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

MÉTODOS DE DETECÇÃO DE FRAUDE EM CARTÕES DE CRÉDITO: UM ESTUDO COMPARATIVO

Luiz Eduardo Piccin

Trabalho de Conclusão de Curso

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

Métodos de detecção de fraude em cartões de crédito: um estudo comparativo

Luiz Eduardo Piccin
Orientador(a): Prof. Dr. Ricardo Felipe Ferreira

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

São Carlos
Abril de 2022

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

Credit card fraud detection methods: a comparative study

Luiz Eduardo Piccin
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
April 2022

Luiz Eduardo Piccin

Métodos de detecção de fraude em cartões de crédito: um estudo comparativo

Este exemplar corresponde à redação final do trabalho de conclusão de curso devidamente corrigido e defendido por Luiz Eduardo Piccin e aprovado pela banca examinadora.

Aprovado em 14 de abril de 2022.

Banca Examinadora:

- Prof. Dr. Ricardo Felipe Ferreira
- Profa. Dra. Daiane Aparecida Zuanetti
- Prof. Dr. Márcio Alves Diniz

Agradecimentos

Agradeço a minha mãe Priscilla e ao meu pai Alexandre por todo amor, ensinamentos e suporte ao longo da minha trajetória. A minha noiva Milena pelo carinho e incentivos desde que nos conhecemos. As amizades construídas na universidade, que certamente levarei para vida, em especial André, Daniel, Hideki, Igor, Luben, Reinaldo, Vitor Gomes e Vitor Ramos. A todo o grupo PET pelos momentos e experiências vivenciadas. Finalmente, agradeço ao meu orientador Ricardo, por compartilhar seu tempo, vasto conhecimento e sua maneira de fazer ciência, sendo essencial em todo o trabalho.

"All models are wrong, but some are useful."

(George Box)

Resumo

A chegada da pandemia mudou radicalmente os hábitos de consumo de bens e serviços, passando a ocorrer quase que exclusivamente no mundo virtual, que por sua vez, no contexto de fraude, possui um maior número de brechas quando comparado ao mundo físico. Esse aumento na quantidade de transações *online* (aprovadas predominantemente com cartão de crédito) resultou em um maior número de fraudes. Na ótica do negócio, é de extrema importância que as companhias sejam capazes de detectar uma transação fraudulenta, evitando prejuízos no relacionamento com o cliente e também perdas financeiras.

Usualmente, no processo de detecção de fraude, existe nos bastidores um modelo preditivo, que se aproxima do ideal quando apresenta alta performance na detecção de transações fraudulentas e isso se estende para as transações legítimas (em termos técnicos, significa observar baixo volume de falsos negativos e positivos). Neste trabalho, propomos comparar a performance de dois classificadores base quando treinados em duas arquiteturas distintas: a versão limitada da regressão logística contra a sua versão não-limitada, ambas com regularização ℓ_1, utilizando como conjunto de treinamento tanto dados balanceados (via *k-means*) e diversificados (via *bagging*) quanto dados desbalanceados. Nos k subconjuntos de treinamento balanceados e diversificados a serem construídos, os classificadores bases são treinados, combinados por uma média ponderada e a previsão final é julgada a partir dessa média. O estudo comparativo é realizado em um cenário de dados reais, em termos da AUC (*Area Under the Curve*) e de outras estatísticas de teste, como o KS (índice Kolmogorov-Smirnov), por exemplo. Os resultados obtidos poderão ser comparados também com outras obras presentes na literatura.

Palavras-chave: *Detecção de fraude, Regressão logística limitada, Regularização ℓ_1, Balanceamento e diversificação do conjunto de treinamento.*

Abstract

The arrival of the pandemic radically changed the consumption habits of goods and services, starting to occur almost exclusively in the virtual world, which in turn, in the context of fraud, has a greater number of loopholes when compared to the physical world. This increase in the number of online transactions (predominantly approved by credit card) has resulted in a greater number of frauds. From the business point of view, it is extremely important that companies are able to detect a fraudulent transaction, avoiding damage to the customer relationship and also financial losses.

Usually, in the fraud detection process, there is a predictive model behind the scenes, which approaches the ideal when it presents high performance in the detection of fraudulent transactions and this extends to legitimate transactions (in technical terms, it means observing a low volume of false negatives and positives). In this work, we propose to compare the performance of two base classifiers when trained in two different architectures: the bounded version of logistic regression against its unbounded version, both with ℓ_1 regularization, using both balanced data (via k-means) and diversified (via bagging) as unbalanced data. On the k balanced and diversified training subsets to be built, the base classifiers are trained, combined by a weighted average and the final prediction is judged from this average. The comparative study is carried out in a real data scenario, in terms of AUC (Area Under the Curve) and other test statistics, such as KS (Kolmogorov-Smirnov), for example. The results obtained can also be compared with other works present in the literature.

Keywords: *Fraud detection, bounded logistic regression, ℓ_1 regularization, balancing and diversifying the training set.*

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

Cartões de crédito tornaram-se a forma mais popular de pagamento tanto para compras *online* quanto para compras *offline*, as quais lidam, diariamente, com milhares de transações fraudulentas.

O crescente número de fraudes impacta negativamente a receita das instituições financeiras ao redor do mundo. Assim, a detecção eficiente dessas fraudes é essencial para manter a confiança nesse sistema de pagamento. Nesse contexto, metodologias para detecção e prevenção à fraude tem sido desenvolvidas. Dentre essas metodologias existem aquelas que utilizam uma abordagem de aprendizagem supervisionada, tais como regressão logística (Shen *et al.*, 2007), árvores de decisão (Sahin *et al.*, 2013), *support vector machine* (Sahin e Duman, 2011), redes neurais (Fanning *et al.*, 1995); e outras que utilizam uma abordagem não-supervisionada como, por exemplo, *restricted Boltzmann machine* (Niu *et al.*, 2019), *k-means* e *density-based spatial clustering of application with noise* - DBSCAN (Beltran, 2019). Niu *et al.* (2019) realizaram um estudo comparativo entre as abordagens, observando resultados melhores na aprendizagem supervisionada. Embora muitos dos métodos propostos tenham alcançado resultados promissores, é ainda um grande desafio detectar prontamente e com precisão transações fraudulentas de cartão de crédito devido ao desbalanceamento das classes, i.e., o número de transações legítimas é muito maior do que as fraudulentas.

Um modelo ideal para detecção de fraude deve ter um bom desempenho tanto na classificação de indivíduos ou objetos pertencentes à classe majoritária (transações legítimas) quanto para aqueles da classe minoritária (transações fraudulentas). Os algoritmos usuais de aprendizagem, em geral, favorecem a classe majoritária e, geralmente, apresentam baixo desempenho na classificação de indivíduos ou objetos da classe minoritária (Wang

et al., 2015). Nesse sentido, uma alternativa, segundo Wang *et al.* (2015), é considerar conjuntos de treinamento equilibrados, obtidos a partir de algoritmos de agrupamento; e que sejam diversificados a partir de técnicas baseadas em árvores de decisão com reamostragem *bootstrap*.

Moraes (2008) utiliza amostras *state-dependent* para diversificar e equilibrar o conjunto de treinamento. Todavia, ela conclui que a estimação dos parâmetros baseada em amostras de treinamento *state-dependent* geram, de acordo com certas estatísticas de teste, classificadores piores do que aqueles cuja estimação dos parâmetros foi conduzida com base em amostras de treinamento que não sofreram diversificação e balanceamento. Além disso, Moraes (2008) propõe utilizar como classificador base a regressão logística limitada, pois o modelo logístico não-limitado não apresenta boa performance quando a variável resposta é extremamente desbalanceada (Cramer, 2004). Na dissertação de mestrado de Moraes (2008), a autora apresenta um estudo comparativo entre a forma limitada e não-limitada do modelo de regressão logístico e conclui, de acordo com certas métricas para comparação de modelos, que a versão limitada do modelo é um classificador melhor do que a sua versão não-limitada.

Por outro lado, em um contexto de *credit scoring*, Wang *et al.* (2015) propõem que as amostras de treinamento sejam, primeiramente, obtidas de forma equilibrada e diversificada através do uso de algoritmos de agrupamento e *bagging*. Em seguida, a regressão logística com regularização ℓ_1 (LASSO) é utilizada como classificador base para avaliar o risco de crédito. Os autores mostram que o algoritmo proposto supera os modelos de pontuação de crédito mais populares, como, por exemplo, árvores de decisão e floresta de decisão em termos da AUC (*Area Under the Curve*) e da estatística F1 (média harmônica da precisão e sensibilidade).

Neste trabalho, propomos utilizar, como classificador base, uma versão limitada da regressão logística com regularização ℓ_1 e, então, comparar a sua performance com a sua versão não-limitada utilizando como conjunto de treinamento tanto dados balanceados quanto dados desbalanceados. Para o balanceamento dos dados, vamos utilizar o algoritmo de agrupamento k-means para dividir a classe majoritária (transações legítimas) em k subgrupos. Para a diversificação, vamos utilizar diferentes subgrupos da classe majoritária (transações legítimas) combinadas com versões *bagging* da classe minoritária (transações fraudulentas) para formar k conjuntos de treinamento. Classificadores bases são, então, treinados com base em cada um dos conjuntos de treinamento. Em seguida,

esses classificadores são agrupados com base em uma média ponderada e a previsão final é julgada a partir dessa média. O estudo comparativo será realizado em um cenário de dados reais, em termos da AUC, da estatística de Kolmogorov-Smirnov (KS), da estatística F1, da acurácia e da soma de quadrados dos resíduos. Os resultados obtidos poderão, portanto, ser comparados com os obtidos em Moraes (2008) e Wang *et al.* (2015).

A principal contribuição deste trabalho é o estudo da performance da versão limitada da regressão logística com regularização ℓ_1. Até onde vai nosso conhecimento, não há na literatura trabalho algum que utilize esse classificador base para detecção de fraudes. Portanto, este trabalho complementa os estudos sobre métodos para detecção de fraude em cartão de crédito.

Este trabalho está organizado da seguinte maneira. No próximo capítulo, apresentamos o funcionamento do mercado de cartões de crédito e como se dão as fraudes nesse ambiente. No Capítulo 3, além de explicar como realizaremos o balanceamento e a diversificação do conjunto de treinamento, estudamos algumas metodologias para detecção de fraude e apresentamos um novo classificador, baseado na versão limitada da regressão logística com regularização ℓ_1, que acreditamos possuir uma boa performance para detectar fraudes em cartão de crédito. As medidas que utilizaremos para avaliar a performance dos classificadores também são apresentadas no Capítulo 3. No Capítulo 4 realizamos um estudo de caso, aplicando todas as metodologias propostas e, por fim, o Capítulo 5 encerra esta monografia com algumas considerações, conclusões e sugestões para estudos futuros.

Capítulo 2

Cartão de crédito e Fraudes

O cartão de crédito, método de pagamento emitido por instituição financeira e consolidado pela sua portabilidade e relativa segurança, vem movimentando cada vez mais o mercado de pagamentos. A Figura 2.1 traz dados levantados pela ABECS (Associação Brasileira das Empresas de Cartão de Crédito e Serviços) mostrando que o montante transacionado por brasileiros (no Brasil e no exterior), nessa modalidade, está crescendo ano após ano.

Entretanto, quando falamos desse meio de pagamento e as altas cifras movimentadas pelo mesmo, não podemos deixar de citar os prejuízos ocasionados pelas fraudes. A definição formal para fraude, segundo o dicionário *Oxford Languages*, é "ato ardiloso, enganoso e de má-fé que tem o objetivo de lesar ou ludibriar outrem para trazer algum tipo de vantagem". No contexto de cartão de crédito, trata-se de uma vantagem financeira que o fraudador exerce sobre a vítima (dono do cartão). A fraude está inserida em um universo complexo, com diferentes naturezas de crimes e penas previstas no Código Penal Brasileiro. Um grande problema é que a fraude se comporta como um negócio rentável e estável para aqueles que a praticam (Hand, 2002).

Cressey (1953) desenvolveu uma teoria conhecida como triângulo da fraude, contemplando três dimensões do comportamento fraudulento: a pressão, a oportunidade e a racionalização. Em maiores detalhes, a pressão, também conhecida como motivação, prescreve a existência de problemas financeiros ocultos, que não podem ser compartilhados. Nisso, surge a oportunidade de resolver secretamente esses problemas, pela violação da confiança financeira. Por fim, a racionalização do ato fraudulento, como necessário e justificável para resolução dos problemas financeiros.

Além disso, com o recente crescimento do *e-commerce*, impulsionado pela pandemia

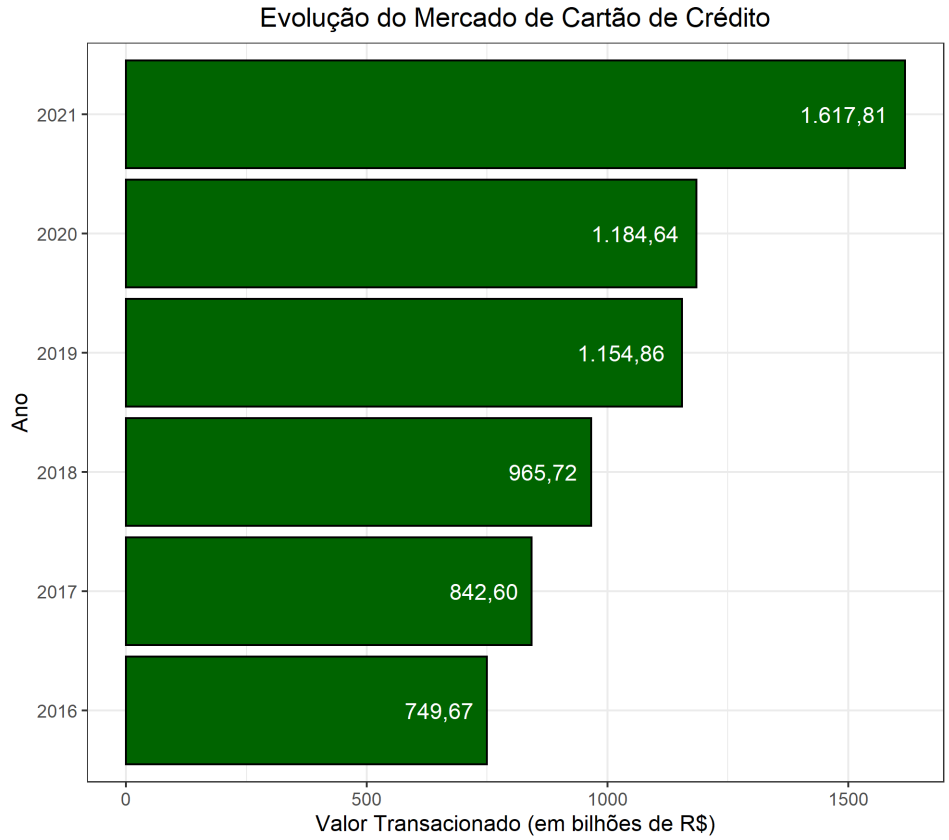

Figura 2.1: Valor transacionado (em bilhões de reais) com cartão de crédito nos últimos anos. Fonte: ABECS.

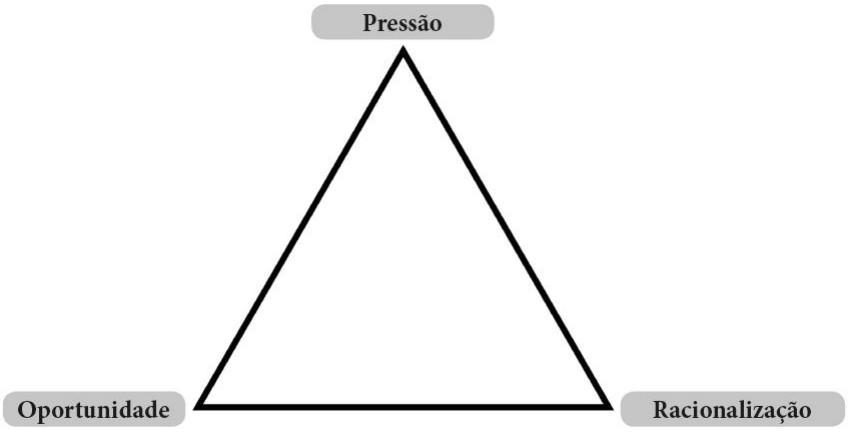

Figura 2.2: Triângulo da fraude construído por Cressey (1953).

do novo coronavírus (uma vez que grande parte das transações passaram a ocorrer no mundo *online*), o uso de cartão de crédito como forma de pagamento aumentou muito a partir de 2020, levando a um grande acréscimo no número de fraudes (Singh *et al.*, 2021).

2.1 Mercado de cartão de crédito: composição e funcionamento

É importante mencionar que há dois tipos principais de cartões de crédito: a versão básica (destinada apenas para pagamentos) e a versão diferenciada (além de permitir pagamentos, oferece benefícios como milhas, descontos, atendimento personalizado, etc.), sendo que a última tem como objetivo fidelizar o cliente, isto é, fazer com que o cartão de crédito seja sua principal forma de pagamento.

Explicamos a seguir a forma geral do funcionamento dessa modalidade de pagamento.

No mercado de cartões existem dois grupos de interessados que garantem a subsistência dessa forma de pagamento:

(i) as pessoas que desejam realizar pagamentos com os cartões;

(ii) os estabelecimentos que desejam aceitar pagamentos com cartões.

Para garantir o pleno funcionamento dessa plataforma de pagamento, podemos elencar cinco agentes principais e suas respectivas responsabilidades:

1. **Bandeiras:** criam as regras da operação (entre emissores e adquirentes), mantém uma rede global de comunicações, promovem políticas e ações de marketing para assegurar a dinâmica do mercado. Os principais exemplos são: Visa e MasterCard;

2. **Emissores:** são os responsáveis por emitir o cartão, manter relacionamento direto com os portadores, e oferecer o crédito a eles. Em geral, os bancos e outras instituições financeiras desempenham essa função;

3. **Adquirentes:** responsáveis por credenciar e manter o relacionamento com os estabelecimentos, capturar as transações, processá-las e fazer a liquidação delas nas contas bancárias dos estabelecimentos. Alguns exemplos são: Cielo, Rede e Stone;

4. **Estabelecimentos:** entidades dispostas a aceitarem pagamentos com cartões. Para isso, é necessário tornar-se cliente de alguma das adquirentes, recebendo a máquina de cartões (para compras presenciais) e/ou acesso aos servidores da adquirente (para compras *online*);

5. **Portadores:** em geral são as pessoas físicas dispostas a pagarem com cartão, tornando-se cliente de algum emissor de cartão. Pessoas jurídicas também podem usufruir dos cartões, solicitando ao emissor o cartão corporativo.

Após explicar sobre os principais agentes que compõem essa modalidade, explicaremos como o pagamento é realizado.

No mundo físico, quando o cartão é acionado (leitura do chip pela máquina), no mesmo instante é transmitido um sinal para o adquirente, que repassa para a bandeira, que por sua vez envia ao emissor. É ele quem decide por aprovar ou recusar a transação (baseando-se em alguns critérios, como limite disponível, status da fatura, etc.), informando ao estabelecimento a decisão final.

No mundo virtual o processo é muito semelhante, com exceção da primeira etapa, já que não há a máquina de cartão. Ao inserir os dados do cartão e confirmar o pagamento da compra, essas informações são repassadas à adquirente através de um *gateway* (intermediário para troca de informações), que as criptografa com intuito de garantir a segurança da informação. Ao chegar na adquirente, o fluxo final é o mesmo que o citado anteriormente.

2.2 Tipos de fraude

No complexo universo de fraudes, essas podem ser divididas em diferentes tipos, sendo os mais comuns listados abaixo:

- **Roubo de identidade** – é o tipo mais comum, onde o fraudador está munido com informações sensíveis de terceiros (conta bancária e/ou cartão de crédito, além da senha de acesso), realizando compras no nome da vítima;

- **Interceptação de mercadorias** – o fraudador altera o endereço de entrega que o comprador legítimo cadastrou no site de compra ou então, conhecendo a data e local de entrega, se passa pelo destinatário;

- **Abuso** – diferente das anteriores, o próprio portador do cartão de crédito realiza a compra, e com intuito de não pagá-la, notifica a instituição alegando que não reconhece a compra, recebendo o estorno;

- **Fraude amiga** – alguém próximo ao titular do cartão, por exemplo parentes ou amigos, com conhecimento das informações necessárias (número do cartão, senha e outros dados bancários) compra algo sem o aval do titular.

Salienta-se que em todos esses tipos listados, transações realizadas no mundo *online* são mais propícias ao fraudador, já que não é possível validar nenhum tipo de documento de identidade do dono do cartão (Moraes, 2008).

2.3 Ciclo da fraude

Alguns autores, como Cristofaro (2006), sugerem um possível e eficiente ciclo de vida para gestão de fraude, dividido em oito estágios. As instituições são responsáveis por encontrar o equilíbrio entre os pontos, além da melhor ordem de execução. Os oito estágios são:

(i) **Intimidação** – ações que visam desestimular o fraudador antes que ele tente a fraude;

(ii) **Prevenção** – enrijecer os protocolos de segurança para que a execução da fraude fique mais difícil. Deve ser utilizado com cuidado, pois um grande engessamento pode gerar desconforto no cliente legítimo;

(iii) **Detecção** – ações que visam implementar métodos para detecção de fraude. É o objetivo do presente estudo;

(iv) **Medidas** – atitudes que evitem (ou minimizem) perdas à instituição (ou seja, impedir que o fraudador continue fraudando);

(v) **Análise** – encontrar fatores relacionados à fraude;

(vi) **Política** – adotar políticas que reduzam a incidência de fraudes;

(vii) **Investigação** – obter informações relevantes para reduzir ou inibir as fraudes, com objetivo máximo de recuperar recursos ou receber restituições;

(viii) **Acusação** – condenar fraudadores em termos legais.

Com uma aplicação assertiva dessas ações é possível atingir resultados eficientes no combate à fraude.

2.4 Impactos da fraude

A fraude tem um grande impacto social e financeiro. Para as instituições, gera-se um grande custo financeiro, dos quais podemos citar: o valor perdido na transação e os gastos com as análises (que vão da construção e implementação do modelo estatístico até as verificações manuais, quando existirem). Há também desgastes de imagem e insatisfação do cliente, os quais são difíceis de serem mensurados.

No Brasil, estima-se a ocorrência de 7 fraudes por minuto, gerando um impacto anual estimado em 3,6 bilhões de reais. No geral, estima-se que 1,34% das transações realizadas são tentativas de fraude. Esse percentual é maior em algumas regiões do país, como no Norte, chegando a 3,52% das transações. Nota-se também que há uma concentração no produto a ser fraudado, sendo que celulares são as principais vítimas, concentrando 4,24% das tentativas de fraudes, devido à sua alta procura no mercado e à facilidade de revenda (ClearSale, 2021).

Capítulo 3

Métodos de detecção de fraude

Como visto no capítulo anterior, a fraude de cartões de créditos têm impactos sociais e econômicos negativos. No ciclo de gestão de fraude, Cristofaro (2006) cita a detecção como um dos estágios. Entretanto, com milhões de transações ocorrendo diariamente, é inviável uma análise humana para todo esse volume, e como a grande maioria se trata de transações legítimas essa checagem seria muito onerosa. É exatamente nessa etapa em que entram os modelos estatísticos, que após serem devidamente calibrados, podem ser aplicados às transações futuras, colaborando para a detecção rápida e eficiente de transações fraudulentas. Nesse sentido, nas próximas seções, estudamos algumas metodologias para a detecção de fraude que apresentaram uma boa performance em estudos anteriores e, além disso, propomos um novo método que acreditamos possuir uma boa solução para o problema em questão.

3.1 Balanceamento e diversificação do conjunto de treinamento

No contexto de *credit scoring*, onde o conjunto de dados é desbalanceado (no caso binário, em estudo, significa que alguma das duas classes possui incidência muito menor que a outra) Wang *et al.* (2015) propuseram um tratamento no conjunto de treinamento (partição do conjunto de dados em que os parâmetros do classificador serão estimados) que fez com que os classificadores (métodos para determinar a qual classe o indivíduo pertence, com base em suas covariáveis) utilizados superassem, em termos da performance da predição (métrica utilizada para avaliar quão bom é o classificador, com base nos acertos e

erros de classificação), aqueles que foram obtidos a partir de um conjunto de treinamento sem a aplicação desse tratamento. O tratamento em questão consiste em equilibrar as classes majoritária (de maior incidência) e minoritária (de menor incidência) do conjunto de dados e diversificá-las. Neste trabalho vamos dividir a classe majoritária (transações legítimas) em k *clusters*, $k \in \mathbb{N}^*$, através do algoritmo *k-means*, e a cada um desses *clusters* será anexada uma amostra *bagging* da classe minoritária (transações fraudulentas). Ao final, teremos k subconjuntos de treinamento balanceados e diversificados.

3.1.1 Algoritmo de agrupamento *k-means*

Em linhas gerais, algoritmos de agrupamento possuem como objetivo, dividir uma coleção de objetos ou indivíduos em subconjuntos de forma que os objetos ou indivíduos que compõem um subconjunto específico sejam similares entre si e sejam suficientemente diferentes daqueles que pertencem a um subconjunto diferente. Dessa forma, espera-se que os objetos ou indivíduos que pertencem a um mesmo subconjunto possuam características ou propriedades similares.

Os subconjuntos formados a partir de algoritmos de agrupamento são popularmente conhecidos na literatura como *clusters*. Essa será a terminologia que vamos utilizar neste trabalho a partir de agora.

Os algoritmos de agrupamento mais populares atribuem diretamente um indivíduo ou objeto a um *cluster*, sem considerar um modelo probabilístico. Nesse sentido, os métodos de agrupamento tem por objetivo obter uma partição $C_1, \ldots, C_k$ de uma coleção de objetos ou indivíduos $\{1, 2, \ldots, n\}$. Em outras palavras, queremos determinar $C_1, \ldots, C_k$ de forma que

$$C_1 \cup \ldots \cup C_k = \{1, 2, \ldots, n\}$$

e

$$C_i \cap C_j = \varnothing, \, \forall i, j \in \{1, 2, \ldots, k\} \text{ com } i \neq j.$$

Um conceito fundamental a todos os objetivos da análise de agrupamento é o de similaridade (ou dissimilaridade) entre dois objetos (ou indivíduos). Para calcular a similaridade (ou dissimilaridade) entre os objetos ou indivíduos, alguns métodos, utilizam alguma medida de distância (métrica).

Definição 3.1 (Métrica) *Seja $p \geq 1$ um número natural. Uma **métrica** em $\mathbb{R}^p$ é uma*

função $d : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ *que associa a cada par ordenado de vetores* $(\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{R}^p \times \mathbb{R}^p$ *um número real* $d(\boldsymbol{x}, \boldsymbol{y})$*, chamado* ***distância*** *entre* $\boldsymbol{x}$ *e* $\boldsymbol{y}$*, de modo que sejam satisfeitas as seguintes condições para quaisquer* $\boldsymbol{x}$*,* $\boldsymbol{y}$ *e* $\boldsymbol{z}$*:*

1. $d(\boldsymbol{x}, \boldsymbol{x}) = 0$*;*
2. *Se* $\boldsymbol{x} \neq \boldsymbol{y}$ *então* $d(\boldsymbol{x}, \boldsymbol{y}) > 0$*;*
3. $d(\boldsymbol{x}, \boldsymbol{y}) = d(\boldsymbol{y}, \boldsymbol{x})$*;*
4. $d(\boldsymbol{x}, \boldsymbol{z}) \leq d(\boldsymbol{x}, \boldsymbol{y}) + d(\boldsymbol{y}, \boldsymbol{z})$*.*

Nesse sentido, suponha que observamos p características numéricas dos objetos ou indivíduos $\{1, 2, \ldots, n\}$. Seja $\boldsymbol{x}_i \in \mathbb{R}^p$ o vetor das características observadas associado ao indivíduo $i \in \{1, 2, \ldots, n\}$. Assim, a similaridade (ou dissimilaridade) entre dois objetos distintos i e j da coleção $\{1, 2, \ldots, n\}$ pode ser medida através de alguma métrica d. Algumas medidas que podem ser utilizadas são a distância Euclidiana, em que d é dada por

$$d(\boldsymbol{x}_i, \boldsymbol{x}_j) := \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2};$$

e a distância de Minkowski, em que d é definida como

$$d(\boldsymbol{x}_i, \boldsymbol{x}_j) := \left(\sum_{k=1}^{p} |x_{ik} - x_{jk}|^m \right)^{1/m}.$$

Note que a distância de Minkowski traz consigo alguns casos especiais, por exemplo: a distância absoluta ou de Mahalanobis, que ocorre quando $m = 1$, a distância Euclidiana quando $m = 2$, e distância de Chebyshev, que é caso em que $m \to \infty$. Neste trabalho, vamos utilizar a distância Euclidiana como medida de similaridade (ou dissimilaridade).

É importante salientar que os métodos de agrupamento podem ser divididos em algumas categorias, sendo as principais:

1. **Métodos baseados em particionamento:** são os mais básicos, geralmente com duas entradas – as observações e o número de *clusters* a serem formados. O método itera de forma que as observações que compõem cada *cluster* sejam atualizadas, baseado em alguma medida de similaridade, até que nenhuma modificação seja necessária;

2. **Métodos hierárquicos:** as observações são agrupadas a cada iteração segundo um critério pré-estabelecido. Podem ser executados na versão aglomerativa (parte da premissa que cada observação seja um grupo, e a cada iteração o grupo mais similar seja combinado, de forma que ao final do processo, reste apenas um grupo, com todas as observações iniciais) ou divisora (inicia com um grupo composto por todas observações, e a cada iteração um grupo é dividido em duas partes da forma mais heterogênea possível até restarem as observações isoladas), criando uma hierarquia entre os dados que pode ser representada por um dendrograma;

3. **Métodos baseados em distribuições:** assumem que as observações são compostas de distribuições. Conforme a distância para o centro da distribuição aumenta, a chance da observação pertencer à distribuição diminui. Vale a recíproca.

Nesta monografia, vamos utilizar um algoritmo de agrupamento baseado em particionamento: o algoritmo *k-means*, que é um dos métodos de agrupamento mais utilizados na literatura e que funciona de forma iterativa. A medida de dissimilaridade d utilizada durante sua execução normalmente é a distância Euclidiana (todavia, é possível executar o algoritmo com outras medidas de distância). É necessário escolher a quantidade de *clusters* que serão formados, que é um número inteiro positivo a ser representado por k. O objetivo é, portanto, encontrar a partição $C_1, \ldots, C_k$ da coleção de objetos ou indivíduos $\{1, 2, \ldots, n\}$ de forma que a soma dos quadrados dentro de cada *cluster*

$$\sum_{l=1}^{k} \frac{1}{|C_l|} \sum_{i,j \in C_l} d^2(\boldsymbol{x}_i, \boldsymbol{x}_j) \tag{3.2}$$

seja a menor possível, em que $|C_l|$ denota o número de elementos do conjunto C_l.

Como existem k^n maneiras de particionar a coleção dos objetos ou indivíduos em k *clusters*, não é possível encontrar analiticamente a partição que minimiza a soma (3.2), ou seja, a solução pode ser encontrada apenas de forma numérica. Um dos algoritmos iterativos que podem ser utilizados para esse fim é denominado Algoritmo de Lloyd. Esse algoritmo consiste nos seguintes passos:

1. Escolher um número k de *clusters*;

2. Escolher os centróides $\boldsymbol{c}_1, \boldsymbol{c}_2, \ldots, \boldsymbol{c}_k$ de cada *cluster*, em que $\boldsymbol{c}_l \in \mathbb{R}^p$, $l = 1, 2, \ldots, k$. As maneiras mais comuns de escolher os centróides iniciais são:

- Criar k vetores aleatórios, e eleger cada um deles como o l-ésimo centróide, $l = 1, 2, \ldots, k$;
- Escolher aleatoriamente por k objetos da coleção, e como na opção anterior, nomear cada um deles como o l-ésimo centróide, $l = 1, 2, \ldots, k$;
- *k-means++* (Arthur e Vassilvitskii, 2006), uma proposta de escolha dos centróides iniciais para aumentar a chance de que o mínimo global seja encontrado. Nele, escolhemos o primeiro centróide $\boldsymbol{c}_1$ aleatoriamente na coleção de objetos e então definimos $\boldsymbol{C} = \{\boldsymbol{c}_1\}$, em que $\boldsymbol{C}$ denota o conjunto do centróides. As escolhas subsequentes levam em consideração uma probabilidade p_i definida por

$$p_i := \frac{D^2(\boldsymbol{x}_i)}{\sum_{j=1}^{n} D^2(\boldsymbol{x}_j)},$$

em que

$$D(\boldsymbol{x}_i) = \min_{\boldsymbol{c} \in \boldsymbol{C}} ||\boldsymbol{x}_i - \boldsymbol{c}||.$$

e $\boldsymbol{x}_i$ é um objeto da coleção. Ou seja, a escolha dos próximos centróides deixa de possuir probabilidade uniforme, passando a ser mais provável a escolha dos objetos mais distantes dos centróides já escolhidos, por possuírem maior p_i. Escolhendo $\boldsymbol{c}_l$, atualizamos

$$\boldsymbol{C} = \boldsymbol{C} \cup \{\boldsymbol{c}_l\},$$

recalculamos as probabilidades associadas aos objetos e escolhemos novo centróide seguindo o mesmo procedimento. Esse processo é repetido até termos escolhidos k centróides, ou seja, $\boldsymbol{C} = \{\boldsymbol{c}_1, \ldots, \boldsymbol{c}_k\}$.

Em seguida, iteramos o algoritmo até a convergência, isto é, o momento em que nenhuma observação necessita trocar de *cluster*. Esse processo se dá pelas seguintes etapas:

3. Associar cada observação ao *cluster* mais próximo pela distância Euclidiana. Defina, portanto, o *cluster* C_r, $r \in \{1, 2, \ldots, k\}$, como sendo

$$C_r := \left\{ i \in \{1, 2, \ldots, n\} : \arg\min_{1 \leq l \leq k} d(\boldsymbol{x}_i, \boldsymbol{c}_l) = r \right\},$$

em que $\boldsymbol{x}_i \in \mathbb{R}^p$ é o vetor das características numéricas observadas no objeto ou indivíduo i.

4. Calcular os novos centróides para cada um dos k *clusters* que foram criados:

$$\boldsymbol{c}_r := \frac{1}{|C_r|} \sum_{i \in C_r} \boldsymbol{x}_i, \quad r = 1, \ldots, k. \tag{3.3}$$

Existe um problema com o algoritmo: não há garantias teóricas que a melhor solução será de fato encontrada, isso porque o algoritmo depende das escolhas iniciais dos centróides. Cada escolha pode levar a um mínimo local diferente, não convergindo, portanto, necessariamente, para um mínimo global. Para contornar esse problema, uma possibilidade é executar o algoritmo várias vezes com diferentes inicializações (James *et al.*, 2013), a serem escolhidos a partir do algoritmo *k-means++*. Dentre as partições resultantes, escolhemos aquela que minimiza a Equação (3.2).

3.1.2 *Bagging*

Podemos utilizar técnicas de reamostragem para melhorar a estimação e a predição de um modelo estatístico (Elith *et al.*, 2008). Como motivação, considere o seguinte problema de regressão. Suponha que ajustamos um modelo para o conjunto de treinamento $\boldsymbol{Z} = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$, em que $\boldsymbol{x}_i \in \mathbb{R}^p$ é o vetor de covariáveis observadas e $y_i \in \mathbb{R}$ é a observação da variável resposta, ambos associados a um objeto ou a um indivíduo i. Seja $\hat{g} : \mathbb{R}^p \to \mathbb{R}$ a função de predição obtida com a estimação dos parâmetros do modelo em questão a partir do conjunto de treinamento $\boldsymbol{Z}$. Assim, se $\boldsymbol{Z}_1$ e $\boldsymbol{Z}_2$ são dois conjuntos de treinamento distintos, e $\hat{g}_1$ e $\hat{g}_2$ são as funções de predição estimadas, respectivamente, a partir dos conjuntos de treinamento, segue, para qualquer nova entrada $\boldsymbol{x} \in \mathbb{R}^p$, que uma função de predição $\hat{g}$ tal que

$$\hat{g}(\boldsymbol{x}) = \frac{\hat{g}_1(\boldsymbol{x}) + \hat{g}_2(\boldsymbol{x})}{2},$$

resulta em

$$E[(Y - \hat{g}(\boldsymbol{x}))^2 | \, \boldsymbol{x}] \leq E[(Y - \hat{g}_l(\boldsymbol{x}))^2 | \, \boldsymbol{x}], \; l = 1, 2,$$

desde que $\hat{g}_1$ e $\hat{g}_2$ sejam não-correlacionadas, não-viesadas e com a mesma variância (Izbicki e dos Santos, 2020). Tal conclusão é válida mesmo quando combinamos B funções de predição. Portanto, é melhor utilizar uma função de predição combinada ao invés de usá-las separadamente, uma vez que o seu risco de predição será sempre menor.

O método *bagging*, cujo nome deriva da expressão *bootstrap aggregation*, utiliza a motivação anterior para criar B estimadores distintos, utilizando B amostras *bootstrap* da

amostra original. Lembramos que a amostra *bootstrap* (não-paramétrica) é obtida através da reamostragem aleatória com reposição dos dados originais (nesse estudo, apenas a classe minoritária passará pelo *bootstrap*, antes de serem acopladas nos subconjuntos de treinamento). Para modelos de regressão, a função de predição final do método é dada por

$$\hat{g}(\boldsymbol{x}) = \frac{1}{B}\sum_{b=1}^{B}\hat{g}_b(\boldsymbol{x}),$$

em que $\boldsymbol{x} \in \mathbb{R}^p$ é uma nova entrada de covariáveis e $\hat{g}_b$ é a função de predição estimada utilizando a b-ésima reamostragem.

Em problemas de classificação, a agregação dos estimadores pode ser feita através da moda, ou seja

$$\hat{g}(\boldsymbol{x}) = \text{moda}(\hat{g}_b(\boldsymbol{x}),\ b = 1, \ldots, B),$$

qualquer que seja a nova entrada de covariáveis $\boldsymbol{x} \in \mathbb{R}^p$. Intuitivamente, uma nova observação $\boldsymbol{x}$ é avaliada em todos os estimadores e a predição final é dada pela categoria predita com maior frequência.

No presente estudo, faremos a agregação dos estimadores por uma média ponderada, como proposto por Wang *et al.* (2015). Em cada um dos k subconjuntos de treinamento construídos por reamostragem, será treinado um classificador base, e sua performance será avaliada em termos de alguma estatística de teste. Seja p_l a performance do l-ésimo classificador, treinado no l-ésimo subconjunto e avaliado nos demais $k-1$ subconjuntos de treinamento e $\alpha_l \in [0, 1]$ um peso associado à performance p_l desse classificador, definido a partir da seguinte função sigmóide

$$\alpha_l = \frac{1}{1 + e^{-p_l}}, \quad l = 1, 2, \ldots, k. \tag{3.4}$$

O classificador final é construído através da agregação pela média ponderada em α_l^* dos classificadores construídos nos k subconjuntos de treinamento, isto é,

$$\hat{g}(\boldsymbol{x}) := \sum_{i=l}^{k}\alpha_l^*\hat{g}_l(\boldsymbol{x}), \tag{3.5}$$

em que $\boldsymbol{x} \in \mathbb{R}^p$ é a nova entrada de covariáveis, e $\alpha_l^* = \frac{\alpha_l}{\sum_{i=l}^{k}\alpha_l}$.

3.1.3 Algoritmo para balanceamento e diversificação

Ao longo da Seção 3.1 introduzimos o tratamento no conjunto de treinamento proposto por Wang *et al.* (2015) e descrevemos as técnicas a serem utilizadas ao longo do processo. Resumindo, temos o seguinte roteiro para realizar o balanceamento e diversificação das transações que compõem o conjunto de treinamento:

1. Dividir a classe majoritária (em nosso caso, transações legítimas) em k *clusters*, através do algoritmo k-*means*. A escolha de k é discutida durante a aplicação da metodologia no conjunto de dados;

2. Gerar k amostras *bootstrap* da classe minoritária (nesse estudo, transações fraudulentas). Esse número k de amostras *bootstrap* deve coincidir com o número de *clusters*;

3. Construir k subconjuntos de treinamento, cada um deles com algum dos *clusters* de transações legítimas unificado com uma amostra *bootstrap* das transações fraudulentas;

4. Em cada um dos k subconjuntos é treinado o classificador base, que terá sua performance aferida nas transações complementares ao k-ésimo subconjunto. A medida é armazenada em α_l, definida na Equação (3.4).

5. Por fim, temos a agregação dos classificadores em um classificador final, algo característico do *bagging*. Podemos dizer que esses classificadores são não-correlacionados (premissa do *bagging*) por utilizar transações legítimas de *clusters* distintos (que tendem ser heterogêneos entre si). Essa agregação é dada pela Equação (3.5).

3.2 Classificação

Genericamente, qualquer método que consiga discriminar indivíduos ou objetos em classes ou categorias de uma variável qualitativa de interesse, é denominado um método de classificação. Existem muitas técnicas de classificação, ou classificadores, que podem ser utilizados para predizer uma variável resposta qualitativa a partir de um conjunto de covariáveis. Neste trabalho, estamos interessados em classificar transações em fraudulentas ou legítimas, a partir do conhecimento de um conjunto de covariáveis relevantes

para o mercado de cartões de créditos. Nesse contexto, existem diversos classificadores que podem ser utilizados para resolver esse problema, os quais podem ser obtidos utilizando abordagens supervisionadas (a verdadeira classe da transação é conhecida e utilizada na construção do classificador) ou não-supervisionadas (a verdadeira classe da transação pode não ser conhecida, e o classificador deve agrupar as transações em grupos homogêneos). Niu *et al.* (2019) compararam os classificadores das duas abordagens no cenário de fraude e identificaram melhor performance naqueles obtidos a partir de uma abordagem supervisionada.

Os classificadores construídos a partir de uma abordagem supervisionada são aqueles que necessitam das transações previamente rotuladas. Eles podem ser obtidos a partir de diferentes metodologias, tais como regressão logística (Cox, 1958), árvore de classificação (Quinlan, 1986), *support vector machine* - SVM (Cortes e Vapnik, 1995) e redes neurais (Fanning *et al.*, 1995). Devido à sua ampla aceitação na literatura estatística e por estar difundida no ramo da modelagem e possibilitar alta interpretabilidade em seus parâmetros, no presente estudo iremos abordar a regressão logística e suas variações.

Antes de apresentar os classificadores que consideramos neste trabalho, vamos estabelecer a notação utilizada e o espaço em que vamos trabalhar. Nesta monografia consideramos o experimento aleatório que consiste em selecionar n, $n \in \mathbb{N}^*$, transações de cartão de crédito de uma certa população finita. Defina a variável aleatória Y que representa se a transação é fraudulenta ou legítima, i.e.,

$$Y = \begin{cases} 1, & \text{se a transação é fraudulenta,} \\ 0, & \text{se a transação é legítima.} \end{cases}$$

Estamos interessados em estimar a probabilidade da transação ser fraudulenta condicionada à observação de um conjunto de covariáveis. Sejam $X_1, X_2, \ldots, X_p$ as covariáveis consideradas no processo de estimação, de forma que a quantidade p de covariáveis seja um número natural menor do que n. Nesse contexto, a classificação das transações em fraudulentas ou legítimas, é realizada, a partir da modelagem da probabilidade da transação ser fraudulenta condicionada à observação do conjunto de covariáveis, i.e.,

$$P(Y = 1|\boldsymbol{x}) = \pi(\boldsymbol{x}), \tag{3.6}$$

em que $\boldsymbol{x} \in \mathbb{R}^p$ é a observação do conjunto de covariáveis e $\pi : \mathbb{R}^p \to [0, 1]$ é uma função

estritamente crescente.

3.2.1 Regressão logística

A regressão logística visa explicar a relação entre uma variável resposta binária dependente de um conjunto de covariáveis explicativas independentes (Hosmer *et al.*, 1989). Assim, seja Y_i a variável aleatória binária, que representa se a transação i, $i \in \{1, 2, \ldots, n\}$, é fraudulenta ou legítima. No modelo de regressão logística, podemos considerar a função π da Equação (3.6) como sendo a função logística. Assim, para cada transação i podemos reescrever a Equação (3.6) da seguinte forma

$$P(Y_i = 1|\boldsymbol{x}_i) = \pi(\boldsymbol{x_i}) := \frac{e^{\beta_0 + \boldsymbol{x}_i\boldsymbol{\beta}^T}}{1 + e^{\beta_0 + \boldsymbol{x}_i\boldsymbol{\beta}^T}}, \tag{3.7}$$

em que $\beta_0 \in \mathbb{R}$ representa o intercepto, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p) \in \mathbb{R}^p$ é um vetor $1 \times p$ dos coeficientes associados a cada uma das covariáveis em estudo e $\boldsymbol{x}_i \in \mathbb{R}^p$ é um vetor $1 \times p$ que traz o valor observado em cada covariável da i-ésima transação. Note que usamos $\boldsymbol{u}^T$ para representar o vetor transposto de um vetor $\boldsymbol{u}$ qualquer.

É possível reescrever a Equação (3.7) na forma linear, aplicando uma transformação logarítmica, ou seja,

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

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

$$\frac{\pi(\boldsymbol{x_i})}{1 - \pi(\boldsymbol{x_i})}$$

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 transação ser fraudulenta.

Os coeficientes $\beta_0, \beta_1, \ldots, \beta_n$ são desconhecidos e precisam ser estimados a partir de um conjunto de treinamento adequado. A obtenção das estimativas desses parâmetros usualmente é realizada através da maximização da função de verossimilhança. Para aplicar este método é necessário assumir que (i) $Y_1, Y_2, \ldots, Y_n$ sejam variáveis aleatórias independentes e identicamente distribuídas (i.i.d.) e (ii) $Y_i|\boldsymbol{X}_i = \boldsymbol{x}_i \sim \text{Bernoulli}(\pi(\boldsymbol{x}_i))$, em que $\boldsymbol{X}_i$ é vetor de covariáveis associadas ao indivíduo i. Tais suposições permitem escrever a

distribuição de $Y_i|\boldsymbol{X_i} = \boldsymbol{x}_i$ da seguinte forma

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

em que $\mathbb{1}$ denota a função indicadora, por exemplo

$$\mathbb{1}_{\{0,1\}}(y_i) = \begin{cases} 1, \text{ se } y_i \in [0,1] \text{ e} \\ 0, \text{ caso contrário,} \end{cases}$$

que faz com que, para cada matriz $\boldsymbol{X}$, de dimensão $n \times p$, que representa as observações das p covariáveis nas n transações; a função de verossimilhança $L(\beta_0, \boldsymbol{\beta}|\boldsymbol{X}) : \mathbb{R}^{p+1} \to \mathbb{R}$ seja dada por

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

Tomando o logaritmo da igualdade, temos

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

Em termos computacionais, encontrar $\hat{\beta}_0$ e $\hat{\boldsymbol{\beta}}$ que maximiza a função ℓ é bem menos custoso do que encontrar $\hat{\beta}_0$ e $\hat{\boldsymbol{\beta}}$ que maximiza a função L. Todavia, os argumentos que maximizam ℓ e L são os mesmos, uma vez que o logaritmo é uma função contínua estritamente crescente. Nesse sentido, os estimadores de máxima verossimilhança são definidos como

$$(\hat{\beta}_0, \hat{\boldsymbol{\beta}}) := \underset{(\beta_0, \boldsymbol{\beta}) \in \mathbb{R}^{p+1}}{\operatorname{argmin}} \ -\ell\left(\beta_0, \boldsymbol{\beta}|\boldsymbol{X}\right). \tag{3.8}$$

Por não possuir forma analítica fechada, recorremos à métodos numéricos, como, por exemplo, o algoritmo de Newton-Raphson, para resolver esse problema de otimização. O artigo de Czepiel (2002) demonstra todos os passos da implementação desse método na regressão logística.

Uma vez que os parâmetros foram estimados, calculamos a probabilidade estimada de

uma transação ser fraudulenta dado qualquer vetor observado das covariáveis, i.e.,

$$\hat{\pi}(\boldsymbol{x}_i) = \frac{\exp(\hat{\beta}_0 + \boldsymbol{x}_i\hat{\boldsymbol{\beta}}^T)}{1 + \exp(\hat{\beta}_0 + \boldsymbol{x}_i\hat{\boldsymbol{\beta}}^T)}.$$

Assim, uma forma de realizar a classificação de uma nova transação j é tomar

$$Y_j = 1 \Leftrightarrow \hat{\pi}(\boldsymbol{x}_j) > c,$$

em que $c \in [0, 1]$ é uma constante a ser fixada.

Alguns estudos mostram que o modelo de regressão logística não apresenta boa performance em discriminar a classe minoritária quando a base de dados é desbalanceada, como é o caso, por exemplo, de bases obtidas para estudar fraudes em cartão de crédito. Nesses casos, duas alternativas são o modelo de regressão logística com LASSO como proposto por Wang *et al.* (2015) e o modelo de regressão logística limitada proposto por Cramer (2004).

3.2.2 Regressão logística com LASSO

Desenvolvido por Tibshirani (1996), o LASSO (do inglês *Least Absolute Shrinkage and Selection Operator)* ou regularização ℓ_1, é um método de regressão que realiza a seleção e a regularização das covariáveis para melhorar a precisão da predição e a interpretabilidade do modelo estatístico. O LASSO consiste em encontrar os coeficientes do Modelo (3.7) que resolva o problema de otimização em (3.8) com uma restrição sobre os coeficientes, dada a partir de uma medida de complexidade dos parâmetros, definida pela soma dos valores absolutos dos coeficientes multiplicada por um parâmetro de regularização, i.e.,

$$\lambda \sum_{j=1}^{p} |\beta_j|,$$

em que $\lambda \in \mathbb{R}_+^*$ é o regularizador escolhido a partir de validação cruzada. Detalhadamente, variamos λ em um intervalo de valores elegíveis, e verificamos sua performance com validação cruzada, que pode ser realizada a partir de dois principais métodos (Izbicki e dos Santos, 2020):

(i) *leave-one-out cross validation* que consiste em ajustar n classificadores utilizando todas as transações com exceção da i-ésima, obtendo o classificador $\hat{g}_{-i}$. A transação

i que ficou de fora é predita pelo classificador $\hat{g}_{-i}$. O valor de λ escolhido é o valor que otimiza a métrica utilizada considerando todas as amostras de validação. Nesse caso, quanto maior n, maior o custo computacional;

(ii) *k-fold cross validation* no qual dividimos as transações em $k \in \mathbb{N}^*$ lotes disjuntos de tamanho uniforme e ajustamos k classificadores (o número k de lotes não possui nenhum vínculo com o número k de *clusters)* deixando o k-ésimo lote de fora, obtendo o classificador $\hat{g}_{-k}$. Similar ao caso anterior, o lote k que não compõe o classificador $\hat{g}_{-k}$ é predito pelo mesmo. O valor de λ escolhido é o valor que otimiza a métrica utilizada considerando todas as amostras de validação. Repare que o *leave-one-out cross validation* é um caso especial do *k-fold cross validation* em que $k = n$.

Adotar esse procedimento para a escolha de λ é importante para evitar *overfitting*, que ocorre quando o λ escolhido é muito baixo, fazendo com que o classificador tenha um desempenho excelente em indivíduos específicos, no caso aqueles que compõem o conjunto de treinamento, perdendo poder de generalização; e *underfitting*, que ocorre quando o λ escolhido é muito alto, restringindo demasiadamente os coeficientes, o que faz o classificador ter um desempenho ruim mesmo nas transações que compõem o conjunto de treinamento.

No caso da regressão logística com LASSO, as estimativas de parâmetros terão que considerar na Equação (3.8) a medida de complexidade de $\boldsymbol{\beta}$, em outras palavras,

$$(\hat{\beta}_0, \hat{\boldsymbol{\beta}}) := \underset{(\beta_0, \boldsymbol{\beta}) \in \mathbb{R}^{p+1}}{\operatorname{argmin}} \left\{ -\ell\left(\beta_0, \boldsymbol{\beta} | \boldsymbol{X}\right) + \lambda \sum_{j=1}^{p} |\beta_j| \right\}. \tag{3.9}$$

Mais uma vez é necessário recorrer à métodos numéricos para encontrar a estimativa dos coeficientes. Devido à sua grande utilidade, a solução do problema de otimização em (3.9) está implementada nos *softwares* estatísticos que serão utilizados neste trabalho.

Analogamente ao que ocorre no caso da regressão logística sem regularização, uma vez que os parâmetros são estimados, é fácil calcular a probabilidade de uma transação ser fraudulenta dada qualquer observação das covariáveis e, então, utilizar essa estimativa para classificar as transações do conjunto de validação.

3.2.3 Regressão logística limitada

Moraes (2008) propõe utilizar como classificador base a regressão logística limitada, pois o modelo logístico não-limitado não apresenta boa performance quando a variável resposta é extremamente desbalanceada (Cramer, 2004). Em sua dissertação de mestrado, Moraes (2008) apresenta um estudo comparativo entre as formas limitada e não-limitada do modelo de regressão logístico e conclui, de acordo com certas métricas para comparação de modelos, que a forma limitada do modelo é um classificador melhor do que a sua forma não-limitada.

A regressão logística limitada é muito semelhante à versão usual, com uma pequena modificação: há um limite superior para a probabilidade de sucesso, dado por $\omega \in [0,1]$. Nesse caso, podemos reescrever a Equação (3.6) que define o modelo da seguinte maneira

$$P(Y_i = 1|\boldsymbol{x}_i) := \omega \frac{e^{\beta_0 + \boldsymbol{x}_i\boldsymbol{\beta}^T}}{1 + e^{\beta_0 + \boldsymbol{x}_i\boldsymbol{\beta}^T}} \mathbb{1}_{[0,1]}(\omega),$$

em que $\omega \in [0,1]$ delimita a probabilidade de sucesso, $\beta_0 \in \mathbb{R}$ representa o intercepto, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p) \in \mathbb{R}^p$ é um vetor $1 \times p$ dos coeficientes associados a cada uma das covariáveis em estudo e $\boldsymbol{x}_i \in \mathbb{R}^p$ é um vetor $1 \times p$ que traz o valor observado em cada covariável da i-ésima transação.

Como na versão não-limitada, abordada na Seção 3.2.1, a obtenção das estimativas dos parâmetros pode ser realizada através da maximização da função de verossimilhança $L(\beta_0, \boldsymbol{\beta}, \omega|\boldsymbol{X}) : \mathbb{R}^{p+1} \times [0,1] \to \mathbb{R}$, definida por

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

Novamente como na versão não-limitada, uma alternativa computacionalmente mais viável é encontrar $\hat{\beta}_0, \hat{\boldsymbol{\beta}}$ e $\hat{\omega}$ que minimiza a negativa da função log-verossimilhança, ou seja,

$$(\hat{\beta}_0, \hat{\boldsymbol{\beta}}, \hat{\omega}) := \underset{\substack{(\beta_0, \boldsymbol{\beta}) \in \mathbb{R}^{p+1} \\ \omega \in [0,1]}}{\operatorname{argmin}} \; -\ell(\beta_0, \boldsymbol{\beta}, \omega|\boldsymbol{X}), \tag{3.10}$$

em que

$$\ell\left(\beta_0, \boldsymbol{\beta}, \omega | \boldsymbol{X}\right) := \log(L(\beta_0, \boldsymbol{\beta}, \omega | \boldsymbol{X})) \\ = \sum_{i=1}^{n} \left[y_i \log\left(\omega \frac{e^{\beta_0 + \boldsymbol{x}_i \boldsymbol{\beta}^T}}{1 + e^{\beta_0 + \boldsymbol{x}_i \boldsymbol{\beta}^T}} \right) + (1 - y_i) \log\left(1 - \omega \frac{e^{\beta_0 + \boldsymbol{x}_i \boldsymbol{\beta}^T}}{1 + e^{\beta_0 + \boldsymbol{x}_i \boldsymbol{\beta}^T}} \right) \mathbb{1}_{\{0,1\}}(y_i) \right].$$

Pelo fato da Equação (3.10) não possuir solução analítica, e não estar implementada nos *softwares* a serem utilizados, iremos recorrer à métodos numéricos como, por exemplo, o método iterativo de Newton-Raphson, para a obtenção das estimativas dos parâmetros, sendo essa uma das contribuições do presente estudo.

Após a estimação dos parâmetros, podemos realizar a classificação das transações em fraudulentas ou legítimas da mesma maneira que fizemos nas duas seções anteriores.

3.2.4 Regressão logística limitada com LASSO

Até onde vai nosso conhecimento, a regressão logística limitada com LASSO ainda não foi abordada pela literatura. Nossa motivação em utilizá-la baseia-se na combinação de dois pontos:

1. A excelente performance da regressão logística limitada em classes extremamente desbalanceadas observada por Cramer (2004) e os resultados obtidos por Moraes (2008), indicando melhor performance da versão limitada quando comparada com a versão usual.

2. As vantagens da regularização ℓ_1, que evita *overfitting* e realiza a seleção de variáveis, podendo resultar em maior poder preditivo. Além disso, podemos citar os resultados obtidos por Wang *et al.* (2015), que utilizam um classificador baseado na regressão logística com LASSO construído a partir de um conjunto de treinamento diversificado e balanceado através dos métodos de diversificação e balanceamento apresentados na Seção 3.1. Tais resultados mostram maior poder discriminatório desse classificador quando comparado a demais métodos como árvore de classificação e *random forests*. É importante salientar que o estudo realizado por Wang *et al.* (2015) também considerou um conjunto de dados com classes desbalanceadas, porém em problemas de *credit scoring*.

De forma análoga ao que fizemos no caso da regressão logística usual com LASSO, na estimação de parâmetros da regressão logística limitada, adicionamos a medida de

complexidade de $\boldsymbol{\beta}$, que é definida pela soma

$$\lambda \sum_{j=1}^{p} |\beta_j|,$$

em que $\lambda > 0$ é um parâmetro escolhido a partir de validação cruzada. Em outras palavras, queremos encontrar

$$(\hat{\beta}_0, \hat{\boldsymbol{\beta}}, \hat{\omega}) := \underset{\substack{(\beta_0, \boldsymbol{\beta}) \in \mathbb{R}^{p+1} \\ \omega \in [0,1]}}{\operatorname{argmin}} \left\{ -\ell(\beta_0, \boldsymbol{\beta}, \omega | \boldsymbol{X}) + \lambda \sum_{j=1}^{p} |\beta_j| \right\}. \tag{3.11}$$

Assim como ocorre no problema de otimização em (3.10), que não possui solução analítica, para encontrar a solução de (3.11) vamos recorrer à métodos numéricos, tais como o já citado método de Newton-Raphson. Uma vez que os parâmetros estão estimados, facilmente conseguimos classificar uma nova transação a partir da comparação da probabilidade estimada condicional da transação ser fraudulenta dada a observação das covariáveis com um ponto de corte. A implementação computacional e o estudo da performance desse classificador é uma segunda contribuição desta monografia.

3.3 Medidas de performance

No presente estudo, iremos construir diferentes classificadores, com base na regressão logística e suas variações. Com os classificadores construídos, vamos compará-los, a fim de determinar qual deles possui melhor desempenho. Para isso, nos basearemos em algumas medidas de performance utilizadas nos estudos que inspiraram este trabalho. Em especial, utilizaremos o KS, utilizado em Moraes (2008), AUC e estatística F1, que foram utilizadas por Wang *et al.* (2015) e também a acurácia e soma dos quadrados dos resíduos. Ressaltamos que para essas medidas, não é necessário a fixação de um único ponto de corte em $P(Y_i = 1|\boldsymbol{x_i})$.

3.3.1 Estatística de Kolmogorov-Smirnov

Uma medida muito utilizada no mercado de crédito é a estatística de Kolmogorov-Smirnov, ou simplesmente KS (Sicsú, 2010). Sua aplicação pode ser estendida no cenário de fraude, medindo o quanto estão separadas as funções de distribuições empíricas da pontuação das transações legítimas e fraudulentas. Sendo assim, definimos $F_f : [0, 1] \to [0, 1]$

como sendo a distribuição empírica das transações fraudulentas e de forma semelhante $F_\ell : [0,1] \to [0,1]$ representando a distribuição empírica das transações legítimas. A estatística KS é, então, dada pela distância absoluta máxima entre as proporções acumuladas ao longo das estimativas obtidas pelo classificador, ou seja:

$$KS = \max_{e \in [0,1]} |F_f(e) - F_\ell(e)|.$$

Em geral, quanto maior o valor da estatística KS (que pode variar entre 0 e 1), melhor a performance do classificador, pois isso reflete a distância entre as distribuições empíricas das classes.

3.3.2 AUC

A medida AUC (do inglês *Area Under Curve*) é derivada da curva ROC (do inglês *Receiver Operating Characteristic*). Na prática, é baseada em duas quantidades bem conhecidas: especificidade e sensibilidade. A seguir, definimos como calcular essas quantidades. Sejam

- FN - Falso negativo (quantidade de transações fraudulentas classificadas como legítimas);
- FP - Falso positivo (quantidade de transações legítimas classificadas como fraudulentas);
- VN - Verdadeiro negativo (quantidade de transações legítimas classificadas como legítimas);
- VP - Verdadeiro positivo (quantidade de transações fraudulentas classificadas como fraudulentas).

Assim,

- Especificidade = $\frac{\text{VN}}{\text{VN+FP}}$ (proporção de transações classificadas como legítimas dado que ela realmente é legítima);
- Sensibilidade = $\frac{\text{VP}}{\text{VP+FN}}$ (proporção de transações classificadas como fraudulentas dado que ela realmente é fraudulenta).

A curva ROC varia o ponto de corte em toda a amplitude das estimativas do modelo, calculando a especificidade e sensibilidade em todo esse intervalo. É considerado no eixo

da abscissa a métrica (1 - especificidade) e no eixo da ordenada apenas a sensibilidade. A AUC é justamente a área abaixo dessa curva. Quanto mais próxima de 1, melhor a performance da previsão.

3.3.3 Estatística F1

A estatística F1, também conhecida como medida-F ou *F-score*, é muito utilizada e aceita na literatura de aprendizagem de máquina. Trata-se de uma média harmônica de duas medidas: precisão e sensibilidade. A sensibilidade já foi definida na Subseção 3.3.2 e, utilizando as quantidades lá definidas, podemos escrever

$$\text{Precisão} = \frac{\text{VP}}{\text{VP+FP}},$$

i.e., das transações classificadas como fraudulentas, a proporção das que realmente são fraudulentas. Com isso, a estatística F1 é definida como

$$\text{F1} := \frac{2}{\text{precisão}^{-1} + \text{sensibilidade}^{-1}}.$$

Ressaltamos que as medidas precisão e sensibilidade dependem das quantidades FN, FP e VP, que por sua vez, necessitam de um ponto de corte para classificar as transações, com base em $\hat{\pi}$. Neste estudo, vamos calcular a estatística F1 com diferentes pontos de corte o valor de c, varrendo todo o intervalo $[0, 1]$. Usaremos como corte para classificação c que maximiza a estatística F1. Quanto maior o valor da estatística F1, melhor será a performance do classificador, por trazer um equilíbrio entre a precisão e sensibilidade, o que estará diretamente ligado à um baixo número de FN e FP.

3.3.4 Acurácia

A acurácia é uma medida que retrata a qualidade geral das classificações do modelo, resumindo o quão próximo predições estão da verdadeira classe da observação. É definida como

$$\text{Acurácia} := \frac{\text{VP} + \text{VN}}{(\text{VP} + \text{FP} + \text{VN} + \text{FN})}.$$

Por depender das classificações, necessitamos de um ponto de corte. Por questões de conveniência, usaremos o mesmo corte que maximizou a estatística F1.

É razoável relembrar que, no caso de conjuntos desbalanceados, uma alta acurácia não quer dizer que o modelo seja bom. Por exemplo, se tivéssemos que 99% das transações fossem legítimas e apenas 1% fraudulentas, e classificarmos todas como legítimas, teríamos uma alta acurácia (os mesmos 99%), porém, um péssimo modelo, que não barraria fraude.

3.3.5 Soma dos Quadrados dos Resíduos

A Soma dos Quadrados dos Resíduos (SQRes) é responsável por retratar a variabilidade da variável dependente que não foi explicada pelo conjunto de variáveis independentes. Em termos práticos, representa as distâncias quadráticas entre os valores observados de Y e seus valores ajustados pelo modelo, escrita como

$$\text{SQRes} = \sum_{i=1}^{n} (Y_i - \hat{\pi}(\boldsymbol{x}_i))^2.$$

Capítulo 4

Aplicação em dados reais

O conjunto de dados a ser utilizado nesta etapa foi disponibilizado por uma instituição financeira, não sendo possível identificar os clientes, servindo apenas para fins de pesquisa e estudos. Por motivos de confidencialidade, as covariáveis serão tratadas com nomes fictícios.

4.1 Descrição da base

O conjunto em questão possui 59 covariáveis, além da variável resposta, indicando fraude ou não. No total, há 68.525 transações, em que 1.804 delas são fraudes (2,63%). Ressaltamos que nas covariáveis contínuas foi aplicada a transformação logarítmica, com intuito de preservar a confidencialidade dos dados. Em relação às covariáveis categóricas, os níveis foram ordenados em ordem crescente em relação ao risco de fraude, isto é, o nível 0 da categoria concentra um menor percentual de fraude que o nível 1 e assim sucessivamente.

4.2 Análise descritiva

Com o auxílio do *software* R, realizamos uma análise descritiva, com a finalidade de observarmos a distribuição de algumas covariáveis em relação à fraude. Trazemos as covariáveis que mais se destacam no sentido de discriminar os tipos de transações.

As Tabelas 4.1 e 4.2 trazem os comportamentos por categoria das covariáveis V1 e V55, respectivamente. Em ambas, percebe-se um menor percentual de fraude na categoria C0. A efeito de comparação, a categoria C2 da covariável V1 possui um percentual de

fraude 11 vezes maior que a categoria C0, enquanto na categoria C2 da covariável V55 temos o dobro de fraude percentual quando comparamos à categoria C0.

Tabela 4.1: Distribuição da covariável V1 em relação ao percentual de fraude.

V1	Transações	% Fraudes
C0	7.629	0,33%
C1	31.196	2,10%
C2	29.700	3,78%

Tabela 4.2: Distribuição da covariável V55 em relação ao percentual de fraude.

V55	Transações	% Fraudes
C0	44.875	2,03%
C1	10.121	2,75%
C2	13.529	4,54%

Em relação as covariáveis contínuas, destacamos V7, V8 e V54. Na Figura 4.1 vemos que a mediana e o terceiro quartil da variável V7 possuem valores bem diferentes de acordo com o tipo de transação, tendo as fraudes maiores valores nessas medidas. Enquanto isso, na Figura 4.2 notamos praticamente os mesmos mínimos e máximos, mas com diferença em relação à mediana, tendo as fraudes maior mediana nessa covariável. Por fim, na Figura 4.3 a maior diferença entre os tipos de transações fica a cargo do terceiro quartil, sendo novamente superior entre as fraudes.

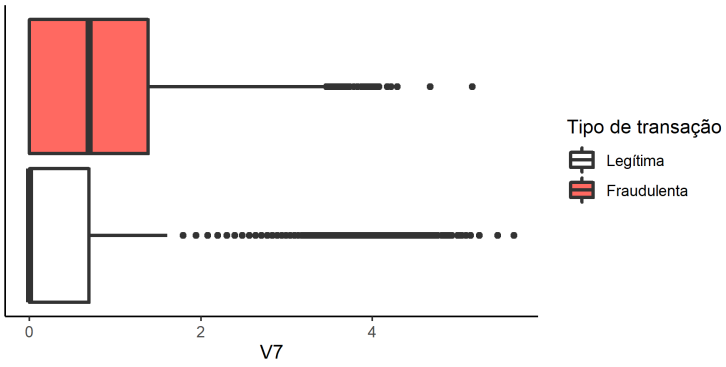

Figura 4.1: Boxplot da covariável V7 de acordo com o tipo de transação.

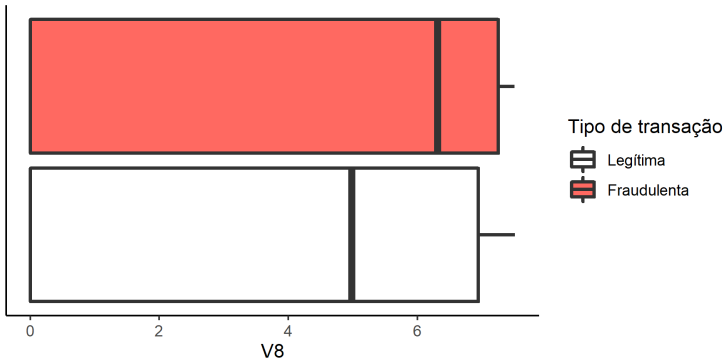

Figura 4.2: Boxplot da covariável V8 de acordo com o tipo de transação.

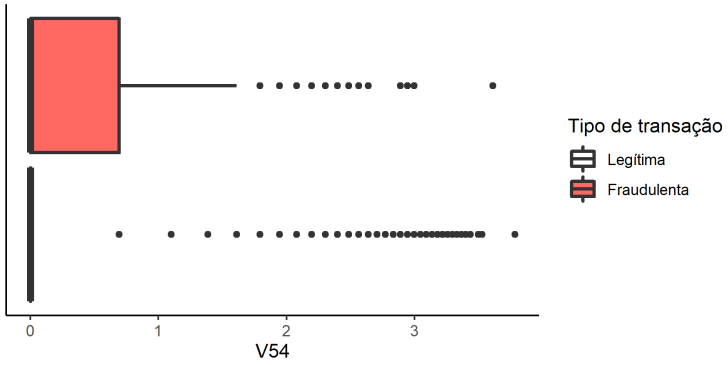

Figura 4.3: Boxplot da covariável V54 de acordo com o tipo de transação.

4.3 Performance dos classificadores

Nesta seção, as metodologias propostas foram aplicadas ao conjunto de dados. Inicialmente, adotamos a configuração natural dos dados, ou seja, com o desbalanceamento. Serão ajustados todos os modelos propostos, e comparados sob diferentes óticas. Posteriormente, será feito o balanceamento e diversificação do conjunto, e mais uma vez as metodologias serão aplicadas.

Os dados serão distribuídos da seguinte forma:

- 70% irão compor o conjunto de treinamento, sob o qual todos os modelos serão ajustados;

- Para o cenário com balanceamento e diversificação, dentro do conjunto de treinamento é feita uma nova separação com 30% (dos 70% iniciais) compondo um conjunto de validação, que auxiliará a encontrar o k ótimo (quantidade de subconjuntos de treinamentos) do balanceamento e diversificação. Ressaltamos que, após

encontrar esse k ótimo, o modelo é reajustado, dessa vez com todos os dados do conjunto de treinamento.

- Os 30% restantes, que não compõem o conjunto de treinamento (e, consequentemente, nem o conjunto de validação) serão o conjunto de teste, no qual a performance dos modelos construídos será analisada.

4.3.1 Conjunto de dados desbalanceado

Como dito anteriormente, os primeiros modelos serão construídos utilizando um conjunto de treinamento desbalanceado. As primeiras técnicas aplicadas foram a regressão logística e a regressão logística limitada.

Para otimizar as funções objetivo, no caso da regressão logística, definida na Equação (3.8), utilizamos a função *glm* do *software R*. Na regressão logística limitada, representada na Equação (3.10), não encontramos nenhuma função pré-configurada. Recorremos à função *optimize* do pacote *scipy* do *software Python* (Scipy, 2022), onde é passado a função objetivo, as restrições dos parâmetros e um chute inicial. Dentro da função *optimize*, existem uma série de métodos de otimização. Escolhemos pelo *L-BFGS-B* (Zhu *et al.*, 1997), pelo fato de ser utilizado por outros algoritmos de aprendizado de máquina.

Na Tabela 4.3 trazemos os resultados das performances de ambos os modelos no conjunto de teste, e nas Tabelas 4.4 e 4.5 as classificações realizadas por cada modelo, com o corte que maximizou a estatística F1 em cada um deles.

Tabela 4.3: Comparativo da performance entre a regressão logística e a regressão logística limitada no conjunto de dados desbalanceado.

Método	KS	AUC	F1	Acurácia	SQRes
Regressão Logística Limitada	0,3143	0,7006	0,1044	0,8807	521,7555
Regressão Logística	0,3130	0,6991	0,1074	0,8820	522,4160

Tabela 4.4: Matriz de confusão das classificações obtidas pela regressão logística no conjunto de dados desbalanceado, com ponto de corte em 0,0512.

		Verdadeira Classe		
		Legítima	Fraudulenta	Total
Classe predita	Legítima	17.985 (89,86%)	397 (73,11%)	18.382
	Fraudulenta	2.029 (10,14%)	146 (26,89%)	2.175
	Total	20.014	543	20.557

Tabela 4.5: Matriz de confusão das classificações obtidas pela regressão logística limitada no conjunto de dados desbalanceado, com ponto de corte em 0,0540.

		Verdadeira Classe		
		Legítima	Fraudulenta	Total
Classe predita	Legítima	17.961 (89,74%)	400 (73,66%)	18.361
	Fraudulenta	2.053 (10,26%)	143 (26,34%)	2.196
	Total	20.014	543	20.557

Note que em 3 das 5 métricas utilizadas, a regressão logística limitada possui performance ligeiramente melhor, justamente nas métricas que não dependem de um ponto de corte. Em relação às classificações realizadas, a regressão logística possui leve vantagem, com maior estatística F1 e acurácia, classificando corretamente um maior número de fraudes.

Salientamos que na Tabela A.1 trazemos um comparativo dos coeficientes obtidos, e as respectivas diferenças entre os métodos.

Os próximos modelos a serem ajustados são aqueles com a regularização ℓ_1 dos coeficientes. Vale lembrar que para realizar o LASSO, as covariáveis precisam estar normalizadas (Tibshirani, 1996).

Explicando um pouco sobre a obtenção dos parâmetros, no caso da regressão logística com LASSO, definida na Equação (3.9), utilizamos a função *glmnet* do pacote *glmnet* do *software R* (Hastie e Qian, 2014). Na regressão logística limitada com LASSO, definida na Equação (3.11), mais uma vez não encontramos nenhuma função pré-configurada, recorrendo à função *optimize* do pacote *scipy* do *software Python* com o método *L-BFGS-B*. Destacamos que essa solução da limitada não é otimizada em questão de tempo, i.e., dependendo o tamanho do conjunto de dados a estimação dos parâmetros pode levar vários minutos (ou até mesmo horas).

Lembramos que nesse tipo de modelo, o parâmetro λ (que é responsável pela regularização), deve ser fixado pelo usuário. Entretanto, uma escolha aleatória pode não levar a bons resultados. Como citado anteriormente, nesse estudo a escolha do λ se dá pelo *k-fold cross validation*, utilizando 5 lotes, sendo a AUC a métrica de referência. Para a regressão logística com LASSO, tal processo é facilmente executado pela função *cv.glmnet* do pacote *glmnet* do *software R*, que varre um intervalo de escolhas razoáveis para λ. Elencamos na Tabela 4.6 os três melhores e os três piores valores de λ testados em termos da AUC. Recordamos que a AUC expressa na Tabela 4.6 é uma média das AUC obtidas em cada um dos 5 lotes de validação.

Tabela 4.6: Amostra de alguns λ testados no LASSO e suas respectivas medidas AUC.

λ	AUC
0,00056	0,69717
0,00062	0,69714
0,00068	0,69712
0,00839	0,64711
0,00921	0,64658
0,01011	0,56137

Pela dificuldade (em questão de tempo até a convergência) de estimar os coeficientes da regressão logística limitada, optamos por não realizar a validação cruzada na regressão logística limitada com LASSO, utilizando o melhor λ da regressão logística com LASSO, no caso $\lambda \approx 0,00056$. Com isso, temos que em ambos os modelos, usaremos $\lambda \approx 0,00056$ e o próximo passo é a obtenção dos parâmetros de cada modelo.

Na Tabela 4.7 trazemos os resultados das performances de ambos os modelos no conjunto de teste, e como no caso anterior, a matriz de confusão de cada classificador, representadas nas Tabelas 4.8 e 4.9, com o corte que maximizou a estatística F1 em cada um dos modelos ajustados.

Tabela 4.7: Comparativo da performance entre os métodos com LASSO no conjunto de dados desbalanceado.

Método	KS	AUC	F1	Acurácia	SQRes
Regressão Logística Limitada com LASSO	0,3218	0,7004	0,1016	0,8701	521,8042
Regressão Logística com LASSO	0,3064	0,7075	0,1112	0,8920	521,4327

Tabela 4.8: Matriz de confusão das classificações obtidas pela regressão logística com LASSO no conjunto de dados desbalanceado, com ponto de corte em 0,0498.

		Verdadeira Classe		
		Legítima	Fraudulenta	Total
Classe predita	Legítima	18.197 (90,92%)	404 (74,40%)	18.601
	Fraudulenta	1.817 (9,08%)	139 (25,60%)	1.956
	Total	20.014	543	20.557

Tabela 4.9: Matriz de confusão das classificações obtidas pela regressão logística limitada com LASSO no conjunto de dados desbalanceado, com ponto de corte em 0,0532.

		Verdadeira Classe		
		Legítima	Fraudulenta	Total
Classe predita	Legítima	17.736 (88,62%)	392 (72,19%)	18.128
	Fraudulenta	2.278 (11,38%)	151 (27,81%)	2.429
	Total	20.014	543	20.557

Diferente do caso anterior, a regressão logística limitada apresentou melhor performance apenas no KS, sendo pior nas demais métricas. Frisamos que em termos da estatística F1 e Acurácia, a logística com LASSO apresentou resultados melhores frente à versão limitada. Isso fica nítido quando comparamos a Tabela 4.8 com a Tabela 4.9, pois a regressão logística com LASSO classifica um menor número de transações como fraudulentas, sem perder tanta precisão, apresentando o menor número de falsos positivos e o maior número de verdadeiros negativos. Entretanto, a afirmação deve ser encarada com cuidado, pois utilizamos o mesmo λ em ambos os métodos, sendo o ótimo para a versão não limitada mas não necessariamente o melhor para a versão limitada.

Nas Tabelas A.2, A.3 e A.4 comparamos os coeficientes obtidos, e as respectivas diferenças entre os métodos. Como nosso objetivo está na performance dos modelos ajustado, não iremos aprofundar na análise de suas covariáveis nesse estudo.

4.3.2 Conjunto de dados balanceado

Nessa segunda etapa do estudo, iremos aplicar as técnicas de balanceamento e diversificação que foram detalhadas ao longo da monografia. Em todos os casos, o balanceamento é feito utilizando o *k-means++* (com distância euclidiana), por ter apresentado melhor resultado em alguns testes iniciais. O primeiro passo é realizar a escolha do número de subconjuntos de treinamentos, que foi denotado por k.

Para tal processo, utilizamos o conjunto de validação citado no início da seção, que em linhas gerais, emula um conjunto de teste. Com o restante do conjunto de treinamento (o que equivale à 49% do conjunto de dados, i.e. 70% de 70%), calculamos os modelos com o k proposto, e mensuramos sua performance no conjunto de validação. Testamos k variando de 2 até 14 (a partir de 15, o algoritmo do *k-means++* começava a divergir), e o ponderamento dos α_l (definido na Equação (3.5)) sendo feito pelo KS e AUC.

Iniciando pela busca do k a ser empregado na regressão logística e regressão logística limitada, salientamos que o processo de validação foi feito exclusivamente na regressão

logística (por possuir uma função de estimação dos parâmetros muito mais eficiente, em termos de tempo, que a *optimize* utilizada na versão limitada), e com isso, o valor de k encontrado como o melhor na regressão logística será replicado para a sua versão limitada.

A Tabela 4.10 traz os resultados das performances obtidas pelos k modelos construídos, na base de validação. Observe que neste cenário, a melhor opção aparenta ser $k = 8$, com o ponderamento dos α_l realizado através da estatística KS. É interessante realçar que para todos os valores de k testados, a performance foi melhor quando se utilizou a estatística KS como entrada para o cálculo de α_l.

Tabela 4.10: Performance dos classificadores de regressão logística combinados na base de validação, ordenados pela soma simples das métricas utilizadas.

k	Ponderação	KS	AUC
8	Via KS	0,2465	0,6402
10	Via KS	0,2434	0,6408
7	Via KS	0,2453	0,6307
13	Via KS	0,2319	0,6365
14	Via KS	0,2205	0,6402
9	Via KS	0,2213	0,6382
6	Via KS	0,2217	0,6272
5	Via KS	0,2184	0,6298
12	Via KS	0,2147	0,6315
2	Via KS	0,2038	0,6158
11	Via KS	0,1945	0,6165
3	Via KS	0,2034	0,6064
4	Via KS	0,1419	0,5773

Dito isso, iremos construir 8 submodelos de regressão logística (e posteriormente de regressão logística limitada), cada um associado à um subconjunto de treinamento (que agora englobam a parcela que antes integrava o conjunto de validação), agregando-os conforme a Equação (3.5) e avaliaremos a performance no conjunto de teste.

Como feito na etapa anterior, na Tabela 4.11 apresentamos os resultados das performances de ambos os modelos no conjunto de teste, e a matriz de confusão de cada classificador, representadas nas Tabelas 4.12 e 4.13, com o corte que maximizou a estatística F1 em cada um dos modelos ajustados.

Tabela 4.11: Comparativo da performance entre a regressão logística e a regressão logística limitada com balanceamento e diversificação.

Método	KS	AUC	F1	Acurácia	SQRes
Regressão Logística	0,2155	0,6210	0,0761	0,6942	11668,2966
Regressão Logística Limitada	0,2154	0,6210	0,0761	0,6942	11668,0866

Tabela 4.12: Matriz de confusão das classificações obtidas pela regressão logística com balanceamento e diversificação, com ponto de corte em 0,8190.

		Verdadeira Classe		
		Legítima	Fraudulenta	Total
Classe predita	Legítima	14.012 (70,01%)	284 (52,30%)	14.296
	Fraudulenta	6.002 (29,99%)	259 (47,70%)	6.261
	Total	20.014	543	20.557

Tabela 4.13: Matriz de confusão das classificações obtidas pela regressão logística limitada com balanceamento e diversificação, com ponto de corte em 0,8190.

		Verdadeira Classe		
		Legítima	Fraudulenta	Total
Classe predita	Legítima	14.011 (70,01%)	284 (52,30%)	14.295
	Fraudulenta	6.003 (29,99%)	259 (47,70%)	6.262
	Total	20.014	543	20.557

De início, vemos uma piora quando comparamos com os modelos construídos sem o balanceamento e diversificação. Entretanto, analisando apenas os que foram construídos neste passo, existe uma enorme semelhança entre as performances dos métodos. Tal semelhança pode ser justificada quando observamos as estimativas dos coeficientes na regressão logística limitada, percebendo que a média dos $\hat{\omega}$ é de 0,9999. Isso pode ser reflexo do balanceamento do conjunto, fazendo com que a limitação não seja necessária. Chama atenção também o fato de que um número muito maior de transações foram classificadas como fraude, quando comparamos com os modelos feitos na base desbalanceada, vide as Tabelas 4.12 e 4.13.

Os próximos, e desta vez últimos, modelos a serem ajustados também serão no conjunto de dados balanceado e diversificado, usando os métodos com regularização LASSO. O itinerário será o mesmo que ao adotado para os métodos sem o LASSO, realizando a escolha do k ótimo para esses modelos conforme a performance no conjunto de validação, e posteriormente construindo o modelo final com todo o conjunto de treinamento. Novamente, a busca por k ocorrerá apenas no modelo não limitado, mais uma vez por questões de eficiência. As únicas mudanças que valem ser ressaltadas, é que por requisitos

do LASSO, as covariáveis estarão normalizadas, e que em cada "submodelo" ocorrerá o *k-fold cross validation*, utilizando 5 lotes, sendo a AUC a métrica de referência para a escolha de λ.

A Tabela 4.14 traz os resultados das performances obtidas pelos k modelos construídos, utilizando a regressão logística com LASSO, na base de validação. Observe que neste cenário, diferentemente do anterior, a melhor opção aparenta ser quando $k = 2$, com o ponderamento dos α_l realizado através pela estatística KS.

Tabela 4.14: Performance dos classificadores com LASSO combinados na base de validação, ordenados pela soma simples das métricas utilizadas.

k	Ponderação	KS	AUC
2	Via KS	0,2161	0,6303
3	Via KS	0,2190	0,6194
4	Via KS	0,2035	0,6107
8	Via KS	0,1835	0,6091
7	Via KS	0,1746	0,5975
13	Via KS	0,1705	0,5936
14	Via KS	0,1754	0,5879
10	Via KS	0,1653	0,5975
11	Via KS	0,1284	0,5877
12	Via KS	0,1310	0,5806
6	Via KS	0,1317	0,5746
5	Via KS	0,1260	0,5741
9	Via KS	0,1228	0,5750

Dessa maneira serão construídos 2 submodelos de regressão logística com LASSO (e posteriormente de regressão logística limitada com LASSO, utilizando o mesmo λ da versão não limitada, como na etapa do conjunto desbalanceado), cada um associado à um subconjunto de treinamento (que novamente englobará a parcela que antes integrava o conjunto de validação), agregando-os conforme a Equação (3.5) e avaliaremos a performance no conjunto de teste.

Na Tabela 4.15 trazemos os resultados das performances de ambos os modelos no conjunto de teste, e a matriz de confusão de cada classificador, representadas nas Tabelas 4.16 e 4.17, com o corte que maximizou a estatística F1 em cada um dos modelos ajustados.

Tabela 4.15: Comparativo da performance entre os métodos com LASSO, balanceamento e diversificação.

Método	KS	AUC	F1	Acurácia	SQRes
Regressão Logística com LASSO	0,2895	0,6737	0,1048	0,8704	3778,6470
Regressão Logística Limitada com LASSO	0,2697	0,6483	0,0830	0,6549	1937,9050

Tabela 4.16: Matriz de confusão das classificações obtidas pela regressão logística com LASSO, balanceamento e diversificação, com ponto de corte em 0,5065.

		Verdadeira Classe		
		Legítima	Fraudulenta	Total
Classe predita	Legítima	17.736 (88,62%)	387 (71,27%)	18.123
	Fraudulenta	2.278 (11,38%)	156 (28,73%)	2.434
	Total	20.014	543	20.557

Tabela 4.17: Matriz de confusão das classificações obtidas pela regressão logística limitada com LASSO, balanceamento e diversificação, com ponto de corte em 0,2295.

		Verdadeira Classe		
		Legítima	Fraudulenta	Total
Classe predita	Legítima	13.141 (65,66%)	222 (40,88%)	13.363
	Fraudulenta	6.873 (34,34%)	321 (59,12%)	7.194
	Total	20.014	543	20.557

Sumarizando os resultados desses modelos, observamos um ganho considerável da versão não limitada em 4 das 5 métricas. A única métrica em que a regressão logística limitada com LASSO se destaca, nesse cenário em que o conjunto foi balanceado e diversificado, é na SQRes, em função do seu parâmetro ω (fazendo com que as probabilidades fiquem mais próximas de 0, que configura a classe predominante do conjunto). Em relação às demais métricas e comparando com todos os modelos construídos, vemos novamente uma piora em relação ao conjunto desbalanceado.

4.4 Discussão

Neste capítulo, realizamos uma aplicação das metodologias abordadas em um conjunto de dados reais, proporcionando diversas comparações entre os métodos. Inicialmente, podemos destacar que o balanceamento e a diversificação não resultaram em ganho de performance neste estudo de caso, já que todos os modelos construídos tiveram performance inferior aos construídos no conjunto desbalanceado.

Ao conduzirmos as comparações em blocos, da versão limitada contra a versão usual, observamos o maior ganho da limitação no conjunto desbalanceado sem a regularização

pelo LASSO. No caso do conjunto balanceado, os resultados foram praticamente os mesmos. Quando adicionamos a regularização com LASSO, a versão usual apresentou melhor performance em ambos os cenários, por possuir melhor métrica em 4 das 5 avaliadas em cada configuração. Mais uma vez, devemos analisar com cuidado pelo fato de utilizar o λ ótimo para a versão usual.

De forma global, na forma em que a aplicação foi conduzida, o melhor modelo apresentado aparenta ser a regressão logística com LASSO, construída sobre o conjunto de dados desbalanceado, por possuir a melhor métrica em 4 de 5 avaliadas (AUC, F1, Acurácia e SRQes). O maior KS observado foi no modelo de regressão logística limitada com LASSO.

Capítulo 5

Considerações Finais

Neste trabalho, estudamos sobre o mercado de cartões de crédito e a ocorrência de fraude nesse meio, assim como seus impactos. Por trazer uma série de prejuízos financeiros e sociais, é de extrema importância o uso de bons modelos estatísticos, para mitigar seus danos. Nesse sentido, propomos diferentes metodologias que apresentaram resultados interessantes em estudos anteriores, sendo elas a regularização LASSO, a regressão logística limitada e um processo de balanceamento e diversificação do conjunto de dados. Finalmente propomos um novo método, a regressão logística limitada com LASSO.

A fim de verificar o desempenho das técnicas estudadas, conduzimos um estudo de caso sobre o conjunto de dados disponibilizado por uma instituição financeira, não sendo possível identificar os clientes (estando de acordo com a Lei Geral de Proteção a Dados Pessoais).

Todos os modelos foram construídos no conjunto de treinamento, que corresponde à 70% dos dados, e avaliados no conjunto de teste (com os 30% restantes). Dividimos as comparações em duas frentes, uma com o conjunto de treinamento desbalanceado, i.e., em sua forma original e a outra com tratamento de balanceamento e diversificação.

Em linhas gerais, os modelos construídos sobre o conjunto desbalanceado apresentou performance superior (de acordo com as métricas utilizadas) aos que foram construídos no conjunto com balanceamento e diversificação. Tal resultado não pode ser encarado com grande surpresa, pois a regressão logística é um método consolidado para a modelagem desse tipo de dados, e suas variações (versão limitada e regularização com LASSO) a tornam ainda mais robusta.

Dentro do bloco dos modelos construídos no conjunto desbalanceado, observamos que quando não utilizamos a regularização com LASSO, a versão limitada apresenta resultados

melhores que a versão usual em termos de KS, AUC e SQRes. Quando adicionamos o LASSO, observamos que a versão limitada apresenta resultados piores que a versão usual, sendo superior apenas na estatística KS. Tal resultado pode ter sido ocasionado pela dupla regularização (o que pode aumentar o viés do modelo) ou pelo fato de utilizar o mesmo λ em ambos métodos.

Em resumo, para esse estudo de caso, o melhor modelo obtido foi o que utiliza a regressão logística com LASSO por possuir as melhores métricas em 4 das 5 avaliadas (AUC, F1, Acurácia e SRQes). Apenas no KS esse modelo não atingiu o maior valor, ficando a cargo da regressão logística limitada com LASSO.

É válido ressaltar que não encontramos algoritmos/funções nos *softwares R* e *Python* relacionados à estimação dos parâmetros da regressão logística limitada e regressão logística limitada com LASSO, nem sobre o processo de balanceamento e diversificação proposto pela Wang *et al.* (2015), sendo essas contribuições extras do estudo para a área de modelagem.

Por fim, para trabalhos futuros seria interessante a otimização da função de estimação da regressão logística limitada e sua versão com LASSO. Em especial, isso possibilitaria a validação cruzada para encontrar o melhor λ, em um espaço tempo aceitável. Além disso, outros estudos de caso poderiam ser conduzidos, para verificar se os resultados seriam os mesmos dos aqui expostos. Em relação ao balanceamento e diversificação, algumas simulações poderiam ser realizadas, para verificar em qual cenário o tratamento apresentaria ganho de performance.

Os algoritmos utilizados para aplicação do estudo podem ser acessados em:
`https://github.com/Piccin-LE/TG_Estatistica/blob/master/codigos.Rmd`.

Referências Bibliográficas

Arthur, D. e Vassilvitskii, S. (2006). k-means++: The advantages of careful seeding. Relatório técnico, Stanford.

Beltran, R. D. (2019). *Detecção de fraudes bancárias utilizando métodos de clustering*. Universidade Federal do Pampa.

ClearSale (2021). Mapa da fraude - resultados 2020. Disponível em https://br.clear.sale/hubfs/marketing/mapa/clearsale_mapa_da_fraude_resultados_2020.pdf.

Cortes, C. e Vapnik, V. (1995). Support vector machine. *Machine learning*, **20**(3), 273–297.

Cox, D. R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, **20**(2), 215–232.

Cramer, J. (2004). Scoring bank loans that may go wrong: A case study. *Statistica Neerlandica*, **58**(3), 365–380.

Cressey, D. R. (1953). *Other people's money; a study of the social psychology of embezzlement*. Free Press.

Cristofaro, E. A. U. (2006). *Uma abordagem bayesiana para análise de fraude de subscrição em telecomunicações*. Universidade Federal de São Carlos.

Czepiel, S. A. (2002). Maximum likelihood estimation of logistic regression models: theory and implementation. *Disponível em czep.net/stat/mlelr.pdf*, **83**.

Elith, J., Leathwick, J. R. e Hastie, T. (2008). A working guide to boosted regression trees. *Journal of animal ecology*, **77**(4), 802–813.

Fanning, K., Cogger, K. O. e Srivastava, R. (1995). Detection of management fraud: a neural network approach. *Intelligent Systems in Accounting, Finance and Management*, **4**(2), páginas 113–126.

Hand, D. J. (2002). Pattern detection and discovery. Em *Pattern detection and discovery*, páginas 1–12. Springer.

Hastie, T. e Qian, J. (2014). Glmnet vignette. *Retrieved June*, páginas 1–30.

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

Izbicki, R. e dos Santos, T. M. (2020). *Aprendizado de máquina: uma abordagem estatística*. Rafael Izbicki.

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

Moraes, D. d. (2008). *Modelagem de fraude em cartão de crédito*. Universidade Federal de São Carlos.

Niu, X., Wang, L. e Yang, X. (2019). A comparison study of credit card fraud detection: Supervised versus unsupervised. *ArXiv preprint arXiv:1904.10604*.

Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, **1**(1), 81–106.

Sahin, Y., Bulkan, S. e Duman, E. (2013). A cost-sensitive decision tree approach for fraud detection. *Expert Systems with Applications*, **40**(15), 5916–5923.

Sahin, Y. G. e Duman, E. (2011). *Detecting credit card fraud by decision trees and support vector machines*. Newswood Limited.

Scipy (2022). Documentação da função *optimize*. Disponível em https://docs.scipy.org/doc/scipy/reference/optimize.html.

Shen, A., Tong, R. e Deng, Y. (2007). Application of classification models on credit card fraud detection. Em *2007 International conference on service systems and service management*, páginas 1–4. IEEE.

Sicsú, A. L. (2010). *Credit Scoring: Desenvolvimento, Implantação, Acompanhamento*. Blucher.

Singh, A., Ranjan, R. K. e Tiwari, A. (2021). Credit card fraud detection under extreme imbalanced data: A comparative study of data-level algorithms. *Journal of Experimental & Theoretical Artificial Intelligence*, páginas 1–28.

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society: Series B (Methodological)*, **58**(1), 267–288.

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

Zhu, C., Byrd, R. H., Lu, P. e Nocedal, J. (1997). Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. *ACM Transactions on mathematical software (TOMS)*, **23**(4), 550–560.

Apêndice A

Tabelas comparativas dos coeficientes

As tabelas a seguir são referentes aos coeficientes obtidos por cada um dos métodos aplicados no conjunto de dados desbalanceado.

Tabela A.1: Comparação dos coeficientes obtidos entre a Regressão Logística e a Regressão Logística Limitada.

Coeficiente	Logística	Logística Limitada	Diferença Absoluta
ω	1,0000	0,0806	0,9194
Intercepto	-3,2862	-0,1001	3,1860
V1:C1	2,0027	2,7893	0,7866
V1:C2	2,5274	3,6779	1,1505
V2:C1	0,0726	0,1773	0,1047
V2:C2	0,1454	0,5343	0,3889
V3	-0,1364	-0,4381	0,3017
V4:C1	0,0239	0,0161	0,0078
V4:C2	0,2001	0,2555	0,0554
V5:C1	0,0016	-0,0941	0,0957
V5:C2	0,0704	0,4999	0,4295
V6:C1	-0,1264	0,0620	0,1884
V6:C2	-0,1808	-0,2875	0,1067
V7	0,3847	0,8411	0,4564
V8	0,0998	0,1721	0,0723
V9	0,6271	0,6382	0,0112
V10	-0,1709	0,0474	0,2183
V11	0,0174	-0,0436	0,0609
V12	-0,2047	-0,0494	0,1553
V13	0,0461	0,1217	0,0756
V14	-0,0752	-0,1364	0,0611
V15	-0,1874	-0,1788	0,0086
V16	-0,7629	-0,8249	0,0620
V17	0,5970	0,8112	0,2142
V18	0,9223	1,2896	0,3673
V19	-0,6596	-0,7983	0,1387
V20	0,3705	0,5332	0,1627
V21	-0,3870	-0,5813	0,1943
V22	0,1234	0,1619	0,0385
V23	0,0000	-0,1387	0,1387
V24	0,1279	0,0516	0,0763
V25	-0,3694	-0,4862	0,1168
V26	-0,3387	-0,4775	0,1388
V27:C1	0,1640	0,5752	0,4112
V28	0,0668	0,2591	0,1923
V29	-0,5422	-0,8313	0,2891
V30:C1	-0,4161	-0,6418	0,2257
V30:C2	-0,4272	-0,7352	0,3080
V31:C1	0,1631	0,0873	0,0758
V32	-0,1163	-0,3481	0,2318
V33	-0,0252	-0,0560	0,0308
V34	-0,2390	-0,2901	0,0511
V35	-0,1862	-0,2885	0,1023
V36	0,1822	0,3646	0,1825
V37	1,3952	1,5196	0,1245
V38	0,4261	0,5906	0,1646
V39	-0,1233	-0,0451	0,0782
V40	-0,9619	-0,8226	0,1394
V41	-0,7833	-0,8371	0,0539
V42	0,4387	0,4452	0,0066
V43	-1,7090	-1,7247	0,0157
V44	0,0528	0,1524	0,0996
V45	-0,6632	-0,7063	0,0431
V46	-1,0879	-0,9803	0,1077
V47	0,0913	0,1166	0,0253
V48	1,3463	1,4465	0,1002
V49:C1	-0,1721	-0,3351	0,1630
V49:C2	-0,2018	-0,7661	0,5643
V50	-0,0508	-0,0885	0,0376
V51:C1	0,4172	0,5953	0,1781
V52	0,0788	0,1416	0,0629
V53	-0,0323	-0,0303	0,0020
V54	-0,0567	-0,1433	0,0867
V55:C1	0,0707	0,1076	0,0369
V55:C2	0,2617	0,5310	0,2693
V56	0,8259	1,0975	0,2716
V57	0,0192	-0,0031	0,0223
V58	-0,0626	-0,1198	0,0572
V59:C1	2,2793	2,4537	0,1744

Tabela A.2: Comparação dos coeficientes obtidos entre a Regressão Logística com LASSO e a Regressão Logística Limitada com LASSO.

Coeficiente	Logística com LASSO	Logística Limitada com LASSO	Diferença Absoluta
ω	1,0000	0,0803	0,9197
Intercepto	-3,8300	-1,0830	2,7470
V1:C1	0,4662	1,2803	0,8141
V1:C2	0,7376	1,7173	0,9797
V2:C1	0,0000	0,0648	0,0648
V2:C2	0,0000	0,1485	0,1485
V3	0,0000	-0,1958	0,1958
V4:C1	0,0000	-0,0020	0,0020
V4:C2	0,0198	0,0732	0,0534
V5:C1	0,0000	-0,0321	0,0321
V5:C2	0,0268	0,1096	0,0829
V6:C1	0,0000	-0,0433	0,0433
V6:C2	0,0000	-0,1233	0,1233
V7	0,2019	0,7467	0,5448
V8	0,2351	0,5241	0,2890
V9	0,0000	0,0963	0,0963
V10	0,0000	-0,0318	0,0318
V11	0,0000	-0,0118	0,0118
V12	0,0000	0,0587	0,0587
V13	0,0000	-0,0420	0,0420
V14	-0,0020	-0,0768	0,0748
V15	0,0096	-0,0343	0,0439
V16	0,0000	-0,1186	0,1186
V17	0,0000	0,1504	0,1504
V18	0,0347	0,2884	0,2536
V19	0,0000	-0,1255	0,1255
V20	0,0000	0,1918	0,1918
V21	0,0000	-0,1581	0,1581
V22	0,0000	0,0786	0,0786
V23	0,0000	-0,1255	0,1255
V24	-0,0637	0,0581	0,1218
V25	0,0000	-0,2933	0,2933
V26	-0,1410	-0,2029	0,0619
V27:C1	0,0000	0,1642	0,1642
V28	0,0000	0,0634	0,0634
V29	-0,1669	-0,2905	0,1236
V30:C1	0,0000	-0,4895	0,4895
V30:C2	-0,0035	-0,5269	0,5234
V31:C1	0,0000	-0,0057	0,0057
V32	-0,0810	-0,2298	0,1488
V33	-0,0623	-0,1521	0,0898
V34	-0,0809	-0,1611	0,0802
V35	0,0000	-0,0694	0,0694
V36	0,0141	0,1382	0,1241
V37	0,0000	0,1067	0,1067
V38	0,0000	0,0465	0,0465
V39	0,0000	-0,0074	0,0074
V40	0,0000	-0,1061	0,1061
V41	-0,0038	-0,0981	0,0943
V42	0,0000	0,0342	0,0342
V43	0,0000	-0,2699	0,2699
V44	0,0000	0,0303	0,0303
V45	0,0000	-0,0680	0,0680
V46	0,0000	-0,2721	0,2721
V47	0,0000	0,0330	0,0330
V48	0,0000	0,4184	0,4184
V49:C1	0,0000	-0,1038	0,1038
V49:C2	0,0000	-0,2047	0,2047
V50	-0,0114	-0,1765	0,1651
V51:C1	0,0130	0,0389	0,0259
V52	0,0000	0,0700	0,0700
V53	-0,0366	-0,0965	0,0599
V54	0,0221	-0,0754	0,0975
V55:C1	0,0099	0,0339	0,0240
V55:C2	0,0954	0,2009	0,1055
V56	0,0141	0,2294	0,2154
V57	0,0000	0,0975	0,0975
V58	-0,0145	-0,3530	0,3386
V59:C1	0,0284	0,1193	0,0909

Tabela A.3: Comparação dos coeficientes obtidos entre a Regressão Logística e a Regressão Logística com LASSO.

Coeficiente	Logística	Logística com LASSO	Diferença Absoluta
Intercepto	-3,2862	-3,8300	0,5438
V1:C1	2,0027	0,4662	1,5365
V1:C2	2,5274	0,7376	1,7898
V2:C1	0,0726	0,0000	0,0726
V2:C2	0,1454	0,0000	0,1454
V3	-0,1364	0,0000	0,1364
V4:C1	0,0239	0,0000	0,0239
V4:C2	0,2001	0,0198	0,1803
V5:C1	0,0016	0,0000	0,0016
V5:C2	0,0704	0,0268	0,0437
V6:C1	-0,1264	0,0000	0,1264
V6:C2	-0,1808	0,0000	0,1808
V7	0,3847	0,2019	0,1828
V8	0,0998	0,2351	0,1353
V9	0,6271	0,0000	0,6271
V10	-0,1709	0,0000	0,1709
V11	0,0174	0,0000	0,0174
V12	-0,2047	0,0000	0,2047
V13	0,0461	0,0000	0,0461
V14	-0,0752	-0,0020	0,0732
V15	-0,1874	0,0096	0,1970
V16	-0,7629	0,0000	0,7629
V17	0,5970	0,0000	0,5970
V18	0,9223	0,0347	0,8876
V19	-0,6596	0,0000	0,6596
V20	0,3705	0,0000	0,3705
V21	-0,3870	0,0000	0,3870
V22	0,1234	0,0000	0,1234
V23	0,0000	0,0000	0,0000
V24	0,1279	-0,0637	0,1917
V25	-0,3694	0,0000	0,3694
V26	-0,3387	-0,1410	0,1976
V271	0,1640	0,0000	0,1640
V28	0,0668	0,0000	0,0668
V29	-0,5422	-0,1669	0,3754
V30:C1	-0,4161	0,0000	0,4161
V30:C2	-0,4272	-0,0035	0,4237
V31:C1	0,1631	0,0000	0,1631
V32	-0,1163	-0,0810	0,0353
V33	-0,0252	-0,0623	0,0371
V34	-0,2390	-0,0809	0,1581
V35	-0,1862	0,0000	0,1862
V36	0,1822	0,0141	0,1680
V37	1,3952	0,0000	1,3952
V38	0,4261	0,0000	0,4261
V39	-0,1233	0,0000	0,1233
V40	-0,9619	0,0000	0,9619
V41	-0,7833	-0,0038	0,7794
V42	0,4387	0,0000	0,4387
V43	-1,7090	0,0000	1,7090
V44	0,0528	0,0000	0,0528
V45	-0,6632	0,0000	0,6632
V46	-1,0879	0,0000	1,0879
V47	0,0913	0,0000	0,0913
V48	1,3463	0,0000	1,3463
V49:C1	-0,1721	0,0000	0,1721
V49:C2	-0,2018	0,0000	0,2018
V50	-0,0508	-0,0114	0,0395
V51:C1	0,4172	0,0130	0,4042
V52	0,0788	0,0000	0,0788
V53	-0,0323	-0,0366	0,0042
V54	-0,0567	0,0221	0,0788
V55:C1	0,0707	0,0099	0,0608
V55:C2	0,2617	0,0954	0,1663
V56	0,8259	0,0141	0,8118
V57	0,0192	0,0000	0,0192
V58	-0,0626	-0,0145	0,0481
V59:C1	2,2793	0,0284	2,2509

Tabela A.4: Comparação dos coeficientes obtidos entre a Regressão Logística Limitada e a Regressão Logística Limitada com LASSO.

Coeficiente	Logística Limitada	Logística Limitada com LASSO	Diferença Absoluta
ω	0,0806	0,0803	0,0003
Intercepto	-0,1001	-1,0830	0,9829
V1:C1	2,7893	1,2803	1,5089
V1:C2	3,6779	1,7173	1,9605
V2:C1	0,1773	0,0648	0,1124
V2:C2	0,5343	0,1485	0,3858
V3	-0,4381	-0,1958	0,2423
V4:C1	0,0161	-0,0020	0,0180
V4:C2	0,2555	0,0732	0,1822
V5:C1	-0,0941	-0,0321	0,0620
V5:C2	0,4999	0,1096	0,3903
V6:C1	0,0620	-0,0433	0,1053
V6:C2	-0,2875	-0,1233	0,1642
V7	0,8411	0,7467	0,0944
V8	0,1721	0,5241	0,3520
V9	0,6382	0,0963	0,5419
V10	0,0474	-0,0318	0,0792
V11	-0,0436	-0,0118	0,0318
V12	-0,0494	0,0587	0,1081
V13	0,1217	-0,0420	0,1637
V14	-0,1364	-0,0768	0,0596
V15	-0,1788	-0,0343	0,1445
V16	-0,8249	-0,1186	0,7063
V17	0,8112	0,1504	0,6608
V18	1,2896	0,2884	1,0012
V19	-0,7983	-0,1255	0,6728
V20	0,5332	0,1918	0,3414
V21	-0,5813	-0,1581	0,4232
V22	0,1619	0,0786	0,0833
V23	-0,1387	-0,1255	0,0132
V24	0,0516	0,0581	0,0065
V25	-0,4862	-0,2933	0,1929
V26	-0,4775	-0,2029	0,2746
V271	0,5752	0,1642	0,4110
V28	0,2591	0,0634	0,1957
V29	-0,8313	-0,2905	0,5408
V30:C1	-0,6418	-0,4895	0,1523
V30:C2	-0,7352	-0,5269	0,2083
V31:C1	0,0873	-0,0057	0,0930
V32	-0,3481	-0,2298	0,1182
V33	-0,0560	-0,1521	0,0961
V34	-0,2901	-0,1611	0,1290
V35	-0,2885	-0,0694	0,2192
V36	0,3646	0,1382	0,2264
V37	1,5196	0,1067	1,4129
V38	0,5906	0,0465	0,5441
V39	-0,0451	-0,0074	0,0376
V40	-0,8226	-0,1061	0,7165
V41	-0,8371	-0,0981	0,7390
V42	0,4452	0,0342	0,4111
V43	-1,7247	-0,2699	1,4548
V44	0,1524	0,0303	0,1221
V45	-0,7063	-0,0680	0,6383
V46	-0,9803	-0,2721	0,7082
V47	0,1166	0,0330	0,0836
V48	1,4465	0,4184	1,0282
V49:C1	-0,3351	-0,1038	0,2313
V49:C2	-0,7661	-0,2047	0,5614
V50	-0,0885	-0,1765	0,0880
V51:C1	0,5953	0,0389	0,5564
V52	0,1416	0,0700	0,0716
V53	-0,0303	-0,0965	0,0661
V54	-0,1433	-0,0754	0,0680
V55:C1	0,1076	0,0339	0,0737
V55:C2	0,5310	0,2009	0,3301
V56	1,0975	0,2294	0,8680
V57	-0,0031	0,0975	0,1006
V58	-0,1198	-0,3530	0,2333
V59:C1	2,4537	0,1193	2,3344