Um estudo comparativo de técnicas de validação cruzada aplicadas a modelos para dados desbalanceados

Luiza Tuler Veloso

Dissertação apresentada
ao
Instituto de Matemática e Estatística
da
Universidade de São Paulo
para
obtenção do título
de
Mestre em Ciências

Programa: Estatística

Orientadora: Profa. Dra. Viviana Giampaoli

Durante o desenvolvimento deste trabalho o autor recebeu auxílio financeiro do CNPq

São Paulo, Fevereiro de 2022

Resumo

Dentro do contexto de modelagem preditiva, a escolha de um modelo perpassa pela avaliação da qualidade das predições por meio do Risco Esperado. Esse risco, no entanto, pode ser subestimado se obtido a partir da mesma amostra utilizada para ajuste do modelo. Para lidar com este problema, surgem estratégias de Validação Cruzada (*Hold-out*, *K-Fold*, *Leave-one-out*, *Bootstrap*) que buscam separar os dados disponíveis em Amostra de Treino, em que o modelo será ajustado, e Amostra de Validação, em que o modelo terá seu desempenho verificado. Ao se tratar de dados desbalanceados, ou seja, dados nos quais o evento de interesse ($Y = 1$) da variável resposta binária ocorre dezenas a milhares de vezes menos do que a outra categoria ($Y = 0$), podem ser necessárias algumas adaptações no processo de modelagem e validação. Em vista disso, este trabalho busca avaliar a maneira com que as técnicas de validação de modelos se comportam conforme o desbalanceamento dos dados para tamanhos distintos de amostra. Para isso, foi realizada revisão das técnicas que possibilitam a validação de modelos e revisão das tratativas e principais dificuldades ao modelar dados desbalanceados. Por fim, as técnicas de validação foram avaliadas, a partir de simulações, para modelos logísticos com correção no viés para dados desbalanceados proposta por King e Zeng [2001] e, posteriormente, foi avaliada a metodologia em estudo de simulação e aplicada em uma base de dados real referente a notificação de casos da Síndrome Inflamatória Multissistêmica (SIM-P) temporalmente associada à COVID-19.

Palavras-chave: Validação Cruzada, Dados desbalanceados, Risco Esperado, SIM-P.

Abstract

Veloso, L. T. **A comparative study of cross validation techniques applied to imbalanced data models**. 2022. 63 f. Dissertação - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2010.

Within the context of predictive modeling, the chosing of a model involves evaluating, through Expected Risk, the quality of predictions. Such risk, however, may be underestimated if obtained from the same sample utilized to adjusting the model. To deal with such problem, Cross Validation strategies (*Hold-out*, *K-Fold*, *Leave-one-out*, *Bootstrap*) emerge, that seek to split the available data in Training Sample, in which the model will be adjusted, and Validation Sample, where the model will have its performance verified. When dealing with imbalanced data, in other words, data in which the event of interest ($Y = 1$) of the binary response variable occurs dozens to thousands of times less than the other category ($Y = 0$), might need some adaptations in the process of modeling and validation. In view of this, this paper seeks to evaluate the way in which model validation techiniques behave, according to the degree of data imbalance and different sample sizes. For such, a review of the techniques that enable the models validation and revision of the approaches and main difficulties when modeling imbalanced data was made. Finally, the validation techniques were evaluated, through simulation studies, for corrected logistic regression applied to imbalanced data, proposed by King e Zeng [2001] and, later, the methodology was assessed in a simulation study, then applied to a real database regarding the notification of cases of Multisystem Inflammatory Syndrome in Children (MIS-C) temporally associated with COVID-19.

Keywords: Cross Validation, Imbalanced Data, Expected Risk, MIS-C.

Sumário

1. Introdução

O Aprendizado Estatístico e a Modelagem Preditiva surgem em um contexto em que se deseja conhecer uma variável Y a partir de outras p variáveis $X = (X_1, X_2, \cdots, X_p)$ que, potencialmente, ajudariam a compreender o comportamento dessa variável de interesse. A ideia por trás da modelagem preditiva, então, está na busca por um modelo $\hat{f}(X)$ que possibilite as melhores predições possíveis de Y a partir do conhecimento exclusivo de X [James *et al.*, 2013].

O processo de definição do modelo $\hat{Y} = \hat{f}(X)$ é realizado a partir de uma amostra dos dados que apresenta valores de Y e X. A escolha do modelo ideal perpassa pelo cálculo do Risco Esperado que irá possibilitar a comparação do que foi predito pelo modelo com o que foi observado na amostra [Borra e Ciaccio, 2010]. Quanto menor for o Risco Esperado estimado melhor é considerado o modelo em questão.

No caso do Risco Esperado ser estimado a partir da mesma amostra utilizada para o ajuste do modelo, é comum lidar com um risco subestimado que acaba induzindo a escolha de modelos com baixa capacidade de generalização. Para contornar este problema, é interessante que o modelo tenha o desempenho verificado para observações inéditas. Sendo assim, surge a necessidade de separar a amostra disponível em um conjunto em que o modelo será ajustado, Amostra de Treino, e outro conjunto em que o modelo terá o desempenho verificado, Amostra de Validação [James *et al.*, 2013]. Este procedimento, tal como abordado por Stone [1974], é denominado Validação Cruzada.

Hastie *et al.* [2008] vão além e discutem que, em um contexto de validação de grandes bases de dados, o ideal seria dividir aleatoriamente a base original em três partes: Amostra de Treino, Amostra de Validação e Amostra de Teste. Dessa maneira, a Amostra de Treino seria utilizada para o ajuste dos modelos, a Amostra de Validação seria utilizada na verificação do desempenho e seleção dos possíveis modelos e a partir da Amostra de Teste seria calculado o erro de predição do modelo final selecionado. Neste trabalho será considerada somente a situação em que a amostra original é particionada em Amostra de Treino e Amostra de Validação.

Estratégias de Validação Cruzada conhecidas e amplamente utilizadas são *Hold-out*, *K-fold*,

Leave-one-out e *Bootstrap*. Estas técnicas distintas podem levar a diferentes conclusões no que tange ao "melhor" modelo e apresentam diferentes vantagens e desvantagens abordadas em mais detalhes no Capítulo 3. Por apresentarem distinções entre si quanto a usabilidade é que se torna emergente o interesse em realizar estudos comparativos entre estes procedimentos em diferentes contextos, tal como feito por Kohavi [1995], Kim [2009], Borra e Ciaccio [2010], Cunha [2019], entre outros.

Quando a variável de interesse Y possui uma natureza binária podem ocorrer situações em que a quantidade de observações pertencentes a classe de interesse ($y = 1$) apareça dezenas a milhares de vezes menos do que a outra classe ($y = 0$), caracterizando desbalanceamento dos dados [King e Zeng, 2001]. Estes tipos de dados, em situações práticas, podem surgir no contexto de detecção de fraudes bancárias [Lima e Pereira, 2015, Moepya *et al.*, 2014, Wei *et al.*, 2013], diagnóstico de doenças [Casañola-Martin *et al.*, 2016, Lusa *et al.*, 2010], análise de sentimento em textos [Lane *et al.*, 2012], análise de conflitos políticos entre países [King e Zeng, 2001], entre outras situações em que se precisa lidar com eventos raros [Al-Ghraibah *et al.*, 2015, Vong *et al.*, 2015].

Se tratando de um evento raro é possível que o processo de modelagem e validação usuais não se apresentem de forma adequada. No Capítulo 4 são debatidas algumas questões com relação a modelagem desses tipos de dados, assim como, métricas de desempenho mais adequadas, estratégias existentes para lidar com o desbalanceamento e possíveis problemas que podem dificultar esse processo de modelagem [Branco *et al.*, 2016, López *et al.*, 2013]. No Capítulo 5 é apresentada uma estratégia que corrige o modelo logístico para que se torne adequado a dados desbalanceados.

É de conhecimento que quanto maior a amostra melhor a performance do modelo independentemente do grau de desbalanceamento e, além disso, sabe-se que o percentual de desbalanceamento que vai possibilitar uma melhor performance das predições varia conforme o tamanho da Amostra de Treino [López *et al.*, 2013]. Além disso, Raeder *et al.* [2012] mostram, neste mesmo contexto, que a métrica e o método de validação escolhidos podem levar a conclusões extremamente diferentes com relação aos modelos testados.

Sendo assim, o objetivo deste trabalho é avaliar o desempenho das principais técnicas de validação cruzada diante o desbalanceamento dos dados considerando cenários distintos de tamanho de amostra e grau de desbalanceamento. Para isso, no decorrer deste processo de análise será considerado para ajuste o Modelo Logístico com correção do viés para dados desbalanceados proposto por King e Zeng [2001].

No Capítulo 2 são abordados conceitos relacionados ao aprendizado estatístico, assim como, evidenciadas a importância e a maneira com que se estima o Risco Esperado a partir do Erro Aparente,

além de expor outras métricas utilizadas para verificação do desempenho de modelos. Os diferentes métodos de validação cruzada, *Hold-out*, *K-fold*, *Leave-one-out* e *Bootstrap* são apresentados no Capítulo 3.

Contemplando a problemática de se modelar um evento tido como raro, no Capítulo 4 são abordadas algumas particularidades deste tipo de dado, problemas que podem dificultar o processo de construção de modelos e algumas nuances dos métodos de validação já abordados no Capítulo 3. Além disso, no Capítulo 5 são explicitadas estratégias específicas de modelagem para dados desbalanceados. O trabalho segue no Capítulo 6 com um estudo de simulação que busca compreender o desempenho de técnicas de validação distintas em diferentes cenários de tamanho de amostra e grau de desbalanceamento.

Feito isso, no Capítulo 7 é realizada aplicação das técnicas abordadas no trabalho em uma base de dados real a respeito da evolução de casos notificados da Síndrome Inflamatória Sistêmica (SIM-P) temporalmente associada à COVID-19. Por fim, as conclusões finais e síntese do trabalho são realizadas no Capítulo 8.

2. Aprendizado Estatístico

Aprendizado estatístico é o termo utilizado para se referir a um conjunto de técnicas utilizadas para modelar e compreender bancos de dados complexos [James *et al.*, 2013]. A ideia está pautada na possibilidade de extrair conhecimento a partir de dados.

O frequente entusiasmo em verificar a relação entre variáveis e realizar a predição de determinados eventos de interesse, que possibilitem o conhecimento de fenômenos e a tomada decisões motivam a construção de modelos estatísticos [James *et al.*, 2013]. Quando é observada a variável do evento de interesse Y, denominada variável resposta, e outras p variáveis X_1, X_2, ..., X_p que assumimos estarem relacionadas à resposta, denominadas variáveis explicativas, é possível escrever a relação de tal forma que

$$Y \approx f(X) + \epsilon,$$

nesse caso, $f(X)$ é uma função fixa de $\mathbf{X} = (X_1, X_2, \ldots, X_p)$ e representa a informação sistemática de X sobre Y, enquanto que ϵ é o termo que faz referência ao erro aleatório e é independente de X.

Geralmente a função $f(\cdot)$ é desconhecida e o interesse dentro da aprendizagem estatística está justamente nas diferentes abordagens para estimação dessa função. Em alguns estudos o interesse ao estimar $f(\cdot)$ está em descobrir de que forma ocorre a relação entre Y e X (objetivo inferencial), em outros o foco é prioritariamente na qualidade da predição de Y a partir dos valores de X (objetivo preditivo). Neste trabalho será dado enfoque à abordagem preditiva.

Nos casos em que as variáveis explicativas podem ser facilmente acessadas e a variável resposta, na qual está o interesse principal, não pode ser obtida de maneira simples é possível predizer Y de tal forma que

$$\hat{Y} = \hat{f}(X),$$

em que $\hat{Y}$ representa o valor predito de Y e $\hat{f}(\cdot)$ a estimativa para $f(\cdot)$.

Conforme abordado por James *et al.* [2013] a acurácia de $\hat{Y}$ na predição de Y está atrelada a dois aspectos, o erro redutível e o erro irredutível. De forma geral, $\hat{f}(\cdot)$ não será impecável na estimação de $f(\cdot)$ o que implica em um erro que pode ser potencialmente reduzido conforme melhorias na acurácia de $\hat{f}(\cdot)$, erro redutível. No entanto, ainda que $\hat{f}(\cdot)$ seja uma estimativa perfeita para $f(\cdot)$ a predição de Y ainda teria um erro atrelado relacionado a ϵ, o erro irredutível.

2.1 Função Perda e Risco Esperado

A ideia, dentro desse contexto, é encontrar a $\hat{f}(\cdot)$ que irá implicar no menor erro possível com relação às predições de Y. Para que seja mensurada a intensidade dessa dissimilaridade define-se a Função Perda $L(Y, \hat{f}(X))$ que penaliza, sob algum critério a ser definido, a distinção verificada entre o valor predito e o real valor de Y. A Função Perda é definida de tal forma que quanto maiores os valores obtidos maior o indicativo de que a $\hat{f}(\cdot)$ não é adequada para predizer Y.

Quando se trata de um problema de classificação em que a variável resposta é caracterizada como categórica, conforme apontado por Hastie *et al.* [2008], a função perda usual é do tipo perda 0-1 sendo definida como

$$L(Y, \hat{f}(X)) = \mathbb{1}(Y \neq \hat{Y}).$$

De acordo com Hastie *et al.* [2008] quando Y se trata de uma variável categórica com apenas duas classes, $y = 1$ denominada evento e $y = 0$ não evento, e $\hat{f}(X)$ representa o número estimado para $Y^* = \mathrm{P}(Y = 1)$ é possível escrever a função perda entropia cruzada que compara Y e $\hat{f}(X)$ da seguinte maneira:

$$L(Y, \hat{f}(X)) = -Y \log[\hat{f}(X)] - (1 - Y) \log[1 - \hat{f}(X)].$$

A partir da definição da Função Perda é possível estabelecer uma medida mais geral denominada Risco Esperado. Essa medida, conforme abordado por Borra e Ciaccio [2010], considera todos os possíveis valores das variáveis explicativas da amostra analisada e é dada por

$$\text{Risco Esperado} = \mathrm{E}_X \, \mathrm{E}_{Y|X}[L(Y, \hat{f}(X))|\hat{f}, X]. \tag{2.1}$$

Por depender do conhecimento da distribuição de X, a medida do Risco Esperado definida pela equação (2.1) pode ser difícil de se calcular. Em vista disso, Borra e Ciaccio [2010] propõem o uso

do estimador Erro Aparente para o Risco Esperado definido como

$$err = \frac{1}{n}\sum_{i=1}^{n} L(y_i, \hat{f}(\mathbf{x_i})), \tag{2.2}$$

ou seja, o Erro Aparente é a média da Função Perda para as n observações da amostra.

Borra e Ciaccio [2010] destacam que, apesar da expressão (2.2) poder ser facilmente calculada, geralmente subestima o Risco Esperado (2.1). A otimalidade deste estimador decorre do uso da mesma amostra para definição do modelo e avaliação do erro do modelo, de tal forma que quanto mais complexo o modelo definido menor costuma ser o erro estimado por (2.2).

Uma alternativa para evitar o problema mencionado é o uso de técnicas de validação que buscam dividir a amostra original $\mathcal{D}$ em amostra de treino $\mathcal{D}_t$ e amostra de validação $\mathcal{D}_v$. Essa divisão ocorre da tal forma que $\hat{f}_t(\cdot)$ é obtida a partir da amostra de treino e o Risco Esperado é estimado como uma média da Função Perda para as v observações presentes na amostra de validação que contempla observações ainda não utilizadas para o ajuste do modelo. Ou seja, o Risco Esperado pode ser dado por

$$err_v = \frac{1}{v}\sum_{i=1}^{v} L(y_i, \hat{f}_t(\mathbf{x_i})), \tag{2.3}$$

em que y_i e $\mathbf{x_i}$ se referem aos valores das variáveis na amostra de validação.

Aspectos mais específicos relacionados à validação de modelos serão abordados detalhadamente no Capítulo 3.

Dessa maneira, James *et al.* [2013] apontam para a existência de dois tipos de erro com relação a amostra considerada, são eles, o erro de treino e o erro de validação. O primeiro diz respeito ao erro médio calculado ao utilizar o modelo estatístico para predizer resultados de observações utilizadas no ajuste do modelo, enquanto o segundo se refere ao erro médio calculado ao utilizar o modelo na predição de observações que não compunham a amostra de treino.

Trade-off entre Viés e Variância

James *et al.* [2013] afirmam que o erro de validação esperado está usualmente atrelado ao viés e à variância dos estimadores. Esta relação ocorre de tal forma que, quanto menores forem o viés e a variância, de forma simultânea, menor é este erro.

Neste caso, o conceito de variância está relacionado ao quanto a $\hat{f}(X)$ pode se alterar ao considerar uma amostra distinta para ajuste do modelo [James *et al.*, 2013]. Diferentes amostras vão implicar em diferentes $\hat{f}(X)$, mas o ideal é que esta variabilidade não seja elevada, principalmente quando se tratam de mudanças sutis na amostra considerada. Hastie *et al.* [2008] apontam que mo-

delos mais complexos, que super-ajustam os dados e possuem menor capacidade de generalização, costumam apresentar variância elevada.

No entanto, o viés está associado a uma possível falha na escolha da complexidade do modelo. Isto é, está relacionado a situações em que é selecionado um modelo mais simples do que seria, de fato, exigido pelo problema. Neste caso, diferentemente da variância, modelos mais complexos vão, geralmente, implicar em um menor viés [James *et al.*, 2013]. Sendo assim, a ideia ao realizar a escolha de modelos por meio de técnicas de validação é buscar parcimônia na escolha da complexidade do modelo para que exista um equilíbrio entre o viés e a variância atrelados às estimativas.

2.2 Métricas de desempenho

Branco *et al.* [2016] apontam que é possível enxergar a obtenção de um modelo estatístico a partir de uma base de dados como sendo um processo de busca guiada por algum critério de preferência com relação a performance do modelo. Segundo Branco *et al.* [2016], a definição da métrica para avaliar o modelo a ser escolhido deve ser pensada no contexto do problema trabalhado, tendo em vista que diferentes métricas podem levar a conclusões distintas.

Quando se trata de um problema de classificação, em geral, o que se modela é a probabilidade de pertencer a determinada classe e a observação é atribuída à classe com maior probabilidade predita. Para situações em que a variável resposta só possui duas categorias é estabelecido um ponto de corte c, usualmente definido de forma a maximizar os acertos do modelo, ver por exemplo [Tamura, 2007]. O ponto de corte c funciona de tal maneira que:

$$\hat{Y} = \begin{cases} 1, & \text{se } \hat{f}(X) > c \\ 0, & \text{se } \hat{f}(X) < c. \end{cases}$$

Estabelecido este limiar é possível ter um panorama do comportamento do modelo quanto a erros e acertos. A seguir são apresentadas algumas das métricas mais usuais.

Conforme o esquema de uma Matriz de Confusão apresentado na Figura 2.1 existem duas maneiras possíveis para a ocorrência de erros no processo de predição: o falso positivo, em que a classe predita de maneira incorreta é $\hat{y} = 1$, definida como classe positiva; e o falso negativo, em que a classe predita incorretamente é $\hat{y} = 0$, intitulada classe negativa [Murphy, 2012].

		Classe Predita		
		+	-	**Total**
Classe Observada	+	Verdadeiro Positivo (VP)	Falso Negativo (FN)	P
	−	Falso Positivo (FP)	Verdadeiro Negativo (VN)	N
	Total	P^*	N^*	

Figura 2.1: *Possíveis resultados de um modelo de classificação*

Métricas escalares

Métricas escalares possibilitam a utilização de um resultado sucinto, a partir de um único valor, que permite a comparação de modelos. Métricas deste tipo costumam estar atreladas à definição de alguns fatores, tal como o ponto de corte, que podem ser um pouco incertos [Branco *et al.*, 2016].

A Acurácia é uma métrica frequentemente utilizada que permite avaliar um modelo simplesmente considerando o percentual de observações classificadas corretamente.

$$\text{Acurácia} = \frac{\mathrm{VP} + \mathrm{VN}}{\mathrm{P} + \mathrm{N}} = \frac{\mathrm{VP} + \mathrm{VN}}{\mathrm{P}^* + \mathrm{N}^*}.$$

Em que VP, VN, P, N, P^* e N^* são definidos tal como exposto na Figura 2.1. Esta medida trata de maneira similar os diferentes erros de classificação possíveis, o que raramente é ideal.

A partir da possibilidade de se apurar diferentes tipos de erros e de acertos são definidas algumas medidas para auxiliar na verificação do desempenho preditivo do modelo. Estas métricas, em geral, se mostram mais adequadas do que a Acurácia porque possibilitam considerar a distribuição dos dados. A maneira com que a Taxa de Verdadeiros Positivos ($\mathrm{VP}_{\%}$), Taxa de Verdadeiros Negativos ($\mathrm{VN}_{\%}$), Taxa de Falsos Positivos ($\mathrm{FP}_{\%}$), Taxa de Falsos Negativos ($\mathrm{FN}_{\%}$) e Precisão são definidas pode ser verificada na Tabela 2.1.

Tabela 2.1: *Medidas importantes para classificação e diagnóstico*

Medida	Definição
Taxa de Verdadeiros Positivos (Sensibilidade ou Recall)	VP/P
Taxa de Verdadeiros Negativos (Especificidade)	VN/N
Taxa de Falsos Positivos	FP/N
Taxa de Falsos Negativos	FN/P
Precisão	VP/P^*

Monitorar de maneira simultânea tantas métricas quanto as expostas na Tabela 2.1 pode se tornar difícil no processo de seleção de modelos, principalmente, pelo *tradeoff* existente entre elas que impede a obtenção simultânea de um "valor ótimo" destas medidas. Sendo assim, novas métricas foram sendo propostas no sentido de sintetizar a avaliação de modelos.

Para um ponto de corte fixo em que são obtidos valores únicos de Precisão e Sensibilidade Murphy [2012] aponta que é possível calcular uma estatística única denominada F1-score. Essa medida é a média harmônica da Precisão e Sensibilidade calculada por

$$\text{F1-score} = \frac{2 \times \text{Precisão} \times \text{Sensibilidade}}{\text{Precisão} + \text{Sensibilidade}},$$

que, conforme apontado por Forman e Scholz [2010], pode ter sua definição simplificada por

$$\begin{aligned}
\text{F1-score} &= \frac{2 \times \text{Precisão} \times \text{Sensibilidade}}{\text{Precisão} + \text{Sensibilidade}} \\
&= 2\frac{\left(\frac{\text{VP}}{\text{P}^*}\right)\left(\frac{\text{VP}}{\text{P}}\right)}{\left(\frac{\text{VP}}{\text{P}^*}\right) + \left(\frac{\text{VP}}{\text{P}}\right)} \\
&= 2\frac{\left(\frac{\text{VP}}{\text{VP+FP}}\right)\left(\frac{\text{VP}}{\text{VP+FN}}\right)}{\left(\frac{\text{VP}}{\text{VP+FP}}\right) + \left(\frac{\text{VP}}{\text{VP+FN}}\right)} \\
&= \frac{2 \times \text{VP}}{2 \times \text{VP} + \text{FP} + \text{FN}}.
\end{aligned}$$

É possível notar que o valor da medida F1-score será alto conforme os valores dos Falsos Positivos e Falsos Negativos forem baixos. Sendo assim, um modelo de classificação mais adequado seria aquele com o maior F1-score possível.

As medidas mencionadas na Tabela 2.1 e a F1-score podem ser úteis principalmente para problemas em que é importante verificar de que maneira ocorreram os erros e acertos ao realizar a predições dentro de cada uma das classes. Ou seja, quando há interesse em modelar problemas em que um Falso Positivo tem importância diferente da de um Falso Negativo, a utilização das medidas mencionadas se mostra útil para auxiliar na seleção do modelo.

Métricas baseadas em gráficos

Conforme mencionado, as medidas tratadas anteriormente (Seção 2.2) dependem do ponto de corte estabelecido para que sejam realizadas as classificações. A redução no ponto de corte pode, por exemplo, implicar na redução de Falsos Negativos e no aumento de Falsos Positivos, e vice versa. Dessa maneira, é importante estabelecer este limiar conforme as características inerentes ao

problema e a gravidade associada a cada tipo de erro [James *et al.*, 2013].

Uma ferramenta gráfica que auxilia na definição do ponto de corte para que seja realizada a classificação é conhecida por Curva ROC [Murphy, 2012]. Este gráfico permite a visualização do comportamento das Taxas de Verdadeiros Positivos e Falsos Positivos conforme é variado o limiar para classificação. Ou seja, cada ponto no gráfico é referente ao desempenho do modelo considerando um diferente critério de classificação.

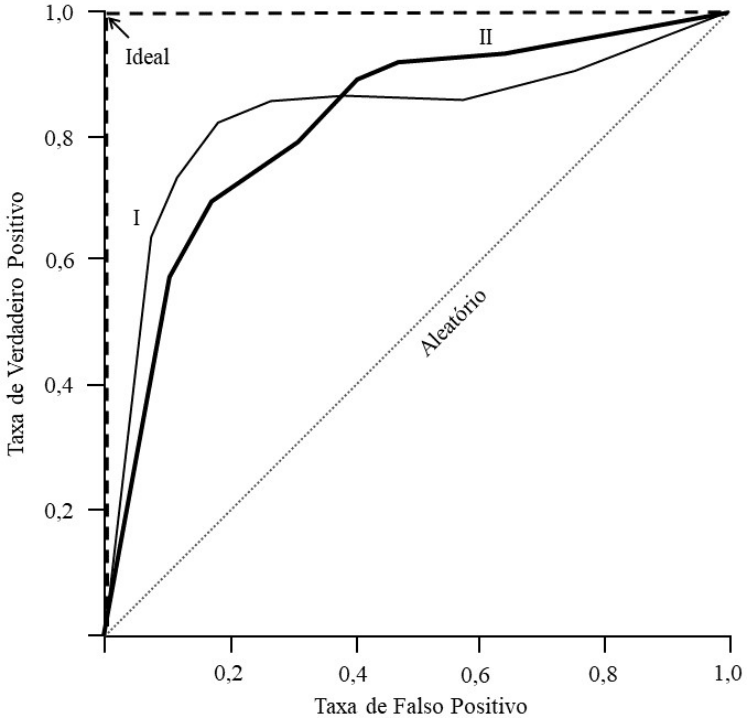


Figura 2.2: *Curva ROC para quatro modelos: ideal, I, II e aleatório.*
Fonte: Elaborada pela autora.

Na Figura 2.2 é possível visualizar um exemplo do que seria a curva ROC para quatro modelos diferentes. Em um gráfico desse tipo quanto mais próxima a curva passa do canto esquerdo (Taxa de Falso Positivos igual a 0) e do limite superior (Taxa de Verdadeiro Positivo igual a 1) melhor é o desempenho do modelo analisado. Além disso, quanto mais a curva se aproxima da diagonal que apresenta um ângulo de 45°, mais próximo o desempenho da classificação do modelo está de um classificador completamente aleatório.

Apesar da curva ROC ser uma solução amplamente utilizada para avaliação de desempenho, comparar diversos modelos por meio desta ferramenta pode ser uma tarefa complicada, principalmente, se uma das curvas não apresenta permanente dominância com relação às demais. Como consequência da dificuldade mencionada e decorrente da característica de que a curva ROC não fornece um valor único para medida de performance é habitualmente utilizada medida referente à área sob a curva ROC, conhecida por AUC, que possibilita a escolha do melhor modelo em média

[Branco *et al.*, 2016].

Segundo Forman e Scholz [2010], o AUC pode, ainda, ser interpretado como sendo a probabilidade empírica de que uma observação positiva ($y_i = 1$) selecionada aleatoriamente apresente um valor de $f(X_i)$ abaixo do valor $f(X_j)$ de uma observação negativa ($y_j = 0$), também selecionada aleatoriamente. Isto é,

$$\text{AUC} = \frac{1}{\text{N} \cdot \text{P}} \sum_{i|y_i=1} \sum_{j|y_j=0} \mathbb{1}(f(X_i), f(X_j)),$$

em que N e P são definidos conforme Figura 2.1 e a função indicadora é tal que,

$$\mathbb{1}(f(X_i), f(X_j)) = \begin{cases} 1, & \text{se } f(X_i) > f(X_j) \\ 1/2, & \text{se } f(X_i) = f(X_j) \\ 0, & \text{se } f(X_i) < f(X_j). \end{cases}$$

Além disso, AUC é uma medida usualmente utilizada, em detrimento da Acurácia, para avaliar modelos de aplicações que envolvem desbalanceamento dos dados. No Capítulo 4 serão abordadas outras métricas para modelos de classificação e discutidas com relação ao desempenho quando utilizadas em contexto de desbalanceamento.

3. Técnicas para validação de modelos

Quando se tem interesse em validar um modelo é essencial avaliá-lo quanto a capacidade de predição. Métodos que envolvem reamostragem são ferramentas indispensáveis para a estatística moderna e amplamente utilizadas no contexto de validação de modelos.

Tendo em vista o que foi abordado no capítulo anterior com relação à estimação do Risco Esperado e a vantagem em considerar observações inéditas específicas para validação, é possível justificar a necessidade do uso de técnicas que possibilitem a divisão da amostra original em amostra de treino e de validação. Sendo assim, a seguir serão abordadas algumas destas técnicas tal como tratado por Cunha [2019].

3.1 Hold-out

Hold-out é um método de validação que consiste na separação das n observações que compõem a amostra original $\mathcal{D}$ em duas partes, uma amostra $\mathcal{D}_t$ de tamanho t para treino e outra amostra $\mathcal{D}_v$ de tamanho v para validação, de tal forma que $n = v + t$. A amostra selecionada para validação representa uma fração p da amostra original de modo que $v = n \cdot p$ e, consequentemente, $t = n \cdot (1-p)$, usualmente o valor considerado é $p = \frac{1}{3}$. Para este método de validação o Risco Esperado, definido em (2.1), pode ser estimado por

$$err_{hop} = \frac{1}{v} \sum_{i=1}^{v} L(y_i, \hat{f}_t(x_i)), \tag{3.1}$$

em que $\hat{f}_t(\cdot)$ representa a função estimada a partir da amostra de treino $\mathcal{D}_t$. Além disso, note que é avaliada, por meio da função perda, a qualidade do ajuste com relação a todas as observações (y_i, X_i) contempladas pela amostra de validação $\mathcal{D}_v$.

Kohavi [1995] menciona que quanto maior o tamanho da amostra de validação, maior o viés atrelado à estimativa obtida por (3.1), no entanto menores amostras de validação implicam em maior variabilidade do estimador (3.1).

Borra e Ciaccio [2010] destacam que os resultados obtidos a partir do *Hold-out* acabam sendo muito dependentes da seleção inicial da amostra de treino e de validação. Uma solução para reduzir a dependência mencionada consiste em repetir o procedimento de separação da amostra uma quantidade R de vezes de modo que em cada repetição r é obtida uma amostra para validação $\mathcal{D}_{vr}$ e outra para treino $\mathcal{D}_{tr}$ distintas. Este procedimento é conhecido por *Repeated Hold-out* e a estimativa do Risco Esperado é definida por

$$err_{rhop} = \frac{1}{R}\sum_{r=1}^{R}\frac{1}{v}\sum_{i=1}^{v} L(y_{ir}, \hat{f}_{tr}(x_{ir})), \tag{3.2}$$

em que $\hat{f}_{tr}(\cdot)$ é a função estimada com base na r-ésima amostra de treino $\mathcal{D}_{tr}$ selecionada e que é avaliada na amostra de validação $\mathcal{D}_{vr}$ análoga.

Kohavi [1995] aponta que ao utilizar o *Repeated Hold-out* a desvantagem é que a suposição de independencia entre as observações da amostra de treino e validação é violada. Além disso, conforme são realizadas as repetições, parte dos dados podem ser subrepresentados.

3.2 K-fold

No método de validação *K-fold* as n observações da amostra original $\mathcal{D}$ são divididas em K conjuntos disjuntos de observações, sendo eles $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_K$, cada um de tamanho m_k aproximadamente igual, tal que $n = \sum_{k=1}^{K} m_k$. A partir disso, a amostra de validação é composta pela partição $\mathcal{D}_k$ enquanto que a amostra de treino engloba as outras $K-1$ partições que não incluem a k-ésima partição, ou seja, o conjunto de treino é dado por $\mathcal{D}_{(-k)} = \{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_{k-1}, \mathcal{D}_{k+1}, \ldots, \mathcal{D}_K\}$. Este processo, conforme esquema da Figura 3.1, é repetido iterativamente K vezes até que cada uma das $k = 1, 2, \ldots, K$ partições da amostra seja considerada como amostra de validação.

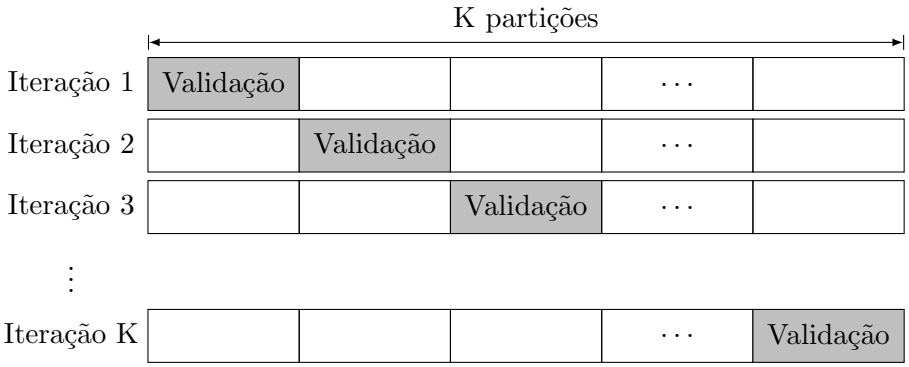


Figura 3.1: *Esquema representativo das K iterações inerente ao procedimento K-fold*
Fonte: Elaborada pela autora.

No *K-fold* o Risco Esperado como definido em (2.1) pode ser estimado por

$$err_{kfK} = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{m_k}\sum_{i=1}^{m_k} L(y_{ik}, \hat{f}_{(-k)}(X_{ik})), \tag{3.3}$$

de tal maneira que $\hat{f}_{(-k)}(\cdot)$ representa o modelo definido por meio da amostra de treino $\mathcal{D}_{(-k)}$ [Borra e Ciaccio, 2010]. Ademais, a função perda é avaliada em todas as m_k observações contempladas pela amostra de validação $\mathcal{D}_{(k)}$ em cada uma das K iterações inerentes ao processo de validação *K-fold*.

O número K de vezes em que a amostra será repartida precisa ser pré-estabelecido. Borra e Ciaccio [2010] mencionam que o viés atrelado às estimativas do erro obtidas pelo método *K-fold* será menor para valores maiores de K. Um maior valor de K implica em maior amostra de treino que, usualmente, está atrelada a maiores variâncias, além disso, são necessárias um maior numero de iterações que acarretam em maior custo computacional.

Note que a cada iteração é determinada uma nova amostra de treino, no entanto, as observações desta amostra não são completamente inéditas com relação às demais. O problema mencionado é um dos principais para este método de validação e implica em amostras de treino que não são independentes entre si e acabam ocasionando uma variabilidade no estimador (3.3) que pode ser elevada [Borra e Ciaccio, 2010].

Conforme mostrado por Burman [1989], o estimador err_{kfK} (Equação 3.3) apresenta viés considerável, principalmente, em situações nas quais o valor estabelecido para K é pequeno. Além disso, para modelos com uma quantidade elevada de parâmetros este viés pode ser ainda maior [Borra e Ciaccio, 2010].

Posto isto, Burman [1989] propõe a seguinte correção para este estimador

$$err_{bkfK} = err_{kfK} + err - err_{kf^+}, \tag{3.4}$$

em que err_{kfK} é o estimador para o método *K-fold* definido em (3.3), err é o erro estimado para a amostra como um todo, desconsiderando qualquer processo de validação, tal como definido na Equação (2.2) e err_{kf^+} diz respeito ao erro calculado para a amostra completa a partir de um modelo ajustado com base nas $\mathcal{D}_{(-k)}$ amostras de treino, ou seja, err_{kf^+} é definido por

$$err_{kf^+} = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{n}\sum_{i=1}^{n} L(y_i, \hat{f}_{(-k)}(X_i)), \tag{3.5}$$

de tal forma que $\hat{f}_{(-k)}(\cdot)$ é a função determinada por meio da amostra de treino $\mathcal{D}_{(-k)}$ e apresenta

função perda avaliada em todas as n observações da amostra trabalhada.

Para o método *K-fold*, assim como para o método *Hold-out*, também é possível pensar em uma versão em que ocorrem repetições. Rodriguez *et al.* [2009] mencionam que esta versão com repetições do *K-fold* pode estabilizar o processo de estimação do erro provocando uma redução na variância do estimador. O método denominado por *Repeated K-fold* consiste, então, em realizar o procedimento usual do *K-fold* um número R de vezes e, por fim, obter a média dos erros estimados em cada uma dessas repetições [Burman, 1989]. Sendo assim, o erro para o *Repeated K-fold* é dado por

$$err_{rkfK} = \frac{1}{R}\sum_{r=1}^{R} err_{kfK}^{(r)}, \tag{3.6}$$

de modo que $err_{kfK}^{(r)}$ representa o erro calculado conforme Equação (3.3) para a r-ésima repetição do processo *K-fold*.

Burman [1989] aponta para a possibildade de ser aplicada a correção para cada uma das repetições realizadas na execução do *Repeated K-fold*. Assim, de maneira análoga, a estimativa do erro para o *Repeated K-fold* em uma versão corrigida pode ser obtida por

$$err_{rbkfK} = \frac{1}{R}\sum_{r=1}^{R} err_{bkfK}^{(r)}, \tag{3.7}$$

em que $err_{bkfK}^{(r)}$ e o valor do erro, para a r-ésima repetição, estimado por (3.4).

3.3 Leave-one-out

O método de validação *Leave-one-out* pode ser visto como um caso particular do *K-fold* em que cada partição da amostra possui apenas uma observação, ou seja, $K = n$. Sendo assim, são obtidas a cada iteração uma amostra de treino $\mathcal{D}_k$ de tamanho $n-1$ e uma amostra de validação $\mathcal{D}_{(-k)}$ de tamanho 1 em que é possível estimar o Risco Esperado (2.1) como

$$err_{loo} = \frac{1}{n}\sum_{k=1}^{n} L(y_k, \hat{f}_{(-k)}(X_k)), \tag{3.8}$$

sendo $\hat{f}_{(-k)}(\cdot)$ o modelo definido a partir das observações com exceção à k-ésima. Além disso, nesse caso, a função perda é avaliada nas n iterações para a única observação k que foi deixada de fora da amostra de treino.

Borra e Ciaccio [2010] mencionam que o estimador obtido por meio do *Leave-one-out*, apesar de apresentar variabilidade elevada, é aproximadamente não viesado. Tendo em vista, ainda, que o número de iterações realizadas no método *Leave-one-out* são tantas quanto o tamanho do conjunto de dados considerado o custo computacional atrelado à execução do método acaba sendo elevado.

3.4 Bootstrap

O *Bootstrap* é uma técnica comumente utilizada para quantificar a incerteza associada a estimadores ou métodos de aprendizagem estatística, no entanto pode ser aplicada em diversos outros contextos. No âmbito da validação de modelos, dentre as n observações disponíveis na amostra original $\mathcal{D}$, são selecionados aleatoriamente e com reposição K conjuntos, sendo eles $\mathcal{B}_1, \mathcal{B}_2, \ldots, \mathcal{B}_K$, cada um de tamanho m. Dessa forma, a partição $\mathcal{B}_k$ é definida como amostra de treino e as demais observações que não compõem $\mathcal{B}_k$ são alocadas na amostra de validação, isto é, o conjunto de validação é definido por $\mathcal{B}_{(-k)} = \{(X_{ik}, y_{ik}) \notin \mathcal{B}_k\}$ e possui tamanho m_k.

Note que, por serem selecionadas com reposição, as observações em $\mathcal{B}_k$ podem aparecer mais de uma vez o que implica em distintos tamanhos de m_k e em $m + m_k \neq n$. Este processo de partição é repetido até que cada uma das K amostras retiradas inicialmente sejam consideradas como amostra de treino. O Risco Esperado (2.1) pode ser obtido por

$$err_{bt} = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{m_k} \sum_{i=1}^{m_k} L(y_{ik}, \hat{f}_k(X_{ik})), \tag{3.9}$$

em que $\hat{f}_k(\cdot)$ é o modelo definido a partir da amostra de treino $\mathcal{B}_k$ da k-ésima iteração. Além disso, a função perda é avaliada em todas as m_k observações da amostra de validação $\mathcal{B}_{(-k)}$.

Apesar do estimador obtido por meio do método *Bootstrap* lidar bem com amostras pequenas devido à baixa variabilidade, é uma técnica que exige muito computacionalmente [Kim, 2009]. Além disso, o estimador do Risco Esperado, tal como proposto em (3.9), apresenta resultados superestimados. O viés deste estimador está atrelado ao fato de que, no método *Bootstrap*, a probabilidade de uma observação i pertencer a amostra de treino $\mathcal{B}_k$ é de $0,632$ para todo k, conforme detalhado a seguir:

$$\mathrm{P}(i \in \mathcal{B}_k) = 1 - \mathrm{P}(i \notin \mathcal{B}_k) = 1 - \left(1 - \frac{1}{m}\right)^m \approx 1 - e^{-1} = 0,632.$$

Dessa forma, espera-se a inclusão de 63,2% das observações da amostra original na amostra de treino e 36,8% na amostra de validação. Para tratar este viés, Efron [1983] propõe versão ponderada

de estimador para o Risco Esperado (2.1) que pode ser calculada por

$$err_{bt632} = 0,632 \cdot err_{bt} + 0,368 \cdot err, \tag{3.10}$$

no qual err_{bt} é calculado conforme Equação (3.9) e err é o estimador Erro Aparente calculado para toda a amostra $\mathcal{D}$ tal como definido pela Equação (2.2).

A ideia por trás da proposta de Efron [1983] é corrigir a estimativa superviesada de err_{bt} por meio da ponderação simultânea da estimativa subviesada err. Esta correção, no entanto, não se mostra eficaz em modelos sobreajustados, ou seja, em que $err = 0$. Neste cenário, a estimativa obtida pela Equação (3.10) se torna subviesada.

Inicialmente, busca-se definir métrica conhecida por taxa de erro não informativo que faz referência ao erro esperado para X e Y na presença de independência. Esta medida, simbolizada por γ pode ser estimada por

$$\hat{\gamma} = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} L(y_i, \hat{f}(X_j)),$$

em que $i, j = 1, 2, \ldots, n$, a função perda é avaliada para todos os pares da amostra e $\hat{f}(\cdot)$ se refere ao modelo definido a partir da amostra completa $\mathcal{D}$.

Sendo assim, Efron [1983] apresentam estimador alternativo, que atribui maior peso para err_{bt} na presença de superajuste. Este peso é denominado Taxa Relativa de Superajuste e é definido por

$$\hat{R} = \frac{err_{bt} - err}{\hat{\gamma} - err},$$

resultando em valores entre 0 e 1, de tal maneira que quanto mais próximo de 1 maior o sobreajuste do modelo.

A partir disto, é definido o estimador err_{bt632+} conforme proposto por Efron e Tibshirani [1995] e exposto a seguir:

$$err_{bt632+} = \hat{w} \cdot err_{bt} + (1 - \hat{w}) \cdot err, \tag{3.11}$$

em que err_{bt} é o estimador definido pela Equação (3.9) e err é o estimador Erro Aparente conforme Equação (2.2). O $\hat{w}$ é calculado por

$$\hat{w} = \frac{0,632}{1 - 0,368 \cdot \hat{R}},$$

de tal forma que, quando se trata de um modelo extremamente superajustado $\hat{R} = 1$ e, consequen-

temente, $\hat{w} = 1$, fazendo com que o estimador err_{bt632+} seja equivalente à err_{bt} (3.9). No caso de $\hat{R} = 0$, é obtido $\hat{w} = 0,632$ e o estimador err_{bt632+} equivale à err_{bt632} (3.10).

4. Dados desbalanceados

A problemática do desbalanceamento surge quando as diferentes categorias das variáveis não estão igualmente representadas na base de dados, nem mesmo de maneira aproximada [Raeder *et al.*, 2012].

Quando se tratam de eventos raros verifica-se dificuldade na construção de modelos para classificação. Este problema, de acordo com o que foi abordado por López *et al.* [2013], está relacionado ao maior interesse na predição da classe minoritária que pode estar associada a casos significantes excepcionais ou que pode ser restrita, simplesmente, por causa da dificuldade em obtenção de observações em que o evento ocorra. Além disso, a maioria dos métodos usuais para construção de modelos consideram uma amostra balanceada, o que pode implicar na escolha de modelos "sub ótimos" que predizem bem somente as observações da classe majoritária [López *et al.*, 2013].

Branco *et al.* [2016] indicam a necessidade de adaptar o processo de obtenção do modelo por meio de estratégias adequadas ao desbalanceamento. Junto a isso, torna-se imprescindível considerar métricas adequadas para avaliação destes modelos.

4.1 Métricas de desempenho

Escolher a métrica de avaliação apropriada desempenha papel chave na tarefa de lidar corretamente com dados desbalanceados. Critérios tradicionais costumam priorizar classes mais frequentes e podem não levar a modelos de classificação ótimos, o que, geralmente, implica em conclusões enganosas priorizando a classe majoritária [López *et al.*, 2013].

A seguir serão comentadas, dentro do contexto do desbalanceamento, características de algumas das métricas já mencionadas na Seção 2.2. Além disso, serão apresentadas métricas que surgiram ou são populares em decorrência dos desafios de mensurar a performance de um modelo na presença de um evento raro.

Métricas Escalares

Apesar da Acurácia (Equação 2.2) ser considerada a métrica mais geral e intuitiva, quando se tratam de dados desbalanceados acaba não sendo uma boa escolha [Haixiang *et al.*, 2017]. Isso é decorrente da característica da acurácia em não considerar a distribuição dos dados, o que acaba priorizando modelos que usualmente não apresentam boa performance na classificação de observações pertencentes a classe minoritária [Kotsiantis *et al.*, 2006].

Haixiang *et al.* [2017] indicam que a Precisão, Sensibilidade e Especificidade (Tabela 2.1) são medidas utilizadas com certa frequência para dados desbalanceados. Já foi abordada na Seção 2.2, no entanto, a dificuldade em realizar um monitoramento simultâneo destas métricas, o que acaba complicando a utilização delas como mensuração da performance de um modelo.

Assim como outras métricas, o F1-score (Equação 2.2) se apresenta como uma alternativa para a dificuldade de monitoramento simultâneo mencionada. Esta medida é muito utilizada e se mostra mais informativa do que a Acurácia no que diz respeito à capacidade de um modelo classificar corretamente a classe minoritária [Branco *et al.*, 2016].

Forman e Scholz [2010] alertam para a possibilidade do F1-score ser indefinido como consequência da indefinição da Precisão ou Sensibilidade. A Precisão será indefinida caso não sejam preditos valores positivos ($VP + FP = 0$), situação possível quando se tratam de amostras de validação pequenas ou desbalanceadas e modelos que sejam extremamente conservadores. A Sensibilidade será indefinida se na base de dados utilizada para validação do modelo não existirem observações positivas ($VP + FN = 0$), o que pode ocorrer em bases de dados extremamente desbalanceadas em que a amostra de validação é selecionada de maneira aleatória [Forman e Scholz, 2010].

A seguir serão exploradas medidas, ainda não mencionadas na Seção 2.2, usuais quando se tratam de dados desbalanceados. Estas medidas não esgotam as possibilidades de métricas para bases de dados que contemplem eventos raros. A saber: G-média (*G-mean*), G-média ajustado (*Adjusted G-mean*), Acurácia Média Balanceada por Classe (*Mean Class-Weighted Accuracy* - CWA) e Coeficiente de Correlação de Matthews (*Matthews Correlation Coefficient* - MCC).

G-mean

A métrica *G-mean* diz respeito à média geométrica da taxa de acerto em cada uma das classes e pode ser dada por:

$$\mathrm{GM} = \sqrt{\frac{\mathrm{VP}}{\mathrm{VP}+\mathrm{FN}} \times \frac{\mathrm{VN}}{\mathrm{VN}+\mathrm{FP}}} = \sqrt{\text{sensibilidade} \times \text{especificidade}}.$$

A ideia está em maximizar a sensibilidade e especificidade de maneira simultânea e equilibrada. Branco *et al.* [2016] comentam que apesar desta ter sido uma medida concebida para mensurar a performance em situações específicas de desbalanceamento, porém ela acaba atribuindo igual importância para os acertos de cada classe. Sendo assim, foi motivada a busca por medidas capazes de informar a contribuição dos acertos de cada classe na performance final do modelo [López *et al.*, 2013].

G-mean Ajustado

Pensando no problema mencionado o *G-mean* Ajustado surge como alternativa ao *G-mean* convencional. Esta medida busca obter a maior sensibilidade possível sem que a especificidade seja muito reduzida [López *et al.*, 2013]. A medida é definida por:

$$\text{GMA} = \begin{cases} \frac{\text{GM}+\text{VP}_{\%}\cdot(\text{FP}+\text{VN})}{1+\text{FP}+\text{VN}}, & \text{se VP}_{\%} > 0 \\ 0, & \text{se VP}_{\%} = 0. \end{cases}$$

Note que, se nenhuma obervação for classificada corretamente na classe positiva, o *G-mean* Ajustado é igual a zero, o que representaria o pior cenário possível de acordo com a métrica.

Mean Class-Weighted Accuracy (CWA)

Pensando nas limitações do *F1-score* em considerar o desempenho do modelo também na classe negativa e do *G-mean* em permitir atribuir maior peso a erros na classe minoritária, surge a métrica *CWA* [Branco *et al.*, 2016]. A construção desta medida é baseada na possibilidade de considerar simultaneamente a sensibilidade e a especificidade permitindo que sejam atribuídos pesos para estas medidas conforme as características do problema tratado. O *CWA* é calculado por:

$$\text{CWA} = w \cdot \text{sensibilidade} + (1 - w) \cdot \text{especificidade},$$

de tal forma que $0 \leq w \leq 1$ e que melhor será o modelo quanto maior for o valor do *CWA*.

Coeficiente de Correlação de Matthews (MCC)

Outra métrica interessante é o MCC que surge como mais uma alternativa de medida não afetada pelo desbalanceamento presente nos dados [Seiffert *et al.*, 2014]. O MCC pode ser obtido por:

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

podendo resultar em valores entre -1 e 1, em que -1 ocorre quando todas as observações são classificadas de forma incorreta e $+1$ quando todas as observações são classificadas de forma correta. A obtenção de um MCC $= 0$ representa uma situação em que a classificação possui desempenho análogo ao de uma classificação completamente aleatória [Seiffert *et al.*, 2014].

Métricas baseadas em gráficos

AUC (Equação 2.2), medida motivada pela curva ROC (Figura 2.2), é uma medida comumente utilizada, em detrimento da Acurácia, para avaliar modelos de aplicações que envolvem desbalanceamento dos dados.

4.2 Estratégias para lidar com o desbalanceamento

As estratégias utilizadas no processo de construção de modelos para dados desbalanceados são divididas por Branco *et al.* [2016] em quatro tipos que serão detalhados mais a frente, são eles:

(i) Alteração nos dados antes de iniciar o processo de aprendizado, pré-processamento dos dados;

(ii) Utilização de métodos de aprendizado especiais;

(iii) Transformações aplicadas nas predições do modelo utilizado;

(iv) Estratégias híbridas que combinam diferentes estratégias de (i) a (iii).

Vale destacar que a efetividade e aplicabilidade de cada uma destas estratégias vai depender das características, informações e, principalmente, necessidades atreladas ao problema tratado. Pode ser que a utilização de uma única estratégia já seja suficiente para contornar a dificuldade oriunda do desbalanceamento em determinada base de dados [Branco *et al.*, 2016].

Pré-processamento dos dados

Estratégias que envolvem o pré-processamento dos dados consistem na busca por mecanismos que possibilitem a utilização dos dados disponíveis de forma a priorizar as necessidades atreladas ao problema e à identificação da classe minoritária. Dessa forma, o ajuste do modelo não é mais realizado na amostra de treino originalmente selecionada e sim na base de dados que resulta da

aplicação de técnicas de pré-processamento nesta amostra de treino. Estas técnicas podem ser divididas entre aquelas que buscam realizar alteração na distribuição dos dados e aquelas que procuram ponderar o espaço de dados [Branco *et al.*, 2016].

Estratégias que realizam alteração na distribuição dos dados consistem na busca por mecanismos que possam alterar a distribuição da variável resposta na base desbalanceada original visando à possibilidade de modelos convencionais focarem nas observações de maior relevância para o problema, aquelas pertencentes à classe minoritária. López *et al.* [2013] apontam que métodos que buscam balancear a distribuição dos dados costumam ser soluções úteis. São eles:

1. Métodos *Undersampling*: alteram a amostra de treino original eliminando algumas das observações, usualmente as que pertencem à classe majoritária, entretanto podem acarretar na perda de alguma informação ou relação importante;

2. Métodos *Oversampling*: alteram a amostra de treino original replicando ou criando novas observações a partir daquelas pré-existentes na amostra, usualmente as observações criadas são pertencentes à classe minoritária, podem implicar em *overfitting* do modelo;

3. Métodos híbridos: combinam as duas estratégias citadas anteriormente.

Uma outra possibilidade para realizar o pré-processamento dos dados envolve a alteração da amostra de treino com base no custo atrelado a cada tipo de classificação incorreta, ou seja, seria realizada uma ponderação do espaço de dados. O modelo escolhido, então, acabaria por evitar erros associados a perdas elevadas [Branco *et al.*, 2016].

Métodos de aprendizado especiais

Métodos de aprendizado especiais dizem respeito a soluções que modificam métodos de aprendizado pré-existentes para melhor atender às particularidades do problema tratado ou até mesmo concepção de técnicas específicas para problemas em que ocorre desbalanceamento. A utilização destes modelos é vantajosa por possibilitar a incorporação de necessidades e peculiaridades do problema diretamente no modelo o que, em geral, implica em modelos melhor compreensíveis e interpretáveis [Branco *et al.*, 2016].

A tarefa de incluir as nuances de cada problema tratado no processo de ajuste do modelo pode não ser tão simples. Além disso, no caso de realizar, por exemplo, a troca da função perda considerada o processo de seleção do modelo pode precisar ser refeito [Branco *et al.*, 2016].

Pós-processamento das predições

Outra possibilidade para tratar o problema do desbalanceamento é a execução de alguma forma de correção das predições obtidas a partir do modelo ajustado. Este tipo de tratamento possibilita que modelos usuais de aprendizado sejam utilizados [Branco *et al.*, 2016].

Kotsiantis *et al.* [2006] apontam como uma das estrategias para lidar com o desbalanceamento, por exemplo, a variação do limiar c para realizar a classificação das observações após a predição da probabilidade de pertencimento à classe minoritária. Dessa maneira, seria possível obter, a partir de um mesmo modelo, classificações distintas que implicariam em desempenhos distintos.

Métodos híbridos

Estratégias híbridas envolvem o uso, de forma combinada, de algumas das estratégias comentadas anteriormente. A ideia por trás de técnicas deste tipo está na tentativa de aproveitar as principais vantagens obtidas por meio de cada um dos métodos mencionados [Branco *et al.*, 2016]

4.3 Problemas que dificultam a modelagem preditiva

Além das dificuldades já mencionadas no processo de predição de um evento raro, existem problemas intrínsecos aos dados que podem contribuir para que seja ainda mais difícil obter predições de qualidade [Branco *et al.*, 2016]. Estes problemas podem aparecer de maneira isolada ou de forma simultânea [López *et al.*, 2013]. Serão abordados brevemente os problemas de Sobreposição de Classes, Amostras Pequenas, Dados Superdimensionados, Dados Ruidosos, Pequenos Disjuntos e *Data Shift*.

Sobreposição de Classes

O problema da Sobreposição de Classes ou da separabilidade das classes ocorre quando determinada região do espaço de dados é composta, na amostra de treino, por uma quantidade igual ou similar de observações das classes positiva e negativa [López *et al.*, 2013]. Esta circunstância faz com que se torne difícil a obtenção de um modelo capaz de distinguir bem a que classe pertencem as observações nesta região de sobreposição [Branco *et al.*, 2016].

Branco *et al.* [2016] apontam que a combinação entre desbalanceamento e Sobreposição de Classes implica em maiores dificuldades no processo de aprendizado estatístico do que quando tratados individualmente. A relação entre estes dois dificultadores vem sendo estudada e algumas

soluções tem sido propostas para lidar simultaneamente com estes problemas [Alejo *et al.*, 2013, García *et al.*, 2006, Prati *et al.*, 2004].

Amostra Pequena

Amostras de treino pequenas acabam se tornando um problema porque a quantidade restrita de dados dificulta a obtenção de modelos que serão capazes de generalizar bem as predições em um contexto de novas observações [López *et al.*, 2013]. Quando este problema aparece associado ao desbalanceamento dos dados o tratamento se torna ainda mais complicado, tendo em vista que a presença de poucos exemplos da classe minoritária dificulta a obtenção de modelos capazes de capturar as características desta classe [Branco *et al.*, 2016].

Nesta situação não há dados o suficiente que possibilitem a compreensão das fronteiras de separação das classes o que se torna ainda mais significativo quando a concentração de exemplos da classe minoritária é tão baixa que pode ser tratada como ruído [López *et al.*, 2013]. López *et al.* [2013] destacam, ainda, que o aprendizado pode ser prejudicado não necessariamente pelo baixo percentual de observações positivas e sim pelo valor total de observações positivas que se torna limitado em decorrência do tamanho amostral.

Dados Superdimensionados

Os dados são superdimensionados quando possuem uma quantidade elevada de potenciais variáveis explicativas. Usualmente é realizada seleção de um subconjunto destas variáveis focando naquelas que melhor contribuiriam para a compreensão da variável resposta. O desafio está justamente no critério a ser utilizado para selecionar esta menor quantidade de variáveis de forma a não perder relações importantes para o problema [Branco *et al.*, 2016].

Quando se tratam de dados desbalanceados, a seleção de variáveis deve ser pensada com cautela. É importante levar em consideração técnicas que tratem o desbalanceamento dos dados de maneira adequada [Chu *et al.*, 2010, Forman, 2003, Wasikowski e Chen, 2009, Zheng *et al.*, 2004].

Dados Ruidosos

São considerados Dados Ruidosos observações que, por algum motivo, são distorcidas, incorretas ou corrompidas. Seiffert *et al.* [2014] apontam que este problema é ainda mais grave quando ocorre em relação a classe que a observação pertence, ou seja, a real classe daquela observação não equivale a classe registrada na base de dados. Neste caso, a utilização de modelos robustos para classificação é uma estratégia inicial para tentar contornar problema dos ruídos presentes [Seiffert *et al.*, 2014].

No caso de ser necessário lidar com dados ruidosos na presença de desbalanceamento o impacto deste ruído pode ser ainda maior para a classe minoritária do que em outros casos [López *et al.*, 2013]. O que López *et al.* [2013] abordam é que, por possuir uma quantidade menor de observações, uma quantidade pequena de observações ruidosas seria suficiente para atrapalhar o processo de aprendizado estatístico das características daquela classe.

Usualmente os modelos são mais sensíveis a dados ruidosos do que ao desbalanceamento. Seiffert *et al.* [2014], no entanto, apontam que conforme a prevalência da classe minoritária é reduzida ocorre impacto negativo considerável na performance do modelo. Além disso, mostram que técnicas de pré-processamento que reamostram os dados apresentam bom desempenho ao lidar com ruídos e desbalanceamento.

Pequenos Disjuntos

A problemática dos Pequenos Disjuntos surge quando é notada, dentro de uma das categorias da variável resposta estudada, a existência de pequenas subdivisões com relação aos valores atribuídos às variáveis explicativas, sub-conceitos pouco representados [López *et al.*, 2013]. Ou seja, Pequenos Disjuntos estão atrelados a um pequeno grupos de observações, dentro de uma mesma classe, em que a variável resposta estaria relacionada às variáveis explicativas de uma maneira muito específica e distinta das demais observações.

López *et al.* [2013] destacam que este é um problema que aparece com frequência no processo de construção de modelos, no entanto, a presença simultânea com o desbalanceamento faz com que se torne um problema ainda mais complexo. Mencionam, ainda, que quando o evento estudado é considerado raro se torna difícil distinguir se aquele pequeno conjunto de observações com comportamento diferenciado diz respeito a observações legítimas a serem tratadas ou se são observações ruidosas.

Data Shift

O problema de *Data Shift* faz referência a situações em que a distribuição dos dados nas amostras de treino e de validação são distintas. Esta distinção costuma ser frequente e não acarreta em prejuízos consideráveis no que tange à performance do modelo. Branco *et al.* [2016], no entanto, apontam que quando se tratam de bases de dados altamente desbalanceadas o *Data Shift* pode provocar perdas intensas na performance do modelo.

De acordo com López *et al.* [2014], é possível dividir o problema de *Data Shift* em três tipos:

1. *Prior Probability Shift*: diz respeito a situações em que a distribuição das classes da variável resposta são distintas entre amostra de treino e amostra de validação. Para casos mais extremos pode implicar em uma amostra de treino que não possui uma das classes da variável;

2. *Covariate Shift*: neste caso, a diferença entre amostras de treino e de validação está na distribuição das variáveis explicativas;

3. *Concept Shift*: ocorre quando a relação entre a variável resposta e as variáveis explicativas muda de uma amostra para outra.

4.4 Validação de modelos

Quando se trata da validação de modelos construídos com o intuito de realizar a predição de eventos raros é possível pensar em algumas adaptações com relação ao que foi apresentado no Capítulo 3. Estas alterações podem ser pensadas desde a maneira com que a partição entre amostra de treino e validação é feita em cada um dos métodos até a métrica que é escolhida para subsidiar o processo de validação.

A seguir serão abordadas técnicas para a seleção das amostras de validação e treino que sabidamente minimizam problemas de *Data Shift* do tipo *Prior Probability Shift*, já mencionado (Seção 4.3), inerente ao processo de modelagem de dados desbalanceados. Além disso, será explicitado de que maneira as métricas *F1-score* e AUC podem ser calculadas concomitantemente a alguns dos possíveis processos de validação.

4.4.1 *Hold-out* Estratificado

Para problemas em que a variável resposta se trata de uma variável categórica pode ser interessante que a distribuição da variável resposta nas amostras de treino $\mathcal{D}_t$ e de validação $\mathcal{D}_v$ se assemelhem à distribuição verificada na amostra original $\mathcal{D}$. Quando se trata de um cenário em que o evento de interesse é considerado raro é especialmente importante garantir que as amostras de treino e de validação contemplem ambas as categorias da variável modelada.

Neste contexto, é possível a utilização do *Hold-out* Estratificado que consiste em realizar a seleção da amostra de validação de forma a selecionar aleatoriamente observações dentro de cada um dos estratos, $y = 1$ e $y = 0$. Sendo assim, torna-se possível garantir a presença de eventos e não eventos nas amostras tal como controlar a prevalência da classe minoritária em cada uma delas para que se assemelhe a da amostra original.

4.4.2 *K-fold* Estratificado

A maneira com que são selecionadas as observações para compor as partições da amostra e execução da validação impactam nas estimativas da performance do modelo [López *et al.*, 2014]. Conforme já mencionado, garantir que a distribuição dos dados seja semelhante em amostra de treino e de validação é uma estratégia interessante.

O *K-fold* Estratificado é uma maneira direta de evitar distribuições distintas da variável resposta nas amostras de treino e validação [López *et al.*, 2014]. Para a versão estratificada do *K-fold* as observações de cada uma das categorias existentes são distribuídas igualmente pelas K partições de forma a garantir que cada uma das partições tenha aproximadamente o mesmo percentual de observações pertencentes a classe minoritária. Por consequência, as amostras de treino e validação também terão distribuição semelhante.

López *et al.* [2014] vão além e propõem adaptação do método *K-fold* que tem como objetivo a obtenção das K partições com distribuições análogas considerando também o conjunto das variáveis explicativas. Essa metodologia é denominada "*Distribution optimally balanced SCV*" (DOB-SCV) e funciona de maneira a alocar as K observações mais semelhantes possíveis, segundo critério dos vizinhos mais próximos, em K partições distintas.

Para o cálculo da medida F1-score no contexto do *K-fold*, independentemente da estratégia utilizada para separação das K partições, Forman e Scholz [2010] recomendam que a estimativa deve ser feita da forma menos viesada possível. A maneira mais adequada, então, consiste em calcular os Verdadeiros Positivos (VP_k), Falsos Positivos (FP_k) e Falsos Negativos (FN_k) em cada uma das $k = 1, 2, \ldots, K$ amostras de validação, para que depois, a partir do valor total destas medidas, seja calculado o F1-score. Ou seja,

$$\mathrm{VP}_{kfK} = \sum_{k=1}^{K} \mathrm{VP}k;$$
$$\mathrm{FP}_{kfK} = \sum_{k=1}^{K} \mathrm{FP}_k;$$
$$\mathrm{FN}_{kfK} = \sum_{k=1}^{K} \mathrm{FN}_k;$$

$$\mathrm{F}_{kfK} = \frac{2 \times \mathrm{VP}_{kfK}}{2 \times \mathrm{VP}_{kfK} + \mathrm{FP}_{kfK} + \mathrm{FN}_{kfK}}.$$

Além disso, Forman e Scholz [2010] apontam que a maneira mais adequada para o cálculo do

AUC no processo de validação *K-fold* consiste em calcular o valor de AUC para as K amostras de validação separadamente para depois realizar o cálculo da média, isto é,

$$\mathrm{AUC}_{kfK} = \frac{1}{K}\sum_{k=1}^{K}\mathrm{AUC}_k,$$

sendo AUC_k o valor de AUC encontrado considerano a k-ésima amostra de validação.

4.4.3 *Leave-one-out*

Quando se trata da validação do tipo *Leave-one-out* não faz sentido pensarmos em estratégias para que a distribuição das amostras de treino e validação sejam similares. Isto se deve a característica deste método de validação que resulta em amostra de validação com apenas uma observação.

O cálculo do F1-Score e AUC pode ser feito de forma análoga ao que foi realizado para o *K-fold*. Ou seja,

$$\mathrm{VP}_{loo} = \sum_{k=1}^{n}\mathrm{VP}_k;$$
$$\mathrm{FP}_{loo} = \sum_{k=1}^{n}\mathrm{FP}_k;$$
$$\mathrm{FN}_{loo} = \sum_{k=1}^{n}\mathrm{FN}_k;$$

$$\mathrm{F}_{loo} = \frac{2\times\mathrm{VP}_{loo}}{2\times\mathrm{VP}_{loo}+\mathrm{FP}_{loo}+\mathrm{FN}_{loo}}.$$

Com relação ao cálculo da métrica AUC, Forman e Scholz [2010] apontam, no contexto do *K-fold*, que partições que não apresentam observações positivas impossibilitam a obtenção do AUC. Pensando no método *Leave-one-out* como um caso particular do *K-fold* ($K = n$), as partições sem valores positivos serão tantas quanto forem o número de observações negativas presentes na amostra.

Dessa forma, o cálculo do AUC se torna viável não mais pelo seu valor médio e sim por meio da interpretação do AUC como probabilidade empírica explicitada na Equação 2.2. Esta abordagem consiste na ordenação dos valores $\hat{f}_{(-k)}(X_k)$ obtidos em cada uma das iterações que vão implicar em uma única curva ROC e, consequentemente, um único valor para AUC denominado AUC_{loo} [Forman e Scholz, 2010].

4.4.4 *Bootstrap* Estratificado

A utilização do método de validação *Bootstrap* na presença de desbalanceamento da variável resposta exige adaptação. Neste caso, é necessário garantir que a distribuição da variável resposta nas amostras de treino e de validação se assemelhem a da amostra original.

Dessa forma, se torna adequada a utilização do método de validação *Bootstrap* em sua versão estratificada. A ideia deste método é a seleção aleatória, com reposição, de observações de cada uma das classes $y = 1$ e $y = 0$. A quantidade de observações selecionadas de cada uma das categorias deve ser tal que a prevalência da classe minoritária na amostra de treino se mantenha a mesma da amostra inicial disponível.

5. Modelagem para Eventos Raros

Assim como já foi discutido, quando se tratam de bases de dados com variável resposta desbalanceada nota-se certa dificuldade na construção de modelos que tenham intuito explicativo ou preditivo da variável resposta. Considerando as particularidades decorrentes do desbalanceamento, King e Zeng [2001] propõem correção do Modelo Logístico baseado na correção do viés dos estimadores e ajuste para redução do vício e erro quadrático médio das probabilidades estimadas. Para uma melhor compreensão da notação, na seção seguinte se apresenta uma breve revisão a respeito dos modelos lineares generalizados.

5.1 Modelos Lineares Generalizados

A classe dos Modelos Lineares Generalizados (MLG), tal como abordado por Paula [2013], é definida de modo que a variável resposta $\mathbf{Y}$ deve pertencer a família exponencial, isto é, a função densidade de cada uma das $Y_1, \cdots, Y_n$ variáveis deve ser do tipo

$$f(y_i; \theta_i, \phi) = \exp[\phi\{y_i\theta_i - b(\theta_i)\} + c(y_i, \phi)], \tag{5.1}$$

o que implica em $\mathrm{E}(Y_i) = \mu_i = b'(\theta_i)$, $\mathrm{Var}(Y_i) = \phi^{-1}V_i$, de modo que $V_i = V(\mu_i) = \dfrac{\partial \mu_i}{\partial \theta_i}$ é a função de variância e $\phi^{-1} > 0$ é o parâmetro de dispersão.

Além de possuir a forma especificada na Equação (5.1), um MLG também é definido pelo preditor linear

$$g(\mu_i) = \eta_i = \mathbf{X}_i^{\intercal}\boldsymbol{\beta},$$

em que $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^{\intercal}$, $p < n$, é um vetor de parâmetros, $\mathbf{X}_i = (x_{i1}, \cdots, x_{ip})^{\intercal}$ são os valores das variáveis explicativas e $g(\cdot)$ é denominada função de ligação que tem como característica ser uma função monótona e diferenciável.

Para ϕ conhecido e respostas independente, o logaritmo da função verossimilhança de um MLG é dado por

$$\mathrm{L}(\boldsymbol{\beta}|\boldsymbol{y}) = \sum_{i=1}^{n} \phi\{y_i\theta_i - b(\theta_i)\} + \sum_{i=1}^{n} c(y_i, \phi).$$

Além disso, a matriz de informação de Fisher para $\boldsymbol{\beta}$ de um MLG é tal que

$$\mathbf{K}_{\beta\beta}(\boldsymbol{\beta}|\boldsymbol{y}) = -\,\mathrm{E}\left(\frac{\partial^2 \mathrm{L}(\boldsymbol{\beta}|y)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\intercal}}\right) = \phi\mathbf{X}^{\intercal}\mathbf{W}\mathbf{X},$$

em que $W = \mathrm{diag}\{\omega_1, \cdots, \omega_n\}$, com $\omega_i = \left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2 \frac{1}{V_i}$, é a matriz de pesos e $\mathbf{X}$ é a matriz de variáveis explicativas $n \times p$ de posto completo em que cada linha é dada por $\mathbf{X_i^{\intercal}}$, $i, \cdots, n$.

Para obtenção das estimativas de $\boldsymbol{\beta}$ é considerado estimador de máxima verossimilhança que é assintoticamente, $n \to \infty$, não viesado. Ou seja, a distribuição assintótica do estimador $\hat{\boldsymbol{\beta}}$ é tal que $\hat{\boldsymbol{\beta}} \sim \mathrm{N}_p(\boldsymbol{\beta}, \mathbf{K}_{\beta\beta}^{-\mathbf{1}}(\boldsymbol{\beta}|\boldsymbol{y}))$.

5.2 Modelo Logístico

O Modelo Logístico é um caso particular de MLG que considera distribuição Binomial para resposta e função de ligação logito $g(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right)$. Ou seja, considere Y_i^* como a proporção de ocorrência de um evento em n ensaios independentes, cada um com probabilidade de ocorrência μ_i, assume-se que $nY_i^* \sim B(n, \mu_i)$ [Paula, 2013]. A função de probabilidade de Y_i^* fica, então, expressa na forma

$$\begin{aligned} f(y_i^*; n, \mu_i) &= \binom{n}{ny_i^*} \mu_i^{ny_i^*} (1-\mu_i)^{n-ny_i^*} \\ &= \exp\left\{\log\binom{n}{ny_i^*} + ny_i^* \log\left(\frac{\mu_i}{1-\mu_i}\right) + n\log(1-\mu_i)\right\}, \end{aligned}$$

em que $0 < \mu_i$, $y^* < 1$. Para reescrever na forma definida em (5.1) é necessário que $\phi = n$, $\theta_i = \log\left(\frac{\mu_i}{1-\mu_i}\right)$, $b(\theta_i) = \log(1 + \exp^{\theta_i})$ e $c(y_i^*, \phi) = \log\binom{\phi}{\phi y_i^*}$, com função de variância $V(\mu_i) = \mu_i(1-\mu_i)$

Quando é considerada função de ligação logito, específica do Modelo Logístico, e fazendo $\mu_i = \mathrm{E}(Y_i^*) = \pi_i$ como a probabilidade de ocorrência do evento, temos por definição que

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \eta_i,$$

e é possível reescrever μ_i como função do preditor linear η_i

$$\mu_i(\eta_i) = \frac{1}{1+e^{-\eta_i}},$$

em que as derivadas de primeira e segunda ordem de $\mu(\eta_i)$ com relação a η_i são tais que

$$\mu_i' = \frac{e^{\eta_i}}{(1+e^{-\eta_i})^2} \text{ e } \mu_i'' = \frac{e^{\eta_i}(1-e^{\eta_i})}{(1+e^{-\eta_i})^3}.$$

Por convenção, para o modelo logístico, assume-se $\phi = 1$ e $\omega_i = n\pi_i(1-\pi_i)$ equivalente à variância da distribuição Binomial. Sendo assim, a matriz de informação de Fisher para o Modelo Logístico é tal que

$$\mathbf{K}_{\beta\beta}(\boldsymbol{\beta}|\boldsymbol{y}) = \mathbf{X}^\intercal \mathbf{W} \mathbf{X},$$

com $\mathbf{W} = diag\{n\pi_1(1-\pi_1), \cdots, n\pi_n(1-\pi_n)\}$

Apesar do Modelo Logístico possuir vasta aplicabilidade, principalmente, por causa da fácil interpretação dos parâmetros, se apresenta como uma solução viesada na presença de eventos raros. A seguir, será abordado proposta realizada por King e Zeng [2001] na tentativa de contornar os problemas associados à utilização de modelo logístico tradicional em bases de dados com resposta desbalanceada.

5.3 Estimador KZ para o Modelo de Regressão Logística

Conforme apontado por King e Zeng [2001], o modelo usual de regressão logística, ao lidar com um evento raro, subestima a probabilidade de ocorrência do evento de interesse. Além disso, é de conhecimento que para amostras finitas $\hat{\boldsymbol{\beta}}$ é um estimador viesado, no contexto de eventos raros, ainda que o tamanho amostral seja razoável, Diniz e Louzada [2013] comentam que os estimadores de máxima verossimilhança permanecem viesados. Dessa forma, King e Zeng [2001] propõem uma estratégia para, na presença de dados desbalanceados, realizar correção nas estimativas dos parâmetros e nas probabilidades preditas. Estes mecanismos são expostos a seguir.

Correção nos parâmetros

Conforme comentado, quando se trata de um evento raro, o estimador de máxima verossimilhança $\hat{\beta}$ é viesado mesmo para grandes amostras. Sendo assim, a ideia é a de construir um novo estimador $\tilde{\beta}$ que seja corrigido pelo vício estimado.

McCullagh e Nelder (1989) abordam que, para qualquer MLG, o vício do estimador do vetor

de parâmetros pode ser estimado por

$$\text{viés}(\hat{\boldsymbol{\beta}}) = \mathbf{K}_{\beta\beta}^{-\mathbf{1}}(\boldsymbol{\beta}|\boldsymbol{y})\mathbf{X}^{\intercal}\mathbf{W}\boldsymbol{\xi},$$

em que $\boldsymbol{\xi}$ é um vetor com o i-ésimo termo $\xi_i = -0,5\frac{\mu_i''}{\mu_i'}Q_{ii}$ com Q_{ii} como o i-ésimo elemento da diagonal principal da matriz $\mathbf{X}\mathbf{K}_{\beta\beta}^{-\mathbf{1}}(\boldsymbol{\beta}|\boldsymbol{y})\mathbf{X}^{\intercal}$, no caso do modelo logístico, $\xi_i = -0,5\frac{1-\exp_i^{\eta}}{1+\exp_i^{\eta}}Q_{ii}$

Dessa forma, para o modelo logístico temos

$$\xi_i = -0,5\left(\frac{1-e^{\eta_i}}{1+e^{\eta_i}}\right)Q_{ii}.$$

Sendo assim, o estimador corrigido pelo vício é tal que

$$\tilde{\beta} = \hat{\beta} - \text{viés}(\hat{\beta}).$$

A matriz de variância e covariância do estimador proposto pode ser escrita de tal forma que

$$V(\tilde{\beta}) = \left(\frac{n}{n+p+1}\right)^2 V(\hat{\beta}),$$

sendo n o tamanho total da amostra considerada para ajuste do modelo e p a quantidade variáveis explicativas que foram levadas em conta neste modelo.

Tendo em vista que $\left(\frac{n}{n+p+1}\right)^2 < 1$ verifica-se que $V(\tilde{\beta}) < V(\hat{\beta})$ o que significa que o estimador em que foi realizada a correção do viés apresentou redução da variabilidade.

Correção no cálculo das probabilidades

Um possível estimador para a probabilidade do evento de interesse $\pi(x_i)$ pode ser obtido a partir de $\tilde{\beta}$, que possui erro quadrático médio inferior a $\hat{\beta}$, de tal forma que

$$\tilde{\pi}(\mathbf{x}_i) = \widehat{\mathrm{P}(Y_i = 1|\tilde{\beta})} = \frac{1}{1+e^{-\mathbf{x}_i\tilde{\beta}}}. \tag{5.2}$$

O estimador $\tilde{\pi}(\mathbf{x_i})$ proposto se mostra mais adequado do que o estimador usual $\hat{\pi}(\mathbf{x_i}) = \widehat{\mathrm{P}(Y_i = 1|\hat{\beta})}$, mas continua sendo um estimador não ótimo por não considerar a incerteza atrelada à estimação de β. Pode-se considerar como incerteza o erro amostral ou o fato do valor de β não ser conhecido e necessitar de estimação.

King e Zeng [2001] indicam que, por desconsiderar a incerteza da obtenção de $\tilde{\beta}$, $\tilde{\pi}(\mathbf{x_i})$ pode continuar gerando estimativas subviesadas da probabilidade $\pi(x_i)$, em geral, bem inferiores a $0,5$.

Se assumirmos que existem valores não observados de uma variável Y^* com densidade logística, $\log\left(\frac{\pi_i}{1-\pi_i}\right) = \eta_i$ que implica em $\mu_i = \frac{1}{1+e^{-\eta_i}}$ tal que, se $Y^* > c$ então $Y = 1$ e se $Y^* \leq c$, $Y^* = 0$. No modelo em questão a probabilidade de um evento é a área sob a curva a direita do limite c, área sombreada mais escura na Figura 5.1. Ignorar a incerteza em relação a β implica em uma distribuição com variância muito pequena e, principalmente para eventos raros, uma área a direita do limiar c também pequena. Incluir na estimação da probabilidade a ideia de incerteza implica em um aumento da variância que, por consequência, amplia a área sob a curva a direita do limiar c gerando maiores estimativas para as probabilidades. Na Figura 5.1 é possível visualizar o aumento da área a direita do limiar c ao considerar a incerteza.

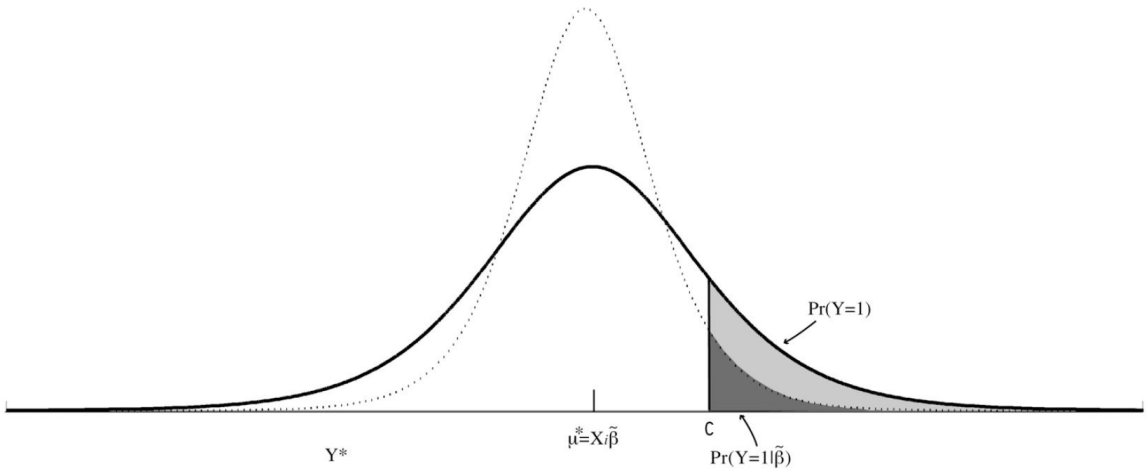


Figura 5.1: *Comparação da distribuição de Y^* considerando β conhecido e igual a $\tilde{\beta}$ (curva pontilhada) e considerando a incerteza atrelada a obtenção do β (curva com linha contínua).*
Fonte: Adaptado de King e Zeng [2001].

Uma maneira de levar em conta a incerteza atrelada a β é considerar uma média dos valores possíveis de tal forma que o calculo da probabilidade esteja condicionado apenas ao valor observado para as variáveis explicativas, ou seja, definimos

$$\pi(\mathbf{x_i}) = \mathrm{P}(Y_i = 1) = \int \mathrm{P}(Y_i = 1|\beta^*)\,\mathrm{P}(\beta^*)d\beta^*, \tag{5.3}$$

de tal forma que $P(\cdot)$ simboliza a incerteza atrelada a β que, em uma perspectiva Bayesiana, pode ser considerada como a densidade a posteriori de $\beta \sim \mathrm{Normal}[\tilde{\beta}, V(\tilde{\beta})]$. Além disso, vale notar que, por uma perspectiva frequentista, a expressão (5.3) pode ser vista como $\mathrm{E}_{\tilde{\beta}}[\mathrm{P}(Y_i = 1|\tilde{\beta})]$

Assim como foi dito, é possível enxergar o estimador definido pela equação (5.3) em uma perspectiva Bayesiana, mas não de maneira completa dado que β é estimado considerando os valores de $\tilde{\beta}$ e $\mathrm{V}(\tilde{\beta})$. Se a informação a priori para β estiver disponível ou for de fácil definição é possível utilizar um método completamente Bayesiano nesta etapa. A estimação Bayesiana de

β com uma priori não informativa é equivalente a estimação usual da regressão logística e, nesse cenário, vimos que $\tilde{\beta}$ é um estimador superior a $\hat{\beta}$.

A expressão (5.3) pode ser calculada por aproximação Monte Carlo ou expandindo em Séries de Taylor. A primeira solução seria obtida ao gerar aleatoriamente valores de β a partir de $\mathrm{P}(\beta)$ e $[1+e^{-\mathbf{x_i}\beta}]^{-1}$ para cada um destes valores, por fim, seria computada a média. A segunda solução é atingida ao realizar a expansão de Taylor até segunda ordem da expressão $\tilde{\pi}(\mathbf{x_i})$ conforme definido em (5.2) em torno de $\tilde{\beta}$ e, em seguida, tomada a esperança, King e Zeng [2001] detalham esse resultado e mostram que o valor obtido é tal que

$$\pi_i(\mathbf{x_i})^* = \mathrm{P}(Y_i = 1) \approx \tilde{\pi}(\mathbf{x_i}) + C_i, \tag{5.4}$$

com

$$C_i = [0,5 - \tilde{\pi}(\mathbf{x_i})]\tilde{\pi}(\mathbf{x_i})[1 - \tilde{\pi}(\mathbf{x_i})]\mathbf{x_i}^{\intercal} V(\tilde{\beta})\mathbf{x_i}.$$

Sendo assim, o estimador com correção para a probabilidade proposto por King e Zeng [2001] é denominado KZ1. Note que, se a incerteza (5.4) atrelada a β for nula o fator de correção C_i será zero, além disso, C_i aumenta conforme a incerteza também aumenta. Outro ponto é que, na presença de alguma incerteza, $[0,5 - \tilde{\pi}(\mathbf{x_i})]$ determina o sinal da constante C_i. No caso de um evento raro, temos $\tilde{\pi}(\mathbf{x_i}) < 0,5$, o que implica em um fator de correção C_i positivo que auxilia na correção de $\tilde{\pi}(\mathbf{x_i})$ que, a princípio, sozinho iria subestimar a probabilidade do evento.

O estimador $\pi_i(\mathbf{x_i})^*$ apresenta menor erro quadrático médio, no entanto os resultados permanecem viesados. Em vista disso, é proposta por King e Zeng [2001] alternativa que tenta de uma vez por todas resolver a problemática do viés. Devido a não linearidade da função logística, apesar do estimador $\tilde{\beta}$ ser aproximadamente não viesado, ou seja, $E(\tilde{\beta}) \approx \beta$, o mesmo não acontece para $\tilde{\pi}_i$ em que $\mathrm{E}(\tilde{\pi}_i)$ não é aproximadamente igual a π_i. De fato, se a equação (5.3) for interpretada como o valor esperado de $\tilde{\beta}$, é possível definir $E_{\tilde{\beta}}(\tilde{\pi}_i) \approx \pi_i + C_i$ de tal forma que o fator de correção C_i pode ser visto como o termo que representa o viés do estimador. Dessa maneira, a subtração do termo de correção faz com que $(\tilde{\pi}_i - C_i)$ seja aproximadamente não viesado enquanto a adição $(\tilde{\pi}_i + C_i)$ implica em um estimador com menor erro quadrático médio.

Denota-se $(\tilde{\pi}_i - C_i)$ como o estimador aproximadamente não viesado e $(\tilde{\pi}_i + C_i)$ como o estimador aproximadamente Bayesiano. Na maior parte dos casos o estimador aproximadamente Bayesiano é mais adequado, mas o estimador aproximadamente não viesado pode ser adequado em casos específicos em que, por exemplo, se tem estudos com uma grande quantidade de bases de dados de

tamanho pequeno a serem combinadas, assim como na meta análise.

Por fim, não há justificativa plausível para se utilizar o estimador tradicional $\hat{\pi}_i$, a menos que a variância dos coeficientes seja aproximadamente nula ou o percentual de eventos observados na amostra seja próximo de 50% (dados aproximadamente balanceados). Nestes casos King e Zeng [2001] mostraram que as vantagens da aproximação proposta é baixo perto do esforço computacional demandado.

6. Estudo de Simulação

É de conhecimento que o método escolhido para validação do modelo pode acarretar em conclusões distintas com relação aos modelos testados [Raeder *et al.*, 2012]. Em vista disso, a partir deste estudo de simulação, busca-se compreender o desempenho de técnicas de validação distintas em diferentes cenários definidos a partir do tamanho de amostra e grau de desbalanceamento da variável resposta em um modelo logístico. O Estudo não contemplou a avaliação do impacto nas métricas de desempenho. O conjunto de dados simulado e os cenários construídos foram realizados com auxílio do *Software R 3.5.3* [R Core Team, 2019].

6.1 Base de dados

A base de dados simulada $\mathcal{B}$ contempla 50 mil observações. A lógica para construção desta base foi tal que, inicialmente, foram gerados valores da variável explicativa $\mathbf{X}_1$ e definidos os valores dos parâmetros da regressão foram obtidos os valores da variável resposta que juntas compõem a base de dados simulada. A maneira na qual este processo ocorreu está descrita em detalhes a seguir.

A variável explicativa única definida segue distribuição normal padrão ($X_1 \sim N(0,1)$), tal como modelo utilizado no estudo de simulação de King e Zeng [2001]. Os valores associados a essa covariável foram obtidos por meio da função do R para geração de valores aleatórios de uma distribuição normal com o comando *rnorm*.

Para a construção da base de dados a ser utilizada no estudo de simulação foi considerada variável resposta com observações independentes e seguindo o modelo logístico dado por,

$$\eta_i = \beta_0 + \beta_1 X_{1i}, \tag{6.1}$$

em que $i = 1, \ldots, 50000$ se refere ao índice da observação e foi fixado $\beta_1 = 1$.

A probabilidade de ocorrência do evento na i-ésima observação é dada por

$$f(X_{1i}) = \frac{e^{\eta_i}}{1 + e^{\eta_i}}.$$

A partir disso, foram considerados diferentes valores fixos conhecidos de β_0 gerando, assim, os valores da variável resposta contidos em $\mathbf{Y} = \{y_i, i = 1, \cdots, 50000\}$ foram gerados pela função *rbinom* do *software* R que possibilitou gerar, aleatoriamente, valores da Distribuições Binomial cujo o parâmetro referente à probabilidade de sucesso de cada observação foi definido tomando como base o resultado $\hat{f}(x_{1i})$ obtido por meio da Equação (6.1). Ou seja,

$$y_i = \text{rbinom}(50000, 1, \hat{f}(x_{1i})), \ (i = 1, \cdots, 50000).$$

A definição do parâmetro β_0, da maneira que o modelo está definido, vai influenciar na probabilidade esperada de eventos na amostra e, consequentemente, no percentual esperado da ocorrência de $Y = 1$. Dito isso, a ideia é gerar oito bases de dados com oito valores distintos de β_0 que irão implicar em oito cenários diferentes de prevalência do evento.

Para possibilitar a geração de bases de dados com diferentes níveis de desbalanceamento, de maneira análoga ao que foi proposto por King e Zeng [2001], foram atribuídos valores distintos ao intercepto do modelo tal como expresso na Tabela 6.1.

Tabela 6.1: *Prevalência do evento na base simulada conforme valor definido para o intercepto.*

$\boldsymbol{\beta_0}$	0	-1	-2	-3	-4	-5	-6	-7
% 1's	$50,27$	$30,58$	$15,52$	$7,05$	$2,77$	$1,09$	$0,42$	$0,138$

Feito isso, foram obtidas 8 bases de dados simuladas com percentuais distintos de eventos. Estas bases de dados contemplam um total de 50 mil observações cada e subsidiaram o estudo de simulação. Para cada uma destas bases de dados foram realizadas todas as etapas do estudo de simulação descritas nas seções seguintes.

6.2 Estimador corrigido para o Modelo de Regressão Logística

A partir da base de dados simulada em que se conhece os valores de β_0 e β_1 do modelo (6.1) considerado para geração dos dados, o intuito é estimar um modelo de regressão logística considerando o que foi proposto por King e Zeng [2001] em relação a correção na estimação dos parâmetros (5.3). Nesse estudo, para a estimação da probabilidade foi considerado o estimador $\tilde{\pi}(x_{1i})$ que não leva em consideração as correções aproximadamente Bayesiana e para ajuste de viés. Esta definição

perpassa pelo interesse do estudo em comparar as técnicas de validações estudadas segundo função perda definida. A avaliação do impacto das correções sugeridas por King e Zeng [2001] para os estimadores de π_i pode ser realizada em trabalho futuro.

Dessa forma, a partir do pacote *zelig* do *software R* foram obtidas as estimativas com correção de viés $\tilde{\beta}_0$ e $\tilde{\beta}_1$. A partir desses valores foi obtida a probabilidade de ocorrência do evento, tal que:

$$\tilde{f}(x_{1i}) = \frac{\exp(\tilde{\beta}_0 + \tilde{\beta}_1 X_{1i})}{1 + \exp(\tilde{\beta}_0 + \tilde{\beta}_1 X_{1i})}.$$

Assim, dado x_{1i}, a ideia é comparar o modelo proposto $\tilde{f}(x_{1i})$ com o modelo original $f(x_{1i})$ a partir da função perda entropia cruzada (2.1). Além disso, buscou-se verificar o Risco Esperado observado e o desempenho das estimativas obtidas por meio das diversas técnicas estudadas e definidas em mais detalhas na seção a seguir.

6.3 Técnicas para validação de modelos

Para a simulação foram consideradas todas as técnicas de validação abordadas no Capítulo 3 na versão estratificada, mais adequada para lidar com desbalanceamento em cada caso, conforme discutido na Seção 4.4. Os estimadores para o Risco Esperado (2.1) considerados nas simulações estão explicitados na Tabela 6.2. Importante ressaltar que todos os métodos foram utilizados considerando as equações destacadas na tabela, mas com os ajustes explicitados na Seção 4.4 no que diz respeito à técnica de seleção de amostra estratificada para que se tornassem mais adequados ao contexto de dados desbalanceados.

Tabela 6.2: *Estimadores estratificados considerados no estudo de simulação.*

Referência		Técnica de Validação	Equação
1.	$\mathbf{err}$	Erro Aparente	2.2
2.	$\mathbf{err_{ho3}}$	*Hold-out* ($p = 1/3$)	3.1
3.	$\mathbf{err_{ho10}}$	*Hold-out* ($p = 1/10$)	3.1
4.	$\mathbf{err_{rho3}}$	*Repeated Hold-out* ($p = 1/3$ e $R = 5$)	3.2
5.	$\mathbf{err_{rho10}}$	*Repeated Hold-out* ($p = 1/10$ e $R = 5$)	3.2
6.	$\mathbf{err_{kf10}}$	*K-fold* ($K = 10$)	3.3
7.	$\mathbf{err_{bkf10}}$	*Burman K-fold* ($K = 10$)	3.4
8.	$\mathbf{err_{rkf10}}$	*Repeated K-fold* ($K = 10$ e $R = 5$)	3.6
9.	$\mathbf{err_{rbkf10}}$	*Repeated Burman K-fold* ($K = 10$ e $R = 5$)	3.7
10.	$\mathbf{err_{loo}}$	*Leave-one-out*	3.8
11.	$\mathbf{err_{bt}}$	*Bootstrap* ($K = 50$ e $m = 2n/3$)	3.9
12.	$\mathbf{err_{bt632}}$	*Bootstrap* Ponderado ($K = 50$ e $m = 2n/3$)	3.10
13.	$\mathbf{err_{bt632+}}$	*Bootstrap* Ponderado ($K = 50$ e $m = 2n/3$)	3.11

Tomando como referência a simulação esquematizada por King e Zeng [2001], será avaliado o

comportamento das diferentes técnicas de validação em amostras de tamanho $n = \{100, 250, 500, 750, 1.000, 1.500, 5.000\}$. O estudo de cada um destes tamanhos de amostra ocorre por meio de simulações de Monte Carlo em que são selecionadas, da base de dados simulada $\mathcal{B}$ com tamanho 50 mil, 100 amostras aleatórias sem reposição de tamanho n pré-estabelecido, o que possibilita a verificação do comportamento das estimativas para cada um dos cenários de tamanho de amostra.

Definido o tamanho n da amostra, as etapas a serem realizadas para cada uma das 100 amostras estratificadas retiradas são descritas a seguir:

1. Ajuste do modelo com correção proposto por King e Zeng [2001] considerando todas as observações contempladas na amostra estratificada de tamanho n sem realizar a separação entre amostra de treino e de validação;

2. A partir do modelo ajustado na primeira etapa, são calculados os valores preditos para Y considerando as observações de uma base de dados de validação grande (25 mil observações) selecionadas de maneira estratificada da base de dados $\mathcal{B}$ e não contempladas na amostra de treino em que o modelo foi ajustado na Etapa 1. Comparando os valores preditos com os verdadeiros valores de Y, por meio da Entropia Cruzada, é obtida uma boa aproximação do Risco Esperado (Equação 2.1);

3. Para a mesma amostra retirada na Etapa 1, é realizada a estimação do Risco Esperado considerando as amostras de validação e teste obtidas por cada uma das técnicas de validação mencionadas na Tabela 6.2, na sua versão estratificada.

Para auxiliar no entendimento das três etapas descritas segue quadro resumo (Figura 6.1) com o esquema realizado para 1 das 100 amostras retiradas para um cenário em que o tamanho da amostra é igual a 250.

Dessa forma, para cada tamanho de amostra, é obtido o Risco Esperado (2.1) a partir da média dos 100 valores resultantes da Etapa 2 descrita anteriormente. Além disso, o viés médio das estimativas obtidas por cada uma das técnicas de validação $j = 1, \ldots, 13$, definidas na Tabela 6.2, pode ser calculado por:

$$\overline{vies_j} = \sum_{i=1}^{100} \frac{err_{ij} - Err_i}{100},$$

em que,

err_{ij} se refere à estimativa do Risco Esperado na amostra i obtida por meio da j-ésima técnica de validação descrita na Tabela 6.2;

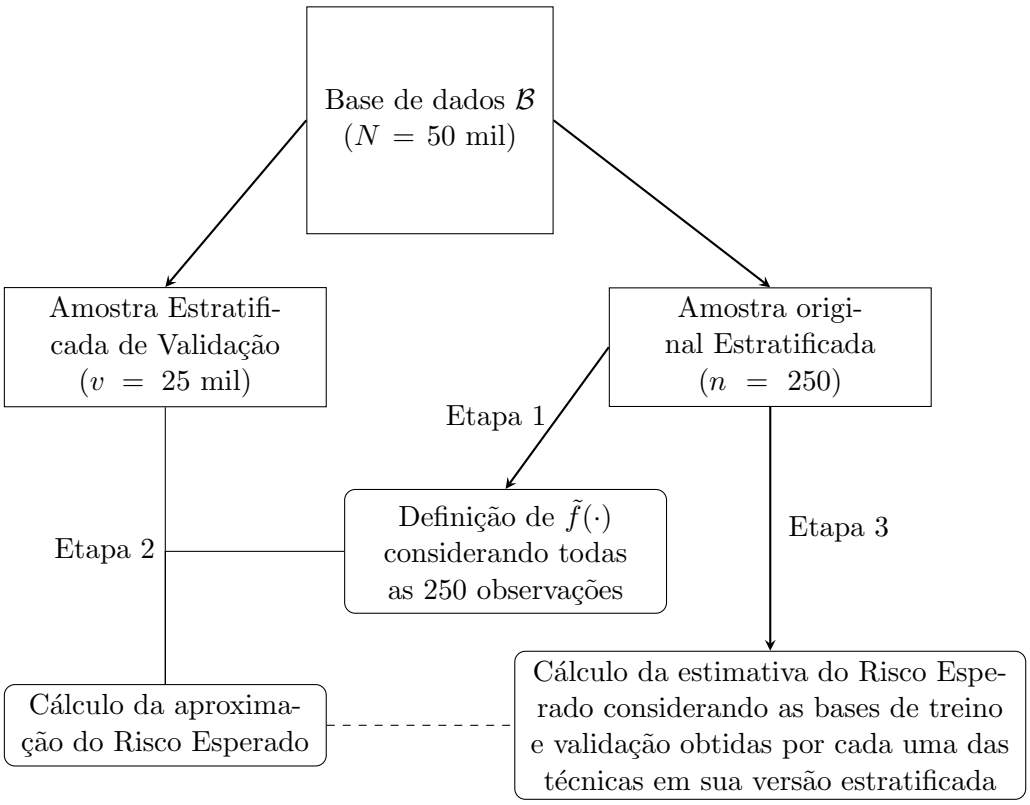


Figura 6.1: *Esquema de 1 repetição do processo de simulação quando $n = 250$.*

Err_i é a aproximação do Risco Esperado obtida utilizando a amostra i.

Além disso, foi obtida a variância amostral das estimativas de cada método a partir de:

$$Var_j = \sum_{i=1}^{100} \frac{(err_{ij} - \overline{err_j})^2}{99},$$

em que $\overline{err_j}$ representa a média dos 100 valores de Risco Esperado estimados por meio do método de validação j.

6.4 Resultados da Simulação

Os resultados obtidos por meio da simulação realizada foram compilados e analisados levando em consideração a distribuição, o viés e a variância das estimativas obtidas. Além disso, foi comparado o comportamento da aproximação do Risco Esperado com as estimativas obtidas pelos diferentes métodos e tamanhos de amostra. Todas essas análises foram replicadas considerando os 8 cenários distintos de prevalência do evento na base simulada (Tabela 6.1).

Vale lembrar que, conforme discutido na Seção 4.3, a construção de modelos a partir de amostras pequenas de treino se torna um tarefa ainda mais complexa na presença de dados desbalanceados. Em vista disso, para alguns cenários de simulação que envolviam pequenas amostras e, principal-

mente, maior grau de desbalanceamento não foi possível obter as estimativas dos parâmetros do modelo logístico visto que não houve convergência. Na Tabela 6.3 é possível ter mais detalhes dos métodos e cenários com maiores problemas no que tange à convergência do modelo. Para os cenários de tamanho de amostra 100, com proporções $0,42\%$ e $0,13\%$, e tamanho de amostra 250, com proporção $0,13\%$, não foi possível obter a estimativa do risco esperado por nenhum dos métodos em nenhuma das 100 repetições e para melhor visualização não aparecem na Tabela 6.3.

Tabela 6.3: *Percentual de repetições em que não foi possível obter o resultado por cenário de simulação.*

Técnica de	$7,05\%$	$2,77\%$		$1,09\%$			$0,42\%$					$0,13\%$			
Validação	100	100	250	100	250	500	250	500	750	1000	1500	500	750	1000	1500
err	0	0	0	13	0	0	2	0	0	0	0	4	9	5	0
err_{ho3}	0	3	1	44	6	0	47	13	4	1	0	38	44	45	15
err_{ho10}	0	1	0	23	0	0	20	0	0	0	0	15	12	9	4
err_{rho3}	2	57	3	98	30	6	100	74	37	9	5	97	99	97	71
err_{rho10}	0	8	0	71	1	0	68	12	2	0	0	75	62	74	16
err_{kf10}	0	4	0	100	0	0	100	2	0	0	0	100	100	100	10
err_{bkf10}	0	4	0	100	0	0	100	2	0	0	0	100	100	100	10
err_{rkf10}	0	10	0	100	0	0	100	4	0	0	0	100	100	100	11
err_{rbkf10}	0	10	0	100	0	0	100	4	0	0	0	100	100	100	11
err_{loo}	0	3	0	100	0	0	100	2	0	0	0	100	100	100	8
err_{bt}	10	49	6	34	27	9	20	22	12	2	0	14	15	10	16
err_{bt632}	10	49	6	34	27	9	20	22	12	2	0	14	15	10	16
err_{bt632+}	10	49	6	34	27	9	20	22	12	2	0	14	15	10	16

No exemplo simulado, os métodos do tipo *Bootstrap* e *Repeated Hold-Out* do tipo err_{rho3} demonstraram dificuldade de convergência do modelo para valores de prevalência iguais ou inferiores a $7,05\%$. Os demais métodos apresentaram dificuldade gradativa da convergência enquanto a prevalência diminuía, para valores menores ou iguais a $2,77\%$. Note que a dificuldade no ajuste do modelo surge também para tamanhos de amostra maiores conforme a prevalência diminui.

Um ponto de destaque para a dificuldade de obtenção da estimativa do Risco Esperado por meio dos métodos do tipo *K-Fold* está relacionado a maneira com que é realizada a partição da amostra. Para err_{kf10}, por exemplo, obtemos 10 partições disjuntas em que, para cada uma delas, se deseja manter a mesma prevalência dos dados como um todo. Em um cenário de desbalanceamento extremo e tamanho de amostra muito pequeno, o número absoluto de eventos presente na amostra pode não ser suficiente para manter esta prevalência. Métodos do tipo *Bootstrap* parecem mais adequados nesse contexto justamente por se tratar de um processo que retira amostras com reposição e conseguir garantir a manutenção da prevalência original.

Em decorrência da dificuldade de ajuste de modelos em amostras pequenas e desbalanceadas acaba não sendo possível a utilização de alguns métodos de validação em cenários mais extremos que implicarão em amostras de treino ainda menores. Em situações assim, pode ser necessário

escolher o método de validação levando em consideração o tamanho da amostra de treino que será gerada e não só um possível desempenho do método.

Nas Figuras 6.2 e 6.3 estão apresentados o comportamento médio, isto é, a média das estimativas, desconsiderando a variabilidade atrelada a cada uma delas. Dessa forma, considerando somente o comportamento médio dos métodos é possível ter uma noção de qual deles mais se assemelha com o risco esperado de acordo com os diferentes tamanhos de amostra definidos.

A partir do gráfico da Figura 6.2 referente ao cenário em que $p = 7,05\%$ é possível observar a confirmação do resultado que se esperava para as estimativas do Risco Esperado obtidas pelo estimador erro aparente (err). Para todos os cenários de prevalência simulados é possível perceber que o comportamento do erro aparente é, em geral de subestimar os valores do risco esperado, se mostrando um estimador inadequado.

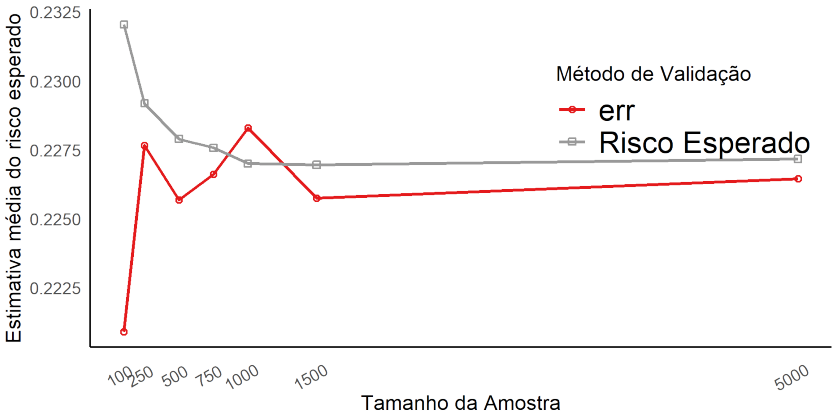
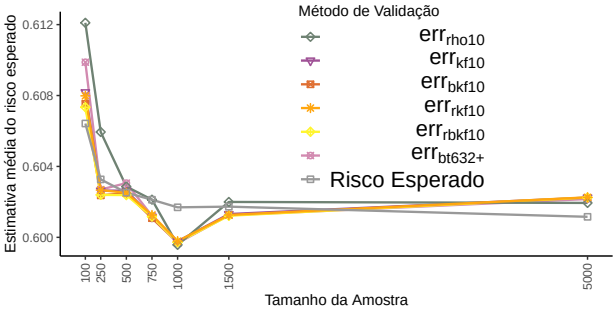


Figura 6.2: *Comparação do erro aparente médio com o Risco Esperado médio aproximado* $p = 7,05\%$

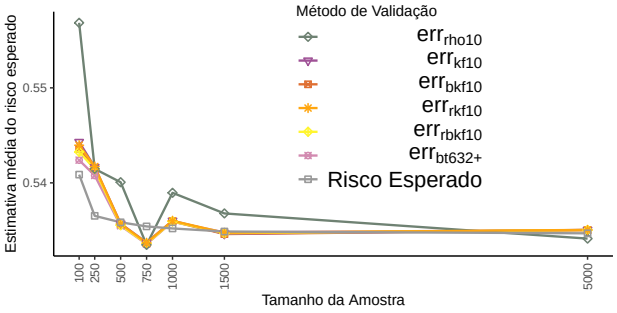
Nos gráficos compilados na Figura 6.3 estão destacadas as estimativas médias dos métodos que mostraram melhor desempenho dentre as categorias *Hold-Out*, *K-fold* e *Bootstrap* em cada um dos cenários de desbalanceamento, isto é, os métodos que apresentaram valor médio estimado mais próximo do real valor do Risco Esperado.

Os métodos do tipo Boostrap ponderados tem um bom desempenho para todos os cenários de prevalência estudados, sendo que aquele que resulta em err_{bt632+} apresentaram desempenho melhor para prevalências superiores ou iguais a $2,77\%$ e para cenário em que a prevalência é inferior a $2,77\%$ a estimativa média obtida por err_{bt632} é a de melhor desempenho. Resultados similares a estes, em todos os cenários de prevalência, tiveram as técnicas: *K-Fold*, *Burman K-fold*, *Repeated K-fold* e *Repeated Burman K-fold*.

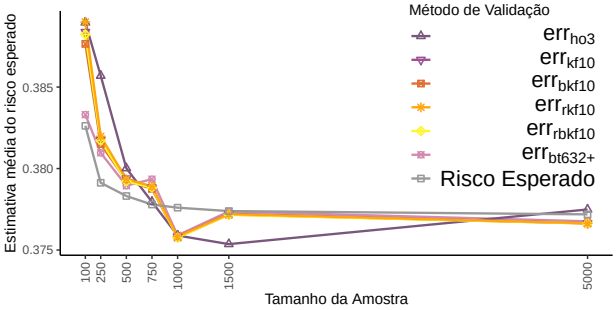
É possível avaliar os gráficos das Figuras 6.4 e 6.5 para verificar o desempenho dos métodos de validação no que diz respeito à variabilidade e ao viés das estimativas. Na Figura 6.4 a linha



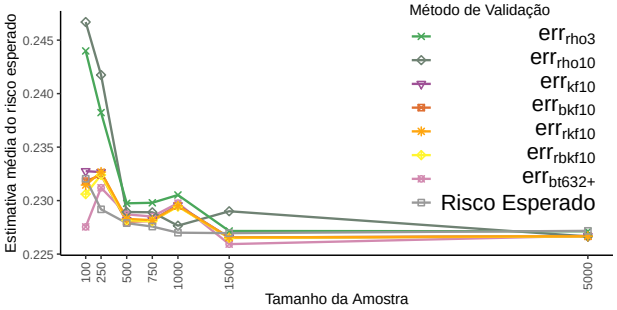
(a) *Prevalência de* $50,27\%$



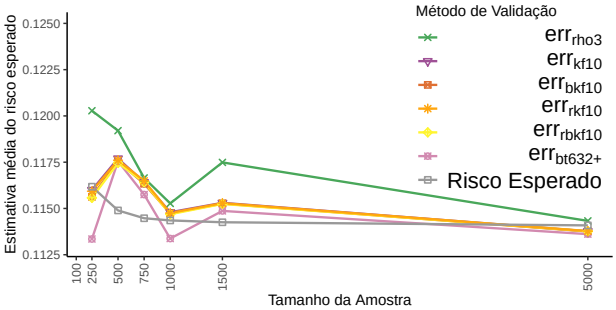
(b) *Prevalência de* $30,58\%$



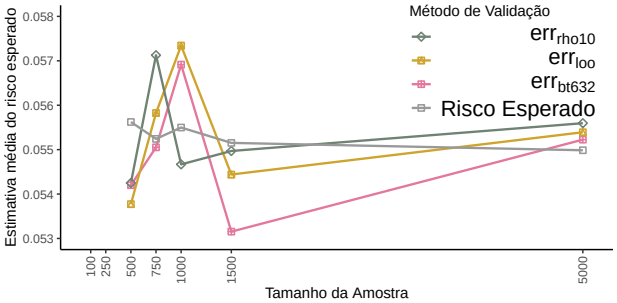
(c) *Prevalência de* $15,52\%$



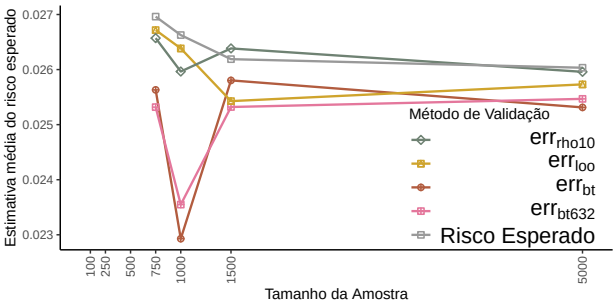
(d) *Prevalência de* $7,05\%$



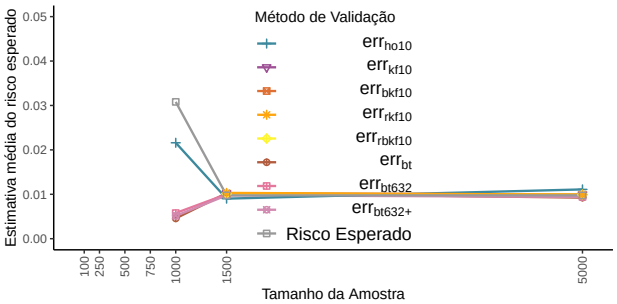
(e) *Prevalência de* $2,77\%$



(f) *Prevalência de* $1,09\%$



(g) *Prevalência de* $0,42\%$



(h) *Prevalência de* $0,13\%$

Figura 6.3: *Comparação das melhores estimativas do Risco Esperado obtidas para diferentes tamanhos de amostra para cada uma das diferentes prevalências.*

pontilhada de referência se trata do Risco Esperado médio e na Figura 6.5 está atrelada à referencia de uma estimativa não viesada, ou seja, Viés = 0. Aqui será explicitada análise detalhada da variabilidade e viés apenas para o tamanho de amostra 1.000. O que se observa para tamanhos maiores de amostra é uma redução considerável da variabilidade e do viés das estimativas.

Para o cenário em que o tamanho da amostra inicial é 1.000 e a prevalência é $0,13\%$ (Figuras 6.4h 6.5h) os gráficos não apresentam informação para os métodos do tipo *Bootstrap* porque não foi possível realizar o ajuste do método nesse cenário. As análises seguirão para os demais métodos.

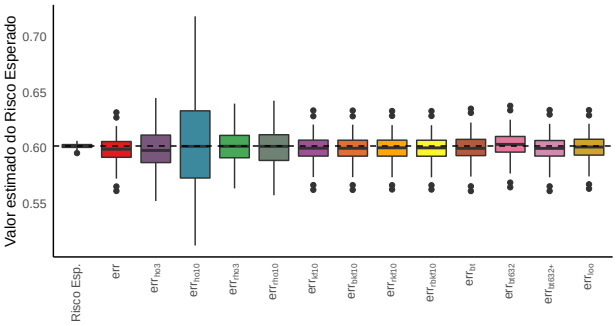
Um ponto de destaque no que tange a variabilidade das estimativas está atrelada aos métodos *Hold-Out* (err_{ho10} e err_{ho3}) que apresentam variabilidade superior a dos demais métodos para todos os cenários de prevalência simulados, sendo err_{ho10} responsável pela maior instabilidade das estimativas. É nítido que a realização do *Repeated Hold-Out*, de fato, auxilia na redução da variabilidade associada às estimativas do método convencional do *Hold-Out*.

No entanto, quando se trata do método *K-fold* não é possível perceber redução nítida da variabilidade das estimativas quando executado o método com repetição. Sendo assim, é possível assumir que, por *K-fold* e *Repeated K-fold* possuírem desempenhos semelhantes, o esforço computacional atrelado a realização do método com repetição não parece valer a pena.

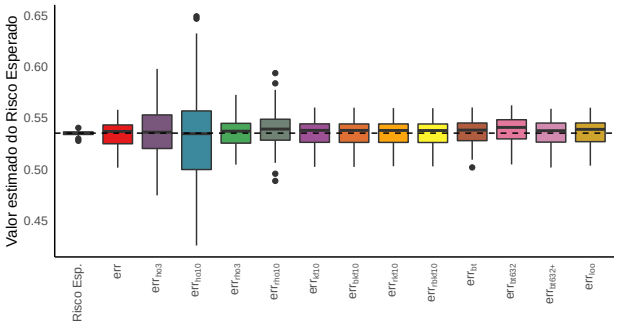
Com relação ao viés, exposto nos gráficos da Figura 6.5 é possível perceber algumas diferenças de comportamento com relação a diferentes cenários de prevalência do evento na amostra. Destaque para a presença de observações de viés extremamente elevados, *outliers*, tanto positivos quanto negativos principalmente para menores percentuais de eventos na amostra.

O Erro Aparente é, no geral, uma estimativa subviesada para o Risco esperado. Essa conclusão fica mais clara, principalmente, nos cenários de prevalência em que a mediana está explicitamente abaixo do valor de referência Viés = 0, são eles: $50,27\%$ e $15,52\%$.

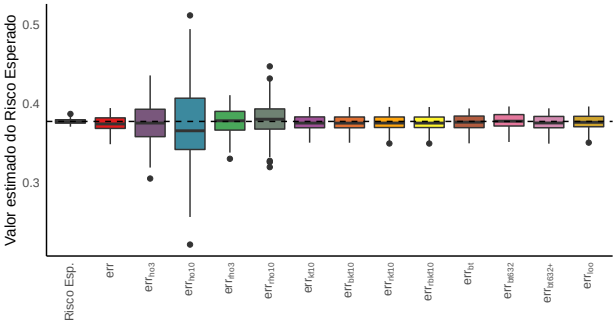
A partir dos diferentes resultados obtidos na etapa de simulação verifica-se que diferentes métodos de validação implicarão em diferentes estimativas do Risco Esperado. Esta diferença pode, em alguns cenários, implicar em diferentes escolhas de modelos.



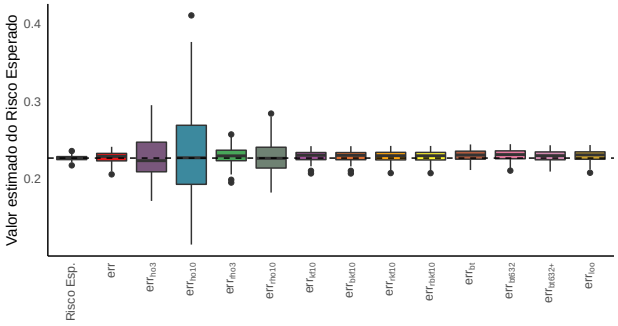
(a) *Prevalência de* $50,27\%$



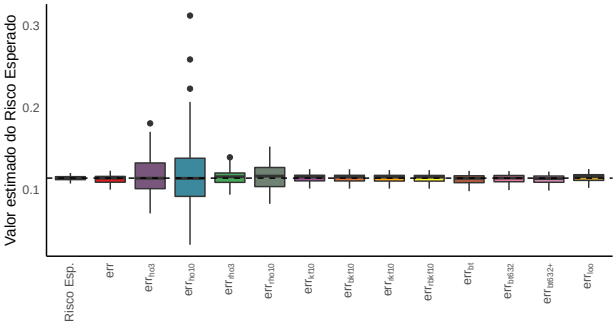
(b) *Prevalência de* $30,58\%$



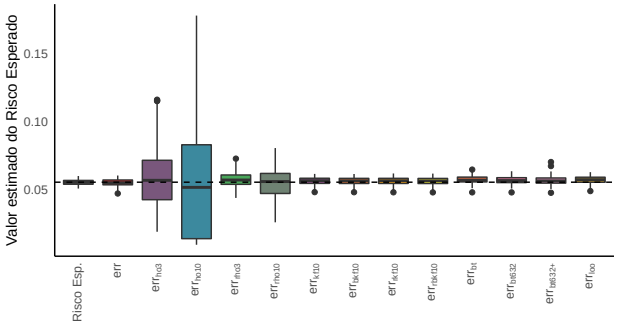
(c) *Prevalência de* $15,52\%$



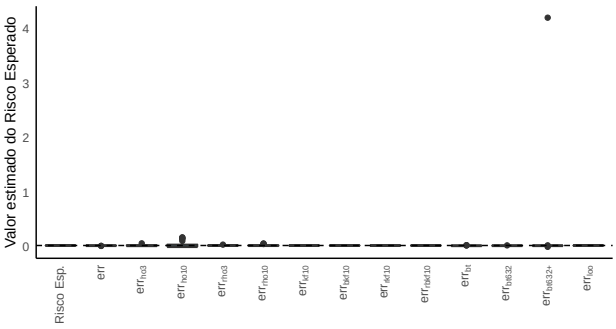
(d) *Prevalência de* $7,05\%$



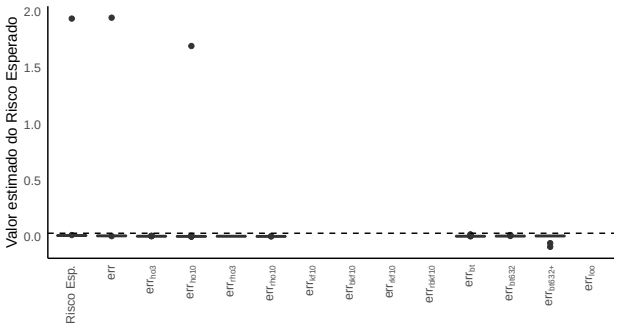
(e) *Prevalência de* $2,77\%$



(f) *Prevalência de* $1,09\%$

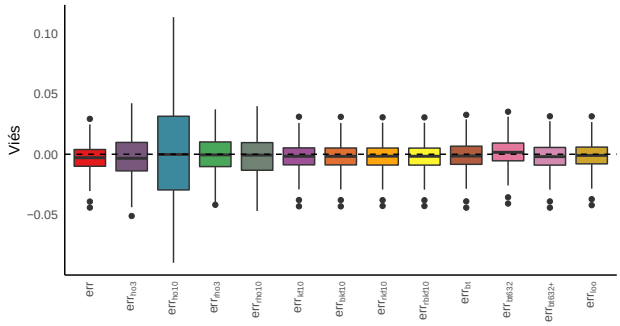


(g) *Prevalência de* $0,42\%$

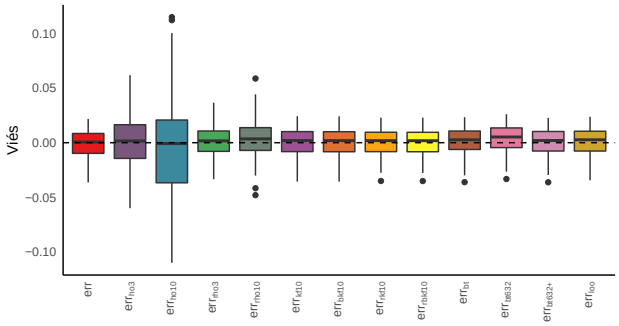


(h) *Prevalência de* $0,13\%$

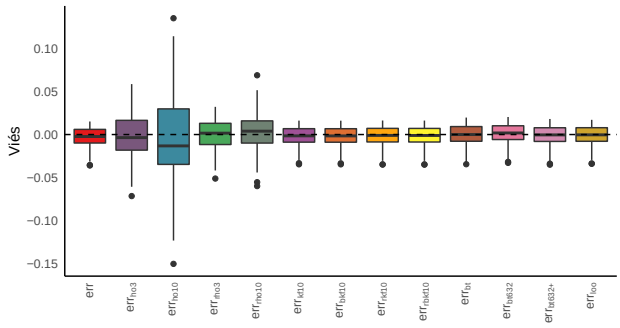
Figura 6.4: *Distribuição das estimativas do Risco Esperado para diferentes prevalências e* $n = 1.000$.



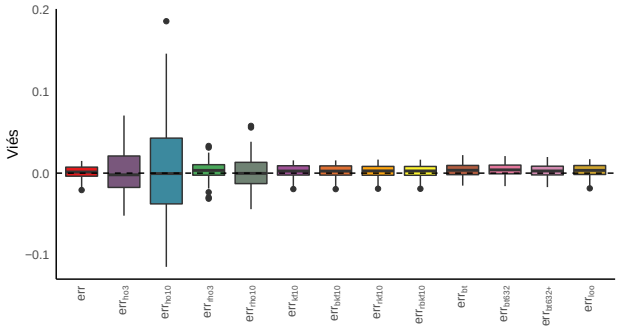
(a) *Prevalência de* $50,27\%$



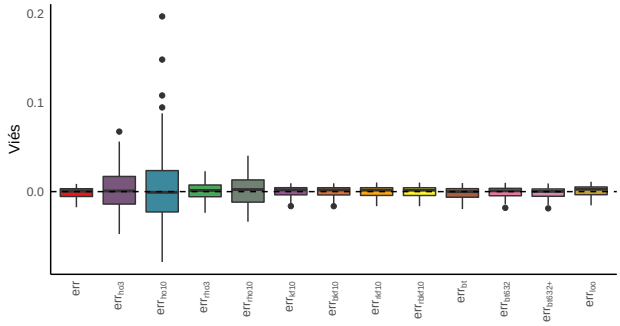
(b) *Prevalência de* $30,58\%$



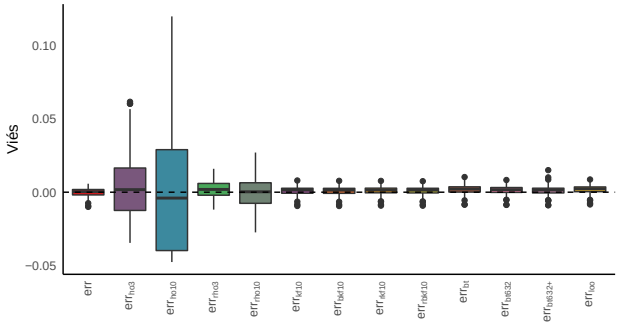
(c) *Prevalência de* $15,52\%$



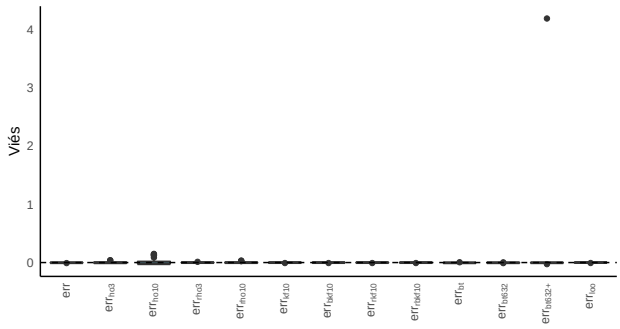
(d) *Prevalência de* $7,05\%$



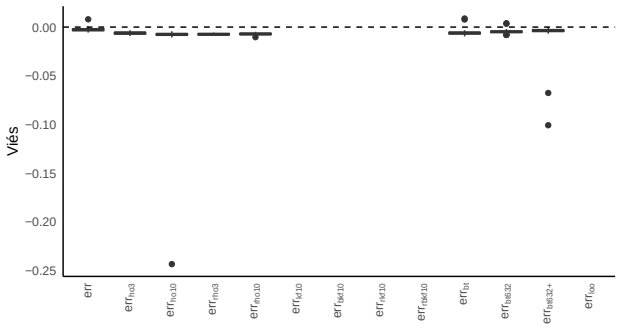
(e) *Prevalência de* $2,77\%$



(f) *Prevalência de* $1,09\%$



(g) *Prevalência de* $0,42\%$



(h) *Prevalência de* $0,13\%$

Figura 6.5: *Distribuição do viés das estimativas para diferentes prevalências e* $n = 1.000$.

7. Aplicação

Por fim, foi realizada aplicação das técnicas abordadas neste trabalho em uma base de dados real a respeito da evolução de casos notificados da Sídrome Inflamatória Multissistêmica Pediátrica (SIM-P) temporalmente associada à COVID-19. A ideia é aplicar as técnicas tratadas ao longo do trabalho e verificar, em dados reais, o comportamento do que foi concluído a partir das simulações realizadas no Capítulo 6. Além disso, ter a oportunidade de verificar as estratégias para lidar com o desbalanceamento em uma aplicação real.

7.1 Descrição dos dados

A base de dados foi disponibilizada pelo Ministério da Saúde via solicitação aberta no Sistema Eletrônico de Informações ao Cidadão (e-SIC) e parceria com a médica responsável Marcela da Costa. Os dados são referentes a um compilado de informações das fichas de notificação preenchidas em todo o Brasil no período de julho de 2020 a junho de 2021 para casos de SIM-P com algum tipo de associação temporal à COVID-19 (evidência de COVID-19 ou história de contato com caso de COVID-19),

Uma notificação se trata da comunicação realizada por profissionais da saúde para autoridades sanitárias a respeito da ocorrência de determinada doença ou agravo. Neste caso, o critério para que um paciente seja passível de notificação como um caso de SIM-P associada à COVID-19 está atrelado à definição de caso preliminar especificada pelo Ministério da Saúde.

Por meio das notificações é possível, além de ter acesso ao quantitativo total de casos preliminares da doença, obter informações mais detalhadas sobre o paciente, respectivo quadro clínico e evolução (alta hospitalar ou óbito). No caso da base de dados utilizada neste trabalho estão contemplados 990 pacientes que totalizam 64 óbitos, 836 pacientes com alta hospitalar e 90 pacientes sem informação da evolução,

Para este estudo serão considerados somente os 821 pacientes notificados que tiveram o diagnóstico final confirmado como SIM-P, possuem registro de diagnóstico de COVID-19 ou contato

com caso confirmado da doença e apresentam informação da evolução do caso (alta hospitalar ou óbito). Neste cenário, os pacientes que evoluíram para óbito se tratam de $6,94\%$ dos pacientes analisados, se enquadrando em um contexto de desbalanceamento da variável resposta,

O intuito do estudo é verificar de que forma informações clínicas do paciente se relacionam com a variável resposta evolução do caso (alta hospitalar ou óbito). Além disso, buscou-se compreender de que maneira as estimativas do Risco Esperado, obtidas por meio de métodos de validação distintos, se comportam para essa base de dados.

A maior parte do questionário a ser respondido para notificação é composto por perguntas que resultam em variáveis categóricas. A escolha das variáveis explicativas para compor o estudo se baseou na completude das informações contidas no questionário, frequência de respostas razoável em cada uma das classes das variáveis, fatores de relevância para a compreensão da doença e verificação de possível associação univariada pelo teste Qui-quadrado. As variáveis consideradas no estudo estão elencadas na Tabela 7.1.

7.2 Modelo proposto

Seja Y_i, $i = 1, 2, \cdots, 821$, a evolução do i-ésimo paciente, ou seja, $y_i = 1$ se o paciente foi a óbito e $y_i = 0$ caso o paciente tenha recebido alta hospitalar. Dessa forma, $Y_i \sim \text{Bernoulli}(f_i(X))$ tal que

$$f_i(X) = \frac{\exp(\eta_i)}{1 + \exp(\eta_i)},$$

em que

$$\eta_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_6 x_{6i}.$$

O parâmetro β_0 faz referência ao efeito basal, os demais parâmetros β_{ji}, $j = 1, \ldots, 6$, estão associados ao efeito com relação a evolução dos pacientes de cada uma das respectivas variáveis detalhadas na Tabela 7.1.

Tabela 7.1: *Variáveis consideradas no modelo*

Variável	Descrição
X_1	Hipotensão arterial ou choque (1 - apresentou; 0 - não apresentou)
X_2	Irritabilidade (1 - apresentou; 0 - não apresentou)
X_3	Complicação: Falência de outros órgãos (1 - apresentou; 0 - não apresentou)
X_4	Complicação: Insuficiência renal aguda (1 - apresentou; 0 - não apresentou)
X_5	Complicação: Necessidade de ventilação invasiva (1 - apresentou; 0 - não apresentou)
X_6	Doença ou condição pré-existente (1 - apresentou; 0 - não apresentou)

Além disso, para cálculo da probabilidade estimada foi considerada versão corrigida proposta

por King e Zeng [2001] e destacada na equação 5.4.

7.3 Resultados e Discussões

O modelo foi ajustado no *software R* a partir do pacote *zelig* considerando o modelo logístico com correção de viés tal qual proposto por King e Zeng [2001] e detalhado no Capítulo 5. Apesar do modelo utilizado já estar implementado no pacote *zelig* o erro padrão das estimativas considera a variância sem correção. Sendo assim, o cálculo do erro padrão e, consequentemente, do p-valor foram obtidos por função própria implementada.

Tabela 7.2: *Estimativa dos parâmetros do modelo*

Parâmetro	Modelo Logístico				Modelo Logístico Corrigido			
	Estimativa	EP[1]	p-valor	RR[2]	Estimativa	EP[1]	p-valor	RR[2]
β_0	$-5,413$	$0,474$	$< 0,001$	-	$-5,258$	$0,470$	$< 0,001$	-
β_1	$0,949$	$0,449$	$0,035$	$2,57$	$0,916$	$0,445$	$0,035$	$2,48$
β_2	$1,240$	$0,388$	$0,001$	$3,42$	$1,213$	$0,384$	$0,001$	$3,32$
β_3	$1,218$	$0,458$	$0,008$	$3,35$	$1,187$	$0,454$	$0,008$	$3,24$
β_4	$1,059$	$0,380$	$0,005$	$2,86$	$1,040$	$0,377$	$0,005$	$2,80$
β_5	$2,477$	$0,418$	$< 0,001$	$11,36$	$2,398$	$0,414$	$< 0,001$	$10,46$
β_6	$1,075$	$0,352$	$0,002$	$2,91$	$1,051$	$0,349$	$0,002$	$2,83$

[1] Erro Padrão (EP).
[2] Risco Relativo (RR) para a i-ésima variável, $i = 1, \ldots, 6$, calculado por $RR = \frac{P(Y=1|x_i=1)}{P(Y=1|x_i=0)}$ mantendo as demais variáveis constantes e iguais a zero.

A partir do que está apresentado na Tabela 7.2 é possível verificar o valor das estimativas obtidas por meio do modelo logístico usual, a exceção do intercepto, levemente maiores do que as obtidas pelo modelo logístico corrigido. Além disso, conforme esperado, verifica-se maiores valores do erro padrão e menores valores de p-valor para o modelo logístico usual que, nesse caso, não impactou na decisão final com relação a significância das variáveis.

Considerando o Modelo Logístico Corrigido obtido é possível chegar em algumas conclusões com relação a fatores de risco para o óbito. Pacientes com hipotensão arterial ou choque, irritabilidade, que apresentam falência de órgãos, insuficiência renal aguda, necessidade de ventilação invasiva e tem alguma doença ou condição pré-existente possuem maior probabilidade de óbito. A probabilidade de óbito de um paciente que apresenta somente uma dessas características de maneira isolada, é de, respectivamente, $2,48$, $3,32$, $3,24$, 2.80, $10,46$ e $2,83$ vezes maior do que um paciente que apresentou nenhuma das características. Sendo assim, a necessidade de ventilação invasiva apresenta maior destaque no que tange ao aumento da probabilidade de óbito indicando ser uma complicação de maior risco que sozinha aumenta em $11,36$ vezes a probabilidade de óbito.

		Classe Predita	
		+	-
Classe Observada	+	VP = 52	FN = 5
	−	FP = 134	VN = 630
	Total	P* = 186	N* = 635

Figura 7.1: *Matriz de confusão do modelo logístico usual ($c = 0,04$) e corrigido ($c = 0,05$).*

Na Tabela 7.2 é possível verificar o risco relativo associado também ao Modelo Logístico usual que é sistematicamente superior à do Modelo Logístico Corrigido.

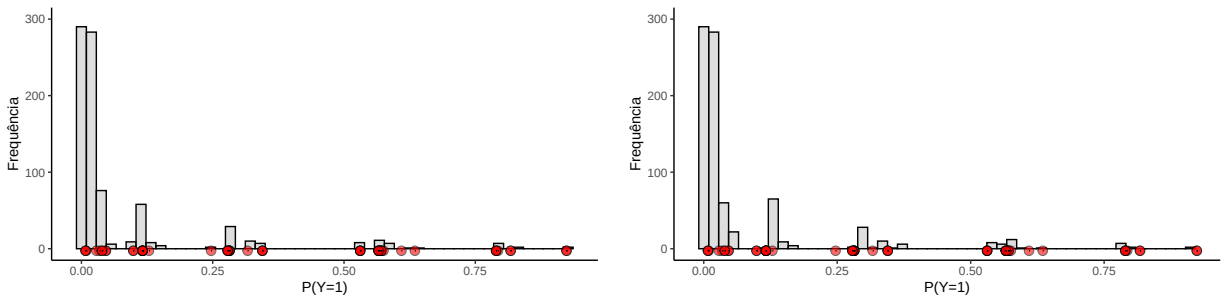
Por meio do auxílio da curva ROC foi possível obter qual seria o limiar mais adequado c para cada um dos modelos obtidos no que tange à maximização do percentual de verdadeiros positivos e verdadeiros negativos. Considerando limiar de $c = 0,04$ para o modelo logístico usual e $c = 0,05$ para o modelo logístico corrigido os resultados das classificações foram os mesmos e, consequentemente, as métricas de desempenho obtidas também foram as mesmas. Na Tabela 7.3 é possível verificar as métricas de desempenho de ambos os modelos e na Figura 7.1 a matriz de confusão obtida.

Tabela 7.3: *Métricas de desempenho do Modelo Logístico usual e corrigido.*

Métrica de Desempenho	Valor
Acurácia	$0,94$
%VP (Sensibilidade/Recall)	$91,23$
%VN (Especificidade)	$82,46$
%FP	$17,54$
%FN	$8,77$
Precisão	$27,96$
AUC	$92,10^*$
F1-score	$0,52$
GM	$0,92$
GMA	$0,91$
CWA ($w = 0,7$)	$0,88$
MCC	$0,44$

(*) A métrica AUC para ambos os modelos apresenta valores aproximadamente iguais.

Apesar de não ser possível verificar diferença na classificação das observações segundo cada um dos modelos, as estimativas obtidas para a probabilidade não são exatamente as mesmas. Assim como esperado, as probabilidades estimadas obtidas por meio do modelo logístico corrigido, ainda



(a) *Modelo Logístico Usual.* **(b)** *Modelo Logístico Corrigido.*

Figura 7.2: *Histograma das probabilidades estimadas.*

que não tão distintas, são sistematicamente maiores do que as do modelo usual. Na Figura 7.2 é possível comparar a distribuição das probabilidades estimadas, os pontos ao longo do eixo x representam casos em que $Y = 1$ foi, de fato, observado.

Além disso, foram obtidas estimativas do Risco Esperado segundo todos os métodos abordados nesse trabalho na sua versão estratificada. O resultado destas estimativas está destacado na Tabela 7.4 e contempla valores entre $0,142015$ e $0,166135$, sendo estes referentes ao err_{ho3} e err_{bt}, respectivamente.

Tabela 7.4: *Estimativas do Risco Esperado*

Técnica de Validação	Estimativa Risco Esperado
1. $\mathbf{err}$	0,150434
2. $\mathbf{err_{ho3}}$	0,142015
3. $\mathbf{err_{ho10}}$	0,183910
4. $\mathbf{err_{rho3}}$	0,167428
5. $\mathbf{err_{rho10}}$	0,148030
6. $\mathbf{err_{kf10}}$	0,158778
7. $\mathbf{err_{bkf10}}$	0,158326
8. $\mathbf{err_{rkf10}}$	0,159241
9. $\mathbf{err_{rbkf10}}$	0,158765
10. $\mathbf{err_{loo}}$	0,150435
11. $\mathbf{err_{bt}}$	0,166135
12. $\mathbf{err_{bt632}}$	0,161354
13. $\mathbf{err_{bt632+}}$	0,160634

Assim como já foi comentado, estimativas obtidas por meio do erro aparente são, em geral, subviesadas. Na Tabela 7.4 temos *err* entre os menores valores estimados para o risco esperado, indicando um valor subestimado. No entanto, de maneira geral, as estimativas obtidas para essa aplicação não apresentaram diferenças tão expressivas.

Em suma, o desenvolvimento da análise de uma variável resposta desbalanceada passa por algumas etapas importantes. Inicialmente é necessário realizar a seleção de variáveis explicativas a serem consideradas no modelo. A partir disso, são obtidas as estimativas corrigidas para os

parâmetros do modelo e calculada a probabilidade estimada, também com correção. Por fim, são calculadas métricas de desempenho adequadas para que se decida pelo melhor modelo possível.

8. Conclusão

O processo de seleção de modelos é mais adequado quando realizado em uma base de dados com observações inéditas. Os métodos de validação cruzada se mostram aliados deste processo de seleção de modelos. Neste trabalho foi apresentada uma revisão dos principais métodos de validação existentes, sendo eles, *Hold-out*, *K-Fold*, *Leave-one-out* e *Bootstrap*.

Além disso, foram destacadas várias questões atreladas ao desbalanceamento da variável resposta, algumas dificuldades e estratégias para contornar esses problemas. Nesse contexto, foram apresentadas versões estratificadas dos métodos de validação convencionais para garantir que a distribuição da variável resposta na amostra original e na amostra de treino fossem equivalentes.

Tendo em vista o viés atrelado às estimativas dos parâmetros do modelo logístico usual, na presença de eventos raros, foi sugerida a utilização da correção proposta por King e Zeng [2001]. Este modelo apresenta resultados mais adequados para respostas desbalanceadas. A utilização desse modelo conjuntamente aos métodos de validação estratificados formam estratégia conjunta para lidar com as dificuldades atreladas ao desbalanceamento. Sendo assim, o estudo desses métodos na versão estratificada simultaneamente à utilização do modelo proposto por King e Zeng [2001] é a principal contribuição desse trabalho que permitiu a análises conjunta destes métodos, gerando novas questões de pesquisa.

Foi possível, por meio do processo de simulação realizado, perceber nitidamente que diferentes métodos de validação irão implicar em diferentes estimações do Risco Esperado. Além disso, nota-se que, em geral, quanto menor o tamanho da amostra mais urgente é a necessidade de considerar algum método que avalie o modelo para observações que não foram contempladas no processo de ajuste do modelo.

Neste estudo métodos do tipo *K-fold*, em geral, apareceram como possibilidades interessantes para cenários em que a prevalência do evento na amostra era de $2,77\%$ ou mais, para prevalências inferiores métodos do tipo *Hold-Out* pareceram mais adequados. Um ponto de destaque está para os métodos do tipo *Bootstrap* e a capacidade de contornar dificuldades inerentes ao processo de

validação de amostras pequenas e desbalanceadas por retirar amostras com reposição.

Por fim, acreditamos que este trabalho sirva de referência para consulta de estratégias capazes de lidar com o desbalanceamento e estudos relacionados à validação de modelos. Alguns pontos que não foram explorados e seriam interessantes para trabalhos futuros são a utilização de métodos de seleção de variáveis mais eficazes para dados desbalanceados e a compreensão das limitações e estratégias ao lidar com desbalanceamento das variáveis explicativas.

Referências Bibliográficas

Al-Ghraibah *et al.***(2015)** Amani Al-Ghraibah, Laura E Boucheron e RT James McAteer. A study of feature selection of magnetogram complexity features in an imbalanced solar flare prediction data-set. Em *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, páginas 557–564. IEEE. Citado na pág. 2

Alejo *et al.***(2013)** Roberto Alejo, Rosa Maria Valdovinos, Vicente García e J Horacio Pacheco-Sanchez. A hybrid method to face class overlap and class imbalance on neural networks and multi-class scenarios. *Pattern Recognition Letters*, 34(4):380–388. Citado na pág. 27

Borra e Ciaccio(2010) Simone Borra e Agostino Di Ciaccio. Measuring the prediction error. a comparison of cross-validation, bootstrap and covariance penalty methods. *Computational statistics & data analysis*, 54(12):2976–2989. Citado na pág. 1, 2, 6, 7, 13, 15, 16

Branco *et al.***(2016)** Paula Branco, Luís Torgo e Rita P. Ribeiro. A survey of predictive modeling on imbalanced domains. *ACM Computing Surveys*, 49(2):1–50. ISSN 0360-0300. doi: 10.1145/2907070. URL https://dx.doi.org/10.1145/2907070. Citado na pág. 2, 8, 9, 12, 21, 22, 23, 24, 25, 26, 27, 28

Burman(1989) Prabir Burman. A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods. *Biometrika*, 76(3):503–514. Citado na pág. 15, 16

Casañola-Martin *et al.***(2016)** Gerardo Casañola-Martin, Teresa Garrigues, Marival Bermejo, Isabel González-Álvarez, Nam Nguyen-Hai, Miguel Ángel Cabrera-Pérez, Huong Le-Thi-Thu *et al.* Exploring different strategies for imbalanced adme data problem: case study on caco-2 permeability modeling. *Molecular diversity*, 20(1):93–109. Citado na pág. 2

Chu *et al.***(2010)** Leilei Chu, Hui Gao e Wenbo Chang. A new feature weighting method based on probability distribution in imbalanced text classification. Em *2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery*, volume 5, páginas 2335–2339. IEEE. Citado na pág. 27

Cunha(2019) João Paulo Zanola Cunha. Um estudo comparativo das técnicas de validação cruzada aplicadas a modelos mistos. mathesis, Universidade de São Paulo, São Paulo. Citado na pág. 2, 13

Diniz e Louzada(2013) Carlos Diniz e Francisco Louzada. Métodos estatısticos para análise de dados de crédito. Em *6th Brazilian Conference on Statistical Modeling in Insurance and Finance, Maresias-SP*. Citado na pág. 35

Efron(1983) Bradley Efron. Estimating the error rate of a prediction rule: improvement on cross-validation. *Journal of the American statistical association*, 78(382):316–331. Citado na pág. 17, 18

Efron e Tibshirani(1995) Bradley Efron e Robert J Tibshirani. *Cross-validation and the bootstrap: Estimating the error rate of a prediction rule.* Tese de Doutorado, Division of Biostatistics, Stanford University. Citado na pág. 18

Forman(2003) George Forman. An extensive empirical study of feature selection metrics for text classification. *Journal of machine learning research*, 3(Mar):1289–1305. Citado na pág. 27

Forman e Scholz(2010) George Forman e Martin Scholz. Apples-to-apples in cross-validation studies: pitfalls in classifier performance measurement. *Sigkdd Explorations*, 12(1):49–57. Citado na pág. 10, 12, 22, 30, 31

García *et al.***(2006)** Vicente García, Roberto Alejo, José Salvador Sánchez, José Martínez Sotoca e Ramón Alberto Mollineda. Combined effects of class imbalance and class overlap on instance-based classification. Em *International Conference on Intelligent Data Engineering and Automated Learning*, páginas 371–378. Springer. Citado na pág. 27

Haixiang *et al.***(2017)** Guo Haixiang, Li Yijing, Jennifer Shang, Gu Mingyun, Huang Yuanyue e Gong Bing. Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications*, 73:220–239. Citado na pág. 22

Hastie *et al.***(2008)** Trevor Hastie, Robert Tibshirani e Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction.* Springer Science & Business Media. Citado na pág. 1, 6, 7

James *et al.***(2013)** Gareth James, Daniela Witten, Trevor Hastie e Robert Tibshirani. *An introduction to statistical learning*, volume 112. Springer. Citado na pág. 1, 5, 6, 7, 8, 11

Kim(2009) Ji-Hyun Kim. Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Computational statistics & data analysis*, 53(11):3735–3745. Citado na pág. 2, 17

King e Zeng(2001) Gary King e Langche Zeng. Logistic regression in rare events data. *Political Analysis*, 9:137–163. Citado na pág. iii, v, 2, 33, 35, 36, 37, 38, 39, 41, 42, 43, 44, 55, 59

Kohavi(1995) Ron Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. Em *Ijcai.* Citado na pág. 2, 13, 14

Kotsiantis *et al.***(2006)** Sotiris Kotsiantis, Dimitris Kanellopoulos e Panayiotis Pintelas. Handling imbalanced datasets: A review. *GESTS International Transactions on Computer Science and Engineering*, 30(1):25–36. Citado na pág. 22, 26

Lane *et al.***(2012)** Peter C.R. Lane, Daoud Clarke e Paul Hender. On developing robust models for favourability analysis: Model choice, feature sets and imbalanced data. *Decision Support Systems*, 53(4):712–718. ISSN 0167-9236. doi: 10.1016/j.dss.2012.05.028. URL https://dx.doi.org/10.1016/j.dss.2012.05.028. Citado na pág. 2

Lima e Pereira(2015) Rafael Franca Lima e Adriano Cesar Machado Pereira. A fraud detection model based on feature selection and undersampling applied to web payment systems. Em *2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, volume 3, páginas 219–222. IEEE. Citado na pág. 2

Lusa *et al.***(2010)** Lara Lusa *et al.* Class prediction for high-dimensional class-imbalanced data. *BMC bioinformatics*, 11(1):523. Citado na pág. 2

López *et al.***(2013)** Victoria López, Alberto Fernández, Salvador García, Vasile Palade e Francisco Herrera. An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information Sciences*, 250:113–141. ISSN 0020-0255. doi: 10.1016/j.ins.2013.07.007. URL https://dx.doi.org/10.1016/j.ins.2013.07.007. Citado na pág. 2, 21, 23, 25, 26, 27, 28

López *et al.***(2014)** Victoria López, Alberto Fernández e Francisco Herrera. On the importance of the validation technique for classification with imbalanced datasets: Addressing covariate shift when data is skewed. *Information Sciences*, 257:1–13. ISSN 0020-0255. doi: 10.1016/j.ins.2013.09.038. Citado na pág. 28, 30

Moepya *et al.***(2014)** Stephen O Moepya, Sharat S Akhoury e Fulufhelo V Nelwamondo. Applying cost-sensitive classification for financial fraud detection under high class-imbalance. Em *2014 IEEE International Conference on Data Mining Workshop*, páginas 183–192. IEEE. Citado na pág. 2

Murphy(2012) Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press. Citado na pág. 8, 10, 11

Paula(2013) Gilberto A. Paula. Modelos de regresão com apoio computacional. IME - USP, 2013. Citado na pág. 33, 34

Prati *et al.***(2004)** Ronaldo C Prati, Gustavo EAPA Batista e Maria Carolina Monard. Class imbalances versus class overlapping: an analysis of a learning system behavior. Em *Mexican international conference on artificial intelligence*, páginas 312–321. Springer. Citado na pág. 27

R Core Team(2019) R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2019. URL https://www.R-project.org. Citado na pág. 41

Raeder *et al.***(2012)** Troy Raeder, George Forman e Nitesh V. Chawla. Learning from imbalanced data: Evaluation matters. Em Dawn E. Holmes e Lakhmi C. Jain, editors, *Data Mining: Foundations and Intelligent Paradigms. Intelligent Systems Reference Library*, volume 23, chapter 12. Springer, Berlin, Heidelberg. Citado na pág. 2, 21, 41

Rodriguez *et al.***(2009)** Juan Diego Rodriguez, Aritz Pérez e Jose Antonio Lozano. Sensitivity analysis of k-fold cross validation in prediction error estimation. *IEEE transactions on pattern analysis and machine intelligence*, 32(3):569–575. Citado na pág. 16

Seiffert *et al.***(2014)** Chris Seiffert, Taghi M. Khoshgoftaar, Jason Van Hulse e Andres Folleco. An empirical study of the classification performance of learners on imbalanced and noisy software quality data. *Information Sciences*, 259:571–595. Citado na pág. 23, 24, 27, 28

Stone(1974) Mervyn Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):111–133. Citado na pág. 1

Tamura(2007) Karin Ayumi Tamura. Modelo logístico multinível: um enfoque em métodos de estimação e predição. mathesis, Universidade de São Paulo, São Paulo. Citado na pág. 8

Vong *et al.***(2015)** Chi-Man Vong, Weng-Fai Ip, Chi-Chong Chiu e Pak-Kin Wong. Imbalanced learning for air pollution by meta-cognitive online sequential extreme learning machine. *Cognitive Computation*, 7(3):381–391. Citado na pág. 2

Wasikowski e Chen(2009) Mike Wasikowski e Xue-wen Chen. Combating the small sample class imbalance problem using feature selection. *IEEE Transactions on knowledge and data engineering*, 22(10):1388–1400. Citado na pág. 27

Wei *et al.***(2013)** Wei Wei, Jinjiu Li, Longbing Cao, Yuming Ou e Jiahang Chen. Effective detection of sophisticated online banking fraud on extremely imbalanced data. *World Wide Web*, 16(4): 449–475. Citado na pág. 2

Zheng *et al.***(2004)** Zhaohui Zheng, Xiaoyun Wu e Rohini Srihari. Feature selection for text categorization on imbalanced data. *ACM Sigkdd Explorations Newsletter*, 6(1):80–89. Citado na pág. 27