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

CLASSIFICAÇÃO BINÁRIA DE DADOS FINANCEIROS EM PROBLEMAS COM CLASSES DESBALANCEADAS

Lucas Fernando de Moura

Trabalho de Conclusão de Curso

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

CLASSIFICAÇÃO BINÁRIA DE DADOS FINANCEIROS EM PROBLEMAS COM CLASSES DESBALANCEADAS

Lucas Fernando de Moura
Orientador: Prof. Dr. Ricardo Felipe Ferreira

Trabalho de Conclusão de Curso apresentado como parte dos requisitos para obtenção do título de Bacharel em Estatística.

São Carlos
Fevereiro de 2024

Federal University of São Carlos
Exact and Technology Sciences Center
Department of Statistics

BINARY CLASSIFICATION OF FINANCIAL DATA IN PROBLEMS WITH IMBALANCED CLASSES

Lucas Fernando de Moura
Advisor: Prof. Dr. Ricardo Felipe Ferreira

Bachelors dissertation submitted to the Department of Statistics, Federal University of São Carlos - DEs-UFSCar, in partial fulfillment of the requirements for the degree of Bachelor in Statistics.

São Carlos
February 2024

Lucas Fernando de Moura

CLASSIFICAÇÃO BINÁRIA DE DADOS FINANCEIROS EM PROBLEMAS COM CLASSES DESBALANCEADAS

Este exemplar corresponde à redação final do trabalho de conclusão de curso devidamente corrigido e defendido por Lucas Fernando de Moura e aprovado pela banca examinadora.

Aprovado em 29 de Janeiro de 2024

Banca Examinadora:

- Prof. Dr. Ricardo Felipe Ferreira
- Profa. Dra. Daiane Aparecida Zuanetti
- Prof. Dr. Renato Jacob Gava

Dedico este trabalho aos meus pais e avó, graças ao esforço deles que cheguei até aqui.

Agradecimentos

Em primeiro lugar, agradeço a Deus por me proporcionar perseverança e sempre me mostrar o caminho correto.

Aos meus pais, Sergio e Ana Paula, por nunca terem medido esforços para me proporcionar um ensino de qualidade durante todo o meu período escolar e por todo incentivo em minhas decisões. Á minha avó, Iraide, por me apoiar em toda vida. Á minha irmã, Taline, por todas as conversas e confiança em mim. Á minha namorada, Marianna, por todo companheirismo e compreensão durante a realização desse trabalho.

Ao meu orientador Ricardo, por todos os ensinamentos, conversas, reflexões e amizade desde o primeiro ano de faculdade. Seus ensinamentos e incentivos foram essenciais para eu chegar até aqui.

Aos meus colegas, por compartilharem comigo tantos momentos de aprendizado, dificuldade e felicidade.

Por fim, reconheço também todos aqueles que contribuíram, de alguma forma, para a realização deste trabalho e para minha formação como pessoa e profissional.

"We cannot live better than in seeking to become better."

(Socrates)

Resumo

Com o intuito de diminuir os riscos e incertezas associados à concessão de crédito, as instituições financeiras estão constantemente explorando métodos para aperfeiçoar o sistema de avaliação creditícia. No mesmo contexto, o crescimento no volume de transações com cartões de crédito resultou no aumento das fraudes, ocasionando perdas bilionárias anuais para as instituições financeiras. Logo, é fundamental que as empresas sejam capazes de detectar efetivamente transações fraudulentas. Uma maneira de minimizar as perdas decorrentes da inadimplência ou da fraude é utilizar métodos estatísticos que gerem resultados próximos à realidade, apresentando uma baixa margem de erro. No entanto, a grande dificuldade na execução desse processo é que esses tipos de dados financeiros são desbalanceados, isto é, observamos uma maior proporção de clientes adimplentes e transações legítimas (grupos majoritários) do que de clientes inadimplentes e transações fraudulentas (grupos minoritários). Esse desequilíbrio acarreta em um viés de classificação, uma vez que os algoritmos de aprendizagem tendem a classificar melhor as observações do grupo majoritário. Nesse sentido, este trabalho tem como proposta realizar um estudo comparativo da performance das máquinas de vetores suporte com a regressão logística na classificação de novas unidades amostrais. Esse estudo será realizado a partir de três conjuntos de dados financeiros com diferentes graus de desbalanceamento, considerando três contextos: (*i*) sem aplicar técnica alguma para lidar com o desbalanceamento dos conjuntos de dados; (*ii*) aplicando técnicas de pré-processamento de dados para lidar com os desbalanceamento dos conjuntos de dados; e (*iii*) utilizando a versão sensível ao custo dos classificadores originais para lidar com o desbalanceamento dos conjuntos de dados. A análise da performance dos classificadores dar-se-á a partir de medidas baseadas na matriz de confusão que tem se mostrado menos sensíveis ao desbalanceamento dos dados, tais como a G-média, o coeficiente de Mathews e o F-Score.

Palavras-chave: *Classificadores sensíveis ao custo, desbalanceamento de classes, máquinas de vetores suporte, métodos de pré-processamento de dados, regressão logística.*

Abstract

In order to mitigate the risks and uncertainties associated with credit granting, financial institutions are constantly exploring methods to enhance the credit evaluation system. In the same context, the growth in credit card transactions has led to an increase in fraud, resulting in billions of dollars in annual losses for financial institutions. Therefore, it is crucial for companies to effectively detect fraudulent transactions. One way to minimize losses due to default or fraud is to use statistical methods that yield results close to reality, presenting a low margin of error. However, the major challenge in executing this process is that such financial data is imbalanced, meaning there is a higher proportion of non-defaulting customers and legitimate transactions (majority groups) than delinquent customers and fraudulent transactions (minority groups). This imbalance leads to a classification bias, as learning algorithms tend to classify observations from the majority group better. In this context, this work aims to conduct a comparative study of the performance of support vector machines and logistic regression in classifying new sample units. This study will be carried out using three financial datasets with different degrees of imbalance, considering three contexts: (*i*) without applying any technique to handle the imbalance of the datasets; (*ii*) applying data preprocessing techniques to handle the imbalance of the datasets; and (*iii*) using the cost-sensitive version of the original classifiers to handle the imbalance of the datasets. The analysis of classifier performance will be based on measures derived from the confusion matrix that have been shown to be less sensitive to data imbalance, such as the G-mean, Matthews correlation coefficient, and F-score.

Keywords: *Cost-sensitive classifiers, class imbalance, suport vector machines, data preprocessing techniques, logistic regression.*

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

O aprendizado estatístico refere-se ao vasto conjunto de ferramentas cujo principal objetivo é aprender padrões a partir dos dados. Essas ferramentas podem ser classificadas como supervisionadas e não-supervisionadas. Grosso modo, as técnicas de aprendizagem supervisionada envolvem a construção de um modelo estatístico para predizer o valor de uma variável de saída a partir do conhecimento de uma ou mais variáveis de entrada. Por outro lado, as técnicas de aprendizagem não-supervisionada buscam aprender relações ou estruturas presentes nos dados a partir de um conjunto de variáveis de entradas. Neste trabalho, estamos interessados em estudar técnicas de aprendizagem supervisionada. Nesse contexto, o problema de aprendizado estatístico 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. Nesta monografia, estamos interessados no problema de classificação binária de dados financeiros.

Muitos problemas de classificação binária exibem um desequilíbrio no número de unidades amostrais pertencentes a cada uma das classes (Kubat *et al.*, 1998; Laradji *et al.*, 2015; Márquez-Vera *et al.*, 2016; Tran e Liatsis, 2016; Zhu *et al.*, 2017). Qualquer conjunto de dados cujas classes sigam distribuições distintas é, tecnicamente, desbalanceado. No entanto, essa não é uma definição muito útil uma vez que, praticamente, nenhum conjunto de dados é perfeitamente balanceado. Nesse sentido, na literatura, um conjunto de dados é dito ser desbalanceado quando uma das classes é significantemente mais numerosa do que a outra, i.e., quando existe uma diferença significativa entre as distribuições das classes (He e Ma, 2013; Fernández *et al.*, 2018; Singh e Khim, 2021). Todavia, não há um consenso em relação ao grau de desequilíbrio que torna significativo o desbalanceamento entre as classes. Neste trabalho, vamos seguir a proposta de He e Ma (2013) e definir que

conjuntos de dados em que a classe majoritária é composta por menos do que 75% das unidades amostrais são ditos possuir um desbalanceamento leve, conjunto de dados em que a classe majoritária é composta por, aproximadamente, 90% das unidades amostrais são ditos possuir um desbalanceamento moderado e os conjuntos de dados cuja classe majoritária é composta por, aproximadamente, 99% das unidades amostrais são ditos possuir um desbalanceamento severo.

Nos problemas de classificação binária, a classe minoritária é a menos representativa e, usualmente, é referida como classe positiva, enquanto a classe majoritária é referida como classe negativa (Fernández *et al.*, 2018). A classe positiva é, em geral, a classe de interesse em muitos problemas práticos e, por ser a menos representativa, esse interesse se torna uma das principais questões em problemas de classificação envolvendo desbalanceamento de classes, uma vez que a classificação incorreta de novas unidades amostrais pertencentes à classe de interesse (minoritária) é mais frequente do que aquelas pertencentes à classe majoritária. Obviamente, gostaríamos de uma alta acurácia na classificação de novas unidades amostrais para ambas às classes. No entanto, na presença de desbalanceamento, os classificadores tendem a ter uma ótima acurácia na classe majoritária e apresentam baixo desempenho na classificação de unidades amostrais pertencentes à 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 das unidades amostrais 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 comparar empiricamente a performance do modelo de regressão logística (James *et al.*, 2013) com a performance da máquina de vetores suporte (Vapnik, 1999) na classificação de novas unidades amostrais em três cenários distintos, nos quais observa-se diferentes graus de desbalanceamento entre as classes envolvidas, além de comparar a performance do classificador logístico com ponto de corte padrão $\frac{1}{2}$ com o classificador logístico com ponto de corte sendo o ponto que maximiza as métricas G-Média e MCC (Coeficiente de correlação de Matthews). Os mecanismos que vamos utilizar para lidar com o desbalanceamento das classes são tanto externos, isto é, técnicas que envolvem o pré-processamento dos dados a partir de métodos de reamostragem, quanto internos, ou seja, técnicas que envolvem mudanças no classificador original.

Os métodos de pré-processamento dos dados consistem em modificar o conjunto de treinamento desbalanceado a fim de produzir uma distribuição mais balanceada das unidades amostrais entre as classes, o que permite que os classificadores performem de maneira similar aos problemas de classificação onde a severidade do desbalanceamento é baixa. 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. Nesta monografia, pretendemos estudar comparativamente a performance da regressão logística e da máquina de vetores suporte na classificação de novas instâncias nos seguintes cenários: (*i*) conjunto de treinamento balanceado por meio do algoritmo de subamostragem Tomek Link + One-Sided-Selection (Tomek, 1976); (*ii*) conjunto de treinamento balanceado por meio de algoritmo de sobreamostragem SMOTE (Chawla *et al.*, 2002); e (*iii*) conjunto de treinamento balanceado por meio do algoritmo híbrido SMOTE + Tomek Link .

Os métodos internos consistem em modificações no classificador original a fim de aliviar

seu viés em relação à classe majoritária, poupando-nos de realizar alterações no conjunto de treinamento (He e Ma, 2013; Fernández *et al.*, 2018). No entanto, para propor uma modificação no algoritmo de aprendizagem é necessário, primeiramente, entender o que atrapalha o seu desempenho. No que diz respeito aos classificadores que apresentam uma alta acurácia na classificação de unidades amostrais pertencentes à classe majoritária e uma baixa acurácia na classe minoritária, possíveis modificações consistem na utilização de diferentes custos para a má classificação de unidades amostrais de modo que a classe minoritária tenha maior penalização do que a classe majoritária. Neste trabalho, pretendemos estudar comparativamente dois métodos internos: as máquinas de vetores suporte sensíveis ao custo e a regressão logística sensível ao custo.

Máquinas de vetores suporte (SVM, do inglês *Support Vector Machine*) foi introduzida por Vapnik (1998), no contexto da teoria de aprendizagem estatística. É um conceito que toma como entrada um conjunto de dados e prediz, para cada nova entrada dada, qual de duas possíveis classes essa entrada pertence. Portanto, o SVM é um classificador binário não-probabilístico. Essencialmente, essa técnica constrói um hiperplano no espaço das variáveis que estão sendo consideradas no estudo através de algum mapa, escolhido *a priori*. Esse hiperplano divide o espaço das variáveis em dois subconjuntos, de tal forma que a separação entre os subconjuntos seja tão ampla quanto possível. As novas entradas são, então, mapeadas no espaço das variáveis e preditas como pertencentes a uma das duas possíveis classes baseada em qual subconjunto são colocadas. Em problemas de classificação em que observa-se um desequilíbrio entre as classes envolvidas, podemos utilizar as máquinas de vetores suporte sensível ao custo para realizar a classificação de novas unidades amostrais. Essa é uma técnica proposta por Veropoulos *et al.* (1999) que modifica o classificador original a partir da atribuição de custos distintos às classes em análise, de modo que errar a classificação de uma observação da classe minoritária seja mais penalizado do que errar uma classificação da classe majoritária.

A regressão logística é um modelo estatístico utilizado para realizar a classificação de uma variável binária a partir de um conjunto de variáveis preditoras. A regressão logística utiliza uma função de ligação, em geral, logística para realizar a transformação linear das variáveis explicativas em uma probabilidade de pertencer à classe positiva, e, a partir da estimativa dessa probabilidade, classificar novas unidades amostrais como pertencentes a uma das duas classes. Em problemas com desbalanceamento entre as classes, podemos utilizar uma versão da regressão logística que também seja sensível ao custo (Shen *et al.*,

2020; Sushma *et al.*, 2023).

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, vamos desenvolver, neste trabalho, um estudo comparativo da performance das máquinas de vetores suporte com a performance da regressão logística na classificação de novas unidades amostrais a partir de três conjuntos de dados financeiros: dois relativos a análise de crédito e outro relativo a detecção de fraude.

O primeiro conjunto de dados é sobre a inadimplência de clientes de cartão de crédito e possui um desbalanceamento leve, sendo 77% das unidades amostrais compostas por clientes adimplentes, disponível em: `http://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients`. O segundo conjunto de dados é sobre inadimplência de crédito e possui um desbalanceamento moderado, sendo 93% das unidades amostrais compostas por clientes adimplentes, disponível em `https://www.kaggle.com/competitions/GiveMeSomeCredit/overview/description`. O terceiro conjunto de dados é sobre fraudes em transações de crédito e possui um desbalanceamento severo, sendo 99% das transações legítimas, esse conjunto foi disponibilizado por uma instituição financeira, não sendo possível identificar os clientes, servindo apenas para fins de pesquisa e estudos e, por motivos de confidencialidade, as covariáveis serão tratadas com nomes fictícios, disponível em `https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud`. Além de comparar a performance dos classificadores sob o efeito de diferentes graus de desbalanceamento, vamos realizar esse estudo considerando três contextos: (i) sem aplicar técnica alguma para lidar com o desbalanceamento dos conjuntos de dados; (ii) aplicando técnicas de

pré-processamento de dados para lidar com o desbalanceamento dos conjuntos de dados; e (*iii*) utilizando a versão sensível ao custo dos classificadores supramencionados para lidar com o desbalanceamento dos conjuntos de dados.

Até onde vai nosso conhecimento, não há trabalho na literatura que tenha comparado a performance das máquinas de vetores suporte com a regressão logística no contexto de dados financeiros da maneira como pretendemos realizar, isto é, sob o efeito de diferentes níveis de desbalanceamento e comparando o ganho adquirido com a aplicação do pré-processamento dos dados com aqueles obtidos através da versão sensível ao custo dos classificadores originais. Portanto, este estudo complementa os estudos sobre métodos de detecção de fraude e métodos para análise de crédito.

Este trabalho está organizado da seguinte maneira. No próximo capítulo, estudamos a metodologia dos classificadores de regressão logística e máquina de vetores suporte que serão utilizados para a classificação de novas unidades amostrais em nossos conjuntos de dados. Ainda neste capítulo, apresentamos as medidas de performance que serão utilizadas para a comparação da performance de cada classificador. No Capítulo 3 estudamos os métodos de pré-processamento, metodologias que nos ajudam a superar a perda de precisão na classificação das observações da classe minoritária por conta do desbalanceamento dos dados. No capítulo 4 apresentamos o aprendizado sensível ao custo, que é uma abordagem de modificação em um nível do algoritmo para lidar com o desequilíbrio de classes. No Capítulo 5, aplicamos as metodologias propostas nos bancos de dados financeiros descritos anteriormente. Por fim, o Capítulo 6 encerra esta monografia com algumas considerações finais acerca de todo trabalho.

Capítulo 2

Classificadores

A classificação é uma tarefa fundamental na área de aprendizado de máquina, na qual o objetivo é classificar observações a partir do conhecimento de um conjunto de variáveis relevantes para a situação. A regressão logística e o SVM são abordagens poderosas para enfrentar esse desafio, fornecendo soluções eficientes e eficazes para problemas de classificação. A regressão logística, baseia-se em um modelo estatístico que busca estimar a probabilidade de uma determinada observação pertencer a uma classe específica dado um conjunto de covariáveis. Enquanto, máquina de vetores suporte, baseia-se em um método de aprendizado de máquina que busca encontrar o hiperplano que melhor separa as diferentes classes no espaço de observações. Afim de comparar os dois diferentes métodos de classificação, fornecemos uma comparação entre os métodos através de um exemplo simples utilizando um dos conjuntos de dados que será utilizado à posteriori no trabalho. A comparação será feita por meio de medidas de performance que serão detalhadas neste capítulo, as quais serão responsáveis pela análise de desempenho dos classificadores.

2.1 Uma visão geral sobre classificação

O termo aprendizagem estatística é relativamente novo, mas muitos dos conceitos subjacentes foram desenvolvidos há muito tempo. No século XIX, Legendre e Gauss introduziram o método dos mínimos quadrados, que deu origem à regressão linear. Fisher, propôs a análise de discriminante linear para prever variáveis qualitativas. Na década de 1970, Nelder e Wedderburn criaram o termo modelos lineares, que englobam métodos como regressão linear e logística. Na década de 1980, avanços computacionais permitiram o uso de métodos não lineares como, por exemplo, as árvores de regressão. Desde

então, com os avanços tecnológicos, novos métodos baseados em abordagens algorítmicas passaram a serem propostos com a finalidade de treinar métodos de inteligência artificial.

O aprendizado estatístico baseia-se em métodos e/ou modelos matemáticos com a finalidade de entender os dados. Essas metodologias podem ser classificadas como supervisionadas e não-supervisionadas. O aprendizado estatístico supervisionado envolve a construção de um modelo para predizer o valor de uma variável de saída a partir do conhecimento de uma ou mais variáveis de entrada. Por outro lado, o aprendizado não-supervisionado busca aprender relações ou estruturas presentes nos dados a partir de um conjunto de variáveis de entrada.

Neste trabalho, estamos interessados em classificar transações em fraudulentas ou legítimas, a partir do conhecimento de um conjunto de variáveis relevantes para o mercado de cartões de crédito. De forma análoga, pretendemos utilizar variáveis do mercado financeiro para classificar clientes como inadimplentes ou adimplentes. Portanto, nesse contexto, estamos interessados em estudar técnicas de aprendizagem supervisionada.

Os classificadores construídos por meio de uma abordagem supervisionada podem ser obtidos a partir de diferentes metodologias, tais como regressão logística e máquina de vetores suporte, por exemplo. Nesse sentido, temos dois tipos de variáveis que compõem o modelo: as variáveis de entrada e a variável de saída. As variáveis de entrada, também denominadas variáveis preditoras, são tipicamente denotadas por $X_1, \ldots, X_p$, em que p é um número inteiro positivo. A variável de saída é tipicamente denotada por Y e também é conhecida como variável resposta. Em um cenário de avaliação de risco de crédito, por exemplo, queremos classificar um cliente como inadimplente e adimplente, ou seja, mau pagador e bom pagador, respectivamente. Para isso, utilizamos a variável de saída como sendo $Y = 1$, se o cliente é inadimplente e $Y = 0$, se o cliente é adimplente. Por fim, as variáveis de entrada são definidas por $X_1, X_2, ..., X_p$, onde, por exemplo, $X_1 =$ histórico de crédito, $X_2 =$ renda, entre outras variáveis relevantes para a análise.

A grosso modo, qualquer método que consiga discriminar indivíduos ou objetos em classes ou categorias de uma variável qualitativa de interesse é denominado um método de classificação. Os problemas de regressão e classificação são bem similares, pois ambos utilizam dados de entrada já observados para prever uma resposta. A diferença é que nos problemas de classificação procura-se estimar a classe ou categoria de uma nova unidade amostral e não um valor numérico, como é o caso da regressão.

Um dos desafios para os métodos de classificação consiste em lidar com dados desba-

lanceados, uma vez que o desbalanceamento afeta diretamente a classificação final. No caso de classificação de operações em fraudulentas ou legítimas, por exemplo, o número de observações da classe de operações legítimas (classe majoritária) é muito superior ao número de observações da classe de operações fraudulentas (classe minoritária) e, por esse motivo, a classificação final acaba sendo viesada a favor da classe majoritária, apresentando um baixo desempenho na classificação de unidades amostrais pertencentes à classe minoritária. Nesse contexto, acabamos classificando erroneamente operações fraudulentas como sendo legítimas, o que impacta negativamente empresas financeiras, uma vez que esses erros prejudicam sua reputação, aumentam os custos operacionais e as expõem a riscos regulatórios. É essencial para a saúde financeira e a sustentabilidade das empresas do ramo financeiro garantir a detecção e a prevenção eficaz de fraudes para proteger seus próprios interesses e os de seus clientes.

A fim de controlar o efeito causado pelo desbalanceamento na classificação de novas unidades amostrais, podemos considerar técnicas de pré-processamento de dados ou a versão sensível ao custo dos classificadores. Portanto, nesta monografia, estamos interessados em comparar empiricamente a performance do modelo de regressão logística, método com ampla aceitação no mercado financeiro, com a performance da máquina de vetores suporte, que é uma metodologia que tem sido utilizada como uma boa alternativa pra lidar com conjuntos de dados desbalanceados. Esse estudo comparativo é realizado em três cenários distintos, nos quais observa-se diferentes graus de desbalanceamento entre as classes envolvidas. Os mecanismos que vamos utilizar para lidar com o desbalanceamento das classes são tanto externos, isto é, técnicas que envolvem o pré-processamento dos dados a partir de métodos de reamostragem, quanto internos, ou seja, técnicas que envolvem mudanças no classificador original.

2.2 Regressão logística

A regressão logística é um modelo de regressão utilizado para modelar a relação de uma variável resposta binária com um conjunto de covariáveis, as quais podem ser quantitativas ou qualitativas. Especificamente, na área financeira, a regressão logística desempenha um papel fundamental na detecção de fraudes e na avaliação de riscos de crédito. No contexto de detecção de fraudes, esse método é aplicado para identificar transações fraudulentas com base em variáveis explicativas como histórico de transações, padrões de gastos e

comportamento do cliente. No caso da avaliação de riscos de crédito, a regressão logística é usada para estimar a probabilidade de inadimplência de um indivíduo ou empresa com base em informações como histórico de crédito, renda, emprego e outros fatores relevantes. A capacidade da regressão logística de lidar com variáveis preditoras binárias e contínuas, juntamente com sua interpretabilidade, torna-a uma ferramenta valiosa na tomada de decisões financeiras relacionadas à fraude e ao crédito.

Sejam p e n números inteiros positivos não-nulos tais que $p < n$. Considere um conjunto de dados compostos por p covariáveis (variáveis preditoras), que descrevem características de n unidades amostrais, e sejam $Y_1, Y_2, ..., Y_n$ variáveis aleatórias independentes definidas em um espaço de probabilidade $(\Omega, \mathcal{F}, P)$ 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 $i = 1, 2, ..., n$.

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 na 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 $j = 1, 2, \ldots, p$. Analogamente, 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

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

Definido um plano amostral, podemos coletar uma amostra com n unidades amostrais e observar em cada uma delas a variável que indica a qual classe tal unidade pertence e também cada uma das p covariáveis especificadas no estudo. Seja $(\boldsymbol{x}_1, y_1)$, $(\boldsymbol{x}_2, y_2)$, ..., $(\boldsymbol{x}_n, y_n)$ 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 $i = 1, 2, \ldots, n$.

Na regressão logística, modelamos a probabilidade de uma nova unidade amostral

pertencer à classe minoritária, condicionada à observação $\boldsymbol{x}$ de um vetor de covariáveis de ordem $1 \times p$, como uma função desse vetor de covariáveis, ou seja,

$$P(Y = 1|\boldsymbol{x}) = \pi(\boldsymbol{x}) := \frac{e^{\beta_0 + \boldsymbol{x}\boldsymbol{\beta}}}{1 + e^{\beta_0 + \boldsymbol{x}\boldsymbol{\beta}}}, \tag{2.1}$$

em que $\beta_0 \in \mathbb{R}$ representa o intercepto, $\boldsymbol{\beta} := (\beta_1, \beta_2, ..., \beta_p)^\top$ 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.

É possível reescrever a equação acima na forma linear, aplicando uma transformação logarítmica, ou seja,

$$\log\left(\frac{\pi(\boldsymbol{x})}{1 - \pi(\boldsymbol{x})}\right) = \beta_0 + \boldsymbol{x}\boldsymbol{\beta}.$$

O lado esquerdo da igualdade anterior é chamado de *log-odds* ou logito, uma vez que a razão

$$\frac{\pi(\boldsymbol{x})}{1 - \pi(\boldsymbol{x})} \tag{2.2}$$

é chamada de *odds*. As *odds* podem assumir qualquer valor não-negativo, de forma que valores de *odds* próximas a 0 ou muito grandes indicam, respectivamente, probabilidade muito baixa ou muito alta da classificação pertencer a classe minoritária.

Neste trabalho, vamos estimar os parâmetros β_0 e $\boldsymbol{\beta}$ do modelo de regressão logística por meio do método da máxima verossimilhança. Para isso, assumiremos que $Y_1, Y_2, ..., Y_n$ são variáveis aleatórias e que $Y_i|\boldsymbol{X}_i = \boldsymbol{x}_i \sim Bernoulli(\pi(\boldsymbol{x}_i))$, em que $\boldsymbol{X}_i$ é o vetor de covariáveis associadas a i-ésima unidade amostral, em que $i = 1, 2, \ldots, n$. Sendo assim, podemos escrever a distribuição de $Y_i|\boldsymbol{X}_i = \boldsymbol{x}_i$ da seguinte forma:

$$P(Y_i = y_i|\boldsymbol{x}_i) = \pi(\boldsymbol{x}_i)^{y_i}(1 - \pi(\boldsymbol{x}_i))^{1-y_i}\mathbb{I}_{\{0,1\}}(y_i) \tag{2.3}$$

em que $\mathbb{I}$ denota a função indicadora, por exemplo,

$$\mathbb{I}_{\{0,1\}}(y_i) = \begin{cases} 1, & \text{se } y_i \in \{0, 1\}, \\ 0, & \text{se caso contrário.} \end{cases}$$

Em um cenário de previsão utilizando regressão logística, é comum dividir o conjunto de dados em conjuntos de treinamento e teste. Essa divisão é feita para avaliar a capacidade do modelo generalizar padrões aprendidos durante o treinamento e para estimar o desempenho do modelo em dados não vistos anteriormente.

Dessa maneira, para cada amostra de treinamento $\mathcal{T} := \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_m, y_m)\}$ com $m < n$ unidades amostrais, podemos definir a função de log-verossimilhança da seguinte maneira

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

Tomando o logaritmo da igualdade anterior, temos

$$\begin{aligned} \ell(\beta_0, \boldsymbol{\beta}|\mathcal{T}) &:= \log(L(\beta_0, \boldsymbol{\beta}|\mathcal{T})) \\ &= \sum_{i=1}^{m}\left[\log\left(1-\frac{e^{\beta_0+\boldsymbol{x}_i\boldsymbol{\beta}}}{1+e^{\beta_0+\boldsymbol{x}_i\boldsymbol{\beta}}}\right) + y_i(\beta_0+\boldsymbol{x}_i\boldsymbol{\beta})\right] \end{aligned}$$

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

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

Não existe uma expressão analítica para as estimativas do parâmetro e, portanto, é necessário utilizar métodos numéricos como o algoritmo de Newton-Raphson para encontrá-los.

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

$$\hat{\pi}(\boldsymbol{x}) := \frac{e^{\hat{\beta}_0+\boldsymbol{x}\hat{\boldsymbol{\beta}}}}{1+e^{\hat{\beta}_0+\boldsymbol{x}\hat{\boldsymbol{\beta}}}}, \tag{2.4}$$

em que $\boldsymbol{x}$ é o vetor de covariáveis observadas em uma nova unidade amostral. Nesse sentido, a classificação $y \in \{0, 1\}$ da nova unidade amostral 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, denominada ponte de corte da classificação.

A seleção adequada do ponto de corte é de extrema importância ao utilizar regressão logística para classificar observações, especialmente quando lidamos com bases de dados desbalanceadas. O ponto de corte padrão é 0.5, ou seja, todas as observações com probabilidade maior ou igual a 0.5 serão classificadas na classe minoritária e as demais na classe majoritária. Em uma base desbalanceada, é possível que o ponto de corte de 0.5 não seja o ideal. Isso ocorre porque o modelo pode ter uma alta taxa de falsos negativos, isto é, classificar erroneamente observações positivas como negativas, ou uma alta taxa de falsos positivos, isto é, classificar erroneamente observações negativas como positivas. Portanto, ao lidar com bases desbalanceadas, a seleção adequada do ponto de corte é crucial para equilibrar as taxas de falsos positivos e falsos negativos. Esse ponto de corte pode ser obtido de diversas maneiras, uma delas é tomar o ponto de corte como sendo a proporção de unidades amostrais da classe minoritária, outra maneira é escolher com base no ponto que maximiza alguma métrica de performance como, por exemplo, G-Média e MCC. Para escolher o ponto de corte dessa maneira, precisamos fazer um *grid* de valores possíveis para o ponto de corte e, então, escolher aquele que maximiza as métricas G-Média e MCC.

2.3 Máquinas de vetores suporte

Máquina de vetores suporte é um método de classificação proposto por Vapnik (1999). Ao contrário da regressão logística, esse método é não probabilístico, ou seja, em nenhum momento estimamos as probabilidades de uma unidade amostral pertencer a classe majoritária ou dela pertencer a classe minoritária. Com isso, essa técnica é baseada na separação linear das classes por hiperplanos, ou seja, obter um hiperplano que separe as observações de acordo com a classe ao qual cada observação pertence.

Um hiperplano é a generalização de plano para diferentes números de dimensões e para diferentes espaços vetoriais. Dado um vetor de parâmetros $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$, o hiperplano $H_{\boldsymbol{\beta}}$ em $\mathbb{R}^p$ é definido, matematicamente, como sendo o conjunto

$$H_{\boldsymbol{\beta}} := \{\boldsymbol{x} \in \mathbb{R}^p : \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p = 0\}.$$

Dizer que, para um dado vetor de parâmetros $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$, a equação

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p = 0 \tag{2.5}$$

define um hiperplano $H_{\boldsymbol{\beta}}$, significa que qualquer vetor $\boldsymbol{x} \in \mathbb{R}^p$ que satisfaz a Equação (2.5), pertence ao hiperplano $H_{\boldsymbol{\beta}}$. Por outro lado, se o vetor $\boldsymbol{x}$ não satisfizer a Equação (2.5), então $\boldsymbol{x}$ não pertence ao hiperplano $H_{\boldsymbol{\beta}}$, e, consequentemente, $\boldsymbol{x}$ poderá fazer com a expressão seja positiva ou negativa. Com isso, o hiperplano $H_{\boldsymbol{\beta}}$ divide o espaço $\mathbb{R}^p$ em dois conjuntos disjuntos:

$$H_{\boldsymbol{\beta}}^{+} := \{\boldsymbol{x} \in \mathbb{R}^p : \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p > 0\}$$

e

$$H_{\boldsymbol{\beta}}^{-} := \{\boldsymbol{x} \in \mathbb{R}^p : \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p < 0\}.$$

Dessa forma, dado um vetor $\boldsymbol{x} \in \mathbb{R}^p$, ou ele pertence ao hiperplano $H_{\boldsymbol{\beta}}$ ou a um dos dois conjuntos disjuntos $H_{\boldsymbol{\beta}}^{+}$ e $H_{\boldsymbol{\beta}}^{-}$.

Em um contexto de classificação, queremos utilizar o conceito de hiperplano para obter uma regra de classificação, ou seja, encontrar um hiperplano que consiga separar as observações de acordo com a classe ao qual cada observação pertence.

Nesse contexto, suponha que seja possível encontrar um hiperplano $H_{\boldsymbol{\beta}}$ que separe perfeitamente o conjunto de treinamento de acordo com a classe ao qual cada cliente pertence. Então, uma possível regra de classificação, baseada no hiperplano de separação $H_{\boldsymbol{\beta}}$, é obtida a partir da análise do sinal da função linear $f : \mathbb{R}^p \to \mathbb{R}$ tal que

$$f(\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p,$$

em que $\boldsymbol{x} \in \mathbb{R}^p$ é o vetor de covariáveis observado e $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$ é o vetor de parâmetros associado ao hiperplano $H_{\boldsymbol{\beta}}$ em questão. Nesse sentido, se $\boldsymbol{x}^* \in \mathbb{R}^p$ é o vetor de características observadas de uma nova unidade amostral, então quando $f(\boldsymbol{x}^*) > 0$ temos $\boldsymbol{x}^* \in H_{\boldsymbol{\beta}}^{+}$, classificando assim a nova unidade amostral na classe minoritária. Do mesmo modo, quando $f(\boldsymbol{x}^*) < 0$, temos $\boldsymbol{x}^* \in H_{\boldsymbol{\beta}}^{-}$, classificando assim a nova unidade amostral na classe majoritária.

De modo geral, se for possível encontrar um hiperplano que separe perfeitamente o conjunto de treinamento em dois lados, então existirá diversos hiperplanos que podem ser obtidos que resultem nessa divisão. Nesse contexto, a metodologia de máquinas de vetores suporte baseia-se no conceito de hiperplano de margens máximas, ou seja, busca

por aquele hiperplano que tem maior margem, isto é, aquele que fica "mais distante" de todos os pontos observados. Os pontos utilizados para definir as margens são chamados de vetores suporte. Dado um hiperplano $H_{\boldsymbol{\beta}}$, definimos a margem do hiperplano, indicada por $M(H_{\boldsymbol{\beta}})$, como sendo a menor dentre todas as distâncias Euclidianas d de um vetor de características observado ao hiperplano, isto é,

$$M(H_\beta) = \min_{1 \leq i \leq n} d\left(\boldsymbol{x}_i, \text{proj}_{H_\beta} \boldsymbol{x}_i\right),$$

em que proj_{H_β} é a projeção ortogonal da observação $\boldsymbol{x}_i$ sobre o hiperplano $H_{\boldsymbol{\beta}}$.

Para obtermos o classificador de margem máxima, basta resolvermos o seguinte problema de otimização:

$$\begin{aligned} &\underset{\boldsymbol{\beta} \in \mathbb{R}^{p+1}}{\text{maximizar}}\ M\left(H_{\boldsymbol{\beta}}\right) \\ &\text{Sujeito a } \sum_{j=1}^{p} \beta_j^2 = 1, \\ y_i(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}) &\geq M\left(H_{\boldsymbol{\beta}}\right) \quad \forall i = 1, \ldots, n, \end{aligned}$$

em que $\beta \in \mathbb{R}^{p+1}$ é o vetor de parâmetros e $M(H_{\boldsymbol{\beta}})$ é a margem definida pelo hiperplano $H_{\boldsymbol{\beta}}$.

Se $M(H_{\boldsymbol{\beta}})$ for maior do que zero, garantimos que para o hiperplano ótimo o classificador de margem máxima classifica todas as observações corretamente.

O classificador de margem máxima utiliza a suposição de que os dados são linearmente separáveis, no entanto em diversas situações isso não ocorre. Uma possível solução para a classificação de dados quando não for possível obter o hiperplano ótimo é utilizar um classificador mais flexível, no qual é permitido que algumas observações extrapolem o limite da margem e até mesmo do hiperplano. Nesse contexto, o classificador de vetores suporte é baseado em um hiperplano que não separa perfeitamente as duas classes. Portanto, buscamos o hiperplano que seja solução do seguinte problema de otimização:

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^{p+1},\boldsymbol{\epsilon}\in\mathbb{R}^{n+1}}{\text{maximizar}} \; M\left(H_{\boldsymbol{\beta}}\right) \tag{2.6}$$

$$\text{Sujeito a } \sum_{j=1}^{p}\beta_j^2 = 1, \tag{2.7}$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}) \geq M\left(H_{\boldsymbol{\beta}}\right)(1-\epsilon_i), \tag{2.8}$$

$$\epsilon_i \geq 0, \sum_{i=1}^{n}\epsilon_i \leq c, \tag{2.9}$$

em que c é um parâmetro real de ajuste não-negativo, $M(H_{\boldsymbol{\beta}})$ é a margem associada ao hiperplano $H_{\boldsymbol{\beta}}$ e ϵ_i são as variáveis de "folga" que permitem o modelo ser mais flexível, ou seja, que permitem observações serem classificadas do lado errado da margem ou até mesmo do hiperplano.

Note que ϵ_i pode ser maior do que um, de modo que é permitido que $y_i(\beta_0+\beta_1 x_{i1}+...+\beta_p x_{ip})$ seja negativo, ou seja, que a i-ésima amostra fique do lado errado do hiperplano. Se $\epsilon_i = 1$, a i-ésima observação está classificada do lado correto da margem; se $0 < \epsilon_i < 1$, a i-ésima observação violou o limite da margem; e se $\epsilon_i \geq 1$, a i-ésima observação está localizada do lado errado do hiperplano. O quão longe ela fica, contudo, é limitado por c, uma vez que ϵ_i não pode ser maior do que c. Assim, c é hiperparâmetro: quanto maior é seu valor, mais se permite que observações estejam do lado "errado" das margens.

Na grande maioria dos conjuntos de dados utilizados em estudos reais, não temos linearidade e, por isso, não conseguimos encontrar um hiperplano, mesmo aplicando a flexibilização do classificador de vetores suporte. Nessa situação, aplicamos uma transformação não-linear arbitrária Φ no vetor de covariáveis das unidades amostrais pertencentes ao conjunto de treinamento, buscando a projeção do espaço original p-dimensional em um espaço de maior dimensionalidade, denominado de espaço de características.

Isso nos leva a um novo conjunto de treinamento $Z^{'} = \{(\Phi(\boldsymbol{x}_1), y_1), \ldots, (\Phi(\boldsymbol{x}_n), y_n)\}$. A obtenção do classificador de vetores suporte a partir desse novo conjunto de treinamento Z' se resume ao mesmo problema de otimização abordado no classificador de vetores suporte, no qual devemos apenas substituir o conjunto das covariáveis.

Podemos reescrever o classificador $f : \mathbb{R}^p \to \mathbb{R}$ da seguinte maneira:

$$f(\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + ... + \beta_d x_d = \beta_0 + \sum_{i=1}^{n}\alpha_i\langle \boldsymbol{x}, \boldsymbol{x}_i\rangle,$$

em que $\langle \boldsymbol{x}, \boldsymbol{x}_i \rangle$ é o produto interno de dois vetores definido como $\langle \boldsymbol{a}, \boldsymbol{b} \rangle = \sum_{k=1}^{p} a_k b_k$ e α_i é um parâmetro associado a i-ésima observação do conjunto de treinamento. Assim, para calcular f isto é, os coeficientes α_i, tudo o que precisamos é do produto interno entre todas as observações. A proposta dos classificadores de máquinas de vetores suporte é a substituição do produto interno anterior por um kernel genérico $k(\boldsymbol{x}, \boldsymbol{x}_i)$, resultando no classificador $f : \mathbb{R}^p \to \mathbb{R}$ tal que

$$f(\boldsymbol{x}) = \beta_0 + \sum_{i \in S} \alpha_i k(\boldsymbol{x}, \boldsymbol{x}_i).$$

Note que, utilizando kernels não é necessário calcular o produto interno, o que torna o processo de obtenção do classificador de máquinas de vetores suporte mais eficiente e menos custoso.

O problema de otimização definido no caso de vetores suporte, pode ser reescrito da seguinte forma:

$$\underset{\boldsymbol{\beta} \in \mathbb{R}^p, \boldsymbol{\beta_0}}{\text{minimizar}} \left(\frac{1}{2} \|\boldsymbol{\beta}\|^2 + C \sum_{i=1}^{n} \epsilon_i \right) \tag{2.10}$$

$$\text{Sujeito a } \epsilon_i \geq 0,\ y_i(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}) \geq (1 - \epsilon_i), \quad i = 1, \ldots, n, \tag{2.11}$$

em que $||\boldsymbol{\beta}||^2 = \sum_{j=1}^{p} \beta_j^2$, ϵ_i são variáveis de folga e C pode ser definido como um parâmetro de custo, que tem a finalidade de ser como um custo de má classificação.

Com a restrição $\sum_{j=1}^{p} \beta_j^2 = 1$ ou $||\boldsymbol{\beta}||^2 = 1$, temos um problema de otimização convexa, que pode ser resolvido pelo método clássico de programação quadrática: os multiplicadores de Lagrange.

Seguindo Ng (2000) a função primal de Lagrange, a qual queremos minimizar, é dada por:

$$\mathcal{L}_{(\beta, \beta_0, \epsilon, \alpha, r)} = \frac{1}{2}||\boldsymbol{\beta}||^2 + C \sum_{i=1}^{n} \epsilon_i - \sum_{i=1}^{n} \alpha \left[y_i(\boldsymbol{x}_i \boldsymbol{\beta} + \beta_0) - (1 - \epsilon_i) \right] - \sum_{i=1}^{n} r_i \epsilon_i, \tag{2.12}$$

em que α_i e r_i são os multiplicadores de Lagrange.

2.4 Medidas de performance

As bases de dados que serão utilizadas neste trabalho apresentam classes desbalanceadas. Nesse cenário, os classificadores tendem a ter uma boa performance na classificação de novas unidades amostrais pertencentes a classe majoritária e uma má performance para aquelas unidades pertencentes à classe minoritária. Nos capítulos seguintes, apresentamos alguns métodos comumente utilizados na literatura para superar essa dificuldade. A fim de comparar qual desses métodos levam a uma melhor performance dos classificadores e qual dos classificadores possui o melhor desempenho, vamos utilizar medidas de performance baseadas na matriz de confusão, são elas: acurácia, sensibilidade, especificidade, valor preditivo positivo, valor preditivo negativo, G-média, coeficiente de Mathews e F1-Score.

A matriz de confusão é uma tabela que resume os resultados de um método de classificação. Na tabela 2.1, apresentamos de modo genérico quais são as quantidades levadas em consideração no resumo da performance dos classificadores. Nessa Tabela, denotamos por

- m o número de clientes no conjunto de teste;
- VP o número de unidades amostrais da classe positiva que foram classificadas corretamente como sendo da classe positiva;
- FP o número de unidades amostrais da classe negativa que foram erroneamente classificadas como sendo da classe positiva;
- FN o número de unidades amostrais da classe positiva que foram classificadas erroneamente como sendo da classe negativa;
- VN é o número de unidades amostrais da classe negativa que foram classificadas corretamente como sendo da classe negativa;
- P é o número de unidades amostrais classificadas como sendo da classe positiva;
- N é o número de unidades amostrais classificadas como sendo da classe negativa;
- p é o número de unidades amostrais pertencentes à classe positiva;
- n é o número de unidades amostrais pertencentes à classe negativa.

Tabela 2.1: Matriz de confusão para classificação binária.

Valores previstos	Valores observados		
pelo classificador	Positivo	Negativo	Total
Positivo	VP	FP	P
Negativo	FN	VN	N
Total	p	n	m

2.4.1 Sensibilidade

A sensibilidade (S) mede a capacidade do modelo em identificar corretamente os casos positivos entre todos os casos verdadeiramente positivos. A fórmula da Sensibilidade é dada por:

$$S = \frac{VP}{p} = \frac{VP}{VP + FN}.$$

2.4.2 Especificidade

A Especificidade (E) mede a proporção de casos negativos que foram corretamente identificados pelo modelo em relação ao total de casos negativos verdadeiros. A fórmula da Especificidade é dada por:

$$E = \frac{VN}{n} = \frac{VN}{VN + FP}.$$

2.4.3 Acurácia

A Acurácia (ACC) mede a proporção de previsões corretas feitas pelo modelo em relação ao número total de amostras. A fórmula da Acurácia é dada por:

$$ACC = \frac{VP + VN}{m}.$$

2.4.4 Valor Preditivo Positivo

O Valor Preditivo Positivo (VPP) ou Precisão mede a proporção de casos positivos previstos corretamente em relação ao total de casos classificados como positivos pelo modelo. A fórmula do Valor Preditivo Positivo é dada por:

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

2.4.5 Valor Preditivo Negativo

O Valor Preditivo Negativo (VPN) mede a proporção de casos negativos que foram corretamente identificados pelo modelo em relação ao total de casos classificados como negativos pelo modelo. A fórmula do Valor Preditivo Negativo é dada por:

$$VPN = \frac{VN}{N} = \frac{VN}{VN + FN}.$$

2.4.6 F1-Score

O F1-Score é calculado pela média harmônica entre a Precisão e a Sensibilidade, sua fórmula é dada por:

$$\text{F1-Score} = 2 \times \frac{\text{Precisão} \times \text{Sensibilidade}}{\text{Precisão} + \text{Sensibilidade}}.$$

O F1-Score busca um equilíbrio entre a capacidade do modelo de evitar falsos positivos e falsos negativos. Quanto maior o valor de F1-Score, melhor será a performance do classificador, por trazer um equilíbrio entre a precisão e a sensibilidade.

2.4.7 G-Média

A métrica G-Média, foi sugerida no estudo de Kubat *et al.* (1997) como uma medida de performance para avaliar a predição em conjunto de dados desbalanceados. Basicamente, calcula-se a média geométrica da sensibilidade e da especificidade, sua fórmula é dada por:

$$\text{G-Média} = \sqrt{\text{Sensibilidade} \times \text{Especificidade}}.$$

A G-Média é uma métrica que considera ambas as taxas de acertos (Sensibilidade e Especificidade), fornecendo uma medida de como o modelo performa em ambas as classes. Por conta disso, uma das principais vantagens da G-média é não ser afetada pelo desbalanceamento das classes. Quando a G-Média é alta, indica que o modelo possui uma boa capacidade de discriminação e desempenho geral em ambas as classes.

2.4.8 Coeficiente de Correlação Matthews

O coeficiente de correlação de Matthews (MCC), proposto por Matthews (1975), é uma medida de performance que calcula a correlação entre o valor observado na amostra de

teste e os valores preditos pelo classificador, a partir da utilização dos valores encontrados na matriz de confusão. A fórmula do MCC é dada por:

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

O Coeficiente de Matthews varia de -1 a 1, onde 1 indica uma previsão perfeita do modelo e -1 indica uma previsão totalmente incorreta ou uma inversão completa das classes. O valor intermediário 0, indica uma ausência de relação entre a previsão e o valor observado, de modo que a predição se assemelha ao acaso.

Uma vantagem do MCC é que ele não é afetado pelo desbalanceamento das classes, uma vez que leva em conta todos os aspectos da matriz de confusão e, por conta disso, avalia o desempenho do classificador em ambas as classes.

2.5 Considerações sobre a performance dos classificadores

Para ilustração da interpretação das medidas de performance dos classificadores, apresentamos a seguir um exemplo, simplificado, utilizando o conjunto de dados de inadimplência de crédito com classes levemente desbalanceadas.

Exemplo 2.13 *Suponha a situação em que desejamos classificar clientes em adimplentes e inadimplentes utilizando apenas as covariáveis X_1 e X_2, que significam, respectivamente, atraso médio no pagamento da fatura e fatura acima do limite. Para isso, o conjunto de dados foi dividido em conjunto de treinamento e conjunto de teste, nas proporções de 70% e 30%, respectivamente. Para realizar a classificação utilizamos o modelo de regressão logística e as máquinas de vetores suporte. Uma comparação das performances de classificação desses métodos na base de teste pode ser observada na Tabela 2.2.*

Tabela 2.2: Medidas de performance para os classificadores de Regressão Logística e Máquina de Vetores Suporte (SVM) quando utilizados com os dados desbalanceados.

	ACC	S	E	VPP	VPN	F1	G-Média	MCC
Regressão Logística	0.79	0.49	0.88	0.53	0.86	0.51	0.66	0.38
SVM	0.80	0.27	0.95	0.63	0.82	0.37	0.50	0.32

Em resumo, quando utiliza-se máquina de vetores suporte o modelo prioriza uma

sensibilidade alta e uma especificidade baixa. Já quando utiliza-se regressão logística, o modelo mantém uma sensibilidade alta, mas não tão elevada quanto o modelo de máquina de vetores suporte, e uma especificidade que é relevantemente maior que a especificidade do modelo de máquinas de vetores suporte.

Em um cenário financeiro, é interessante encontrar a maior sensibilidade possível, ou seja, encontrar a maior taxa de classificação correta para clientes inadimplentes, uma vez que esses clientes trazem prejuízos para instituições financeiras. Portanto, nos próximos capítulos, vamos estudar métodos para lidar com o desbalanceamento de classes e melhorar a classificação de unidades amostrais pertencentes à classe minoritária.

Nesse exemplo, o modelo de regressão linear foi superior às máquinas de vetores suporte, uma vez que ele apresentou, comparativamente, melhores valores das medidas de performance elencadas neste estudo.

Capítulo 3

Métodos de pré-processamento de dados

Em muitos problemas de classificação binária, as classes podem estar desbalanceadas, ou seja, uma classe pode ter muito mais exemplos do que outra, como é o caso das bases presentes em nosso estudo. Isso pode ocasionar um viés de classificação, onde o classificador pode favorecer a classe majoritária e não conseguir generalizar bem para a classe minoritária. Consequentemente, o classificador não será tão preciso para a classe minoritária, pois estará viesado para a classe majoritária. Os métodos de pré processamento dos dados ajudam a resolver esses problemas, melhorando o desempenho geral do modelo.

Para isso, o capítulo tem o objetivo de introduzir uma visão geral dos métodos de reamostragem, explicando como a reamostragem pode ser usada para criar conjuntos de dados balanceados, onde as classes minoritárias são aumentadas ou as classes majoritárias são reduzidas.

Além disso, o capítulo aborda o método de subamostragem Tomek Link, um dos métodos mais comumente usados para eliminar ruídos e reduzir a sobreposição de dados, detalhando como esse método identifica e remove pares de observações de classes opostas que estão próximas e podem causar confusão aos algoritmos de aprendizado de máquina. Ademais, abordaremos o método de sobreamostragem SMOTE, uma técnica eficaz para "clonar" novas observações da classe minoritária, aumentando sua representatividade no conjunto de dados, detalhando como o método gera amostras sintéticas, considerando as características das observações existentes, e como ele pode melhorar a precisão dos modelos de aprendizado de máquina. Por fim, discutiremos o método híbrido combinando Tomek Link com SMOTE, que combina as vantagens da subamostragem pelo método Tomek

Link e da sobreamostragem pelo método SMOTE.

3.1 Uma visão geral sobre métodos de reamostragem

Em muitos problema reais, principalmente na área financeira, encontramos conjuntos de dados cujas classes são desbalanceadas. Nesse trabalho, temos dois problemas reais envolvendo conjuntos de dados desbalanceados: o conjunto de crédito em que observamos uma maior proporção de clientes adimplentes (grupo majoritário) do que clientes inadimplentes (grupo minoritário); e o conjunto de fraude em que observamos mais transações legítimas (grupo majoritário) do que transações fraudulentas (grupo minoritário).

Em geral, os classificadores não performam bem em classes desbalanceadas, uma vez que são geralmente a favor da classe majoritária e, assim, apresentam baixo desempenho na classificação das unidades amostrais pertencentes à classe minoritária.

Nesse sentido, os métodos de reamostragem são de grande importância no contexto de classificação utilizando conjunto de dados desbalanceados, pois tem o objetivo de realizar o balanceamento do conjunto de treinamento antes da classificação de novas unidades amostrais, não deixando o desbalanceamento dos dados impactar o desempenho do classificador.

Dentre os métodos mais difundidos na aplicação de reamostragem, temos subamostragem, sobreamostragem e, por fim, métodos híbridos. Basicamente, na subamostragem reduzimos o conjunto de treinamento a partir da exclusão de observações da classe majoritária a fim de tornar esse conjunto mais balanceado. Já na sobreamostragem, aumentamos o conjunto de treinamento a partir da inclusão de réplicas de observações pertencentes à classe minoritária. Por fim, nos método híbridos estamos interessados em fazer a combinação dos dois métodos mencionados anteriormente.

3.2 Método de subamostragem Tomek Link

O método de subamostragem é simples de ser aplicado, contudo, é comum ser acompanhado de uma perda considerável de informação útil para a classificação, levando a uma baixa performance do classificador. Estes métodos podem ser adequados quando o conjunto de dados é levemente desbalanceado, mas não é recomendado quando observa-se um desbalanceamento severo das classes. Um exemplo de método de submamostragem é

o Tomek Link, cujo é algoritmo é

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.

É importante observar que esse algoritmo não necessariamente nos fornece classes igualmente balanceadas uma vez que sua função é eliminar da classe majoritária aquelas observações muito similares a observações pertencentes à classe minoritária. Dessa forma, o algoritmo Tomek Link se preocupa na melhora da discriminação das unidades amostrais em relação às classes consideradas e suaviza, de certa maneira, o problema de perda de informação inerente à retirada de muitas observações da base de dados.

3.3 Método de subamostragem One Sided Selection

O método One Sided Selection, conhecido como *OSS*, é uma técnica que combina o método Tomek Link (3.2) e a regra do método *Condensed Nearest Neighbor*, conhecido como *CNN*. Basicamente, os Tomek Links, aquelas observações que estão na zona confusa do conjunto de dados são removidos da classe majoritária e o método *CNN* é usado para remover exemplos similares que estão longe do limite de decisão. O algoritmo do *OSS* consiste em:

1. Seja S o conjunto de treinamento original;

2. Defina, inicialmente, C como o conjunto da classe minoritária de S;

3. Defina C^- como o conjunto da classe majoritária de S;

4. Retire um caso aleatoriamente de C^- e insira-o em C;

5. Para todo $x \in C^-$, determine o vizinho $y \in C$ mais próximo de x.

   (a) Se a classe de x é diferente da classe de y, então retire x de C^- e insira-o em C;

   (b) Senão, mantenha x em C^-;

6. Retire de C todos os exemplos da classe majoritária;

7. Defina S' como o novo conjunto de treinamento e faça

$$S' := C \cup C^-.$$

8. Retorne S'.

3.4 Método de sobreamostragem SMOTE

Em uma clássica técnica de sobreamostragem, a classe minoritária é replicada 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 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.

O método de sobreamostragem é simples de ser implementado e 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 treinamento, mas não generalize tão bem para observações do conjunto de teste. 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 eficiente é o SMOTE (do inglês *Synthetic Minority Oversampling Technique*), cujo algoritmo é

1. Defina N, onde N é um número real positivo, como sendo a proporção de sobreamostragem desejada, de forma que N multiplicado pelo número de observações da classe minoritária será a quantidade de observações sintéticas resultantes do algoritmo;

2. Defina o valor K, $K \in \mathbb{N}$, como sendo a quantidade de vizinhos mais próximos que será utilizada pelo algoritmo;

3. Inicia-se o processo iterativo para criação das unidades sintéticas, da seguinte forma:

   (a) Seleciona-se aleatoriamente uma observação $\boldsymbol{x}_i$ pertencente a classe minoritária;

   (b) Encontra-se os K vizinhos mais próximos de $\boldsymbol{x}_i$, pertencentes a classe minoritária, a partir de uma medida de distância, por exemplo a distância euclidiana;

   (c) Seleciona-se aleatoriamente, com repetição, N observações do conjunto dos K vizinhos mais próximos;

   (d) Em seguida calcula-se a diferença entre o vetor de características da observação selecionada com o vetor de cada um um dos N vizinhos selecionados;

   (e) Por fim, para cada uma das N observação no item (c) gera-se uma nova observação sintética somando-se ao vetor de característica da observação selecionada a diferença, obtida no passo anterior, multiplicada por um valor entre 0 e 1, gerado de uma distribuição uniforme. A definição de uma nova observação sintética $\boldsymbol{s}_{ij}$ é dada pela seguinte expressão:

$$\boldsymbol{s}_{ij} = \boldsymbol{x}_i + unif(0,1) \times (\boldsymbol{x}_{ij} - \boldsymbol{x}_i)^2 \qquad i = 1, \ldots, n \text{ e } j = 1, \ldots, N,$$

   em que x_{ij} representa o j-ésimo vizinho selecionado da i-ésima observação;

4. Repita o processo selecionando outras obervações do conjunto minoritário até que a quantidade de amostras geradas seja equivalente a multiplicação de N pelo número de linhas da classe minoritária.

3.5 Método combinando Tomek Link e One Sided Selection com SMOTE

O SMOTE + Tomek Link + One Sided Selection consiste basicamente na aplicação conjunta dos métodos SMOTE com o Tomek Link e One Sided Selection. Nesse sentido, o algoritmo funciona de forma que, após aplicarmos a subamostragem One Sided Selection, aplicamos a técnica Tomek Link para limpar, ainda mais, a zona de confusão e, em seguida,

aplicamos o método de sobreamostragem SMOTE afim de criar observações sintéticas com o objetivo de balancear as classes no conjunto de dados. Os métodos híbridos são os que, em geral, levam a melhores resultados. Contudo, essa abordagem exige pré-processamento mais dedicado para os dados em estudo e pode apresentar as desvantagens inerentes aos outros três métodos.

3.6 Considerações sobre o impacto dos métodos de pré-processamento dos dados

No conjunto de treinamento utilizado para o Exemplo 2.13, as classes estão distribuídas de maneira desbalanceada. Afim de contornar essa situação e, consequentemente, melhorar o desempenho do modelo de classificação, aplicaremos os métodos de pré-processamento dos dados discutidos anteriormente. Para o conjunto de treinamento, a proporção de observações em cada uma das classes de interesse é apresentada na Tabela 3.1.

Tabela 3.1: Distribuição das classes no conjunto de treinamento anterior à aplicação dos métodos de pré-processamento dos dados.

Adimplentes	Inadimplentes
16355	4645

Primeiramente, aplicando o método Tomek Link, ou seja, um método de subamostragem que baseia-se em eliminar observações pertencentes à classe majoritária que são muito similares à observações pertencentes à classe minoritária observamos, na Tabela 3.2, que o nível de redução da classe majoritária é pequeno. Isso acontece pois a função do algoritmo é justamente eliminar observações similares que podem confundir o modelo, e não necessariamente fornecer classes igualmente balanceada.

Tabela 3.2: Distribuição das classes no conjunto de treinamento posterior à aplicação do método de subamostragem Tomek Link.

Adimplentes	Inadimplentes
16313	4645

Aplicando agora o método SMOTE, ou seja, um método de sobreamostragem que baseia-se em criar novas observações diferentes das já existentes na classe minoritária, observamos que a distribuição das classe, agora, é dada de maneira balanceada.

Tabela 3.3: Distribuição das classes no conjunto de treinamento posterior à aplicação do método de sobreamostragem SMOTE.

Adimplentes	Inadimplentes
16355	16350

Por fim, aplicamos o método combinado Tomek Link com SMOTE, ou seja, um método híbrido que baseia-se em utilizar os dois métodos de pré-processamento para otimizar o balanceamento. Assim, balanceamos as classes e reduzimos o número de observações que podem confundir o classificador no momento de classificar determinada observação. A distribuição das classes após a aplicação do método híbrido é apresentado na Tabela 3.4.

Tabela 3.4: Distribuição das classes no conjunto de treinamento posterior à aplicação do método híbrido.

Adimplentes	Inadimplentes
16313	16303

A partir disso, podemos treinar os classificadores nos conjuntos de treinamento obtidos com os métodos de pré-processamento. Nesse contexto, podemos comparar os resultados a partir das medidas de performance discutidas no capítulo anterior e verificar seus efeitos na performance dos classificadores utilizados. Os resultados na base teste são apresentados na Tabela 3.5.

Tabela 3.5: Medidas de performance para os classificadores de Regressão Logística e Máquina de Vetores Suporte (SVM) quando utilizados com os dados desbalanceados e pré-processados pelos métodos Tomek Link, SMOTE e ambos combinados.

	ACC	S	E	VPP	VPN	F1	G-Média	MCC
Regressão Logística								
Tomek Link	0.79	0.50	0.87	0.53	0.86	0.51	0.66	0.38
SMOTE	0.79	0.49	0.88	0.53	0.86	0.51	0.65	0.38
Híbrido	0.79	0.49	0.87	0.53	0.86	0.51	0.65	0.38
Desbalanceado	0.79	0.49	0.88	0.53	0.86	0.51	0.66	0.38
SVM								
Tomek Link	0.80	0.27	0.95	0.63	0.82	0.37	0.50	0.32
SMOTE	0.79	0.49	0.87	0.53	0.86	0.51	0.65	0.38
Híbrido	0.79	0.49	0.87	0.53	0.86	0.51	0.65	0.38
Desbalanceado	0.80	0.27	0.95	0.63	0.82	0.37	0.50	0.32

A priori, é importante ressaltar que algumas medidas de performance, em conjunto de dados desbalanceados, são mais confiáveis que outras. Como foi dito na Seção 2.4, as medidas G-Média e Coeficiente de Correlação Matthews (MCC) são medidas que não são afetadas com o desbalanceamento das classes, ou seja, em nosso contexto, são medidas

mais confiáveis que sensibilidade e especificidade, uma vez que essas são medidas afetadas pelo desbalanceamento das bases. Além disso, os resultados são derivados de um exemplo reduzido, isto é, utilizando apenas 2 covariáveis.

De maneira geral, a performance dos classificadores que foram submetidos ao método de pré-processamento Tomek Link é semelhante às performances dos classificadores desbalanceados, assim como os classificadores que foram submetidos ao método de pré-processamento SMOTE são semelhantes, em performance, aos classificadores que foram submetidos ao método híbrido. Isso acontece uma vez que o método de pré processamento Tomek Link tem o objetivo de eliminar da classe majoritária observações muito similares a observações pertencentes à classe minoritária e como vimos na Tabela 3.2, não existem tantas observações para serem retiradas. Consequentemente, o classificador no qual utilizamos Tomek Link obtém um desempenho muito semelhante ao classificador que não aplicamos método de pré-processamento. Já para os métodos híbrido e SMOTE, essa semelhança ocorre por conta que o método híbrido é a aplicação dos dois métodos, primeiramente aplicando o Tomek Link para retirar as observações similares pertencentes a classe majoritária e, posteriormente, aplicando o método SMOTE para criar observações na classe minoritária parecidas com as observações já existentes nessa classe. Novamente, o método Tomek Link não encontrou muitas observações para serem retiradas da base, com isso, o balanceamento das classes ficou bastante semelhante, ocasionando valores iguais das medidas de performance para esses classificadores, nesses casos. Além de todos esses pontos, a simplicidade do exemplo em trabalhar com apenas 2 covariáveis é um fator que pode afetar a performance dos métodos de pré-processamento dos dados.

Quando comparamos o classificador via regressão logística com o classificador via SVM, ambos com a utilização do Tomek link, observamos que o via regressão logística possui um desempenho, para esse exemplo, superior ao via SVM, uma vez que a taxa de classificação correta da classe minoritária (sensibilidade) é significativamente maior no modelo de regressão logística, cerca de $0,23$ pontos percentuais, além das medidas G-Média e MCC, que são medidas mais confiáveis para casos desbalanceados, também serem maiores. Olhando para os classificadores que utilizaram os métodos SMOTE ou híbrido, vemos que tanto para o classificador via regressão logística como para o classificador via SVM, a performance do classificador é a mesma. Isso ocorre por conta da simplicidade do exemplo proposto, o qual contém apenas duas covariáveis que possuem uma forte correlação com a variável resposta, ou seja, são covariáveis que conseguem discriminar

satisfatoriamente as observações em suas verdadeiras classes, ver Figuras 3.1 e 3.2. Como o método SMOTE acaba criando novas observações para a classe minoritária em torno das observações já existentes e as covariáveis discriminam bem as observações em suas classes, os classificadores não apresentam diferentes comportamentos com esse pré-processamento dos dados.

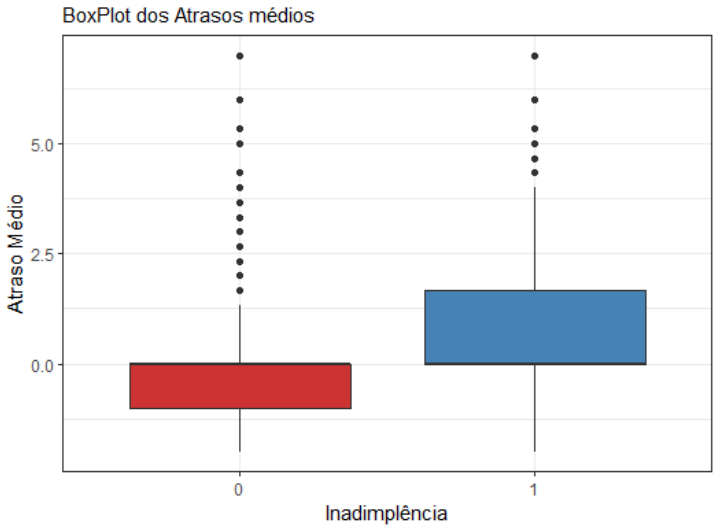

Figura 3.1: BoxPlot dos atrasos médios para o conjunto de dados.

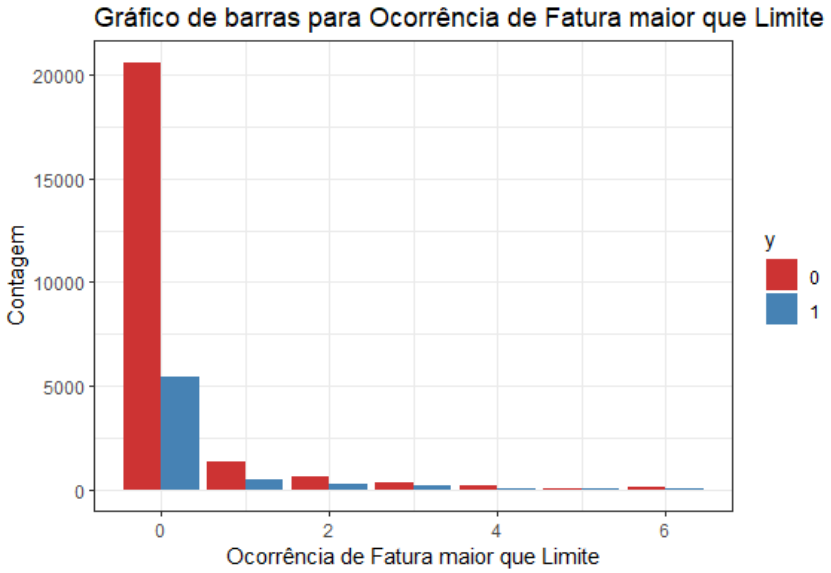

Figura 3.2: Gráfico de barras para ocorrência de faturas maior que o limite para o conjunto de dados.

Capítulo 4

Classificadores sensíveis ao custo

Em muitos problemas reais, além do desbalanceamento dos dados, as classes podem ter diferentes importâncias relativas, ou seja, errar na classificação de uma classe pode ser mais custoso do que errar na outra. Em nosso contexto, classificar clientes inadimplentes como adimplentes e transações fraudulentas como legítimas são erros que causam grandes prejuízos às instituições financeiras. Nesse contexto, os algoritmos sensíveis ao custo estão inseridos para utilizar diferentes custos de má classificação para as classes, de forma que a classificação incorreta na classe minoritária tenha maior penalização do que na classe majoritária.

4.1 Aprendizado sensível ao custo

Aprendizado sensível ao custo é uma abordagem onde executamos modificações nos classificadores para lidar com o desequilíbrio de classes. Ao invés de usar a avaliação padrão baseada em erro, introduz-se um custo para a classificação errônea, visando minimizar o risco condicional. Ao penalizar significativamente erros em algumas classes, podemos aumentar a importância delas durante o treinamento do classificador. Os custos associados às classificações errôneas podem originar-se de diversos aspectos relacionados a um problema da vida real e podem ser fornecidos por um especialista do assunto ou aprendidos durante a fase de treinamento do classificador. Na literatura, existem duas visões distintas sobre classificadores sensíveis ao custo, o custo associado às variáveis e o custo associado às classes.

O custo associado às variáveis é o cenário no qual adquirir observações dessa variável está conectado a um custo específico, ou seja, há uma dificuldade em adquirir essa ob-

servação, seja essa dificuldade monetária, temporal, entre outras. Nesse aspecto da aprendizagem sensível ao custo, o objetivo é criar um classificador que obtenha o melhor desempenho preditivo possível, utilizando características que possam ser obtidas ao menor custo possível, buscando equilibrar o desempenho do classificador e o custo das variáveis utilizadas. O custo associado às classes é o cenário em que cometer erros na classificação de determinada classe está conectada a um custo elevado, por exemplo, classificar um cliente inadimplente como adimplente traz um prejuízo grande para uma instituição financeira, consequentemente esse erro acaba sendo mais custoso quando comparado à classificar um cliente adimplente como inadimplente. Nesse aspecto da aprendizagem sensível ao custo, o objetivo é treinar um classificador de forma a focar nas classes que possuem custos mais elevados associado à elas, tratando-as de forma diferenciada durante o procedimento de treinamento.

Os custos são definidos em forma de matriz de custos, o método mais utilizado é chamado de função de perda $0-1$, que atribui valor 0 a uma instância classificada corretamente e valor 1 a uma classificada incorretamente. Como a função de perda $0-1$ usa o mesmo custo associado a uma classificação incorreta para todas as classes consideradas, ela é altamente suscetível a distribuições de classes desbalanceadas. A aprendizagem sensível ao custo tem justamente o objetivo de amenizar esse problema, adaptando uma função de perda diferente, com custos distintos associados a cada classe. Ao penalizar mais significativamente os erros de uma determinada classe, forçamos o procedimento de treinamento do classificador a focar nas observações provenientes dessa classe. O custo esperado (risco condicional) de classificar uma observação x como pertencente à classe i pode ser expresso como:

$$R(i|x) = \sum_{j=1}^{M} P(j|x) \cdot C(i,j),$$

em que $C(i,j)$ é o custo de classificar a observação x como da classe i dado que ela pertence a classe j e $P(j|x)$ é a estimativa de probabilidade de classificar a observação x como pertencente a classe j de um conjunto de M classes.

Para um problema de duas classes, um classificador sensível ao custo classificará a observação dada x como pertencente à classe positiva se e somente se:

$$P(0|x) \cdot (C(1,0) - C(0,0)) \leq P(1|x) \cdot (C(0,1) - C(1,1)).$$

Sob a suposição de que $C(0,0) = C(1,1) = 0$, um classificador sensível ao custo classificará a observação x como pertencente à classe positiva se e somente se:

$$P(0|x) \cdot C(1,0) \leq P(1|x) \cdot C(0,1).$$

Seguindo o fato de que $P(0|x) = 1 - P(1|x)$, podemos obter um limiar p^* para classificar uma observação x como pertencente à classe positiva se $P(1|x) \geq p^*$, onde $p^* = \frac{C(1,0)}{(C(1,0)+C(0,1))}$.

Algoritmos de aprendizagem sensível ao custo podem ser divididos em dois grupos principais: (i) algoritmos com abordagens diretas, que consistem em introduzir o custo de classificação incorreta no procedimento de treinamento do classificador e (ii) algoritmos de meta-aprendizado, que consistem em modificar os dados de treinamento ou as saídas do classificador, não modificando o algoritmo. Nesse caso, o custo de classificação incorreta pode ser incluído no pré-processamento, com o objetivo de modificar o conjunto de treinamento e no pós-processamento, com o objetivo de modificar as saídas do classificador.

A eficácia dos algoritmos sensíveis ao custo dependem dos custos fornecidos, custos muito baixos não permitirão ajustar corretamente os limites de classificação, enquanto custos muito altos levarão à perda de capacidade de generalização na classe restante. Em caso de dados desbalanceados, custos mal inseridos podem causar um viés a favor da classe majoritária. Existem duas maneiras de obter a matriz de custo: (i) por meio de um especialista, ou seja, alguém que tem conhecimento elevado no problema e pode estimar custos mais realistas às classificações incorretas e (ii) por meio de estimativas utilizando o conjunto de treinamento. A abordagem heurística mais popular para estimar o custo em problemas de dados desbalanceados é utilizar o nível de desbalanceamento como uma forma direta de estimar o custo.

4.1.1 Regressão logística sensível ao custo

A essência dos modelos de classificação de crédito e de detecção de fraudes é a mineração de dados e o aprendizado estatístico de algoritmos a partir de informações históricas das unidades amostrais para uma previsão mais precisa das probabilidades de inadimplência do solicitante do empréstimo e das probabilidades de fraude em transações financeiras. A regressão logística é muito utilizada na área financeira para tais objetivos.

No entanto, esse classificador faz uso de uma função de perda logarítmica. A função de perda logarítmica pode efetivamente medir o desvio entre o valor previsto e o valor real; ou seja, quanto maior o desvio, maior o valor da função perda. No entanto, a função de perda logarítmica não reflete o pensamento sensível ao custo. Para isso, devemos atribuir diferentes custos para o erro de classificação de novas unidades amostrais para as diferentes classes. Nesse sentido, a função de perda sensível ao custo para a regressão logística é

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

em que C^{+} é o custo de classificação incorreta para classe minoritária e C^{-} o custo de classificação incorreta para majoritária.

4.1.2 Máquina de vetores suporte sensível ao custo

Uma das causas para o baixo desempenho do classificador de máquinas de vetores suporte em um cenário de desequilíbrio de classe, é que na busca de diminuir o termo de penalidade o hiperplano é enviesado para classe minoritária. Umas das suposições para ocorrência de tal efeito é que o custo é fixo para ambas as classes. Nesse sentido, Veropoulos *et al.* (1999) propôs em seu estudo utilizar diferentes custos de má classificação para as classes, de forma que a classificação incorreta na classe minoritária tenha maior penalização do que na classe majoritária. Basicamente, para tal realização a formulação primal lagrangiana, dada pela Equação (2.12), sofre a seguinte modificação:

$$\begin{aligned}\mathcal{L}_{(\boldsymbol{\beta},\beta_0,\epsilon,\alpha,r)} = \frac{1}{2}\|\boldsymbol{\beta}\|^2 + C^{+}\sum_{\{i|y_i=+1\}}^{n}\epsilon_i + C^{-}\sum_{\{i|y_i=-1\}}^{n}\epsilon_i - \\ \sum_{i-1}^{n}\alpha_i\left[y_i\left(\boldsymbol{x_i^t}\boldsymbol{\beta}+\beta_0\right)-(1-\epsilon_i)\right] - \sum_{i=1}^{n} r_i\epsilon_i,\end{aligned}$$

em que $\epsilon_1, \ldots \epsilon_m$ são variáveis de "folga", α_i e r_i são os multiplicadores de Lagrange, e $\|\cdot\|$ denota a norma ℓ_2. Note que denotamos por C^{+} o custo de classificação incorreta para classe minoritária e C^{-} o custo de classificação incorreta para classe majoritária.

É direto mostrar que a forma dual lagrangiana segue a mesma que a exposta no caso do classificador original, porém com as seguintes restrições para α_i: se $y_i = +1$, então

$0 \leq \alpha_i \leq C^+$ e se $y_i = -1$, temos $0 \leq \alpha_i \leq C^-$.

Capítulo 5

Aplicações em dados financeiros

Nesta etapa, realizamos um estudo comparativo da performance das máquinas de vetores suporte com a performance da regressão logística na classificação de novas unidades amostrais a partir de três conjuntos de dados financeiros: dois relativos a análise de crédito e outro relativo a detecção de fraude. O primeiro conjunto de dados é sobre a inadimplência de clientes de cartão de crédito e possui um desbalanceamento leve, sendo 77% das unidades amostrais compostas por clientes adimplentes. O segundo conjunto de dados é sobre inadimplência de crédito e possui um desbalanceamento moderado, sendo 93% das unidades amostrais compostas por clientes adimplentes. O terceiro conjunto de dados é sobre fraudes em transações de crédito e possui um desbalanceamento severo, sendo 99% das transações legítimas. Além de analisar o desempenho dos classificadores em resposta a diferentes graus de desbalanceamento, também avaliamos a eficácia deles ao considerar técnicas para lidar com o desbalanceamento dos conjuntos de dados. Nesse sentido, mitigamos o efeito do desbalanceamento tanto a partir de técnicas de pré-processamento de dados quanto a partir da versão sensível ao custo dos classificadores supramencionados. Os métodos de pré-processamento considerados neste trabalho são o algoritmo Tomek Link + One Sided Selection (métodos de subamostragem), o algoritmo SMOTE (método de sobreamostragem) e o combinado desses algoritmos (método híbrido). Para medir o desempenho dos classificadores empregados, utilizamos medidas obtidas a partir da matriz de confusão tais como o coeficiente de correlação de Matthews (MCC), a G-Média, a sensibilidade, a especificidade, o valor preditivo positivo (VPP) e o valor preditivo negativo (VPN). Inicialmente, vamos apresentar a análise de forma individual para cada conjunto de dados, e, no final do capítulo, apresentamos uma discussão destacando os principais resultados obtidos.

5.1 Conjunto de dados de inadimplência de crédito com classes levemente desbalanceadas

O primeiro conjunto de dados considerado neste trabalho, compreende em informações individuais e de pagamento da fatura de cartão de crédito de clientes de um grande banco de Taiwan, durante o ano de 2005. Os dados foram extraídos de um estudo conduzido por Yeh e Lien (2009) e podem ser encontrados no repositório de dados de aprendizado de máquina da Universidade da Califórnia, Irvine (UCI), acessível em: http://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients.

Esse conjunto contém registros de 30.000 clientes, dos quais 6.636 (22, 12%) são clientes inadimplentes e 23.364 (77, 88%) são clientes adimplentes. A variável resposta é binária, indicando se um cliente é inadimplente (1) ou adimplente (0). Além disso, tínhamos à disposição 23 covariáveis, das quais foram consideradas neste estudo apenas algumas ou combinações delas, escolhidas e criadas com o intuito de otimizar a utilização das variáveis e potencializar os resultados a serem obtidos com a aplicação das metodologias propostas. Nesse sentido, as variáveis preditoras consideradas neste estudo são:

- Atraso Médio: Quantidade média de atrasos dos pagamentos considerando os 3 meses mais recentes;
- Quantidade de atrasos: Número de vezes que o cliente atrasou o pagamento, considerando os seis meses de referência do estudo;
- Proporção de pagamento: Proporção do valor pago em relação ao valor da fatura, considerando os meses de agosto, julho, junho e maio;
- Pagamento maior que fatura: Quantidade de vezes em que o pagamento foi maior do que o valor da fatura do cartão de crédito.

Destacamos que as variáveis referentes a valores monetários são expressos em dólares taiwanês.

5.1.1 Análise descritiva e exploratória dos dados

Antes da aplicação dos métodos propostos, realizamos a padronização das variáveis e uma análise descritiva e exploratória dos dados, a fim de identificar padrões de comportamento dos clientes adimplentes e inadimplentes em relação às variáveis utilizadas no

modelo. Durante a análise descritiva das variáveis, notamos a presença de pontos discrepantes, isto é, observações atípicas que podem influenciar no resultado das análises. Logo, para lidar com essas observações, aplicamos um tratamento que consiste em encontrar o valor que representa o quantil $98,5\%$ de cada variável e, em seguida, substituir o valor das observações que ultrapassem tal quantil, pelo valor do quantil encontrado. O mesmo tratamento foi aplicado para o limite inferior, que nesse caso, substituímos os valores das observações cuja característica observada fosse inferior ao quantil $1,5$.

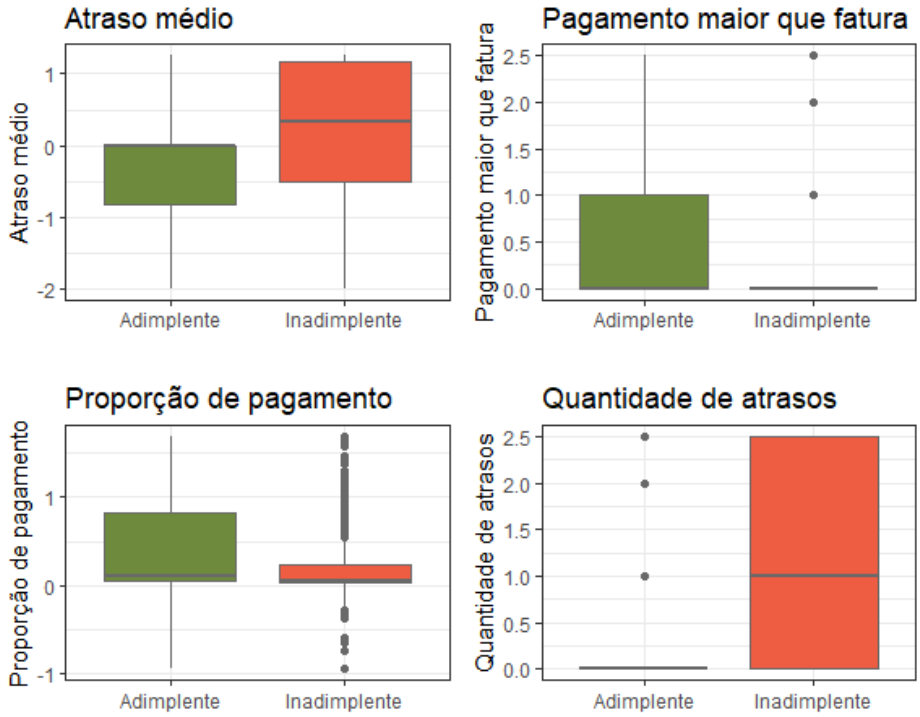

Figura 5.1: Os boxplots à esquerda representam a distribuição dos clientes adimplentes (caixa verde) e inadimplentes à direita (caixa laranja) para as variáveis padronizadas atraso médio, pagamento maior que fatura, proporção de pagamento e quantidade de atrasos, respectivamente.

Com a análise dos *boxplots* presentes na Figura 5.1, temos um indicativo de que as variáveis, por si só, são suficientes para discriminar satisfatoriamente os clientes em adimplentes e inadimplentes. Notamos que as variáveis pagamento maior do que a fatura e quantidade de atrasos conseguem distinguir muito bem os clientes em suas classes. Em particular, para a segunda variável, relacionada ao pagamento maior do que fatura, vemos que praticamente todos os clientes adimplentes não possuem pagamento maior do que a fatura e os que possuem são considerados pontos discrepantes, enquanto para a quarta variável, relacionada a quantidade de atrasos, notamos que praticamente todos os clientes adimplentes não possuem atrasos em seus pagamentos e aqueles que possuem são consi-

derados *outliers*. Por outro lado, as variáveis proporção de pagamentos e atraso médio também discriminam suficientemente bem os clientes adimplentes dos clientes inadimplentes. Em particular, para a terceira variável, relacionada a proporção de pagamentos, notamos os clientes adimplentes possuem proporções de valores pagos em relação ao valor da fatura, considerando os meses de agosto, julho, junho e maio, maiores do que as proporções dos clientes inadimplentes. Por fim, para a primeira variável, relacionada a atraso médio, vemos que não é uma variável que discrimina tão bem comparada as demais, mas consegue distinguir, de maneira satisfatória, os clientes em suas classes, tendo os clientes adimplentes médias de atrasos do pagamento, considerando os 3 meses mais recentes, menores do que os clientes inadimplentes. De maneira geral, temos um indicativo de que nesse conjunto de dados, as covariáveis elencadas para este trabalho conseguem discriminar de maneira satisfatória satisfatoriamente os clientes em adimplentes e inadimplentes. Esse indicativo sugere que o os classificadores podem possuir uma performance satisfatória mesmo o conjunto de dados sendo desbalanceado.

5.1.2 Resultados

Nesta etapa, realizamos a aplicação das metodologias discutidas anteriormente utilizando o conjunto de dados sobre inadimplência de clientes de cartão de crédito. Em outras palavras, vamos conduzir a versão completa da análise iniciada no Exemplo 2.13. Inicialmente, dividimos o conjunto de dados em dois subconjuntos: o conjunto de treinamento e o conjunto de teste. A composição desses subconjuntos foi determinada por meio de uma seleção aleatória sem reposição do conjunto de dados original, de modo que 70% das observações pertencentes ao conjunto de dados original fossem destinadas ao conjunto de treinamento e 30% ao conjunto de teste. Essa divisão foi realizada de modo a preservar as proporções de clientes adimplentes e inadimplentes presentes na base de dados original. Essa divisão ajuda a avaliar a capacidade do modelo de generalização, ou seja, aplicar o conhecimento adquirido durante o treinamento a instâncias não vistas anteriormente. Esta abordagem ajuda a prevenir a sobreajustagem, onde o classificador se adapta demais aos dados de treinamento, mas não generaliza bem para novas observações. Portanto, a divisão em conjuntos de treinamento e teste é fundamental para garantir a robustez e a eficácia dos classificadores.

Ressaltamos que os procedimentos de análises foram realizados utilizando os *softwares* *R* e *Python*. A regressão logítica é um método de classificação probabilístico, enquanto as

máquinas de vetores suportes são um método não-probabilístico. A avaliação da adequabilidade do modelo e a análise da significância das variáveis são aspectos importantes no desenvolvimento de modelos probabilísticos. Nesse sentido, observamos que antes e depois da 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 à adequabilidade do modelo. No entanto, não vamos detalhar essas informações, pois o foco do trabalho é comparar o poder preditivo dos classificadores propostos tanto em um cenário no qual o conjunto de dados está desbalanceado, quanto considerando considerando técnicas estatísticas para lidar com o desbalanceamento do conjunto de dados, seja por meio da versão sensível ao custo dos classificadores ou através de técnicas de reamostragem. Assim, permitimos uma leitura mais fluida e focada nas discussões que gostaríamos de tratar nesta monografia.

Neste trabalho, avaliamos a performance dos classificadores por meio de medidas obtidas a partir da matriz de confusão. Uma vez que a regressão logística estima a probabilidade de pertencimento a uma determinada classe com base em uma função logística, é necessário comparar a probabilidade estimada com um ponto de corte a fim de classificar uma nova instância. Nesse sentido, utilizamos como ponto de corte, o argumento que maximiza as métricas G-Média e MCC.

Tabela 5.1: Performance da regressão logística e das máquinas de vetores suporte na classificação das instâncias pertencentes ao conjunto de teste da base de dados de inadimplência de crédito com classes levemente desbalanceadas.

	ACC	S	E	VPP	VPN	F1	G-Média	MCC	Tempo
Logística									
Corte 1/2	0.8013	0.3460	0.9306	0.5863	0.8336	0.4352	0.5675	0.3409	0.07
Desbalanceado	0.7594	0.5997	0.8048	0.4660	0.8762	0.5245	0.6947	0.3720	0.07
Subamostragem	0.7653	0.5876	0.8158	0.4754	0.8744	0.5256	0.6923	0.3757	2.17
SMOTE	0.7568	0.6032	0.8004	0.4619	0.8766	0.5232	0.6948	0.3696	1.23
Híbrido	0.7619	0.5947	0.8094	0.4698	0.8755	0.5249	0.6937	0.3735	3.44
Sensível ao custo	0.7542	0.6102	0.7951	0.4583	0.8777	0.5234	0.6965	0.3691	0.02
SVM									
Desbalanceado	0.8034	0.3184	0.9412	0.6061	0.8293	0.4175	0.5474	0.3362	49.94
Subamostragem	0.8028	0.3300	0.9371	0.5984	0.8312	0.4254	0.5560	0.3386	42.04
SMOTE	0.7347	0.6554	0.7572	0.4340	0.8855	0.5222	0.7044	0.3630	161.69
Híbrido	0.7352	0.6549	0.7580	0.4347	0.8855	0.5225	0.7046	0.3636	122.22
Sensível ao custo	0.7333	0.6590	0.7545	0.4326	0.8862	0.5223	0.7050	0.3630	280

Na Tabela 5.1, apresentamos os resultados da performance da regressão logística e das máquinas de vetores suporte para todos os cenários propostos no trabalho. Nas colunas, estão representadas as medidas de performance utilizadas (ver Seção 2.4), em que ACC = Acurácia, S = Sensibilidade, E = Especificidade, VPP = Valor Preditivo Positivo, VPN

= Valor Preditivo Negativo, F1 = F1-Score, G-Média = G-Média, MCC = Coeficiente de Correlação de Matthews e Tempo = Tempo de processamento em segundos. Nas linhas, estão dispostos os cenários nos quais os classificadores foram ajustados. Os resultados estão dispostos em dois blocos, um trata-se da regressão logística e o outro trata-se das máquinas de vetores suporte.

Para aplicação do classificador de máquina de vetores suporte operamos a função *svm* do pacote *e1071*, em que utilizamos como parâmetros o *Kernel* Radial/Gaussiano com $\sigma^2 = 0,25$ (valor *default* da função *svm*, que é calculado como $\frac{1}{NV}$, em que NV é número de variáveis do conjunto de treinamento) e o parâmetro de custo igual a um ($C^- = 1$). Os mesmos parâmetros foram utilizados para aplicação do classificador de máquinas de vetores suporte sensível ao custo, entretanto nesse caso o parâmetro de custo para classe adimplente foi mantido como um, enquanto que o custo para classe inadimplente foi $3,2$ ($C^+ = 3,2$). Esse valor foi encontrado a partir de um *grid* de valores que foram testados, no qual escolhemos aquele que maximizou as medidas MCC e G-Média. Já para a regressão logística sensível ao custo, utilizamos o parâmetro de custo de forma análoga à máquina de vetores suporte sensível ao custo, isto é, o parâmetro de custo para a classe adimplente como sendo igual a 1 e para a classe inadimplente como sendo igual a $3,6$. Esse valor também foi encontrado a partir de um *grid* de valores que foram testados, no qual escolhemos aquele que maximizou as medidas MCC e G-Média. Para balanceamento do conjunto de dados aplicamos a técnica de SMOTE no conjunto de treinamento, definindo $N = 2$ e $K = 5$. Ao final do processo de criação das observações sintéticas, o novo conjunto de treinamento ficou com 32.705 observações, das quais 16.355 (50%) são da classe adimplente e 13.350 (50%) da classe inadimplente. Por outro lado, ao aplicarmos a técnica Tomek Link + One Sided Selection com $K = 1$, o conjunto subamostrado ficou com 18.559 observações, das quais 13.914 (75%) são da classe adimplente e 4.645 (25%) da classe inadimplentes. É importante ressaltar que essa técnica de subamostragem não balanceia o conjunto de dados, uma vez que o algoritmo não encontra mais nenhum vizinho próximo para ser retirado do conjunto de dados, ele é encerrado. Portanto, essa técnica de reamostragem apenas retira as observações que poderiam ser possíveis confundidores para o classificador, i.e., as observações que se encontram na fronteira das classes. Por fim, ao aplicarmos a técnica de reamostragem híbrida, o conjunto amostrado ficou com 27.802 observações, das quais 13.914 (50%) são da classe adimplente e 13.888 (50%) são da classe inadimplente.

Analisando os resultados referentes à regressão logística, observamos que exceto no cenário em que utilizamos como ponto de corte o valor padrão igual a $\frac{1}{2}$, os resultados são similares. Notamos um melhor desempenho do classificador com ponto de corte padrão em termos de acurácia ($\approx$ 80%), especificidade ($\approx$ 93%) e valor predito positivo ($\approx$ 58%) enquanto em termos das demais medidas de performance, o classificador possui uma melhor performance quando escolhemos um ponto de corte personalizado e alguma técnica para lidar com o desbalanceamento é aplicada. Nos cenários em que escolhemos o ponto de corte e aplicamos técnicas para lidar com o desbalanceamento dos dados temos, aproximadamente, a acurácia igual a 75%, a sensibilidade igual a 60%, a especificidade igual a 80%, o valor preditivo positivo igual a 47%, o valor preditivo negativo igual a 87%, a medida F1 igual a 52%, a G-média igual a 69% e o coeficiente de Mathews igual a 37%. Esses resultados sugerem que os métodos de reamostragem e a versão sensível ao custo da regressão logística não trazem uma melhora significativa à performance dos classificadores em termos de predição nesse cenário. Conclusão que vai de encontro aos indicativos encontrados durante a análise descritiva dos dados, no qual observamos que as variáveis preditoras por si só já conseguem discriminar satisfatoriamente os clientes em adimplentes e inadimplentes, e, portanto, o efeito dos métodos de reamostragem e da versão sensível ao custo da regressão logística sobre a performance dos classificadores parece ser mitigado por essa capacidade discriminatória inerente às covariáveis.

Por outro lado, analisando os resultados referentes às máquinas de vetores suporte, observamos que os resultados do cenário desbalanceado são semelhantes aos do cenário com dados subamostrados e que os resultados do cenário em que utilizamos a versão sensível ao custo das máquinas de vetores suporte são semelhantes aos cenários em que os dados foram sobreamostrados pela técnica SMOTE, seja a partir da sua aplicação pura ou híbrida. Em particular, nos dois primeiros cenários (desbalanceado e com subamostragem), observamos um melhor desempenho do classificador em termos de acurácia ($\approx$ 80%), especificidade ($\approx$ 94%) e valor preditivo positivo ($\approx$ 83%). Esses resultados revelam que, na ausência de abordagens eficientes para lidar com o desbalanceamento de classes, o classificador tende a favorecer a classe majoritária, resultando em um desempenho inferior na classificação da classe minoritária. De fato, o método de subamostragem adotado neste trabalho não realiza o balanceamento das classes, excluindo uma quantidade reduzida de unidades de treino, o que não tem um impacto significativo no nível de desbalanceamento. Nos três outros cenários (SMOTE, híbrido e sensível ao custo),

observamos um melhor desempenho do classificador em termos de sensibilidade ($\approx 65\%$), do valor preditivo negativo ($\approx 85\%$), da medida F1 ($\approx 52\%$), da G-média ($\approx 70\%$) e do coeficiente de Mathews ($\approx 36\%$). Com base nesses resultados, temos um indicativo de que a atribuição de pesos diferentes às classes durante o treinamento ou o uso da técnica SMOTE permite um melhor aprendizado da classe minoritária pelas máquinas de vetores suporte, levando a um aumento da sua performance em classificar clientes inadimplentes.

Vale destacar que, nos três últimos cenários (SMOTE, híbrido e sensível ao custo), embora os resultados relativos à performance das máquinas de vetores suporte em termos de acurácia, especificidade e valor preditivo positivo tenha sido inferior aos resultados obtidos nos dois primeiros cenários (desbalanceado e subamostragem), tais resultados são similares aos da performance da regressão logística. De fato, no que diz respeito à capacidade do classificador de generalização para a classe minoritária, as abordagens adotadas neste trabalho para lidar com o desbalanceamento das classes, garantiram uma consideração mais equitativa de ambas as classes durante o treinamento e, consequentemente, levaram as máquinas de vetores suporte a terem um desempenho preditivo análogo ao da regressão logística.

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, observamos uma performance similar de ambos os classificadores em todos os cenários. Embora com uma ligeira diferença, as máquinas de vetores suporte sensível ao custo possuem o melhor desempenho (G-média $\approx 71\%$ e $MCC \approx 36\%$) e as máquinas de vetores suporte sem balanceamento o pior desempenho (G-média $\approx 54\%$ e $MCC \approx 33\%$). Todavia, é importante ressaltar que a regressão logística sensível ao custo, além de possuir um desempenho similar às máquinas de vetores suporte sensível ao custo (G-média $\approx 70\%$ e $MCC = 37\%$), se destaca em termos de tempo de processamento (0,02).

Em suma, para esse conjunto de dados, nossa análise sugere que a aplicação de métodos para lidar com o desbalanceamento de classes pode resultar em um aprimoramento geral do desempenho tanto da regressão logística quanto das máquinas de vetores suporte na classificação de novas instâncias.

5.2 Conjunto de dados de inadimplência de crédito com desbalanceamento moderado das classes

O segundo banco de dados considerado neste trabalho foi obtido de uma competição publicada em https://www.kaggle.com/competitions/GiveMeSomeCredit/overview/description, cujo objetivo era a construção de um algoritmo capaz de prediz com alta acurácia a probabilidade de um cliente se tornar inadimplente. Dado o caráter competitivo deste projeto e a natureza dos dados fornecidos pelos organizadores, a disponibilidade limitada de informações sobre a origem dos dados foi percebida. No entanto, é implicitamente compreendido que esses dados dizem respeito a informações cadastrais e características de pagamento de clientes associados a alguma instituição financeira específica.

O conjunto de dados original já é dividido em treinamento e teste com 150.000 e 101.504 observações, respectivamente. Pelo fato do conjunto de teste omitir a informação referente a inadimplência, utilizamos apenas o conjunto de treinamento em nosso estudo, o qual, por sua vez, será dividido em novos conjuntos de treinamento e teste. Dos 150.000 clientes que compõem essa base de dados, 139.974 (93.32%) são adimplentes e 10.026 (6.68%) são inadimplentes. A variável resposta é binária, indicando se um cliente é inadimplente (1) ou adimplente (0). Temos à disposição 10 covariáveis, das quais foram consideradas neste estudo apenas algumas ou combinações delas, escolhidas e criadas com o intuito de otimizar a utilização das variáveis e potencializar os resultados a serem obtidos com a aplicação das metodologias propostas. Nesse sentido, as variáveis preditoras consideradas neste estudo são:

- Proporção limite utilizado: 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;
- Idade: Idade dos clientes;
- Atraso 30-59 dias: Número de vezes que o cliente ficou 30-59 dias atrasados, mas não pior nos últimos 2 anos;
- Proporção pagamento de dívidas: Pagamentos de dívidas mensais, pensão alimentícia, custos de vida, dividios pela renda bruta mensal;
- Quantidade empréstimos ativos: Número de empréstimos ativos (parcela como

empréstimo de carro ou hipoteca) e linhas de crédito (por exemplo, cartões de crédito);

- Quantidade empréstimos imobiliário: Número de empréstimos hipotecários e imobiliários, incluindo linhas de crédito para aquisição de habitação.

5.2.1 Análise descritiva e exploratória dos dados

Análogo à análise descritiva e exploratória dos dados conduzida para o primeiro banco de dados (ver Seção 5.1.1), realizamos a análise descritiva com o intuito de identificar padrões de comportamento dos clientes adimplentes e inadimplentes em relação às variáveis preditoras utilizadas e aplicamos o mesmo tratamento aos pontos discrepantes para evitar para mitigar uma possível influência de tais observações nos resultados de classificação, além de padronizar as variáveis.

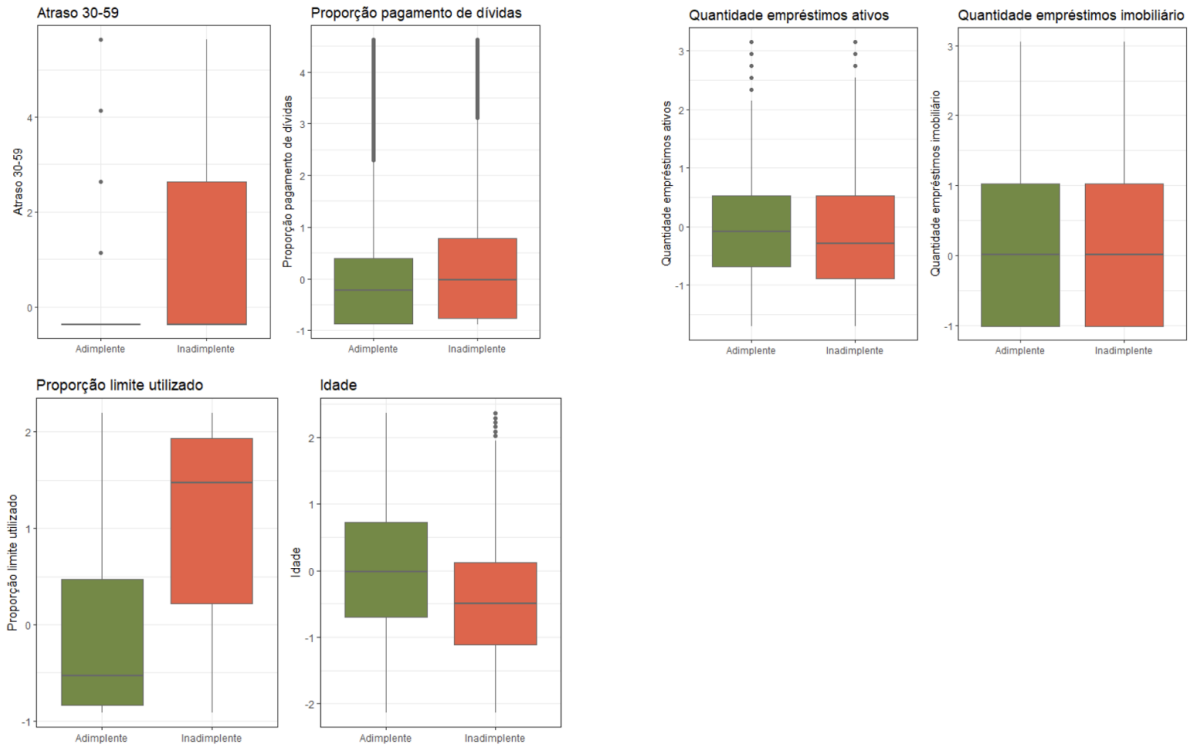

Figura 5.2: Os boxplots à esquerda representam a distribuição dos clientes adimplentes (caixa verde) e inadimplentes à direita (caixa laranja) para as variáveis padronizadas Atraso 30-59, Proporção pagamento de dívidas, Quantidade empréstimos ativos, Quantidade empréstimos imobiliário, Proporção limite utilizado e Idade.

Com a análise dos *boxplots* presentes na Figura 5.2, temos um indicativo de que algumas variáveis, por si só, conseguem discriminar os clientes em adimplentes e inadimplentes, uma vez que possuem padrões de comportamento diferentes nas duas classes, mas, por outro lado, também existem variáveis que podem encontrar dificuldades em classificar um novo cliente em sua verdadeira classe, pois possuem padrões de comportamento semelhantes nas duas classes. Nesse sentido, notamos que a variável *Atraso 30-59* é a que melhor distingue as observações em suas classes, pois praticamente, todos os clientes adimplentes possuem número de atrasos igual a zero. Por outro lado, as variáveis *Proporção pagamentos de dívidas*, *Quantidade empréstimos ativos* e *Quantidade empréstimos imobiliário* não conseguem discriminar tão bem os clientes adimplentes dos inadimplentes, uma vez que as observações das variáveis estão distribuídas de maneira similar nas duas classes. Com isso, é possível que essas variáveis encontrem dificuldades no momento da classificação de novos clientes. Por fim, as variáveis *Proporção limite utilizado* e *Idade* são variáveis que conseguem discriminar, de maneira satisfatória, os clientes em suas verdadeiras classes, mas que podem encontrar dificuldades na classificação de novos clientes, uma vez que observa-se uma discrepância moderada no padrão de comportamento dessas variáveis. Em particular, para a variável *Proporção limite utilizado*, em geral, os clientes adimplentes possuem saldo total em cartões de crédito e linhas de crédito pessoais menores do que os clientes adimplentes, enquanto para a variável *Idade*, em geral, os clientes adimplentes possuem idades superiores comparado aos clientes inadimplentes. De maneira geral, temos um indicativo de que algumas das covariáveis elencadas para este trabalho conseguem discriminar, de maneira satisfatória, os clientes em suas classes verdadeiras, porém é esperado que encontrem dificuldades classificando observações provenientes da classe minoritária, uma vez que o nível de desbalanceamento é moderado e existem variáveis que apresentam padrões semelhantes em seus boxplots.

5.2.2 Resultados

Inicialmente, dividimos o conjunto de dados em dois subconjuntos: o conjunto de treinamento e conjunto de teste. A composição desses subconjuntos foi determinada por meio de uma seleção aleatória sem reposição do conjunto de dados original, de modo que 70% das observações pertencentes ao conjunto de dados original fossem destinadas ao conjunto de treinamento e 30% ao conjunto de teste. Essa divisão foi realizada de modo a preservar as proporções de clientes adimplentes e inadimplente presentes na base

de dados original. Os procedimentos de análises foram realizados em *R* e *Python*. Além disso, seguindo de maneira análoga à análise dos resultados do primeiro banco de dados (ver Subseção 5.1.2), não vamos detalhar os resultados sobre o ajuste e adequabilidade dos modelos a fim de deixar a leitura mais fluida e focada nas discussões sobre o poder preditivo dos classificadores. As medidas de performance utilizadas foram as mesmas da seção anterior e a obtenção do ponto de corte foi por meio do ponto de maximiza as métricas G-Média e MCC.

Tabela 5.2: Performance da regressão logística e das máquinas de vetores suporte na classificação das instâncias pertencentes ao conjunto de teste da base de dados de inadimplência de crédito com desbalanceamento moderado das classes.

	ACC	S	E	VPP	VPN	F1	G-Média	MCC	Tempo
Logística									
Corte 1/2	0.9333	0.0987	0.9938	0.5403	0.9382	0.1669	0.3132	0.2105	1.31
Desbalanceado	0.8457	0.5908	0.8642	0.2402	0.9667	0.3415	0.7146	0.3068	1.31
Subamostragem	0.8548	0.5659	0.8757	0.2485	0.9652	0.3454	0.7040	0.3073	16.98
SMOTE	0.8450	0.5944	0.8632	0.2399	0.9670	0.3419	0.7163	0.3077	4.86
Híbrido	0.8782	0.4931	0.9061	0.2761	0.9609	0.3540	0.6684	0.3076	19.77
Sensível ao custo	0.8780	0.4917	0.9061	0.2757	0.9608	0.3533	0.6675	0.3068	1.2
SVM									
Desbalanceado	0.9328	0.0416	0.9974	0.5474	0.9347	0.0774	0.2038	0.1374	3683
Subamostragem	0.9331	0.0905	0.9943	0.5369	0.9377	0.1549	0.3000	0.2007	3388
SMOTE	0.7620	0.7503	0.7628	0.1868	0.9767	0.2992	0.7565	0.2897	7066
Híbrido	0.7598	0.7503	0.7605	0.1854	0.9767	0.2973	0.7554	0.2877	6092
Sensível ao custo	0.7782	0.7358	0.7812	0.1963	0.9760	0.3100	0.7582	0.2985	7258

Na Tabela 5.2, apresentamos os resultados da performance da regressão logística e das máquinas de vetores suporte para todos os cenários propostos no trabalho. Nas colunas, estão representadas as medidas de performance utilizadas (ver Seção 2.4), em que ACC = Acurácia, S = Sensibilidade, E = Especificidade, VPP = Valor Preditivo Positivo, VPN = Valor Preditivo Negativo, F1 = F1-Score, G-Média = G-Média, MCC = Coeficiente de Correlação de Matthews e Tempo = Tempo de processamento em segundos. Nas linhas, estão dispostos os cenários nos quais os classificadores foram ajustados. Os resultados estão dispostos em dois blocos, um trata-se dos modelos ajustados pela da regressão logística e o outro trata-se dos modelos ajustados pela das máquinas de vetores suporte.

Para aplicação do classificador de máquina de vetores suporte operamos a função *svm* do pacote *e1071*, em que utilizamos como parâmetros o *Kernel* Radial/Gaussiano com $\sigma^2 = 0,166$ (valor *default* da função *svm*, que é calculado como $\frac{1}{NV}$, em que NV é número de variáveis do conjunto de treinamento) e o parâmetro de custo igual a um ($C^- = 1$). Os mesmos parâmetros foram utilizados para aplicação do classificador de máquinas de vetores suporte sensível ao custo, entretanto nesse caso o parâmetro de custo para classe

adimplente foi mantido como 1, enquanto que o custo para classe inadimplente foi 14 ($C^+ = 14$). Esse valor foi encontrado a partir de um *grid* de valores que foram testados, no qual escolhemos aquele que maximizou as medidas MCC e G-Média. Já para a regressão logística sensível ao custo, utilizamos os parâmetros de custo de forma análoga à máquina de vetores suporte sensível ao custo, isto é, o parâmetro de custo para a classe adimplente como sendo igual a 1 e para a classe inadimplente como sendo igual a 14. Esse valor também foi encontrado a partir de um *grid* de valores que foram testados, no qual escolhemos aquele que maximizou as medidas MCC e G-Média. Para o balanceamento do conjunto de dados, aplicamos a técnica de SMOTE no conjunto de treinamento, definindo $N = 2$ e $K = 5$. Ao final do processo de criação das observações sintéticas, o novo conjunto de treinamento ficou com 195.704 observações, das quais 98.012 (50%) são da classe adimplente e 97.692 (50%) da classe inadimplente. Por outro lado, ao aplicarmos a técnica Tomek Link + One Sided Selection $K = 1$, o conjunto subamostrado ficou com 101.173 observações, das quais 94.195 (93%) são da classe adimplente e 6.978 (7%) são da classe inadimplente. Lembramos que essa técnica de subamostragem não balanceia o conjunto de dados, retirando apenas as observações que poderiam ser possíveis confundidores para o classificador. Por fim, ao aplicarmos a técnica de reamostragem híbrida o conjunto amostrado ficou com 188.398 observações, das quais 94.195 (50%) são da classe adimplente e 94.203 (50%) são da classe inadimplente.

Analisando os resultados referentes à regressão logística, observamos que exceto no cenário em que utilizamos como ponto de corte o valor padrão igual a $\frac{1}{2}$, os resultados são similares. Notamos um melhor desempenho do classificador com ponto de corte padrão em termos de acurácia ($\approx 93\%$), especificidade ($\approx 99\%$) e valor predito positivo ($\approx 54\%$) enquanto em termos das demais medidas de performance, o classificador possui uma melhor performance quando alguma técnica para lidar com o balanceamento é aplicada. Em particular, nesse último caso, temos aproximadamente a sensibilidade igual a 55%, valor preditivo negativo igual a 96%, G-média igual a 69%, F1 igual a 35% e o coeficiente de Mathews igual a 30%. Portanto, esses resultados sugerem que a utilização do ponto de corte padrão em conjuntos de dados desbalanceados interfere na acurácia do classificador na classificações de novas observações provenientes da classe minoritária.

Ainda no que diz respeito aos resultados da regressão logística com ponto de corte obtido a partir da G-média e do MCC, podemos separar os resultados em dois blocos: (i) modelos com cenário desbalanceado, com subamostragem e com sobreamostragem

SMOTE e (ii) modelos com cenário de amostragem híbrida e algoritmos sensível ao custo. Esses blocos foram definidos de acordo com a similaridade dos resultados. Observamos um melhor desempenho do bloco (i) em termos de sensibilidade ($\approx 57\%$), valor preditivo negativo ($\approx 96\%$) e G-média ($\approx 71\%$) enquanto observamos um melhor desempenho dos classificadores do bloco (ii) em termos de acurácia ($\approx 87\%$), especificidade ($\approx 90\%$), valor predito positivo ($\approx 27\%$), F1 ($\approx 35\%$) e do coeficiente de Mathews ($\approx 30\%$). Esses resultados sugerem que o método de reamostragem híbrida, assim como a versão sensível ao custo dos classificadores aumentaram, mesmo que de forma mínima, a performance dos classificadores em termos de predição.

Por outro lado, analisando os resultados referentes às máquinas de vetores suporte, observamos que os resultados do cenário desbalanceado são semelhantes aos do cenário com dados subamostrados e que os resultados do cenários em que utilizamos a versão sensível ao custo das máquinas de vetores suporte são semelhantes aos cenários em que os dados foram sobreamostrados pela técnica SMOTE, seja a partir da sua aplicação pura ou híbrida. Em particular, nos dois primeiros cenários (desbalanceado e com subamostragem), observamos um melhor desempenho do classificador em termos de acurácia ($\approx 93\%$), especificidade ($\approx 99\%$) e valor predito positivo ($\approx 54\%$). Esses resultados revelam que, na ausência de abordagens eficientes para lidar com o desbalanceamento de classes, o classificador tende a favorecer a classe majoritária, resultando em um desempenho inferior na classificação da classe minoritária. De fato, o método de subamostragem adotado neste trabalho não realiza o balanceamento das classes, excluindo uma quantidade reduzida de unidades de treino, o que não tem um impacto significativo no nível de desbalanceamento. Nos três outros cenários (SMOTE, híbrido e sensível ao custo), observamos um melhor desempenho do classificador em termos de sensibilidade ($\approx 75\%$), do valor preditivo negativo ($\approx 97\%$), da medida F1 ($\approx 30\%$), da G-média ($\approx 75\%$) e do coeficiente de Mathews ($\approx 29\%$). Com base nesses resultados, temos um indicativo de que a atribuição de pesos diferentes às classes durante o treinamento ou o uso da técnica SMOTE permite um melhor aprendizado da classe minoritária pelas máquinas de vetores suporte, levando a um aumento da sua performance em classificar clientes inadimplentes.

Vale destacar novamente que, nos três últimos cenários (SMOTE, híbrido e sensível ao custo), embora os resultados relativos à performance das máquinas de vetores suporte em termos de acurácia, especificidade e valor preditivo positivo tenha sido inferior aos resultados obtidos nos dois primeiros cenários (desbalanceado e subamostragem), tais resultados

são similares aos da performance da regressão logística. De fato, no que diz respeito à capacidade do classificador de generalização para a classe minoritária, as abordagens adotadas neste trabalho para lidar com o desbalanceamento das classes, garantiram uma consideração mais equitativa de ambas as classes durante o treinamento e, consequentemente, levaram as máquinas de vetores suporte a terem um desempenho preditivo análogo ao da regressão logística.

Analisando as métricas de performance que condensam os valores das métricas individuais e, por conseguinte, expressam o desempenho global do classificador, observamos uma performance similar de ambos os classificadores em todos os cenários. As máquinas de vetores suporte sensível ao custo possuem o melhor desempenho (G-média $\approx 75\%$ e $MCC \approx 29\%$) e as máquinas de vetores suporte sem balanceamento o pior desempenho (G-média $\approx 20\%$ e $MCC \approx 13\%$).

Em suma, para esse conjunto de dados, nossa análise sugere que a aplicação de métodos para lidar com o desbalanceamento de classes pode resultar em um aprimoramento geral do desempenho tanto da regressão logística quanto das máquinas de vetores suporte na classificação de novas instâncias.

5.3 Conjunto de dados de fraude em cartão de crédito com desbalanceamento severo das classes

O terceiro conjunto de dados considerado neste trabalho, compreende informações de transações de cartão de crédito em setembro de 2013 detentores europeus de cartão de crédito e que foi disponibilizado por uma instituição financeira, não sendo possível identificar os clientes, servindo apenas para fins de pesquisa e estudos e, por motivos de confidencialidade, as covariáveis serão tratadas com nomes fictícios. Esse conjunto de dados pode ser acessado em https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud.

Esse conjunto de dados contém 284.807 transações, das quais 284.315 ($99,82\%$) são transações legítimas e apenas 492 (0.18%) são transações fraudulentas, sendo considerado um conjunto de dados com desbalanceamento severo das classes. A variável resposta é binária, indicando se uma transação é fraudulenta (1) ou legítima (0). Além disso, temos à disposição as seguintes covariáveis:

- Time: Tempo relativo a primeira transação;
- Amount: Valor da transação;
- $V1, \cdots, V28$: Numérica.

Afim de otimizar a utilização das variáveis e potencializar os resultados a serem obtidos com a aplicação da metodologia proposta, vamos padronizar as variáveis *Time* e *Amount* da seguinte maneira:

- Para a variável *Time*:

$$time_{new} = \frac{time_{old} - \text{média}}{\text{variância}}$$

- Para a variável *Amount*:

$$amount_{new} = \frac{amount_{old} - \text{mediana}}{Q_3 - Q_1}$$

Além disso, foi necessário realizar o tratamento de observações discrepantes, para mitigar uma possível influência de tais observações nos resultados de classificação. Nesse caso, substituímos os valores dos pontos discrepantes pelo valor que representa o limite mais próximo dele, ou seja, se é uma observação acima do limite superior, substituímos por esse limite. Caso seja uma observação abaixo do limite inferior, substituímos pelo valor do próprio limite inferior.

Como as covariáveis foram tratadas com nomes fictícios e possuíram seus valores alterados por motivos de confidencialidade, a análise descritiva e exploratória dos dados não será feita, uma vez que perdemos o poder interpretativo dessa análise.

5.3.1 Resultados

Inicialmente dividimos o conjunto de dados em dois subconjuntos: o conjunto de treinamento e conjunto de teste. A composição desses subconjuntos foi determinada por meio de uma seleção aleatória sem reposição do conjunto de dados original, de modo que 70% das observações pertencentes ao conjunto de dados original fossem destinadas ao conjunto de treinamento e 30% ao conjunto de teste. Além disso, seguindo de maneira análoga à análise dos resultados dos demais bancos de dados (Seções 5.1.2 e 5.2.2), os procedimentos foram realizados em *R* e *Python*, as medidas de performance analisadas

foram as mesmas e a obtenção do ponto de corte foi por meio do ponto que maximizava as métricas G-Média e MCC.

Tabela 5.3: Performance da regressão logística e das máquinas de vetores suporte na classificação das instâncias pertencentes ao conjunto de teste da base de dados fraude em cartão de crédito com desbalanceamento severo das classes.

	ACC	S	E	VPP	VPN	F1	G-Média	MCC	Tempo
Logística									
1/2	0.9994	0.7805	0.9998	0.8731	0.9996	0.8297	0.8890	0.8305	5.85
Desbalanceado	0.9995	0.7837	0.9998	0.9062	0.9996	0.8405	0.8852	0.8425	5.85
Subamostragem	0.9995	0.7837	0.9998	0.9133	0.9996	0.8436	0.8852	0.8458	710
SMOTE	0.9988	0.8310	0.9990	0.6089	0.9997	0.7028	0.9112	0.7108	15.61
Híbrido	0.9988	0.8310	0.9990	0.6059	0.9997	0.7008	0.9112	0.7090	718
Sensível ao custo	0.9988	0.8243	0.9991	0.6321	0.9996	0.7155	0.9075	0.7213	2.86
SVM									
Desbalanceado	0.9995	0.7837	0.9998	0.8923	0.9996	0.8345	0.8852	0.8345	347
Subamostragem	0.9995	0.7837	0.9998	0.8923	0.9996	0.8345	0.8842	0.8345	943
SMOTE	0.9976	0.8040	0.9979	0.4075	0.9996	0.5409	0.8957	0.5409	1729
Híbrido	0.9976	0.8040	0.9979	0.4061	0.9996	0.5396	0.8957	0.5704	2181
Sensível ao custo	0.9985	0.8310	0.9987	0.5418	0.9997	0.6560	0.9110	0.6703	168

Na Tabela 5.3, apresentamos os resultados da performance da regressão logística e das máquinas de vetores suporte para todos os cenários propostos no trabalho. Nas colunas, estão representadas as medidas de performance utilizadas (ver Seção 2.4), em que ACC = Acurácia, S = Sensibilidade, E = Especificidade, VPP = Valor Preditivo Positivo, VPN = Valor Preditivo Negativo, F1 = F1-Score, G-Média = G-Média, MCC = Coeficiente de Correlação de Matthews e Tempo = Tempo de processamento em segundos. Nas linhas, estão dispostos os cenários nos quais os classificadores foram ajustados. Os resultados estão dispostos em dois blocos, um trata-se dos modelos ajustados pela da regressão logística e o outro trata-se dos modelos ajustados pela das máquinas de vetores suporte.

Para aplicação do classificador de máquina de vetores suporte operamos a função *svm* do pacote *e1071*, em que utilizamos como parâmetros o *Kernel* Radial/Gaussiano com $\sigma^2 = 0,033$ (valor *default* da função *svm*, que é calculado como $\frac{1}{NV}$, em que NV é número de variáveis do conjunto de treinamento) e o parâmetro de custo igual a um ($C^- = 1$). Os mesmos parâmetros foram utilizados para aplicação do classificador de máquinas de vetores suporte sensível ao custo, entretanto nesse caso o parâmetro de custo para classe adimplente foi mantido como um, enquanto que o custo para classe inadimplente foi 578.5 ($C^+ = 578.5$). Esse valor foi encontrado a partir de um *grid* de valores que foram testados, no qual escolhemos aquele que maximizou as medidas MCC e G-Média. Já para a regressão logística sensível ao custo, utilizamos o parâmetro de custo de forma análoga à máquina de vetores suporte sensível ao custo, isto é, o

parâmetro de custo para a classe adimplente como sendo um e para a classe inadimplente como sendo 578.5. Esse valor, também, foi encontrado a partir de um *grid* de valores que foram testados, no qual escolhemos aquele que maximizou as medidas MCC e G-Média. Para balanceamento do conjunto de dados aplicamos a técnica de SMOTE no conjunto de treinamento, definindo $N = 2$ e $K = 5$. Ao final do processo de criação das observações sintéticas, o novo conjunto de treinamento ficou com 1398.024 observações, das quais 199.020 (50%) são da classe adimplente e 199.004 (50%) da classe inadimplente. Por outro lado, ao aplicarmos a técnica Tomek Link + One Sided Selection $K = 1$, o conjunto subamostrado ficou com 197.094 observações, das quais 196.750 (99%) são da classe adimplente e 344 (1%) são da classe inadimplente. Lembramos que essa técnica de subamostragem não balanceia o conjunto de dados, retirando apenas as observações que poderiam ser possíveis confundidores para o classificador. Por fim, ao aplicarmos a técnica de reamostragem híbrida, o conjunto híbrido ficou com 393.483 observações, das quais 196.750 (50%) são da classe adimplente e 196.733 (50%) são da classe inadimplente.

A partir da análise dos resultados da Tabela 5.3, notamos que, tanto a regressão logística quanto as máquinas de vetores suporte tiveram um desempenho preditivo similar em todos os cenários de acordo com as medidas de performance adotadas neste trabalho. Para ambos os classificadores, notamos que a aplicação de técnicas para lidar com o desbalanceamento dos dados aumenta a acurácia de ambos os classificadores na classe positiva preservando a acurácia na classe negativa, i.e., observa-se um aumento da sensibilidade e pouca mudança na especificidade, o que leva, consequentemente, a um aumento na G-média. Em particular, na regressão logística, nos dois primeiros cenários observamos um melhor desempenho do classificador em termos de especificidade ($\approx 99\%$), valor predito positivo ($\approx 91\%$), F1 ($\approx 84\%$) e coeficiente de Mathews ($\approx 84\%$). Nos três outros cenários (SMOTE, híbrido e sensível ao custo), observamos um melhor desempenho do classificador em termos de acurácia ($\approx 99\%$), sensibilidade ($\approx 83\%$), valor predito negativo ($\approx 99\%$) e G-Média ($\approx 91\%$). Analogamente, nas máquinas de vetores suporte, observamos nos dois primeiros cenários um melhor desempenho do classificador em termos de especificidade ($\approx 99\%$), valor predito positivo ($\approx 89\%$), F1 ($\approx 83\%$) e coeficiente de Mathews ($\approx 83\%$). Nos três outros cenários (SMOTE, híbrido e sensível ao custo), observamos um melhor desempenho do classificador em termos de acurácia ($\approx 99\%$), sensibilidade ($\approx 80\%$) e G-Média ($\approx 90\%$).

Diferente do que ocorre nas duas primeiras bases de dados, onde o nível de desbalan-

ceamento não é tão severo, observamos que o MCC é maior nos casos em que não houve a geração de observações sintéticas ou que não houve penalização para classe minoritária. Esse comportamento é esperado, pois o MCC leva em conta verdadeiros positivos, verdadeiros negativos, falsos positivos e falsos negativos e, portanto, tanto a introdução de exemplos sintéticos pelo SMOTE quanto a definição de um alto custo para classe minoritária pode aumentar o número de falsos positivos ou falsos negativos, impactando negativamente o valor do MCC. Portanto, acreditamos que a G-Média é uma medida mais justa para avaliar o desempenho geral dos classificadores, uma vez que é mais robusta à geração exarcebada de observações sintéticas e à definição de custos altos para classe minoritária. Nesse sentido, observamos um melhor desempenho nos três últimos casos de cada classificador (SMOTE, híbrido e sensível ao custo).

Em suma, para esse conjunto de dados, nossa análise sugere que a aplicação de métodos para lidar com o desbalanceamento de classes pode resultar em uma maior capacidade em identificar observações provenientes da classe minoritária, em ambos os classificadores.

5.4 Discussão

A partir da análise dos resultados para os três conjuntos de dados abordados no trabalho, observamos que a aplicação de técnicas para lidar com o desbalanceamento dos dados aumenta a acurácia de ambos os classificadores na classe positiva e não afeta drasticamente seu desempenho na classe negativa. Esse fato é comprovado ao observamos que a G-média aumenta nos cenários em que as técnicas para lidar com o desbalanceamento foram aplicadas. Esse resultado sugere que as técnicas de reamostragem, as versões sensíveis ao custo e a escolha de pontos de corte ótimos surtem efeito na classificação de instâncias da classe minoritária e, de maneira geral, melhoram o desempenho do classificador. Lembramos que o fato de o MCC, na base onde o nível de desbalanceamento é severo, ser maior nos casos em que não houve a geração de observações sintéticas ou que não houve penalização para a classe minoritária já era esperado, uma vez que a introdução de exemplos sintéticos ou definição de um alto custo para a classe minoritária pode aumentar o número de falsos positivos e negativos, impactando negativamente o valor do MCC.

Os resultados ainda relevam que no caso da regressão logística, ao utilizarmos o ponto de corte padrão $\frac{1}{2}$, ela encontra dificuldade em classificar instâncias da classe minoritária.

Em particular, principalmente nos conjunto de dados com desbalanceamento leve e moderado, observamos que a utilização do ponto de corte padrão faz com que a capacidade do classificador em identificar instâncias da classe minoritária seja afetado negativamente. Logo, utilizar um ponto de corte que otimize alguma medida de performance de interesse, leva a um aumento da acurácia na classe minoritária e, por conseguinte, um aumento na medida G-média do classificador logístico, o que leva a uma melhora no seu desempenho na classificação de novas instâncias. Nesse contexto, podemos destacar uma das vantagens das máquinas de vetores suporte que por não ser um classificador probabilístico não carece da definição de pontos de corte para realizar a classificação de novas unidades amostrais.

O trabalho permite concluir também que ao ajustar adequadamente os pesos das classes é possível dar mais importância à classe minoritária durante o treinamento, aumentando a performance dos classificadores na classe positiva. Essa performance se equipara àquela obtida com o método de sobreamostragem SMOTE em sua versão pura e híbrida, o que sugere que a escolha entre métodos de reamostragem e versões sensíveis ao custo depende das características específicas do seu conjunto de dados e dos objetivos do seu problema de aprendizado de máquina. Ambas as abordagens se mostraram úteis para lidar com o desbalanceamento de classes.

Portanto, a partir dos resultados obtidos, notamos que o desempenho de ambos os classificadores performam de maneira similar. Embora as máquinas de vetores suporte tenham os classificadores com melhor desempenho nas bases com desbalanceamento leve e moderado ressaltamos que a utilização da regressão logística tem algumas vantagens, como o fato de sua aplicação com variáveis categóricas já estar difundida, a possibilidade de análise da contribuição de cada variável no resultado obtido e, por fim, o tempo de processamento ser significativamente menor do que os classificadores obtidos via máquina de vetores suporte.

Capítulo 6

Considerações finais

Neste trabalho, abordamos a classificação de novas unidades amostrais em situações de análise de crédito e detecção de fraude, cenários onde as instituições financeiras têm interesse e sempre buscam aprimorar os resultados obtidos. Por conta do alto interesse em aprimorar esses resultados, os problemas de classificação estão cada vez mais embasados em métodos analíticos obtidos a partir de modelos e algoritmos estatísticos. Além disso, observamos que nessas situações, é comum os conjuntos de dados serem desbalanceados, isto é, a quantidade observações de uma classe se sobrepor a outra.

Nesse sentido, a metodologia desenvolvida no decorrer dessa monografia, focou em estudar dois classificadores: (*i*) Regressão Logística (ver Seção 2.2) e (*ii*) Máquina de vetores suporte (SVM) (ver Seção 2.3). O primeiro é um método de classificação que se baseia em probabilidades para separar as observações em suas classes, enquanto o segundo é um método de classificação não probabilístico que se baseia na construção de um hiperplano para separar as observações em suas respectivas classes. Além disso, o estudo abrange métodos de pré-processamento dos dados (ver Capítulo 3), que tem o objetivo de superar a perca de performance dos classificadores quando utilizados em dados desbalanceados. Por fim, para lidar com o desbalanceamento dos conjuntos de dados, o estudo contemplará os algoritmos sensíveis ao custo (ver Capítulo 4) que buscam aliviar o problema de desbalanceamento adaptando diferentes funções de perda, com diferentes custos associados a cada classe.

A partir dos resultados obtidos, observamos que o nível de desbalanceamento dos dados afeta a performance de um classificador e utilizar um valor padrão $\frac{1}{2}$ como ponto de corte pode afetar a capacidade desse em identificar observações provenientes da classe minoritária. Logo, a utilização de um ponto de corte que maximiza as medidas G-Média

e MCC leva a um aumento da acurácia na classe positiva. Concluímos também que a utilização de técnicas de reamostragem e versões sensíveis ao custo contribuem para uma melhor performance do classificador, levando à uma melhora na capacidade de identificar uma observação proveniente da classe minoritária. De modo geral, ambos os classificadores performam de maneira similar e apesar de as máquinas de vetores suporte possuírem os classificadores com desempenho ligeiramente superior nas base com desbalanceamento leve e moderado, ressaltamos que sua utilização deve ser avaliada, dado que apresenta algumas desvantagens em relação à regressão logística, como a falta de flexibilidade para o uso de variáveis preditoras qualitativas, a impossibilidade de analisar a contribuição das variáveis no processo de classificação e o alto custo computacional para convergência do algoritmo.

Ressaltamos que as conclusões obtidas neste trabalho estão embasadas nos conjuntos de dados utilizados para aplicação, podendo ser diferentes caso analisadas em outros cenários. Nesse sentido, considerar um estudo de simulação para criar uma sensibilidade maior relativa às técnicas consideradas para lidar com o desbalanceamento das classes é uma extensão natural desse trabalho. Além disso, é possível considerar outras maneiras de lidar com o desbalanceamento de classes tais como métodos *ensemble*. Os métodos de aprendizado de máquina costumam atingir um platô, i.e, um ponto em que o desempenho do modelo estabiliza e não melhora significativamente, mesmo com o aumento do tamanho do conjunto de dados ou a complexidade do modelo. Uma maneira de superar essa dificuldade é considerar métodos de aprendizagem profunda, cuja aplicação nos conjuntos de dados aqui considerados também enxergamos como uma extensão desse trabalho.

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. *Google Scholar Google Scholar Digital Library Digital Library*.

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 *Machine Learning: ECML-97: 9th European Conference on Machine Learning Prague, Czech Republic, April 23–25, 1997 Proceedings 9*, páginas 146–153. Springer.

Kubat, M., Holte, R. C. e Matwin, S. (1998). Machine learning for the detection of oil spills in satellite radar images. *Machine Learning*, **30**, 195–215.

Laradji, I. H., Alshayeb, M. e Ghouti, L. (2015). Software defect prediction using ensemble learning on selected features. *Information and Software Technology*, **58**, 388–402.

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.

Márquez-Vera, C., Cano, A., Romero, C., Noaman, A. Y. M., Mousa Fardoun, H. e Ventura, S. (2016). Early dropout prediction using data mining: a case study with high school students. *Expert Systems*, **33**(1), 107–124.

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.

Ng, A. (2000). Cs229 lecture notes. *CS229 Lecture Notes*, **1**(1). Part V.

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

Shen, F., Wang, R. e Shen, Y. (2020). A cost-sensitive logistic regression credit scoring model based on multi-objective optimization approach. *Technological and Economic Development of Economy*, **26**(2), 405–429.

Singh, S. e Khim, J. (2021). Statistical theory for imbalanced binary classification. *ArXiv E-prints*, páginas arXiv–2107.

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.

Sushma, S., Prasanna Kumar, S. C. e Assegie, T. A. (2023). A cost-sensitive logistic regression model for breast cancer detection. *The Imaging Science Journal*, páginas 1–9.

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

Tran, Q. D. e Liatsis, P. (2016). Raboc: An approach to handle class imbalance in multimodal biometric authentication. *Neurocomputing*, **188**, 167–177.

Vapnik, V. (1999). *The nature of statistical learning theory*. Springer science & business media.

Vapnik, V. N. (1998). *Statistical learning theory*. J. Wiley.

Veropoulos, K., Campbell, I. e Cristianini, N. (1999). Controlling the sensitivity of support vector machines. Em *Proceedings of the International Joint Conference on Artificial Intelligence, Stockholm, Sweden (IJCAI99)*, páginas 55 – 60. Other: Workshop ML3.

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.

Zhu, B., Baesens, B. e vanden Broucke, S. K. (2017). An empirical comparison of techniques for the class imbalance problem in churn prediction. *Information Sciences*, **408**, 84–99.