**CURSO DE CIÊNCIAS CONTÁBEIS**

**APLICAÇÃO DE REGRESSÃO LOGÍSTICA PARA ANÁLISE DO PERFIL DE RISCO DE CRÉDITO DE CLIENTES DE BAIXA RENDA DE UMA ADMINISTRADORA DE CARTÕES DE CRÉDITO BRASILEIRA**

JOSÉ ANTÔNIO BELQUIOR GOMES

GUAXUPÉ

DEZEMBRO /2023

SUMÁRIO

[1- INTRODUÇÃO 3](#_Toc145171581)

[2- OBJETIVOS 6](#_Toc145171582)

[2.1 - Objetivo Geral 6](#_Toc145171583)

[2.2 - Objetivos Específicos 6](#_Toc145171584)

[3- JUSTIFICATIVA 6](#_Toc145171585)

[4- REVISÃO TEÓRICA 7](#_Toc145171586)

[4.1 - *Crédit Scoring* 7](#_Toc145171587)

[4.2 - Regressão Logística 7](#_Toc145171588)

[4.3 - 8](#_Toc145171589)

[5- METODOLOGIA 8](#_Toc145171590)

[5.1 - Dados 10](#_Toc145171591)

[6- CONCLUSÃO 11](#_Toc145171592)

[7- REFERÊNCIAS 12](#_Toc145171593)

# 1- INTRODUÇÃO

Para Sicsú (2009), a crise econômica mundial deflagrada em 2008 teve como principal causa de origem o não controle de liberação de crédito e avaliação de potenciais riscos. A falta de cuidado ao realizar a análise do perfil de crédito dos americanos de baixa renda que adquiriram casas próprias acabou por levar a uma crise econômica que abalou todo o mundo.

Pessoas que não tinham o perfil adequado conseguiram com facilidade hipotecas, que com o passar do tempo não puderam mais arcar com as parcelas. Com a inadimplência, os bancos realizaram a tomada dos imóveis, porém os mesmos não valiam mais do que uma fração do crédito originalmente concedido.

Quando essa bolha estourou, segundo Bresser-Pereira (2009), seus efeitos foram devastadores, sendo comparáveis somente à famosa crise de 1929, devido à amplitude e severidade dos impactos socioeconômicos nos Estados Unidos e no resto do mundo.

Justamente em 2008, estava começando uma revolução tecnológica que moldaria os pilares da integração homem máquina para sempre. A chegada da terceira geração do iPhone ditava como os smartphones trariam produtos e serviços a qualquer momento. Aliado às modernas redes móveis, de qualquer lugar com sinal se tornou possível realizar qualquer operação na internet. Os serviços financeiros apenas acompanharam essa tendência, permitindo digitalizar cada vez mais processos que antes dependiam de uma visita a uma agência.

No cenário recente, tendo em vista essa grande digitalização de serviços financeiros, houve a entrada de diversos novos *players* no Sistema Financeiro Brasileiro (SFB), o que aumentou a concorrência no mercado de crédito, acarretando correr mais riscos para a liberação de crédito (FIORDELISI, MARQUES-IBANEZ e MOLYNEUX, 2011).

No Brasil, o Banco Central (Bacen) é o responsável pela regulamentação, oficialização e fiscalização das instituições financeiras. Em relação à análise de risco, o Bacen se utiliza dos critérios e metodologia estabelecidos no acordo de Basileia (Basel, 2011), que é um tratado envolvendo os países mais ricos do mundo, cujo objetivo é lidar com o mercado financeiro e evitar novas crises sistêmicas como a já citada crise de 2008.

Conhecendo-se os efeitos extremamente negativos que uma análise de crédito mal executada – ou a falta dela – podem acarretar, vê-se que a metodologia para se estabelecer o risco de crédito de potenciais clientes é um dos temas mais relevantes para instituições financeiras e o mercado financeiro em geral.

Segundo Tavares (2009), o crédito permite a elevação do padrão de vida dos consumidores e do lucro das empresas. Essa liberação é realizada caso a instituição se sinta segurança de que o capital cedido será pago e voltará com acréscimo de juros. Quanto mais confiáveis forem os processos e métricas empregados nessa decisão, menor as chances de o cliente incidir em perdas, e, consequentemente, a taxa de inadimplência será menor (Marcelino, 2012).

Modelos estatísticos que visam realizar essa análise e calcular o risco de inadimplência de um potencial crédito baseado no perfil do cliente, são comumente denominados *Credit Scoring*, ou Pontos de Crédito, que visam ordenar os clientes quanto à probabilidade do cumprimento das obrigações, quanto mais pontos, menor o risco de inadimplência, portanto, melhor o perfil do cliente.

Modelos de *Credit Scoring* reduzem drasticamente o custo e tempo da análise da concessão do crédito, podendo inclusive serem feitos de forma completamente automatizada. Para que o processo seja assertivo, no entanto, necessita-se que os dados dos clientes estejam os mais corretos e atualizados possível, para poder traçar uma correlação entre esses dados e a probabilidade de o cliente manter os pagamentos em dia.

Segundo Lewis (1992), cada dado do perfil, ou combinação de dados semelhantes, possui um peso, e essa combinação de pesos permite chegar num grau de risco, que, a partir de uma nota de corte, podem indicar se os clientes são possivelmente “bons” ou “maus”, chegando então à decisão: liberar ou negar o crédito requisitado.

Stevenson (1986),entende que “A correlação mede a força, ou grau, de relacionamento entre duas variáveis; a regressão dá uma equação que descreve o relacionamento em termos matemáticos.”

Isso significa que é possível a partir do processamento dos dados de uma base de dados de clientes com diversos perfis já estabelecidos (adimplentes e inadimplentes), se estabelecer a correlação entre seus dados e o seu perfil, permitindo assim encontrar a equação que representa esse comportamento. Com essa equação, é possível predizer o resultado estimado usando os dados de um novo cliente em potencial ou de um cliente já presente que deseja adquirir mais crédito. Quanto maior a precisão dos dados, maior a confiabilidade da predição dada pelo cálculo da equação, portanto, com um banco de dados bem alimentado, a instituição pode incluir como grande auxiliar a regressão para realizar a análise de risco de crédito de seus clientes.

# 2- OBJETIVOS

## 2.1 - Objetivo Geral

O objetivo geral do presente trabalho é realizar o estudo dos dados de uma administradora de cartões de crédito para a implementação e melhora do seu sistema de análise de concessão de crédito. A partir de um conjunto de dados que compõem o perfil desses clientes extraídos da base de dados da companhia, pretende-se realizar o processamento de regressão logística para determinar se o cliente pode ser considerado como potencial bom ou mau pagador.

## 2.2 - Objetivos Específicos

Baseado no histórico da companhia, que sempre atuou com as classes C, D e E, é de interesse da mesma a sua abrangência sobre a região nordeste do país. Na conjuntura atual, a administradora possui uma quantidade de clientes captados na região, porém ainda não possui uma base sólida de conhecimento para poder ter um perfilamento mais preciso e confiável.

Devido a esse cenário, serão tomados como base da análise a base de dados de clientes dessa região.

Deverão ser levantados quais os principais dados dos perfis e uma análise de quais que oferecem o melhor peso para a aplicação da regressão. O objetivo é chegar ao melhor set de variáveis possível, baseado na acurácia que essas informações combinadas terão no momento da análise de crédito.

# 3- JUSTIFICATIVA

Visto a importância econômica que uma boa análise de crédito representa para as instituições financeiras, os estudos de metodologias para a implantação e aprimoramento das análises são primordiais para a saúde das companhias que atuam nesse ramo.

A taxa de inadimplência pode ser o fator decisivo para a viabilidade de produtos e serviços que uma empresa do ramo financeiro oferece, e pode, segundo Marcelino (2012), colocar em risco toda a operação, caso essa taxa seja muito alta.

Chegar a um cenário de equilíbrio pode ser desafiador, principalmente se, como no estudo de caso dessa companhia, o perfil alvo já seja considerado de maior risco. Segundo Lana (2015), quanto menor a renda do indivíduo, maior a sua exclusão financeira, que é justamente a remoção de seu perfil como potencial para a concessão de crédito. Esse processo leva a gerar ainda mais desigualdade.

Paradoxalmente, os perfis das classes mais baixas são considerados de maior risco pela maior possibilidade de não contarem com recursos para arcar com suas obrigações, e isso leva ao crescimento da desigualdade social, uma vez que o não acesso ao crédito limita ainda mais as possibilidades financeiras dessa população.

Portanto, o aprofundamento da análise do perfil de crédito dessas classes pode levar à descoberta de um nicho de potenciais consumidores, que são, segundo Lana (2015), ignorados e pouco explorados pelo sistema financeiro.

Isso deve ser aliado a um robusto processo de análise, uma vez que não apenas é a análise de crédito de maior risco, mas também por se tratar de perfis inéditos para o mercado financeiro, pois são pessoas que, em grande parte, nunca tiveram acesso aos serviços financeiros ou sequer foram bancarizadas.

# 4- REVISÃO TEÓRICA

## 4.1 - *Crédit Scoring*

O *Credit Scoring* é um processo cujo objetivo é estabelecer um modelo quantitativo baseado em uma série de características para distinguir dois subgrupos de solicitantes nesse caso, os candidatos dignos de crédito e os de risco (Crook, et al., 2007).

Geralmente envolve algumas etapas, como o pré-processamento dos dados para deixa-los na mesma escala e remover ruídos. Métodos estatísticos são sempre empregados para poder realizar esses processos, tornando os dados e as predições matematicamente confiáveis.

## 4.2 - REGRESSÃO LOGÍSTICA

A regressão logística é um modelo de classificação que estima a probabilidade de um certo dado pertencer a uma classificação binária, sendo o mesmo considerado positivo ou negativo nessa classificação. A regressão logística é um dos modelos estatísticos mais populares utilizados em algoritmos de classificação (Abdou et al, 2016).

A regressão pode ser realizada tanto com dados quantitativos, como idade, renda, número de dependentes, assim como dados qualitativos, como gênero, naturalidade, classificação socioeconômica.

A regressão logística pode ser realizada estimando-se os parâmetros da seguinte equação:

log (Pgi / (1 – Pgi)) = α + β1K1i + β2K2i + β3K3i + . . .

Onde α, β1, β2, β3, … são os coeficientes do modelo e os Kji representam os valores das propriedades ou variáveis j do indivíduo i sendo analisado, e Pgi representa a probabilidade do indivíduo i ter um bom ou mau perfil de crédito.

Portanto, a probabilidade de um indivíduo possuir um dado perfil de crédito pode ser obtido por:

Pgi = [exp(α + β1K1i + β2K2i + β3K3i + . . .)]/[1 + exp(α + β1K1i + β2K2i + β3K3i + . . .)]

Os parâmetros são obtidos pelo método da máxima verossimilhança. O valor de Pgi então pode ser classificado acima de um ponto de corte, sendo considerado um bom perfil, ou abaixo disso, sendo considerado um mau perfil. Esse ponto de corte é determinado pela instituição, definindo o risco que ela deseja correr no momento de classificar seus perfis.

## 4.3 - Matriz de confusão

A matriz de confusão, segundo Duda e Stork (2001), é uma matriz de duas dimensões que apresenta de forma sintética o resultado da classificação de um dado realizado por um algoritmo de predição, sendo dividido entre as classificações corretas e incorretas. É utilizado para dados binários, sendo positivo ou negativo. Essa representação tem por objetivo apresentar o desempenho da classificação. As informações são divididas entre:

Verdadeiro positivo (true positive — TP): quando a classificação é realizada corretamente para um caso positivo

Falso positivo (false positive — FP): quando a classificação erroneamente classifica o dado como positivo.

Falso verdadeiro (true negative — TN): quando a classificação determina corretamente o dado como negativo.

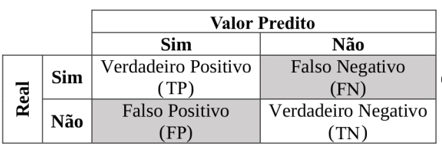
Falso negativo (false negative — FN): quando a classificação erroneamente classifica o dado como negativo.

Figura 1: Representação da Matriz de Confusão. Fonte: Diego Nogare[[1]](#footnote-1)

# 5- METODOLOGIA

A análise de perfil de crédito foi realizada com dados provenientes de uma administradora de cartões de crédito brasileira, que atua em todo o território nacional e possui em sua base mais de 7 milhões de clientes, sendo cerca de 600 mil clientes ativos mensalmente. Além da base de clientes compradores, a empresa, integrante de um grupo econômico, possui sua própria adquirente e todos os processos para o credenciamento de pontos de venda, automação para integração das vendas no cartão e a emissão e envio dos cartões aos clientes.

Por possuir uma atuação holista no ramo de meios de pagamento, a companhia enfrenta inúmeros desafios, como a avaliação do perfil tanto dos clientes quanto dos lojistas que venderão com o cartão, evitar fraudes, e manter a taxa de inadimplência controlada, assim como taxas competitivas junto aos lojistas.

Por ser uma empresa que atua nesse ramo há mais de 20 anos, durante esse tempo foram constantemente aperfeiçoados os processos de análise e liberação do crédito, atuando nas frentes tecnológica e estratégica.

Essa expertise propiciou a criação de um produto de aprovação instantânea. Esse produto permite que clientes com bom perfil tenham crédito liberado de maneira imediata e totalmente automatizada.

Clientes podem comparecer a certos lojistas da rede credenciada e fazer o pedido do cartão. Caso o processo libere automaticamente, ele já pode realizar a compra e sair do estabelecimento tendo adquirido o que queria em questão de minutos.

Mesmo com tanto tempo de atuação e constante atualização dos procedimentos de análise, manter a taxa de inadimplência controlada é um desafio. Constantemente surgem novos fatores e variáveis a serem analisadas.

Quando a região é conhecidamente mais de baixa renda, existe maior dificuldade na liberação de crédito, visto que a renda é um dos principais fatores causadores da inadimplência. Pessoas com baixa renda dificilmente têm um histórico abrangente de crédito, ou até mesmo o nome livre de dívidas.

Portanto, a maior dificuldade é a atuação com poucos dados sobre os clientes. Sendo considerado os perfis de baixa renda, os mesmos, como já dito, não apresentam histórico significativo, ou dados considerados muitas vezes fundamentais, como renda fixa e comprovante de renda como holerite.

Devido a isso, a companhia libera linhas de crédito menores sem comprovação de renda, que seria um pilar fundamental para garantir um bom perfil. O sistema possui a informação de quantidade de salários mínimos mensal ganhos pelo cliente, porém essa informação para a grande maioria dos casos não é comprovada, sendo apenas declarada pelo cliente.

## 5.1 - Dados

Para o processamento de regressão logística, foram considerados dados de clientes recentes da região nordeste do Brasil. A escolha por essa região se deu por dois fatores: primeiro por ser uma região que começou a ser incorporada recentemente pela administradora na sua rede de credenciados, e, consequentemente, a trazer clientes regionais, e segundo por ser uma região ainda pouco conhecida no cenário financeiro, visto que o foco acaba sempre sendo nas capitais e há muitas áreas com potencial para receber linhas de crédito.

Foram consideradas as seguintes variáveis dos clientes analisados: idade, renda em quantidade de salários mínimos (não comprovada, apenas declarada pelo cliente), estado e cidade em que o cliente mora, limite atual do cartão, sexo, se o cliente em algum momento teve registro de inadimplência nos órgãos de crédito, não necessariamente por inclusão da própria administradora, e se o cliente em algum momento foi considerado inadimplente ou foi protestado pela própria administradora. Esse último dado não contabiliza de o cliente está atualmente protestado, uma vez que tal dado é o considerado como classificador do perfil do cliente, o que contaminaria a análise. Foi considerado se em algum momento do passado o cliente esteve ou não protestado, independente do status atual do mesmo.

Foi considerado uma amostragem de vinte mil clientes, selecionados aleatoriamente, para que fossem processados na regressão. Dessa amostragem, 75% foi separado para o treinamento do algoritmo, enquanto 25% foram processados para se medir a acurácia do modelo.

Durante a fase de levantamento dos dados, foram testados diversos dados e analisada a precisão de predição do algoritmo. Foi observado que, quanto mais variáveis com boa correlação ao objetivo adicionadas, melhor foi a acurácia, contanto que esses dados apresentassem um peso compatível. Foram descartados dados como o total de compras realizado pelo cliente, uma vez que não se observou uma correlação com o fato de o cliente ter ou não um bom perfil de crédito. Ao se aplicar variáveis não correlacionadas, a acurácia diminuía, logo, foram retiradas do set de variáveis final.

Abaixo, a distribuição dos dados utilizados na amostragem:

Gráfico 1: Distribuição da amostragem de clientes por idade. Fonte: do autor

# 

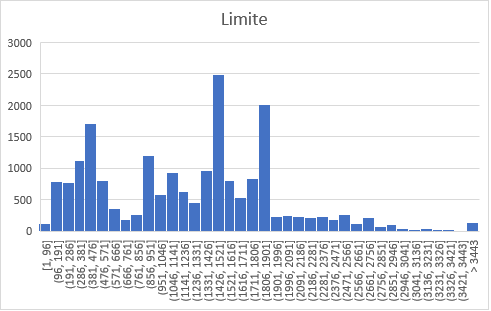


Gráfico 3: Distribuição da amostragem de clientes pelo limite do cartão

Gráfico 2: Distribuição da amostragem de clientes por estado

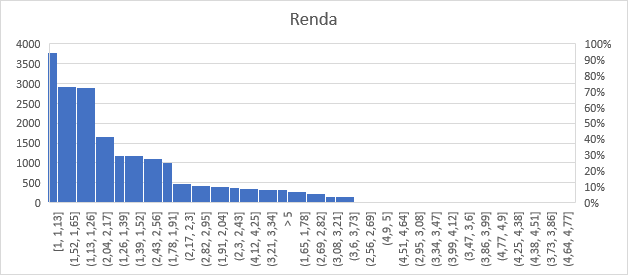


Gráfico 4: Distribuição da amostragem de clientes pela renda em quant. de salários mínimos

Gráfico 6: Distribuição entre clientes que foram em algum momento inadimplentes

Gráfico 5: Distribuição da amostragem de clientes por sexo

Gráfico 8: Distribuição da amostragem de clientes em algum momento presentes nos órgãos de crédito

Gráfico 7: Distribuição da amostragem de clientes em algum momento protestados

## 5.2 - PROCESSAMENTO DOS DADOS

Os dados emitidos diretamente do banco de dados da administradora foram processados em um algoritmo de regressão logística, cujo objetivo seria responder: o cliente possui ou não um bom perfil de crédito? Para definir como bons ou ruins, foi considerado se os clientes da amostragem estão ou não atualmente protestados na administradora, ou seja, são clientes que não cumpriram com suas obrigações financeiras para com a administradora, vindo a ter seus cartões bloqueados e inseridos nos registros dos principais órgãos de crédito.

Os dados foram salvos no formato CSV, os dados booleanos, assim como o sexo, foram contabilizados como 0 ou 1. A última coluna do CSV é o status do cliente, sendo considerado como bom (0) ou ruim (1), indicando se o cliente está ou não protestado.

O algoritmo realiza as seguintes tarefas:

- Lê os dados constantes do arquivo CSV;

- Separa a última coluna desses dados como sendo o resultado, e o restante como as variáveis a serem analisadas;

- Realiza mais uma separação, dessa vez entre o set de treino e o set de análise, sendo que para essa amostragem, ficou determinado 75% para treino, e 25% para análise de predição;

- Escalonamento dos dados, para que os mesmos fiquem dentro de uma escala uniforme, para evitar distorções no processamento;

- Treinamento dos dados no algoritmo de Regressão Logística com o set de 75%, assim como predição do set de 25% dos mesmos;

- Cálculo e exibição da Matriz de Confusão;

- Cálculo e impressão da acurácia.

Quanto maior a acurácia da predição, melhor, pois indica que o algoritmo seria eficiente em classificar um potencial cliente como bom ou ruim levando em consideração as variáveis utilizadas.

# 6- CONCLUSÃO

Ao realizar a predição na amostragem usada nesse estudo, observou-se uma acurácia de 80,18%, ou seja, a cada cinco casos, o algoritmo conseguiu predizer corretamente quatro casos. A acurácia foi considerada muito boa tendo em vista o cenário onde não há tantas variáveis para serem utilizadas e dados importantes como um comprovante de renda não são pedidos. Abaixo, a matriz de confusão para os dados:

|  |  |
| --- | --- |
| 2443 | 424 |
| 1555 | 578 |

Tabela 1: Matriz de Confusão da amostragem testada. Fonte: do autor.

Podemos ver que dos 5000 registros do set de predição, temos um total de 3021 classificados corretamente, sendo 2443 classificados como bons perfis e 578 classificados como maus perfis de crédito. Em contrapartida, houveram 1555 registros erroneamente classificados como bons, assim como 424 erroneamente classificados como ruins.

É desafiador a análise de crédito de clientes que são de um nicho completamente novo não só para a administradora aqui estudada, mas para o mercado em geral. Mesmo com um percentual de acerto de 80%, a análise classificou erroneamente 38,89% dos perfis considerados bons. Do ponto de vista da liberação de crédito, tal porcentual pode causar um impacto enorme na taxa de inadimplência. O cenário é mais favorável quando se classifica bons perfis erroneamente como ruins, uma vez que se perde o custo de oportunidade de fornecer crédito para um possível bom cliente.

Dados os resultados, o algoritmo foi implantado como parte dos dados analisados para liberação de crédito dos clientes que caem para a análise manual. Apesar de não ter o peso final na decisão, a predição da regressão logística para esse nicho de clientes tem uma significativa importância, sendo atrelado a outros pontos que já eram analisados durante o processo.

# 7- REFERÊNCIAS

SICSÚ, J. Para além das políticas de resgate. Revista de Economia Política, vol. 29, nº 1 (113), pp. 144-145, janeiro-março/2009

BRESSER-PEREIRA, L.C. Crise e recuperação da confiança. Revista de Economia Política, vol. 29, nº 1 (113), pp. 133-134, janeiro-março/2009

FIORDELISI, F.; MARQUES-IBANEZ, D.; MOLYNEUX, P. Efficiency and risk in European banking. Journal of Banking & Finance, v. 35, n. 5, p. 1315-1326, 2011.

Basel Committee on Banking Supervision (BCBS). (2011). Basel III: A global regulatory framework for more resilient banks and banking systems, Basel. Disponível em: http://www.bis.org/publ/bcbs189.htm Acesso em outubro de 2022.

STEVENSON, W. J. (1986). Estatística aplicada à administração. São Paulo: Harbra, p. 341

DUDA, R. O.; STORK, D. G. (2001). Pattern classification. Nova Iorque: Willey. 2ª edição.

TAVARES, M.D.C. (2009). A crise financeira atual. Paper Itamaraty

MARCELINO, J. A. (2012). Credit scoring: uma ferramenta para análise de crédito em uma instituição de microcrédito produtivo e orientado

LEWIS, E. M. (1992). An introduction to credit scoring. Fair, Isaac and Company.

LANA, T. P. (2015) Exclusão financeira e sua relação com a pobreza e a desigualdade de renda no Brasil - Rio de Janeiro: BNDES, 2015. 292 p.

CROOK, J.N., EDELMAN, D.B., TOMAS, L.C. 2007. Recent developments in consumer credit risk assessment. European Journal of Operational Research, 183, 1447-1465.

ABDOU, H. A., TSAFACK, M. D. D., NTIM, C. G., & BAKER, R. D. (2016). Predicting creditworthiness in retail banking with limited scoring data. Knowledge-Based Systems, 103, 89-103. Disponível em: https://doi.org/10.1016/j.knosys.2016.03.023. Acesso em agosto de 2023.

1. https://diegonogare.net/2020/04/performance-de-machine-learning-matriz-de-confusao/ [↑](#footnote-ref-1)