



**INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DO PIAUÍ
CAMPUS URUÇUÍ
CURSO LICENCIATURA PLENA EM MATEMÁTICA**

GEAN PEREIRA DAMACENO

**APLICAÇÃO DE REGRESSÃO LOGÍSTICA PARA PREVISÃO DE INADIMPLENTES
DE CARTÃO DE CRÉDITO**

Uruçuí

2023

Gean Pereira Damaceno

**APLICAÇÃO DE REGRESSÃO LOGÍSTICA PARA PREVISÃO DE INADIMPLENTES
DE CARTÃO DE CRÉDITO**

Trabalho de Conclusão de Curso (artigo científico) apresentado como exigência parcial para obtenção do diploma do Curso de Licenciatura plena em Matemática do Instituto Federal de Educação, Ciência e Tecnologia do Piauí, Campus Uruçuí .

Orientador: Prof^a. Dr. Ernandes Guedes Moura.
Coorientador: Prof. Me. Wanderson de Sousa Leite.

URUÇUÍ

2023

**APLICAÇÃO DE REGRESSÃO LOGÍSTICA PARA PREVISÃO DE
INADIMPLENTES DE CARTÃO DE CRÉDITO**

Gean Pereira Damaceno¹
Ernandes Guedes Moura²

Agradecimentos

A Deus, por me conceder saúde, educação, conhecimento, responsabilidade e uma oportunidade incrível de conquistar meus objetivos durante todos os meus anos de estudo.

A minha Família, especialmente a minha mãe, que me apoiou diariamente, contribuindo para que fosse possível realizar este sonho.

Aos amigos Ramon Ferreira da Silva, Dheyson Sandes Costa, Aldeane da Silva Dias, Cynthia Lima de Moraes e Lorrán Carreiro, que sempre estiveram ao meu lado, pela amizade incondicional e pelo apoio demonstrado ao meu longo percurso de todo o curso e pela amizade verdadeira.

Em especial ao meu professor Dr. Ernandes Guedes Moura, pela orientação e a forma com que conduziu o trabalho com paciência e dedicação, sempre disponível a compartilhar todo o seu vasto conhecimento.

Aos meus Professores(a), Esp. Marcílio de Macêdo Vieira, Me. Verônica Elizeu de Araújo Fernandes, Me. Mayara Danyelle Rodrigues de Oliveira, Me. Francisco Nordman Costa Santos, Me. César Marcos do Nascimento Lucas, Me. Erimar dos Santos Oliveira, Me. Gabriel dos Santos Pinto, Me. Ana Paula, Dr. Bruno Ribeiro de mesquita, Esp. Aloisio Campos,

Esp. Celmar Lustosa; ao Instituto Federal de Educação Ciência e Tecnologia do Piauí, Campus - Uruçuí - IFPI, por disponibilizar uma estrutura de ensino de qualidade e a oportunidade da graduação, para mim, foi uma honra poder concluir meu sonho na instituição.

E a todos aqueles que estiveram conosco em sala de aula, durante os estágios e que contribuíram de alguma forma para que eu chegasse até este momento, me apoiando e com um incentivo de ir mais longe na vida acadêmica, não permitindo que eu desistisse e tornando mais leve a caminhada desta conquista.

1 Graduando do curso Licenciatura em matemática. Instituto Federal de educação ciência e tecnologia do Piauí – Campus Uruçuí. Geandamaceno030@gmail.com.

2 Professor do campus Uruçuí do Instituto Federal do Piauí, Licenciado em Matemática, doutor em Estatística experimental e agropecuária. ernandes.guedes@ifpi.edu.br.

RESUMO

A previsão de inadimplentes é um desafio importante para as instituições financeiras, especialmente no setor de cartões de crédito. No presente trabalho, explora-se aplicação de regressão logística como uma técnica para prever a probabilidade de inadimplência de clientes de cartão de crédito, uma vez que, essa técnica é geralmente empregada para predizer eventos binários. Utilizando um conjunto de dados históricos contendo informações socioeconômicos e transacionais dos clientes, construímos um modelo de regressão logística para identificar os principais fatores que contribuem para a inadimplência e avaliar sua capacidade de previsão, via matriz de confusão que confronta dos dados reais e previstos numa pequena tabela, além de calcular algumas métricas de qualidade de ajuste do modelo como acurácia, sensibilidade e especificidade. Por fim, foi construída uma curva ROC para avaliar o desempenho do modelo de classificação à medida que seu limiar aumenta.

Palavras-chave: Estatística; regressão logística; linguagem R; Curva Roc; Matriz de confusão.

ABSTRACT

Predicting defaults is an important challenge for financial institutions, especially in the credit card industry. In the present work, the application of logistic regression is explored as a technique to predict the probability of default of credit card customers, since this technique is generally used to predict binary events. Using a set of historical data containing clients' socioeconomic and transactional information, we built a logistic regression model to identify the main factors that contribute to default and evaluate its predictive capacity, via a confusion matrix that compares actual and predicted data in a small table, in addition to calculating some metrics of model fit quality such as accuracy, sensitivity and specificity. Finally, a ROC curve was constructed to evaluate the performance of the classification model as it your threshold increases.

Keywords: Statistic; logistic regression; R language; Roc Curve; confusion matrix

Data de aprovação: 26/06/2023

1 INTRODUÇÃO

A regressão logística é uma técnica estatística que tem por base a regressão linear e foi desenvolvida para lidar com problemas de classificação. A regressão logística tem várias aplicações no mundo real em muitos setores diferentes. As primeiras aplicações da Regressão Logística se deram em estudos biomédicos, mas, com o passar dos anos, a regressão logística se tornou uma técnica amplamente utilizada em diversas áreas, como ciências sociais, marketing e finanças, (AGRESTI, 2002). A Regressão Logística se tornou uma ferramenta popular na área financeira para classificar clientes.

A análise de risco de crédito é uma preocupação constante para as instituições financeiras, especialmente para aquelas que emitem cartões de crédito. (FERNANDES, 2016). No setor de cartões de crédito, a inadimplência representa um risco financeiro significativo para as instituições em termos de perdas financeiras. A capacidade de prever antecipadamente quais clientes têm maior probabilidade de se tornarem inadimplentes permite que as instituições tenham insights que ajudam a tomar decisões preventivas que minimizam os riscos. A regressão logística é uma técnica estatística comumente utilizada para modelar a relação entre variáveis independentes e uma variável dependente binária, tornando-a adequada para prever a inadimplência.

No presente estudo, serão utilizados dados históricos de clientes, como número de cartões de crédito em uso, renda média mensal, idade, frequência de uso do cartão, entre outros. A análise dos resultados pode ser relevante para as instituições financeiras melhorarem a sua análise de risco de crédito e, conseqüentemente, tomar decisões mais acertadas na concessão de crédito aos seus clientes. Além disso, a aplicação da regressão logística em outros setores, como o de marketing e vendas, também pode ser muito útil para identificar fatores que contribuem para o sucesso ou fracasso de campanhas ou estratégias de negócio.

O crescimento das economias trouxe consigo a expansão das linhas de crédito. Contudo, os alvos deste benefício são famílias que muitas vezes não têm educação financeira necessária para lidar com todos estes produtos financeiros, levando-as a contrair cada vez mais dívidas (ZEFERINO, 2021). Dessa forma, torna-se cada vez mais importante a utilização de metodologias que permitam a estas empresas, concedentes de crédito, entender melhor sua carteira de clientes para oferecer melhores produtos e evitar grandes prejuízos. Gonzales (2018) afirma que aplicações de análise de crédito, por exemplo, podem utilizar a regressão logística para calcular a probabilidade de um indivíduo ser merecedor da concessão de crédito. Informações pessoais como salário, emprego, tipo de moradia, são exemplos de possíveis variáveis que podem ser levadas em consideração. Isto permite não apenas a estimação dessa probabilidade, mas também como traçar o perfil de um bom ou mau pagador.

Há diversas técnicas de mineração de dados, cada uma com suas particularidades, a escolha da tarefa a ser utilizada é um importante passo para a obtenção de resultados satisfatórios dentro de um determinado estudo, isso só é possível com a compreensão da mesma e de sua

capacidade. Neste contexto, o presente estudo tem como objetivo aplicar a regressão logística em dados de clientes de cartão de crédito, com o intuito de identificar os fatores que mais influenciam a probabilidade de inadimplência com suas obrigações financeiras e avaliar sua capacidade de previsão do modelo.

2 Revisão Bibliográfica

Modelos de regressão logística vêm sendo bastante aplicado em várias áreas de conhecimento em especial, na área da saúde e, mais recentemente, em áreas de negócio e Marketing. Assim, em abordagens estatísticas e de aprendizagem de máquina para problemas de previsão em que a variável alvo de interesse é discreta, por exemplo, inadimplência ou adimplência, levaram a uma variedade de aplicações em risco de crédito (FITZPATRICK e MUES, 2016). Dessa forma, a regressão logística parece acomodar bem esse cenário, juntamente com outros métodos de aprendizado de máquinas.

A regressão logística é útil para situações nas quais se deseja prever a presença ou ausência de uma característica ou resultado com base em valores de um conjunto de variáveis preditoras. É importante ressaltar que a regressão logística é semelhante a um modelo de regressão linear, mas é adequada para modelos em que a variável dependente é categórica. Há portanto, três tipos de regressão linear:

- I Regressão logística binária: Caso em que a resposta é, como o próprio nome sugere, é uma variável binária ou dicotômica.
- II Regressão logística multinomial: Caso em que a variável dependente tem três ou mais resultados possíveis;
- III Regressão logística ordinal: Caso em que a variável resposta também tem três ou mais resultados possíveis, mas neste caso, esses valores têm uma ordem definida.

Para o escopo desse estudo, foi empregado a regressão logística binária, dado que o objetivo foi prever inadimplentes ou não. Segundo Gonzales (2018) a regressão logística é uma técnica estatística que tem por objetivo produzir, a partir de um conjunto de dados históricos, um modelo que permita a predição de valores tomados por uma variável categórica, frequentemente binária, em função de uma ou mais variáveis independentes. Dessa forma, o modelo prevê a ocorrência de um evento binário (sim/não, sucesso/fracasso etc.) em relação a um conjunto de variáveis explicativas. A regressão logística é uma extensão da regressão linear, porém, em vez de prever uma variável contínua, ela trabalha com uma variável discreta.

2.1 O modelo logística

A regressão logística faz parte de uma classe de modelos na estatística denominado de modelo linear generalizado (GLM) é uma generalização da regressão linear. A formulação geral de um GLM consiste em três componentes principais: uma função de ligação, uma função de distribuição e um linear preditor. A função de ligação relaciona a média da variável resposta com o preditor linear. Ela pode ser escolhida de acordo com a natureza da variável resposta. No contexto desse estudo, a função de ligação apropriada é a função logit. Já a função de distribuição descreve a distribuição da variável resposta, levando em consideração suas características específicas. Para o presente estudo, em que a variável resposta é binária, comumente se utiliza a distribuição binomial. Por fim, o preditor linear é uma combinação linear das variáveis independentes com os coeficientes do modelo.

A função logit é definida como o logaritmo natural da razão entre a probabilidade de sucesso p e a probabilidade de fracasso $1 - p$. A função logit é expressa da seguinte forma:

$$\text{Logit}(p) = \log\left(\frac{p}{1-p}\right) \quad (1)$$

Em que p é a probabilidade de sucesso estimada no modelo de regressão logística. Já o preditor linear é dado pela seguinte equação:

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \varepsilon_i = \eta \quad (2)$$

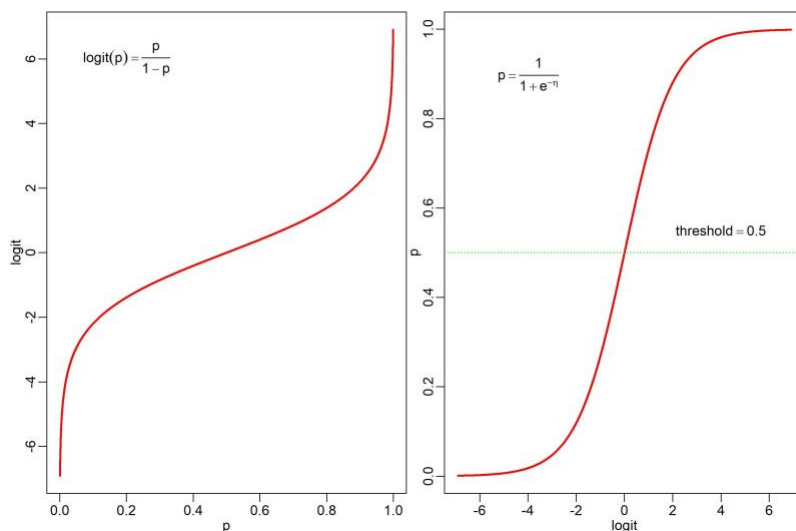
Note que a função logit é utilizada para modelar a relação entre as variáveis independentes e a probabilidade de ocorrência de um evento na variável resposta (binária).

Das equações as equações (1) e (2), tem-se que:

$$\log = \frac{p}{1-p} = \eta \Rightarrow p = \frac{e^\eta}{1+e^\eta} \quad (3)$$

Esses resultados graficamente é dado por:

Figura 1: Curva logit



Fonte: autoria própria

Na figura (1) a linha tracejada em verde é o ponto de corte (threshold) em que, a partir desse ponto de corte o modelo classifica os indivíduos como pertencente à classe positiva de referência, caso contrário, classifica como pertencente a outra classe, ou seja, à classe negativa. É importante ressaltar que a classe positiva é a classe de interesse do estudo, por exemplo, se o interesse for prever inadimplentes, essa será considerada classe positiva para o estudo. Vale ressaltar que geralmente se organiza os resultados previstos e reais em uma tabela de classificação, denominada de matriz de confusão. Para Fernandes *et.al* (2021), o pesquisador deve analisar a matriz de confusão, pois ela fornece uma medida da capacidade preditiva do modelo.

Tabela 1: Matriz de Confusão genérica

Real/Previsto	Classe A	Class B
Classe A	VP	FN
Classe B	FP	VN

Fonte: autoria própria

A tabela (1) permite uma análise mais detalhada dos erros cometidos pelo modelo. Ela possui quatro elementos principais:

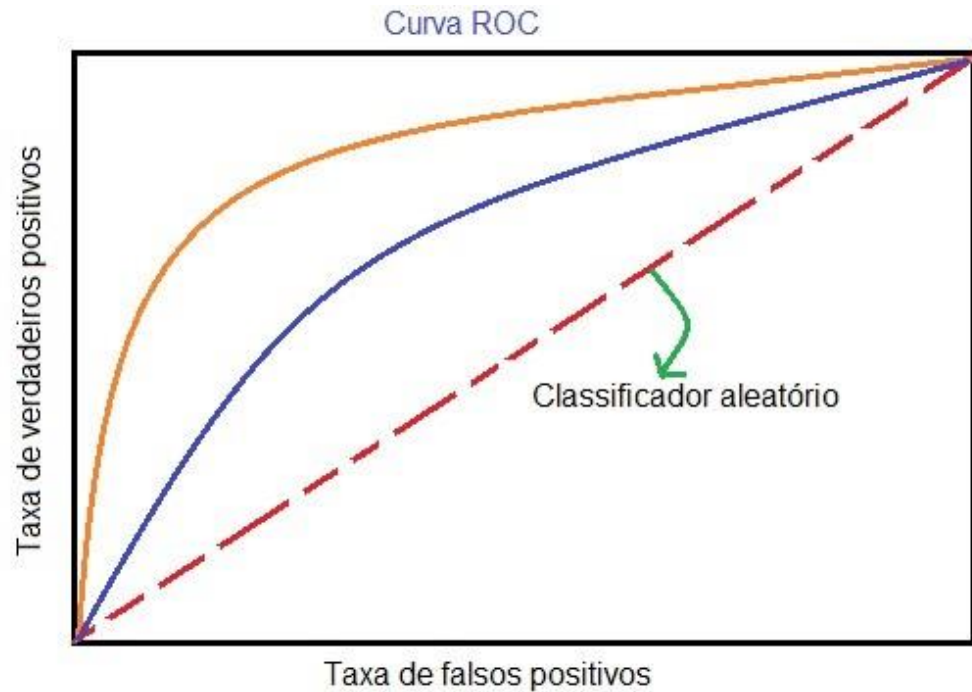
- 1 **VP** significa verdadeiros positivos;
- 2 **FN** significa falsos negativos;
- 3 **FP** significa falsos positivos;
- 4 **VN** significa falsos verdadeiros negativos.

Como visto acima, a partir da tabela (1), pode-se obter várias métricas de qualidade do modelo. As mais utilizadas são sensibilidade, especificidade e acurácia.

- **Sensibilidade** é dado por $VP/(VP+FN)$
- **especificidade** é dado por $VN/(VN+FP)$
- **acurácia** é dado por $(VP + VN)/(VP+VN+FP+FN)$

Uma outra maneira de avaliar um modelo de regressão logística, é por meio de uma curva ROC. Uma curva ROC é uma representação gráfica do desempenho do modelo de classificação. A gráfico abaixo ilustra duas curvas ROC, representando o resultados de dois modelos distintos.

Figura 2: Curva ROC genérica



Fonte: autoria própria

A curva ROC permite avaliar o desempenho de modelos em diferentes pontos de corte e determinar um equilíbrio entre a taxa de verdadeiros positivos e a taxa de falsos positivos. Quanto mais próxima a curva ROC estiver do canto superior esquerdo do gráfico, melhor será o desempenho do modelo. Para a figura (2) por exemplo, o modelo representado pela curva em laranja, é melhor do que o modelo representado pela curva em azul.

3 Metodologia

3.1 Coleta de Dados

O conjunto de dados utilizado para esse estudo, foi extraído da plataforma de dados aberto Kaggle (<https://www.kaggle.com/datasets>), o qual contém informações relacionadas a clientes de uma instituição financeira que estão em atraso no pagamento de suas faturas de cartão de crédito, ou seja, informações de inadimplentes. Para o presente estudo, esse conjunto de dados foi usado para modelagem preditiva utilizando regressão logística. Essa análise, tem por objetivo prever a probabilidade de inadimplência de novos clientes com base em seus atributos históricos.

O conjunto inicial de dados, tinha dimensão 30000 por 24, ou seja, é composto por 23 atributos e uma variável resposta (0= inadimplente e 1= adimplente) de 30000 indivíduos (clientes). Após a coleta dos dados, foi realizado um pré-processamento de limpeza, retirada de redundância e inconsistências e, devidas transformações de dados. Nessa etapa, foi constatado 35 linhas redundantes, reduzindo a dimensão dos dados para 29965 linhas por 24 colunas. Percebe-se que, caso não fosse corrigido, poderia interferir diretamente nos resultados das análises. A etapa acima, é de suma importância para que as inferências sejam coerentes com a realidade. Assim, só após essa etapa inicial de tratamento de dados, é que, então os dados são submetidos às técnicas de análise.

3.2 Modelo

Neste trabalho, foi empregado o modelo de regressão logística que é um modelo estatístico particularmente empregado na previsão de variáveis binárias. Porém, para entender o modelo logístico, é importante entender primeiro um modelo de regressão linear.

3.3 Regressão linear

A regressão linear parte do pressuposto de que existe uma relação linear funcional entre as variáveis, ou seja, que os valores da variável resposta podem ser expressos como uma combinação linear das variáveis independentes ou covariáveis, multiplicadas por coeficientes específicos, denominados parâmetros da regressão linear. O modelo de regressão linear simples é dado por:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad (4)$$

em que y_i é a resposta para o indivíduo i , β_0 é o intercepto, β_1 é a taxa de variação da reta de regressão e ε_i é o erro associado a resposta i .

Embora a regressão linear simples seja muito utilizada, em muitas situações específicas, ela não é suficiente para modelar o fenômeno estudado, pois a regressão linear simples assume uma relação linear direta entre uma única variável independente a uma variável dependente ou resposta. Todavia, a regressão linear múltipla permite incluir várias variáveis independentes ao modelo de regressão linear. Assim, o modelo de regressão linear múltipla é apenas uma extensão da regressão linear simples. O modelo de regressão linear múltipla é dado por:

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \varepsilon_i \quad (5)$$

em que y_i é a resposta para o indivíduo i , $(\beta_0, \beta_1, \dots, \beta_k)$ são os parâmetros de regressão, $(x_1, x_2, \dots, x_k)$ são as variáveis independentes e ε_i é o erro associado a resposta i .

3.4 Regressão logística

Regressão logística é uma técnica estatística mais popular para modelagem preditiva de eventos binários. Assim, como regressão linear, é um modelo linear, o que significa que também é definida por uma equação linear, ou seja, uma equação da forma:

$$g(y_i) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \varepsilon_i = \eta \quad (6)$$

em que $(\beta_0, \beta_1, \dots, \beta_k)$ são as constantes (parâmetros).

Portanto, a regressão logística compartilha com uma série de características com os modelos de regressão linear, mas com a particularidade de ser empregada em cenários de resposta binária. Na regressão logística, a variável dependente é transformada por meio da função logística para que esteja no intervalo entre 0 e 1, representando a probabilidade de ocorrência do evento. Para o emprego da regressão logística, é utilizado uma função de ligação denominada logit. Logit é definido como logaritmo natural da chance, ou ainda, $\text{logit} = \log\left(\frac{p}{1-p}\right)$, que resulta na seguinte equação geral do modelo:

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \varepsilon_i = \eta \quad (7)$$

onde p é a probabilidade de ocorrência do evento (variável dependente), $x_1, x_2, \dots, x_k$ são as variáveis independentes e $\beta_0, \beta_1, \dots, \beta_k$ são os coeficientes estimados pelo modelo.

Sendo que:

- Quando $p \rightarrow \infty$, $\text{logit} \rightarrow 0$;
- Quando $p \rightarrow 1$, $\text{logit} \rightarrow +\infty$;
- Não se define quando $p = 0$ ou quando $p = 1$.

Uma outra forma alternativa, é isolar o p na equação (7), como segue:

$$p = \frac{1}{1+e^{-n}} \quad (8)$$

Essa equação é conhecida na estatística como transformação inversa do logito. É importante lembrar que a regressão logística é um modelo probabilístico e, portanto, a transformação inversa do logito é usada para recompor as probabilidades da variável dependente. Dado isto, geralmente utiliza-se um limiar para classificar as observações como pertencentes a uma das categorias do evento binário com base nessas probabilidades. Para o contexto desse estudo, para um limiar (threshold) maior que 0.5 o cliente é classificado como inadimplente.

3.5 Implementação da análise

As análises foram realizadas utilizando a linguagem de programação Python 3, utilizando a biblioteca scikit-learn (v0.24.2) do Python, por meio do pacote LogisticRegression.

4 Resultados e Discussão

Para treinar o modelo de regressão logística utilizou-se 70% do conjunto de dados e, após o treino, o modelo foi avaliado usando os outros 30% do conjunto de dados como teste. Calculou-se métricas como acurácia, sensibilidade e especificidade para avaliar o desempenho geral do modelo. Além disso, foi construído uma matriz de confusão para visualizar as previsões corretas e incorretas do modelo e, por fim, foi plotado uma curva ROC. A tabela abaixo, representa de forma resumida o desempenho do modelo ao classificar clientes de um cartão de crédito como inadimplente ou adimplente.

Tabela 2: Matriz de confusão

Real/Previsto	Positivo	Negativo
Positivo	4935 (VP)	2043 (FN)
Negativo	950 (FP)	1062 (VN)

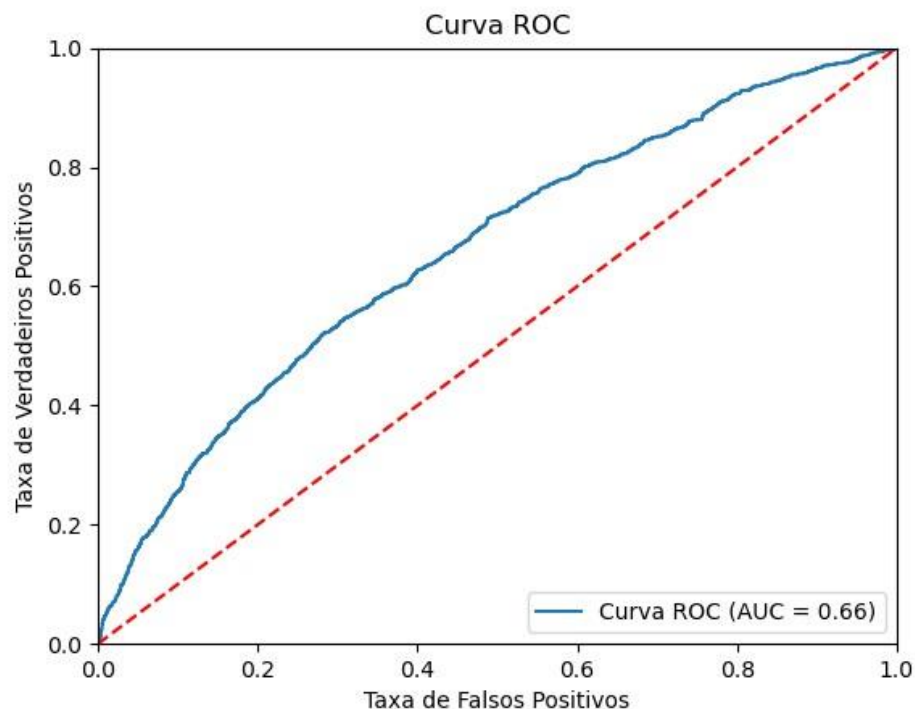
Fonte: autoria própria

De acordo com tabela (2) fica evidente os valores de:

- 1 Verdadeiro Positivo (VP): Representa o número de casos corretamente classificados como positivos pelo modelo. Esses são os casos em que o modelo previu corretamente a classe positiva, isto é, o modelo classificou como inadimplente e de fato, o cliente era inadimplente.
- 2 Verdadeiro Negativo (VN): Representa o número de casos corretamente classificados como negativos pelo modelo. Esses são os casos em que o modelo previu corretamente a classe negativa, isto é, o modelo classificou como adimplente e de fato, o cliente era adimplente.
- 3 Falso Positivo (FP): Representa o número de casos erroneamente classificados como positivos pelo modelo. Esses são os casos em que o modelo previu a classe positiva, mas a classe real é negativa, isto é, o modelo classificou erroneamente como inadimplente, mas o cliente era adimplente. Em estatística, esse erro é denominado de erro do tipo I.
- 4 Falso Negativo (FN): Representa o número de casos erroneamente classificados como negativos pelo modelo. Esses são os casos em que o modelo previu a classe negativa, mas a classe real é positiva, isto é, o modelo classificou como adimplente, mas o cliente era inadimplente. Em estatística esse erro é denominado de erro do tipo II.

A partir desses elementos, é possível calcular várias métricas para avaliar o desempenho do modelo. Iremos abordar as métricas mais empregadas, tais como: sensibilidade, especificidade e acurácia. A sensibilidade é a proporção de verdadeiros positivos em relação a todos os casos reais positivos. A sensibilidade para esse estudo, foi de 0.7, ou seja, a capacidade do modelo de identificar corretamente os casos positivos foi de 70%. Já a especificidade é a proporção de verdadeiros negativos em relação a todos os casos reais negativos. A especificidade do modelo ajustado foi de 0.53, ou seja, a capacidade do modelo de identificar corretamente os casos negativos foi de 53%. Por fim, a acurácia é a proporção de classificações corretas em relação a todos os casos. A acurácia foi de 0.55, isso indica que a capacidade geral do modelo de classificar corretamente os casos é de 55%. Para melhorar a compreensão, foi plotado uma curva que ilustra o desempenho do modelo, essa curva é denominada como curva ROC (do inglês: receiver operating characteristic curve).

Figura 3: Curva ROC



Fonte: autoria própria

A figura (3) é composta pelo eixo x (taxa de falso positivo) e pelo eixo y (taxa de verdadeiro positivo). A curva mostra como o modelo se comporta em diferentes limiares de classificação. Uma curva ROC de um modelo ideal é quando a curva passa próximo ao canto superior esquerdo, indicando uma alta taxa de verdadeiros positivos e uma baixa taxa de falsos positivos. Em outras palavras, a curva ROC mostra a relação entre a sensibilidade e a especificidade em diferentes pontos de corte. Assim, pela curva ROC acima, é possível observar que o modelo foi mediano, pois a curva ROC não se aproximou muito do canto superior direito. É importante ressaltar que a curva ROC é muito utilizada para comparação de modelos, pois o modelo que tiver a curva ROC mais arredondada em direção ao canto superior esquerdo é melhor, em outras palavras, o modelo que tiver a maior área sob sua respectiva curva, é melhor. A área sob a curva ROC é denominada de AUC (do inglês: Area Under the Curve) e varia de 0 a 1, sendo que um valor mais próximo de 1 indica um modelo com melhor capacidade de distinguir entre as classes. Uma AUC de 0.5 indica que o modelo tem desempenho semelhante a uma escolha ao acaso. De acordo com Patri *et.al*(2008), a análise ROC prevê uma avaliação mais rica do que simplesmente avaliar o modelo de classificação a partir de uma única métrica.

5 Conclusão

Demonstramos a aplicação da regressão logística na previsão de inadimplentes de cartão de crédito. Os resultados obtidos mostraram que o modelo de regressão logística pode fornecer previsões coerentes com a realidade. Isso pode ajudar as instituições financeiras a tomar decisões de forma mais segura e implementar medidas preventivas para reduzir o risco de inadimplência. Embora a regressão logística seja uma técnica eficaz para previsão de inadimplentes, é importante ressaltar que outros métodos de aprendizado de máquina também podem ser explorados. Além disso, a qualidade e a representatividade dos dados, bem como um pré-processamento adequado

dos dados, são fundamentais para a construção de modelos precisos. Portanto, é crucial realizar uma análise cuidadosa dos dados e considerar o contexto específico da aplicação antes de implementar qualquer modelo de previsão de inadimplência. Dessa forma, para trabalhos futuros é interessante ajustar outros modelos de classificação para comparar a acurácia preditiva e, assim, adotar o melhor modelo para realizar previsões em novos conjuntos de dados.

6 Referências

- Agresti, A. Categorical Data Analysis. 2ª ed. Hoboken: John Wiley & Sons. 2002.
- FERNANDES, Antônio Alves Tôres et al. Leia este artigo se você quiser aprender regressão logística. Revista de Sociologia e Política, v. 28, 2021. 710 p.
- FERNANDES, Renato Proença. Análise de Risco de Crédito a Particulares: contribuição de variáveis psicológicas e comportamentais. 2016. Tese de Doutorado. FEUC.
- FITZPATRICK, Trevor; MUES, Christophe. An empirical comparison of classification algorithms for mortgage default prediction: evidence from a distressed mortgage market. European Journal of Operational Research, v. 249, n. 2, p. 427-439, 2016.
- Fernando, LPérez: <http://leg.ufpr.br/~lucambio/CE225/202111S/MLGExemplo08.html>. Acesso em: 30 abril. 2023
- GONZALES, Leandro de Azevedo. Regressão Logística e suas Aplicações. São Luís. P(01-46). 2018. Disponível: <https://monografias.ufma.br/jspui/bitstream/123456789/3572/1/LEANDRO-GONZALEZ.pdf>. Acesso: 01 maio. 2023
- Scikit-learn. (2021). LogisticRegression [Software package]. Disponível em https://scikitlearn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html
- ZEFERINO, Júlio César da Silva. Análise do risco de crédito utilizando regressão logística. Varginha. P.(0129), 2021. Disponível em: file:///D:/ARQUIVOS%20WINDOWS/Downloads/TCP_Julio_Zeferino_Corrigido.pdf. Acesso: 30 de abril de 2023
- PRATI, Ronaldo Cristiano et al. Curvas ROC para avaliação de classificadores. Revista IEEE América Latina, v. 6, n. 2, p. 215-222, 2008.