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

Uma proposta de escore de risco de crédito via modelagem do tempo e saldo devedor no momento da inadimplência

Marcos Augusto Daza Barbosa

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

Brasília - DF, 2021

Marcos Augusto Daza Barbosa

Uma proposta de escore de risco de crédito via modelagem do tempo e saldo devedor no momento da inadimplência

Relatório apresentado ao Departamento de Estatística, Instituto de Exatas, Universidade de Brasília, como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

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

Brasília - DF, 2021

Resumo

Avaliar o risco associado a um cliente, no contexto de concessão de crédito, é de muito interesse por parte das instituições financeiras. Este trabalho propôs um escore capaz de discriminar clientes conforme o risco, baseado no saldo devedor, com a utilização da técnica de Análise de Sobrevivência. A metodologia foi ilustrada em um conjunto de dados simulados e os resultados evidenciam que o escore de risco proposto é adequado, para classifcar os clientes, conforme o risco de gerar um prejuízo financeiro.

Palavras-chave: *Credit Scoring*, empréstimo, financiamento, Análise Sobrevivência, finanças, modelagem estatística, Modelo de Riscos Proporcionais de Cox, censura, risco, cliente, classificação, SAC.

Abstract

Evaluating the risk associated to a client, in the credit granting context, is interesting for financial institutions. The aim of this work was to propose a score to discriminate clients by their risk, which is based on the balance due, with the utilization of the technique of Survival Analysis. The methodology was illustrated on a simulated dataset and the results evidence that the proposed score risk is appropriate to classify the clients, by their risk of generating a financial prejudice.

Keywords: *Credit Scoring*, loan, financing, Survival Analysis, finance, statistical modeling, Cox Proportional-Hazards Model, censored data, risk, client, classification, System of Constant Amortization.

Lista de ilustrações

Figura 1 – Exemplo de Função de Sobrevivência 5
Figura 2 – Exemplo de Curva ROC 20
Figura 3 – Função de sobrevivência estimada por Kaplan-Meier 25
Figura 4 – Função de sobrevivência estimada do cliente com $X_1 = 0,3505$, $X_2 = 0,5844$, $X_3 = 0$, $V = 83332,69$ e $N = 42$ 27
Figura 5 – Funções de densidade e médias do escore por classificação de risco 27
Figura 6 – Curva ROC 29

Lista de tabelas

Tabela 1 – Exemplo de Matriz de Confusão 18
Tabela 2 – Medidas-resumo de cada variável quantitativa 24
Tabela 3 – Distribuição das variáveis categóricas 24
Tabela 4 – Resultados dos coeficientes 25
Tabela 5 – Teste de tendência dos resíduos de Schoenfeld 25
Tabela 6 – Tempo e Função de Sobrevivência Base Estimada 26
Tabela 7 – 20 clientes com maiores escores de risco 28
Tabela 8 – 20 clientes com menores escore de risco 29
Tabela 9 – Matriz de confusão no cenário com tolerância moderada ao erro 30
Tabela 10 – Métricas de performance no cenário com tolerância moderada ao erro 30
Tabela 11 – Matriz de confusão no cenário com tolerância baixa ao erro 31
Tabela 12 – Métricas de performance no cenário com tolerância baixa ao erro 31
Tabela 13 – Resultados das métricas de *performance* 37

Sumário

1 Introdução 1
2 Revisão de Literatura 3
2.1 Notação 3
2.2 Censura 3
2.3 Função densidade de probabilidade 4
2.4 Função de sobrevivência 4
2.5 Função de risco e função de taxa de falha acumulada 5
2.6 Relação entre as funções 6
2.7 Estimador de Kaplan-Meier 6
2.8 Estimador de Nelson-Aalen 7
2.9 Método de estimação por máxima verossimilhança 7
3 Metodologia 9
3.1 Modelo de Riscos Proporcionais de Cox 9
3.1.1 Estimação dos parâmetros no Modelo de Riscos Proporcionais de Cox . . 10
3.1.2 Estimação de funções relacionadas com $h_0(t)$ 11
3.1.3 Interpretação dos parâmetros no Modelo de Riscos Proporcionais de Cox 11
3.1.4 Adequação do Modelo de Riscos Proporcionais de Cox 12
3.1.4.1 Avaliação da suposição de riscos proporcionais 12
3.1.4.2 Avaliação geral do ajuste do Modelo de Cox 14
3.1.4.3 Avaliação de outros aspectos do Modelo de Cox 14
3.2 Modelagem de Risco e Escore proposto 15
3.2.1 Introdução à modelagem de risco 15
3.2.2 Escore proposto que considera a informação do tempo 16
3.2.3 Critério de acurácia do modelo 17
3.3 Sobre o conjunto de dados simulados 20
4 Resultados 24
4.1 Ajuste do Modelo de Cox 25
4.2 Cálculo do escore de risco e classificação dos indivíduos conforme risco 26
4.2.1 Ilustração com exemplo de um cliente 26
4.2.2 Ilustração com a amostra de dados 27
5 Conclusão 33

Referências 35
APÊNDICE A Resultados das métricas de performance 37

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, pois a rentabilidade está relacionada ao número de empréstimos feitos e à porcentagem de clientes que regularizam suas posições. Portanto, é ponto pacífico afirmar que uma medida que quantifica o risco de um determinado cliente é uma ferramenta útil e necessária nesse mercado.

Ao longo dos anos, diversos modelos estatísticos surgiram para avaliar esse risco. Nesse contexto, há os Modelos de *Credit Scoring*, cujo objetivo é avaliar a possibilidade de ocorrência de perdas associadas ao não cumprimento pelo tomador de suas obrigações financeiras. Contudo, a maioria desses modelos têm como base a metodologia de Regressão Logística, que tem a limitação de analisar apenas a probabilidade do cliente honrar seu compromisso e não incorporam a informação do tempo. Por exemplo, é interessante saber qual seria o tempo estimado até que esse evento de interesse (inadimplência ou atraso no pagamento) ocorra, que é o principal foco da técnica de Análise de Sobrevivência.

De acordo com Colosimo e Giolo (2006), esse campo da Estatística tem crescido de maneira acelerada nas últimas duas décadas, pois apresenta aplicações em diversas áreas, como experimentos biomédicos e industriais, demografia, seguros e, neste caso, no setor financeiro de concessão de crédito.

A importância de estudar o tempo até inadimplência se deve ao fato de que essa informação está diretamente associada ao valor do saldo devedor (montante total que ainda falta ser pago). Neste trabalho, será proposto um escore de risco que incorpora a informação do tempo e o valor do saldo devedor. De modo mais específico, a variável tempo irá corresponder ao número de meses (completos) até o possível primeiro atraso de pagamento de alguma parcela do empréstimo e será estudada por meio do Modelo de Riscos Proporcionais de Cox (COX, 1972).

A metodologia proposta neste trabalho será ilustrada em um conjunto de dados simulados por meio do programa R (R Core Team, 2020).

2 Revisão de Literatura

Nesta seção, será descrita uma revisão de literatura de Análise de Sobrevivência. Nessa técnica estatística, a variável resposta costuma ser o tempo até a ocorrência de um evento de interesse. Esse evento pode ser a morte de um paciente, a falha de um equipamento industrial, atraso em algum pagamento, etc.

2.1 Notação

As fórmulas apresentadas durante este trabalho possuem notação específica da técnica de Análise Sobrevivência. As principais notações são:

- δ: indicador de censura;
- $f(t)$: função densidade de probabilidade;
- $S(t)$: função de sobrevivência;
- $h(t)$: função de risco (ou função taxa de falha);
- $H(t)$: função de risco (ou função taxa de falha) acumulada;
- $\boldsymbol{\beta}$: vetor de parâmetros.

2.2 Censura

A principal característica de dados de sobrevivência é a presença de censura (COLOSIMO; GIOLO, 2006), que são observações incompletas ou parciais que acontecem quando, por alguma razão, o acompanhamento do indivíduo em estudo foi interrompido. Segundo Aniceto (2011), mesmo sendo incompletas, as observações censuradas trazem algumas informações dos indivíduos e, por isso, devem ser incluídas na análise estatística. Se retiradas, os resultados do estudo podem se tornar viciados. Vale destacar que caso não existisse censura no conjunto de dados, técnicas clássicas da Estatística poderiam ser usadas, como Análise de Regressão ou Delineamento e Análise de Experimentos.

Dessa forma, existe a necessidade da introdução de uma variável extra na análise, que indica se o valor do tempo de sobrevivência de um determinado indivíduo é, ou não, um tempo de falha. Essa variável, geralmente representada por δ, é conhecida como variável indicadora de censura e é expressa por:

$$\delta = \begin{cases} 1, & \text{se o tempo é observado (falha),} \\ 0, & \text{se o tempo é censurado.} \end{cases} \tag{2.1}$$

Os tipos de censuras podem ser divididos em:

- **Censura à direita:** é caracterizada pelo tempo de ocorrência do evento estar à direita do tempo registrado, ou seja, ser maior do que o valor que foi registrado. Alguns mecanismos de censura à direita são:
  - **Censura do Tipo I** - O estudo será terminado após um período de tempo fixo, e ao final dele, uma ou mais observações em estudo não falharam. Esse tempo deve ser determinado antes do início do estudo;
  - **Censura do Tipo II** - O estudo termina após ocorrer o evento de interesse em um número fixo de indivíduos. O número de falhas deve ser determinado antes do início do estudo;
  - **Censura aleatória:** São todos os casos em que as observações não experimentam o evento de interesse por motivos não controláveis.
- **Censura à esquerda:** ocorre quando o tempo registrado é maior do que o tempo de falha. O evento de interesse já aconteceu quando o indivíduo foi observado;
- **Censura intervalar:** ocorre em estudos em que os elementos tem acompanhamento periódico. O evento de interesse ocorrem em um intervalo de tempo.

2.3 Função densidade de probabilidade

A função densidade de probabilidade é definida como o limite da probabilidade de um indivíduo experimentar o evento de interesse em um intervalo de tempo $[t; t + \Delta t)$ por unidade de Δt (comprimento do intervalo), ou simplesmente por unidade de tempo. Para uma variável aleatória T contínua, a função densidade de probabilidade é expressa por (LEE; WANG, 2003):

$$f(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t}, t \geq 0. \tag{2.2}$$

2.4 Função de sobrevivência

A função de sobrevivência, denotada por $S(t)$, é definida como a probabilidade de um indivíduo sobreviver a um tempo t, ou seja, a probabilidade de um indivíduo não falhar até um certo tempo t. Ela é expressa por:

$$S(t) = P(T > t) = \int_t^\infty f(x)dx, t \geq 0. \tag{2.3}$$

De maneira análoga, a função de distribuição acumulada é definida como a probabilidade de um indivíduo não sobreviver ao tempo t, ou seja, $F(t) = 1 - S(t)$.

Vale ressaltar as seguintes propriedades da função de sobrevivência:

- $\lim_{t\to 0} S(t) = 1$;
- $\lim_{t\to \infty} S(t) = 0$.

Por último, é apresentado um gráfico, como exemplo, para ilustrar o comportamento dessa função.

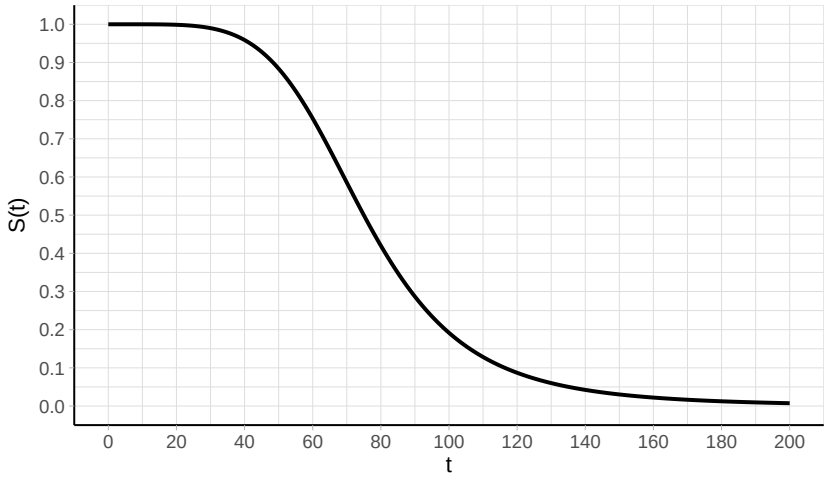

Figura 1 – Exemplo de Função de Sobrevivência

2.5 Função de risco e função de taxa de falha acumulada

A função de risco (ou função taxa de falha) é definida como o limite da probabilidade de um indivíduo falhar no intervalo de tempo $[t; t + \Delta t)$, assumindo que este mesmo indivíduo sobreviveu até o tempo t, dividida pelo comprimento do intervalo e é representada por (LAWLESS, 2011):

$$h(t) = \lim_{\Delta t\to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}, t \geq 0. \tag{2.4}$$

Pode ainda ser expressa em termos da função de densidade de probabilidade e da função de sobrevivência, isto é,

$$h(t) = \frac{f(t)}{S(t)}. \tag{2.5}$$

A função de taxa de falha acumulada é obtida a partir da função de risco (2.4) e é representada por:

$$H(t) = \int_0^t h(u)du. \tag{2.6}$$

Essa função fornece a taxa de falha acumulada e auxilia na escolha da distribuição utilizada na parametrização.

2.6 Relação entre as funções

É possível encontrar relações matemáticas entre as funções apresentadas, que serão úteis para os cálculos nesse trabalho. Algumas dessas relações serão apresentadas a seguir:

$$f(t) = \frac{\partial F(t)}{\partial t}. \tag{2.7}$$

Uma vez que a função de distribuição acumulada é $F(t) = 1 - S(t)$, segue que:

$$f(t) = \frac{\partial [1 - S(t)]}{\partial t} = -S^{'}(t). \tag{2.8}$$

Levando em conta $f(t) = -S^{'}(t)$ e (2.5), tem-se que

$$h(t) = -\frac{S^{'}(t)}{S(t)} = -\frac{\partial [\log S(t)]}{\partial t}. \tag{2.9}$$

Portanto,

$$\log S(t)|_0^t = -\int_0^t h(u)du. \tag{2.10}$$

Por meio das equações (2.6) e (2.10), segue que:

$$S(t) = \exp[-H(t)]. \tag{2.11}$$

As equações (2.5) e (2.11) serão comuns para se realizarem os cálculos.

2.7 Estimador de Kaplan-Meier

O estimador não-paramétrico de Kaplan-Meier é uma adaptação da função de sobrevivência empírica que, na ausência de censuras, é definida como:

$$\hat{S}(t) = \frac{\text{nº de observações que não falharam até o tempo } t}{\text{nº total de observações no estudo}}. \tag{2.12}$$

Quando existe censura no estudo, o estimador (2.12) não é adequado. Por isso, é indicado a utilização de outros estimadores que consideram as censuras observadas.

Suponha um estudo com n indivíduos, k falhas distintas nos tempos $t_{(1)} < t_{(2)} < \cdots < t_{(k)}$ e algumas considerações:

- Considere tantos intervalos de tempo quantos forem o número de falhas distintas. Os limites dos intervalos de tempo são os tempos de falha da amostra;
- d_j : o número de falhas em $t_{(j)}, j = 1, \ldots, k$;

- n_j : o número de indivíduos sob risco em $t_{(j)}$, ou seja, os indivíduos que não falharam e não foram censurados até o instante imediatamente anterior a $t_{(j)}$.

Sob as considerações acima, o estimador de Kaplan-Meier é definido por:

$$\hat{S}(t) = \prod_{j:t_{(j)} \leq t} \left(\frac{n_j - d_j}{n_j} \right). \tag{2.13}$$

2.8 Estimador de Nelson-Aalen

O estimador de Nelson-Aalen é uma opção para estimar a função de risco acumulada e é definido por:

$$\hat{H}_{NA}(t) = \sum_{j:t_j < t} \left(\frac{d_j}{n_j} \right), \tag{2.14}$$

em que d_j é o número de falhas em t_j, n_j é o número de indivíduos sob risco em t_j e $j = 1, \ldots, k$.

Ao considerar (2.14) e a relação $H(t) = -[log(S(t))]$, o estimador não-paramétrico para a função de sobrevivência é definido por:

$$\hat{S}_{NA}(t) = \exp\{-\hat{H}_{NA}(t)\}. \tag{2.15}$$

2.9 Método de estimação por máxima verossimilhança

Quando se trabalha com modelagem estatística, é necessário estimar os parâmetros do modelo proposto. Para a técnica de Análise de Sobrevivência, tem-se que: sendo $t_1, \ldots, t_n$ uma amostra aleatória observada de tamanho n da variável aleatória T com função de densidade $f(t|\theta)$ e incorporando observações censuradas, em que θ é o vetor de parâmetros, a função de verossimilhança é definida como:

$$L(\theta) \propto \prod_{i=1}^{n} [f(t_i;\theta)]^{\delta_i} [S(t_i;\theta)]^{1-\delta_i}, \tag{2.16}$$

em que δ_i é a variável indicadora de falha definida em (2.1).

Os estimadores de máxima verossimilhança são os valores de θ que maximizam $L(\theta)$ que podem ser obtidos ao resolver o sistema de equações:

$$\frac{\partial \log L(\theta)}{\partial \theta} = 0.$$

3 Metodologia

3.1 Modelo de Riscos Proporcionais de Cox

O modelo de riscos proporcionais de Cox (COX, 1972) é um dos mais utilizados na análise de dados de sobrevivência e é caracterizado por não assumir uma distribuição de probabilidade para o tempo. Além disso, por meio dessa modelagem, é possível avaliar o tempo até a ocorrência de um evento de interesse, considerando-se a presença de variáveis explicativas.

Assumindo que os tempos $t_i, i = 1, \ldots, n$, são independentes, o modelo de Cox estabelece que as covariáveis agem multiplicaticamente na função de risco do indivíduo, isto é:

$$h_i(t|\mathbf{x}_i) = h_0(t)\exp(\boldsymbol{\beta}^T\mathbf{x}_i), \tag{3.1}$$

em que h_0 representa o risco base, que é uma função não negativa e arbitrária (componente não paramétrico), isto é, o risco do indivíduo i quando todas as covariáveis são nulas ($\mathbf{x}_i = 0$), enquanto que $\boldsymbol{\beta}$ é o vetor de parâmetros desconhecidos e $\mathbf{x}_i$ é o vetor de covariáveis observadas do indivíduo i de dimensão p.

A denominação desse modelo como sendo de riscos proporcionais é justificada pelo fato de que a razão entre as funções de risco de dois indivíduos, i e l, tal que $i \neq l$, não depende de t, isto é,

$$\frac{h(t|\mathbf{x}_i)}{h(t|\mathbf{x}_l)} = \frac{h_0(t)\exp(\boldsymbol{\beta}^T\mathbf{x}_i)}{h_0(t)\exp(\boldsymbol{\beta}^T\mathbf{x}_l)} = \exp[\boldsymbol{\beta}^T(\mathbf{x}_i - \mathbf{x}_l)]. \tag{3.2}$$

Sendo assim, a suposição primordial para o uso desse modelo é que as funções de taxas de falha sejam proporcionais.

Considerando a função de risco base $h_0(t)$ e as expressões (2.6) e (2.11)

$$H_0(t) = \int_0^t h_0(u)du \tag{3.3}$$

e

$$S_0(t) = \exp\left(-\int_0^t h_0(u)du\right) \tag{3.4}$$

Sendo assim, pode-se concluir que a função de sobrevivência associada ao modelo de Cox é:

$$
\begin{aligned}
S(t|\mathbf{x}) &= \exp\left(-\int_0^t h(u|\mathbf{x})du\right) \\
&= \exp\left(-\int_0^t h_0(u)\exp(\boldsymbol{\beta}^T\mathbf{x})du\right) \\
&= \exp\left(-\exp(\boldsymbol{\beta}^T\mathbf{x})\int_0^t h_0(u)du\right) \\
&= \exp(-\exp(\boldsymbol{\beta}^T\mathbf{x})H_0(t)) \\
&= \left[e^{-H_0(t)}\right]^{\exp(\boldsymbol{\beta}^T\mathbf{x})} \\
&= [S_0(t)]^{\exp(\boldsymbol{\beta}^T\mathbf{x})}. \qquad (3.5)
\end{aligned}
$$

3.1.1 Estimação dos parâmetros no Modelo de Riscos Proporcionais de Cox

A inclusão do componente não paramétrico impede o uso do método de máxima verossimilhança, que é normalmente o mais utilizado para a estimação de parâmetros desconhecidos.

Sendo assim, COX (1975) elaborou a função de verossimilhança parcial para estimar o vetor de parâmetros desconhecidos $\boldsymbol{\beta}^T = (\beta_1, \beta_2, \ldots, \beta_k)$. Esta função considera que a probabilidade da i-ésima observação vir a falhar no tempo t_i, dado as observações sob risco em t_i, é:

$$
\frac{h_i(t_i|\mathbf{x}_i)}{\sum_{j\in R(t_i)} h_j(t_i|\mathbf{x}_j)} = \frac{h_0(t_i)\exp(\boldsymbol{\beta}^T\mathbf{x}_i)}{\sum_{j\in R(t_i)} h_0(t)\exp(\boldsymbol{\beta}^T\mathbf{x}_j)} = \frac{\exp(\boldsymbol{\beta}^T\mathbf{x}_i)}{\sum_{j\in R(t_i)}\exp(\boldsymbol{\beta}^T\mathbf{x}_j)}, \qquad (3.6)
$$

em que $R(t_i) = R_i$ é conjunto de todos os indivíduos ainda em risco no tempo t_i.

Sendo assim, tem-se que a função de verossimilhança será expressa por:

$$
L(\boldsymbol{\beta}) = \prod_{i=1}^{n}\left[\frac{\exp(\boldsymbol{\beta}^T\mathbf{x}_i)}{\sum_{j\in R_i}\exp(\boldsymbol{\beta}^T\mathbf{x}_j)}\right]^{\delta_i}. \qquad (3.7)
$$

Aqui, δ_i é o indicador de censura que assume o valor 1, se t_i é um tempo de falha ou 0, se t_i é um tempo censurado. Os indivíduos censurados são inseridos na função de verossimilhança parcial $L(\boldsymbol{\beta})$ por meio do conjunto de Risco R_i e contribuem, para esta função, somente enquanto permancem em risco. A estimação do vetor de paramêtros $\boldsymbol{\beta}$ é feita por meio da maximização do logaritmo da função $L(\boldsymbol{\beta})$. Portanto, resolve-se o seguinte sistema de equações escores:

$$
0 = \frac{\partial log L(\boldsymbol{\beta})}{\beta_k} = \sum_{i=1}^{n}\delta_i\left[x_{ik} - \frac{\sum_{j\in R_i} x_{jk}\exp(\boldsymbol{\beta}^T\mathbf{x}_j)}{\sum_{j\in R_i}\exp(\boldsymbol{\beta}^T\mathbf{x}_j)}\right], k = 1, \ldots, p. \qquad (3.8)
$$

No caso de empates nos valores observados, Breslow (1972) e Peto (1972) propõem a seguinte função de verossimilhança parcial modificada:

$$
L(\boldsymbol{\beta}) = \prod_{i=1}^{n}\frac{\exp(\boldsymbol{\beta}^T\mathbf{s}_i)}{[\sum_{j\in R_i}\exp(\boldsymbol{\beta}^T\mathbf{x}_j)]^{d_i}}, \qquad (3.9)
$$

em que $\mathbf{s}_i$ é o vetor formado pela soma das correspondentes p covariáveis para os indivíduos que falham no mesmo tempo t_i e d_i é o número de falhas neste mesmo tempo.

3.1.2 Estimação de funções relacionadas com $h_0(t)$

É importante estimar as funções $H_0(t)$ e $S_0(t)$, que estão relacionadas com $h_0(t)$.

Uma vez que $h_0(t)$ é de natureza não-paramétrica, não é possível estimá-lo por meio da função de verossimilhança.

Sendo assim, o estimador proposto para a função de risco base acumulada é dado por:

$$\hat{H}_0(t_i) = \sum_{i:t_i \leq t} \frac{d_i}{\sum_{j \in R_j} \exp(\hat{\boldsymbol{\beta}}^T \mathbf{x}_j)}, \tag{3.10}$$

em que d_i é o número de falhas no tempo $t_{(i)}$ e R_j é o conjunto de todos os indivíduos ainda em risco no tempo $t_{(j)}$ e $\hat{\boldsymbol{\beta}}$ é o estimador de $\boldsymbol{\beta}$ obtido em (3.8). Este estimador é conhecido como estimador de Nelson-Aalen-Breslow.

E, consequentemente, o estimador para a função de sobrevivência base (3.4), é

$$\hat{S}_0(t) = \exp[-\hat{H}_0(t)]. \tag{3.11}$$

E, a partir de (3.5), tem-se que a função de sobrevivência estimada, para um indivíduo com vetor de covariáveis $\mathbf{x}_i$, é expressa por:

$$\hat{S}(t|\mathbf{x}_i) = [\hat{S}_0(t)]^{\exp(\hat{\boldsymbol{\beta}}^T \mathbf{x}_i)}. \tag{3.12}$$

3.1.3 Interpretação dos parâmetros no Modelo de Riscos Proporcionais de Cox

Por meio dos coeficientes $\boldsymbol{\beta}$, é possível avaliar o efeito das covariáveis sobre a taxa de falha.

Considerando (3.2), que retrata a propriedade de riscos proporcionais, a razão entre as funções de risco para dois indivíduos (i e l) que possuem os mesmo valores para as covariáveis, com exceção na q-ésima covariável, é expressa por:

$$\frac{h(t|\mathbf{x}_i)}{h(t|\mathbf{x}_l)} = \frac{\exp(\beta_q x_{iq})}{\exp(\beta_p x_{lq})} = \exp[\beta_q (x_{iq} - x_{lq})]. \tag{3.13}$$

No caso de variáveis dicotômicas, por exemplo, $x_{iq} = 1$ (masculino) e $x_{lq} = 0$ (feminino), a função de risco do indivíduo masculino é $\exp(\beta_q)$ vezes a função de risco do indivíduo do sexo feminino, com as demais variáveis fixas.

3.1.4 Adequação do Modelo de Riscos Proporcionais de Cox

Para investigar a adequação do Modelo de Riscos Proporcionais de Cox, é importante avaliar a suposição de proporcionalidade dos riscos e o seu ajuste geral.

3.1.4.1 Avaliação da suposição de riscos proporcionais

Para validar essa suposição há três estratégias relevantes:

Método gráfico descritivo

1. Dividir o conjunto de dados em j estratos conforme as j categorias de alguma covariável;
2. Estimar $\hat{H}_{0j}(t)$ para cada estrato j e obter, em seguida, as curvas de $\log[\hat{H}_{0j}(t)]$ versus t, ou $\log(t)$.

Espera-se que, se as curvas de $\log[\hat{H}_{0j}(t)]$ versus t, ou $\log(t)$ apresentarem diferenças constantes no tempo, isto é, aproximadamente paralelas, a suposição de riscos proporcionais é válida. Para cada covariável, essas curvas podem ser encontradas e, dessa maneira, averiguar se há alguma covariável que esteja violando essa suposição primordial. No caso de variável contínua, é recomendado agrupá-la em um número pequeno de categorias, para depois executar o método mencionado.

Resíduos de Schoenfeld (SCHOENFELD, 1982)

Se o i-ésimo indivíduo com vetor de covariáveis $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T$ falhar, tem-se para esse indivíduo um vetor de resíduos de Schoenfeld $\mathbf{r}_i = (r_{i1}, x_{i2}, \ldots, r_{ip})^T$, em que cada componente r_{iq}, para $q = 1, \ldots, p$, definido por:

$$r_{iq} = x_{iq} - \frac{\sum_{j \in R(t_i)} x_{jk} \exp(\hat{\boldsymbol{\beta}}^T \mathbf{x}_j)}{\sum_{j \in R(t_i)} \exp(\hat{\boldsymbol{\beta}}^T \mathbf{x}_j)}. \tag{3.14}$$

Uma forma padronizada dos resíduos de Schoefeld é bastante usada e expressa por:

$$s_i^* = [I(\hat{\boldsymbol{\beta}})]^{-1} \times \mathbf{r}_i, \tag{3.15}$$

em que $I(\hat{\boldsymbol{\beta}})$ é a matriz de informação observada.

A utilização desses resíduos padronizados de Schoenfeld para avalidar a suposição de riscos proporcionais é baseado em um resultado, apresentado em (GRAMBSCH; THERNEAU, 1994), que considera o modelo expresso por:

$$h(t) = h_0(t) \exp[\mathbf{x}^T \boldsymbol{\beta}(t)], \tag{3.16}$$

com a restrição de que $\boldsymbol{\beta}(t) = \boldsymbol{\beta}$, como uma forma alternativa de representar o modelo de Cox.

Considerações sobre esse resíduo:

- A restrição $\boldsymbol{\beta}(t) = \boldsymbol{\beta}$ implica na proporcionalidade dos riscos;
- Quando $\boldsymbol{\beta}(t)$ não é constante, o efeito de uma ou mais covariáveis no risco pode variar com o tempo;
- O gráfico $\boldsymbol{\beta}_q(t)$ versus t deve ser uma linha horizontal;
- Segundo Grambsch e Therneau (1994), sugere-se o gráfico $s^*_{iq} + \hat{\beta}_q$ versus t. Inclinação zero mostra evidências a favor da suposição de riscos proporcionais.

Testes de Hipótese

O coeficiente de correlação de Pearson (ρ) entre os resíduos padronizados de Schoenfeld e $g(t)$ (uma função no tempo), para cada covariável, é uma medida possível para avaliar a suposição de riscos proporcionais. Valores de ρ próximos de zero demonstram que não há evidências para rejeição dessa suposição.

Um teste para a hipótese global de proporcionalidade dos riscos, isto é, levando em conta todas as covariáveis do modelo, pode ser executado, por meio da estatística de teste:

$$T = \frac{(g - \bar{g})^T S^* I S (g - \bar{g})}{d \sum_k (g_k - \bar{g})^2}, \tag{3.17}$$

em que:

- I é a matriz de informação de Fisher observada;
- d é o número de falhas;
- $S^* = dRI^{-1}$, tal que R é a matriz $d \times p$ dos resíduos de Schoenfeld não padronizados.

Sob a hipótese nula de proporcionalidade dos riscos, T possui distribuição qui-quadrado com p graus de liberdade. Portanto, valores de $T > \chi^2_{p,1-\alpha}$ mostram evidências contra a suposição de riscos proporcionais.

É interessante também realizar esse teste para uma determinada covariável, isto é, um teste local. Para a q-ésima covariável, $q = 1, \ldots, p$, a estatística de teste correspondente é:

$$T_q = \frac{d[\sum_k (g_k - \bar{g}) s^*_{qk}]^2}{I_q^{-1} \sum_k (g_k - \bar{g})^2}, \tag{3.18}$$

em que I_q^{-1} é o q-ésimo elemento da diagonal do inverso da matriz de informação de Fisher observada.

Sob a hipótese nula de proporcionalidade dos riscos para a q-ésima covariável, T_q possui distribuição qui-quadrado com 1 grau de liberdade. Desse modo, valores de $T_q > \chi^2_{1,1-\alpha}$ mostram evidências contra a suposição de riscos proporcionais para a covariável x_q.

3.1.4.2 Avaliação geral do ajuste do Modelo de Cox

De acordo com Colosimo e Giolo (2006), o resíduo de Cox-Snell auxilia a examinar o ajuste global do modelo final e é definido por:

$$\hat{e}_i = \hat{H}_0(t_i) \exp\left[\sum_{k=1}^{p} x_{ik}\hat{\beta}_k\right], i = 1, \ldots, n. \tag{3.19}$$

Caso o modelo esteja bem ajustado, os resíduos e_i's podem ser interpretados como uma amostra censurada de uma distribuição exponencial padrão e, desse modo, o gráfico $\hat{H}(e_i)$ versus e_i deveria se aproximar de uma reta.

Vale ressaltar que há uma desvantagem em utilizar esses resíduos: ele não indicam o tipo de falha detectado pelo modelo, quando o gráfico de $\hat{H}(e_i)$ versus e_i não é linear.

3.1.4.3 Avaliação de outros aspectos do Modelo de Cox

Ademais, outros aspectos do Modelo de Riscos Propocionais de Cox também podem ser avaliados por meio dos resíduos Martingal e Deviance.

Resíduo Martingal

O resíduo Martingal foi introduzido em processos de contagem, depois reescrito para modelos de regressão de análise de sobrevivência, na forma:

$$\hat{m}_i = \delta_i - \hat{e}_i, \tag{3.20}$$

em que δ_i é a variável indicadora de censura e $\hat{e}_i$ é o resíduo de Cox-Snell definido na Equação (3.19). O gráfico de x_{iq} versus $\hat{m}_i$ é indicado para examinar a melhor forma funcional para a covariável x_q em um modelo de regressão.

Resíduo Deviance

Os resíduos *deviance* nos modelos de regressão paramétricos são definidos por

$$\hat{d}_i = sinal(\hat{m}_i)\left[-2\left(\hat{m}_i + \delta_i \log(\delta_i - \hat{m}_i)\right)\right]^{1/2}, \tag{3.21}$$

em que $\hat{m}_i$ é o resíduo Martingal definido na equação (3.20). Esse resíduo facilita a detecção de pontos atípicos. Se o modelo for apropriado, o resíduo Deviance deveria apresentar um comportamento aleatório em torno de zero.

O gráfico dos resíduos Martingal (explicados na seção 3.1.4.3) ou Deviance, versus o tempo, é uma forma de verificar a adequação do modelo ajustado e auxiliam na detecção de observações atípicas (COLOSIMO; GIOLO, 2006).

3.2 Modelagem de Risco e Escore proposto

Nesta seção, será apresentada uma explicação sobre importantes conceitos relacionados a modelagem de risco e, além disso, será proposto um escore que avalia o risco de um determinado cliente, com a incorporação da informação do tempo e do saldo devedor, a partir do primeiro atraso de pagamento de alguma parcela do empréstimo.

3.2.1 Introdução à modelagem de risco

Desde os primórdios do sistema financeiro, é possível observar o seguinte fenômeno: a tomada de um certo valor monetário, com o intuito de utilizá-lo para um determinado fim e seu devido retorno acrescido de juros para o credor. Esse fenômeno pode ser denominado como concessão de crédito e pode ocorrer nas mais diversas situações, como um empréstimo feito, para um pequeno agricultor, que visa adquirir máquinas para seu modesto empreendimento ou, indo para um outro extremo, a concessão de um capital, para um cliente que pretende construir um complexo residencial de luxo.

Por meio desse fenômeno, a criação ou desenvolvimento de negócios é fomentada, uma vez que uma quantia financeira relevante é capturada para empreender. Muitas vezes, sem a utilização desse instrumento, não seria possível seguir com algum projeto, pois não haveria capital suficiente.

Sendo assim, é notório que a concessão de crédito está correlacionada com a expansão e prosperidade da economia.

Dito isso, é importante ressaltar o aspecto do risco que é intrínseco, em maior ou menor grau, a todo empréstimo. O risco, no contexto deste trabalho, consiste na probabilidade de um indívuo não cumprir com suas obrigações financeiras que foram determinadas no contrato.

Isso pode ocorrer devido a diversas causas. Em uma perspectiva *macro*, é possível que fatores externos, como uma crise financeira que o país está enfrentando ou decisões políticas arbitrárias que prejudicam determinado setor econômico dificultem, de modo grave, o retorno do capital tomado pelo cliente. Outro motivo, em um ponto de vista *micro*, poderia ser justamente o fato de que o empreendimento auxiliado pela concessão de crédito não foi bem sucedido e o cliente dependia disso para honrar seus cumprimentos financeiros.

Portanto, é de interesse das instituições financeiras avaliar o risco associado ao cliente, antes de realizar a concessão de crédito, com o objetivo de evitar casos de inadimplência e, consequentemente, tornar essa operação lucrativa, pois a rentabilidade está relacionada ao número de empréstimos feitos e à porcentagem de clientes que conseguiram regularizar suas posições.

Nesse contexto, ao longo dos anos, os Modelos de *Credit Scoring* surgiram como um instrumento estatístico, para quantificar o risco de crédito e para predizer se um determinado cliente, dada uma regra de classificação, pode ser considerado como um bom ou mau pagador.

Além dessa característica principal, essa ferramenta possibilita aos credores um meio rápido, automático, padronizado e objetivo para tomar decisões, conforme exposto em Machado (2015). Antes do surgimento dessa técnica, a análise do risco de crédito era feito por um analista, com base predominante em sua experiência, o que implicava em um alto grau de subjetividade.

A partir da década de 1990, os Modelos de *Credit Scoring* começaram a ser amplamente utilizados. Com o desenvolvimento acelerado dos mercados financeiros e avanço tecnológico, o crédito assume uma posição de destaque na economia e esse setor se torna cada vez mais competitivo. Para ter uma certa vantagem, as grandes instituições financeiras costumam negociar com uma quantidade altíssima de clientes por meio de uma margem de lucro pequena em suas operações.

Sendo assim, é fundamental que os credores assumam riscos de maneira inteligente e estratégica, para que a saúde financeira do negócio não seja prejudicada.

3.2.2 Escore proposto que considera a informação do tempo

Uma parcela considerável dos Modelos de *Credit Scoring*, principalmente nos primórdios de sua implementação, têm uma base metodológica na Regressão Logística, a qual possui uma limitação de analisar apenas a probalidade de um determinado cliente honrar seu compromisso financeiro, sem a utilização da informação do tempo. É intuitivo pensar que um cliente que atrasa um pagamento em alguma das primeiras prestações possui um risco maior quando comparado com um cliente que atrasa nas últimas prestações. Então, o tempo é um dado importante.

Sendo assim, a técnica de Análise de Sobrevivência entra em destaque, pois esta é capaz de estudar o tempo estimado até que o evento de interesse ocorra (inadimplência ou atraso no pagamento). Além disso, por meio dela, é possível considerar censuras, ou seja, registros em que o evento de interesse não foi observado.

Como afirmado anteriormente, é de fundamental importância categorizar um cliente conforme seu risco. Por exemplo, cliente de baixo risco ou de alto risco. A predição dessa classificação pode ser feita por meio de um escore de risco que, neste trabalho, irá avaliar o saldo devedor do cliente e o tempo até o possível primeiro atraso de pagamento de parcela do empréstimo/financiamento. Visto que saldo devedor está diretamente associado ao tempo de vigência do contrato, esse escore de risco será baseado na estimativa da função de sobrevivência obtida, pelo ajuste do Modelo de Riscos Proporcionais de Cox.

Sob as suposições de que o saldo devedor decresce, proporcionalmente, à medida que o tempo passa (Sistema de Amortização Constante - SAC), que o tomador do empréstimo não adianta o pagamento das prestações e que o tempo é medido em meses completos (t discreto), o escore de risco proposto neste trabalho é definido por:

$$ES = \sum_{t=0}^{N} [1 - \hat{S}(t)] \left[1 - \frac{t}{N}\right] V, \tag{3.22}$$

em que $\hat{S}(t)$ é a função de sobrevivência estimada pelo Modelo de Riscos Proporcionais de Cox (3.12), N é a vigência do contrato e V é o valor do empréstimo. Vale dizer que o cliente deve, em teoria, pagar uma prestação a cada mês inteiro que passa.

Neste trabalho, optou-se em estimar a função de sobrevivência por meio do modelo de riscos proporcionais de Cox, porque é um dos modelos de regressão mais populares em Análise de Sobrevivência. No entanto, a definição do escore de risco (3.22) pode também ser aplicado para estimativas paraméricas de $S(t)$, considerando o tempo discreto (NAKAGAWA; OSAKI, 1975; NAKANO; CARRASCO, 2006; NEKOUKHOU; BIDRAM, 2015; BRUNELLO; NAKANO, 2015; CARDIAL; FACHINI-GOMES; NAKANO, 2020).

É importante ressaltar o seguinte aspecto sobre o escore proposto: o número em si calculado não possui intrinsecamente uma interpretação e pode variar de maneira bastante severa, quando parâmetros, como valor e número de prestações do empréstimo são alterados.

O escore proposto para um determinado cliente possui sua relevância, quando utilizado em uma análise comparativa com os escores dos outros clientes do conjunto de dados e, nessa situação, quanto maior o seu valor maior seria o seu risco e, analogamente, quanto menor o seu valor menor seria o risco.

O cálculo dos escores de risco propostos serão implementados por meio do *software R*, com a utilização da biblioteca *survival* para realizar o ajuste e diagnóstico do Modelo de Riscos Proporcionais de Cox.

3.2.3 Critério de acurácia do modelo

Após o ajuste do modelo e o cálculo do escore proposto para toda a base de dados, é possível separar os clientes em duas categorias de risco: baixo ou alto. A definição dessas categorias é bastante dependente da dispersão e da magnitude dos valores e de número de prestações dos empréstimos. Cabe ao analista financeiro adotar o critério que tenha sentido negocial para a sua tarefa.

Neste trabalho, será adotada a seguinte regra de decisão baseada no saldo devedor: os clientes que, possivelmente, no primeiro atraso de pagamento, possuem um saldo devedor menor ou igual a um determinado valor são classificados como de baixo risco e os clientes que possuem um saldo devedor acima desse valor, nessa situação, como de alto risco.

Após essa classificação, de maneira análoga, seria definido um corte ótimo para o escore proposto, com o objetivo de definir as predições de categorias de risco. De maneira detalhada, esse procedimento consiste em, sequencialmente, selecionar valores do escore e, para cada iteração, computar as métricas de *performance*. No final, o analista financeiro seria capaz, por exemplo, de olhar em uma tabela todas essas informações e adotar o corte ótimo que cumpra com os seus fundamentos financeiros.

As métricas de performance do escore proposto correspondem às tradicionais medidas utilizadas em modelos binários de classificação, que foram renomeadas para o contexto de

concessão crédito. Essas métricas são obtidas, a partir da Matriz de Confusão, apresentada na Tabela 1.

Tabela 1 – Exemplo de Matriz de Confusão

		Predição do Risco	
		Baixo (-)	Alto (+)
Risco Verdadeiro	Baixo (-)	TN (True Negative)	FP (False Positive)
	Alto (+)	FN (False Negative)	TP (True Positive)

A partir dessa matriz de confusão, as seguintes métricas são apresentadas:

$$\begin{aligned}
\text{Acurácia} &= \frac{TN + TP}{TN + FP + FN + TP} \\
\text{Precisão de Alto Risco} &= \frac{TP}{TP + FP} \\
\text{Precisão de Baixo Risco} &= \frac{TN}{TN + FN} \\
\text{Erro da Precisão de Alto Risco} &= 1 - \text{Precisão de Alto Risco} \\
\text{Erro da Precisão de Baixo Risco} &= 1 - \text{Precisão de Baixo Risco} \\
\text{Sensibilidade de Alto Risco} &= \frac{TP}{TP + FN} \\
\text{Sensibilidade de Baixo Risco} &= \frac{TN}{TN + FP} \\
\text{Erro da Sensibilidade de Alto Risco} &= 1 - \text{Sensibilidade de Alto Risco} \\
\text{Erro da Sensibilidade de Baixo Risco} &= 1 - \text{Sensibilidade de Baixo Risco}
\end{aligned}$$

As interpretações para as métricas são apresentadas a seguir:

- A acurácia é medida que corresponde a proporção de acerto do modelo considerando toda a amostra dos dados;
- A precisão de alto risco é medida que corresponde a proporção de clientes que, de fato, são de alto risco, dado que modelo prediz alto risco;
- A precisão de baixo risco é medida que corresponde a proporção de clientes que, de fato, são de baixo risco, dado que modelo prediz baixo risco;
- A sensibilidade de alto risco é a medida que corresponde a proporção de acerto quando o modelo prediz alto risco, dado que a amostra é apenas de clientes que, de fato, são de alto risco;

- A sensibilidade de baixo risco é a medida que corresponde a proporção de acerto quando o modelo prediz baixo risco, dado que a amostra é apenas de clientes que, de fato, são de baixo risco;
- O erro da sensibilidade de alto risco é a proporção do número de predições com classificação baixo risco, dado que a amostra era de apenas clientes de alto risco;
- O erro da sensibilidade de baixo risco é a proporção do número de predições com classificação alto risco, dado que a amostra era de apenas clientes de baixo risco.

Algumas considerações são importantes a serem feitas sobre a *performance* da classificação com a utilização do escore.

Consideração 1 (Em relação ao erro da sensibilidade)

- **Erro 1**: Erro da Sensibilidade de Alto Risco;
- **Erro 2**: Erro da Sensibilidade de Baixo Risco.

Como afirmado anteriormente, as instituições financeiras costumam lidar com uma quantidade alta de clientes e com uma baixa margem de lucro. Sendo assim, conceder o capital de modo amplo, com uma porcentagem baixíssima de maus empréstimos é uma estratégia rentável, nesse cenário.

Portanto, o Erro 1 é pior que o Erro 2, caso o objetivo seja identificar, com um alto grau de certeza, quais são os clientes de alto risco, evitando conceder crédito para esses indivíduos. Isso implica que é mais importante buscar uma sensibilidade de alto risco elevada e maior que a sensibilidade de baixo risco.

Consideração 2 (Em relação ao erro da precisão)

- **Erro 1**: Erro da Precisão de Baixo Risco;
- **Erro 2**: Erro da Precisão de Alto Risco.

Nessa linha de raciocínio, é evidente que emprestar capital para alguém que é de alto risco, dado que modelo classificou como baixo risco é um erro pior que deixar de conceder crédito para alguém de baixo risco dado que o modelo classificou como alto risco, pois, no primeiro erro, implicaria em tomar um risco muito grande do cliente se tornar inadimplente. Isso implica que é mais importante buscar uma precisão de baixo risco elevada e maior que a precisão de alto risco.

Além dessas métricas de *performance*, é possível utilizar a curva ROC (*Receiver Operating Characteristic Curve*), que possibilita avaliar, em um gráfico com dois eixos, a sensibilidade de alto risco e o erro da sensibilidade de baixo risco. A partir dessa curva, é possível obter a área

sob a curva ROC (AUC - *Area Under the Curve*), que é uma medida utilizada para mensurar o poder de classificação. O AUC varia entre 0 e 1, em que 1 significa uma classificação perfeita e $0,5$ significa que o escore de risco não tem capacidade de separação de classe (área sob a reta identidade).

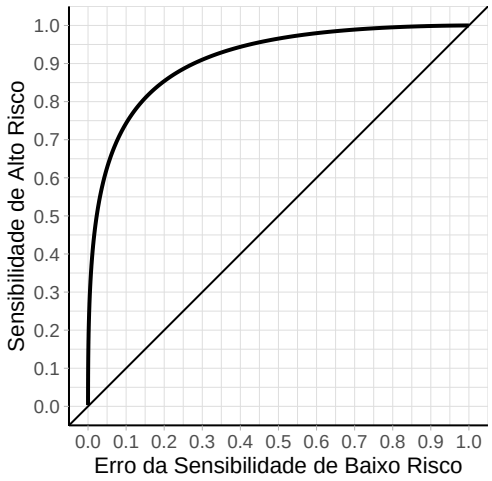

Figura 2 – Exemplo de Curva ROC

3.3 Sobre o conjunto de dados simulados

O foco deste trabalho, como dito anteriormente, é o escore e seu desempenho na classificação do risco. No intuito de estudá-lo em um cenário mais controlado, o Modelo de Riscos Proporcionais de Cox foi ajustado em um conjunto de dados simulados.

Esse conjunto de dados possui $n = 100.000$ observações e, inicialmente, três covariáveis foram geradas com as seguintes distribuições de probabilidade:

- $X_1 \sim \mathcal{N}(\mu = 5,\, \sigma^2 = 1)$;
- $X_2 \sim \mathcal{N}(\mu = 1,\, \sigma^2 = 4)$;
- $X_3 \sim \mathrm{B}ernoulli(p = 0,2)$.

Essas covariáveis correspondem às informações dos clientes que poderiam ser observadas em um conjunto de dados reais.

Além disso, outras variáveis foram geradas: o valor do empréstimo (V) e o número de prestações a pagar pelo cliente (N), com as seguintes distribuições de probabilidade:

- $V \sim \mathcal{N}(\mu = 10^5,\, \sigma^2 = 10^8)$;
- $N \sim \mathrm{P}oisson(\lambda = 50)$.

Para gerar o tempo de sobrevivência t, o processo é mais complexo, pois deve-se levar em consideração a possibilidade de censura. Além disso, o tempo não pode ser maior que o número de prestações. O Algoritmo 1 apresenta o procedimento utilizado para gerar as amostras com censura aleatória, em que Y é a distribuição do tempo de sobrevivência desejado, C é a distribuição dos tempos de censura, N é a distribuição do número de prestações de contrato do empréstimo/financimento e δ é o indicador de censura. Todas essas informações são utilizadas, ao longo do algoritmo, para definir o tempo de sobrevivência t, que é de fato utilizado no ajuste do modelo, conforme descrito a seguir:

Algoritmo 1 Algoritmo para gerar amostras com censura aleatória

1: **Passo 1**: Obter os valores dos paramêtros da distribuição do tempo de censura (C), a partir dos parâmetros da distribuição do tempo de sobrevivência desejado (Y);
2: **Passo 2**: Gerar $Y_1, Y_2, \ldots, Y_n$ da distribuição de interesse;
3: **Passo 3**: Gerar $C_1, C_2, \ldots, C_n$ da distribuição de censura, com os parâmetros definidos no Passo 1;
4: **Passo 4**: Gerar $N_1, N_2, \ldots, N_n$ com distribuição $N_i \sim \text{P}oisson(\lambda)$;
5: **Passo 5**: Para todo $i \in \{1, \ldots, n\}$:

- **Passo 5.1**: Se $Y_i < C_i$ e $Y_i \leq N_i$, fazer $\delta_i \leftarrow 1$ e $t_i \leftarrow Y_i$;
- **Passo 5.2**: Se $Y_i > C_i$ e $C_i \leq N_i$, fazer $\delta_i \leftarrow 0$ e $t_i \leftarrow C_i$;
- **Passo 5.3**: Se $Y_i = C_i$ e $Y_i \leq N_i$, fazer $\delta_i \sim \text{B}ernoulli(p = 0,5)$ e $t_i \leftarrow Y_i$;
- **Passo 5.4**: Se $min(Y_i, C_i) > N_i$, fazer $\delta_i \leftarrow 0$ e $t_i \leftarrow N_i$.

A distribuição do tempo de sobrevivência desejado (Y) foi gerado por meio de um modelo de regressão exponencial, considerando a função de ligação *log*. Essa escolha é justificada pelo fato do modelo exponencial ser popular em dados de Análise de Sobrevivência e, além disso, por apresentar uma estrutura de riscos proporcionais (COLOSIMO; GIOLO, 2006), que é coerente com o Modelo de Riscos Proporcionais de Cox implementado neste trabalho.

Sendo assim, os valores da distribuição de interesse (Y) foram gerados, a partir de uma distribuição Exponencial ($Exp(\lambda_1)$), em que:

$$\lambda_1 = \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3). \tag{3.23}$$

Neste trabalho, considerou-se $\beta_0 = -0,8$, $\beta_1 = -0,7$, $\beta_2 = -0,2$ e $\beta_3 = -0,1$.

De forma similar ao procedimento apresentado por Biazatti e Nakano (2020), para a distribuição Weibull (a exponencial é um caso particular da Weibull), os tempos de censura (C) foram gerados a partir de uma distribuição Exponencial ($Exp(\lambda_2)$). O valor de λ_2 depende do percentual de censura na amostra $p \times 100\%$ e é obtido pela seguinte fórmula:

$$\lambda_2 = \left(\frac{p}{1-p}\right) \lambda_1. \tag{3.24}$$

Neste trabalho, foi utilizado $p = 0,10$.

Tanto os valores da distribuição do tempo de sobrevivência desejado (Y) quanto os tempos de censura (C) gerados foram discretizados (tomou-se o maior valor inteiro menor ou igual ao gerado).

Como descrito anteriormente, é de interesse saber o saldo devedor do cliente. Neste trabalho, essa informação corresponde ao valor monetário que falta ser pago pelo cliente, a partir do primeiro atraso de pagamento da parcela. É importante ressaltar que, em dados simulados, o saldo devedor pode ser obtido até para os casos de censura, quando o tempo censurado é menor que o número de prestações. No caso de tempo censurado maior que o número de prestações, o saldo devedor é igual a zero (nesta situação, o cliente não atrasa durante a vigência do empréstimo/financiamento). Dito isso, pode-se definir o saldo devedor como:

$$saldo_{devedor} = \begin{cases} V - \frac{(t-1)\,V}{N}, & \text{se } \delta = 1, \\ V - \frac{min(Y,N)\,V}{N}, & \text{se } \delta = 0, \end{cases} \tag{3.25}$$

em (3.25), t é o tempo de sobrevivência (meses completos) até o primeiro atraso de pagamento de parcela do empréstimo, gerado pelo Algoritmo 1, Y é o valor gerado da distribuição de interesse, N é o numéro de prestações do contrato de empréstimo a pagar pelo cliente, V é o valor do crédito concedido e δ é o indicador de censura.

Em resumo, a estrutura do conjunto de dados simulados, no final, é a seguinte:

- X_1: primeira covariável gerada aleatoriamente;
- X_2: segunda covariável gerada aleatoriamente;
- X_3: terceira covariável gerada aleatoriamente;
- t: tempo, em meses completos, até o primeiro atraso de pagamento de parcela do empréstimo;
- δ: variável indicadora de censura;
- N: número de prestações do contrato de empréstimo a pagar pelo cliente;
- V: valor do crédito concedido;
- $saldo_{devedor}$: saldo devedor do cliente.

4 Resultados

Como apresentado na Seção 3.3, a metodologia proposta neste trabalho foi ilustrada com base em uma amostra de 100.000 clientes e considerando 3 covariáveis. Foi utilizada a regra de decisão que classifica os clientes com saldo devedor abaixo de $9000,00$, como baixo risco e, caso contrário, como de alto risco. Vale ressaltar que outros cortes de saldo devedor poderiam ser utilizados, para classificar os clientes. Neste trabalho, foi considerado que o valor do corte escolhido do saldo devedor é condizente, pois representa uma fração pequena ($< 10\%$), em relação à media dos valores de contrato de empréstimo/financiamento.

As Tabelas 2 e 3 apresentam uma análise descritiva da amostra considerada:

Tabela 2 – Medidas-resumo de cada variável quantitativa

Variável	**Mín**	**1º Quartil**	**Mediana**	**3º Quartil**	**Máx**	**Média**	**Desvio Padrão**
X_1	0,35	4,32	5,00	5,68	9,50	5,00	1,01
X_2	-7,73	-0,36	1,00	2,36	9,11	1,00	2,01
N	18	45	50	55	82	50,02	7,05
V	57339,00	93302,00	100001,00	106738,00	143629,00	100050,00	9956,71
$saldo_{devedor}$	0	0	0	59344,00	143629,00	27626,00	36258,94

Tabela 3 – Distribuição das variáveis categóricas

Variável	**Nível**	**Frequência Simples**	**Frequência Relativa**
X_3	0	80004	80,00%
	1	19996	20,00%
δ	0	56417	56,42%
	1	43583	43,58%
Risco	Alto	56770	55,77%
	Baixo	43230	43,23%
$saldo_{devedor} > 0$	Não	54493	54,49%
	Sim	45507	45,51%

Pela Tabela 2, é evidente que os valores do crédito concedido possuem uma dispersão razoavelmente baixa, não há valores de contratos de milhões de reais e também não há valores de poucas centenas de reais. É interessante ressaltar também que o saldo devedor só passa a assumir valores maiores que zero a partir da mediana.

Além disso, pela Tabela 3, nota-se que mais da metade dos clientes tiveram tempo censurado ($56,42\%$) e classificação risco alto (55,77%). Em relação a última variável indicadora de atraso, $54,49\%$ não atrasaram pagamento de prestação do empréstimo/financiamento (bons pagadores).

A Figura 3 apresenta a estimativa de Kaplan-Meier da função de sobrevivencia do tempo até a inadimplência da amostra.

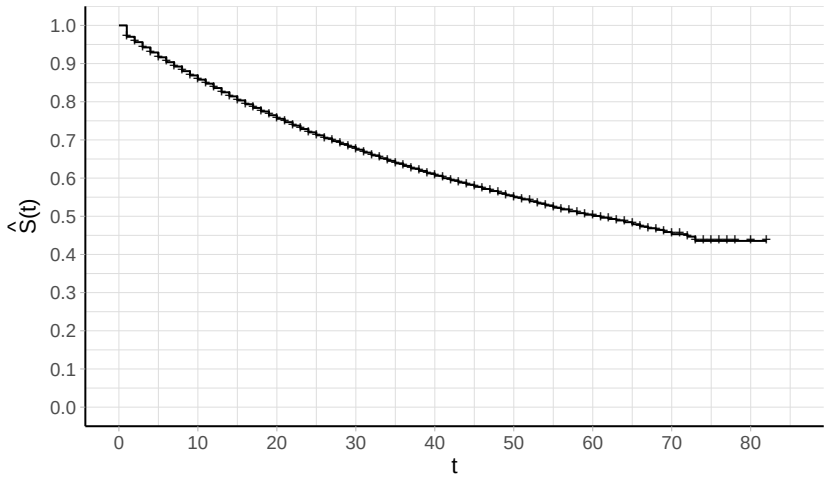

Figura 3 – Função de sobrevivência estimada por Kaplan-Meier

Nesta amostra, foi observado um tempo médio de 52,1 meses e um tempo mediano de 61,10 meses (ambos obtidos por meio da estimativa de Kaplan-Meier), aproximadamente.

4.1 Ajuste do Modelo de Cox

Após a construção do conjunto de dados simulados, foi ajustado o Modelo de Riscos Proporcionais de Cox. Na Tabela 4, os resultados são encontrados:

Tabela 4 – Resultados dos coeficientes

Parâmetro	**Estimativa**	**exp(Estimativa)**	**Erro Padrão(Estimativa)**
β_1	$-0,699$	$0,497$	$0,005$
β_2	$-0,198$	$0,821$	$0,002$
β_3	$-0,109$	$0,897$	$0,012$

Uma vez que os dados foram criados de modo artificial e considerando a estrutura de riscos proporcionais, era esperado que as estimativas tivessem p-valores virtualmente iguais a zero. Além disso, a Tabela 5 apresenta o teste de tendências dos resíduos de Schoenfeld, indicando que a suposição de proporcionalidade dos riscos do Modelo de Riscos Proporcionais de Cox foram atendidas.

Tabela 5 – Teste de tendência dos resíduos de Schoenfeld

Variável/Global	$\boldsymbol{\chi^2}$	**Graus de Liberdade**	**p-valor**
X_1	$0,65$	1	$0,42$
X_2	$1,25$	1	$0,26$
X_3	$7,92 \times 10^{-5}$	1	$0,99$
Global	$2,10$	3	$0,55$

4.2 Cálculo do escore de risco e classificação dos indivíduos conforme risco

4.2.1 Ilustração com exemplo de um cliente

É interessante tomar um exemplo de algum cliente do conjunto de dados para ilustrar a metodologia. Sendo assim, considere um cliente com as seguintes características:

- $X_1 = 0,3505$, $X_2 = 0,5844$ e $X_3 = 0$;
- Valor concedido de crédito igual a $V = 83332,69$;
- Número de prestações igual a $N = 42$.

Com essas informações e com a Tabela 6, que apresenta a função de sobrevivência base estimada, para cada tempo até o fim do período de pagamento, é possível calcular o escore de risco desse indivíduo.

Tabela 6 – Tempo e Função de Sobrevivência Base Estimada

t	$\hat{S}_0(t)$	t	$\hat{S}_0(t)$	t	$\hat{S}_0(t)$
0	1,000	14	0,849	28	0,729
1	0,978	15	0,840	29	0,721
2	0,967	16	0,830	30	0,713
3	0,957	17	0,822	31	0,705
4	0,946	18	0,813	32	0,698
5	0,936	19	0,804	33	0,690
6	0,926	20	0,795	34	0,682
7	0,917	21	0,786	35	0,675
8	0,907	22	0,777	36	0,668
9	0,897	23	0,768	37	0,660
10	0,888	24	0,760	38	0,653
11	0,879	25	0,752	39	0,646
12	0,869	26	0,744	40	0,639
13	0,859	27	0,737	41	0,632
⋮	⋮	⋮	⋮	42	0,626

Para esse cliente, a função de sobrevivência estimada é dada por

$$\begin{aligned}
\hat{S}(t|\mathbf{x}) &= \hat{S}_0(t|\mathbf{x})^{\exp(\hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3)} \\
&= \hat{S}_0(t|\mathbf{x})^{\exp(-0,699 X_1 - 0,198 X_2 - 0,109 X_3)} \\
&= \hat{S}_0(t|\mathbf{x})^{\exp(-0,699 \times 0,3505 \, - 0,198 \times 0,5844 \, - 0,109 \times 0)} \\
&= \hat{S}_0(t|\mathbf{x})^{\exp(-0,3607)}, \qquad (4.1)
\end{aligned}$$

em que $\hat{S}_0(t|\mathbf{x})$ é a função de sobrevivência base estimada apresentada na Tabela 6.

A Figura 4 apresenta a função de sobrevivência estimada (pelo Modelo de Riscos Proporcionais de Cox), para esse cliente:

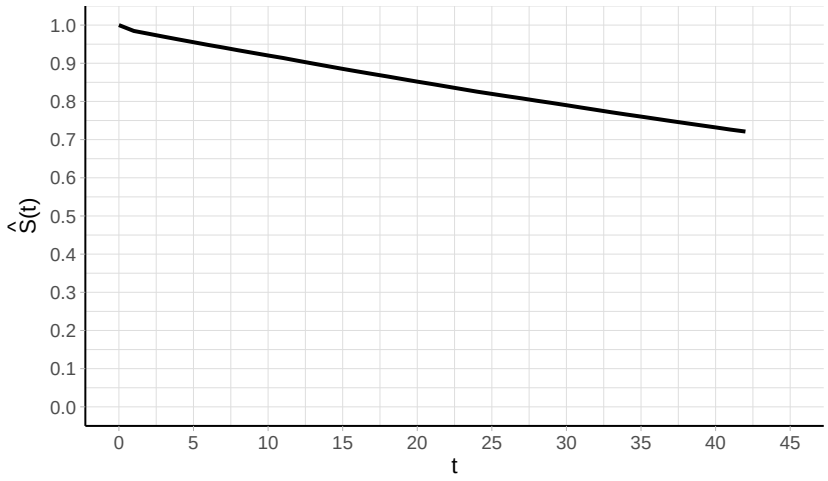

Figura 4 – Função de sobrevivência estimada do cliente com $X_1 = 0,3505$, $X_2 = 0,5844$, $X_3 = 0$, $V = 83332,69$ e $N = 42$

Para esse cliente, o escore de risco computado foi $ES = 184074,1$. Além disso, é interessante notar que a probabilidade de sobrevivência é igual a, aproximadamente, 72,11%, no tempo de 42 meses completos.

4.2.2 Ilustração com a amostra de dados

É importante também avaliar o comportamento do escore, levando em consideração a amostra de dados integralmente.

A Figura 5 apresenta a distribuição empírica e as médias dos escores de risco, representadas por linhas traçadas verticalmente, com todos os indivíduos da amostra, conforme a classificação do risco (baixo/alto).

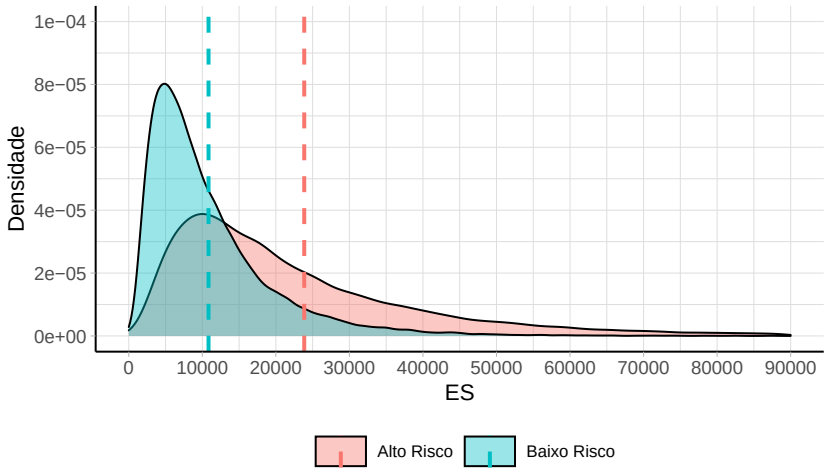

Figura 5 – Funções de densidade e médias do escore por classificação de risco

Apesar da interseção considerável das curvas, nota-se que, como esperado os valores de ES são maiores para aqueles clientes classificados como alto risco.

A instituição financeira, por meio desse escore de risco, poderia ordenar (*ranking*) a base de dados de oportunidades de clientes e identificar quais são os de perfil mais arriscado e quais os de perfil mais seguro e, em seguida, elaborar políticas de concessão de crédito customizadas para cada grau de risco.

As Tabelas 7 e 8 ilustram esses *rankings*: a Tabela 7 com os 20 clientes com os maiores escores de risco e a Tabela 8 com os clientes com os menores escores de risco.

Tabela 7 – 20 clientes com maiores escores de risco

X_1	X_2	X_3	t	δ	N	V	$saldo_{devedor}$	ES
2,13	-4,69	0	1	1	64	100307,48	100307,48	404152,99
1,32	-1,31	1	1	1	64	115279,75	115279,75	383214,25
2,33	-4,74	0	2	1	57	100550,76	98786,71	288652,26
2,21	-2,53	0	9	1	65	103070,10	90384,55	273420,10
2,61	-3,57	0	1	0	68	99325,11	93482,45	268428,55
2,28	-1,64	0	3	1	63	131480,41	127306,42	266068,98
1,91	-2,67	0	2	1	55	109848,12	107850,88	265454,24
1,22	-1,70	0	3	1	51	95816,96	92059,44	262494,99
2,19	-1,86	1	1	1	70	106136,48	106136,48	261304,31
2,25	-5,91	0	3	1	44	109000,71	104046,13	252452,27
2,56	-4,96	0	4	1	52	115242,86	108594,23	250701,49
2,34	-2,71	0	5	1	62	106158,89	99309,93	243939,67
2,73	-4,43	0	1	1	61	100523,01	100523,01	238735,71
1,10	-0,82	0	1	1	48	104533,85	104533,85	236168,14
1,74	-2,39	0	2	1	52	102068,93	100106,07	235453,78
2,40	-4,98	0	6	1	52	95551,15	86363,54	230660,75
2,17	-2,93	0	10	0	55	107082,96	87613,33	228284,81
2,33	-2,45	0	25	1	60	108803,82	65282,29	225807,30
2,16	-2,11	0	1	1	55	122215,31	122215,31	225516,57
1,76	-0,63	0	7	1	57	114626,10	102560,20	223455,75

Tabela 8 – 20 clientes com menores escore de risco

X_1	X_2	X_3	t	δ	N	V	$saldo_{devedor}$	ES
6,20	3,04	0	18	0	18	106024,50	0,00	517,24
7,77	5,33	0	40	0	40	106616,40	0,00	507,92
7,56	4,74	0	37	0	37	93479,84	0,00	500,54
7,57	4,45	1	38	0	38	90202,12	0,00	477,59
6,33	8,00	0	31	0	31	100024,68	0,00	470,65
6,41	7,71	1	33	0	33	98096,75	0,00	469,15
7,67	6,12	1	42	0	42	106410,04	0,00	459,65
9,50	1,90	0	52	0	52	97642,54	0,00	457,59
8,30	6,05	1	50	0	50	111285,95	0,00	438,96
8,58	0,78	1	34	0	34	95819,52	0,00	417,63
7,88	7,20	0	45	0	45	105957,52	0,00	405,85
9,17	3,49	0	54	0	54	83486,48	0,00	386,81
8,16	4,06	0	37	0	37	95693,99	0,00	385,24
7,46	6,36	1	42	0	42	79606,21	0,00	379,16
8,25	4,61	1	46	0	46	78281,86	0,00	362,23
9,18	2,49	1	47	0	47	91929,69	0,00	352,60
7,17	4,92	1	30	0	30	82224,95	0,00	333,37
8,19	7,32	0	44	0	44	108922,16	0,00	314,31
7,98	8,92	0	43	0	43	111106,72	0,00	259,14
8,85	5,13	0	43	0	43	94672,61	0,00	253,76

Note que os 20 maiores escores são de indivíduos de alto risco com saldos devedores bastante elevados e os 20 menores, de baixo risco, com saldo devedor igual a zero.

Por fim, a curva ROC foi adotada para estudar o poder de classificação do escore de risco e é apresentada na Figura 6.

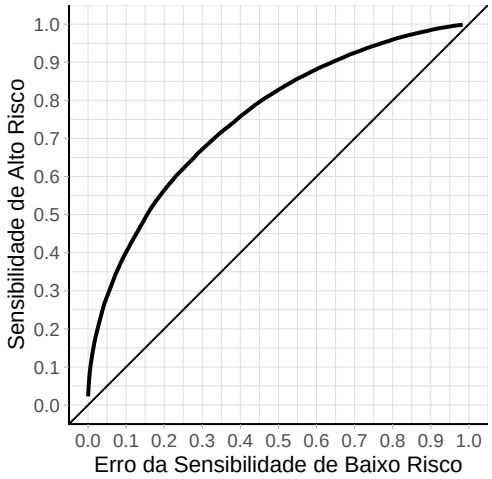

Figura 6 – Curva ROC

Neste exemplo, foi observado AUC (área sob a curva ROC) igual a $75,55\%$, o que indica um moderado poder classificatório.

É evidente que um aumento na sensibilidade de alto risco implica em um aumento no

erro da sensibilidade de baixo risco. Cabe ao analista financeiro escolher um corte que cumpra com o critério de aceitação do risco para o seu negócio.

Com a utilização da regra de decisão que classifica os clientes com saldo devedor abaixo de $9000,00$, como baixo risco e, caso contrário, como de alto risco, diversos pontos de corte para o escore de risco poderiam ser utilizados, de acordo com o critério aceitação de risco (Apêndice A). Dois cenários relevantes serão ilustrados:

Cenário com tolerância moderada ao erro

Apesar da AUC ter apresentado um resultado moderado (75,55%), como visto na Seção 4.2.1, já era previsto que não é possivel obter uma clara e perfeita distinção dos escores de risco entre os clientes de baixo e os clientes de alto risco, o que é evidenciado também na interseção das áreas sob as curvas na Figura 5. Como explicado na Seção 3.2.3, é mais importante obter-se:

- Uma sensibilidade de alto risco elevada e maior que a sensibilidade de baixo risco;
- Uma precisão de baixo risco elevada e maior que precisão de alto risco.

É possível determinar um corte, com o intuito de possuir uma tolerância moderada ao erro da precisão de baixo risco e ao erro da sensibilidade de alto risco. Sendo assim, para um corte $ES > 10287,28$, os seguintes resultados seriam encontrados:

Tabela 9 – Matriz de confusão no cenário com tolerância moderada ao erro

		Predição do Risco	
		Baixo (-)	Alto (+)
Risco Verdadeiro	Baixo (-)	34343	22427
	Alto (+)	10657	32573

A Tabela 10 apresenta as métricas de performance, considerando o ponto de corte $ES > 10287,28$.

Tabela 10 – Métricas de performance no cenário com tolerância moderada ao erro

Acurácia	**Sensibilidade**		**Precisão**	
	Baixo Risco	Alto Risco	Baixo Risco	Alto Risco
66,92%	60,49%	75,35%	76,32%	59,22%

Com esse corte, apesar das métricas de precisão de alto risco (59,22%) e de sensibilidade de baixo (60,49%) assumirem valores não tão altos, as métricas de precisão de baixo risco (76,32%) e de sensibilidade de alto risco (75,35%) estão elevadas, o que é mais importante, conforme foi descrito na Seção 3.2.3

Portanto, para esse corte adotado, em 75,35% das vezes em que os clientes eram, de fato, alto risco, modelo classificou corretamente e, em cerca de 76,32% das vezes, o cliente era

de baixo risco, dado que o modelo classificou como baixo. É evidente que há uma margem moderada de erro para a precisão de baixo risco (23,68%) e para a sensibilidade de alto risco (24,65%).

Cenário com tolerância baixa ao erro

Caso o objetivo seja evitar rigorosamente o risco de conceder crédito, para clientes de alto risco, e identificar com grande grau de certeza quais são esses clientes, é possível adotar o corte $ES > 4691,15$. Para esse ponto de corte adotado, os seguintes resultados foram encontrados:

Tabela 11 – Matriz de confusão no cenário com tolerância baixa ao erro

		Predição do Risco	
		Baixo (-)	Alto (+)
Risco Verdadeiro	Baixo (-)	12871	43899
	Alto (+)	2129	41101

A Tabela 12 apresenta as métricas de performance, considerando o ponto de corte $ES > 4691,15$.

Tabela 12 – Métricas de performance no cenário com tolerância baixa ao erro

Acurácia	**Sensibilidade**		**Precisão**	
	Baixo Risco	Alto Risco	Baixo Risco	Alto Risco
53,97%	22,67%	95,08%	85,81%	48,35%

Com esse corte, apesar das métricas de precisão de alto risco (48,35%) e de sensibilidade de baixo risco (22,67%) assumirem valores baixos, as métricas de precisão de baixo risco (85,81%) e de sensibilidade de alto risco (95,08%) estão bastante elevadas.

Portanto, para esse corte adotado, em 95,08% das vezes em que os clientes eram, de fato, alto risco, modelo classificou corretamente, o que demonstra uma capacidade muito boa de identificação para esses clientes arriscados. Além disso, 85,81% dos clientes classificados como baixo risco pelo modelo são, de fato, baixo risco, o que indica uma elevada confiabilidade nas predições de baixo risco.

A desvantagem é o fato do corte de escore adotado tender a classificar a maioria dos clientes como alto risco, o que implicaria em analisar pouquíssimas oportunidades de concessão de crédito, para os clientes classificados como de baixo risco.

5 Conclusão

O presente trabalho teve como objetivo propor um escore de risco capaz de discriminar clientes conforme o seu risco de gerar um prejuízo financeiro devido à inadimplência. Este escore de risco foi baseado no saldo devedor do cliente, no momento em que ocorre o primeiro atraso de pagamento de prestação do empréstimo/financiamento.

O escore de risco proposto tem como base a função de sobrevivência, que foi estimada, neste trabalho, por meio do Modelo de Riscos Proporcionais de Cox. Ademais, esse escore de risco também pode ser obtido a partir de estimativas paramétricas da função de sobrevivência, considerando modelos para tempos discretos.

A metodologia proposta foi ilustrada por meio de dados simulados no R e os resultados obtidos mostraram que o escore proposto se mostrou adequado para classificar clientes com alto ou baixo risco de gerar um prejuízo financeiro.

Referências

ANICETO, M. C. *Análise de sobrevivência aplicada ao risco de crédito*. Monografia (Bacharel) — Universidade de Brasília, Brasília, 2011.

BIAZATTI, E. C.; NAKANO, E. Y. Uma proposta de orientação para o uso de modelos contínuos em dados de sobrevivência discretos. *REMAT: Revista Eletrônica da Matemática*, v. 6, n. 2, p. e4002–e4002, 2020.

BRESLOW, N. Contribution to the discussion of cox (1972). *Journal of the Royal Statistical Society Series B*, v. 34, p. 216–217, 1972.

BRUNELLO, G. H. V.; NAKANO, E. Y. Inferência bayesiana no modelo weibull discreto em dados com presença de censura. *TEMA (São Carlos)*, SciELO Brasil, v. 16, p. 97–110, 2015.

CARDIAL, M. R. P.; FACHINI-GOMES, J. B.; NAKANO, E. Y. Exponentiated discrete weibull distribution for censored data. *REVISTA BRASILEIRA DE BIOMETRIA*, v. 38, n. 1, p. 35–56, 2020.

COLOSIMO, E. A.; GIOLO, S. R. *Análise de Sobrevivência Aplicada*. São Paulo: Editora Blucher, 2006.

COX, D. R. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, Wiley Online Library, v. 34, n. 2, p. 187–202, 1972.

GRAMBSCH, P. M.; THERNEAU, T. M. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, Oxford University Press, v. 81, n. 3, p. 515–526, 1994.

LAWLESS, J. F. *Statistical models and methods for lifetime data*. [S.l.]: John Wiley & Sons, 2011. v. 362.

LEE, E. T.; WANG, J. *Statistical Methods for Survival Data Analysis*. [S.l.]: John Wiley & Sons, 2003. v. 476.

MACHADO, A. R. Collection scoring via regressão logística e modelo de riscos proporcionais de cox. 2015.

NAKAGAWA, T.; OSAKI, S. The discrete weibull distribution. *IEEE transactions on reliability*, IEEE, v. 24, n. 5, p. 300–301, 1975.

NAKANO, E. Y.; CARRASCO, C. G. Uma avaliação do uso de um modelo contínuo na análise de dados discretos de sobrevivência. *Trends in Computational and Applied Mathematics*, v. 7, n. 1, p. 91–100, 2006.

NEKOUKHOU, V.; BIDRAM, H. The exponentiated discrete weibull distribution. *Sort*, Institut d'Estadística de Catalunya, v. 39, p. 127–146, 2015.

PETO, R. Contribution to the discussion of cox (1972). *J. Roy. Statist. Soc. Ser. B Stat. Methodol*, v. 34, p. 205–207, 1972.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2020. Disponível em: <https://www.R-project.org/>.

SCHOENFELD, D. Partial residuals for the proportional hazards regression model. *Biometrika*, Oxford University Press, v. 69, n. 1, p. 239–241, 1982.

APÊNDICE A – Resultados das métricas de performance

Tabela 13 – Resultados das métricas de *performance*

Corte	**Escore**	**Acurácia**	**Sensibilidade**		**Precisão**	
			Baixo Risco	Alto Risco	Baixo Risco	Alto Risco
1%	1551,4177	0,4408	0,0163	0,9983	0,9270	0,4359
2%	1962,3073	0,4493	0,0326	0,9966	0,9260	0,4396
3%	2271,4804	0,4572	0,0483	0,9941	0,9143	0,4430
4%	2541,6474	0,4650	0,0640	0,9915	0,9085	0,4465
5%	2779,0810	0,4728	0,0797	0,9891	0,9054	0,4501
6%	3007,3499	0,4801	0,0949	0,9859	0,8982	0,4534
7%	3215,0982	0,4869	0,1097	0,9821	0,8897	0,4565
8%	3414,6176	0,4941	0,1249	0,9789	0,8861	0,4600
9%	3600,1043	0,5010	0,1398	0,9754	0,8819	0,4634
10%	3793,3529	0,5080	0,1547	0,9719	0,8784	0,4668
11%	3980,1487	0,5147	0,1695	0,9681	0,8747	0,4702
12%	4159,9582	0,5214	0,1842	0,9643	0,8714	0,4737
13%	4339,2662	0,5277	0,1985	0,9600	0,8669	0,4770
14%	4517,1876	0,5335	0,2125	0,9551	0,8615	0,4801
15%	4691,1477	0,5397	0,2267	0,9508	0,8581	0,4835
16%	4863,8625	0,5456	0,2407	0,9459	0,8539	0,4868
17%	5040,3958	0,5518	0,2549	0,9415	0,8514	0,4904
18%	5208,7933	0,5575	0,2688	0,9366	0,8478	0,4938
19%	5382,2254	0,5629	0,2824	0,9313	0,8438	0,4971
20%	5541,6509	0,5685	0,2961	0,9262	0,8405	0,5005
21%	5715,9510	0,5743	0,3100	0,9214	0,8381	0,5042
22%	5890,3773	0,5793	0,3232	0,9156	0,8341	0,5074
23%	6067,6649	0,5843	0,3365	0,9098	0,8305	0,5108
24%	6237,2409	0,5895	0,3499	0,9043	0,8276	0,5144
25%	6412,5933	0,5943	0,3629	0,8982	0,8240	0,5177
26%	6591,8695	0,5994	0,3761	0,8925	0,8213	0,5214
27%	6772,0257	0,6045	0,3895	0,8869	0,8190	0,5252
28%	6933,5940	0,6091	0,4023	0,8806	0,8157	0,5287
29%	7115,8945	0,6136	0,4151	0,8743	0,8126	0,5323
30%	7300,2217	0,6179	0,4277	0,8677	0,8093	0,5359
31%	7486,1748	0,6224	0,4405	0,8613	0,8066	0,5397

32%	7662,6018	0,6270	0,4533	0,8551	0,8042	0,5436
33%	7843,6395	0,6307	0,4654	0,8478	0,8007	0,5470
34%	8032,6492	0,6348	0,4778	0,8409	0,7978	0,5508
35%	8223,5398	0,6385	0,4898	0,8336	0,7945	0,5544
36%	8423,1677	0,6423	0,5020	0,8265	0,7916	0,5583
37%	8620,5642	0,6457	0,5138	0,8189	0,7884	0,5619
38%	8828,4606	0,6493	0,5258	0,8115	0,7855	0,5658
39%	9019,7820	0,6529	0,5378	0,8041	0,7828	0,5698
40%	9218,3736	0,6561	0,5494	0,7962	0,7797	0,5736
41%	9435,4754	0,6591	0,5609	0,7881	0,7766	0,5775
42%	9636,2511	0,6616	0,5719	0,7794	0,7730	0,5809
43%	9847,1705	0,6642	0,5829	0,7708	0,7696	0,5846
44%	10068,8938	0,6669	0,5942	0,7625	0,7666	0,5886
45%	10287,2786	0,6692	0,6049	0,7535	0,7632	0,5922
46%	10525,6006	0,6711	0,6154	0,7441	0,7595	0,5957
47%	10743,0165	0,6734	0,6263	0,7352	0,7564	0,5997
48%	10985,7020	0,6760	0,6374	0,7267	0,7538	0,6041
49%	11230,7381	0,6786	0,6485	0,7181	0,7513	0,6087
50%	11474,1515	0,6807	0,6592	0,7090	0,7484	0,6130
51%	11718,2397	0,6824	0,6695	0,6994	0,7452	0,6171
52%	11967,7131	0,6843	0,6799	0,6900	0,7423	0,6215
53%	12211,2614	0,6861	0,6903	0,6805	0,7394	0,6259
54%	12483,0081	0,6879	0,7007	0,6711	0,7367	0,6307
55%	12748,0035	0,6896	0,7111	0,6615	0,7339	0,6355
56%	13024,2475	0,6903	0,7205	0,6508	0,7304	0,6394
57%	13315,0960	0,6915	0,7303	0,6405	0,7274	0,6439
58%	13605,3842	0,6931	0,7405	0,6308	0,7248	0,6493
59%	13904,2699	0,6941	0,7502	0,6204	0,7218	0,6541
60%	14211,5789	0,6956	0,7603	0,6106	0,7194	0,6599
61%	14535,8744	0,6967	0,7701	0,6003	0,7167	0,6654
62%	14883,3353	0,6969	0,7791	0,5889	0,7134	0,6700
63%	15211,9704	0,6976	0,7886	0,5782	0,7106	0,6756
64%	15576,0638	0,6979	0,7976	0,5669	0,7075	0,6808
65%	15927,3151	0,6981	0,8066	0,5557	0,7045	0,6863
66%	16316,5405	0,6980	0,8153	0,5440	0,7013	0,6916
67%	16722,2215	0,6979	0,8240	0,5322	0,6982	0,6972
68%	17117,2861	0,6974	0,8324	0,5201	0,6949	0,7027
69%	17520,5255	0,6965	0,8404	0,5076	0,6915	0,7078
70%	17967,2168	0,6951	0,8480	0,4944	0,6877	0,7124
71%	18408,6438	0,6935	0,8554	0,4809	0,6839	0,7169

72%	18880,4379	0,6922	0,8631	0,4679	0,6805	0,7224
73%	19389,2302	0,6906	0,8705	0,4545	0,6769	0,7277
74%	19912,3337	0,6893	0,8781	0,4414	0,6736	0,7338
75%	20462,5804	0,6879	0,8857	0,4282	0,6704	0,7405
76%	21015,2658	0,6860	0,8928	0,4144	0,6669	0,7465
77%	21608,8028	0,6844	0,9002	0,4010	0,6637	0,7537
78%	22240,6076	0,6826	0,9075	0,3874	0,6605	0,7612
79%	22901,7537	0,6804	0,9143	0,3732	0,6570	0,7683
80%	23628,2285	0,6777	0,9208	0,3586	0,6534	0,7751
81%	24363,1604	0,6751	0,9273	0,3440	0,6499	0,7827
82%	25170,9815	0,6720	0,9333	0,3288	0,6461	0,7896
83%	25989,4968	0,6686	0,9391	0,3133	0,6423	0,7967
84%	26920,9312	0,6652	0,9450	0,2978	0,6386	0,8047
85%	27913,9439	0,6621	0,9510	0,2827	0,6352	0,8147
86%	29049,8099	0,6587	0,9568	0,2672	0,6316	0,8249
87%	30184,2685	0,6542	0,9617	0,2504	0,6275	0,8325
88%	31501,0168	0,6494	0,9663	0,2333	0,6234	0,8404
89%	32929,9742	0,6446	0,9709	0,2162	0,6193	0,8497
90%	34545,3158	0,6396	0,9752	0,1988	0,6152	0,8594
91%	36316,0283	0,6345	0,9796	0,1814	0,6111	0,8713
92%	38225,5055	0,6289	0,9834	0,1633	0,6068	0,8825
93%	40478,4238	0,6229	0,9870	0,1448	0,6025	0,8944
94%	43243,5719	0,6163	0,9900	0,1256	0,5979	0,9050
95%	46551,9043	0,6098	0,9930	0,1065	0,5934	0,9206
96%	50976,1238	0,6025	0,9954	0,0865	0,5886	0,9345
97%	56704,0505	0,5945	0,9972	0,0657	0,5836	0,9473
98%	65556,6585	0,5862	0,9987	0,0445	0,5785	0,9625
99%	83023,6133	0,5774	0,9997	0,0227	0,5733	0,9830