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

VITÓRIA DE OLIVEIRA SILVA

DETECÇÃO DE FRAUDES NA UTILIZAÇÃO DE CARTÕES USANDO A TÉCNICA DE REGRESSÃO LOGÍSTICA: UMA APLICAÇÃO COM DADOS DESBALANCEADOS

PRESIDENTE PRUDENTE
2021/2022

VITÓRIA DE OLIVEIRA SILVA

DETECÇÃO DE FRAUDES NA UTILIZAÇÃO DE CARTÕES USANDO A TÉCNICA DE REGRESSÃO LOGÍSTICA: UMA APLICAÇÃO COM DADOS DESBALANCEADOS

Relatório Final apresentado ao Curso de Graduação em Estatística da FCT/Unesp para aproveitamento na disciplina Trabalho de Conclusão de Curso.

Orientadora: Profa. Dra. Olga L. Anglas R. Tarumoto

Co-orientador: Prof. Dr. Mário H. Tarumoto

PRESIDENTE PRUDENTE

2021/2022

S586d

Silva, Vitória de Oliveira

Detecção de fraudes na utilização de cartões usando a técnica de Regressão Logística: uma aplicação com dados desbalanceados / Vitória de Oliveira Silva. -- Presidente Prudente, 2022

61 p. : il., tabs.

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

Orientadora: Olga L. Anglas R. Tarumoto

Coorientador: Mario H. Tarumoto

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

Sistema de geração automática de fichas catalográficas da Unesp. Biblioteca da Faculdade de Ciências e Tecnologia, Presidente Prudente. Dados fornecidos pelo autor(a).

Essa ficha não pode ser modificada.

TERMO DE APROVAÇÃO

Vitória de Oliveira Silva

DETECÇÃO DE FRAUDES NA UTILIZAÇÃO DE CARTÕES USANDO A TÉCNICA DE REGRESSÃO LOGÍSTICA: UMA APLICAÇÃO COM DADOS DESBALANCEADOS

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:

Orientadora: _____

Profa. Dra. Olga L. Anglas R. Tarumoto

Departamento de Estatística

Coorientador: _____

Prof. Dr. Mário H. Tarumoto

Departamento de Estatística

_____

Profa. Dra. Miriam Rodrigues Silvestre
Departamento de Estatística

_____

Prof. Dr. Fernando Antônio Moala
Departamento de Estatística

Presidente Prudente, 26 de março de 2022.

RESUMO

Com o passar dos anos, o número de fraudes em cartões de crédito e débito vem crescendo e as maneiras como os fraudadores atuam são inovadas diariamente. Isso se dá por conta da ascensão do uso de cartões como forma de pagamento, que acompanha o avanço da tecnologia. Para identificar as fraudes e as atuações por trás delas, é necessário recorrer às estratégias, estudos e técnicas estatísticas que ajudarão a prever e detectar as ocorrências de fraudes. Uma situação comum nestes casos, é que a proporção de fraudes é muito pequena comparada a não fraude, consequentemente, os dados se tornam desbalanceados e necessitam ser tratados. No presente trabalho foram utilizados os métodos de *Oversampling* e *Undersampling* para balancear os dados utilizados e a técnica de Regressão Logística para detectar transações realizadas em cartões de crédito e débito que possuem cunho fraudulento. Para a aplicação, foram utilizados dados sintéticos gerados por um simulador, o qual se baseia em uma amostra de dados reais. Observou-se um severo desbalanceamento dos dados, tendo em vista que apenas 1,3% da base, após os devidos tratamentos, eram transações fraudulentas. Assim, foram feitas três aplicações do modelo, sendo uma com os dados desbalanceados e as outras duas usando os métodos de balanceamento, e notou-se que o *Undersampling* foi o método que apresentou melhores resultados.

Palavras-chaves: Detecção de Fraude, Dados desbalanceados, Regressão Logística

ABSTRACT

Over the years, the number of credit and debit card frauds has been growing and the ways fraudsters act are innovated daily. This is due to the rise in the use of cards as a form of payment, which accompanies the advancement of technology. To identify frauds and the actions behind them, it is necessary to resort to strategies, studies and statistical techniques that will help to predict and detect fraud occurrences. A common situation in these cases is that the proportion of fraud is very small compared to non-fraud, consequently, the data becomes unbalanced and needs to be treated. In the present work, the Oversampling and Undersampling techniques were used to balance the data used and the Logistic Regression technique to detect fraudulent credit and debit card transactions. For the application, synthetic data generated by a simulator were used, which is based on a sample of real data. There was a severe imbalance in the data, considering that only 1.3% of the base, after the proper treatment, were fraudulent transactions. Thus, three applications of the model were made, one with unbalanced data and the other two using balancing methods, and it was noted that Undersampling was the method that presented the best results.

Keywords: Fraud Detection, Unbalanced Data, Logistic Regression

LISTA DE FIGURAS

FIGURA 1 – PRIMEIRA VERSÃO EM PAPEL DO CARTÃO *DINERS CLUB* 13

FIGURA 2 – FLUXO DE UMA TRANSAÇÃO NO CARTÃO 14

FIGURA 3 – FUNÇÃO DO MODELO DE REGRESSÃO LOGÍSTICA 21

FIGURA 4 – COMPORTAMENTOS DA CURVA ROC 30

FIGURA 5 – AUC 31

FIGURA 6 – *UNDERSAMPLING* 35

FIGURA 7 – MÉTODO SMOTE 36

FIGURA 8 – *OVERSAMPLING* 36

FIGURA 9 - GRÁFICO DE BARRAS DA VARIÁVEL “TYPE” 41

FIGURA 10 – DISTRIBUIÇÃO DAS VARIÁVEIS DUMMYS “TYPE” 43

FIGURA 11 – CURVA ROC 47

FIGURA 12 – DISTRIBUIÇÃO DAS TRANSAÇÕES FRAUDULENTAS E NÃO FRAUDULENTAS PRÉ E PÓS BALANCEAMENTO NEARMISS 48

FIGURA 13 – CURVA ROC NEARMISS 50

FIGURA 14 – DISTRIBUIÇÃO DAS TRANSAÇÕES FRAUDULENTAS E NÃO FRAUDULENTAS PRÉ E PÓS BALANCEAMENTO SMOTE 52

FIGURA 15 – CURVA ROC SMOTE 54

LISTA DE TABELAS

TABELA 1 - POSSÍVEIS VALORES DO MODELO DE REGRESSÃO LOGÍSTICA PARA VARIÁVEL INDEPENDENTE DICOTÔMICA. 19

TABELA 2 - MATRIZ DE CONFUSÃO 28

TABELA 3 - MATRIZ DE CONFUSÃO PARA DETECÇÃO DE TRANSAÇÕES FRAUDULENTAS ... 29

TABELA 4 – DISTRIBUIÇÃO DOS CLIENTES EM RELAÇÃO AO NÚMERO DE DEPENDENTES .. 32

TABELA 5 – DISTRIBUIÇÃO DOS CLIENTES EM RELAÇÃO A RENDA MENSAL..... 32

TABELA 6 – DISTRIBUIÇÃO DOS CLIENTES EM RELAÇÃO AO VÍNCULO EMPREGATÍCIO..... 32

TABELA 7 – ESTIMATIVA DOS PARÂMETROS E TESTE WALD 33

TABELA 8 - DISTRIBUIÇÃO DA VARIÁVEL “ISFRAUD”..... 39

TABELA 9 - DISTRIBUIÇÃO DA VARIÁVEL “ISFLAGGEDFRAUD” 40

TABELA 10 - DISTRIBUIÇÃO DA VARIÁVEL “TYPE” 40

TABELA 11 - MEDIDAS DE POSIÇÃO DA VARIÁVEL “AMOUNT”..... 41

TABELA 12 - DISTRIBUIÇÃO DA VARIÁVEL “TYPE” 42

TABELA 13 – ESTIMATIVAS DOS PARÂMETROS RL COM DADOS DESBALANCEADOS 45

TABELA 14 – MATRIZ DE CONFUSÃO RL COM DADOS DESBALANCEADOS 46

TABELA 15 – MÉTRICAS DE DESEMPENHO DAS PREDIÇÕES DO MODELO 46

TABELA 16 – ESTIMATIVAS DOS PARÂMETROS NEARMISS..... 48

TABELA 17 – MATRIZ DE CONFUSÃO NEARMISS..... 49

TABELA 18 – MÉTRICAS DE DESEMPENHO DAS PREDIÇÕES DO MODELO NEARMISS 49

TABELA 19 – ODDS RATIO NEARMISS 51

TABELA 20 – ESTIMATIVAS DOS PARÂMETROS SMOTE..... 53

TABELA 21 – MATRIZ DE CONFUSÃO SMOTE..... 53

TABELA 22 – MÉTRICAS DE DESEMPENHO DAS PREDIÇÕES DO MODELO SMOTE 54

TABELA 23 – ODDS RATIO SMOTE 55

SUMÁRIO

1 INTRODUÇÃO 10

2 HISTÓRICO DOS CARTÕES 13

2.1 FUNCIONAMENTO DOS CARTÕES 14

3 FRAUDES 15

3.1 TIPOS DE FRAUDES 15

3.2 PREVENÇÃO E DETECÇÃO DE FRAUDE 16

4 REGRESSÃO LOGÍSTICA 18

4.1 ESTIMATIVAS DE MÁXIMA VEROSSIMILHANÇA 21

4.2 TESTES DE AJUSTE DO MODELO 23

4.2.1 Teste da Razão de Verossimilhança 23

4.2.2 Teste de Wald 24

4.2.3 Teste de Score 25

4.3 INTERVALO DE CONFIANÇA 25

4.4 SELEÇÃO DAS VARIÁVEIS 26

4.5 AVALIAÇÃO DO DESEMPENHO DE PREDIÇÃO DO MODELO 27

4.5.1 Matriz de Confusão 27

4.5.2 Curva ROC e AUC 29

4.6 EXEMPLO DE APLICAÇÃO 31

4.6.1Classificação de inadimplência 31

5 DADOS DESBALANCEADOS 34

5.1 MÉTODO DE SUBAMOSTRAGEM OU *UNDERSAMPLING* 34

5.2 MÉTODO DE SOBREAMOSTRAGEM OU *OVERSAMPLING* 35

5.3 MÉTODO DE ADAPTAÇÃO DE ALGORITMOS 37

6 APLICAÇÃO 38

6.1 DESCRIÇÃO DOS DADOS E ANÁLISES DESCRITIVAS 38

6.1.1 Variável Resposta "isFraud" 39

6.1.2 isFlaggedFraud 39

6.1.3 type 40

6.1.4 amount 41

6.2 TRATAMENTO DOS DADOS 42

6.3 APLICAÇÕES DO MODELO E RESULTADOS 44

6.3.1 Regressão Logística com os dados desbalanceados 45

6.3.2 Regressão Logística após balancear os dados utilizando a técnica de Undersampling 47

6.3.3 Regressão Logística após balancear os dados utilizando a técnica de Oversampling 52

7 CONCLUSÃO FINAL 57

REFERÊNCIAS 59

1 Introdução

Na década passada, o cartão de crédito se tornou o principal meio de pagamento nacional e internacional, uma vez que com o avanço da tecnologia e da economia, esse item tem aumentado o poder de compra e se tornado indispensável no cotidiano de muitas pessoas (BOLTON e HAND, 2002). Segundo Beraldi (2014), o portador possui maior facilidade ao realizar suas compras pois é atribuído ao cliente um limite pelo emissor, que muitas vezes ultrapassa sua renda mensal, além de possibilitar o parcelamento em compras com valores mais elevados.

Além do cartão de crédito, também há o cartão de débito que possui grande importância pois está diretamente interligado ao saldo que o portador possui em sua instituição financeira, uma vez que ao utilizar o cartão, o dinheiro é debitado no mesmo instante da conta, descartando a necessidade de ter cédulas em mãos.

Segundo a Associação Brasileira de Cartões de Crédito e Serviços (ABECS), em 2020 foi transacionado R$1,18 trilhões no cartão de crédito, apresentando um aumento de 2,6% em relação a 2019. Já no cartão de débito, foi transacionado R$762,4 bilhões apresentando um aumento de 14,8% quando comparado ao ano anterior.

Através destes meios eletrônicos, é possível realizar transações presenciais e não presenciais (*e-commerce*). Em decorrência da pandemia houve um crescimento muito expressivo nas compras virtuais, e segundo a ABECS, no Brasil houve um aumento de 19,8% de compras realizadas no e-commerce em 2020 quando comparado a 2019, concentrando R$388,2 bilhões.

Esse aumento do fluxo de compras em cartões, dá margem ao aumento no número de fraudes, que por definição do dicionário Michaelis é o “Ato de má-fé que tem por objetivo fraudar ou ludibriar alguém”. Cada fraude causa uma situação muito complexa, por isso é necessário que aconteçam diversos processos para manter um equilíbrio entre o volume de fraudes e o impacto na receita do estabelecimento atingido (BERALDI, 2014).

Inicialmente, é de suma importância que os estabelecimentos tenham uma espécie de manual com orientações que conscientizem o cliente a proteger seus dados, como por exemplo, evitar acessar links estranhos, usar senhas fortes, não

compartilhar a senha com terceiros, entre outras. Essas orientações dificultam o roubo dos dados pelos fraudadores e diminui a possibilidade de ocorrência de fraudes.

No entanto, todos os cuidados citados não são suficientes para anular a existência de fraudes. Os fraudadores sempre encontram maneiras de conseguir os dados, seja persuadindo a vítima ou conseguindo de alguma forma desconhecida, roubar seus dados e realizarem a fraude. Logo, existe a necessidade de ações voltadas a estudos e identificação dos perfis de usuários que causaram fraude com o intuito de evitar e barrar futuras transações por sujeitos com perfil semelhante ao do fraudador.

Tais ações devem ser tomadas por parte dos emissores, portadores, adquirentes, estabelecimentos e bandeiras. Os emissores são as instituições financeiras que emitem os cartões; os portadores são aqueles que possuem e realizam compras em seu cartão; os adquirentes, também chamados de credenciadores, são responsáveis pela transição do dinheiro do portador para o estabelecimento onde o cartão será utilizado além de instalar e realizar manutenções nas maquinetas[1]; os estabelecimentos são empresas que são credenciadas e aceitam pagamentos através de cartão de crédito e débito; as bandeiras possuem a função de fazer o intermédio entre emissores e adquirentes dos estabelecimentos comerciais que aceitarão cartão como forma de pagamento.

Para mitigar uma fraude, é preciso seguir duas vertentes: prevenção e detecção de fraude. A prevenção visa utilizar sistemas e estratégias para evitar que a fraude aconteça, por outro lado, a detecção visa utilizar estudos e criar ferramentas capazes de identificar com agilidade a ocorrência de fraude.

A cada dia observa-se a evolução dos estudos para detectar uma fraude, no entanto, os fraudadores conseguem acompanhar tal evolução e desenvolvem novas técnicas para não serem barrados novamente, atribuindo assim um grande desafio aos profissionais que precisam constantemente buscar meios de combater as fraudes.

Uma das possíveis ferramentas estatísticas que podem ser utilizadas para diferenciar o perfil de fraudadores e não fraudadores, é a Regressão Logística, que é uma das mais indicadas para encontrar a probabilidade de uma transação ser fraudulenta ou não, e ainda segundo Agresti (2002), é o modelo mais importante para dados cuja resposta é categórica.

---

[1] Também denominada como *POS (point of sales)*, é o terminal que captura os dados do cartão via chip ou tarja magnética (OLIVEIRA, 2015)

Na maioria dos casos em que se trabalha com cenários de fraude, a quantidade de transações fraudulentas é muito inferior à quantidade de transações não fraudulentas. Em decorrência desse comportamento, os dados ficam desbalanceados e atrapalham a detecção, sendo necessário utilizar métodos para balanceá-los antes de aplicar o modelo. Alguns métodos bastante utilizados e efetivos, são os métodos de subamostragem também conhecido como *undersampling* e sobreamostragem ou *oversampling*. A subamostragem reduz a quantidade de transações efetivas até que a proporção de transações fraudulentas e não fraudulentas seja de 1:1. Já o método de sobreamostragem faz o processo reverso, aumenta a quantidade de transações efetivas até que a proporção seja de 1:1.

O atual trabalho tem como principal objetivo, aplicar o modelo de Regressão Logística para detectar transações em cartão de crédito e débito com características fraudulentas, a fim de identificar variáveis que tenham relevância para o processo de fraude. Tem o intuito ainda, de, estudar o comportamento de transações fraudulentas, utilizar os métodos *undersampling* e *oversampling* para identificar o impacto no modelo ao balancear os dados e realizar um estudo mais aprofundado da técnica de Regressão Logística.

Para o desenvolvimento do trabalho, está sendo utilizado uma base de dados sintéticos retirada do site *Kaggle.com,* que utiliza um simulador para gerar as informações, no entanto, mantendo a estrutura de dados o mais próximo possível da realidade.

Nos capítulos subsequentes é apresentado o embasamento teórico explorando os conceitos referentes aos cartões e as fraudes associadas a eles, juntamente com conceitos e aplicações do modelo de Regressão Logística.

Posteriormente, é realizada a descrição dos dados e análises descritivas a fim de ter conhecimento das variáveis que estão sendo utilizadas. Na sequência, são realizados alguns tratamentos dos dados e por fim são apresentadas as aplicações do modelo de Regressão Logística com os dados desbalanceados e após balanceá-los aplicando as técnicas de subamostragem e sobreamostragem.

2 Histórico dos Cartões

O conceito de cartão de crédito se iniciou na década de 1920 nos Estados Unidos, quando varejistas criaram a logística de cartões de crédito para os clientes mais ricos, pois estes sempre pagavam suas contas em dia, com o intuito de aumentar as vendas e simultaneamente fidelizá-los. Nessa época, os cartões eram emitidos em um pedaço de metal pelos próprios varejistas.

Em 1950 foi criado o cartão *Diners Club*, a primeira versão em papel. Essa versão possuía o nome do portador e dos estabelecimentos filiais, e inicialmente seria usado em restaurantes com projeto de ser usado no futuro em estabelecimento para comprar bens. Quem possuía o cartão recebia as cobranças mensalmente ou anualmente e pagavam uma taxa anual pelo uso.

Figura 1 – Primeira versão em papel do cartão *Diners Club*



Fonte: http://www.fraudes.org/images/Diners_1950_card.jpg.

Em 1954, o empresário tcheco Hanus Tauber comprou a franquia da *Diners Club*, fez sociedade com o empresário Horácio Klabin e foi o iniciador dos cartões de crédito no Brasil.

Na década de 1980, surgiu o primeiro cartão de débito no Brasil lançado pelo Bradesco.

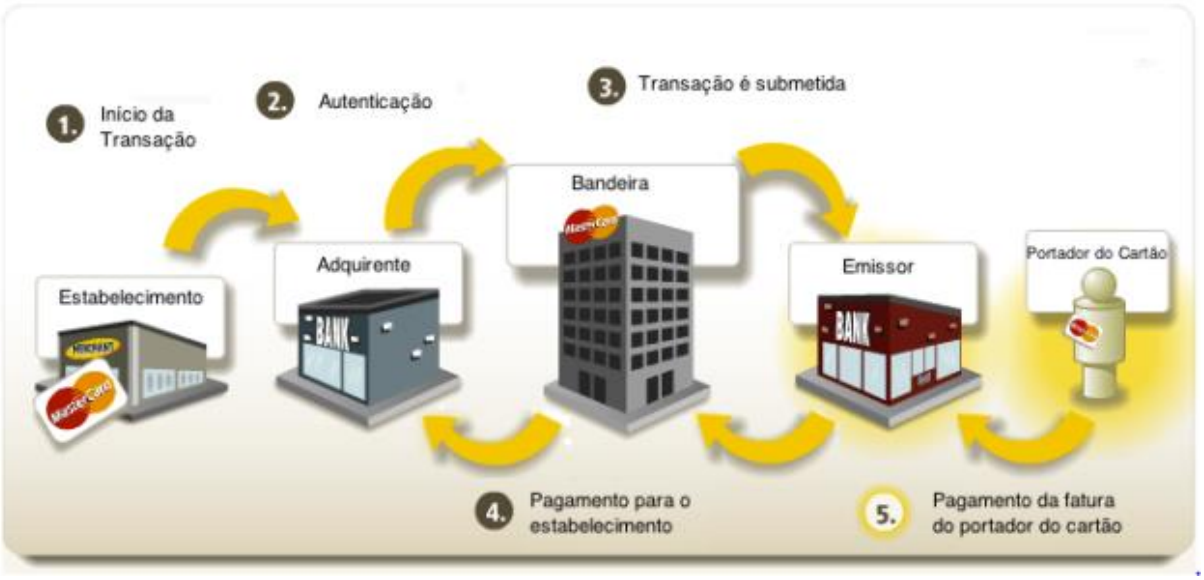
Desde então, com o passar dos anos, os métodos foram se aperfeiçoando, empresas e bandeiras foram surgindo e a expansão do uso de cartões aconteceu gradativamente.

2.1 Funcionamento dos cartões

Para que uma transação seja efetuada com sucesso, existem operadores por trás com papel fundamental nessa efetuação. Os operadores são: emissores, portadores, adquirentes, estabelecimentos e bandeiras.

Dada a existência de todos operadores, ocorre um fluxo os envolvendo para a realização da transação através dos cartões. Esse fluxo se inicia quando o portador realiza uma compra, e no momento em que ele insere o cartão na maquineta, o estabelecimento detecta a transação, em seguida a transação é submetida para a bandeira, a qual faz uma análise antes de aprová-la e então envia para o emissor que faz o pagamento para o adquirente e o mesmo faz o pagamento para o estabelecimento, assim o portador consegue finalizar a compra e recebe em sua fatura o valor transacionado.

Figura 2 – Fluxo de uma transação no cartão



Fonte: Oliveira (2016, p.6)

3 Fraudes

Com o crescimento do uso de cartões de crédito e débito, e o avanço da tecnologia, o número de fraudes vem crescendo exponencialmente e as maneiras como elas vêm acontecendo são renovadas rapidamente. Desse modo, a cada mês se escuta falar em uma nova forma de fraude e em muitos casos é totalmente desconhecida por quem se torna vítima.

A incidência de fraudes atinge diretamente os emissores e os estabelecimentos, que arcam com as despesas e perdem milhões de reais por ano como consequência dessa incidência. Dessa forma, é importante que os emissores busquem incessantemente por novas estratégias e ações, visando diminuir a ocorrência de transações fraudulentas.

3.1 Tipos de fraudes

Existem diversos tipos de fraude, mas há alguns que são mais frequentes no mercado, que são eles: fraude efetiva, fraude amigável, autofraude, extravio, invasão de conta, cartão perdido ou roubado e falsidade ideológica.

- **Fraude efetiva:** é cometida por fraudadores com diversas habilidades e experiências, os quais agem sozinhos e/ou em quadrilha, efetuando o roubo de dados pessoais do portador, conseguindo assim efetuar diversas transações fraudulentas.
- **Fraude amigável:** quando alguém efetua uma compra ou transaciona no cartão de um conhecido, sem o consentimento dele.
- **Autofraude:** é realizada pelo próprio portador do cartão, ou seja, ele faz uma compra e ao ter o produto em mãos, aciona o emissor e alega que não realizou aquela compra e então é ressarcido.
- **Extravio:** também conhecido como *golpe do motoboy* é quando o cartão de crédito é roubado no trajeto da entrega para o portador.
- **Invasão de conta**: se dá quando o fraudador consegue roubar informações e senhas pessoas da vítima e a partir daí consegue realizar pagamentos, fazer transferências, solicitar empréstimos e realizar diversas fraudes no nome da pessoa.

- **Cartão perdido ou roubado:** ocorre quando a pessoa perde seu cartão ou é roubado, e então quem consegue ter acesso a esse cartão realiza diversas fraudes com ele.
- **Falsidade ideológica:** é a realização de fraudes após a solicitação ou alteração de um cartão com os dados de alguém, sem seu consentimento.

3.2 Prevenção e detecção de fraude

Para combater fraudes, é necessário recorrer à duas vertentes que fazem grande diferença quando há bons investimentos e boas execuções: prevenção e detecção de fraude. Prevenção e detecção de fraude estão diretamente relacionadas, uma vez que se a prevenção falhar, a detecção entra em ação (BERALDI, 2014). No entanto, é necessário ter conhecimento da diferença entre os dois métodos (BOLTON e HAND, 2002).

Prevenção à fraude tem como objetivo evitar que a fraude aconteça através de medidas e estratégias, como construção de regras, grupos de trabalhos, construção de *onboardings*[2] mais consistentes etc. Por outro lado, a detecção de fraude tem como objetivo identificar a fraude o mais rápido possível e barrar a ocorrência de fraudes com o mesmo comportamento através de ações e ferramentas estatísticas.

Trabalhar com detecção de fraude implica em uma grande dificuldade, pois a proporção de transações fraudulentas comparadas às efetivas são muito pequenas, isto é, existem muitas transações não fraudulentas para cada uma fraudulenta (BOLTON e HAND, 2002), com isso é preciso analisar o comportamento das transações, e ao identificar as que possuem cunho fraudulento, realizar um trabalho minucioso para afirmar quais daquelas apontadas de fato são fraudes, com o intuito de ser eficaz e evitar atingir os bons clientes.

Existem diversos métodos estatísticos, os quais são utilizados para detecção de fraude. Esses métodos visam estimar a probabilidade de uma transação ser considerada como fraude e são divididos em duas classificações: método supervisionado e não supervisionado. O método supervisionado consiste em utilizar dados de fraudes e não fraudes, com o intuito de construir modelos que possam ser utilizados para verificar as chances das novas transações serem fraudulentas ou não.

[2] Primeiros contatos entre cliente e empresa, visando a melhor e mais segura experiência para o cliente.

Por outro lado, o método não supervisionado busca informações pessoais dos clientes que se comportam de maneira anormal, sendo um *outlier* com potencial de ser uma fraude (BOLTON e HAND, 2002).

Dado isso, inúmeras técnicas vêm sendo utilizadas no mercado como forma de detectar uma fraude. Técnicas de *clustering* ou agrupamento (SANTIAGO, 2014) e ferramentas de controle de qualidade (BERALDI, 2014) são bons exemplos do método não supervisionado. Em contrapartida, técnicas como Redes Neurais, Redes Bayesianas, Árvore de decisão e Regressão Logística são muito utilizadas no método supervisionado (SANTIAGO, 2014).

4 Regressão Logística

Os métodos existentes de regressão têm o propósito de construir um modelo para entender a relação entre um conjunto de variáveis independentes, também chamadas de variáveis explicativas, com uma variável dependente, também conhecida como variável resposta.

Os modelos mais comuns e utilizados são os de Regressão Linear e Regressão Logística. No modelo de Regressão Linear a variável resposta é continua e os parâmetros são estimados através do método de mínimos quadrados ou de máxima verossimilhança. No modelo Logístico, a variável resposta é categórica e os parâmetros são estimados pelo método de máxima verossimilhança. Neste último modelo não é exigida a normalidade dos erros e a homogeneidade da variância, como no caso do modelo de regressão linear (OLIVEIRA, 2015).

O método de Regressão logística é utilizado frequentemente para problemas de classificação, onde a variável resposta tem como característica ser dicotômica ou multinomial e as variáveis explicativas assumem resultados qualitativos e quantitativos.

Desta maneira, é uma técnica que atende as necessidades da vertente de detecção de fraudes, pois quando a variável resposta é binária e pode assumir as seguintes classificações: transações fraudulentas e transações efetivas (ou não fraudulenta) e as variáveis explicativas podem ser tanto qualitativas quanto quantitativas. Assim, poderá trazer bons resultados.

Na Regressão Logística Binária, a variável resposta Y segue distribuição Bernoulli, onde:

$$\pi(x) = P(Y = 1|X = x) = 1 - P(Y = 0|X = x) \quad (4.1)$$

e assume dois valores:

Sucesso: Representado pelo valor “1”, é a ocorrência do evento de interesse, ou seja, para o estudo em questão, a transação ser fraudulenta.

Fracasso: Representado pelo valor “0”, é a ocorrência do evento, o qual não é de interesse, isto é, para o estudo em questão, a transação efetiva (ou não fraudulenta).

Para aplicar a técnica de Regressão Logística, é de suma importância a média condicional, isto é, o valor médio da variável resposta em relação ao valor de cada variável explicativa. Ela varia de 0 a 1, ou seja, $0 \leq E(Y|x) \leq 1$ e é representada por:

$$E(Y|x) = \beta_0 + \beta_1 x \quad (4.2)$$

onde:

β representa os coeficientes;

x representa a variável explicativa ou em casos que há mais de uma, representará as variáveis explicativas X_i, sendo $i = 1,2,\dots,p$.

Quando $\pi(x) = E(Y|x)$, o modelo de regressão é dado por:

$$\pi(x) = \frac{e^{(\beta_0+\beta_1 x)}}{1+e^{(\beta_0+\beta_1 x)}}. \quad (4.3)$$

Os coeficientes β_0, β_1 são estimados pelo método de máxima verossimilhança através do conjunto de dados, e os valores de $\hat{\beta}_0$,$\hat{\beta}_1$ maximizam o logaritmo da função de máxima verossimilhança.

Para interpretar os parâmetros da Regressão Logística, é feito uma comparação da probabilidade de sucesso com a probabilidade de fracasso através da função *odds ratio* (AGRESTI, 2007), também chamada de razão das chances, que se dá por:

$$g(x) = \frac{\pi(x)}{1-\pi(x)} = \frac{\frac{e^{(\beta_0+\beta_1 x)}}{1+e^{(\beta_0+\beta_1 x)}}}{1-\frac{e^{(\beta_0+\beta_1 x)}}{1+e^{(\beta_0+\beta_1 x)}}} = \frac{\frac{e^{(\beta_0+\beta_1 x)}}{1+e^{(\beta_0+\beta_1 x)}}}{\frac{1}{1+e^{(\beta_0+\beta_1 x)}}} = e^{\beta_0+\beta_1 x}. \quad (4.4)$$

Quando a variável independente é dicotômica, tem-se que se:

$x = 1$, então $\frac{\pi(1)}{1-\pi(1)}$,

$x = 0$, então $\frac{\pi(0)}{1-\pi(0)}$.

E portanto, pode-se seguir a tabela 1.

Tabela 1 - Possíveis valores do Modelo de Regressão Logística para variável independente dicotômica.

	Variável independente X	
Variável Y	$X = 1$	$X = 0$
$Y = 1$	$\pi(1) = \frac{e^{(\beta_0+\beta_1)}}{1+e^{(\beta_0+\beta_1)}}$	$\pi(0) = \frac{e^{\beta_0}}{1+e^{\beta_0}}$
$Y = 0$	$1-\pi(1) = \frac{1}{1+e^{(\beta_0+\beta_1)}}$	$1-\pi(0) = \frac{1}{1+e^{(\beta_0+\beta_1)}}$
Total	1	1

Fonte: Hosmer e Lemeshow (2000)

Ao aplicar o logaritmo, denota-se:

$$g(1) = \ln \frac{\pi(1)}{1-\pi(1)} \tag{4.5}$$

e

$$g(0) = \ln \frac{\pi(0)}{1-\pi(0)}. \tag{4.6}$$

O *odds ratio* (razão das chances) é denotado por Ψ, cujo logaritmo natural é dado por:

$$\ln(\Psi) = \ln\left[\frac{\pi(1)/1-\pi(1)}{\pi(0)/1-\pi(0)}\right] = g(1) - g(0). \tag{4.7}$$

Assim, tem-se que o *odds ratio* (razão das chances) para variável independente dicotômica é:

$$\Psi = \frac{\dfrac{e^{(\beta_0+\beta_1)}}{1+e^{(\beta_0+\beta_1)}} \Big/ \dfrac{1}{1+e^{(\beta_0+\beta_1)}}}{\dfrac{e^{\beta_0}}{1+e^{\beta_0}} \Big/ \dfrac{1}{1+e^{(\beta_0+\beta_1)}}} = \frac{e^{(\beta_0+\beta_1)}}{e^{\beta_0}} = e^{\beta_1} \tag{4.8}$$

e o logaritmo é denota por:

$$\ln(\Psi) = \ln\left[e^{\beta_1}\right] = \beta_1 \tag{4.9}$$

Ψ normalmente tem distribuição assimétrica. A assimetria amostral de $\Psi = 0{,}5$ ocorre pois ela varia de 0 a ∞, apresentando valor neutro apenas quando $\Psi = 1$. Para grandes amostras, Ψ terá distribuição normal, logo, o intervalo de confiança do *odds ratio* (razão das chances) de $100\ (1-\alpha)\%$ é denotado por:

$$e^{[\hat{\beta}_1 \pm z_{1-\alpha/2}\widehat{SE}(\hat{\beta}_1)} \tag{4.10}$$

onde $\widehat{SE}(\hat{\beta}_p) = \sqrt{\left[\hat{\sigma}^2(\hat{\beta}_p)\right]}$

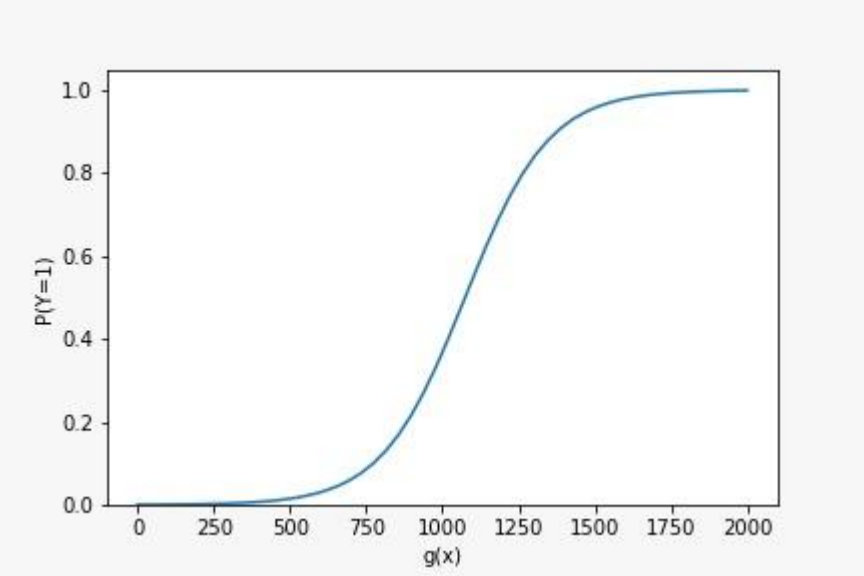
A curva para $\pi(x)$ segue um formato de “S” na Regressão Logística e é determinada pelo parâmetro β(AGRESTI, 2007). Assim, se

$\beta > 0$ a curva irá subir

$\beta < 0$ a curva irá descer

$\beta = 0$ a curva se tornará uma linha reta horizontal

Figura 3 – Função do modelo de Regressão Logística



Fonte: Própria (2021)

4.1 Estimativas de Máxima Verossimilhança

No modelo de Regressão Logística, a estimação dos parâmetros é feita através do método da máxima verossimilhança. Os valores de $\hat{\beta}_i$ são aqueles que maximizam o logaritmo da função de verossimilhança e representam os estimadores de β_i.

Ao considerar uma amostra independente (x_i, y_i) de tamanho n, onde x_i representa o valor da *i*-ésima observação da variável independente e y_i representa o valor da variável resposta dicotômica, a $P(Y = 1|x)$ é uma distribuição Bernoulli com probabilidade de sucesso de $\pi(x)$.

Se $y_i = 1$, a função de verossimilhança será contribuída por $\pi(x_i)$, enquanto se $y_i = 0$, a função de verossimilhança será contribuída por $1 - \pi(x_i)$. A quantidade $\pi(x_i)$ denota o valor de $\pi(x)$ calculado em x_i. Assim, a função de verossimilhança expressa para (x_i, y_i) é dada por:

$$f(y_i\, \pi(x_i)) = \pi(x_i)^{y_i}[1 - \pi(x_i)]^{1-y_i}. \tag{4.11}$$

Dado que as observações são independentes, a função de verossimilhança é expressa por:

$$L(\beta) = \prod_{i=1}^{n} f\big(y_i\, \pi(x_i)\big) = \prod_{i=1}^{n} \pi(x_i)^{y_i}[1 - \pi(x_i)]^{1-y_i}. \tag{4.12}$$

Para determinar os parâmetros $\hat{\beta}$ que maximizam $L(\beta)$ é necessário aplicar o logaritmo na expressão (4.12), assim tem-se:

$$l(\beta) = \ln[L(\beta)] = \sum_{i=1}^{n} \{y_i \ln[\pi(x_i)] + (1 - y_i) \ln[1 - \pi(x_i)]\}, \qquad (4.13)$$

em seguida é necessário derivar a função (4.13) em relação aos elementos do vetor β, os quais são funções dos logaritmos presentes na equação. Sabe-se que:

$$\pi(x) = \frac{e^{(\beta_0 + \beta_1 x)}}{1 + e^{(\beta_0 + \beta_1 x)}}$$

então, substituindo $\pi(x)$ em (4.13), tem-se:

$$l(\beta) = \sum_{i=1}^{n} y_i \ln\left(\frac{e^{(\beta_0 + \beta_1 x_i)}}{1 + e^{(\beta_0 + \beta_1 x_i)}}\right) + (1 - y_i) \ln\left[1 - \left(\frac{e^{(\beta_0 + \beta_1 x_i)}}{1 + e^{(\beta_0 + \beta_1 x_i)}}\right)\right]$$

$$= \sum_{i=1}^{n} \{y_i \left[\ln\left(e^{(\beta_0 + \beta_1 x_i)}\right) - \ln\left(1 + e^{(\beta_0 + \beta_1 x_i)}\right] + \left[-\ln(1 + e^{(\beta_0 + \beta_1 x_i)}\right) + y_i \ln(1 + e^{(\beta_0 + \beta_1 x_i)})\}$$

$$= \sum_{i=1}^{n} y_i \ln(e^{(\beta_0 + \beta_1 x_i)}) - \ln(1 + e^{(\beta_0 + \beta_1 x_i)}). \qquad (4.14)$$

Para encontrar os valores de β que maximizam $L(\beta)$, é necessário diferenciá-lo em relação a cada parâmetro, e igualar a zero. Assim serão encontradas $p + 1$ equações de verossimilhança para os $p + 1$ parâmetros que se deseja estimar, também conhecidas como equações de estimação. Essas equações são dadas por:

$$\sum_{i=1}^{n} y_i - \pi(x_i) = 0, \qquad (4.15)$$

e

$$\sum_{i=1}^{n} x_i[y_i - \pi(x_i)] = 0. \qquad (4.16)$$

Para a Regressão Logística, as equações de verossimilhança são não-lineares em β, dessa forma requerem a utilização de métodos iterativos para encontrar a solução, necessitando de auxílio computacional para se chegar aos resultados (HOSMER e LEMESHOW, 2000).

4.2 Testes de Ajuste do Modelo

Quando o modelo possui mais de uma variável explicativa, com o intuito de prever a condição de interesse, é necessário testar a significância de cada uma delas para a contribuição na formação do valor resposta, cujo intuito é o de identificar a qualidade do modelo e garantir a confiabilidade do mesmo.

Através de gráficos é possível identificar a qualidade do modelo estimado e o quanto ele se ajusta aos dados observados, no entanto existem medidas que ajudam a avaliar a inclusão de cada variável no modelo, como por exemplo, o teste da Razão de Verossimilhança, teste de Wald e teste de score.

4.2.1 Teste da Razão de Verossimilhança

Segundo Hosmer e Lemeshow (2000), o teste da Razão de Verossimilhança normalmente é o mais utilizado e consiste em comparar os valores observados da variável resposta com os valores preditos para dois modelos:

- Modelo atual: modelo com a variável a ser analisada inclusa;
- Modelo saturado: modelo sem a variável a ser analisada inclusa.

Assim, a importância das variáveis será dada por uma medida de significância estatística dos coeficientes β que acompanham as variáveis na função resposta (OLIVEIRA, 2015). Diferente da Regressão Linear onde os erros seguem distribuição normal e seria utilizado um teste F, na Regressão Logística os erros seguem a distribuição binomial e por isso é feito o teste da Razão de Verossimilhança e em cada etapa do teste, quanto maior for a alteração no logaritmo da verossimilhança em relação ao modelo saturado, maior será a importância da variável.

O teste da Razão de Verossimilhança se baseia na estatística *deviance* e é dada por:

$$D = -2\ln\left[\frac{Verossimilhança\ do\ modelo\ atual}{Verossimilhança\ do\ modelo\ saturado}\right] = -2\sum_{i=1}^{n}\left[y_i \ln\left(\frac{\hat{\pi}_i}{y_i}\right) + (1-y_i)\ln\left(\frac{1-\hat{\pi}_i}{1-\ y_i}\right)\right]. \quad (4.17)$$

Para estimar a significância de uma variável de interesse, os valores de D dos modelos com e sem a variável são comparados e então obtém-se a estatística G, expressa por:

$$G = D(modelo\ sem\ a\ variável) - D(modelo\ com\ a\ variável),$$

ou

$$G = -2\ln\left[\frac{verossimilhança\ sem\ a\ variável}{verossimilhanca\ com\ a\ variável}\right]. \quad (4.18)$$

Quando há uma única variável independente e ela não está no modelo, o estimador de máxima verossimilhança de α é $\ln\left(\frac{n_1}{n_0}\right)$ (TSUCHIYA, 2002), onde:

- $n_1 = \sum y_i$;
- $n_0 = \sum(1 - y_i)$.

E o valor preditivo é constante. Assim, tem-se que G é:

$$G = -2\ln\left[\frac{\left(\frac{n_1}{n}\right)^{n_1}\left(\frac{n_0}{n}\right)^{n_0}}{\prod_{i=1}^{n}\hat{\pi}_i^{y_i}(1-\hat{\pi}_i)^{(1-y_i)}}\right],$$

ou

$$G = 2\left\{\sum_{i=1}^{n}[y_i\ln(\hat{\pi}_i) + (1-y_i)\ln(1-\hat{\pi}_i)] - [n_1\ln(n_1) + n_0\ln(n_0) - n\ln(n)]\right\}. \quad (4.19)$$

A hipótese a ser testada é:

$$\begin{cases} H_0: \beta_1 = 0 \\ H_1: \beta_1 \neq 0 \end{cases}$$

Sob a hipótese nula, G terá distribuição qui-quadrado com 1 grau de liberdade.

4.2.2 Teste de Wald

O teste de Wald testa a significância dos parâmetros do modelo através da estimativa de máxima verossimilhança do parâmetro β_k com estimativa do erro padrão $\hat{SE}$. A estatística do teste W, sob a hipótese de que $\beta_k = 0$, terá distribuição normal padrão e é expressa por:

$$W = \frac{\hat{\beta}_k}{\widehat{SE}(\hat{\beta}_k)} \quad (4.20)$$

a hipótese a ser testada é:

$$\begin{cases} H_0: \beta_k = 0 \\ H_1: \beta_k \neq 0 \end{cases}$$

e o p-valor é P($|Z| > W$) onde z segue distribuição normal e rejeita-se a hipótese nula se $Z_{\frac{\alpha}{2}}$, em valor absoluto, for maior que W(TSUCHIYA, 2002).

No entanto, pesquisadores afirmam que o teste de Wald as vezes rejeita um coeficiente significativo, se mostrando não ser tão assertivo.

4.2.3 Teste de Score

O teste de Score se baseia na teoria de distribuição de derivadas do log de verossimilhança, porém não é encontrado facilmente em pacotes computacionais, isso faz com que ele seja limitado

. Sabe-se que:

- $\hat{\pi} = \frac{n_1}{n} = \overline{y}$
- $\overline{y}(1-\overline{y})\sum(x_i - \overline{x})^2$ é a variância estimada

Assim, a estatística do teste ST é dada por:

$$ST = \frac{\sum_i^n x_i(y_i - \overline{y})}{\sqrt{\overline{y}(1-\overline{y})\sum(x_i - \overline{x})^2}} \tag{4.21}$$

a hipótese a ser testada é:

$$\begin{cases} H_0: \beta_1 = 0 \\ H_1: \beta_1 \neq 0 \end{cases}$$

e o p-valor é P($|Z| > ST$) < valor limite p , onde Z expressa a variável aleatória da distribuição normal (TSUCHIYA, 2002).

4.3 Intervalo de Confiança

Coeficiente de inclinação:

$$\hat{\beta}_1 \pm \frac{Z_{1-\alpha/2}}{\widehat{SE}(\hat{\beta}_1)}. \tag{4.22}$$

Intercepto:

$$\hat{\beta}_0 \pm \frac{Z_{1-\alpha/2}}{\widehat{SE}(\hat{\beta}_0)}. \tag{4.23}$$

Função *Logit*:

$$\hat{g}(x_i) \pm \frac{Z_{1-\alpha/2}}{\sqrt{\hat{V}\left(\hat{g}(x_i)\right)}} = \hat{\beta}_0 + \beta_1 x_i \pm \frac{Z_{1-\alpha/2}}{\sqrt{\hat{V}\left(\hat{\beta}_0\right) + x_i^2 \hat{V}\left(\hat{\beta}_1\right) + (2x_i Cov(\hat{\beta}_0, \hat{\beta}_1))}}. \tag{4.24}$$

4.4 Seleção das variáveis

Quando é considerado o modelo com mais de uma variável explicativa para prever a existência da condição de interesse, é importante saber o quanto cada uma das variáveis presentes no modelo irá contribuir para a variável resposta, com o intuito de retirar aquelas que contribuem muito pouco ou até mesmo nada para que o modelo seja parcimonioso (OLIVEIRA, 2015). Quanto maior o número de variáveis inclusas, maior será o erro padrão estimado e mais dependente dos dados o modelo será (HOSMER e LEMESHOW, 2000).

Segundo Hosmer e Lemeshow (2000), existem várias etapas que podem auxiliar na seleção de variáveis em um modelo de Regressão Logística. As principais são:

1. O processo é iniciado com uma análise univariada para cada uma das variáveis presentes no modelo. Para variáveis nominais, ordinais e contínuas, é sugerido que faça uma tabela de contingência com valores $(y = 0{,}1)$ versus k níveis da variável independente, no entanto é importante tomar cuidado com a variável independente pois ela pode atribuir uma célula zero na tabela de contingência atribuindo uma estimativa pontual univariada para uma das razões de chances igualada a zero ou infinito.

2. Logo após, é feita uma análise multivariada com as variáveis selecionadas. As variáveis cujo teste univariado tiver um valor de p-valor < $0{,}25$ são candidatas a entrarem no modelo multivariado junto com as variáveis julgadas importantes clinicamente. É recomendado utilizar $\alpha = 0{,}25$, pois o uso mais comum, em que $\alpha = 0{,}05$, muitas vezes falha na verificação da significância de uma variável.

3. Em seguida, deve-se verificar a importância de cada variável que foi incluída no modelo, através do cálculo da estatística Wald e comparar cada coeficiente estimado com o coeficiente do modelo univariado. As variáveis que não mostrarem contribuição para o modelo baseado nesses critérios devem ser retiradas do modelo e então um novo modelo deve ser ajustado. Assim, o novo modelo deve ser comparado com o modelo antigo através do teste de razão de verossimilhança. Além do que, os coeficientes estimados para as variáveis restantes devem ser comparados com aqueles do modelo completo. Esse processo se repete até que todas as variáveis significativas

sejam inclusas no modelo.

4. Após a obtenção do modelo com as variáveis julgadas essenciais, é importante considerar a interação entre elas. Cada interação é incluída no modelo principal e então é feito uma comparação entre o modelo de interações e o modelo principal. A partir daí as interações são selecionadas e é ajustado um novo modelo. Esse modelo é comparado com o modelo principal e todo o processo se repete até que seja encontrado um modelo completo com as interações.

Para realizar a seleção de variáveis, são utilizados métodos que têm como apoio o teste da Razão de Verossimilhança. O *Stepwise, forward selection* e *backward selection* são alguns exemplos (OLIVEIRA, 2015).

O *Stepwise* é um processo de inclusão e exclusão das variáveis de acordo com sua significância com base em uma regra de decisão fixa e é realizado através de um algoritmo estatístico. Dado que, na Regressão logística, os erros seguem distribuição binomial e a significância é obtida por meio do teste qui-quadrado da razão da verossimilhança, para qualquer etapa do processo a variável mais significativa será aquela que produzir maior mudança no log da verossimilhança em relação ao modelo sem a variável (HOSMER e LEMESHOW, 2000).

O processo de exclusão e inclusão se dá pelos métodos de *backward selection* e *forward selection* respectivamente. O método de *backward selection* implica em remover variáveis que foram incluídas nos passos do processo e ajustar o modelo para avaliar a importância das variáveis retiradas. Já o *forward selection* tem o papel de incluir as variáveis.

4.5 Avaliação do Desempenho de Predição do Modelo

Para avaliar o desempenho de predição do modelo de classificação é necessário utilizar algumas métricas, como por exemplo Matriz de Confusão, Acurácia, Recall, Curva ROC entre outras.

4.5.1 Matriz de Confusão

A Matriz de Confusão é uma tabela que traz frequências referentes a cada classe preditiva e observada do modelo, e possibilita o desenvolvimento de algumas

métricas de avaliação de desenvolvimento. A Matriz de Confusão é composta por quatro medidas classificadas pelo modelo:

- Verdadeiro Positivo (VP): classificação correta da ocorrência de um evento dado que ele realmente acorreu.
- Falso Positivo (FP): classificação incorreta da ocorrência de um evento dado que ele não acorreu.
- Falso Negativo (FN): classificação incorreta da não ocorrência de um evento dado que ele ocorreu.
- Verdadeiro Negativo (VN): classificação correta da não ocorrência de um evento dado que ele realmente não ocorreu.

Dada as medidas, elas são distribuídas da seguinte maneira:

Tabela 2 - Matriz de Confusão

	Valores Observados	
Valores Preditos	Ocorrência do Evento	Não ocorrência do Evento
Ocorrência do Evento	Verdadeiro Positivo	Falso Positivo
Não ocorrência do Evento	Falso Negativo	Verdadeiro Negativo

Fonte: Autoria própria (2021)

A partir da Matriz de Confusão é possível calcular algumas métricas que irão ajudar na avaliação, como:

Acurácia: traz a assertividade do modelo em todas as previsões possíveis e denotada pela razão entre o somatório das previsões corretas sobre o somatório de todas as previsões,

$$Acurácia = \frac{(VP + VN)}{(VP + VN + FP + FN)}. \quad (4.25)$$

Sensibilidade ou Recall: traz a proporção de eventos identificados corretamente pelo modelo, avaliando o quão bom o modelo é para prever verdadeiros positivos, e é denotada pela razão entre verdadeiros positivos sobre o somatório de verdadeiros positivos e falsos negativos.

$$Recall = \frac{VP}{(VP + FN)}. \quad (4.26)$$

Especificidade: traz a proporção de eventos identificados corretamente pelo modelo, avaliando o quão bom o modelo é para prever verdadeiros negativos, e é denotada pela razão entre verdadeiros negativos sobre o somatório de falsos positivos e verdadeiros negativos.

$$Especificidade = \frac{VN}{(VN + VP)}. \tag{4.27}$$

Precisão: traz a proporção de eventos classificados corretamente de uma categoria em relação a todas as classificações feitas dessa mesma categoria, e é denotada pela razão entre verdadeiros positivos sobre o somatório de verdadeiros positivos e falsos positivos.

$$Precisão = \frac{VP}{(VP + FP)}. \tag{4.28}$$

Coeficiente de Correlação de Matthews: traz um compilado de todas as possibilidades de medidas presentes na matriz de confusão além de ser uma métrica indispensável ao se tratar de dados desbalanceados.

$$MCC = \frac{VP * VN - FP * FN}{\sqrt{(VP + FP)(VP + FN)(VN + FP)(VN + FN)}}. \tag{4.29}$$

Assim, ao adaptar a matriz de confusão para a detecção de fraudes em cartões, tem-se a matriz de confusão descrita na tabela 3.

Tabela 3 - Matriz de confusão para Detecção de Transações Fraudulentas

	Valores Observados	
Valores Preditos	Realmente fraudulenta	Realmente efetiva
Classificação como fraudulenta	Verdadeiro Positivo	Falso Positivo
Classificação como efetiva	Falso Negativo	Verdadeiro Negativo

Fonte: Autoria própria (2021)

4.5.2 Curva ROC e AUC

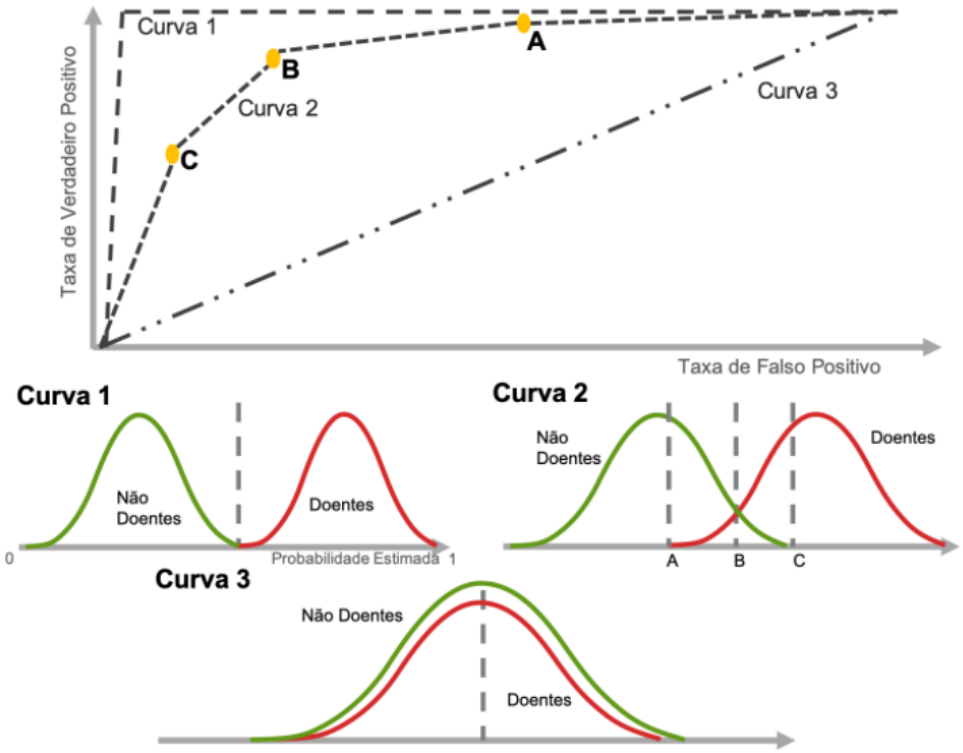
Ao trabalhar com um modelo de classificação, em que a variável resposta é binária, é possível visualizar o desempenho do modelo através da curva ROC (*Receiver Operating Characteristic*), a qual utiliza um ponto de corte onde os valores inferiores a ele são classificados como não ocorrência do evento e os valores superiores a ele são classificados como ocorrência do evento (FREITAS, 2019), e

para determinar o ponto de corte, utiliza-se a relação entre a sensibilidade e a especificidade.

A sensibilidade é a taxa de verdadeiros positivos, ou seja, ela avalia o quão bom o modelo é em prever a ocorrência de um evento dado que, de fato, ele aconteceu. Já a especificidade é a taxa de verdadeiros negativos, ou seja, ela avalia o quão o modelo é em prever a não ocorrência de um evento dado que, de fato, ele não aconteceu.

O ponto de corte, também compreendido como otimização da sensibilidade em relação a especificidade é o ponto da curva mais próximo do canto superior esquerdo do gráfico. Na figura 4 é expresso o comportamento desse gráfico em diferentes situações.

Figura 4 – Comportamentos da curva ROC

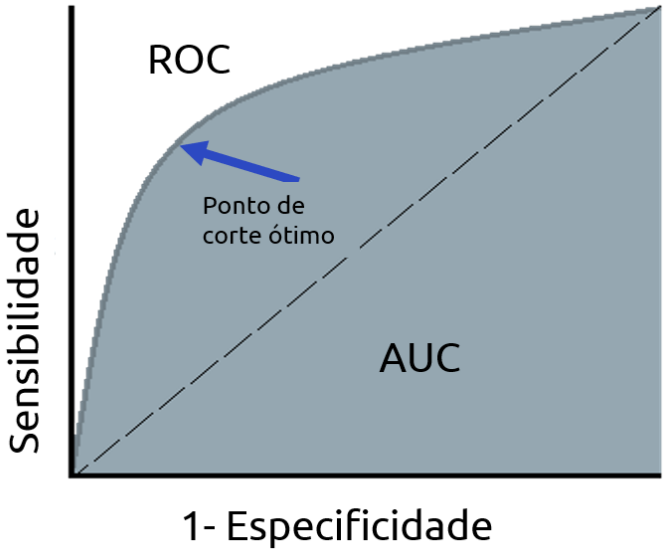


Fonte: Ciência e negócios (2020)

Ao analisar as curvas, nota-se que a curva 1 representa um gráfico perfeito, onde a taxa de verdadeiros positivos está em 100%, já a curva 2 representa um comportamento mais comum na modelagem, onde há algumas predições erradas, mas no geral se saiu bem e por fim a curva 3, a qual representa um modelo com desempenho ruim e que não agrega ao estudo desejado.

Para facilitar a interpretação do gráfico dado que na prática, muitas vezes a curva não fica sutil dificultando a visualização e interpretação, é utilizado a métrica AUC (*Area Under the Curve*), a área situada abaixo da curva ROC. A AUC varia de 0 a 1, e quanto maior ela for melhor será a qualidade e capacidade de previsão do modelo.

Figura 5 – AUC



Fonte: Flai (2020)

4.6 Exemplo de Aplicação

Para ver como a Regressão Logística se comporta na prática, será apresentado um exemplo.

4.6.1Classificação de inadimplência

Para avaliar se um crédito deve ou não ser concedido a um cliente, uma concessionária fez uma análise histórica com o objetivo de estimar a probabilidade do cliente ser classificado como adimplente ou inadimplente, utilizando uma amostra aleatória de 92 clientes que solicitaram o crédito nos últimos 12 meses. Para essa análise, considerou-se o número de dependentes (ND), a renda mensal (RM) e se o cliente possui ou não um vínculo empregatício (VE), onde 1 possui vinculo e 0 não possui vinculo. Foi observado ainda deste cadastro, se os clientes foram classificados como adimplente (0) e inadimplente (1) (CARVO, 2018).

Ao realizar uma análise descritiva dos dados, notou-se que 51 clientes foram classificados com 0, ou seja, 55% do total foram considerados como adimplentes, enquanto que 45% do total foram classificados como inadimplentes.

Em seguida, foram feitas análises para verificar a distribuição dos clientes em relação as variáveis, assim tem-se:

Tabela 4 – Distribuição dos clientes em relação ao Número de Dependentes

Número de Dependentes	Frequência	%
0	4	4%
1	23	25%
2	32	35%
3	23	25%
4	9	10%
5	1	1%
Total	92	100%

Fonte: (CARVO, 2018)

Tabela 5 – Distribuição dos clientes em relação a renda mensal

Renda Mensal	Frequência	%
(1,2]	21	23%
(2,3]	32	35%
(3,4]	20	22%
(4,5]	11	12%
(5,6]	7	7%
(6,7]	0	0%
(7,8]	0	0%
(8,9]	1	1%
Total	92	100%

Fonte: (CARVO, 2018)

Tabela 6 – Distribuição dos clientes em relação ao Vínculo Empregatício

Vínculo Empregatício	Frequência	%
1	50	54%
0	42	46%
Total	92	100%

Fonte: (CARVO, 2018)

Ao observar as Tabelas 4, 5 e 6, nota-se que 35% dos clientes possui 2 dependentes, 35% ganham entre 2 a 3 salários mínimos e 54% possui vínculo empregatício.

Dado o conhecimento do comportamento dos dados através das análises descritivas, foram feitas as estimativas dos parâmetros através do método de Máxima Verossimilhança e foram obtidos os resultados apresentados na Tabela 7:

Tabela 7 – Estimativa dos Parâmetros e Teste Wald

	Estimativa	Erro Padrão	Estatística W	P-valor
Intercepto	1.4776	1.6569	0.892	0.372501
RM	-1.8824	0.4885	-3.853	0.000117 ***
ND	0.8596	0.3857	2.228	0.025854 *
VE	2.8221	0.8521	3.312	0.000926 ***

Fonte: (CARVO, 2018)

Nota-se que todos os p-valores são menores que 5% (nível de significância), logo conclui-se que todos são significantes, com exceção do intercepto.

Através das estimativas é possível calcular a função *logit*, que é dada por:

$$\hat{g}(x_i) = 1{,}4776 - 1{,}8824 * RM + 0{,}8596 * ND + 2{,}8221 * VE$$

e ao calcular as razões das chances, tem-se as seguintes interpretações:

- Renda mensal

$$e^{(-1{,}8824)} = 0{,}15222.$$

Significa que chance de o cliente ser inadimplente diminui em 85% a cada unidade acrescida na renda mensal.

- Número de dependentes

$$e^{(0.8596)} = 2{,}36221.$$

Significa que a chance de o cliente ser inadimplente aumenta em 136% a cada acréscimo de dependente.

- Vínculo empregatício

$$e^{(2.8221)} = 16.81212.$$

Significa que a chance de um cliente ser inadimplente aumenta em 1581% ao não ter vínculo empregatício.

5 Dados Desbalanceados

Um conjunto de dados é considerado desbalanceado quando há uma diferença notória no número de dados presente em uma ou mais classes em relação às outras (BARELLA, 2015). Quando há essa discrepância, chama-se de classes minoritárias aquelas que representam a minoria dos dados e de classes majoritárias, a maioria.

O desbalanceamento de dados é encontrado em diversas áreas de atuação como estudos de doenças raras, identificação de inadimplência de clientes, detecção de fraudes, entre outros. Logo, se trata de situações onde os dados são compostos por muitas informações referentes as classes majoritárias e pouquíssimas das classes minoritárias. E nota-se que os dados estudados no atual trabalho são altamente desbalanceados.

Ao se deparar com conjuntos de dados desse tipo, é necessário utilizar técnicas para balanceá-los, caso contrário, a modelagem fará com que as classes minoritárias não sejam reconhecidas favorecendo as classes majoritárias, que por consequência não resultará em classificadores assertivos (BARELLA, 2015) e sofrerá o Paradoxo da Acurácia.

O Paradoxo da Acurácia se trata de uma falsa impressão da qual o modelo se ajusta bem devido à alta acurácia, quando na verdade reflete uma falha na predição do mesmo, pois os parâmetros não são capazes de distinguir o que está na classe minoritária em relação as classes majoritárias, resultando em classificações errôneas e previsões sem utilidades.

Alguns métodos bastante utilizados para contornar essa situação são os métodos de *undersampling* (subamostragem) e *oversampling* (sobreamostragem), os quais utilizam mecanismos que alteram a estrutura original dos dados a fim de balanceá-los, bem como o método de adaptação de algoritmos que faz adaptações com o cenário em estudo ao invés de tentar balancear os dados.

5.1 Método de subamostragem ou *undersampling*

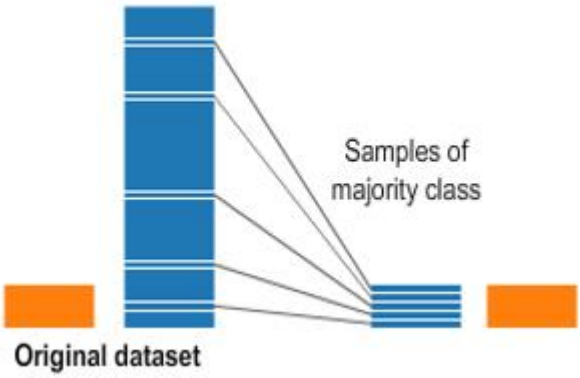
O método de subamostragem, também conhecido como *undersampling*, reduz a quantidade de dados da classe majoritária como intuito de diminuir a diferença entre as classes até que a proporção entre elas seja de 1:1, e pode ser feito de duas formas:

- Retirando elementos aleatoriamente
- Realizando uma fusão entre uma ou mais observações da classe majoritária e transformando em uma.

A primeira forma ocasiona uma perda maior de dados, o que muitas vezes pode ser um problema, assim é mais recomendado utilizar a segunda forma já que ela tem uma perda menor dos dados, e para utiliza-la existe alguns métodos como NearMiss, K-NN (*K-Nearest Neighbors Algorithm* ou K-ésimo Vizinho Mais Próximo), entre outros, em que suas implementações se dão através de pacotes computacionais.

De modo geral, interpreta-se a técnica *undersampling* conforme a figura 6.

Figura 6 – *Undersampling*



Fonte: Medium (2021)

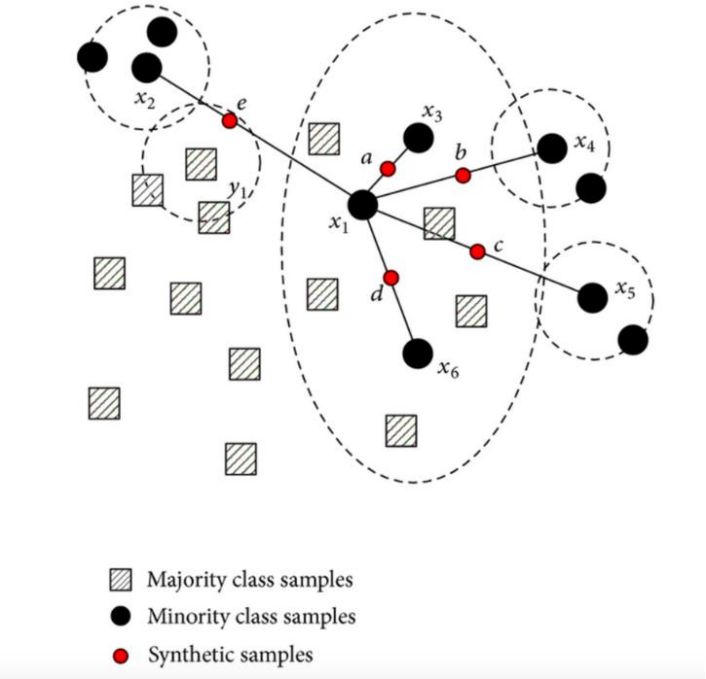
5.2 Método de sobreamostragem ou *oversampling*

O método de sobreamostragem ou oversampling aumenta a quantidade de dados da classe minoritária com o intuito de diminuir a diferença entre as classes até que a proporção entre elas seja de 1:1, e para obter esse aumento existem três formas:

- Os dados já existentes são replicados aleatoriamente
- O conjunto todo é replicado
- Novos dados são gerados sinteticamente

Dentre as formas citadas acima, a terceira conta com uma técnica mais sofisticada conhecida como SMOTE. Essa técnica tem como objetivo gerar informações artificiais através do K-NN baseadas nos dados da classe minoritária, logo ele utiliza informações já existentes e que são próximas, isto é, elementos vizinhos, para gerar uma nova informação, de acordo com a figura 7.

Figura 7 – Método SMOTE



Fonte: Medium (2021)

Para obter um resultado ainda mais assertivo, usa-se a técnica ADASYN, a qual é uma versão aprimorada do SMOTE (HE et al., 2008 *apud* BARELLA, 2015) e tem como principal objetivo ponderar as novas informações geradas, criando novos pontos entre elas, a fim de aumentar a variância da amostra e tornar os dados mais espaçados.

De modo geral, pode-se interpretar a técnica *oversampling* conforme a figura 8.

Figura 8 – *Oversampling*



Fonte: Medium (2021)

5.3 Método de adaptação de algoritmos

Por fim, o método de adaptação de algoritmos se diferencia dos métodos de subamostragem e sobreamostrem pois ele não visa modificar a base de dados, mas sim adaptar as técnicas de classificações já existentes em um conjunto de dados que são desbalanceados.

Essa adaptação é feita através de diferentes funções de custos que irão penalizar os erros de classificação atribuídos à ambas classes, majoritárias e minoritárias, isso fará com que o modelo tenha um aprendizado mais rígido e o aprendizado minimize “o custo total de erros de classificação e a quantidade de exemplos classificados erroneamente em situações de alto custo de erro” (BARELLA, 2015, p. 16).

Portanto, ao utilizar umas das técnicas quando se trabalha com dados desbalanceados, é possível modelar e obter resultados mais assertivos que irão ser capazes de fazer previsões.

6 Aplicação

Na seção 6.1 será feita uma descrição dos dados utilizados no trabalho, bem como análises descritas, já na seção 6.2 serão descritos os tratamentos dos dados que foram necessários e por fim, na seção 6.3 terá a aplicação do modelo com os dados desbalanceados e balanceados.

6.1 Descrição dos dados e análises descritivas

Para realização do trabalho, está sendo utilizada uma base de dados retirada do site *Kaggle.com*, que é um site de competição de ciência de dados composto por bases de dados públicos, códigos já realizados, comunidades com diversas discussões e minicursos, voltados para o aprendizado de máquina.

A base escolhida possui dados sintéticos gerados em um simulador de pagamento móvel chamado *PaySim*, o qual gera dados artificiais de pagamento móvel baseados em um conjunto de dados reais de maneira a torná-los bastante semelhantes aos dados originais. A amostra de transações reais utilizada é referente a transações que aconteceram em um país africano durante um mês. Foi necessário recorrer a estes dados sintéticos, tendo em vista que dificilmente um trabalho acadêmico terá acesso a um conjunto de dados reais na área de fraude devido a sensibilidade das informações.

A base de dados é composta por 6.362.620 linhas e 11 variáveis.

As variáveis presentes na base de dados são:

- step – mapeamento de uma unidade de tempo, o que corresponde a 1 hora (744 etapas)
- type – método de pagamento
- amount – montante da transação na moeda local
- nameOrig – Identificação do cliente que iniciou a transação
- oldbalanceOrg – Saldo inicial pré-transação
- newbalanceOrig – Saldo pós transação
- nameDest – Destinatário da transação
- oldbalanceDest – destinatário do saldo inicial pré-transação. Observação: Não há informações para clientes que possui M como inicial da identificação.
- newbalanceDest – destinatário do saldo pós transação. Observação: Não há informações para clientes que possui M como inicial da identificação.

- isFraud - Transações feitas pelos fraudadores dentro da simulação.
- isFlaggedFraud - Tentativa ilegal, onde é considerado transações com valores maior ou igual a 200.000 em uma única transação.

Para entender o comportamento das variáveis é importante fazer uma análise descritiva dos dados. Assim, temos algumas análises para as variáveis isFraud isFlaggedFraud, type e amount realizadas no software Python, bem como todo o desenvolvimento do atual trabalho.

6.1.1 Variável Resposta “isFraud”

A variável resposta para este trabalho é a “isFraud”, a qual classifica as transações realizadas transação como fraude e não fraude, onde:

- 0 representa a transação efetiva, ou seja, não fraude
- 1 representa a transação fraudulenta, ou seja, fraude.

A distribuição de cada uma, é apresentada na Tabela 8.

Tabela 8 - Distribuição da variável “isFraud”

	Classificação		Total
	0	1	
Quantidade	6.354.407	8.213	6.362.620
%	99,87%	0.13%	100%

Fonte: Autoria própria (2021)

Pode-se notar que o volume total de transações é alto, porém quando segmenta por fraude e não fraude, apenas 0,13% é classificado como fraude, mesmo que represente um número alto de transações (8.213 transações).

6.1.2 isFlaggedFraud

A variável “isFlaggedFraud” classifica as transações com valores maior ou igual a 200.000 (moeda do país) em uma única como tentativa ilegal onde:

- 0 representa uma transação legal
- 1 representa uma transação ilegal

A distribuição destes dados é apresentada na Tabela 9.

Tabela 9 - Distribuição da variável “isFlaggedFraud”

	Classificação		Total
	0	1	
Quantidade	6.362.604	16	6.362.620
%	99,99975%	0.00025%	100%

Fonte: Autoria própria (2021)

Ao observar a quantidade de transações consideradas ilegal, nota-se que é baixíssimo, representando praticamente 0% do total, pois o valor que sinaliza a ocorrência de uma transação ilegal é bastante alto.

6.1.3 type

A variável “type” classifica a forma em que a transação foi feita e possui 5 categorias:

- PAYMENT
- TRANSFER
- DEBIT
- CASH-IN
- CASH-OUT

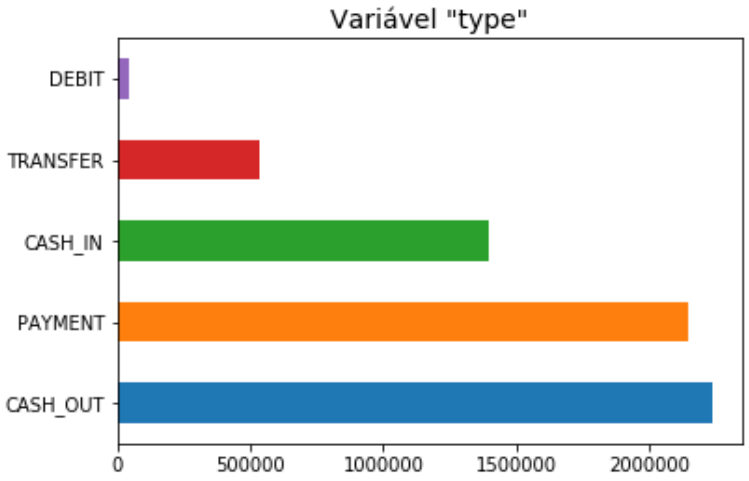
A distribuição de cada uma, é apresentada na Tabela 10 e na Figura 9.

Tabela 10 - Distribuição da variável “type”

Classificação	Quantidade	%
PAYMENT	2.151.495	34%
TRANSFER	532.909	8%
DEBIT	41.432	1%
CASH-IN	1.399.284	22%
CASH-OUT	2.237.500	35%
Total	6.362.620	100%

Fonte: Autoria própria (2021)

Figura 9 - Gráfico de barras da variável “type”



Fonte: Autoria própria (2021)

Nota-se que dentre as cinco classificações, a quantidade de transações está concentrada em três classificações, que são elas: Cash-out (35%), Payment (34%) e Cash-in (22%), as quais juntas representam 91% do total.

6.1.4 amount

A variável “amount” é o montante da transação na moeda local, ou seja, o valor transacionado. Algumas estatísticas referentes a ela são apresentadas na Tabela 11.

Tabela 11 - Medidas de posição da variável “amount”

Medidas	Valor na moeda do País
Média	179.861,904
Desvio Padrão	603.858,231
Valor mínimo	0,000
1ºQuartil	13.389,570
Mediana	74.871,940
3ºQuartil	208.721,478
Valor Máximo	92.445.516,640

Fonte: Autoria própria (2021)

Ao observar as medidas, nota-se que o desvio padrão é bem alta. Por outro lado, a média é muito maior que a mediana, possivelmente influenciada pelo valor máximo, que é muito alto.

6.2 Tratamento dos dados

Através das análises descritivas e identificação do comportamento dos dados, se fez necessário adentrar algumas análises para identificar qualquer anomalia que viesse a atrapalhar o desenvolvimento e retirá-las da base.

Inicialmente foi verificado se haviam dados nulos na base, dado que eles atrapalham a predição do modelo e não dizem nada por não terem informação, no entanto não havia, porém a variável “nameDest”, a qual classifica quem recebe uma transação através de um código, possui letras iniciais e para aquelas classificações que iniciam com “M” não há informações nas variáveis que dizem o saldo do recebedor pré e pós recebimento, logo podem ser interpretadas como valores nulos, assim todas as linhas com iniciais “M” foram removidas da base de dados, resultando em uma base de 4.211.125 milhões de linhas.

Em seguida foram observados os *“data types”* das variáveis a fim de tratar todas aquelas que fossem categóricas, já que dificultam a execução do código do modelo no software Python. Dentre as 11 variáveis, 3 são do tipo categóricas: “type”, “nameOrig” e “nameDest”. E para avaliar o comportamento delas, foi gerado a Tabela 12 com o total de informações, a quantidade de classificações distintas que cada variável possui, a classificação que mais aparece em cada uma e a frequência das mesmas.

Tabela 12 - Distribuição da variável “type”

	type	nameOrig	nameDest
Total	4.211.125	4.211.125	4.211.125
Quantidade	4	4.207.035	571.961
Top	CASH_OUT	C1462946854	C1286084959
Frequência	2.237.500	3	113

Fonte: Autoria própria (2022)

Observa-se que a variável type possui 4 classificações distintas, pois todas as informações com classificação “PAYMENT” foram excluídas no momento em que as classificações iniciadas com “M” da variável “nameDest” foram retiradas da base. Já as variáveis “nameOrig” e “nameDest” possuem muitas classificações distintas e as classificações que aparecem mais vezes em ambas possuem uma frequência muito

baixa, assim nota-se que essas variáveis não irão contribuir para o modelo, logo elas não serão inclusas.

Ainda na linha de definir as variáveis que entrarão no modelo, a variável “step” também não será inclusa pois ela representa a unidade de tempo, através de horas totalizando um mês (774/31 = 24), porém dado que não se sabe quais unidades representam por exemplo os finais de semana, de qual mês do ano se trata e por não haver segundos, as quais são informações que influenciam diretamente na identificação da fraude, não agregará ao modelo, logo não será utilizada.

E além das variáveis “nameOrig”, “nameDest” e “step”, também não será utilizada a variável “isFlaggedFraud” pois a volumetria de fraudes presentes nela é extremamente baixa então não irá ajudar o modelo a discriminar o estudo de interesse.

Por fim, foi feito uma transformação da única variável categórica a ser utilizada, a variável “type”, em variáveis dummys, as quais ficaram dispostas de acordo com a figura 10.

Figura 10 – Distribuição das variáveis dummys “type”

	CASH_IN	CASH_OUT	DEBIT	TRANSFER
2	0	0	0	1
4221113	0	1	0	0
4221114	1	0	0	0
4221115	1	0	0	0
4221116	1	0	0	0
...	...	...	...	...
3049066	0	0	1	0
6016400	0	0	1	0
6016398	0	0	1	0
4423725	0	0	1	0
3165946	0	0	1	0

[4211125 rows x 4 columns]

Fonte: Autoria própria (2022)

Portanto, as variáveis definidas para serem utilizadas no modelo, além da variável resposta “isFraud” foram “amount”, “oldbalanceOrg”, “newbalanceOrig”, “oldbalanceDest”, “newbalanceDest”, “CASH_IN”, “CASH_OUT”, “DEBIT” e “TRANSFER”. Dessa forma, sabe-se que CASH_IN é o recebimento do dinheiro, CASH_OUT é o pagamento via cartão de crédito, DEBIT é o pagamento via cartão de débito e TRANSFER é uma transferência direta.

Em decorrência do grande volume de dados, ficou um pouco inviável trabalhar com a volumetria inteira, logo foi preciso utilizar uma amostra aleatória simples para facilitar o processo. Para chegar no tamanho ideal da amostra foi utilizado a fórmula:

$$n = \frac{NZ^2p(1-p)}{Z^2p(1-p) + e^2(N-1)} \quad (6.1)$$

onde:

N = População total

Z = Nível de confiança desejado

e = Erro

p = Proporção de transações fraudulentas

Então, calculando os valores e substituindo na fórmula 6.1, temos o seguinte:

$$n = \frac{4211125 * 1{,}96^2 * 0{,}00195031(1-0{,}00195031)}{1{,}96^2 * 0{,}00195031(1-0{,}0019031) + 0{,}0001^2(4211125-1)} = 635011{,}039. \quad (6.2)$$

Logo foi utilizado uma amostra de 635.011 linhas, mantendo todas as fraudes já que possui uma volumetria muito baixa e se removêssemos parte dela, provavelmente perderíamos insumo para trazer bons resultados. Assim, a proporção de fraudes passou de 0,00195031 para 0,01293363, isto é, teve um aumento, mas ainda assim representa muito pouco do todo, portanto pode-se dizer que os dados permaneceram desbalanceados.

6.3 Aplicações do modelo e resultados

Dado o desbalanceamento dos dados, aplicar a regressão logística sem visar balanceá-los resultará em um modelo que não é capaz de distinguir a classe minoritária, no atual estudo, a fraude. Logo, para avaliar a importância desta tratativa para obter um bom modelo e comparar o que melhor se adapta, serão feitas três aplicações:

- Regressão Logística com os dados desbalanceados
- Regressão Logística após balancear os dados utilizando a técnica de *undersampling* pelo método NearMiss
- Regressão Logística após balancear os dados utilizando a técnica de *oversampling* pelo método SMOTE

Para os três casos a base foi dividida em duas partes, sendo uma para treinar o modelo, isto é, para estimar os parâmetros e outra para testar as predições e avaliar se quando o modelo for submetido a outros dados, ele conseguirá discriminar a fraude

de não fraude. A base de treino representa 70% do total, resultando em 444.507 linhas e a base de teste, 30% do total, resultando em 190.504.

6.3.1 Regressão Logística com os dados desbalanceados

Sabe-se que as variáveis selecionadas incialmente para a construção do modelo foram, além da variável resposta “isFraud”, “amount”, “oldbalanceOrg”, “newbalanceOrig”, “oldbalanceDest”, “newbalanceDest”, “CASH_IN”, “CASH_OUT”, “DEBIT” e “TRANSFER”.

Ao gerar o modelo inicial, as variáveis “oldbalanceOrg”, “newbalanceOrig”, “CASH_IN”, “DEBIT” e “TRANSFER” não foram estatisticamente significativas, com p-valores > 0,05, logo foram retiradas para obter um modelo mais assertivo. Assim, as estimativas dos parâmetros com as variáveis que permaneceram se dão pela tabela 13.

Tabela 13 – Estimativas dos parâmetros RL com dados desbalanceados

	coef	std err	z	P>|z|	[0.025	0.975]
intercept	-5.0858	0.031	-161.940	0.000	-5.147	-5.024
amount	3.356e-06	5.29e-08	63.500	0.000	3.25e-06	3.46e-06
newbalanceDest	-2.262e-06	4.91e-08	-50.072	0.000	-2.55e-06	-2.36e-06
oldbalanceDest	1.752e-06	4.96e-08	35.308	0.000	1.65e-06	1.85e-06
CASH_OUT	0.9103	0.036	25.154	0.000	0.839	0.981

Fonte: Autoria própria (2022)

Nota-se que todos os p-valores são estatisticamente significativos ao nível de significância de 5%, os coeficientes em sua maioria são números bem pequenos, e ao analisar os intervalos de confiança, nenhum está contido o 0, logo pode-se seguir com as análises.

No entanto, para afirmar que o modelo está bom e poder tirar conclusões, é necessário avaliar seu desempenho através da matriz de confusão, das métricas e da Curva ROC.

Tendo em vista as estimativas dos parâmetros, considerando um ponto de corte de 0.5 e utilizando a base de teste, temos a seguinte matriz de confusão:

Tabela 14 – Matriz de confusão RL com dados desbalanceados

Valores Preditos	Valores Observados	
	Realmente fraudulenta	Realmente efetiva
Classificação como fraudulenta	408	67
Classificação como efetiva	2.053	187.976

Fonte: Autoria própria (2022)

Ao observar a matriz de confusão nota-se que o modelo previu muito bem transações efetivas, dado que ele acertou 99%, em contrapartida previu pouquíssimas fraude, ou seja, ele não foi capaz de identificar o comportamento da fraude em meio a maioria de transações efetivas. Dado esse resultado, calculou-se a acurácia, *recall e MCC* para ajudar nas análises e retornou-se:

Tabela 15 – Métricas de desempenho das predições do modelo

Métrica	Resultado
Acurácia	99%
Recall	16%
MCC	37%

Fonte: Autoria própria (2022)

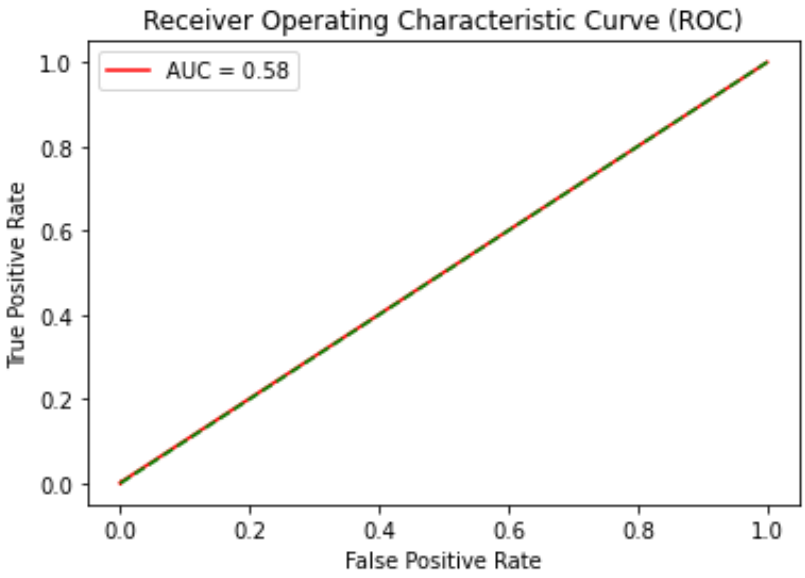
Se este modelo fosse avaliado apenas pela acurácia, poderia afirmar que é um modelo que prevê muito bem, dado que ele erra apenas 1% das classificações observadas, no entanto quando as outras métricas são observadas nota-se que na verdade, o modelo está sofrendo paradoxo da acurácia, pois possui uma acurácia altíssima e um *Recall* de 16%, ou seja, de todas as predições realizadas para prever as transações fraudulentas, 16% foram predições correta, logo o modelo considerou a maioria das transações como efetivas e não retornou a resposta de interesse.

Através do *MCC* igual a 37% é possível entender o motivo da discrepância entre as métricas e o porquê o modelo não ter identificado as transações fraudulentas, que é o desbalanceamento dos dados. Essa métrica aponta quando há esse tipo de problema na base de dados de acordo com seu resultado, então quanto menor for, maior atenção será necessária quanto ao desbalanceamento das informações.

Para visualizar o desempenho do modelo ao diferenciar as transações fraudulentas das transações não fraudulentas, baseado na taxa de verdadeiros

positivos e falsos positivos, gerou-se a curva ROC e a AUC, com ponto de corte de 0,5.

Figura 11 – Curva ROC



Fonte: Autoria própria (2022)

Nota-se que a AUC foi de 0,58, ou seja, está muito próxima da linha de corte observada, reafirmando o quanto o modelo é não satisfatório e que não servirá para detectar as transações fraudulentas, as quais são de interesse do atual trabalho, portanto é um modelo dispensável.

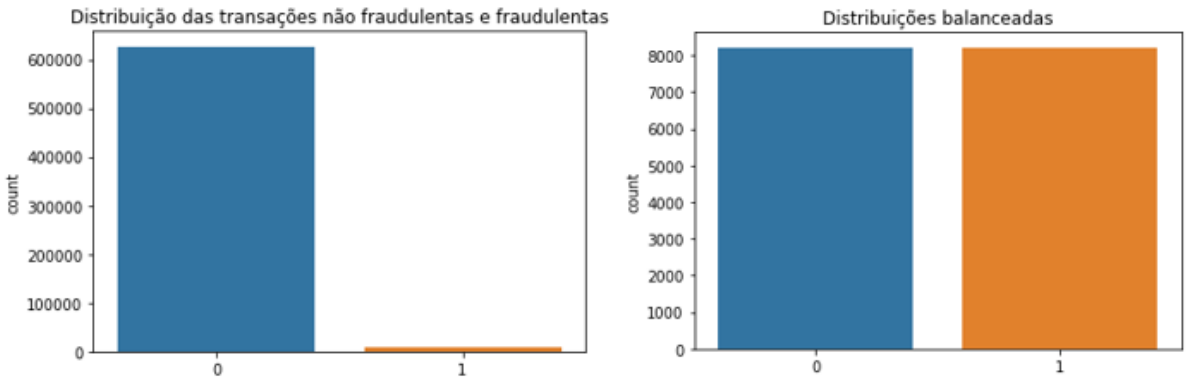
6.3.2 Regressão Logística após balancear os dados utilizando a técnica de *Undersampling*

A técnica de *undersampling* reduz a volumetria da classe majoritária, isto é, das transações não fraudulentas até que ela se iguale a volumetria da classe minoritária, ou seja, das transações fraudulentas. Isso irá balancear os dados fazendo com que a proporção seja a mesma para ambas classes, e dado as formas possíveis de realizar essa redução, foi escolhido o método NearMiss pois ele faz uma fusão entre uma ou mais observações da classe majoritária e transforma em uma nova informação, evitando maior perda dos dados.

Tendo em vista que a quantidade de transações fraudulentas na base é de 8.213, ao aplicar essa técnica passou-se a trabalhar com uma base de 16.426 linhas, onde 70% (11.498) dos dados foram separados para treino e 30% (4.928) dos dados foram reservados para testar o modelo.

Na figura 12 tem-se a distribuição das transações fraudulentas e não fraudulentas antes de aplicar a técnica e após aplicação, já com os dados balanceados.

Figura 12 – Distribuição das transações fraudulentas e não fraudulentas pré e pós balanceamento NearMiss



Fonte: Autoria própria (2022)

Assim como na primeira aplicação, as variáveis selecionadas incialmente para a construção do modelo, além da variável resposta “isFraud”, foram “amount”, “oldbalanceOrg”, “newbalanceOrig”, “oldbalanceDest”, “newbalanceDest”, “CASH_IN”, “CASH_OUT”, “DEBIT” e “TRANSFER”.

Ao gerar o modelo inicial, as variáveis “oldbalanceOrg”, “CASH_IN”, “DEBIT” e “TRANSFER” não foram estatisticamente significativas, com p-valores > 0,05, logo foram retiradas do modelo. Assim, as estimativas dos parâmetros com as variáveis que permaneceram estão dispostas na tabela 16.

Tabela 16 – Estimativas dos parâmetros NearMiss

	coef	std err	z	P>|z|	[0.025	0.975]
intercept	-1.2682	0.061	-20.700	0.000	-1.388	-1.148
amount	2.555e-05	1.02e-06	25.055	0.000	2.36e-05	2.75e-05
newbalanceOrig	-5.047e-06	2.53e-07	-19.925	0.000	-5.54e-06	-4.55e-06
newbalanceDest	-7.296e-06	1.01e-06	-7.203	0.000	-9.28e-06	-5.31e-06
oldbalanceDest	1.066e-05	1.04e-06	10.212	0.000	8.61e-06	1.27e-06
CASH_OUT	-1.8131	0.078	-23.192	0.000	-1.966	-1.660

Fonte: Autoria própria (2022)

Observa-se que todos os p-valores são inferiores ao nível de 5% de significância e não há intervalos em que o 0 está contido.

Desta maneira, a função *Logit* é dada por:

$$\hat{g}(x_i) = -1.2682 + 2.555e-05(amount) - 5.047e-06(newbalanceOrig) - 7.296e-06(newbalanceDest) + 1.066e-05(oldbalanceDest) - 1.8131(CASH_OUT). \quad (6.3)$$

Agora serão feitas as análises de desempenho do modelo para avalia-lo e poder tirar conclusões.

Tendo em vista as estimativas dos parâmetros, considerando um ponto de corte de 0.5 e utilizando a base de teste, temos a seguinte matriz de confusão:

Tabela 17 – Matriz de confusão NearMiss

	Valores Observados	
Valores Preditos	Realmente fraudulenta	Realmente efetiva
Classificação como fraudulenta	2.069	80
Classificação como efetiva	421	2.358

Fonte: Autoria própria (2022)

Ao observar a matriz de confusão nota-se que o modelo previu muito bem tanto transações efetivas quanto transações não efetivas, resultando em uma taxa global de classificação igual a 90%, ou seja, o modelo está errando apenas 10% das predições. Dado esse resultado, calculou-se a acurácia, *recall e MCC* para ajudar nas análises e retornou-se:

Tabela 18 – Métricas de desempenho das predições do modelo NearMiss

Métrica	Resultado
Acurácia	90%
Recall	83%
MCC	80%

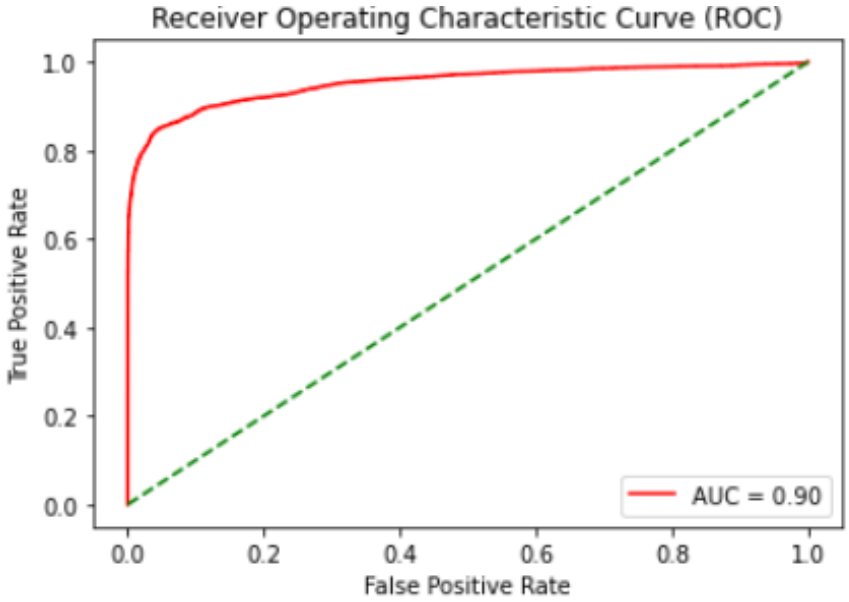
Fonte: Autoria própria (2022)

Além de uma acurácia ótima, agora o modelo também possui um *recall* alto, de 83%, o que é muito bom para os dados que estão sendo estudados, pois esse *recall* diz que, de todas as fraudes preditas pelo modelo, 83% dessas estão corretas. E

através da métrica *MCC* é possível confirmar o quão bom ele está, já que quanto maior o *MCC*, melhor balanceados os dados a serem modelados estão, ou seja, tem-se a afirmação de que os dados estão adequados.

Para visualizar o quão bom o modelo foi para diferenciar as transações fraudulentas das transações não fraudulentas, baseado na taxa de verdadeiros positivos e falsos positivos, gerou-se a curva ROC e a AUC, com ponto de corte de 0,5.

Figura 13 – Curva ROC NearMiss



Fonte: Autoria própria (2022)

Nota-se que a AUC foi de 0,90, ou seja, está bem acima da linha de corte, o que implica em um valor bastante satisfatório para modelos de classificação, logo conclui-se que acerta 90% das predições obtidas através da base de dados utilizada no momento de estimar os parâmetros do modelo.

Para concluir a interpretação, calculou-se as *Odds Ratio* estimadas, dispostas na tabela 19.

Tabela 19 – Odds Ratio NearMiss

Parâmetros	*Odd Rate*
amount	1.00003
newbalanceOrig	0.99999
newbalanceDest	0.99999
oldbalanceDest	1.00001
CASH_OUT	0.16315

Fonte: Autoria própria (2022)

Ao observar as *Odds Ratio,* nota-se que a “amount”, variável que representa o valor transacionado, impacta de maneira significativa as chances de incidência de fraude, ou seja, quanto maior for o valor transacionado, maior será as chances daquela transação ser uma fraude, logo pode-se dizer que, a cada unidade acrescida no valor a ser transacionado, as chances de ser uma transação fraudulenta aumentam em 0.003%.

Já o destinatário do saldo após transação e o saldo após aquela transação, representados respectivamente pelas variáveis “newbalanceOrig” e “newbalanceDest”, ao serem aumentados em uma unidade, diminui em 0,001% as chances de ocorrência de fraude, isto é, ao analisar o destinatário e observar o saldo dele após transações feitas, quanto maior for seu saldo, menor as chances das transações relacionadas serem fraudulentas.

A “oldbalanceDest”, variável que representa o destinatário do saldo inicial pré-transação também impacta as chances de incidência de fraude, ou seja, quanto maior for o saldo já existente naquela conta, maior será as chances de ocorrência de fraude, aumentando cerca de 0,001%as chances de fraude após o acréscimo de unidade nessa variável.

Por fim, a variável “CASH_OUT”, a qual representa a forma de pagamento, impacta as chances de incidência de fraude de forma bem mais expressiva que as demais variáveis, logo se o tipo de transação for um CASH_OUT, a chance de fraude é reduzida em 84%. Portanto pode-se concluir que o valor a ser transacionado, saldos iniciais e saldos pós transações, e tipo de pagamento podem dizer muito sobre uma transação ser efetiva ou não.

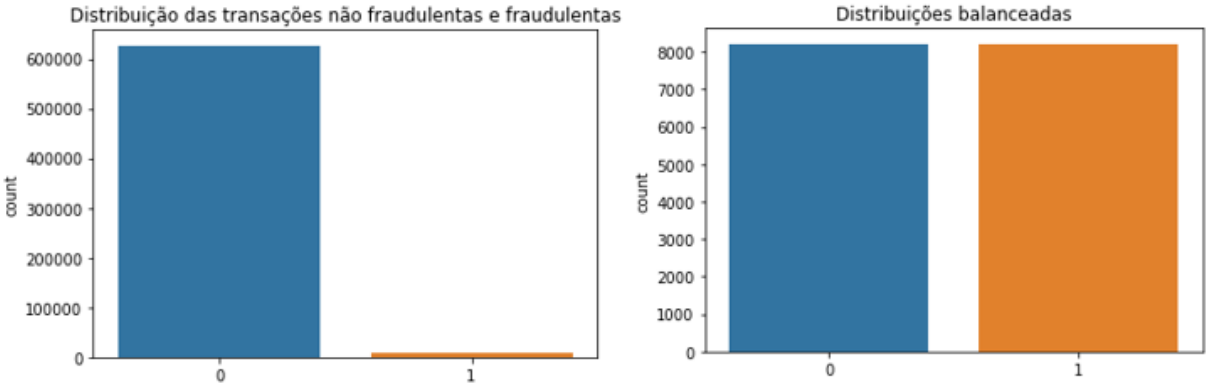
6.3.3 Regressão Logística após balancear os dados utilizando a técnica de *Oversampling*

A técnica de *oversampling* aumenta a volumetria da classe minoritária, isto é, das transações fraudulentas até que ela se iguale a volumetria da classe majoritária, ou seja, das transações fraudulentas. Esta técnica é outra forma de balancear os dados fazendo com que a proporção seja a mesma para ambas classes, e dado as opções disponíveis para realizar esse aumento, foi escolhido o método SMOTE pois ele gera informações artificiais através do K-NN para criar novas informações que se assemelham com os dados já existentes

Sabe-se que a quantidade de transações não fraudulentas na base representa a maioria dos dados, com 626.798 linhas, assim ao aplicar essa técnica passou-se a trabalhar com uma base de 1.253.596 informações, onde 70% (877.517) dos dados foram separados para treino e 30% (376.079) dos dados foram reservados para testar o modelo.

Na figura 14 tem-se a distribuição das transações fraudulentas e não fraudulentas antes de aplicar a técnica e após aplicação, já com os dados balanceados.

Figura 14 – Distribuição das transações fraudulentas e não fraudulentas pré e pós balanceamento SMOTE



Fonte: Autoria própria (2022)

Assim como nas aplicações anteriores, as variáveis selecionadas incialmente para a construção do modelo, além da variável resposta “isFraud”, foram “amount”, “oldbalanceOrg”, “newbalanceOrig”, “oldbalanceDest”, “newbalanceDest”, “CASH_IN”, “CASH_OUT”, “DEBIT” e “TRANSFER”.

Ao gerar o modelo inicial, as variáveis “oldbalanceOrg”, “CASH_IN”, “DEBIT” e “TRANSFER” não foram estatisticamente significativas, com p-valores > 0,05, logo foram retiradas do modelo. Assim, as estimativas dos parâmetros com as variáveis que permaneceram estão dispostas na tabela 20.

Tabela 20 – Estimativas dos parâmetros SMOTE

	coef	std err	z	P>|z|	[0.025	0.975]
intercept	-0.5963	0.005	-114.135	0.000	-0.606	-0.586
amount	4.146e-06	1.84e-08	225.300	0.000	4.11e-06	4.18e-06
newbalanceOrig	-1.28e-06	7.26e-09	-176.215	0.000	-1.29-e06	-1.27e-06
newbalanceDest	-3.103e-06	1.8e-08	-172.434	0.000	-3.14e-06	-3.07e-06
oldbalanceDest	2.872e-06	1.82e-08	157.908	0.000	2.84e-06	2.91e-06
CASH_OUT	0.2881	0.006	47.769	0.000	0.276	0.300

Fonte: Autoria própria (2022)

Observa-se que todos os p-valores são inferiores ao nível de 5% de significância e não há intervalos em que o 0 está contido.

Desta maneira, a função *Logit* é dada por:

$$\hat{g}(x_i) = -0.5963 + 4.146e-06(amount) - 1.28e-06(newbalanceOrig) - 3.103e-06(newbalanceDest) + 2.872e-06(oldbalanceDest) + 0.2881(CASH_OUT). \quad (6.4)$$

Agora os mesmos passos realizados nas aplicações anteriores serão feitos. Logo, na sequencia tem-se as análises de desempenho do modelo para avalia-lo e interpreta-lo.

Tendo em vista as estimativas dos parâmetros, considerando um ponto de corte de 0.5 e utilizando a base de teste, temos a seguinte matriz de confusão:

Tabela 21 – Matriz de confusão SMOTE

	Valores Observados	
Valores Preditos	Realmente fraudulenta	Realmente efetiva
Classificação como fraudulenta	116.153	24.871
Classificação como efetiva	72.036	163.019

Fonte: Autoria própria (2022)

Ao observar a matriz de confusão nota-se que o modelo previu corretamente boa parte dos eventos, resultando em uma taxa global de classificação igual a 74%, o que é bem aceitável. Dado esse resultado, calculou-se a acurácia, *recall e MCC* para ajudar nas análises e retornou-se:

Tabela 22 – Métricas de desempenho das predições do modelo SMOTE

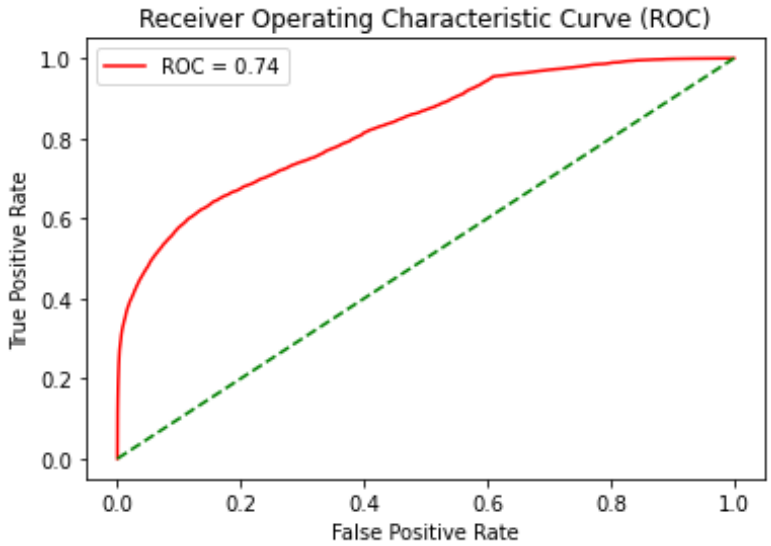
Métrica	Resultado
Acurácia	74%
Recall	62%
MCC	50%

Fonte: Autoria própria (2022)

Além de uma acurácia bastante aceitável, o modelo possui um *recall* de 62%, ou seja, de todas as fraudes preditas pelo modelo, 62% dessas estão corretas, então 3/5 das transações fraudulentas estão sendo preditas corretamente. E através da métrica *MCC* é possível confirmar, que de fato, o modelo está bom dado que 50% é um valor mediano, então mesmo com um volume grande de dados, sabe-se que eles estão adequados para serem utilizados em modelagem.

Para visualizar o quão bom o modelo foi para diferenciar as transações fraudulentas das transações não fraudulentas, baseado na taxa de verdadeiros positivos e falsos positivos, gerou-se a curva ROC e a AUC, com ponto de corte de 0,5.

Figura 15 – Curva ROC SMOTE



Fonte: Autoria própria (2022)

Nota-se que a AUC foi de 0,74, ou seja, está acima da linha de corte, o que implica em um valor bastante satisfatório para modelos de classificação, logo conclui-se que acerta 74% das predições obtidas através da base de dados utilizada no momento de estimar os parâmetros do modelo.

Para concluir a interpretação, calculou-se as *Odds Ratio* estimadas, dispostas na tabela 23.

Tabela 23 – Odds Ratio SMOTE

Parâmetros	*Odd Rate*
amount	1.000004
newbalanceOrig	0.99999
newbalanceDest	0.99999
oldbalanceDest	1.000002
CASH_OUT	1.33389

Fonte: Autoria própria (2022)

Ao observar as *Odds Ratio,* nota-se que a “amount”, assim como na aplicação anterior, impacta de maneira significativa as chances de incidência de fraude, ou seja, a cada unidade acrescida no valor a ser transacionado, as chances de ser uma transação fraudulenta aumentam em 0,0004%.

A “newbalanceOrig” e “newbalanceDest”, imprimiram basicamente os mesmos resultados, isto é, quanto maior for o saldo naquela conta após uma transação, menor será as chances de ocorrência de fraude, diminuindo cerca de 0.001% as chances de fraude após o acréscimo de unidade nessas variáveis.

O destinatário do saldo inicial pré-transação (“oldbalanceDest”) também impacta as chances de incidência de fraude, no entanto de maneira a aumenta-las, ou seja, quanto maior for o saldo já existente naquela conta, maior será as chances de ocorrência de fraude, aumentando cerca de 0,0002% as chances de fraude após o acréscimo de unidade nessa variável.

Todavia, a variável “CASH_OUT”, a qual representa uma forma de pagamento, imprimiu um resultado um pouco diferente do modelo anterior, isto é, impacta as chances de incidência de fraude de forma bem mais expressiva, porém aumentando as chances de ocorrência de fraude. Podemos dizer que se a transação for um “CASH_OUT”, a chance de inadimplência aumenta em 33%.

Portanto pode-se concluir, basicamente o mesmo que o modelo anterior, isto é, tanto o valor a ser transacionado, quanto saldos iniciais, saldos pós transações, e tipo de pagamento podem dizer muito sobre uma transação ser efetiva ou não.

7 Conclusão final

No atual trabalho, atingiu-se o objetivo de estudar a aplicação da técnica de Regressão Logística para detectar transações fraudulentas além de obter conhecimento sobre como lidar, tratar e trabalhar com dados desbalanceados para chegar em um modelo eficiente e que será capaz de prever novas fraudes futuramente, ao pensar em cenários no mercado de trabalho.

Através das três aplicações realizadas ao longo do trabalho, Regressão Logística com os dados desbalanceados; Regressão Logística após balancear os dados utilizando a técnica de *undersampling* pelo método NearMiss; Regressão Logística após balancear os dados utilizando a técnica de *oversampling* pelo método SMOTE, foi possível entender a importância de utilizar dados balanceados para modelagem de interesse.

Na primeira aplicação, em que os dados não foram balanceados, foi obtido uma acurácia de 99%, no entanto o *recall* foi extremamente baixo, o *MCC* se comportou da mesma forma e a AUC praticamente igual ao ponto de corte de 0,5, ou seja, o modelo não conseguiu identificar as transações fraudulentas pois ele considerou a maioria das transações como boas, consequentemente sofreu um paradoxo da acurácia. Diante disso, caso esse modelo fosse utilizado para distinguir fraude e não fraude, ele daria a falsa impressão de estar tudo bem já que as transações fraudulentas seriam classificadas como transações efetivas.

Na segunda aplicação, em que os dados foram balanceados utilizando *undersampling*, todas as métricas obtidas foram bem altas, ou seja, o modelo ficou bom e conseguiu atender o estudo de interesse, prevendo boa parte das transações fraudulentas, logo balancear os dados fez total diferença nos resultados finais dos modelos gerados.

Na terceira aplicação, em que os dados foram balanceados utilizando *oversampling*, todas as métricas obtidas também foram altas, porém foram um pouco mais baixas do que na segunda aplicação, refletindo o comportamento de modelos utilizados com dados reais, isto porque a quantidade de dados foi muito maior, ou seja, utilizar esse método para balancear os dados também fez diferença no resultado final.

Em suma, as aplicações mostraram que ambas formas de balancear os dados funcionam, inclusive trouxeram os mesmos resultados, exceto pela variável

“CASH_OUT”, provavelmente por conta da quantidade de dados, afinal são 16.426 “contra” 1.253.596 linhas, e o que vai definir qual método a ser utilizado é o objetivo e o custo operacional caso esteja lidando com bases muito grandes.

Referências

AGRESTI, A. **Categorical Data Analysis.** 2nd ed. Ney Jersey: John Wiley & Sons, 2002.

AGRESTI, A. **An Introduction To Categorical Data Analysis.** 2nd ed. Ney Jersey: John Wiley & Sons, 2007.

BARELLA, V. H. **Técnicas para o problema de dados desbalanceados em classificação hierárquica.** 2015. 103 f. Dissertação (Mestrado) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2015.

BERALDI, F. **Atualização dinâmica de modelo de regressão logística binária para detecção de fraudes em transações eletrônicas com cartão de crédito**. 2014. 156 f. Dissertação (Mestrado) – Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2014.

BOLTON, R.; HAND, D. **Statistical Fraud Detection: A Review**. Statistical Science. vol. 17. No. 3. 235-255, 2002.

CARVO, T. T. **Técnicas estatísticas aplicadas na analítica de aprendizagem.** 2018. 65 p. Monografia (Graduação) – Universidade Estadual Paulista, Presidente Prudente, 2018.

COMO LIDAR COM DADOS DESBALANCEADOS EM PROBLEMAS DE CLASSIFICAÇÃO. Disponível em: https://medium.com/data-hackers/como-lidar-com-dados-desbalanceados-em-problemas-de-classifica%C3%A7%C3%A3o-17c4d4357ef9. Acesso em: 06 jan. 2022.

FRAUDE. Disponível em: https://michaelis.uol.com.br/moderno-portugues/busca/portugues-brasileiro/fraude. Acesso em: 29 abr. 2021.

FREITAS, P. H. **Regressão Logística na modelagem da probabilidade de vitória em jogos de futebol americano.** 2019. 54 p. Monografia (Graduação) – Universidade Federal de Uberlândia, Faculdade de Matemática, Uberlândia-MG, 2019.

HOSMER, D. W. LEMESHOW, S. **Applied Logistic Regression.** 2nd ed. Ney Jersey: John Wiley & Sons, 2000.

KAGGLE. Disponível em: https://kaggle.com/. Acesso em: 16 abr. 2021.

OLIVEIRA, P. H. M. A. **Detecção de fraude em cartões: um classificador baseado em regras de associação e regressão logística.** 2015. 100 f. Dissertação (Mestrado) – Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2015.

PARADÓXO DA ACURÁCIA. Disponível em: https://medium.com/turing-talks/paradoxo-da-acur%C3%A1cia-56baa75334f2. Acesso em: 06 jan. 2022.

QUAL A MELHOR MÉTRICA PARA AVALIAR OS MODELOS DE MACHINE LEARN. Disponível em: https://www.flai.com.br/juscudilio/qual-a-melhor-metrica-para-avaliar-os-modelos-de-machine-learning/. Acesso em: 13 jan. 2022.

SANTIAGO, G. P. **Um Processo para Modelagem e Aplicação de Técnicas Computacionais para Detecção de Fraudes em Transações Eletrônicas.** 2014. 130 f. Dissertação (Mestrado) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2014.

SYNTHETIC FINANCIAL DATASETS FOR FRAUD DETECTION. Disponível em: https://www.kaggle.com/ntnu-testimon/paysim1. Acesso em: 27 mar. 2021.

TSUCHIYA, I. **Regressão Logística Aplicada na Análise Espacial de Dados Arqueológicos.** 2002. 103 f. Dissertação (Mestrado) – Universidade Estadual Paulista, Presidente Prudente, 2002.