Gabriela Machado Moura

# Regressão Logística aplicada a análise de risco de crédito

Brasil
2018

*Este trabalho é dedicado ao meu maior amor, minha mãe. E ao meu pai, que a saudade só não é maior que o meu amor.*

# Agradecimentos

Primordialmente, gostaria de agradecer a minha orientadora Raquel Nicolette, pela honra de ter sido sua orientanda, pela sua disponibilidade e por ter acolhido minha pesquisa, por todo conhecimento e experiência compartilhada. Parafraseando Novos Baianos 'E pela lei natural dos encontros eu deixo e recebo um tanto' pois eu sei que você gosta de música boa e com certeza amadureci muito com este trabalho e principalmente com você, profissionalmente e como pessoa.

Gostaria de agradecer a minha mãe, Rosângela Machado Moura, que foi e é minha base e inspiração, por todo seu apoio e motivação para que eu pudesse alcançar esta conquista. Mesmo distantes nunca estivemos tão próximas, obrigada por tudo, meu maior amor, minha melhor amiga, te amo.

Ao meu irmão, Guilherme Machado Moura, que há 1.500km de distância me fazia rir até doer a barriga nos momentos que eu mais precisei.

Aos meus avós paternos e maternos, principalmente ao meu avô José Benevenuto Machado que esteve ao meu lado, tanto que até sonhava que se atrasava pra aula.

Ao meu companheiro, Marcos Ramis por estar do meu lado e mais do ninguém sabe o que essa tal Matemática me fez fazer, mas mais do que isso por sua parceria, te amo amor. E também por se tornar minha família e ter emprestado a sua, obrigada Família Ramis por toda hospitalidade, acolhimento e almoços especiais (haha).

A professora Suzi Samá, por ter plantado a semente da Estatística em mim.

*"A educação é um elemento importante na luta pelos direitos humanos. É o meio para ajudar os nossos filhos e as pessoas a redescobrirem a sua identidade e, assim, aumentar o seu auto-respeito. Educação é o nosso passaporte para o futuro, pois o amanhã só pertence ao povo que prepara o hoje."(Malcolm X)*

# Resumo

A concessão de crédito tem um papel fundamental na economia de um país. Os modelos de *Credit Scoring* fazem a estimativa da probabilidade de um solicitante de crédito se tornar inadimplente com base nas suas informações pessoais e financeiras. Nesse sentido, este trabalho tem por objetivo desenvolver um modelo de *Credit Scoring* ('pontuação de crédito') utilizando a técnica estatística de Regressão Logistíca, com a finalidade de classificar pessoas físicas tomadoras de crédito como adimplentes ou inadimplentes. O modelo desenvolvido foi aplicado em dois conjuntos de dados, sendo um destes cedidos por um microoperadora de crédito do estado do Rio Grande do Sul, que apresentou resultado excelente no poder de discriminação do modelo, alcançando uma taxa de acerto geral de 97%. Bem como, a aplicação em um conjunto de dados clássicos que obteve um poder de discriminante aceitável e uma taxa de acerto geral de 72%.

**Palavras-chave**: Regressão Logística. Risco de Crédito. Credit Scoring, Estatística, Matemática Aplicada.

# Lista de ilustrações

# Lista de tabelas

# Sumário

# 1 Introdução

A expansão de crédito no Brasil foi uma ferramenta significativa para o desenvolvimento socioeconômico do país, desde o Plano Real em 1994 aumentou consideravelmente a quantidade de crédito concedido as famílias, o que é essencial na economia de um país pelo seu impacto no Produto Interno Bruto – PIB. Instituições financeiras disponibilizam crédito em troca de um ganho sobre o capital emprestado.

As Crises do petróleo em 1979 geraram a Crise da divida externa em 1982 e a alta inflação atingiu valores de até 227% ao ano. Posteriormente, na democracia a criação de diversos planos fracassados na tentativa do governo intervir na economia, Plano Cruzado 1986, Plano Bresser 1987, Plano Verão 1989, Plano Collor I e II 1990, Plano Brady 1993, que somente fomentava o crescimento inflacionário com a desestabilização da moeda, que trocava constantemente, pelo menos quatro vezes num período de dez anos, fez com que a inflação anual acumulada no mês anterior ao Plano Real atingisse cerca de 5.150,0%. [1]

Os quinze anos de hiperinflação no Brasil denunciavam um grande problema econômico do país, que era o excesso governamental e a impressão descontrolada de dinheiro, era preciso um Plano inovador que estabilizasse a moeda e a inflação no país, então se pensou em um novo plano econômico, o Plano Real implantado no governo de Itamar Franco, baseado no Tripé Macroeconômico (responsabilidade fiscal, metas para inflação, câmbio flutuante), porém desta vez antes de colocar em vigor, o governo preparou a economia para receber essa nova moeda e constituir as reformas necessárias, criou-se então uma moeda virtual URV (unidade real de valor) a qual primeiro a sociedade se adaptou até a substituição então pela nova moeda, o Real, que criou uma segurança na moeda estabilizando-a e consequentemente o controle da inflação.

A redução de pobreza e desigualdade propiciou a capacidade das pessoas se planejarem frente ao futuro e entrarem na sociedade do consumo. Com isso estabeleceu-se um novo cenário na economia do país, de confiança, no qual propunha uma estabilidade maior, aumentando assim na demanda por crédito e o aumento de financiamentos, dado a diminuição de juros e taxas.

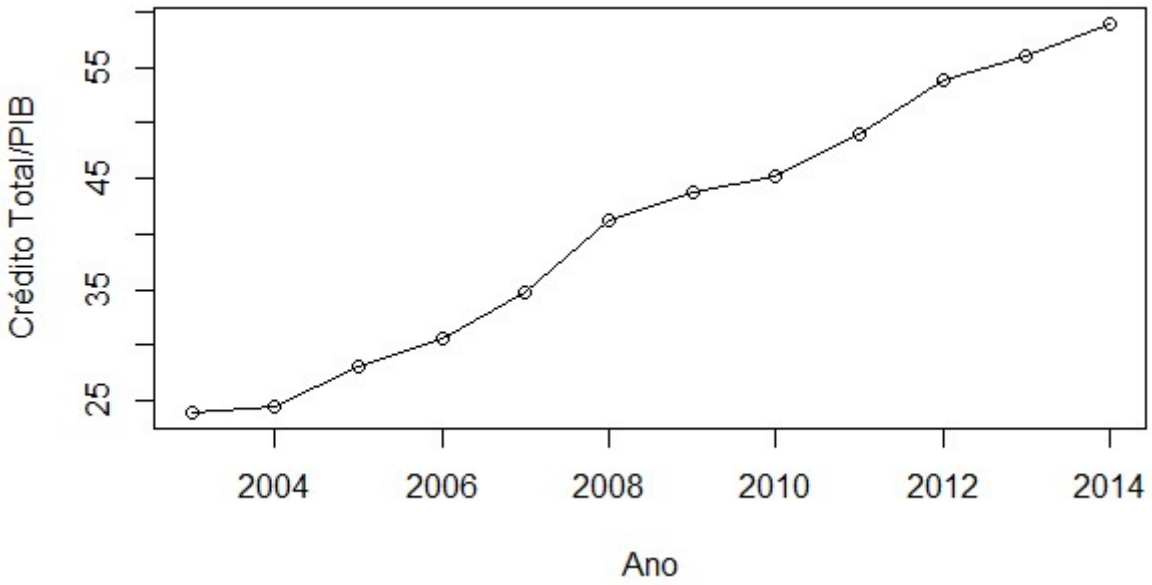
Contudo, quem mais sofreu com a queda da inflação foram os bancos, que obtinham seus lucros principalmente da receita inflacionária provinientes da instabilidade da moeda, os bancos tiveram que buscar outros mecanismos para obter lucros e novas receitas, expandindo assim emergencialmente a concessão de crédito e empréstimos financeiros (ROSA, 2000).

Como podemos observar na Figura 1 desde os anos 2000 a importância do crédito

[1] A referências para essas taxas de inflação é IGP-DI/FGV. Veja, http://www.fgv.br.

total em relação ao PIB, sendo sua contribuição de 27%, após 2006 percebe-se uma elevação no crescimento do crédito total que de 30,8% e elevou-se em 2007 para 35,2%. Recentemente o crédito total é responsável por contribuir com mais da metade do PIB do país.

Figura 1 – Crédito total em relação ao PIB(%) brasileiro 2003-2014




Fonte: Banco Central do Brasil (2014)

Após o Plano Real, no Brasil percebe-se o crescimento elevado da concessão do crédito sem a análise adequada, devido à mudança repentina de cenário, o setor não se preparou e alguns dos maiores bancos não resistiram e quebraram. A dificuldade na análise de crédito impulsionou os desenvolvimentos de modelos que dêem suporte ao analista de crédito, em busca de diminuir as perdas do setor bancário, devido ao alto nível de inadimplência que ocorrera.

Os primeiros modelos que foram desenvolvidos na década de 1970 começaram a se popularizar como suporte aos analistas e gerentes de crédito. A análise de concessão de crédito até o século XX era baseada exclusivamente por gerentes de créditos ou analistas (THOMAS, 2000). O que ocasionava em uma análise relativa, dado que em uma mesma instituição uma pessoa poderia receber ou não a concessão de crédito, dependendo do analista que examinasse o pedido.

As instituições financeiras necessitam de cuidado na análise do perfil do cliente, bem como na tomada de decisão de conceder ou não o crédito, visto que "qualquer erro na decisão de concessão pode significar que em uma única operação haja a perda do ganho obtido em dezenas de outras bem sucedidas", correspondendo a sobrevivência da empresa, "analisar uma proposta de negócio e comparar o custo de conceder com o custo de negar a operação" (STEINER, *et al.* , 1999).

Segundo Lewis (1992) disponibilizar crédito ao consumidor é um empreendimento essencial, pois é rentável para empresa, assim quanto mais ampla for a disponibilidade desta ferramenta ao consumidor, mais rentável a empresa será. Nesse sentindo, intrínseca a concessão do crédito temos o risco da probabilidade de inadimplência.

Com o desenvolvimento da análise descriminante por Fisher (1936), a qual a partir de características disponíveis de um indivíduo, gera um modelo de classificação, no qual permite inferir a que população este indivíduo pertence. O que propiciou os primeiros modelos de *credit scoring* (pontuação de crédito), que criam uma pontuação de crédito a fim de ordenar ou classificar os clientes frente a probabilidade de pagar o empréstimo concedido, a probabilidade de risco de crédito. As análises de *credit scoring* são baseadas em modelos estatísticos que fazem o uso de técnicas multivariadas que possibilitam que se analise o comportamento de crédito de um conjunto de indivíduos.

Contudo, o uso desta ferramenta estatística ao invés da experiência de um analista ou gerente de crédito, não foi bem aceita inicialmente, somente com o aumento da demanda de solicitantes de créditos, evidenciou-se a inviabilidade de analisar cada pedido individualmente. Agregando mais agilidade na tomada de decisão gerando diminuição de custos e poder preditivo, os modelos de credit scoring se popularizaram, sendo o mais utilizado atualmente (HAND  HENLEY, 1997).

Com base nisso, o presente trabalho utiliza os modelos de *Credit Scoring* e tem por objetivos:

## 1.1 Objetivos

**Objetivos gerais**

1. Desenvolver um modelo de *Credit Scoring* por meio de Regressão Logística.
2. Aplicar o modelo desenvolvido em dados clássicos.
3. Estudo de caso aplicando o modelo desenvolvido em dados reais.
4. Analisar o modelo na classificação dos clientes para a concessão do crédito bancário.

**Objetivos específicos**

Como objetivos específicos parte-se da análise direta das variáveis:

- Determinar bases de dados e as variáveis a serem utilizadas.
- Identificar as variáveis com o maior poder discriminante entre os clientes.
- Atribuir pesos para as mesmas;

- Estabelecer critérios de eficiência e qualidade.

Um desafio aqui encontrado foi associar os conceitos econômicos com os matemáticos, desta forma este trabalho está dividido da seguinte forma: no Capítulo 2 são apresentadas os principais conceitos da área econômica aqui utilizados, o Capítulo 3 apresenta toda a fundamentação matemática. A parte prática de análise de *credit scoring* é dada no Capítulo 4 e um estudo de caso é apresentado no Capítulo 5.

# 2 Revisão Bibliográfica

Nesta seção serão abordados os conceitos aplicados à economia, fundamentais para o entendimento deste trabalho. Inicialmente tem-se o conceito de Crédito, a definição de Probabilidade de Risco de Crédito, bem como a caracterização de um modelo *credit scoring*.

## 2.1 Crédito

A concessão de crédito depende de duas partes, a *credora* e a *devedora*. A credora é aquela que empresta o dinheiro a uma pessoa ou instituição, por isso crê em que a contraparte devedora devolva o dinheiro com um prêmio de risco, chamado juros. Dentre as inúmeras definições de crédito, a origem desta palavra vem do latim creditu, a qual significa "eu acredito" ou "confio". Conceder crédito é confiar, acreditar na contraparte devedora com sustentações nas informações disponíveis sobre o seu passado e o presente, e principalmente a perspectiva a cerca do futuro, no qual é intrínseco o risco e a incerteza. Neste trabalho adotaremos crédito de acordo com Schrickel (1995):

> "Todo ato de vontade ou disposição de alguém de destinar ou ceder, temporariamente, parte de seu patrimônio a um terceiro, com a expctativa de que esta parcela volte à sua posse integralmente depois de decorrido o tempo estipulado."(p.9)

O conceito de crédito pode ser aplicado em:

- Compras à prazo → Instituições comercias. (Exemplo: Lojas)
- Concessão de empréstimo →Instituições Financeiras. (Exemplo: Bancos)

Neste trabalho iremos abordar o conceito de crédito do ponto de vista das Instuições Financeiras ao conceder crédito a pessoas físicas. No sistema bancário, crédito significa fornecer para o cliente (captador de recursos) um financimento ou empréstimo, frente a um cadastro pré-aprovado para o comprimento da promessa de pagamento futura.

## 2.2 Probabilidade de Risco de Crédito

Weerthof (2011) define risco no setor bancário quando este concede crédito e não recebe o reembolso integral ou parcial do acordado. Para Gitman (1997) risco é a possibilidade de um prejuízo financeiro. Corroboram Caouette *et al.* (2000), "se credito

pode ser definido como a expectativa de recebimento de uma soma em dinheiro em um prazo determinado, então Risco de Crédito é a chance que esta expectativa não se concretize". Ademais Lewis (1992) destaca a imprevisibilidade do futuro, no qual resulta o fato de que tanto do ponto de vista lógico ou por testes, nem todas as dívidas serão pagas como o acordado.

Nesse sentido, conclui-se que nem todos os contratos irão ser pagos, por isso o banco aplica a sua taxa de juros, sendo que uma parcela desta é o serviço que o banco lhe presta. Porém, boa parte destes juros é composto pelo fator de risco que esta instituição financeira possui ao realizar este empréstimo, pois nem todos os contratos serão cumpridos. Desta forma para garantir este percentual de perda as pessoas que pagaram devidamente cobrem o custo das pessoas que não pagaram, garantindo assim o lucro dos bancos. O risco então esta relacionado a instabilidade de possíveis retornos.

No âmbito financeiro, para avaliação do risco de crédito, as instituições utilizam principalmente técnicas qualitativas e quantitativas, sendo isso um levantamento de dados com o objetivo de avaliar as probabilidades envolvidas na negociação. Técnicas qualitativas dependem de um analista ou gerente de crédito para fazer o julgamento do cliente para concessão do crédito, alicerçadas geralmente da teoria dos 5 C's do crédito (caráter, capital, capacidade, colateral e condições). Bem como a técnica quantitativa, que por meio dos dados dos clientes utiliza métodos estatísticos e econométricos a fim de analisar o risco de crédito. Essa a técnica mais utilizada ultimamente, até mesmo conjunta à técnica qualitativa como suporte a tomada de decisão de gerentes ou analistas de crédito.

## 2.3 Modelos de *Credit Scoring*

Na modelagem de risco de crédito, há duas principais vertentes: a primeira modela o risco em carteira de crédito, pessoas jurídicas, empresas, que não abordaremos nesse trabalho; e a segunda modelagem, a qual contempla o nosso estudo, o risco para concessão de créditos no varejo, em geral para pessoas físicas, que denominamos técnicas baseadas em *Credit Scoring*.

Os modelos de *Credit Scoring* são um processo baseado nas informações do solicitante de crédito, das quais originam variáveis e que por meio de técnicas estatísticas passam a ter pontuações, que combinadas formam *scores*. O *score* é a mensuração da credibilidade solicitante de crédito, um ponto de corte, no qual procura prever quais serão os possíveis "bons" e "maus" pagadores (LEWIS, 1992, p.1). De acordo com Saunders (2000) esta classificação dos clientes de crédito pode ser tanto quanto bons e maus, adimplentes e inadimplentes, desejáveis, ou não, dependendo da modelagem do problema.

A pontuação do *Credit Scoring* pode ser interpretada como a probabilidade de risco de crédito, risco de perda. Além disso, a equação da modelagem deste problema

gera indicadores quantitativos das chances que esse cliente não cumpra com o acordo, se torne inadimplente. Cada instituição financeira tem como base suas próprias premissas; variáveis estabelecidas para decidir sobre o crédito e o risco que estão dispostos a correr. A mensuração de cada uma dessas variáveis carrega pesos e delimita a política de crédito de cada instituição.

O conjunto de critérios, variáveis e de procedimentos definidos que devem ser aplicados para analisar e dimensionar o risco dos devedores, criam a política de crédito da instituição, que com o auxilio do modelo de *credit scoring* oferece suporte ao gestor ou analista de crédito no processo. Dentre as técnicas estatísticas mais utilizadas na modelagem de *credit scoring* destacam-se: Regressão Linear, Análise Discriminante, Redes Bayesianas, Redes Neurais, Regressão Logística e Análise de Sobrevivência (HARRISON; ANSELL, 2002; ANDREEVA, 2003).

# 3 Fundamentação Matemática

## 3.1 Modelos Lineares

A teoria estatística denominada Modelo Lineares explora relações aditivas entre variáveis preditivas e uma variável resposta. O modelo linear, juntamente com os modelos de análise de variância, formam um grande núcleo clássico de modelos lineares. Ademais Wheelan (2016) ressalta que análises de Regressão, em particular linear, são as mais populares e importantes ferramentas estatísticas para encontrar padrões significativos em grandes conjuntos de dados.

O termo Regressão tem origem no trabalho de Bruni *apud* Galton (1885). No século $XIX$ Galton investigou a relação entre alturas de pais e filhos, descobriu sem surpresas que pais altos tendem a ter filhos altos, do mesmo modo que pais baixos tendem a ter filhos baixos

Um importante propósito da Regressão é explorar a dependência de uma variável em relação as outras. Na Regressão Linear Simples a média de uma variável aleatória simples $y$ é modelada como função de outra variável observável $x$ pela relação:

$$E(y) = a + bx$$

Assim quando a variável resposta $y$ está associada a uma única variável preditiva numérica $x$ por meio de uma equação de uma reta $f(x, \theta) = a + bx$ fala-se em Regressão Linear Simples, sendo que tal modelo pode ser facilmente extendido para incorporar duas ou mais variáveis preditivas, o qual é chamado de Regressão Linear Múltipla, já o modelo de Regressão aplicado em problemas de classificação quando a variável de interesse (resposta) é binária, que com base em um conjunto de observações modela uma predição desta variável a partir da relação com as variáveis explicativas.

Nesse trabalho usaremos o modelo de Regressão Logística.

## 3.2 Regressão Logística

O modelo de Regressão Logística é semelhante ao modelo de Regressão Linear, sendo este estabelece uma relação entre as variáveis explicativas e a probabilidade de ocorrer ou não o fenômeno estudado, o que permite criar uma variável binária para estimar a probabilidade de classificarmos (1) sucesso (0) fracasso. A variável de interesse é expressa da sequinte forma:

Seja $Y_i$ uma variável binária que assume dois valores

$$Y_i = \begin{cases} 1 & = \text{sucesso} \\ 0 & = \text{fracasso} \end{cases}$$

Dado uma amostra ($x_k$,$m_k$,$y_k$), $k = 1, 2, ..., n$ com $n$ observações independentes, onde:

- $x_k$ valor da variável explicativa
- $m_k$ número de ensaios
- $y_i$ quantidade de clientes bons em $m_k$ ensaios
- $n$ tamanho da amostra

Assim, a variável resposta tem Distribuição de Probabilidade Binomial $Y_i \sim \mathrm{B}(\mathrm{m}_i, \pi_i)$, tal que:

$$P(Y_i = y_i) = \begin{pmatrix} m_i \\ y_i \end{pmatrix} \pi_i^{y_i} (1 - \pi_i)^{m_i - y_i} \tag{3.1}$$

Para adequarmos a resposta média ao modelo linear usamos a função de ativação

$$\pi_i = \frac{e^{g(x)}}{1 + e^{g(x)}} \tag{3.2}$$

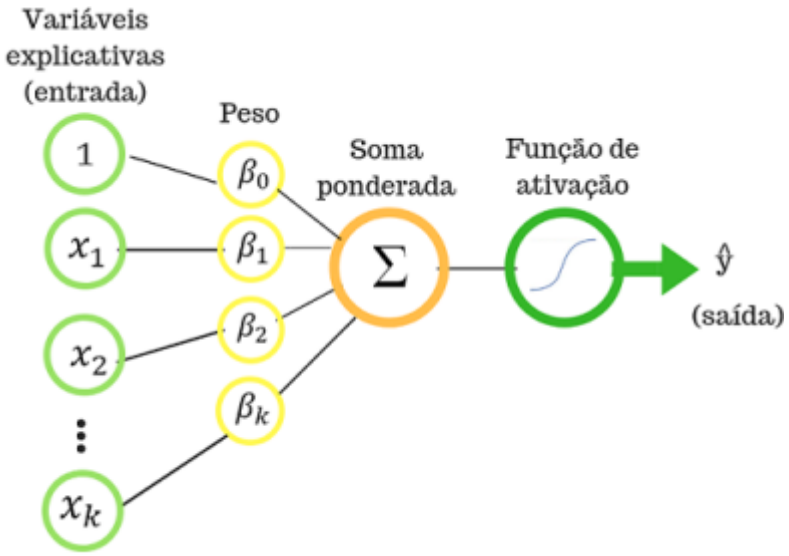
onde $g(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$

$x$ é um vetor, no qual o primeiro elemento é constante 1 e as variáveis independentes do modelo $x = (1, x_1, x_2, ..., x_k)$

$\beta$ é o vetor de parâmetros associados a cada variável independente $\beta = (\beta_0, \beta_1, \beta_2, ..., \beta_k)$

Os coeficientes $\beta$ são estimados pelo método da verossimilhança com base no conjunto de dados, no qual por meio de interações encontra uma combinação de coeficientes que maximiza a probabilidade da amostra ter sido observada. Ao fixarmos uma combinação de $\beta$ e variarmos o valor de $x$, percebe-se que o formato da curva logística possui um comportamento probabilístico em formato da letra 'S', sendo esta uma característica da Regressão Logística (HOSMER E LEMESHOW, 2000).

Figura 2 – Diagrama esquemático do modelo matemático da Regressão Logística




Fonte: Elaborado pelas autoras

## 3.3 Função de Verossimilhança

Seja a função

$$P(Y_i = y_1, y_2, ..., y_n | \beta_0, \beta_1, ..., \beta_k) = \prod_{i=1}^{n} \begin{pmatrix} m_i \\ y_i \end{pmatrix} \pi_i^{y_i} (1 - \pi_i)^{m_i - y_i} =$$

$$= \prod_{i=1}^{n} \begin{pmatrix} m_i \\ y_i \end{pmatrix} \pi_i^{y_i} (1 - \pi_i)^{m_i} (1 - \pi_i)^{-y_i} = \prod_{i=1}^{n} \begin{pmatrix} m_i \\ y_i \end{pmatrix} \pi_i^{y_i} \left( \frac{\pi_i}{1 - \pi_i} \right)^{y_i} (1 - \pi_i)^{m_i}$$

Sendo assim obtemos

$$P(Y_i = y_1, y_2, ..., y_n | \beta_0, \beta_1, ..., \beta_k) = \prod_{i=1}^{n} \begin{pmatrix} m_i \\ y_i \end{pmatrix} \pi_i^{y_i} \left( \frac{\pi_i}{1 - \pi_i} \right)^{y_i} (1 - \pi_i)^{m_i}$$

Aplicando o logaritmo neperiano em ambos os lados da equação, temos:

$$L((\beta_0, \beta_1, ..., \beta_k) | (x_i, m_i, y_i)) = \sum_{i=1}^{n} ln \left( \frac{\pi_i}{1 - \pi_i} \right)^{y_i} + ln \left( 1 - \pi_i \right)^{m_i} =$$

$$L((\beta_0, \beta_1, ..., \beta_k) | (x_i, m_i, y_i)) = \sum_{i=1}^{n} y_i ln \left( \frac{\pi_i}{1 - \pi_i} \right) + m_i ln \left( 1 - \pi_i \right)$$

Substituindo

$$\pi_i = \frac{e^{g(x)}}{1 + e^{g(x)}}$$

$$L((\beta_0,\beta_1,...,\beta_k)|(x_i,m_i,y_i)) = \sum_{i=1}^{n} y_i ln\left(\frac{\frac{e^{g(x)}}{1+e^{g(x)}}}{1-\frac{e^{g(x)}}{1+e^{g(x)}}}\right) + m_i ln\left(1-\frac{e^{g(x)}}{1+e^{g(x)}}\right) =$$

$$= \sum_{i=1}^{n} y_i ln\left(\frac{e^{g(x)}}{1+e^{g(x)}}(1+e^{g(x)})\right) + m_i ln\left(\frac{1+e^{g(x)}-e^{g(x)}}{1+e^{g(x)}}\right) = \sum_{i=1}^{n} y_i ln\left(e^{g(x)}\right) + m_i ln\left(\frac{1}{1+e^{g(x)}}\right) =$$

$$= \sum_{i=1}^{n} y_i ln\left(e^{g(x)}\right) + m_i\left(ln(1) - ln(1+e^{g(x)})\right) = \sum_{i=1}^{n} y_i ln\left(e^{g(x)}\right) - m_i\left(ln(1+e^{g(x)})\right)$$

Substituindo

$$g(x) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k$$

$$L((\beta_0,\beta_1,...,\beta_k)|(x_i,m_i,y_i)) = \sum_{i=1}^{n} y_i ln\left(e^{\beta_0+\beta_1 x_1+...+\beta_k x_k}\right) - m_i ln(1+e^{\beta_0+\beta_1 x_1+...+\beta_k x_k}) =$$

$$= \sum_{i=1}^{n} y_i(\beta_0+\beta_1 x_1+...+\beta_k x_k)ln(e) - m_i ln(1+e^{\beta_0+\beta_1 x_1+...+\beta_k x_k}) =$$

$$= \sum_{i=1}^{n} y_i(\beta_0+\beta_1 x_1+...+\beta_k x_k) - m_i ln(1+e^{\beta_0+\beta_1 x_1+...+\beta_k x_k})$$

Sendo assim, temos:

$$L((\beta_0,\beta_1,...,\beta_k)|(x_i,m_i,y_i)) = \sum_{i=1}^{n} y_i(\beta_0+\beta_1 x_1+...+\beta_k x_k) - m_i ln(1+e^{\beta_0+\beta_1 x_1+...+\beta_k x_k})$$

## 3.4 Estimação dos parâmetros

A estimação dos parâmetros do modo feito pelo método da máxima verossimilhança é dada de modo que os estimadores $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k)$ maximizem o logaritmo da função de verossimilhança. Assim para maximizar a função de verossimilhança basta derivar em relação aos parâmentros do modelo.

$$L(\beta_0,\beta_1,...,\beta_k) = \sum_{i=1}^{n} y_i(\beta_0+\beta_1 x_1+...+\beta_k x_k) - m_i ln(1+e^{\beta_0+\beta_1 x_1+...+\beta_k x_k})$$

Derivando em relação aos parâmetros

$$\frac{\partial L(\beta_0, \beta_1, ..., \beta_k)}{\partial \beta_0} = \sum_{i=1}^{n} y_i - m_i ln\left(\frac{e^{\beta_0+\beta_1 x_1+...+\beta_k x_k}}{1+e^{\beta_0+\beta_1 x_1+...+\beta_k x_k}}\right)$$

$$\frac{\partial L(\beta_0, \beta_1, ..., \beta_k)}{\partial \beta_1} = \sum_{i=1}^{n} y_i x_1 - m_i x_1 ln\left(\frac{e^{\beta_0+\beta_1 x_1+...+\beta_k x_k}}{1+e^{\beta_0+\beta_1 x_1+...+\beta_k x_k}}\right)$$

$$\vdots$$

$$\frac{\partial L(\beta_0, \beta_1, ..., \beta_k)}{\partial \beta_k} = \sum_{i=1}^{n} y_i xk - m_i x_k ln\left(\frac{e^{\beta_0+\beta_1 x_1+...+\beta_k x_k}}{1+e^{\beta_0+\beta_1 x_1+...+\beta_k x_k}}\right)$$

Ao igualar a zero têm-se:

$$\sum_{i=1}^{n} y_i - m_i ln\left(\frac{e^{\hat{\beta}_0+\hat{\beta}_1 x_1+...+\hat{\beta}_k x_k}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 x_1+...+\hat{\beta}_k x_k}}\right) = 0$$

E assim sucessivamente para cada uma das derivadas anteriores. Onde $(\hat{\beta}_0 + \hat{\beta}_1 x_1 + ... + \hat{\beta}_k x_k)$ são os estimadores dos parâmetros $(\beta_0 + \beta_1 x_1 + ... + \beta_k x_k)$. O conjunto resultante de equações exige o uso de métodos numéricos iterativos para a sua solução.

## 3.5 Função Logística

Ao se fazer uma análise de Regressão Logística o problema que se tem em mente é o de classificação, ou seja, o valor que é retornado sempre será entre 0 e 1.

Diferente da Regressão Linear, a Regressão Logística não retorna uma reta que melhor se ajusta aos dados, mas sim uma curva em formato de 'S' que melhor se ajusta ao modelo.

Assim a função de ligação é a função logística ou sigmóide. Esta função é defenida por:

$$P(Y = 1) = \frac{e^{g(x)}}{1 + e^{g(x)}}$$

com $g(x) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k)$

apenas reajustando os termos, tem-se:

$$P(Y = 1) = \frac{\frac{e^{g(x)}}{e^{g(x)}}}{\frac{1}{e^{g(x)}} + \frac{e^{g(x)}}{e^{g(x)}}} = \frac{1}{\frac{1}{e^{g}(x)} + 1} = \frac{1}{1 + e^{-g(x)}}$$

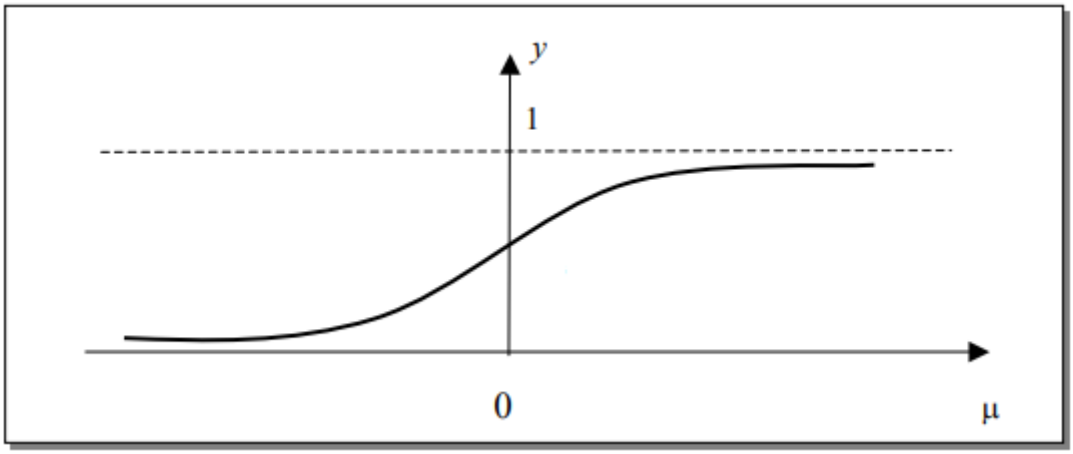
Assim

$$P(Y = 1) = \frac{1}{1 + e^{-g(x)}}$$

Observa-se pela Figura 3 o comportamento da função sigmóide, analisando sua epistemologia 'sigm ($\varsigma$) sigma é uma letra grega em que uma de suas variações parece e representa o 'S' e 'oid' é formato, por isso seu nome, pois essa função tem a forma de um 'S'.

A função sigmóide atribui a regressão logística um alto grau de generalidade.

Figura 3 – Função logística – Sigmóide




Fonte: Adaptado de Guimarães e Neto (2002)

a) Quando $g(x) \to +\infty$, então $P(Y = 1) \to 1$

b) Quando $g(x) \to -\infty$, então $P(Y = 1) \to 0$

Evidentemente indepentende do valor inserido na função sigmóide ela sempre retorna valores entre 0 e 1, nunca será zero e nem 1. Desta forma, se pode estimar a probabilidade direta da ocorrência de um evento $P(Y = 1)$, pode-se estimar a não ocorrência deste evento $P(Y = 0)$, sendo seu complementar:

$$P(Y = 0) = 1 - P(Y = 1)$$

Esta característica de 0 e 1 da função sigmóide é como se ela desligasse e ligasse sendo assim uma função de ativação. A Regressão Logística retorna a classe que o objeto pertence, mas também a probabilidade de pertencimento desde objeto.

## 3.6 Função de erro (entropia cruzada)

A função erro em Regressão Logística, sempre será uma comparação entre o valor original ($y$) e o valor previsto ($\hat{y}$). Naturalmente o objetivo é minimizar a função de entropia cruzada, pois como a sigmóide adicionou a não linearidade ao sistema, a função é descrita como o logaritmo da verossimilhança:

$$L = -y \ln(\hat{y}) - (1 - y) \ln(1 - \hat{y})$$

Para entendermos esta função custo precisamos analisar ela em dois casos: 1º) Quando $y = 1$ então a segunda parcela da função se anula e obtemos $(-1) \ln(\hat{y}) = 0$. A função sempre retorna valores entre zero e um e o logaritmo destes valores são sempre negativos, por isso a função é multiplicada por $(-1)$ assim obtemos um erro sempre positivo. 2º) No entando, ao analisarmos a função quando $y = 0$ o que obtemos é $(-1) \ln(1 - \hat{y}) = 0$, nos resultando também valores entre zero e um.

Sendo assim, custo total do erro é o somatório de todos os erros divididos por $m$ que é a quantidade de ensaios na nossa base de dados, para a regressão logística, a função de entropia cruzada é dada por:

$$\frac{1}{m} \sum_{i=1}^{m} -y \ln(\hat{y}) - (1 - y) \ln(1 - \hat{y})$$

## 3.7 Razão de Chances (odds ratio)

A Razão de Chances (*odds ratio* - O.R) compara a chande de dois eventos, e é definida como a razão entre a chance de ocorrer um evento em um grupo e a chance de ocorrer o mesmo evento em outro grupo. Sejam dois grupos 'A' e 'B' e as probabilidades de um evento em cada um destes respectivamente 'p' e 'q', a razão de chances é obtida por:

$$O.R = \frac{\frac{p}{1-p}}{\frac{q}{1-q}} = \frac{p(1-q)}{q(1-p)}$$

Assim, tem-se:

$$O.R = \frac{p(1-q)}{q(1-p)}$$

As razões de chances são constantes, não importando os valores que as outras variáveis independentes tomem. Outro aspecto interessante é:

$O.R = 1 \rightarrow$ indica que o evento é igualmente provável em ambos os grupos

$O.R < 1 \rightarrow$ indica que a probabilidade de ocorrer o evento é menor no primeiro grupo 'A' do que no que no segundo grupo 'B'

$O.R > 1 \rightarrow$ indica que o evento tem maior probabilidade de ocorrer no primeiro grupo 'A'

## 3.8 Teste Wald

O teste Wald é um teste estatístico paramétrico que testa se cada coeficiente é significativamente diferente de zero. Desta forma, este teste verifica se cada uma das variáveis independentes apresenta uma relação estatísticamente significativa com a variável dependente. Hipótese do teste:

$$H_0 : \beta_k = 0$$
$$H_1 : \beta_k \neq 0$$

com $k = 0, 1, ..., k$

$$W = \frac{\hat{\beta}}{\sqrt{var(\hat{\beta})}}$$

## 3.9 Curva ROC (Receiver Operating Characteristic)

Na área de risco de crédito uma das técnicas mais utilizadas para avalidar o desempenho do modelo é a curva ROC, a qual obtemos gerando um gráfico da especificidade e sensibilidade (taxas de acerto) das previsões do modelo e considerando diferentes pontos de corte no modelo. Segundo Hosmer e Lemeshow (2000) a regra geral para avaliação do resultado da área sob a curva ROC de models de *credit scoring* é dada por:

$$área < 0,7 \rightarrow \text{baixa discriminação}$$
$$0,7 \leq área < 0,8 \rightarrow \text{discriminação aceitável}$$
$$0,8 \leq área < 0,9 \rightarrow \text{discriminação excelente}$$
$$área > 0,9 \rightarrow \text{discriminação excepcional}$$

## 3.10 Método de seleção das variáveis

A seleção das variáveis do modelo é baseada em algum algoritmo que verificam a importância de dada variável e a sua inclusão ou não no modelo. Assim tem-se de forma bastante difundida estes três métodos aqui apresentados

- **Método *enter*** todas as variáveis pré-selecionadas são forçadas a ficar no modelo, não tem exclusão de variável insignificante.

- **Método *forward*** cada variável é adicionada individualmente, sendo a primeira a que adiciona maior poder de explicação ao modelo e assim sussetivamente até que nenhuma das variáveis restantes aumente o poder de explicação do modelo.

- **Método *backward*** contrário do *forward* ele começa com todas as variáveis e retira individualmente a variável que adiciona o menor poder de explicação ao modelo até que restem somente as variáveis que expliquem siginificamente uma parcela da nossa variável dependente.

- **Método *stepwise*** incorpora os modelos *forward* e *backward*, inicia com o *forward* porém a cada variável adicionada as variáveis anteriores são revisadas e verifica-se se seu poder de explicação do modelo permanece significante.

Aqui será apresentado somente o algoritmo *stepwise*

- Inicia com o *forward*

1. Ordenar as variáveis preditoras em ordem crescente
2. Ajustar o modelo com a primeira variável da lista
3. Testar sua significância
4. - Se é significativa:
   a) Salva a variável no modelo
   b) Retira a variável da lista
   c) Volta para o passo 2.
   - Se não, para.
5. Possível lista de variáveis

- Passa para o *backward*

1. Calcula a estatística F parcial para todas as variáveis selecionadas no passo 5. do algoritmo *forward*
2. Escolhe a variável com menor valor
3. Testa sua significância:
   - Se a variável é significatente fica no modelo
   - Se não, sai do modelo e o procedimento para

Repetir *forward* e *backward* até chegar a um modelo que já foi escolhido antes, o modelo escolhido é o do passo anterior.

### 3.10.1 Critério de informação de Akaike (AIC)

O AIC é defenido por:

$$AIC = -2\ln(L_p) + 2[(p+1)+1]$$

onde $L_p$ é a função de máxima verossimilhança do modelo e $p$ é o número de variáveis explicativas no modelo. Como busca-se sempre o menor valor do AIC, o critério de informação de Akaike penaliza os modelos com muitas variáveis, pois quanto mais variáveis maior será o valor do AIC.

# 4 Análise *Credit Scoring* utilizando Regressão Logística

Uma das metodologias mais utilizadas na construção de modelos *credit scoring* é a Regressão Logística, como técnica estatística para a classificação de grupos (THOMAS, 2000). Diante disso, a construção do modelo se baseia na variável de interesse que é expressa da seguinte forma:

$$Y_i = \begin{cases} 1 & \text{se o cliente for adimplente} \\ 0 & \text{se o cliente for inadimplente} \end{cases}$$

## 4.1 *Software R*

Para este trabalho foi utilizado o *Software R* para criação dos modelos de *credit scoring* e análise de dados. O *R* é uma linguagem orientada a objetos que associada a um ambiente integrado possibilita a manipulação e análise dos dados, gerar gráficos e realizar cálculos. O *R* não é programa estatístico, porém se tornou uma importante ferramenta quando falamos em análise e manipulação de dados, pela sua capacidade de permitir rotinas com os mesmos, como modelagem linear e não linear, análise de séries temporais, de sobrevivência, testes paramétricos e não paramétricos, estatística espaciais e simulações. Todas estas funcionalidades com um domínio livre, público e de código aberto motivando assim muitas contribuições de pesquisadores de diversas áreas.

## 4.2 Construção do modelo

Os conjuntos de dados são seccionados em dois: treinamento e validação. Pois o conjunto treinamento é utilizado para construir o modelo e o conjunto de validação é reservado para avaliar a performance do modelo, testar o seu ajuste. As etapas desenvolvidas para a construção do modelo de Regressão Logística para a análise de *credit scoring* está apresentada no algoritmo abaixo.

**Algoritmo Regressão Logística**

1. Transformar
    - variáveis categóticas → fatores
2. Seccionar dados

- TREINAMENTO – 60%
- VALIDAÇÃO – 40%

3. Inferir a variável $y$ no TREINAMENTO
4. Selecionar as variáveis – *Stepwise*
5. Reestruturar o modelo dado as variáveis selecionadas pelo *stepwise*
6. Cálculo de O.R, Teste Wald, Parâmetros
7. Aplicar o modelo selecionado em VALIDAÇÃO
8. Avaliar a performance do modelo

## 4.3 Aplicação do modelo

Inicialmente aplicamos a Regressão Logística na base de dados *German Credit Data*, disponibilizada pela Universidade da California-Irvin UCI em seu repositório *Machine Learning Repository's*. Optou-se por esta base de dados por já ter sido explorada em outros estudos como Karcher e Cipparrone (2009), West (2000), Hsieh (2005) e entre outros o qual nos propicía uma maior confiabilidade nos resultados.

### 4.3.1 Descrição dos dados '

Este conjunto de dados contém informações financeiras e pessoais em relação a 1.000 solicitantes de crédito, destes 700 foram categorizados como bons candidatos e 300 como maus candidatos. As variáveis contidas na base conforme as Tabelas 1 e 2, são qualitativas e numéricas, ao todo somam vinte variáveis e mais uma de saída a qual nos informa se o solicitante é um "bom" ou "mau" futuro pagador. Assim, possibilitando aplicar o experimento.

Tabela 1 – Descrição das variáveis de características bancárias

| **Variável** | **Descrição da Variável** | **Tipo de Variável** | **Nº de Categorias** | **Categorias** |
|---|---|---|---|---|
| Risco | Variável Resposta | Categórica | 2 | Adimplente e Inadimplente |
| ContaBancaria | Status da conta corrente existente | Categórica | 4 | 1:$x < 0$,2: $0 \leq x < 200$ , 3: $x \geq 200$ e 4:Sem conta corrente (não neste banco) |
| TempEmp | Duração do empréstimo em meses | Numérica | - | - |
| Historico | Histórico de Crédito | Categórica | 5 | 1:Nenhum crédito tomado, 2:Todos os créditos deste banco foram devidamentes pagos, 3:Créditos existentes pagos até agora, 4:Atraso no pagamento no passado e 5: Conta crítica / outros créditos existentes (não neste banco) |
| Proposito | Propósito/finalidade | Categórica | 10 | 0:Compra carro novo, 1:Compra carro usado, 2:Móveis, 3:Rádio / televisão, 4:Eletrodomésticos, 5:Educação, 6:Período de férias, 7:Reciclagem,8:Negócios e 9:Outros |
| Montante | Valor do empréstimo | Numérica | - | - |
| Poupanca | Poupança/Títulos | Categórica | 5 | 1:$x < 100$, 2:$100 \leq x < 500$ , 3: $500 \leq x < 1000$, 4:$x \geq 1000$ e 5:Desconhecido/sem conta poupança |
| Fiador | Outros devedores/fiador | Categórica | 3 | 1:Nenhum, 2:Co-requerente, 3:Fiador |
| CreditosBanco | Número de créditos existentes neste banco | Numérica | - | - |
| Taxa | Taxa de juros em % do valor do empréstimo | Numérica | - | - |
| Planos | Outros planos de parcelamento | Categórica | 3 | 1:Bancos, 2:Lojas e 3:Nenhum |

Tabela 2 – Descrição das variáveis de características pessoais

| **Variável** | **Descrição da Variável** | **Tipo de Variável** | **Nº de Categorias** | **Categorias** |
|---|---|---|---|---|
| Ocupacao | Emprego | Categórica | 4 | 1:Desempregado/não qualificado, 2:Empregado sem qualificação, 3:Empregado qualificado/funcionário público e 4:Gerência/autônomo/funcionário altamente qualificado |
| TempoOcu | Emprego atual desde | Categórica | 5 | 1:Desempregado,2: $x < 1$ ano, 3: $1 \leq x < 4$ anos ,4: $4 \leq x < 7$ anos e 5: $x \geq 7$ anos |
| Casa | Tipo de moradia | Categórica | 3 | 1:Aluguel, 2:Própria e 3:Moradia Cedida |
| TempoMoradia | Tempo na moradia | Categórica | 4 | 1:$x > 1$ ano, 2: $1 \leq x < 2$anos, 3: $2 \leq x < 4$anos, 4:$x \geq 4$ anos |
| EstadoCivil | Status pessoal e sexo | Categórica | 5 | 1:Homem:divorciado/separado, 2:Mulher: divorciada/separada/casada, 3:Homem: solteiro, 4:Homem: casado / viúvo 5: e Mulher: solteira |
| Bens | Bens/Propriedade | Categórica | 4 | 1:Imóvel, 2:Seguro de vida,3: Carro ou outros, 4:Nenhum |
| Idade | Idade em anos | Numérica | - | - |
| NumDep | Número de Dependentes | Numérica | - | - |
| Fone | Telefone próprio | Categórica | 2 | 0:Sim e 1:Não |
| Estrangeiro | Trabalhador estrangeiro | Categórica | 2 | -0:Sim e 1:Não |

### 4.3.2 Análise Exploratória dos dados

Este conjunto de dados contém informações financeira com relação a 1.000 solicitantes de crédito, destes 700 foram categorizados como bons candidatos e 300 como maus candidatos. O perfil que pode ser traçado com base nas Tabelas 3 e 4 do tomador de crédito é que $39,4\%$ não possui conta bancária na instituição credora, sendo que 53% possuem créditos pagos. Quanto as garantias como poupança, bens e fiadores, $60,3\%$ possuem menos de 100 unidades monetárias na poupança, $90,7\%$ não possuem fiadores somente $15,4\%$ não possuem nenhum bem. Ao analisarmos o tipo de moradia, $71,4\%$ possuem moradia própria. A idade do solicitante de crédito varia de 19 à 75 anos, com média de 35 anos, somente 25% desses tem mais de 42 anos.

Tabela 3 – Reseumo dos dados *'German Credit data'*

| Risco | ContaBancaria | Poupanca | Historico | Proposito | TempoOcu |
|---|---|---|---|---|---|
| Inadimplente:300 | 1:274 | 1:603 | 0: 40 | 3 :280 | 1:62 |
| Adimplente :700 | 2:269 | 2:172 | 1: 49 | 0 :234 | 2:172 |
| | 3: 63 | 3:63 | 2:530 | 2 :181 | 3:339 |
| | 4:394 | 4:48 | 3: 88 | 1 :103 | 4:174 |
| | | 5:183 | 4:293 | 9 : 97 | 5:253 |
| | | | | 6 : 50 | |
| | | | | (Other): 55 | |

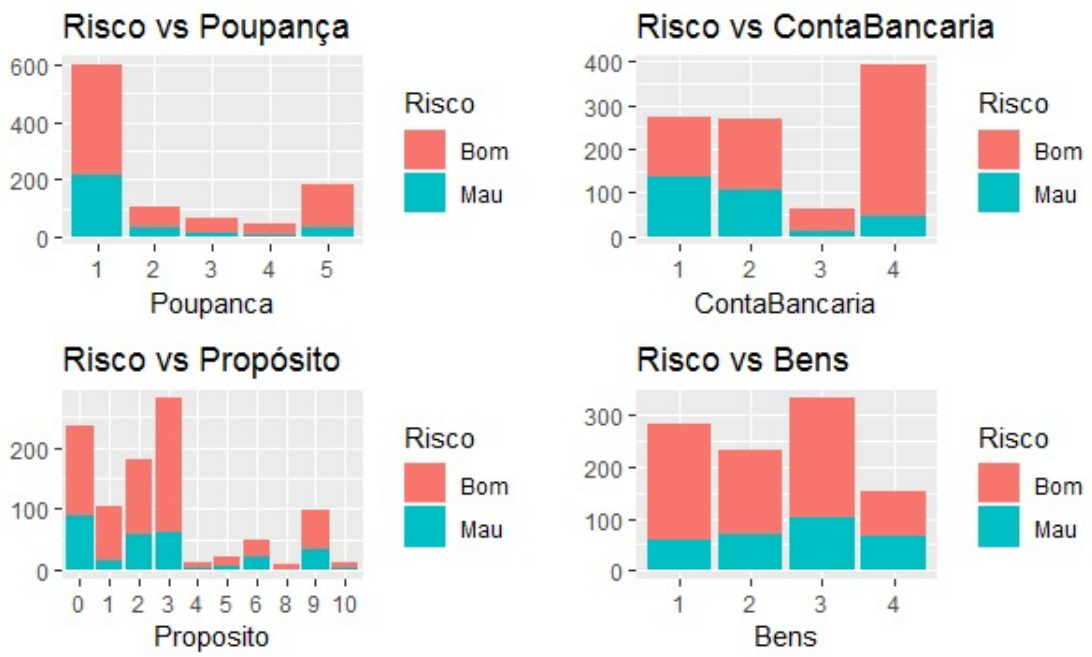
Tabela 4 – Resumo dos dados *'German Credit data'*

| EstadoCivil | Planos | Casa | Fiador | Ocupacao | Bens | Fone | Estrangeiro |
|---|---|---|---|---|---|---|---|
| 1:50 | 1:139 | 1:179 | 1:907 | 1: 22 | 1:282 | 1:596 | 1:963 |
| 2:310 | 2: 47 | 2:714 | 2:41 | 2:200 | 2:232 | 2:404 | 2: 37 |
| 3:548 | 3:814 | 3:107 | 3:52 | 3:630 | 3:332 | | |
| 4:92 | | | | 4:148 | 4:154 | | |

A Figura 4 apresenta o gráfico da relação da variável resposta (Risco) com quatro variáveis explicativas: Poupanca, ContaBancaria,Proposito e Bens. Pode-se observar por meio da variável Poupança que os maus pagagores poupam pouco pois se concentram em sua maioria (217) na categoria 1:($x < 100$) ou não poupam nada representado a categoria 5. Ao analisamos o gráfico referente a Risco vs ContaBancaria, os bons pagadores se concentram na categoria 4, ou seja, não possuem conta corrente neste banco. Já os maus pagadores concentram-se na categoria 1, ou seja estão com a conta corrente negativa ou possuem menos de 200 unidades monetárias.

Na relação da variável Risco vs Proposito Figura 4 pode-se perceber que os bons pagadores e maus pagadores possuem os mesmos propósitos, sendo eles: 0:Carro novo, 2:Móveis e 3: Rádio/Televisão, sendo este último o mais frequente nos propósitos dos solicitantes de crédito.Da mesma forma, em relação aos Risco vs Bens, os bons pagadores possuem algum tipo de bem, sejam eles 1:Imóvel, 2:Seguro de vida ou 3:Carro ou outros.

Figura 4 – Gráfico de Risco vs Outras variáveis




Fonte: Elaborado pelas autoras

Em contraste com bons pagores se encontra os maus pagadores que tendem a possuir 3:Carros ou não possuem nenhum tipo de bem.

### 4.3.3 Regressão Logística

Para a estimação do modelo de Regressão Logística, utilizou-se a amostra TREINAMENTO de 600 casos divididos 70% na categoria de bons e 30% na categoria de maus clientes. Das 20 variáveis independentes, de acordo com o método *stepwise* somente 12 variáveis foram selecionadas como variáveis significativas sendo elas: ContaBancaria, TempoEmp, Historico, Proposito, Poupanca, Taxa, EstadoCivil, Fiador, Bens, Idade, Planos e Fone.

A Tabela 5 apresenta as variáveis selecionadas e as estatísticas geradas pelo modelo logístico, considerando nível de significância de 5%. Sendo a função matemática do modelo dada por:

$$P(Y=1) = \frac{e^{g(x)}}{1+e^{g(x)}}$$

onde

$g(x) = -1,773 + 0,33ContaBancaria_2 + 1,333ContaBancaria_3 + 1,441ContaBancaria_4 - 0,03TempEmp + 0,127Historico_1 + 1,031Historico_2 + 0,94Historico_4 + 1,329Proposito_1 + 0,855Proposito_2 + 1,076Proposito_3 + 0,314Proposito_4 + 0,582Proposito_5 - 0,337Proposito_6 + 1,263Proposito_7 + 0,945Proposito_9 + 1,377Proposito_10 + 1,066Poupanca_2 + 1,093Poupanca_3 + 1,452Poupanca_4 + 1,138Poupanca_5 - 0,163Taxa + 0,22EstadoCivil_2 + 0,843EstadoCivil_3 + 0,95EstadoCivil_4 - 0,914Fiador_2 + 1,02Fiador_3 - 0,975Bens_2 - 0,86Bens_3 - 0,998Bens_4 + 0,018Idade - 0,302Planos_2 + 0,518Planos_3 + 0,426Fone_2 + 1,285Estrangeiro_2$

As variáveis destacadas em negrito na Tabela 5 são as mais siginificativas no modelo, considerando o nível de significância ($\alpha = 0,05$), sendo elas: Conta Bancaria ($x \geq 200$ e Sem conta corrente (não neste banco)), Tempo do Empréstimo (em meses) , Histórico (Nenhum crédito tomado,Atraso no pagamento no passado), Propósito (compra carro novo, Compra carro usado, Móveis,Negócios), Poupança ($100 \leq x < 500$ ,Desconhecido/sem conta) e Bens (Seguro de vida, carro ou outros e nenhum).

Tabela 5 – Modelo de aprovação de crédito

| Variável | Coef. estimado | Erro-padrão | O.R Wald | Teste | P-valor |
|---|---|---|---|---|---|
| (Intercept) | -1.773 | 0.97621 | 0.17 | -1.817 | 0.069225 |
| ContaBancaria2 | 0.333 | 0.292 | 1.39 | 1.119 | 0.263220 |
| **ContaBancaria3** | **1.333** | **0.523** | **3.79** | **2.548** | **0.010844** |
| **ContaBancaria4** | **1.441** | **0.306** | **4.23** | **4.697** | **0.0000026** |
| **TempoEmp** | **-0.03** | **0.106** | **0.97** | **-2.869** | **0.004113** |
| Historico1 | 0.127 | 0.707 | 1.14 | 0.180 | 0.856820 |
| **Historico2** | **1.031** | **0.515** | **2.80** | **2.000** | **0.45463** |
| Historico3 | 0.94 | 0.583 | 2.56 | 1.613 | 0.10675 |
| **Historico4** | **1.948** | **0.558** | **7.02** | **3.489** | **0.000485** |
| **Proposito1** | **1.329** | **0.469** | **3.78** | **2.836** | **0.004569** |
| **Proposito2** | **0.855** | **0.346** | **2.35** | **2.470** | **0.013494** |
| **Proposito3** | **1.076** | **0.340** | **2.94** | **3.161** | **0.001574** |
| Proposito4 | 0.314 | 0.909 | 1.37 | 0.346 | 0.729328 |
| Proposito5 | 0.582 | 0.749 | 1.79 | 0.777 | 0.436940 |
| Proposito6 | -0.337 | 0.590 | 0.71 | -0.572 | 0.567446 |
| Proposito8 | 1.263 | 1.205 | 3.54 | 1.048 | 0.294604 |
| **Proposito9** | **0.945** | **0.448** | **2.57** | **2.111** | **0.034754** |
| Proposito10 | 1.377 | 1.172 | 3.97 | 1.176 | 0.239723 |
| **Poupanca2** | **1.066** | **0.413** | **2.90** | **2.580** | **0.009867** |
| Poupanca3 | 1.093 | 0.576 | 2.98 | 1.896 | 0.057938 |
| Poupanca4 | 1.452 | 0.824 | 4.27 | 1.761 | 0.078232 |
| **Poupanca5** | **1.138** | **0.352** | **3.12** | **3.228** | **0.001247** |
| Taxa | -0.163 | 0.108 | 0.85 | -1.512 | 0.130590 |
| EstadoCivil2 | 0.22 | 0.502 | 1.25 | 0.443 | 0.657812 |
| EstadoCivil3 | 0.843 | 0.495 | 2.32 | 1.702 | 0.088703 |
| EstadoCivil4 | 0.95 | 0.644 | 2.58 | 1.471 | 0.141184 |
| Fiador2 | -0.914 | 0.553 | 0.40 | -1.652 | 0.098601 |
| Fiador3 | 1.02 | 0.649 | 2.77 | 1.571 | 0.116138 |
| **Bens2** | **-0.975** | **0.351** | **0.38** | **-2.775** | **0.005524** |
| **Bens3** | **-0.86** | **0.322** | **0.42** | **-2.658** | **0.007854** |
| **Bens4** | **-0.998** | **0.427** | **0.37** | **-2.336** | **0.019515** |
| Idade | 0.018 | 0.011 | 1.02 | 1.645 | 0.100023 |
| Planos2 | -0.302 | 0.549 | 0.74 | -0.550 | 0.582048 |
| Planos3 | 0.518 | 0.332 | 1.68 | 1.561 | 0.118455 |
| Fone2 | 0.426 | 0.245 | 1.53 | 1.737 | 0.082426 |
| Estrangeiro2 | 1.285 | 0.860 | 3.62 | 1.494 | 0.135109 |

Fonte: Elaborado pelas autoras

O impacto de cada variável explicativa do modelo pode ser explicado ao analisar o seu coeficiente. Os coeficientes positivos são características que produzem um aumento na probabilidade do cliente não se tornar inadimplente. Estas indicam as características dos clientes que individualmente favorecem a redução do risco de inadimplência, que neste estudo foram:

- Conta Bancária ($0 \leq x < 200$, $x \geq 200$ e Sem conta corrente (não neste banco))
- Histórico (Nenhum crédito tomado, Todos os créditos deste banco foram devidamentes pagos, Créditos existentes pagos até agora, Atraso no pagamento no passado)
- Propóstito (Compra carro novo, Compra carro usado, Móveis,Rádio / televisão,

Eletrodomésticos, Período de férias,Reciclagem, Negócios e Outros)

- Poupança ( $100 \leq x < 500$ ,$500 \leq x < 1000$ ,$x \geq 1000$ e Desconhecido/sem conta)
- Estado Civil (Mulher: divorciada/separada/casada,
  Homem: solteiro,Homem: casado / viúvo)
- Fiador (Fiador)
- Idade
- Outros planos (Nenhum)
- Telefone (Não)
- Estrangeiro (Não)

Por outro lado, temos as variáveis com coeficientes negativos que produzem uma redução na probabilidade do cliente se tornar um bom pagador, ou seja, reduzem a probabilidade do cliente não se tornar inadimplente. Estes indicam as características dos clientes que individualmente que aumentam o risco de inadimplência, sendo estes:

- Tempo do Empréstimo (em meses)
- Propósito (Reformas)
- Taxa
- Fiador (Co-requerente)
- Bens (Seguro de vida, carro ou outros e nenhum)
- Outros planos (Lojas)

Sendo assim, quanto maior o tempo de empréstimo que um cliente solicita maior a probabilidade dele se tornar inadimplente ao longo deste empréstimo, bem como a taxa de juros quanto maior for, a probabilidade de inadimplencia aumenta.

### 4.3.4 Avaliação da performance do modelo

Com o conjunto de dados separados para validação podemos fazer uma análise da performance do modelo, esta análise busca julgar a eficiência do modelo quando utilizado dados inéditos.

O modelo de *credit scoring* desenvolvido por meio de Regressão Logística apresentou o percentual de acerto de classificação geral de 72%, sendo assim, o modelo está bem acurado e apresentou bons resultados de classificação. De acordo com Selau e Ribeiro

(2009) especialistas consideram bons os modelos de *credit scoring* com taxa de acerto acima de 65% .

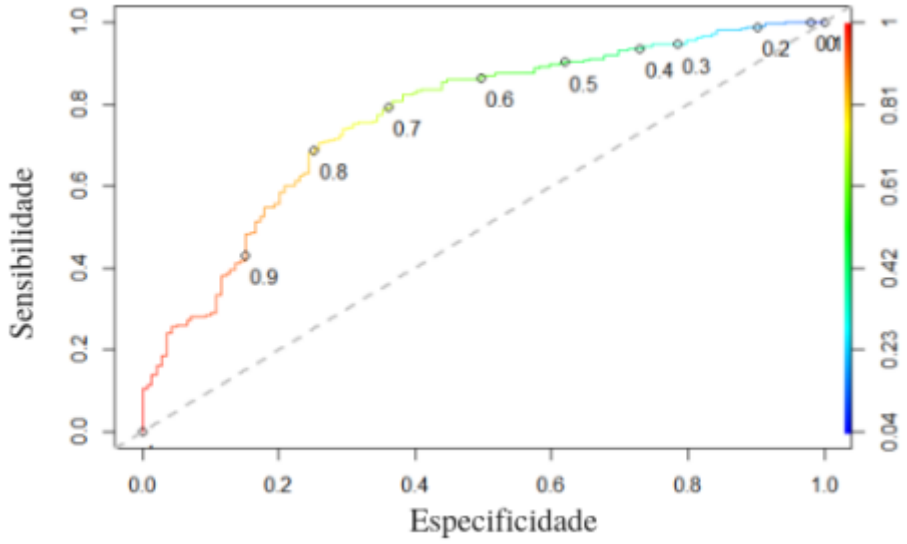
Tabela 6 – Classificação dos casos

| | | Observado | | |
|---|---|---|---|---|
| | | Inadimplente | Adimplente | Pocentagem correta |
| Estimado | Inadimplente | 99 | 72 | 0,66 |
| | Adimplente | 40 | 189 | 0,72 |
| | Porcentagem total | | | 0,72 |

Fonte: Elaborado pelas autoras

A sensibilidade,a capacidade do modelo classificar Tabela 6 o cliente como adimplente quando ele realmente é adimplente foi de 0,72 , sendo assim o modelo classificou dos 261 clientes adimplentes 189 (72%) corretamente. Bem como, a especifidade é a capacidade de classificar como inadimplente quando ele realmente é inadimplente foi de 0,58 evidenciando assim que o modelo classificou dos 139 clientes inadimplentes, 99 (66%) corretamente. A (Figura 5) exibe o gráfico da sensibilidade e especificidade do modelo de Regressão Logística.

Figura 5 – Curva ROC de probabilidades




Fonte: Elaborado pelas autoras

Ao analisarmos a curva ROC Figura 5 do modelo de Regressão Logística verificamos que o indicador da área sob a curva ROC é de 0,766 nível de significância ($p < 0,05$ ). De acordo com Fávero *et al.* (2009) e Hosmer e Lemeshow (2000), o modelo tem poder discriminatório aceitável quando a área da curva estiver entre 0,7 e 0,8, como a área sob a curva do modelo está dentro do intervalo citado pelos autores, podemos concluir que o modelo tem poder discriminatório aceitável. Ademais, como verificamos que o ajuste do modelo é aceitável consequentemente a acurária do modelo também é aceitável no poder de classificação.

# 5 Estudo de Caso - Risco de crédito em uma microoperadora do RS

Após a aplicação da análise dos modelos de *credit scoring* a um conjunto de dados considerado clássico para este tipo de análise, pois já foi objeto de estudo de diversos trabalhos como já citado anteriormente, optou-se por realizar um estudo de caso com dados cedidos por uma microoperadora de crédito do estado do Rio Grande do Sul – RS, com estabelecimentos comerciais em quatro diferentes municípios do estado. Por razão de segurança e sigilo comercial, o nome da microoperadora de crédito não será divulgado.

## 5.1 Construção do modelo

As etapas desenvolvidas para a construção do modelo de Regressão Logística para a análise de *credit scoring* deste estudo de caso seguem o algoritmo apresentado na sessão 4.2, seguindo os mesmos percentuais para os conjuntos de treinamento e validação.

## 5.2 Aplicação do modelo

Analisamos duas modalidades de crédito oferecidas pela microoperadora, nas quais foi diagnosticado haver inadimplência, desta forma tem-se:

- **CDC** (Crédito Direto ao Consumidor)
  Modalidade de crédito pessoal para o consumidor não consignável.
- **Privado**
  Modalidade de crédito pessoal para o consumidor consignável.

### 5.2.1 Descrição dos dados

Os dados recebidos pelo microoperador contêm informações pessoais e financeiras em relação a 3.230 solicitantes de crédito (Privado e CDC), desdes apenas 70 $(2,17\%)$ foram classificados como inadimplentes e 3.160 $(97,83\%)$ como adimplentes. No presente conjunto temos oito variáveis explicativas e uma de resposta que estão apresentadas na Tabela 7, sendo elas categóricas e numéricas.

Tabela 7 – Descrição das variváveis dos dados da Microoperada de Crédito

| **Variável** | **Descrição da Variável** | **Tipo de Variável** | **Nº de categorias** | **Categorias** |
|---|---|---|---|---|
| Risco | Variável Resposta | Categórica | 2 | 0:Inadimplente e 1:Adimplente |
| Sexo | Gênero | Categórica | 2 | 0:Feminino 1:Masculino |
| Idade | Idade em anos | Numérica | - | - |
| Tipo | Tipo do empréstimo | Categórica | 2 | 0:Privado e 1:CDC |
| EpocaAno | Época do ano em que foi feito o empréstimo | Numérica | - | 1:(Janeiro, Fevereiro e Março) 2:(Abril, Maio e Junho) 3:(Julho, Agosto e Setembro) e 4:(Novembro, Outbro e Dezembro) |
| ValorParcela | Valor da parcela | Numérica | - | - |
| TotalParcelas | Total de parcelas | Numérica | - | - |
| Pagas | Total de parcelas pagas | Numérica | - | - |
| Montante | Valor total do empréstimo | Numérica | - | - |

Fonte: Elaborado pelas autoras

### 5.2.2 Análise exploratória de dados

A cartela de clientes desta microoperadora nos foi relatado informalmente que se concentra em aposentados e pensionistas o que pode ser uma evidência que as idades dos solicitantes de crédito não apresenta *outliers*, valores discrepantes, evidenciado pela média e a mediana com valores muito próximos.

Nas Tabelas 8 e 9 é apresentado o resumo estatístico dos dados cedidos pela microoperadora de crédito. Podemos observar que o número de clientes inadimplentes $(2,17\%)$ é muito inferior ao número de clientes adimplentes $(97,83\%)$, bem como o número de pessoas de sexo feminino possuem maior frequencia $(62\%)$ na cartela de clientes, igualmente os tipos de crédito ao consumidor, o crédito consignável privado se destaca $(66,3\%)$ nas operações realizadas na microoperadora.

Tabela 8 – Resumo dos dados Microoperadora de crédito

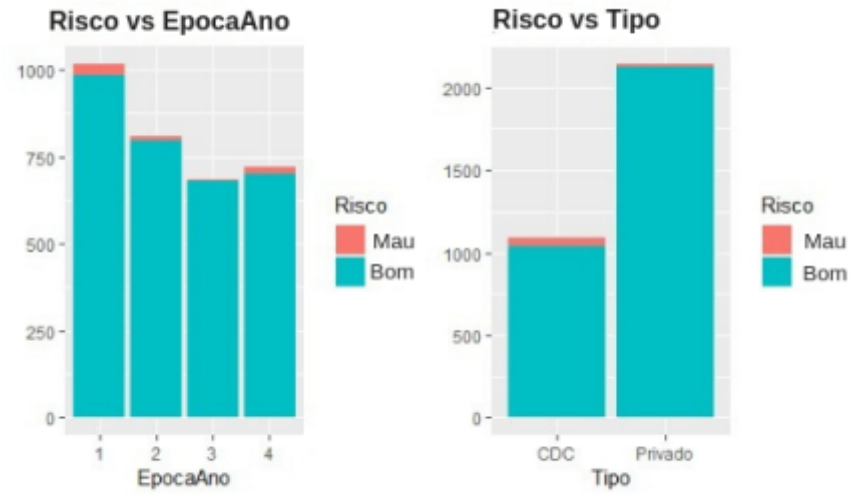
| Sexo | Idade | Tipo | ValorParcela | EpocaAno | Risco |
|---|---|---|---|---|---|
| Feminino :2005 | Min. :35.00 | CDC :1087 | Min. : 9.00 | 1:1017 | Inadimplente: 70 |
| Masculino:1225 | 1st Qu.:62.00 | Privado:2143 | 1st Qu.: 28.79 | 2: 810 | Adimplente :3160 |
| | Median :66.00 | | Median : 55.01 | 3: 684 | |
| | Mean :65.95 | | Mean : 84.32 | 4: 719 | |
| | 3rd Qu.:71.00 | | 3rd Qu.: 105.24 | | |
| | Max. :78.00 | | Max. :1112.97 | | |

Tabela 9 – Resumo dos dados Microoperadora de crédito

| TotalParcelas | Montante | Pagas |
|---|---|---|
| Min. : 6.00 | Min. : 283.1 | Min. : 1.00 |
| 1st Qu.:72.00 | 1st Qu.: 1987.2 | 1st Qu.: 10.00 |
| Median :72.00 | Median : 3721.8 | Median : 17.00 |
| Mean :69.34 | Mean : 5841.2 | Mean : 21.02 |
| 3rd Qu.:72.00 | 3rd Qu.: 7234.6 | 3rd Qu.: 30.00 |
| Max. :72.00 | Max. :80133.8 | Max. :243.00 |

Na Figura 6 pode-se observar que no gráfico Risco vs EpocaAno os bons pagadores se distribuem em todas épocas do ano. Em contrapartida, os maus pagadores se concentram no primeiro trimestre do ano (janeiro, fevereiro, março) e no último trimestre do ano (outubro, novembro, dezembro). Já ao analisar o gráfico Risco vs Tipo pode-se perceber que os bons pagadores em sua maioria se concentram no tipo de crédito consignável (Privado). Por outro lado em contraste com os bons pagadores, os maus pagadores se acumulam no tipo de crédito nao consignável (CDC).

Figura 6 – Gráfico de Risco vs Outras variáveis




Fonte: Elaborado pelas autoras

### 5.2.3 Regressão Logística

A construção do modelo de Regressão Logística para a classificação de risco de crédito foi igual ao aplicado no conjunto de dados *German credit data*, bem como foi utilizado as mesmas etapas do algoritmo exibido no início desde capítulo.

Aplicou-se o algoritmo desenvolvido no cojunto de dados com 1.938 objetos, TREINAMENTO. Incialmente o conjunto possuia 8 variáveis explicativas que após a aplicação do método *stepwise* desconsiderou apenas a variável explicativa Idade, que pode ser justificado por ser uma carcterística comum entre os solicitantes de crédito desta microoperadora. A seleção indicou 7 variáveis como significativas para o modelo, sendo elas: Sexo, Tipo, ValorParcela, EpocaAno, TotalParcelas, Monante e Pagas.

A Tabela 10 apresenta as variáveis selecionadas e as estatísticas geradas pelo modelo logístico, considerando nível de significância de 5%. Sendo a função matemática do modelo dada por:

$$P(Y = 1) = \frac{e^{g(x)}}{1 + e^{g(x)}}$$

onde

$$g(x) = -2,77 + 0.66Sexo_1 + 1.76Tipo + 0,02ValorParcela + 0,28EpocaAno_2 + 1,25EpocaAno_3 - 0,1EpocaAno_4 + 0,06TotalParcelas - 0,0002Montante + 0,08Pagas$$

O impacto de cada variável explicativa do modelo pode ser explicado ao analisar o seu coeficiente. Os coeficientes positivos são características que produzem um aumento na probabilidade do cliente não se tornar inadimplente. Estes indicam as características dos

Tabela 10 – Modelo de aprovação de crédito base de dados Microoperadora

| | Coeficiente estimado | Erro-padrão | OR | Teste Wald | P-valor |
|---|---|---|---|---|---|
| (Intercept) | -2.77 | 1.819 | 0.06 | -1.524 | 0.1276 |
| Sexo1 | 0.66 | 0.405 | 1.94 | 1.630 | 0.1030 |
| **TipoPrivado** | **1.76** | **0.360** | **5.84** | **4.893** | **9.94e-07** |
| ValorParcela | 0.02 | 0.012 | 1.02 | 1.446 | 0.1481 |
| EpocaAno2 | 0.28 | 0.413 | 1.33 | 0.694 | 0.4878 |
| **EpocaAno3** | **1.25** | **0.647** | **3.51** | **1.940** | **0.0523** |
| EpocaAno4 | -0.1 | 0.389 | 0.85 | -0.427 | 0.6694 |
| **TotalParcelas** | **0.06** | **0.025** | **1.06** | **2.236** | **0.0254** |
| Montante | -0.0002 | 0.0001 | 1.00 | -1.339 | 0.1805 |
| **Pagas** | **0.08** | **0.019** | **1.08** | **4.107** | **4.01e-05** |

clientes que individualmente favorecem a redução do risco do risco de inadimplência, que neste estudo de caso foram:

- Sexo (Masculino)
- Tipo (Privado)
- Valor Parcela
- Época do ano ( (Abril, Maio e Julho) e (Junho, Agosto e Setembro) )
- Total de Parcelas
- Parcelas Pagas

Uma observação importe é que o coeficiente mais significativo que contribui para o aumento da probabilidade do cliente não se tornar inadimplente Tipo (Privado) que representa a modalidade de crédito privado consignado é coerente, pois de fato, as parcelas são descontadas diretamente na folha de pagamento e de fato na discussão apresentada na sessão anterior o tipo de crédito privado foi considerado uma característica dos bons pagadores.

Por outro lado, temos as variáveis com coeficientes negativos que produzem uma redução na probabilidade do cliente se tornar um bom pagador. Estes indicam as características dos clientes que individualmente que aumentam o risco de inadimplência, sendo estas:

- Época do ano (Novembro, Outubro e Dezembro)
- Montante

Nessa lógica, os clientes que buscam por empréstimos no final do ano tendem a se tornar inadimplentes, bem como quanto maior for o valor do empréstimo o risco de inadimplência aumenta. O coeficiente estimado para o Montante esta bem próximo de zero, seu Teste Wald não demonstrou diferença significativa, e a sua O.R é igual a um o que indica que o montante com valores elevados é igualmente provável em ambos os grupos (adimplentes e inadimplentes). Por isso nos faz questionar se esta variável é realmente relevante no modelo .

### 5.2.4 Avaliação da performance do modelo

Posterior a construção do modelo, o mesmo foi aplicado em um conjunto inédito de dados VALIDAÇÃO, o qual englobava 1.292 clientes, destes 27 $(2,09\%)$ eram inadimplentes e 1.265 $(97,9\%)$ adimplentes. Visto que a proporção de inadimplentes é muito inferior do que a de adimplentes, pois neste momento estamos trabalhando com dados reais. O modelo atingiu as expectativas com o percentual de acerto geral de 97%, apresentando uma excelente performance na classificação e está bem ajustado.

Tabela 11 – Classificação dos casos

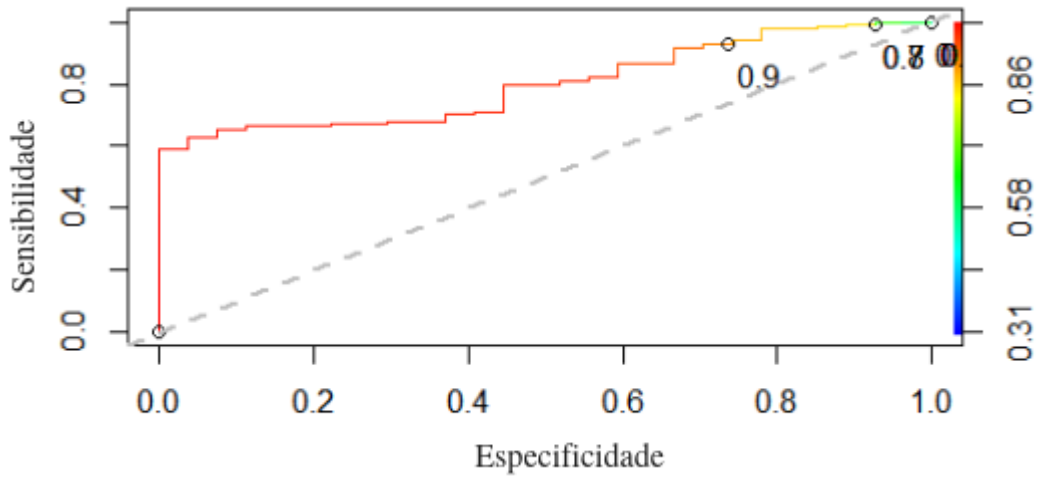
| | | Observado | | |
|---|---|---|---|---|
| | | Inadimplente | Adimplente | Pocentagem correta |
| Estimado | Inadimplente | 2 | 12 | 0,07 |
| | Adimplente | 25 | 1.253 | 0,99 |
| | Porcentagem total | | | 0,97 |

Fonte: Elaborado pelas autoras

Em contra partida, a especificidade do modelo considerando a não proporcionalidade dos dados está coerente, visto que o modelo possuia muito mais dados de adimplentes do que de inadimplentes, apresentando resultados inferiores, o que era esperado, dos 27 inadimplente no conjunto o modelo classificou corretamente apenas 2 (7%). De acordo com Moreira e Selau (2014) há evidencias de que manter a proporcionalidade na amostra influencia na capacidade preditiva do modelo, bem como maior percentual de maus pagadores na amostra permite a melhora na identificação deste perfil.

Visto que no conjunto VALIDAÇÃO possui mais clientes adimplentes o modelo dispõe de superioridade na identificação destes. Portando a sensibilidade do modelo apresentou bons resultados, dos 1.265 clientes adimplentes o modelo classificou corretamente

Figura 7 – Curva ROC de probabilidades Microoperadora




Fonte: Elaborado pelas autoras

1253 (99%). A Figura 7 apresenta o gráfico da curva ROC, a área sob a curva ROC é baseada no cálculo de sensibilidade e especificidade calculadas em relação ao *score*.

O modelo de Regressão Logística verifica que o indicador de área sob a curva ROC Figura 7 é de 0,80 com nível de significância ($p < 0{,}05$). Nesse perspectiva, o modelo apresenta poder de discriminação excelente e o modelo está adequado e bem ajustado.

# 6 Considerações Finais

Este trabalho deu-se com o intuito de atingir as propostas pela ênfase em Economia Matemática proposta no curso de Bacharelado em Matemática Aplicada. Sendo asim, objetivou-se desenvolver um modelo de *credit scoring* com o uso da técnica estatística de Regressão Logística para discriminar as características de um cliente, pessoa física, que produzem um aumento ou diminuição na probabilidade de risco de crédito.

Para alcançar este objetivo algumas etapas foram realizadas, inicialmente a seleção das variáveis pelo método *stepwise* permitiu identificar as variáveis com maior poder discriminante entre o grupo de cliente adimplentes e inadimplentes. Sequêncialmente a reestruturação do modelo com as variáveis indicadas permitiu a atribuição de pesos nas mesmas.

A aplicação em conjuntos de dados clássicos, que já foram testados e explorados anteriormente como a base de dados '*German Credit Data*', bem estruturados no qual você já conhece o comportamento propicia uma estabilidade maior nos resultados. .

Os resultados obtidos no conjunto de dados '*German Credit Data*' que possuia maior proporcionalidade nos dados em relação a clientes adimplentes e inadimplentes se mostrou superior na classificação de clientes inadimplentes, o que nos causa menos riscos, afinal apenas um cliente identificado incorretamente como adimplente visto que ele será inadimplente, pode arruinar com os lucros obtidos em muitas classificações corretas de adimplentes.

O modelo apresentou bom desempenho com taxa de acerto geral de 72%, e classificação correta de inadimplentes de 66%, bem como o indicador de avaliação do modelo da área sob a curva ROC salientou que o modelo tem aceitável poder de classificação.

O estudo de caso dos dados cedidos pela microoperadora de crédito possibilitou uma aplicação ao mundo real, no qual nem sempre temos as proporções sugeridas estatísticamente. Por outro lado, a aplicação em dados reais permite a reflexão da dualidade da prática e teoria, pois ao trabalhar com dados reais confrontamos desafios.

Em suma, a modelagem *credit scoring* para o estudo de caso revelou resultados satisfatórios. Visto que o conjunto possuia poucos clientes inadimplentes para desenvolver um maior poder de identificação deste perfil, pela baixa frequencia de inadimplentes na amostra o modelo classificou corretamente apenas 7% dos clientes inadimplentes,e 99% dos adimplentes. Sendo a sua taxa de acerto geral de 97% o que nos oportuniza um bom modelo.

Ademais o indicador de avaliação do modelo da área sob a curva ROC corrobora

que o modelo apresenta excelente poder de discriminação. A aplicação em dados reais atribuiu a minha formação como bacharel Matemática Aplicada uma maturidade maior para trabalhar com os desafios da aplicação da Economia Matemática em problemas desconhecidos e reais.

Como passos futuros, almejo aplicar outros métodos estatísticos neste conjunto de dados. De posse do comportado destes dados com a técnica de Regressão Logística, a aplicação em técnicas como Redes Neurais e Análise Discriminante, propiciariam outro questionamento, em busca do modelo mais indicado para a classificação de clientes adimplentes e inadimplentes.

# Referências

ANDREEVA, G. *European generic scoring models using logistic regression and survival analysis*. In: YOUNG OR CONFERENCE, 2003, Bath. Anais... Bath: Young OR, 2003.

BANCO CENTRAL DO BRASIL. **Relatório de Economia Bancária e Crédito - 2014**. Disponível em: <http://www.bcb.gov.br/pec/depep/spread/rebc_2014.pdf>. Acesso em: abril 2018. 16

BRUNI. E. S. Uso de regressão logística para precificação de *Credit Default Swaps*. (Monografia de graduação) Universidade federal de São Paulo, São Paulo, 2007. 23

CAOUETTE, J., ALTMANO, E.; NARAYANAN, P. **Gestão do risco de crédito**. Rio de Janeiro: Qualitymark, 2000. 19

DUA, D., KARRA, E. *UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science*. 2017. Disponível em: <http://archive.ics.uci.edu/ml> Acesso em: Abril de 2018.

FÁVERO, L. P. **Análise de dados:** modelagem multivariada para tomada de decisões. Rio de Janeiro: Elsevier, 2009. 43

FISHER, R. A. *The use of multiple measurements in taxonomic problems*. Annals of Eugenics, 7, 179-188. 1936. 17

GITMAN, L. J. **Princípios de Administração Financeira**. São Paulo: Harbra.1997. 19

GUIMARÃES, I. A. NETO, A. C. Reconhecimento de padrões: Metodologias estatísticas em crédito ao consumidor. RAE-eletrônica, Volume 1, Número 2, jul-dez/2002. Disponível em: <http://www.rae.com.br/eletronica/.cfm?FuseAction=ArtigoID=1215Secao=FINANÇAS2Volume=1Numero=2 Ano=2002>. Acesso em: novembro 2018.

HAND, D. J.; HENLEY, W. E. *Statistical Classification Methods in Consumer Credit Scoring: a Review*. **Journal of Royal Statistical Society:** Series A, n. 160, p. 523-541 Londres: Royal Statistical Society. 1997

HARRISON, T.; ANSELL, J*Customer retention in the insurance industry: using survival analysis to predict cross-selling opportunities. Journal of Financial Services Marketing*, London, v. 6, n. 3, p. 229-239, 2002.

HSIEH, N. *Hybrid mining approach in the design of credit scoring models.***Expert Systems with Applications**. 28. 655-665. 10.1016/j.eswa.2004.12.022. 34

HOSMER, D, W., LEMESHOW, S. *Applied Logistic Regression, 2nd ed. New York: John Wiley Sons*, 2000. 30, 43

KARCHER, C.; CIPPARRONE, F. A. M. **Redes Bayesianas aplicadas à análise do risco de crédito**.Universidade de São Paulo, São Paulo, 2009. Disponível em: < http://www.teses.usp.br/teses/disponiveis/3/3142/tde-25052009-162507/ >. Acesso em: abril 2018. 34

SELAU, L. P. R.; RIBEIRO, J. L. D. Uma sistemática para construção e escolha de modelos de previsão de risco de crédito. Revista Gestão e Produção, v. 16, n. 3, p. 398–413. 2009. 41

LEWIS, E. M. **An Introduction to Credit Scoring**. San Rafael: Fair Isaac and Co., Inc. 1992. 17, 20

MOREIRA, P. D.; SELAU, L. P. R. Comparação do desempenho de modelos de Credit Scoring utilizando diferentes composições amostrais de grupos de clientes. (Monografia) 2014. 50

ROSA, P.T.M. *Modelos de credit scoring:* Regressão Logística, CHAID e Real. Dissertação de Mestrado, Departamento de Estatística,Universidade de São Paulo, São Paulo, 2000. 15

SAUNDERS, A. Medindo o risco de crédito: novas abordagens para o *value at risk* e outros paradigmas. Rio de Janeiro: Qualitymark, 2000. 20

SCHRICKEL, W. K. Análise de Crédito: Concessão e Gerência de Empréstimos, São Paulo: Atlas. 1995. 19

SOUZA, A. L. Redes Bayesianas: Uma introdução aplicada a Credit Scoring. In: Simpósio Nacional de Probabilidade e Estatística (SINAPE), São Paulo, 2010.

STEINER, M.T.A.; CARNIERI, C.; KOPITTKE, B.H.; STEINER NETO, P.J. Sistemas especialistas probabilísticos e redes neurais na análise do crédito bancário. **Revista de Administração da Universidade de São Paulo (RAUSP)**, São Paulo, v.34, n.3, p.56-67, jul./set. 1999.

THOMAS, L. C. *A Survey of Credit and Behavioural Scoring: Forecasting Financial Risk of Lending to Consumers*, International Journal of Forecasting, v. 16, n. 2, p. 149-172, Londres: Elsevier. 2000. 16, 33

WEERTHOF, R.V. Programa de Gestão de Risco do SAS. Instituto na Europa. 2011. Disponível em: <http://www.gestaoderisco.no.sapo.pt/GestaoRisco.html>. Acesso em: março 2018. 19

WEST, D. *Neural network credit scoring models*. **Computers  Operations Research**, v. 27, n. 11-12, p. 1131-1152, 2000.

WHEELAN, C. Estatística: O que é, para que serve, como funciona. Zahar, 2016. 34

23

# Anexos

# ANEXO A – Código Regressão Logística modelo *credit scoring*

```
#Apresentação das variáveis e pacotes
install.packages("xtable")
install.packages("readxl")
install.packages("ggplot2")
install.packages("ROCR")
library(xtable)
library(readxl)
library(ggplot2)
library(ROCR)

        data <- read_excel("data.xls")

        View(data)

#Apresenta a estrutura do DataFrame
str(data)

        colnames(data)

#Transforma em fatores as variaveis categoricas e "dummies"

        data$Risco <- as.factor(data$Risco)

        data$Proposito <- as.factor(data$Proposito)

        data$ContaBancaria <- as.factor(data$ContaBancaria)

        data$Historico <- as.factor(data$Historico)

        data$Poupanca <- as.factor(data$Poupanca)

        data$TempoOcu <- as.factor(data$TempoOcu)

        data$EstadoCivil <- as.factor(data$EstadoCivil)

        data$Fiador <- as.factor(data$Fiador)

        data$Bens <- as.factor(data$Bens)

        data$Planos <- as.factor(data$Planos)
```

```
        data$Casa <- as.factor(data$Casa)
        data$Ocupacao <- as.factor(data$Ocupacao)
        data$Fone <- as.factor(data$Fone)
data$Estrangeiro <- as.factor(data$Estrangeiro)
str(data)
# Separar o conjunto de dados em dados p teste e validacao
#indices obtidos apos a aleatorizacao ordena = sort(sample(nrow(data), nrow(data)*.6))
#Dados para o treinamento treinamento<-data[ordena,]
#Dados para a validacao validacao<-data[-ordena,]
#Regressao Logistica modelo.completo <- glm(Risco ~.,family=binomial,data=treinamento)

#Abordagem Stepwise para selecao de variaveis
stepwise <- step(modelo.completo,direction="both")

        stepwise$formula
#Modelo com as variaveis indicadas pelo Stepwise

        stepwise <- glm(stepwise$formula, family=binomial,data=treinamento)

#Resume os resultados do modelo
summary(stepwise)

#Calcula a razÃ£o de chances
razao<-exp(cbind(OR = coef(stepwise), confint(stepwise)))
razao
xtable(razao)
#Faz a previsao para a base de validaco (probabilidade)
        predito<-predict(stepwise,validacao,type="response")
        pred = prediction(predito, validacao$Risco)
        corte<-as.numeric(performance(pred, "auc")@y.values)
#score validacao data set
        validacao$score<-predict(stepwise,type='response',validacao)
pred<-prediction(validacao$score, validacao$Risco)
```

```
perf <- performance(pred,"tpr","fpr")

plot(perf) plot(perf, colorize=TRUE) #adicionar
plot(perf, colorize=TRUE, print.cutoffs.at=seq(0,1,by=0.1), text.adj=c(-0.2,1.7))
abline( a =0, b = 1, lwd = 2, lty = 2, col = "gray")


#Escolhe quem vai ser "1"e quem vai ser "0"
predito<-ifelse(predito>=corte,1,0)


#Compara os resultados tab<-table(predito,validacao$Risco)
tab
xtable(tab)
taxaacerto<-(tab[2,2]+tab[1,1])/sum(tab)
taxaacerto
```