Universidade Federal de Alfenas - UNIFAL
Instituto de Ciências Sociais Aplicadas - ICSA

Júlio César da Silva Zeferino

Análise do risco de crédito utilizando regressão logística

Varginha
2021

Universidade Federal de Alfenas - UNIFAL

Instituto de Ciências Sociais Aplicadas - ICSA

Júlio César da Silva Zeferino

Análise do risco de crédito utilizando regressão logística

Trabalho de conclusão do Piepex apresentado ao Instituto de Ciências Sociais Aplicadas da Universidade Federal de Alfenas como requisito parcial à obtenção do título de Bacharel em Ciência e Economia.

Orientador: Profª Drª Patrícia de Siqueira Ramos

Varginha

2021

Universidade Federal de Alfenas - UNIFAL

Instituto de Ciências Sociais Aplicadas - ICSA

Júlio César da Silva Zeferino

Análise do risco de crédito utilizando regressão logística

A banca examinadora abaixo-assinada, aprova o trabalho de conclusão do PIEPEX (TCP), apresentado como parte dos requisitos para a obtenção do grau de Bacharel em Ciência e Economia pelo Instituto de Ciências Sociais Aplicadas da Unifal-MG.

Trabalho Aprovado em: __ / __ / ____

Profª Drª Patrícia de Siqueira Ramos — Assinatura: _____
Universidade Federal de Alfenas

Profª Drª Adriele Aparecida Pereira — Assinatura: _____
Universidade Federal de Alfenas

Profª Drª Gislene Araújo Pereira — Assinatura: _____
Universidade Federal de Alfenas

Varginha

2021

Agradecimentos

Aos meus pais que me apoiaram diariamente, contribuindo para que fosse possível realizar este sonho.

Aos amigos Alex, Cláudio e Mariah, que sempre estiveram ao meu lado, pela amizade incondicional e pelo apoio demonstrado ao longo de todo o curso.

A professora Patrícia, pela orientação e a forma com que conduziu o trabalho com paciência e dedicação, sempre disponível a compartilhar todo o seu vasto conhecimento.

Ao Instituto de Ciências Sociais Aplicadas da Unifal-MG, por disponibilizar uma estrutura de ensino de qualidade e a oportunidade da graduação.

A todos aqueles que contribuíram de alguma forma para que eu chegasse até este momento, me incentivando, não permitindo que eu desistisse e tornando mais leve a caminhada desta conquista.

A Deus, pela oportunidade incrível de conquistar meus objetivos durante todos os meus anos de estudo.

Resumo

A aplicação de métodos da estatística multivariada no mercado tem se tornado fator de competitividade. Eles podem ser utilizados desde para entender as variações e explicar fenômenos até mesmo para realizar previsões, auxiliando no planejamento estratégico das empresas. O objetivo deste trabalho é demonstrar uma aplicação do modelo de regressão logística para a previsão de inadimplência em uma operadora de crédito. Contudo, os dados utilizados para o ajuste do modelo estavam com as classes desbalanceadas. Dessa forma, tornou-se necessário a utilização do método SMOTE, que emprega o método dos vizinhos mais próximos para criar dados da classe minoritária até que ela seja balanceada à classe majoritária. Após o ajuste dos modelos, com e sem a utilização do método SMOTE e, com base nas métricas de avaliação utilizadas, pode-se concluir que o modelo criado com dados balanceados apresentou melhor desempenho.

Palavras-chave: regressão logística, estatística multivariada, SMOTE

Abstract

An application of multivariate statistical methods in the market has become a reference factor. They can be used since to understand the variations and to explain the characteristics of the elements even to make forecasts, helping in the strategic planning of the companies. The objective of this work is to demonstrate an application of the logistic regression model for a default forecast in a credit provider. However, the data used to adjust the model was with an unbalanced class. Thus, it was necessary to use the SMOTE method, using the method of the k-nearest neighbors to create data from the minority class until it is balanced with the main class. After adjusting the models, with and without the use of the SMOTE method, it is concluded that the model presented a better performance with balanced data.

key words: logistic regression, multivariate statistics, SMOTE

Sumário

1. Introdução

O crescimento das economias trouxe consigo a expansão das linhas crédito. Contudo, os alvos deste benefício são famílias que muitas vezes não têm a alfabetização financeira necessária para lidar com todos estes produtos financeiros, levando-as a contrair cada vez mais dívidas. Dessa forma, torna-se cada vez mais importante a utilização de metodologias que permitam a estas empresas, concedentes de crédito, entender melhor sua carteira de clientes para oferecer melhores produtos e evitar grandes prejuízos.

Segundo Mingoti (2005), a análise multivariada se refere a um conjunto de métodos estatísticos que podem ser utilizados em cenários onde várias variáveis são medidas simultaneamente em cada elemento amostral. Dado o elevado número de variáveis, estas podem ser correlacionadas tornando as análises ainda mais complexas caso se pretenda utilizar os métodos mais comuns da estatística univariada.

Uma das características mais importantes de um gestor está na sua capacidade de utilizar de diversas ferramentas para obter uma visão ampla do seu negócio. Partindo deste princípio, a análise multivariada de dados torna-se imprescindível em um cenário de constantes mudanças, em que a geração e coleta de dados tornam-se o novo "petróleo" e as empresas que conseguem lidar melhor com estes dados têm um salto em eficiência e competitividade.

Com a regressão logística, aliada a outras técnicas estatísticas, pode-se obter um modelo para identificar a probabilidade de que um cliente deixe de pagar sua fatura do próximo mês a partir de uma análise de diversas características da carteira de clientes da operadora. Dado este contexto, a questão que se pretende responder com este trabalho é: é possível antecipar a inadimplência dos consumidores e, dessa forma, buscar alternativas para reduzir seu impacto?

A presente pesquisa está dividida em cinco partes. A primeira é a presente seção; em seguida tem-se uma introdução sobre a forma de modelagem utilizada e o método SMOTE para amenizar o efeito do desbalanceamento das classes; a terceira apresenta os dados utilizados e aspectos metodológicos da aplicação do modelo; o quarto apresenta os resultados obtidos através do modelo criado e, por fim tem-se algumas considerações quanto aos resultados da pesquisa e sugestões para trabalhos futuros.

2. Referencial teórico

2.1. Modelos Lineares Generalizados

Os modelos lineares generalizados (GLMs, na sigla em inglês) são uma extensão do modelo linear clássico dado por:

$$\boldsymbol{Y = Z\beta + \varepsilon,}$$

em que **Y** é uma matriz de $(n \times 1)$ observações, **Z** é uma matriz de dimensão $(n \times p)$ associada a um vetor $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\top}$ de parâmetros e $\boldsymbol{\varepsilon}$ é um vetor de erros aleatórios com dimensão $(n \times 1)$ e uma distribuição que se supõe $N \sim (\mathbf{0}, \boldsymbol{\sigma^2})$. Estas hipóteses implicam que $\boldsymbol{E(Y|Z) = \mu}$, com $\boldsymbol{\mu} = \mathbf{Z\beta}$, isto é, o valor esperado da variável resposta é uma função linear das covariáveis (MCCULLAGH; NELDER, 1989).

Esta extensão é feita em duas direções. Por um lado, a distribuição considerada não precisa ser necessariamente normal, podendo ser qualquer distribuição da família exponencial; por outro lado, embora se mantenha a estrutura de linearidade, a função que relaciona o valor esperado e o vetor de covariáveis pode ser qualquer função diferencial (MCCULLAGH; NELDER, 1989).

Os GLMs, são considerados uma das mais importantes classes de modelos, além da regressão. São caracterizados por dois componentes principais: uma distribuição de probabilidade ou família (binomial, no caso da regressão logística) e uma função de ligação mapeando a resposta até os preditores (logito, ainda no caso da regressão logística) (HASTIE, 2009).

Vale observar que a função de ligação é aquela que especifica uma transformação não linear utilizada para modelar respostas em que a variável dependente relaciona-se com as variáveis explicativas de forma não linear (PINO, 2007).

Há casos em que será necessário o uso de modelos GLMs com classes binárias. Quando são modeladas probabilidades de ocorrência será necessário um modelo que tenha uma escala limitada entre 0 e 1. Dessa forma, faz mais sentido modelar em uma escala transformada, que é basicamente o que é feito no modelo logit, que será tratado adiante (DALGAARD, 2008).

2.2. Regressão Logit Binomial

Considere um vetor **y** $(n \times 1)$ com observações sobre uma variável de resposta binária (o valor 1 indica a presença de determinada característica qualitativa e 0 a ausência desta característica). O modelo logit faz a suposição que a probabilidade de se observar $\boldsymbol{y} = 1$ dado uma variável $x_i(x_{i1}, \dots, x_{ip})^T$ é dada pela função logística de uma combinação linear de **x**:

$$p(x_i) = P(y_i = 1 \mid x_i) = \frac{\exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij})}{1 + \exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij})}.$$

Esta equação implica na probabilidade de ausência da variável y que será dada por:

$$1 - p(x_i) = P(y_i = 0 \mid x_i) = \frac{1}{1 + \exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij})},$$

O que implica que:

$$\log\left\{\frac{p(x_i)}{1 - p(x_i)}\right\} = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}.$$

Assim, o modelo logit é equivalente a um modelo log-linear para a razão de chances $p(x_i) \,/\, \{1 - p(x_i)\}$. De forma simples, a razão de chances se refere às chances de determinado evento acontecer dada a influência de determinada variável comparada à chance de o evento acontecer sem a influência desta. Isto é, a razão da chance de um evento ocorrer em um grupo e a chance de ocorrer em outro grupo (MESQUITA, 2014).

Dessa forma, um valor positivo de β_j indica uma variável explicativa x_j que favorece a presença da característica, uma vez que melhora as chances. Um valor de β_j igual a 0 corresponde à ausência de um efeito desta variável sobre o aparecimento da característica qualitativa (HÄRDLE; SIMAR, 2015).

O modelo logit deve ser ajustado usando estimação de máxima verossimilhança (MLE, da sigla em inglês). A estimação de máxima verossimilhança é um processo que tenta encontrar o modelo que mais provavelmente produziu os dados observados.

Em um sentido muito geral, o método da máxima verossimilhança gera valores para os parâmetros desconhecidos que maximizam a probabilidade de obter o conjunto de dados observado. Para aplicar esse método, primeiro é necessário construir uma função, chamada função de verossimilhança. Esta função expressa a probabilidade dos dados observados em função dos parâmetros desconhecidos. Os estimadores de probabilidade máxima desses parâmetros são escolhidos para serem os valores que maximizam essa função. Assim, os estimadores resultantes são aqueles que melhor concordam com os dados observados (HOSMER; LEMESHOW, 2000).

Para observações independentes e identicamente distribuídas, a função de máxima verossimilhança é dada por:

$$L(\beta_0, \beta) = \prod_i^n p(x_i)^{y_i} \{1 - p(x_i)\}^{1-y_i} .$$

O objetivo é estimar um valor β que maximiza a função $L(\beta_0, \beta)$, para isso a manipulação matemática pelo log da verossimilhança é mais fácil. O log da verossimilhança é definido pelo problema de maximização não linear a seguir:

$$(\hat{\beta_0}, \hat{\beta}) = argmax_{\hat{\beta_0}, \hat{\beta}} (\beta_0, \beta) ,$$

em que:

$$\log L(\beta_0, \beta) = \sum_{i=1}^{n} [y_i \log p(x_i) + (1 - y_i) \log\{1 - p(x_i)\}] .$$

Para encontrar o valor de β é preciso diferenciar $L(\beta_0, \beta)$ com relação a β_0 e β_1 e o conjunto resultante é igualado a 0. Essas equações são chamadas de equações de verossimilhança (HOSMER; LEMESHOW, 2000).

Vale dizer que, como a probabilidade está sendo modelada diretamente, isso por si só determinará a variabilidade do resultado binário. Não existe um parâmetro de variação (erro), como na distribuição normal (DALGAARD, 2008). Ademais, quando se tratam de pressupostos, a regressão logística é muito flexível se comparada a outras técnicas estatísticas porque não requer que as variáveis preditoras possuam distribuição normal ou que mantenham

relacionamento linear com a variável dependente e que haja homogeneidade de variâncias em cada um dos grupos da variável dependente (TABACHNICK; FIDELL, 2012).

Contudo, alguns pressupostos são necessários: i) a variável dependente deve ser categórica e binária; ii) as observações devem ser independentes; iii) as variáveis explicativas devem ser linearmente correlacionadas à razão de chances; iv) deve haver ausência de outliers e; v) deve haver ausência de multicolinearidade.

A multicolinearidade pode ser um problema no ajuste do modelo de regressão em que as variáveis independentes ou explicativas possuem relações lineares exatas ou aproximadamente exatas. Esta situação pode causar diversos problemas nas estimativas dos parâmetros e desvios em seu comportamento (MILOCA; CONEJO, 2013).

A diferença entre as regressões logística e linear é que na primeira a variável dependente é disposta em categorias e a resposta é expressa como a probabilidade de ocorrência, enquanto na segunda, a variável dependente é contínua e a resposta é um valor numérico (PINO, 2007).

Dentre as principais vantagens do modelo logit pode-se destacar a facilidade no uso de variáveis categóricas para problemas que envolvam a estimação de probabilidades, os parâmetros do modelo costumam explicar bem o fenômeno de estudo, oferecendo um bom ajuste com menos parâmetros se comparado aos modelos lineares e à grande quantidade de pacotes estatísticos para a implementação da técnica (MESQUITA, 2014).

2.3. O método SMOTE

Faz parte da rotina diária dos pesquisadores que trabalham com análise de dados lidar com conjuntos de dados desbalanceados, ou seja, quando os grupos estudados possuem quantidades diferentes de observações. Uma das técnicas utilizadas para lidar com este tipo de situação é replicar os dados existentes da classe minoritária até que haja um pareamento com a classe majoritária, esta é chamada de *oversampling*. Também é possível realizar o processo contrário, isto é, a remoção de amostras classe majoritária até que haja um balanceamento com a classe minoritária (*undersampling*). Contudo, estas técnicas costumam ser muito atacadas pela comunidade científica, já que podem aumentar o viés do classificador para uma classe específica (MACHADO, 2009).

Para lidar com este problema, Chawla et al. (2002) desenvolveram o método SMOTE (*Synthetic Minority Oversampling Technique*), que consiste em gerar dados sintéticos (não duplicados) da classe minoritária a partir de vizinhos. A Figura 1 mostra como este método funciona, encontram-se os vizinhos mais próximos da classe minoritária (x1, x2 e x3) para cada

amostra das classificações. Em seguida, é necessário traçar uma reta entre o ponto de origem e o vizinho para que seja possível identificar a localização das observações sintéticas (a, b, c e d), escolhidas de forma aleatória (VAZ, 2019).

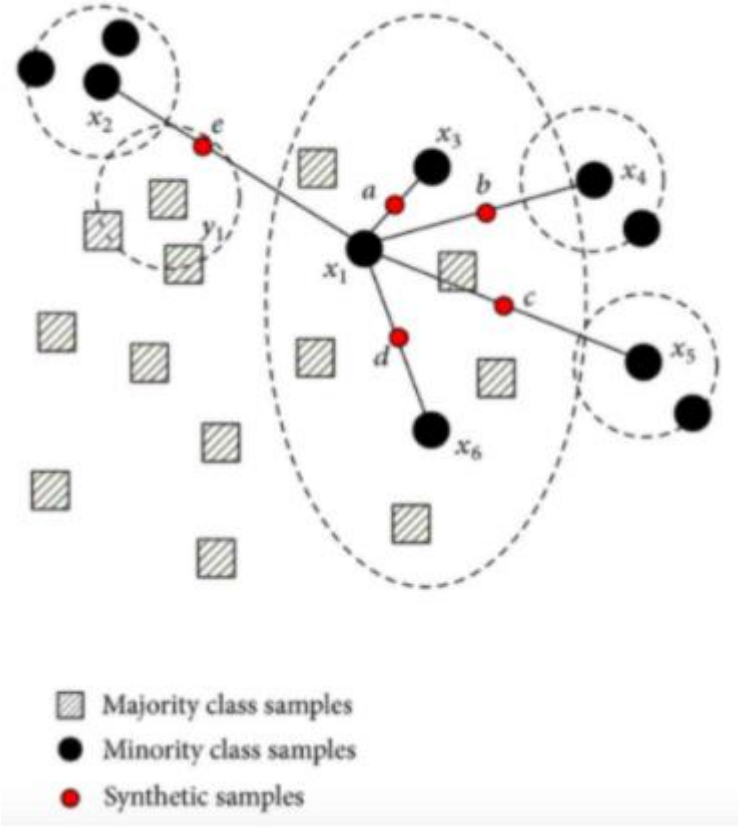

Figura1. Funcionamento do método SMOTE

Fonte: (VAZ, 2019)

Dessa forma, o método SMOTE apresenta-se como uma solução robusta para conseguir lidar com conjuntos de dados desbalanceados. Na próxima seção, apresentam-se os aspectos metodológicos desta pesquisa.

3. Metodologia

3.1. O conjunto de dados

Utiliza-se um conjunto de dados hospedados no site UCI Machine Learning que foi disponibilizado pelo pesquisador I-Cheng Yeh do Departamento de Gestão da Informação da

Universidade Chung Hua, em Taiwan. O conjunto de dados pode ser encontrado em https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients.

Segundo o pesquisador, o conjunto de dados foi coletado em outubro de 2005 em um banco muito importante de Taiwan com o objetivo de analisar o comportamento dos portadores de cartão de crédito do banco.

Vale salientar que conjunto de dados utilizado teve por objetivo apenas atender ao estudo de caso desta pesquisa já que, por pertencer ao ano de 2005, as informações contidas podem não representar o atual cenário econômico mundial.

Os dados representam 30.000 observações, sendo que 22,5% são de clientes adimplentes do banco. A variável resposta escolhida foi o pagamento padrão da fatura em outubro de 2005 (Sim = 1 e Não = 0). Além disso, foram coletadas 23 variáveis explicativas relacionadas ao cliente, que se encontram na Tabela 1.

Tabela 1. Variáveis explicativas do conjunto de dados

Variável	Descrição	Valor assumido
cred	Valor do crédito concedido	Valor em NT$ (Dólar NT)
sexo	Sexo do cliente	1=Masculino, 2=Feminino
edu	Grau acadêmico do cliente	1=Pós, 2=Graduado, 3=Ensino Médio, 4=Outros
est_civil	Estado civil do cliente	1=Casado, 2=Solteiro, 3=Outros
idade	Idade do cliente	Anos
hist_p6	Status do pagamento em abril de 2005	-1=Correto, 1=Atraso de 1 mês, 2=Atraso de 2 meses, 3=Atraso de 3 meses, 4=Atraso de 4 meses, 5=Atraso de 5 meses, 6=Atraso de 6 meses, 7=Atraso de 7 meses, 8=Atraso de 8 meses, 9=Atraso de 9 meses ou mais.
hist_p5	Status do pagamento em maio de 2005	
hist_p4	Status do pagamento em junho de 2005	
hist_p3	Status do pagamento em julho de 2005	
hist_p2	Status do pagamento em agosto de 2005	
hist_p0	Status do pagamento em setembro de 2005	
fatura_6	Valor da fatura em abril de 2005	Valor em NT$ (Dólar NT)
fatura _5	Valor da fatura em maio de 2005	
fatura _4	Valor da fatura em junho de 2005	

fatura _3	Valor da fatura em julho de 2005	
fatura _2	Valor da fatura em agosto de 2005	
fatura _1	Valor da fatura em setembro de 2005	
valorpago_6	Valor do pagamento em abril de 2005	Valor em NT$ (Dólar NT)
valorpago_5	Valor do pagamento em maio de 2005	
valorpago_4	Valor do pagamento em junho de 2005	
valorpago_3	Valor do pagamento em julho de 2005	
valorpago_2	Valor do pagamento em agosto de 2005	
valorpago_1	Valor do pagamento em setembro de 2005	

Fonte dos dados: DUA; GRAFF, 2019

Dadas as características do conjunto de dados utilizado, faz-se necessário tratar sobre a forma com que os modelos estatísticos foram aplicados.

3.2. O modelo

A pesquisa apresentada neste trabalho pode ser considerada de natureza descritiva concentrada em um estudo de caso. As pesquisas descritivas trabalham com base em dados ou fatos colhidos da própria realidade, e "buscam conhecer as diversas relações que ocorrem na vida social política, econômica e demais aspectos do comportamento humano, tanto do indivíduo tomado isoladamente como de grupos e comunidades mais complexas" (CERVO; BERVIAN; SILVA, 2007).

Utiliza-se para a aplicação do método as bibliotecas para manipulação de dados e aplicação de métodos estatísticos da linguagem de programação Python.

Antes de iniciar o processo de modelagem, foi realizada uma análise visando entender melhor o conjunto de dados e selecionar possíveis variáveis explicativas. Além disso, faz-se necessário investigar se há dependências entre essas variáveis, pois existem situações em que essas dependências são significativas, causando efeitos nocivos de multicolinearidade.

Dada a problemática, utiliza-se o fator VIF (*variance inflation factor*) para a identificação dessas correlações. Se nenhum fator for correlacionado, os VIFs de todas as variáveis devem ser 1, representando a ausência de multicolinearidade. Se o VIF apresentar um resultado acima de 10 pode-se presumir que houve inconformidades na estimação dos

coeficientes. A solução mais simples é remover as variáveis altamente correlacionadas (AKINWANDE; DIKKO; SAMSON, 2015).

A Tabela 2 apresenta as variáveis que apresentaram um VIF consideravelmente alto e foram retiradas do modelo.

Tabela 2. Valores VIF das variáveis explicativas que foram retiradas do modelo

Variável	Valor VIF
Idade	11,22
Valor da fatura em setembro de 2005	20,82
Valor da fatura em agosto de 2005	38,22
Valor da fatura em julho de 2005	31,78
Valor da fatura em junho de 2005	29,70
Valor da fatura em maio de 2005	36,08
Valor da fatura em abril de 2005	21,42

Fonte dos dados: DUA; GRAFF, 2019.

Em seguida, faz se necessário uma análise minuciosa dos outliers. Esta foi feita a partir da plotagem de gráficos boxplot, ou diagramas de caixa. Para os boxplots serem construídos, é preciso calcular os quartis 25% (Q1) e 75% (Q3). Em seguida, calcula-se a faixa interquartil (FIQ). Optou-se pela remoção dos outliers e a seguinte regra foi utilizada: todos os dados acima de Q3 + 1,5 FIQ e abaixo de Q1 + 1,5 FIQ foram removidos.

Removidas as variáveis com alta correlação e aplicando um pré-processamento no conjunto de dados para atender aos pressupostos do modelo, os dados foram divididos em conjuntos de treino e teste, com aproximadamente 70% e 30% dos dados, respectivamente. Utilizando os dados de treino obtém-se o melhor ajuste do modelo a partir da seleção das variáveis mais significativas. Com os dados de teste é possível calcular as métricas de avaliação (COELHO, 2018).

Assim como observado pela Figura 2, a variável resposta está desbalanceada nas classes podendo prejudicar a estimação dos parâmetros do modelo. Dessa forma, pode-se utilizar o método SMOTE na busca por um melhor ajuste. Contudo, esta metodologia não lida com variáveis categóricas e por este motivo o modelo foi reduzido a 14 variáveis (excluindo-se o sexo e o grau acadêmico e já desconsiderando aquelas que apresentaram um VIF alto e foram removidas antes do ajuste do primeiro modelo). Com os dados balanceados, novamente realiza-

se uma divisão dos dados em treino e teste e ajusta-se um novo modelo selecionando as variáveis mais significativas.

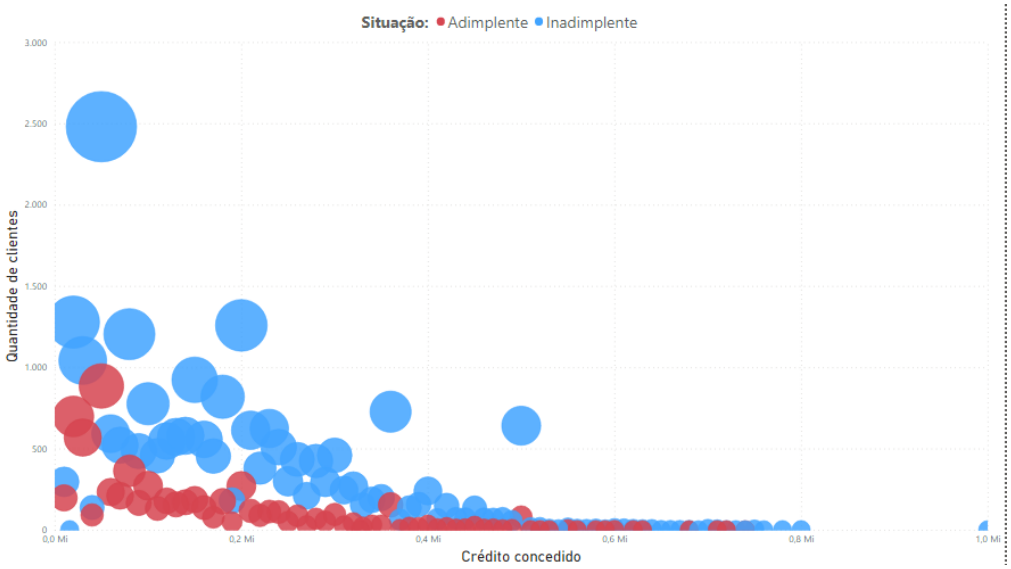

Figura 2. Quantidade de clientes por classe

Fonte dos dados: DUA; GRAFF, 2019.

3.3. Métricas de avaliação

Após o ajuste do modelo, o primeiro passo é avaliar a significância de cada variável, por este motivo são utilizados testes de hipóteses com o objetivo de observar se a variável explicativa é realmente correlacionada com a variável dependente.

Além disso, faz-se necessária uma análise dos coeficientes do modelo, estes apresentam o comportamento da probabilidade de ocorrência de determinado evento à medida que varia uma unidade da variável explicativa, se o coeficiente for positivo quanto maior for o seu valor maior o impacto no aumento da probabilidade de ocorrência do evento de estudo (MESQUITA, 2014).

Outra interpretação muito utilizada se refere à razão de chances, que analisa o impacto de cada coeficiente sobre a chance de o evento ocorrer em determinado grupo ou outro. Para isso, basta calcular a exponencial do próprio coeficiente para obter o impacto que ele exerce sobre a razão de chance (MESQUITA, 2014).

A avaliação do desempenho dos modelos será feita com base em 4 métricas: a acurácia, a F-Score, a área sob a curva ROC e uma análise gráfica da matriz de confusão. A acurácia

(ACC) do modelo conta a quantidade de acertos que os modelos tiveram independente da classe:

$$ACC = \frac{ACERTOS}{TOTAL}$$

Dois conceitos se fazem necessários: precisão e revocação. A precisão pode ser entendida como a proporção de predições positivas que estão corretas, ou seja, quão bem o modelo predisse os valores positivos, nenhum exemplo negativo é incluído. Já a revocação, mostra o quanto de informação relevante foi recuperada com relação ao total de informações relevantes, nenhum exemplo positivo é deixado de fora (MATOS et al., 2009). Contudo é necessário ter cuidado já que quanto mais aumentamos os acertos (melhorar a precisão), menos estamos dispostos a errar (aumentar a revocação).

Contudo, por serem medidas de difícil interpretação, o F-Score (também chamado de F1 ou medida F) combina as duas medidas e apresenta uma visão geral da qualidade do modelo, quanto maior o F-Score, mais apurado é o modelo. Sua fórmula é dada por:

$$F_\alpha = \frac{1}{\frac{1-\alpha}{REVOCAÇÃO} + \frac{\alpha}{PRECISÃO}}, 0 < \alpha < 1$$

Por fim, a matriz de confusão realiza uma comparação entre os valores reais e preditos pelo classificador. Já a curva ROC, como pode ser visto na Figura 3, é uma representação gráfica dos valores de precisão (eixo x) e revocação (eixo y) para os diferentes pontos da validação. Quanto maior a área sob a curva ROC, melhor será o desempenho do modelo.

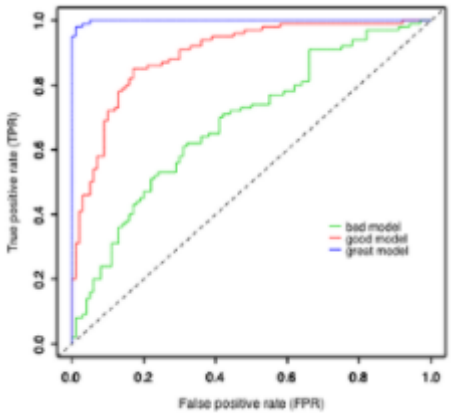

Figura 3. Curva ROC

Fonte: GURGEL, 2021

4. Resultados e discussão

4.1 Análise descritiva dos dados

O Gráfico 1 apresenta um gráfico de colunas com a frequência observada em cada classe da variável resposta, que é o pagamento padrão da fatura. De acordo com o Gráfico 1, a maior parte dos clientes do conjunto de dados está inadimplente. O número de inadimplentes representa cerca de 77% do total de clientes.

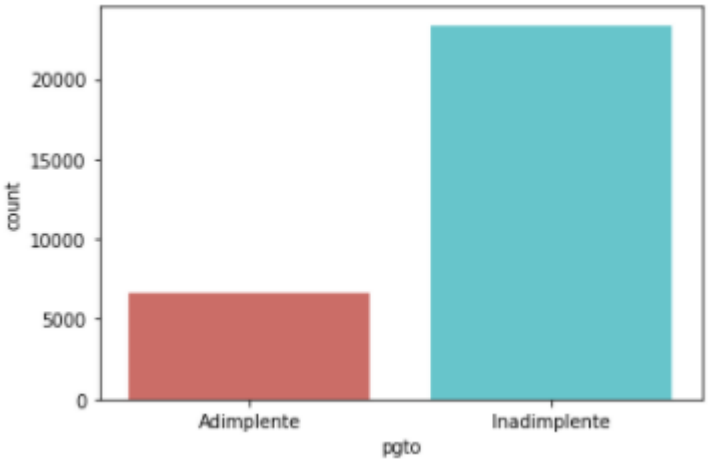

Gráfico 1. Classificação dos clientes por situação

Fonte dos dados: DUA; GRAFF, 2019.

A Tabela 3 mostra um resumo das principais estatísticas da variável crédito concedido. Ao analisar os dados sobre o crédito concedido, 95% dos clientes conseguiram um empréstimo

de até NT$430.000,00. Além disso, em média, o valor do crédito é de NT$167.484,00. O Gráfico 2 mostra a frequência com que os valores de crédito estão distribuídos.

Tabela 3. Análise do crédito concedido (valores em Dólar NT)

Mínimo	Máximo	Média	Desvio Padrão	Mediana
NT$ 10.000	NT$ 1.000.000	NT$ 167.484,00	NT$ 129,747	NT$ 140.000

Fonte dos dados: DUA; GRAFF, 2019.

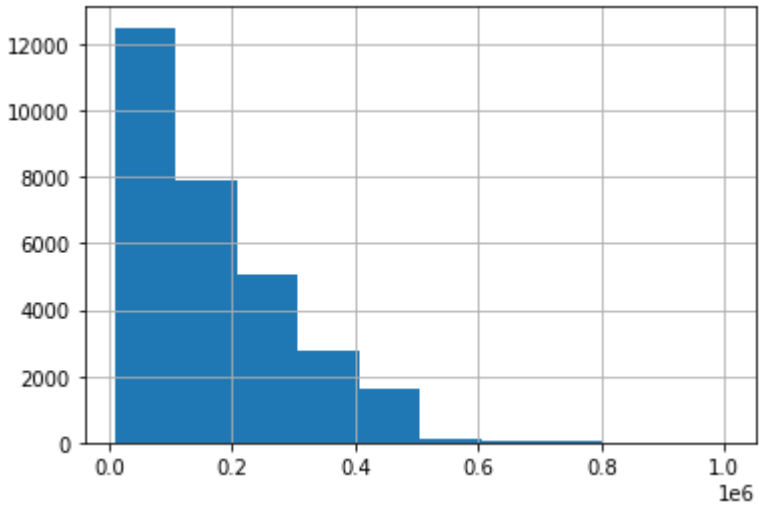

Gráfico 2. Distribuição de frequência do valor do crédito concedido

Fonte dos dados: DUA; GRAFF, 2019.

A amplitude de idades dos clientes está entre 21 a 79 anos. Contudo, 95% dos clientes têm até 53 anos e, assim como pode ser observado no Gráfico 3, em média os clientes têm entre 29 a 34 anos de idade.

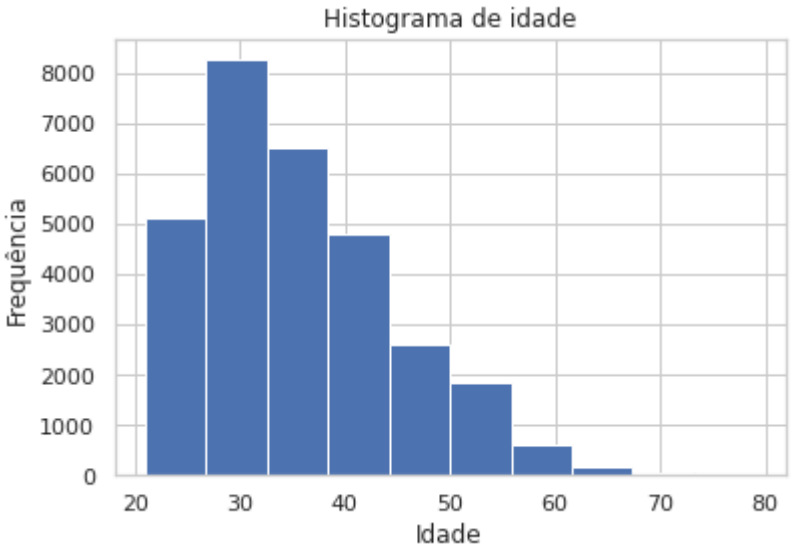

Gráfico 3. Histograma de frequência das idades dos clientes

Fonte dos dados: DUA; GRAFF, 2019.

O conjunto de dados analisado contém 60,40% dos clientes pertencentes ao sexo feminino e, ao analisar o Gráfico 4, nota-se que embora a maior parte dos homens e mulheres estão inadimplentes e haja mais mulheres que homens no conjunto de dados, a diferença proporcional entre adimplentes e inadimplentes é muito pequena comparando os dois sexos.

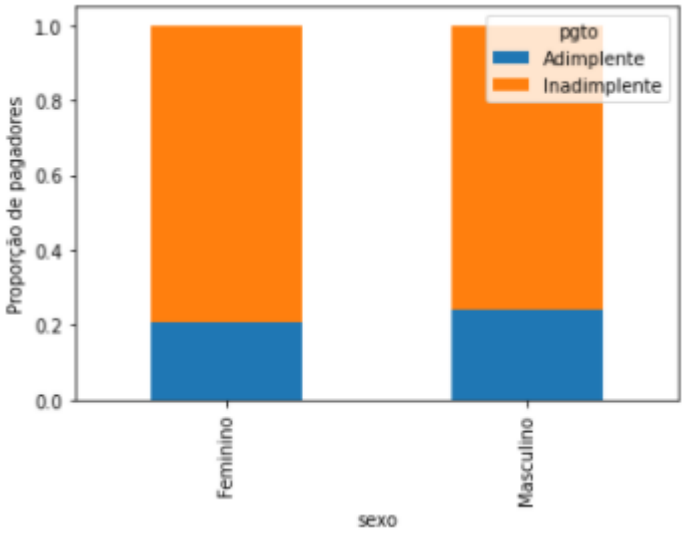

Gráfico 4. Status dos clientes por sexo

Fonte dos dados: DUA; GRAFF, 2019.

Outro resultado importante se dá com relação ao grau de escolaridade. A análise descritiva do Gráfico 5 mostra que, quanto maior o grau acadêmico do cliente, maior a proporção de clientes inadimplentes.

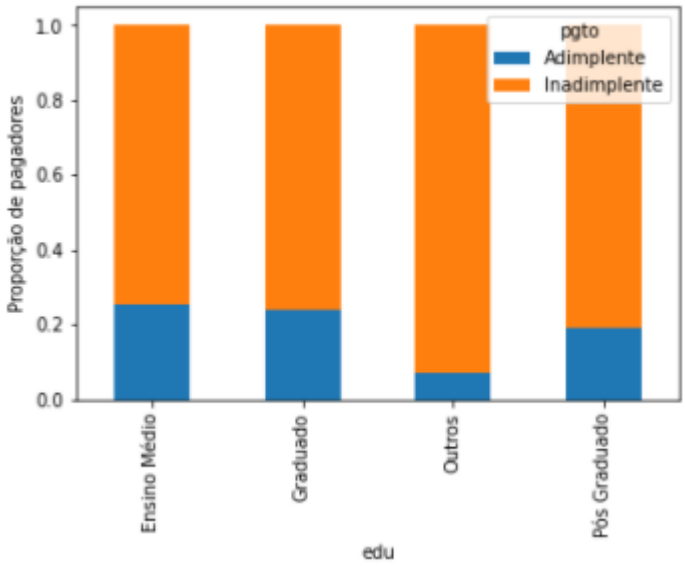

Gráfico 5. Status dos clientes por grau acadêmico

Fonte dos dados: DUA; GRAFF, 2019.

Por fim, ao analisar o estado civil dos clientes no Gráfico 6, nota-se que a maior parte dos adimplentes são os divorciados, seguidos pelos casados.

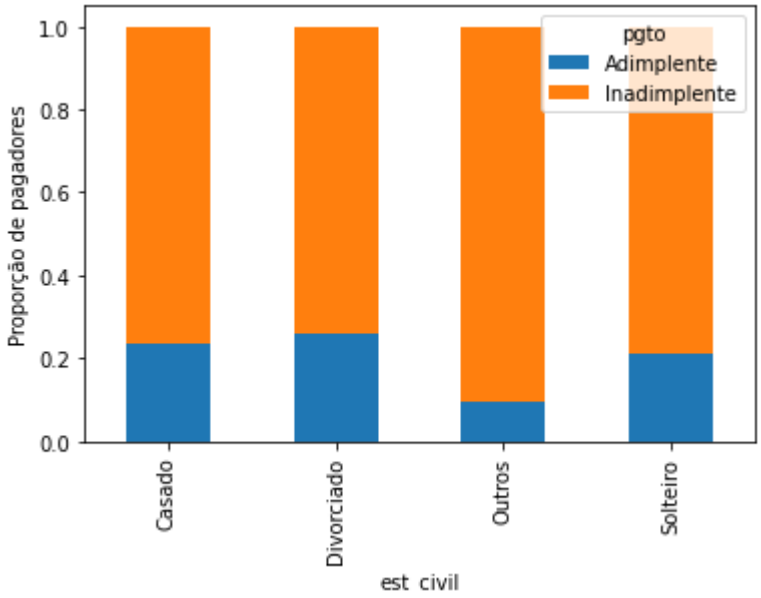

Gráfico 6. Situação de pagamento por estado civil

Fonte dos dados: DUA; GRAFF, 2019.

4.2. Resultados do modelo

Ao avaliar os coeficientes do modelo com as variáveis dependentes desbalanceadas (Tabela 4), nota-se que o histórico de pagamento nos meses de julho (hist_p3), agosto (hist_p2) e setembro (hist_p0) de 2005 tem impacto positivo sobre as chances de haver o pagamento da próxima fatura. Dessa forma, conclui-se que quanto mais faturas atrasadas o cliente tiver, maiores as chances de que ele cumpra o compromisso de pagar a fatura do próximo mês.

No mês de setembro, este impacto é de 1,89 vezes mais chances de adimplência para cada unidade acrescida.

Outro grupo de variáveis que também se apresentaram significativas para o modelo foram os valores pagos nas faturas dos últimos 2 meses (valor_pago1 e valor_pago2) e o valor pago no mês de abril de 2005 (valor_pago6). Estas têm um impacto negativo sobre as chances de adimplência.

No mês de setembro, para cada unidade acrescida no valor da fatura, aumenta em 0,78 vezes as chances de adimplência. Já no mês de agosto esta proporção é de 0,82, um número parecido com a do mês de abril que é de 0,84 vezes.

Tabela 4. Resultados do modelo com a variável resposta desbalanceada (modelo 1)

variável	β	*odds ratio*
intercepto	-1,3698	
hist_p0	0,64	1,89
hist_p2	0,09	1,09
hist_p3	0,16	1,18
valor_pago1	-0,25	0,78
valor_pago2	-0,2	0,82
valor_pago6	-0,17	0,84

Fonte dos dados: DUA; GRAFF, 2019.

A Tabela 5 apresenta os coeficientes do modelo ajustado com dados balanceados. Avaliando os resultados nota-se que, quanto maior o valor do crédito concedido (cred), menores as chances de o cliente ser adimplente. Ainda, a cada unidade acrescida no valor do crédito, as chances de se tornar adimplente será 0,94 vezes menor.

Além disso, o status de pagamento nos últimos 3 meses (hist_p0, hist_p2, hist_p3) influenciam positivamente na efetivação do pagamento no próximo mês. Quanto maior o atraso do cliente, maiores as chances de que ele pague a próxima fatura. O mais impactante deles é o de setembro de 2005, em que a cada unidade aumentada na variável aumenta-se em 1,79 vezes as chances de adimplência no próximo mês.

Por fim, quanto maior os valores pagos nas faturas dos últimos 6 meses (valor_pago1, valor_pago2, valor_pago3, valor_pago4, valor_pago5, valor_pago6), tem-se um impacto negativo sobre a chance de adimplência do cliente. Quanto maior o valor da fatura em setembro de 2005, por exemplo, a chance de o cliente ser adimplente torna-se 0,74 vezes menor para cada unidade acrescida.

Tabela 5. Resultados dos modelo com a variável resposta balanceada (modelo 2)

variável		*odds ratio*
intercepto	-0,191	
cred	-0,06	0,94
hist_p0	0,58	1,79
hist_p2	0,08	1,08
hist_p3	0,12	1,13
valor_pago1	-0,31	0,74
valor_pago2	-0,2	0,82
valor_pago3	-0,03	0,97
valor_pago4	-0,07	0,93
valor_pago5	-0,06	0,94
valor_pago6	-0,14	0,87

Fonte dos dados: DUA; GRAFF, 2019.

De acordo com as métricas consideradas nesta pesquisa, ao comparar os dois modelos (Tabela 6) é possível observar que, embora o modelo com dados desbalanceados apresente uma taxa de acerto maior, 81%, o F-score mostra que a qualidade do modelo com os dados balanceados é melhor. Contudo, como os modelos levam em conta variáveis explicativas distintas não se pode afirmar que a melhora da qualidade do modelo se deve apenas ao balanceamento das classes.

Tabela 6. Métricas de avaliação

Métrica	Resultados - Modelo 1	Resultados – Modelo 2
Acurácia	0,81	0,66
Precisão	0,73	0,37
Recall	0,23	0,67
F-Score	0,36	0,47
AUCROC	0,61	0,66

Fonte dos dados: DUA; GRAFF, 2019.

Esta situação também pode ser confirmada pela matriz de confusão dos dois modelos, representados nas Figuras 4 e 5. Como os dados estão desbalanceados para a classe dos inadimplentes, o modelo tem dificuldades para acertar os clientes adimplentes. Observa-se que a dificuldade é amenizada quando os dados foram balanceados via método SMOTE.

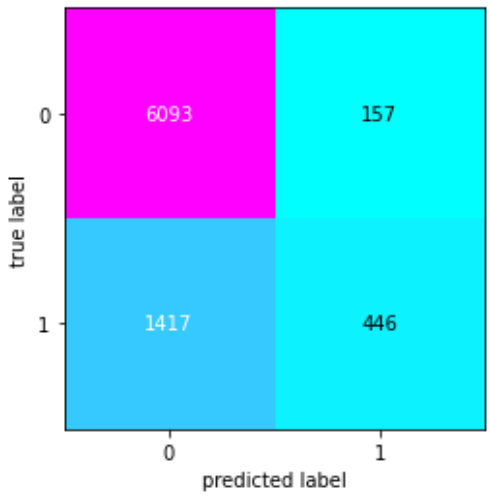

Figura 4. Matriz de confusão – Modelo 1

Fonte: DUA; GRAFF, 2019.

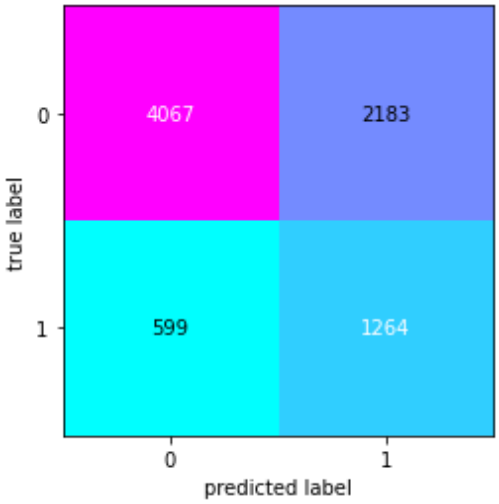

Figura 5. Matriz de confusão – Modelo 2

Fonte: DUA; GRAFF, 2019.

Por fim, ao analisar a curva ROC dos dois modelos nota-se pelo Gráfico 8 que o modelo com dados balanceados tem um melhor ajuste, apresentando uma área sobre a curva ROC maior comparado ao modelo criado com dados desbalanceados (Gráfico 7).

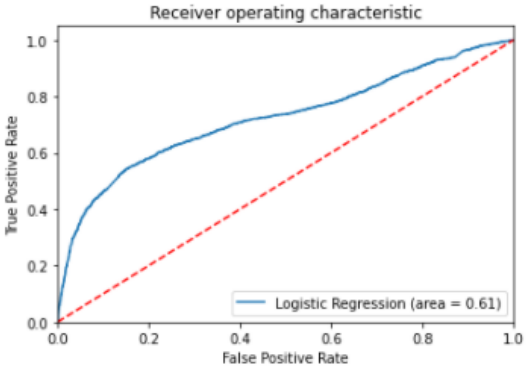

Gráfico 7. Curva ROC – Modelo 1

Fonte dos dados: DUA; GRAFF, 2019.

Gráfico 8. Curva ROC – Modelo 2

Fonte dos dados: DUA; GRAFF, 2019.

5. Considerações Finais

Considerando que o objetivo deste trabalho foi apresentar um exemplo de aplicação do método de regressão logística no mercado de crédito, observa-se que através do modelo é possível identificar as principais características dos clientes que têm mais chances de se tornarem inadimplentes.

O modelo desbalanceado apresentou os status de pagamento dos últimos três meses, valor pago nos últimos dois meses e o valor pago no mês de abril de 2005 como significativos. O primeiro grupo de variáveis com um impacto positivo sobre a variável dependente e as variáveis relacionadas ao valor do pagamento com um impacto negativo.

Já o modelo balanceado, apresenta com um impacto negativo sobre a classificação de adimplência o valor do crédito concedido e o valor pago nos últimos seis meses. O status no pagamento dos últimos três meses, assim como no modelo com dados desbalanceados, tem impacto positivo sobre o pagamento padrão no próximo mês.

Outro ponto extremamente importante trata-se da forma com que o problema do desbalanceamento dos dados foi resolvido. De acordo com as métricas de avaliação escolhidas, o método SMOTE apresentou melhores resultados para o ajuste do segundo modelo, embora não se possa afirmar que esta melhora se deu apenas pelo balanceamento da variável resposta. Percebe-se que um conjunto de dados com a variável dependente desbalanceada pode tornar o modelo enviesado, isto é, o modelo tende a acertar a classe majoritária em detrimento da classe minoritária.

Comparando os dois modelos, foi possível concluir que o modelo com dados balanceados tem um melhor ajuste comparado ao modelo com a classe desbalanceada. Embora

o modelo desbalanceado apresente uma taxa de acerto maior, ele é enviesado para a classe majoritária.

A aplicação de modelos para a previsão de inadimplência pode ser extremamente relevante não apenas para as empresas concedentes de crédito como para a criação de políticas públicas, já que é possível entender melhor as características dos possíveis clientes inadimplentes e buscar soluções para amenizar a situação. Como sugestão, outras técnicas de classificação e novos conjuntos de dados poderiam ser utilizados na busca por um melhor ajuste do modelo.

6. Referências bibliográficas

AKINWANDE, M. O.; DIKKO, H. G.; SAMSON, A. Variance Inflation Factor: As a Condition for the Inclusion of Suppressor Variable(s) in Regression Analysis. **Open Journal of Statistics**, v. 05, n. 07, 2015.

CERVO, A. L.; BERVIAN, P. A.; SILVA, R. **Metodologia Científica**. 6. ed. São Paulo: Pearson Prentice Hall, 2007.

CHAWLA, N. V.; BOWYER, K. W.; HALL, L. O.; KEGELMEYER. SMOTE: **Synthetic Minority Over-sampling Technique**. JAIR, v. 16, p.321-357, 2002.

COELHO, G. **Treino é treino, teste é teste**. Disponível em: <https://gusrabbit.com/intuition/treino-teste/#:~:text=A divisão em datasets de,melhores parâmetros para o mesmo>. Acesso em: 17 mar. 2021.

DALGAARD, P. **Introductory Statistics With R**. 2. ed. New York: Springer, 2008.

DUA, D.; GRAFF, C. **UCI Machine Learning Repository: Data Sets**. Disponível em: <https://archive.ics.uci.edu/ml/index.php>. Acesso em: 13 mar. 2021.

GURGEL, G. K. Como Avaliar Seu Modelo de Classificação. **Turing Talks, 2021**. Disponível em https://medium.com/turing-talks/como-avaliar-seu-modelo-de-classifica%C3 %A7%C3% A3o-acd2a03690e Acesso em 14 de março de 2021

HÄRDLE, W. K.; SIMAR, L. **Applied multivariate statistical analysis, fourth edition**. 4. ed. Berlin: Springer Berlin Heidelberg, 2015.

HASTIE, T. et al. **Springer Series in Statistics: The Elements of Statistical Learning**. New York: Sprienger series in statistics, 2009. v. 27

HOSMER, D. W.; LEMESHOW, S. **Applied logistic regression. 2nd Edition**. 2. ed. New York: Wiley, 2000.

MACHADO, E. L. Um estudo de limpeza em base de dados desbalanceada e com sobreposição de classes. **Area**, 2009.

MATOS, P. et al. **Relatório Técnico "Métricas de Avaliação"**. Universidade Federal de Sao Carlos. **Anais**...2009.

MCCULLAGH, P.; NELDER, J. A. Generalized Linear Models. In: CHAPMAN AND HALL (Ed.). . **Statistics in the 21st Century**. 2. ed. New York: CRC Press, 1989.

MESQUITA, P. S. B. **Um modelo de regressão logística para avaliação dos programas de pós-graduação no Brasil**. [s.l.] Universidade Estadual do Norte Fluminense, 2014.

MILOCA, S. A.; CONEJO, P. D. **Multicolinearidade em Modelos de Regressao**. XXII SEMANA ACADÊMICA DA MATEMÁTICA. **Anais**...Cascavel: 2013

MINGOTI, S. A. **Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada**. Belo Horizonte: Editora UFMG, 2005.

PINO, F. A. MODELOS DE DECISÃO BINÁRIOS: uma revisão. **Revista de Economia Agrícola**, v. 54, n. 1, 2007.

TABACHNICK, B. G.; FIDELL, L. S. **Using multivariate statistics (6th ed.)**. New York: Harper Collins, 2012.

VAZ, A. L. **Como lidar com dados desbalanceados em problemas de classificação**.

Disponível em: <https://medium.com/data-hackers/como-lidar-com-dad os-desbalanceados-em-problemas-de-classificação-17c4d4357ef9>. Acesso em: 10 mar. 2021.