Universidade Federal de São Carlos
Centro de Ciências Exatas e de Tecnologia
Departamento de Estatística

ESTRATÉGIAS DE REAMOSTRAGEM APLICADAS EM PROBLEMAS DE CLASSIFICAÇÃO BINÁRIA DE DADOS FINANCEIROS

Rafael Setti Riedel Sturaro
Orientador: Prof. Dr. Ricardo Felipe Ferreira

Trabalho de Conclusão de Curso

Universidade Federal de São Carlos
Centro de Ciências Exatas e de Tecnologia
Departamento de Estatística

ESTRATÉGIAS DE REAMOSTRAGEM APLICADAS EM PROBLEMAS DE CLASSIFICAÇÃO BINÁRIA DE DADOS FINANCEIROS

Rafael Setti Riedel Sturaro
Orientador: Prof. Dr. Ricardo Felipe Ferreira

Trabalho de Conclusão de Curso apresentado ao Departamento de Estatística da Universidade Federal de São Carlos - DEs-UFSCar, como parte dos requisitos para obtenção do título de Bacharel em Estatística.

São Carlos
Agosto de 2023

Resumo

Muitos problemas de classificação binária exibem classes desbalanceadas, o que consiste em uma das duas classes ser significantemente mais numerosa do que a outra. A classe minoritária é a menos representativa, possuindo uma quantidade bem menor de indivíduos ou objetos do que a classe majoritária. Na conjuntura de análise de dados financeiros, os bancos de dados podem ser desbalanceados. Por exemplo, em análise de crédito os bancos de dados contém, em geral, mais observações referentes a clientes adimplentes (classe majoritária) do que em relação aos clientes inadimplentes (classe minoritária). Da mesma maneira, em detecção de fraude os bancos dados geralmente contém mais informações referentes a transações legítimas (classe majoritária) do que em relação a transações fraudulentas (classe minoritária). Esse desbalanceamento acarreta em um viés de classificação, já que os algoritmos de aprendizagem tendem a classificar melhor as observações do grupo majoritário. Nesse sentido, esse trabalho tem como proposta abordar estratégias de reamostragem (dentre elas, métodos de subamostragem, sobreamostragem e métodos híbridos) para classificar, através da regressão logística, os clientes solicitantes de crédito em adimplentes ou inadimplentes e transações em legítima ou fraudulentas. Pretendemos estudar comparativamente a performance da regressão logística na classificação de novas instâncias nos seguintes cenários: (i) conjunto de treinamento balanceado por meio de técnicas de subamostragem; (ii) conjunto de treinamento balanceado por meio de técnicas de sobreamostragem; e (iii) conjunto de treinamento balanceado por meio de técnicas híbridas de reamostragem. Vamos conduzir esse estudo comparativo em dois contextos: (i) utilizando todas as variáveis do conjunto de dados; e (ii) realizando a seleção de variáveis a partir das estimativas de máxima verossimilhança dos coeficientes da regressão logística com regularização ℓ_1.

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

O problema de aprendizagem estatística refere-se ao vasto conjunto de ferramentas cujo principal objetivo é aprender padrões a partir dos dados. Entre os vários contextos de aprendizagem, a aprendizagem supervisionada ocorre com frequência em diversas aplicações com dados reais. Grosso modo, as técnicas de aprendizagem supervisionada envolvem a construção de um modelo estatístico para predizer uma variável de saída baseada em uma ou mais variáveis de entrada. O problema é definido como sendo uma regressão se a variável de saída é numérica, e como sendo uma classificação se a variável de saída é categórica.

Muitos problemas de classificação binária exibem classes desbalanceadas, o que consiste em uma das duas classes ser significantemente mais numerosa do que a outra. A classe minoritária é a menos representativa, possuindo uma quantidade bem menor de indivíduos ou objetos do que a classe majoritária. Classificadores que performam bem com classes balanceadas geralmente falham quando as classes são desbalanceadas, pois são, em geral, a favor da classe majoritária e, geralmente, apresentam baixo desempenho na classificação de indivíduos ou objetos da classe minoritária (Wang et al., 2015). O desenvolvimento de técnicas confiáveis para distinguir corretamente a classe minoritária permanece uma área de pesquisa desafiadora (He e Ma, 2013; Krawczyk, 2016; Fernández et al., 2018). Muitas técnicas têm sido desenvolvidas para melhorar a performance dos classificadores quando expostos a classes desbalanceadas. Essas técnicas podem ser categorizadas em quatro grupos: (i) técnicas a nível do algoritmo de aprendizagem, que consiste em abordagens que buscam reformular os algoritmos de classificação para otimizar diferentes métricas de performance (Dembczynski et al., 2013); (ii) técnicas a nível dos dados, que consistem em abordagens baseadas na reamostragem dos dados a fim de

balancear a distribuição dos indivíduos ou objetos entre as classes (Chawla et al., 2002; Batista et al., 2004; Fernández et al., 2018; Napierała et al., 2010; Stefanowski e Wilk, 2008); (*iii*) técnicas sensíveis ao custo que consistem em abordagens que incorporam tanto modificações nos algoritmos de classificação quanto transformações no conjunto de dados, ambas levando em consideração custos de classificação incorreta (e, possivelmente, outros tipos de custo) (Chawla et al., 2008; Ling et al., 2006; Zhang et al., 2008); e (*iv*) técnicas baseadas na combinação de um conjunto de algoritmos de aprendizagem (Galar et al., 2011; Polikar, 2006).

Neste trabalho, vamos estudar, em um contexto de classes desbalanceadas, a performance do modelo de regressão logística (James et al., 2013) na classificação de novos indivíduos ou objetos em três cenários distintos. O mecanismo que utilizaremos para fixar esses cenários e, consequentemente, lidar com o desbalanceamento das classes são algumas técnicas a nível dos dados que fazem uso de métodos de reamostragem. Essas técnicas consistem em modificar o conjunto de treinamento desbalanceado a fim de produzir uma distribuição mais balanceada dos indivíduos ou objetos entre as classes, o que permite que os classificadores performem de maneira similar aos problemas de classificação balanceada. Na literatura especializada, muitos estudos têm mostrado empiricamente que, para os diversos tipos de classificadores, o balanceamento do conjunto de dados a partir de técnicas de reamostragem melhoram significantemente a performance desses classificadores quando comparado ao cenário em que não houve esse pré-processamento dos dados (Chawla et al., 2008; Estabrooks et al., 2004; García et al., 2012). Nesse sentido, uma das principais vantagens dessas técnicas é que elas não dependem do classificador escolhido.

As técnicas de reamostragem podem ser classificadas em três grupos: (*i*) técnicas de subamostragem, que consiste em abordagens que eliminam instâncias, em geral, da classe majoritária (Liu et al., 2008; Wilson, 1972); (*ii*) técnicas de sobreamostragem, que consiste em abordagens que criam instâncias, em geral, da classe minoritária (Chawla et al., 2002); e (*iii*) técnicas híbridas que combinam ambas as abordagens de reamostragem. Ao mesmo tempo que as técnicas de reamostragem criam uma distribuição mais balanceada dos dados, elas sofrem algumas desvantagens. Por exemplo, a maior desvantagem das técnicas de subamostragem é que elas podem descartar dados que sejam potencialmente úteis, fazendo com que decisões sobre os indivíduos ou objetos que estejam na fronteira entre a classe majoritária e minoritária sejam difíceis de ser tomada, resultando em uma perda da performance da classificação. Para as técnicas de sobreposição, uma vez que são

feitas cópias de instâncias já existentes, muitos autores concordam que essa abordagem pode levar à ocorrência de sobreajuste, o que afeta a performance da classificação. A fim de superar essas limitações, técnicas mais sofisticadas de reamostragem foram desenvolvidas. Nesse sentido, pretendemos estudar comparativamente a performance da regressão logística na classificação de novas instâncias nos seguintes cenários: (i) conjunto de treinamento balanceado por meio de técnicas de subamostragem; (ii) conjunto de treinamento balanceado por meio de técnicas de sobreamostragem; e (iii) conjunto de treinamento balanceado por meio de técnicas híbridas de amostragem.

Conjuntos de dados desbalanceados são muito comuns na área financeira, por exemplo, em classificação de crédito e detecção de fraude. A fim de diminuir os riscos e as incertezas que envolvem a concessão do crédito, instituições financeiras estão sempre buscando maneiras de aprimorarem seu processo de análise de créditos, sendo modelos de classificação de crédito umas das principais metodologias que corrobora com esse objetivo. Nessa mesma linha, instituições financeiras estão sempre aprimorando algoritmos que sejam capazes de detectar efetivamente transações fraudulentas. Dessa forma, uma maneira de minimizar as perdas decorrentes da inadimplência ou da fraude é utilizar classificadores que consigam ter bom desempenho na classificação de bons e maus clientes e de transações legítimas e fraudulentas. Porém, na execução desse processo nos deparamos com um grande desafio, em que observamos uma maior proporção de clientes adimplentes e transações legítimas (grupos majoritários) do que clientes inadimplentes e transações fraudulentas (grupos minoritários), o que pode impactar a performance dos classificadores. Nesse sentido, para o estudo de comparação que pretendemos desenvolver neste trabalho, vamos aplicar as metodologias supramencionadas em dois conjuntos de dados financeiros, um relativo a análise de crédito e outro relativo a detecção de fraude. Além disso, vamos conduzir esse estudo comparativo em dois contextos: (i) utilizando todas as variáveis do conjunto de dados; e (ii) realizando a seleção de variáveis a partir das estimativas de máxima verossimilhança dos coeficientes da regressão logística com regularização ℓ_1.

A principal contribuição deste trabalho é o estudo comparativo de métodos de reamostragem para classificação de dados desbalanceados. Embora, para muitos classificadores, não seja necessário o pré-processamento dos dados, pois muito deles são capazes de lidar diretamente com conjunto de dados desbalanceados; não há uma regra que nos diga qual é a melhor estratégia.

Este trabalho está organizado da seguinte maneira. No próximo capítulo, apresentamos os algoritmos de reamostragem que serão utilizados para o balanceamento do conjunto de treinamento. No Capítulo 3, explicamos como realizaremos a classificação de novas instâncias a partir da regressão logística e explicamos como as variáveis serão selecionadas a partir das estimativas de máxima verossimilhança dos coeficientes da regressão logística com regularização ℓ_1. As medidas que utilizaremos para avaliar a performance do classificador proposto também são apresentadas no Capítulo 3. No Capítulo 4, comparamos a performance da classificação do modelo logístico com e sem o pré-processamento dos dados. Para isso, utilizamos dois conjuntos de dados financeiros com níveis de desbalanceamento distintos. O Capítulo 5 encerra esta monografia com algumas considerações finais, conclusões e sugestões para estudos futuros.

Capítulo 2

Algoritmos de reamostragem

Em muitos problemas reais, encontramos conjuntos de dados cujas classes são desbalanceadas, isto é, onde há uma distribuição desproporcional das unidades amostrais nas diferentes classes da variável de interesse. Por exemplo, em classificação de crédito, observamos uma maior proporção de clientes adimplentes (grupo majoritário) do que clientes inadimplentes (grupo minoritário); e em detecção de fraude, observamos mais transações legítimas (grupo majoritário) do que transações fraudulentas (grupo minoritário). O desbalanceamento das classes pode impactar na performance de classificadores. Em geral, classificadores que performam bem com classes balanceadas falham quando as classes são desbalanceadas, pois são geralmente a favor da classe majoritária e, geralmente, apresentam baixo desempenho na classificação das unidades amostrais pertencentes à classe minoritária. Nesse sentido, vamos utilizar técnicas de reamostragem para realizar o balanceamento do conjunto de treinamento antes de realizar a classificação de novas instâncias. Neste trabalho, vamos estudar três técnicas de reamostragem: Tomek Link (subamostragem), SMOTE (sobreamostragem) e SMOTE + Tomek Link (híbrido). Apresentamos tais técnicas em mais detalhes nas seções seguintes.

2.1 Classificação para conjunto de dados desbalanceados

Classificação estatística é um problema recorrente no contexto de aprendizagem de máquina. Um dos desafios para os métodos de classificação consiste em lidar com dados desbalanceados que ao serem processados pelo algorítimo de classificação passam por

dificuldades ao encontrar a presença de uma classe com número de observações muito superior a outra. A classe com mais observações é chamada de **majoritária** e a outra, com menos observações, é chamada de classe **minoritária**.

Pensemos, por exemplo, em um cenário de operações de cartão de crédito em que a grande maioria das operações são legítimas e a outra pequena parte é composta por operações fraudulentas. Nessa área de atuação, é bastante difundido o desenvolvimento de classificadores para detectar operações fraudulentas. Nesse contexto, as transações genuínas formam a classe majoritária e as transações fraudulentas formam a classe minoritária.

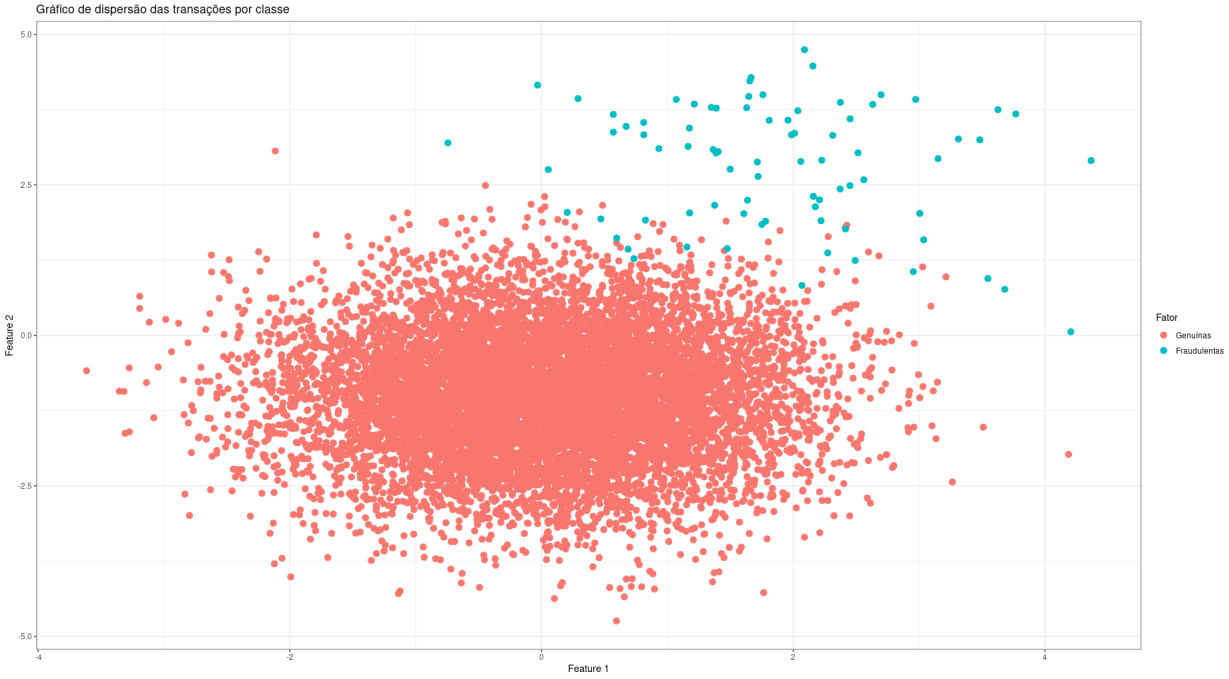

Figura 2.1: Gráfico de dispersão para transações de cartão de crédito. Os pontos em vermelho representam as transações legítimas e os pontos azuis, as transações fraudulentas.

Na Figura 2.1, observa-se a presença de uma classe majoritária, representada em vermelho e uma classe minoritária representada em azul. A dificuldade que o algoritmo encontra, por exemplo, em um contexto de detecção de fraude, é classificar como legítimo transações que na realidade são fraudulentas. Por outro lado, os algoritmos não encontram dificuldades em classificar como legítimas transações que de fato são legítimas. Portanto, os classificadores, em um contexto de dados desbalanceados, tendem a ser viesados em favor da classe majoritária, apresentando um baixo desempenho na classe majoritária.

Outro exemplo para esse tipo de problemática é o caso da classificação de clientes solicitante de crédito em adimplentes e inadimplentes. Similar ao exemplo anterior, nesse

caso, tem-se a presença de uma classe majoritária (adimplentes) e uma classe minoritária (inadimplentes). Uma das diferenças entre o que ocorre em problema de detecção de fraudes e concessão de crédito é o tamanho das classes. No caso da detecção fraudes, as classes são mais desbalanceadas do que aquelas do cenário de concessão de crédito. Essa diferença no desbalanceamento da amostra é ilustrada na Figura 2.2.

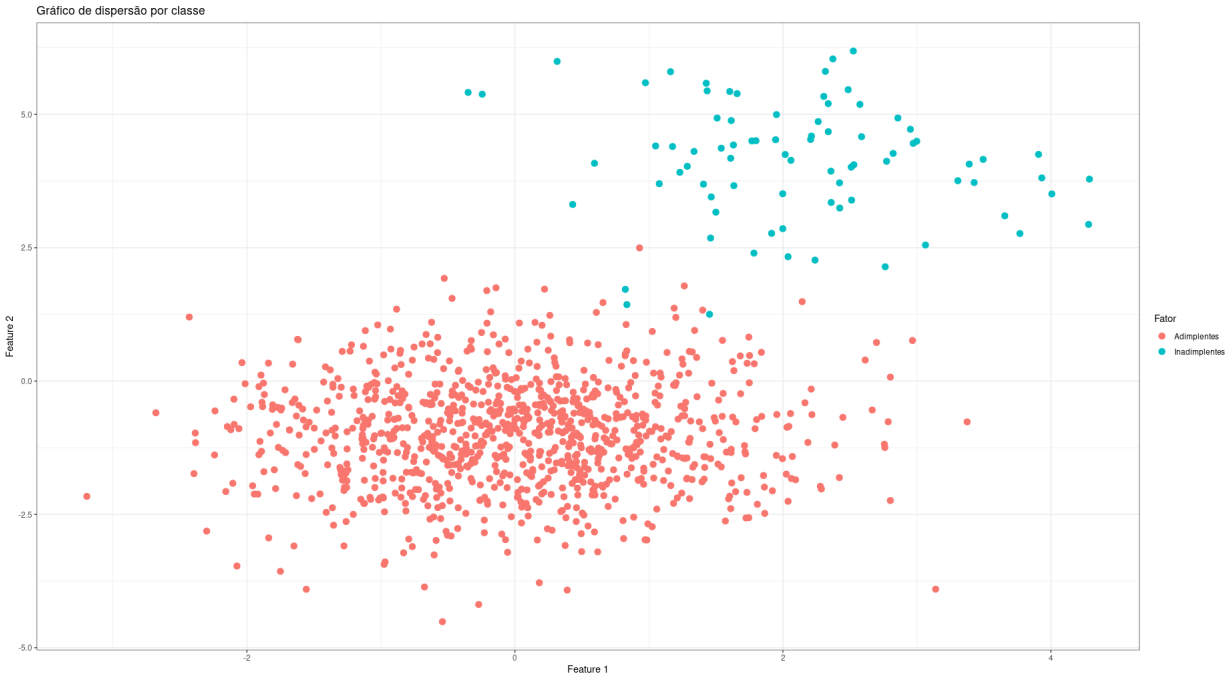

Figura 2.2: Gráfico de dispersão para operações de concessão crédito. Os pontos em vermelho representam os clientes adimplentes e os pontos azuis, os clientes inadimplentes.

Muitas técnicas tem sido desenvolvidas para melhorar a performance de classificadores em um contexto de dados desbalanceados. Essas técnicas podem ser classificadas em:

- Técnicas a **nível de algorítimo** (ou interna): Consiste em abordagens que buscam reformular o algoritmo para equilibrar o erro oriundo da presença de classes desbalanceadas;

- Técnicas a **nível de dados** (ou externa): consiste em métodos que pré-processam os dados da amostra antes de aplicar da classificação, utilizando de reamostragem a fim de encontrar uma amostra mais balanceada;

- Técnicas **sensíveis ao custo**: buscam realizar uma mescla entre os dois métodos citados anteriormente, ambos levando em consideração custos de classificações incorretas.

Além das 3 técnicas citadas, existe uma 4ª alternativa que vem da combinação de vários classificadores ponderados por sua performance. Neste trabalho, serão estudadas as técnicas de reamostragem que são técnicas a nível de dados aplicados em um contexto de concessão de crédito.

2.2 Algoritmos de reamostragem

Classificadores que performam bem com classes balanceadas geralmente falham quando as classes são desbalanceadas, pois são, em geral, a favor da classe majoritária e, geralmente, apresentam baixo desempenho na classificação de indivíduos ou objetos da classe minoritária. Conjuntos de dados desbalanceados são muito comuns na área financeira, por exemplo, em classificação de crédito, em que observamos uma maior proporção de clientes adimplentes (grupo majoritário) do que clientes inadimplentes (grupo minoritário), o que pode impactar a performance dos classificadores. Nesse sentido, para o estudo de comparação que pretendemos desenvolver neste trabalho, vamos aplicar as técnicas de reamostragem. Em geral, os algoritmos de reamostragem apresentam um bom desempenho quando incluídos em problemas de classificação e exigem uma menor capacidade computacional para serem implementados. Dentre os métodos mais difundidos na aplicação de reamostragem, temos **subamostragem**, **sobreamostragem** e, por fim, **métodos híbridos**, esse último busca equilibrar técnicas utlizadas em cada uma das duas outras metodologias mencionadas.

2.2.1 Subamostragem

Quando realizamos subamostragem, removemos unidades amostrais pertencentes à classe majoritária a fim de balancear o conjunto de treinamento. Contudo, escolher aleatoriamente as unidades amostrais a serem removidas pode não ser uma boa estratégia, pois pode vir acompanhado de uma perda considerável de informação útil para a classificação, levando a uma baixa performance do classificador Dessa forma, alguns métodos de subamostragem foram desenvolvidos. Nesse trabalho, usamos o método de subamostragem Tomek Link, proposto por Tomek (1976).

O método Tomek Link tem como principal objetivo eliminar unidades amostrais da classe majoritátia que sejam similares a unidades amostrais da classe minoritária. Nesse sentido, podemos afirmar que estamos interessados em eliminar unidades amostrais que

estejam em uma zona de confusão.

Se $d : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}_+$ é uma distância, então o algoritmo Tomek Link é dado de acordo com os seguintes passos:

1. Considere $d(\boldsymbol{x}^*, \tilde{\boldsymbol{x}})$ a distância euclidiana de $\boldsymbol{x}^*$ (associado a uma unidade amostral da classe majoritária) para $\tilde{\boldsymbol{x}}$ (associado a uma unidade amostral da classe minoritária);

2. Se não há uma observação $(\boldsymbol{x}_k, y)$ que satisfaça a seguinte condição

$$d(\boldsymbol{x}^*, \boldsymbol{x}_k) < d(\boldsymbol{x}^*, \tilde{\boldsymbol{x}}) \text{ ou } d(\tilde{\boldsymbol{x}}, \boldsymbol{x}_k) < d(\boldsymbol{x}^*, \tilde{\boldsymbol{x}}),$$

   então, o par $(\boldsymbol{x}^*, \tilde{\boldsymbol{x}})$ é dito ser um Tomek Link. Assim que formado todos os pares Tomek Link presentes, elimina-se os indivíduos da classe majoritária para cada um desses pares.

Ou seja, o algorítmo busca os elementos amostrais que sejam mais semelhantes entre si de acordo com as covariáveis do modelo e que divirjam em suas classes.

2.2.2 Sobreamostragem

A sobreamostragem consiste em métodos que aumentam o conjunto de treinamento a partir da inclusão de réplicas de observações pertencentes à classe minoritária. Este método também é simples de ser implementado e, diferente dos métodos de subamostragem, não apresenta perda de informação que pode ser útil para a classificação. Todavia, tem a desvantagem de ajustar um classificador que performe bem no conjunto de validação, mas não generalize tão bem para clientes desconhecidos. Além disso, há um custo computacional adicional quando o desbalanceamento entre as classes é grande. Estes métodos tendem a ser mais eficientes justamente quando é observada um desbalanceamento severo entre as classes. Um exemplo de método de sobremamostragem que vem mostrando ser bastante eficiência é o SMOTE (do inglês *Synthetic Minority Oversampling Technique*) que foi proposto por Chawla et al. (2002).

Em uma clássica técnica de sobreamostragem, a classe minoritária é duplicada a partir dos dados originais da população. Contudo, enquanto balanceia as classes dentro do conjunto de treinamento, esse método não traz nenhuma nova informação relevante quanto

a variabilidade dos novos dados, não agregando suficientemente para o modelo de classificação. O método SMOTE, no entanto, consiste em utilizar do KNN (do inglês, K *Nearest Neighbour*) para gerar indivíduos ou objetos sintéticos para a classe minoritária da forma como descrita a seguir.

Considere o conjunto

$$\mathcal{C} := \{(\boldsymbol{x}_i, y_i) \in \mathbb{R}^p \times \{0,1\} : i = 1, 2, \ldots, n\}$$

que representa uma amostra observada com N indivíduos, em que $\boldsymbol{x}_i \in \mathbb{R}^p$ é o vetor de covariáveis observadas relativo a i-ésima unidade amostral e $y_i \in \{0,1\}$ é a sua respectiva classe. Defina também o subconjunto $\mathcal{H}$ de $\mathcal{C}$ que contém todos os indivíuos da classe minoritária, isto é,

$$\mathcal{H} := \{(\boldsymbol{x}_i, 1) \in \mathbb{R}^p \times \{0,1\} : i = 1, 2, \ldots, n_1\}.$$

Se $d : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}_+$ é uma distância, então o algoritmo SMOTE é dado de acordo com os seguintes passos:

1. Define-se n_{min} como sendo o número total de observações presentes na classe minoritária após o processamento do algoritmo;

2. Em seguida, identifica-se um valor de K para performar o KNN a fim de estabelecer n_{min} conforme foi definido;

3. Performa-se o KNN da seguinte forma:

   (a) Seleciona-se aleatoriamente um indivíduo $(\boldsymbol{x}^*, 1) \in \mathcal{H}$;

   (b) A partir da observação selecionada aleatoriamente em $\mathcal{H}$, encontra-se os K vizinhos mais próximos $\tilde{\boldsymbol{x}}_1, \tilde{\boldsymbol{x}}_2, \ldots, \tilde{\boldsymbol{x}}_k \in \mathcal{C}$ de forma que, sem perda de generalidade,

   $$d(\boldsymbol{x}^*, \tilde{\boldsymbol{x}}_1) \leq d(\boldsymbol{x}^*, \tilde{\boldsymbol{x}}_2) \leq \cdots \leq d(\boldsymbol{x}^*, \tilde{\boldsymbol{x}}_K) \leq d(\boldsymbol{x}^*, \tilde{\boldsymbol{x}}),$$

   qualquer que seja $\tilde{\boldsymbol{x}} \in \mathcal{C}$.

   (c) E, por fim, para cada $k = 1, 2, \ldots, K$, gera-se uma nova obervação sintética $(\boldsymbol{x}_s, 1) \in \mathbb{R}^p \times \{0,1\}$ de forma que

   $$d(\boldsymbol{x}^*, \boldsymbol{x}_s) = u + d(\boldsymbol{x}^*, \tilde{\boldsymbol{x}_k}),$$

em que $u \in (0, 1]$ é um valor aleatório gerado a partir de uma variável aleatória U com distribuição uniforme em $(0, 1]$.

(d) Repetir o processo para outras observações aleatorimente selecionadas do conjunto $\mathcal{H}$ composta pelos indivíduos da classe minoritária até atingir o valor N_{min} fixado.

Basicamente, o algoritmo busca gerar indivíduos sintéticos que sejam próximos dos atuais indivíduos da classe minoritária. Assim, o algoritmo procura corrigir o problema de *overfitting* que pode vir a ocorrer no processo de classificação e também agrega mais a respeito da variabilidade, diferente do exemplo em que replicamos todos os indivíudos presentes na classe minoritária.

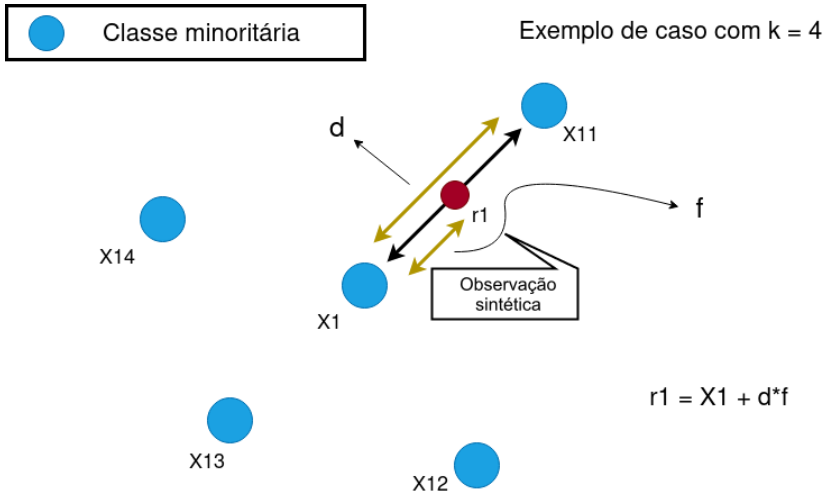

Figura 2.3: Ilustração do algoritmo SMOTE.

A Figura 2.3 mostra como se dá o processo de sintetização de novos elementos dentro da classe minoritária.

2.2.3 Métodos híbridos de reamostragem

Os métodos híbridos consistem em uma combinação dos dois métodos mencionados anteriormente. Estes métodos são os que, em geral, levam a melhores resultados. Contudo, esta abordagem exige pré-processamento mais dedicado para os dados em estudo e pode apresentar as desvantagens inerentes aos outros dois métodos. Um exemplo de método híbrido é a combinação do SMOTE com o Tomek Link, ambos já apresentados nesse trabalho.

O SMOTE + Tomek Link consiste basicamente na aplicação conjunta dos métodos SMOTE com o Tomek Link.

O algoritmo SMOTE + Tomek Link é dado de acordo com os seguintes passos:

1. Primeiramente, performa-se o algoritmo do SMOTE;

2. Em seguida, performa-se o Tomek Link.

Nesse sentido, o algoritmo funciona de forma que, após performar a sobreamostragem com o método SMOTE, em seguida, aplica da subamostragem a partir da busca pelos pares Tomek Link a fim de remover os indivíduos da classe majoritária presente em cada um desses pares.

Capítulo 3

Classificação e seleção de variáveis

Em dados financeiros, uma má gestão da concessão de crédito e de detecção de fraudes pode acarretar em perdas significativas de capital. Nesse contexto, podemos utilizar modelos estatísticos afim de predizer, baseado nas covariáveis disponíveis, se a transação é legítima ou fraudulenta (no caso de fraudes) ou se o cliente é adimplente ou inadimplente (no caso de crédito). Nesse sentido, baseado em um histórico de estudos nessa área, nesse capítulo vamos abordar técnicas que já se mostraram ser uma boa solução para esse tipo de problema.

3.1 Regressão logística

A regressão logística é um modelo de regressão que visa estudar a relação de uma variável binária com um conjunto de covariáveis (Hosmer et al., 1989), as quais podem ser quantitativas ou qualitativas.

Sejam p e n números naturais não-nulos tais que $p < n$. Considere um conjunto de dados composto por p covariáveis, que descrevem características de n unidades amostrais, e sejam $Y_1, Y_2, \ldots, Y_n$ variáveis aleatórias independentes definidas em um espaço de probabilidade (Ω, F, P) tais que

$$Y_i = \begin{cases} 1, & \text{se a } i\text{-ésima unidade amostral pertence à classe minoritária,} \\ 0, & \text{se a } i\text{-ésima unidade amostral pertence à classe majoritária.} \end{cases}$$

para todo $i = 1, 2, \ldots, n$.

Defina $\boldsymbol{X}$ como sendo uma matriz real de ordem $n \times p$ tal que x_{ij} é o elemento que está

na i-ésima linha e j-ésima coluna da matriz $\boldsymbol{X}$ e representa o valor da j-ésima covariável para a i-ésima unidade amostral, em que $i = 1, 2, \ldots, n$ e $j = 1, 2, \ldots, p$. Dessa forma, definimos $\boldsymbol{Y}$ como sendo um vetor aleatório de ordem $n \times 1$ tal que Y_i é o elemento que está na i-ésima linha do vetor $\boldsymbol{Y}$ e representa a variável aleatória binária que indica se a i-ésima unidade amostral pertence à classe majoritária ou minoritária, em que $i = 1, 2, \ldots, n$. De forma matricial, podemos escrever, portanto,

$$\boldsymbol{X} := \begin{bmatrix} x_{11} & \ldots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \ldots & x_{np} \end{bmatrix} \quad \text{e} \quad \boldsymbol{Y} := \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}.$$

Definido um plano amostral, podemos coletar uma amostra com n unidades amostrais e observar em cada um delas a variável indicadora Y e cada uma das p covariáveis especificadas no estudo. Seja $(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_n, y_n)$ a amostra observada, em que $y_i \in \{0, 1\}$ e $\boldsymbol{x}_i$ é um vetor real de ordem $1 \times p$ tal que cada entrada do vetor representa o valor de uma das p covariáveis observada na i-ésima unidade amostral, em que $i = 1, 2, \ldots, n$. Utilizando um conjunto de treinamento, isto é, um subconjunto da amostra observada, desejamos encontrar um classificador que tenha boa performance em discriminar novas unidades amostrais entre as classes majoritária e minoritária. Para tal, vamos utilizar a regressão logística, que tem sido, na prática, um dos métodos mais utilizados para detecção de inadimplência e fraude. Assim, modelamos a probabilidade de uma nova unidade amostral pertencer à classe minoritária, condicionada às observações das covariáveis consideradas no estudo, como uma função dessas covariáveis, ou seja,

$$P(Y = 1|\boldsymbol{x}) = \pi(\boldsymbol{x}) := \frac{e^{\beta_0 + \boldsymbol{x}^T \boldsymbol{\beta}}}{1 + e^{\beta_0 + \boldsymbol{x}^T \boldsymbol{\beta}}},$$

em que $\beta_0 \in \mathbb{R}$ representa o intercepto, $\boldsymbol{\beta} := (\beta_1, \beta_2, ..., \beta_p)^T$ é um vetor real de ordem $p \times 1$ que representa os coeficientes associados a cada uma das p covariáveis consideradas no estudo e $\boldsymbol{x}$ é o vetor real de ordem $1 \times p$ que representa o valor de cada uma das p covariáveis observadas na nova unidade amostral em questão.

Neste trabalho, vamos estimar os parâmetros β_0 e $\boldsymbol{\beta}$ do modelo de regressão logística ou, equivalentemente, treinar o classificador logístico, utilizando o método da máxima verossimilhança. Para isso, assumiremos que $Y_1, Y_2, ..., Y_n$ são variáveis aleatórias independentes e $Y_i|\boldsymbol{X_i} = \boldsymbol{x_i} \sim Bernoulli(\pi(\boldsymbol{x}_i))$, em que $\boldsymbol{X_i}$ é o vetor de covariáveis associadas

a i-ésima unidade amostral. Sendo assim, podemos escrever a distribuição de $Y_i|\boldsymbol{X_i} = \boldsymbol{x}_i$ da seguinte forma:

$$P(Y_i = y_i|\boldsymbol{x_i}) = \pi(\boldsymbol{x_i})^{y_i}(1 - \pi(\boldsymbol{x_i}))^{1-y_i}\mathbb{I}_{\{0,1\}}(y_i)$$

em que $\mathbb{I}$ denota a função indicadora.

Dessa maneira, para cada matriz $\boldsymbol{X}$, de dimensão $n \times p$, que representa as observações das p covariáveis das n transações, resultando em uma função de verossimilhança L dada por

$$\begin{aligned} L(\beta_0, \boldsymbol{\beta}|\boldsymbol{X}) &:= \prod_{i=1}^{n} \left[\pi(\boldsymbol{x_i}^{y_i}(1 - \pi(\boldsymbol{x_i}))^{1-y_i}\mathbb{I}_{\{0,1\}}(y_i)\right] \\ &= \prod_{i=1}^{n} \left[\left(\frac{e^{\beta_0 + \boldsymbol{x}_i^T\boldsymbol{\beta}}}{1 + e^{\beta_0 + \boldsymbol{x}_i^T\boldsymbol{\beta}}} \right)^{y_i} \left(1 - \frac{e^{\beta_0 + \boldsymbol{x}_i^T\boldsymbol{\beta}}}{1 + e^{\beta_0 + \boldsymbol{x}_i^T\boldsymbol{\beta}}} \right)^{1-y_i} \mathbb{I}_{\{0,1\}}(y_i) \right]. \end{aligned}$$

Dessa forma, a função ℓ de log-verossimilhança é dada por

$$\ell(\beta_0, \boldsymbol{\beta}|\boldsymbol{X}) := \sum_{i=1}^{n} \left[\log \left(1 - \frac{e^{\beta_0 + \boldsymbol{x}_i^T\boldsymbol{\beta}}}{1 + e^{\beta_0 + \boldsymbol{x}_i^T\boldsymbol{\beta}}} \right) + y_i(\beta_0 + \boldsymbol{x}_i^T\boldsymbol{\beta}) \right],$$

e, consequentemente, os estimadores de máxima verossimilhança são definidos da seguinte forma

$$(\hat{\beta}_0, \hat{\boldsymbol{\beta}}) := \arg \min_{\beta_0 \in \mathbb{R}, \boldsymbol{\beta} \in \mathbb{R}^p} -\ell(\beta_0, \boldsymbol{\beta}|\boldsymbol{X}).$$

O classificador logístico obtido através do método de máxima verossimilhança é, portanto, definido da seguinte maneira

$$\hat{\pi}(\boldsymbol{x}^*) := \frac{e^{\hat{\beta}_0 + (\boldsymbol{x}^*)^T\hat{\boldsymbol{\beta}}}}{1 + e^{\hat{\beta}_0 + (\boldsymbol{x}^*)^T\hat{\boldsymbol{\beta}}}}, \tag{3.1}$$

em que $\boldsymbol{x}^*$ é o vetor de covariáveis observadas em uma nova unidade amostral. Nesse sentido, a classificação $y^* \in \{0, 1\}$ do novo cliente é dada a partir da seguinte regra de decisão:

$$y^* = 1 \Leftrightarrow \hat{\pi}(\boldsymbol{x}^*) \geq c,$$

em que $c \in [0, 1)$ é uma constante pré-fixada.

3.2 Regressão logística com regularização ℓ_1 dos coeficientes

Ao estabelecer uma restrição aos coeficientes do classificador é possível que haja uma redução na variância desses estimadores ao passo que um aumento insignificante ocorra no seu viés, o que pode levar a um aumento na precisão da classificação de unidades amostrais que não tenham sido utilizadas no conjunto de treinamento. Além disso, é possível que uma ou mais covariáveis envolvidas no estudo não estejam de fato associadas com a variável resposta (variável que indica a qual classe a unidade amostral pertence). Incluir variáveis irrelevantes no modelo aumenta a variabilidade dos valores preditos e leva a uma complexidade desnecessária dificultando, por exemplo, a interpretabilidade do classificador. Nesse sentido, vamos utilizar a regularização ℓ_1 nos coeficientes do modelo de regressão logística a fim de realizar a seleção de variáveis e aumentar a precisão da classificação.

Nesse contexto, vamos estimar os parâmetros β_0 e $\boldsymbol{\beta}$ do modelo de regressão logística ou, equivalentemente, treinar o classificador logístico, utilizando o método da máxima verossimilhança e impondo uma regularização sobre os coeficientes de modo a reduzir suas estimativas a zero. Assim, considere $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_m, y_m)$ como sendo o conjunto de treinamento. Nesse caso, os estimadores de máxima verossimilhança são definidos da seguinte forma

$$(\hat{\beta}_0, \hat{\boldsymbol{\beta}}) := \arg \min_{\beta_0 \in \mathbb{R}, \boldsymbol{\beta} \in \mathbb{R}^p} \left\{ -\ell(\beta_0, \boldsymbol{\beta} | \boldsymbol{X}) + \lambda \sum_{i=1}^{p} |\beta_i| \right\},$$

em que $\beta_0 \in \mathbb{R}$ e $\boldsymbol{\beta}$ são os coeficientes do modelo e $\lambda > 0$ é um parâmetro escolhido a partir de validação cruzada (James et al., 2013) de modo a diminuir o risco de predição. Quando λ é suficientemente grande, alguns coeficientes em $\hat{\boldsymbol{\beta}}$ podem se igualar a zero, ou seja, além de comprimir as estimativas dos coeficientes, esse tipo de regularização zera as estimativas de alguns coeficientes, fazendo com que algumas variáveis sejam excluídas do modelo, tornando-o mais eficiente e mais simples de ser interpretado.

O classificador logístico, neste caso, é definido da mesma maneira que na equação (3.1). Todavia, agora, leva em consideração apenas as covariáveis que foram selecionadas.

3.3 Medidas de performance

Tendo em vista o pré-processamento dos dados utilizados para fazer este trabalho, o objetivo final é entender, a partir de conjuntos de dados reais, para qual dos métodos de reamostragem, o classificador logístico obteve a melhor performance. Para avaliar a performance da classificação da regressão logística, vamos utilizar medidas baseadas na matriz de confusão, a área sobre a curva ROC e a estatística de Komolgorov-Smirnov.

3.3.1 Matriz de confusão

Em problemas de classificação, quando estamos estudando duas classes, a matriz de confusão é uma matriz com duas linhas e duas colunas que informa o número de verdadeiros positivos, falsos negativos, falsos positivos e verdadeiros negativos. As linhas da matriz representam os valores observados das classes e nas colunas os valores previstos, ou vice e versa. Nesse sentido, a matriz de confusão é um caso especial de tabela de contigência em duas dimensões. A denominação *matriz de confusão* vem do fato que é fácil ver, a partir dessa matriz, se o classificador está confundindo as duas classes durante o processo de classificação. Um exemplo de matriz de confusão é dado pela Tabela 3.3.1.

Classe prevista pelo classificador	Valores observados		
	Positivo	Negativo	Total
Positivo	VP	FP	PP
Negativo	FN	VN	PN
Total	P	N	n

Tabela 3.1: Um exemplo de matriz de confusão.

Na matriz de confusão apresentada anteriormente, temos n como o número de unidades amostrais presentes no conjunto de validação, VP é o número de unidades amostrais que são verdadeiros positivos, FP é o número de unidades amostrais que são falsos positivos, FN é o número de unidades amostrais que são falsos negativos, VN é o número de unidades amostrais que são verdadeiros negativos, P é o número de unidades amostrais pertencentes à classe positivo, N é o número de unidades amostrais pertencentes à classe negativo, PP é o número de unidades amostrais classificadas como positivas e PN é o número de unidades amostrais classificadas como negativas.

3.3.2 Medidas de performance baseadas na matriz de confusão

Na análise preditiva, a matriz de confusão é frequentemente utilizada, pois permite uma análise mais detalhada da performance do processo de predição, uma vez que não se baseia simplesmente na proporção de classificações corretas (acurácia). A acurácia, por exemplo, pode levar a conclusões enganosas quando o conjunto de dados é desbalanceado. Nesse sentido, trazemos a seguir algumas medidas baseadas na matriz de confusão que serão utilizadas para avaliar a performance do processo de classificação proposto neste trabalho.

Levando em consideração as mesmas nomenclaturas utilizadas no exemplo de matriz de confusão dado na subseção anterior, temos as seguintes métricas:

- VP: é o número de unidades amostrais que são verdadeiros positivos;
- VN: é o número de unidades amostrais que são verdadeiros negativos;
- FP: é o número de unidades amostrais que são falsos positivos;
- FN: é o número de unidades amostrais que são falsos negativos.

A partir dessas métricas, definimos as seguintes medidas de performance:

- **Acurácia (ACC)**: é uma medida que quantifica a proporção de unidades amostrais que foram classificadas corretamente. Matematicamente,

$$ACC := \frac{VN + VP}{n}.$$

- **Taxa de erro (TE)**: é uma medida que quantifica a proporção de unidades amostrais que foram classificadas incorretamente. Matematicamente,

$$TE := \frac{FN + FP}{n} = 1 - ACC.$$

- **Sensibilidade (S)**: é uma medida que quantifica a proporção de unidades amostrais positivas que foram classificadas corretamente. Matematicamente,

$$S := \frac{VP}{VP + FN} = \frac{VP}{P}.$$

- **Especificadade (E)**: é uma medida que quantifica a proporção de unidades amostrais negativas que foram classificadas corretamente. Matematicamente,

$$E := \frac{VN}{VN + FP} = \frac{VN}{N}.$$

- **Valor predito positivo (VPP)**: é uma medida que quantifica a proporção das unidades amostrais classificadas como positivos que foram classificadas corretamente. Matematicamente,

$$VPP := \frac{VP}{VP + FP} = \frac{VP}{PP}.$$

- **Valor Preditivo Negativo (VPN)**: é uma medida que quantifica a proporção de unidades amostrais classificadas como negativas que foram classificados corretamente. Matematicamente,

$$VPN := \frac{VN}{VN + FN} = \frac{VN}{PN}.$$

Vale ressaltar que como todos as medidas representam proporções, todas valem no mínimo 0 e no máximo 1. Com exceção da taxa de erro, que indica uma boa performance da classificação para valores próximos de 0, todas as outras medidas de performance listadas anteriormente indicam uma boa performance de classificação para valores próximos de 1.

As medidas definidas anteriormente são altamente afetadas pelo desbalanceamento das classes. Nesse sentido, a literatura tem proposto medidas mais adequadas para avaliar a performance dos classificadores a fim de superar o problema gerado pela baixa performance dos classificadores na classificação de unidades amostrais pertencentes à classe minoritária. As métricas utilizadas neste trabalho são: a G-média e o coeficiente de Matthews.

- G-média: A G-média foi sugerida por Kubat et al. (1997) e é calculada a partir da média geométrica entre a sensibilidade e especificidade. Matematicamente,

$$\text{G-Média} = \sqrt{S \times E}.$$

A G-média será alta quando ambas medidas, sensibilidade e especificidade, forem altas ou quando a diferença entre elas for pequena. Por outro lado, a G-média

será baixa se a diferença entre as medidas for grande. Assim, o classificador será penalizado caso seu poder de discriminação entre as classes seja baixo.

- Coeficiente de Matthews (MCC): O coeficiente de Matthews (MCC), proposto por Matthews (1975), é uma métrica que calcula a correlação entre o valor observado na amostra de teste e os valores preditos pelo classificador, conforme indicados na matriz de confusão. É uma medida que retorna valores entre -1 e 1, sendo que o valor 1 indica alta relação entre a previsão e o valor observado, ou seja, o classificador performa parecido com o real. Valores próximos a -1 indicam uma relação inversa do classificador com as reais classes. Valores próximos a 0 indicam um classificador problemático, com poder preditivo quase nulo. Matematicamente, o MCC é dado tal que

$$\text{MCC} = \frac{VP \times VN - FP \times FN}{\sqrt{(VP + FP) \times (VP + FN) \times (VN + FP) \times (VN + FN)}}$$

Note que o MCC considera as 4 medidas da matriz de confusão VP, VN, FP e FN, consequentemente, o cálculo do MCC considera que as duas classes tem igual relevância. Vale ressaltar que só obtemos um valor alto para o MCC se duas das medidas da matriz de confusão (VP e VN) forem altas e as outras duas (FP e FN) forem baixas.

3.3.3 Área embaixo da curva ROC (AUC)

A curva ROC é um gráfico utilizado para avaliar o desempenho de um classificador binário. Essa curva é criada plotando a taxa de verdadeiros positivos em relação à taxa de falsos positivos considerando uma sequência de pontes de cortes utilizados na classificação das unidades amostrais do conjunto de validação. A taxa de verdadeiros positivos é a sensibilidade (S) e a taxa de falsos positivos também é conhecida como probabilidade de alarme falso e pode ser calculada como $1 - E$, em que E representa a especificidade. Na Figura 3.2, podemos ver um exemplo genérico de curva ROC.

O nome ROC tem um caráter histórico e vem do inglês *Receiver Operating Characteristic*. Foi um método originalmente desenvolvido para operadores de radar militar.

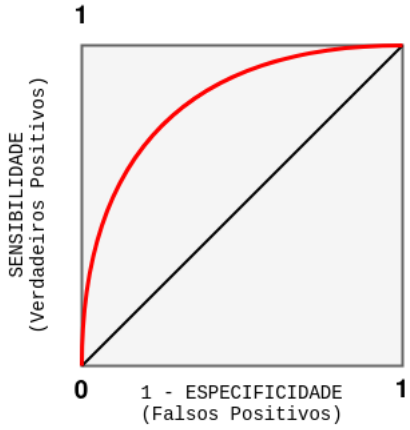

Figura 3.1: Um exemplo de curva ROC.

O desempenho geral de um classificador, levando em consideração todos os pontes de corte possíveis, é dado pela área sob a curva ROC. Vamos denotar essa área por AUC, do inglês *Area Under Curve*. Uma curva ROC ideal é aquela que contorna o canto superior esquerdo da Figura 3.2, então quanto maior a área sob a curva ROC, maior será a AUC e, portanto, melhor será o classificador.

Dessa forma, interpreta-se o valor da AUC de maneira que valores próximos a 0 indicam um classificador que erra muito, já uma AUC com valores próximos a 1 indica um classificador com alta taxa de acerto e baixa taxa de erro.

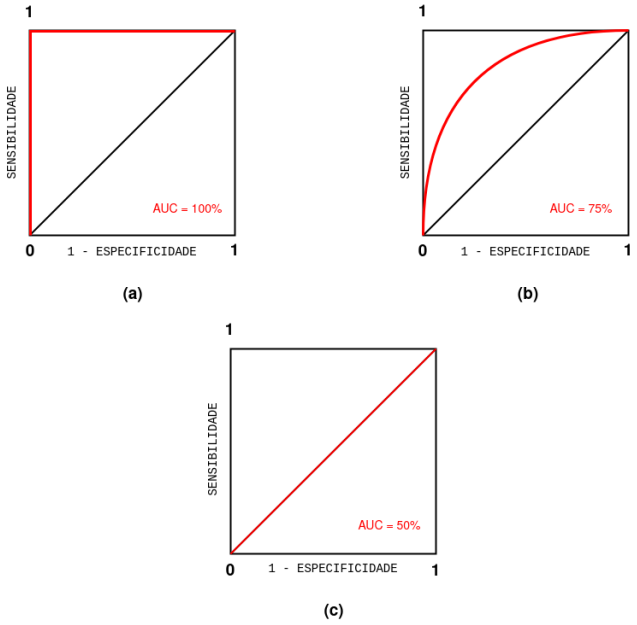

Figura 3.2: Áre sob a curva ROC (AUC). (a) classificador ideal com $AUC = 100\%$, (b) classificador bom com $AUC = 75\%$ e (c) classificador sem capacidade significativa de separar as unidades amostrais entre as classes com $AUC = 50\%$.

3.3.4 Estatística de Kolmogorov-Smirmnov (KS)

Na área financeira, uma medida de performance muito difundida para avaliar o quão bom é um classificador é a estatística de Kolmogorov-Smirnov (KS) (Sicsú, 2010). Essa estatística quantifica a distância entre a função de distribuição empírica dos valores preditos pertencentes à classe positiva e a função de distribuição empírica dos valores preditos pertencentes à classe negativa. Sendo assim, definimos $F_p : [0,1] \to [0,1]$ como sendo a distribuição empírica dos valores preditos classificadas como positivos e, de forma semelhante, definimos $F_n : [0,1] \to [0,1]$ a distribuição empírica dos valores preditos classificadas como negativos. Nesse sentido, definimos $\mathcal{V}$ como o conjunto de validação, e os subconjuntos $\mathcal{P}$ e $\mathcal{N}$ como, respectivamente, o conjunto das unidades amostrais do conjunto de validação que pertencem à classe positivas e o conjunto das unidades amostrais do conjunto de validação que pertencem à classe negativas. Para cada unidade amostral $i \in \mathcal{P}$, está associada uma estimativa $\hat{\theta}_p(i)$ da probabilidade dessa unidade amostral pertencer à classe positiva e para cada unidade amostral $j \in \mathcal{N}$, está associada uma estimativa $\hat{\theta}_n(j)$ dessa unidade amostral pertencer à classe negativa. Assim, para cada $c \in [0,1]$, podemos definir as funções de distribuições empíricas da seguinte maneira

$$F_p(c) := \frac{\left|i \in \mathcal{P} : \hat{\theta}_p(i) \leq c\right|}{|\mathcal{P}|} \quad \text{e} \quad F_n(c) := \frac{\left|j \in \mathcal{N} : \hat{\theta}_n(j) \leq c\right|}{|\mathcal{N}|},$$

em que $|\quad|$ denota a cardinalidade do conjunto.

A estatística KS é dada pela distância absoluta máxima entre as proporções acumuladas ao longo das estimativas observadas pelo classificador em análise, isto é,

$$KS := \max_{c \in [0,1]} \left|F_p(c) - F_n(c)\right|.$$

Essa estatística varia entre 0 e 1 e quanto maior o valor da estatística KS, melhor a performance do classificador, uma vez que maior será a distância entre as distribuições empíricas das classes.

Capítulo 4

Aplicações em dados reais

Nessa etapa, aplicamos o modelo de regressão logística em conjuntos de dados reais e comparamos sua performance na classificação de novas instâncias em quatro cenários distintos: (i) sem considerar qualquer método de pré-processamento dos dados; (ii) utilizando o algoritmo Tomek Link para realizar a remoção de pares que representam casos de sobreposição ou ambiguidade na fronteira de decisão do classificador, onde uma instância de uma classe é cercada por instâncias de outra classe; (iii) utilizando o algoritmo SMOTE para sintetizar novas instâncias da classe minoritária ao criar amostras sintéticas através da combinação de exemplos existentes e (iv) combinando o algoritmo Tomek Link com algoritmo SMOTE. Além disso, também avaliaremos a performance da regressão logística nesses cenários considerando uma regularização ℓ_1 dos coeficientes do modelo de regressão. Para medir o desempenho dos classificadores empregados, utilizaremos o AUC, KS, coeficiente de correlação de Matthews (MCC), G-Média e medidas fundadas na matriz de confusão, como sensibilidade, especificidade, valor preditivo positivo (VPP) e valor preditivo negativo (VPN). Inicialmente, abordaremos a análise de forma individual para cada conjunto de dados, e finalizaremos o capítulo sumarizando os principais resultados obtidos.

4.1 Inadimplência de clientes de cartão de crédito

É comum que aplicações de classificação de crédito apresentem conjuntos de dados altamente desbalanceados, onde a classe majoritária (adimplentes) domina e a classe minoritária (inadimplentes) possui um número substancialmente menor de unidades amostrais. Essa disparidade pode levar a problemas de desempenho em classificadores con-

vencionais, que tendem a favorecer a classe majoritária. No entanto, para uma classificação eficaz, é crucial que o modelo seja capaz de lidar igualmente bem com ambas as classes. Nesse contexto, técnicas de reamostragem têm se mostrado eficazes. Um exemplo desse cenário é observado no primeiro banco de dados examinado, compreende informações individuais e de pagamento da fatura de cartão de crédito de clientes de um grande banco em Taiwan, durante o ano de 2005. Os dados foram coletados por meio de um estudo conduzido por Yeh e Lien (2009) e estão disponíveis no repositório de aprendizado de máquina da Universidade da Califórnia, Irvine (UCI), acessível em: http://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients. Este conjunto contém registros de 30.000 clientes, dos quais 6.636 $(22,12\%)$ são classificados como inadimplentes e 23.364 $(77,88\%)$ como adimplentes em relação aos pagamentos. A variável resposta é binária, indicando se um cliente é inadimplente (1) ou adimplente (0). Além disso, temos à disposição as seguintes covariáveis:

- X_1: Valor do crédito concedido: inclui tanto o crédito atitular da conta quanto o crédito familiar (dependentes);
- X_2: Gênero (1 = masculino; 2 = feminino);
- X_3: Educação (1 = pós-graduação; 2 = universidade; 3 = ensino médio; 4 = outros);
- X_4: Estado civil (1 = casado; 2 = solteiro; 3 = outros);
- X_5: Idade (ano);
- Histórico de pagamentos anteriores. Acompanhamento do histórico de pagamentos mensais considerando o período de abril a setembro de 2005 da seguinte forma: X_6 = status do pagamento em setembro de 2005; X_7 = status do pagamento em agosto de 2005; ...; X_{11} = status do pagamento em abril de 2005. Em que a escala de medição para status de amortização é: -1 = pagar devidamente; 1 = atraso no pagamento por um mês; 2 = atraso no pagamento por dois meses; ...; 8 = atraso no pagamento por oito meses; 9 = atraso no pagamento por nove meses ou superior.
- $X_{12} - X_{17}$: Valor da fatura do cartão de crédito. X_{12} = Valor da fatura do cartão de crédito em setembro de 2005; X_{13} = Valor da fatura do cartão de crédito em agosto de 2005; ...; X_{17} = Valor da fatura do cartão de crédito em abril de 2005.

- $X_{18} - X_{23}$: Valor do pagamento anterior. X_{18} = valor pago em setembro de 2005; X_{19} = valor pago em agosto de 2005; ...; X_{23} = valor pago em abril de 2005.

É importante ressaltar que as variáveis referentes a valores monetários são expressos em dólares taiwaneses.

Com o intuito de otimizar a utilização das variáveis e potencializar os resultados a serem obtidos com a aplicação da metodologia proposta, vamos selecionar e criar de forma criteriosa variáveis a partir das variáveis originais disponíveis no conjunto de dados. Tal procedimento será embasado por experiências anteriores em modelagem de dados de crédito. Assim, o conjunto de dados que utilizaremos para aplicação da metodologia será composto pelas seguintes variáveis:

- Idade: Em anos

- Atraso Médio: Quantidade média de atrasos do pagamento considerando os 3 meses mais recentes, isto é,
$$\frac{X_6 + X_7 + X_8}{3};$$

- Quantidade de atrasos: Quantidade de vezes que a fatura do cliente estava com algum status de atraso ao fim do mês, isto é
$$\sum_{i=6}^{11} \mathbb{I}(X_i \geq 0)$$

- Quantas vezes o cliente atrasou o pagamento, considerando os seis meses de referência do estudo, ou seja,
$$\sum_{i=6}^{11} \mathbb{I}(X_i \geq 0)$$

- Tendência crescente do gasto: Quantas vezes o cliente aumentou seu gasto de um mês para o outro, considerando os seis meses de referência do estudo, isto é,
$$\sum_{i=12}^{17} \mathbb{I}(X_i \geq X_{i-1})$$

- Fatura acima do Limite: Quantidade de vezes que a fatura mensal terminou o mês acima do limite, isto é,
$$\sum_{i=12}^{17} \mathbb{I}(X_i \geq X_1)$$

- Diferença entre fatura e pagamento: Total absoluto da diferença entre o valor da fatura e do pagamento em cada um dos meses, isto é,

$$\sum_{i=12}^{17} X_i - \sum_{18}^{23} X_i$$

4.1.1 Análise descritiva e exploratória de dados

Como nosso objetivo do estudo envolve a classificação dos indivíduos entre os estados de adimplência ou inadimplência, faremos uma análise descritiva e exploratória dos dados a fim de obter indícios preliminares do padrão de comportamento dos clientes adimplentes e inadimplentes de acordo com as covariáveis consideradas no estudo.

Primeiramente, apresentamos um gráfico de barras para ilustrar a porcentagem de unidades amostrais pertencentes a cada uma das classes da variável resposta Y.

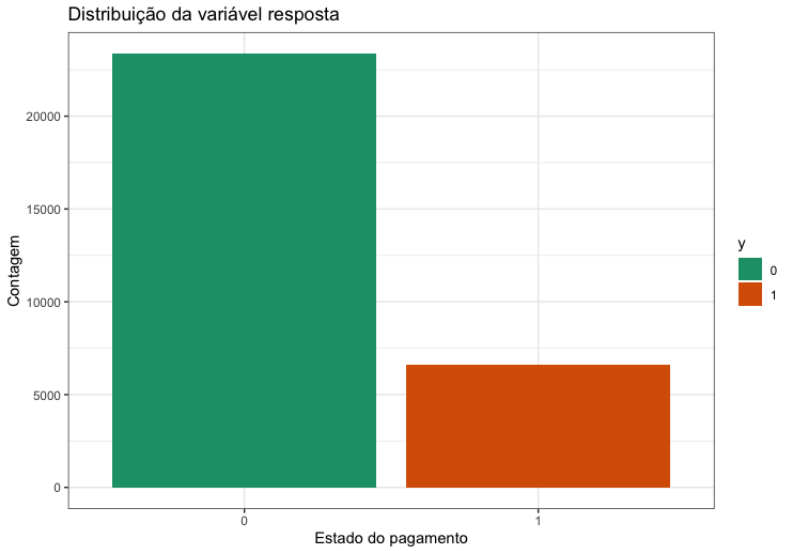

Figura 4.1: **Distribuição da variável indicadora de inadimplência**. As barras representam o número de unidades amostrais pertencentes à classe de adimplentes (caixa verde) e à classe de inadimplentes (caixa vermelha).

Tabela 4.1: Quantidade, em números absolutos e porcentagens, de unidades amostrais em cada uma das classes da variável resposta Y.

	Adimplentes	Inadimplentes
Totais	23.364 ($\approx$ 78%)	6.636 ($\approx$ 22%)

De acordo com a Figura 4.1 e a Tabela 4.1, observamos a presença de um desbalanceamento entre as classes da variável resposta. Nesse conjunto de dados, o nível de desbalanceamento é em torno de 78% das instâncias pertencentes à classe majoritária

e 22% delas pertencentes à classe minoritária. Na área financeira, esse é um nível de desbalanceamento considerado moderado. Ele ocorre quando a diferença entre as classes majoritária e minoritária é mais significativa, mas ainda não é extremamente desproporcional. Pode ser caracterizado por uma relação de $70 - 30$ ou $80 - 20$ entre as classes majoritária e minoritária, respectivamente.

Agora, vamos comparar a distribuição de cada covariável em cada uma das classes da variável resposta Y. Primeiramente, vamos observar como se dá a distribuição das idades na classe majoritária (clientes adimplentes) e na classe minoritária (clientes inadimplentes), conforme indicam os *boxplots* da Figura 4.2.

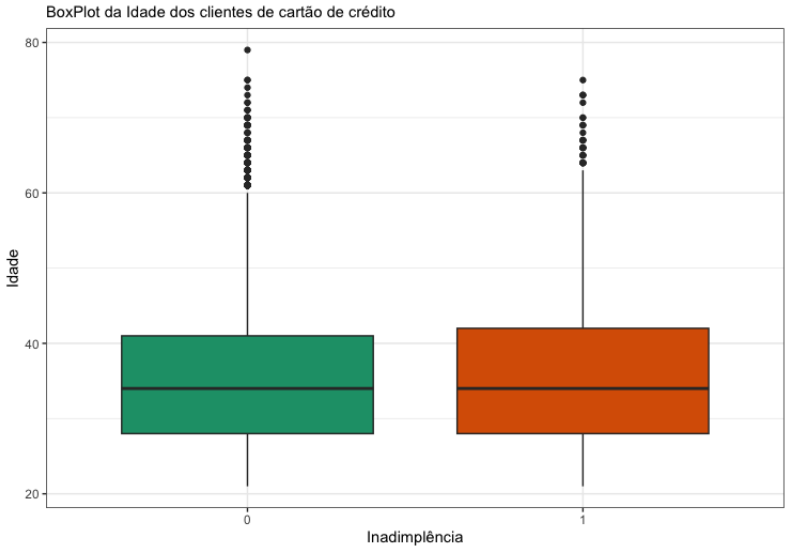

Figura 4.2: **Distribuição da variável idade de acordo com as classes da variável indicadora de inadimplência.** Os boxplots representam a distribuição dos clientes adimplentes (caixa verde) e inadimplentes (caixa vermelha) de acordo com as suas idades.

Tabela 4.2: Estatísticas resumo da variável idade de acordo com as classes da variável indicadora de inadimplência.

	Mínimo	1º Quartil	Mediana	Média	3º Quartil	Máximo
Inadimplente	21,00	28,00	34,00	35,73	42,00	75,00
Adimplente	21,00	28,00	34,00	35,42	41,00	79,00

Com respeito aos *boxplots* apresentados na Figura 4.2 e às medidas resumos apresentadas na Tabela 4.2, é possível identificar que a variabilidade entre o primeiro e o terceiro quartil é ligeiramente maior para a classe de clientes inadimplente. Contudo, as distribuições não aparentam ter uma grande diferença entre elas. Portanto, temos um indicativo de que a idade não é um fator de discriminação para inadimplência dos clientes,

no sentido que saber a idade de um cliente não é um indicativo da propensão desse cliente se tornar inadimplente.

Na Figura 4.3 e na Tabela 4.3, vamos analisar a diferença de comportamento do atraso médio dos clientes adimplentes com o dos clientes inadimplentes.

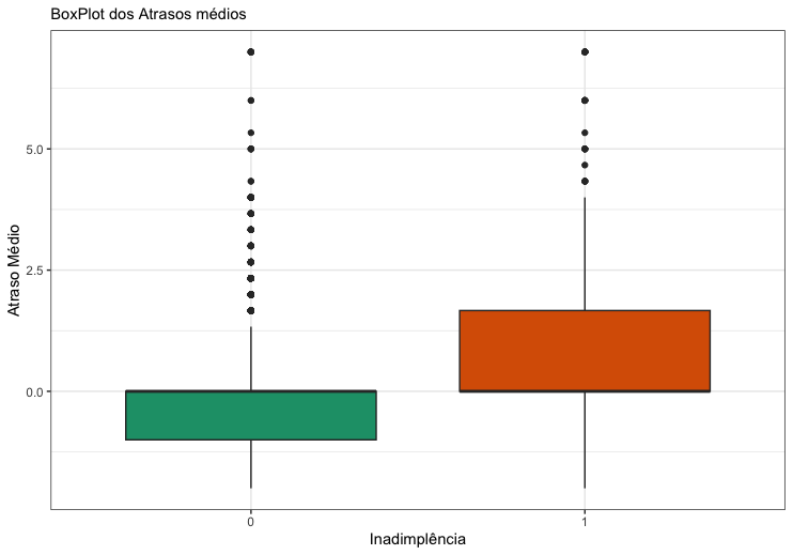

Figura 4.3: **Distribuição da variável atraso médio de acordo com as classes da variável indicadora de inadimplência**. Os boxplots representam a distribuição dos clientes adimplentes (caixa verde) e inadimplentes (caixa vermelha) de acordo com os seus atrasos médios

Tabela 4.3: Estatísticas resumos da variável atraso médio de acordo com as classes da variável indicadora de inadimplência.

	Mínimo	1º Quartial	Mediana	Média	3º Quartil	Máximo
Inadimplente	-2,0000	0,0000	0,0000	0,4962	1,6667	7,0000
Adimplente	-2,0000	-1,0000	0,0000	-0,2765	0,0000	7,0000

Quanto às distribuições da variável atraso médio, notamos uma igualdade no valor mediano, mínimo e máximo. Contudo, para os os clientes adimplentes, a mediana e o terceiro quartil se coincidem. Já para os inadimplentes, o primeiro quartil e a mediana se coincidem, indicando que a maioria dos atrasos médios superiores a 0 meses estão presentes na classe dos inadimplentes, o contrário do que ocorre para os clientes adimplentes. Portanto, temos um indicativo de que a variável atraso médio seja um fator de discriminação para inadimplência de clientes.

Na Figura 4.4 e na Tabela 4.4, vamos analisar a diferença de comportamento da diferença média entre fatura e pagamento dos clientes adimplentes com a dos clinetes inadimplentes.

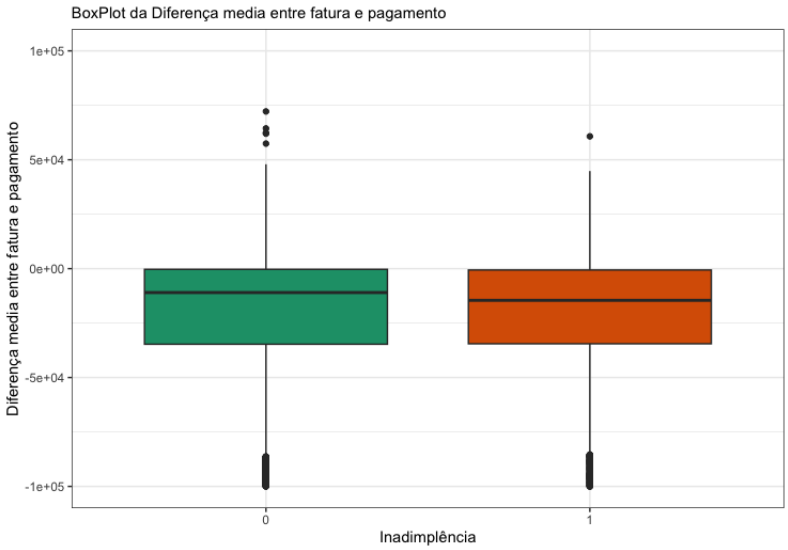

Figura 4.4: **Distribuição da variável diferença média entre fatura e pagamento de acordo com as classes da variável indicadora de inadimplência**. Os boxplots representam a distribuição de clientes adimplentes (caixa verde) e inadimplentes (caixa vermelha) de acordo com a diferença média entre fatura e pagamento.

Tabela 4.4: Estatísticas resumos da variável diferença média entre fatura e pagamento de acordo com as classes da variável indicadora de inadimplência.

	Mínimo	1º Quartil	Mediana	Média	3º Quartil	Máximo
Adimplente	-502.915	-48.310	-17.985	-40.142	-1.326	114.000
Inadimplente	-686.013,3	-51.863,0	-16.585,2	-39.576,6	-660,8	445.252,3

Como a diferença é feita entre a quantidade paga no mês e o valor da fatura do mesmo mês, é esperado que a o valor pago ao mês é, ao menos menor do que o valor da fatura. Ou seja, é esperado que para os clientes inadimplentes os valores flutuem em torno de valores mais negativos quando comparado ao grupo dos clientes adimplentes. Sendo assim, observa-se que, em geral, os valores dessa diferença estão posicionados mais abaixo do terceiro quartil quando para os inadimplentes. Entretanto, a diferença média é maior (em módulo) para os adimplentes.

Na Figura 4.5 e na Tabela 4.5, vamos analisar a diferença de comportamento da tendência de gastos crescentes dos clientes adimplentes com a dos clientes inadimplentes.

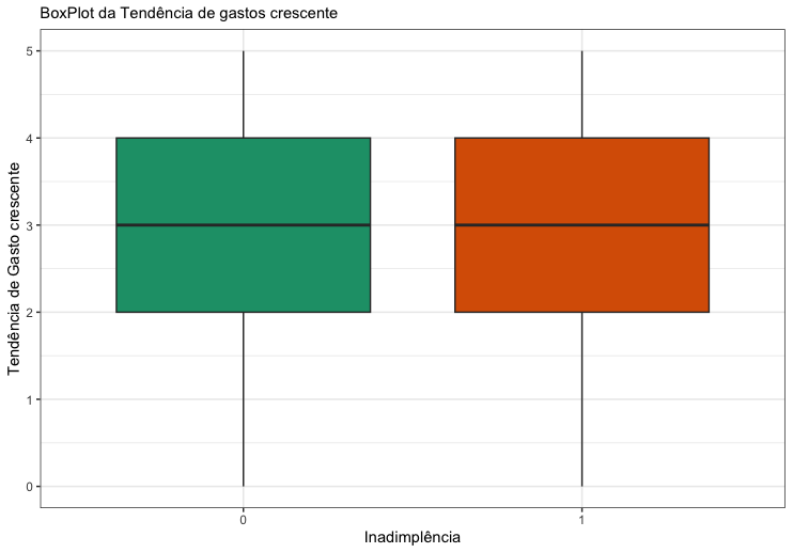

Figura 4.5: **Distribuição da variável tendência de gastos crescente de acordo com as classes da variável indicadora de inadimplência**. Os boxplots representam a distribuição de clientes adimplentes (caixa verde) e inadimplentes (caixa vermelha) de acordo com as tendências de gastos crescentes.

Tabela 4.5: Estatísticas resumos da variável tendência de gastos crescente de acordo com as classes da variável indicadora de inadimplência.

	Mínimo	1º Quartil	Mediana	Média	3ª Quartil	Máximo
Inadimplente	0,000	2,000	3,000	2,985	4,000	5,000
Adimplente	0,000	2,000	3,000	2,713	4,000	5,000

Nesse caso, as distribuição não aparentam ter uma grande diferença entre elas. Portanto, temos um indicativo de que a tendência de gastos crescentes não é um fator de discriminação para inadimplências de clientes.

Na Figura 4.6 e na Tabela 4.6, vamos analisar a diferença de comportamento da quantidade média de atrasos dos clietes adimplentes com a dos clientes inadimplentes.

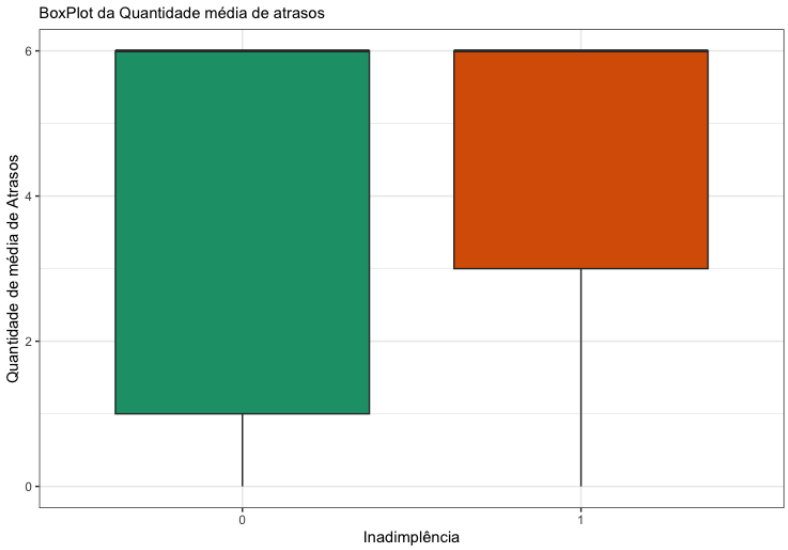

Figura 4.6: **Distribuição da variável quantidade média de atrasos de acordo com as classes da variável indicadora de inadimplência**. Os boxplots representação a distribuição de clientes adimplentes (caixa verde) e inadimplentes (caixa vermelha) de acordo com as quantidades médias de atrasos.

Tabela 4.6: Estatísticas resumos da variável quantidade média de atrasos de acordo com as classes da variável indicadora de inadimplência.

	Mínimo	1º Quartil	Mediana	Média	3º Quartil	Máximo
Inadimplente	0.000	3.000	6.000	4.471	6.000	6.000
Adimplente	0.000	1.000	6.000	3.907	6.000	6.000

Nesse caso, vemos igualdade nos valores medianos, de terceiro quartil e máximos. A média de atrasos dos inadimplentes e a posição do primeiro quartil também está mais acima quando comparado entre os grupos.

Para a variável fatura maior do que, que conta a quantidade de vezes que a fatura mensal do cliente fechou acima do limite do cartão disponibilizado no banco, nao há diferenças muito grande quanto a distribuição dos quartis e das estatísticas resumo. Por conta disso, para evitar confusão, foi optado não incluir no trabalho esses resultados. Para ilustrar a semelhança nas distribuições, olhemos para o seguinte gráfico de barras:

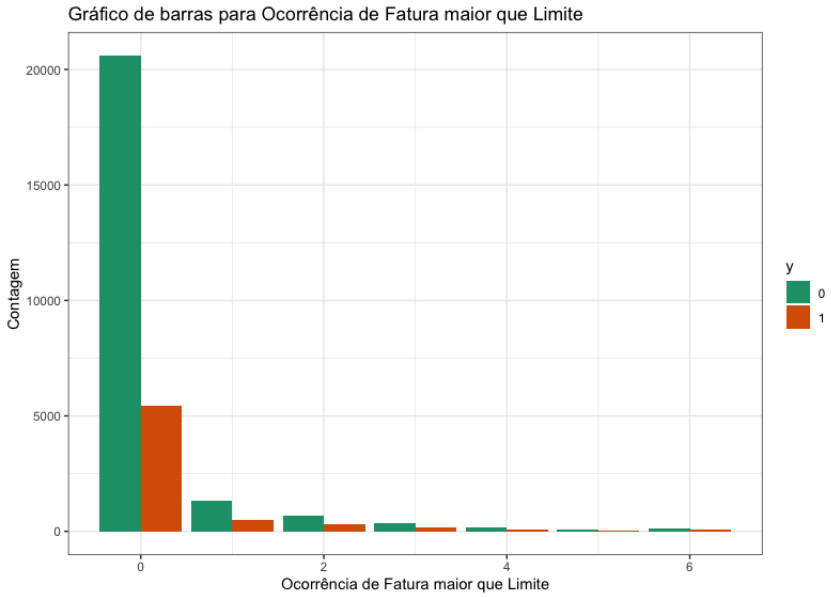

Figura 4.7: **Gráfico de barras da quantidade de vezes que a fatura foi superior ao limite dividido por grupos.** As barras representam o número de unidades amostrais pertencentes à classe adimplentes (caixa verde) e à classe de inadimplentes (caixa vermelha).

Na Figura 4.8, vemos a presença de uma correlação baixa entre todas as covariáveis com excessão da correlação entre a quantidade de atrasos totais e a média de atrasos nos últimos 3 meses. A presença de covariáveis pouco correlacionadas é um bom indicativo que não será necessário contornar possíveis problemas de multicolinearidade ao estabelecer o modelo.

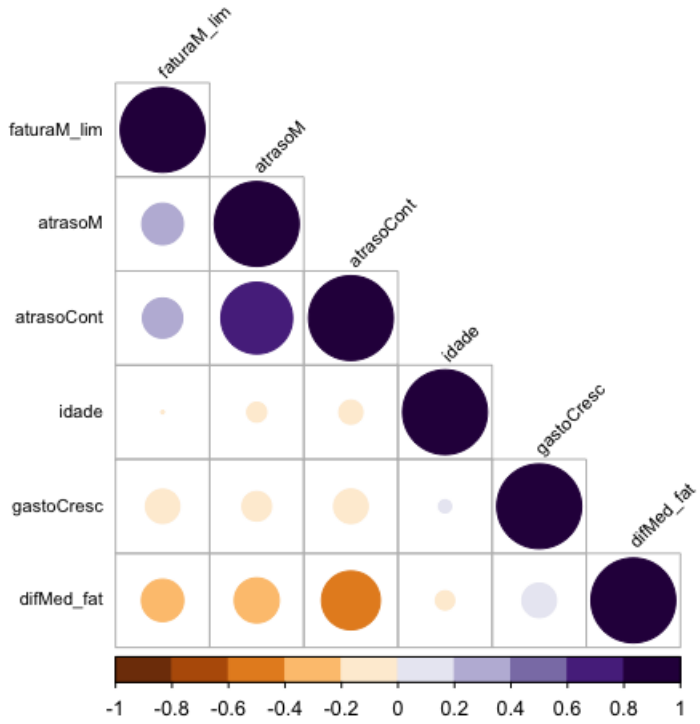

Figura 4.8: Gráfico de correlação das covariáveis presentes no modelo.

Na Tabela 4.7, nota-se que atrasoCont é a variável que mais infla a variância, indicando ser a variável mais correlacionada dentre as covariáveis do modelo escolhido. Contudo, por convenção, um índice $VIF < 3$ já passa a ser suficientemente pequeno para conduzir um modelo de regressão sem ter de contornar o problema da multicolinearidade.

Tabela 4.7: **Tabela com os valores dos VIFs obtidos**. Sendo X1 = Idade, X2 = Atraso médio, X3 = Quantidade média de atrasos, X4 = Tendência de gasto crescente, X5 = Fatura maior que o limite, X6 = Diferença média entre fatura e pagamento

X1	X2	X3	X4	X5	X6
1,015600	1,956766	2,322020	1,052089	1,119818	1,347660

4.1.2 Resultados

Nesta fase, colocamos em prática as abordagens discutidas previamente utilizando o conjunto de dados sobre inadimplência de clientes de cartão de crédito. Inicialmente, efetuamos uma padronização do conjunto de dados, com o objetivo de uniformizar todas as variáveis, exceto a variável de resposta, para uma mesma escala. Nesse processo, as variáveis são ajustadas para possuírem média zero e desvio padrão um. Em seguida, dividimos o conjunto de dados em dois subconjuntos conhecidos como conjunto de trei-

namento e conjunto de teste. A formação desses subconjuntos foi determinada por meio de uma seleção aleatória sem reposição do conjunto de dados original, assegurando que 80% das observações do conjunto de dados inicial fossem destinadas ao conjunto de treinamento, enquanto os 20% restantes fossem alocados ao conjunto de teste. É importante ressaltar que essa seleção foi realizada mantendo as proporções iniciais de clientes adimplentes e inadimplentes presentes na base de dados. A principal razão para essa divisão é proporcionar recursos para avaliar a eficácia do classificador em um ambiente independente. No conjunto de treinamento, treinamos o classificador, enquanto no conjunto de teste avaliamos sua capacidade de classificar novas observações de forma precisa.

Frisamos que todos procedimentos de análises foram realizados no *software R*. Para aplicação da regressão logística, primeiramente, consideramos todas as covariáveis discutidas na subseção anterior. Ao ajustar o modelo completo, obtivemos as estimativas dos coeficientes do modelo de regressão e p-valores conforme indicado na Tabela 4.8.

Tabela 4.8: Estimativas dos coeficientes obtidos no modelo logístico sem seleção de variáveis e sem aplicação de qualquer método de reamostragem.

	Estimativa	Erro padrão	Estatística	p-valor
(Intercepto)	-1.2405	0.0796	-15.58	0.0001
Idade	0.0054	0.0016	3.36	0.0008
Atraso Médio	1.0668	0.0207	51.59	0.0001
Quantidade de Atrasos	-0.2096	0.0098	-21.45	0.0001
Tendencia Gasto Crescente	0.2122	0.0117	18.13	0.0001
Fatura acima do Limite	0.1082	0.0170	6.39	0.0001
Diferença média de fatura	0.0001	0.0001	1.92	0.0548

De acordo com os testes de hipóteses individuais, a amostra trás evidências de que todas as covariáveis possuem um efeito significativo sobre a resposta esperada quando considerado o modelo completo e um nível de significância de pelo menos 6%.

A partir da Figura 4.1.2, temos um indicativo de que as hipóteses do modelo logístico estão sendo satisfeitas de acordo com a amostra observada, salvo a presença de alguns outliers.

É importante observar que após a aplicação do pré-processamento dos dados, as estimativas, os p-valores dos testes e os gráficos da análise de diagnóstico sofreram alterações. Todavia, sem mudar as nossas conclusões com a relação a adequabilidade do modelo. Uma vez que o foco deste trabalho é comparar o poder preditivo do modelo lógistico sem e com o pré-processamento dos dados, não trouxemos aqui tais resultados para permitir uma leitura mais fluida e focada nas discussões que gostaríamos de tratar nesta monografia.

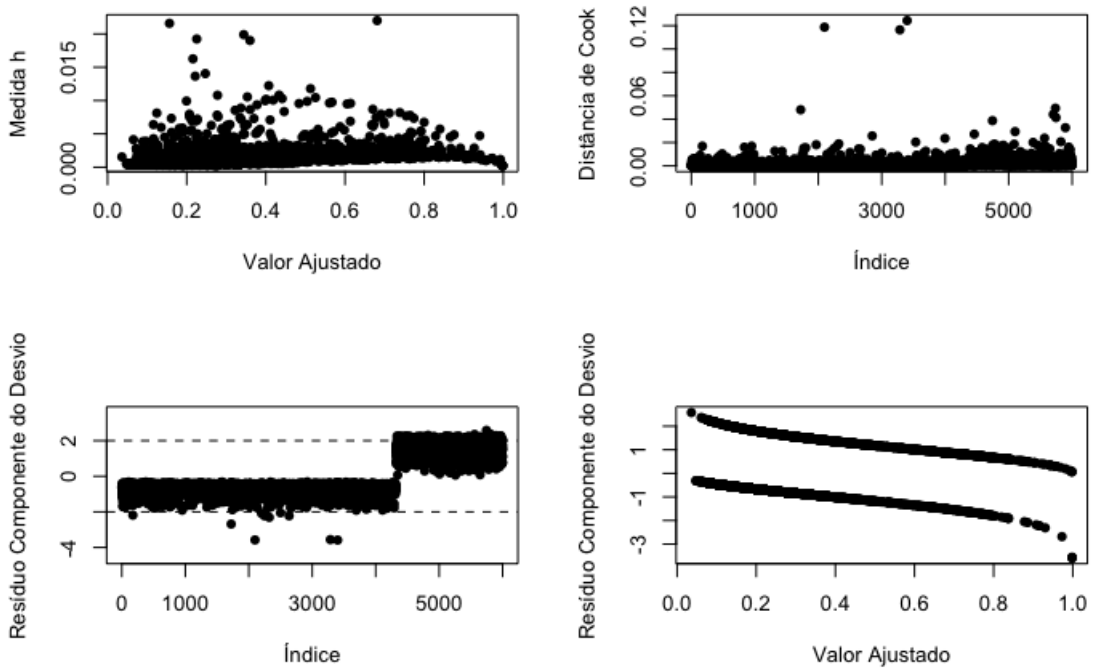

Figura 4.9: Análise de diagnóstico para o modelo logístico geral.

Tabela 4.9: Medidas de performance aplicadas para método de reamostragem com e sem seleção de variáveis em que A: Modelo sem reamostragem, B: Modelo com Tomek Link, C: Modelo com SMOTE, D: Modelo com SMOTE e Tomek Link.

		KS	AUC	ACC	ESP	SEN	VPN	VPP	G-MÉDIA	MCC
S/ SELEÇÃO	A	0,3669	0,7283	**0,7702**	**0,8481**	0,4962	0,8556	0,4812	0,6489	0,3405
	B	0,3919	0,7427	0,7519	0,8121	0,5399	**0,8614**	0,4494	0,6704	0,3313
	C	0,3782	0,7345	0,6888	0,7800	0,5977	0,6597	**0,7309**	0,6807	**0,3840**
	D	**0,4014**	**0,7453**	0,6835	0,7444	**0,6226**	0,6636	0,7090	**0,6813**	0,3698
LASSO	A	0,3713	0,7191	**0,7890**	**0,8754**	0,4847	0,8567	0,5250	0,6533	0,3708
	B	0,3939	0,7337	0,7751	0,8492	0,5143	**0,8602**	0,4921	0,6630	0,3579
	C	0,3728	0,7268	0,6846	0,8341	**0,5352**	0,6421	0,7634	**0,6711**	**0,3870**
	D	**0,3996**	**0,7417**	0,6775	0,8500	0,5050	0,6320	**0,7710**	0,6634	0,3782

Na Tabela 4.9, apresentamos os resultados da aplicação da regressão logística no conjunto de teste. Nas colunas, estão representados as medidas utilizadas para medir o desempenho dos classificadores, em que ACC = Acurácia, SEN = Sensibilidade, VPP = Valor Preditivo Positivo, VPN = Valor Preditivo Negativo, G-MÉDIA = G-Média e MCC = Coeficiente de Correlação de Matthews. Nas linhas, estão dispostos os cenários nos quais a regressão logística foi ajustada. Os resultados estão apresentados em dois blocos, um trata-se do modelo completo com todas as covariáveis incluídas e o outro trata-se do modelo reduzido, apenas com as covariáveis selecionadas pelo LASSO. A marcação em negrito, destaca qual dos cenários teve melhor desempenho para as medidas avaliadas em cada um dos blocos. Para a aplicação da técnica SMOTE, definimos $K = 5$. Vale ressaltar que não foi realizado nenhuma alteração no conjunto de teste.

Primeiramente, observamos que estabelecemos que tanto no cenário sem processamento dos dados quanto nos cenários em que aplicamos algum método de pré-processamento dos dados, escolhemos o ponto de corte para realizar a classificação como sendo o argumento que resultada na estatística KS.

Analisando o modelo completo, identificamos que, de forma geral, o cenário híbrido demonstrou os melhores resultados em termos de desempenho da regressão logística no conjunto de teste, de acordo com as métricas consideradas.

Ao observar as medidas de performance individualmente, destacamos que a acurácia da regressão logística foi mais elevada nos cenários sem pré-processamento de dados e naquele em que o método Tomek Link foi aplicado. Nesses casos, aproximadamente 75% dos clientes foram classificados corretamente. Nos cenários C e D, essa proporção foi de aproximadamente 68%. No entanto, nos cenários A e B, onde a acurácia foi maior, o conjunto de treinamento continua desbalanceado, com uma grande proporção de clientes adimplentes, o que leva o modelo logístico a ter uma boa performance na classificação de clientes pertencentes a essa classe. Esse fato pode ser evidenciado quando analisamos a especificidade e o valor predito negativo (VPN) desses cenários. No caso da especificidade, segue que aproximadamente 83% dos clientes pertencentes a classe majoritária foram classificados corretamente como adimplentes, enquanto, no caso do VPN, segue que 86% dos clientes que foram classificados como adimplentes, pertenciam de fato a classe majoritária. A classificação correta de clientes da classe majoritária, eleva, portanto, o valor da acurácia.

Por outro lado, nos cenários A e B, a proporção de clientes inadimplentes é baixa, o que

leva o modelo logístico a ter uma má performance na classificação de clientes pertencentes a essa classe. Esse fato pode ser evidenciado quando analisamos a sensibilidade e o valor predito positivo (VPP) desses cenários. No caso da sensibilidade, segue que aproximadamente 52% dos clientes pertencentes à classe minoritária foram classificados corretamente como inadimplentes, enquanto, no caso do VPP, segue que aproximadamente 46% dos clientes classificados como inadimplentes pertenciam de fato a classe minoritária. Nos cenários C e D, observamos valores maiores de sensibilidade e especificidade quando comparados aos obtidos nos cenários A e B. Temos aproximadamente uma sensibilidade de 61% e um VPP de aproximadamente 72%. De fato, nos cenários C e D, a base foi balanceada pelo método SMOTE, que criou instâncias sintéticas de clientes inadimplentes, aumentando sua proporção e, consequentemente, permitindo um melhor aprendizado da classe minoritária pelo modelo logístico, levando a um aumento da sua performance em classificar clientes inadimplentes.

A partir dessas observações, fica claro que as métricas de acurácia, sensibilidade, especificidade, VPP e VPN são altamente sensíveis ao desequilíbrio das classes. Portanto, interpretá-las isoladamente pode levar a conclusões equivocadas. Para mitigar essa questão, foram desenvolvidas métricas de performance agregadas, como o coeficiente de Matthews e a G-média, que condensam os valores das métricas individuais em uma única medida, expressando o desempenho global do classificador. Com base nessas métricas, os cenários C e D foram os mais bem-sucedidos. Em ambos os cenários, temos uma base de treino balanceada.

Ao analisar o caso com modelo reduzido aplicando LASSO, observamos que, em todos os cenários, a performance do modelo logístico se manteve próxima àquela do modelo completo com todas as variáveis. Os cenários A e B mostraram um melhor desempenho em termos de acurácia, especificidade e VPN, enquanto os cenários C e D se destacaram em sensibilidade e VPP. As métricas de G-média e coeficiente de Matthews indicaram que os cenários C e D, ambos balanceados, alcançaram um desempenho global superior. Também à respeito do LASSO, vale mencionar que apenas no cenário A o LASSO efetuou a seleção de variáveis, em que zerou os coeficientes associados as covariáveis "Idade" e "Diferença média entre fatura".

Em suma, para esse conjunto de dados, nossa análise sugere que a aplicação de métodos de balanceamento de classes pode resultar em um aprimoramento geral do desempenho do modelo logístico na classificação de novas instâncias.

4.2 Inadimplência de crédito

Esse segundo conjunto de dados foi obtido de uma competição publicada em https://www.kaggle.com/competitions/GiveMeSomeCredit/overview/description cujo objetivo a construção de um algoritmo capaz de predizer a probabilidade de um cliente vir a se tornar inadimplente. Por se tratar de uma competição e os dados serem disponibilizados pelos organizadores, não foi possível encontrar muitas informações a respeito da coleta desses dados, porém fica subentendido que estes são referentes a algumas informações cadastrais e de características de pagamento de clientes de um determinado banco. O arquivo original contém uma base de treinamento com 150.000 observações e uma base de teste contendo 101.504 observações. Para esse estudo iremos utilizar apenas a base de treinamento fornecida, pois a base de teste omite a informação referente a inadimplência, logo a base de treinamento fornecida será nosso conjunto de dados. Dos 150.000 clientes listados na base de treinamento, 139.974 (93.32%) clientes são adimplentes e 10.026 (6.68%) inadimplentes. Nesse estudo, a variável de interesse é o status de adimplência do cliente que é representada por uma variável binária, indicando se um cliente é inadimplente (1) ou adimplente (0). Além disso, temos à disposição as seguintes covariáveis:

- RevolvingUtilizationOfUnsecuredLines: Saldo total em cartões de crédito e linhas de crédito pessoais, exceto imóveis e sem dívidas parceladas, como empréstimos de carro, dividido pela soma dos limites de crédito;
- Age: Idade do cliente;
- NumberOfTime30-59DaysPastDueNotWorse: Número de vezes que o cliente ficou de 30-59 dias atrasado, mas não pior nos últimos 2 anos;
- DebtRatio: Pagamentos de dívidas mensais, pensão alimentícia, custos de vida, divididos pela renda bruta mensal;
- MonthlyIncome: Renda mensal;
- NumberOfOpenCreditLinesAndLoans: Número de empréstimos ativos (parcela como empréstimos de carro ou hipoteca) e linhas de crédito (por exemplo, cartões de crédito);
- NumberOfTimes90DaysLate: Número de vezes que o cliente ficou com 90 dias ou mais de atraso;

- NumberRealEstateLoansOrLines: Número de empréstimos hipotecários e imobiliários, incluindo linhas de crédito para aquisição de habitação;
- NumberOfTime60-89DaysPastDueNotWorse: Número de vezes que o cliente ficou 60-89 dias atrasado, mas não pior nos últimos 2 anos;
- NumberOfDependents: Número de dependentes na família excluindo o cliente (cônjugue, filhos etc.);
- SeriousDlqin2yrs: Cliente com 90 dias de atraso ou pior (Sim ou não).

Com o intuito de otimizar a utilização das variáveis e potencializar os resultados a serem obtidos com a aplicação da metodologia proposta, elegemos apenas algumas variáveis para seguir com as análises. Seguindo os mesmos critérios utilizados para o conjunto de dados anterior, tal procedimento será embasado por experiências anteriores em modelagem de dados de crédito. Dentre as variáveis disponíveis, selecionamos e renomeamos as variáveis da seguinte maneira:

- RevolvingUtilizationOfUnsecuredLines - Prop_limite_utlizado;
- Age - Idade;
- NumberOfTime30-59PastDueNotWors - Atraso30_59 dias;
- DebtRatio - Prop_divida_renda;
- NumberOfOpenCreditLinesAndLoan - Qnt_emprestimos_ativos;
- NumberRealEstateLoansOrLine - Qnt_emprestimo_imobiliario.

4.2.1 Análise descritiva e exploratória de dados

Assim como a Subseção 4.2.1, realizamos uma análise descritiva e exploratória dos dados a fim de obter indícios preliminares do padrão de comportamento dos clientes adimplentes e inadimplentes de acordo com as covariáveis consideradas no estudo.

Começamos apresentando um gráfico de barras para ilustrar a porcentagem de unidades amostrais pertencentes a cada uma das classes da variável resposta Y.

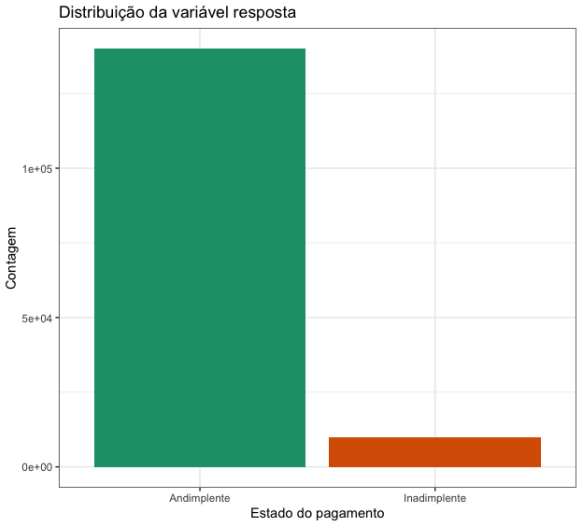

Figura 4.10: **Distribuição da variável indicadora de inadimplência**. As barras representam o número de unidades amostrais pertencentes à classe de adimplentes (caixa verde) e à classe de inadimplentes (caixa vermelha).

Tabela 4.10: Quantidade, em números absolutos e porcentagens, de unidades amostrais em cada uma das classes da variável resposta Y.

	Totais
Adimplente	139.973 ($\approx 93\%$)
Inadimplente	10.026 ($\approx 7\%$)

De acordo com a Figura 4.10 e a Tabela 4.10, observamos a presença de um desbalanceamento entre as classes da variável resposta. Nesse conjunto de dados, o nível de desbalanceamento é em torno de 93% das instâncias pertencentes à classe majoritária e 7% delas pertencentes à classe minoritária. Na área financeira, esse é um nível de desbalanceamento considerado severo. Ele ocorre quando há uma disparidade substancial entre o número de instâncias nas classes majoritária e minoritária. Pode ser caracterizado por uma relação $90 - 10$, $99 - 1$ entre as classes majoritária e minoritária, respectivamente.

Agora, vamos comparar a distribuição de cada covariável em cada uma das classes da variável resposta Y. Primeiramente, vamos observar como se dá a distribuição das idades na classe majoritária (clientes adimplentes) e na classe minoritária (clientes inadimplentes), conforme indicam os *boxplots* da Figura 4.11.

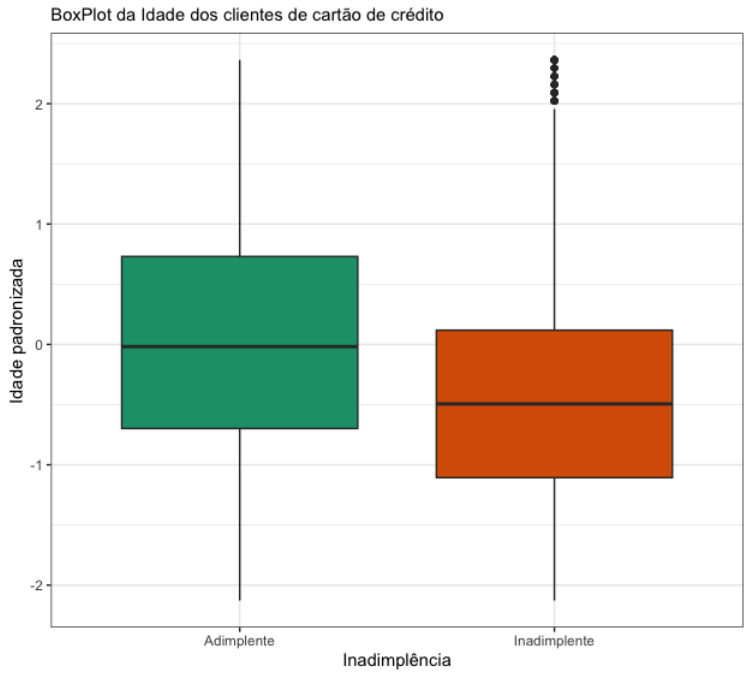

Figura 4.11: **BoxPlot da Idade de acordo com as classes da variável indicadora de inadimplência**. Os boxplots representam a distribuição de clientes adimplentes (caixa verde) e inadimplentes (caixa vermelha) de acordo com as tendências de gastos crescentes.

Tabela 4.11: Estatísticas resumo da variável Idade com as classes da variável indicadora de inadimplência.

	Mínimo	1º Quartil	Mediana	Média	3º Quartil	Máximo
Inadimplente	-2.1276	-1.1069	-0.4944	-0.4321	0.1180	2.3636
Adimplente	-2.1276	-0.6986	-0.0181	0.0309	0.7304	2.3636

Na Figura 4.11, é possível notar que, em geral, a idade dos clientes que foram observados como inadimplentes é menor quando comparado aos clientes adimplentes. A Tabela 4.11 confirma o que foi observado no BoxPlot, tendo equivalência apenas no mínimo e no máximo. Portanto, temos um indicativo de que a idade é um fator de discriminação para inadimplência dos clientes, no sentido que saber a idade de um cliente é um indicativo da propensão desse cliente se tornar inadimplente.

Na Figura 4.12 e na Tabela 4.12, vamos analisar a diferença de comportamento da proporção do limite utilizado dos clientes adimplentes com a dos clientes inadimplentes.

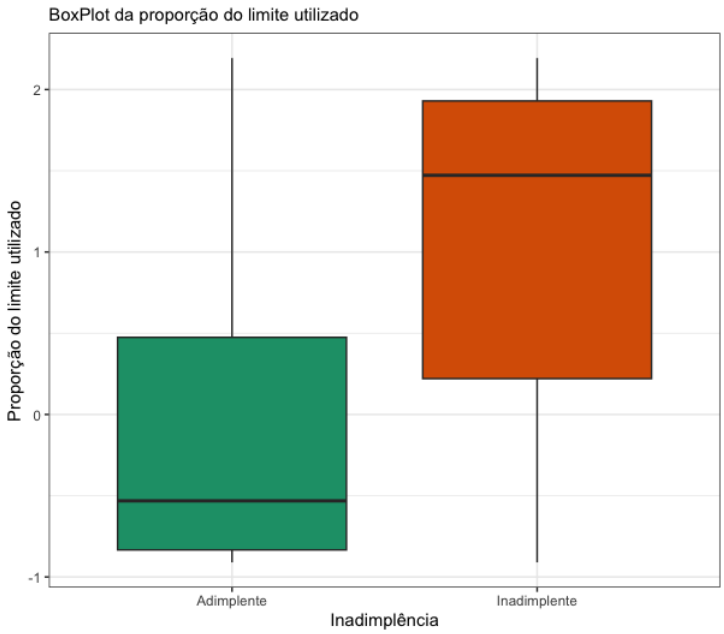

Figura 4.12: **BoxPlot da proporção do limite utilizado de acordo com as classes da variável indicadora de inadimplência**. Os boxplots representam a distribuição de clientes adimplentes (caixa verde) e inadimplentes (caixa vermelha) de acordo com as tendências de gastos crescentes.

Tabela 4.12: Estatísticas resumo da variável proporção do limite utilizado com as classes da variável indicadora de inadimplência.

	Mínimo	1º Quartil	Mediana	Média	3º Quartil	Máximo
Inadimplente	-0.9101	0.2207	1.4720	1.0493	1.9296	2.1936
Adimplente	-0.9101	-0.8335	-0.5316	-0.0752	0.4748	2.1936

De acordo com a Figura 4.12 e a Tabela 4.12, é possível observar que, assim como para a variável idade, comparativamente a distribuição da proporção do limite utilizado é diferente de um grupo para o outro. Fato evidenciado pela diferença entre as estatísticas resumo de um grupo com a do outro, com exceção do mínimo e do máximo que são equivalentes. Portanto, temos um indicativo de que a variável proporção do limite utilizado seja um fator de discriminação para inadimplência de clientes.

Na Figura 4.13 e na Tabela 4.13, vamos analisar a diferença de comportamento da quantidade de empréstimos imobiliários dos clientes adimplentes com a dos clientes inadimplentes.

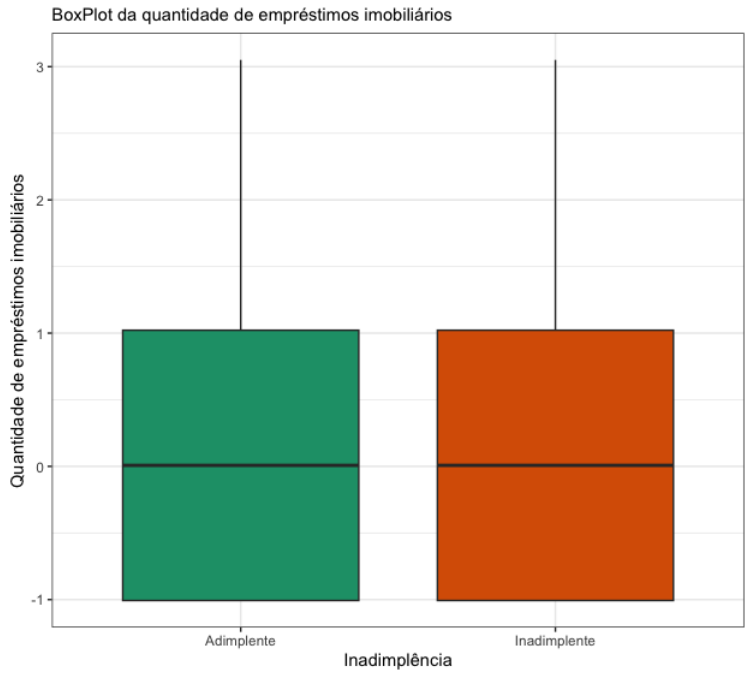

Figura 4.13: **BoxPlot da quantidade de empréstimos imobiliários de acordo com as classes da variável indicadora de inadimplência**. Os boxplots representam a distribuição de clientes adimplentes (caixa verde) e inadimplentes (caixa vermelha) de acordo com as tendências de gastos crescentes.

Tabela 4.13: Estatísticas resumo da variável quantidade empréstimos imobiliários com as classes da variável indicadora de inadimplência.

	Mínimo	1º Quartil	Mediana	Média	3º Quartil	Máximo
Inadimplente	-1.006889	-1.006889	0.007513	-0.074036	1.021916	3.050722
Adimplente	-1.006889	1.006889	0.007513	0.005303	1.021916	3.050722

De acordo com a Figura 4.13 e a Tabela 4.13, não é possível observar uma diferença significativa na distribuição da quantidade de empréstimo imobiliários entre os dois grupos. Portanto, temos um indicativo de que a variável quantidade de empréstimos imobiliários seja um fator de discriminação para inadimplência de clientes.

Na Figura 4.14 e na Tabela 4.14, vamos analisar a diferença de comportamento do número de atrasos de pagamento entre 30 e 59 dias dos clientes adimplentes com o dos clientes inadimplentes.

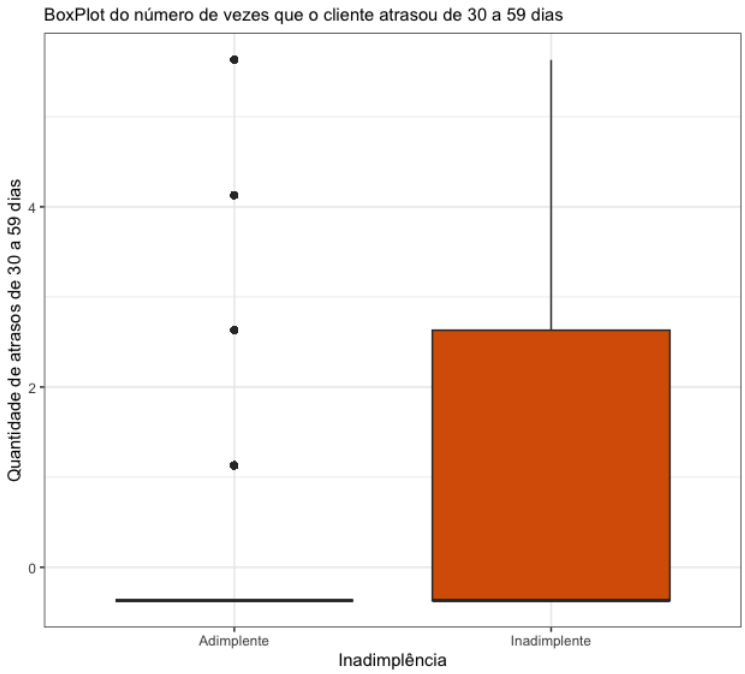

Figura 4.14: **BoxPlot da quantidade de vezes que o cliente atrasou o pagamento entre 30 e 59 dias de acordo com as classes da variável indicadora de inadimplência**. Os boxplots representam a distribuição de clientes adimplentes (caixa verde) e inadimplentes (caixa vermelha) de acordo com as tendências de gastos crescentes.

Tabela 4.14: Estatísticas resumo da variável quantidade de vezes que o cliente atrasou o pagamento entre 30 e 59 dias com as classes da variável indicadora de inadimplência.

	Mínimo	1º Quartil	Mediana	Média	3º Quartil	Máximo
Inadimplente	-0.3687	-0.3687	-0.3687	1.0758	2.6306	5.6300
Adimplente	-0.3687	-0.3687	-0.3687	0.0771	-0.3687	5.6300

Com respeito à quarta covariável presente no estudo, o número de vezes que o cliente atrasou o pagamento entre 30 e 59 vezes, é possível notar uma severa diferença na distribuição dessa quantidade em função do status de adimplência como mostra a Figura 4.14 e confirmado pela Tabela 4.14. Vale ressaltar que os valores de mínimo, primeiro quartil e mediana coincidem, contudo é esperado, pois o valor mediano representa também o valor mínimo, o qual indica que o cliente não atrasou. Portanto, temos um indicativo de que a variável quantidade de atrasos do cliente é um fator de discriminação para inadimplência de clientes.

Na Figura 4.15 e na Tabela 4.15, vamos analisar a diferença de comportamento da quantidade de empréstimos ativos dos clientes adimplentes com a dos clientes inadimplentes.

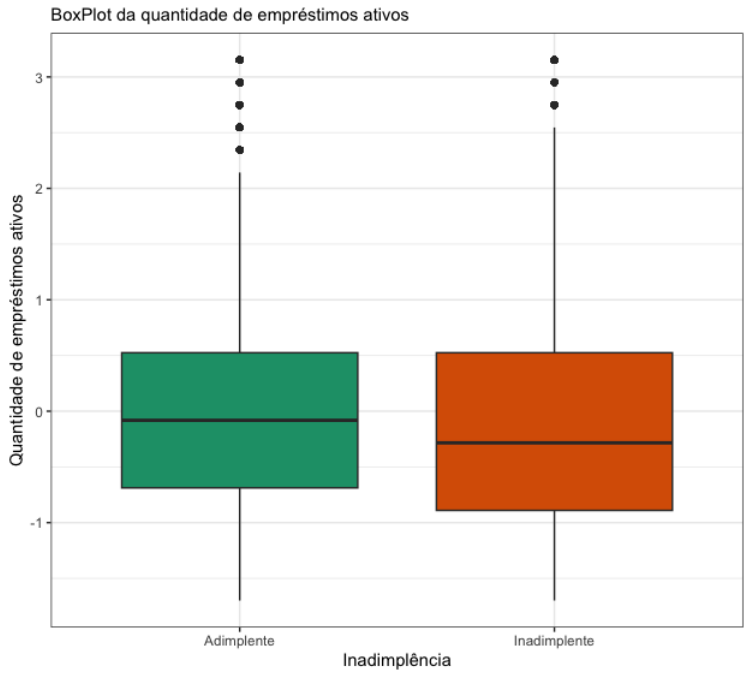

Figura 4.15: **BoxPlot da quantidade de empréstimos ativos de acordo com as classes da variável indicadora de inadimplência**. Os boxplots representam a distribuição de clientes adimplentes (caixa verde) e inadimplentes (caixa vermelha) de acordo com as tendências de gastos crescentes.

Tabela 4.15: Estatísticas resumo da variável quantidade de empréstimos ativos com as classes da variável indicadora de inadimplência.

	Mínimo	1º Quartil	Mediana	Média	3º Quartil	Máximo
Inadimplente	-1.6990	-0.8903	-0.2838	-0.1177	0.5248	3.1530
Adimplente	-1.6990	-0.6881	-0.0817	0.0084	0.5248	3.1530

Com respeito à quantidade de empréstimos tomados para cada um dos grupos, é possível notar uma certa similaridade nas distribuições entre os grupos. Supreendentemente, a mediana da quantidade de empréstimos efetuados pelo grupo dos adimplentes é maior que para o grupo dos inadimplentes como mostra a Figura 4.15 e a Tabela 4.15. Um dos motivos para isso acontecer é uma eventual oferta maior de crédito aos clientes que possuem um bom histórico bancário.

Na Figura 4.16 e na Tabela 4.16, vamos analisar a diferença de comportamento da proporção da dívida em relação a renda dos clientes adimplentes com a dos clientes inadimplentes.

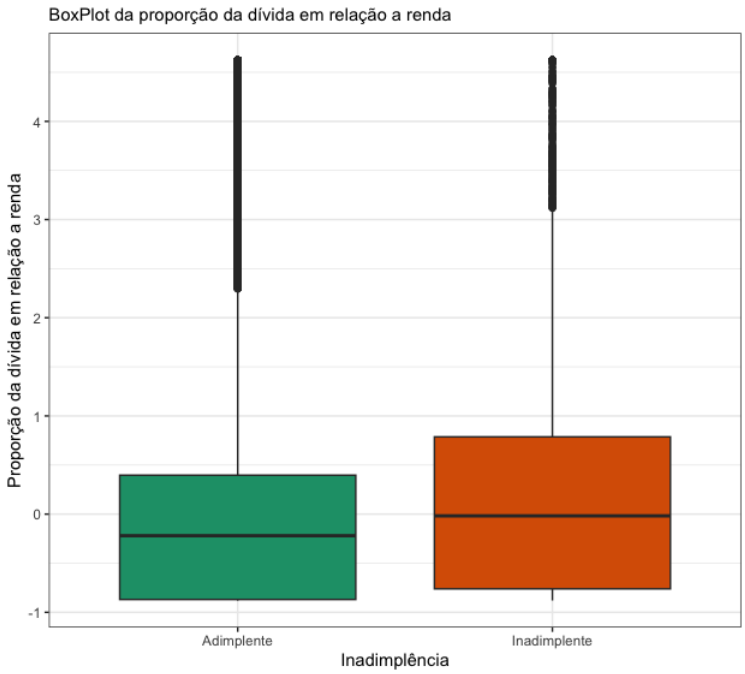

Figura 4.16: **BoxPlot da proporção da dívida em relação a renda de acordo com as classes da variável indicadora de inadimplência**. Os boxplots representam a distribuição de clientes adimplentes (caixa verde) e inadimplentes (caixa vermelha) de acordo com as tendências de gastos crescentes.

Tabela 4.16: Estatísticas resumo da variável proporção da dívida em relação a renda com as classes da variável indicadora de inadimplência.

	Mínimo	1º Quartil	Mediana	Média	3º Quartil	Máximo
Inadimplente	-0.87981	-0.76104	-0.01775	0.25763	0.78701	4.62582
Adimplente	-0.87981	-0.86949	-0.21992	-0.01845	0.39650	4.62582

Para nossa última covariável presente no estudo, a dívida em relação a renda do cliente, é possível notar que, de acordo com a Figura 4.16 e a Tabela 4.16, a dívida em relação a renda do grupo dos inadimplentes está distribuída em torno de valores maiores, onde o primeiro, segundo e terceiro quartil são superiores para o grupo dos inadimplentes.

Da mesma forma que para o conjunto de dados anterior, é necessário que atentemos para a correlação entre as covariáveis afim de nos precaver de eventuais problemas relacionados a multicolinearidade. Sendo assim, observemos a matriz de correlação ilustrada pela Figura 4.17. De acordo com essa figura e com a Tabela 4.17, observamos que não é esperado que encontremos problemas realacionados a multicolinearidade, uma vez que todos os valores do VIF são aproximadamente menores do $1,5$. Como, por convenção, é delimitado um valor limite de $VIF = 10$, não é eseperado que tenhamos problemas de multicolinearidade.

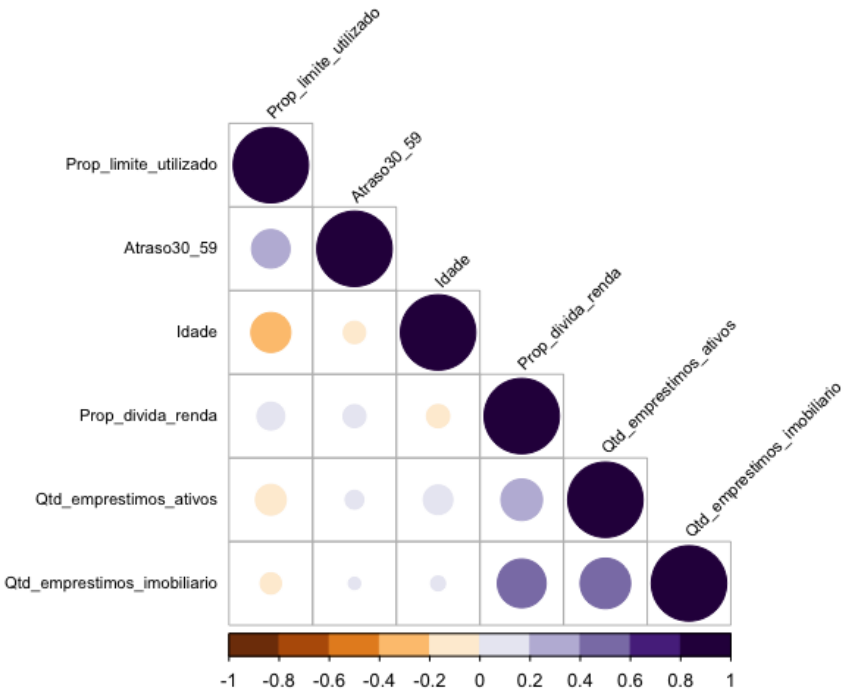

Figura 4.17: Gráfico de correlação das covariáveis presentes no modelo.

Tabela 4.17: **Tabela com os valores do VIF obtidos para o segundo conjunto de dados**. Sendo X1 = Proporção do limite utilizado, X2 = Idade, X3 = Número de vezes que o cliente atrasou de 30 a 59 dias, X4 = Proporção da dívida em relação a renda, X5 = Quantidade de empréstimos imobiliários, X6 = Quantidade de empréstimos ativos

X1	X2	X3	X4	X5	X6
1,206542	1,116702	1,064570	1,290159	1,544583	1,492966

4.2.2 Resultados

Para obter os resultados para nosso segundo conjunto de dados, fizemos uso das mesmas técnicas de reamostragem utilizadas anteriormente. Vale ressaltar que dessa vez, como o grau de desbalanceamento é maior, é esperado que obtenhamos resultados diferentes.

Também é necessário pontuar que, por conta da alta demanda computacional para execução dos métodos de reamostagem propostos, foi necessário obter uma amostra do nosso conjunto de dados original para prosseguirmos com os algorítmos. Essa amostra foi estabelecida com o número $N = 30.000$ (mesmo tamanho do primeiro conjunto de dados que estudamos) e manteve o mesmo grau de desbalanceamento da base original (em torno de $6,7\%$). Em seguida, dividimos o conjunto de dados em dois subconjuntos conhecidos como conjunto de treinamento e conjunto de teste. A formação desses subconjuntos foi determinada por meio de uma seleção aleatória sem reposição do conjunto de dados original, assegurando que 80% das observações do conjunto de dados inicial fossem destinadas ao conjunto de treinamento, enquanto os 20% restantes fossem alocados ao conjunto de teste. É importante ressaltar que essa seleção foi realizada mantendo as proporções iniciais de clientes adimplentes e inadimplentes presentes na base de dados. Da mesma forma que no conjunto de dados anterior, realizamos essa divisão para avaliar a eficácia do classificador em um ambiente independente. No conjunto de treinamento, treinamos o classificador, enquanto no conjunto de teste avaliamos sua capacidade de classificar novas observações de forma precisa.

Para aplicação da regressão logística, primeiramente, consideramos todas as covariáveis discutidas na subseção anterior. Ao ajustar o modelo completo, obtivemos as estimativas dos coeficientes do modelo de regressão e p-valores conforme indicado na Tabela 4.18.

Tabela 4.18: Estimativas dos coeficientes obtidos no modelo logístico sem seleção de variáveis e sem aplicação de qualquer método de reamostragem.

	Estimativa	Erro padrão	Estatística	p-valor
(Intercepto)	-3.2438	0.0187	-173.38	0.0001
Proporção do limite utilizado	0.8089	0.0142	57.13	0.0001
Idade	-0.2538	0.0157	-16.21	0.0001
Quantidade de atrasos entre 30 e 59 dias	0.4328	0.0086	50.57	0.0001
Proporção da dívida em relação a renda	0.1322	0.0134	9.89	0.0001
Quantidade de empréstimos ativos	0.0544	0.0162	3.35	0.0008
Quantidade de empréstimos imobiliários	-0.0359	0.0163	-2.21	0.0274

De acordo com os testes de hipóteses individuais, a amostra trás evidências de que

todas as covariáveis possuem um efeito significativo sobre a resposta esperada quando considerado o modelo completo e um nível de significância de pelo menos 3%.

A partir da Figura 4.18, temos um indicativo de que as hipóteses do modelo logístico estão sendo satisfeitas de acordo com a amostra observada, salvo a presença de alguns outliers.

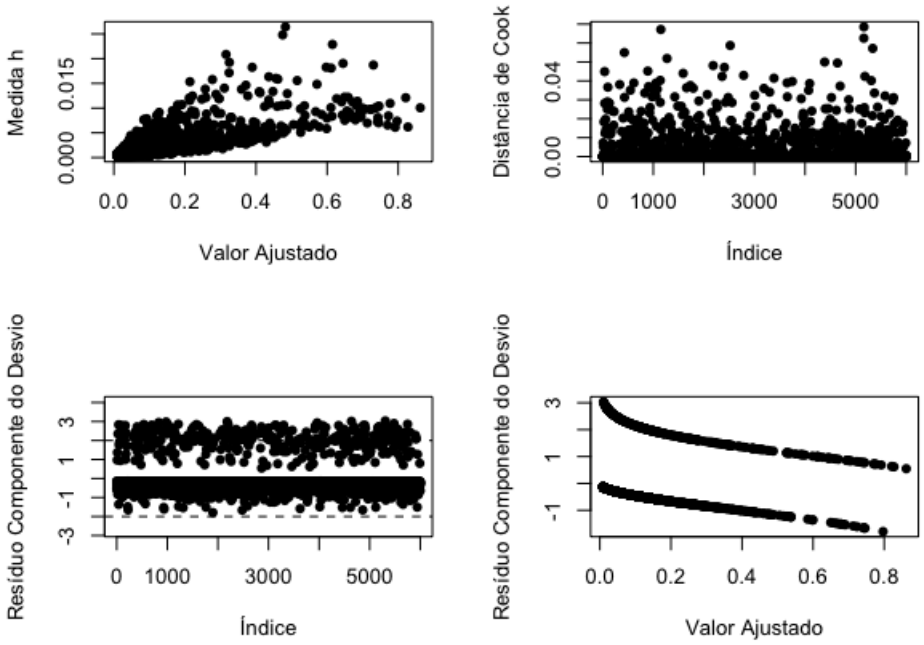

Figura 4.18: Análise de diagnóstico para o modelo logístico geral.

É importante relembrar que após a aplicação do pré-processamento dos dados, as estimativas, os p-valores dos testes e os gráficos da análise de diagnóstico sofreram alterações. Todavia, sem mudar as nossas conclusões com a relação a adequabilidade do modelo. Da mesma maneira que fizemos para o primeiro conjunto de dados, não trouxemos aqui tais resultados, uma vez que o foco deste trabalho é comparar o poder preditivo do modelo lógistico sem e com o pré-processamento dos dados.

Da mesma maneira que fizemos para o primeiro conjunto de dados, a constante que delimitará o ponto de corte foi o valor da probabilidade que resulta no valor do KS, ou seja, o valor de probabilidade que mais distancia a distribuição empírica das duas classes. Vale mencionar que esse ponto de corte baseado no KS sempre flutua próximo ao grau de desbalanceamento do conjunto de dados.

Na Tabela 4.19, apresentamos os resultados da aplicação da regressão logística no conjunto de teste. Nas colunas, estão representados as medidas utilizadas para medir o

Tabela 4.19: Medidas de performance aplicadas para método de reamostragem com e sem seleção de variáveis em que A: Modelo sem reamostragem, B: Modelo com Tomek Link, C: Modelo com SMOTE, D: Modelo com SMOTE e Tomek Link.

Modelo/Performance		KS	AUC	ACC	ESP	SEN	VPN	VPP	G-MÉDIA	MCC
S/ SELEÇÃO	A	0,5053	0,8245	0,7283	0,7245	**0,7805**	**0,9785**	0,1707	0,7520	0,2744
	B	0,5049	0,8251	0,7597	0,7438	0,7608	0,9761	0,1843	0,7522	0,2845
	C	0,5077	0,8255	0,7509	**0,7505**	0,7569	0,9770	0,1805	0,7537	0,2827
	D	**0,5161**	**0,8265**	**0,7714**	0,7415	0,7737	0,9751	**0,2001**	**0,7574**	**0,3005**
LASSO	A	0,5014	0,8210	0,7247	0,7206	**0,7805**	**0,9784**	0,1687	0,7500	0,2714
	B	0,5045	0,8218	**0,7597**	0,7346	0,7700	0,9778	0,1741	0,7521	0,2768
	C	0,5052	0,8247	0,7432	**0,7631**	0,7417	0,9773	0,1767	0,7524	0,2788
	D	**0,5129**	**0,8263**	0,7563	0,7556	0,7563	0,9760	**0,1914**	**0,7560**	**0,2927**

desempenho dos classificadores, em que ACC = Acurácia, SEN = Sensibilidade, VPP = Valor Preditivo Positivo, VPN = Valor Preditivo Negativo, G-MÉDIA = G-Média e MCC = Coeficiente de Correlação de Matthews. Nas linhas, estão dispostos os cenários nos quais a regressão logística foi ajustada. Os resultados estão apresentados em dois blocos, um trata-se do modelo completo com todas as covariáveis incluídas e o outro trata-se do modelo reduzido, apenas com as covariáveis selecionadas pelo LASSO. A marcação em negrito, destaca qual dos cenários teve melhor desempenho para as medidas avaliadas em cada um dos blocos. Para a aplicação da técnica SMOTE, novamente definimos $K = 5$. Vale ressaltar que não foi realizado nenhuma alteração no conjunto de teste.

Dessa vez, diferente do primeiro conjunto de dados, os cenários balanceados e desbalanceados não seguiram o mesmo padrão que anteriormente. Então, dessa vez, faremos a análise isolada para cada método separadamente.

A respeito dos resultados observados em termos de precisão do classificador, foi possível obervar que, de maneira geral, a classificação realizada sob o cenário D de reamostragem SMOTE + Tomek Link foi a que gerou os melhores resultados. Como tanto o KS, AUC, G-MÉDIA e MCC são medidas que buscam sumarizar a matriz de confusão como um todo, é possível notar que em D foram obtidos os melhores valores para todas essas medidas.

Ao avaliar a acurácia, ao contrário do primeiro conjunto de dados, verificou-se que as acurácias mais elevadas foram alcançadas após a aplicação do pré-processamento dos dados. Em todos os casos dos métodos B, C e D, a acurácia superou a do método A. Vale destacar que a acurácia mais alta foi registrada no método D, onde o conjunto de dados estava balanceado e não houve seleção de variáveis. Em D, aproximadamente 77% dos indivíduos foram classificados corretamente.

Supreendemente, a sensibilidade observada para a regressão logística, sem quaisquer

pré-processamento dos dados foi a maior dentre todos os métodos, incluindo os métodos com LASSO. De acordo com a sensibilidade observada no método A, temos que 78% dos clientes pertencentes à classe minoritária foram classificados corretamente como inadimplentes. A regressão logística sem nenhum método de reamostragem também foi a que obteve o maior VPN - Assim como era esperado. A especificidade foi maior para o método C.

Entretando, se o método A foi a melhor para os valores de sensibilidade, foi o pior para o VPP, onde aproximadamente 17% dos clientes classificados como inadimplentes pertenciam de fato a classe minoritária. Essa diferença não foi tão impactante comparado aos outros modelos (onde todos eles não performaram bem). Contudo, é possível observar que o método D foi o que obteve os melhores resultados para essa medida, o que nos fornece mais um argumento a favor desse cenário.

Ao analisar o caso com modelo reduzido aplicando LASSO, novamente observamos que, em todos os cenários, a performance do modelo logístico se manteve próxima àquela do modelo completo com todas as variáveis e vale ressaltar que para todos os modelos comparados até agora, incluindo o primeiro conjunto de dados, o LASSO tinha removido no máximo duas variáveis em apenas 1 dos modelos. Dessa vez, no cenário B, o LASSO zerou o coeficiente de 4 variáveis, sendo elas: "Idade", "Proporção de dívida sobre renda", "Quantidade de empréstimos ativos" e "Quantidade de empréstimos imobiliários". Sendo assim, permaneceu nesse modelo apenas as variáveis "Proporção do limite utilizado" e "Atraso entre 30 e 59 dias", além do intercepto. Nota-se, também, que o poder preditivo do cenário B com Lasso não foi muito diferente dos demais de maneira geral. Inclusive, foi preverível quando comparado a outros modelos. Dessa forma, esse modelo pode ser uma alternativa mais simples em relação aos demais.

O LASSO, em geral, não nos forneceu grandes diferenças dos resultados observados. Na maioria das vezes, quando comparamos as medidas obtidas para o mesmo método com e sem LASSO, o método de seleção de variáveis acaba performando um pouco pior. É possível observar que a performance dos modelos sob a performance do LASSO foi um espelho da performance sem seleção de variáveis, onde todas as melhores medidas foram as mesmas para os dois métodos, com exceção da acurácia, que no LASSO o cenário B obteve a melhor acurácia e sem LASSO o método D teve a melhor acurácia.

4.3 Discussão

Neste estudo, aplicamos a mesma abordagem de pré-processamento de dados a dois conjuntos de dados distintos. Com base nos resultados obtidos, optamos pelo **método híbrido de reamostragem** para ambos os conjuntos, o qual mostrou-se capaz de aumentar, em geral, o poder preditivo do nosso algoritmo de classificação além de estabelecer um balanceamento pleno entre as classes.

Para o primeiro conjunto de dados, observamos que os algoritmos melhoraram um pouco mais as medidas que eram de maior interesse, como a Sensibilidade e o VPP. Por outro lado, no segundo conjunto de dados, a sensibilidade, considerando o ponto de corte fixado, já apresentava valores satisfatórios, mesmo sem a aplicação de quaisquer métodos de reamostragem.

Embora nossos resultados tenham sido potencializados pelas técnicas de pré processamento de dados propostas nesse trabalho, é necessário nos atentarmos que, de maneira geral, a alteração dos resultados não foi de uma magnitude muito alta. Em alguns casos, a preferência por um modelo em relação ao outro, venceu por diferenças na segunda casa decimal.

Essa baixa diferença na performance de algumas medidas sugere o que alguns autores defendem: Em conjuntos de dados desbalanceados, somos capazes de contornar esse grau de desbalanceamento apenas alterando o valor de corte para a classificação de acordo com o grau de desbalanceamento.

No nosso caso, para toda a classificação realizada, foi determinado como ponto de corte, o valor da probabilidade que distância ao máximo as distribuições empíricas das classes, essa distância associada a esse valor de probabilidade é o valor da estatísticas KS. Sendo assim, padronizando esse valor dinâmico do ponto de corte, já foi possível observar que a qualidade do classificador melhora bastante por si só.

O nível de severidade do desbalanceamento não afetou tanto nossa classificação, inclusive, medidas mais gerais a respeito da performance dos modelos apontaram que os mesmos performaram melhor em nosso segundo conjunto de dados, em que o nível de desbalanceamento era de aproximadamente $6,7\%$ comparado a aproximadamente 22% do primeiro conjunto de dados.

Dessa forma, baseado nos resultados observados, embora possamos dizer que alterar o nível de corte é uma solução bastante simples e que se mostrou efetiva, não podemos des-

cartar que, para ambos os conjuntos de dados, foi preferível a realização de algum método de reamostragem sob o modelo original em quase todas as medidas de performance, o que argumenta em favor da utilização desses métodos para refinamento do modelo.

Um ponto que também precisa de atenção (esse, que argumenta a favor da alteração do ponto de corte) é a capacidade computacional exigida para a realização dos métodos de reamostragem propostos. O método SMOTE não exigiu muito computacionalmente, até porque o algorítmo é mais simples e atua apenas dentro da classe minoritária, gerando indivíduos similares as características observadas de cada um. Já no método Tomek Link, foi exigido uma alta capacidade computacional para realização do método, já que ele calcula a distância 1 a 1 de todos os elementos entre si para, em seguida, observar o valor referente a variável resposta.

Capacidade computacional essa que não fomos capazes de obter para realização desse trabalho com os dados completos. Para performarmos os métodos de pré-processamento de dados nesse segundo conjunto de dados, foi necessário que fosse tirada uma amostra desses dados, resultando em uma base de dados de tamanho $N = 30.000$, 5 vezes menor que a base original, que continha $N = 150.000$ observações.

Na literatura, o método mais reconhecido e amplamente documentado para reamostragem em contextos de classificação é o SMOTE. Além da efetividade em balancear o conjunto de treinamento, os resultados provenientes do uso do SMOTE também se revelaram bastante satisfatórios. Contudo, ainda que tenham sido favoráveis, esses resultados não conseguiram atingir o mesmo patamar de desempenho apresentado pelo método híbrido. Porém, é importante ressaltar que, devido à sua menor demanda computacional, o SMOTE representa uma alternativa bastante interessante e, por esse motivo, é amplamente utilizado no contexto de classificação.

Ainda em relação ao SMOTE, é importante mencionar que ele se destaca por ser o único método que busca atingir o balanceamento completo do conjunto de treinamento. Em comparação, o método Tomek Link corrige parcialmente o desbalanceamento, mas não alcança a mesma plenitude proporcionada pelo SMOTE.

Em resumo, os métodos de reamostragem aplicados aos dois conjuntos de dados demonstraram a capacidade de aprimorar a eficiência do nosso classificador em termos preditivos. Dessa forma, observamos que o objetivo de melhorar a classificação foi atingido, entretanto, o custo computacional pode ser um impeditivo para implementação desses métodos dependendo do cenário. Assim, para pesquisadores com recursos computaci-

onais suficientes que buscam refinar seus modelos, esses métodos se mostraram como alternativas eficientes que cumpriram com o propósito estabelecido desde o início do estudo.

Capítulo 5

Considerações Finais

Em resumo, nossa análise sugere que a aplicação de métodos de balanceamento de classes pode resultar em um aprimoramento geral do desempenho do modelo logístico na classificação de novas instâncias. Essa conclusão enfatiza a relevância de considerar abordagens de balanceamento ao desenvolver modelos preditivos. Todavia, ressaltamos, que o custo computacional pode ser alto a depender do tamanho do base a ser analisada.

Os próximos passos desse trabalho são:

- Avaliar o impacto do balanceamento das classes da variável de interesse em um contexto em que o nível de desbalanceamento é bem severo, numa proporção, por exemplo, $99 - 1$ entre unidades amostrais da classe majoritária e minoritária.

- Considerar métodos que de fato realizem o balanceamento das classes da variável resposta a partir de subamostragem. Nesse trabalho, embora o algoritmo Tomek Link seja considerado uma método de subamostragem ele apenas realiza uma limpeza, excluindo da classe majoritária aquelas unidades amostrais muito similares a unidades da classe minoritária.

- Realizar um estudo de sensibilidade do tempo computacional para execução dos algoritmos de acordo com o número de variáveis presentes no modelo.

- Comparar a eficácia dos métodos de reamostragem com outras metodologias utilizadas para lidar com desbalanceamento de classes tais como classificadores sensíveis ao custo e métodos *ensemble*.

Referências Bibliográficas

Batista, G. E., Prati, R. C. e Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD Explorations Newsletter, **6**(1), 20–29.

Chawla, N. V., Bowyer, K. W., Hall, L. O. e Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, **16**, 321–357.

Chawla, N. V., Cieslak, D. A., Hall, L. O. e Joshi, A. (2008). Automatically countering imbalance and its empirical relationship to cost. Data Mining and Knowledge Discovery, **17**(2), 225–252.

Dembczynski, K., Jachnik, A., Kotlowski, W., Waegeman, W. e Hüllermeier, E. (2013). Optimizing the f-measure in multi-label classification: Plug-in rule approach versus structured loss minimization. Em International Conference on Machine Learning, páginas 1130–1138. PMLR.

Estabrooks, A., Jo, T. e Japkowicz, N. (2004). A multiple resampling method for learning from imbalanced data sets. Computational Intelligence, **20**(1), 18–36.

Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B. e Herrera, F. (2018). Learning from imbalanced data sets, volume 10. Springer.

Galar, M., Fernandez, A., Barrenechea, E., Bustince, H. e Herrera, F. (2011). A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), **42**(4), 463–484.

García, V., Sánchez, J. S. e Mollineda, R. A. (2012). On the effectiveness of preproces-

sing methods when dealing with different levels of class imbalance. Knowledge-Based Systems, **25**(1), 13–21.

He, H. e Ma, Y. (2013). Imbalanced learning: foundations, algorithms, and applications.

Hosmer, D. W., Jovanovic, B. e Lemeshow, S. (1989). Best subsets logistic regression. Biometrics, páginas 1265–1270.

James, G., Witten, D., Hastie, T. e Tibshirani, R. (2013). An introduction to statistical learning, volume 112. Springer.

Krawczyk, B. (2016). Learning from imbalanced data: open challenges and future directions. Progress in Artificial Intelligence, **5**(4), 221–232.

Kubat, M., Holte, R. e Matwin, S. (1997). Learning when negative examples abound. Em European Conference on Machine Learning, páginas 146–153. Springer.

Ling, C. X., Sheng, V. S. e Yang, Q. (2006). Test strategies for cost-sensitive decision trees. IEEE Transactions on Knowledge and Data Engineering, **18**(8), 1055–1067.

Liu, X.-Y., Wu, J. e Zhou, Z.-H. (2008). Exploratory undersampling for class-imbalance learning. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), **39**(2), 539–550.

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of t4 phage lysozyme. Biochimica et Biophysica Acta (BBA)-Protein Structure, **405**(2), 442–451.

Napierała, K., Stefanowski, J. e Wilk, S. (2010). Learning from imbalanced data in presence of noisy and borderline examples. Em International Conference on Rough Sets and Current Trends in Computing, páginas 158–167. Springer.

Polikar, R. (2006). Ensemble based systems in decision making. IEEE Circuits and Systems Magazine, **6**(3), 21–45.

Sicsú, A. L. (2010). Credit Scoring: desenvolvimento, implantação, acompanhamento. Blucher.

Stefanowski, J. e Wilk, S. (2008). Selective pre-processing of imbalanced data for improving classification performance. Em International Conference on Data Warehousing and Knowledge Discovery, páginas 283–292. Springer.

Tomek, I. (1976). Two modifications of cnn. IEEE Transactions Systems, Man and Cybernetics, **6**, 769–772.

Wang, H., Xu, Q. e Zhou, L. (2015). Large unbalanced credit scoring using lasso-logistic regression ensemble. PloS One, **10**(2), e0117844.

Wilson, D. L. (1972). Asymptotic properties of nearest neighbor rules using edited data. IEEE Transactions on Systems, Man, and Cybernetics, (3), 408–421.

Yeh, I.-C. e Lien, C.-h. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert systems with applications, **36**(2), 2473–2480.

Zhang, S., Liu, L., Zhu, X. e Zhang, C. (2008). A strategy for attributes selection in cost-sensitive decision trees induction. Em 2008 IEEE 8th International Conference on Computer and Information Technology Workshops, páginas 8–13. IEEE.