UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Câmpus de Presidente Prudente

THAMIRES DOS SANTOS MOTA

APLICAÇÃO DO MODELO DE REGRESSÃO LOGÍSTICA EM SITUAÇÕES DE DADOS COM DESBALANCEAMENTO SEVERO

PRESIDENTE PRUDENTE
2025

THAMIRES DOS SANTOS MOTA

APLICAÇÃO DO MODELO DE REGRESSÃO LOGÍSTICA EM SITUAÇÕES DE DADOS COM DESBALANCEAMENTO SEVERO

Relatório Final de Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Estatística da FCT/UNESP para aproveitamento na disciplina Trabalho de Conclusão de Curso.
Orientador: Prof. Dr. Sérgio Minoru Oikawa
Coorientador: Prof. Dr. Mário Hissamitsu Tarumoto

PRESIDENTE PRUDENTE
2025

M917a

Mota, Thamires dos Santos

Aplicação do modelo de Regressão Logística em situações de dados com desbalanceamento severo / Thamires dos Santos Mota. -- Presidente Prudente, 2025

55 p. : il., tabs.

Trabalho de conclusão de curso (Bacharelado - Estatística) - Universidade Estadual Paulista (UNESP), Faculdade de Ciências e Tecnologia, Presidente Prudente

Orientador: Sérgio Minoru Oikawa

Coorientador: Mário Hissamitsu Tarumoto

1. Regressão Logística. 2. Dados Desbalanceados. 3. Detecção de Fraudes. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Dados fornecidos pelo autor(a).

TERMO DE APROVAÇÃO

THAMIRES DOS SANTOS MOTA

APLICAÇÃO DO MODELO DE REGRESSÃO LOGÍSTICA EM SITUAÇÕES DE DADOS COM DESBALANCEAMENTO SEVERO

Relatório de Final de Trabalho de Conclusão de Curso aprovado como requisito para obtenção de créditos na disciplina Trabalho de Conclusão do curso de graduação em Estatística da Faculdade de Ciências e Tecnologia da Unesp, pela seguinte banca examinadora:

gov.br Documento assinado digitalmente
SERGIO MINORU OIKAWA
Data: 12/12/2025 12:42:57-0300
Verifique em https://validar.iti.gov.br

Orientador: _____
Prof. Dr. Sérgio Minoru Oikawa
Departamento de Estatística

gov.br Documento assinado digitalmente
MARIO HISSAMITSU TARUMOTO
Data: 12/12/2025 14:14:23-0300
Verifique em https://validar.iti.gov.br

Coorientador: _____
Prof. Dr. Mário Hissamitsu Tarumoto
Departamento de Estatística

gov.br Documento assinado digitalmente
EDILSON FERREIRA FLORES
Data: 12/12/2025 08:25:46-0300
Verifique em https://validar.iti.gov.br

_____
Prof. Dr. Edilson Ferreira Flores
Departamento de Estatística

Presidente Prudente, 08 de dezembro de 2025.

AGRADECIMENTOS

Agradeço primeiramente a Deus por ter me sustentado em cada etapa desta caminhada, e me permitindo chegar até esta conquista tão importante.

Aos meus pais, Reginaldo e Gislaine, deixo minha gratidão mais profunda. Obrigada por cada gesto de amor, por todas as oportunidades que me proporcionaram e por serem minha base e meu maior incentivo desde sempre.

Ao meu irmão, Danilo, agradeço pela parceria, pelas conversas que me acalmaram e por estar presente em cada etapa da minha jornada, dividindo comigo momentos de preocupação e de alegria.

Ao meu namorado, Luca, sou imensamente grata pelo carinho, pela paciência e pelo incentivo constante. Obrigada por compreender as minhas ausências, por vibrar com cada pequena vitória e por estar ao meu lado em todos os momentos desta jornada.

Aos meus amigos, que tornaram os dias mais leves com palavras de apoio, risadas e companheirismo sincero. Obrigada por acreditarem em mim e por celebrarem cada etapa comigo.

Ao meu orientador, Prof. Dr. Mário, expresso minha sincera gratidão pela orientação dedicada, pelos conselhos, pela disponibilidade e pelos ensinamentos que contribuíram não só para este trabalho, mas também para minha formação profissional.

Estendo meus agradecimentos a todos os professores do curso de Estatística da UNESP de Presidente Prudente, que foram fundamentais na construção do conhecimento que levo comigo.

A todos que, direta ou indiretamente, fizeram parte desta trajetória, deixo meu muito obrigada. Este trabalho é também resultado do apoio e carinho que recebi ao longo do caminho.

RESUMO

Este trabalho tem como objetivo investigar a importância do uso de técnicas de balanceamento em modelos de regressão logística aplicados à detecção de fraudes em bases de dados desbalanceadas. Esta forma de balanceamento dos dados será comparada à proposta apresentada por Assunção, Izbicki e Prates (2024). O estudo utiliza duas bases de dados com diferentes proporções de casos de fraude, buscando analisar como o desequilíbrio afeta a performance dos modelos. O modelo de regressão logística foi escolhido por sua aplicação em problemas de classificação binária e principalmente pela possibilidade de interpretação dos parâmetros do modelo. Serão aplicadas diversas métricas de avaliação para analisar a capacidade dos modelos em identificar corretamente as fraudes. Este trabalho pretende contribuir para a compreensão do impacto das técnicas de balanceamento e avaliação da melhor técnica a ser utilizada na melhoria da detecção de fraudes, destacando a importância da escolha adequada das métricas de desempenho em cenários com dados desbalanceados.

Palavras-chave: Regressão logística, desequilíbrio de classes, detecção de fraudes, balanceamento de dados.

ABSTRACT

This work aims to investigate the importance of applying data balancing techniques to logistic regression models used for fraud detection in imbalanced datasets. These balancing approaches are compared with the strategy proposed by Assunção, Izbicki, and Prates (2024). The study employs two datasets with different proportions of fraudulent cases, in order to analyze how class imbalance affects model performance. Logistic regression was chosen due to its suitability for binary classification problems and, in particular, for the interpretability of its parameters. Several evaluation metrics are applied to assess the ability of the models to correctly identify fraudulent transactions. This work seeks to contribute to a better understanding of the impact of balancing techniques and to evaluate the most effective strategy for improving fraud detection, highlighting the importance of choosing appropriate performance metrics in scenarios with imbalanced data.

Keywords: Logistic regression, class imbalance, fraud detection, data balancing.

SUMÁRIO

1 INTRODUÇÃO 9

2 REFERENCIAL TEÓRICO 11

2.1 Regressão Linear Simples e Múltipla 11

2.2 Regressão Logística 13

2.3 Medidas de performance 15

2.4 Técnicas de Amostragem 18

2.5 Técnicas alternativas em situações de dados desbalanceados. 19

3 Conjunto de dados 20

3.1 Metodologia 20

3.2 Descrição dos dados 22

3.2.1 Descrição do primeiro conjunto de dados 22

3.2.2 Descrição do segundo conjunto de dados 23

3.3 Análise exploratória do primeiro conjunto de dado 23

3.4 Análise exploratória do segundo conjunto de dados 27

4 RESULTADOS E DISCUSSÃO 31

4.1 Resultados do primeiro conjunto de dados 31

4.2 Resultados do segundo conjunto de dados 42

4 CONCLUSÕES 51

REFERÊNCIAS 54

1 INTRODUÇÃO

No contexto atual de crescente digitalização das transações financeiras, a detecção de fraudes tornou-se um dos principais desafios enfrentados por empresas, instituições financeiras e consumidores. Com o aumento exponencial do volume de transações eletrônicas, seja por meio de plataformas de e-commerce, bancos digitais ou sistemas de pagamento online, a quantidade de atividades fraudulentas também aumentou significativamente, o que acarreta grandes prejuízos econômicos tanto para as empresas quanto para os indivíduos (SERASA EXPERIAN, 2024). Muitas pessoas são vítimas de fraudes financeiras diariamente, o que pode resultar em perdas financeiras, roubo de identidade e danos à confiança nas plataformas digitais (IBM, 2023). Apesar do crescente número de tentativas de fraudes, a proporção de fraudes em comparação às transações legítimas continua sendo muito menor, o que resulta em dados desbalanceados. De acordo com alguns autores e publicações em sites de internet, esse desbalanceamento é um grande desafio para os modelos estatísticos e de aprendizado de máquina, pois a quantidade limitada de informações relacionadas às fraudes dificulta a capacidade desses modelos de fazer previsões precisas.

A regressão logística é uma técnica estatística de classificação amplamente utilizada em diversos campos, inclusive na detecção de fraudes financeiras, devido à sua capacidade de prever a probabilidade de ocorrência de um evento binário, como fraude (evento positivo) ou transação legítima (evento negativo). No entanto, um dos maiores desafios na aplicação dessa técnica reside no desequilíbrio das classes. Essa disparidade pode resultar em modelos enviesados que priorizam a classe majoritária, o que, por sua vez, leva a uma alta taxa de falsos negativos, ou seja, fraudes que passam despercebidas pelo modelo.

Para contornar esse problema, uma das abordagens propostas é reorganizar a base de dados, buscando equilibrar as proporções entre fraudes e não fraudes. Existem duas estratégias principais para isso: o *oversampling* e o *undersampling*. O *oversampling* visa aumentar a proporção de fraudes por meio da replicação dos dados da classe minoritária, ou seja, criando amostras de fraudes para tentar balancear o conjunto de dados. Por outro lado, o *undersampling* foca em reduzir a classe majoritária, retirando uma amostra aleatória das transações legítimas, de modo que o número de observações fique semelhante entre as duas classes. Apesar destas indicações, Assunção *et al.* (2024), discutem a não necessidade de realização destas técnicas no contexto de modelos de aprendizado de máquina (*machine learning*), o artigo sugere que apenas realizando o ajuste do limiar de decisão (*cutoff*) do modelo, conseguem obter resultados

semelhantes ou até melhores do que os obtidos através de estratégias tradicionais de aumentação, como o *oversampling*.

Diante disso, este trabalho tem como objetivo principal estudar e avaliar o desempenho da regressão logística na detecção de fraudes financeiras em um cenário de dados desbalanceados, investigando se a aplicação de técnicas de balanceamento pode melhorar sua capacidade preditiva, aplicar técnicas de *oversampling*, *undersampling* e ajuste do limiar de decisão, para ajustar a base de dados, comparar os resultados obtidos antes e depois dessas técnicas e avaliar métricas como sensibilidade, precisão e taxa de falsos negativos.

Assim, parte-se da hipótese de que a aplicação dessas técnicas pode melhorar o desempenho da regressão logística em bases desbalanceadas, conforme sugerido por Chawla et al. (2002).

Após a reorganização dos dados, o modelo de Regressão Logística será ajustado e analisado em termos de sua capacidade de prever fraudes. O objetivo é verificar se, com o tratamento adequado dos dados, o modelo consegue melhorar sua sensibilidade na detecção de fraudes sem comprometer a acurácia geral.

A relevância deste estudo se justifica pela crescente preocupação com a segurança das transações financeiras no Brasil. De acordo com o "Relatório de Identidade Digital e Fraude 2024" da Serasa Experian, quatro em cada dez brasileiros já foram vítimas de golpes, sendo que a perda média estimada por fraude ultrapassa dois mil reais. Além disso, o número de empresas preocupadas com a recorrência de fraudes aumentou significativamente nos últimos anos. Esses dados ressaltam a importância de se desenvolver soluções eficientes de detecção de fraudes, tanto para reduzir prejuízos financeiros quanto para fortalecer a confiança nas plataformas digitais.

Essa pesquisa também contribui com iniciativas globais voltadas para o desenvolvimento sustentável. A melhoria da segurança digital está diretamente alinhada aos Objetivos de Desenvolvimento Sustentável da ONU, como o ODS 8, que promove crescimento econômico e trabalho decente, o ODS 9, voltado à inovação e infraestrutura digital resiliente, e o ODS 16, que incentiva instituições eficazes e justas. Dessa forma, esta pesquisa visa apoiar o desenvolvimento de uma sociedade mais segura, justa e economicamente estável diante dos desafios impostos pela era digital.

Na parte introdutória deste Trabalho de Conclusão de Curso, foi apresentada as dificuldades de estimação da probabilidade de fraude e a justificativa do trabalho. Na seção 2, foram realizados estudos teóricos dos modelos utilizados neste trabalho, como o modelo de regressão linear, regressão logística e os principais métodos de tratamento de dados

desbalanceados. Em seguida, na seção 3, são descritos os dados utilizados e os testes estatísticos aplicados para a análise exploratória, como o teste t para variáveis contínuas e o teste qui-quadrado para variáveis categóricas, com o objetivo de investigar possíveis diferenças entre transações fraudulentas e não fraudulentas.

Na seção 4, são apresentados os resultados e as discussões referentes aos modelos ajustados, contemplando a aplicação das diferentes técnicas de tratamento de desbalanceamento e a comparação de desempenho entre os modelos treinados e avaliados. Por fim, a seção final reúne as conclusões do trabalho, destacando as principais contribuições, limitações e sugestões para estudos futuros.

2 REFERENCIAL TEÓRICO

2.1 Regressão Linear Simples e Múltipla

A regressão linear é uma das técnicas estatísticas mais antigas e amplamente utilizadas para modelar relações entre variáveis. Seu principal objetivo é descrever e quantificar a relação entre uma variável dependente (Y) e uma ou mais variáveis independentes $(X_1, X_2, \dots, X_p)$, por meio de uma equação de natureza linear. Essa técnica permite a realização de previsões e inferências estatísticas.

A origem da regressão linear remonta ao século XIX, quando Francis Galton observou que a altura dos filhos tendia a se aproximar da média entre a altura dos pais e da população. Esse fenômeno foi chamado por ele de “regressão para a média”, originando o termo "regressão" (GALTON, 1886). Posteriormente, pesquisadores como Karl Pearson e Ronald Fisher contribuíram significativamente para a formalização da técnica e sua aplicação em contextos estatísticos mais amplos (MONTGOMERY; PECK; VINING, 2012).

No modelo de regressão linear simples se estuda a relação entre uma variável resposta (Y) e uma única variável explicativa (X). A forma geral desse modelo é expressa pela equação:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, i = 1, 2, \dots, n$$

Nessa equação, β_0 representa o intercepto, ou seja, o valor esperado de Y quando $X = 0$, β_1 é o coeficiente angular da reta, que indica a taxa de variação de Y em relação a X e por fim ε_i que representa o erro aleatório, associado à i-ésima observação que é assumido como tendo média zero e variância constante σ^2. Os coeficientes β_0 e β_1 são estimados com base nos

dados observados por meio do método dos mínimos quadrados ordinários, que consiste em minimizar a soma dos quadrados dos resíduos:

$$min_{\beta_0,\beta_1} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

sendo que os valores ajustados $\hat{Y}_i$ são obtidos por:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

Para que os resultados obtidos por meio da regressão linear sejam válidos, é necessário que algumas suposições sejam atendidas. As principais são: a relação entre as variáveis deve ser linear, os erros devem ser independentes, os erros devem ter variância constante (homoscedasticidade) e por fim os erros devem seguir uma distribuição normal.

Quando há mais de uma variável explicativa, utiliza-se o modelo de regressão linear múltipla, cuja equação é generalizada para:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + \varepsilon_i, i = 1,2,\ldots,n$$

A regressão linear múltipla é amplamente utilizada na seleção de covariáveis, isto é, na identificação das variáveis explicativas que mais contribuem para explicar a variação da variável dependente Y. Entre os principais métodos de seleção de variáveis, destacam-se o *Forward Selection*, o *Backward Elimination* e o *Stepwise Selection*. O método *Forward* inicia com o modelo nulo (que contém apenas o intercepto) e, a cada etapa, adiciona ao modelo a variável que mais melhora o ajuste, com base em critérios como menor valor de p associado ao teste t, maior aumento no $R^2{}_{ajustado}$, ou maior redução no AIC ou BIC.

O método Backward parte do modelo completo, com todas as variáveis, e remove iterativamente aquelas que menos contribuem, seguindo critérios semelhantes. O método Stepwise combina os dois anteriores: a cada iteração, além de adicionar uma nova variável, também verifica se alguma das já incluídas deve ser retirada, refinando o modelo passo a passo.

Apesar de sua ampla utilização, a regressão linear, seja na forma simples ou múltipla, não é adequada em situações em que a variável dependente é categórica, especialmente binária (como no caso de classificar uma transação como fraudulenta ou não). Isso ocorre porque esse tipo de modelo pressupõe que a variável resposta seja contínua e normalmente distribuída, o que não se aplica a dados que assumem apenas valores 0 e 1. Além disso os pressupostos de homoscedasticidade e normalidade dos resíduos são violados quando se tenta ajustar uma regressão linear a variáveis binárias, resultando em estimativas viesadas e testes não confiáveis.

Diante dessas limitações, torna-se necessário recorrer a outros tipos de regressão, que variam conforme a natureza da variável dependente. Entre os principais modelos, destacam-se

a regressão linear simples e múltipla, aplicadas quando a variável resposta é contínua; a regressão logística, adequada para variáveis categóricas binárias, pois modela a probabilidade de ocorrência de um evento; a regressão polinomial, utilizada quando há indícios de uma relação não linear entre as variáveis; a regressão de Poisson, voltada para dados de contagem; a regressão log-linear, utilizada em tabelas de contingência; e a regressão de Cox, aplicada em análises de sobrevivência, quando o interesse é modelar o tempo até a ocorrência de determinado evento.

No contexto deste trabalho, a regressão logística apresenta-se como a técnica mais adequada, uma vez que o objetivo é classificar as transações como fraudulentas ou legítimas, caracterizando um problema de natureza binária. Esse modelo é mais apropriado pois, diferentemente da regressão linear, modela diretamente a probabilidade de ocorrência de um evento dentro do intervalo [0,1], permitindo interpretações probabilísticas consistentes. Dessa forma, a regressão logística representa uma extensão natural da regressão linear para situações em que a variável resposta assume apenas dois valores possíveis, sendo amplamente aplicada em estudos de prevenção e detecção de fraudes financeiras.

2.2 Regressão Logística

A regressão logística tem suas raízes no século XIX, quando Pierre François Verhulst introduziu a função logística para modelar o crescimento populacional. No entanto, sua aplicação estatística em modelagem de variáveis binárias ganhou destaque ao longo do século XX, especialmente nas áreas de biostatística e epidemiologia.

A regressão logística, como conhecemos hoje, foi significativamente desenvolvida por Joseph Berkson na década de 1940, que criou o termo "logit" em 1944.

Hosmer e Lemeshow, em seu livro Applied Logistic Regression, destacam que o desenvolvimento e a aplicação da regressão logística foram impulsionados pela necessidade de modelar relações entre variáveis categóricas e pela evolução dos softwares estatísticos, que tornaram a implementação do modelo mais acessível.

A regressão logística é um método estatístico utilizado para resolver problemas de classificação binária (0 ou 1), em que o objetivo é prever a probabilidade de um determinado evento ocorrer (classe 1) ou não (classe 0), baseado em um conjunto de variáveis explicativas. Esse modelo usa a função logística (ou sigmoid) para transformar uma combinação linear de

variáveis preditivas em uma probabilidade limitada entre 0 e 1, facilitando a tomada de decisão entre duas classes (Hosmer; Lemeshow, 2000).

A função logística é dada por:

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

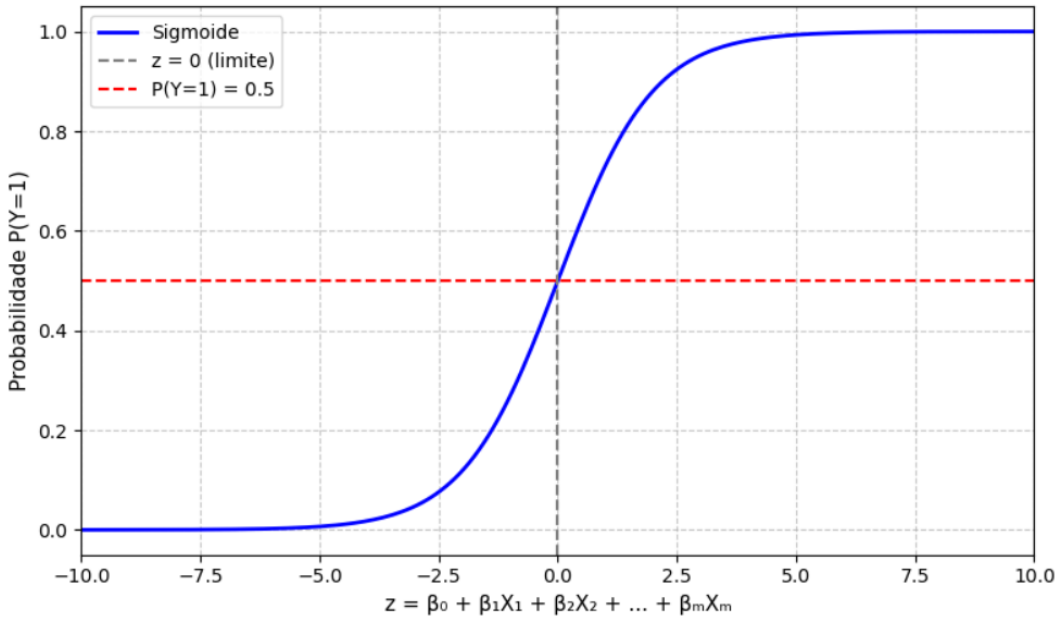
- Onde z representa uma combinação ponderada das variáveis independentes do modelo, ou seja, $z = \beta_0 + \beta_1 X_1 + \cdots + \beta_m X_m$.
- β_0 é o intercepto do modelo.
- $\beta_1, \beta_2, \ldots, \beta_m$ são os coeficientes das variáveis independentes $X_1, X_2, \ldots, X_m$.
- e é a constante de Euler, base do logaritmo natural.

O valor de $\sigma(z)$ é a probabilidade estimada de o evento de interesse ocorrer. Ou seja, o modelo prediz:

- $P(Y = 1) = \sigma(z)$
- $P(Y = 0) = 1 - \sigma(z)$

Essa função garante que a probabilidade seja limitada ao intervalo $[0, 1]$, e é usada para modelar a relação entre as variáveis explicativas e a variável dependente binária.

Figura 1 – Curva Sigmoide gerada com z variando de -10 a 10



Fonte: Elaborado pelo autor (2025).

A Figura 1 ilustra a curva sigmoide, também conhecida como função logística, utilizada no modelo de regressão logística. Essa função é responsável por transformar a combinação

linear das variáveis independentes em valores de probabilidade que variam entre 0 e 1. Observa-se que a curva possui formato em "S", característico da função logística. Quando z assume valores muito negativos, a probabilidade de ocorrência do evento positivo tende a zero; à medida que z aumenta, a probabilidade cresce de forma não linear, aproximando-se de 1. O ponto central, onde $z = 0$, corresponde a uma probabilidade de 0.5, representando o limite de decisão do modelo. Esse comportamento permite interpretar os coeficientes da regressão logística em termos de variação na probabilidade de ocorrência de um evento, tornando a técnica especialmente adequada para problemas de classificação binária.

No cenário de fraudes financeiras, a regressão logística é amplamente utilizada. No entanto, esses modelos enfrentam um grande desafio quando os dados estão desbalanceados, ou seja, quando a quantidade de transações legítimas é muito superior ao número de fraudes. Isso dificulta a identificação de fraudes, já que o modelo pode dar maior peso à classe majoritária (transações legítimas), reduzindo sua eficácia. Esse desbalanceamento pode ser amenizado por meio de técnicas de reamostragem; entre essas técnicas, destacam-se a amostragem por conglomerado e a amostragem estratificada.

2.3 Medidas de performance

A avaliação da performance de um modelo de regressão logística é uma etapa essencial, especialmente quando lidamos com bases de dados desbalanceadas, como é o caso de problemas de detecção de fraudes. Nesses contextos, utilizar apenas a acurácia pode gerar conclusões equivocadas, pois um modelo que apenas prevê a classe majoritária pode apresentar uma acurácia alta, mas falhar completamente em identificar os casos de fraude (classe minoritária). Por isso, é importante considerar diferentes métricas que avaliem não só o número total de acertos, mas também a capacidade do modelo de identificar corretamente os casos positivos e negativos. Essas métricas são baseadas na matriz de confusão, que é utilizada para organizar os resultados da classificação em quatro categorias.

Tabela 1 – Matriz de confusão.

		Classe Real	
		Transação Legitima	Fraude
Classe Prevista	Transação Legitima	**Verdadeiro Negativo (VN)**	**Falso Negativo (FN)**
	Fraude	**Falso Positivo (FP)**	**Verdadeiro Positivo (VP)**

Fonte: Elaborado pelo autor (2025).

A Tabela 1 representa a matriz de confusão, ou seja, quando o modelo previu corretamente uma fraude temos o verdadeiro positivo (VP), falso positivo (FP) quando o modelo indicou uma fraude quando na verdade era uma transação legítima, falso negativo (FN) o modelo deixou de identificar uma fraude, o que é comum em dados desbalanceados, e por fim o verdadeiro negativo (VN) o modelo previu corretamente uma transação legítima. Com base nesses elementos, são definidas as principais métricas de avaliação.

A acurácia representa a proporção total de previsões corretas em relação ao total de observações.

$$Acurácia = \frac{VP + VN}{VP + VN + FP + FN}$$

Apesar de ser uma métrica bastante utilizada, sua interpretação em bases desbalanceadas deve ser feita com cautela. Um valor alto de acurácia pode simplesmente refletir o acerto na classe majoritária, sem que o modelo esteja aprendendo a identificar a classe minoritária.

A precisão representa a proporção de casos que foram classificados como positivos (fraudes) e que realmente pertenciam à classe positiva.

$$Precisão = \frac{VP}{VP + FP}$$

Essa métrica é especialmente útil quando o custo de um falso positivo é elevado. Por exemplo, bloquear indevidamente uma transação legítima pode causar incômodos ao cliente e prejuízos à instituição. Por isso, uma boa precisão indica que o modelo erra pouco ao classificar algo como fraude. Em termos práticos, valores de precisão acima de 70% já são considerados aceitáveis, e valores acima de 80% são bastante satisfatórios. No entanto, como a precisão só avalia os acertos dentro das previsões positivas, ela não informa quantas fraudes o modelo deixou de identificar.

Por isso, é necessário também considerar o Recall ou Sensibilidade, que mede a capacidade do modelo de detectar todos os casos reais da classe positiva.

$$Recall = \frac{VP}{VP + FN}$$

Neste caso, estamos interessados em saber qual proporção das fraudes reais foi corretamente identificada pelo modelo. Um Recall alto significa que o modelo consegue "enxergar" a maior parte dos casos de interesse. Essa métrica é crucial em problemas de fraude, pois deixar de identificar um caso (erro tipo FN) pode ser mais grave do que um alarme falso (erro tipo FP).

A especificidade indica a proporção de casos negativos que foram corretamente classificados como tal, ou seja, quantas transações legítimas foram corretamente reconhecidas pelo modelo como não fraudulentas.

$$\text{Especificidade} = \frac{\text{VN}}{\text{VN} + \text{FP}}$$

É comum utilizar uma métrica que combine a Precisão com a Sensibilidade, o F1-score. Essa métrica é a média harmônica entre precisão e recall e busca um equilíbrio entre os dois aspectos.

$$F1 - score = \frac{Precisão \cdot Recall}{Precisão + Recall}$$

O F1-score é especialmente útil em bases desbalanceadas, pois penaliza casos em que o modelo tem uma métrica muito boa e outra muito baixa. Um bom F1-score indica que o modelo consegue detectar fraudes com razoável acerto e, ao mesmo tempo, sem gerar muitos falsos negativos.

Outra forma de avaliar o desempenho global do modelo é por meio da curva ROC (Receiver Operating Characteristic) e da AUC (Área sob a Curva ROC). A curva ROC representa a relação entre a taxa de verdadeiros positivos (Recall) e a taxa de falsos positivos, para diferentes limiares de classificação. A área sob essa curva, chamada de AUC-ROC, varia de 0 a 1 e indica a capacidade do modelo de diferenciar corretamente entre as classes positiva e negativa. Um modelo com AUC = 0,5 possui desempenho aleatório, enquanto AUC = 1 representa separação perfeita entre as classes.

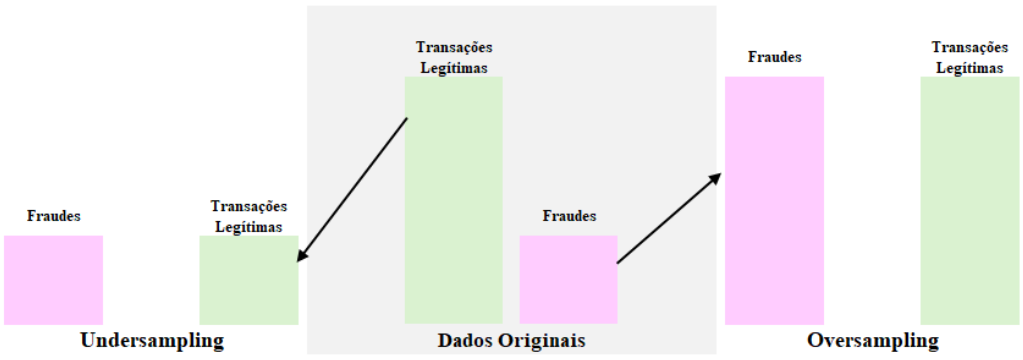
2.4 Técnicas de Amostragem

Em problemas de classificação com dados desbalanceados, como na detecção de fraudes financeiras, é comum que existem muitas observações da classe majoritária (transações legítimas) e poucas da classe minoritária (fraudes). Esse desequilíbrio faz com que modelos estatísticos aprendam mais sobre a classe majoritária e acabem cometendo muitos falsos negativos, ou seja, não identificando fraudes. Para lidar com esse problema, existem estratégias como *oversampling*, *undersampling* e, mais recentemente, o ajuste do *cutoff*, proposto como alternativa por Assunção, Izbicki e Prates (2024).

O *oversampling* é uma técnica que aumenta a quantidade de exemplos da classe minoritária (fraudes), para que o modelo possa aprender melhor seus padrões. A forma mais simples é o *oversampling* aleatório com reposição, ou seja, um mesmo exemplo pode ser sorteado mais de uma vez, dessa forma amostras da classe minoritária (fraudes) são replicadas até que haja um número semelhante ao da classe majoritária.

O *undersampling* elimina parte da classe majoritária (transações legítimas), até equilibrar com a classe minoritária, isso é feito por amostragem aleatória sem reposição, ou seja, cada exemplo da classe majoritária é sorteado apenas uma vez e não pode se repetir.

Figura 2 – Ilustração dos métodos de *undersampling* e *oversampling*.



Fonte: Elaborado pelo autor (2025).

A Figura 2 ilustra as estratégias de *undersampling* e *oversampling* aplicadas a conjuntos de dados desbalanceados. No centro da imagem, observa-se o conjunto de dados original, onde há uma disparidade clara entre as transações legítimas (representadas em verde claro) e as fraudes (representadas em rosa), evidenciando o desbalanceamento entre as classes.

À esquerda, o processo de *undersampling* é representado pela redução da quantidade de transações legítimas, de forma a equilibrar o número de observações com a quantidade de fraudes. Essa técnica seleciona uma amostra da classe majoritária (legítima), preservando todas as fraudes, para gerar um novo conjunto balanceado.

À direita, é observado o *oversampling*, onde a classe minoritária (fraudes) é aumentada por meio da replicação de observações, até que seu número se iguale ao da classe majoritária. Essa abordagem amplia artificialmente a presença de fraudes no conjunto de dados, mantendo todas as observações legítimas.

Ambas as estratégias visam corrigir o desbalanceamento entre classes, aumentando a sensibilidade dos modelos preditivos na detecção de fraudes e minimizando o viés em direção à classe mais frequente.

2.5 Técnicas alternativas em situações de dados desbalanceados.

Alternativas mais recentes vêm sendo propostas na literatura, como a modificação do critério de decisão do modelo em vez da base de dados. Uma dessas alternativas é o ajuste do *cutoff*, discutido por ASSUNÇÃO, IZBICKI e PRATES (2024) em seu artigo "Is Augmentation Effective in Improving Prediction in Imbalanced Datasets?".

Em modelos como a regressão logística, o resultado do modelo é uma probabilidade associada à ocorrência do evento de interesse, nesse estudo, a fraude. Para transformar essa probabilidade em uma decisão categórica (fraude ou não fraude), é definido um *cutoff* (ou limiar de decisão), que determina o ponto a partir do qual a observação será classificada como positiva (classe 1). O valor padrão desse limiar de decisão é 0,5, ou seja:

$$\hat{P}(Y = 1 \mid X) > 0{,}5$$

Classifica – se como pertencente à classe positiva (fraude), caso contrário, classifica-se como pertencente à classe negativa (transação legítima).

No entanto esse valor não é adequado em conjuntos de dados desbalanceados, isso ocorre porque o modelo tende a favorecer a classe majoritária, resultando em baixa sensibilidade (capacidade de identificar corretamente as fraudes) e alta especificidade (capacidade de identificar corretamente as transações legítimas). Dessa forma muitas fraudes podem ser classificadas erradas, como transações legítimas.

Para resolver esse problema, os autores propõem ajustar o *cutoff* de forma a maximizar a acurácia balanceada, que é a média entre a sensibilidade (detecção correta das fraudes) e a

especificidade (detecção correta das transações legítimas). A fórmula para determinar o *cutoff* ideal é:

$$g^*(x) = \begin{cases} 1, & se\ P(Y = 1 \mid x) \geq\ P(Y = 1) \\ 0, & caso\ contrário \end{cases}$$

$g^*(x)$ é a função de decisão do modelo, ela define se uma observação x será classificada como fraude ou não fraude. Nesta função, $P(Y = 1 \mid x)$ é a probabilidade estimada de que a observação x pertença a classe positiva (fraude), dada pelas predições do modelo. Por outro lado, $P(Y = 1)$ é a proporção de fraudes no conjunto de dados (classe minoritária), ou seja, a taxa base do evento raro.

Essa função é apresentada no Corolário 1 do artigo (p. 5), que classifica como fraude se a probabilidade prevista for maior ou igual à proporção de fraudes no banco de dados, caso contrário, classifique como transação legítima. Por exemplo, se apenas 5% das transações forem fraudes, o *cutoff* ideal será $c = 0{,}05$. Assim qualquer transação com $\hat{P}(Y = 1 \mid X) \geq 0{,}05$ será classificada como fraude.

Essa abordagem busca encontrar o valor de cutoff que equilibra a capacidade do modelo de identificar corretamente ambas as classes, sem a necessidade de técnicas de *oversampling* ou *undersampling*. Os autores mostram, teoricamente e com exemplos, que ajustar o *cutoff* pode produzir resultados comparáveis ou superiores aos obtidos com técnicas de amostragem.

3 Conjunto de dados

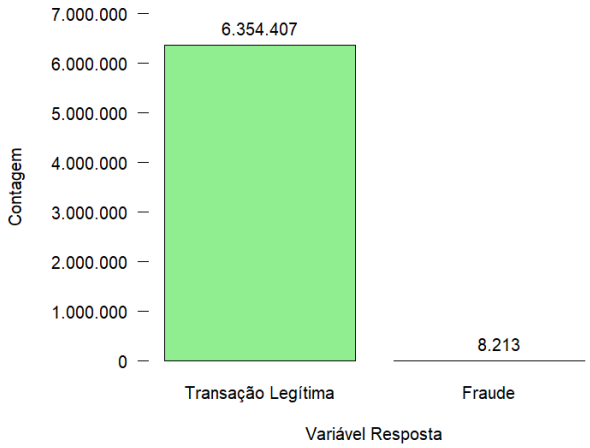
3.1 Metodologia

O objetivo deste estudo vai além da aplicação do modelo de regressão logística, buscando investigar o impacto de diferentes métodos de reamostragem nos dados e avaliar se essas técnicas aprimoram a capacidade preditiva do modelo na detecção de fraudes.

Os dados utilizados são retirados do Kaggle[1], que apresentam uma proporção muito pequena de fraudes em relação às transações legítimas. Esse desbalanceamento representa um dos maiores desafios na aplicação do modelo de regressão logística, uma vez que os modelos tendem a não captar adequadamente a classe minoritária (fraudes) sem intervenções adequadas de balanceamento.

[1]Kaggle é uma plataforma online de ciência de dados e aprendizado de máquina que oferece base de dados públicos que na maior parte são dados simulados a partir de dados reais.

Figura 3 – Exemplo de desbalanceamento de dados de fraude.



Fonte: Elaborado pelo autor (2025).

O gráfico ilustrado na Figura 3 representa um exemplo de desbalanceamento de dados no contexto de fraudes financeiras. Ele exibe duas classes: "Fraude" (1) e "Não Fraude" (0), representando transações fraudulentas e não fraudulentas, respectivamente. Observa-se uma enorme diferença na quantidade de registros entre as duas categorias, com a classe "Não Fraude" tendo 6.354.407 (99,87% da base total) observações, enquanto a classe "Fraude" possui apenas 8.213 (0,12% da base total), A partir desse gráfico, fica claro que será necessário aplicar técnicas de balanceamento de classes. Na fase de pré-processamento, os dados passaram por um processo de limpeza, que incluiu a remoção de valores ausentes.

Para o desenvolvimento do trabalho, as bases de dados serão divididas em duas bases. Uma delas contendo 70% dos dados (base de treino), para a as estimativas dos parâmetros do modelo, fase conhecida também como fase de treinamento. A verificação da performance do modelo, será realizada na base de dados restante (30%) conhecida como base de teste.

O modelo de regressão logística foi inicialmente ajustado sem qualquer tipo de balanceamento, de modo a avaliar seu desempenho com o conjunto de dados original, altamente desbalanceado. Após a obtenção das estimativas dos parâmetros do modelo nas quatro situações, sem balanceamento, com ajuste do *cutoff*, com *oversampling* e com *undersampling*, os resultados foram comparados para identificar qual abordagem apresentou melhor capacidade preditiva na detecção de fraudes. Esta capacidade preditiva foi testada na base de dados de teste.

Por fim, as métricas de avaliação foram utilizadas para verificar se as técnicas de reamostragem melhoraram a detecção da classe minoritária sem comprometer o desempenho geral do modelo. O trabalho foi desenvolvido utilizando RStudio e Python.

3.2 Descrição dos dados

Neste trabalho foram utilizadas duas bases de dados distintas, ambas relacionadas à detecção de fraudes em transações financeiras, e que são dados simulados retirados do Kaggle. De acordo com a descrição dos dados, estas bases foram simuladas com o objetivo de proporcionar aprendizado ao usuário para na construção de modelos estatísticos. O objetivo principal é avaliar a performance de modelos de regressão logística na identificação de transações fraudulentas, especialmente em cenários com diferentes níveis de desbalanceamento entre as classes. Em ambos os conjuntos de dados, a variável resposta é binária, indicando se uma determinada transação foi fraudulenta (1) ou legítima (0).

3.2.1 Descrição do primeiro conjunto de dados

A primeira base de dados utilizada é composta por um número elevado de observações, totalizando 6.362.620 registros. O objetivo é prever a probabilidade de um cliente ser uma fraude. A variável resposta é binária sendo codificada como 1 para fraude e 0 para transação legítima. Trata-se de uma base altamente desbalanceada, já que a proporção de casos positivos (fraude) é muito pequena.

A Tabela 2 contém a descrição das variáveis explicativas e da variável resposta.

Tabela 2 – Descrição das variáveis.

Variáveis	Descrição
STEP	Representa o tempo, onde 1 step é igual a 1 hora
TYPE	Tipo de transação online realizada
AMOUNT	Valor da transação
NAMEORIG	Protocolo do cliente
OLDBALANCEORG	Saldo da conta antes da transação
NEWBALANCEORIG	Saldo da conta após a transação
NAMEDEST	Destinatário da transação
OLDBALANCEDEST	Saldo inicial do destinatário antes da transação
NEWBALANCEDEST	O novo saldo do destinatário após a transação
ISFRAUD	Variável resposta que indica se a transação foi ou não fraudulenta

Fonte: Elaborado pelo autor (2025).

3.2.2 Descrição do segundo conjunto de dados

A segunda base, embora menor em número de observações (1.000.000 registros), apresenta uma proporção mais equilibrada entre as classes. A variável resposta também é binária, indicando a ocorrência de fraude (1) ou não (0), porém, neste caso, a taxa de fraude é de aproximadamente 8%, o que representa um cenário menos severamente desbalanceado comparada à primeira base de dados. Isso permite avaliar como o grau de desbalanceamento impacta na performance dos modelos.

A Tabela 3 contém a descrição das variáveis explicativas e da variável resposta.

Tabela 3 – Descrição das variáveis

Variáveis	Descrição
DISTANCE FROM HOME	A distância de casa onde a transação ocorreu
DISTANCE FROM LAST TRANSACTION	A distância desde a última transação
RATIO TO MEDIAN PURCHASE PRICE	Proporção entre o preço de transação realizado e o preço médio de compra
REPEAT RETAILER	A transação ocorreu no mesmo varejista ou não
USED CHIP	A transação ocorreu por meio de chip (cartão de crédito) ou não
USED PIN NUMBER	A transação ocorreu usando o número PIN ou não
ONLINE ORDER	A transação é um pedido online ou não
ISFRAUD	Variável resposta que indica se a transação foi ou não fraudulenta

Fonte: Elaborado pelo autor (2025).

3.3 Análise exploratória do primeiro conjunto de dado

O primeiro conjunto de dado analisado neste trabalho é composto por transações simuladas, geradas com o objetivo de representar situações reais de fraudes financeiras. Para melhor entendimento dos dados foi realizado uma análise descritiva das variáveis.

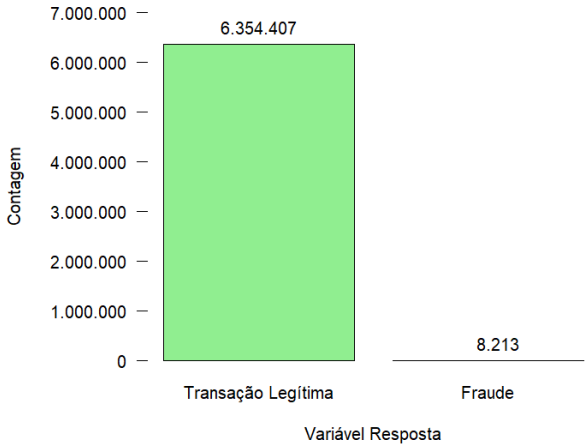
Tabela 4 – Análise Descritiva das variáveis numéricas

	step	amount	oldbalanceOrg	newbalanceOrig	oldbalanceDest	newbalanceDest	isFraud
Média	243,40	179.861,90	833.883,10	855.113,67	1.100.701,67	1.224.996,40	0,00
Desvio Padrão	142,33	603.858,23	2.888.242,67	2.924.048,50	3.399.180,11	3.674.128,94	0,04
Mínimo	1,00	0,00	0,00	0,00	0,00	0,00	0,00
25%	156,00	13.389,57	0,00	0,00	0,00	0,00	0,00
50%	239,00	74.871,94	14.208,00	0,00	132.705,66	214.661,44	0,00
75%	335,00	208.721,48	107.315,18	144.258,41	943.036,71	1.111.909,25	0,00
Máximo	743,00	92.445.516,64	59.585.040,37	49.585.040,37	356.015.889,35	356.79.278,92	1,00

Fonte: Elaborado pelo autor (2025).

É observado na Tabela 4 as estatísticas resumidas, como média, desvio padrão, valores mínimos, máximos e quartis de cada uma das variáveis. Observa-se uma variância muito grande em relação ao valor da transação, sendo o valor mínimo de transação foi 0 (zero) e o valor máximo de mais de 92 milhões. O mesmo ocorre com outras variáveis.

Figura 4 – Gráfico de barras da variável resposta que identifica o desfecho (fraude e não fraude).



Fonte: Elaborado pelo autor (2025).

A Figura 4 exibe a distribuição das transações quanto à ocorrência ou não de fraude, com base na variável isFraud. O eixo X indica se a transação foi fraudulenta (1) ou não (0), enquanto o eixo Y mostra a contagem de transações. Observa-se que a grande maioria das transações não foi fraudulenta. Especificamente, o total de transações legítimas é de 6.354.407, enquanto as transações classificadas como fraudes somam apenas 8.213. Essa diferença expressiva evidencia o forte desbalanceamento da base de dados, onde menos de 0,13% das transações são fraudulentas. Esse tipo de desbalanceamento é comum em problemas reais de detecção de fraude, já que fraudes são eventos raros. Contudo, ele representa um desafio importante na modelagem, pois modelos de classificação tendem a favorecer a classe majoritária (não fraude) se nenhuma medida for tomada para compensar esse desequilíbrio.

Tabela 5 – Diferença das médias para variáveis contínuas.

	step	amount	oldbalanceOrg	newbalanceOrig	oldbalanceDest	newbalanceDest
isFraud						
1	368,41	1.467.967,30	1.649.667,61	192.392,63	544.249,62	1.279,707,62
0	243,24	178.197,04	832.828,71	855.970,23	1.101.420,87	1.224,925,68
Diferença das médias	125,18	1.289.770,26	816.838,89	-663.577,60	-557.171,26	54.781,93

Fonte: Elaborado pelo autor (2025).

A Tabela 5 retrata as médias das variáveis contínuas separadas conforme a ocorrência de fraude (isFraud = 1) ou não (isFraud = 0), bem como a diferença entre essas médias. A análise das diferenças nos permite identificar padrões que distinguem transações fraudulentas das legítimas.

O valor médio das transações fraudulentas é expressivamente maior (R$ 1.467.967,30) quando comparado ao das transações não fraudulentas (R$ 178.197,04), com uma diferença superior a R$ 1.2 milhão. Isso sugere que as fraudes geralmente envolvem quantias significativamente mais elevadas. Além disso, o saldo da conta de origem antes da transação (oldbalanceOrg) é, em média, mais alto nos casos de fraude, indicando que fraudadores podem preferir alvos com maior disponibilidade de recursos. Após a transação, o saldo médio (newbalanceOrig) nas transações fraudulentas é muito inferior ao das transações legítimas, o que pode indicar uma retirada quase total do valor disponível.

Em relação à conta de destino, o saldo anterior à transação (oldbalanceDest) é menor nos casos de fraude, o que pode apontar para o uso de contas recém-criadas ou pouco movimentadas. Já o saldo posterior à transação (newbalanceDest) é levemente superior nas transações fraudulentas, mas a diferença é pequena, o que pode indicar tentativas de disfarçar a operação ou movimentações imediatas após o recebimento do valor.

Esses padrões evidenciam que variáveis como valor da transação, saldo antes e depois nas contas de origem e destino podem ser relevantes para a detecção de fraudes e, portanto, devem ser levadas em consideração na modelagem preditiva.

Para comparar as diferenças entre transações fraudulentas e não fraudulentas, foi realizado o teste t nas variáveis contínuas do conjunto de dados. O teste t é uma técnica estatística que verifica se há diferença significativa entre as médias de dois grupos, considerando a variabilidade dos dados. Nesse caso, os grupos comparados foram as transações classificadas como fraude (ISFRAUD = 1) e não fraude (ISFRAUD = 0). Foram analisadas as seguintes variáveis contínuas: STEP (tempo em horas desde o início da observação), AMOUNT (valor da transação), OLDBALANCEORG (saldo da conta de origem antes da transação), NEWBALANCEORIG (saldo da conta de origem após a transação), OLDBALANCEDEST (saldo da conta destinatária antes da transação) e NEWBALANCEDEST (saldo da conta destinatária após a transação). Variáveis categóricas e identificadores, como TYPE, NAMEORIG e NAMEDEST, não foram incluídos nesse teste.

Tabela 6 – Teste T para comparação de médias das variáveis numéricas entre as categorias de desfecho (fraude e não fraude).

Variável	Estatística T	p-valor	Diferença das médias
step	52,41	< 0,05	125,18
amount	48,61	< 0,05	1.289.770,26
oldbalanceOrg	20,86	< 0,05	816.838,89
newbalanceOrig	30,55	< 0,05	-663.577,60
oldbalanceDest	-15,12	< 0,05	-557.171,26
newbalanceDest	1,27	> 0,05	54.781,93

Fonte: Elaborado pelo autor (2025).

Os resultados indicam que, com exceção da variável NEWBALANCEDEST, todas as demais apresentaram diferenças médias estatisticamente significativas entre os grupos, com p-valores inferiores a 0,05. Isso significa que a média dessas variáveis difere de forma relevante entre transações fraudulentas e não fraudulentas.

Mais especificamente, a variável STEP apresentou uma diferença média positiva, sugerindo que as fraudes tendem a ocorrer em horários diferentes das transações legítimas. O valor médio das transações fraudulentas (AMOUNT) foi significativamente maior, evidenciando que fraudes geralmente envolvem valores mais elevados. O saldo da conta origem antes da transação (OLDBALANCEORG) também foi maior nas fraudes, enquanto o saldo após a transação (NEWBALANCEORIG) apresentou redução significativa, indicando retirada de recursos. Quanto à conta destinatária, o saldo inicial (OLDBALANCEDEST) foi menor nas transações fraudulentas, possivelmente indicando que os valores são enviados para contas com pouco saldo, muitas vezes associadas a esquemas fraudulentos. Contudo, o saldo após a transação (NEWBALANCEDEST) não apresentou diferença significativa, sugerindo que essa variável não é um indicador relevante para distinguir fraudes.

Para investigar a relação entre o tipo de transação (type) e a ocorrência de fraude (isFraud), foi realizado o teste qui-quadrado de independência. Esse teste avalia se existe associação significativa entre duas variáveis categóricas, comparando as frequências observadas com as frequências esperadas sob a hipótese de independência. A variável type representa o tipo de operação financeira realizada, com categorias como CASH_IN, CASH_OUT, DEBIT, PAYMENT e TRANSFER. A variável isFraud indica se a transação foi classificada como fraude (1) ou não (0).

Tabela 7 – Tabela de contingência entre o tipo de transação e desfecho (fraude e não fraude)

Tipo de Transação	Não Fraude	Fraude	Total
CASH_IN	1.399.284	0	1.399.284
CASH_OUT	2.233.384	4.116	2.237.500
DEBIT	41.432	0	41.432
PAYMENT	2.151.495	0	2.151.495
TRANSFER	528.812	4.097	532.909
Total	6.354.407	8.213	6.362.620

Fonte: Elaborado pelo autor (2025).

Os resultados mostraram um valor da estatística qui-quadrado de 22.082,54 com 4 graus de liberdade e p-valor inferior a 0,0001, indicando forte evidência de associação entre o tipo de transação e a ocorrência de fraude. Em outras palavras, a distribuição dos tipos de transação difere significativamente entre operações fraudulentas e não fraudulentas. Ao observar as frequências, percebe-se que as transações do tipo TRANSFER e CASH_OUT concentram praticamente todos os casos de fraude. Enquanto isso, os demais tipos de transação, como CASH_IN, DEBIT e PAYMENT, não apresentam fraudes no conjunto de dados analisado. Isso sugere que fraudes estão fortemente associadas a operações de transferência de valores ou saques.

3.4 Análise exploratória do segundo conjunto de dados

O segundo conjunto de dado analisado neste trabalho também é composto por transações simuladas, geradas com o objetivo de representar situações reais de fraudes financeiras. Para melhor entendimento dos dados foi realizado uma análise descritiva das variáveis apresentadas pela Tabela 8.

Tabela 8 – Análise Descritiva das variáveis númericas

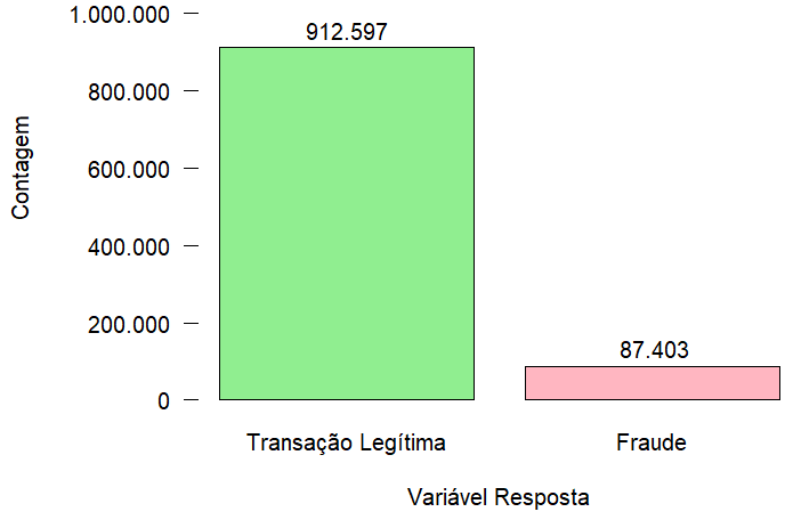
	distance_from_home	distance_from_last_transaction	ratio_to_median_purchase_price
Média	26,63	5,04	1,82
Desvio Padrão	65,39	25,84	2,80
Mínimo	0,00	0,00	0,00
25%	3,88	0,30	0,48
50%	9,97	1,00	1,00
75%	25,74	3,36	2,10
Máximo	10.632,72	11.851,10	267,80

	repeat_retailer	used_chip	used_pin_number	online_order	fraud
Média	0,88	0,35	0,10	0,65	0,09
Desvio Padrão	0,32	0,48	0,30	0,48	0,28
Minimo	0,00	0,00	0,00	0,00	0,00
25%	1,00	0,00	0,00	0,00	0,00
50%	1,00	0,00	0,00	1,00	0,00
75%	1,00	1,00	0,00	1,00	0,00
Maximo	1,00	1,00	1,00	1,00	1,00

Fonte: Elaborado pelo autor (2025).

Na tabela 8 são apresentadas as estatísticas resumidas, como média, desvio padrão, valores mínimos, máximos e quartis de cada uma das variáveis.

Figura 5 – Gráfico de barras da variável resposta que identifica o desfecho (fraude e não fraude).



Fonte: Elaborado pelo autor (2025).

É observado na Figura 5 a distribuição da variável Fraud, que indica se uma transação foi ou não classificada como fraudulenta. Observa-se que, do total de transações analisadas, 912.597 (aproximadamente 91,3%) são não fraudulentas (Fraud = 0), enquanto apenas 87.403 (cerca de 8,7%) foram classificadas como fraudes (Fraud = 1). A visualização foi feita em escala logarítmica no eixo y, o que facilita a observação de diferenças expressivas entre as categorias.

Tabela 9 – Diferença das médias para variáveis contínuas.

	distance_from_home	distance_from_last_transaction	ratio_to_median_purchase_price
isFraud			
1	66,26	12,71	6,01
0	22,83	4,30	1,42
Diferença das médias	43,43	8,41	4,58

Fonte: Elaborado pelo autor (2025).

Na Tabela 9 observa-se a comparação das médias de três variáveis contínuas entre transações fraudulentas e não fraudulentas, com o objetivo de identificar possíveis padrões de comportamento associados às fraudes. As variáveis analisadas foram distance_from_home, que representa a distância da transação em relação à residência do usuário, distance_from_last_transaction, que mede a distância em relação à última transação realizada e ratio_to_median_purchase_price, que corresponde à razão entre o valor da compra e o valor mediano das compras anteriores daquele usuário. Observa-se que, para transações classificadas como fraude, a média da distance_from_home foi de 66,26, enquanto para transações legítimas foi de 22,83, resultando em uma diferença de 43,43 unidades. Já a média da distance_from_last_transaction foi de 12,71 para fraudes e 4,30 para não fraudes, com diferença de 8,41. Por fim, a razão entre o valor da compra e a mediana das compras anteriores (ratio_to_median_purchase_price) apresentou média de 6,01 em transações fraudulentas, contra 1,42 em transações legítimas, com uma diferença de 4,58. Esses resultados indicam que transações fraudulentas tendem a ocorrer a distâncias maiores da residência do cliente e da última transação registrada, além de envolverem valores proporcionalmente mais altos em relação ao histórico de compras do usuário.

Tabela 10 – Teste de comparação de médias - Teste t-student.

Variável	Estatística T	p-valor	Diferença das médias
distance_from_home	94,84	< 0,05	43,43
distance_from_last_transaction	51,27	< 0,05	8,41
ratio_to_median_purchase_price	242,07	< 0,05	4,58

Fonte: Elaborado pelo autor (2025).

Para verificar se havia diferença significativa nas médias das variáveis contínuas entre transações fraudulentas e não fraudulentas, foi realizado o teste t de Student. Os resultados obtidos para as variáveis distance_from_home, distance_from_last_transaction e ratio_to_median_purchase_price. Todas as variáveis apresentaram p-valor inferior a 0,05, indicando que as diferenças observadas entre as médias são estatisticamente significativas ao nível de 5%. A variável distance_from_home apresentou uma estatística t de 94,84 e diferença média de 43,43 unidades entre os dois grupos, sugerindo que transações fraudulentas tendem a ocorrer a distâncias significativamente maiores da residência do usuário. A variável distance_from_last_transaction apresentou uma estatística t de 51,27, com diferença média de 8,41, indicando que fraudes também ocorrem mais frequentemente a distâncias maiores da última transação registrada. Já a variável ratio_to_median_purchase_price apresentou uma estatística t de 242,07 e diferença média de 4,58, revelando que fraudes estão associadas a compras com valores proporcionalmente mais altos em relação à mediana de compras anteriores do usuário.

Com o objetivo de avaliar a associação entre a variável resposta isFraud e um conjunto de variáveis explicativas categóricas, foi aplicado o teste do qui-quadrado de independência. Esse teste permite verificar se existe dependência estatística entre duas variáveis categóricas, ou seja, se a distribuição dos valores de uma variável difere significativamente de acordo com os níveis da outra. As variáveis analisadas foram: repeat_retailer (indica se o varejista já foi utilizado anteriormente pelo cliente), used_chip (informa se a transação foi realizada com chip), used_pin_number (informa se o número PIN foi utilizado) e online_order (indica se a transação foi realizada online). Essas variáveis foram selecionadas por representarem características operacionais da transação que podem estar associadas a comportamentos típicos de fraude.

Tabela 11 – Teste Qui-Quadrado para verificar a relação entre as categóricas e o desfecho.

Variável	Estatística Qui-quadrado	p-valor
repeat_retailer	1,83	> 0,05
used_chip	3,717	< 0,05
used_pin_number	10,057	< 0,05
online_order	36,852	< 0,05

Fonte: Elaborado pelo autor (2025).

Como pode ser observado na Tabela 11, o teste qui-quadrado realizado para verificar a relação das variáveis de categóricas com o desfecho, os resultados revelam que três dessas variáveis possuem associação estatisticamente significativa com a variável resposta isFraud, dado que apresentaram valores de p inferiores a 0,05. São elas: used_chip (qui-quadrado = 3,717), used_pin_number (qui-quadrado = 10,057) e online_order (qui-quadrado = 36,852). Isso indica que a ocorrência de fraude está relacionada ao uso de chip, uso de senha e à realização do pedido online. Por outro lado, a variável repeat_retailer apresentou um valor de p superior a 0,05 (qui-quadrado = 1,83), não evidenciando associação significativa com fraudes.

4 RESULTADOS E DISCUSSÃO

4.1 Resultados do primeiro conjunto de dados

Antes do ajuste do modelo, realizou-se uma etapa inicial de preparação da base de dados. Primeiramente, foram removidas as variáveis nameOrig, nameDest e step, por não apresentarem relevância para o processo de modelagem e por não contribuírem para a capacidade preditiva do modelo. Em seguida, a base de dados foi dividida em dois subconjuntos utilizando amostragem aleatória simples: 70% dos registros foram destinados ao treinamento, 30% para o teste do modelo. Após essa etapa, o modelo de Regressão Logística foi ajustado sem a aplicação de qualquer técnica de balanceamento, mantendo-se a distribuição natural das classes. A seguir, são apresentados os resultados obtidos para esse primeiro modelo.

Tabela 12 – Coeficientes estimados do modelo de Regressão Logística (sem balanceamento).

Variável	Estimativa (β)	Erro-padrão	z-value	p-value	Significância
(Intercept)	-29,420	117,500	-0,250	0,802	
typeCASH_OUT	23,860	117,500	0,203	0,839	
typeDEBIT	-2,727	1195,000	-0,002	0,998	
typePAYMENT	4,064	208,600	0,019	0,984	
typeTRANSFER	25,500	117,500	0,217	0,828	
amount	0,000	0,000	-39,111	<0,001	***
oldbalanceOrg	0,000	0,000	59,839	<0,001	***
newbalanceOrig	0,000	0,000	-63,067	<0,001	***
oldbalanceDest	0,000	0,000	47,787	<0,001	***
newbalanceDest	0,000	0,000	-51,302	<0,001	***

Fonte: Elaborado pelo autor (2025).

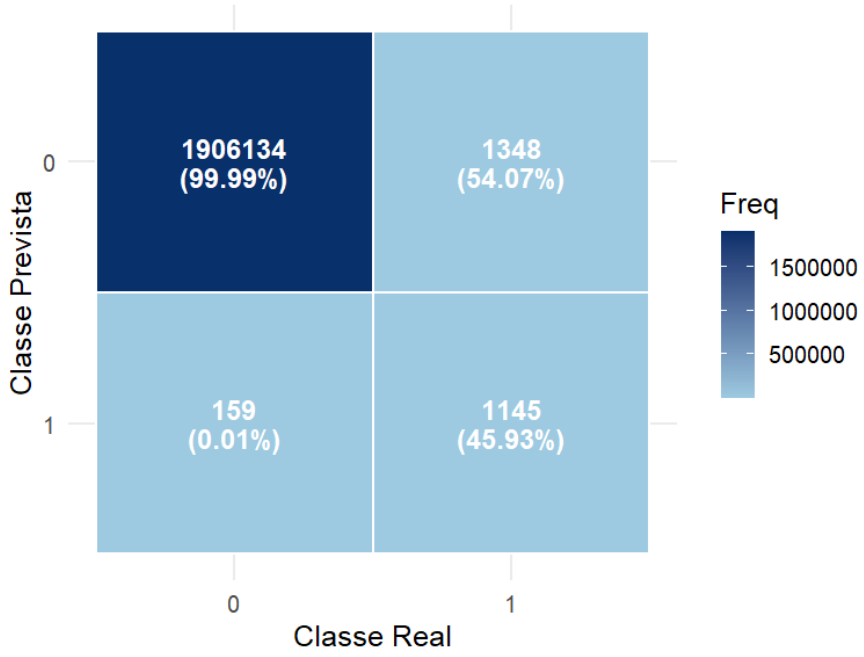
Na Tabela 12 observa-se os coeficientes estimados pelo modelo de Regressão Logística ajustado ao conjunto de treinamento sem a aplicação de técnicas de balanceamento. As variáveis categóricas relacionadas ao tipo de transação (CASH_OUT, DEBIT, PAYMENT e TRANSFER) não se mostraram estatisticamente significativas, indicando que, essas categorias não contribuem de forma relevante para explicar a ocorrência de fraude neste conjunto de dados.

Por outro lado, as variáveis contínuas amount, oldbalanceOrg, newbalanceOrig, oldbalanceDest e newbalanceDest apresentaram coeficientes altamente significativos ($p < 0,001$). Essas variáveis representam informações financeiras diretamente ligadas ao fluxo de valores nas contas envolvidas, o que justifica a importância delas na detecção de padrões associados à fraude.

Em relação às medidas globais de ajuste, o modelo apresentou AIC igual a 30.391. O AIC não avalia qualidade absoluta, mas serve como critério comparativo entre modelos: quanto menor o valor, melhor o equilíbrio entre ajuste e complexidade.

Dessa forma, após a interpretação dos coeficientes e do ajuste global, prossegue-se para a avaliação do desempenho preditivo do modelo no conjunto de teste, adotando-se o *cutoff* padrão de 0,5.

Figura 6 – Matriz de Confusão – Conjunto de Teste (*cutoff* = 0,5).



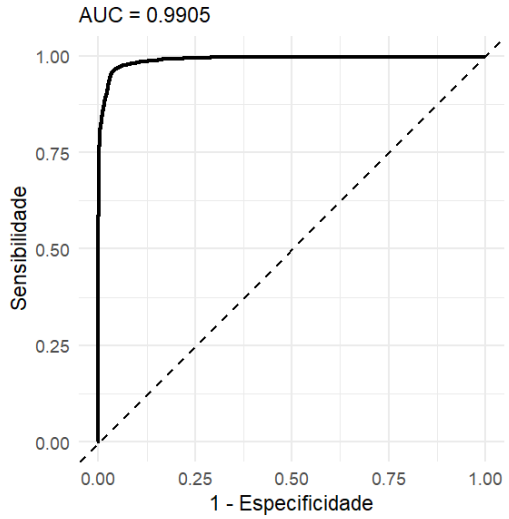
Fonte: Elaborado pelo autor (2025).

A partir da matriz de confusão construída para o conjunto de teste, observa-se que o modelo classificou corretamente 1.906.134 transações não fraudulentas (classe 0), representando 99,99% desse grupo. Em relação às transações fraudulentas (classe 1), 1.145 foram reconhecidas corretamente (45,93%), enquanto 1.348 foram incorretamente rotuladas como legítimas (54,07%). Esses resultados indicam que, embora o modelo apresente desempenho exemplar na identificação da classe majoritária, ainda enfrenta limitações na detecção da classe minoritária.

As métricas reforçam esse comportamento. A acurácia está elevada (99,92%), reflexo direto do desbalanceamento da base, no qual a classe 0 domina o conjunto de dados. A especificidade atinge valor quase perfeito (99,99%), confirmando que o modelo praticamente não rotula transações legítimas como fraude. A precisão, igual a 87,81%, indica que a maioria das instâncias previstas como fraude realmente pertence a essa classe, o que é um ponto positivo.

No entanto, o F1-Score de 60,31%, que combina precisão e sensibilidade, evidencia uma limitação importante: embora as previsões de fraude realizadas pelo modelo sejam confiáveis, a sensibilidade continua baixa, fazendo com que muitas fraudes reais passem despercebidas. Isso confirma que o *cutoff* de 0,5 não é adequado para maximizar a detecção de fraudes nesse cenário.

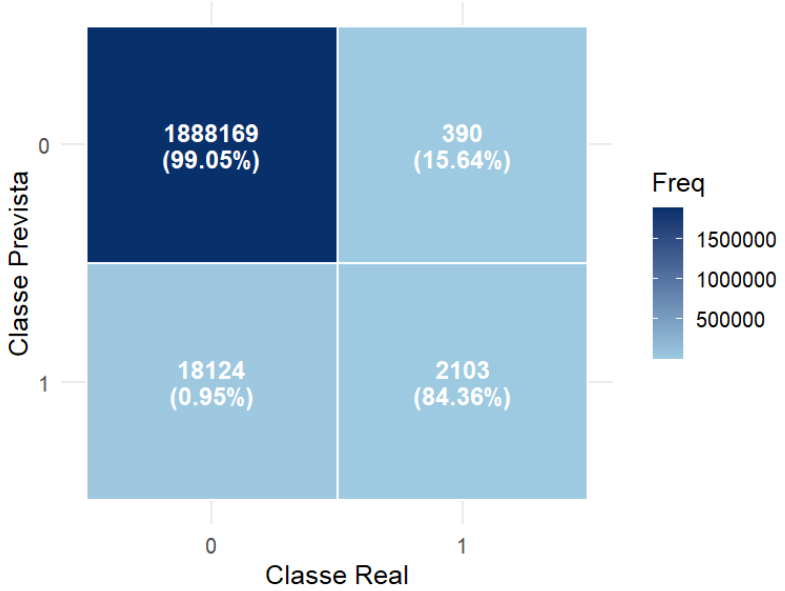
Figura 7 – Curva ROC – Conjunto de Teste (*cutoff* = 0,5).



Fonte: Elaborado pelo autor (2025).

A curva ROC reforça os resultados observados. A curva está próxima ao canto superior esquerdo, e o valor do AUC = 0.9905 indica excelente capacidade discriminativa do modelo ao ranquear as instâncias de acordo com sua probabilidade estimada de fraude. Ou seja, o modelo consegue separar bem as classes em termos de probabilidade, ainda que o *cutoff* padrão não explore plenamente esse potencial. Na continuidade da análise, apresentam-se os resultados considerando o novo *cutoff* de 0,01.

Figura 8 – Matriz de Confusão – Conjunto de Teste (*cutoff* = 0,01).



Fonte: Elaborado pelo autor (2025).

A matriz de confusão obtida com o *cutoff* reduzido para 0,01 evidencia uma mudança importante no comportamento do modelo em relação aos resultados anteriores. Com esse novo limiar, o modelo passa a identificar um número maior de transações fraudulentas. Das instâncias realmente pertencentes à classe 1, 2.103 foram corretamente classificadas como fraude (84,36%), representando um aumento expressivo na sensibilidade. A adoção de um *cutoff* menor resultou em menos falsos negativos, evidenciando um ganho importante na detecção de fraudes que antes passavam despercebidas.

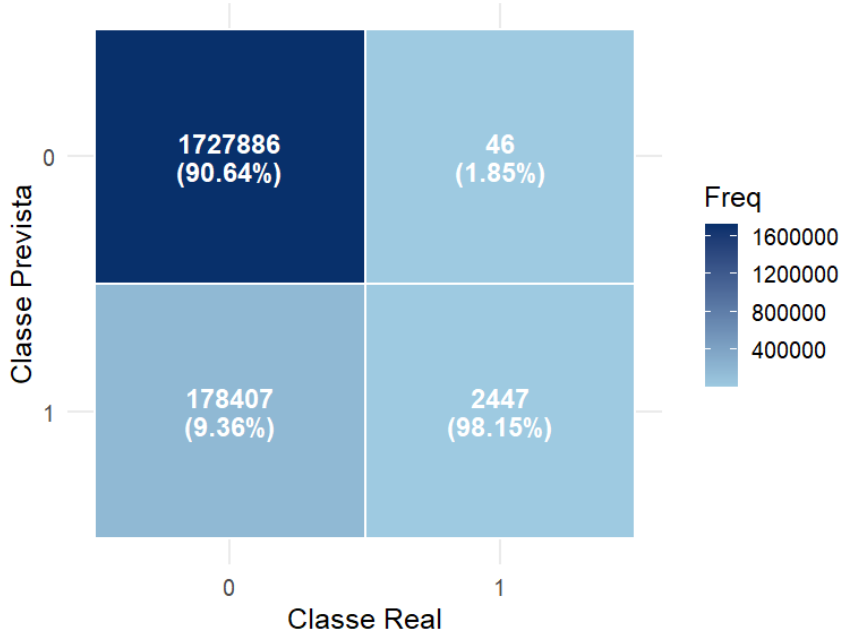
Apesar desse avanço na detecção de fraudes, a classificação de transações não fraudulentas permanece majoritariamente correta, 1.888.169 casos da classe 0 (99,05%) foram corretamente rotuladas como legítimas, enquanto 18.124 foram classificadas como fraude (0,95%). Esse padrão indica que, embora o modelo tenha se tornado mais sensível, ele ainda mantém um bom desempenho na identificação da classe majoritária.

As métricas numéricas reforçam essas observações. A acurácia, apesar da mudança no *cutoff*, permanece elevada (99,03%). A especificidade apresenta leve queda, mas ainda em nível elevado (99,05%), o que significa que a maioria das transações legítimas continua sendo identificada corretamente.

A precisão da classe fraudulenta, porém, atinge 10,4%, indicando que apenas cerca de 10% das transações classificadas como fraude são realmente fraudulentas. Essa redução era esperada, uma vez que o *cutoff* mais baixo favorece o aumento dos falsos positivos. Já o F1-Score, que combina precisão e sensibilidade, assume valor 18,51%, refletindo que, embora a capacidade de identificar fraudes tenha melhorado, há perda considerável de precisão nessas previsões.

Na etapa seguinte, são apresentados os resultados obtidos no conjunto de teste utilizando um *cutoff* igual a 0,0012, correspondente à proporção de fraudes observada na base de dados. A adoção desse limiar segue a recomendação de, Assunção, Izbicki e Prates (2024), que sugerem o uso da prevalência da classe minoritária como ponto de corte inicial em cenários de forte desbalanceamento.

Figura 9 – Matriz de Confusão – Conjunto de Teste (*cutoff* = 0,0012).



Fonte: Elaborado pelo autor (2025).

A matriz de confusão obtida com o *cutoff* igual a 0,0012, valor correspondente à proporção de fraudes observada na base de dados, conforme sugerido por Assunção, Izbicki e Prates (2024), evidencia uma mudança expressiva no comportamento do modelo. Com esse limiar extremamente reduzido, o modelo passa a classificar como fraude praticamente todas as transações cuja probabilidade estimada seja mínima, elevando substancialmente a sensibilidade.

Das transações realmente fraudulentas, 2.447 foram corretamente identificadas (98,15%), indicando que o modelo se torna altamente eficaz em detectar fraudes quando utiliza o *cutoff* baseado na prevalência da classe minoritária. No entanto, essa forte ampliação da sensibilidade resulta em um aumento considerável no número de falsos positivos: 178.407 transações legítimas (9,36%) foram classificadas como fraude. Ainda assim, 1.727.886 transações legítimas (90,64%) foram devidamente reconhecidas como não fraudulentas.

As métricas numéricas refletem esse equilíbrio delicado entre sensibilidade e precisão. A acurácia cai para 90,65%, a especificidade, que mede a capacidade do modelo de identificar corretamente as transações legítimas, também diminui para 90,64%, permanecendo compatível com o percentual de classificações corretas da classe majoritária.

Por outro lado, a precisão atinge um valor muito baixo (1,35%), indicando que apenas cerca de 1,3% das transações rotuladas como fraude são, de fato, fraudulentas.

Consequentemente, o F1-Score, que integra precisão e sensibilidade, assume valor reduzido (2,67%).

Dando continuidade às análises, procedeu-se à aplicação da técnica de *oversampling* no conjunto de treinamento. Para isso, a classe minoritária (fraudes) foi aumentada até atingir a mesma volumetria da classe majoritária, de forma que ambas as classes passassem a possuir quantidades equivalentes de observações

Com o conjunto balanceado, foi ajustado um novo modelo de Regressão Logística utilizando exclusivamente o treino resultante do *oversampling*. Embora o processo de otimização não tenha convergido, ainda assim foi possível obter as estimativas dos coeficientes.

Tabela 13 – Coeficientes estimados do modelo de Regressão Logística (*oversampling*).

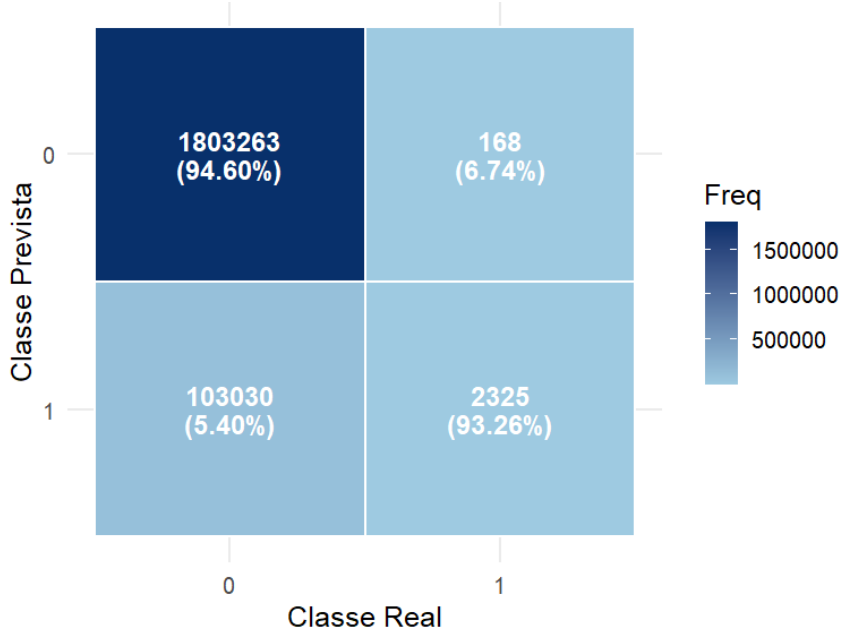
Variável	Estimativa (β)	Erro-padrão	z-value	p-value	Significância
(Intercept)	-42,420	0,461	-92,104	<0,001	***
typeCASH_OUT	41,980	0,461	91,108	<0,001	***
typeDEBIT	-25210,000	393500,000	-0,064	0,949	
typePAYMENT	-19,900	40,150	-0,496	0,620	
typeTRANSFER	43,600	0,461	94,506	<0,001	***
amount	0,000	0,000	-95,505	<0,001	***
oldbalanceOrg	0,000	0,000	900,863	<0,001	***
newbalanceOrig	0,000	0,000	-806,567	<0,001	***
oldbalanceDest	0,000	0,000	372,846	<0,001	***
newbalanceDest	0,000	0,000	-377,256	<0,001	***

Fonte: Elaborado pelo autor (2025).

Observa-se que, nesse cenário, as variáveis relacionadas ao tipo de transação (CASH_OUT e TRANSFER) tornam-se altamente significativas, com coeficientes positivos de grande magnitude, indicando forte associação dessas categorias com a ocorrência de fraude quando o modelo é treinado em uma base balanceada. As variáveis financeiras (amount, oldbalanceOrg, newbalanceOrig, oldbalanceDest e newbalanceDest) também permanecem estatisticamente significativas, ainda que com coeficientes de pequena magnitude em termos absolutos, o que é compatível com a escala monetária elevada das variáveis. O modelo com *oversampling* apresentou AIC igual a 2.638.860.

Após ajustar o modelo utilizando o conjunto de treinamento balanceado por *oversampling*, avaliou-se seu desempenho no conjunto de teste real mantendo o *cutoff* padrão de 0,5.

Figura 10 – Matriz de Confusão – Conjunto de Teste (*oversampling*).

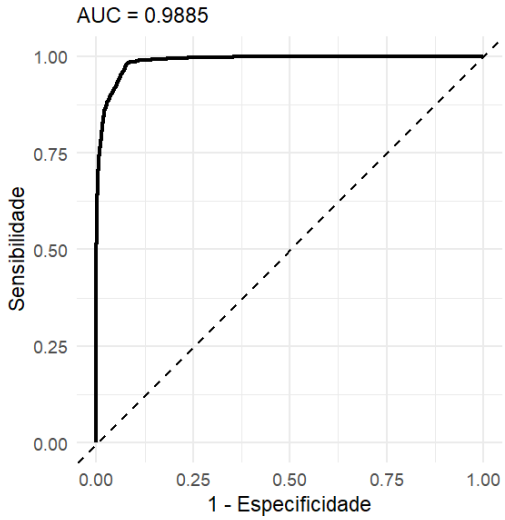


Fonte: Elaborado pelo autor (2025).

O modelo passou a identificar uma quantidade maior de transações fraudulentas corretamente, alcançando 2.325 verdadeiros positivos, o que corresponde a uma sensibilidade de 93,26%.

Entretanto, essa melhora na sensibilidade ocorreu acompanhada de um aumento substancial no número de falsos positivos, com 103.030 transações legítimas classificadas incorretamente como fraude. Esse comportamento explica a precisão extremamente baixa observada, igual a 2,21%. A acurácia registrada, de 94,59%, permanece elevada principalmente pelo grande volume de transações legítimas no conjunto de teste. O F1-Score, que sintetiza a relação entre precisão e sensibilidade, apresentou valor reduzido (4,31%), consequência direta da baixa precisão, apesar da alta sensibilidade.

Figura 11 – Curva ROC – Conjunto de Teste (*oversampling*).



Fonte: Elaborado pelo autor (2025).

A curva ROC obtida confirma que o modelo treinado com *oversampling* possui boa capacidade discriminatória, apresentando uma área sob a curva igual a 0,9885.

Após avaliar o desempenho do modelo sem balanceamento e com *oversampling*, foi aplicada a técnica de *undersampling* ao conjunto de treinamento. Nesse procedimento, a classe majoritária (transações legítimas) foi reduzida aleatoriamente até atingir aproximadamente o mesmo volume da classe minoritária, resultando em um conjunto de dados substancialmente menor, porém balanceado.

Assim como observado no *oversampling*, o modelo ajustado ao conjunto subamostrado também não convergiu completamente. Ainda assim, o procedimento produziu estimativas válidas para os coeficientes.

Tabela 14 – Coeficientes estimados do modelo de Regressão Logística (*undersampling*).

Variável	Estimativa (β)	Erro-padrão	z-value	p-value	Significância
(Intercept)	-4,722	0,400	-11,814	<0,001	***
typeCASH_OUT	4,335	0,405	10,700	<0,001	***
typeDEBIT	-21,43	50460,000	0,000	0,9997	
typePAYMENT	-21,6	6353,000	-0,003	0,9973	
typeTRANSFER	5,827	0,414	14,066	<0,001	***
amount	-0,000	0,000	-2,436	0,015	*
oldbalanceOrg	0,000	0,000	32,442	<0,001	***
newbalanceOrig	-0,000	0,000	-32,262	<0,001	***
oldbalanceDest	0,000	0,000	27,796	<0,001	***
newbalanceDest	-0,000	0,000	-28,120	<0,001	***

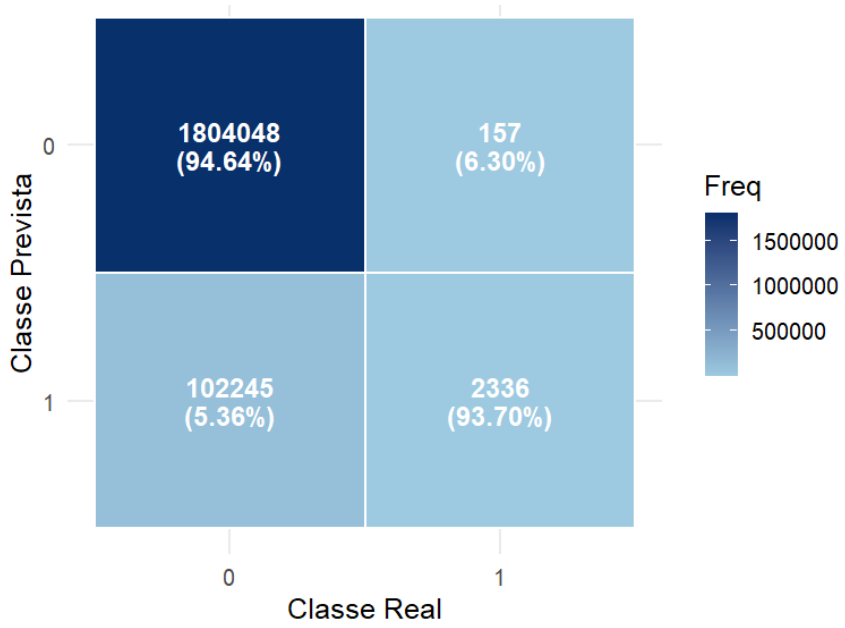
Fonte: Elaborado pelo autor (2025).

As estimativas obtidas mostram que, apesar da falta de convergência, o modelo consegue identificar relações importantes entre algumas variáveis e a probabilidade de fraude. As categorias CASH_OUT e TRANSFER, por exemplo, mantêm coeficientes positivos significativos, indicando maior associação com transações fraudulentas, comportamento semelhante ao observado nos outros modelos.

As variáveis categóricas DEBIT e PAYMENT permanecem não significativas. Já as variáveis contínuas relacionadas aos saldos e ao valor da transação, embora apresentem coeficientes muito pequenos devido à sua escala original, continuam altamente significativas, o que indica que mudanças nessas quantias influenciam consistentemente a chance de uma transação ser fraudulenta. O modelo apresentou AIC = 3364,5.

A seguir, são apresentados os resultados do modelo ajustado com *undersampling* quando aplicado ao conjunto de teste utilizando o *cutoff* padrão de 0,5.

Figura 12 – Matriz de Confusão – Conjunto de Teste (*undersampling*).



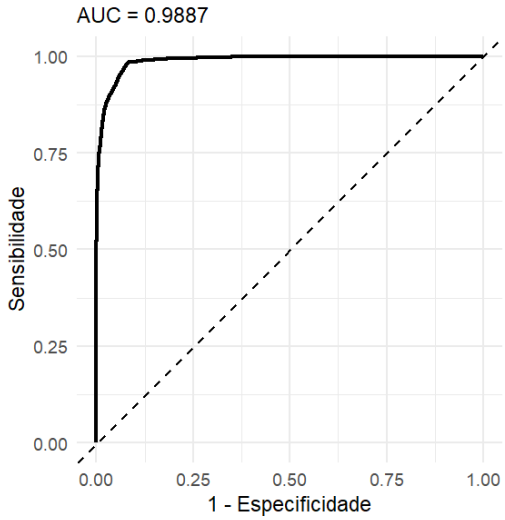
Fonte: Elaborado pelo autor (2025).

A matriz de confusão evidencia que o modelo foi capaz de identificar corretamente 93,70% das transações fraudulentas presentes no conjunto de teste, indicando um bom desempenho em termos de sensibilidade. Por outro lado, a taxa de falsos positivos elevada, resultando em uma precisão baixa (2,23%).

A acurácia atingiu 94,64%, valor fortemente influenciado pelo alto volume de transações legítimas na base. A especificidade também foi de 94,64%, indicando que o modelo

manteve boa capacidade de reconhecer observações da classe majoritária. Já o recall apresentou valor de 93,7%, compatível com o observado na matriz de confusão, mostrando que a técnica de *undersampling* ampliou a capacidade do modelo de capturar fraudes. Contudo, o F1-Score permaneceu baixo (4,36%), reflexo direto da combinação entre recall elevado e precisão muito reduzida.

Figura 13 – Curva ROC – Conjunto de Teste (*undersampling*).



Fonte: Elaborado pelo autor (2025).

A curva ROC obtida para o modelo ajustado com *undersampling* no conjunto de teste indica que, o modelo apresenta boa capacidade discriminativa. A área sob a curva (AUC = 0,9887) mostra que o classificador consegue separar adequadamente as classes na maior parte dos limiares de decisão, mantendo um desempenho significativamente superior ao de um modelo aleatório, representado pela linha tracejada que marca a diagonal de referência.

O comportamento da curva, que rapidamente se aproxima da região superior esquerda do gráfico, reflete a elevada sensibilidade observada nas métricas, evidenciando que o modelo é eficiente em identificar transações fraudulentas.

Ao comparar os métodos aplicados na primeira base de dados, observou-se que o forte desbalanceamento entre as classes afeta diretamente o desempenho do modelo de regressão logística. O modelo sem qualquer tratamento manteve alta acurácia, mas apresentou dificuldade em identificar as transações fraudulentas. As técnicas de *oversampling* e *undersampling*, por sua vez, aumentaram a capacidade de detecção de fraudes, elevando o recall e reduzindo a quantidade de falsos negativos. Além disso, o *oversampling* envolve maior custo computacional, enquanto o *undersampling* reduz o tamanho efetivo da base de treinamento. Por

fim, o ajuste do *cutoff* mostrou-se uma alternativa simples e eficiente, permitindo melhorar a detecção sem alterar a estrutura dos dados. Assim, cada abordagem apresentou ganhos e limitações, e a escolha adequada depende do equilíbrio desejado entre sensibilidade e precisão no contexto de aplicação.

4.2 Resultados do segundo conjunto de dados

Para o segundo conjunto de dados, realizou-se inicialmente a divisão da amostra em 70% para o treinamento e 30% para o teste, utilizando amostragem aleatória simples. Com o conjunto de treinamento definido, ajustou-se o modelo de Regressão Logística sem a aplicação de qualquer técnica de balanceamento, de modo a preservar a distribuição original das classes e avaliar o desempenho do modelo diante do desbalanceamento natural da base.

Tabela 15 – Coeficientes estimados do modelo de Regressão Logística (sem balanceamento).

Variável	Estimativa (β)	Erro-padrão	z-value	p-value	Significância
(Intercept)	-10,343	0,052	-198,800	<0,001	***
distance_from_home	0,015	0,000	153,730	<0,001	***
distance_from_last_transaction	0,026	0,000	89,100	<0,001	***
ratio_to_median_purchase_price	0,862	0,003	254,370	<0,001	***
repeat_retailer	-0,621	0,019	-32,910	<0,001	***
used_chip	-1,046	0,015	-71,700	<0,001	***
used_pin_number	-13,454	0,187	-72,110	<0,001	***
online_order	6,631	0,044	149,760	<0,001	***

Fonte: Elaborado pelo autor (2025).

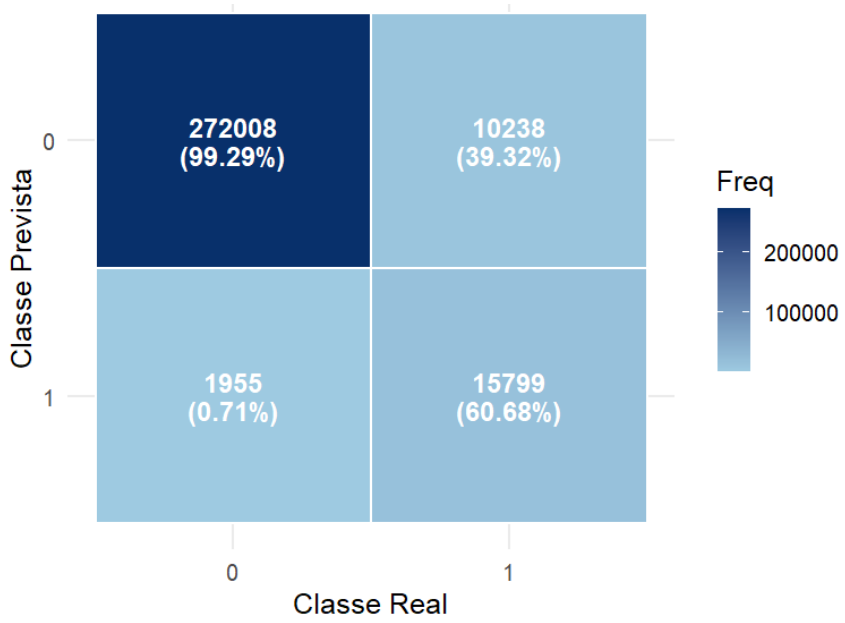
Os resultados obtidos para o modelo ajustado sem qualquer técnica de balanceamento indicam que todas as variáveis incluídas apresentaram significância estatística ao nível de 1%, evidenciando que possuem efeito relevante na probabilidade estimada de fraude.

As variáveis contínuas relacionadas ao comportamento transacional, distance_from_home, distance_from_last_transaction e ratio_to_median_purchase_price, apresentaram coeficientes positivos, indicando que aumentos nessas medidas elevam a chance de uma transação ser fraudulenta. Em particular, o ratio_to_median_purchase_price exibiu o maior coeficiente entre as variáveis contínuas, sugerindo forte influência do valor relativo da compra em relação ao padrão do cliente.

Entre as variáveis categóricas, observa-se que repeat_retailer, used_chip e used_pin_number apresentaram coeficientes negativos, significando que transações realizadas

em estabelecimentos conhecidos, com uso de chip ou com número de pin tendem a reduzir a probabilidade estimada de fraude, comportamento coerente com a literatura sobre segurança transacional. Por outro lado, a variável online_order apresentou coeficiente bastante elevado e positivo, indicando forte associação entre compras online e a ocorrência de fraude, o que também é compatível com o comportamento usual desse tipo de delito. O valor do AIC (188.544) representa o nível de ajuste do modelo à base de treinamento.

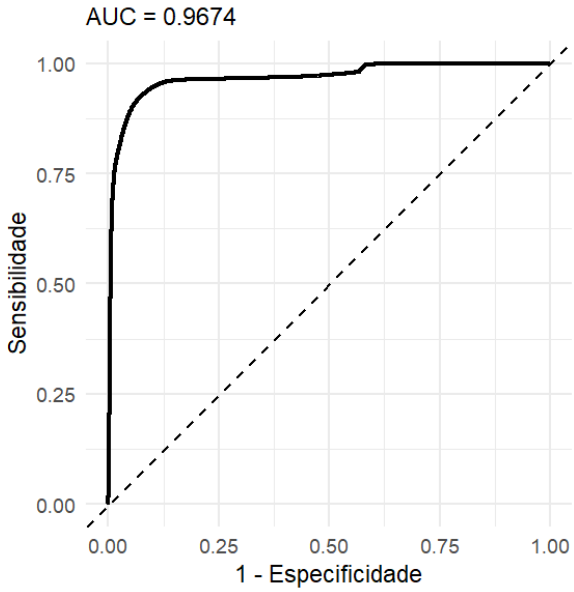
Figura 14 – Matriz de Confusão – Conjunto de Teste (*cutoff* = 0,5).



Fonte: Elaborado pelo autor (2025).

O modelo ajustado com o *cutoff* padrão de 0,5 apresentou desempenho consistente no conjunto de teste. A matriz de confusão mostra que 272.008 transações legítimas (99,29%) foram corretamente classificadas como não fraude, evidenciando a elevada especificidade do modelo. Entre as transações fraudulentas, 15.799 casos (60,68%) foram identificados corretamente, enquanto 10.238 fraudes (39,32%) não foram detectadas, permanecendo como falsos negativos, um comportamento esperado em bases onde a classe positiva é minoritária. A precisão de 88,99% indica que, entre todas as transações previstas como fraude, a maior parte realmente corresponde a casos fraudulentos, reduzindo o risco de alarmes indevidos. O F1-Score de 72,16%, por sua vez, demonstra um equilíbrio adequado entre precisão e sensibilidade, refletindo uma capacidade razoável de identificar fraudes sem gerar excesso de falsos positivos. Por fim, a acurácia global de 95,94% confirma o bom desempenho geral do modelo, embora influenciada pela predominância de transações legítimas.

Figura 15 – Curva ROC – Conjunto de Teste (*cutoff* = 0,5).

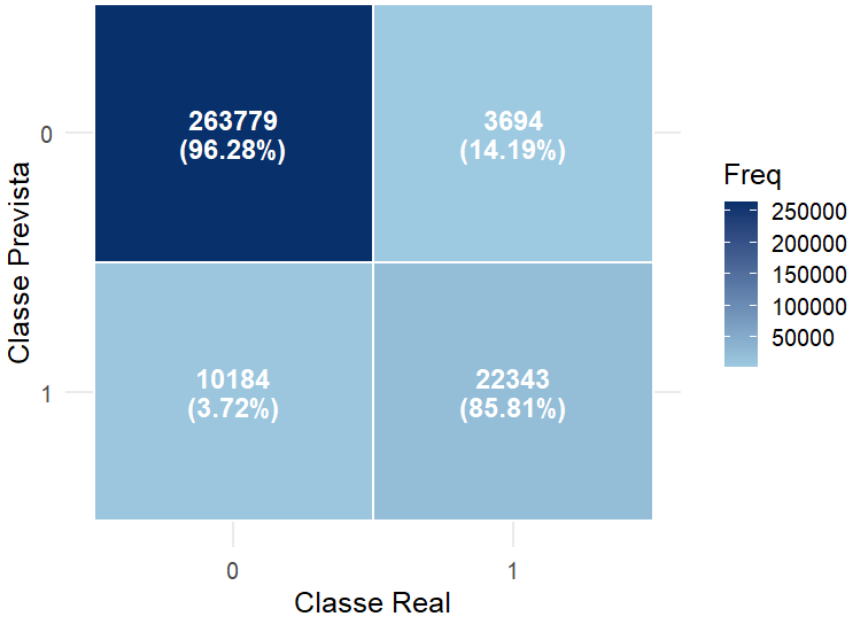


Fonte: Elaborado pelo autor (2025).

Observa-se que a curva apresentada na Figura 15 se mantém bastante elevada desde os primeiros incrementos em 1 – especificidade, atingindo valores próximos de 1,0 de sensibilidade com baixo aumento na taxa de falsos positivos. O AUC de 0,9674 reforça essa conclusão, uma vez que representa uma capacidade discriminativa excelente. Esse valor demonstra que, ao escolher aleatoriamente uma transação fraudulenta e uma não fraudulenta, o modelo tem aproximadamente 96,7% de chance de atribuir maior probabilidade de fraude ao caso realmente fraudulento.

Em seguida, o modelo foi avaliado no conjunto de teste utilizando um *cutoff* de 0,2, permitindo analisar o impacto de um limiar mais baixo sobre a detecção de fraudes.

Figura 16 – Matriz de Confusão – Conjunto de Teste (*cutoff* = 0,2).



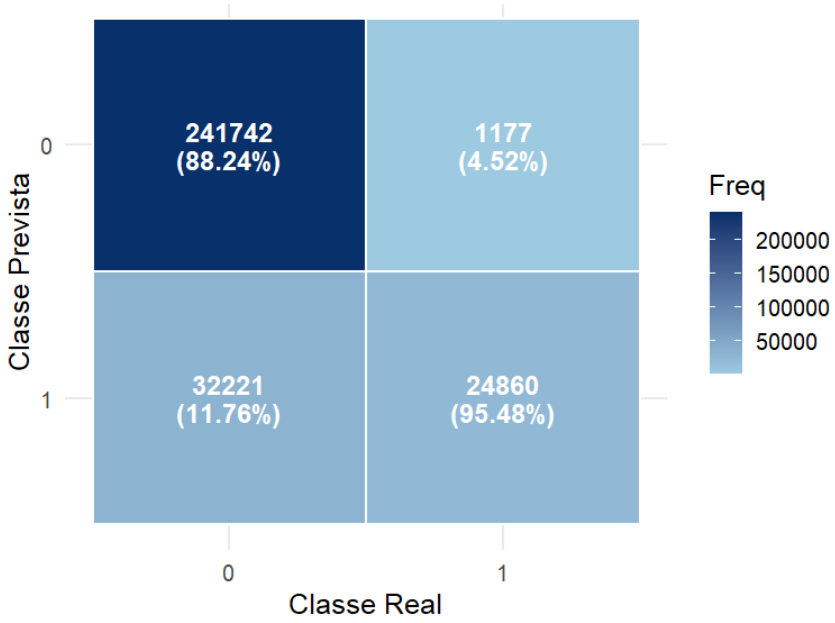
Fonte: Elaborado pelo autor (2025).

A aplicação do *cutoff* igual a 0,2 no conjunto de teste resultou em um desempenho distinto daquele observado com o *cutoff* padrão. A redução do limiar de classificação aumentou a capacidade do modelo em identificar transações fraudulentas, refletida na maior proporção de verdadeiros positivos (85,81%). Como consequência, houve também um incremento do número de falsos positivos, o que reduziu a precisão para 68,69%. Ainda assim, o F1-Score alcançou 76,3%, indicando um equilíbrio mais favorável entre precisão e sensibilidade quando comparado ao *cutoff* de 0,5. A especificidade permaneceu elevada (96,28%), evidenciando que a maior parte das transações legítimas continuou sendo corretamente identificada, apesar do aumento moderado de falsos alarmes. A acurácia geral, de 95,37%, manteve-se em nível semelhante ao observado anteriormente, reforçando que a alteração do *cutoff* não comprometeu substancialmente o desempenho global do modelo.

A seguir, o modelo foi avaliado no conjunto de teste utilizando um *cutoff* igual a 0,08, valor que corresponde à proporção de fraudes observada no conjunto de dados. Esse limiar permite examinar o comportamento do modelo quando o ponto de decisão é ajustado para refletir diretamente a prevalência da classe minoritária. A matriz de confusão, ilustrada na Figura 17, mostra uma sensível ampliação na detecção de fraudes, com 95,48% de verdadeiros positivos, ao mesmo tempo em que se observa um aumento considerável no número de falsos positivos, resultando em uma precisão de 43,55%. A especificidade reduziu-se para 88,24%, refletindo esse maior volume de classificações incorretas para a classe legítima. O F1-Score,

igual a 59,82%, indica uma relação mais equilibrada entre precisão e sensibilidade quando comparado ao *cutoff* anterior, embora a acurácia global tenha diminuído para 88,87% em função do crescimento dos falsos positivos.

Figura 17 – Matriz de Confusão – Conjunto de Teste (*cutoff* = 0,08).



Fonte: Elaborado pelo autor (2025).

Após avaliar o modelo sem balanceamento e com diferentes valores de *cutoff*, aplicou-se a técnica de *oversampling* ao conjunto de treinamento. Nesse procedimento, a classe minoritária foi replicada até atingir a mesma volumetria da classe majoritária, resultando em um conjunto de dados artificialmente balanceado.

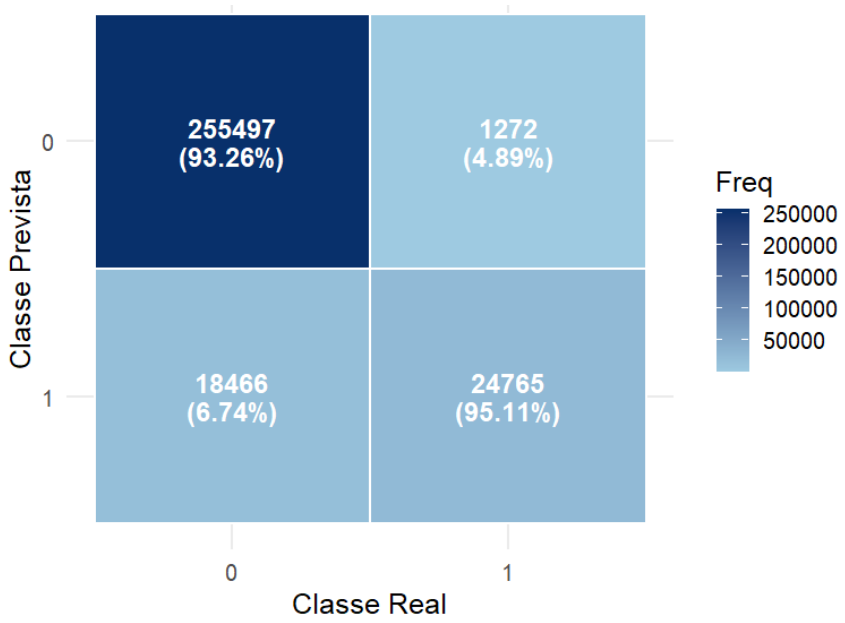
Tabela 16 – Coeficientes estimados do modelo de Regressão Logística (*oversampling*).

Variável	Estimativa (β)	Erro-padrão	z-value	p-value	Significância
(Intercept)	-7,645	0,020	-385,200	<0,001	***
distance_from_home	0,029	0,000	352,300	<0,001	***
distance_from_last_transaction	0,050	0,000	239,800	<0,001	***
ratio_to_median_purchase_price	1,212	0,002	490,600	<0,001	***
repeat_retailer	-1,444	0,012	-124,100	<0,001	***
used_chip	-1,193	0,008	-142,900	<0,001	***
used_pin_number	-9,838	0,057	-171,900	<0,001	***
online_order	5,020	0,016	316,900	<0,001	***

Fonte: Elaborado pelo autor (2025).

As estimativas obtidas após a aplicação do *oversampling* mostram que todas as variáveis permaneceram estatisticamente significativas, reforçando sua relevância para a predição de fraude. O valor do AIC (532.310) é maior que o observado no modelo sem balanceamento, o que é esperado, pois o *oversampling* aumenta o tamanho da base de treinamento e modifica a estrutura estatística dos dados.

Figura 18 – Matriz de Confusão – Conjunto de Teste (*oversampling*).

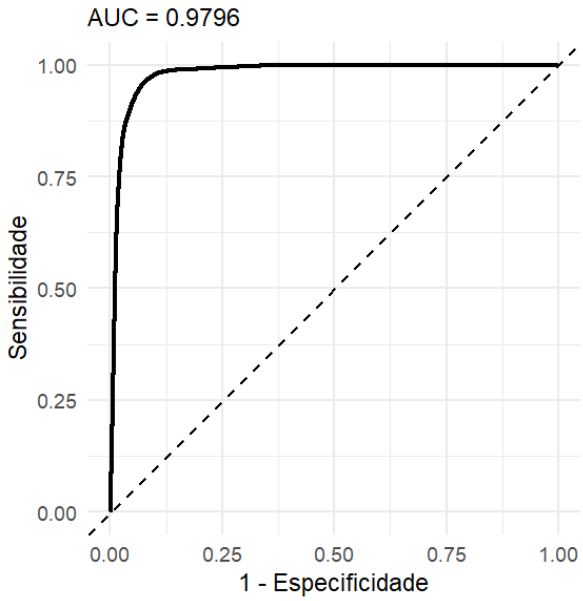


Fonte: Elaborado pelo autor (2025).

A aplicação do modelo treinado com *oversampling* ao conjunto de teste, considerando o *cutoff* padrão de 0,5, resultou em melhorias relevantes na detecção de fraudes em comparação ao modelo ajustado sem balanceamento. Na Figura 18 observa-se que o modelo identificou corretamente 95,11% das transações fraudulentas, refletindo um recall elevado. Esse aumento na capacidade de detecção ocorre em conjunto com uma redução da precisão, que passou para 57,29%, indicando que uma parcela significativa das transações classificadas como fraudulentas não correspondia à classe positiva. A especificidade (93,26%) também diminuiu em relação ao modelo original, o que era esperado devido ao maior número de falsos positivos gerado pela estratégia de balanceamento.

Apesar dessa perda de precisão, o F1-Score atingiu 71,5%. A acurácia geral, de 93,42%, permaneceu em nível adequado, ainda que naturalmente impactada pelo maior volume de classificações incorretas da classe majoritária.

Figura 19 – Curva ROC – Conjunto de Teste (*oversampling*).



Fonte: Elaborado pelo autor (2025).

O valor de AUC = 0,9796, muito próximo de 1, confirma excelente capacidade discriminativa: significa que, ao comparar uma transação fraudulenta com uma legítima selecionadas aleatoriamente, o modelo tem cerca de 98% de chance de atribuir maior probabilidade ao caso realmente fraudulento.

Após os ajustes anteriores, aplicou-se a técnica de *undersampling* ao conjunto de treino, reduzindo a quantidade de observações da classe majoritária até igualá-la ao volume da classe minoritária.

Tabela 17 – Coeficientes estimados do modelo de Regressão Logística (*undersampling*).

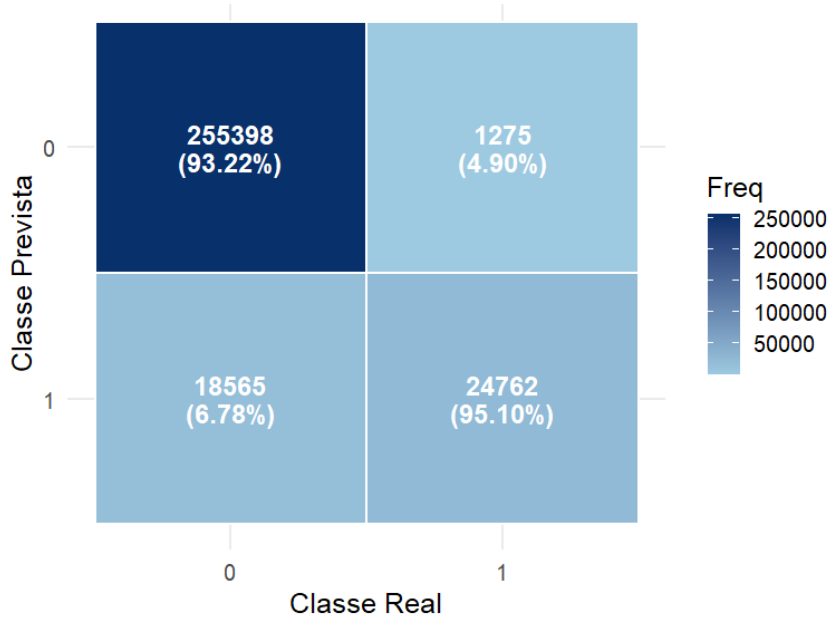
Variável	Estimativa (β)	Erro-padrão	z-value	p-value	Significância
(Intercept)	-7,509	0,063	-119,540	<0,001	***
distance_from_home	0,029	0,000	108,770	<0,001	***
distance_from_last_transaction	0,047	0,001	72,290	<0,001	***
ratio_to_median_purchase_price	1,200	0,008	152,710	<0,001	***
repeat_retailer	-1,408	0,037	-38,040	<0,001	***
used_chip	-1,157	0,027	-43,240	<0,001	***
used_pin_number	-9,888	0,186	-53,050	<0,001	***
online_order	4,910	0,050	97,900	<0,001	***

Fonte: Elaborado pelo autor (2025).

A Tabela 17 exibe as estimativas do modelo de Regressão Logística ajustado após a aplicação do *undersampling*. Observa-se que todas as variáveis, com exceção de nenhuma,

apresentam significância estatística ao nível de 1%, indicando forte evidência de que contribuem para a distinção entre transações legítimas e fraudulentas dentro do cenário balanceado artificialmente. O valor do AIC para o modelo foi de 50.670, indicando boa qualidade de ajuste dentro do cenário do *undersampling*.

Figura 20– Matriz de Confusão – Conjunto de Teste (*undersampling*).



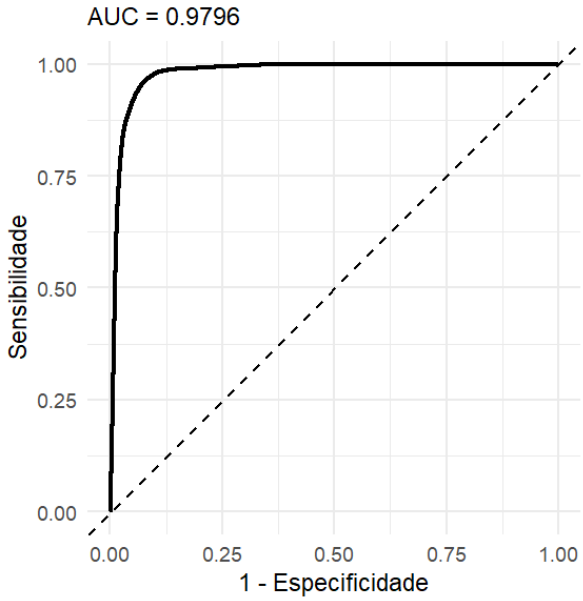
Fonte: Elaborado pelo autor (2025).

A avaliação do modelo ajustado com *undersampling* no conjunto de teste evidencia que o balanceamento artificial reduziu substancialmente o viés em favor da classe majoritária, resultando em um desempenho mais equilibrado entre as métricas de interesse. A matriz de confusão mostra que o modelo identificou corretamente 24.762 transações fraudulentas (95,10%), indicando elevada sensibilidade. Entretanto, esse ganho ocorre acompanhado de um aumento considerável no número de falsos positivos (18.565 casos), o que reduz a capacidade do modelo de distinguir adequadamente fraudes de transações legítimas.

As métricas confirmam esse comportamento: a sensibilidade (Recall) foi de 95,1%, evidenciando excelente detecção das fraudes presentes na base, enquanto a especificidade, de 93,2%, demonstra que o modelo também mantém bom desempenho na identificação de transações legítimas, embora inferior ao observado no modelo sem balanceamento. A precisão foi de 57,15%, refletindo que pouco mais da metade das transações classificadas como fraude de fato pertenciam a essa classe. O F1-Score, que sintetiza precisão e recall, atingiu 71,4%, valor considerado satisfatório e coerente com o objetivo de maximizar a recuperação de fraudes

em contextos de alto risco operacional. A acurácia geral, de 93,39%, permanece elevada, mas sua interpretação deve ser relativizada devido ao desbalanceamento original da base.

Figura 21 – Curva ROC – Conjunto de Teste (*undersampling*).



Fonte: Elaborado pelo autor (2025).

A curva ROC reforça esses resultados, apresentando uma área sob a curva (AUC = 0,9796), o que indica excelente capacidade discriminativa. O formato da curva, com rápida ascensão próxima ao eixo vertical, demonstra que o modelo mantém forte habilidade de separar as duas classes, mesmo quando treinado em um conjunto artificialmente balanceado.

De modo geral, os resultados obtidos para o segundo conjunto de dados evidenciam diferenças importantes entre as estratégias avaliadas. O ajuste apenas do *cutoff* mostrou-se uma alternativa simples e de baixo custo computacional, capaz de alterar significativamente o equilíbrio entre precisão e sensibilidade sem modificar o processo de estimação do modelo.

4 CONCLUSÕES

Este trabalho teve como objetivo avaliar o desempenho do modelo de Regressão Logística na detecção de fraudes em dois conjuntos de dados com níveis distintos de desbalanceamento e investigar o impacto de diferentes estratégias para lidar com essa característica: ajuste do *cutoff*, *oversampling* e *undersampling*. Os objetivos propostos foram integralmente atingidos, permitindo analisar como cada técnica influencia o desempenho do modelo em cenários com proporções distintas de classe minoritária.

A partir das análises realizadas ao longo deste trabalho, foi possível avaliar de forma abrangente o desempenho da regressão logística na detecção de fraudes em dois conjuntos de dados com diferentes níveis de desbalanceamento entre as classes. Em ambos os casos, foram considerados quatro cenários principais: modelo sem qualquer tratamento de desbalanceamento (*cutoff* padrão), modelo com ajuste do *cutoff*, modelo com *oversampling* e modelo com *undersampling*. A avaliação foi conduzida com base em métricas clássicas de classificação, acurácia, sensibilidade (recall), especificidade, precisão, F1-Score, bem como na área sob a curva ROC (AUC) e na interpretação detalhada das matrizes de confusão, com foco na capacidade de captura de fraudes.

No primeiro conjunto de dados, caracterizado por um desbalanceamento extremo entre transações legítimas e fraudulentas, o modelo de regressão logística ajustado sem qualquer intervenção apresentou elevada acurácia e especificidade, mas desempenho limitado na detecção da classe minoritária. A partir daí, os ajustes no *cutoff* e a aplicação das técnicas de *oversampling* e *undersampling* mostraram-se fundamentais para aumentar a sensibilidade, permitindo que uma fração significativamente maior das fraudes fosse identificada. Entre as abordagens avaliadas, o ajuste adequado do *cutoff* destacou-se como a estratégia mais interessante do ponto de vista prático: obteve métricas muito próximas e até melhores às dos modelos reamostrados, inclusive com AUC elevada e melhora consistente na captura de fraudes, sem exigir a modificação da base de dados nem o aumento expressivo do custo computacional.

No caso das técnicas de reamostragem aplicadas a esse primeiro conjunto, tanto o *oversampling* quanto o *undersampling* foram capazes de melhorar a detecção de fraudes em relação ao modelo original, com curvas ROC de excelente desempenho e F1-Scores compatíveis. O *oversampling*, ao ampliar a classe minoritária, favoreceu a aprendizagem de padrões de fraude, mas à custa de maior custo computacional, dado o aumento do volume de

dados no treinamento. Já o *undersampling*, ao reduzir a classe majoritária, produziu métricas de desempenho semelhantes às obtidas com *oversampling*, porém com menor custo computacional, uma vez que o modelo é ajustado em uma amostra consideravelmente menor. Em bases de grande porte, como a primeira utilizada neste estudo, essa diferença de custo torna o *undersampling* uma alternativa atrativa quando se busca um compromisso entre desempenho e eficiência computacional.

No segundo conjunto de dados, caracterizado por um desbalanceamento menos acentuado, observou-se que todas as técnicas, apresentaram comportamento muito semelhante ao observado na primeira base. O ajuste do *cutoff* e as técnicas de reamostragem também produziram melhorias consistentes, mas sem diferenças substanciais entre si, evidenciando que, mesmo com níveis distintos de desbalanceamento, as estratégias mantêm desempenho estável e eficiente.

De forma geral, os resultados indicam que não existe uma "melhor técnica" universal, a escolha da estratégia mais apropriada depende do contexto, do grau de desbalanceamento e das restrições operacionais envolvidas. Quando se opta pelo ajuste do *cutoff* adota-se uma postura mais conservadora, voltada a barrar o maior número possível de fraudes com uma técnica simples, de fácil implementação e baixo custo computacional. Essa abordagem mostrou-se interessante porque, em ambos os conjuntos de dados, foi capaz de produzir métricas comparáveis às obtidas com *oversampling* e *undersampling*, preservando a estrutura original dos dados e facilitando a interpretação do modelo.

Por outro lado, em bases muito grandes, como a primeira analisada, a comparação entre *oversampling* e *undersampling* sugere uma vantagem prática para o *undersampling*: ainda que o *oversampling* apresente desempenho muito bom na curva ROC e na captura de fraudes, o *undersampling* alcançou resultados similares em termos de sensibilidade, precisão e F1-Score, com custo computacional inferior. Dessa forma, quando se considera o volume de dados e a necessidade de treinar modelos de maneira recorrente, o *undersampling* se mostra uma opção interessante para conciliar desempenho e eficiência.

Em síntese, as análises desenvolvidas ao longo do trabalho evidenciam que as diferentes estratégias de tratamento de desbalanceamento, ajuste de *cutoff*, *oversampling* e *undersampling*, são capazes de melhorar a performance da regressão logística em distintos níveis de desequilíbrio entre as classes, tanto em bases altamente desbalanceadas quanto em cenários mais moderados. A escolha final da técnica deve considerar não apenas os indicadores estatísticos, como AUC, F1-Score e matrizes de confusão, mas também aspectos práticos, como custo computacional, volume de dados disponível, tolerância a falsos positivos e exigências

operacionais do sistema de detecção de fraudes. Essa visão integrada, construída a partir das evidências empíricas obtidas nas duas bases analisadas, reforça a importância de combinar rigor estatístico com critérios de viabilidade prática na seleção de modelos e estratégias de balanceamento para aplicações reais de prevenção a fraudes.

REFERÊNCIAS

ASSUNÇÃO, Gabriel O.; IZBICKI, Rafael; PRATES, Marcos O. ***Is Augmentation Effective in Improving Prediction in Imbalanced Datasets?.*** Disponível em: https://jds-online.org/journal/JDS/article/1390/info. Acesso em: 16 out. 2024.

CHAWLA, N. V.; BOWYER, K. W.; HALL, L. O.; KEGELMEYER, W. P. ***SMOTE: Synthetic Minority Over-sampling Technique***. **Journal of Artificial Intelligence Research**, v. 16, p. 321-357, 2002. Disponível em: https://arxiv.org/abs/1106.1813. Acesso em: 8 set. 2024.

FARIAS, Ariston Dias de. ***Análise e construção de variáveis para a melhoria de modelo de regressão logística de concessão de crédito do Banco de Brasília (BRB)***. 2019. Disponível em: https://bdm.unb.br/handle/10483/34258. Acesso em: 15 set. 2024.

GALTON, Francis. ***Regression towards mediocrity in hereditary stature***. The Journal of the Anthropological Institute of Great Britain and Ireland, 1886.

HOSMER, David W.; LEMESHOW, Stanley. ***Applied Logistic Regression***. New York: Wiley-Blackwell, 2000.

HUAYANAY, Alex de la Cruz. ***Modelos de regressão para resposta binária na presença de dados desbalanceados***. 2019. Dissertação (Mestrado em Estatística) – Universidade de São Paulo (USP), Instituto de Ciências Matemáticas e de Computação, e Universidade Federal de São Carlos (UFSCar), Departamento de Estatística. Disponível em: https://repositorio.ufscar.br/handle/ufscar/11103. Acesso em: 10 out. 2024.

IBM. ***Estudo da IBM demonstra impacto de fraudes financeiras e o comportamento das vítimas em diferentes regiões do mundo***. Disponível em: https://brasil.newsroom.ibm.com/2022-08-10-Estudo-da-IBM-demonstra-impacto-de-fraudes-financeiras-e-o-comportamento-das-vitimas-em-diferentes-regioes-do-mundo. Acesso em: 30 out. 2024.

KAGGLE. Disponível em: https://kaggle.com/. Acesso em: 9 set. 2024.

McCULLAGH, Peter; NELDER, John *A*. ***Generalized linear models***. 2. ed. London: Chapman and Hall, 1989.

MONTGOMERY, Douglas C.; PECK, Elizabeth A.; VINING, G. Geoffrey. ***Introduction to linear regression analysis***. 5. ed. Hoboken: John Wiley & Sons, 2012.

SERASA EXPERIAN. ***Relatório de fraude da Serasa Experian: 4 em cada 10 brasileiros já foram vítimas de golpes e preocupação de empresas aumentou 58% em um ano***. Disponível em: https://www.serasaexperian.com.br/sala-de-imprensa/prevencao-a-fraude/relatorio-de-fraude-da-serasa-experian-4-em-cada-10-brasileiros-ja-foram-vitimas-de-golpes-e-preocupacao-de-empresas-aumentou-58-em-um-ano/. Acesso em: 30 out. 2024.

SERASA EXPERIAN. ***Relatório de Identidade Digital e Fraude 2024***. Disponível em: https://tinyurl.com/yf3frwx2. Acesso em: 30 out. 2024.

SILVA, Vitória de Oliveira. ***Detecção de fraudes na utilização de cartões usando a técnica de regressão logística: uma aplicação com dados desbalanceados***. Trabalho de Conclusão de Curso (Bacharelado em Estatística) – Universidade Estadual Paulista (UNESP), Faculdade de Ciências e Tecnologia, Presidente Prudente, 2022. Disponível em:

https://repositorio.unesp.br/server/api/core/bitstreams/468daaec-9842-4d49-a097-2f841975e3ab/content. Acesso em: 10 set. 2024.

UN. ***Objetivos de Desenvolvimento Sustentável***. Disponível em: https://brasil.un.org/pt-br/sdgs. Acesso em 12 set. 2024.