Universidade de Brasília - UnB
Instituto de Ciências Exatas - IE
Departamento de Estatística - EST

Modelo de Credit Scoring via Regressão de Poisson

Rodrigo Dantas Berçott

Orientador: Prof. **Dr. Eduardo Yoshio Nakano**

Brasília - DF, 2022

Rodrigo Dantas Berçott

Modelo de Credit Scoring via Regressão de Poisson

Orientador(a): Prof. Eduardo Yoshio Nakano

Trabalho de Conclusão de Curso apresentado para o Departamento de Estatística da Universidade de Brasília como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Brasília - DF, 2022

Resumo

O objetivo desse trabalho foi propor um escore de risco com base no modelo de regressão de Poisson para classificação de bons e maus e clientes. A metodologia proposta foi ilustrada por meio de um conjunto de dados obtido na literatura sobre clientes solicitantes de crédito. Os resultados mostraram que o escore de risco proposto é útil para a classificação dos clientes e considerado bom pelos analistas, apresentando uma taxa de acertos geral de 76,57%. Esse valor se aproximou muito da taxa de acertos obtida pelo escore de risco baseado no modelo logístico (76,71%), que é atualmente o modelo mais popular para a modelagem de risco. Assim, a modelagem de risco via modelo de regressão Poisson se mostrou uma boa alternativa para a classificação de clientes.

Palavras-chaves: escore de risco, regressão de Poisson, regressão logística.

Abstract

The objective of this work was to propose a risk score based on the Poisson regression model to classify good and bad customers. The proposed methodology was illustrated through a set of data obtained from the literature on credit requesting customers. The results showed that the proposed risk score is useful for classifying customers and considered good by analysts, with an overall hit rate of 76.57%. This value was very close to the hit rate obtained by the risk score based on the logistic model (76.71%), which is currently the most popular model for risk modeling. Thus, risk modeling via Poisson regression model proved to be a good alternative for classifying customers.

Keywords: risk score, Poisson regression, logistic regression.

Lista de Figuras

Lista de Tabelas

Sumário

1 Introdução

O crédito ao consumidor, segundo Lewis (1992), é um negócio essencial. O desafio é tornar o crédito largamente disponível, assim, tantas pessoas quanto possível terão a oportunidade de usar essa poderosa ferramenta. Desta forma, em todo crédito que envolve uma expectativa de retorno financeiro, existe um risco que é a probabilidade de que esse crédito não seja devolvido. Assim, as entidades e as instituições financeiras que intermediam as relações entre credores e devedores necessitam do controle de risco. O sistema financeiro é único e precisa ser preservado, pois a circulação de ativos e investimentos por pessoas, empresas e governos é aqui realizada.

O risco de crédito pode ser avaliado de forma subjetiva pela análise qualitativa que consiste numa avaliação de um analista, mas não quantifica o risco. Ou pode ser medido de forma objetiva por meio de uma metodologia quantitativa. Esse processo de avaliação de risco vem passando por mudança e aprimoramento nos últimos anos. Os métodos qualitativos que se baseiam apenas na análise de um especialista estão perdendo espaço para os métodos que quantificam o risco de uma forma mais objetiva (BRITO; NETO, 2006). Dito isso, para Sicsu (2010), medir o risco de crédito usando técnicas quantitativas tem uma série de vantagens:

- Consistência nas decisões: se submetermos uma mesma solicitação de crédito a diferentes analistas, poderemos obter diferentes avaliações subjetivas e isso não ocorrerá se for aplicado um modelo quantitativo de *credit scoring* porque o escore será o mesmo independente do analista, agência ou filial do credor;
- Decisões rápidas: os recursos computacionais permitem que o escore de risco seja computado quase que instantaneamente;
- Decisões adequadas: o conhecimento das probabilidades de perda permite calcular perdas e ganhos esperados com as operações e os clientes podem ser divididos em classes de risco conforme seu escore;
- Permite verificar o grau que a instituição atende aos requisitos de órgãos reguladores;
- Estabelecer uma linguagem comum entre quem decide o crédito;
- Permite definir níveis de alçada para concessão de crédito.

Durand (1941) iniciou os estudos de modelos de *credit scoring* que afirmou que a Análise de Discriminante poderia ser utilizada para separar os bons e maus empréstimos. Nos anos de 1980, escores de riscos passaram a ser desenvolvidos por meio da Regressão Logística, mas atualmente os modelos de risco de crédito passaram a ser muito populares e

são largamente utilizados. Dito isso, há cada vez mais técnicas estatísticas sendo utilizadas como ferramenta para modelar risco. Dentre elas estão Árvores de Decisão, Redes Neurais, Análise de Sobrevivência, entre outros.

Neste contexto, o objetivo desse trabalho é propor um escore de risco utilizando a regressão de Poisson para taxas e comparar os resultados com a Regressão Logística quanto a sua capacidade preditiva de classificar bons e maus empréstimos. A metodologia proposta será aplicada em um conjunto de dados obtido na literatura e os resultados obtidos serão comparados com aqueles encontrados por meio de um escore de risco utilizando regressão logística. Todas as análises serão realizadas pelo *software R* na versão 4.1.1 (R Core Team, 2021).

2 Revisão de Bibliografia

2.1 Regressão de Poisson

A regressão de Poisson também é conhecida como Modelo Log-Linear de Poisson, faz parte da família de Modelos Lineares Generalizados (MLG) e é adequada para a modelagem de variáveis que envolvam dados de contagem ou taxas.

A distribuição de Poisson pode ser escrita como:

$$P(Y = y) = \frac{\mu^y e^{-\mu}}{y!}, y = 0, 1, 2, ..., \tag{2.1.1}$$

em que Y é a variável aleatória que representa o número de ocorrências e μ é o parâmetro que representa o valor esperado de Y. O efeito das variáveis explicativas na variável resposta Y é modelado através do parâmetro μ.

No modelo de regressão linear normal, o valor médio da resposta na presença de k variáveis explicativas, $x' = (1, x_1, x_2, ..., x_k)$ é dado por: $E(y|x) = x'\beta$, em que $\beta' = (\beta_0, \beta_1, ..., \beta_k)$ é o vetor dos parâmetros desconhecidos.

No entanto, essa mesma representação não é possível no modelo de regressão de Poisson, visto que sua média, μ, é positiva. Sabendo que a função logarítmica é a função de ligação natural para o modelo Poisson, o modelo log-linear é considerado:

$$log(\mu) = \beta_0 + \sum_{j=1}^{k} \beta_j x_j, \tag{2.1.2}$$

que resulta em:

$$\mu = \mu(x) = e^{x'\beta'}, \tag{2.1.3}$$

em que $\beta' = (\beta_0, \beta_1, ..., \beta_k)$ é vetor de parâmetros associado ao vetor de covariáveis $x' = (1, x_1, ..., x_k)$.

2.1.1 Estimação de máxima verossimilhança

Ao considerar uma amostra aleatória, a função de verossimilhança para n observações do modelo de regressão de Poisson é dada por:

$$L(\beta) = \prod_{i=1}^{n} \frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!}. \tag{2.1.4}$$

O *log* da verossimilhança para esse modelo é:

$$\ell(\beta) = log[L(\beta)] = \sum_{i=1}^{n}(y_i log(\mu(x_i)) - \mu(x_i) - log(y_i!)). \tag{2.1.5}$$

Segundo a ligação dada por (2.1.3):

$$\ell(\beta) = \sum_{i=1}^{n}(y_i x_i'\beta - e^{x_i'\beta} - ln(y_i!)). \tag{2.1.6}$$

O Estimador de Máxima Verossimilhança (EMV) de β da equação é obtido resolvendo o seguinte sistema de equações:

$$\begin{cases} \dfrac{\partial\ell(\beta)}{\partial(\beta_0)} = 0 \\ \dfrac{\partial\ell(\beta)}{\partial(\beta_1)} = 0 \\ ... \\ \dfrac{\partial\ell(\beta)}{\partial(\beta_k)} = 0 \end{cases}$$

cuja solução pode ser obtida por meio do método de *Newton-Rapshon* ou outros métodos computacionais.

2.1.2 Qualidade de ajustamento

A fim de avaliar a adequação do modelo de regressão de Poisson, deve-se primeiro olhar as estatísticas descritivas básicas para os dados de contagem de eventos. Caso a média da contagem e a variância forem significativamente diferentes (equivalente em uma distribuição de Poisson), o modelo provavelmente será superdisperso ou subdisperso.

A opção de análise para avaliar a qualidade de ajuste de um modelo de Poisson é utilizar a medida AIC, a estatística de Qui-Quadrado de Pearson e a *deviance* D.

2.1.3 AIC (*Akaike Information Criterion*)

A informação de *Akaike* (AIC) é uma métrica que tem por base o logaritmo de verossimilhança e é uma maneira de selecionar um modelo de um conjunto de modelos. O AIC é definido por:

$$AIC = -2[Log(L) - u]. \tag{2.1.7}$$

Em que u é o número de parâmetros do modelo e L o valor da verossimilhança para o modelo estimado.

O AIC permite comparar modelos alinhados ou não e quanto menor for o seu valor, menor será a informação perdida e, por consequência, melhor será o ajuste do modelo.

2.1.4 Teste de qualidade do ajuste

O desempenho geral do modelo ajustado pode ser medido por dois testes qui-quadrado diferentes: a *deviance* e o qui-quadrado de Pearson.

A *deviance* é uma estatística que avalia a significância dos coeficientes estimados. A função para a distruição de Poisson segundo Dobson (2002) é:

$$D = 2\sum_{i=1}^{n}[y_i log\left(\frac{y_i}{\hat{\mu}_i}\right) - (y_i - \hat{\mu}_i)] \sim \chi^2_{n-u}, \tag{2.1.8}$$

em que:

- y_i é o número de eventos do indivíduo i, $i = 1, 2, ...n$;
- n é o número de observações;
- $\hat{\mu}_i$ é a média de eventos do indivíduo i ajustada pelo modelo de Poisson;
- Quando o modelo está adequado, a estatística segue uma distribuição qui-quadrado com $n - u$ graus de liberdade. Com u sendo o número de parâmetros estimados.

A estatística de qui-quadrado de Pearson para o modelo de Poisson é uma medida importante na avaliação do modelo ajustado. Segundo Dobson (2002):

$$\chi^2 = \sum_{i=1}^{n}\frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}. \tag{2.1.9}$$

Essa equação, assim como a *deviance*, possui distribuição qui-quadrado χ^2 com $n - u$ graus de liberdade.

2.1.5 Teste de Wald

O teste de Wald tem como objetivo avaliar a relação de significância de cada variável explicativa incluída no modelo.

As hipóteses são: $\begin{cases} H_0 : \beta_j = 0 \\ H_1 : \beta_j \neq 0, \ com \ j = 0, 1, ..., k \end{cases}$

sendo a estatística do teste, sob H_0:

$$W_j = \frac{\hat{\beta_j}^2}{\widehat{var}(\hat{\beta_j})} \sim \chi^2_{(1)}, \tag{2.1.10}$$

em que:

- $\hat{\beta_j}$ é o estimador de máxima verossimilhança de β_j;
- $\widehat{var}(\hat{\beta_j})$ é o estimador de máxima verossimilhança de sua respectiva variância;
- sob a hipótese nula, a estatística de Wald segue uma distribuição qui-quadrado com 1 grau de liberdade

2.1.6 Teste da razão de verossimilhança

Um teste simples para testar o ajuste geral do modelo é o teste da razão de verossimilhança (LRT - Likelihood Ratio Test). Dado que L_1 é a função de verossimilhança maximizada de um modelo completo e L_0 como a maximização do modelo sem a covariável $j(j = 1, ..., k)$. A estatística é:

$$LRT = -2log\left(\frac{L_0}{L_1}\right) = -2[log(L_0) - log(L_1)]. \tag{2.1.11}$$

Sob a hipótese H_0 que $\beta_j = 0$, a estatística se aproxima da distribuição de χ^2 com 1 grau de liberdade.

2.1.7 Superdispersão

A distribuição de Poisson assume que a esperança e a variância são iguais, mas, na prática, os dados quase nunca refletem essa suposição. Quando é observado (como costuma ser o caso) que a variância é maior do que a média se tem a superdispersão no modelo . Esse é um problema que afeta sua interpretação.

A estatística de Pearson mostrada em (2.1.9) pode ser usada como teste de superdispersão e uma maneira simples de ajustar a superdispersão é estimar o parâmetro de dispersão dentro do modelo.

2.1.8 Pseudo R^2

O R-quadrado (R^2) é uma medida que representa a proporção da variância para uma variável dependente explicada por uma variável independente em um modelo de

regressão. (PENNSTATE, 2019)

No modelo de regressão de Poisson, o valor de R^2 do modelo linear não pode ser usado. Para isso é comum usar o *pseudo* R^2 que é definido como:

$$R^2 = \frac{\ell(\hat{\beta}_0) - \ell(\hat{\beta})}{\ell(\hat{\beta}_0)} = 1 - \frac{D(\hat{\beta})}{D(\hat{\beta}_0)}, \tag{2.1.12}$$

em que $\ell(\hat{\beta}_0)$ é o logaritmo da verossimilhança do modelo quando apenas o intercepto é incluído e $D(\hat{\beta})$ é a *deviance* apresentado em (2.1.8). O *pseudo* R^2 vai de 0 a 1, sendo 1 o ajuste perfeito.

2.1.9 Análise de resíduos

Resíduo bruto

O resíduo bruto é a diferença entre a resposta observada e a estimativa do valor do modelo. Como falado anteriormente, a esperança e a variância na distribuição de Poisson são iguais, mas é esperado que no resíduo bruto as variâncias não sejam iguais. Isso faz com que a interpretação seja dificultada. A fórmula para o resíduo bruto estimado é:

$$\hat{r_i} = y_i - \exp\{\mathbf{X}_i\hat{\beta}\}, \tag{2.1.13}$$

em que:

- y_i é a resposta observada do i-ésimo indivíduo $i = 1, 2, ..., n$;
- X_i é o vetor de covariáveis do indivíduo i;
- $\hat{\beta}$ é o EMV dos coeficientes do modelo.

Resíduo de Pearson

O resíduo de Pearson corrige a diferença das variâncias do resíduo bruto. É dado por:

$$\hat{p_i} = \frac{\hat{r_i}}{\sqrt{\hat{\phi}\exp\{\mathbf{X}_i\beta\}}}, \tag{2.1.14}$$

em que:

- $\hat{\phi} = \frac{1}{n-u}\sum_{i=1}^{n}\frac{(y_i - \exp\{\mathbf{X}_i\hat{\beta}\})^2}{\exp\{\mathbf{X}_i\hat{\beta}\}}$ é o parâmetro de dispersão para controlar a super dispersão;

- $\hat{r_i}$ é o resíduo bruto apresentado em (2.1.13).

É representado por um gráfico com $\hat{y_i}$ *vs* $\hat{phi_i}$. Isso significa que, se o modelo estiver correto, os resíduos de Pearson devem ter dispersão constante.

Resíduo de *deviance*

Os resíduos *deviance* são úteis na identificação de padrões de covariáveis não ajustados. A deviance do modelo é uma estatística de qualidade do ajuste que está baseada na função de log-verossimilhança. Está definido como:

$$\hat{d}_j = \texttt{sinal}(y_i - \exp\{\mathbf{x}_i\hat{\beta}\})\sqrt{D_i}. \tag{2.1.15}$$

Em que:

- y_i é o valor da resposta do i-ésimo padrão de covariáveis;
- $\exp\{\mathbf{x}_i\hat{\beta}\}$ é o valor ajustado do i-ésimo termo de covariáveis;
- D_i é a *deviance* do i-ésimo termo em (2.1.8).

2.2 Regressão Logística

A regressão logística pertence aos modelos lineares generalizados e é um recurso que nos permite estimar a probabilidade associada à ocorrência de determinado evento em face de um conjunto de variáveis explanatórias.

É constituído pelos modelos em que a variável dependente pode ser associada a uma variável aleatória Bernoulli. Segundo Hosmer e Lemeshow (2013), a principal diferença entre modelos de regressão logística e regressão linear é a distribuição da variável resposta. A regressão logística substitui a distribuição Normal da variável resposta pela distribuição de Bernoulli.

Nos modelos de regressão logística, a variável dependente é de natureza binária usada para estimar a probabilidade de classificar (0) sucesso e (1) fracasso e as variáveis independentes podem ser categóricas ou não.

Seja Y uma variável binária que assume dois valores:

$$Y = \begin{cases} 0, \ sucesso \\ 1, \ fracasso, \end{cases}$$

com probabilidade de ocorrência de eventos: $P(Y = 0) = (1 - \pi)$, $P(Y = 1) = \pi$.

2.2.1 Formulação do modelo de regressão logística

Neste modelo de regressão, a quantidade chave é o valor médio da variável resposta dado o valor da variável independente. É chamado de valor médio condicional que na regressão linear é dado por:

$$E[Y|x] = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k, \tag{2.2.1}$$

Em que $\beta' = \beta_0, \beta_1, ..., \beta_k$ são os coeficientes do modelo associado às variáveis explicativas $x' = (1, x_1, x_2, ..., x_k)$.

Na regressão logística, no caso em que a variável resposta assume 2 valores distintos (0 e 1), tem-se:

$$\pi(x_i) = \frac{e^{(\beta_0 + \beta_1 x_1 ... + \beta_k x_k)}}{1 + e^{(\beta_0 + \beta_1 x_1 + ... + \beta_k x_k)}}. \tag{2.2.2}$$

A transformação fundamental nos modelos de regressão logística é a transformação *logit* que tem o objetivo de linearizar o modelo:

$$logit(\pi(x_i)) = log\left(\frac{\pi(x_i)}{1 - \pi(x_i)}\right) = \beta_0 + \beta_1 x_{1i} + ... + \beta_k x_{ki}. \tag{2.2.3}$$

Sendo que no modelo simples tem-se:

$$g(x) = log\left(\frac{\pi(x)}{1 - \pi(x)}\right). \tag{2.2.4}$$

$$g(x) = log\left(\frac{\frac{exp(\beta_0 + \beta_1 x)}{1 + exp(\beta_0 + \beta_1 x)}}{1 - \frac{exp(\beta_0 + \beta_1 x)}{1 + exp(\beta_0 + \beta_1 x)}}\right) = log\left(\frac{\frac{exp(\beta_0 + \beta_1 x)}{1 + exp(\beta_0 + \beta_1 x)}}{\frac{1}{1 + exp(\beta_0 + \beta_1 x)}}\right), \tag{2.2.5}$$

$$g(x) = log(exp(\beta_0 + \beta_1 x) = \beta_0 + \beta_1 x. \tag{2.2.6}$$

Essa transformação assume a propriedade do modelo de regressão linear em que a função *logit* é linear nos parâmetros.

2.2.2 Razão de Chances

A razão de chances (OR, *Odds Ratio* em inglês) é a medida de associação entre uma exposição e um resultado. Representa a probabilidade de que um resultado ocorra dada uma determinada exposição em comparação com as chances do resultado ocorrer na ausência dessa exposição. É uma forma de interpretar os parâmetros da regressão logística.

A razão $\frac{\pi(x)}{1-\pi(x)}$ é chamada de *odds*. Logo temos segundo Agresti (1990):

$$Odds_1 = \frac{\frac{exp(\beta_0+\beta_1 X)}{1+exp(\beta_0+\beta_1 X)}}{1-\frac{exp(\beta_0+\beta_1 X)}{1+exp(\beta_0+\beta_1 X)}}$$

e

$$Odds_2 = \frac{\frac{exp(\beta_0+\beta_1 (X+1))}{1+exp(\beta_0+\beta_1 X)}}{1-\frac{exp(\beta_0+\beta_1 (X+1))}{1+exp(\beta_0+\beta_1 (X+1))}},$$

o que resulta em:

$$\frac{Odds_2}{Odds_1} = exp(\beta_1). \tag{2.2.7}$$

Assim, tem-se que $exp(\beta_1)$ é a razão de chances de uma variação unitária da covariável X.

2.2.3 Estimação dos parâmetros

Sicsu (2010), afirma que para estimar os parâmetros da regressão logística utiliza-se o método de máxima verossimilhança. Esse método maximiza a função de verossimilhança e será associada à distribuição de probabilidade de Bernoulli.

A função de distribuição $Y \sim Bernoulli(p)$ é dada por:

$$P(Y = y|X) = \pi(x)^y (1-\pi(x))^{(1-y)}$$

.

Considerando uma amostra de n indivíduos, a função de verossimilhança é:

$$L(\beta|Y,X) = \prod_{i=1}^{n} \pi(x_i)^{y_i} (1-\pi(x_i))^{(1-y_i)}$$

.

A log-verossimilhança é definida por:

$$\ell(\beta) = log[L(\beta)] = \sum_{i=1}^{n} Y_i log\left(\frac{\pi(x_i)}{1-\pi(x_i)}\right) + \sum_{i=1}^{n} log(1-\pi(x_i)).$$

Substituinto $log\left(\frac{\pi(x_i)}{1-\pi(x_i)}\right)$ por $\beta_0 + \beta_1 x_{1i} + ... + \beta_k x_{ki}$ tem-se:

$$\ell(\beta) = \sum_{i=1}^{n} Y_i(\beta_0 + \beta_1 x_{1i} + ... + \beta_k x_{ki}) - \sum_{i=1}^{n} log(1 + e^{(\beta_0+\beta_1 x_{1i}+...+\beta_k x_{ki})}).$$

Para completar o procedimento de máxima verossimilhança deriva-se em relação a cada parâmetro e iguala a zero. Assim, o estimador pode ser obtido resolvendo o sistema abaixo:

$$\begin{cases} \frac{\partial \ell(\beta)}{\partial(\beta_0)} = 0 \\ \frac{\partial \ell(\beta)}{\partial(\beta_1)} = 0 \\ ... \\ \frac{\partial \ell(\beta)}{\partial(\beta_k)} = 0 \end{cases}$$

As equações de verossimilhança para regressão logística são não-lineares em β, e assim, necessitam usar métodos numéricos para solucioná-las, como o método de *Newton-Raphson*. (CZEPIEL, 2002).

2.2.4 Seleção de variáveis

Segundo Sicsu (2010), os métodos de seleção das variáveis mais utilizados são:

- *Forward selection*: a seleção inicia-se com um modelo somente com o intercepto e as variáveis são selecionadas e adicionadas ao modelo, uma a uma. A seleção interrompe quando a inclusão de qualquer nova variável não implicar melhoria do poder discriminador do modelo.
- *Backward elimination*: a seleção inicia-se com um modelo contendo todas as variáveis disponíveis. Variáveis são excluídas gradativamente, uma a uma, até que a exclusão de qualquer variável comprometa o poder discriminador do modelo.
- *Stepwise (forward)*: este método é uma mescla das duas técnicas anteriores. As variáveis são gradativamente adicionadas ao modelo. Após a inclusão de uma nova variável, é verificado se variáveis incluídas anteriormente podem ser excluídas devido à entrada da nova variável. Este é o método mais utilizado de seleção de variáveis.

2.2.5 Teste de *Hosmer-Lemeshow*

O teste Hosmer-Lemeshow serve para adequação do ajuste a modelos de regressão logística. É usado com frequência em modelos de previsão de risco. O teste avalia se as taxas de eventos observadas correspondem ou não às taxas de eventos esperadas em G subgrupos da população do modelo. É aplicado apenas para variáveis binárias.

Os dados são primeiro reagrupados ordenando as probabilidades previstas e formando o número de grupos sendo que a quantidade de grupos mais utilizados é $G = 10$ A estatística segue uma distribuição Qui-quadrado.

As hipóteses a serem testadas são:

$$\begin{cases} H_0 : O\ modelo\ se\ ajusta\ bem\ aos\ dados \\ H_1 : O\ modelo\ não\ se\ ajusta\ bem\ aos\ dados \end{cases}$$

Calcula-se as frequências esperadas para $Y = 1$, que é a soma das probabilidades estimadas de todos os componentes do grupo e para $Y = 0$ que é dada por 1 menos a probabilidade do outro grupo (CORRAR; PAULO, 2007). A estatística do teste segue aproximadamente uma distribuição Qui-quadrado com $G-2$ graus de liberdade, e é dada por:

$$H = \sum_{g=1}^{G} \frac{(O_g - N_g\pi_g)^2}{N_g\pi_g(1-\pi_g)} \sim \chi^2_{(G-2)}, \tag{2.2.8}$$

em que:

- N_g: frequência total de pessoas no G-ésimo grupo, $G = 1, 2, .., G$;
- O_g: frequência total de resultados de evento no G-ésimo grupo;
- π_g: probabilidade média estimada prevista de um resultado de eventos para o G-ésimo grupo.

2.2.6 Análise de resíduos

Resíduo bruto

O resíduo bruto é a diferença entre a resposta e a estimativa do valor do modelo. É dado por:

$$\hat{r_i} = y_i - \hat{\pi}_i. \tag{2.2.9}$$

Em que:

- y_i é a resposta observada do i-ésimo indivíduo $i = 1, 2, ..., n$;
- $\hat{\pi}_i$ é o valor ajustado do i-ésimo indivíduo.

Deve-se fazer gráfico do resíduo e comparar com o gráfico dos valores ajustados afim de observar uma propagação que aumenta com os valores ajustados (embora não proporcionalmente).

Resíduo de Pearson

O resíduo de Pearson corrige a variância diferente dos resíduos brutos, dividindo pelo desvio padrão e dado por:

$$\hat{p_i} = \frac{\hat{r_i}}{\sqrt{\hat{\pi}_i(1-\hat{\pi}_i)}}, \tag{2.2.10}$$

em que:

- r_i é o resíduo bruto dado pela equação (2.2.9);
- $\hat{\pi}_i$ é o valor ajustado do i-ésimo indivíduo.

Assim como no resíduo bruto, o resíduo de Pearson compara com o modelo ajustado. Isso significa que, se o modelo estiver correto, os resíduos de Pearson devem ter dispersão constante.

Resíduo de *deviance*

Assim como na regressão de Poisson, o resíduo de *deviance* é a soma dos quadrados desses resíduos. É dado por:

$$\hat{d_i} = \pm\sqrt{2\left(y_i \log\left(\frac{y_i}{\hat{\pi}_i}\right) + (1-y_i)\log\left(\frac{1-y_i}{1-\hat{\pi}_i}\right)\right)}. \tag{2.2.11}$$

O resíduo de *deviance* pode ser calculado como raiz quadrada de duas vezes a diferença entre a probabilidade logarítmica da observação i no modelo saturado e a probabilidade logarítmica da observação i no modelo ajustado.

A soma dos quadrados dos resíduos de deviance somam a deviance residual que é um indicador de ajuste do modelo.

Se um resíduo de *deviance* for extraordinariamente grande, convém verificar se houve um erro ao rotular esse ponto de dados.

2.2.7 Curva ROC

A curva ROC (em inglês, *Receiver Operating Characteristc*) mensura a capacidade de predição do modelo proposto através das predições da sensibilidade e da especificidade. Segundo Fawcett (2006), esta técnica serve para visualizar, organizar e classificar o modelo com base na performance preditiva. Pode ser feita por meio de um gráfico que permite estudar a variação da sensibilidade e especificidade para diferentes pontos de quebra.

Deve-se considerar um ponto de corte C e comparar cada probabilidade estimada com o valor de C. O valor mais utilizado para C é $0,5$ (HOSMER; LEMESHOW, 2013).

Segundo Hosmer e Lemeshow (2013), a regra geral para avaliação do resultado da área sob a curva ROC (AUC) de modelos de *credit scoring* é dada por:

Tabela 1: Índice ROC e classificação

Índice	**Classificação**
$AUC = 0,5$	Não há discriminação
$0,7 \leq AUC < 0,8$	Discriminação aceitável
$0,8 \leq AUC < 0,9$	Discriminação excelente
$AUC \geq 0,9$	Discriminação excepcional

A curva ROC é um gráfico de Sensibilidade (taxa de verdadeiros positivos) *vs* 1 - Especificidade (taxa de falsos negativos), sendo a Especificidade a taxa de verdadeiro positivo. Permite evidenciar os valores nos quais existe otimização da Sensibilidade em função da Especificidade correspondente ao ponto que se encontra mais próximo do canto superior esquerdo do diagrama, uma vez que o indício de verdadeiro positivo é 1 e o de falso positivo é 0.

3 Crédito, risco de crédito e escore de risco

Este capítulo apresenta um compilado teórico do que será abordado neste trabalho de conclusão de curso.

3.1 Crédito, risco de crédito e avaliação do risco de crédito

Crédito, segundo Schrickel (1995), é todo ato de vontade ou disposição de alguém de destacar ou ceder, temporariamente do seu patrimônio à um terceiro, com a expectativa de que esta parcela volte a sua posse integralmente após decorrido o tempo estipulado. Com o fato de envolver a expectativa do retorno de patrimônio, é necessário pensar que todo crédito tem um risco, sendo esse risco a possibilidade de que esta expectativa não se cumpra.

Quando se fala em crédito ao consumidor, Santos (2000) diz que essa expressão pode ser entendida como uma forma de comércio em que uma pessoa física pega dinheiro, bens ou serviços e se compromete a pagar por isso. É nada mais que um financiamento destinado a todos os consumidores, ou seja, qualquer um que quiser fazer compras ou aquisições de produtos, serviços ou bens de uma forma parcelada. Mas isso não quer dizer que o crédito será liberado indistintamente para todos que solicitam, a decisão de liberar ou não o crédito será medida pelo risco.

Risco é qualquer situação que pode afetar a capacidade de atingir objetivos, como a possibilidade de prejuízo financeiro, Lima (2002) diz que "no risco, as probabilidades de ocorrência de um dado evento são conhecidas enquanto na incerteza não há dados para calcular estas probabilidades".

A concessão de crédito é uma decisão sob condições de incertezas, ou seja, sempre há a possibilidade de perder. Caso o credor possa estimar a probabilidade de que essa perda aconteça, sua decisão será mais confiável. O objetivo dos modelos de *credit scoring* é prever a probabilidade que o crédito incorra em perda para o credor. A probabilidade de perda de uma operação é denominada risco de crédito (SICSU, 2010).

Risco de crédito é a medida numérica da incerteza em relação ao recebimento de um valor contratado ou compromissado que deverá ser pago por um tomador de empréstimo. (DUARTE, 1999)

Devido a esse fato, um cliente que causa perdas não aceitáveis ao credor será considerado um "mau cliente". Caso contrário será considerado um "bom cliente" ou, até mesmo, "intermediário".

A avaliação do risco de crédito é o principal ponto para a concessão dele, caso a

empresa avalie mal, poderá perder dinheiro e isso pode acontecer tanto por aceitar um mau cliente, tanto por recusar um bom cliente. Com isso, caso as empresas tenham uma boa avaliação vão levar vantagem por não ficarem vulneráveis em relação às concorrentes. Há duas formas de avaliar o risco potencial de um cliente: por meio de uma análise subjetiva que envolve uma análise de cliente mais qualitativa por meio de um cadastro que, segundo Schrickel (1995), é avaliado caráter, capacidade, capital e condições atuais do momento da economia. Já a outra forma é por meio de uma análise quantitativa que se faz com o desenvolvimento de modelos que são denominados *credit scoring* que é o nome dado no mercado para as fórmulas de cálculo.

3.2 *Credit Scoring*

Modelos de *credit scoring* são sistemas que atribuem pontuações às variáveis de decisão de crédito de um proponente mediante a aplicação de técnicas estatísticas. Esses modelos visam a segregação de características que permitem distinguir os bons dos maus clientes (LEWIS, 1992).

Caouette, Altman e Narayanan (2000) consideram que muitos elementos entram na construção de crédito. Primeiro, devem ser postuladas as relações entre as variáveis que parecem afetar o risco de inadimplência. Depois, com base no corpo de dados, deve ser empregado um conjunto de ferramentas para derivar um modelo formal. Por fim, uma série de testes deve ser aplicada para determinar se o modelo tem o desempenho esperado.

Os modelos de *credit scoring*, a partir de uma equação gerada através de variáveis de operação de crédito formam um escore de crédito que tem finalidade de quantificar o risco de crédito. (SICSU, 2010)

Saunders (2000) diz que esse escore pode ser usado para a classificação de créditos como adimplentes ou inadimplentes, bons ou maus, desejáveis ou não, de acordo com a pontuação obtida por cada crédito. A forma que a informação gerada é utilizada para a decisão de conceder ou não o crédito é atribuição dos gestores de crédito.

Figura 1: Resumo dos passos para o *credit scoring*

Fonte: autor

A diferença de um modelo de *credit scoring* para uma análise subjetiva é que,

no primeiro caso, os fatores de seleçãoF se dão por intermédio de técnicas e processos estatísticos e, com isso, fornecem indicadores quantitativos para liberação do crédito, como as chances de inadimplência de um cliente. Caouette, Altman e Narayanan (2000) afirmam que embora estes sistemas sejam usados para decisões sobre a concessão ou não de crédito, que está centrada na avaliação do risco de crédito ou inadimplência, algumas instituições utilizam-no para determinação do tamanho do crédito a ser concedido.

3.3 Definição escore de risco

O escore de risco pode ser estimado por meio do modelo de regressão logística, sendo a sua grandeza equivalente ao valor calculado do preditor linear do modelo (MACHADO, 2015). No modelo de regressão Poisson com função de ligação log, quanto maior o valor do preditor linear, maior a taxa de ocorrência do evento de interesse, isso implica, no contexto desse trabalho, maior probabilidade do cliente inadimplir.

Desta forma, assim como no modelo de regressão logística, o preditor linear $x'\beta$ pode também ser o escore de risco para o modelo de regressão Poisson, isto é:

$$ER = X'\beta = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k. \tag{3.3.1}$$

4 Resultados

4.1 Dados e Apresentação do problema

O problema apresentado a seguir será propor um escore de risco por meio da Regressão de Poisson e compará-lo com o escore por meio da Regressão Logística.

Neste trabalho, considerado um desfecho dicotômico, representado pela variável resposta Y:

$$Y = \begin{cases} 0, \ \textit{se o cliente for adimplente} \\ 1, \ \textit{se o cliente for inadimplente} \end{cases}$$

Para a construção do modelo, o banco de dados será dividido em dois: treino e teste. O conjunto treino é usado para construção do modelo e representa 70% do total de dados. Por outro lado, os dados de teste são apresentados após a criação do modelo e usados para simular previsões reais, ou seja, permite que o desempenho real seja verificado e testar o ajuste, representa 30% dos dados.

O desenvolvimento do modelo de regressão de Poisson e logística será feito nas seguintes etapas: divisão dos dados em treino e teste, aplicação do modelo de Poisson e logístico com o método *backward* de seleção de variáveis nos dados de treino, testar a aplicabilidade do modelo, aplicar o modelo nos dados teste, avaliar e comparar o desempenho dos 2 modelos.

Os dados utilizados no trabalho serão o *German Credit Data*, disponível pela Universidade da California-Irvine (UCI) em seu repositório: *Machine Learning Repository's*. Foi optado esse banco de dados porque ele é muito usado em trabalhos de análise de risco e é um banco confiável.

Esse conjunto de dados possui 1000 solicitantes de crédito, sendo que desses, 700 foram considerados bons pagadores e 300 foram considerados maus pagadores. O banco possui 21 variáveis sendo 8 numéricas e 13 categóricas

A Tabela 2 possui a descrição de todas as variáveis presentes no banco, com isso, é possível aplicar o modelo desejado com mais confiança de modo que seja mais provável ter resultados mais condizentes com a realidade.

Tabela 2: Dicionário das Variáveis

Variável	**Descrição**	**Tipo**	**Categorias**
Default_status	Variável resposta	Categórica	Adimplente e Inadimplente
Status.of.existing.checking.account	Status da conta corrente existente	Categórica	$A11 : x < 0$ $A12 : 0 \leq x < 200$ $A13 : x \geq 200$ A14: sem conta
Duration.in.month	Duração do empréstimo em meses	Numérica	-
Credit.history	Histórico de Crédito	Categórica	A30: nenhum crédito tomado A31: todos os créditos foram pagos A32: créditos existentes pagos até agora A33: atraso no pagamento no passado A34: conta crítica / outros créditos existentes (não neste banco)
Purpose	Propósito ou finalidade	Categórica	A40: compra carro novo A41: compra carro usado A42: móveis A43: rádio/televisão A44: eletrodomésticos A45: reparos A46: educação A48: reciclagem A49: negócios A410: outros
Credit.amount	Valor do empréstimo	Numérica	-
Savings.account.bond	Poupança/Títulos	Categórica	$A61 : x < 100$ $A62 : 100 \leq x < 500$ $A63 : 500 \leq x < 1000$ $A64 : x \geq 1000$ A65: desconhecido/sem conta poupança
Present.employment.since.	Emprego atual desde	Categórica	A71: desempregado $A72 : x < 1$ *ano* $A73 : 1 \leq x < 4$ *anos* $A74 : 4 \leq x < 7$ *anos* $A75 : x \geq 7$ *anos*
Installment.rate.in.percentage.of.disposable.income	Taxa de prestação em percentagem do rendimento disponível	Numérica	-
Personal.status.and.sex	Estado civil e sexo	Categórica	A91: homem divorciado/separado A92: mulher divorciada/separada/casada A93: homem solteiro A94: homem casado/viúvo A95: mulher solteira
Other.debtors...guarantors	Outros devedores/fiadores	Categórica	A101: nenhum A102: co-requerente A103: fiador
Present.residence.since	Tempo de morador desde	Categórica	$1 : x < 1$ *ano* $2 : 1 \leq x < 2$ *anos* $3 : 2 \leq x < 4$ *anos* $4 : x \geq 4$ *anos*
Property	Propriedade	Categórica	A121: imobiliária A122: contrato de poupança/seguro de vida da sociedade de construção A123: carro ou outro A124: desconhecido
Age.in.Years	Idade em anos	Numérica	-
Other.installment.plans	Outros planos de parcelamento	Categórica	A141: banco A142: lojas A143: nenhum
Housing	Tipo de moradia	Categórica	A151: aluguel A152: própria A153: cedida/de graça
Number.of.existing.credits.at.this.bank.	Número de créditos existentes neste banco	Numérica	-
Job_status	Ocupação	Categórica	A171: desempregado/não qualificado A172: empregado sem qualificação A173: empregado qualificado/funcionário público A174: gerência/autônomo/alta qualificação/policial
Number.of.people.being.liable.to.provide.maintenance.for.	Número de pessoas responsáveis pela manutenção	Numéricia	-
Telephone	Telefone próprio	Categórica	A191: não A192: sim
foreign_worker	Estrangeiro	Categórica	A201: não A202: sim

4.2 Análise descritiva e dados covariáveis

Nessa seção será apresentado a análise descritiva e a análise de correlação dos dados do banco escolhido.

As Tabelas 3 e 4 representam uma análise descritiva das variáveis numéricas e categóricas.

Tabela 3: Medidas-resumo das variáveis numéricas

Variável	**Mín**	**Mediana**	**Máx**	**Média**	**Desvio Padrão**
Duration.in.month	4	18	72	20,90	12,06
Credit.amount	250	2320	18424	3271	2822,737
Installment.rate.of.disposable.income	1,00	3,00	4,00	2,973	1,119
Age.in.Years	19	33	75	35,55	11,38
Number.of.existing.credits	1	1	4	1,407	0,578

Tabela 4: Distribuição das variáveis categóricas

Variável	**Nível**	**Frequência Simples**	**Frequência Relativa**
Default_status	0	700	70,00%
	1	300	30,00%
Credit.history	Nenhum tomado	40	4,00%
	Todos pagos	49	4,90%
	Créditos existentes pagos	530	53,00%
	Atraso no passado	88	8,8%
	Conta crítica/outro banco	293	29,3%
Purpose	Compra carro novo	234	23,4%
	Compra carro usado	103	10,3%
	Móveis	181	18,1%
	Rádio/Televisão	280	28,0%
	Eletrodoméstico	12	1,2%
	Reparos	22	2,2%
	Educação	50	5%
	Reciclagem	9	0,9%
	Negócios	97	9,7%
	Outros	12	1,2%
Savings.account	$x < 100$	603	60,3%
	$100 \leq x < 500$	103	10,3%

Continua

Tabela 4 – *Continuação*

Variável	**Nível**	**Frequência Simples**	**Frequência Relativa**
	$500 \leq x < 1000$	63	6,3%
	$x \geq 1000$	48	4,8%
	Desconhecido	183	18,3%
President.resident	Menos que 1 ano	130	13,0%
	De 1 a 2 anos	308	30,8%
	De 2 a 4 anos	149	14,9%
	Mais de 4 anos	413	41,3%
Present.employment	Desempregado	62	6,2%
	Menos de 1 ano	172	17,2%
	De 1 a 4 anos	339	33,9%
	De 4 a 7 anos	174	17,4%
	Mais de 7 anos	253	25,3%
Personal.status.and.sex	Homem divorciado	50	5,0%
	Mulher divorciada/casada	310	31,0%
	Homem solteiro	548	54,8%
	Homem casado/viúvo	92	9,2%
Other.debtors	Nenhum	907	90,7%
	Co-requerente	41	4,1%
	Fiador	52	5,2%
Property	Imobiliária	282	28,2%
	Poupança/seguro de vida	232	23,2%
	Carro ou outro	332	33,2%
	Desconhecido	154	15,4%
Other.installment.plans	Banco	139	13,9%
	Lojas	47	4,7%
	Nenhum	814	81,4%
Housing	Aluguel	179	17,9%
	Casa própria	713	71,3%
	Cedida/de graça	108	10,8%
Job_status	Desempregado/não qualificado	22	2,2%
	Empregado sem qualificação	200	20,0%
	Empregado qualificado	630	63,0%
	Outros	148	14,8%
Telephone	Não	596	59,6%
	Sim	404	40,4%

Pela Tabela 3 é possível ver que os dados numéricos possuem bastante dispersão e muitos valores outliers como a duração do empréstimo que vai de um mínimo de 4 meses e um máximo de 72 meses. A diferença de idade dos solicitantes de crédito indo de 19 a 75 e o de valor do empréstimo que varia muito.

A Tabela 4 mostra os dados categóricos e suas frequências simples e relativa. Vale destacar que a maioria dos solicitantes possuíam créditos existentes (53,0%). Quando olha-se para o propósito do empréstimo, a maior parte pede para compra de carro novo (23,4%) e rádio/televisão (28%). A grande parte dos solicitantes possuem casa própria (71,3%) e são homens solteiros (54,8%).

O Figura 2 apresenta a relação entre a variável resposta (se o cliente foi adimplente ou não) e algumas variáveis explicativas.

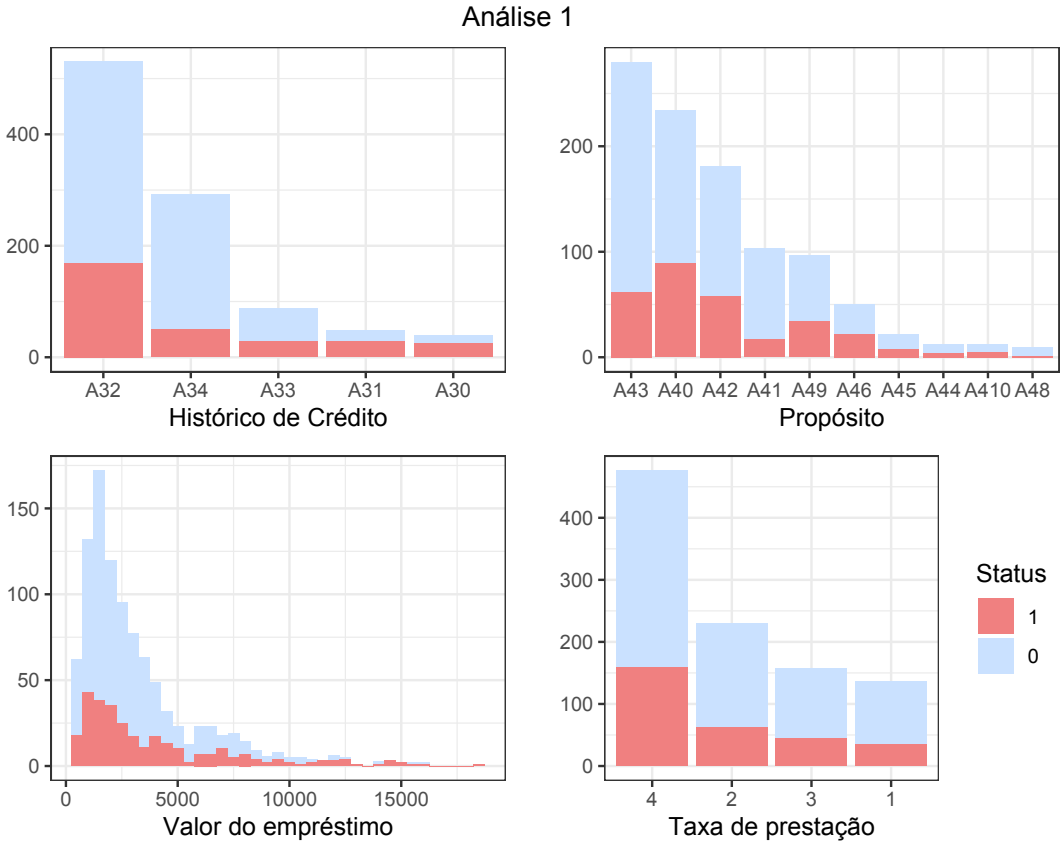

Figura 2: Análise descritiva - gráficos de variáveis em relação a variável resposta

A partir da observação da Figura 2 é possível observar que no primeiro gráfico a grande parte dos solicitantes adimplentes e inadimplentes possuem créditos existentes pagos. No segundo, a maioria das pessoas adimplentes pediram empréstimo com o propósito de comprar móveis, eletrodomésticos e carro novo, de outro modo, os inadimplentes tive-

ram o propósito de comprar carro novo. Ao olhar o terceiro gráfico é possível ver que o valor do empréstimo segue a mesma tendência entre clientes adimplentes e inadimplentes. O quarto mostra que a maioria dos solicitantes pede a maior taxa.

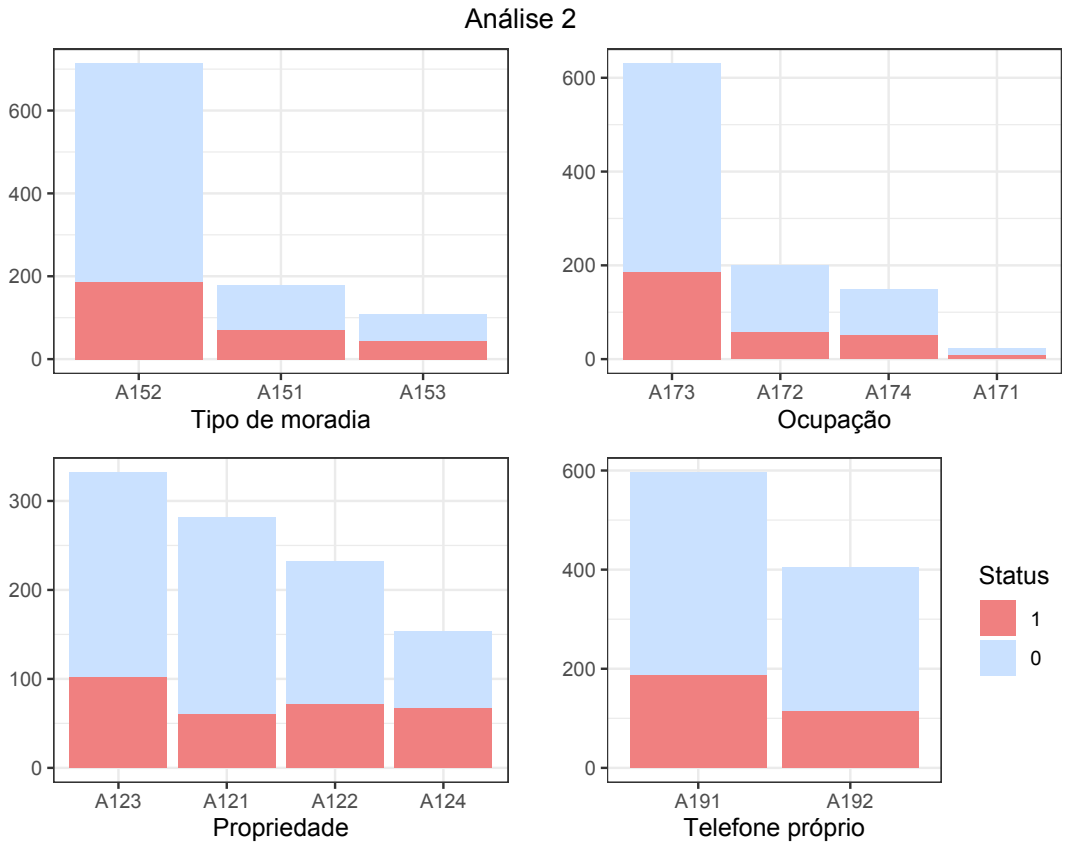

Figura 3: Análise descritiva - gráficos de variáveis em relação a variável resposta

Na Figura 3 pode-se observar que a maioria dos solicitantes possuem casa própria, a grande parte das pessoas também são empregadas. O terceiro gráfico dessa análise mostra que a maioria dos adimplentes e dos inadimplentes possuem bens sendo o carro o bem possuído com a maior frequência. O último gráfico dessa análise mostra uma tendência parecida entre os clientes adimplentes e inadimplentes quando se tem telefone próprio ou não.

A Figura 4 mostra o gráfico de correlação das variáveis numéricas, nele é possível ver que nenhuma variável é muito correlacionada sendo que a maior correlação é da variável Duração com a variável Quantidade com o valor de $0,62$.

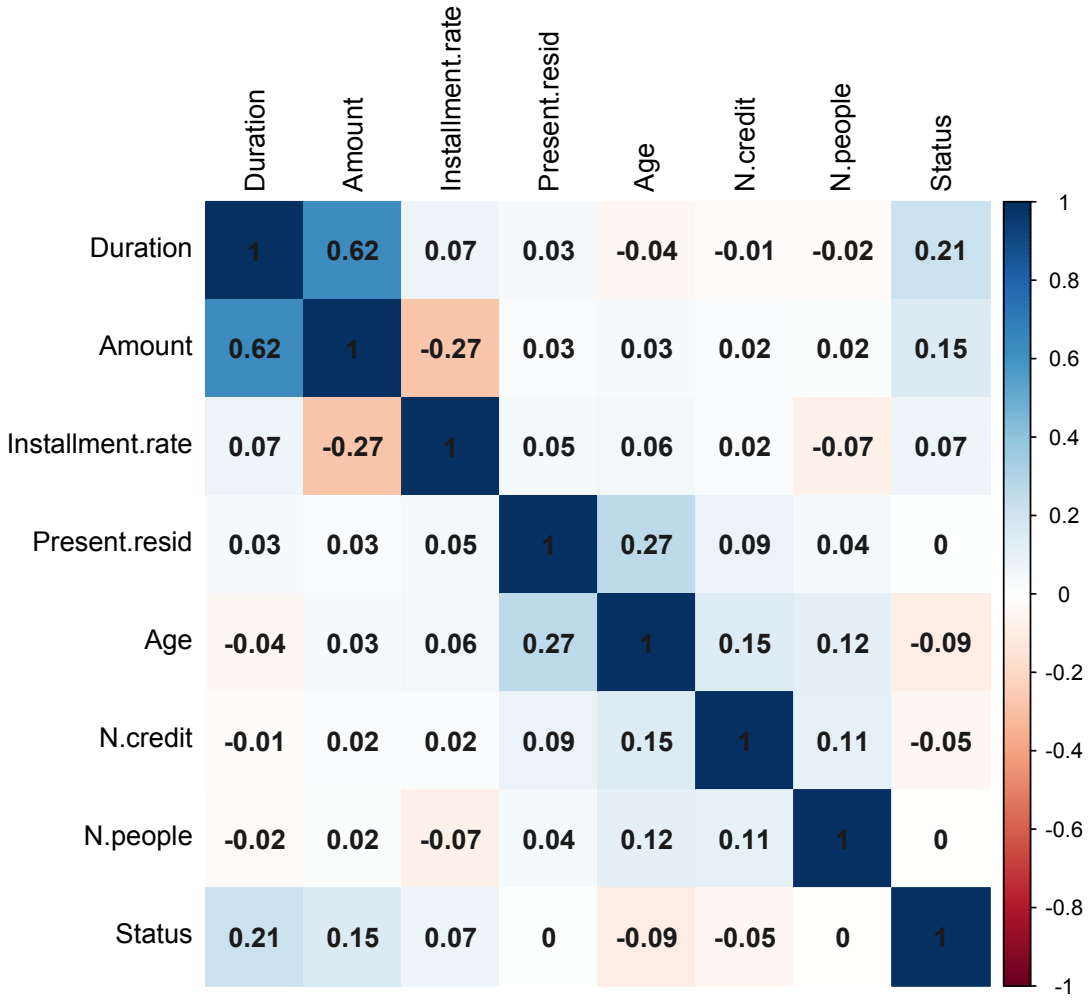

Figura 4: Gráfico de Correlações

4.3 Escore de risco via Regressão de Poisson

Nessa seção será definido o escore de risco por meio da regressão de Poisson, como falado anteriormente, o banco de dados é dividido em treino e teste.

4.3.1 Ajuste do modelo

Para a seleção das variáveis foi usado o comando *Anova* do *Software R*. O procedimento iniciou com o modelo completo, com todas as varíaveis e foram retiradas, uma a uma, aquelas não significativas (que apresentou o maior p-valor) até que restassem somente variáveis significativas (método de seleção *backward*). O nível de significância definido foi de 5%. Com isso, o modelo final ficou:

$$g(x) = -0,994136 - 0,183644x_1 - 0,637593x_2 - 1,433932x_3 + 0,016904x_4 - 0,75434x_5 - 0,548831x_6 - 0,541173x_7 - 0,950532x_8 + 0,135585x_9. \tag{4.3.1}$$

Em que:

- x_1: Status.of.existing.checking.accountA12;
- x_2: Status.of.existing.checking.accountA13;
- x_3: Status.of.existing.checking.accountA14;
- x_4: Duration.in.month;
- x_5: Credit.historyA31;
- x_6: Credit.historyA32;
- x_7: Credit.historyA33;
- x_8: Credit.historyA34;
- x_9: Installment.rate.of.disposable.income.

Esses dados mostram que das 21 variáveis apenas 4 foram significativas. Neste modelo de regressão, os parâmetros positivos aumentam o valor de $\mu(x)$, tendo assim uma relação direta. Por sua vez, os parâmetros negativos fazem com que o valor de $\mu(x)$ diminua à medida que x aumenta, observando assim, uma relação decrescente. A Tabela 5 mostra que as variáveis significativas foram:

Tabela 5: Estimativa dos coeficientes do modelo de regressão de Poisson

Variável	**Estimativa***	**IC 95%**		**P-valor**
(Intercept)	-0,994136	[-1,693078111;	-0,33505310]	0,00405
Status.of.existing.checking.accountA11	0	-		-
Status.of.existing.checking.accountA12	-0,183644	[-0,489983746;	0,12026662]	0,23732
Status.of.existing.checking.accountA13	-0,637593	[-1,358750082;	-0,03025401]	0,05723
Status.of.existing.checking.accountA14	-1,433932	[-1,886278251;	-1,01453936]	9,53e-11
Duration.in.month	0,016904	[0,006584218;	0,02691222]	0,00110
Credit.historyA30	0	-		-
Credit.historyA31	-0,075434	[-0,698201880;	0,55041466]	0,81142
Credit.historyA32	-0,548831	[-1,008145443;	-0,03785512]	0,02581
Credit.historyA33	-0,541173	[-1,202116234;	0,10842253]	0,10254
Credit.historyA34	-0,950532	[-1,509260812;	-0,36644301]	0,00104
Installment.rate.in.percentage.of.disposable.income	0,135585	[0,008730930;	0,26695567]	0,03930

*resultado 0 são os níveis de referência.

4.3.2 Critérios de ajuste do modelo de Poisson

A qualidade do ajuste do modelo final selecionado será feito por informações AIC e por testes de hipóteses, que por sua vez tem o objetivo de validar pela significância.

Tabela 6: Comparação AIC - Poisson

Modelo	**AIC**
Completo	878,21
Final	835,8

A Tabela 6 mostra o comparativo da informação de *Akaike* (AIC) e como foi definido: permite comparar modelos alinhados ou não. Visto que o modelo final, com menos varíaveis, teve um valor do AIC menor que o modelo completo por isso ele foi melhor alinhado.

Agora será feito o teste da razão de verossimilhança para validação do modelo de Regressão de Poisson.

Testes de razão de verossimilhança são usados para comparar a qualidade de ajuste de dois modelos estatísticos. O teste compara dois modelos aninhados hierarquicamente para determinar se adicionar complexidade ao seu modelo (ou seja, adicionar mais parâmetros) torna seu modelo significativamente mais preciso. Os "modelos hierarquicamente aninhados" significam simplesmente que o modelo complexo difere apenas do modelo mais simples (ou "aninhado") pela adição de um ou mais parâmetros. Em resumo, o teste fala se adicionar parâmetros beneficia o modelo ou se é melhor ficar com um modelo mais simples.

As hipóteses são:

$$\begin{cases} H_0 : Melhor\ usar\ o\ modelo\ simples \\ H_1 : Melhor\ usar\ o\ modelo\ complexo. \end{cases}$$

Tabela 7: Teste razão de máxima verossimilhança: constante - Poisson

	Graus de liberdade	**Log Verossimilhança**	**Graus de liberdade**	**Qui-quadrado**	**P-valor**
1	10	-407,90			
2	1	-462,83	-9	109,87	0,000

O primeiro teste será comparar o modelo final com a constante. Pela Tabela 7 é possível concluir que o modelo final é significativamente melhor do que o modelo apenas com a constante, ou seja, sem as covariáveis.

Tabela 8: Teste razão de máxima verossimilhança: modelo inicial - Poisson

	Graus de liberdade	**Log Verossimilhança**	**Graus de liberdade**	**Qui-quadrado**	**P-valor**
1	10	-407,90			
2	49	-390,11	39	35,582	0,6266

O próximo teste é comparar o modelo final com o modelo inicial que possui todas as 21 variáveis do banco de dados. Pelos resultados na Tabela 8 é possível concluir que não faz sentido adicionar mais variáveis ao modelo final, porque possuem a mesma verossimilhança e como o modelo final possui menos variáveis foi o escolhido para os cálculos.

Tabela 9: Teste qui-quadrado - Poisson

Variável	**DF**	**Deviance**	**DF residual**	**Deviance residual**	**P-valor**
NULL			699	505,67	
Status.of.existing.checking.account	3	75,742	696	429,93	2,512e-16
Duration.in.month	1	15,207	695	414,72	9,634e-05
Credit.history	4	14,526	691	400,19	0,005792
Installment.rate.in.percentage.of.disposable.income	1	4,397	690	395,80	0,035993

Outro teste é o de adequação de ajuste da qui-quadrado que adiciona uma variável por vez e começa de cima para baixo, indo de um modelo sem nenhuma variável para o completo. Pela Tabela 9 é possível concluir que todas as variáveis quando adicionadas de uma a uma trazem ganho significativo quando $\alpha = 5\%$.

O último teste para a validação do modelo é o de Hosmer e Lemeshow que considera a hipótese estatística de que as classificações em grupo previstas são iguais as observadas, ou seja, é um teste de ajuste do modelo aos dados. Tomando como base 10 grupos, a Tabela 10 mostra os valores esperados e observados.

Tabela 10: Hosmer e Lemeshow - Poisson

	Observado	Estimado
Grupo	$Y = 0$ — $Y = 1$	$Y = 0$ — $Y = 1$
[0,0432; 0,0752]	68 — 3	66,56248 — 4,437524
(0,0752; 0,0968]	67 — 2	63,15218 — 5,847821
(0,0968; 0,122]	57 — 13	62,43084 — 7,569155
(0,122; 0,179]	60 — 10	59,90185 — 10,098152
(0,179; 0,261]	53 — 17	54,51421 — 15,485789
(0,261; 0,332]	52 — 18	49,50259 — 20,497405
(0,332; 0,375]	50 — 23	46,98121 — 26,018792
(0,375; 0,459]	35 — 34	40,09567 — 28,904329
(0,459; 0,567]	30 — 38	32,53497 — 35,465031
(0,567; 1,25]	18 — 52	14,32400 — 55,676001

O teste de Hosmer e Lemeshow tem como hipóteses:

$$\begin{cases} H_0 : O\ modelo\ ajusta\ bem\ aos\ dados \\ H_1 : O\ modelo\ não\ ajusta\ bem\ aos\ dados. \end{cases}$$

Ao calcular o teste por meio do comando *hoslem.test* tem-se: $\chi^2 = 11,909$ com 8 graus de liberdade e p-valor $= 0,1553$. Os resultados não rejeitam H_0 e pode-se concluir que os valores se ajustam bem ao modelo.

A partir dos resultados dos testes feitos conclui-se que o modelo escolhido está bem ajustado e validado.

4.3.3 Obtenção do escore de risco

Para obter o escore de risco no modelo de regressão Poisson com função de ligação log, quanto maior o valor do preditor linear, maior a taxa de ocorrência do evento de interesse, isso implica maior probabilidade do cliente inadimplir.

Há 2 tipos de erro possíveis, um quando o modelo classifica os bons pagadores como maus e outro quando os maus pagadores são classificados como bons. Visto que o erro mais grave do modelo seria categorizar os maus pagadores no lugar de bons pagadores, foi definido um ponto de corte que o controle. Logo esse ponto de corte foi definido de forma que a probabilidade desse erro fosse de no máximo 20%, resultando em 0,40. Assim, clientes com $Escore\ de\ risco \leq 0,40$ foram classificados como Bons Pagadores e aqueles com $Escore\ de\ risco > 0,40$ como Maus Pagadores.

Com a amostra de treino é possível verificar os erros e acertos de um modelo de Regressão de Poisson pela matriz de confusão.

Tabela 11: Classificação - Poisson

	Observado		
Predito	Bom	Mau	% de acertos
Bom	419	93	81,8%
Mau	71	117	62,23%
Total	490	210	76,57%

Segundo Picinni e Oliveira (2003): "Modelos de credit scoring com taxas de acerto acima de 65% são considerados bons por especialistas". Logo, como o modelo teve uma taxa de acerto global de 76,57%, ele foi considerado um bom modelo.

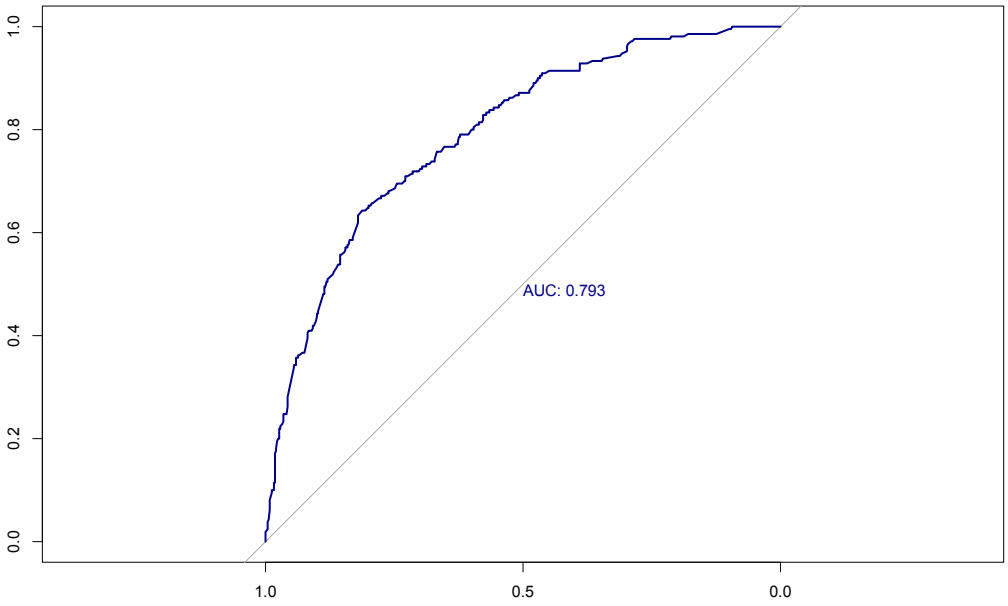

Figura 5: Curva ROC - Poisson

Avaliando a curva ROC, presente na Figura 5, por meio da regressão de Poisson verifica-se que a área da curva corresponde a aproximadamente $0,793$ o que, segundo Hosmer e Lemeshow (2013) está próximo de ser um poder de discriminação excelente.

Tabela 12: Classificação: teste - Poisson

	Observado		
Predito	Bom	Mau	% de acertos
Bom	170	45	79,07%
Mau	40	45	53,0%
Total	210	90	71,67%

Como é possível ver pelo banco teste, a taxa de acerto global permaneceu próxima do banco treino com 71,67% de acerto e o erro controlado chegou próximo dos 80%. O valor da AUC na amostra teste teve uma leve queda para 0,722. A Tabela 13 abaixo mostra um resumo da amostra treino e teste.

Tabela 13: Medidas treino e teste - Poisson

Medida	Treino	Teste
Taxa de acerto	76,57%	71,67%
AUC	79,3%	72,2%

4.4 Escore de risco via Regressão Logística

Agora, nesta parte do trabalho, será feita a análise de risco de crédito usando a regressão logística que é um dos métodos mais utilizados e confiáveis de definição do escore. Esse modelo vai ser usado como forma de comparação entre o que foi feito via regressão de Poisson por isso será feito o model com as mesmas variáveis do modelo de regressão de Poisson, com isso, a comparação será mais justa.

4.4.1 Ajuste do modelo

Com as variáveis sendo iguais a da regressão de Poisson, a equação da regressão logística ficou:

$$g(x) = -0,122827 - 0,325131x_1 - 1,031631x_2 - 2,063608x_3 + 0,033675x_4 - 0,149683x_5 - 1,387592x_6 - 1,418623x_7 - 1,977860x_8 + 0,234622x_9. \tag{4.4.1}$$

Em que:

- x_1: Status.of.existing.checking.accountA12;
- x_2: Status.of.existing.checking.accountA13;
- x_3: Status.of.existing.checking.accountA14;
- x_4: Duration.in.month;
- x_5: Credit.historyA31;
- x_6: Credit.historyA32;
- x_7: Credit.historyA33;
- x_8: Credit.historyA34;
- x_9: Installment.rate.of.disposable.income.

A Tabela 14 apresenta as estimativas dos coeficientes do modelo de regressão logística com o Intervalo de Confiança (IC) definido a 95%.

Tabela 14: Estimativa dos coeficientes do modelo de regressão logística

Variável	**Estimativa**	**IC 95%**	**P-valor**
(Intercept)	-0,122827	[-1,19687905; 0,98376340]	0,82443
Status.of.existing.checking.accountA11	0	-	-
Status.of.existing.checking.accountA12	-0,325131	[-0,76294022; 0,11004136]	0,14384
Status.of.existing.checking.accountA13	-1,031631	[-1,86231873; -0,27670010]	0,01010
Status.of.existing.checking.accountA14	-2,0636082	[-2,59152972; -1,56266572]	3,19e-15
Duration.in.month	0,033675	[0,01869834; 0.04894431]	1,23e-05
Credit.historyA30	0	-	-
Credit.historyA31	-0,149683	[-1,34849780; 1,03867153]	0,80474
Credit.historyA32	-1,387592	[-2.32071274; -0.51847260]	0,00236
Credit.historyA33	-1.418623	[-2,51504000; -0,38018604]	0,00887
Credit.historyA34	-1,977860	[-2,96276405; -1,05214169]	4,46e-05
Installment.rate.in.percentage.of.disposable.income	0,234622	[0,06552567; 0,40780104]	0,00713

A Tabela 14 mostra o resultado dos coeficientes significativos para a formulação do escore de risco do modelo final da regressão logística, nele possui a estimativa, o Intervalo de Confiança e o p-valor. Assim como na regressão de Poisson, todas as 4 variáveis também foram significativas no modelo logístico.

4.4.2 Critérios de ajuste do modelo logístico

O ajuste será feito apenas com o teste de Hosmer e Lemeshow para saber se os dados se ajustam bem ao modelo de regressão Logística.

O teste de Hosmer e Lemeshow tem como hipóteses:

$$\begin{cases} H_0 : O\ modelo\ ajusta\ bem\ aos\ dados \\ H_1 : O\ modelo\ não\ ajusta\ bem\ aos\ dados. \end{cases}$$

Ao calcular o teste foi possível ver que: $\chi^2 = 11,19$ com 8 graus de liberdade e $p - valor = 0,1912$. Os resultados concluem que não rejeita H_0, ou seja, não há indícios que o modelo logístico não se ajusta bem aos dados.

4.4.3 Obtenção do escore de risco

Como dito anteriormente, o escore de risco pode ser estimado por meio do modelo de regressão logística, sendo a sua grandeza equivalente ao valor calculado do preditor linear do modelo (MACHADO, 2015).

O grande objetivo de determinar o escore de risco é controlar o erro de categorizar os maus clientes em bons em até 20%. Para que isso fosse possível foi determinado o ponto de corte em 0,45. Assim, clientes com $Escore\ de\ risco \leq 0,45$ foram classificados como Bons Pagadores e aqueles com $Escore\ de\ risco > 0,45$ como Maus Pagadores. A demonstração é feita por meio da matriz de confusão.

Tabela 15: Classificação - Logística

	Observado		
Predito	Bom	Mau	% de acertos
Bom	427	100	81,02%
Mau	63	110	63,58%
Total	490	210	76,71%

O modelo de regressão logística teve um desempenho bem avaliado pelos especialistas que consideram 65% como um bom resultado. (PICINNI; OLIVEIRA, 2003). A taxa de acerto global foi de 76,71% como mostra a Tabela 15. O objetivo falado anteriormente era o de controlar o erro grave de considerar maus pagadores em bons e isso foi feito muito bem com 81,02% de acerto.

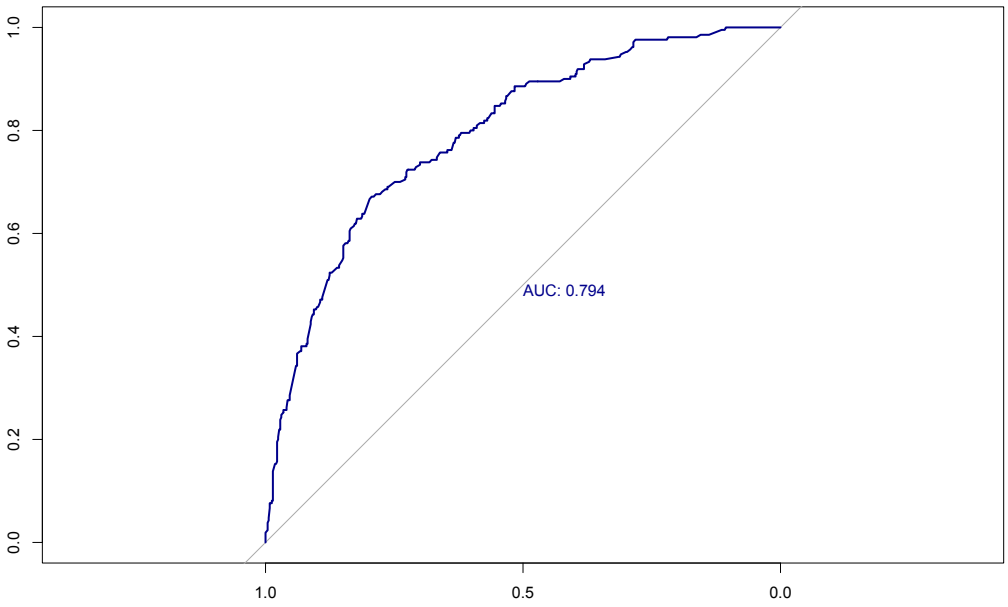

Figura 6: Curva ROC - Logístico

A curva ROC, representada na Figura 6, mostra que a área da curva explica 79% dos dados pela AUC (área embaixo da curva), que segundo Hosmer e Lemeshow (2013) está próximo de ter um poder de discriminação excelente.

Tabela 16: Classificação: teste - Logística

	Observado		
Predito	Bom	Mau	% de acertos
Bom	175	51	77,43%
Mau	35	39	52,7%
Total	210	90	71,3%

Os dados de teste mostram que pela Tabela 16, os dados estão próximos do banco de treino com 71,3% de acerto global.

Tabela 17: Medidas treino e teste - Logística

Medida	Treino	Teste
Taxa de acerto	76,71%	71,3%
AUC	79,4%	71,9%

A Tabela 17 mostra um comparativo entre a taxa de acerto e a AUC dos bancos de treino e de teste. É possível ver que a taxa de acerto permaneceu bem próxima nos

dois bancos, mas por ter menos variáveis ficou um pouco abaixo e isso foi visto também no AUC, em que teve uma proximidade, mas um pouco mais baixo do que no banco treino.

4.5 Comparação dos dois modelos apresentados

O objetivo desse trabalho sempre foi a comparação do escore de risco do modelo de Poisson com o modelo logístico, a Tabela 18 mostra os dados comparativos entre os dois modelos pela matriz de confusão.

Tabela 18: Comparação entre os modelos - amostra Treino

Modelo	Taxa de acerto	AUC	AIC
Poisson	76,57%	79,3%	835,8
Logístico	76,71%	79,4%	686,72

Pela Tabela 18 é possível notar que os dois modelos tiveram um acerto global muito próximo com uma diferença mínima. O modelo de regressão logística e de regressão de Poisson tiveram um excelente desempenho na AUC que mede a discriminação entre as classes estudadas, neste caso, bons e maus clientes com apenas 0,1% de diferença entre os dois modelos.

A maior diferença entre os dois modelos foi na medida AIC, em que o modelo logístico teve um resultado consideravelmente menor que o de Poisson com uma diferença de quase 150. Quanto menor o valor maior a sua acurácia, que indica um melhor percentual de acertos e menor percentual de Falsos Positivos e os dados mostram que o modelo de regressão logística teve um resultado melhor.

Um teste que pode ser feito para modelos de *credit scoring* é o teste de Kolmogorov-Smirnov (KS). É um teste não paramétrico para determinar se duas amostras foram extraídas da mesma população. Segundo Picinni e Oliveira (2003), o teste de Kolmogorov-Smirnov é utilizado no mercado financeiro como um dos indicadores de eficiência de modelos de Credit Scoring, sendo que o mercado considera um bom modelo aquele que apresente um valor de KS igual ou superior a 30.

Tabela 19: % KS dos modelos

Modelo	KS
Poisson	30,95%
Logístico	30,47%

A Tabela 19 mostra que ambos os modelos tiveram um resultado satisfatório

acima dos 30% com uma diferença muito pequena entre os dois, isso mostra que os dois modelos são bem adequados.

5 Considerações Finais

O objetivo do trabalho foi propor um escore de risco com base em um modelo de regressão de Poisson para classificação de bons e maus clientes em um banco chamado *German Credit Data* disponível na internet, no repositório de *Machine Learning Repository's* da Universidade da California-Irvine (UCI).

Os resultados do modelo de regressão de Poisson mostram que o escore de risco proposto é considerado pelos especialistas como um bom resultado, apresentando uma taxa de acertos global em 76,57%. O resultado se aproxima muito da taxa de acertos obtida pelo escore de risco baseado no modelo logístico que atualmente é o modelo mais popular para a modelagem de risco com 76,71%.

Para fazer a comparação entre os dois modelos foram usados 4 diferentes indicadores, a taxa de acerto, a AUC, a AIC e a medida do teste de KS. Os resultados dos indicadores usados não obtiveram uma diferença significativa entre o modelo de regressão logística e o modelo de regressão de Poisson, apenas para o valor de AIC que a diferença foi maior.

Como tópicos de estudos futuros é possível aplicar o modelo de regressão de Poisson, com todos os indicadores sendo aceitáveis pelos especialistas e muito próximo ao modelo logístico. Uma taxa de acerto alta, um erro considerado grave baixo e com uma discriminação perto de ser excelente. Caso o estudo seja feito com uma base de dados maior é provável que o modelo se saia ainda melhor.

A regressão de Poisson se mostrou bastante eficaz para a previsão do evento estudado, pois teve uma taxa de acerto alta, um erro grave baixo e uma discriminação perto de ser excelente. Por mais que o modelo necessite ser ajustado rotineiramente, por conta de fatores externos que podem influenciar a disponibilidade de crédito. A modelagem de risco via modelo de regressão Poisson se mostrou uma boa alternativa para a classificação de clientes.

Referências

AGRESTI, A. *Categorical Data Analysis*. [S.l.]: New York: John Wiley, 1990.

BRITO, G. A. S.; NETO, A. A. Modelo de classificação de risco de crédito de empresas. *USP: Revista de Contabilidade e Finanças*, 2006.

CAOUETTE, J. B.; ALTMAN, E.; NARAYANAN, P. *Gestão do Risco de Crédito*. [S.l.]: Qualitymark, 2000.

CORRAR, S. L.; PAULO, E. *Análise multivariada para cursos de administração, ciências contábeis e economia*. [S.l.]: São Paulo: Atlas, 2007.

CZEPIEL, S. A. Maximum likelihood estimation of logistic regression models: Theory and implementation. 2002.

DOBSON, A. J. *An introduction to generalized linear models Second Edition*. [S.l.]: Chapman Hall/CRC, 2002.

DUARTE, A. M. J. Gerenciamento de riscos corporativos. *São Paulo: Bolsa de Mercadorias & Futuros*, 1999.

DURAND, D. *Risk Elements in Consumer Instalment Financing Technical Edition*. [S.l.]: NBER, 1941.

FAWCETT, T. An introduction to roc analysis. Pattern Recognition Letters, 2006.

HOSMER, D. W.; LEMESHOW, S. *Applied Logistic Regression*. [S.l.]: Wiley, 2013.

LEWIS, E. M. An introduction to credit scoring. *San Rafael: Fair Isaac and Co., Inc.*, 1992.

LIMA, J. D. de. A análise econômico-financeira de empresas sob a Ótica da estatísca multivaria dissertação de mestrado. *Curitiba: Universidade Federal do Paraná*, 2002.

MACHADO, A. R. Collection scoring via regressão logística e modelo de riscos proporcionais de cox. Dissetação de mestrado - Universidade de Brasília, 2015.

PENNSTATE. Poisson regression. 2019.

PICINNI, R.; OLIVEIRA, G. *Mineiração de Critério de Credit Scoring Utilizando algorítmos genéricos*. [S.l.]: VI Simpósio Brasileiro de Automação Inteligente, Bauru, SP, 2003.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2021. Disponível em: ⟨https://www.R-project.org/⟩.

SANTOS, J. A. *Análise de Crédito: Empresas e Pessoas Físicas*. [S.l.]: atlas, 2000.

SAUNDERS, A. *Medindo o Risco de Crédito - Novas Abordagens para o Value at Risk e Outros Paradigmas*. [S.l.]: Qualitymark, 2000.

SCHRICKEL, W. K. *Análise de Crédito: Concessão e Gerência de Empréstimos*. [S.l.]: atlas, 1995.

SICSU, A. L. *Credit Scoring*. [S.l.]: Blucher, 2010.