UNIVERSIDADE FEDERAL DO PARANÁ
SETOR DE CIÊNCIAS EXATAS
DEPARTAMENTO DE ESTATÍSTICA
CURSO DE ESTATÍSTICA

Bruno Henrique Abreu

Análise de sobrevivência aplicada a modelos de crédito

CURITIBA
2019

UNIVERSIDADE FEDERAL DO PARANÁ
SETOR DE CIÊNCIAS EXATAS
DEPARTAMENTO DE ESTATÍSTICA
CURSO DE ESTATÍSTICA

Bruno Henrique Abreu

Análise de sobrevivência aplicada a modelos de crédito

Trabalho de Conclusão de Curso apresentado à disciplina Laboratório B do Curso de Estatística do Setor de Ciências Exatas da Universidade Federal do Paraná, como exigência parcial para obtenção do grau de Bacharel em Estatística.

Orientadora: Profa. Dra. Suely Ruiz Giolo

CURITIBA
2019

AGRADECIMENTOS

Agradeço a Deus.

Agradeço à minha família, pelo apoio, especialmente à minha Mãe Iara.

Agradeço aos meus colegas de faculdade, pela amizade e parceria em tempos difíceis.

Agradeço à minha noiva, Renata, pela paciência e apoio em todos os momentos.

Agradeço à minha orientadora, Profª Drª Suely Ruiz Giolo, pela paciência em ensinar e pela colaboração para que este trabalho fosse realizado.

Agradeço à Profª Drª Fernanda B. Rizzato, que aceitou participar da banca deste trabalho.

"*Democracia é oportunizar a todos o mesmo ponto de partida.*

Quanto ao ponto de chegada, depende de cada um".

(*Fernando Sabino*)

RESUMO

A indústria bancária usualmente utiliza o modelo de regressão logística para modelar a probabilidade de um cliente se tornar *mau* pagador, desprezando uma informação importante que é o tempo até o cliente se tornar inadimplente. Dada a oportunidade de melhorar as predições para a tomada de decisão, o objetivo deste trabalho foi o de utilizar métodos de análise de sobrevivência para modelar o tempo até a aquisição do crédito (propensão ao crédito), bem como o tempo até a inadimplência (risco de crédito). Para isso, foram utilizados o modelo de Cox e o modelo de mistura logito-Cox. A base de dados utilizada, que contém 74 variáveis de 890.000 clientes, foi disponibilizada pela empresa *Lending Club*, que é uma empresa *peer-to-peer*, fundada nos Estados Unidos, com o objetivo principal de ligar mutuários a investidores. Neste trabalho, foi analisada uma amostra correspondendo aos anos de 2005, 2006 e 2007, totalizando 92.347 registros de clientes. Desse total, 44.713 realizaram a contratação de algum produto, sendo que 12% apresentaram inadimplência. Após realizar uma análise univariada das 74 variáveis (para excluir as com valores zeros e *missing)*, bem como avaliar a correlação de Pearson entre as variáveis contínuas e, ainda, estimar, para as variáveis categóricas, a curva de sobrevivência via o estimador de Kaplan-Meier, restaram 10 variáveis para o processo de ajuste dos modelos. Os modelos considerados para analisar a propensão ao crédito e o risco de crédito apresentaram ajustes satisfatórios, sendo possível combiná-los para a tomada de decisão. A análise combinada dos modelos permitiu a identificação de clientes com probabilidade alta de aquisição de algum produto de crédito e baixa probabilidade de inadimplência.

Palavras-chave: *Credit Scoring*, Inadimplência, Modelo de Cox, Modelo de Mistura.

Sumário

1 INTRODUÇÃO

Considerando que a capacidade assertiva de tomar boas decisões é o objetivo comum em qualquer área de interesse, não poderia ser diferente no setor financeiro quando o assunto é a análise de crédito. É necessário que a operação em questão seja rentável para o credor e, para isso, mecanismos são criados para auxiliar na tomada de decisão. Por esse motivo, sabendo da importância do controle de risco associado às operações de crédito, modelos estatísticos são necessários para avaliar o risco de inadimplência em tais situações.

Historicamente, nos últimos anos, o mercado de crédito brasileiro vem apresentando significativas taxas de crescimento. De acordo com dados emitidos pelo Banco Central do Brasil (BCB, 2017), as concessões no crédito direcionado aumentaram 8,3% nas operações com pessoas físicas, enquanto foi registrada uma alta de 9% nas operações com empresas. Nesse mesmo período, o saldo das operações de crédito em meio ao sistema financeiro subiu 0,5% em maio, para R$ 3,107 trilhões.

Em especial na área financeira, o controle de risco é o que permite a sobrevivência das instituições. Sem o controle do prejuízo e da inadimplência, seriam insustentáveis a concessão de crédito e a obtenção de lucros. Para lidar com isso no âmbito da globalização dos mercados monetários e tratar da instabilidade financeira, foi assinado, em 1988, o 1º acordo de Basiléia, que tem por objetivo orientar e regulamentar a atuação das instituições financeiras com a finalidade de garantir menor risco nas operações, maximizar a capacidade de absorção de crises e, principalmente, manter a transparência no funcionamento operacional dessas instituições, estabelecendo regras como: reservas mínimas de capital, controle de índice de liquidez e métodos de classificação de risco para clientes.

No mundo contemporâneo, a concessão de crédito continua sendo um dos mecanismos mais rentáveis para o sucesso e desenvolvimento das instituições financeiras. Com isso, torna-se necessário controle rígido em relação à previsão e controle da inadimplência, para que situações deficitárias sejam evitadas e o mercado possa manter sua atuação sem grandes distorções em relação à taxa de juros.

Com o avanço das transações financeiras e a necessidade de uma análise padronizada que abrangesse todos os tópicos necessários para viabilização da concessão ou não do crédito, é que a adoção de modelos de crédito (*credit scoring models*) está em crescente uso, haja visto que eles proporcionam: automatização das análises, consistência

nas decisões tomadas e, principalmente, aumento no volume de análises e capacidade de monitorar e administrar o risco de uma carteira de crédito.

Em 1933, com a primeira publicação da revista *Econometrica* (Cleveland), houve uma intensificação na aplicação e desenvolvimento de métodos estatísticos para, dentre outros objetivos, testar teorias econômicas, avaliar e implementar políticas comerciais, estimar relações econômicas e dar suporte à concessão de crédito, unificando o conhecimento qualitativo teórico e empírico desses problemas econômicos.

Os primeiros modelos de *credit scoring* foram desenvolvidos entre os anos 40 e 50, com a metodologia básica aplicada a esse tipo de problema orientada por métodos de discriminação produzidos por Fisher (1936). Foi de Durand (1941) o primeiro trabalho conhecido que utilizou análise discriminante para um problema de crédito. Nele, o autor utilizou as técnicas desenvolvidas por Fisher para discriminar bons e maus empréstimos.

Em 1984, o livro *Risco e Recompensa: o negócio de crédito ao consumidor* apresentou as primeiras menções ao modelo de *credit scoring*, sendo esse um modelo de escore com base em dados cadastrais dos clientes e comportamentais históricos. Este tipo de modelo é utilizado nas decisões de concessão do crédito ao solicitante. Por outro lado, o modelo denominado *behaviour scoring* se baseia em dados comportamentais provenientes da utilização do crédito cedido; sua utilização ocorre nas decisões de manutenção e/ou renovação de linhas e produtos para os já clientes. Ainda, o modelo *collection scoring* se baseia em dados comportamentais de clientes inadimplentes, sendo utilizado nas decisões de priorização de estratégias de cobranças.

É possível mencionar que, usualmente, os modelos de *credit scoring* se dividem em dois grupos distintos. Um deles, trata especialmente dos novos clientes, cujo objetivo é conceder crédito utilizando, a priori, referências cadastrais tais como: escolaridade, idade, estado civil, tempo de trabalho e histórico financeiro. No outro grupo, estão os modelos conhecidos como *behaviour scoring*, desenvolvidos para coordenar o crédito daqueles que já são clientes. Neste modelo, além dos dados cadastrais, o registro e o histórico comportamental dos clientes no sistema financeiro, incluindo a apresentação de restritivos de crédito, a pontualidade em pagamentos ou o alto comprometimento da renda, são analisados criteriosamente. Por agrupar um número maior de variáveis e possuir informações mais qualificadas é que, neste grupo, existe a possibilidade de administrar os modelos com maior grau de distinção entre maus e bons pagadores do que os do primeiro grupo.

Estes e vários outros modelos são utilizados como as principais ferramentas dentro do ciclo de crédito, que contempla as etapas de propensão, concessão de crédito, manutenção e cobrança. São, assim, utilizados durante todo o ciclo de vida de um tomador de crédito, dando suporte às inúmeras instituições financeiras no mundo.

Diversas técnicas de análise estatística são empregadas para o desenvolvimento dos modelos de *credit scoring*. Dentre elas, destacam-se: algoritmos genéticos, regressão logística, redes neurais, análise discriminante e, recentemente, análise de sobrevivência. Em grande parte, as técnicas utilizadas para a construção dos modelos fornecem a probabilidade do cliente se tornar um mau pagador, porém não levam em conta o fato de que, apesar dos dados utilizados na modelagem serem naturalmente discretos, o relacionamento do cliente com a instituição é contínuo, a partir do seu cadastro. Com isso, a técnica de análise de sobrevivência apresenta uma vantagem em relação às outras técnicas, pois resulta em uma resposta temporal, com condições de predizer quando ocorrerá o evento de interesse, especificamente, nesse caso, determina a possibilidade real do cliente se tornar ou não um mau pagador.

Nos últimos anos, dado a importância desses modelos, diversos autores trabalharam na construção de modelos de risco de crédito mais robustos e assertivos. A ideia de empregar análise de sobrevivência para construir modelos de escore de crédito foi introduzida por Narain (1992) e, posteriormente, desenvolvida por Thomas et al. (1999). A motivação para a utilização de técnicas de análise de sobrevivência na área de crédito tem seus fundamentos na tomada de uma decisão mais completa, ou seja, não apenas estimar a probabilidade de um cliente se tornar um mau pagador, mas também estimar quando este evento poderá ocorrer.

Neste contexto, um dos objetivos deste trabalho foi o de, inicialmente, aplicar um modelo de análise de sobrevivência no cenário de propensão ao crédito (momento em que a concessão do crédito ao proponente está sob análise) com o intuito de que tal modelo possa auxiliar na decisão sobre a concessão. Para tanto, foi utilizado o modelo de Cox.

Por outro lado, no cenário de risco de crédito (isto é, após a concessão do crédito), foi utilizado o modelo de fração de imunes (denominado, neste trabalho, "modelo de fração de adimplentes") com a finalidade de avaliar o risco de inadimplência dos proponentes que tiveram o empréstimo concedido. Em ambos os modelos, foram utilizadas diversas variáveis (informações) relacionadas ao cliente no momento da concessão do crédito, bem como variáveis comportamentais e macroeconômicas.

De modo geral, o trabalho apresenta a seguinte estrutura. O Capítulo 2 traz uma breve revisão da fundamentação teórica sobre os modelos de *credit scoring* focando o referencial teórico principalmente na técnica de análise de sobrevivência. Nos Capítulos 3 e 4, são descritos os dados e os métodos utilizados para a análise dos mesmos, assim como são apresentados os principais resultados obtidos com base nos modelos ajustados. Por fim, tem-se, no Capítulo 5, algumas considerações finais.

2 REVISÃO DE LITERATURA

Neste capítulo, é apresentada uma breve revisão do processo de construção dos modelos de crédito, com ênfase nas técnicas de análise de sobrevivência.

Considerando a importância do entendimento dos agentes determinantes presentes no risco de crédito, inúmeros estudos foram realizados ao longo dos últimos anos com o objetivo de predizer e mensurar o risco por meio de modelos estatísticos mais assertivos capazes de calcular a inadimplência associada a um empréstimo. Destacam-se neste âmbito os chamados modelos de *credit scoring*.

Entre as várias técnicas estatísticas empregadas na construção dos modelos de predição de risco de crédito, destaca-se a análise de sobrevivência. Para Almeida (2008), além de predizer se um cliente é um potencial inadimplente, há também a necessidade de predizer quando esse evento poderá acontecer. Por isso, a técnica estatística que mais se adequa a modelagem temporal é a análise de sobrevivência (BELOTTI; CROOK, 2009). A análise de sobrevivência é uma das áreas da estatística que mais cresceu nas últimas décadas, o que, em parte, se deve ao aprimoramento das técnicas estatísticas combinado com computadores cada vez mais velozes (COLOSIMO; GIOLO, 2006).

De acordo com Thomas et al. (2002), todas as decisões relacionadas à análise de concessão de crédito eram realizadas, até o início do século XX, somente pelo julgamento subjetivo dos analistas. No entanto, os autores afirmam que, após a publicação, em 1936, da técnica de Análise Linear Discriminante, desenvolvida por Fisher, a Estatística passou a ser pensada e aplicada para descrever as principais características de bons e maus pagadores. Os primeiros modelos de *credit scoring* foram desenvolvidos por Durand (1941), com o objetivo de demandar os proponentes quanto à probabilidade de pagar o capital emprestado. Devido à agilidade no processo de tomada de decisão, menor custo, maior objetividade e, também, poder preditivo mais apurado, é que os modelos de *credit scoring* foram se popularizando, sendo muito utilizados atualmente (HAND; HENLEY, 1997).

Os modelos de *credit scoring* fazem uso de técnicas estatísticas juntamente com algoritmos matemáticos para aferir a probabilidade de que determinado evento aconteça. Aplicando-se as fórmulas, o sistema atribui uma pontuação específica para cada característica do proponente/cliente para prever um determinado resultado.

No entanto, é necessário atentar-se que apenas as informações obtidas do *credit score* não determinam o sucesso de um modelo adotado por uma instituição financeira; por medida preventiva é necessário manter um acompanhamento com o foco de monitorar continuamente a posição dos clientes. Este acompanhamento contínuo é conhecido como *behaviour score* e concentra-se no conhecimento e acompanhamento das operações dos clientes durante o relacionamento com a instituição. Tais informações envolvem as principais características sobre os hábitos na utilização de crédito, bem como o tempo de relacionamento e o perfil do cliente, dentre outros. Essas informações são geradas e transformadas em dados constantemente, de modo que tal processo torna a modelagem em si uma atividade dinâmica que necessita de revisão permanente, visto que o seu gerenciamento é valioso e utiliza de ferramentas que auxiliam e viabilizam a tomada de decisão assertiva com o objetivo de minimizar riscos e maximizar resultados. Desse modo, é natural que as empresas, principalmente as atuantes no mercado financeiro, adotem e invistam em estratégias e técnicas que auxiliem os sistemas de análise de crédito, sendo a Análise Sobrevivência uma das mais recentes neste processo.

Em linhas gerais, a Análise de Sobrevivência consiste de um conjunto de técnicas estatísticas utilizadas com o objetivo de predizer a probabilidade de um evento ocorrer no tempo t. O tempo decorrido desde um tempo inicial predeterminado até o evento de interesse é denominado tempo de falha. No caso dos modelos de risco de crédito, o tempo de falha compreende o período desde a contratação do empréstimo até a ocorrência do evento de interesse (por exemplo, a inadimplência).

Entre as principais características relevantes de um modelo no contexto de análise de sobrevivência, destaca-se a presença de censura, ou seja, a existência de informação parcial da variável resposta. No cenário das instituições financeiras, no que tange à análise de concessão ao crédito, a censura pode ocorrer justamente pelo fato de o evento de interesse não ter sido observado no final do período de acompanhamento ou pelo contrato ter sido liquidado ao longo do estudo. Embora as censuras forneçam informação parcial da variável resposta, elas são importantes e devem ser levadas em consideração na estimação do modelo. Segundo Colosimo e Giolo (2006), as observações censuradas fornecem informações sobre o tempo de vida e, sendo assim, a omissão das censuras no cálculo das estatísticas de interesse pode produzir conclusões viciadas.

Neste contexto, a análise de sobrevivência permite identificar quais variáveis afetam e viabilizam o risco de ocorrência de determinado fenômeno, com a presença de censuras ocorrendo, geralmente, pelo fato de alguns clientes abandonarem a carteira ou não experimentarem o evento de interesse sob estudo.

Algumas funções são muito utilizadas nos estudos envolvendo dados de sobrevivência. Uma delas, é a função densidade de probabilidade, 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 (LEE, 1992). Denotando por T a variável aleatória tempo, usualmente contínua e não negativa, a função densidade de probabilidade fica expressa por

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

com $f(t) \geq 0$, para todo t, e com a área abaixo da curva $f(t)$ igual a 1.

Outra função bastante utilizada para descrever a variável aleatória T é a função de sobrevivência, definida como a probabilidade de um indivíduo não falhar (ou do evento não ocorrer) até determinado tempo t (LAWLESS, 1982), ou seja,

$$S(t) = P(T > t) = 1 - \int_0^t f(x)dx,$$

com $f(x)$ a função densidade de probabilidade.

No capítulo a seguir, são descritos os dois modelos utilizados neste trabalho no contexto de análise de sobrevivência. O foco do primeiro modelo está na propensão ao crédito (concessão do crédito solicitado pelo proponente), e o do segundo no risco de crédito (risco de inadimplência após a concessão do crédito). O uso combinado desses dois modelos (de propensão ao crédito e de risco de crédito) tem como finalidade encontrar o perfil de cliente com maior propensão ao empréstimo (concessão do crédito solicitado) e com baixo risco de inadimplência.

3 MATERIAL E MÉTODOS

3.1 Material

3.1.1 Conjunto de Dados

A base de dados utilizada neste trabalho foi extraída do site da *Lending Club* <https://www.lendingclub.com/>, que é uma empresa *peer-to-peer*, fundada nos Estados Unidos, com o objetivo de ligar mutuários a investidores, mudando a experiência que as pessoas têm em relação ao acesso ao crédito. Diferente dos bancos tradicionais, a *Lending Club* não precisa seguir as regras estabelecidas pelo órgão regulador responsável pelas operações e, sendo assim, consegue trabalhar com taxas de juros mais atraentes.

A base de dados contém 74 variáveis, classificadas em 5 grupos, conforme descrito no Quadro 1. Visando a otimização computacional, uma vez que a base completa possui 890.000 registros, foram selecionados os contratos cujos proponentes tiveram o começo do seu relacionamento com a *Lending Club* nos anos de 2005, 2006 e 2007, totalizando uma amostra de 92.347 registros.

Quadro 1 – Descrição dos cinco grupos de variáveis disponíveis na base de dados utilizada

Grupo	Chave	Cadastral	Comportamental	Histórica	Operação
Quantidade	**5**	**9**	**15**	**21**	**24**
Descrição	Conjunto de um ou mais atributos, que permite identificar unicamente uma entidade no conjunto de entidades	Conjunto de atributos pessoais que caracterizam o indivíduo.	Conjunto de atributos referente ao comportamento do cliente após a efetivação do empréstimo	Conjunto de atributos referente ao comportamento do cliente em empréstimos passados	Conjunto de atributos referente a condição da operação de crédito atual
Exemplo	**ID:** Um identificador atribuído pelo LC para a listagem de empréstimos.	**zip_code:** Os primeiros 3 números do código postal	**last_pymnt_amnt:** Valor total do último pagamento recebido.	**mths_since_last_delinq:** O número de meses desde a última inadimplência do tomador.	**issue_d:** O mês em que o empréstimo foi iniciado.

Fonte: O autor (2019).

Inicialmente, para o ajuste do modelo de propensão ao crédito no contexto de análise de sobrevivência, foram utilizados os 92.347 contratos, tendo sido fixado o acompanhamento dos mesmos por um período de até 8 anos. Ao final deste período, foi registrado a aquisição do produto (isto é, do uso do crédito) para 48% deles. Ou seja, 48% de falhas e 52% de censuras, como pode ser observado no registro cumulativo em função do tempo (em anos) mostrado na Tabela 1. Além disso, foi registrado o tempo até a aquisição do produto ou até o final do período de acompanhamento fixado (para os que não adquiriram nenhum produto).

Tabela 1 – Registro cumulativo da aquisição do produto (crédito) em função do tempo (em anos)

Tempo até a aquisição do produto	Registros	Porcentagem	Registros Acumulados	Porcentagem Acumulada
0 —| 2 anos	164	0%	164	0%
2 —|4 anos	1276	1%	1440	1%
4 —|6 anos	8159	9%	9599	10%
6 —|8 anos	35114	38%	44713	48%
8 —|10 anos	43443	47%	88156	95%
Acima de 10 anos	4191	5%	92347	100%

Fonte: O autor (2019).

Na sequência, para ajustar o modelo de risco de crédito, também no contexto de análise de sobrevivência, foram utilizados todos os contratos ativados dentro do prazo de 8 anos, o que corresponde a 44.713 contratos. Para estes 44.713 contratos, foi registrado os que apresentaram e os que não apresentaram inadimplência (o que resultou em 12% de inadimplência e 88% de adimplência), bem como o tempo até a inadimplência ou até o término do período de acompanhamento (para os adimplentes).

3.1.2 Recursos Computacionais

Para a análise dos dados, que compreendeu a análise exploratória dos dados e o ajuste dos modelos de Cox e de fração de adimplentes (descritos na seção a seguir), foi utilizado o *software* R, versão 3.5.1 (R CORE TEAM, 2017). Os principais pacotes utilizados foram: *glm*, *survival*, *smcure e survivalROC*.

3.2 Métodos

Nesta seção, são apresentados os modelos utilizados, no contexto de análise de sobrevivência, para a análise dos dados descritos na Seção 3.1.1.

3.2.1 Modelo de Propensão ao Crédito no Contexto de Análise de Sobrevivência

O termo análise de sobrevivência se refere a uma área da estatística destinada à análise de dados em que a variável resposta é o tempo até a ocorrência de um evento de interesse. Como, geralmente, os indivíduos são acompanhados por um período de tempo preestabelecido, o evento de interesse pode não ocorrer nesse período para alguns dos indivíduos, o que caracteriza a presença de censuras. Nesses casos, o tempo registrado para tais indivíduos denominam-se tempos censurados.

No contexto do modelo de propensão ao crédito, o evento de interesse para os dados analisados neste trabalho corresponde à aquisição de algum produto de crédito pelo cliente no período de até 8 anos. Em consequência, a não aquisição de produtos de crédito no período mencionado corresponde às censuras. Para cada cliente *i*, *i* = 1, ..., *n*, a informação registrada consiste da tripla (t_i, δ_i, $\mathbf{x}_i$), com t_i o tempo decorrido desde o início do relacionamento com a *Lending Club* até a aquisição de algum produto, e $\delta_i = 1$, se ocorreu a aquisição e $\delta_i = 0$, se não ocorreu. Ainda, o vetor $\mathbf{x}_i = (x_{i1}, ..., x_{ip})$ corresponde aos valores observados das p variáveis $\mathbf{X} = (X_1, ..., X_p)$.

Algumas funções as quais se tem interesse em estimar para dados dessa natureza são: a função de sobrevivência e a função taxa instantânea de falha, expressas para o indivíduo *i*, respectivamente, por

$$S(t \mid \mathbf{x}_i) = P(T > t \mid \mathbf{x}_i) \tag{1}$$

e

$$\lambda(t \mid \mathbf{x}_i) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t+\Delta t \mid T \geq t; \mathbf{x}_i)}{\Delta t}. \tag{2}$$

A função em (1) fornece a probabilidade do indivíduo *i* não falhar até certo tempo t, ou seja, a probabilidade dele sobreviver ao tempo t, enquanto a função em (2) pode ser vista como a probabilidade "aproximada" do indivíduo que ainda não experimentou o evento até o tempo t, vir a experimentá-lo no instante imediatamente posterior a t.

Para estimar a função em (1), em um contexto não paramétrico, é comum utilizar o estimador não paramétrico de Kaplan-Meier (KAPLAN; MEIER, 1958), expresso por

$$\hat{S}(t) = \prod_{j:t_j < t} \left(\frac{n_j - d_j}{n_j}\right) = \prod_{j:t_j < t} \left(1 - \frac{d_j}{n_j}\right),$$

em que d_j denota o número de falhas em t_j e n_j o número o indivíduos sob risco em t_j .

Anterior ao ajuste do modelo de propensão ao crédito, o estimador de Kaplan-Meier foi utilizado, neste trabalho, para estimar as curvas de sobrevivência associadas às respectivas categorias de cada variável X_k (k = 1, ..., p). O intuito dessa análise foi a de verificar a influência individual de cada uma das p variáveis sobre o tempo até a contratação de algum produto pelo cliente. Após essa análise inicial, foi ajustado o modelo de Cox (COX, 1972), cujas funções de sobrevivência e de risco, para o indivíduo *i*, são dadas por

$$S(t \mid \boldsymbol{x}_i) = [S_0(t)]^{\exp(\beta_1 x_{i1} + \ ... \ + \beta_p x_{ip})} \quad \text{e} \quad \lambda(t \mid \boldsymbol{x}_i) = \lambda_0(t)\exp\left(\beta_1 x_{i1} + \ ... \ + \beta_p x_{ip}\right),$$

com $S_0(t)$ a função de sobrevivência de base, $\lambda_0(t)$ a função risco de base e $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)$ o vetor de parâmetros desconhecidos associados às variáveis $\mathbf{X} = (X_1, ..., X_p)$.

Para estimar o vetor de parâmetros $\boldsymbol{\beta}$, Cox (1975) propôs a função de verossimilhança parcial, definida como

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{k} \frac{\exp\{\boldsymbol{x}_i'\boldsymbol{\beta}\}}{\sum_{j\in R(t_i)} \exp\{\boldsymbol{x}_j'\boldsymbol{\beta}\}} = \prod_{i=1}^{n} \left(\frac{\exp\{\boldsymbol{x}_i'\boldsymbol{\beta}\}}{\sum_{j\in R(t_i)} \exp\{\boldsymbol{x}_j'\boldsymbol{\beta}\}} \right)^{\delta_i}$$

em que δ_i denota o indicador de falha e $R(t_i)$ o conjunto de indivíduos sob risco em t_i.

Os valores de $\boldsymbol{\beta}$ que maximizam a função de verossimilhança parcial, podem ser obtidos resolvendo-se a equação

$$U(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \left[x_i - \frac{\sum_{j\in R(t_i)} x_j \exp\{\boldsymbol{x}_j'\hat{\beta}\}}{\sum_{j\in R(t_i)} \exp\{\boldsymbol{x}_j'\hat{\beta}\}} \right] = 0.$$

A função de verossimilhança parcial atribui tempos de sobrevivência contínuos e, sendo assim, ignora a possibilidade de empates nos valores observados. Como pode ocorrer empates nos tempos de falha ou censura, usa-se a convenção de que a censura ocorreu após a falha para definir as observações incluídas no conjunto de indivíduos sob risco $R(t_i)$.

Nos casos em que há empates, foram propostas aproximações para a função de verossimilhança parcial. Uma delas, proposta por Breslow (1972), é dada por

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{k} \frac{\exp\{\boldsymbol{S}_i'\boldsymbol{\beta}\}}{\left[\sum_{j\in R(t_i)} \exp\{\boldsymbol{x}_j'\boldsymbol{\beta}\}\right]^{d_i}},$$

em que $\boldsymbol{S}_i$ corresponde ao vetor da soma das p covariáveis para os indivíduos que falharam no mesmo tempo t_i (i = 1, ..., k) e d_i ao número de falhas no tempo observado.

As propriedades assintóticas dos estimadores de máxima verossimilhança parcial foram estudadas por Cox (1975), Tsiatis (1981) e Andersen e Gill (1982), tendo em vista a necessidade de construir intervalos de confiança e testar hipóteses. Utilizando a relação entre os tempos de falha e *martingais*, verificaram que os estimadores são consistentes e assintoticamente normais. Sendo assim, é possível realizar inferências utilizando a estatística de Wald, o teste da razão de verossimilhanças e a estatística escore.

Dado que nenhuma forma paramétrica é assumida para $\lambda_0(t)$, Breslow (1972) propôs um estimador para a função risco de base acumulada, $\Lambda_0(t)$, expresso por

$$\widehat{\Lambda}_0(t) = \sum_{j:t_j<t} \frac{d_j}{\sum_{l\in R_j} \exp\{\boldsymbol{x}_l'\widehat{\boldsymbol{\beta}}\}}, \qquad (3)$$

em que d_j denota o número de falhas em t_j e R_j o conjunto de indivíduos sob risco em t_j.

A partir de (3), as funções de sobrevivência, podem ser estimadas por

$$\hat{S}_0(t) = \exp\{-\widehat{\Lambda}_0(t)\}$$

e

$$\hat{S}(t \mid \boldsymbol{x}) = \exp\{-\widehat{\Lambda}(t \mid \boldsymbol{x})\}.$$

Apesar de o modelo de regressão de Cox ser flexível, verificar sua adequação aos dados é imprescindível para realizar predições mais acuradas. Para tanto, são utilizados os resíduos de Cox-Snell (1968) definidos, para i = 1, ..., n, por

$$\hat{e}_i = \widehat{\Lambda}_0(t_i) \exp\left\{\sum_{k=1}^{p} x_{ik}\widehat{\beta_k}\right\},$$

em que $\widehat{\Lambda}_0(t_i)$ é a função risco de base acumulada estimada em t = t_i.

Para concluir que o modelo se ajusta bem aos dados, os resíduos $\hat{e}_i$ (i = 1, ..., n) devem ser vistos como uma amostra censurada de uma distribuição exponencial padrão, de modo que o gráfico $\widehat{\Lambda}(\hat{e}_i)$ versus $\hat{e}_i$ deve ser aproximadamente uma reta.

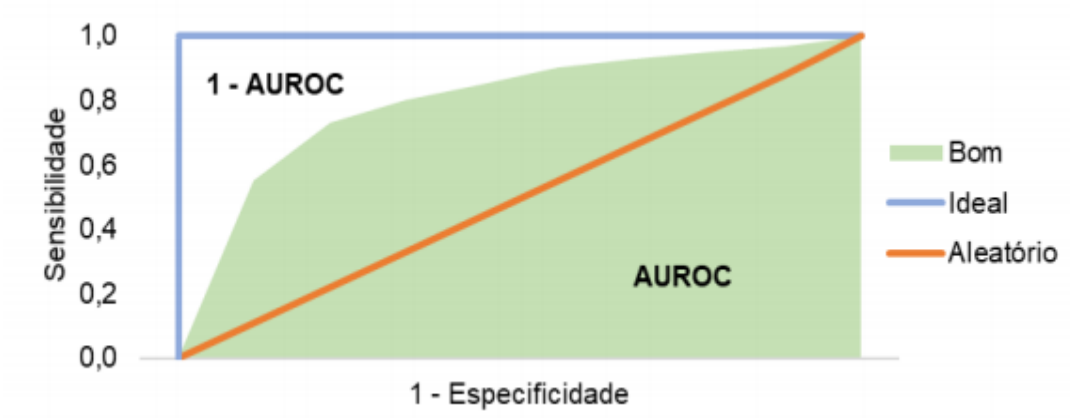
Adicionalmente, é preciso verificar a validade da suposição de taxas (ou riscos) proporcionais, uma vez que a violação dessa suposição pode acarretar em sérios vícios na estimação dos coeficientes do modelo (STRUTHERS; KALBFLEISH, 1986). Segundo Colosimo e Giolo (2006), um gráfico simples para avaliar tal suposição consiste em dividir os dados em m estratos de acordo com as categorias de uma dada covariável X_k e, então, estimar $\widehat{\Lambda}_{0j}(t)$ para j = 1, ..., m. Se a suposição de riscos proporcionais for válida, as curvas do logaritmo de $\widehat{\Lambda}_{0j}(t)$ *versus t,* devem apresentar diferenças aproximadamente constantes ao longo do tempo. Curvas não paralelas, indicam violação da suposição.

Outro método gráfico utilizado para avaliar a suposição de riscos proporcionais se baseia nos resíduos padronizados de Schoenfeld (SCHOENFELD, 1982), denotados por s_{iq}, para i = 1, ..., d e q = 1, …, p, com d o número de falhas e p o número de covariáveis. Grambsch e Therneau (1994) sugeriram os gráficos $s_{iq} + \hat{\beta}_q$ *versus* t, q = 1, …, p, que devem apresentar o comportamento de uma linha horizontal (reta com inclinação nula) para que haja evidências a favor da suposição de proporcionalidade dos riscos.

A obtenção de medidas estatísticas também pode auxiliar a verificar a suposição de taxas proporcionais. Uma delas, é calcular o coeficiente de correlação de Pearson (ρ) entre os resíduos padronizados de Schoenfeld e t para cada covariável (ou para as categorias de cada covariável, se esta for categórica). Valores de ρ próximos de zero indicam a não rejeição da suposição de riscos proporcionais (COLOSIMO; GIOLO, 2006).

Quanto ao poder preditivo do modelo de Cox, pode-se utilizar a área abaixo da curva ROC para vários tempos ao longo do período de seguimento. Essa metodologia foi proposta por Heagerty e Zheng (2005), consistindo de uma extensão daquela utilizada em regressão logística, a qual se baseia em duas medidas conhecidas como *sensibilidade* e *especificidade*. No contexto de *credit scoring*, elas são definidas da seguinte maneira: *sensibilidade* = probabilidade de um indivíduo ser classificado como mau pagador, dado que ele é realmente *mau* pagador, e *especificidade* = probabilidade de um indivíduo ser classificado como *bom* pagador, dado que é realmente *bom* pagador. A curva ROC é construída dispondo-se no eixo horizontal os valores de (1 - *especificidade*) e, no eixo vertical, a *sensibilidade.* Quanto mais a curva ROC se aproximar do canto superior esquerdo, bem como quanto mais a área abaixo da curva (AUROC) se aproximar de 1, melhor o poder preditivo do modelo. A Figura 1 ilustra a representação da curva ROC.

Figura 1 – Ilustração da representação gráfica de uma curva ROC



Fonte: Bojanowski e Lolatto (2018).

No contexto de análise de sobrevivência, a curva ROC é construída em diferentes tempos a fim de se observar o poder preditivo do modelo ao longo do tempo.

3.2.2 Modelo de Risco de Crédito no Contexto de Análise de Sobrevivência

No contexto de risco de crédito, foi considerado o modelo com sobreviventes de longa duração, conhecido na área médica por modelo de fração de cura (ou de imunes ao evento). Em particular, foi adotado o modelo de mistura com fração de imunes (KUK; CHEN, 1992, CORBIÈRE; JOLY, 2007), denominado no contexto dos dados deste trabalho de modelo de mistura com fração de adimplentes. De acordo com Diniz e Louzada

(2012), tal modelo pode ser considerado como uma forma de tratar o tempo até a ocorrência de um problema de pagamento de crédito, quando existe uma possível "imunidade" (fração de imunes) em relação ao evento dentro de um período específico do horizonte de previsão.

A fração de imunes mencionada deve ser explicada de forma que não cause um estranhamento com as censuras. Portanto, é importante estabelecer a distinção entre o que é fração de imunes e o que é censura. Por exemplo, em um estudo em que os indivíduos são acompanhados por um período de tempo considerado longo e, mesmo assim, o evento não ocorre para parte deles, denomina-se fração de imunes a proporção de indivíduos na qual o evento não ocorrerá, mesmo se eles forem observados por mais tempo. Por outro lado, a parcela de indivíduos, usualmente muito pequena, na qual o evento não foi observado após um longo período de acompanhamento, mas que possivelmente ocorreria, caso fossem acompanhados por mais tempo, caracterizam as censuras.

Sendo assim, os modelos de mistura para dados de longa duração foram propostos para acomodar essas situações. Uma dessas situações, está associada aos dados de *credit scoring*, tendo em vista que há a possibilidade de se observar uma fração de clientes imunes ao evento (ou seja, não suscetíveis ao evento de interesse). Por exemplo, para situações relacionadas à inadimplência, existe uma fração elevada de clientes adimplentes que permanecerão adimplentes, mesmo se observados por um longo período de tempo. Este fato, motivou o uso do modelo citado neste trabalho.

Para o ajuste do modelo mencionado, a informação registrada, no contexto de risco de crédito (risco de o cliente se tornar inadimplente após efetivação do empréstimo), consiste da tripla (t_i, δ_i, $\mathbf{x}_i$), com t_i o "tempo até a inadimplência" e $\delta_i = 1$, se ocorreu a inadimplência e $\delta_i = 0$, se não ocorreu. Ainda, $\mathbf{x}_i = (x_{i1}, ..., x_{ip})$ corresponde ao vetor de valores observados das p variáveis explicativas $\mathbf{X} = (X_1, ..., X_p)$.

Quanto à formulação do modelo de mistura com fração de adimplentes, considere que na população de clientes sendo acompanhados exista um grupo "suscetível" ao evento (inadimplentes) e outro "não suscetível" ao evento (adimplentes). De acordo com Berkson e Gage (1952), a expressão do modelo de mistura fica dada por

$$S_p(t \mid \boldsymbol{x}, \boldsymbol{z}) = (1 - p(\boldsymbol{z})) + p(\boldsymbol{z}) S_U(t \mid \boldsymbol{x}) \qquad (4)$$

sendo $0 < p(\mathbf{z}) < 1$ a probabilidade de inadimplência e $S_U(t \mid \boldsymbol{x})$ a probabilidade de um cliente suscetível à inadimplência não apresentá-la no tempo t, com $\boldsymbol{z}$ e $\boldsymbol{x}$ correspondendo aos vetores de variáveis influenciando $p(\mathbf{z})$ e $S_U(t \mid \boldsymbol{x})$, respectivamente.

Para $p(\boldsymbol{z})$ e $S_U(t\,|\boldsymbol{x})$, foram considerados a função logística e o modelo de Cox, de modo que

$$p(\boldsymbol{z}) = \frac{\exp(\boldsymbol{z'\beta})}{1 + \exp(\boldsymbol{z'\beta})} \quad \text{e} \quad S_U(t\,|\boldsymbol{x}) = [\,S_0(t)]^{\exp(\boldsymbol{x'\gamma})},$$

com $S_0(t)$ denotando a função de sobrevivência de base, e **β** e **γ** os vetores de parâmetros associados às variáveis **Z** e **X**, respectivamente.

Para estimação dos vetores **β** e **γ**, considere U a variável que indica se o indivíduo é suscetível (U = 1) ou não suscetível (U = 0) ao evento, com probabilidade $p(\boldsymbol{z})$. Desse modo, dado $\mathbf{u} = (u_1, \ldots, u_n)$ e (t_i, δ_i, $\mathbf{x}_i$, $\mathbf{z}_i$), a função de verossimilhança, segundo Cai et al. (2012), fica expressa por

$$\mathrm{L}(\boldsymbol{\beta}, \boldsymbol{\gamma}, S_{\mathbf{0}}(t)) = \prod_{i=1}^{n} [1 - p(\mathbf{z}_i)]^{1-u_i}\, p(\boldsymbol{z}_i)^{u_i}\, \lambda(t_i \mid U = 1, \mathbf{x_i})^{\delta_i u_i}\, S(t_i \mid U = 1, \mathbf{x_i})^{u_i}$$

em que λ(.) corresponde à função taxa de falha. Para maximização do logaritmo desta função de verossimilhança é, em geral, utilizado o algoritmo EM.

Para verificar a adequação do modelo, foi feita a comparação da curva estimada pelo modelo com aquela estimada pelo Kaplan-Meier para cada perfil (combinação das categorias das covariáveis), bem como calculada a correlação de Pearson entre as curvas.

Ao final das análises, foi realizada uma comparação dos resultados do modelo de propensão ao crédito (modelo de Cox) com os do modelo de risco de crédito (modelo de mistura logito-Cox), com a finalidade de identificar o perfil de clientes com alta propensão ao crédito e baixo risco de inadimplência.

3.2.3 Procedimento de Seleção de Variáveis

O processo de seleção de covariáveis tem por finalidade a identificação e o ajuste de um modelo parcimonioso (que seja de fácil compreensão e com número reduzido de parâmetros), mas capaz de se ajustar satisfatoriamente aos dados. Em estudos que envolvem um número elevado de covariáveis, pode ser útil o uso de algum algoritmo de seleção para a identificação de um modelo adequado.

Neste trabalho, o processo de seleção compreendeu algumas etapas. Inicialmente, foi realizado uma análise descritiva univariada das variáveis, a fim de identificar e excluir variáveis com percentual elevado de zeros, valores repetidos e dados faltantes.

A seguir, foi avaliada a correlação das variáveis remanescentes com a variável resposta. Para tanto, foi utilizado, para as variáveis contínuas, o coeficiente de correlação de Pearson e, para as variáveis categóricas, as curvas de sobrevivência estimadas a partir do estimador de Kaplan-Meier, descrito na Seção 3.2.1. O Valor da Informação, descrito na Seção 3.2.4, também foi utilizado.

Na etapa final, as variáveis remanescentes foram consideradas em cada um dos modelos (de propensão ao crédito e de risco de crédito) e, então, com o auxílio dos procedimentos de seleção *stepwise* e *backward*, foram identificadas as que permaneceram em cada um dos modelos, ou seja, as que apresentaram efeito significativo.

3.2.4 Valor da Informação

O valor da informação (IV) é amplamente utilizado na área financeira com o objetivo de auxiliar na seleção de variáveis preditoras para uma resposta binária. Para uma dada variável X, o cálculo do IV (SIDDIQI, 2006, p. 79-83), também conhecido por força da variável preditora X, é obtido por

$$IV = \sum_{k=1}^{K} (g_k - b_k)\ \log\left(\frac{g_k}{b_k}\right),$$

em que K > 2 corresponde a contagem de níveis de X e g_k e b_k são valores positivos, para todo k = 1,…, K, correspondendo às frequências relativas de clientes (utilizadas em formato decimal) classificados como "bons" e "maus" pagadores em cada nível de X.

O valor da informação é apropriado para um preditor X com número modesto de níveis, tipicamente abaixo de 20 e sem células com frequência zero. Os preditores com intervalos de valores "contínuos" (por ex., dólares, distâncias) devem ser categorizados.

Pode-se dizer que $\log(g_k/b_k)$ mede o desvio entre as distribuições de g e b, enquanto $(g_k - b_k)$ mede a importância do desvio. Considerando, por exemplo, duas razões para g_k/b_k iguais a 0,02/0,01 e 0,2/0,1, a razão 0,02/0,01 é considerada menos importante no cálculo de IV, uma vez que ela é ponderada pela diferença $(0{,}02 - 0{,}01)$.

Com base no IV, utiliza-se a seguinte regra prática para classificar a força de X como preditor para uma resposta binária: *i*) menor que 0,02 = imprevisível; *ii*) 0,02 a 0,1 = fraco; *iii*) 0,1 a 0,3 = médio; e *iv*) 0,3 ou maior = forte.

3.2.5 Peso da evidência

É muito comum, na indústria bancária, categorizar as variáveis contínuas, de modo que seja possível verificar se a variável discrimina o risco e faz sentido para o negócio. Para auxiliar nessa categorização, é utilizado o "peso da evidência", que mede a força que cada categoria de uma dada variável tem em discriminar entre bons e maus clientes, ou seja, mensura se há diferença expressiva entre bons e maus clientes em cada categoria (SIDDIQI, 2006). A medida peso da evidência (do inglês, *weight of evidence* = *woe*), é calculada, para cada categoria considerada, por

$$WOE = \ln(ODDS),$$

em que a ODDS (chance), no caso do modelo de propensão, consiste da razão entre os clientes que adquiriram algum produto e os que não adquiriram e, no caso do modelo de risco, consiste da razão entre os clientes inadimplentes e os adimplentes.

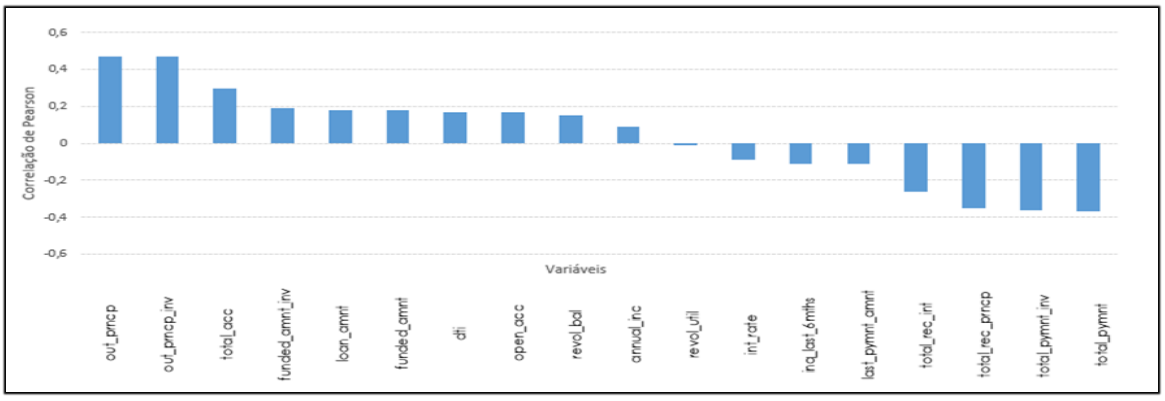
Para o modelo de Cox, os resultados obtidos nesta etapa (frequências, percentuais e *woe*'s), para cada uma das categorias das variáveis, encontram-se na Tabela A1 do Apêndice A. De modo análogo, os resultados obtidos para o modelo de mistura logito-Cox encontram-se na Tabela A2, Apêndice A.

4 RESULTADOS E DISCUSSÃO

4. 1 Resultado das Etapas de Seleção de Variáveis

Para o modelo de propensão ao crédito, a Figura 2 mostra a correlação de Pearson entre 18 variáveis contínuas e a variável resposta tempo até a aquisição do produto. Dentre elas, cinco apresentaram valores entre -0,15 e 0,15 e foram, portanto, excluídas. Quanto às variáveis categóricas, foi excluída apenas uma variável.

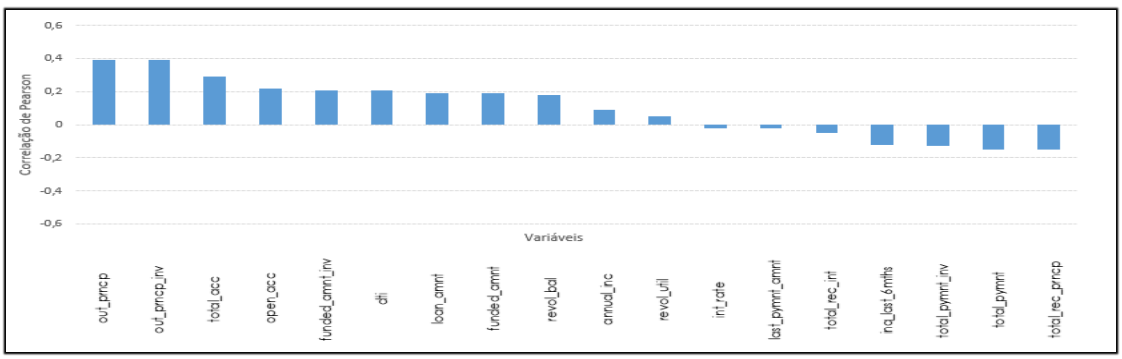
Figura 2 – Correlação de Pearson entre 18 variáveis contínuas e o tempo até a aquisição do produto



Fonte: O autor (2019).

Para o modelo de risco de crédito, a Figura 3 apresenta a correlação de Pearson entre as mesmas 18 variáveis e a variável resposta tempo até a inadimplência. De acordo com o mesmo critério utilizado anteriormente, 7 (sete) delas foram excluídas por terem apresentado correlação entre -0,15 e 0,15. Quanto às variáveis categóricas, foi também excluída apenas uma variável.

Figura 3 - Correlação de Pearson entre as 18 variáveis contínuas e o tempo até a inadimplência



Fonte: O autor (2019).

Ao final desta etapa do processo de seleção, foram excluídas 28 variáveis (contínuas e categóricas). A Tabela 2, a seguir, mostra os valores de informação (IV) obtidos para as variáveis remanescentes nesta etapa do processo de seleção.

Tabela 2 – IV das variáveis para os modelos de propensão (à esquerda) e de risco (à direita)

IV	Variável
0,364	revol_bal
0,302	total_acc
0,293	initial_list_status
0,250	funded_amnt_inv
0,247	funded_amnt
0,247	loan_amnt
0,241	dti
0,220	total_rec_prncp
0,202	out_prncp
0,190	out_prncp_inv
0,156	faixa_tot_cur
0,145	faixa_mths_since
0,106	emp_length
0,080	open_acc
0,064	total_rec_int
0,050	term
0,049	home_ownership
0,047	total_pymnt
0,042	total_pymnt_inv
0,014	grade
0,005	Regiao

IV	Variável
0,942	dti
0,614	funded_amnt_inv
0,524	loan_amnt
0,523	funded_amnt
0,324	out_prncp
0,323	out_prncp_inv
0,231	grade
0,167	revol_bal
0,107	total_rec_prncp
0,068	faixa_total_rev
0,061	faixa_tot_cur
0,053	initial_list_status
0,030	total_acc
0,024	term
0,021	open_acc
0,013	emp_length
0,006	home_ownership
0,003	faixa_mths_since
0,003	total_pymnt

Fonte: O autor (2019).

Após o cálculo do IV, foi calculado o coeficiente de correlação entre as variáveis contínuas, para evitar multicolinearidade. Ao final, apenas as 10 variáveis com o maior IV e com baixa correlação entre elas (variáveis contínuas) foram consideradas nos modelos

Vale salientar que o processo de seleção foi similar para ambas as situações (propensão ao crédito e risco de crédito). Porém, o conjunto de variáveis selecionadas para cada uma das duas situações foi diferente, como pode ser observado nas Tabelas A1 e A2 do Apêndice A, que mostram as variáveis selecionadas.

4. 2 Resultados do Modelo de Propensão ao Crédito: Modelo de Cox

Para ajustar o modelo de Cox, o conjunto de dados contendo informações de 92.347 clientes foi dividido em duas partes: uma delas com 80% (o que corresponde a 73.877 clientes) foi considerada para proceder ao ajuste do modelo, e a outra com 20% (18.470 clientes) foi considerada para proceder à validação do modelo.

Inicialmente, foram consideradas as 10 variáveis mais relevantes no modelo e, então, com o auxílio dos métodos de seleção *stepwise* e *backward* foi selecionado o melhor modelo, que apresentou valor do AIC = 524.377,6 (menor dentre todos os considerados) e concordância igual a 0,804 (maior dentre todas).

Dentre as 10 variáveis consideradas, 8 (oito) delas apresentaram efeito significativo na presença das demais no modelo. As estimativas obtidas para estas 8 variáveis podem ser vistas na Tabela 3.

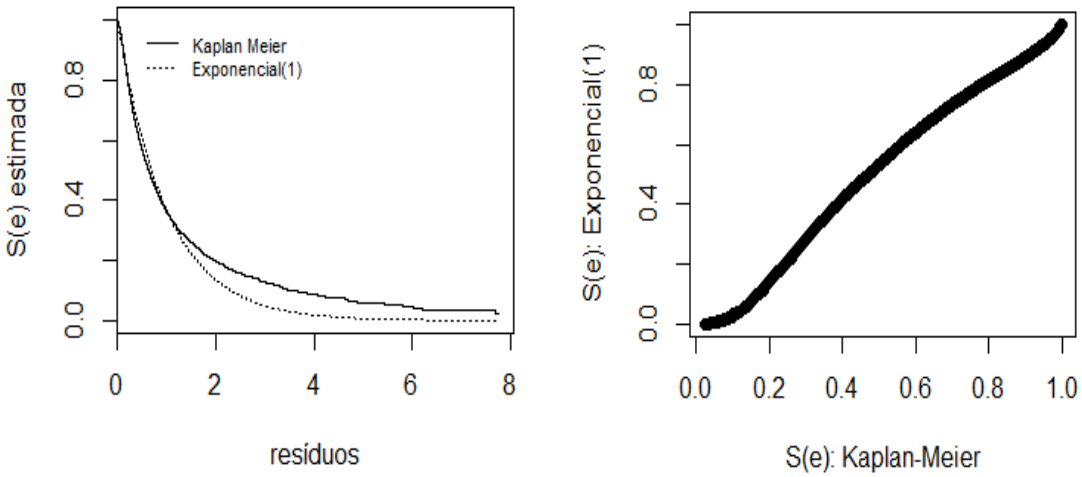
Tabela 3 – Estimativas associadas as oito variáveis significativas para o modelo de Cox

Variável	Descrição	Categoria	Estimativa	p-valor
total_rec_prncp	O montante total comprometido pelos investidores para esse empréstimo naquele momento	0 - 958	-2,158	< 0.001
		958 - 2854	-0,636	< 0.001
		> 5950	-1,219	< 0.001
funded_amnt_inv	O montante total comprometido pelos investidores para esse empréstimo naquele momento	11750 - 15500	-1,245	< 0.001
		15525 - 35000	-1,617	< 0.001
		7000 - 9600	-0,776	< 0.001
		9625 - 11748	-0,994	< 0.001
total_rec_int	Juros recebidos até o momento	0 - 361	-1,223	< 0.001
		361 - 718	-0,787	< 0.001
		718 - 1253	-0,490	< 0.001
total_acc	O número total de linhas de crédito atualmente no arquivo de crédito do mutuário - Pré-Aprovado	14 - 17	-0,351	< 0.001
		18 - 22	-0,526	< 0.001
		23 - 106	-0,886	< 0.001
initial_list_status	O status de listagem inicial do empréstimo. Valores possíveis são - W, F	w	-0,519	< 0.001
tot_cur	Saldo atual total de todas as contas	0 - 17577	-0,420	< 0.001
revol_bal	Saldo total de crédito rotativo	11249 - 15024	-0,118	< 0.001
		15025 - 284435	-0,305	< 0.001
		7827 - 11248	-0,067	< 0.001
dti	Razão utilizando os pagamentos mensais totais da dívida do mutuário sobre o total das obrigações de dívida	13.9 - 25.4	-0,135	< 0.001
		25.5 - 54.4	-0,277	< 0.001

Fonte: O autor (2019).

A adequação global do modelo foi avaliada utilizando-se os resíduos de Cox-Snell. Para tanto, foram obtidos os gráficos das curvas $\hat{S}(\hat{e}_i)_{km}$ *versus* $\hat{S}(\hat{e}_i)_{\exp}$ e dos pares de pontos $(\hat{S}(\hat{e}_i)_{km}\ ;\ \hat{S}(\hat{e}_i)_{exp})$, mostrados na Figura 4. Com base em ambos os gráficos, é possível dizer que o modelo de ajustou razoavelmente bem aos dados, pois os resíduos apresentaram distribuição próxima da exponencial padrão.

Figura 4 – Análise gráfica dos resíduos de Cox-Snell associados ao modelo de Cox



Fonte: O autor (2019).

Para avaliar a suposição de riscos proporcionais, foram utilizados os resíduos padronizados de Schoenfeld, mencionados na Seção 3.2.1. A Tabela 4 mostra os valores dos coeficientes de correlação ρ associados às categorias das variáveis que permaneceram no modelo. Como todos os valores de ρ se encontram próximos de zero, há evidências de não violação da suposição de riscos proporcionais.

Vale mencionar, que quando o tamanho da amostra for grande (como é o caso dos dados analisados neste trabalho), deve-se avaliar a suposição de riscos proporcionais apenas com base nos valores de ρ, tendo em vista que os testes indicarão, em geral, violação da suposição, como pode ser observado pelos valores p apresentados na Tabela 4. Ou seja, a violação da suposição sugerida pelo teste é decorrente do tamanho amostral grande; não há, nesse caso, violação da suposição pois os valores de ρ estão próximos de zero.

Em relação ao poder preditivo do modelo de Cox, este foi avaliado por meio da curva ROC, que se mostrou satisfatória para os tempos considerados, evidenciando bom poder preditivo do modelo. Foram obtidas três curvas ROC (nos tempos 4, 6 e 8 anos), que apresentaram AUROC = 0,89, 0,87 e 0,82, respectivamente, podendo tais curvas serem visualizadas no Apêndice B (Figura B1). Ainda em relação ao poder preditivo do modelo, foram também obtidas três curvas ROC para a base de validação (18.470 registros), que também apresentaram comportamento satisfatório (Figura B2, Apêndice B), com valores AUROC = 0,86, 0,85 e 0,80 nos tempos 4, 6 e 8 anos, respectivamente.

Tabela 4 – Correlação ρ (rho) entre os resíduos padronizados de Schoenfeld e os tempos

Variável	Categoria	rho	p-valor
total_rec_prncp	*0 - 958*	*0,084*	*< 0.001*
	958 - 2854	*0,027*	*< 0.001*
	> 5950	*0,103*	*< 0.001*
funded_amnt_inv	*11750 - 15500*	*0,078*	*< 0.001*
	15525 - 35000	*0,080*	*< 0.001*
	7000 - 9600	*0,067*	*< 0.001*
	9625 - 11748	*0,090*	*< 0.001*
total_rec_int	*0 - 361*	*-0,047*	*< 0.001*
	361 - 718	*-0,008*	*0.018*
	718 - 1253	*0,018*	*0.004*
total_acc	*14 - 17*	*0,091*	*< 0.001*
	18 - 22	*0,109*	*< 0.001*
	23 - 106	*0,144*	*< 0.001*
initial_list_status	*w*	*0,203*	*< 0.001*
tot_cur	*0 - 17577*	*0,164*	*< 0.001*
revol_bal	*11249 - 15024*	*0,058*	*< 0.001*
	15025 - 284435	*0,068*	*< 0.001*
	7827 - 11248	*0,053*	*< 0.001*
dti	*13.9 - 25.4*	*0,079*	*< 0.001*
	25.5 - 54.4	*0,113*	*< 0.001*

Fonte: O autor (2019).

4. 3 Resultados do Modelo de Risco de Crédito: Modelo de Mistura Logito-Cox

Para ajustar o modelo de mistura logito-Cox, o conjunto de dados contendo as informações de 43.713 clientes também foi dividido em duas partes. Uma para o ajuste do modelo de mistura, com 80% (35.770 clientes), e outra com 20% (8.943 clientes) para a validação do modelo.

A seleção de variáveis foi realizada de modo manual, pois o pacote *smcure*, que ajusta o modelo de mistura, não tem implementado um método computacional de seleção de variáveis, como o *stepwise* ou o *backward*. Após o ajuste de diversos modelos, o modelo que melhor se ajustou aos dados foi o com as estimativas apresentadas na Tabela 5. No componente associado à função de sobrevivência $S_U(t\,|\boldsymbol{x})$, assim como no componente associado à probabilidade p(**z**), foram observadas quatro variáveis com efeito significativo em cada componente, sendo somente uma delas comum aos dois componentes.

Tabela 5 – Estimativas das variáveis significativas do modelo de mistura logito-Cox

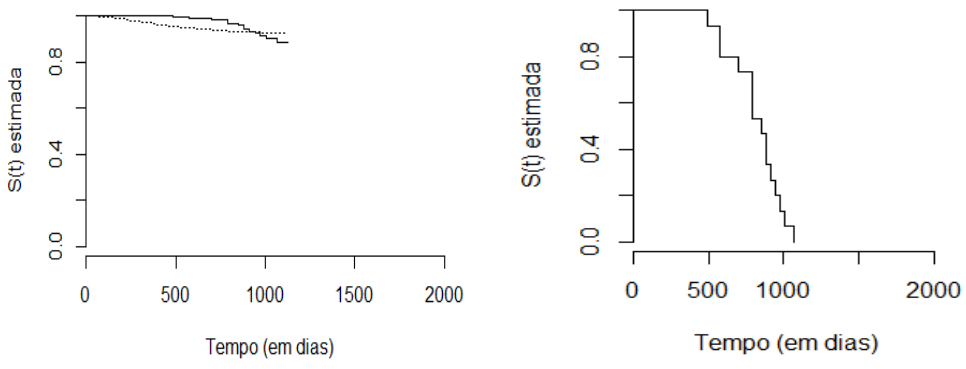
Sobrevivência (Cox)				
Variável	**Descrição**	**Categoria**	**Estimativa**	**p-valor**
funded_amnt_inv	*O montante total comprometido pelos investidores para esse empréstimo naquele momento*	*> 8400*	*1,513*	*< 0.001*
out_prncp	*Capital remanescente para parte do valor total financiado pelos investidores*	*> 1914*	*-2,112*	*< 0.001*
initial_list_status	*O status de listagem inicial do empréstimo. Valores possíveis são - W, F*	*w*	*0,200*	*< 0.001*
dti	*Razão utilizando os pagamentos mensais totais da dívida do mutuário sobre o total das obrigações de dívida*	*18.8 - 54.4*	*-0,142*	*< 0.001*

Regressão Logística - Link: Logito				
Variável	**Descrição**	**Categoria**	**Estimativa**	**p-valor**
Intercepto			*-3,910*	*< 0.001*
total_rec_prncp	*Valor recebido até o momento*	*0 - 1950*	*8,975*	*< 0.001*
		1950 - 3758	*4,343*	*< 0.001*
		3758 - 7310	*2,175*	*< 0.001*
out_prncp	*Capital remanescente para parte do valor total financiado pelos investidores*	*> 1914*	*-0,142*	*< 0.001*
grade	*LC atribuiu nota de empréstimo*	*A*	*0,960*	*< 0.001*
		B	*1,430*	*< 0.001*
		C	*1,626*	*< 0.001*
		D	*1,864*	*< 0.001*
		E	*2,185*	*< 0.001*
		F	*2,886*	*< 0.001*
tot_cur	*Saldo atual total de todas as contas*	*0 - 108696*	*0,068*	*< 0.001*
		> 108696	*-0,485*	*< 0.001*

Fonte: O autor (2019).

A adequação do modelo ajustado foi realizada comparando, para alguns perfis, as estimativas da curva de Kaplan-Meier com as do modelo. A Figura 5 mostra o comparativo dessas curvas para o perfil 2 disposto na Tabela 6, bem como a curva $S_U(t|\boldsymbol{x})$ estimada para os inadimplentes associada ao mesmo perfil. Para 20 perfis selecionados (Tabela 6), a correlação foi elevada (acima de 0.90), mostrando boa adequação do modelo.

Figura 5 – Comparativo das curvas $S_p(t|\boldsymbol{z},\boldsymbol{x})$ estimadas pelo modelo e por Kaplan-Meier (à esquerda) e estimativa da curva $S_U(t|\boldsymbol{x})$ para os inadimplentes (à direita)



Fonte: O autor (2019).

Tabela 6 – Correlação entre as curvas estimadas pelo modelo de mistura e pelo estimador de Kaplan-Meier para 20 perfis de clientes

Estrato	x11	x66	x77	x22	x51	x52	x53	x81	x82	x83	x84	x85	x86	x10a	x10b	n	correlação
1	1	0	0	1	0	0	0	1	0	0	0	0	0	0	0	613	0.973
2	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0	463	0.973
3	0	0	0	1	0	0	1	1	0	0	0	0	0	0	1	420	0.994
4	1	0	0	1	0	0	0	1	0	0	0	0	0	0	1	415	0.958
5	0	0	0	1	0	0	1	1	0	0	0	0	0	0	0	372	0.997
6	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	350	0.96
7	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	327	0.947
8	1	0	0	1	0	0	0	0	1	0	0	0	0	0	1	312	0.947
9	1	0	0	1	0	0	0	0	0	1	0	0	0	0	0	299	0.973
10	0	0	0	1	0	0	1	0	1	0	0	0	0	0	0	283	0.998
11	0	0	0	1	0	0	1	0	1	0	0	0	0	0	1	278	0.993
12	1	0	0	1	0	0	0	1	0	0	0	0	0	1	0	271	0.972
13	1	1	1	0	1	0	0	0	1	0	0	0	0	0	0	269	0.999
14	1	0	1	1	0	0	0	1	0	0	0	0	0	0	0	252	0.995
15	0	1	0	0	1	0	0	0	1	0	0	0	0	0	0	250	0.994
16	1	1	0	0	0	0	1	0	1	0	0	0	0	0	0	242	0.997
17	0	1	1	0	1	0	0	0	1	0	0	0	0	0	0	238	0.998
18	0	1	1	1	1	0	0	1	0	0	0	0	0	0	0	234	0.997
19	0	1	1	1	1	0	0	0	1	0	0	0	0	0	0	233	0.997
20	1	1	0	1	0	0	1	0	1	0	0	0	0	0	0	233	0.996

Fonte: O autor (2019).

Nota: x11 = empréstimo naquele momento; x66 = capital remanescente; x77 = status inicial do empréstimo; x22 = razão entre pagamentos realizados pela dívida ativa x51 a x53 = valor recebido; x81 a x86 = grau atribuído ao empréstimo; x10a e x10b = saldo de todas as contas.

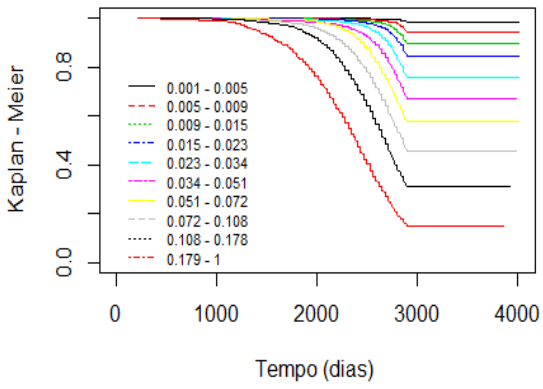
4. 4 Combinação dos Resultados do Modelo de Propensão e do Modelo de Risco

Na indústria bancária, é usual a combinação de dois ou mais escores para a tomada de decisão. Nesse caso, foi combinado os escores de propensão ao crédito e o de risco de crédito, de tal modo a encontrar o melhor perfil de cliente, isto é, clientes com alta probabilidade de propensão ao crédito e baixa probabilidade de inadimplência.

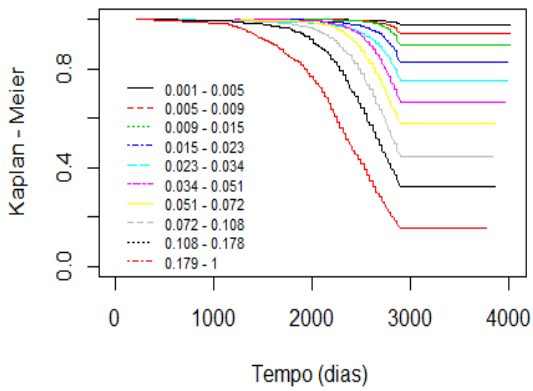
Para encontrar a "régua" de tomada de decisão para o modelo de propensão ao crédito, foi utilizado o exponencial exp($\boldsymbol{x}'\boldsymbol{\beta}$) do modelo de Cox. Para avaliar se esta régua utilizada estaria de acordo com o objetivo, o escore foi dividido em 10 decis e foi plotado as curvas de sobrevivência estimadas pelo estimador de Kaplan-Meier, conforme mostra o Gráfico (a) da Figura 6. Na base de dados de validação do modelo, foi observado comportamento análogo para as curvas, como pode ser visto no Gráfico (b) da Figura 6.

Figura 6 – Curvas estimadas por Kaplan-Meier para as 10 faixas (decis) de escore consideradas

(a) base de dados - ajuste do modelo



(b) base de dados - validação do modelo



Fonte: O autor (2019).

A partir das curvas de sobrevivência mostradas na Figura 6, observa-se que quanto maior o escore mais rápido ocorre a queda da curva estimada, evidenciando que a régua proposta está de acordo com o objetivo. No contexto dos dados analisados, a "régua" indica que os clientes com escore exp($\boldsymbol{x}'\boldsymbol{\beta}$) mais elevados são os que apresentaram perfil com maior propensão à concessão do crédito solicitado.

Nota-se que o escore proposto considerou apenas a parte paramétrica do modelo de Cox. Contudo, também é possível estimar, a partir do modelo ajustado, a probabilidade do evento ocorrer (isto é, de ocorrer a concessão do crédito) em certos tempos de interesse. Sendo assim, tal probabilidade, dada por

$$\hat{F}(\, t \mid \boldsymbol{x}) = 1 - \hat{S}(t \mid \boldsymbol{x}) = 1 - [\widehat{S_0}(t)]^{\exp(\boldsymbol{x}'\widehat{\boldsymbol{\beta}})},$$

foi estimada para alguns tempos associados à pior e à melhor faixa de escores. Os tempos escolhidos foram: a moda e os quartis de cada faixa de escore.

As Tabelas 7 e 8 mostram, para a pior e melhor faixa de escores, respectivamente, as probabilidades de concessão de crédito estimadas em 6 tempos, sendo possível observar que as probabilidades são menores na pior faixa e maiores na melhor faixa. Por exemplo, no tempo 3.804 dias, a probabilidade de aquisição do crédito por um cliente classificado na pior faixa de escores foi estimada em 0,03 (3%), enquanto para um cliente classificado na melhor faixa, tal probabilidade foi estimada em 0,88 (88%).

Tabela 7 – Probabilidades associadas à pior faixa de escores

Pior faixa de escore: 0.001 - 0.005			
Quartil	**Tempo até a aquisição**	**Probabilidade**	**escore**
Min	*1979*	*0,00*	*0,005*
Q25	*3316*	*0,01*	*0,001*
Q50	*3590*	*0,04*	*0,005*
Moda	*3712*	*0,03*	*0,004*
Q75	*3804*	*0,03*	*0,004*
Max	*4016*	*0,01*	*0,001*

Fonte: O autor (2019).

Tabela 8 – Probabilidades associadas à melhor faixa de escores

Melhor faixa de escore: 0.179 - 1			
Quartil	**Tempo até a aquisição**	**Probabilidade**	**escore**
Min	*274*	*0,00*	*0,324*
Q25	*2037*	*0,17*	*0,269*
Q50	*2373*	*0,52*	*0,402*
Moda	*2556*	*0,52*	*0,223*
Q75	*2679*	*0,72*	*0,276*
Max	*3804*	*0,88*	*0,236*

Fonte: O autor (2019).

Por fim, foi calculado a taxa de aquisição (concessão do crédito) em cada faixa de escores, a fim de verificar se a aquisição aumentou conforme aumentou a faixa de escores. Os resultados obtidos estão dispostos na Tabela 9, sendo possível observar que a taxa de aquisição aumenta à medida que o escore cresce.

Tabela 9 – Taxa de aquisição para cada faixa de escore

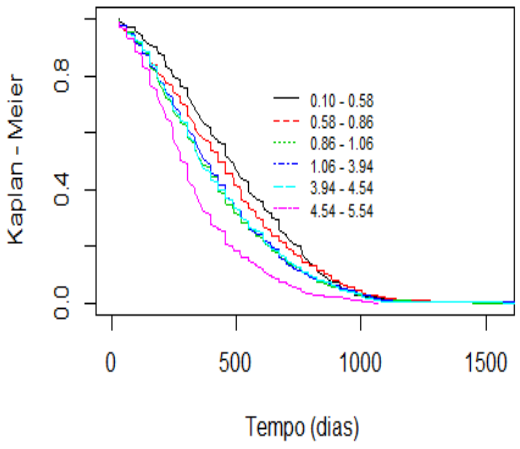
Faixa escore	Faixa de tempo	Taxa de aqusição
0.001 - 0.005	*1613 - 4016*	*0,5*
0.005 - 0.009	*1641 - 4016*	*1,7*
0.009 - 0.015	*1278 - 4016*	*3,0*
0.015 - 0.023	*820 - 4016*	*4,5*
0.023 - 0.034	*606 - 4016*	*7,2*
0.034 - 0.051	*274 - 3985*	*9,6*
0.051 - 0.073	*1156 - 4016*	*12,5*
0.073 - 0.108	*272 - 3985*	*15,9*
0.108 - 0.178	*303 - 3926*	*20,2*
0.179 - 1	*213 - 3863*	*24,9*

Fonte: O autor (2019).

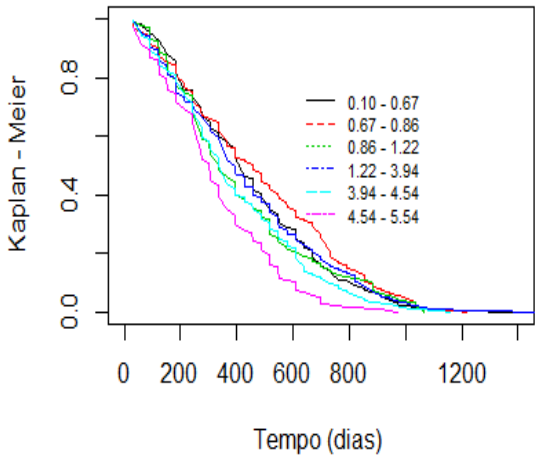
Para o modelo de risco de crédito (modelo de mistura logito-Cox), foi inicialmente utilizado o "escore" obtido a partir do exponencial da soma dos coeficientes associado ao componente $S_U(t \mid \boldsymbol{x})$ (isto é, associado ao componente que considera somente os clientes inadimplentes). Para avaliar a viabilidade dessa "régua", os escores foram divididos em 6 faixas e, para cada uma delas, foi plotada a correspondente curva de sobrevivência estimada por Kaplan-Meier, como mostra o Gráfico (a) da Figura 7. Curvas similares foram obtidas para a base de validação do modelo, podendo ser vistas no Gráfico (b) da Figura 7.

Figura 7 - Curvas de sobrevivência estimadas por Kaplan-Meier para as seis faixas de escore

(a) Base de dados - ajuste do modelo



(b) Base de dados - validação do modelo



Fonte: O autor (2019).

É possível observar, a partir da Figura 7 (a), que as faixas de escore "1,06 a 3,94″ e "3,94 a 4,54″ ficaram bem próximas. Como o escore está sendo calculado apenas para os inadimplentes (12 % da base), os perfis em cada faixa de escore ficaram parecidos. Como na base de validação, o volume de inadimplente é ainda menor (1.086), porém com taxa de inadimplência similar (12%), perfis mais parecidos ficaram mais evidentes, porém mesmo assim, as curvas para faixas de escore extremos são diferentes, indicando que quando maior o escore, mais rápido ocorre a inadimplência.

Da mesma forma que o escore proposto para o modelo de propensão utiliza apenas a parte paramétrica do modelo de Cox, o escore para o modelo de risco foi calculado com base apenas nos clientes inadimplentes (componente $S_U(t \mid \boldsymbol{x})$ do modelo). Para considerar tanto os inadimplentes quanto os adimplentes, pode ser utilizada a função $S_p(t \mid \boldsymbol{z}, \boldsymbol{x})$.

Desse modo, foi calculada a probabilidade $\hat{S}_p(t \mid \mathbf{x}, \mathbf{z})$ para um determinado perfil da pior faixa de escore e outro da melhor faixa de escore. Na melhor faixa, com tempo de inadimplência de 90 dias, a probabilidade de um determinado perfil se tornar mau pagador foi de 0,001; porém, para o outro perfil na pior faixa, com tempo de inadimplência de 914 dias, a probabilidade foi de 0,03. Nota-se que como o volume de inadimplentes é pequeno, não há intersecção de tempos e perfis entre as faixas de escores, impossibilitando comparações entre as faixas de escores para os mesmos tempos de inadimplência.

Combinando os escores, conforme mostrado na Tabela 10, é possível encontrar uma subpopulação com alta probabilidade de aquisição de crédito e baixa probabilidade de inadimplência, dado que o cliente já é inadimplente. A justificativa em construir um escore para os inadimplentes e combiná-lo com outros escores se deve ao fato de que muitas instituições financeiras procuram oportunidades, até mesmo em clientes inadimplentes, devido à concorrência no mercado de crédito com as chamadas *Fintechs*.

Tabela 10 - Combinação do escore de propensão com o escore de risco de crédito

		Modelo de Propensão						
		0.001 - 0.018	**0.018 - 0.031**	**0.031 - 0.051**	**0.051 - 0.08**	**0.08 - 0.133**	**0.133 - 1**	**Total**
Modelo Risco	**0.10 - 0.58**	*2,3%*	*2,1%*	*2,7%*	*3,8%*	*3,4%*	*2,7%*	*17,0%*
	0.58 - 0.86	*1,9%*	*3,7%*	*3,6%*	*2,5%*	*3,5%*	*5,9%*	*21,1%*
	0.86 - 1.06	*2,1%*	*2,7%*	*2,0%*	*2,0%*	*2,0%*	*2,0%*	*12,9%*
	1.06 - 3.94	*3,4%*	*2,6%*	*3,7%*	*4,0%*	*4,2%*	*4,2%*	*22,2%*
	4.94 - 4.54	*2,5%*	*2,5%*	*2,9%*	*2,9%*	*2,7%*	*1,7%*	*15,3%*
	4.54 - 5.54	*4,4%*	*3,1%*	*1,8%*	*1,4%*	*0,8%*	*0,1%*	*11,6%*
	Total	*16,7%*	*16,7%*	*16,7%*	*16,7%*	*16,6%*	*16,7%*	*100,0%*

Fonte: O autor (2019).

Como o modelo de risco apresentou pouca variabilidade devido ao escore ter sido calculado com base apenas nos inadimplentes, a combinação foi realizada em 6 faixas de escores e não em 10 faixas, como usualmente é realizado. Tomando-se, então, as duas melhores faixas de escores do modelo de propensão e do modelo de risco observa-se, na Tabela 10, a existência de uma subpopulação de 5,3% que apresenta alta probabilidade de aquisição de crédito e baixa probabilidade de inadimplência, dado que o cliente já é inadimplente. Essa subpopulação pode ser revista dependendo do apetite de risco e do perfil de cliente que a instituição financeira deseja atacar.

Como o escore associado ao modelo de risco de crédito foi calculado com base apenas nos clientes inadimplentes, a "régua de crédito" se restringiu aos clientes que realizaram a aquisição do produto e tiveram algum tipo de inadimplência no decorrer do tempo. Contudo, é possível construir uma "nova régua" para o modelo de risco que leve em conta tanto os clientes inadimplentes quanto os adimplentes. Tal "régua" consiste em obter as probabilidades $\hat{S}_p(t \mid \boldsymbol{z}, \boldsymbol{x})$ associadas ao modelo de mistura logito-Cox e, então, dividi-las em 10 faixas como mostrado na Tabela 11.

Combinando, agora, o escore de propensão ao crédito com este novo escore, tendo como alvo as melhores faixas de escore de cada modelo, pode-se encontrar a subpopulação com alta probabilidade de propensão ao crédito e baixa probabilidade de inadimplência, conforme mostrado na Tabela 11. Nota-se que como é possível alterar o tempo para realizar a combinação dos escores, o escore de propensão foi calculado para $t = 8$ anos até a aquisição de algum produto, e o escore de crédito para $t = 1$ ano até ficar inadimplente.

Tabela 11 - Combinação do escore de propensão (em t = 8 anos até a aquisição) com o escore de risco de crédito (em t = 1 ano até a inadimplência)

		Modelo de Propensão										
		0.001 - 0.012	**0.012 - 0.02**	**0.02 - 0.028**	**0.028 - 0.037**	**0.037 - 0.048**	**0.048 - 0.059**	**0.059 - 0.078**	**0.078 - 0.104**	**0.105 - 0.157**	**0.158 - 1**	**Total Geral**
Modelo de Risco	**0.379 - 0.792**	*0,2%*	*0,6%*	*1,0%*	*1,2%*	*1,3%*	*1,3%*	*1,0%*	*1,0%*	*1,3%*	*1,2%*	*10,0%*
	0.793 - 0.86	*0,0%*	*0,9%*	*0,7%*	*1,2%*	*1,4%*	*1,7%*	*1,4%*	*1,0%*	*1,1%*	*0,7%*	*10,1%*
	0.86 - 0.869	*0,5%*	*1,1%*	*1,5%*	*1,5%*	*1,6%*	*1,4%*	*1,0%*	*1,1%*	*0,4%*	*0,0%*	*10,1%*
	0.869 - 0.893	*0,0%*	*0,9%*	*1,7%*	*1,8%*	*1,2%*	*1,6%*	*1,3%*	*0,8%*	*0,9%*	*0,2%*	*10,3%*
	0.893 - 0.925	*0,4%*	*0,8%*	*1,5%*	*1,2%*	*1,5%*	*0,9%*	*1,0%*	*0,8%*	*0,5%*	*0,8%*	*9,6%*
	0.925 - 0.95	*1,1%*	*1,1%*	*1,1%*	*0,8%*	*0,9%*	*0,7%*	*0,7%*	*1,3%*	*0,8%*	*1,4%*	*10,0%*
	0.95 - 0.963	*2,4%*	*2,4%*	*1,4%*	*0,7%*	*0,6%*	*0,4%*	*0,3%*	*0,4%*	*0,8%*	*1,1%*	*10,4%*
	0.963 - 0.983	*2,5%*	*1,5%*	*0,5%*	*0,6%*	*0,4%*	*0,5%*	*1,0%*	*0,5%*	*0,7%*	*1,4%*	*9,6%*
	0.983 - 0.99	*2,2%*	*0,4%*	*0,4%*	*0,3%*	*0,4%*	*0,4%*	*0,8%*	*1,4%*	*2,0%*	*1,7%*	*10,0%*
	0.99 - 1	*0,6%*	*0,3%*	*0,5%*	*0,5%*	*0,8%*	*1,1%*	*1,7%*	*1,6%*	*1,4%*	*1,5%*	*10,0%*
	Total Geral	*10,0%*	*10,0%*	*10,2%*	*9,8%*	*10,0%*	*10,0%*	*10,1%*	*9,9%*	*10,0%*	*10,0%*	*100,0%*

Fonte: O autor (2019).

Para os tempos considerados na Tabela 11 (8 anos até a aquisição e 1 ano até a inadimplência), observa-se a existência de uma subpopulação de 14,5% de clientes com alta probabilidade de aquisição do produto e baixa probabilidade de inadimplência.

Combinando novamente os escores, alterando o tempo até a inadimplência para 1,5 ano, obtém-se a configuração exibida na Tabela 12, que mostra uma subpopulação de 11,5% com alta probabilidade de aquisição e baixa probabilidade de inadimplência.

Tabela 12 - Combinação do escore de propensão (t = 8 anos até a aquisição) com o escore de risco de crédito (t = 1,5 ano até a inadimplência)

		Modelo de Propensão										
		0.001 - 0.008	**0.008 - 0.015**	**0.015 - 0.024**	**0.024 - 0.033**	**0.033 - 0.046**	**0.046 - 0.058**	**0.058 - 0.074**	**0.074 - 0.097**	**0.097 - 0.135**	**0.136 - 0.535**	**Total Geral**
Modelo de Risco	**0.308 - 0.808**	*0,0%*	*0,3%*	*0,6%*	*0,9%*	*1,0%*	*1,7%*	*1,3%*	*1,8%*	*1,5%*	*1,4%*	*10,7%*
	0.808 - 0.819	*0,1%*	*0,6%*	*2,0%*	*1,6%*	*1,4%*	*1,6%*	*1,3%*	*1,3%*	*0,9%*	*0,6%*	*11,7%*
	0.82 - 0.828	*0,0%*	*0,2%*	*0,9%*	*1,0%*	*1,7%*	*1,3%*	*1,5%*	*0,9%*	*1,0%*	*0,6%*	*9,3%*
	0.829 - 0.874	*0,1%*	*0,2%*	*0,7%*	*1,0%*	*1,7%*	*1,1%*	*1,1%*	*1,0%*	*1,0%*	*0,4%*	*8,3%*
	0.874 - 0.91	*0,0%*	*0,3%*	*0,8%*	*1,3%*	*1,6%*	*1,4%*	*1,5%*	*1,3%*	*0,9%*	*0,9%*	*10,2%*
	0.911 - 0.947	*0,1%*	*1,3%*	*1,5%*	*1,7%*	*0,6%*	*0,9%*	*1,1%*	*0,7%*	*1,0%*	*1,5%*	*10,6%*
	0.95 - 0.973	*0,9%*	*2,2%*	*1,3%*	*0,9%*	*0,5%*	*0,7%*	*0,6%*	*0,7%*	*0,7%*	*0,8%*	*9,5%*
	0.973 - 0.983	*2,6%*	*2,2%*	*0,9%*	*0,4%*	*0,3%*	*0,6%*	*0,4%*	*0,8%*	*0,8%*	*1,1%*	*10,2%*
	0.983 - 0.99	*4,2%*	*1,6%*	*0,4%*	*0,3%*	*0,3%*	*0,1%*	*0,3%*	*0,4%*	*0,9%*	*1,0%*	*9,7%*
	0.99 - 1	*1,9%*	*0,9%*	*0,7%*	*0,6%*	*0,8%*	*0,6%*	*0,6%*	*1,1%*	*1,1%*	*1,5%*	*9,9%*
	Total Geral	*10,0%*	*10,0%*	*10,0%*	*10,0%*	*10,0%*	*10,0%*	*9,9%*	*10,0%*	*10,0%*	*10,0%*	*100,0%*

Fonte: O autor (2019).

Para finalizar, os escores foram combinados alterando o tempo até a inadimplência para 3 anos, como mostrado na Tabela 13. Para essa situação, tem-se 7,2% de clientes com alta probabilidade de aquisição e baixa probabilidade de inadimplência.

Tabela 13 - Combinação do escore de propensão (t = 8 anos até a aquisição) com o escore de risco de crédito (t = 3 anos até a inadimplência)

		Modelo de Propensão										
		0.435 - 0.746	**0.747 - 0.823**	**0.823 - 0.876**	**0.877 - 0.93**	**0.931 - 0.956**	**0.956 - 0.969**	**0.97 - 0.973**	**0.973 - 0.983**	**0.983 - 0.99**	**0.99 - 1**	**Total Geral**
Modelo de Risco	**0.001 - 0.009**	*0,0%*	*0,0%*	*0,1%*	*0,3%*	*0,5%*	*1,1%*	*0,1%*	*2,0%*	*4,2%*	*1,9%*	*10,0%*
	0.009 - 0.02	*0,1%*	*0,5%*	*0,4%*	*0,7%*	*1,8%*	*1,4%*	*0,1%*	*1,9%*	*2,3%*	*1,0%*	*10,0%*
	0.02 - 0.041	*0,8%*	*1,4%*	*1,3%*	*1,6%*	*1,2%*	*0,6%*	*0,1%*	*0,9%*	*0,5%*	*1,4%*	*9,9%*
	0.041 - 0.067	*1,5%*	*1,0%*	*1,3%*	*1,6%*	*0,7%*	*0,9%*	*0,5%*	*0,4%*	*0,5%*	*1,6%*	*10,0%*
	0.067 - 0.098	*2,4%*	*1,2%*	*1,0%*	*1,1%*	*0,9%*	*0,5%*	*0,9%*	*0,5%*	*0,4%*	*0,9%*	*10,0%*
	0.098 - 0.138	*1,9%*	*0,9%*	*0,6%*	*1,0%*	*1,2%*	*1,2%*	*1,5%*	*0,5%*	*0,6%*	*1,0%*	*10,4%*
	0.139 - 0.186	*1,0%*	*1,0%*	*0,4%*	*0,8%*	*0,9%*	*1,4%*	*1,9%*	*0,7%*	*0,5%*	*0,8%*	*9,6%*
	0.186 - 0.244	*1,1%*	*0,9%*	*0,9%*	*0,5%*	*1,3%*	*1,1%*	*2,0%*	*1,2%*	*0,3%*	*0,7%*	*10,0%*
	0.246 - 0.329	*0,6%*	*1,3%*	*1,5%*	*0,7%*	*1,0%*	*1,0%*	*1,9%*	*1,0%*	*0,3%*	*0,7%*	*10,0%*
	0.337 - 1	*0,7%*	*2,6%*	*2,6%*	*0,5%*	*0,6%*	*0,8%*	*1,1%*	*0,9%*	*0,0%*	*0,1%*	*9,9%*
	Total Geral	*10,0%*	*10,8%*	*10,2%*	*9,0%*	*10,0%*	*10,0%*	*10,1%*	*10,1%*	*9,8%*	*10,0%*	*100,0%*

Fonte: O autor (2019).

Dos resultados apresentados, é possível notar que quanto maior o tempo até a inadimplência, menor é a população alvo. Logo, para a adoção de uma estratégia de concessão de crédito e de risco de crédito, há de se considerar: o perfil da população varia de acordo com o produto oferecido, a rentabilidade desejada, o menor risco tomado, dentre outros aspectos de negócio que podem alterar a população alvo.

Como o objetivo deste trabalho foi o de combinar o escore de propensão ao crédito com o de risco de crédito, ficou evidente que o tempo até a ocorrência do evento altera a população alvo. Utilizando, desse modo, as expressões dos escores de propensão e de risco de crédito propostos neste trabalho (nos tempos os quais se tem interesse), ficou clara a possibilidade de maior assertividade na tomada de decisão.

5 CONSIDERAÇÕES FINAIS

Novas técnicas estatísticas podem auxiliar na tomada de decisões mais assertivas na indústria bancária. Devido à fácil interpretação e fácil aplicação, a regressão logística é usada amplamente para calcular a probabilidade de inadimplência, desprezando, contudo, o tempo até o evento de interesse, uma informação que pode auxiliar na tomada de decisão.

Como alternativa ao modelo de regressão logística, modelos no contexto de análise de sobrevivência podem viabilizar novas oportunidades para discriminar entre bons e maus pagadores, ainda mais com tantas informações disponíveis na era do *Big Data.*

Nesse contexto, o modelo de Cox foi considerado, neste trabalho, para obtenção de um escore de propensão ao crédito alternativo/complementar ao obtido pelo modelo de regressão logística. O ajuste do modelo foi satisfatório, tendo em vista que as faixas de escore consideradas discriminaram os clientes com maior propensão ao crédito. O escore foi calculado utilizando o exponencial da soma dos coeficientes do modelo de Cox.

No contexto de risco de crédito, foi considerado o modelo de mistura logito-Cox, que também apresentou ajuste satisfatório. Com base no modelo citado, foram então propostos dois escores de risco de crédito: o primeiro (que considera apenas os clientes inadimplentes) foi calculado como sendo o exponencial da soma dos coeficientes associado ao componente $S_U(t \mid \boldsymbol{x})$, e o segundo (que considera tanto os adimplentes quanto os inadimplentes) foi calculado como sendo igual a probabilidade $S_p(t \mid \boldsymbol{z}, \boldsymbol{x})$.

Combinando o escore de propensão ao crédito com o primeiro escore de risco de crédito proposto, observou-se uma subpopulação de 5,3% com alta probabilidade de aquisição de produto de crédito e baixa probabilidade de inadimplência, dado que o cliente já era inadimplente. Por outro lado, combinando o escore de propensão ao crédito com o segundo escore de risco de crédito proposto, observou-se que, fixando o tempo de aquisição em até 8 anos e variando o tempo de inadimplência em 1, 1,5 e 3 anos, a subpopulação pertencente às melhores faixas de escores diminuiu de 14,5% para 7,2%, evidenciando que o tempo até a ocorrência do evento altera a população alvo.

Vale salientar que para a tomada de decisões mais assertivas, o tempo até a ocorrência do evento se mostrou, sem dúvida, como uma informação muito rica. No entanto, é importante mencionar que informações relacionadas ao negócio como, por exemplo, a aceitação do risco e a rentabilidade em determinados produtos, além de outros aspectos do negócio, devem ser levadas em consideração para elevar o poder dos escores construídos com base no modelo de Cox e no modelo de mistura logito-Cox.

REFERÊNCIAS

ALMEIDA, M. P. *Estimação bayesiana em modelos de sobrevivência: Uma aplicação em Credit Scoring.* 2008. 78f. Dissertação (Mestrado em Estatística) – Programa de Pós-Graduação em Matemática e Estatística, Universidade Federal do Pará, Pará.

ANDERSEN, P. K., GILL, R. Cox's regression model for counting processes: a large sample study. *Annals of Statistics*, v. 10, p. 1100-1200, 1982.

BANCO CENTRAL DO BRASIL (BCB). *Sistema gerenciador de séries temporais – v2.1*, [Online]. Disponível em: <www3.bcb.gov.br/sgspub> Acesso em: 12 mai. 2019.

BANCO CENTRAL DO BRASIL (BCB). *Relatório anual 2017* [Online]. Disponível em: <http://www.bcb.gov.br/pec/boletim/banual2012/rel2012p.pdf> Acesso em: 12 mai. 2019.

BANASIK, J.; CROOK, J.; THOMAS, L. Not if but when will borrowers default. *The Journal of the Operational Research Society*, v. 50, n. 12, p. 1185-1190, 1999.

BOJANOWSKI, D. Z; LOLATTO, G. A. *Regressão logística e modelos com fração de cura em um estudo sobre clientes inadimplentes*, Trabalho de Conclusão de Curso. 29f., Graduação em Estatística, UFPR, 2018.

BELLOTTI, T.; CROOK, J. Credit scoring with macroeconomic variables using survival analysis. *The Journal of the Operational Research Society*, v. 60, n. 12, p. 1669-1707, 2009.

BRESLOW, N. E. Discussion of Professor Cox's paper. *Journal of the Royal Statistical Society B*, v. 34, p. 2166-217, 1972.

BERKSON, J.; GAGE, R. Survival curve for cancer patients following treatment. *Journal of the American Statistical Association*, v. 47, p. 501–515, 1952.

CAI, C, ZOU, Y, PENG, Y, ZHANG, J. smcure: an R-package for estimating semiparametric mixture cure models. *Comput. Methods Programs Biomed.,* v. 108, n. 3, p. 1255–1260, 2012.

COX, D. R. Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B*, v. 34, p. 187-220, 1972.

COX, D. R Partial likelihood. *Biometrika*, v. 62, p. 269-76, 1975.

COX, D. R., SNELL, E. J. A general definition of residuals. *Journal of the Royal Statistical Society, Series B*, v. 30, p. 248-275, 1968.

COLOSIMO, E. A; GIOLO, S. R. *Análise de sobrevivência aplicada*. São Paulo: Blucher, 2006. 392p.

CORBIÈRE, F.; JOLY, P. A SAS macro for parametric and semiparametric mixture cure models. *Comput. Methods Programs Biomed.*, v.85, n. 2, p. 173-180, 2007.

DURAND, D. *Risk elements in consumer installment financing*. New York: NBER, 1941.

DINIZ, C.; LOUZADA, F. *Métodos estatísticos para análise de dados de crédito.* Minicurso: In 6th Brazilian Conference on Statistical Modelling in Insurance and Finance, Maresias, São Paulo, 2013. 124 p.

EFRON, B. The efficiency of Cox's likelihood function for censored data. *Journal of the American Statistical Association*, v. 72, p. 557-565, 1977.

FISHER, R. A. The use multiple measurements in taxonomic problems. *Annals of Eugenics*, v. 7, p. 179-188, 1936.

FAREWELL, V. T.; PRENTICE, R. L. The approximation of partial likelihood with emphasis on case-control studies. *Biometrika*, v. 67, p. 273-279, 1980.

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

HAND, D. J.; HENLEY, W. E. Statistical classification methods in consumer credit scoring: a review. *Journal of the Royal Statistical Society, Series A*, v.160, n. 3, p. 523–541, 1997.

HEAGERTY, P. J.; ZHENG, Y. Survival model predictive accuracy and ROC curves. *Biometrics*, v. 61, p. 92-105, 2005.

KAPLAN, E. L; MEIER, P. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, v. 53, p. 457-481, 1958.

KUK, A. Y. C.; CHEN, C. H. A mixture model combining logistic regression with proportional hazards regression. *Biometrika*, v. 79, n. 3, p. 531-541, 1972.

LAWLESS, J. *Statistical models and methods for lifetime data*. New York: John Wiley & Sons, 1982.

LEE, E. T. *Statistical methods for survival data analysis*.2nd ed. New York: John Wiley & Sons, 1992.

NARAIN, B. *Survival analysis and the credit granting decision.* In: Thomas, L.C., Crook, J.N. and Edelman, D.B., Eds., Credit Scoring and Credit Control, OUP, Oxford, p. 109-121, 1992.

R CORE TEAM. 2017. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. Disponível em: <https://www.R-project.org/>. Acesso em 20 abr. 2019.

SIDDIQI, N. *Credit risk scorecards: developing and implementing intelligent credit scoring*. SAS Institute, 2006.

STRUTHERS, C. A.; KALBFLEISH, J. D. Misspecified proportional hazards models. *Biometrika*, v. 73, p. 363-369, 1986.

SCHOENFELD, D. Partial residuals for the proportional hazard regression model. *Biometrika*. v. 69, p. 239-241, 1982.

THOMAS, L. C.; EDELMAN, D. B.; CROOK, J. N. *Credit scoring and its applications*. Siam: Philadelphia, 2002.

TSIATIS, A. A. A large sample study of Cox's regression model. *Annals of Statistics*, v. 9, p. 93-108, 1981.

APÊNDICES

APÊNDICE A – Descrição das variáveis remanescentes após as etapas inicias de seleção e valores obtidos do WOE (*weight of evidence*)

Tabela A1 – Descrição das variáveis e *woe* por categoria – modelo de propensão ao crédito (modelo de Cox)

Variável	Categoria	# total	% total	woe
funded_amnt_inv **Comportamento @ O montante total comprometido pelos investidores para esse empréstimo naquele momento**	*0 - 7000*	*19772*	*27%*	*-34,58*
	7000 - 9600	*10144*	*14%*	*-20,34*
	9625 - 11748	*10716*	*15%*	*-7,94*
	11750 - 15500	*14801*	*20%*	*20,08*
	15525 - 35000	*18444*	*25%*	*42,75*
dti **Comportamento @ Razão utilizando os pagamentos mensais totais da dívida do mutuário sobre o total das obrigações de dívida**	*0 - 13.8*	*25976*	*32%*	*-25,20*
	13.9 - 25.4	*33241*	*45%*	*2,84*
	25.5 - 54.4	*14660*	*23%*	*43,24*
revol_bal **Comportamento @ Saldo total de crédito rotativo**	*0 - 7826*	*33246*	*45%*	*-21,32*
	7827 - 11248	*14777*	*20%*	*-5,66*
	11249 - 15024	*11081*	*15%*	*12,09*
	15025 - 284435	*14773*	*20%*	*51,07*
total_acc **Comportamento @ O número total de linhas de crédito atualmente no arquivo de crédito do mutuário - Pré-Aprovado**	*01 - 13*	*23761*	*34%*	*-51,43*
	14 - 17	*14486*	*21%*	*-8,28*
	18 - 22	*14626*	*21%*	*19,12*
	23 - 106	*17735*	*25%*	*74,98*
total_rec_prncp **Comportamento @ Valor recebido até o momento**	*0 - 958*	*18470*	*26%*	*175,85*
	958 - 2854	*22163*	*32%*	*37,62*
	2854 - 595	*14775*	*21%*	*-59,95*
	> 5950	*14639*	*21%*	*-117,49*
total_rec_int **Operação @ Juros recebidos até o momento**	*0 - 361*	*16744*	*26%*	*113,96*
	361 - 718	*13395*	*21%*	*24,46*
	718 - 1253	*13394*	*21%*	*-22,43*
	1253 - 20983	*20093*	*32%*	*-66,61*
emp_length **Cadastro @ Tempo de emprego. Os valores possíveis são entre 0 e 10, onde 0 significa menos de um ano e 10 significa dez ou mais anos.**	*Até 7 anos*	*46414*	*73%*	*-17,19*
	8 anos	*3572*	*6%*	*25,02*
	Acimo de 8 anos	*13640*	*21%*	*61,57*
initial_list_status **Histórica @ O status de listagem inicial do empréstimo. Valores possíveis são - W, F**	*f*	*32495*	*51%*	*-48,21*
	w	*31131*	*49%*	*61,53*
mths_since **Histórica @ O número de meses desde a última inadimplência do tomador.**	*0 - 15*	*4918*	*8%*	*44,14*
	16 - 56	*13421*	*21%*	*43,14*
	57 - 110	*4572*	*7%*	*84,26*
	Missing	*40715*	*64%*	*-25,26*
tot_cur **Comportamento @ Saldo atual total de todas as contas**	*0 - 17577*	*12391*	*19%*	*-11,54*
	> 17577 or missing	*51235*	*81%*	*2,86*

Fonte: O autor (2019).

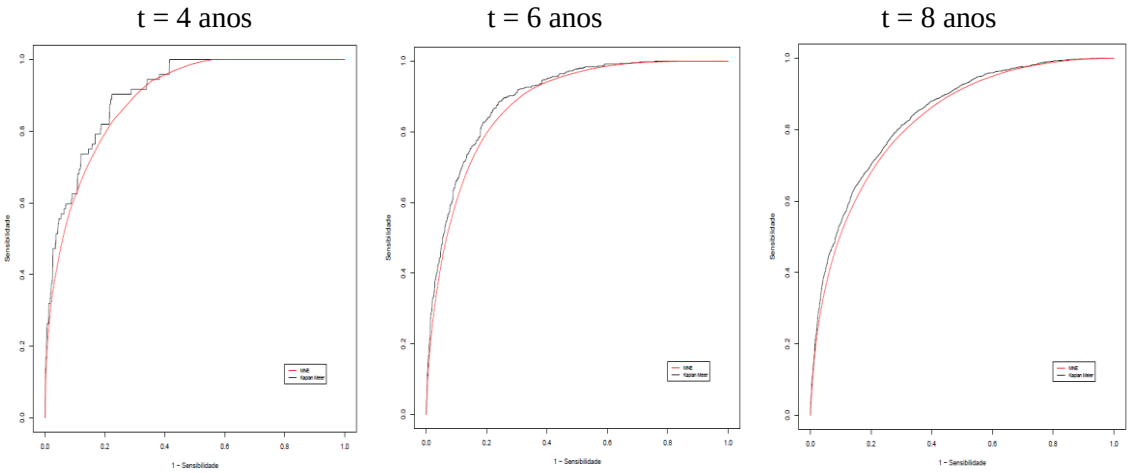
Tabela A2 – Descrição das variáveis e *woe* por categoria - modelo de risco de crédito (modelo de mistura logito-Cox)

Variável	Categoria	# total	% total	woe
funded_amnt_inv **Comportamento @ O montante total comprometido pelos investidores para esse empréstimo naquele momento**	*0 - 8400*	*12776*	*41%*	*7,82*
	8425 - 35000	*18683*	*59%*	*-5,02*
dti **Comportamento @ Razão utilizando os pagamentos mensais totais da dívida do mutuário sobre o total das obrigações de dívida**	*0 - 18.7*	*19083*	*61%*	*8,66*
	18.8 - 54.4	*12376*	*39%*	*-12,04*
revol_bal **Comportamento @ Saldo total de crédito rotativo**	*0 - 5374*	*9386*	*30%*	*-4,47*
	5375 - 519324	*22073*	*70%*	*1,96*
total_acc **Comportamento @ O número total de linhas de crédito atualmente no arquivo de crédito do mutuário - Pré-Aprovado**	*01 - 22*	*25111*	*80%*	*-2,30*
	23 - 79	*6348*	*20%*	*9,63*
total_rec_prncp **Comportamento @ Valor recebido até o momento**	*0 - 1950*	*7050*	*24%*	*-62,80*
	1950 - 3758	*6017*	*20%*	*-27,67*
	3758 - 7310	*8049*	*27%*	*25,58*
	7311 - 35000	*8628*	*29%*	*129,39*
out_prncp **Comportamento @ Capital remanescente para parte do valor total financiado pelos investidores**	*0 - 1914*	*14518*	*46%*	*-52,61*
	> 1914	*16941*	*54%*	*89,99*
initial_list_status **Histórica @ O status de listagem inicial do empréstimo. Valores possíveis são - W, F**	*f*	*11898*	*38%*	*32,59*
	w	*19561*	*62%*	*-15,63*
Variável: grade **Operação @ LC atribuiu nota de empréstimo**	*A*	*3911*	*11%*	*113,50*
	B	*10228*	*29%*	*40,82*
	C	*10393*	*29%*	*-0,94*
	D	*6798*	*19%*	*-35,24*
	E	*3053*	*9%*	*-52,45*
	F	*1128*	*3%*	*-83,62*
	G	*259*	*1%*	*-109,01*
Variável: total_rev_hi_lim **Comportamento @ Total de crédito elevado / limite de crédito rotativo**	*Missing*	*5170*	*14%*	*-48,34*
	0 - 12700	*10827*	*30%*	*-4,38*
	12700 - 24600	*12295*	*34%*	*9,70*
	> 24600	*7478*	*21%*	*35,59*
Variável: tot_cur **Comportamento @ Saldo atual total de todas as contas**	*0 - 108696*	*23539*	*66%*	*5,19*
	> 108696	*7061*	*20%*	*28,28*
	Missing	*5170*	*14%*	*-48,34*

Fonte: O autor (2019).

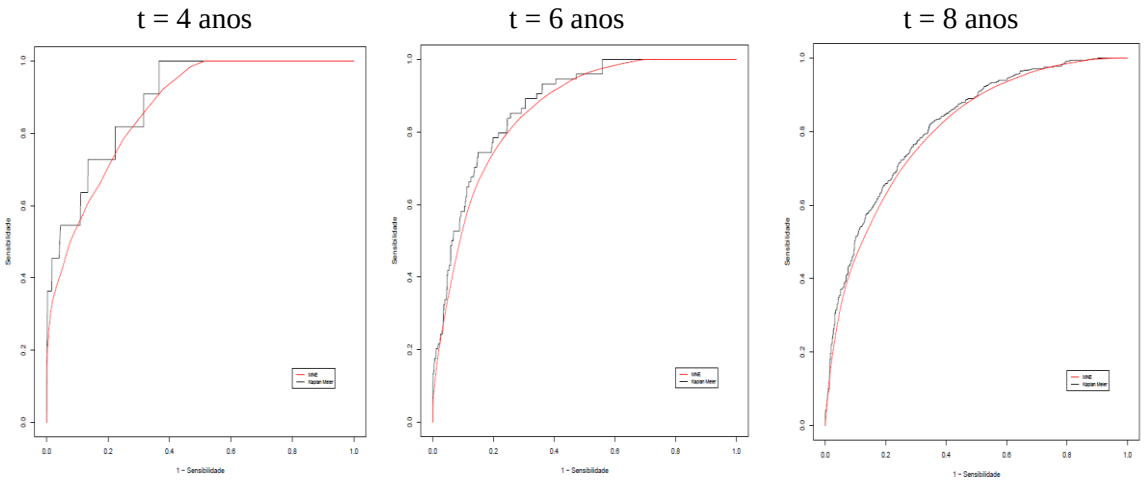
APÊNDICE B – Curvas ROC associadas ao modelo de Cox em diferentes tempos

Figura B1 – Curva ROC nos tempos 4, 6 e 8 anos associadas ao modelo de Cox



Fonte: O autor (2019).
Nota: área abaixo da curva (AUROC): 0.89, 0.87 e 0.82, respectivamente.

Figura B2 – Curva ROC nos tempos 4, 6 e 8 anos associadas ao modelo de Cox (validação)



Fonte: O autor (2019).
Nota: área abaixo da curva (AUC): 0.86, 0.85 e 0.80, respectivamente.

APÊNDICE C – Comandos utilizados no pacote *smcure* para ajustar o modelo de mistura

Lendo e preparando os dados para análise

```
#########################################################################################################
## Obs: a função factor não se aplica a esse pacote; logo se há variáveis categóricas, é necessário construir as
## variáveis dummy para usá-las no modelo
#########################################################################################################

require(smcure)
amostra <- readRDS(file="Amostra_tmp_inad.RData")
attach(amostra)

tempo<-tempo_perf_dias_v2
cens<-bad_new
x11<-ifelse(x1=="0 - 8400",0,1)
x22<-ifelse(x2=="<  18.7",0,1)
x33<-ifelse(x3=="0 - 5374",0,1)
x44<-ifelse(x4=="1 - 20",0,1)
x51<-ifelse(x5=="0 - 1943",1,0)
x52<-ifelse(x5=="1943 - 3750",1,0)
x53<-ifelse(x5=="3750 - 7263",1,0)
x66<-ifelse(x6=="0 - 1914",0,1)
x77<-ifelse(x7=="f",0,1)
x81<-ifelse(x8=="B",1,0)
x82<-ifelse(x8=="C",1,0)
x83<-ifelse(x8=="D",1,0)
x84<-ifelse(x8=="E",1,0)
x85<-ifelse(x8=="F",1,0)
x86<-ifelse(x8=="G",1,0)
x91<-ifelse(x9=="0 - 12700",1,0)
x92<-ifelse(x9=="12700 - 24600",1,0)
x93<-ifelse(x9=="Missing",1,0)
x10a<-ifelse(x10=="108696 - 887964",1,0)
x10b<-ifelse(x10=="Missing",1,0)

dados<-as.data.frame(cbind(tempo,cens,x11,x22,x33,x44,x51,x52,x53,x66,x77,x81,x82,x83,x84,x85,
                                                                    x86,x91,x92,x93,x10a,x10b))
```

Ajustando o modelo de mistura semi-paramétrico: logístico-Cox

```
msm<-smcure(Surv(tempo,cens) ~ x11 + x66 + x77 + x22,
            cureform= ~ x51 + x52 + x53 + x66 + x81 + x82 + x83 + x84 + x85 + x86 + x10a + x10b,
            link="logit", data=dados, model="ph", Var=TRUE, emmax=300, eps=1e-3, nboot=50)
names(msm)
printsmcure(msm, Var=TRUE)
coefsmcure(msm)
```

Função de sobrevivência de base S0(t) -> corresponde ao objeto s das saídas

```
su0<-as.data.frame(cbind(msm$Time, msm$s))        # (tempos, S0(t))
names(su0)<-c("Time","S0")
i<-order(su0$Time)
sbase<-su0[i,]
plot(sbase$Time,sbase$S0, type="s", xlab="Tempos", ylab="S0(t)" )
```

Spop(t) predita para um perfil de indivíduos

```
predm<-predictsmcure(msm,newX = cbind(c(1),c(0),c(1),c(0)),
                     newZ = cbind(c(1),c(0),c(0),c(0),c(1),c(0),c(0),c(0),c(0),c(0),c(0),c(1)),
                     model="ph")

#########################################################################################################
## Obs: newX=(x11,x66,x77,x22) = cbind(c(1),c(0),c(1),c(0))
##      newZ=(x51,x52,x53,x66,x81,x82,x83,x84,x85,x86,x10a,x10b) = acima
#########################################################################################################

plotpredictsmcure(predm, model="ph")

head(predm$prediction)                          # Spop(t) para o perfil definido
predm$newuncureprob                             # uncure probability para o perfil definido
```

 # Spop(t) estimada para dois perfis de indivíduos

```
predm<-predictsmcure(msm, newX=cbind(c(1,0),c(0,0),c(1,1),c(0,0)),
  newZ=cbind(c(1,0),c(0,1),c(0,0),c(0,0),c(1,0),c(0,0),c(0,0),c(0,0),c(0,0),c(0,0),c(0,0),c(1,1)),
  model="ph")
plotpredictsmcure(predm, model="ph")

##############################################################################################################
##  Obs: em newX = cbind(c(1,0),c(0,0),c(1,1),c(0,0)), o 1º valor em cada c(,) corresponde ao valor das covariáveis
##  (x11,x66,x77,x22) associadas ao 1º perfil e o 2º valor em cada c(,) corresponde ao valor das covariáveis
##  (x11,x66,x77,x22) associadas ao 2º perfil.  Idem para newZ.
##############################################################################################################

head(predm$prediction)                          # Spop(t) para os dois perfis definidos
predm$newuncureprob                             # uncure probability para os dois perfis definidos
```

APÊNDICE D – Comandos utilizados no pacote *survival* para ajustar o modelo de Cox

```
#############################################################################################################
## Obs: Utilizando o pacote STEP para encontrar o melhor modelo
#############################################################################################################

mod0 <- coxph(Surv(Tempo_aquisicao_dias,status) ~ 1, data=base, x = F, method="breslow")

step(mod0,~x1+x2+x3+x4+x5+x6+x7+x8+x10, direction ="both", test="Chisq")

fit <- coxph(Surv(Tempo_aquisicao_dias,status)~ x5 + x1 + x6 + x4 + x8 + x10 + x3 + x2,
             data=base, x = F, method="breslow")


#############################################################################################################
## Obs: Calculando os resíduos de Cox Snell
#############################################################################################################

resm<-resid(fit,type="martingale")
res<-base$status - resm
ekm <- survfit(Surv(res, base$status)~1)

par(mfrow=c(1,1))
plot(ekm, mark.time=F, conf.int=F, xlab="resíduos", ylab="S(e) estimada", ylim=c(0,1))
res<-sort(res)
exp1<-exp(-res)
lines(res, exp1, lty=3)
legend(0.3, 1, lty=c(1,3), c("Kaplan Meier","Exponencial(1)"),
       lwd=1, bty="n", cex=0.7)

st<-ekm$surv
t<-ekm$time
sexp1<-exp(-t)
plot(st, sexp1, xlab="S(e): Kaplan-Meier", ylab= "S(e): Exponencial(1)", pch=16)


#############################################################################################################
## Obs: Verificando suposição de riscos proporcionais – resíduos padronizados de Schoenfeld
#############################################################################################################

ro <- cox.zph(fit)

#############################################################################################################
## Obs: curva ROC
#############################################################################################################

temp <- 365 * c(4,6,8)
require(survival)
require(survivalROC)
data=leucc, method="breslow")
pi_cox<-fit$linear.predictors                # marcador Mi(t)
cut<-c(temp)                                 # tempos fixados para predição das AUC
AUC<-matrix(0,length(cut),3)                 # matriz com estimativas AUC(t)
par(mfrow=c(1,length(cut)),3)
for(i in 1:length(cut)){
  cutoff <- cut[i]
  ic.1= survivalROC(Stime=db_cat$Tempo_aquisicao_dias, status=db_cat$status, marker = pi_cox,
                    predict.time = cutoff, method="NNE", lambda=0.02)
  AUC[i,1]<-ic.1$AUC

  ic.2= survivalROC(Stime=db_cat$Tempo_aquisicao_dias, status=db_cat$status, marker = pi_cox,
                    predict.time = cutoff, method="KM")
  AUC[i,2]<-ic.2$AUC; AUC[i,3]<-cut[i]

  plot(ic.2$FP,ic.2$TP, type="l", xlim=c(0,1), ylim=c(0,1), ylab = "Sensibilidade",
       xlab = "1 - Sensibilidade")                                  # ROC método KM
  lines(ic.1$FP,ic.1$TP, type="l", xlim=c(0,1), ylim=c(0,1), col=2)   # ROC método NNE
  legend(0.8, 0.15, c("MNE","Kaplan Meier"),
         lwd=1, bty="l", cex=0.7,col = c(2,1))
}
```