Universidade de Brasília
Departamento de Estatística

Escore de risco via regressão logística ordinal

Filipe Oliveira do Vale Ribeiro

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
2022

Filipe Oliveira do Vale Ribeiro

Escore de risco via regressão logística ordinal

Orientador(a): Prof.º 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
2022

Agradecimentos

Gostaria de agradecer ao Professor Doutor Eduardo Yoshio Nakano pela sua orientação e críticas e a todo corpo docente que me formou desde as etapas mais simples da escola aos meus atuais formadores na universidade.

Um muito obrigado também aos meus colegas de curso pelos dias e noites de estudos e pelos diversos momentos de alegria na UnB. E um agradecimento muito especial á minha família por terem me dado todas as condições para ter chegado aonde cheguei.

Resumo

O objetivo deste trabalho foi propor um escore de risco com base no modelo de regressão logística ordinal para classificação de clientes tomadores de crédito. Uma das principais vantagens de se considerar respostas com três ou mais níveis de risco é poder definir dois ou mais pontos de cortes para a classificação de risco. Isso permite controlar os dois principais erros cometidos nesse tipo de modelagem: taxa de falsos negativos e falsos positivos. A metodologia proposta foi ilustrada por meio de um conjunto de dados obtido na literatura e os resultados mostraram que o escore de risco proposto é útil para a classificação dos clientes, apresentando uma taxa de acertos geral de 70% ao limitar os falsos negativos e falsos positivos em 10% e 20%, respectivamente. Assim, o escore de risco proposto neste trabalho se mostrou uma boa alternativa para a classificação de clientes.

Abstract

The objective of this work was to propose a risk score based on the ordinal logistic regression model for classifying credit applicants. One of the main advantages of considering responses with three or more levels of risk is being able to define two or more cut-off points for the risk classification. This allows controlling the two main mistakes made in this type of modeling: false negative and false positive rate. The proposed methodology was illustrated through a set of data obtained in the literature and the results showed that the proposed risk score is useful for classifying customers, presenting an overall hit rate of 70% by limiting false negatives and false positives at 10% and 20%, respectively. Thus, the risk score proposed in this work proved to be a good alternative for classifying customers.

Lista de Tabelas

Lista de Figuras

Sumário

1 Introdução

A concessão de crédito é um fenômeno bastante recorrente em economias desenvolvidas. A alocação desse capital em empreendimentos rentáveis estimula o crescimento da economia de um país, uma vez que produtos e serviços são criados para atender as demandas do mercado. Contudo, credores se preocupam com o risco envolvido nessa operação, visto que os clientes podem não conseguir honrar o pagamento de seus empréstimos, por diversos motivos. Sendo assim, é de interesse dessas instituições financeiras avaliar o risco associado ao cliente, antes de fazer a concessão do crédito, para evitar casos de inadimplência e, assim, tornar o negócio lucrativo. Sob esse cenário, surgiram os Modelos de *Credit Scoring*, como ferramenta capaz de quantificar o risco de crédito envolvido em uma operação.

Historicamente, os modelos de *Credit Scoring* foram iniciados pelos estudos de Durand (1941), que reconheceu que a Análise Discriminante poderia ser utilizada para separar bons e maus empréstimos. Na década de 80, escores de riscos foram desenvolvidos por meio da Regressão Logística e desde então têm se tornado populares e, atualmente, são largamente utilizados. Em geral, os escores de risco são obtidos por meio de modelos que consideram um desfecho dicotômico (o cliente é classificado como bom ou mal pagador, por exemplo). No entanto, existem situações em que o desfecho do estudo pode ter três ou mais categorias que seguem uma ordenação de gravidade (ou risco). Exemplo do banco de dados (*German Credit*) onde o *status* da conta é ordenado em 5 categorias (Sem histórico de dívidas - Sem dívidas atuais - Pagamentos em dia - Pagamentos atrasados - Conta crítica).

Neste contexto, o objetivo deste trabalho é propor um escore de risco com base no modelo de regressão logística ordinal e utilizá-lo para a classificação de clientes. Mais especificamente, este trabalho apresenta uma revisão dos modelos de regressão logística binária e multinomial (nominal e ordinal), propor um escore de risco para classificar bons e maus clientes e verificar a sua acurácia por meio de dados simulados. Por fim, a metodologia apresentada neste trabalho será ilustrada em um conjunto de dados sobre *status* de contas de crédito na Alemanha (*German credit*).

2 Metodologia

2.1 Regressão logística clássica

A regressão logística busca, a parti de um modelo matemático, explicar a relação entre uma variável resposta Y e uma ou mais variáveis preditora. Considerando a Regressão Logística Clássica, a variável resposta assume somente dois valores qualitativos, comumente denotados por "sucesso" e "fracasso", sendo representada por uma variável indicadora binária de valores 0 e 1.

Sendo assim, tem-se Y, uma variável binaria com distribuição de Bernoulli com probabilidades $P(Y = 1) = \pi$ de sucesso e $P(Y = 0) = (1 - \pi)$ de fracasso. Sua média é $E(Y) = \pi$. Seja $\mathbf{X}$ um vetor de variáveis explicativas $(X_1, X_2, ..., X_k)$, correspondendo as variáveis independentes do estudo. Agora, $\pi = P(Y = 1|\mathbf{x})$ denota a probabilidade de sucesso para os valores específicos das variáveis explicativas.

Considerando os modelos de regressão linear, $\mu = E(Y|x)$ é uma função linear de $\mathbf{X}$, em que $0 < E(Y|x) < 1$. Para uma resposta binária, um modelo análogo seria:

$$E(Y_i|\mathbf{x}_i) = \pi_i = \alpha_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik}. \tag{2.1.1}$$

ou simplesmente:

$$E(Y_i|\mathbf{x}_i) = \pi_i = \mathbf{x}_i'\beta. \tag{2.1.2}$$

com $\beta' = [\alpha_0 \quad \beta_1 \quad ... \quad \beta_k]$ sendo o vetor de coeficientes de regressão a serem estimados e $\mathbf{x}_i = (1, x_{i1}, x_{i2}, ..., x_{ik})$ é o vetor de variáveis explicativas do i-ésimo individuo da amostra , $i = 1, ..., n$.

Entretanto, considerar o modelo de regressão linear quando a variável resposta é binaria traz os seguintes problemas (KUTNER, 2005)

- **Não normalidade dos erros**
  Como o erro $\epsilon_i = Y_i - (\alpha_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik})$ pode assumir somente dois valores:

  $$Y_i = 1 : \epsilon_i = 1 - (\alpha_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik}).$$

  $$Y_i = 0 : \epsilon_i = -(\alpha_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik}).$$

  Assim, o modelo de regressão linear não é apropriado, por supor normalidade dos erros.

- **Heterocedasticidade**

Levando em consideração Y_i que tem uma distribuição de Bernoulli e os erros $\epsilon_i = (Y_i - \pi_i)$ tem que :

$$Var(Y_i) = \pi_i(1 - \pi_i) = E(Y_i)[1 - E(Y_i)].$$

Como π_i é uma constante , tem-se que $Var(Y_i) = Var(\epsilon_i)$. Com isso,

$$Var(\epsilon_i) = E(Y_i)[1 - E(Y_i)].$$

$$= (\alpha_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik})(1 - \alpha_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik}).$$

que depende de cada variável explicativa x_i, consequentemente a variância dos erros dependem de cada nível de $\mathbf{X}$.

- **Restrição no modelo**
  Como o que está sendo modelado são probabilidades, teremos a seguinte restrição para a resposta média do modelo $E(Y_i) = \pi_i$:

  $$0 \leq \pi_i \leq 1.$$

  Funções de resposta lineares como a do modelo de regressão linear podem não atender satisfatoriamente a essa restrição, o que seria uma impropriedade matemática.

Com as dificuldades apontadas acima, usar o modelo de regressão linear para modelar probabilidades não é uma boa escolha. Frequentemente isso implica em transformações que acabam por tornar a resposta esperada pelo modelo em uma função não linear.

Considerando os modelos de regressão não linear. A transformação mais comumente aplicada utiliza a função exponencial para garantir valores compreendidos no intervalo (0, 1). Para uma resposta binária, um modelo análogo seria:

$$E(Y_i|\mathbf{x}_i) = \pi_i = \frac{\exp(\alpha_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik})}{1 + \exp(\alpha_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik})}. \tag{2.1.3}$$

ou simplesmente:

$$E(Y_i|\mathbf{x}_i) = \pi_i = \frac{\exp(\mathbf{x}_i'\beta)}{1 + \exp(\mathbf{x}_i'\beta)}. \tag{2.1.4}$$

com $\beta' = [\alpha_0 \quad \beta_1 \quad ... \quad \beta_k]$ sendo o vetor de coeficientes de regressão a serem estimados e $\mathbf{x}_i = (1, x_{i1}, x_{i2}, ..., x_{ik})$ é o vetor de variáveis explicativas do i-ésimo individuo da amostra, $i = 1, ..., n$. Visto que $\mathbf{x}'\beta$ pode assumir qualquer valor real enquanto π_i está restrito ao intervalo (0, 1), o objetivo da transformação foi atingido.

A transformação de π_i utilizada para a obtenção da forma aditiva é chamada de transformação logito e é definida da seguinte forma:

$$logit(\pi_i) = log\left(\frac{\pi_i}{1-\pi_i}\right) = \mathbf{x}_i'\beta, \quad i = 1, ..., n. \tag{2.1.5}$$

A transformação em (2.1.5) é contínua, linear nos seus parâmetros e pode assumir qualquer valor real. Dessa forma, várias propriedades assumidas por um modelo de regressão linear são satisfeitas.

Um dos fatores que tornaram a regressão logística mais interessante e contribuiu para sua popularização é a interpretação simples e útil dos coeficientes do modelo Kutner (2005). A interpretação é baseada na razão de chances ($\frac{\pi_i}{1-\pi_i}$), que é a divisão da probabilidade de sucesso pela probabilidade de fracasso. Com isso, um aumento u em alguma variável explicativa x_k ocasionará que a chance de sucesso estimado anteriormente a este incremento seja multiplicado por $\exp(\beta_k)^u$, mantidas as demais variáveis explicativas do modelo constantes.

A estimação dos parâmetros do modelo é feita pelo método de Máxima Verossimilhança. Assumindo que cada uma das n respostas da amostra é uma variável de Bernoulli independente, podemos representar sua distribuição de probabilidade como:

$$P(Y_i = 1|\mathbf{x}_i) = \pi_i.$$

$$P(Y_i = 0|\mathbf{x}_i) = 1 - \pi_i.$$

$$P_{Y_i}(y_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i} \quad y_i = 0, 1; \quad i = 1, ..., n. \tag{2.1.6}$$

Como as n observações Y_i são independentes, a função de probabilidade conjunta é dada por:

$$P(y_1, ..., y_n) = \prod_{i=1}^{n} P_{Y_i}(y_i) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i}. \tag{2.1.7}$$

Substituindo π_i pela equação (2.1.4) obtemos a expressão da função de verossimilhança (KUTNER, 2005):

$$L(\beta|y, \mathbf{X}) = \prod_{i=1}^{n} \left(\frac{exp(\mathbf{x}_i'\beta)}{1+exp(\mathbf{x}_i'\beta)}\right)^{y_i} \left(\frac{1}{1+exp(\mathbf{x}_i'\beta)}\right)^{1-y_i}. \tag{2.1.8}$$

As estimativas de máxima verossimilhança dos parâmetros são os valores que maximizam a função (2.1.8). Não há formas analíticas fechadas para o cálculo dessas estimativas, devendo ser utilizado algoritmos numéricos para tal fim, como por exemplo

o de Newton-Raphson (KUTNER, 2005).

Para testar a significância dos parâmetros $\beta_j,\ j = 1, 2, ..., k$ estimados do modelo, com as hipóteses de

$$H_0 : \beta_j = 0.$$

$$H_1 : \beta_j \neq 0.$$

São realizados os testes de Wald e o de Razão de Verossimilhança. A estatística do teste de Wald é dada por Agresti (2019):

$$Z = \frac{\hat{\beta}}{SE(\hat{\beta})};\quad Z \sim N(0,1). \tag{2.1.9}$$

Sendo $SE(\hat{\beta})$ erro padrão de $\hat{\beta}$ e Z com aproximação para a Distribuição Normal (0,1). Equivalentemente pode-se usar Z^2 com a aproximação da distribuição Qui-quadrado(χ^2_1) com 1 grau de liberdade.

O teste de Razão de verissimilhança é dado por Agresti (2019):

$$Q_L = 2log(\ell_1/\ell_0) = 2\left[log(\ell_1) - log(\ell_0)\right] = 2(L_1 - L_0);\quad Q_L \sim \chi^2_1. \tag{2.1.10}$$

Sendo L_0 o máximo da função de log-verossimilhança sob H_0 e L_1 o máximo da função log-verossimilhança sob H_1. Com a aproximação de Q_L para a distribuição Qui-quadrado(χ^2_1) com 1 grau de liberdade.

Para testa o ajuste do modelo é feito o teste de Hosmer e Lemeshow, que avalia o modelo ajustado comparando as frequências observadas e as esperadas, propondo dois tipos de agrupamento que se baseam nas probabilidades estimadas.

Primeiramente, as observações são classificadas e os eventos de probabilidade são estimados. As observações são, então, divididos em cerca de 10 grupos. Seja N o número total de indivíduos e M o número de alvo de indivíduos para cada grupo e dada por:

$$M \cong [0.1N + 0.5].$$

O número de grupos pode ser menor do que 10 se houver menos do que 10 padrões de variáveis explicativas. Devendo haver pelo menos três grupos mínimos para que a estatística de Hosmer-Lemeshow possa ser determinada.

A estatística proposta, por meio de simulação, segue uma distribuição Qui-quadrado quando não ha replicação em qualquer uma das subpopulações e é dada por Hosmer David W.; Lemeshow Stanley; Sturdivant (2013):

$$H = \sum_{g=1}^{G} \frac{(O_g - N_g \pi g)^2}{N_g \pi_g (1 - \pi_g)}. \tag{2.1.11}$$

Onde:

- N_g é a frequência total de indivíduos no g-ésimo grupo $g = 1, 2, ..., G$.
- O_g é a frequência total de resultados de eventos no g-ésimo grupo .
- π_g é a probabilidade média estimada previsto de um resultado de eventos para o *g-ésimo* grupo.

A estatística de Hosmer-Lemeshow é comparada com uma distribuição Qui-quadrado com $(G - 2)$ graus de liberdade. Maiores valores da estatística do teste em relação ao p-valor indicam uma falta de ajuste do modelo.

A informação de *Akaike* (AIC) é uma estatística que tem por base o logaritmo de verossimilhança e é uma maneira de selecionar um modelo de um conjunto de modelos (AKAIKE, 1974) e é definida por

$$AIC = -2[\log(L) - \mu], \tag{2.1.12}$$

em que μ é o número de parâmetros do modelo e L o valor da verossimilhança para o modelo estimado. O AIC permite a comparação de modelos alinhados ou não. Quanto menor for seu valor, menor será a informação perdida e, por consequência, melhor será o ajustamento do modelo.

Quanto aos métodos interativos de seleção de variáveis, existem 3 que se destacam (SISCU, 2010):

- *Forward selection*: 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 (backward)*: este método é uma mescla das duas técnicas anteriores. As variáveis são gradativamente excluídas do modelo. Após a exclusão de uma nova variável, é verificado se variáveis excluídas anteriormente podem ser incluídas devido à entrada da nova variável. Este é o método mais utilizado de seleção de variáveis.

2.2 Regressão Logística Multi-Categórica

Como visto anteriormente a regressão logística é usualmente usada para modelar a relação entre uma variável resposta dicotômica e um conjunto de variáveis independentes. Entretanto em muitas situações a variável dependente possui mais de dois níveis e até pode existir uma ordenação entre os níveis. Nesse contexto, uma das abordagens é usar uma generalização da regressão logística, chamada de regressão logística multi-categórica.

2.2.1 Regressão Nominal Clássica

Seja J o número de categorias de uma variável resposta Y, e $[\pi_1, \pi_2, ..., \pi_J]$ as respectivas probabilidades, satisfazendo $\sum_{j=1}^{J} \pi_j = 1$. Com n observações independentes, a probabilidade de todas as formas que essas n observações podem se associar às J categorias pode ser especificada por uma distribuição de probabilidade multinomial (AGRESTI, 2019).

A i-ésima observação pode ser escrita a partir de J variáveis respostas binárias, $Y_{i1}, ..., Y_{iJ}$, onde:

$$Y_{ij} = \begin{cases} 1, & \textit{se a i-ésima resposta está na categoria j} \\ 0, & \textit{caso contrário} \end{cases}$$

Visto que somente uma categoria pode ser selecionada para a i-ésima variável resposta, teremos:

$$\sum_{j=1}^{J} Y_{ij} = 1 \quad i = 1, 2, ..., n.$$

Considerando a categoria J como (nível) de referência. A probabilidade, π_{ij}, da categoria j ser selecionada para a i-ésima resposta é dada por:

$$\pi_{ij} = P(Y_{ij} = 1) = \frac{\exp(\mathbf{x}_i' \beta_j)}{1 + \sum_{j=1}^{J-1} \exp(\mathbf{x}_i' \beta_j)}. \tag{2.2.1}$$

Para o caso da regressão logística binária, tem-se que $J = 2$. Se denotarmos $Y_i = 1$ se a i-ésima resposta for da Categoria 1, e $Y_i = 0$ se a i-ésima resposta for da Categoria 2, então:

$$\pi_i = \pi_{i1} \quad e \quad 1 - \pi_i = \pi_{i2}.$$

Para a regressão logística binária, o *Logit* de π_i é modelado utilizando um preditor linear. Como há somente 2 categorias na regressão logística binária, o *Logit* de fato compara a probabilidade da resposta ser da Categoria 1 com a probabilidade da resposta

ser da Categoria 2:

$$logit(\pi_i) = log\left[\frac{\pi_i}{1-\pi_i}\right] = log\left[\frac{\pi_{i1}}{\pi_{i2}}\right] = logit(\pi_{i1,2}) = \mathbf{x}_i'\beta_{1,2}.$$

Generalizando para as J categorias, tem-se a combinação de $\binom{J}{2} = J(J-1)/2$ pares de categorias para comparação, e consequentemente $J(J-1)/2$ preditores lineares. Por exemplo com 3 categorias, $J = 3$, resulta em 3 comparações e consequentemente 3 modelos de regressão logística a serem estimados:

$$logit(\pi_{i1,2}) = \log\left[\frac{\pi_{i1}}{\pi_{i2}}\right] = \mathbf{x}_i'\beta_{1,2},$$

$$logit(\pi_{i1,3}) = \log\left[\frac{\pi_{i1}}{\pi_{i3}}\right] = \mathbf{x}_i'\beta_{1,3} \quad \text{e}$$

$$logit(\pi_{i2,3}) = \log\left[\frac{\pi_{i2}}{\pi_{i3}}\right] = \mathbf{x}_i'\beta_{2,3}.$$

Ou seja, os modelos logísticos multi-categóricos nominais utilizam simultaneamente todos os pares de categorias, especificando a chance da resposta estar em uma categoria em comparação a outra.

No entanto, não é necessário desenvolver todos os $J(J-1)/2$ modelos logísticos. Na prática, uma categoria é escolhida como base e então todas as demais categorias são comparadas a ela. A escolha da categoria base, também chamada de categoria (nível) de referência, é arbitrária, (AGRESTI, 2019). Por exemplo, utilizando a categoria J como referência, é necessário considerar apenas as $J-1$ comparações a essa categoria. O *Logit* para a j-ésima comparação é (KUTNER, 2005):

$$logit(\pi_{ijJ}) = \log\left[\frac{\pi_{ij}}{\pi_{iJ}}\right] = \mathbf{x}_i'\beta_{jJ} \quad j = 1, 2, ..., J-1. \tag{2.2.2}$$

Como todas as comparações são feitas com a categoria J, a equação acima pode ser escrita com $\pi_{ij,J} = \pi_{ij}$ e $\beta_{jJ} = \beta_j$ ou seja:

$$logit(\pi_{ij}) = \log\left[\frac{\pi_{ij}}{\pi_{iJ}}\right] = \mathbf{x}_i'\beta_j \quad j = 1, 2, ..., J-1. \tag{2.2.3}$$

O motivo para se considerar apenas os $J-1$ *Logit* está no fato de que qualquer outro *Logit* pode ser obtido através deles (KUTNER, 2005). Por exemplo, a comparação das Categorias 1 e 2 (com a terceira categoria sendo a de referência) pode ser obtida por:

$$\log\left[\frac{\pi_{i1}}{\pi_{i2}}\right] = \log\left[\frac{\pi_{i1}}{\pi_{i3}} \times \frac{\pi_{i3}}{\pi_{i2}}\right].$$

$$= \log\left[\frac{\pi_{i1}}{\pi_{i3}}\right] - \log\left[\frac{\pi_{i2}}{\pi_{i3}}\right].$$
$$= \mathbf{x}_i'\beta_1 - \mathbf{x}_i'\beta_2.$$

Generalizando, para comparar as categorias f e p temos:

$$\log\left[\frac{\pi_{if}}{\pi_{ip}}\right] = \mathbf{x}_i'(\beta_f - \beta_p). \tag{2.2.4}$$

Os coeficientes do Modelo Nominal (compostas pelos $J-1$ modelos logísticos) podem ser interpretados da mesma maneira dos coeficientes da Regressão Logística Clássica, por meio da razão de chances. Deve-se ficar atento apenas em relação as parâmetros que estão sendo analisados e suas categorias (AGRESTI, 2013).

Dadas as $J-1$ expressões logísticas em (2.2.3), obtêm-se as $J-1$ expressões diretas para as probabilidades de cada categoria em temo dos $J-1$ preditores lineares $\mathbf{x}'\beta_j$. As expressões resultantes são (KUTNER, 2005):

$$\pi_{ij} = \frac{\exp(\mathbf{x}_i'\beta_j)}{1 + \sum_{j=1}^{J-1}\exp(\mathbf{x}_i'\beta_j)}. \tag{2.2.5}$$

A estimação dos $J-1$ vetores de parâmetros β_j, com $j = 1, ..., J-1$, é feita simultaneamente pelo método de máxima verossimilhança. Para isso é necessário obter a função de verossimilhança dos dados.

A fim de explicação, suponha que há $J = 3$ categorias e a terceira categoria é a escolhida para a i-ésima resposta. Consequentemente, para o i-ésimo caso teremos:

$$Y_{i1} = 0, \quad Y_{i2} = 0 \text{ e } \ Y_{i3} = 1.$$

E a probabilidade dessa resposta é :

$$P(Y_{i1} = 0, Y_{i2} = 0, Y_{i3} = 1) = \pi_{i3} = [\pi_{i1}]^0 \times [\pi_{i2}]^0 \times [\pi_{i3}]^1$$
$$= \prod_{j=1}^{3}[\pi_{ij}]^{y_{ij}}.$$

Assim, para n observações independentes e J categorias, a função de probabilidade conjunta é dada por Kutner (2005):

$$P(y_1, ..., y_n) = \prod_{i=1}^{n} P_{Y_i}(y_i) = \prod_{i=1}^{n}\left[\prod_{j=1}^{J}[\pi_{ij}]^{y_{ij}}\right]. \tag{2.2.6}$$

Como $\pi_{iJ} = 1 - \sum_{j=1}^{J-1} \pi_{ij}$ e $y_{iJ} = 1 - \sum_{j=1}^{J-1} y_{ij}$, a expressão fica:

$$P(y_1, ..., y_n) = \prod_{i=1}^{n} P_{Y_i}(y_i) = \prod_{i=1}^{n} \left[\left(\prod_{j=1}^{J-1} [\pi_{ij}]^{y_{ij}} \right) \left(\left[1 - \sum_{j=1}^{J-1} \pi_{ij} \right]^{1-\sum_{j=1}^{J-1} y_{ij}} \right) \right].$$

Substituindo π_{ij} por (2.2.5), obtém-se a expressão da função de verossimilhança desejada, considerando J como referência Kutner (2005):

$$P(y_1, ..., y_n | \beta_1, \beta_2, ..., \beta_{J-1}) =$$

$$= \prod_{i=1}^{n} \left[\prod_{j=1}^{J-1} \left(\frac{\exp(\mathbf{x}_i' \beta_j)}{1 + \sum_{k=1}^{J-1} \exp(\mathbf{x}_i' \beta_k)} \right)^{y_{ij}} \right] \left(\frac{1}{1 + \sum_{k=1}^{J-1} \exp(\mathbf{x}_i' \beta_k)} \right)^{1-\sum_{j=1}^{J-1} y_{ij}}, \qquad (2.2.7)$$

em que $\beta = [\alpha_0, \beta_1, ..., \beta_k]$ são os coeficientes de regressão a serem estimados e $x_i = (1, x_{i1}, ..., x_{ik})$ é o vetor de covariáveis do indivíduo i.

As estimativas dos parâmetros em β são os valores que maximizam essa função. Não existem formas analíticas fechadas para o cálculo dessas estimativas, devendo ser utilizado algoritmos numéricos como por exemplo o de Newton-Raphson (KUTNER, 2005).

2.2.2 Regressão Logística Ordinal Clássica

Variáveis categóricas ordinais são importantes em muitas áreas de estudo, principalmente em situações em que medidas exatas não são possíveis. Com isso a regressão ordinal clássica é utilizada para modelar a relação entre uma variável resposta com níveis ordenados entre si com as variáveis independentes.

As variáveis ordinais podem ser originadas de formas diferentes: variáveis contínuas agrupadas e variáveis categóricas naturalmente ordenadas. A primeira forma é pela categorização de uma variável contínua. Por exemplo, o tempo de relacionamento de um cliente bancário pode ser medido de forma ordinal por meio da categorização: “menos de 2 anos”, “2 a 5 anos” e “mais de 5 anos”.

A segunda forma consiste na avaliação de uma informação não quantificável, casos em que uma medida precisa nem sempre é possível, associada a níveis de uma escala ordinal, realizando coleção de categorias naturalmente ordenadas. Como ilustração, temos o risco de ocorrência de inadimplência, que pode ser classificado como “alto”, “médio” ou “baixo”.

Uma variável categoria é referida como ordinal ao invés de intervalar quando há uma ordem clara das categorias, mas as distancias absolutas entre elas são desconhecidas. Assim, para muitas variáveis categorias ordinais, é sensato imaginar a existência de uma

variável continua subjacente. Para se aproximar da escola subjacente é frequentemente útil associar um conjunto "razoável" de scores às categorias.

2.2.3 Modelos Cumulativos

A relação de ordem entre as classes das variáveis dependente faz com que a tarefa de modelar a probabilidade de ocorrência de uma das suas classes seja feita em termos de probabilidade acumuladas (AGRESTI, 2019). A probabilidade cumulativas de uma variável Y é a probabilidade de Y assumir valores iguais ou menores que um determinado ponto.

Seja Y uma variável ordinal com J classes, a probabilidade de se observar uma classe inferior ou igual a j, para um determinado vetor de observações das variáveis independentes $\mathbf{x}$, é:

$$P(Y \leq j|\mathbf{x}) = \pi_1 + ... + \pi_j, \quad j = 1, ..., J; \tag{2.2.8}$$

com $\pi_1 = P(Y = 1|\mathbf{x})$, $\pi_2 = P(Y = 2|\mathbf{x})$, ... , $\pi_J = P(Y = J|\mathbf{x})$. Consequentemente como as classes são ordenáveis, as probabilidades acumuladas refletem a ordenação natural $P(Y \leq 1|\mathbf{x}) \leq P(Y \leq 2|\mathbf{x}) \leq ... \leq P(Y \leq J - 1|\mathbf{x})$. Modelos para probabilidades cumulativas não utilizam a última categoria, $P(Y \leq J|\mathbf{x})$ visto que ela é necessariamente igual a 1 (informação referente à ultima classe é redundante).

A variável ordinal pode ser interpretada como a operacionalização de uma outra variável continua não medida (latente), como vimos anteriormente. Assim, a variável manifesta(Y) resulta do "corte" da variável latente (Y^*) em J classes ordinais e mutuamente exclusivas.

Suponha $-\infty = \alpha_0 \leq \alpha_1 \leq ... \leq \alpha_J = \infty$ os pontos de corte da escala contínua de Y^*, dado que a variável resposta observada satisfaz:

$$Y = j \;\text{ se }\; \alpha_{j-1} \leq Y^* \leq \alpha_j, \quad j = 1, 2, ..., J. \tag{2.2.9}$$

Em outras palavras, observa-se Y na categoria j quando a variável latente cair no j-ésimo intervalo de valores.

Suponha agora que a varável latente de Y^* é determinada pelas variáveis explicativas de forma linear:

$$Y^* = \mathbf{x}'\beta + \epsilon, \tag{2.2.10}$$

onde $\beta = (\beta_1, ..., \beta_k)$ é o vetor de parâmetros e ϵ é uma variável aleatória com distribuição

F.

Desses fatos, segue que a distribuição de probabilidade da variável observada de Y é dada por:

$$P(Y \leq j|\mathbf{x}) = F(\alpha_j - \mathbf{x}'\beta). \tag{2.2.11}$$

A demonstração da equação (2.2.11) é dada por:

$P(Y \leq j|x) = P(Y = 1|x) + P(Y = 2|x) + ... + P(Y = j|x) = P(\alpha_0 \leq Y^* \leq \alpha_1|x) + P(\alpha_1 \leq Y^* \leq \alpha_2|x) + ... + P(\alpha_{j-1} \leq Y^* \leq \alpha_j|x) = F_{Y^*|x}(\alpha_j) - F_{Y^*|x}(\alpha_0)$

Como $\alpha_0 = -\infty$, temos $F_{Y^*|x}(\alpha_0) = 0$. Assim $F_{Y^*|x}(\alpha_j) = P(\mathbf{x}'\beta + \epsilon \leq \alpha_j) = P(\epsilon \leq \alpha_j - \mathbf{x}'\beta) = F(\alpha_j - \mathbf{x}'\beta)$ onde F é a função de distribuição da variável aleatória ϵ.

O inverso da função F, isto é , F^{-1} é designada função de ligação ($Link$), por faze a associação linear entre a parte aleatória do modelo, $P(Y \leq k)$, e a parte sistemática ($\mathbf{x}'\beta$). Ou seja,

$$Link(P(Y \leq j)) = \alpha_j - \mathbf{x}'\beta. \tag{2.2.12}$$

Várias são as opções para ser usar como função de ligação, cujo uso no modelo ordinal é recomendável de acordo com o tipo de distribuição de probabilidade que as classes da variável dependente apresentam. Esta escolha deve ser feita com cuidado, pois uma escolha inapropriada pode comprometer a significância do modelo e sua capacidade preditiva Agresti (2019). As cinco principais funções de ligação estão descritas na Tabela 1 (AGRESTI, 2013).

Tabela 1: Principais funções de ligação para um modelo de regressão ordinal

Nome	Função Link (F^{-1})
Logit	$\log\left[\frac{P(Y\leq j)}{P(Y\geq j)}\right]$
Complemento Log-log	$\log(-\log(1 - P(Y \leq j))$
Log-log negativo	$-\log(-\log(P(Y \leq j)))$
Cauchit	$\tan(\pi(P(Y \leq j) - 0.5))$
Probit	$\phi^{-1}(P(Y \leq j))$, onde ϕ é a função de distribuição da N(0,1)

Na prática, a função de ligação *Logit* é a mais utilizada, devido a sua interpretação interessante dos coeficientes do modelo e da sua matemática simples. Onde sua interpretação se da, no caso binário, pela razão de chances de uma categoria , comumente chamada de sucesso, se da em relação a outra, frequentemente chamada de fracasso. O que pode ser generalizado para o caso onde possuir mais de 2 categorias da variável resposta.

Essa será a abordagem explorada nesse trabalho.

O modelo é proposto através de uma analogia com a regressão logística usual, de forma que o *Logit* das probabilidades cumulativas são (KUTNER, 2005):

$$Logit[P(Y_i \leq j|\mathbf{x})] = \log\left[\frac{P(Y_i \leq j|\mathbf{x})}{1 - P(Y_i \leq j|\mathbf{x})}\right] = \alpha_j - \beta_1 x_{i1} - ... - \beta_k x_{ik} \quad , j = 1, ..., J-1. \tag{2.2.13}$$

Consequentemente,

$$P(Y_i \leq j|\mathbf{x}) = \frac{\exp(\alpha_j - \mathbf{x}_i'\beta)}{1 + \exp(\alpha_j - \mathbf{x}_i'\beta)} \quad , j = 1, ..., J-1. \tag{2.2.14}$$

O modelo ordinal acima definido permite estimar o logaritmo da probabilidade de a variável resposta tomar os valores de classes inferiores ou iguais a j, comparativamente com a probabilidade de tomar os valores das classes superiores a j.

Para exemplificar, considere $J = 4$, o modelo usa $Logit[P(Y_i \leq 1|\mathbf{x})] = \log[\pi_1/(\pi_2 + \pi_3 + \pi_4)]$ e $Logit[P(Y_i \leq 2|\mathbf{x})] = \log[(\pi_1 + \pi_2)/(\pi_3 + \pi_4)]$ e assim sucessivamente. Cada logito cumulativo utiliza todas as categorias da variável resposta.

Note que os coeficiente de regressão $\beta = (\beta_1, ..., \beta_k)$ em (2.2.13) não apresentam índice j, obrigando o modelo a pressupor que os efeitos das variáveis independentes sobre o $P(Y_i \leq j)$ é igual para todas as classes Kutner (2005). Assim a resposta observada em cada classe apenas se encontra para a direita ou para a esquerda, em função de α_j. Isso resulta em um modelo mais parcimonioso. Para um $\beta_k > 0$, um aumento em algum X_k, mantendo as demais variáveis explicativas constantes, resulta na diminuição da probabilidade de a variável resposta tomar valores de ordem inferiores ou iguais a j ou seja, quando X_k aumenta, Y aumenta. Já para um $\beta_k < 0$, quando X_k aumenta, Y diminui.

A interpretação do modelo pode usar as razões de chances para as probabilidades cumulativas e os seus complementos. Para dois valores de x_1 e x_2 de uma das variáveis explicativas $\mathbf{X}_k$ do estudo, a razão de chances comparando as probabilidades cumulativas, para todas as classes da variável dependente é dada por (mantendo as demais variáveis explicativas constantes):

$$\frac{P(Y \leq j|X_k = x_2)/P(Y > j|X_k = x_2)}{P(Y \leq j|X_k = x_1)/P(Y > j|X_k = x_1)}. \tag{2.2.15}$$

O logaritmo dessa razão de chances é a diferença entre os logitos cumulativos para esses dois valores de X_k. Isso é igual a $-\beta_k(x_2 - x_1)$. Se $x_2 - x_1 = 1$, a chance da variável resposta assumir valores menores para qualquer categoria é multiplicado por $e^{-\beta_k}$

para cada unidade acrescida em X_k .

Os parâmetros do modelo $(\alpha_1, ..., \alpha_{J-1})$ e β são estimados simultaneamente pelo método de Máxima Verossimilhança. Para isso, é necessário a obtenção da função de verossimilhança para os dados, lembrando que o modelo pressupõe que as curvas de probabilidade das $J-1$ classes da variável resposta são iguais para todas as classes e são calculadas de forma cumulativa.

Se baseando de (2.2.6), para n observações independentes e J categorias, a função de verossimilhança é dada por (AGRESTI, 2013):

$$P(y_1, .., y_n|\alpha_1, .., \alpha_{J-1}, \beta) = \prod_{i=1}^{n} P_{Y_i}(y_i) = \prod_{i=1}^{n} \left[\prod_{j=1}^{J} [\pi_{ij}]^{y_{ij}} \right] = \prod_{i=1}^{n} \left[\prod_{j=1}^{J} [P(Y_i \leq j|x) - P(Y_i \leq j-1|x)]^{y_{ij}} \right]. \tag{2.2.16}$$

Visto que $P(Y_i \leq J|x) = 1$, $P(Y_i \leq 0|x) = 0$ e $P(Y_i \leq j|x) = \frac{\exp(\alpha_1 - x_i'\beta)}{1+\exp(\alpha_1 - x_i'\beta)}$, $j = 1, ..., J-1$ por (2.2.14) encontra-se a expressão desejada da função de verossimilhança, em termos de $\alpha_1, ..., \alpha_{J-1}$ e β:

$$P(y_1, ..., y_n|\alpha_1, ..., \alpha_{J-1}, \beta) =$$

$$\prod_{i=1}^{n} \left[\left(\frac{\exp(\alpha_1 - x_i'\beta)}{1+\exp(\alpha_1 - x_i'\beta)} \right)^{y_{i1}} \left(\prod_{j=2}^{J-1} \left(\frac{\exp(\alpha_j - x_i'\beta)}{1+\exp(\alpha_j - x_i'\beta)} - \frac{\exp(\alpha_{j-1} - x_i'\beta)}{1+\exp(\alpha_{j-1} - x_i'\beta)} \right)^{y_{ij}} \right) \left(\frac{1}{1+\exp(\alpha_{J-1} - x_i'\beta)} \right)^{y_{iJ}} \right]. \tag{2.2.17}$$

As estimativas de máxima verossimilhança são os valores dos parâmetros que maximizam (2.2.17). Não existem formas analíticas fechadas para o cálculo dessas estimativas, devendo ser utilizado algoritmos numéricos como por exemplo o de Newton-Raphson (KUTNER, 2005).

3 Modelagem de risco de crédito

3.1 Introdução

No decorrer de toda a história de desenvolvimento econômico e social das sociedades, o crédito é um dos principais fatores a serem considerados, por ser uma forma que os agentes sociais de diferentes meios pudessem de alguma forma ter uma expansão econômica.

Sob essa perspectiva financeira:

> O crédito corresponde a um valor monetário disponibilizado ao tomador de recursos financeiros, em forma de empréstimo ou financiamento, por um período previamente pactuado, com a promessa de pagamento futuro, ao qual é acrescido uma remuneração, denominada juros. Consequentemente o risco é inerente ao processo de concessão de crédito, uma vez que existem incertezas quanto ao futuro das quantias emprestadas (MACHADO, 2015).

Segundo Santos (2011), "o risco é definido pela incerteza de retorno de um investimento perante a possibilidade de um evento possível, futuro e incerto, autônomo à vontade do investidor e cuja ocorrência poderá causar prejuízos". Nesse sentido, o risco de crédito está ligado a fatores internos e externos ao concessor que podem prejudicar a recuperação do montante emprestado. Para o Banco Central do Brasil, conforme Art. 2o. da Resolução 3.721/2009, risco de crédito é definido como a possibilidade de ocorrência de perdas associadas ao não cumprimento pelo tomador ou contraparte de suas respectivas obrigações financeiras nos termos pactuados, à desvalorização de contrato de crédito decorrente da deterioração na classificação de risco do tomador, à redução de ganhos ou remunerações, às vantagens concedidas na renegociação e aos custos de recuperação.

E dentro da esfera do risco, existem diversos aspectos a serem analisados, como

- Risco do Cliente - associado aos C's do Crédito:
    1. Capacidade-habilidade em pagar. Diz respeito aos meios financeiros para honrar com os compromissos assumidos;
    2. Colateral-garantia;
    3. Caráter-confiabilidade e "vontade" de pagar;
    4. Condição-condições ambientais externas, internas e indicadores econômicos;
    5. Capital-reservas e patrimônio.
- Risco da Operação - envolve características do produto, prazo, formas de pagamento, garantia e preço;
- Risco de Carteira - relacionado ao conjunto de clientes e tipos de negócios ;
- Risco de Administração de Crédito - compreende o acompanhamento do crédito concedido

Nesse cenário, surgiram os Modelos de *Credit Scoring*, como ferramenta capaz de quantificar o risco de credito envolvido em uma operação de forma automatizada, padrozinada e objetiva.

Os Modelos de *Credit Scoring* (CS) utilizam-se de algorítimos matemáticos e técnicas estatísticas para calcular a probabilidade de que determinado evento aconteça. Aplicando fórmulas, o sistema atribui pontuação específica para cada característica do proponente/cliente para prever um resultado.

Historicamente, os modelos de *Credit Scoring* foram iniciados pelos estudos de Durand (1941) na área de financiamento ao consumidor após a Grande depressão nos EUA. O projeto foi pioneiro na utilização da Estatística como ferramenta para análise de risco de crédito. Nesse projeto, foi utilizada a Análise de Discriminante desenvolvida por Fisher(1936) para identificar bons e maus empréstimos. Nesse contexto, a pesquisa de Durand pode ser considerada como o ponto de partida para futuros estudos que considerem o desenvolvimento de metodologia de suporte à concessão de crédito.

No início dos anos 1950, Bill Fair e Earl Isaac criaram a primeira Companhia para consultoria em métodos de *scoring*, onde utilizavam dados históricos para melhorar as decisões nos negócios. Posteriormente, em 1958 foi vendido o primeiro Sistema de *Credit Scoring* para a área de Cartão de Crédito. Esse fato é considerado o segundo passo importante para a história dos modelos de *scoring*. Entretanto, o sucesso da Companhia e sua finalidade comercial não implicaram em desenvolvimento de literatura sobre o tema, uma vez que o conhecimento tornou-se valioso e pouco exibido pelas empresas.

Apesar dos Modelos de *Credit Scoring* representarem uma melhoria em relação às análises de risco de crédito, houve dificuldades que impediam o seu crescimento, como relutância dos executivos, limitações tecnológicas, obstáculos no desenvolvimento e implementação dos modelos e, segundo Myers e Forgy (1963), a falta de estatísticos para propagar-se na área de crédito e fazer o trabalho de transformar essa ideia em uma ferramenta operacional bem sucedida e útil. Diante do exposto, apesar do crédito continuar em expansão nos EUA, poucos estudos sobre *Credit Scoring* foram produzidos até os anos 1960.

A partir de 1960, outras pesquisas relevantes foram publicadas, como :

- Desenvolvimento de Sistemas Numéricos de Avaliação de Crédito, (MYERS; FORGY, 1963). Eles se empenharam em verificar a eficácia das fórmulas preditivas de *scoring* e dessa forma introduziram o conceito de amostra *hold-out*, ou seja, diferente daquela utilizada para a modelagem. Esse fato foi importante pois existe a chance de um modelo distinguir bons e maus clientes na base original mas não ser preditivo em outras amostras;

- Conceitos e Utilização de Técnicas de *Credit Scoring*, (WEINGARTNER, 1966). O autor ressaltou a importância de teste antes da utilização dos escores de crédito e sugeriu uma nova técnica de validação: aplicar a fórmula a clientes inadimplentes para verificar se os escores são baixos;

- Índices Financeiros, Análise de Discriminante e Previsão de Falência de Empresas, (ALTMAN, 1968). Introdução dos modelos de *Scoring* para empresas;
- Um modelo de *Credit Scoring* para Empréstimos Comerciais, (ORGLER, 1970). Propôs um modelo para avaliar periodicamente a qualidade dos empréstimos já concedidos.

A partir de 1970, com o aquecimento da economia e consequentemente aumento da demanda de crédito, muitas instituições financeiras nos EUA cresceram de forma insustentável, uma vez que não conseguiam manter uma forma lucrativa. Ao mesmo tempo, a reconstrução da Europa pós Guerra contribuíram para que Modelos de *Credit Scoring* fossem reconhecidos como uma indústria. Desde o inicio dos anos 1990, os Modelos de *CS* tornaram-se o principal mecanismo para avaliação de risco na concessão de vários tipos de empréstimos, sendo as decisões tomadas sem intervenção do responsável pelo pedido da avaliação.

A partir da divulgação do Acordo de Basileia II ocorrida em 2004, os Modelos de *CS* tornaram-se ainda mais importantes, uma vez que o documento destacou a utilização de técnicas que permitam ás instituições e supervisores avaliar corretamente os vários riscos que os bancos enfrentam. Muitas organizações desenvolveram melhores modelos ou modificaram os já existentes para estar em conformidade com as novas regras e com as melhores práticas de mercado, dado que os reguladores forçaram regras mais rigorosas sobre o desenvolvimento, implementação e validação dos modelos internos utilizados para estimar capital a ser provisionado.

Com o contínuo desenvolvimento e crescimento dos mercados financeiros, o crédito tornou-se ainda mais importante na economia. Com a globalização e a sofisticação dos meios de comunicação, como a internet, os consumidores buscam ofertas de crédito mais atrativas. Por isso as instituições buscam desenvolver eficientes ferramentas para avaliar e controlar os riscos.

Modelos de *Credit Scoring*, inicialmente utilizados apenas para decisão de conceder ou não determinado valor ou limite, hoje fazem parte de todo o clico do crédito, estando presente em cada etapa da gestão estratégica de riscos.

3.2 Escore de risco

A mensuração de risco de crédito é o processo de quantificar a credibilidade de um solicitante de crédito, por meio de variáveis explicativas que irão classificar "bons" e "maus" pagadores. Objetivo dessa classificação previa é poder prever comportamentos que possam indicar padrões de inadimplência e assim evitar maiores prejuízos e perda de bons clientes para a instituição financeira.

O modelo de escore de risco será obtido utilizando os resultados de uma regressão logística ordinal com J categorias. Considerando que as categorias 0, 1, 2, ..., $J-1$ estão ordenadas do menor ao maior risco (isto é, a Categoria 0 é a de menor risco e a Categoria $J-1$ é a de maior risco), um escore de risco pode ser definido por Nakano (2010):

$$\text{Escore} = IG_J(P) = \frac{\left(p_1 + 2p_2 + ... + (J-1)p_{(J-1)}\right)}{J-1} \tag{3.2.1}$$

onde, o vetor $P = (p_0, p_1, ..., p_{J-1})$ é o vetor de probabilidades resultante da aplicação de uma regressão logística ordinal em uma unidade amostral e J corresponde número de categorias da variável resposta.

3.3 Classificação dos clientes pelo escore de risco

A classificação dos clientes é um das etapas mais importantes para as instituições financeiras. Com ajuda dessa classificação, que postura e estrategias são tomadas em relação as concessões de credito aos solicitantes. Levando isso em consideração , a classificação dos clientes pelo escore de risco é feito com o objetivo de maximizar os lucros e minimizar os riscos.

Considerando os pontos citados anteriormente, os critérios de classificação dos clientes pelo escore tem os seguintes objetivos :

- Minimizar o erro de classificar maus solicitantes como um risco baixo, assim evitando concessões com risco elevado
- Minimizar o erro de classificar bons solicitantes como um risco alto, assim evitando perda de clientes
- Maximizar o acerto total da classificação dos clientes

3.4 Avaliação da acurácia do modelo

Validação cruzada :

Essa técnica consiste em dividir os dados em duas amostras de mesmo tamanho para serem trabalhados, sendo uma amostra de estimação e outra de validação. A subamostra de construção do modelo é usada para estimação dos parâmetros; já a subamostra de validação, tem a função de validar os parâmetros e verificar o poder de predição do modelo construído. Esse processo avalia quantitativamente a capacidade de previsão do modelo frente a outras observações

Total de acertos:

Corresponde ao numero de classificações certas do modelo para a variável resposta em relação com a variável explicativa.

Sensibilidade:

Corresponde a probabilidade do modelo alocar o indivíduo i na categoria K dado que ele pertence à essa categoria. Exemplo, seja X o *status* de um cliente (1 = mal pagador, 2 = bom pagador) e Y é o diagnóstico do modelo de risco (positivo, quando o modelo classifica o cliente como de alto risco e ; negativo quando, o modelo classifica o cliente como de baixo risco). A sensibilidade será definida pela probabilidade do diagnóstico do modelo acerta a classificação do risco como alto ($Y = +1$) de um mal pagador (AGRESTI, 2019), isto é,

$$Sensibilidade = P(Y = 1|X = 1).$$

Especificidade:

Corresponde a probabilidade do modelo não alocar o indivíduo i na categoria K dado que ele não pertence à essa categoria K. Exemplo, seja X o *status* de um cliente (1 = mal pagador, 2 = bom pagador) e Y é o diagnóstico do modelo de risco (positivo, quando o modelo classifica o cliente como de alto risco e ; negativo quando, o modelo classifica o cliente como de baixo risco). A especificidade será definida pela probabilidade do diagnóstico do modelo classificar como baixo risco ($Y = -1$) um bom pagador (AGRESTI, 2019), isto é,

$$Especificidade = P(Y = -1|X = 2).$$

Falso Positivo:

Corresponde quando o modelo assume categoria sucesso a variável resposta quando a observação pertence a categoria fracasso. Exemplo, seja X o *status* de um cliente (1 = mal pagador, 2 = bom pagador) e Y é o diagnóstico do modelo de risco (positivo, quando o modelo classifica o cliente como de alto risco e ; negativo quando, o modelo classifica o cliente como de baixo risco). Falso positivo seria quando o diagnóstico do modelo de risco classificar como alto risco um bom pagador .

Falso Negativo:

Corresponde quando o modelo assume categoria fracasso a variável resposta quando a observação pertence a categoria sucesso. Exemplo, seja X o estado de um cliente (1 = mal pagador, 2 = bom pagador) e Y é o diagnóstico do modelo de risco (positivo, quando o modelo classifica o cliente como de alto risco e ; negativo quando, o

modelo classifica o cliente como de baixo risco). Falso negativo seria quando o diagnóstico do modelo de risco classificar como baixo risco um mal pagador.

Matriz de confusão:

Corresponde a uma tabela entre os valores reais e os valores preditos pelo modelo, que relata o número de falsos positivos, falso negativos, verdadeiros positivos e verdadeiros negativos. Exemplo, seja X o *status* de um cliente (1 = mal pagador, 2 = bom pagador) e Y é o diagnóstico do modelo de risco (positivo, quando o modelo classifica o cliente como de alto risco e ; negativo quando, o modelo classifica o cliente como de baixo risco). A matriz de confusão seria construída da forma :

Tabela 2: Matriz de confusão

	Valores reais (X)	
Diagnóstico do modelo de risco (Y)	Mal pagador	Bom pagador
Alto risco	Verdadeiro positivo	Falso positivo
Baixo risco	Falso negativo	Verdadeiro negativo

O número total de acertos é dado pela soma da diagonal principal da matriz de confusão.

4 Ilustração da metodologia proposta

4.1 Descrição da aplicação

Este capítulo apresenta a aplicação da metodologia proposta neste trabalho para a classificação de clientes.

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 representam 70% do total de dados. Enquanto, os dados de teste, que representam 30% dos dados, são apresentados após a criação do modelo e usados para simular previsões reais, permitindo que o desempenho real seja verificado e testar o ajuste.

O desenvolvimento do modelo de regressão logística será feito nas seguintes etapas: divisão dos dados em treino e teste; aplicação do modelo logístico com os métodos $Stepwise/Backward/Forward$ de seleção de variáveis nos banco de treino; testar a aplicabilidade e a ajustabilidade do modelo; aplicar o modelo no banco de teste; avaliar as métricas de desempenho.

Os dados utilizados no trabalho serão os *German Credit Data*, disponível pela Universidade da California-Irvin(UCI) em seu repositório: *Machine Learning Repository's*. Foi selecionado esse banco de dados por ser muito usado em trabalhos de análise de risco e é um banco razoável.

Esse conjunto de dados possui 1000 solicitantes de crédito, com 21 variáveis sendo 8 numéricas e 13 categóricas. A Tabela 3 apresenta a descrição de todas as variáveis presentes no banco:

Tabela 3: Dicionário das Variáveis

Variável	**Descrição**	**Tipo**	**Observação:**
Default Status	Status padrão	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
Durantio in month	Duração em meses	Numérica	
Credit History	Histórico de Crédito / Variável resposta	Categórica	A30: Nenhum crédito recebido/ Todos os créditos pagos devidamente A31 : Todos os créditos neste banco foram devidamente pagos A32 : Créditos existentes pagos devidamente até agora A33 : Atraso no pagamento no passado A34 : Conta crítica/ Outros créditos existentes (não neste banco)
Purpose	Propósito ou Finalidade	Categórica	A40 : Comprar carro novo A41 : Comprar carro usado A42 : Comprar móveis/ equipamentos A43 : Comprar rádio/televisão A44 : Comprar eletrodomésticos A45 : Reparos A46 : Educação A47 : Férias 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	Residência atual desde	Numérica	
Propety	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 / Emprego	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	Numérica	
Telephone	Telefone próprio	Categórica	A191 : Não A192 : Sim
Foreign worker	Trabalhador estrangeiro	Categórica	A201 : Não A202 : Sim

A variável resposta *Credit History* possui 5 categorias ordenas. Como as categorias A30 até A32 possuem rótulos parecidos e sua presença no banco de dados é menor em relação a outras optou-se por juntar as 3 para uma só categoria com objetivo de facilitar a interpretação e modelagem do problema. Com isso a variável resposta agora passará a ter 3 níveis ordenados (do menor ao maior risco), sendo :

- A32 : Nenhum crédito recebido/ Todos os créditos pagos devidamente /Créditos existentes pagos devidamente até agora ;
- A33 : Atraso no pagamento no passado ;

- A34 : Conta crítica/ Outros créditos existentes (não neste banco).

Em relação ao campo *Default Status*, por ser uma variável que já classifica o solicitante entre adimplente e inadimplente. Sera desconsiderada no modelo por corresponder em parte aos objetivos do projeto de classificar os clientes entre os níveis da variável resposta.

4.2 Análise descritiva e dados covariáveis

Nessa seção será apresentado a análise descritiva e análise de correlação dos dados do banco escolhido. Sera utilizado tabelas e gráficos para demonstração.

As Tabelas 4 e 5 representam uma análise descritiva das variáveis numéricas e categóricas.

Tabela 4: Medidas resumo de cada variável Numérica

Variável	Mín	Mediana	Máx	Média	Desvio Padrão
Duration in month	4	18	72	20,9	12,06
Credit Amount	250	2320	18424	3271	2822,74
Installment rate of disposable income	1	3	4	2,973	1,12
Present residence since	1	3	4	2,845	1,10
Age in Years	19	33	75	35,55	11,38
Number of existing credits	1	1	4	1,407	0,58
Number of people being liable to provide maintenance for	1	1	2	1,155	0,36

Tabela 5: Medidas resumo de cada variável categórica

Variável	Nível	Frequência Simples	Frequência Relativa
Status of existing crecking account	A11 : x <0	274	27,4%
	A12 : 0 $\leq x$<200	269	26,9%
	A13 : x $\geq$ 200	63	6,3%
	A14 : Sem conta	394	39,4%
Credit History	A32: Nenhum crédito recebido/ Todos os créditos pagos devidamente ou Todos os créditos neste banco foram devidamente pagos ou Créditos existentes pagos devidamente até agora	619	61,9%
	A33 : Atraso no pagamento no passado	88	8,8%
	A34 : Conta crítica/ Outros créditos existentes (não neste banco)	293	29,3%
Purpose	A40 : Comprar carro novo	234	23,4%
	A41 : Comprar carro usado	103	10,3%
	A42 : Comprar móveis/ equipamentos	181	18,1%
	A43 : Comprar rádio/televisão	280	28,0%
	A44 : Comprar eletrodomésticos	12	1,2%
	A45 : Reparos	22	2,2%
	A46 : Educação	50	5%
	A47 : Férias	0	0%
	A48 : Reciclagem	9	0,9%
	A49 : Negócios	97	9,7%
	A410 : Outros	12	1,2%
Savings accounts/bonds	A61 : x <100	603	60,3%
	A62 : 100 $\leq x$<500	103	10,3%
	A63 : 500 $\leq x$<1000	63	6,3%
	A64 : x $\geq$ 1000	48	4,8%
	A65 : Desconhecido/sem conta poupança	183	18,3%
Present Employment	A71 : Desempregado	62	6,2%
	A72 : x <1 ano	172	17,2%
	A73 : 1 $\leq x$<4*anos*	339	33,9%
	A74 : 4 $\leq x$<7*anos*	14	17,4%
	A75 : x $\geq$ 7*anos*	253	25,3%
Personal status and sex	A91 : Homem divorciado/separado	50	5,0%
	A92 : Mulher divorciada/separada/casada	310	31,0%
	A93 : Homem solteiro	548	54,8%
	A94 : Homem casado/viúvo	92	9,2%
	A95 : Mulher solteira	0	0%
Other debtors	A101 : Nenhum	907	90,7%
	A102 : Co-requerente	41	4,1%
	A103 : Fiador	52	5,2%
Property	A121: Imobiliária	282	28,2%
	A122: Contrato de poupança/seguro de vida da sociedade de construção	232	23,2%
	A123: Carro ou outro	332	33,2%
	A124: Desconhecido	154	15,4%
Other installment plans	A141 : Banco	139	13,9%
	A142 : Lojas	47	4,7%
	A143 : Nenhum	814	81,4%
Housing	A151 : Aluguel	179	17,9%
	A152 : Própria	713	71,3%
	A153 : Cedida/de graça	108	10,8%
Job status	A171 : Desempregado/não qualificado	22	2,2%
	A172 : Empregado sem qualificação	200	20,0%
	A173 : Empregado qualificado/funcionário público	630	63,0%
	A174 : Gerência/autônomo/alta qualificação/policial	148	14,8%
Telephone	A191 : Não	596	59,6%
	A192 : Sim	404	40,4%
Foreign worker	A201 : Sim	963	96,3%
	A202 : Não	37	3,7%

Pela Tabela 4 é possível ver que os dados numéricos possuem bastante dispersão e muitos valores outliers como *Credit Amount* que vai de mínimo de 250 a um máximo de 18424, com enfase no desvio padrão que corresponde um valor próximo da média. Comportamento parecido são das variáveis *Durantio* e *Age* que tem uma amplitude alta, média e medianas com valores próximo mostrando uma centralidade , a diferença fica em magnitude do desvio padrão em relação a média.

A Tabela 5 mostra os dados categóricos e suas frequências simples e relativa. Entre as principais informações se observa que mais de 60% dos solicitantes tem menos de 100 DM(Marco alemão) de poupança. Quando se leva em consideração o motivo do empréstimo, a maior parte pede para compra de carro novo (23,4%) e rádio/televisão (28%). Também se observa que grande parte dos solicitantes possuem casa própria (71,3%), se encontram no caso de emprego qualificado ou da gestão publica (63%), são homens (69%) e em grande maioria estrangeiros (96,3%).

A Figura 1 apresenta a relação entre a variável resposta e algumas variáveis explicativas.

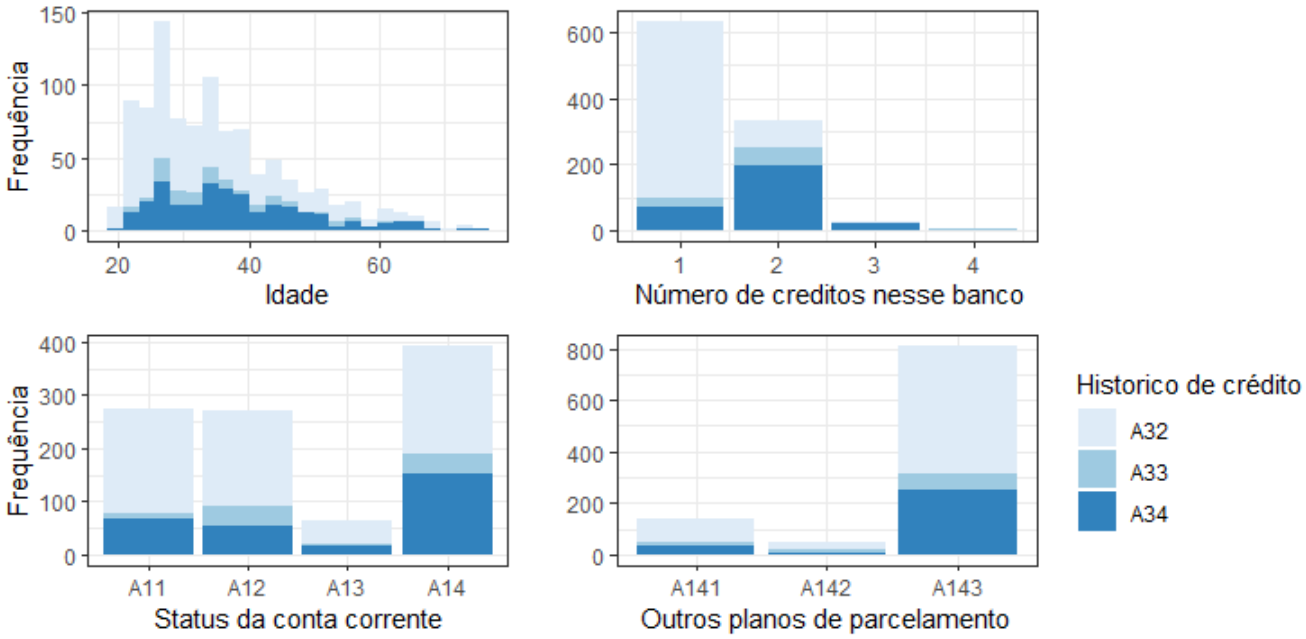

Figura 1: Relação entre o Histórico de Crédito e algumas variáveis explicativas

Pelo primeiro gráfico observa-se que ao ponto que a idade vai aumentando a porcentagem da categoria de menor risco vai diminuindo , mostrando indícios de que a idade tem relação com a variável dependente. Analisando o segundo gráfico é possível observar que ao ponto que o numero de créditos existentes aumenta a porcentagem da categoria de maior risco cresce em relação as demais. Considerando o 3º e 4º gráficos , o comportamento da variável resposta pelas variáveis *Status* e *Planos de parcelamento*

mostrou indícios de semelhança e que maior parte dos solicitantes de maior risco não possuem uma conta bancaria e nenhum outro tipo de plano de parcelamentos.

A Figura 2 apresenta 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 *Durantion* com a variável *Amount* com o valor de 0,62.

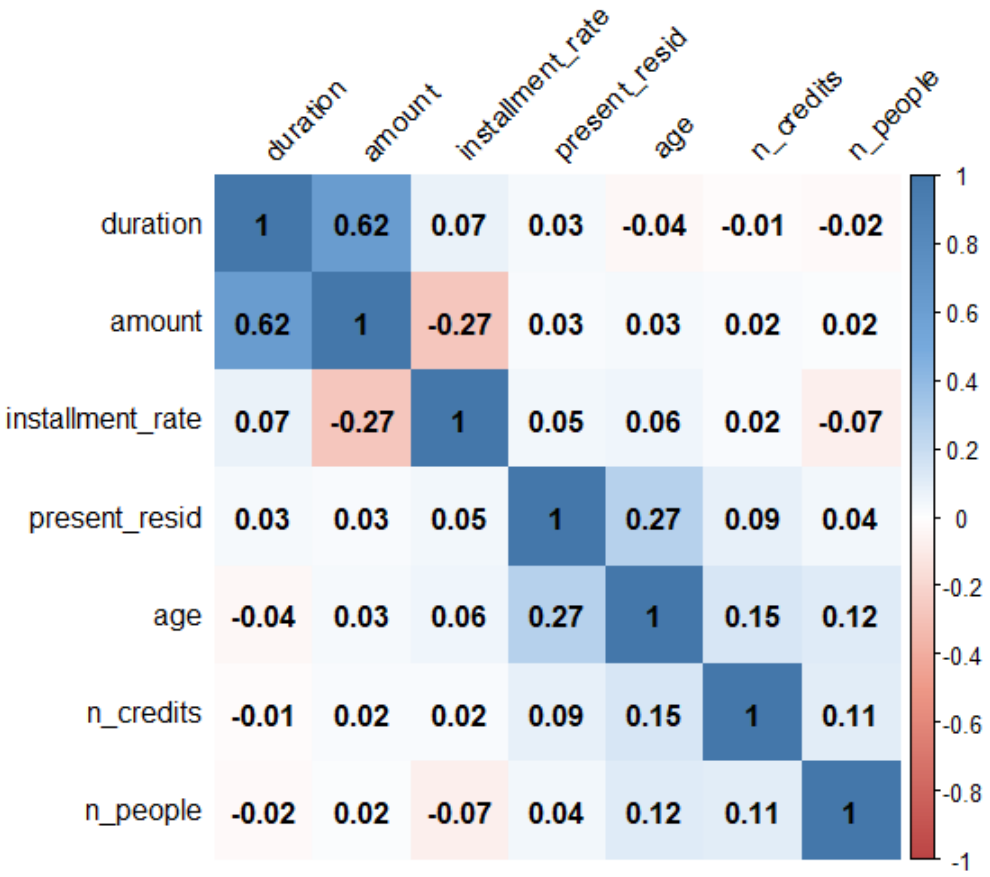

Figura 2: Gráfico de Correlações

4.3 Ajuste do modelo de regressão ordinal

Agora, nesta seção do trabalho, será feito ajuste do modelo de regressão logística ordinal. Como dito previamente, o banco de dados é dividido em "treino" e "teste", outro ponto é a transformação das variáveis categorias em *dummies* fazendo com que o banco possua agora 46 variáveis.

Para a seleção das variáveis foi usado o comando *stepAIC* do *R* que gera de forma automática quais são significativas pelo critério de *AIC*. O modelo final , após teste automáticos e testes manuais ficou (usando a 3º categoria como base):

$$Logit[P(Y_i \leq j|\mathbf{x})] = \log\left[\frac{P(Y_i \leq j|\mathbf{x})}{1 - P(Y_i \leq j|\mathbf{x})}\right] = \alpha_j - \beta_1 x_{i1} - ... - \beta_k x_{ik} \quad j = 1, ..., J-1.$$

$$Logit[P(Y_i \leq 1|\mathbf{x})] = 6,0416 - (0,157x_1 + 0,780x_2 + 0,833x_3 + 0,032x_4 + 0,297x_5 + 0,756x_6 + 2,308x_7). \tag{4.3.1}$$

$$Logit[P(Y_i \leq 2|\mathbf{x})] = 6,6189 - (0,157x_1 + 0,780x_2 + 0,833x_3 + 0,032x_4 + 0,297x_5 + 0,756x_6 + 2,308x_7). \tag{4.3.2}$$

Em que :

- x_1 : Status of existing checking account A12;
- x_2 : Status of existing checking account A13;
- x_3 : Status of existing checking account A14;
- x_4 : Age in years;
- x_5 : Other installments plans A142 ;
- x_6 : Other installments plans A143 ;
- x_7 : number of credits.

Foi feito o teste da Razão de Verossimilhança com todas as variáveis explicativas isoladamente em comparação com o modelo só com intercepto para testar as hipóteses :

$$H_o : \beta_j = 0$$

$$H_1 : \beta_j \neq 0$$

E todos os *p-valores* deram significância menor que $\alpha = 20\%$, assim tendo indícios para rejeitar a hipótese nula e assim mantendo as variáveis no modelo.

Tabela 6: Resultados dos coeficientes - Teste de Wald

Variável	**Estimativa**	**Erro padrão**	**Estatística t**	**P-valor**
Status of existing checking account A12	0,156592	0,264276	0,592532	0,553494
Status of existing checking account A13	0,780235	0,385819	2,022282	0,043147
Status of existing checking accountA14	0,833335	0,233572	3,567789	0,00036
Age in years	0,031921	0,007732	4,128553	3,65E-05
Other installments plans A142	0,296802	0,458434	0,647426	0,517356
Other installments plans A143	0,755978	0,271536	2,78408	0,005368
Number of credits	2,308124	0,178619	12,92203	3,38E-38
Intercept A32—A33	6,041576	0,548444	11,01584	3,21E-28
Intercept A33—A34	6,618896	0,560526	11,80836	3,53E-32

A Tabela 6 mostra o resultado dos coeficientes significativos para a formulação do escore de risco do modelo final de regressão logística ordinal , nele é possui o erro padrão, a estatística t e o p-valor. Como foi possível ver, *Status of existing checking account A12* e *Other installments plans A142* não foram significativamente ($\alpha = 5\%$) diferentes de 0. Entretanto como as variáveis foram significativas no Teste de Razão de Verossimilhança, sera mantido as categorias *Status of existing checking account A12* e *Other installments plans A142* no modelo final.

Tabela 7: Estimativas pontuais e intervalares (95% de confiança)

Variável	**Value**	**Inferior**	**Superior**
Status of existing checking account A12	0,156592	-0,36139	0,674573
Status of existing checking account A13	0,780235	0,024029	1,53644
Status of existing checking account A14	0,833335	0,375534	1,291135
Age in years	0,031921	0,016767	0,047075
Other installments plans A142	0,296802	-0,60173	1,195334
Other installments plans A143	0,755978	0,223767	1,288188
Number of credits	2,308124	1,95803	2,658217
Intercept A32 - A33	6,041576	4,966625	7,116527
Intercept A33-A34	6,618896	5,520264	7,717527

A Tabela 7 representa o intervalo de confiança a 95% dos parâmetros do modelo da regressão logística. Pela quantidade pequena de observações em algumas variáveis foi possível ver que há um intervalo grande em algumas variáveis.

4.4 Critérios de ajuste do modelo logístico

Para o ajuste do modelo final selecionado serão feitos por informações AIC e Resíduo de *deviance* e por testes de hipóteses. Que por sua vez tem o objetivo de validar pela significância.

Tabela 8: Comparação AIC - *Deviance*

Modelo	AIC	Deviance
Modelo completo	1022,623	930,6227
Modelo final	978,6331	960.6331

A Tabela 8 mostra o comparativo entre o modelo completo, que consiste em todas as variáveis, e o modelo final, o proposto nesse trabalho. Visto que o modelo final apresentou um resultado melhor de AIC e um valor levemente superior de *Deviance*, podemos optar pelo modelo final sem perda de informações significativas.

Agora será feito o teste de razão de verossimilhança para validação do modelo de regressão logística ordinal.

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

As hipóteses são:

$H_0 =$ Não existe diferença significativa entre o modelo proposto e o simples

$H_1 =$ Existe diferença significativa entre o modelo proposto e o simples

Tabela 9: Teste razão de máxima verossimilhança : constante - proposto

Modelo	DF	LogLink	DF	Chisq	P-valor
Modelo simples	2	-615,89	2		
Modelo final	9	-480,32	7	271,16	<2.2e-16

O primeiro teste será comparar o modelo final com o modelo apenas com o intercepto (constante). Pela Tabela 9 é possível concluir que o modelo final é significativamente

($\alpha = 10\%$) diferente do que o modelo apenas com a constante , ou seja sem as covariáveis. Com isso temos evidencias para o ajuste do modelo aos dados.

Tabela 10: Teste razão de máxima verossimilhança : modelo inicial - proposto

Modelo	DF	LogLink	DF	Chisq	P-valor
Modelo final	9	-480,32			
Modelo inicial	46	-465,31	37	30,01	0,7856

O segundo teste será comparar o modelo final com o modelo inicial que contem todas as variáveis explicativas do banco. Pela Tabela 10 é possível concluir que o modelo final não apresenta diferença significativa ($\alpha = 10\%$) em relação ao modelo inicial. Com isso podemos ficar com o modelo mais simples sem perda de informações.

Por último, o teste de Hosmer e Lemeshow será feito para ver as classificações em grupo são iguais as observadas.

As hipóteses são :

$$H_0 = \text{O modelo ajusta bem aos dados}$$

$$H_1 = \text{O modelo não ajusta bem aos dados}$$

Ao calcular o teste foi observado a estatística, com base 10 grupos, $\chi^2 = -2,5011$ com 8 graus de liberdade e $p\text{ -}valor\ = 1$. Os resultados concluem que não ha evidencias para rejeitar a hipótese nula, ou seja não há indícios que o modelo não se ajusta bem aos dados. Consequentemente, o modelo escolhido apresentou bons resultados e será considerado bem ajustado.

4.5 Obtenção do escore de risco e classificação dos indivíduos segundo a metodologia proposta

O Escore de Risco apresentado em (3.2.1.) foi calculado para cada cliente da amostra de acordo com as duas probabilidades estimadas de pertencer a uma das 3 categorias de risco.

$$\text{Escore} = IG_3(P) = \frac{(p_{A33} + 2p_{A33})}{2}.$$

Em que p_{A33} e p_{A34} são as probabilidades estimadas dos clientes pertencerem aos Históricos de crédito A33 e A34 respectivamente.

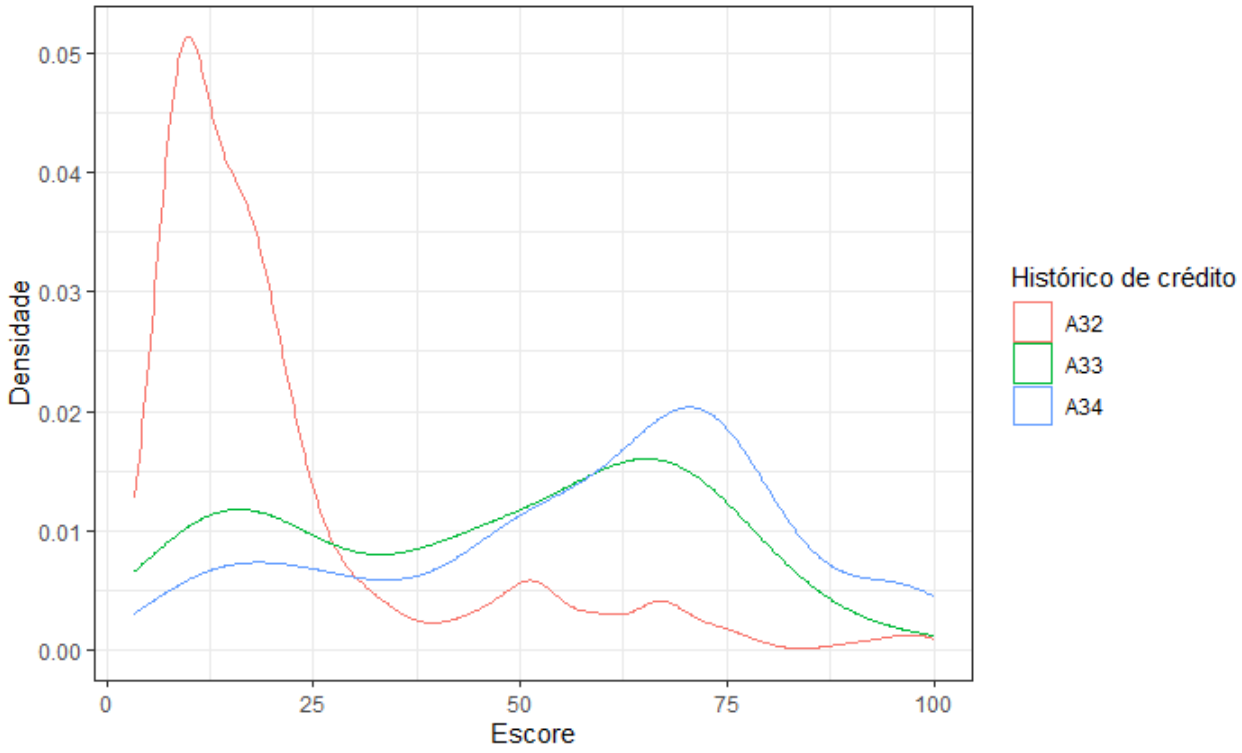

Figura 3: Gráfico da densidade dos escores

Na Figura 3 é apresentado a densidade da distribuição dos escores para as 3 categorias da variável resposta. Pelo gráfico observa-se que a categoria de menor risco teve um pico entre os escores mais baixos mostrando uma diferenciação entre as demais categorias. Enquanto as categorias de "médio" e "alto risco" tiveram um comportamento mais semelhante com a categoria mais grave tendo seu pico entre os escores mais altos.

O grande objetivo de determinar o escore de risco é minimizar: o erro de categorizar maus clientes como uma categoria de baixo risco e o erro de categorizar bons clientes como uma categoria de alto risco. Para que isso fosse possível foi determinado o porto de corte em 29 e 53, isto é, Clientes com escore de risco menor do que 29 foi classificado como "baixo risco (A32)", clientes com $29 \leq ER \leq 53$ como "médio risco (A33)" e clientes com $ER > 53$ foram classificados como "alto risco (A34)". Esses pontos de corte foram definidos de forma a controlar os erros 1 e 2 (probabilidade do erro 1 até 10% e probabilidade do erro 2 até 20%). A matriz de confusão dado esses cortes é apresentada na Tabela 11.

Tabela 11: Classificação - banco treino

	Predito			
Observado	A32	A33	A34	Total
A32	344	41	38	423
A33	19	13	29	61
A34	38	41	137	216
% de acerto	85,78%	13,68%	67,16%	70,57%
% de erro principais	9,48%		18,63%	

O modelo de regressão logística ordinal teve um desempenho razoável com $70,57\%$ de acerto global, chegando a $85,78\%$ na categorização dos solicitantes de menor risco e uma porcentagem bem próxima do acerto global para os solicitantes de maior risco. Como mencionado anteriormente o principal objetivo era obter um escore que classificasse de maneira adequada as categorias de maior risco para as instituições financeiras.

Tabela 12: Classificação - banco teste

	Predito			
Observado	A32	A33	A34	Total
A32	164	16	16	196
A33	11	4	12	27
A34	18	16	43	77
% de acerto	84,97%	11%	60,56%	70%
% de erro principais	9,33%		22,54%	

Ao analisar a Tabela 12 percebe-se que os dados estão próximos do banco de treino com 70% de acerto global e com acerto nas categorias bem próximo. Mostrando que o modelo se comportou bem a novos dados e tendo desempenho satisfatório.

5 Considerações finais

Com o aquecimento da economia e a oferta de crédito, as instituições financeiras viram a necessidade de estruturar processos objetivos e eficientes para a gestão dos risco em todas as etapas do ciclo de crédito. Entretanto, com a expansão do crédito, vem também o aumento da inadimplência, objeto de preocupação deste trabalho.

A proposta do trabalho foi desenvolver um escore de risco e averiguar sua capacidade preditiva de classificação de solicitantes entre categorias de risco ordenadas. A escolha da regressão logística foi pelo fato dela ser muito presente e conhecida no meio financeiro.

Esse trabalho fez o uso de um banco de dados chamado *German Credit Data* disponível na internet, no repositório de *Machine Learning Repository's* da Universidade da California-Irvin (UCI), nele foi possível o desenvolvimento do modelo de *Credit Scoring*: o escore de risco por meio da regressão logística ordinal.

Os resultados observados mostraram que o escore de risco proposto neste trabalho se mostrou adequado para classificação de clientes quando há mais do que duas categorias da variável resposta. A vantagem de se considerar mais do que dois níveis de classificação é permitir mais do que um ponto de corte, possibilitando o controle simultâneo das taxas de falso positivo (classificar um cliente como baixo risco, quando na realidade ele é um mal pagador) e falso negativo (classificar um cliente como alto risco, quando na realidade ele é um bom pagador). Ambos os erros devem ser evitados pela instituição, pois o primeiro resulta em fazer um financiamento para um mal pagador (e então assumir o risco de perdas) e o segundo é deixar de fazer um financiamento para um bom pagador (perda do cliente).

Referências

AGRESTI, A. *Categorical data analysis*. 3ed.. ed. [S.l.]: Wiley, 2013. 17, 20, 22

AGRESTI, A. *An Introduction to Categorical Data Analysis*. 3rd. ed. [S.l.]: John Wiley & Sons, 2019. (Wiley Series in Probability and Statistics). 13, 15, 16, 19, 20, 27

AKAIKE, H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, v. 19, n. 6, p. 716–723, 1974. 14

ALTMAN, E. I. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance*, 1968. 25

CELLA, L. O. G. *Regressão Ordinal Bayesiana*. Dissertação (Mestrado) — Universidade de Brasília, 2013.

DUA, D.; GRAFF, C. *UCI Machine Learning Repository- Banco German Credit*. 2017. Disponível em: ⟨http://archive.ics.uci.edu/ml⟩.

DURAND, D. *Risk Elements in Consumer Instalment Financing*. [S.l.]: National Bureau of Economic Research, 1941. 9, 24

HOSMER DAVID W.; LEMESHOW STANLEY; STURDIVANT, R. X. *Applied logistic regression*. 3. ed. ed. [S.l.]: Wiley, 2013. (Wiley series in probability and statistics). 13

KUTNER, C. J. N. J. N. W. L. M. H. *Applied linear statistical models*. 5th ed. ed. [S.l.]: McGraw-Hill Irwin, 2005. 10, 12, 13, 16, 17, 18, 21, 22

MACHADO, A. R. *Collection Scoring via Regressão Logística e Modelo de Riscos Proporcionais de Cox*. Dissertação (Mestrado) — Universidade de Brasília, 2015. 23

MYERS, J. H.; FORGY, E. W. The development of numerical credit evaluation systems. *Journal of the American Statistical Association*, Taylor & Francis, 1963. Disponível em: ⟨https://www.tandfonline.com/doi/abs/10.1080/01621459.1963.10500889⟩. 24

NAKANO, E. Y. *Soluções bayesianas para alguns problemas clássicos com dados discretos*. Tese (Doutorado) — Universidade de São Paulo, 2010. 26

ORGLER, Y. E. A credit scoring model for commercial loans. *Journal of Money, Credit Banking (Ohio State University Press) 2, 435-445*, 1970. 25

SANTOS, W. de Almeida Aguiar Yamamoto; Edson Aparecida de Araújo Querido Oliveira ; Vilma da S. O gerenciamento de risco de crédito em um banco de varejo: um estudo do segmento pessoas físicas. *XV Encontro Latino Americano de Iniciação Científica e XI Encontro Latino Americano de Pós-Graduação*, 2011. 23

SISCU, A. L. *Credit Scoring*. [S.l.]: Blucher, 2010. 14

WEINGARTNER, H. *Concepts and Utilization of Credit-Scoring Techniques*. [S.l.]: Banking 58,51-54, 1966. 24