mais rápida e eficiente (ECKERT ET AL., 2020). Ela reduziu custos de administração e operação (MORTEZA ET AL., 2011) e aumentou o desempenho de diversos modelos de negócio (ORDANINI ET AL., 2010). Somente em 2020 o Brasil registrou 17 milhões de novos consumidores online impulsionando os seguintes ramos do comércio: Cuidados Pessoais, Bebês, Cuidados com a Casa e Moda (NIELSENIQ, 2022).

Dentro da realidade exposta anteriormente, o e-commerce possui três principais desafios que seriam: compreender as necessidades dos clientes, superar desafios de segurança e confiabilidade e se adaptar às volatilidades do mercado global (MOREIRA, 2016). Dentre os três desafios o que requer maior atenção é o segundo, pois garantir a confiabilidade e a segurança das informações e prevenir fraudes são ações cruciais para manter a saúde financeira de qualquer organização. Ao mencionar questões relacionadas à segurança, privacidade e confiabilidade nas transações online evidencia-se que os consumidores aumentaram seu nível de cautela e passaram a buscar mais informações sobre os comerciantes e as organizações aumentaram seus esforços na busca da mitigação de prejuízos gerados por compras fraudulentas (FEITOSA E GARCIA, 2016).

No Brasil, durante o primeiro semestre de 2023, 2 milhões de tentativas de fraudes no e-commerce foram identificadas, essa quantidade corresponde a R$ 2,5 milhões e as áreas mais lesadas são: eletroeletrônicos, eletrodomésticos, acessórios, automotivo e Beleza (CLEARSALES, 2023). Diante deste cenário, De Souza e colaboradores (2023), afirmam que existem diversos meios de detecção e prevenção de fraude, contudo, as técnicas mais utilizadas atualmente são os modelos de aprendizado de máquina “machine learning”. Machine learning pode ser definido como um conjunto de técnicas estatísticas que visam obter modelos matemáticos que gerencie o comportamento de uma função através de um processo denominado “treinamento” (BOCHIE ET AL., 2020). O treinamento por sua vez consiste em uma dinâmica de autoajuste de parâmetros baseado em amostras “Samples” dotadas de múltiplos atributos “feature”. Zhang e colaboradores (2021) defendem que as principais técnicas de machine learning aplicadas a prevenção de fraudes são: Redes Neurais rasas, Support Vector Machine, Cadeia de Markov e Regressão Logística.

Fernandes et al. (2021) argumentam que a regressão logística é a melhor técnica para ser aplicada em casos em que a variável de resposta é dicotômica, ou seja, possui apenas 2 opções de resposta (fraude ou não fraude). Isso ocorre porque a regressão logística é uma variação dos modelos lineares generalizados (MLG) onde os valores da variável dependente se limitam aos resultados 0 ou 1. Segundo Kleinbaum e colaboradores (2008) essa limitação de resultado é a principal causa da alta popularidade desta técnica. A utilização da regressão logística para detecção de fraude é um tema que pode ser encontrado em diversos segmentos dentro da literatura acadêmica moderna. Mendonça e colaboradores (2021) aplicam esta técnica para identificar a probabilidade de ocorrências de fraudes corporativas em instituições bancárias brasileiras e obtiveram resultados satisfatórios. Vaseli e colaboradores (2021) aplicaram a regressão logística para identificação de fraudes contábeis em empresas contidas na bolsa de Teerã (irã), o modelo desenvolvido atendeu os requisitos básicos de multicolinearidade e heterocedasticidade e obteve um desempenho satisfatórios que comprovou a hipóteses dos autores.

Considerando tamanha relevância deste tema, este trabalho propõe o desenvolvimento de um modelo estatístico de machine learning para identificar pedidos fraudulentos. O modelo criado possui como método estatístico a regressão logística cuja função é indicar a probabilidade do pedido ser uma fraude. Tal experimento se justifica devido ao fato de a aplicação da técnica escolhida já ter se mostrado eficiente em outros segmentos, contudo, vale ressaltar que o resultado da aplicação deste método em bases de dados oriundas de um e-commerce voltado ao ramo da beleza e bem-estar ainda é desconhecido.

2 REFERENCIAL TEÓRICO

Inicia-se o referencial teórico pela regressão logística.

2.1 Regressão logística

Merghadi e colaboradores (2020) afirmam que a regressão logística (RL) é um dos 10 modelos de machine learning mais populares e que sua origem está na estatística. segundo Dias Filho e companheiro (2007) a RL surgiu na década de 1960 em resposta ao desafio de realização de previsão em casos cujo a variável dependente fosse dicotômica. Na década de 1970 surge o conceito de MLG concebido por Nelder e Wedderburn (1972), este consiste em um conjunto de modelos confirmatórios de regressão lineares, não lineares e exponenciais, a RL então passa ser classificada como um dos vários modelos da MLG. Mukhoty e companheiros (2023) relatam que os modelos generalizados são utilizados para estimar tanto valores qualitativos, quanto quantitativos, e em concordância com esta afirmação, Zabor e companheiros (2022) demonstram que o processo utilizado para estimação consiste na inserção de covariáveis em uma função ou equação. Os modelos lineares generalizados são representados matematicamente da seguinte forma:

$$(1)\ \eta_i = \alpha + \beta_1 * X_{1i} + \beta_2 * X_{2i} + \ldots + \beta_k * X_{ki}$$

Na função acima, conforme Fávero e colaboradores (2017), o segundo elemento da função $(\alpha + \beta_1 * X_{1i} + \beta_2 * X_{2i} + \ldots + \beta_k * X_{ki})$ é denominado "logito" e na literatura é amplamente representado pela letra "z". No logito o "α" representa uma constante do modelo, os "β's" são os parâmetros estimados de cada variável explicativa, os "Xs" são as próprias variáveis explicativas e o subscrito "i" representa as observações. Já no primeiro elemento da equação o "η" representa a função canônica de ligação, que no modelo RL assume a seguinte forma matemática:

$$(2)\ \eta = \ln\left(\frac{P_i}{1-P_I}\right)$$

Na função canônica de ligação o "P" representa a probabilidade de ocorrência do evento de interesse.

Hosmer e Lemeshow (1980), trabalharam algebricamente a relação entre o logito e a função canônica, o desenvolvimento matemático proposto:

$$(3)\qquad \eta_i = z$$

$$(4) \quad \ln\left(\frac{P_i}{1-P_I}\right) = z$$

$$(5) \quad \frac{P_i}{1-P_I} = e^z$$

$$(6) \quad P_i = (1- P_i) * e^z \quad \rightarrow \quad P_i = e^z - P_i e^z$$

$$(7) \quad \frac{P_i}{P_i} = \frac{e^z - P_i e^z}{P_i}$$

$$(8) \quad 1 = \frac{e^z}{P_i} - e^z$$

$$(9) \quad \frac{e^z}{1+e^z} = P_i$$

$$(10) \quad P_i = \frac{1}{1+e^{-z}}$$

Dado o desenvolvimento acima definiu-se a expressão geral do modelo de regressão logística como:

$$(11)\ P_i = \frac{1}{1+ e^{-(\alpha+\beta_1*X_{1i}+\beta_2*X_{2i}+\ldots+\beta_k*X_{ki})}}$$

Fávero e colaboradores (2017) argumentam que em decorrência do fato da variável dependente ser sempre qualitativa e com apenas 2 categorias o modelo RL possui duas características importantes:

(i) A distribuição estatística em que a variável dependente se adere é a distribuição Bernoulli. A função densidade probabilidade (DPF) da distribuição Bernoulli é representada matematicamente da seguinte forma:

$$(12)\ P(Y_i) = P_i^{Y_i} * (1-P)^{1-y_i}$$

(ii) A estimação dos parâmetros da equação de probabilidade é realizada por meio do estimador por verossimilhança (likelihood estimation). O estimador por verossimilhança aplicado a DPF da distribuição Bernoulli é representado pelo seguinte modelo matemático:

$$(13)\ L = \prod_{i=1}^{n} \quad [P_i^Y * (1-P)^{1-Y}]$$

Quando aplicado o estimador por verossimilhança na expressão geral do modelo de regressão logística obtém-se a seguinte expressão:

$$(14)\ L = \prod_{i=1}^{n} \quad [(\frac{1}{1+e^{-z}})^{y_i} * (\frac{1}{1+e^z})^{1-y_i}]$$

É importante ressaltar que na prática busca-se a maximização da somatória e não da produtória das observações por esta razão altera-se a equação para o logaritmo da função estimadora (log likelihood function) (FAVERO ET AL., 2017). O modelo matemático que representa do logaritmo do estimador por verossimilhança é:

$$(15)\ LL = \sum_{i=1}^{n} \quad = \{[(y_i) * ln(\frac{1}{1+e^{-z}})] + [(1- y_i) * ln(\frac{1}{1+e^z})]\} = max$$

Olmuş e colaboradores (2022) destacam que o modelo RL possui como vantagem o fato de não existir uma relação linear entre a variável dependente e as explicativas, este fato elimina a necessidade de aderência do erro à distribuição normal e o risco da presença de homocedasticidade, por outro lado, os fatores que afetam a performance deste modelo são: (i) o tamanho da amostra, que pode tornar a previsão enviesada, (ii) o excesso de "não eventos" na amostra, que pode tornar o modelo tendencioso e, (iii) a presença de multicolinearidade, que eleva variância e torna determinadas variáveis independentes estatisticamente insignificantes.

3 MÉTODO

Inicia-se o método pelo conjunto de dados.

3.1. Conjunto de dados

O conjunto de dados utilizado no desenvolvimento do modelo possui 13216 observações e 15 variáveis (Quadro 1)

Quadro 1 - Descrição do conjunto de dados

Variável	Tipo	Função no modelo	Detalhe
Fraude	Dicotômica	Variável dependente	5206 eventos e 8010 não eventos
Pedido	Identificadora	Nenhuma	Consiste no código de identificação do pedido
Qtd_itens_normal	Numérica contínua	Variável independente	Representa a quantidade de item do pedido
Valor_nfe	Numérica contínua	Variável independente	Representa o valor total do pedido
Meio_Captacao	Policotómica 8 níveis	Variável independente	Representa o método utilizado para captação do pedido e está dividido em 8 categorias
Data_pedido	Data	Nenhuma	Consistente na data de inserção do pedido no sistema (excluída do modelo)
Dia_semana	Policotómica 7 níveis	Variável independente	Representa o dia da semana em que o pedido foi inserido no sistema
Hora_pedido	Numérica contínua	Variável independente	Consiste na hora em que o pedido foi inserido no sistema
Forma_pagamento	Policotómica 3 níveis	Variável independente	Se refere a forma de pagamento escolhida pelo cliente e está dividida em 3 categorias
Plano_Pagamento	Policotómica 25 níveis	Variável independente	È uma subcategoria da forma de pagamento e está dividida em 25 categorias
Complemento_end1	Dicotômica	Variável independente	Consiste na presença ou ausência de informações complementares de endereço no cadastro do cliente
Macroregião	Policotómica 15 níveis	Variável independente	Trata-se da macroregião do estado em que o endereço de cliente está localizado
Cidade	Policotómica 543 níveis	Nenhuma	Cidade do endereço cliente (excluída do modelo)
Referência_endereço	Dicotômica	Variável independente	Consiste na presença ou ausência de informações de referência do endereço no cadastro do cliente
Primeiro_pedido	Dicotômica	Variável independente	Consiste na informação se esse pedido é ou não o primeiro pedido do cliente

Fonte: autores

Das 15 variáveis presentes no conjunto de dados, três não foram consideradas no modelo, sendo elas: a variável “Pedido” pois representa apenas o id de identificação do pedido, a variável “data_pedido” pois as observações do banco de dados foram registradas apenas no ano de 2023 impossibilitando a identificação de padrão relacionado a meses e a variável “cidade” pois sua grande quantidade de categoria impossibilitava a depuração do modelo.

Nos Quadros 2, 3 e 4 é possível identificar detalhes estatísticos sobre cada uma das variáveis aplicadas no modelo desenvolvido neste trabalho.

Quadro 2 - Medidas estatísticas das variáveis quantitativas

Variáveis quantitativas		
Medida estatística	Qtd_itens_normal	Hora_pedido
Min	1.00	0.0001389
1st Qu.	5.00	0.4582755
Median	7.00	0.6092535
Mean	14.91	0.5990546
3rd Qu	11.00	0.5990546
Max.	10652.00	0.9995139

Fonte: autores

Quadro 3 - Medidas estatísticas das variáveis politômicas

Variáveis Policotômicas		
Meio_captacao SGI API :8823 Portal RE - Stargate:1630 SGI Call Center :1602 SGI SV: 406 Pedido consolidado: 403 Loja Virtual: 344 (Other) : 8	Dia_semana domingo: 797 quarta-feira: 2356 quinta-feira :2212 sábado:1083 segunda-feira:2439 sexta-feira:1882 terça-feira:2447	Plano_pagamento Boleto - 2X - 30 Dias: 8474 Boleto - 2X - 21 Dias:1223 Boleto - 1X - 30 Dias:1187 Cartão Crédito - De 1x à 4x: 683 Boleto - 1X - 21 Dias: 437 Boleto à Vista: 252 (Other): 960
Macrorregião Metropolitana de São Paulo:7189 Ribeirão Preto:1155 Campinas:1008 Macro Metropolitana Paulista: 820 Vale do Paraíba Paulista: 548 São José do Rio Preto: 429 (Other):2067	Forma_pagamento Boleto: 11969 Múltiplas formas de Pagamento: 8 Outro: 1239	Cidade SAO PAULO: 3432 GUARUJA: 479 GUARULHOS: 470 GUARIBA: 276 PRAIA GRANDE: 257 CAMPINAS: 254 (Other): 8048

Fonte: autores

Quadro 4 - Medidas estatísticas das variáveis dicotômicas

Variáveis Dicotômicas				
Proposição	Fraude	Referência	Complemento_endereço	Primeiro_pedido
Verdadeiro	5206	6441	7808	6140
Falso	8010	6775	5408	7076

Fonte: autores

A organização de origem dos dados é uma líder em e-commerce no Brasil, e possui presença em mais de 50 países. Este grupo surgiu na década de 1970 e hoje possui 10 marcas diferentes, mais de 4 mil lojas físicas e mais 15 mil colaboradores.

Com relação aos dados obtidos é importante ressaltar que devido às restrições relacionadas a compliance e imposições legais oriundas da lei geral proteção de dados (LGPD), todos os dados pessoais e sensíveis relacionados aos clientes foram descaracterizados.

3.2. Ferramentas

A principal ferramenta de modelagem utilizada para o desenvolvimento do modelo foi o software R e seu ambiente de desenvolvimento Rstudio. No ambiente do Rstudio os pacotes utilizados foram: readxl, dplyr; jtools, fastDummies, PerformanceAnalytics, correlation, ggrepel, ggplotr, tidyverse, performance, car, effects, caret, caret, ggplot2, ROCR, plotly, pROC, cowplot. É importante ressaltar que houve uma pequena utilização do MS Excel para validar exportação do dataset.

3.3. Método de análise e validação

Para garantir que o modelo possua grau de eficiência em pelo menos uma variável independente estatisticamente significante foi aplicado o procedimento defendido por Fávero e companheiros (2009) de utilização do teste de distribuição Qui-quadrado de Pearson (2) no nível de significância de 5 % e confiabilidade de 95%. Plackett e companheiros (1983) definem a equação matemática do teste Qui-quadrado como:

$$(16)\ \chi^2 = \sum_{i=1}^{i} \quad \sum_{j=1}^{j} \quad = \frac{(o_{ij} - E_{ij})}{E_{ij}}$$

Já os parâmetros calculados pelo modelo foram validados pelo teste estatístico z de Wald, cuja equação matemática é definida como:

$$(17)\ z_a = \frac{a}{s.e.(a)} \qquad z_\beta = \frac{\beta}{s.e.(\beta)}$$

No cálculo 17 s.e. significa Standard error (erro -padrão).

Para avaliar a qualidade do ajuste do modelo e compará-lo com outros foi utilizado o Critério de informação Bayesiano (BIC) que foi proposto por Schwarz (1978), esse indicador é calculado por meio da seguinte equação.

$$(18)\ BIC = -2 * log\ f(x_n|\theta) + p * log\ n$$

No modelo definido acima $f(x_n|\theta)$ Representa a função de verossimilhança, o "p" é número de parâmetros e o "n" é o tamanho da amostra.

Para confirmar a eficiência do modelo foi utilizado uma matriz de confusão pela qual foi mensurados os seguintes indicadores:

$$(19)\ sensibilidade\ = \frac{verdadeiro\ positivo}{verdadeiro\ positivo + falso\ negativo}$$

$$(20)\ especificidade\ = \frac{verdadeiro\ negativo}{verdadeiro\ negativo + falso\ positivo}$$

$$(21)\ Acurácia\ = \frac{verdadeiro\ positivo + verdadeiro\ Negativo}{total\ de\ observações}$$

É importante ressaltar que o cutoff selecionado foi de 50%. Como indicador independente do cutoff foi utilizado o indicador de Área embaixo da curva ROC.

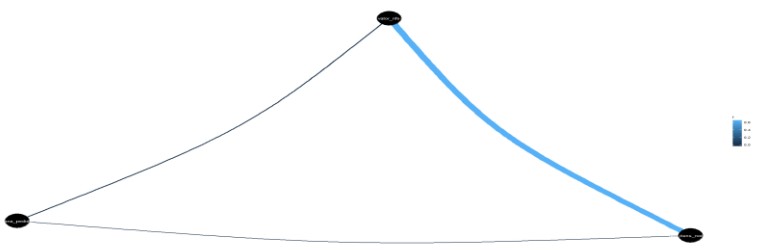
4 RESULTADOS E DISCUSSÃO

Inicia-se os resultados e discussão pela multicolinearidade.

4.1. Multicolinearidade

Como defendido no subtópico Regressão logística, para realização de um modelo RL deve-se primeiramente verificar a presença de multicolinearidade nos dados. Para realizar esta checagem utilizou-se o coeficiente de correlação de Pearson em que se observou a correlação entre as variáveis Hora_pedido e Valor_nfe (-0,039), correlação entre as variáveis Hora_pedido e Qtd_itens_normal (0,026) e correlação entre as variáveis Qtd_itens_normal e Valor_nfe (0,65).

Figura 1 - Correlação entre as variáveis quantitativas



Fonte: autores

4.2. Tratamento das variáveis politômicas , Stepwise e seleção do modelo

Após a checagem da correlação o dataset foi dividido em duas bases de dados, uma para treino e a outra para teste; esta separação foi realizada aleatoriamente com a proporção de 70% e 30%, respectivamente, mantendo essa proporção entre dados eventos (fraude) e não evento (não fraude).

Como resultado deste processo obteve-se a base de treino com 9251 observações, sendo 3644 eventos (fraude) e 5607 não evento (não fraude). Já por outro lado a base de teste ficou com 3965 observações, sendo 1562 evento (fraude) e 2403 não evento (não fraude).

Com o conjunto de dados dividido, criou-se o primeiro modelo que apresentou os seguintes resultados:

Modelo 01

a) Null deviance: 12404.9 on 9250 degrees of freedom;
b) Residual deviance: 4144.4 on 9193 degrees of freedom;
c) AIC: 4260.4;
d) log Lik.' -2072.192 (df=58);
e) BIC = 4674.068.

Ao identificar a grande diferença entre os valores do parâmetro “null deviance” e “residual deviance” foi possível estabelecer que pelo menos uma variável explicativa possui influência na variável resposta.

Utilizando o parâmetro do teste de distribuição Qui-quadrado de Pearson no nível de significância de 5% e confiabilidade de 95% e o teste estatístico z de Wald foi possível identificar que 16 categorias se mostraram estatisticamente significante, logo foi necessário aplicar o procedimento stepwise a fim de eliminar do modelo as variáveis não significantes.

Para a realização do procedimento stepwise em datasets que possuem variáveis politômicas faz-se necessário a realização de processo de tratamento das variáveis, este é popularmente conhecido como “dummização”, após a aplicação desta técnica a base de treino passou a ter 63 variáveis.

Após a realização do tratamento das variáveis politômicas aplicou-se o procedimento stepwise, criando assim um segundo modelo com 25 variáveis

O modelo 02 apresentou 20 categorias / variáveis com significância estatística a 95% de confiabilidade e 5 variáveis sem significância. É importante ressaltar que mesmo com 5 variáveis não significante esta é a configuração que maximiza a estimação por verossimilhança segundo o procedimento stepwise.

O segundo modelo apresentou os seguintes resultados:

Modelo 02:

a) Null deviance: 12404.9 on 9250 degrees of freedom;
b) Residual deviance: 4185.3 on 9226 degrees of freedom;
c) AIC: 4235.263;
d) log Lik.' -2092.631 (df=25);

e) BIC = 4413.575.

Comparando os dois modelos foi possível identificar que a nível estatístico de ambos são bem parecidos, esta semelhança se reflete nos resultados exposto na Quadro 5, contudo, utilizando o critério de informação Bayesiano (BIC) conforme defendido no tópico anterior, e considerando a capacidade de processamento de variáveis optou-se por seguir o experimento com o modelo 2 cujo BIC é menor e possui menos variáveis.

Quadro 5 - Comparação estatística entre os dois modelos

Parâmetros estatístico	Modelo 01	Modelo 02
Null deviance	12404.9	12404.9
Residual deviance	4144.4	4185.3
AIC	4260.4	4235.263
log Lik	-2072.192	-2092.631
BIC	4674.068	4413.575

Fonte: autores

Zellner e companheiro (2004) afirmam que em comparações de modelos de regressão logística aquele que possui menos variáveis pode ser preferível devido menor consumo de processamento. No trabalho apresentado por Fu e companheiros (2020) o procedimento stepwise também é aplicado no objetivo de evitar os erros causados pela multicolinearidade e aumentar o intervalo de confiança das variáveis estatisticamente significativas, como resultado, foi possível observar uma redução considerável das variáveis no dataset, assim como o resultado apresentado neste trabalho.

Outra obra referencial é a pesquisa desenvolvida por Zhang (2016) onde também são utilizados como critérios de seleção do melhor modelo o AIC e o BIC. Com base nestes conteúdos é possível afirmar que os critérios utilizados neste trabalho para seleção do melhor modelos é adequado e corresponde aos métodos mais atuais defendidos na literatura.

4.3. Performance do modelo

Com o modelo definido, realizou-se o processo de tratamento das variáveis politômicas e stepwise na base de teste a fim de se igualar a quantidade de variáveis entre as bases, em seguida aplicou-se o modelo selecionado a fim de verificar sua performance. Na matriz de confusão a seguir é possível verificar o resultado das previsões realizadas pelo modelo (Quadro 6).

Quadro 6 - Matriz de confusão modelo 02

Modelo 02		Referência	
		Não Fraude	Fraude
Previsão	Não Fraude	2098	36
	Fraude	305	1526

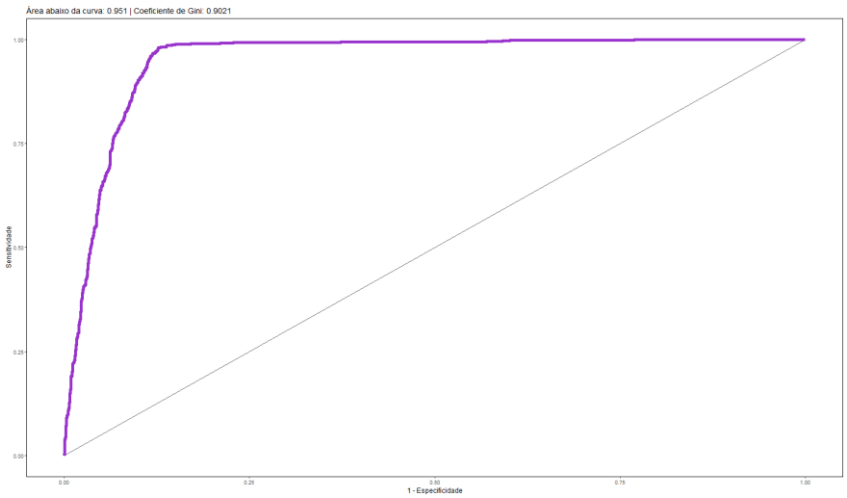
Fonte: autores

Para criação da matriz de confusão mostrada na Tabela 6 utilizou-se o cutoff de 0.5, conforme defendido no tópico “Material e método”, e por meio deste parâmetro obteve-se os seguintes resultados performáticos:

a) Accuracy : 0.914;
b) 95% CI : (0.9048, 0.9225);
c) No Information Rate : 0.6061 ;
d) P-Value [Acc > NIR] : < 2.2e-16;
e) Kappa : 0.825;
f) Mcnemar's Test P-Value : < 2.2e-16;
g) Specificity : 0.8731;
h) Sensitivity : 0.9770;
i) Pos Pred Value : 0.9831;
j) Neg Pred Value : 0.8334;
k) Prevalence : 0.6061;
l) Detection Rate : 0.5291;
m) Detection Prevalence : 0.5382;
n) Balanced Accuracy : 0.9250;
o) 'Positive' Class : FALSE ;
p) Área abaixo da curva ROC: 0.951;
q) Coeficiente de GINI: 0.9021.

É possível verificar graficamente a curva ROC do modelo na Figura 2.

Figura 2 - Curva ROC do modelo



Fonte: autores

Conforme mencionado na introdução deste estudo, a regressão logística é uma das técnicas de machine learning mais utilizadas e estudadas, dentro dessa realidade, Xiahou e colaboradores (2022) a aplicaram em um estudo de comparação entre a Regressão logística e outros métodos, o resultado encontrado foi uma acurácia máxima de 90,98% e a área abaixo da curva ROC de 96,3%. Em outro contexto, mais ainda no campo da pesquisa acadêmica, Huang e companheiros (2020) desenvolveram um modelo de previsão que quando combinado com técnicas de mineração de dados atinge uma acurácia máxima de 78,37%.

Já Gholamnia e companheiros (2020) aplicaram a regressão logística em uma realidade de previsão de incêndios, neste trabalho a única métrica utilizada para validação foi o critério de área abaixo da curva ROC cujo resultado foi de 68,9%. Todos os estudos citados nesse parágrafo possuem resultados próximos ou inferiores aos resultados apresentados nesta obra, contudo, é importante ressaltar que tais pesquisas são consideradas altamente relevantes no meio acadêmico, com base nisso pode-se afirmar que o resultado deste trabalho está dentro da expectativa de diferentes pesquisas acadêmicas podendo assim contribuir consideravelmente para o aprimoramento do tema de aprendizado de máquina.

5 CONSIDERAÇÕES FINAIS

O comércio eletrônico é um dos principais motores da economia mundial do século XXI. Sua rápida expansão nos últimos 25 anos reestruturou as relações entre cliente e fornecedor, além de mudar completamente o ambiente comercial global. Contudo, é importante ressaltar que este advento da modernidade não trouxe consigo apenas benesses, novos desafios de segurança e confiabilidade também surgiram. Diante desse cenário, diversos meios de prevenção de fraudes foram desenvolvidos, atualmente as técnicas mais avançadas são os modelos de aprendizado de máquina “machine learning”. Neste trabalho foi desenvolvido um modelo estatístico de machine learning cujo objetivo é identificar, por meio da técnica de Regressão logística, a probabilidade de um pedido ser fraude.

O dataset utilizado nesta pesquisa possui 13216 observações e 15 variáveis, contudo, apenas 12 foram consideradas no modelo, são elas 1 variável dependente dicotômica, 3 independentes quantitativas e 8 independentes qualitativas que juntas possuem 64 categorias.

Para identificar o nível de confiabilidade do modelo foi aplicado o teste de distribuição Qui-quadrado de Pearson no nível de significância de 5% e o teste estatístico z de Wald, como resultado obteve-se o número de 16 categorias estatisticamente significante, tornando assim necessário a realização do procedimento de tratamento de variáveis politômicas (dummização) e stepwise. Após a aplicação de tais procedimentos o modelo final passou a considerar como dados de entrada 25 categorias, atingindo assim um AIC de 4235.263, um Log lik de -2092.631 com 25 graus de liberdade e um BIC de 4413.575.

Com relação ao desempenho do modelo o resultado obtido foi de 91,4% de acurácia, 87,31% de especificidade e 97,7% de sensibilidade, tal performance resultou em uma área abaixo da curva ROC de 95,1 % e um Coeficiente de GINI de 90,21%. Com base nesses números é possível afirmar que o objetivo da pesquisa foi atingido, pois o modelo desenvolvido é altamente eficaz na prevenção de fraudes.

Com base nos parágrafos anteriores é possível declarar que este trabalho possui duas principais contribuições para a sociedade, a primeira consiste na promoção e divulgação da utilização de técnicas de machine learning para resolução de problemas reais e cotidianos no e-commerce, e a segunda consiste no esclarecimento e fomento do processo de desenvolvimento de modelos de regressão logística, bem como dos seus principais parâmetros de validação e indicadores de performance.

Como limitação desta pesquisa destacam-se 3 ocorrências, a primeira consiste na ausência de comparação deste modelo com outras técnicas de machine learning, tal realidade não contribui para o entendimento de técnicas mais modernas e não fomenta o debate acadêmico de comparação de eficiência entre diferentes ferramentas de previsão.

Já a segunda ocorrência é o fato de que o modelo tem como variável de entrada estatisticamente significante a macrorregião de residência do cliente e isto pode ser considerado uma forma de discriminação. Por último, evidencia-se como limitação deste trabalho o fato de o modelo não ter sido enviado para o ambiente de produção, tal conjuntura não permite a mensuração real da eficiência do modelo.

Para os próximos estudos considera-se realizar a comparação entre diferentes técnicas de previsão a fim de se identificar a mais precisa, além disso, pretende-se utilizar outro software estatísticos a fim de se obter um script mais sucinto e com mais interfaces gráficas.

No mais, os autores do trabalho destacam que embora a literatura acadêmica nesta área do conhecimento seja abundante, faltam exemplos de experimentos em ambientes corporativos e industriais de médio e pequeno porte, outra percepção que os autores destacam é a baixa quantidade de autores sul-americanos e africanos nos acervos acadêmicos, tais circunstâncias podem ser fatores desmotivadores para iniciantes no tema pois dependendo da origem do estudante o conteúdo pesquisado pode parecer muito distante do cotidiano dele.

Para finalizar os autores consideram que todo o conhecimento adquirido no decorrer da pesquisa é de extrema importância e contribuirá consideravelmente para o desenvolvimento profissional e acadêmico.

6 REFERÊNCIAS

BOCHIE, Kaylani et al. Aprendizado profundo em redes desafiadoras: **Conceitos e aplicações. Sociedade Brasileira de Computação**, 2020.

ClearSale. (2023). **Mapa da fraude 1º semestre de 2023**. disponivel em: https://br.clear.sale/hubfs/marketing/CRM/Mapa%20da%20Fraude%201%C2%BA%20semestre%202023/MapaDaFraude2023-1Sem_Ebook.pdf. Acesso em 8 de set. 2023

DE SOUZA, Daniel Henrique Miguel; BORDIN JR, Claudio J. Detecção de fraude de cartão de crédito por meio de algoritmos de aprendizado de máquina. , v. 15, n. 1, p. 1-11, 2023.

CORRAR, Luiz; PAULO, Edilson; DIAS FILHO, José Maria. Análise multivariada para os cursos de administração, ciências contábeis e economia. 2007.

ECKERT, Alex; MILAN, Gabriel Sperandio; TONI, Deonir de. Intenção de recompra no contexto de compras on-line. **Perspectivas em Ciência da Informação**, v. 24, p. 25-50, 2020.

FÁVERO, Luiz Paulo; BELFIORE, Patrícia. Manual de análise de dados: **estatística e modelagem multivariada com Excel®, SPSS® e Stata®**. Elsevier Brasil, 2017.

FÁVERO, Luiz Paulo Lopes et al. Análise de dados: modelagem multivariada para tomada de decisões. 2009.

FEITOSA, Douglas de Lima; GARCIA, Leandro Sumida. Sistemas de reputação: um estudo sobre confiança e reputação no comércio eletrônico brasileiro. **Revista de administração contemporânea**, v. 20, n. 1, p. 84-105, 2016.

FERNANDES, Antônio Alves Tôrres et al. Leia este artigo se você quiser aprender regressão logística. **Revista de Sociologia e Política**, v. 28, p. 006, 2021

FU, Liping; WANG, Yuhui; HE, Lanping. Factors associated with the psychological health of caregiving older parents and support from their grown children: results from the China health and retirement longitudinal study. **International Journal of Environmental Research and Public Health**, v. 17, n. 2, p. 556, 2020.

GHOLAMNIA, Khalil et al. Comparisons of diverse machine learning approaches for wildfire susceptibility mapping. **Symmetry**, v. 12, n. 4, p. 604, 2020.

GUJARATI, Damodar N.; PORTER, Dawn C. **Basic econometrics**. McGraw-hill, 2009.

HOSMER, David W.; LEMESBOW, Stanley. Goodness of fit tests for the multiple logistic regression model. **Communications in statistics-Theory and Methods**, v. 9, n. 10, p. 1043-1069, 1980.

HUANG, Jia-Yen; LIU, Jin-Hao. Using social media mining technology to improve stock price forecast accuracy. **Journal of Forecasting**, v. 39, n. 1, p. 104-116, 2020.

IMDADULLAH, Muhammad; ASLAM, Muhammad; ALTAF, Saima. mctest: An R Package for Detection of Collinearity among Regressors. **R J**., v. 8, n. 2, p. 495, 2016.

KLEINBAUM, David G. et al. **Logistic regression**. New York: Springer-Verlag, 2008.

MERGHADI, Abdelaziz et al. Machine learning methods for landslide susceptibility studies: A comparative overview of algorithm performance. **Earth-Science Reviews**, v. 207, p. 103225, 2020.

MENDONÇA, Júlio César Gomes et al. Transação com partes relacionada como instrumento de fraudes corporativas em bancos brasileiros. **Revista Contabilidade Vista & Revista**, v. 32, n. 3, p. 195-216, 2021.

MOREIRA, Robson Antonio. O Comércio Eletrônico, os métodos de pagamentos e os mecanismos de segurança. **Refas-Revista Fatec Zona Sul**, v. 3, n. 1, p. 16-30, 2016.

MORTEZA; ARIAS-ARANDA, Daniel; BENITEZ-AMADO, Jose. Adoption of e-commerce applications in SMEs. **Industrial Management & Data Systems**, v. 111, n. 8, p. 1238-1269, 2011.

MUKHOTY, Bhaskar; DEY, Debojyoti; KAR, Purushottam. Corruption-tolerant algorithms for generalized linear models. In: **Proceedings of the AAAI Conference on Artificial Intelligence**. 2023. p. 9243-9250.

NELDER, John A; LEE, Youngjo. Hierarchical generalized linear models. **Journal of the Royal Statistical Society Series B: Statistical Methodology**, v. 58, n. 4, p. 619-656, 1972.

NielsenIQ. (2022). **A evolução do e-commerce a nível mundial**. disponivel em: https://nielseniq.com/global/pt/insights/analysis/2022/a-evolucao-do-e-commerce-a-nivel-mundial/. acesso em 8 de set. de 2023

Novikova, Olha, and Kuan Zhang. "Analyses of the E-Commerce Development in the World and China." **Modern Economics**. 2020.

OLMUŞ, Hülya; NAZMAN, Ezgi; ERBAŞ, Semra. Comparison of penalized logistic regression models for rare event case. **Communications in Statistics-Simulation and Computation**, v. 51, n. 4, p. 1578-1590, 2022.

ORDANINI, Andrea; RUBERA, Gaia. How does the application of an IT service innovation affect firm performance? A theoretical framework and empirical analysis on e-commerce. **Information & Management**, v. 47, n. 1, p. 60-67, 2010.

PLACKETT, Robin L. Karl Pearson and the chi-squared test. I**nternational statistical review/revue internationale de statistique**, p. 59-72, 1983.

SCHWARZ, Gideon. Estimating the dimension of a model. **The annals of statistics**, p. 461-464, 1978.

VASELI, Saeed et al. DISCOVERING CORPORATE FRAUD AND ACCOUNTANT FAILURE: CAUSES AND SOLUTIONS. **Lex Humana (ISSN 2175-0947)**, v. 13, n. 2, p. 190-214, 2021.

XIAHOU, Xiancheng; HARADA, Yoshio. B2C E-commerce customer churn prediction based on K-means and SVM. **Journal of Theoretical and Applied Electronic Commerce Research,** v. 17, n. 2, p. 458-475, 2022.

YU, Ying et al. E-commerce logistics in supply chain management: Practice perspective. **Procedia Cirp**, v. 52, p. 179-185, 2016.

ZABOR, Emily C. et al. Logistic regression in clinical studies. **International Journal of Radiation Oncology* Biology* Physics**, v. 112, n. 2, p. 271-277, 2022.

Zellner, D., Keller, F., & Zellner, G. E. (2004). Variable selection in logistic regression models. **Communications in Statistics-Simulation and Computation**, 33(3), 787-805.

ZHANG, Xinwei et al. HOBA: A novel feature engineering methodology for credit card fraud detection with a deep learning architecture. **Information Sciences**, v. 557, p. 302-316, 2021.

ZHANG, Xiaoli. AHI: growth and improvement of cross-border e-commerce using IVAS. **International Journal of Cooperative Information Systems**, v. 32, n. 03, p. 2150011, 2023.

ZHANG, Zhongheng. Variable selection with stepwise and best subset approaches. **Annals of translational medicine**, v. 4, n. 7, 2016.