MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DO PARANÁ
SETOR DE CIÊNCIAS EXATAS
DEPARTAMENTO DE ESTATÍSTICA
CURSO DE ESTATÍSTICA

DANIEL DANTAS

ERICSSON A DONADIA

COMPARAÇÃO ENTRE AS TÉCNICAS DE REGRESSÃO LOGÍSTICA, ÁRVORE DE DECISÃO, *BAGGING* E *RANDOM FOREST* APLICADAS A UM ESTUDO DE CONCESSÃO DE CRÉDITO

CURITIBA/PR

2013

DANIEL DANTAS

ERICSSON A DONADIA

COMPARAÇÃO ENTRE AS TÉCNICAS DE REGRESSÃO LOGÍSTICA, ÁRVORE DE DECISÃO, *BAGGING* E *RANDOM FOREST* APLICADAS A UM ESTUDO DE CONCESSÃO DE CRÉDITO

Trabalho de conclusão de curso apresentado à disciplina CE078 - Laboratório de Estatística do curso de Estatística do Setor de Ciências Exatas, Universidade Federal do Paraná.

Orientador: Prof. Dr. Cesar Augusto Taconeli.

CURITIBA/PR

2013

TERMO DE APROVAÇÃO

DANIEL DANTAS

ERICSSON ANGELI DONADIA

COMPARAÇÃO ENTRE AS TÉCNICAS DE REGRESSÃO LOGÍSTICA, ÁRVORE DE DECISÃO, *BAGGING* E *RANDOM FOREST* APLICADAS A UM ESTUDO DE CONCESSÃO DE CRÉDITO

Trabalho apresentado como requisito parcial à obtenção do grau de Bacharel em Estatística no curso de graduação em estatística, pela seguinte banca examinadora:

_____

Prof. Dr. Cesar Augusto Taconeli

Orientador - Departamento de Estatística, UFPR

_____

Profª. Drª. Suely Ruiz Giolo

Departamento de Estatística, UFPR

Curitiba, 05 de agosto de 2013

AGRADECIMENTOS – DANIEL DANTAS

Exprimo o meu, de coração e verdadeiro, “MUITO OBRIGADO”:

À Jesus e Maria pelas bênçãos e graças recebidas.

À minha mãe Vera e minha avó Aparecida, pelo amor, educação e orações.

Ao meu pai Carlos pelo carinho e pela confiança depositada em mim.

À minha noiva Karen, meu alicerce, pelo seu amor, carinho, conselhos e incentivos.

Às minhas irmãs e irmãos pela compreensão da minha ausência como irmão e amigo.

Ao meu amigo e dupla de TCC, Ericsson Angeli Donadia, pelos esforços jamais negados, sem os quais esta conquista não seria possível e por toda a sua inteligência e sabedoria que fazia o difícil virar fácil e o confuso ficar claro.

À todos os meus amigos que me ajudaram nos estudos para que este momento chegasse, em especial o Bruno Dias dos Santos que na maior parte do curso foi quem somou e multiplicou nossas forças.

À todos os professores do curso que se dedicaram para nos ensinar, em especial:

- O professor César A. Taconeli que acreditou em nosso potencial aceitando nos orientar, e que sempre nos ensinou de maneira clara e com propriedade, além de dedicar seus finais de semana para que este trabalho chegasse ao fim;
- A professora Suely R. Giolo que aceitou ser a banca examinadora e por ter uma das melhores didáticas, facilitando nosso aprendizado;
- O professor Anselmo Chaves Neto, que, carinhosamente, nos ensina a parte teórica, prática e cultural da Estatística.
- Os professores do LEG: Elias T. Krainski, Wagner H. Bonat e Walmes M. Zeviani, que sempre sanam nossas dúvidas mesmo fora do horário de aula.

À Dona Elza, Seu Alcides e Vilmar, pessoas queridas, que nos alegram e nos dão forças para trilharmos a vida no curso.

Aos meus amigos que fiz em todas as empresas que trabalhei, em especial aos do HSBC, que me assistiram e apoiaram nestes últimos, e mais sofridos, meses.

AGRADECIMENTOS – ERICSSON ANGELI DONADIA

Inúmeras pessoas foram fundamentais para que este trabalho fosse realizado, a todas elas deixo os meus sinceros agradecimentos. No entanto, algumas destas contribuíram de forma tão intensa que precisam ser citadas:

Minha mãe, Maria Dulce, meu porto seguro, base da minha educação e caráter. Que mesmo longe, sempre esteve ao meu lado me dando apoio e tranquilidade;

Minha vó, Olinda, quem me acompanha desde o meu nascimento, quem sempre cuidou de mim quando estive enfermo e quem sempre me inclui em suas orações;

A Luise, pela contribuiu efetivamente para este trabalho, me ensinando que “onde”, “através” e “quantia” não podem ser utilizados a torto e a direito. Pelo exemplo de força de vontade, por sempre me apoiar e motivar para que eu continuasse meu caminho. Uma linda mulher, a quem eu desejo do meu lado por toda a vida;

A todos os professores e servidores da UFPR, que com muito trabalho e amor à profissão contribuíram para o meu amadurecimento acadêmico. Destes, gostaria de destacar a Dona Elsa e os professores: Anselmo Chaves Neto, Walmes M. Zeviani, Ary E. Sabbag Jr e a professora Suely R. Giolo, que além da sua contribuição em sala de aula, aceitou o convite para compor a banca avaliadora deste trabalho;

A todos os amigos que fiz dentro da sala de aula e levarei para a vida. Os quais me proporcionaram momentos alegres e energizantes durante todo o curso, fazendo assim com que eu nunca tivesse dúvidas das minhas escolhas;

Ao meu irmão Allisson, que, entre outras contribuições, manteve a televisão em um volume aceitável durante todo o tempo em que estive produzindo este trabalho;

Ao Professor Cesar A. Taconeli, que mostrou confiança no nosso trabalho e prontamente aceitou o convite para nos orientar. Fazendo-o com muita dedicação e competência, por vezes trabalhando aos finais de semana para isso, por muitas outras vezes fazendo com que a gente trabalhasse nos finais de semana;

Ao Daniel Dantas, um amigo, que com muita competência, persistência e trabalho duro dividiu comigo o peso da cruz durante toda essa jornada;

A Deus.

EPÍGRAFE

“Se quiser ir rápido, vá sozinho. Se quiser ir longe, vá acompanhado."

RESUMO

Com o aumento pela busca por crédito e a necessidade de uma resposta rápida a este pedido, faz-se necessária a utilização de métodos para mensuração do risco de crédito. Portanto, esse trabalho objetivou comparar diferentes técnicas para este fim, sendo elas: árvore de decisão, regressão logística e ponderação de modelos (*bagging* e *random forest*). Para isso, foi utilizada uma base de dados contendo 1000 observações, que foi dividida em 80% para o ajuste e 20% para a validação do modelo. Após o ajuste dos modelos, a fim de comparar as diferentes técnicas, foram utilizados os indicadores de desempenho: Índice de Kolmogorov-Smirnov, área abaixo da Curva ROC, sensibilidade, especificidade e poder preditivo. Ao fim do estudo, constatou-se que todas as técnicas se adequam à classificação de risco de crédito, porém a regressão logística mostrou-se superior às árvores de decisão. Além disso, o método *bagging* evidenciou melhores ganhos aplicados às árvores do que aos modelos de regressão logística e o *random forest* mostrou-se similar ao bagging aplicado às árvores de decisão para a base de dados utilizada neste trabalho.

Palavras-chave: Concessão de crédito, Regressão Logística, Árvore de Decisão, Ponderação de modelos, *Bagging*, *Random Forest*.

ABSTRACT

With the increase on the demand for credit and the need for a prompt response to this request, the uses of techniques for measuring credit risk are necessary. Therefore, this study aimed to compare different techniques for this purpose, these being: decision tree, logistic regression and aggregation (bagging and random forest). For this, it was used a database containing 1000 observations which was divided into a subset for models induction, with 80% of the data, and to validate, a subset containing the 20% remaining observations. After adjusting the models, for comparison purposes, diagnostic measures such as the Kolmogorov-Smirnov test, area under the ROC curve (AUC), sensitivity, specificity and accuracy were used. At the end of the study, it was found that all techniques fit the credit risk classification, but the logistic regression was superior to decision trees. Furthermore, the bagging method showed better gains applied to decision trees than to logistic regression models and the random forest appeared similar to bagging applied to the decision tree for the database used in this work.

Keywords: Lending, Logistic Regression, Decision Tree, Aggregation, Bootstrap, Bagging, Random Forest.

SUMÁRIO

1. INTRODUÇÃO

Instituições financeiras são intermediadoras entre superavitários, que querem investir seus recursos em troca de rendimentos, e deficitários, que tomam crédito pagando juros a essas instituições. Em toda concessão de crédito há um risco associado à inadimplência do tomador. Para mitigar esse risco, o agente cedente deve avaliar o potencial de retorno do tomador, identificando se o mesmo possui idoneidade e capacidade financeira suficiente para amortizar a dívida que pretende contrair.

De acordo com a Pesquisa Nacional de Endividamento e Inadimplência do Consumidor – PEIC (CNC, 2013), em junho de 2013, cerca de 63% dos brasileiros estavam endividados em produtos como: cheque pré-datado, cartões de crédito, carnês de lojas, empréstimos pessoais, prestações de carro e seguros (Figura 1). Ainda de acordo com a PEIC, 20,3% estavam com dívidas ou contas em atraso e 7,2% não teriam condições de pagar suas dívidas. A procura por crédito no Brasil vem aumentando ao longo dos meses como mostra a Figura 1 e a principal dívida em junho/2013 é devido ao uso cartão de crédito (76,2%), seguido de carnês (17,1%), financiamento de carros (11,4%), crédito pessoal (10%), cheque especial (6,1%), entre outros mostrados na Figura 2.

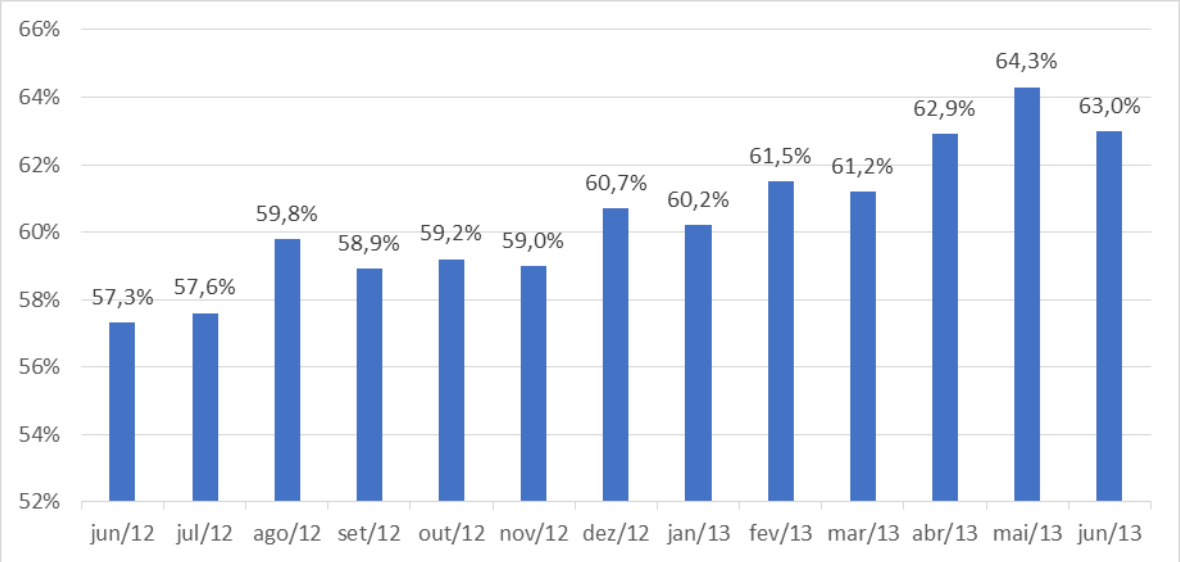

Figura 1 - Percentual de Famílias Endividadas (% do total) entre cheque pré-datado, cartões de crédito, carnês de lojas, crédito pessoal, prestações de carro e seguros.

Fonte: (CNC, 2013).

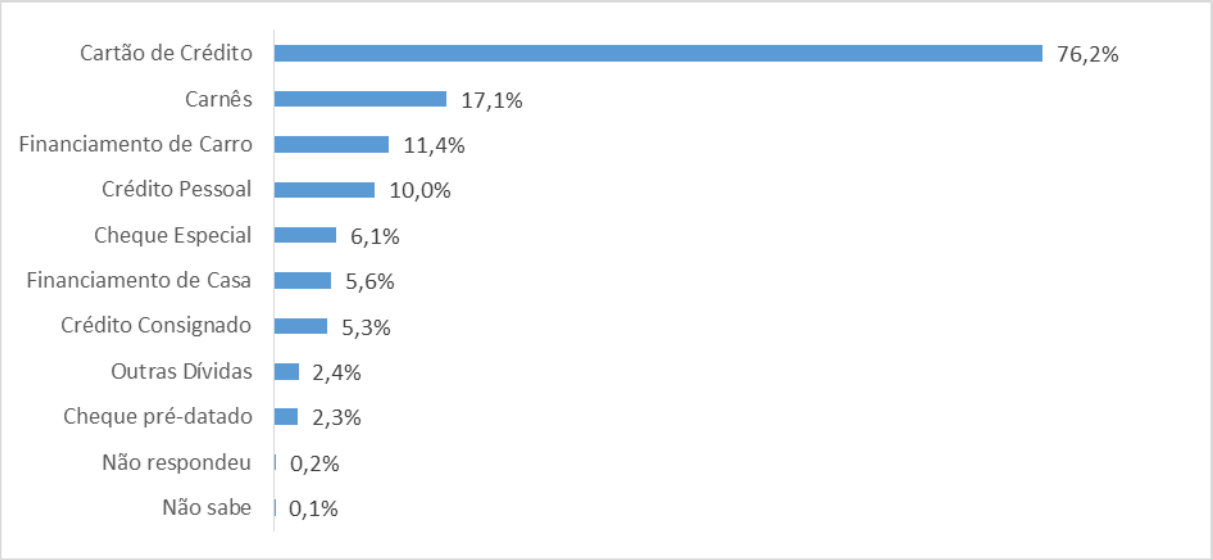

Figura 2 - Tipo de dívida - Junho/2013.

Fonte: (CNC, 2013).

Segundo Santos (2000), o processo de análise e concessão de crédito recorre ao uso de duas técnicas: a técnica subjetiva e a técnica objetiva (ou estatística). A primeira diz respeito à técnica baseada no julgamento humano, principalmente na habilidade e experiência de cada agente de crédito. No entanto, essa técnica demanda tempo e tem custo elevado. A segunda se baseia em métodos estatísticos, que levam em consideração o histórico de créditos concedidos pela instituição para identificar os perfis de bons e maus pagadores, classificando assim os novos proponentes quanto ao risco de crédito (ROSA, 2000).

O aumento da procura por crédito e da concorrência entre instituições financeiras, a exigência por uma melhor qualidade do serviço e a agilidade na decisão do crédito, fazem com que os modelos estatísticos sejam indispensáveis. Com o intuito de reduzir os riscos e aumentar a rentabilidade, diferentes tipos de modelos são utilizados no problema de crédito, sendo alguns deles: as regressões logística e linear, análise de sobrevivência, redes probabilísticas, árvores de classificação, algoritmos genéticos e redes neurais (LOUZADA NETO; DINIZ, 2012).

O objetivo geral deste trabalho é comparar os métodos estatísticos de regressão logística (HOSMER; LEMESHOW, 1989) e árvore de decisão (BREIMAN *et al.*, 1984) na classificação de clientes de acordo com o risco na concessão de crédito. Adicionalmente são utilizadas técnicas de ponderações de modelos, *bagging*

(BREIMAN, 1996) e *random forest* (BREIMAN, 2001), a fim de comparar a capacidade preditiva resultante com as dos métodos individuais. E para que o objetivo geral fosse cumprido, alguns objetivos específicos foram traçados:

- Realizar uma revisão de literatura contemplando métodos de regressão logística, árvore de decisão e alguns de seus algoritmos, técnicas de ponderação de modelos e a aplicação destas técnicas à área financeira;
- Ajustar modelos de regressão logística e árvore de decisão que possam ser utilizados na classificação de risco na concessão de crédito;
- Aplicar métodos de ponderação (*bagging* e *random forest*) nos modelos supracitados;
- Confrontar os resultados dos modelos obtidos de acordo com suas capacidades preditivas baseando-se em técnicas estatísticas;
- Desenvolver uma regra de pontuação de crédito que possa ser utilizada para classificar um novo cliente quanto ao risco de inadimplência.

2. FUNDAMENTAÇÃO TEÓRICA

Técnicas estatísticas de classificação têm por objetivo, basicamente, predizer determinado comportamento de um indivíduo com base em suas características que, de alguma forma, estão correlacionadas com a variável resposta de interesse. Neste trabalho, o interesse é classificar o indivíduo em *bom* ou *mau para o crédito*, baseando-se em informações disponíveis no momento de sua solicitação. As técnicas aplicadas serão abordadas nos tópicos seguintes.

2.1 REGRESSÃO LOGÍSTICA

O modelo paramétrico de regressão logística constitui um método de classificação supervisionada e trata-se de um caso particular dos modelos lineares generalizados (MCCULLAGH; NELDER, 1989), definido pela distribuição binomial com função de ligação canônica (*logit*), apropriado para a modelagem de resposta binária ou categórica. A predição na regressão logística é feita a partir da ponderação das variáveis explicativas (**X**) com base no efeito que cada uma exerce em relação à ocorrência da variável resposta (**Y**), possibilitando assim que se estime a probabilidade de ocorrência do evento de interesse (sucesso ou fracasso, pagar ou não pagar, etc.). No contexto ao qual os dados deste estudo pertencem, ou seja, tomadores de crédito, este evento corresponde ao desempenho creditício dos indivíduos e assume apenas duas possíveis respostas, *bom* ou *mau para o crédito*.

Seja $\boldsymbol{x}_i = \left(x_{i1}, x_{i2}, x_{i3}, \dots, x_{ip}\right)^T$ o vetor dos valores observados de variáveis explicativas (por exemplo: renda, idade, UF, saldo em conta corrente, etc.) para cada indivíduo i, em que $i = (1, 2, \dots, n)$ e n o número total de indivíduos na amostra. A probabilidade de sucesso para o i-ésimo indivíduo é representada por $\pi_i = P(Y_i = 1|\boldsymbol{x_i})$ e a probabilidade de fracasso por $1 - \pi_i = P(Y_i = 0|\boldsymbol{x}_i)$. O modelo de regressão logística pode ser definido da seguinte forma:

$$Y_i|\pi_i \sim binomial\ (n_i, \pi_i).$$

Por aplicar uma *transformação* definida pelo logaritmo neperiano da razão entre π_i e $1 - \pi_i$ tem-se um modelo linear, representado na forma

$$\ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} \quad , i = 1,2,\ldots,n.$$

em que β_i são parâmetros desconhecidos que devem ser estimados, e

$$\pi_i = P(Y = 1|\boldsymbol{x}_i) = \frac{\exp\left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}\right)}{1 + \exp\left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}\right)} \quad , i = 1,2,\ldots,n.$$

Tal *transformação* é denominada *função de ligação canônica* (*logit*) associada ao modelo binomial e permite que π_i, pertencente ao intervalo (0,1), tenha um correspondente no intervalo (-∞ , +∞) com características lineares nas covariáveis (HOSMER; LEMESHOW, 1989).

Portanto, no contexto do estudo, π_i pode ser interpretado como a probabilidade de um proponente ao crédito ser um *bom pagador* dado as características que possui, representadas por $\boldsymbol{x}_i$. No caso da atribuição da categoria *mau pagador*, as interpretações são análogas.

Os parâmetros do modelo usualmente são estimados pelo método da máxima verossimilhança (MV) e a contribuição dos preditores pode ser avaliada pelo teste da razão de verossimilhança (TRV) ou pelo teste de Wald. A função de log-verossimilhança é dada por:

$$l(\boldsymbol{\pi}; \boldsymbol{y}) = \sum_{i=1}^{n} \left[y_i \, log \, \pi_i + (n_i - y_i) \log(1 - \pi_i) + log \binom{n_i}{y_i} \right] \quad ,$$

e o algoritmo para obter a estimativa de máxima verossimilhança é denominado algoritmo dos mínimos quadrados ponderados iterativo (MQPI), como segue para o caso geral dos modelos lineares generalizados:

$$\mathcal{J}^{(m-1)} \boldsymbol{b}^{m} = \mathcal{J}^{(m-1)} \boldsymbol{b}^{(m-1)} + \boldsymbol{U}^{(m-1)},$$

em que m refere-se à iteração; $\boldsymbol{b}^{(m)}$ é o vetor de parâmetros estimados na iteração m; $\mathcal{J}^{(m-1)}$ é a matriz de informação na iteração $m-1$; $\boldsymbol{U}^{(m-1)}$ é o vetor *Score* com as derivadas parciais do logaritmo da função de verossimilhança com relação aos parâmetros desconhecidos β. Seguem as expressões de $\boldsymbol{U}$ e $\mathcal{J}$ específicos para o modelo de regressão logística:

$$\boldsymbol{U} = (U_0, U_1, \ldots, U_p)'$$

$$U_0 = \frac{\delta l(\boldsymbol{\pi}; \boldsymbol{y})}{\delta \beta_0} = \sum_{i=1}^{n} \left\{ y_i - n_i \left[\frac{\exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}} \right] \right\} = \sum_{i=1}^{n} (y_i - n_i \pi_i)$$

$$U_1 = \frac{\delta l(\boldsymbol{\pi}; \boldsymbol{y})}{\delta \beta_1} = \sum_{i=1}^{n} \left\{ y_i - n_i \left[\frac{\exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}} \right] \right\} = \sum_{i=1}^{n} x_i (y_i - n_i \pi_i)$$

$$\vdots$$

$$U_p = \frac{\delta l(\boldsymbol{\pi}; \boldsymbol{y})}{\delta \beta_p} = \sum_{i=1}^{n} \left\{ y_i - n_i \left[\frac{\exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}} \right] \right\} = \sum_{i=1}^{n} x_i (y_i - n_i \pi_i)$$

$$\mathcal{J} = \begin{bmatrix} \sum_{i=1}^{n} n_i \pi_i (1 - \pi_i) & \sum_{i=1}^{n} n_i x_i \pi_i (1 - \pi_i) \\ \sum_{i=1}^{n} n_i x_i \pi_i (1 - \pi_i) & \sum_{i=1}^{n} n_i x_i^2 \pi_i (1 - \pi_i) \end{bmatrix}.$$

Outras funções de ligação bastante utilizadas na análise de dados binários são: a função Probito, que corresponde à inversa da distribuição Normal e a função complemento log-log (HOSMER; LEMESHOW, 1989).

2.1.1 Métodos para Seleção de Variáveis

Segundo Hosmer e Lemeshow (1989), para um modelo ser considerado adequado, além de apresentar um bom ajuste, deve apresentar uma compreensão prática e ser parcimonioso. Para isso, existem métodos de *seleção de variáveis* que reduzem o número de variáveis explicativas, selecionando apenas as que mais contribuem para a explicação da variável resposta. Um algoritmo usado para seleção de variáveis a serem incluídas no modelo é o *stepwise*, que tem por objetivo selecionar variáveis que maximizam o ajuste com o menor número de variáveis empregadas. Partindo de um modelo construído sem qualquer variável explicativa (modelo nulo), o método seleciona variáveis para serem acrescentadas (*forward stepwise*) ou, partindo de um modelo construído com todas as variáveis (modelo saturado), o método seleciona variáveis para serem eliminadas (*backward stepwise*) do modelo. Ainda há outro método, denominado *both stepwise*, que combina os métodos *forward* e *backward*. A cada passo em que variáveis são adicionadas ou removidas do modelo, um novo modelo é ajustado gerando alguma *medida de qualidade de ajuste* (por

exemplo: AIC, BIC, etc.). Ao final, é selecionado o modelo que apresentou o melhor resultado nessa medida (MILLER, 1990). A *medida de ajuste* que foi utilizada no trabalho foi o critério de informação de Akaike (AIC), que pondera a qualidade de ajuste do modelo com a quantidade de parâmetros estimados no modelo (BOZDOGAN *et al*., 1987) e sua expressão é definida da seguinte forma:

$$AIC = -2 \ln L(\boldsymbol{\beta}) + 2p$$

em que:

$L(\beta)$ é o valor da máxima verossimilhança maximizada (FORSTER, 2000) e

p é o número de parâmetros estimados no modelo.

Alguns critérios preliminares para exclusão de variáveis:

- P-valor < 0,20: para selecionar variáveis para a construção do modelo, inicialmente é feita uma análise bivariada testando a relação de cada variável explicativa com a variável resposta. As variáveis explicativas que apresentarem p-valor menor que 0,20, para algum teste estatístico apropriado, são consideradas para possível inclusão no modelo (VICTORA *et al.*, 1997). Pode-se aplicar o teste Qui-quadrado de Pearson para variáveis explicativas categóricas (nominais e ordinais) e, com os valores que são encontrados na tabela ANOVA, o teste F para numéricas.
- Critério baseado na curva ROC: ajusta-se um modelo de regressão logística com apenas uma variável explicativa por meio de uma base selecionada para o ajuste dos modelos e, então, constrói-se uma curva ROC (ver Seção 2.4.1) utilizando os dados de uma amostra selecionada para a validação dos modelos. Se a área compreendida abaixo desta curva (*AUC*) for maior que 0,50, então a variável é considerada para possível inclusão no modelo (KOUKOUVINOS; PARPOULA, 2012).

2.2 ÁRVORES DE DECISÃO

Árvores de decisão configuram métodos que utilizam uma representação gráfica baseada em árvores, cujo objetivo é identificar grupos de indivíduos com

características de interesse em comum. Para tal, é utilizado um método recursivo que divide a amostra inicial em subamostras, baseando-se em resultados observados das variáveis explicativas e em suas interações. Formam-se, assim, grupos para os quais a variável resposta apresenta comportamento homogêneo dentro dos grupos e heterogêneo entre eles (BREIMAN *et al.*, 1984).

Uma árvore de decisão é chamada de Árvore de Classificação se a variável resposta for categórica, ou Árvore de Regressão, se numérica (TACONELI, 2008). Neste trabalho, serão induzidas árvores de classificação, pois a variável resposta é dicotômica (*bom ou mau para o crédito*).

O processo de indução de árvores é iniciado por meio de uma amostra, denominada *nó raiz*, que é dividida em subamostras, denominadas *nós filhos* ou *nós intermediários*. Essas subamostras quando subdivididas são chamadas de nós pais, pois geram *nós filhos*. Quando uma subamostra não puder mais ser subdivida segundo algum critério de parada, é então denominada de *nó final* ou *nó folha*. Esse processo é dito recursivo devido a cada subamostra gerar novas subamostras. A estrutura de uma árvore de decisão está exemplificada na Figura 3.

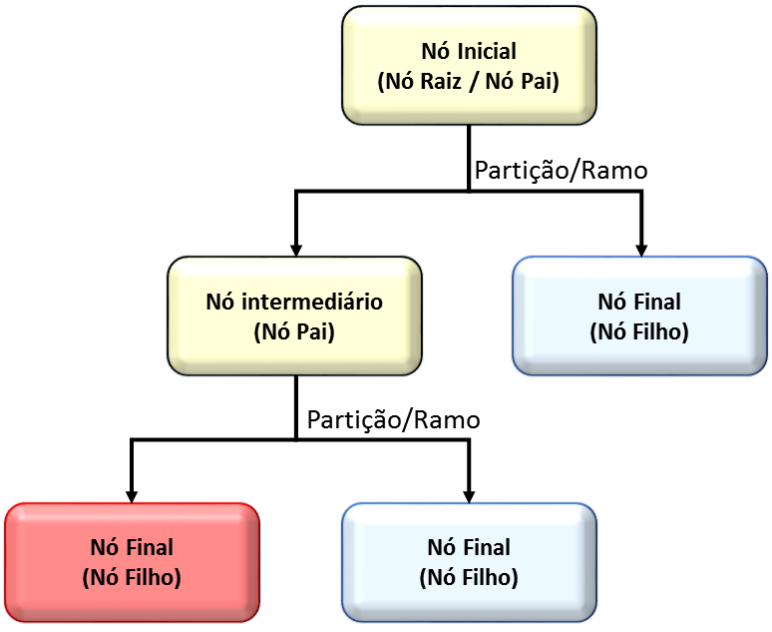

Figura 3 - Estrutura de uma árvore de decisão.

Fonte: Elaborado pelos autores.

2.2.1 Algoritmos de indução de árvores de decisão

Entre os algoritmos para indução de árvores de decisão mais citados nas bibliografias estão:

- CHAID (*Chi-square Automatic Interaction Detection)* (KASS, 1980):
    - Método de partição: Constrói-se uma tabela de contingência com *r (linhas) por c* (colunas) entre variável dependente e variáveis independentes, categorizando em classes as variáveis contínuas. Em seguida, realiza-se um conjunto de testes estatísticos, que variam de acordo com o tipo da variável resposta, agregando as classes de cada variável explicativa até restarem apenas duas, de modo a descobrir o melhor agrupamento de classes. Repete-se para todas as variáveis e a que apresentar a menor probabilidade de significância (*p-valor),* ajustada pelo método de Bonferroni, é então escolhida.
    - Critério de parada: Investiga todas as variáveis até ser encontrada uma partição significativa em cada nó. Portanto, o algoritmo não continua a agrupar as demais categorias, o que pode impedir que uma divisão melhor seja encontrada para aquela variável.
    - Vantagens: Interrompe o crescimento da árvore antes do problema de *overfitting* ocorrer, ou seja, não há tratamento de "poda" (explicado na Seção 2.2.3), gerando ganhos em tempo de processamento.
    - Desvantagens: Necessita de grandes quantidades de dados para assegurar que a quantidade de observações dos nós folhas seja representativa. É possível que o algoritmo não encontre a melhor divisão do nó, pois ele interrompe a busca de partições quando encontra a primeira partição significativa.

- CART (*Classification and Regression Trees)* (BREIMAN *et al.*, 1984):
    - Método de partição: Divide as variáveis de forma binária, ou seja, sempre em dois nós, com base em alguma medida de impureza, como por exemplo o índice de Gini, Entropia, etc. (explicado na Seção 2.2.2), para tornar os subconjuntos de dados cada vez mais homogêneos em relação a variável resposta.

- Critério de parada: Enquanto as divisões trouxerem ganhos em relação à medida de impureza, o algoritmo continua. Existem outros critérios de parada, definidos pelo usuário, que são explicados na Seção 2.2.3.
- Vantagens: Não precisa realizar qualquer tipo de categorização, pois o algoritmo não faz restrições quanto às escalas das variáveis explicativas, podendo estas serem numéricas (discretas ou contínuas), ordinais e nominais. Por utilizar partições binárias, as variáveis podem aparecer em diferentes níveis do modelo, permitindo reconhecer diversas interações com outras variáveis.
- Desvantagens: Por utilizar partições binárias, pode aumentar a complexidade da árvore (muitos níveis de profundidade), o que pode dificultar a apresentação e interpretação dos resultados.

- ID3 (*Iteractive Dichotomizer 3)* (QUINLAN, 1986):
  - Método de partição: A escolha da variável a ser particionada é feita utilizando o método de *Ganho de Informação*, que busca maximizar a medida de impureza (no caso do ID3 usa-se a medida da entropia - Seção 2.2.2) dos *nós filhos* relativo ao *nó pai*.
  - Critério de parada: Enquanto as divisões aumentarem o Ganho de Informação, o algoritmo continua.
  - Vantagens: Simplicidade. Processo de construção de fácil compreensão de seu funcionamento.
  - Desvantagens: Imutável. Uma vez construída a árvore, não se pode eficientemente reutilizar a árvore sem a reconstruir. Não lida com variáveis contínuas, a não ser que sejam discretizadas. Não trata valores desconhecidos, ou seja, todos os dados da amostra para a construção da árvore devem ter valores conhecidos para todos as variáveis. Não inclui nenhum método de poda (explicado na Seção 2.2.3).

- C4.5 (QUINLAN, 1993):
  - Método de partição: Procura sobre um conjunto de variáveis, aquela que “melhor” divide os dados em relação à variável resposta, guiado pela medida estatística de razão do ganho de informação proposta em

Quinlan (1993) que se mostrou superior ao ganho da informação, citado no ID3.

- Critério de parada: Enquanto as divisões aumentarem a razão de ganho de informação, o algoritmo continua.
- Vantagens: Lida com variáveis categóricas (nominais ou ordinais) e numéricas (discretas ou contínuas). Trata valores desconhecidos, mas não os usa para o cálculo de entropia e de ganho. Por usar a medida de razão do ganho de informação, gera árvores mais precisas e menos complexas. Combate o problema de *overfitting,* ou seja, possui método de poda que faz uma busca na árvore, de baixo para cima, e transforma em nós folha aqueles ramos que não apresentam nenhum ganho significativo. Realiza validação cruzada com dois ou mais grupos (*v-fold* ou validação *Jacknife)* diminuindo a estimativa do erro cometido pelo classificador (WITTEN; FRANK, 2005).
- Desvantagens: Não considera valores faltantes para o cálculo da razão de ganho de informação.

O algoritmo escolhido para a indução das árvores de decisão deste trabalho foi o CART. Essa escolha foi motivada pelo fato do algoritmo não fazer restrição quanto à natureza das variáveis explicativas, por gerar resultados de fácil interpretação e ter facilidade em encontrar interações entre as variáveis. Além disso, lida sem dificuldades com dados faltantes, fato comum quando se trata de concessão de crédito.

2.2.2 Critérios de partição

No algoritmo CART um *nó pai* sempre dará origem a dois *nós filhos* (partição binária). Essa partição é feita por meio de uma regra aplicada em uma variável explicativa selecionada, que poderá ser lida da seguinte forma: ‘SE <condição> ENTÃO <decisão>’. Tal condição se dá de maneiras diferentes dependendo da escala da variável selecionada, como apresentado a seguir:

- Seja X_i uma variável explicativa categórica, sendo $U = \{A, B, C, \dots\}$ as categorias observadas na amostra. As partições candidatas baseiam-se em regras do tipo:

$x_i \in S$, para todo $S \subset U$.

Por exemplo, se a variável selecionada para realizar a partição for “estado civil”, essa compreendendo “solteiro”, “casado”, “divorciado” e “viúvo”, uma possível regra seria ‘SE casado ENTÃO *nó filho* 1, SE solteiro, divorciado ou viúvo ENTÃO *nó filho* 2’;

- Seja X_i uma variável explicativa categórica ordenável, sendo $U = \{A_1, A_2, A_3, \dots, A_k\}$ as categorias observadas na amostra. As partições, neste caso, baseiam-se em regras do tipo:

$x_i \in S$, sendo $S = \{A_1, A_2, A_3, \dots, A_j\}$, para todo $\mathrm{j} = 1, 2, \dots, \mathrm{k}$.

Por exemplo, se a variável utilizada para realizar a partição for “nível de escolaridade”, compreendendo: “sem estudo”, “ensino fundamental”, “ensino médio” e “ensino superior”, tem-se uma ordem lógica nessa variável que deve ser seguida, mesmo ela não sendo numérica. Uma possível partição seria ‘SE sem estudo ou ensino fundamental ENTÃO *nó filho* 1, SE “ensino médio” ou “ensino superior” ENTÃO *nó filho* 2’;

- Seja X_i uma variável explicativa numérica, com valores amostrados $x_{i1}, x_{i2}, x_{i3}, \dots, x_{in}$. As partições candidatas tem a seguinte forma:

$x_i \leq x_{ij}$, para todo $j = 1, 2, 3, \dots, n$.

Por exemplo, se temos a variável “idade” compreendendo indivíduos de 20 a 70 anos, qualquer valor presente na amostra entre esses dois números é candidato a ponto de divisão, sendo que indivíduos com idade superior a esse formariam um *nó filho* e com idade inferior o outro.

Para selecionar qual variável e de que forma essa irá dividir o nó, tem-se que identificar a partição que produz maior redução de impureza. Para tal, faz-se necessária a utilização de alguma medida de impureza. Segundo Fonseca (1994), entre os critérios disponíveis para tal estão o método da entropia, o critério de Gini, o

método da paridade, a técnica de Laplace e também a escolha randômica. Dentre estes optou-se pela utilização do critério de Gini, visto que é bastante utilizado, mais favorável à otimização numérica, além de ser a medida padrão do algoritmo CART, que está implementado no *software* R (HASTIE *et al.*, 2001).

O índice Gini, avaliado em um nó *t,* é dado por:

$$gini(t) = 1 - \sum_{i=1}^{k} {p_i}^2$$

em que:

p_i: proporção de indivíduos de cada classe do nó *t*, *i* = 1,2,..., k e

k: número de classes da variável resposta.

O índice Gini assume seu valor máximo quando todas as classes da variável resposta possuem igual distribuição dentro do nó, e mínimo quando este é composto por apenas uma classe.

Depois de dividir t em dois subconjuntos t_1 e t_2 com tamanhos n_1 e n_2, o índice Gini dos dados divididos é definido como:

$$gini_{split} = \frac{n_1}{n} gini(t_1) + \frac{n_2}{n} gini(t_2).$$

A regra selecionada para a primeira partição da amostra é aquela que maximizar, entre todas as candidatas, a redução do índice de Gini do *nó pai* para os *nós filhos*. A partir desta regra, a amostra inicial será dividida em duas subamostras, nas quais o mesmo procedimento é aplicado e assim sucessivamente, até que um critério de parada pré-estabelecido, conforme será discutido na Seção 2.2.3, seja atingido.

Também pode ser aplicado custo de má classificação (será discutido na Seção 2.2.4) no processo de divisão do nó, substituindo o índice de Gini pelo custo de má classificação. Deste modo, a variável selecionada para dividir o nó será aquela que apresentar maior redução no custo do *nó pai* para os *nós filhos*.

2.2.3 Seleção da árvore

O algoritmo CART, inicialmente, expande a árvore até que a quantidade máxima de *nós finais* seja alcançada (exaustão) respeitando um critério de parada pré-estabelecido, o qual é definido, por exemplo, por um número mínimo de indivíduos em um nó folha ou um número mínimo de indivíduos para a partição de um nó. Esta expansão pode ocasionar partições que pouco contribuem para a explicação da variável resposta e, por vezes, refletem apenas ruídos ou erros. Então, faz-se necessário desfazer essas partições, a fim de garantir interpretabilidade, estabilidade e acurácia na árvore.

Denomina-se poda o processo que consiste em desfazer as partições que menos contribuem para a explicação da variável resposta, resultando em uma sequência de árvores de diferentes tamanhos (número de *nós finais*) que variam do tamanho máximo atingido pela árvore original até a menor árvore possível, isto é, apenas o nó raiz, e a cada uma dessas árvores está relacionado um nível de complexidade.

Esse processo parte da definição de uma função do tipo custo-complexidade (BREIMAN *et al.*, 1984). Sejam T_{MAX} a maior árvore construída inicialmente e $\tilde{T}$ o conjunto de *nós finais* para uma subárvore T qualquer de T_{MAX}. Sejam, ainda, $\left|\tilde{T}\right|$ o número de *nós finais* de T e $\alpha \geq 0$ uma constante real denominada *parâmetro de complexidade.* Define-se a função de *custo-complexidade* como

$$R_{\alpha}(T) = R(T) + \alpha\left|\tilde{T}\right|,$$

sendo $R(T) = \sum_{t \in T} \phi(t)$ o custo associado à taxa de má-classificação da árvore T e $\phi(t)$ é o índice de impureza de Gini para o nó t.

Quanto maior for a árvore, menor será o erro de classificação que ela apresentará, dado que o algoritmo procura separar a amostra em nós cada vez mais homogêneos, ficando assim muito específico para o conjunto de indução (overfitting). Logo, essa árvore estendida excessivamente pode não apresentar um ajuste tão bom na hora de predizer novas amostras. Por este motivo, para encontrar a melhor árvore entre todas as criadas no processo de poda é utilizada a validação cruzada (WITTEN; FRANK, 2005). Essa técnica tem como objetivo fornecer medidas de qualidade

preditiva baseadas em dados não utilizados no ajuste do modelo por meio de construções de sucessivos modelos em que, em cada um, uma fração dos dados seja separada para posterior validação.

Todo esse processo pode ser representado por meio de uma curva de custo-complexidade (Figura 4).

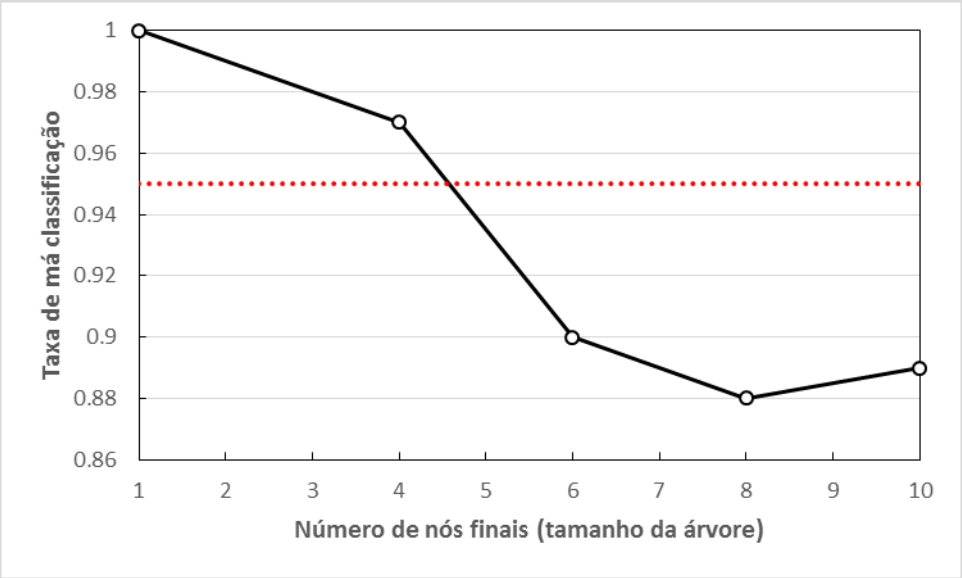

Figura 4 – Representação gráfica para Validação Cruzada.

Fonte: Elaborado pelos autores.

No eixo horizontal do gráfico se encontram os números de *nós finais* das árvores e o eixo vertical apresenta o erro relacionado a cada uma das árvores de diferentes tamanhos. A linha tracejada marca o menor erro encontrado somado ao seu desvio padrão, estimados no processo de validação cruzada. Neste caso, esse erro foi o da árvore de tamanho oito, com um valor de 0,89 e desvio padrão de 0,05. A árvore eleita como de tamanho ideal será a menor árvore que apresentar um erro abaixo da linha tracejada. No gráfico acima a árvore selecionada seria a de tamanho seis.

2.2.4 Classificação dos *nós finais*

A classificação dos *nós finais* se dá a partir da proporção da variável resposta presente em cada nó folha da árvore selecionada. Em certos casos, o nó é classificado de acordo com a frequência predominante da variável resposta, ou seja, se tal *nó final* possui mais de 50% de indivíduos ditos *bons*, então esse nó será classificado como *bom*, e o contrário se aplica. Outra forma de classificação dos *nós finais* da árvore é incorporar *custos de má classificação*, considerando o prejuízo ao classificar um indivíduo como *mau* sendo ele *bom* ou então classificar um indivíduo como *bom* sendo ele *mau*; em outras palavras, aceitar um indivíduo *mau* ou rejeitar um indivíduo *bom*.

Para a aplicação da segunda forma citada, foi adotado o custo de 500 *Deutsche Mark* (*DM*) ao aprovar o crédito para um *mau* pagador, e 100 *DM* ao não aprovar o crédito para um *bom* pagador. Nesses termos, um *nó final* só será classificado como *bom* se apresentar uma frequência superior a 83,3% de *bons*, pois a partir dessa proporção espera-se obter lucro com a aprovação dos indivíduos pertencentes ao nó, caso contrário espera-se prejuízo. A título de exemplo, na Tabela 1 estão expostos alguns cenários com nós de diferentes proporções de *bons* e *maus* para o crédito e o lucro esperado caso o crédito para estes seja aprovado.

Tabela 1 - Exemplo do lucro esperado ao aprovar o crédito aos indivíduos classificados como “bom para o crédito” em cada nó.

% Bom	% Mau	Lucro esperado
78,0%	22,0%	-3200
80,0%	20,0%	-2000
82,0%	18,0%	-800
83,3%	16,7%	0
84,0%	16,0%	400
86,0%	14,0%	1600
88,0%	12,0%	2800
90,0%	10,0%	4000

Fonte: Elaborado pelos autores.

Para a classificação de um novo indivíduo, é preciso alocá-lo a um dos *nós finais* da árvore de acordo com suas características, expressas por meio de seu vetor de variáveis explicativas, de acordo com o conjunto de regras que compõem a árvore.

Uma vez alocado a um *nó final*, a mesma regra usada para a classificação do nó se aplica à classificação desse novo indivíduo.

2.3 PONDERAÇÃO DE MODELOS

A seleção do melhor modelo envolve, além da aplicação de conceitos e técnicas estatísticas, considerações subjetivas. Com isso, a conclusão ao final do estudo, se não considerada a incerteza devido à escolha do modelo, pode acarretar em uma subestimação da variabilidade de quantidades de interesse e/ou inferência super-otimista ou viciada (BUCKLAND et. al., 1997). Com o intuito de mitigar esses vícios e incertezas podem ser aplicados os métodos de ponderação de modelos (HASTIE *et al*., 2001).

A ponderação de modelos, segundo Buckland *et al.* (1997), é dada pela ponderação de estimativas de alguns parâmetros comuns a todos os modelos em estudo, sendo que os pesos desta ponderação são obtidos a partir do uso de critérios de informação ou do método *bootstrap* (EFRON; TIBSHIRANI,1993). A ponderação de modelos assume uma situação em que são considerados K modelos com o objetivo de estimar o parâmetro de interesse. Por fim, cada modelo ajustado fornecerá uma estimativa deste parâmetro e um peso para esta estimativa.

Introduzida por Efron (1979), a técnica de reamostragem *bootstrap* tenta realizar o que seria desejável na prática: repetir o experimento, se tal fosse possível. Para isso, assume a amostra inicial como população do estudo, permitindo que se realizem diversas reamostragens com reposição, gerando pseudo-amostras, a partir das quais se pode estimar e construir intervalos de confiança para diferentes parâmetros da população de interesse, buscando maior precisão e menor vício para as estimativas (EFRON; TIBSHIRANI,1993).

2.3.1 Bagging

O *bagging* (*bootstrap aggregating*) (BREIMAN, 1996) é um método de ponderação de modelos utilizado para aumentar a estabilidade e a precisão de

técnicas de predição. O método consiste na geração de uma série de pseudo-amostras a partir da técnica de reamostragem *bootstrap*, na qual para cada uma dessas amostras será ajustado um modelo preditivo (regressão logística, árvore de decisão, etc.). A predição neste método é dada a partir de uma agregação de todas as predições individuais geradas.

Segundo Breiman (1996) e Optiz e Maclin (1999), realizar uma combinação de preditores de modelos de classificação, em geral, fornece um erro de predição ou classificação menor do que os modelos individuais utilizados para construí-los. Bühlmann e Yu (2002) confirmam a proposição de Breiman (1996) de que o *bagging* reduz a variância e também o erro quadrático médio das predições, para o caso de árvores de decisão.

O algoritmo *Bagging*, aplicados a métodos preditivos, é dado a seguir:

1. Gerar *B* amostras *bootstrap* a partir dos dados originais;
2. Para cada uma das amostras *bootstrap*, ajustar um modelo. Para o caso das árvores de decisão, não realizar o processo de poda;
3. Predizer novos dados agregando as predições de todos os *B* modelos, o que pode ser feito com base na média das probabilidades estimadas fornecidas por cada modelo. Em caso de respostas qualitativas, também é usual a utilização da classificação por proporção de votos. Neste caso conta-se quantas vezes o indivíduo foi classificado em cada classe. Se a proporção de vezes que o indivíduo for classificado em uma classe for maior que um valor preestabelecido (tal como 50%), então será classificado nela.

2.3.2 Random Forest

O algoritmo para induzir uma *Random Forest* foi desenvolvido por Leo Breiman e Adele Cutler. O método combina a ideia do *bagging* e a seleção randômica de variáveis explicativas no processo de indução da árvore. Essa seleção trata-se de um sorteio feito a cada nó da árvore, selecionando aleatoriamente algumas variáveis candidatas para dividir este nó. Com a utilização dessa técnica, diferentes conjuntos de variáveis poderão aparecer em níveis distintos em cada uma das árvores. Com isso, a técnica se torna mais sensível a interações entre as variáveis, além de resultar

em árvores decorrelacionadas, devido ao sorteio aleatório das variáveis candidatas a dividir o nó feito a cada partição (BREIMAN, 2001).

Em muitos problemas, o método *Random Forest* tem apresentado um nível de ajuste altíssimo comparado com outros algoritmos. Esta técnica é executada eficientemente em grandes bases de dados e possibilita o processamento de milhares de variáveis, dispensando a necessidade de exclusões, pois, por meio da seleção de variáveis, remove aquelas redundantes ou indesejáveis sem prejudicar o desempenho de classificação.

O algoritmo *Random Forest* é dado a seguir:

1. Gerar *B* amostras *bootstrap* a partir dos dados originais;
2. Para cada uma das amostras *bootstrap*, induzir uma árvore sem poda (tamanho máximo) com a seguinte modificação: em cada nó, selecionar aleatoriamente um número *m* das variáveis explicativas a serem candidatas para dividir o nó e, dentre estas, escolher a que melhor particiona. (O *Bagging* é um caso particular do *Random Forest* quando este critério não é utilizado, ou seja, quando todas as variáveis são candidatas a dividir cada um dos nós).
3. Predizer novos dados agregando as predições de todos os *B* modelos, o que pode ser feito com base na média das probabilidades estimadas fornecidas por cada modelo. Em caso de respostas qualitativas, também é usual a utilização da classificação por proporção de votos. Neste caso, conta-se quantas vezes o indivíduo foi classificado em cada classe. Se a proporção de vezes que o indivíduo for classificado em uma classe for maior que um valor preestabelecido (tal como 50%), então será classificado nela.

2.4 CRITÉRIOS PARA AVALIAÇÃO DOS MODELOS

Os modelos de concessão de crédito têm como principal objetivo discriminar os indivíduos classificados como “bons para o crédito” dos “maus para o crédito”. Existem alguns métodos padrões na estatística que permitem mensurar e comparar o

desempenho preditivo de modelos contextualizados neste problema. A seguir, serão apresentadas medidas de performance para auxiliar na seleção do modelo adequado.

2.4.1 Curva ROC

Oliveira e Andrade (2002) sugerem a utilização da curva ROC para avaliar a performance de modelo preditivos por se tratar de uma técnica bastante útil para mensurar a capacidade preditiva de modelos de risco de crédito, baseando-se nos conceitos de *sensibilidade* e *especificidade*. Essa técnica também é utilizada para identificar possíveis pontos de corte[1], possibilitando a escolha do ponto que gera o maior poder preditivo para o modelo ou maior lucro esperado, o qual utiliza o custo da má classificação descrito na Seção 2.2.4.

Para a construção da curva ROC se faz uso das medidas de *sensibilidade* e *especificidade* e a partir delas pode-se obter o *poder preditivo* do modelo. Essas medidas são descritas a seguir:

- A sensibilidade é a probabilidade de acerto na previsão da ocorrência de um evento. Logo, seria o modelo predizer que o cliente é *bom para o crédito,* quando ele de fato é;
- A especificidade é a probabilidade de acerto na previsão da não ocorrência de um evento. Logo, seria o modelo predizer que o cliente é *mau para o crédito,* quando ele de fato é;
- O poder preditivo do modelo é definido pela probabilidade de acerto nas predições desse modelo.

Uma forma de se estabelecer e visualizar o cálculo amostral dessas medidas é por meio de uma tabela 2x2 denominada matriz de confusão (LOUZADA NETO; DINIZ, 2012), que é representada na Tabela 2.

[1] Ponto de corte indica que indivíduos com valores preditos iguais ou maiores a esse são classificados, por exemplo, como *bons pagadores* e abaixo desse valor como *maus pagadores.*

Tabela 2 - Matriz de Confusão.

	Real		
Previsão	**M**au	**B**om	Total
mau	m_M	m_B	**m**
bom	b_M	b_B	**b**
Total	**M**	**B**	**n**

Fonte: Elaborado pelos autores.

Assim, tem-se que:

- $Sensibilidade = b_B/B$;
- $Especificidade = m_M/M$;
- $Poder\ preditivo = (m_M + b_B)/n$.

Para a construção da curva ROC, são calculados as medidas de sensibilidade e especificidade considerando todos os possíveis pontos de corte para a predição do modelo. A curva é obtida com a construção de um gráfico onde no eixo vertical apresenta-se a “Sensibilidade” e no eixo horizontal “1 – Especificidade”, como mostra a Figura 5. A área formada sob essa curva mede a capacidade de discriminação do modelo, e quanto mais a curva estiver próxima do canto superior esquerdo, maior será a área abaixo da curva e o poder preditivo do modelo.

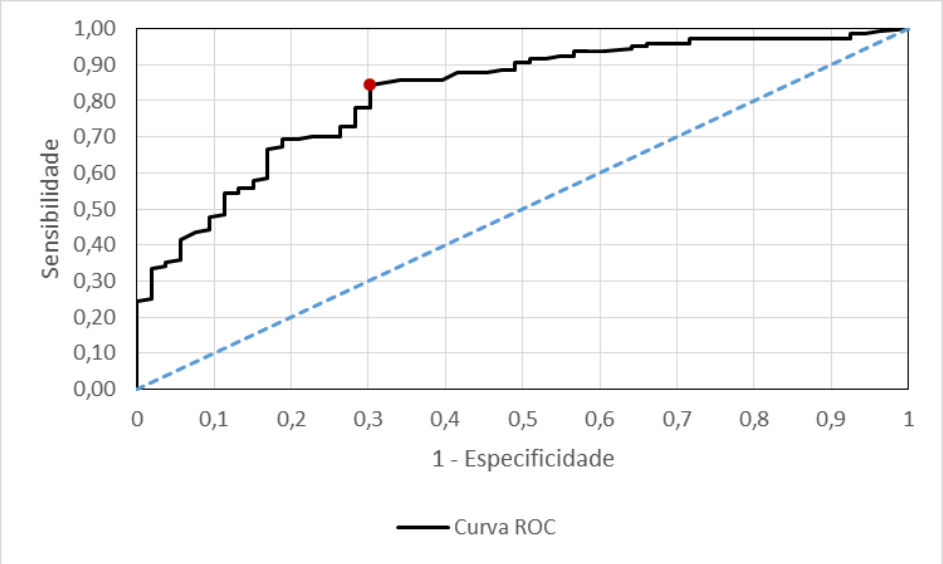

Figura 5 – Ilustração da Curva ROC.

Fonte: Elaborado pelos autores.

2.4.2 Índice Kolmogorov-Smirnov (KS)

Trata-se de uma técnica não paramétrica para determinar se duas amostras pertencem a uma mesma população (SIEGEL,1975). No entanto, para o contexto de concessão de crédito, essa técnica é utilizada para quantificar a capacidade preditiva do modelo, descrevendo o quanto os grupos de bons e maus pagadores são diferentes com relação a probabilidade estimada por esse modelo, o que possibilita a identificação de qual modelo apresenta melhor discriminação (ALVES, 2008).

Ainda no contexto de concessão de crédito, para o cálculo do índice KS é exigido que a coluna de probabilidades preditas para os indivíduos bons ou a de maus seja ordenada. Após, constrói-se uma curva da distribuição acumulada de bons pagadores e outra para os maus pagadores. Então, é calculada a diferença absoluta entre as duas curvas para todos os seus pontos e a estatística KS é dada pela maior dessas diferenças, como ilustra a Figura 6.

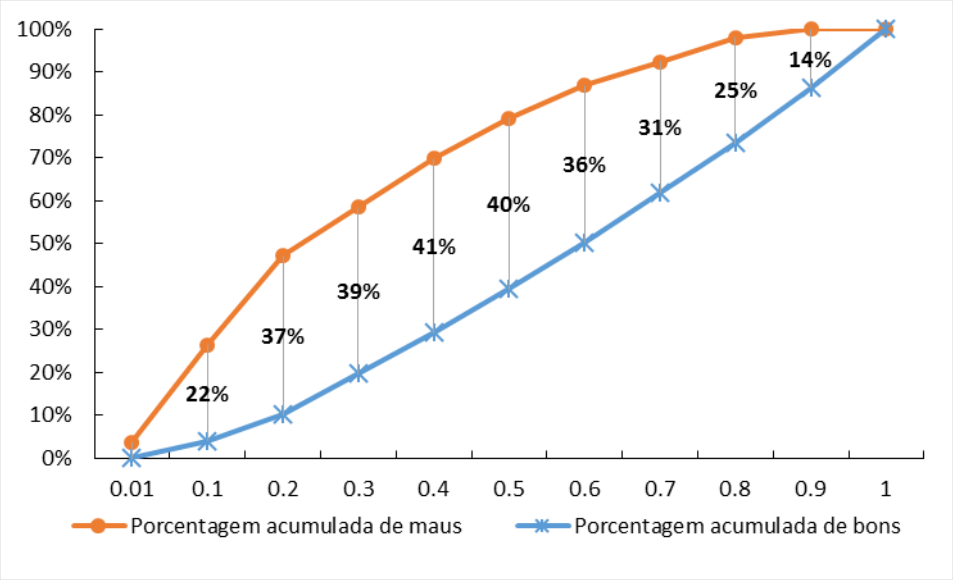

Figura 6 – Ilustração para o cálculo do índice KS.

Fonte: Elaborado pelos autores.

Na Figura 6, a maior diferença entre as curvas acumuladas de *bons* e *maus* pagadores é de 41%. Logo, esse é o valor do KS para o modelo preditivo testado.

Em outros contextos, o KS é utilizado como um teste estatístico de hipóteses para verificar se duas amostras foram geradas por uma mesma distribuição, sob um nível de significância estatística α, e então é verificada essa similaridade com o apoio

de uma tabela cujos valores dependem também do tamanho das amostras (ver SIEGEL, 1975, p. 309-310). Como neste trabalho a amostra é consideravelmente grande, a tendência é que todos os modelos rejeitem a hipótese de igualdade nas distribuições. Por essa razão, será considerado o melhor modelo aquele que possuir o maior valor da estatística no teste, pois este resultado indica uma separação maior entre bons e maus pagadores.

3. MATERIAIS

Como ferramentas de auxílio para a realização deste trabalho foram utilizados, basicamente, três *softwares:* R (R CORE TEAM, 2012), SAS (SAS INSTITUTE INC, 2012) e Microsoft Excel 2013 (MICROSOFT, 2013). Os pacotes computacionais disponíveis para serem executados no *software* R e que foram utilizados neste trabalho são: rpart (THERNEAU *et al*., 2013) para indução das árvores de decisão, ROCR (SING *et al.,* 2005) para construção das curvas ROC e randomForest para aplicação da técnica *Random Forest* (LIAW; WIENER, 2002).

O banco de dados que foi utilizado neste trabalho contempla informações de uma instituição financeira alemã, no ano de 1994, disponibilizado por Bache e Linchman (2013). A *base inicial* contém 1.000 observações de clientes, compreendendo 20 variáveis explicativas e uma variável resposta, a qual é categórica dicotômica, construída com base em informações de pagamento dos clientes, indicando se estes são *bons* ou *maus para o crédito*, distribuída em 700 bons (70%) e 300 maus (30%). A lista de variáveis explicativas, suas descrições e frequências amostrais são apresentadas no Capítulo 5, nas Tabelas 3 e 4.

Para que fosse possível aplicar as metodologias propostas e quantificar a qualidade de ajuste dos modelos, a base inicial foi dividida, de forma aleatória, em duas amostras, uma para o *ajuste* e outra para a *validação* do modelo ajustado, compreendendo o percentual de 80% e 20%, respectivamente. A base de ajuste é composta por 800 indivíduos, sendo que 553 (69,1%) foram considerados como *bons para o crédito* e 247 (30,9%) como *maus para o crédito*. Por sua vez, a base de validação é composta por 200 indivíduos, dos quais 147 (73,5%) foram considerados como *bons para o crédito* e 53 (26,5%) como *maus para o crédito*.

4. MÉTODOS

Com o objetivo de encontrar a árvore de decisão que melhor se ajusta aos dados, por meio do algoritmo CART, foi induzida, inicialmente, uma árvore de decisão (denominada ***T1***) utilizando como critério de parada um tamanho mínimo necessário para que o nó fosse dividido e um tamanho mínimo estabelecido para compor os *nós finais*, respectivamente 7% e 3% do total da base. Aplicando em *T1* o processo de poda, descrito na Seção 2.2.3, chegou-se a uma segunda árvore ***(T2),*** vale ainda ressaltar que, o processo de poda foi executado 300 vezes para cada árvore, por nem sempre apresentar o mesmo resultado, e foi selecionado o tamanho que por mais vezes apareceu.

Adicionalmente foi aplicado o custo referente à má classificação de um indivíduo (explicado na Seção 2.2.4) na indução de uma árvore de decisão, respeitando os critérios de parada citados acima (***T3***). Posteriormente, também foi aplicado o processo de poda em *T3*, resultando em uma árvore denominada ***T4***.

As árvores serão apresentadas em ilustrações que indicam como cada partição é feita e a proporção de *maus* e *bons* que constituem cada *nó final*, assim como a proporção do total da amostra inicial que cada nó representa.

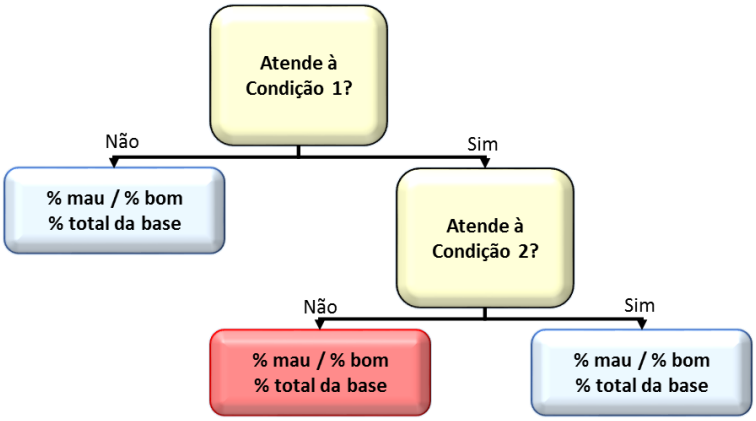

Figura 7 - Exemplo da ilustração das árvores de decisão deste trabalho.

Fonte: Elaborado pelos autores.

Para o caso de regressão logística, também com o objetivo de encontrar um modelo adequado, foram utilizados os métodos de seleção de variáveis, o algoritmo “both stepwise”, o critério de inclusão “p-valor < 0,20” e “critério baseado na curva

ROC", citados na Seção 2.1, resultando em seis modelos diferentes de regressão logística como seguem:

- Modelo com todas as variáveis (***M1*)**;
- Modelo com todas as variáveis aplicado "both stepwise" (***M2***);
- Modelo aplicado o critério de inclusão "p-valor < 0,20" (***M3***);
- Modelo aplicado o critério de inclusão "p-valor < 0,20" e posteriormente o algortimo "both stepwise" ***(M4)***;
- Modelo aplicado "Critério da curva ROC" (***M5***);
- Modelo aplicado "Critério da curva ROC" e posteriormente o algoritmo "both stepwise" (***M6***);

Ademais, no modelo logístico que apresentou melhores resultados, foram acrescentadas interações duplas entre variáveis, evidenciadas em duas diferentes árvores de decisão, uma delas considerando custos e a outra não. É importante ressaltar que foram testadas todas as interações dessas duas árvores, mas apenas as que evidenciaram melhor resultado foram mantidas no modelo.

- Modelo que apresentou melhores resultados com adição de interações da árvore que não considera custos (***M7***).
- Modelo que apresentou melhores resultados com adição de interações da árvore que considera custos (***M8***).

O método *bagging* foi aplicado nas duas técnicas mencionadas acima (regressão logística e árvore de decisão). Como citado na Seção 2.3.1, esse método consiste em diversas reamostragens feitas na base de ajuste. Logo, para sua aplicação se faz necessário definir uma quantidade suficiente de reamostragens que garanta estabilidade nos resultados. Para tal, o *bagging* foi aplicado utilizando diferentes valores para *B*, variando de 10 até 500. Com tais resultados foi averiguado a partir de qual valor de *B* o erro de classificação se torna estável.

Para que se chegasse a bons resultados aplicando o método Bagging em árvores de decisão, vários parâmetros foram testados. Como para essa técnica não se aplica poda nas árvores, testou-se vários valores distintos de limitadores do tamanho final de cada árvore, chegando a um *nó final* de no mínimo 3% da amostra e

um *nó pai* de pelo menos 7%. Assim, foram ajustados quatro modelos diferentes, conforme descritos a seguir:

- BT1 – Árvores de decisão sem a aplicação de custo de má classificação e predições feitas a partir da média de predição de cada modelo (Seção 2.3.1);
- BT2 – Árvores de decisão sem a aplicação de custo de má classificação e com predições feitas a partir da proporção de votos (Seção 2.3.1);
- BT3 – Árvores de decisão com a aplicação de custo de má classificação e predições feitas a partir da média de predição de cada modelo;
- BT4 – Árvores de decisão com a aplicação de custo de má classificação e com predições feitas a partir da proporção de votos.

O ajuste da regressão logística com a utilização do *bagging* foi feito de quatro modos distintos. Inicialmente sem a utilização de nenhum critério de seleção de variáveis e com a predição dada baseada na média das probabilidades estimadas fornecidas por cada modelo (***BM1***). O ***BM2*** foi ajustado de forma análoga ao primeiro mas com suas estimativas dadas por meio de proporção de votos. Para o ajuste do ***BM3*** foi feita a seleção de variáveis por meio do critério da Curva ROC e suas estimativas dadas com base na média das probabilidades estimadas fornecidas por cada um dos modelos. O ***BM4*** foi ajustado de forma análoga ao *BM3*, mas suas estimativas dadas por meio de proporção de votos.

Antes de aplicar o método de *Random Forest*, é necessário, assim como no *bagging*, definir o valor de *B* que garanta a estabilidade dos resultados. Então, foram induzidas diversas árvores de decisão, por meio do método de *Random Forest*, variando o valor de *B* entre 1 e 500. Com as taxas de má classificação dessas árvores foi possível identificar o número *B* necessário para que o método fique estável.

Outro parâmetro a ser definido é a quantidade de variáveis explicativas candidatas a partição de cada nó sorteadas aleatoriamente, conforme descrito na Seção 2.3.2. Desta maneira, foram ajustados 4 modelos de *random forest* variando essa quantidade em: 2, 5, 10 e 15 variáveis, sendo esses modelos denominados: ***RF1***, ***RF2***, ***RF3***, ***RF4***, respectivamente.

Devido a limitações no pacote utilizado, que não disponibiliza a probabilidade predita, e sim a classificação que o indivíduo teve em cada uma das árvores, a classificação final se deu por meio de proporção de votos (ver Seção 2.3.2).

Para mensurar a qualidade de ajuste atingida por cada um dos métodos foram utilizadas algumas técnicas. Para verificar o quanto o modelo consegue discriminar *bons* e *maus* foi utilizado o índice Kolmogorov-Smirnov, tanto na base de ajuste quanto na de validação. As medidas de sensibilidade, especificidade e poder preditivo também foram utilizadas, possibilitando assim, a partir de um ponto de corte estabelecido, verificar o grau de acerto do modelo na predição de *bons pagadores, maus pagadores* além do grau de acerto total. A partir dessas medidas pôde-se construir a curva ROC, que é uma forma visual de avaliar o ajuste. Além disso, foi calculada a área abaixo da curva utilizando a base de validação. Como a proposta do trabalho é ajustar um modelo para a concessão de crédito, faz-se necessário medir o lucro esperado para a aplicação prática de cada técnica, logo, este valor também é analisado no trabalho. Adicionalmente, para a técnica de regressão logística, foram verificados os valores do AIC, que é uma medida de qualidade do ajuste que o penaliza por meio do número de parâmetros estimados.

5. RESULTADOS

5.1 ANÁLISE DESCRITIVA

As Tabelas 3 e 4 apresentam uma análise descritiva de todas as variáveis explicativas disponíveis na base inicial (1000 observações), segmentadas em numéricas e categorizadas, respectivamente.

Para as variáveis numéricas, apresentadas na Tabela 3, foram calculados a média e o desvio padrão, de acordo com a variável resposta, *bom* ou *mau para o crédito.* Ainda de acordo com essa tabela, pode-se perceber que o grupo de indivíduos *maus para o crédito* possui uma idade média inferior, em média buscam crédito mais alto e com empréstimos mais longos, quando comparado com o grupo de *bons para o crédito.*

Para as variáveis categorizadas, a Tabela 4 mostra a proporção da variável resposta para cada categoria, assim como suas quantidades brutas. Analisando a variável BENS, pode-se perceber que a categoria onde mais se acumula clientes classificados como *bons para o crédito* é a dos indivíduos que possuem imóvel quitado, e a categoria que proporcionalmente mais se acumulam *maus* é aquela dos indivíduos que não possuem bens. Nota-se na variável TP_EMPREGO que a proporção de bons e maus é bem similar em todas as categorias, dando sentido que, na Alemanha em 1994, o risco de conceder um empréstimo para um indivíduo desempregado, pouco se diferia do risco de conceder o crédito para um indivíduo que possuía um emprego altamente qualificado.

Tabela 3 - Análise descritiva das variáveis numéricas presentes na base inicial deste estudo.

VARIÁVEIS NUMÉRICAS	DEFINIÇÃO	DOMÍNIO	MAU (0)		BOM (1)	
			média	desv. pad.	média	desv. pad.
IDADE	Idade do proponente	De 19 até 75 anos	34.0	11.2	36.2	11.4
VLR_CREDITO	Valor do crédito	De 250 até 18424 DM	3938.1	3535.8	2985.5	2401.5
TMP_EMPRESTIMO	Tempo de duração do empréstimo	De 4 até 72 meses	24.9	13.3	19.2	11.1

Fonte: Elaborado pelos autores

Tabela 4 - Análise descritiva das variáveis categóricas presentes na base inicial deste estudo.

VARIÁVEIS CATEGORIZADAS	DEFINIÇÃO	DOMÍNIO	MAU (0)		BOM (1)	
			#	%	#	%
TMP_EMPREGO	Tempo no emprego	Desempregado	23	8%	39	6%
		menor que 1 ano	70	23%	102	15%
		de 1 a 3 anos	104	35%	235	34%
		de 4 a 6 anos	39	13%	135	19%
		maior ou igual a 7 anos	64	21%	189	27%
TMP_RESIDENCIA	Tempo de residência	menor ou igual a 1 ano	36	12%	94	13%
		maior que 1 ano até 2 anos	97	32%	211	30%
		maior que 2 anos até 3 anos	43	14%	106	15%
		maior que 3 anos	124	41%	289	41%
RENDA_COMPROMETIDA	Percentual de renda que o crédito comprometerá	maior ou igual a 10%	34	11%	102	15%
		maior que 10% até 20%	62	21%	169	24%
		maior que 20% até 30%	45	15%	112	16%
		maior que 30% até 40%	159	53%	317	45%
INVESTIMENTO	Valor do investimento em conta poupança ou título de capitalização	menor que 100 DM	32	11%	151	22%
		de 100 até 500 DM	217	72%	386	55%
		de 500 até 1000 DM	34	11%	69	10%
		maior ou igual a 1000 DM	11	4%	52	7%
		Desconhecido / Não possui	6	2%	42	6%
STATUS_CONTA	Situação da conta corrente se existente	menor que 0 DM	135	45%	139	20%
		de 0 até menor que 200 DM	105	35%	164	23%
		maior que 200 DM	14	5%	49	7%
		Sem conta corrente	46	15%	348	50%
NR_CREDITO	Número de créditos existentes no banco	Um	200	67%	433	62%
		Dois	92	31%	241	34%
		Três	6	2%	22	3%
		Quatro	2	1%	4	1%
DEPENDENTES	Número de dependentes	Um	254	85%	591	84%
		Dois	46	15%	109	16%

Fonte: Elaborado pelos autores

Tabela 4 – Continuação.

VARIÁVEIS CATEGORIZADAS	DEFINIÇÃO	DOMÍNIO	MAU (0)		BOM (1)	
			#	%	#	%
SEXO_ESTADO_CIVIL	Sexo e estado civil	Feminino	109	36%	201	29%
		Masculino: Solteiro	146	49%	402	57%
		Masculino: Casado / Viúvo	25	8%	67	10%
		Masculino: Divorciado / Separadc	20	7%	30	4%
ESTRANGEIRO	Se o proponente é estrangeiro	Não	4	1%	33	5%
		Sim	296	99%	667	95%
MOTIVO_CREDITO	Motivo da solicitação do crédito	Carro Novo	89	30%	145	21%
		Carro Usado	17	6%	86	12%
		Móveis / Equipamentos	58	19%	123	18%
		Rádio / Televisão	62	21%	218	31%
		Aparelhos domésticos	4	1%	8	1%
		Reparos	8	3%	14	2%
		Educação	22	7%	28	4%
		Treinamento / Curso	1	0%	8	1%
		Trabalho / Negócios	34	11%	63	9%
		Outros	5	2%	7	1%
TP_EMPREGO	Tipo de emprego	Desempregado	7	2%	15	2%
		Emprego não qualificado	56	19%	144	21%
		Emprego qualificado	186	62%	444	63%
		Emprego altamente qualificado	51	17%	97	14%
TP_MORADIA	Tipo de moradia	Alugada	70	23%	109	16%
		Própria	186	62%	527	75%
		Gratuita	44	15%	64	9%
BENS	Bens materiais que o proponente possui	Imóvel quitado	60	20%	222	32%
		Imóvel financiado	71	24%	161	23%
		Outros bens	102	34%	230	33%
		Nenhum bem / Desconhecido	67	22%	87	12%
GARANTIA	Garantia para o crédito	Nenhum	272	91%	635	91%
		Co-requerente	18	6%	23	3%
		Fiador	10	3%	42	6%
TELEFONE	Conta telefônica no nome do proponente	Não possui	187	62%	409	58%
		Possui	113	38%	291	42%
OUTROS_FINANCIAMENTOS	Instituição onde possui outros financiamentos	Em bancos	57	19%	82	12%
		Em lojas	19	6%	28	4%
		Não possui	224	75%	590	84%
HIST_CREDITO	Histórico de Crédito	Nunca tomou empréstimo	25	8%	15	2%
		Todos créditos pagos	28	9%	21	3%
		Empréstimo ativo sem atraso	169	56%	361	52%
		Atrasou alguma vez	28	9%	60	9%
		Empréstimo ativo com atraso	50	17%	243	35%

Fonte: Elaborado pelos autores.

Com base no estudo descritivo, duas variáveis foram desconsideradas no ajuste dos modelos, sendo elas HIST_CREDITO e NR_CREDITO. A primeira por não apresentar as proporções esperadas de acordo com a experiência adquirida na prática, como mostra a Figura 8, em que é possível observar que a categoria que apresenta a menor proporção de maus clientes é justamente “Empréstimo ativo com atraso”, representando apenas 17%, enquanto que os indivíduos da faixa “Empréstimo ativo sem atraso” apresentam um índice de 32% de maus pagadores. Ainda de acordo com essa variável, um indivíduo que nunca tomou empréstimo tem maior tendência a ser um mau pagador (63%) do que um indivíduo que já atrasou o pagamento de um

empréstimo por vez anterior (32%). Fatos que, aparentemente, não se refletem na prática.

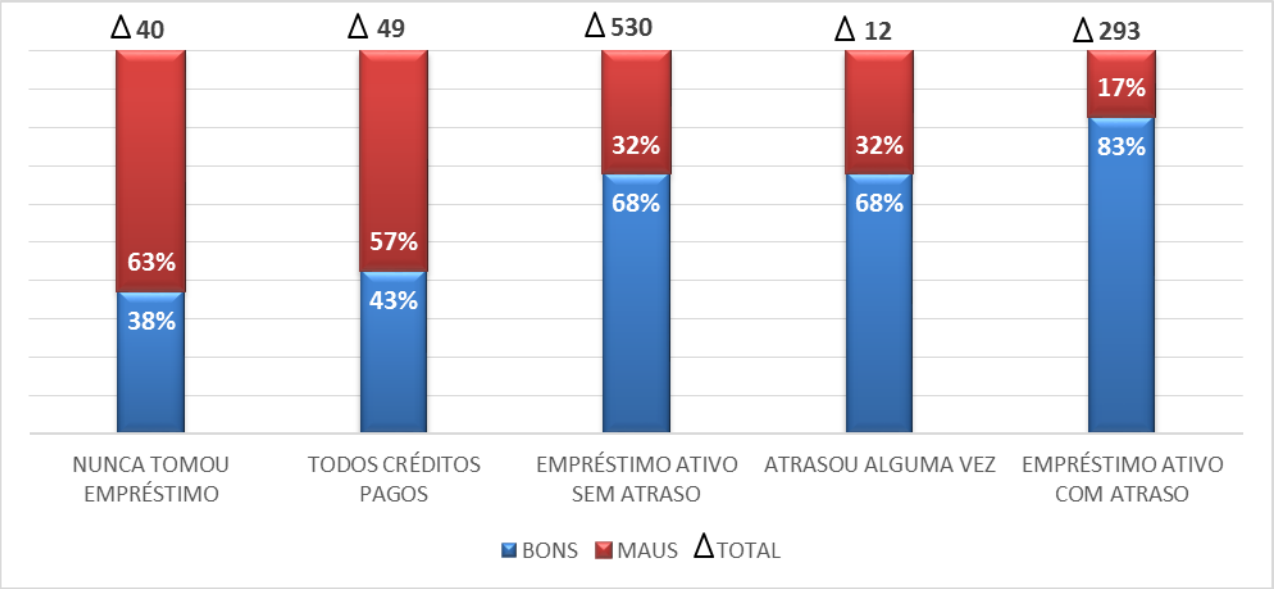

Figura 8 – Distribuição da variável referente ao histórico de crédito do cliente na base inicial.

Fonte: Elaborado pelos autores.

Quanto à variável explicativa NR_CREDITO, composta por: 1, 2, 3 ou 4 créditos tomados, esta apresentou restrições no ajuste do *bagging*, pois sua última categoria apresenta uma frequência muito baixa (apenas seis indivíduos), então, quando nenhum desses vinham a ser sorteados em uma das amostras *bootstrap*, a predição para essa categoria ficava indisponível. Uma solução possível seria uma nova categorização, agrupando as faixas 3 e 4. No entanto, desta forma seriam agrupadas as faixas extremas, de maior e menor proporção de *maus clientes (Figura 9)*, o que diminuiria o poder preditivo da variável, além de não ser aconselhável. Logo, optou-se por não levá-la para o ajuste dos modelos.

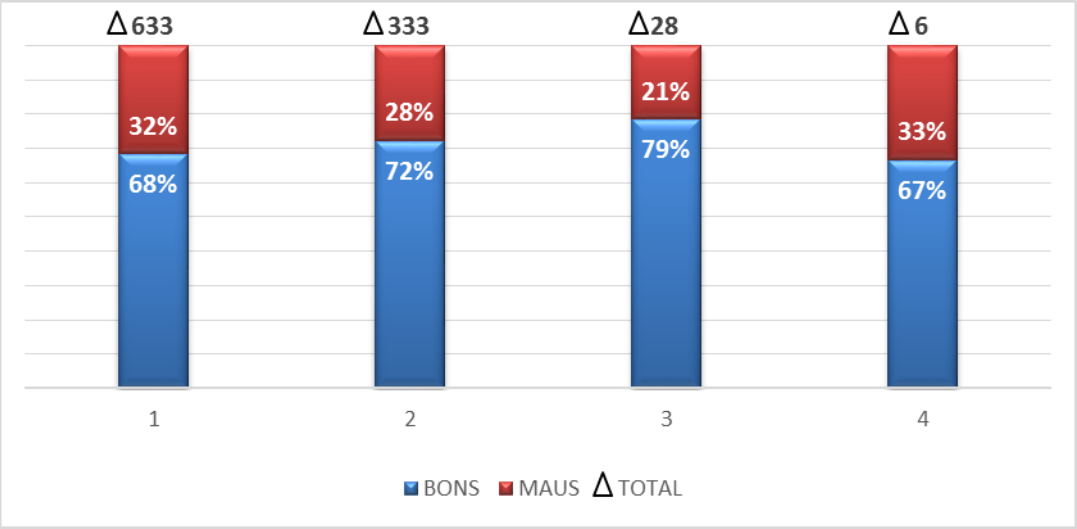

Figura 9 – Distribuição da variável referente ao número de créditos tomados pelo cliente na base inicial.

Fonte: Elaborado pelos autores.

5.2 RESULTADOS DAS ÁRVORES DE DECISÃO

No Capítulo 4 foram descritos os modelos e os métodos aplicados para a indução das árvores de decisão, assim como interpretou-se a ilustração das árvores. Então, com o auxílio do *software* R, seguem os resultados.

A árvore T1, que foi induzida utilizando todas as variáveis e com limitação para os *nós pais* de 7% e *nós filhos* de 3% do total da base, resultou em uma árvore que utiliza 5 variáveis distintas (STATUS_CONTA, TMP_EMPRESTIMO, MOTIVO_CREDITO, INVESTIMENTO e VLR_CREDITO) com 7 *nós finais*, como ilustrado na Figura 10.

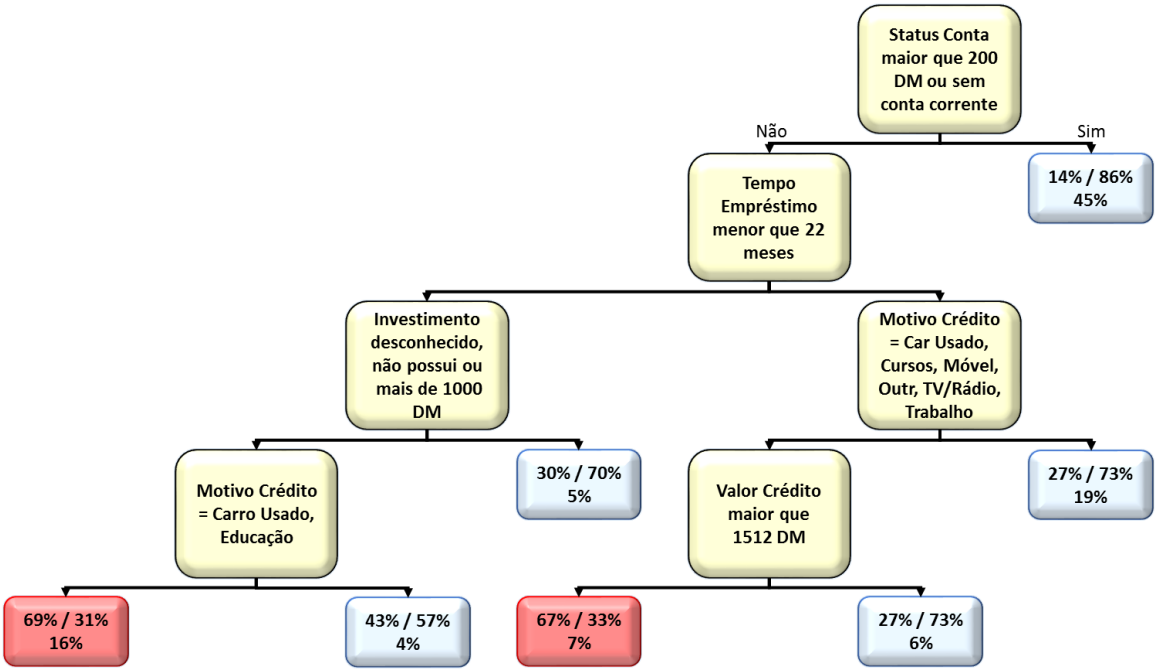

Figura 10 – Árvore de decisão T1.

Fonte: Elaborado pelos autores.

O processo de poda aplicado em T1, necessário para induzir a árvore T2, está representado na curva de custo-complexidade apresentada na Figura 11, que indica que a árvore T1 deve ser podada para permanecer com 6 *nós finais*. Portanto, a árvore T2 possui 6 *nós finais* e uma regra que utiliza as mesma 5 variáveis utilizadas na árvore T1, como ilustrado na Figura 12.

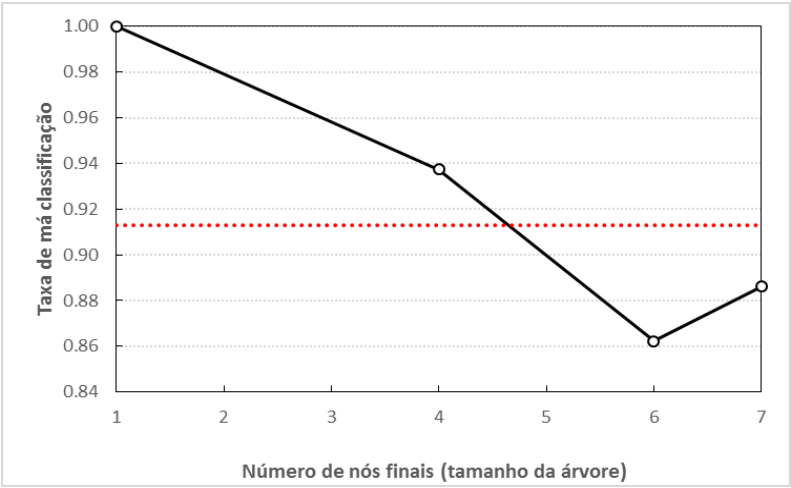

Figura 11 – Curva de custo-complexidade resultante do processo de poda aplicado em T1.

Fonte: Elaborado pelos autores.

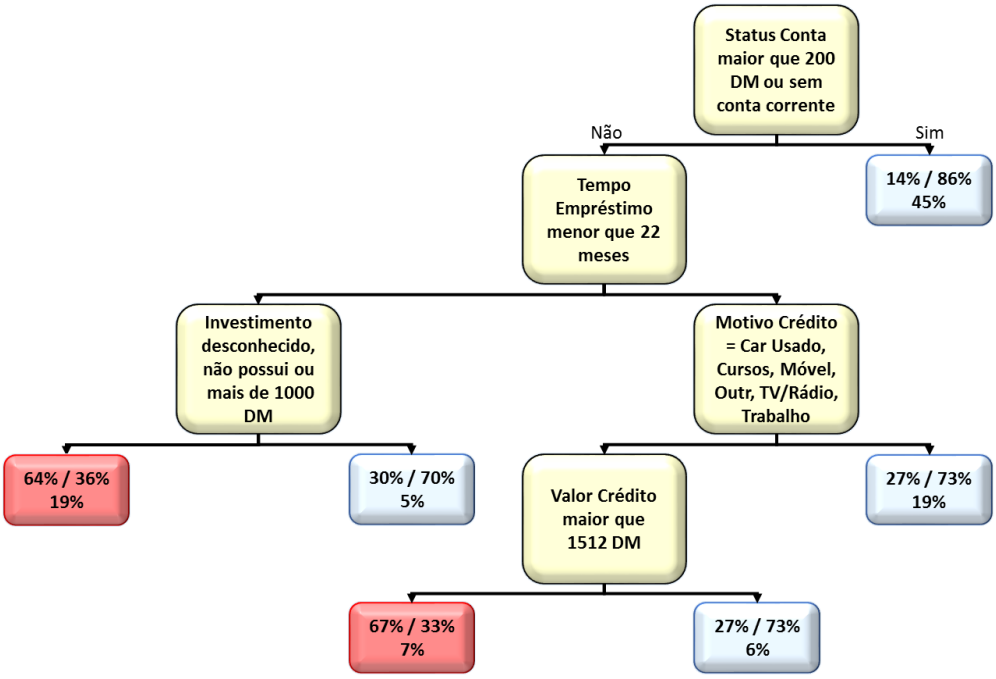

Figura 12 – Árvore de decisão T2.

Fonte: Elaborado pelos autores.

A árvore T3, em que foi aplicado o custo referente à má classificação de um indivíduo em sua construção, respeitando os critérios de parada de 7% para o *nó pai* e 3% para o *nó filho*, resultou em uma regra que utiliza 8 variáveis distintas (STATUS_CONTA, BENS, OUTROS_FINANCIAMENTOS, MOTIVO_CREDITO, IDADE, TMP_EMPRESTIMO, SEXO_ESTADO_CIVIL e INVESTIMENTO) com 10 *nós finais*, como ilustrado na Figura 13.

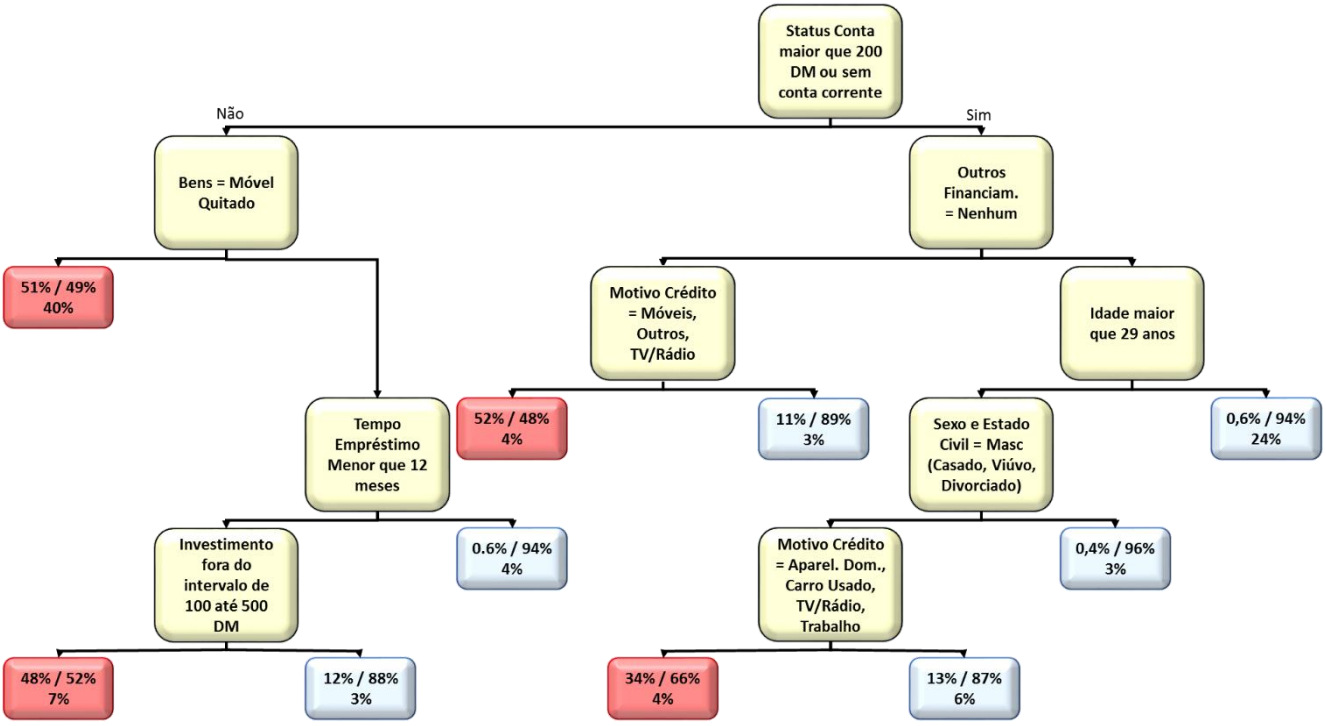

Figura 13 – Árvore de decisão T3.

Fonte: Elaborado pelos autores.

O processo de poda aplicado em T3, necessário para induzir a árvore T4, está representado na Figura 14 que indica que a árvore T3 deve ser podada para permanecer com 3 *nós finais*. Portanto, a árvore T4 possui 3 *nós finais* e uma regra que utiliza 2 variáveis distintas (STATUS_CONTA e OUTROS_FINANCIAMENTOS), como ilustrado na Figura 15.

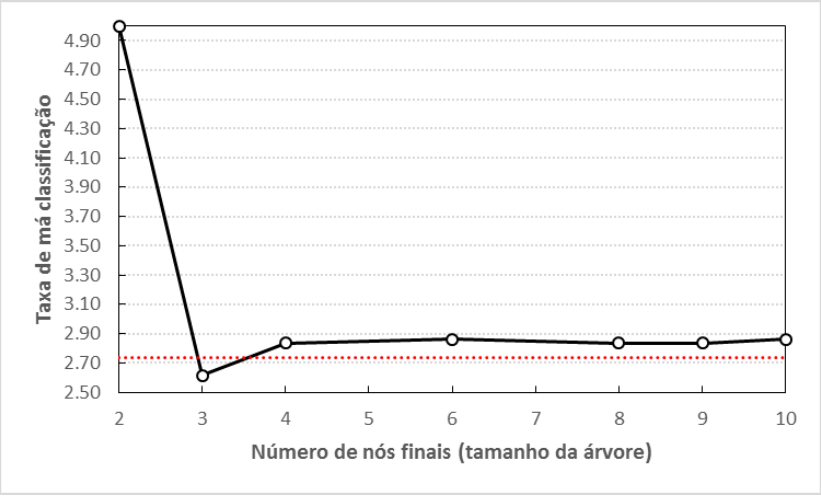

Figura 14 – Curva de custo-complexidade resultante do processo de poda aplicado em T3.

Fonte: Elaborado pelos autores.

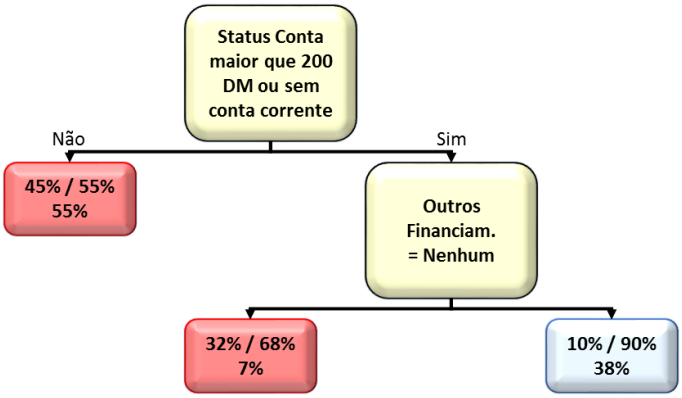

Figura 15 – Árvore de decisão T4.

Fonte: Elaborado pelos autores.

Para cada árvore de decisão foram calculados e representados na Tabela 5 o índice KS (calculado para as bases de ajuste e validação), a área abaixo da curva ROC e o número de *nós finais*. A Tabela 6 mostra alguns indicadores para o ponto de corte que gera o menor custo de má classificação na base de validação para cada árvore de decisão induzida.

Tabela 5 – Indicadores das árvores de decisão - Parte 1.

Modelo	KS Ajuste	KS Validação	AUC	N° de nós finais
T1	41%	38%	0,73	7
T2	41%	38%	0,72	6
T3	49%	42%	0,68	10
T4	37%	37%	0,70	3

Fonte: Elaborado pelos autores.

Tabela 6 – Indicadores das árvores de decisão - Parte 2.

Modelo	Ponto de corte	Sensibilidade	Especificidade	Poder	Custo de má classificação
T1	0,86	0,58	0,79	0,64	11700
T2	0,86	0,58	0,79	0,64	11700
T3	0,66	0,59	0,83	0,66	10500
T4	0,90	0,47	0,87	0,58	11300

Fonte: Elaborado pelos autores.

Por meio da Tabela 5 é possível identificar que a árvore T3 apresentou o melhor índice KS, tanto para a base de ajuste (49%) quanto para a de validação (42%), porém foi a que resultou em uma regra que utiliza 10 *nós finais*, sendo essa a maior

quantidade entre todas as árvores induzidas. Por outro lado, a árvore T3 apresentou menor AUC (0,68) e a árvore T1 apresentou maior AUC (0,73). Para entender por que a árvore T3 apresentou o menor AUC, sendo que é a que melhor discrimina os bons dos maus pagadores, são apresentadas na Figura 16 as Curvas ROC de cada árvore.

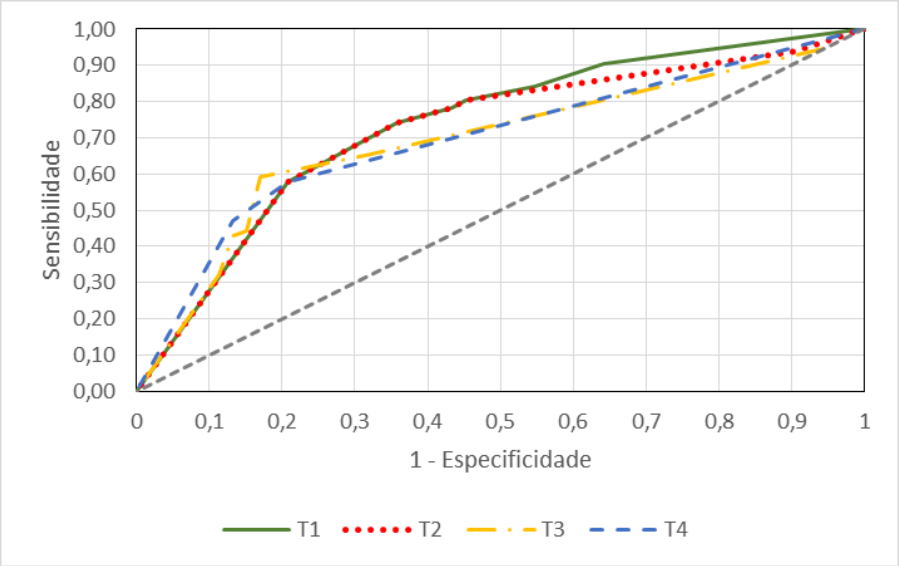

Figura 16 – Curva ROC para as árvores de decisão T1, T2, T3 e T4.

Fonte: Elaborado pelos autores.

Por meio da Figura 16 é possível identificar que a árvore T3 tem um bom desempenho no início da curva, o que explica ter apresentado o maior índice KS, mas à medida que vai aumentando a sensibilidade, vai perdendo poder preditivo. Porém, isso não afeta o desempenho geral desta árvore, pois como mostra a Tabela 6, o ponto de corte para esta árvore foi definido em 0.66, que está localizado na melhor região da sua Curva ROC encontrada, com sensibilidade de 0.59 e especificidade de 0.83, e que, consequentemente, é a árvore que gera o menor custo de má classificação (10.500 DM). Portanto, a árvore T3 é considerada como a mais adequada.

5.3 RESULTADOS DA REGRESSÃO LOGÍSTICA

A técnica de regressão logística (Seção 2.1), foi utilizada para o ajuste de 8 modelos distintos (apresentados na Tabela 7), conforme descrito no Capítulo 4. A

Tabela 8 mostra os resultados dos critérios de avaliação utilizados para comparar tais modelos.

Tabela 7 – Modelos ajustados para regressão logística.

Modelo	Seleção de variáveis
M1	Nenhum
M2	Stepwise
M3	P-valor < 0.20
M4	P-valor < 0.20 e Stepwise
M5	Curva ROC
M6	Curva ROC e Stepwise

Fonte: Elaborado pelos autores.

Tabela 8 – Indicadores dos modelos de regressão logística – Parte 1.

Modelo	KS Ajuste	KS Validação	AUC	AIC	N° de parâmetros
M1	55%	45%	0,79	821	41
M2	51%	45%	0,78	808	31
M3	51%	47%	0,80	816	35
M4	50%	43%	0,78	811	30
M5	50%	52%	0,81	830	32
M6	50%	52%	0,80	827	28

Fonte: Elaborado pelos autores.

A Tabela 9 mostra alguns indicadores para o ponto de corte que gera o menor custo de má classificação na base de validação para cada modelo de regressão logística.

Tabela 9 – Indicadores dos modelos de regressão logística – Parte 2.

Modelo	Ponto de corte	Sensibilidade	Especificidade	Poder	Custo de má classificação
M1	0,90	0,37	0,98	0,54	9700
M2	0,82	0,48	0,89	0,59	10600
M3	0,76	0,64	0,83	0,69	9800
M4	0,83	0,48	0,91	0,59	10200
M5	0,84	0,47	0,94	0,60	9300
M6	0,86	0,41	0,96	0,56	9600

Fonte: Elaborado pelos autores.

Pode-se destacar da Tabela 9 que o modelo M5 (seleção de variáveis pela curva ROC) foi o que apresentou o menor custo de má classificação. Logo, é o que apresenta maior lucro esperado. Esse ainda apresentou maior AUC (igual a 0,81), e avaliando o KS para a base de validação, apresenta um valor de 52%, similar ao modelo M6, sendo os maiores valores encontrados. Entretanto, esse modelo apresenta os menores valores para o KS de ajuste e os maiores valores para o AIC, o que indica que, apesar de ser um excelente modelo de predição, não foi o que melhor se ajustou aos dados originais.

Em busca de um modelo ainda melhor, interações duplas evidenciadas nos resultados das árvores de decisão T1 e T3, foram incluídas no modelo de melhor predição, o M5, chegando a dois novos modelos (apresentados na Tabela 10), e seus resultados estão apresentados na Tabela 11.

Tabela 10 – Modelos ajustados para regressão logística utilizando as interações duplas evidenciadas nas árvores de decisão.

Modelo	Seleção de variáveis	Interações	Interações entre as variáveis
M7	Curva ROC	T1	TMP_EMPRESTIMO*INVESTIMENTO
M8	Curva ROC	T3	STATUS_CONTA*OUTROS_FINANC. STATUS_CONTA*BENS

Fonte: Elaborado pelos autores.

Tabela 11 – Indicadores dos modelos de regressão logística acrescidos de interações – Parte 1.

Modelo	KS Ajuste	KS Validação	AUC	AIC	N° de parâmetros
M7	51%	54%	0,82	830	36
M8	54%	51%	0,80	832	47

Fonte: Elaborado pelos autores.

A Tabela 12 mostra alguns indicadores para o ponto de corte que gera o menor custo de má classificação para cada modelo de regressão logística acrescidos das interações.

Tabela 12 – Indicadores dos modelos de regressão logística acrescidos de interações – Parte 2.

Modelo	Ponto de corte	Sensibilidade	Especificidade	Poder	Custo de má classificação
M7	0,72	0,67	0,83	0,71	9400
M8	0,80	0,56	0,94	0,67	7900

Fonte: Elaborado pelos autores.

Os dois modelos acrescidos das interações apresentaram ganhos. O M7, ao qual foram acrescentadas interações encontradas na árvore que não envolvia custo para sua construção, apresentou ganhos de 2% no KS de validação e 0,1 no AUC. Já o modelo M8, ao qual foram incorporadas interações encontradas na árvore construída utilizando custos, apresentou um ganho no lucro esperado, baixando o custo de má classificação de 9.300 DM para 7.900 DM. Esses modelos, apesar de apresentarem predições mais eficientes, necessitam estimar um número mais elevado de parâmetros, o que resultou em altos valores para o AIC.

Para visualizar o aumento do AUC quando se acrescentou interações no modelo M5, está apresentada a Figura 17 com a curva ROC dos modelos M5 e M7.

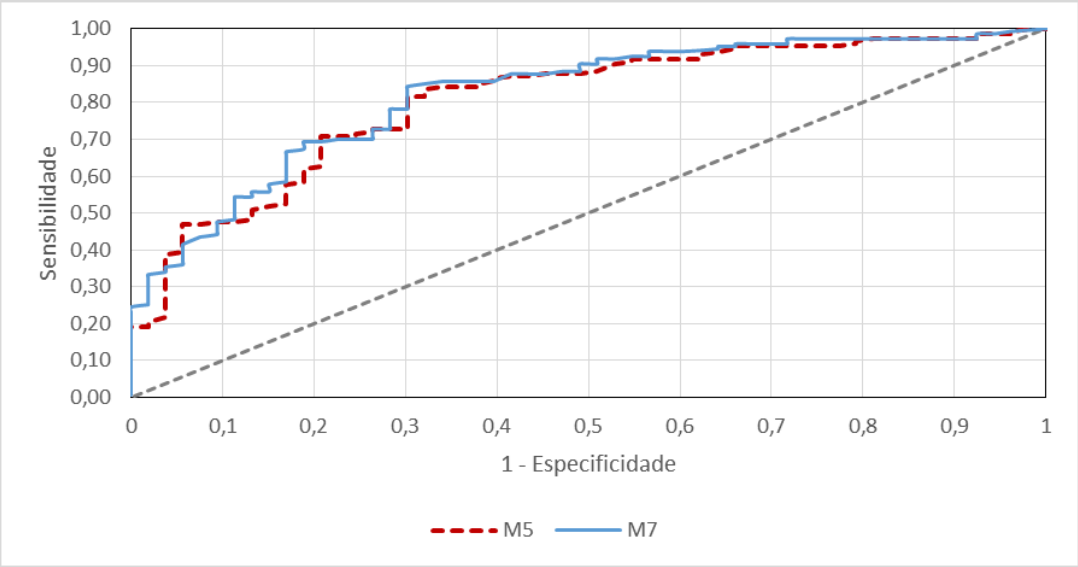

Figura 17 - Curva ROC dos modelos M5 e M7.

Fonte: Elaborada pelos autores.

A fim de identificar o motivo pelo qual o modelo M8 apresenta um melhor lucro mesmo sendo inferior nos critérios de curva ROC e KS, está apresentado na Figura 18 o cálculo do KS para ambos os modelos.

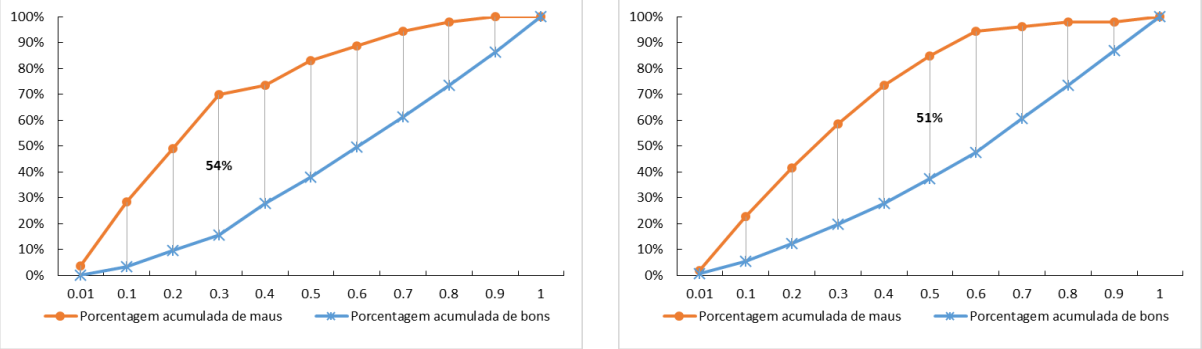

Figura 18 – Representação gráfica do índice KS dos modelos M7 (esquerdo) e M8 (direito).

Fonte: Elaborada pelos autores.

Ressaltando que o custo de má classificação de um indivíduo considerado *mau para o crédito* é 5 vezes a de um indivíduo *bom para o crédito (500 DM contra 100 DM), p*ercebe-se, na Figura 18, que para o modelo M8 o índice KS aparece quando o acumulado de *maus pagadores* está em um nível bem elevado (aproximadamente 90%), o que possibilita aprovar uma boa quantidade de *bons pagadores,* aprovando uma quantidade baixíssima de *maus*. Já para o modelo M7 o índice KS acontece em um momento em que o acúmulo de *maus pagadores* ainda não está tão alto (70%), logo, se esse fosse utilizado como ponto de corte, apesar de se aprovar quase 90% dos *bons pagadores,* o custo de má classificação seria alto, pois se estaria aprovando errado aproximadamente 30% dos *maus*.

5.4 RESULTADOS DA PONDERAÇÃO DE MODELOS

5.4.1 Bagging na Árvore de Decisão

Para a aplicação do *bagging* fez-se necessário definir uma quantidade suficiente que garanta estabilidade nos resultados. Para tal, o *bagging* foi aplicado utilizando diferentes valores para *B*, variando de 10 até 500. Os valores encontrados foram representados em um gráfico para que se pudesse averiguar em qual *B* o erro

de classificação se torna estável. Definiu-se então que, para os dados deste trabalho, o erro se estabiliza a partir de *B* igual a 400, como apresentado na Figura 19.

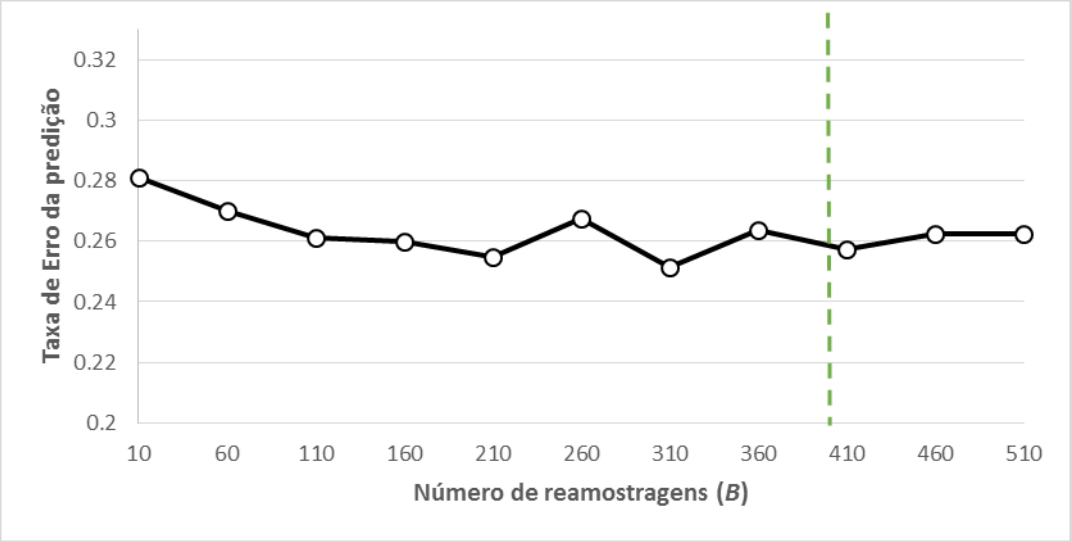

Figura 19 - Estudo para definir a quantidade de reamostragens para *Bagging.*

Fonte: Elaborado pelos autores.

Os modelos de árvore de decisão com aplicação do método *bagging* (apresentados na Tabela 13), descritos no Capítulo 4, foram gerados e seus resultados estão apresentados na Tabela 14.

Tabela 13 – Modelos de árvore de decisão com aplicação do Bagging.

Modelo	Critério de partição	Classificação final
BT1	Índice Gini	Predição média
BT2	Índice Gini	Proporção de votos
BT3	Custo de má classificação	Predição média
BT4	Custo de má classificação	Proporção de votos

Fonte: Elaborado pelos autores.

Tabela 14 – Indicadores do *bagging* aplicados na árvore de decisão – Parte 1.

Modelo	KS Ajuste	KS Validação	AUC
BT1	63%	41%	0,78
BT2	64%	42%	0,78
BT3	62%	49%	0,79
BT4	53%	45%	0,78

Fonte: Elaborada pelos autores.

Definindo o ponto de corte que gera o menor custo de má classificação para a base de validação, chegou-se aos resultados apresentados na Tabela 15.

Tabela 15 – Indicadores do bagging aplicados na árvore de decisão – Parte 2.

Modelo	Ponto de corte	Sensibilidade	Especificidade	Poder	Custo de má classificação
BT1	0,85	0,46	0,91	0,58	10400
BT2	0,96	0,48	0,91	0,60	10100
BT3	0,70	0,63	0,85	0,69	9400
BT4	0,91	0,48	0,92	0,60	9700

Fonte: Elaborada pelos autores.

Com os resultados, pode-se perceber que nenhum dos diferentes métodos de predição utilizados, com base na média de predição de cada árvore gerada no *bagging* e definido por proporção de votos (ambos descritos na Seção 2.3.1), se mostrou melhor que o outro, dado que, para as árvores construídas sem adição de custo de má classificação (BT1 e BT2), a classificação por proporção de votos apresentou melhores resultados (BT2). Já para as árvores que utilizaram custo de má classificação em sua construção (BT3 e BT4), a predição por meio da média foi superior (BT3).

O melhor modelo encontrado aplicando o *bagging* na árvore de decisão foi o BT3, por ter apresentado menor custo (9.400 DM), bem como o maior índice de KS para a amostra de validação (43%) e AUC (0,79). Ainda é possível destacar que o modelo BT3 apresentou uma sensibilidade bastante superior aos demais modelos (0,63), o que indica um número superior de indivíduos com crédito aprovado.

5.4.2 Bagging na Regressão Logística

Os modelos de regressão logística com aplicação do método *bagging* (apresentados na Tabela 16), descritos no Capítulo 4, foram gerados e seus resultados estão apresentados na Tabela 17.

Tabela 16 – Modelos de regressão logística com aplicação do método *Bagging*.

Modelo	Seleção de variáveis	Classificação final
BM1	Nenhum	Predição média
BM2	Nenhum	Proporção de votos
BM3	Curva ROC	Predição média
BM4	Curva ROC	Proporção de votos

Fonte: Elaborado pelos autores.

Tabela 17 – Indicadores do *bagging* aplicados na árvore de decisão – Parte 1.

Modelo	KS Ajuste	KS Validação	AUC
BM1	59%	45%	0,78
BM2	59%	43%	0,77
BM3	55%	52%	0,82
BM4	55%	53%	0,80

Fonte: Elaborada pelos autores.

Definindo o ponto de corte que gera o menor custo de má classificação para a base de validação, chegou-se à Tabela 18.

Tabela 18 – Indicadores do bagging aplicados na árvore de decisão – Parte 2.

Modelo	Ponto de corte	Sensibilidade	Especificidade	Poder	Custo de má classificação
BM1	0,91	0,38	0,98	0,54	9600
BM2	0,95	0,52	0,87	0,62	10500
BM3	0,86	0,50	0,94	0,62	8900
BM4	0,95	0,54	0,89	0,63	9800

Fonte: Elaborada pelos autores.

Com os resultados, pode-se perceber que o método de predição utilizado nos modelos BM1 e BM3, que são baseados na média de predição de cada modelo gerado

com a aplicação do *bagging*, apresentou melhor resultado do que aquele em que é definido por meio de proporção de votos (BM2 e BM4).

O melhor modelo encontrado aplicando o *bagging* na regressão logística foi o BM3, apresentando menor custo (8.900 DM), o maior AUC (0.82), bem como o segundo maior índice KS para a amostra de validação (52%) contra 53% do maior índice KS (BM4).

5.4.3 Random Forest

Foram induzidas diversas árvores de decisão, por meio do método de *Random Forest*, variando o valor de *B* entre 1 e 500, que resultaram na Figura 20. Por meio desta figura, é possível perceber, para os dados deste trabalho, que o erro se estabiliza a partir de *B* igual a 300, logo, esse foi o número utilizado para todas as *random forest* induzidas.

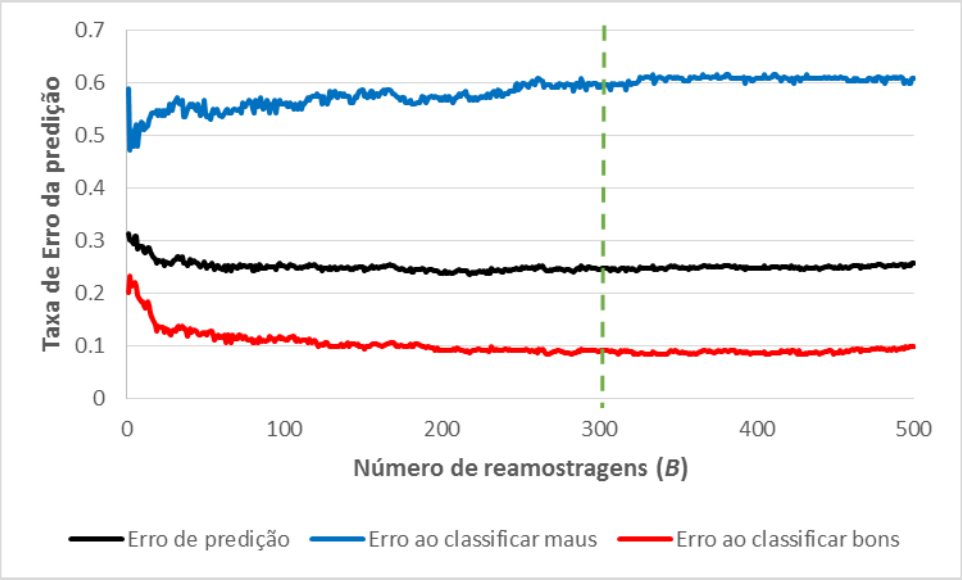

Figura 20 - Estudo para definir a quantidade de reamostragens para Random Forest.

Fonte: Elaborado pelos autores.

Os modelos de *random forest* gerados (apresentados na Tabela 19), descritos no Capítulo 4, têm seus resultados apresentados na Tabela 20.

Tabela 19 – Modelos ajustados para *Random Forest*.

Modelo	Critério de Partição	Classificação final	Sorteio
RF1	Índice Gini	Proporção de votos	2
RF2	Índice Gini	Proporção de votos	5
RF3	Índice Gini	Proporção de votos	10
RF4	Índice Gini	Proporção de votos	15

Fonte: Elaborado pelos autores.

Tabela 20 – Indicadores do *random forest* – Parte 1.

Modelo	KS Ajuste	KS Validação	AUC	Numero de variáveis disponiveis em cada nó
RF1	45%	46%	0,79	2
RF2	47%	43%	0,79	5
FR3	46%	45%	0,78	10
RF4	46%	41%	0,78	15

Fonte: Elaborada pelos autores.

Definindo o ponto de corte que gera o menor custo de má classificação para a base de validação, chega-se à Tabela 21.

Tabela 21 – Indicadores do *random forest* – Parte 2.

Modelo	Ponto de corte	Sensibilidade	Especificidade	Poder	Custo de má classificação
RF1	0,81	0,37	0,98	0,54	9700
RF2	0,82	0,40	0,96	0,55	9800
FR3	0,84	0,39	0,94	0,54	10500
RF4	0,86	0,36	0,96	0,52	10400

Fonte: Elaborada pelos autores.

Percebe-se, com os resultados, que a diferença nos indicadores dos modelos gerados por diferentes números de variáveis disponíveis para dividir cada nó se mostrou bem sutil. Avaliando o índice KS para a amostra de validação e AUC de cada modelo, vê-se uma pequena diminuição desses índices conforme aumenta-se o número de variáveis disponíveis para dividir cada nó, assim como, um aumento do custo de má classificação. O modelo RF1 foi considerado o mais adequado pois apresentou o menor custo. Portanto, na Seção 5.5, este modelo será comparado aos

melhores modelos encontrados nas outras técnicas aplicadas neste trabalho. Essa escolha também foi motivada pelo fato do modelo RF1 ser o que utilizou o menor número de variáveis disponíveis para dividir cada nó, sendo assim o que mais se diferencia do método *bagging* aplicado a árvores de decisão.

5.5 COMPARAÇÃO DOS MODELOS

Definido o melhor modelo de cada uma das técnicas aplicadas, chegou-se a 6 modelos finais, sendo eles:

- T3 – Árvore de decisão construída com a utilização de custos;
- M5 – Modelo de regressão logística com variáveis incluídas pelo critério da curva ROC;
- M8 – Modelo M5 acrescido de interações entre variáveis explicativas evidenciadas na Árvore T3.
- BT3 – *Bagging* aplicado a árvores de decisão utilizando custos com classificação final por meio de média.
- BM3 – *Bagging* aplicado à regressão logística com variáveis explicativas selecionadas por meio do critério da curva ROC e custos, considerando a classificação final por meio de média.
- RF1 – *Random Forest* com 2 variáveis explicativas candidatas a cada partição do nó e com classificação final por meio de proporção de votos.

Tais modelos estão cotejados a seguir:

Tabela 22 – Indicadores dos melhores modelos encontrados – Parte 1.

Modelo	KS Validação	AUC
T3	42%	0,68
M5	52%	0,81
M8	51%	0,80
BT3	49%	0,79
BM3	52%	0,82
RF1	46%	0,79

Fonte: Elaborada pelos autores.

Tabela 23 – Indicadores dos melhores modelos encontrados – Parte 2.

Modelo	Ponto de corte	Sensibilidade	Especificidade	Poder	Custo de má classificação
T3	0,66	0,59	0,83	0,66	10500
M5	0,84	0,47	0,94	0,60	9300
M8	0,80	0,56	0,94	0,67	7900
BT3	0,70	0,63	0,85	0,69	9400
BM3	0,86	0,50	0,94	0,62	8900
RF1	0,81	0,37	0,98	0,54	9700

Fonte: Elaborada pelos autores.

Percebe-se um emparelhamento entre as técnicas. No entanto, pode-se ressaltar que o *bagging* apresentou melhores resultados que a aplicação da técnica de forma individual, principalmente para árvore de decisão (comparando T3 com BT3), elevando um KS de validação de 42% para 49% e a AUC de 0,68 para 0,79 (representado na Figura 21), além de diminuir o custo de 10.500 DM para 9.400 DM.

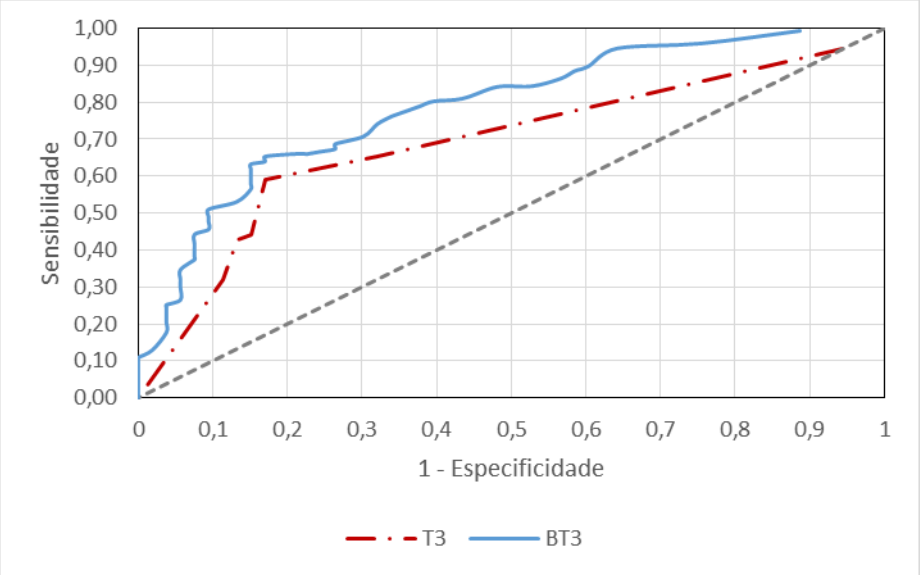

Figura 21 – Curva ROC para os modelos T3 e BT3.

Fonte: Elaborado pelos autores.

A melhora gerada pelo *bagging* na técnica de regressão logística (M5 contra BM3) foi mais modesta que para árvores de decisão, elevando AUC de 0,81 para 0,82 (apresentado na Figura 22), e passando o custo de 9.300 DM para 8.900 DM. Um indicativo é que a técnica de regressão logística possui uma estabilidade maior que árvores de decisão.

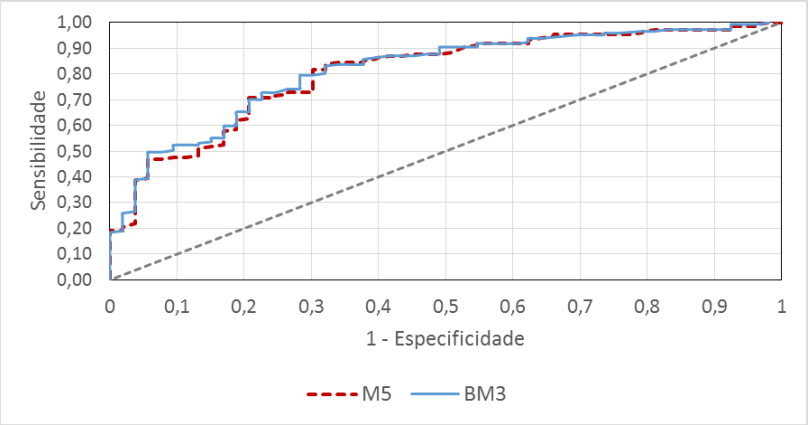

Figura 22 – Curva ROC para os modelos M3 e BM3.

Fonte: Elaborado pelos autores.

Apesar do modelo BM3 apresentar resultados melhores que M5 (seu similar em que não foi utilizado o *bagging*), ele não foi capaz de superar o modelo M8, ao qual foi incluída interação, que apresentou um custo de apenas 7.900 DM, mil a menos que o BM3, além de apresentar valores similares aos melhores modelos para todos os indicadores testados. Outra vantagem do modelo M8 é uma maior facilidade na interpretação, pois mesmo possuindo 47 parâmetros, tem uma interpretação mais simples em relação aos do *bagging*. Portanto, este foi selecionado como o melhor modelo encontrado.

6. CONCLUSÃO

O objetivo deste trabalho foi comparar o método estatístico de regressão logística e árvore de decisão na classificação de clientes de acordo com o risco na concessão de crédito. Ademais, foram utilizadas técnicas de ponderação de modelos, *bagging* e *random forest*, a fim de comparar a capacidade preditiva resultante com as dos métodos individuais.

De acordo com os resultados obtidos, válidos para os dados deste estudo, foi visto que os modelos de árvore de decisão são mais simples de se ajustar, dado que não é necessário uma pré-seleção de variáveis, uma vez que o algoritmo busca as variáveis mais significativas disponíveis. Outra vantagem vista para este modelo é sua fácil apresentação e interpretação dos resultados, podendo ser lido por pessoas sem um avançado conhecimento técnico. Entretanto, esse foi o método que apresentou os piores resultados quanto à capacidade preditiva. A aplicação do custo de má classificação trouxe resultados positivos para este modelo dado que, deste modo, um nó só será dividido caso um dos seus *nós filhos* possa ser classificado como *bom*, ou seja, concentrando uma proporção de *bons* e *maus* pagadores que gere lucro esperado.

Para ajustar os modelos de regressão logística uma maior complexidade no processo de construção é exigida, uma vez que é recomendado realizar uma seleção prévia das variáveis, já que neste trabalho os métodos de seleção de variáveis mostraram-se eficientes quanto ao ganho na predição. Entretanto, é um método que apresentou um alto nível de poder discriminativo, mostrando-se superior aos modelos de árvore de decisão, o que gerou um maior ganho financeiro. Diferentemente da árvore, este é um método que não encontra interações entre variáveis automaticamente, pois é necessário incluí-las no modelo para serem avaliadas quanto a sua significância. Devido ao alto número de variáveis explicativas, avaliar todas as possíveis interações entre elas seria inviável, então, a forma encontrada para simplificar a busca de interações significativas foi adicionar apenas as interações evidenciadas pelas árvores de decisão ao modelo de regressão logística e, assim, testar sua significância. Deste modo, identificando as interações significativas, que foram evidenciadas na árvore induzida com a utilização de custo, e as incluindo no

modelo que utilizou a seleção de variáveis pelo critério da curva ROC, chegou-se ao modelo que gerou um maior lucro esperado.

O método de ponderação de modelos *bagging* aplicado à árvore de decisão gerou resultados melhores que os obtidos com as árvores individuais, alcançando resultados similares aos obtidos pelos modelos de regressão logística, exceto para o KS da base de ajuste, chegando aos melhores patamares. Aplicando o *bagging* aos modelos de regressão logística, pouca melhora foi vista, o que indica que os modelos de regressão logística são mais estáveis e robustos que os de árvores de decisão. Por fim, o outro método de ponderação de modelos, *random forest*, apresentou ganho comparado a árvore de decisão individual. Ademais, o menor custo de má classificação foi gerado pelo modelo que utilizou apenas duas variáveis disponíveis para divisão de cada nó. No entanto, os resultados do *random forest* foram similares ao do *bagging* aplicado à árvore de decisão.

Com este estudo foi possível concluir que todos os métodos são aplicáveis à concessão de crédito. No entanto, o modelo de regressão logística foi o que se mostrou mais adequado, por atingir um alto nível de desempenho preditivo. Quando esses modelos são aplicados de forma adequada, impactam diretamente na operação e lucratividade da empresa, gerando vantagem competitiva.

REFERÊNCIAS

ALVES, M. C. **Estratégias para o desenvolvimento de modelos de Credit Score com inferência de rejeitados**, 2008. São Paulo, Brasil: Dissertação de Mestrado. Instituto de Matemática e Estatística – Universidade de São Paulo.

BACHE, K.; LINCHMAN, M. **UCI Machine Learning Repository.** Irvine, CA: University of California, School of Information and Computer Science, 2013. Disponível em < http://archive.ics.uci.edu/ml/datasets/Statlog+(German+Credit+Data) > Acesso em: 20 mar. 2013.

BOZDONGAN, H. Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. **Psychometrika**. v.52, n.3, p.345-370, 1987.

BREIMAN, L.; FRIEDMAN, J. H.; OLSHEN, R. A.; STONE, C. J. **Classification and Regression Trees**; California: Wadsworth International, 1984. 358p.

BREIMAN, L. Bagging predictors. **Machine Learning**, v. 24, n. 2, p. 123-140, 1996.

BREIMAN, L. Random Forest. **Machine Learning**, v. 45, n. 1, p. 5 - 32, 2001.

BUCKLAND, S. T.; BURNHAM, K. P.; AUGUSTIN, N. H. Model Selection: An Integral Part of Inference, **Biometrics**. v.53 p.603-618, 1997.

BÜHLMANN, P.; YU, B. Analyzing Bagging. **The Annals of Statistics**, v.30, 2002. 927-961.

CNC – Confederação Nacional do Comércio. **Pesquisa de Endividamento e Inadimplência do Consumidor**, 2013. Disponível em: < http://www.ibegi.org.br/indicadores.php> Acesso em: 17 jul. 2013.

EFRON, B. Bootstrap Methods: Another Look at the Jackknife. **Annals of Statistics**. v. 7, p. 1-26, 1979.

EFRON B.; TIBSHIRANI R. J. **An Introduction to the Bootstrap**. New York: Chapman and Hall, 1993. 642p.

FONSECA, J. **Indução de árvores de decisão**. Tese de Mestrado, Lisboa, 1994.

FORSTER, M. R. Key Concepts in Model Selection: Performance and Generalizability. **Journal of Mathematical Psychology,** v.44, p. 205-231, 2000.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statistical Learning.** New York: Springer, 2001. 731p.

HOSMER, D. W.; LEMESHOW, S. **Applied logistic regression**, 2 ed. New York: John Wiley & Sons, 1989**.**

KASS, G. An exploratory technique for investigating large quantities of categorical data. **Applied Statistics**. v. 29, n. 2, p. 119-127, 1980.

KOUKOUVINOS, C.; PARPOULA, C. Variable Selection and Computation of the Prior Probability of a Model via ROC Curves Methodology. Journal of Data Science, v.10, p. 653-672, 2012.

LIAW A.; WIENER M,. Classification and Regression by randomForest. **R News**, v.2, n.3, p.18-22, 2002.

LOUZADA NETO, F.; DINIZ, C. A. R. Modelagem Estatística para Risco de Crédito, **Resumo do 20º SINAPE**, Paraíba, 2012, p. 178.

MCCULLAGH, P.; NELDER, J, A. **Generalized Linear Models**, 2.ed. London: Chapman and Hall, 1989. 532p.

MICROSOFT. Microsoft Excel. Redmond, Washington: Microsoft, 2013. Computer Software.

MILLER, A. J. **Subset selection in regression**. Washtington DC: Chapman and Hall, 1990. 229p.

OLIVEIRA, José G.C.; ANDRADE, Fábio W.M. Comparação entre medidas de performance de modelos de credit scoring. **Tecnologia de Crédito**, n. 33, p. 35-47, 2002.

OPITZ, D; MACLIN, R. Popular ensemble methods: An empirical study, **Journal of Artificial Intelligence Research** 11: 169-198, 1999.

QUINLAN, J. R. Introduction of decision trees. **Machine Learning**, vol. 1, n.1, p. 81-106, 1986.

QUINLAN, J. R. **C4.5: Programs for machine learning**. San Mateo: Morgan Kaufmann Publishers, 1993. 302p.

R CORE TEAM. **R: A language and enviroment for statistical computing**. Vienna, Austria, 2012. Disponível em: http://www.r-project.org, Acesso em: 20 mar. 2013.

ROSA, P. T. M. **Modelos de Credit Scoring: Regressão Logística, CHAID e REAL.** São Paulo: 2000. Dissertação (Mestrado em Estatística). Departamento de Estatística. Universidade de São Paulo. IME/USP.

SANTOS, J. O. dos. **Uma Contribuição ao Estudo de Fatores Sistemáticos Influenciadores da Inadimplência de Pessoas Físicas em Empréstimos Bancários**, 2000. Tese (Doutorado), EAESP-FGV, São Paulo, 2000.

SAS INSTITUTE INC. **SAS language**: Reference, version 6. 4.ed. Cary: SAS Institute, 2012. v.2.

SIEGEL, S. **Estatística Não-paramétrica Para as Ciências do Comportamento**. São Paulo: McGraw-Hill, 1975. 350p.

SING T.; SANDER O.; BEERENWINKEL N.; LENGAUER T. ROCR: visualizing Classifier performance in R. **Bioinformatics**. v. 21, n. 20, p. 7881, 2005.

TACONELI, C. A. **Árvores de classificação multivariadas fundamentadas em coeficientes de dissimilaridade e entropia**, 2008. Tese (Doutorado), Escola Superior de Agricultura “Luiz de Queiroz” – USP, Piracicaba, 2008.

THERNEAU T.; ATKINSON B.; RIPLEY B. **rpart: Recursive Partitioning**. R package version 4.1-1, 2013. Disponível em < http://CRAN.R-project.org/package=rpart > Acesso em: 25 mar. 2013.

VICTORA C. G.; HUTTLY S. R.; FUCHS S. C.; OLINTO M. T. A. The role of conceptual frameworks in epidemiological analysis: a hierarchical approach**. Int J Epidemiol**, v.26, p.224-227, 1997.

WITTEN, I. H. & FRANK, E. **Data Mining: Practical Machine Learning Tools and Techniques**. Morgan Kaufmann, 2005. 525p.

APÊNDICES

Exemplo dos Códigos R, de maneira simplificada, utilizados neste trabalho para ajustar os modelos.

```
#-------------------------------------------------------------------------------------------#
# CONSTRUINDO AS ÁRVORES - CONSIDERANDO CUSTOS #
#-------------------------------------------------------------------------------------------#
require(rpart)
#---------------- loss matrix ----------------------------------#
lmat <- matrix(c(0,1,5,0), nrow=2, ncol=2, byrow=F)
#---------------------------------------------------------------#
#-------- t3: Arvore Com Custo Reduzida pelos nss pais e filhos --------#
#----- Critério de parada pré estabelecido -----#
pc_leaf   <- 0.03; pc_father <- 0.07
#-----------------------------------------------------#
t3 <- rpart(IN_BOM~., parms=(list(loss=lmat)), control = rpart.control(minsplit=dim(da80)[1]*pc_father,minbucket=dim(da80)[1]*pc_leaf),  data=da80)
print(t3); printcp(t3)
#-------------------------------------------------------------------------------------------#
# AJUSTANDO REGRESSÃO LOGÍSTICA #
#-------------------------------------------------------------------------------------------#
m1 <- glm(IN_BOM~., family=binomial(link="logit"), data=da80)
#-------------------------------------------------------------------------------------------------------------#
# BAGGING PARA ÁRVORES DE DECISÃO
#-------------------------------------------------------------------------------------------------------------#
#----- Critério de parada pré estabelecido -----#
pc_leaf   <- 0.03
pc_father <- 0.07
#-----------------------------------------------------#
da_list<- vector(mode="list", length=n); mo_list<- vector(mode="list", length=n)

for(j in 1:400){
  #---------------- Gera a amostra bootstrap ---------------------#
  var_boot <- sample(seq(1:dim(da80)[1]), size=dim(da80)[1], replace=TRUE)
  da_boot <- da80[var_boot,]
```

```
  #----------------- Constrói a árvore a partir da amostra gerada ----------------------#
  t01 <- rpart(IN_BOM~.,
               control = rpart.control(minsplit=dim(da80)[1]*pc_father,minbucket=dim(da80)[1]*pc_leaf),
               data=da_boot)

  #------------------ Guarda Bases e Modelos --------------------#
  da_list[[j]] <- da_boot
  mo_list[[j]] <- t01
}
#-------------------------------------------------------------------------------------------------------------#
# BAGGING PARA REGRESSÃO LOGÍSTICA
#-------------------------------------------------------------------------------------------------------------#
 da_list<- vector(mode="list", length=n); mo_list<- vector(mode="list", length=n)
#-------------------------------------------------------------------------------------------------------------#
for(j in 1:400){
  var_boot <- sample(seq(1:dim(da80)[1]), size=dim(da80)[1], replace=TRUE)
  #----------------- Gera a amostra bootstrap----------------------#
  da_boot <- da80[var_boot,]
  #----------------- Constrói a árvore a partir da amostra----------------------#
  m01 <- glm(IN_BOM~., family=binomial(link="logit"), data=da_boot)
  #------------------Guarda Bases e Modelos --------------------------------#
  da_list[[j]] <- da_boot
  mo_list[[j]] <- m01
}
#-------------------------------------------------------------------------------------------------------------#
# RANDOM FOREST
#-------------------------------------------------------------------------------------------------------------#
require(randomForest)
 rf <- randomForest(IN_BOM~., ntree=300, mtry=2, importance=TRUE, data=da80)
```