Insper Instituto de Ensino e Pesquisa
Programa de Mestrado Profissional em Administração

DANIEL ABREU VASCONCELLOS DE PAULA

MODELOS PARA CLASSIFICAÇÃO DE RISCO DE CRÉDITO E PREVISÃO DE LUCRATIVIDADE EM UMA COOPERATIVA DE CRÉDITO

São Paulo
2017

Daniel Abreu Vasconcellos de Paula

MODELOS PARA CLASSIFICAÇÃO DE RISCO DE CRÉDITO E PREVISÃO DE LUCRATIVIDADE EM UMA COOPERATIVA DE CRÉDITO

Dissertação apresentada ao Programa de Mestrado Profissional em Administração do Insper Instituto de Ensino e Pesquisa, como parte dos requisitos para obtenção do título de Mestre em Administração.

Área de concentração: Estratégia

Orientador: Prof. Rinaldo Artes
Co-Orientador: Prof. Fábio José Ayres

São Paulo
2017

FOLHA DE APROVAÇÃO

Daniel Abreu Vasconcellos de Paula
MODELOS PARA CLASSIFICAÇÃO DE RISCO DE CRÉDITO E PREVISÃO DE LUCRATIVIDADE EM UMA COOPERATIVA DE CRÉDITO

Dissertação apresentada ao Programa de Mestrado Profissional em Administração do Insper Instituto de Ensino e Pesquisa, como requisito parcial para obtenção do título de Mestre em Administração.
Área de concentração: Estratégia

Aprovado em:

Aos meus pais

AGRADECIMENTOS

Agradeço a todos que me incentivaram e me apoiaram nessa etapa, dentre os quais eu destaco:

- Meu amor, Valéria, pelo apoio, paciência e compreensão.
- Meu grande amigo e parceiro, Marcio, profissional admirável e pessoa dotada de generosidade e bom-humor ímpares.
- Os professores Rinaldo Artes e Fabio Ayres pelas contribuições fundamentais para a realização deste trabalho e sobretudo pelas lições de sabedoria.
- Meus sogros, pelo acolhimento e incentivo.
- Minhas sobrinhas e sobrinhos que me orgulham.
- Ana Ceceli, minha gestora na GM Financial, a quem sou muito grato pela compreensão e apoio.

RESUMO

O controle do risco de crédito e a oferta de produtos financeiros com taxas acessíveis são fatores de gestão determinantes para a sustentabilidade das cooperativas de crédito. Para obterem vantagem competitiva, estas instituições devem se posicionar no setor bancário apresentando vantagens de custo e acessibilidade. Embora as cooperativas de crédito sejam sociedades sem fins lucrativos, os seus objetivos dependem da gestão eficiente de recursos e do risco de crédito das operações, alinhados com princípios doutrinários do cooperativismo. Modelos de *credit scoring* e *profit scoring* são ferramentas que ajudam a melhorar a eficiência das cooperativas de crédito aprimorando a alocação de capital para concessão de empréstimos. Enquanto modelos de *credit scoring* são concebidos para estimar a probabilidade de *default*, modelos de *profit scoring* são concebidos para estimar a lucratividade do cliente com base em fatores comportamentais e demográficos. O presente trabalho aborda estas duas modelagens com a utilização do método de *machine learning* do tipo *random forests* e do método tradicional de regressão logística, com base em dados comportamentais e demográficos observados por um período de dois anos e fornecidos por uma cooperativa de crédito localizada no Brasil. Como benefícios esperados pelo uso destas técnicas podem-se citar: a aquisição de conhecimento sobre a lucratividade potencial dos associados, o direcionamento mais eficaz de recursos para segmentos de cooperados com características semelhantes, e a utilização de métodos objetivos para a mitigação de riscos de crédito na decisão de aprovação de novas operações.

Os modelos estimados pelo método *random forests* mostraram-se superiores aos modelos estimados com a regressão logística. Além disso, o trabalho identificou como variáveis preditoras relevantes: modalidade da operação, *rating* de julgamento subjetivo para risco de crédito, renda, tempo de relacionamento, taxa de juros da operação, histórico de inadimplência e o prazo da operação.

Palavras chave: cooperativas de crédito, *credit scoring, profit scoring*, regressão logística, *random forests*, *machine learning*

ABSTRACT

Credit risk control and the offering of financial products with lower rates are determinant factors for the credit unions sustainable growth and to protect their position in the banking sector. Although credit unions are non profit, the economic objectives shared by their members depend on the management of resources and from the risk of operations, aligned with non profit principles. Models of credit scoring and profit scoring are tools that help to improve the efficiency of credit unions in the capital allocation to lending transactions.

While credit scoring models are created to estimate the probability of default, profit scoring models are created to estimate the client profitability based on behavior and demographic factors. This study intends to compare these two models using the machine learning method, random forests type and the traditional logistic regression method, using behavior and demographic data observed during two years and provided by a credit union located in Brazil. Expected benefits through the use of these techniques we can conclude: the knowledge acquisition about the members potential profitability, the most effective allocation of resources to support the financial planning of cooperative segments with similar profiles, and the use of objective methods to risk mitigation of financial losses in credit operations.

The estimated models by the random forests method appear to be superior to the models estimated by the logistic regression method. Besides that, this study identified as significant predictors variables: purpose of loan, judgemental rating of the application, income, credit history length, borrower's interest rate, delinquency history and the loan term.

Key words: credit unions, credit scoring, profit scoring, logistic regression, random forests, machine learning

SUMÁRIO EXECUTIVO

Introdução

O setor de varejo bancário brasileiro possui participantes das esferas pública e privada que atuam em uma grande rede de agências e praças bancárias. Contudo, existem segmentos da sociedade e de determinadas regiões do país que têm dificuldade de acesso ao crédito. As cooperativas de crédito podem fornecer alternativas para estes grupos obterem empréstimos e financiamentos com condições que possibilitem o desenvolvimento econômico coletivo.

Para ser viável, uma cooperativa de crédito precisa ter resultado positivo na intermediação financeira entre os seus sócios-cooperados. Os recursos oriundos do resultado desta atividade são necessários para garantir a sustentação operacional da instituição.

Modelos de *credit scoring* e *profit scoring* são ferramentas que melhoram o processo de análise de crédito e reduzem as perdas por inadimplência ao introduzir critérios objetivos e consistentes para o aceite de propostas de empréstimos, financiamentos e renovação de limites para operações de crédito.

Entre as técnicas disponíveis para modelagem de *credit scoring e profit scoring*, os algoritmos de aprendizado de máquina, ou *machine learning*, têm apresentado capacidade para modelar relações não-lineares e produzir resultados mais confiáveis que os resultados produzidos por técnicas estatísticas mais convencionais como a regressão logística e a análise discriminante.

Proposta

Este trabalho apresenta instrumentos com o objetivo de melhorar o processo de análise de crédito da cooperativa estudada. Esta instituição utiliza a análise subjetiva de crédito, no qual a qualidade creditícia da operação solicitada é julgada por um analista de crédito com base em documentos fornecidos pelo associado e relatórios de *bureau* de crédito. Este tipo de procedimento possui problemas inerentes à capacidade de julgamento do analista, produzindo resultados potencialmente enviesados e inconsistentes.

Para aprimorar este processo, é sugerida a análise objetiva de crédito com variáveis estudadas na literatura de risco de crédito e a construção de modelos de *credit scoring*

e *profit scoring* comparando técnicas tradicionais com a técnica de *machine learning* do tipo *random forest.*

Contexto

Para a construção dos modelos de *credit scoring* e *profit scoring* são utilizados dados de empréstimos e financiamentos realizados para pessoas físicas nos anos de 2015 e 2016. As variáveis dependentes dos modelos de *credit scoring* e *profit scoring* são respectivamente a probabilidade de *default* (PD) e a taxa interna de retorno (TIR) de operações de crédito. As variáveis independentes são características comportamentais e demográficas dos tomadores, e a classificação subjetiva de risco. Na análise descritiva dos dados foram levantadas as principais estatísticas das variáveis dos modelos e realizados testes para verificação de multicolinearidade.

Evidências obtidas

Os modelos logísticos foram elaborados com três diferentes arranjos de variáveis para avaliar o efeito relativo da análise subjetiva de crédito com relação às demais vari áveis utilizadas neste trabalho. A introdução de variáveis comportamentais e demográficas aumentou consideravelmente o desempenho dos modelos de *credit scoring.*

Verificou-se que a técnica *random forest* apresentou desempenho superior ao modelo de regressão logística para *credit scoring* e também ao modelo de mínimos quadrados ordinários com erro padrão robusto para *profit scoring*.

Foram construídos modelos logísticos para verificar se o comportamento das variáveis independentes corresponde ao comportamento observado em estudos anteriores na área de risco de crédito. De maneira geral, os resultados estão consistentes com os efeitos esperados. As variáveis “modalidade”, “rating”, “renda”, “tempo de relacionamento”, “taxa de juros”, “histórico de atrasos” e “prazo da operação” se mostraram significantes para a previsão de *default*.

Implicações práticas

Os resultados deste trabalho sugerem que a análise objetiva de crédito com modelos determinísticos para classificação de risco e previsão de lucratividade é benéfica para as cooperativas de crédito. A adoção destas ferramentas, substituindo ou complementando o processo de decisão de concessão de crédito, aumenta a

eficiência da gestão dos ativos da cooperativa e permite o tratamento igualitário e transparente dos associados que necessitam de empréstimos e financiamentos.

LISTA DE TABELAS

LISTA DE FIGURAS

LISTA DE ABREVIATURAS

BCB – Banco Central do Brasil
SFN – Sistema Financeiro Nacional
TIR – Taxa Interna de Retorno
PD – Probabilidade de *Default*
ROC – Característica de Operação do Receptor (*Receiver Operating Characteristics)*
KS – Estatística Kolmogorov-Smirnov
CART – Árvores de Classificação e Regressão (*Classification and Regression Trees)*
VIF – Fator de Inflação de Variância (*Variance Inflation Factor*)
CAGR – Taxa de Crescimento Anual Composta (*Compound Annual Growth Rate*)
CHAID – Detecção de Interação Automática do Qui-quadrado – (*Chi-square Automatic Interaction Detector)*
MSE – Erro Quadrático Médio (*Mean Square Error*)
KNN – k vizinhos mais próximos (*k-nearest neighbors*)
RAROC – Retorno Ajustado ao Risco no Capital (*Risk-Adjusted Return on Capital*)

SUMÁRIO

1. INTRODUÇÃO

O objetivo das sociedades cooperativas de crédito é atender a demanda por serviços financeiros dos seus sócios com a utilização de recursos aportados pelos mesmos, e, por meio da ajuda mútua com condições acessíveis, elevar os rendimentos financeiros da coletividade associada ao longo do tempo (POLÔNIO, 1999).

Silva Filho (2002) preconiza que a gestão de uma sociedade cooperativa de crédito deve utilizar instrumentos de avaliação de resultados e desempenho como meios para aumentar a eficiência do processo de tomada de decisões, e, por consequência, da estratégia para gestão.

Os resultados excedentes da cooperativa devem retornar aos associados ao final de cada exercício, seja sobre a forma de reinvestimentos em prol da cooperativa, ou proporcionalmente às operações realizadas por cada sócio, conforme previsto no estatuto destas sociedades (GERIZ, 2010).

Embora o lucro não seja o principal objetivo destas sociedades, para se manter competitiva no setor bancário a gestão do segmento de cooperativas de crédito não deve diferir das demais instituições financeiras com relação às seguintes características: a utilização de controles de liquidez e solvência, a busca por economias de escala e a gestão eficiente dos ativos financeiros (SILVA FILHO, 2002). Neste contexto, a modelagem do risco de crédito (modelos de *credit scoring*), e a previsão da lucratividade dos empréstimos (modelos de *profit scoring*) são recursos úteis para assegurar um melhor desempenho das cooperativas.

Modelos de *credit scoring* são amplamente empregados na indústria financeira para controle de risco de crédito. Por meio desta ferramenta determina-se a probabilidade de *default* (PD) do tomador de crédito com base em observações passadas de clientes com características comportamentais e demográficas semelhantes. Uma vez prevista esta probabilidade, é definido um nível tolerável de risco de clientes, minimizando as perdas decorrentes da inadimplência (LEWIS, 1992).

Modelos de *profit scoring* são utilizados para prever a lucratividade de um cliente ou operação. O conceito desta abordagem baseia-se no reconhecimento de que tomadores de crédito adimplentes podem não gerar receitas suficientes para compensar os custos associados com a manutenção das suas contas, ao passo que tomadores inadimplentes podem ser rentáveis se eles contratam crédito ativamente e

honram a maior parte dos seus compromissos (SANCHEZ-BARRIOS; ANDREEVA; ANSELL, 2016). Com efeito, na metade dos anos 90 começaram a ser divulgados os primeiros trabalhos acadêmicos de *profit scoring* para tomada de decisão com as perspectivas de maximização de rentabilidade e controle eficaz do risco de crédito dos tomadores (CROOK; EDELMAN; THOMAS, 2007).

Dentre as decisões que esta modelagem permite endereçar estão: a escolha do nível de risco adequado para seleção de operações rentáveis ao longo do tempo e a formulação de estratégias para aquisição e retenção de clientes rentáveis (SANCHEZ-BARRIOS; ANDREEVA; ANSELL, 2016).

Este trabalho propõe a identificação de variáveis associadas ao risco de perdas financeiras em operações de empréstimos em uma cooperativa de crédito, por meio da construção de modelos de *credit scoring* e *profit scoring.* Tais modelos são construídos com os objetivos de: aumentar a eficiência da gestão de risco de crédito; sustentar a posição competitiva da cooperativa; estabelecer critérios objetivos e alinhados com o interesse majoritário dos sócios para aprovação das operações de crédito contribuindo, por exemplo, para a redução de conflitos de agência na decisão de concessão de empréstimos. Os modelos foram construídos e validados com a base de dados fornecida por uma cooperativa de crédito contendo informações sobre empréstimos ativos nos anos de 2015 e 2016.

Um segundo objetivo é comparar a eficácia de diferentes abordagens no desenvolvimento de modelos: a técnica tradicional de regressão logística e o algoritmo *random forests*.

Os principais benefícios esperados pelo emprego dessas abordagens são a contribuição para o aumento da sustentabilidade da cooperativa de crédito no longo prazo com economias de escala no processo de análise de crédito e a redução de perdas por incumprimento de obrigações.

Esta dissertação está organizada na seguinte ordem: esta introdução; a revisão de literatura e os comportamentos esperados das variáveis dos modelos (capítulo 2); a metodologia utilizada (capítulo 3); a descrição dos dados utilizados e o resultado dos modelos (capítulo 4); as conclusões do estudo e as sugestões de pesquisa (capítulo 5). Em seguida são apresentados os anexos e as referências bibliográficas.

2. REVISÃO DE LITERATURA

O cooperativismo moderno surgiu em 1844 na cidade de Rochdale, na Inglaterra, durante a Revolução Industrial. Um grupo de tecelões fundou uma cooperativa baseada em princípios éticos e comportamentais que são a base do cooperativismo contemporâneo (PINHEIRO, 2008).

Pinheiro (2008) descreve que a instituição pioneira que serviu de modelo para atividade cooperativista de crédito surgiu em 1847 na Alemanha e foi instituída por Friedrich Wilhelm Raiffeisen. As cooperativas fundadas por Raiffeisen tinham como principais características: a responsabilidade ilimitada e solidária dos associados, ou seja, o associado responde por dívidas contraídas pela organização entregando os seus bens particulares; o peso equânime do voto por associado independentemente da sua participação societária; a não distribuição de sobras, excedentes ou dividendos.

Em 1856, Herman Schulze criou a primeira cooperativa de crédito urbana da Alemanha. As cooperativas fundadas por Herman Schulze diferenciam-se das cooperativas de Raiffeisen por permitir que as sobras retornem ao associado em volume proporcional à sua participação no capital social. As cooperativas que seguem estas regras atualmente são conhecidas como cooperativas com modelo Schulze-Delitzsch (PINHEIRO, 2008).

Segundo Soares e Melo Sobrinho (2008), os primeiros registros do cooperativismo de crédito no Brasil datam de 1902, quando, por inciativa do padre suíço Theodor Amstad, foi criada, no Rio Grande do Sul, a Sociedade Cooperativa Caixa de Economia e Empréstimos de Nova Petrópolis. Esta sociedade é uma cooperativa do tipo Raiffeisen que existe até os dias de hoje. Após esta iniciativa pioneira, foram criadas outras cooperativas de crédito para atender a população rural.

No âmbito regulatório, sociedade cooperativa é o conjunto de pessoas que promovem ajuda mútua por meio da prestação de serviços, com forma institucional e natureza jurídica específicas, reguladas no Brasil pela Lei nº 5.764/71.

As cooperativas de crédito são instituições que prestam serviços financeiros aos seus associados e são constituídas sob a forma de sociedade cooperativa sem fins lucrativos. As mesmas são habilitadas a prestar aos seus associados praticamente todos os serviços financeiros ofertados por um banco comercial (PINHEIRO, 2008).

Para ter acesso aos serviços e produtos de uma cooperativa de crédito, o interessado deve adquirir direitos de propriedade comprando quotas-partes do capital social da instituição. Os sócios possuem direito sobre os resíduos (ou seja, as sobras líquidas financeiras provenientes da atividade da cooperativa) que são distribuídos proporcionalmente à quantidade de operações realizadas por cada indivíduo, não havendo proporcionalidade entre capital investido e a distribuição de sobras, salvo deliberação de novas regras na Assembléia Geral (GERIZ, 2010).

As cooperativas de crédito são instituições financeiras reguladas pelo Banco Central do Brasil através da Resolução nº 4.434 de 2015 normatizada pelo Conselho Monetário Nacional. Nesta regulação estão definidos os dispositivos normativos para abertura, operação, modificação e dissolução de sociedades cooperativas de crédito. Os atos societários deliberados pelos cooperados referentes à eleição de membros da diretoria e Conselho Fiscal, reforma do estatuto social, fusão, incorporação ou desmembramento, dissolução voluntária e liquidação da sociedade passam a ter validade apenas com aprovação do Banco Central do Brasil. A natureza da responsabilidade dos sócios pode ser limitada ou ilimitada, conforme determinem os estatutos (GERIZ, 2010).

No Brasil, as cooperativas de crédito têm apresentado crescimento expressivo nos últimos anos, assim como a sua representatividade no setor bancário. A Tabela 1 ilustra o crescimento das cooperativas em ativos e patrimônio líquido em comparação ao segmento bancário consolidado.

Tabela 1: Crescimento do Cooperativismo de Crédito no Brasil (R$ Bilhões)

Segmento	Agregados Patrimoniais	2011	2012	2013	2014	2015	CAGR
Cooperativas de Crédito	Patrimônio Líquido	16	19	23	27	32	19%
	(% crescimento)		*21%*	*18%*	*20%*	*17%*	
	Ativos	86	104	124	151	183	21%
	(% crescimento)		*20%*	*20%*	*21%*	*21%*	
	Depósitos	38	47	56	69	83	21%
	(% crescimento)		*23%*	*20%*	*22%*	*21%*	
	Operações de Crédito	36	46	58	68	76	20%
	(% crescimento)		*26%*	*27%*	*18%*	*12%*	
Segmento Bancário (excluindo Cooperativas)	Patrimônio Líquido	371	439	433	459	481	7%
	(% crescimento)		*18%*	*-1%*	*6%*	*5%*	
	Ativos	4.274	4.981	5.456	6.199	6.863	13%
	(% crescimento)		*17%*	*10%*	*14%*	*11%*	
	Depósitos	1.595	1.647	1.751	1.830	1.928	5%
	(% crescimento)		*3%*	*6%*	*5%*	*5%*	
	Operações de Crédito	1.667	1.947	2.273	2.538	2.990	16%
	(% crescimento)		*17%*	*17%*	*12%*	*18%*	
Representatividade das Cooperativas de Crédito	Patrimônio Líquido	4,1%	4,2%	5,0%	5,6%	6,2%	
	Ativos	2,0%	2,0%	2,2%	2,4%	2,6%	
	Depósitos	2,3%	2,8%	3,1%	3,6%	4,1%	
	Operações de Crédito	2,1%	2,3%	2,5%	2,6%	2,5%	

Fonte: Banco Central do Brasil

Os principais agregados patrimoniais do segmento de cooperativas de crédito (patrimônio líquido, ativos, depósitos e carteira de crédito) tiveram crescimento médio anual (calculado pelo indicador CAGR – *compound annual growth rate*) em torno de 20% ao ano entre 2011 e 2015. Já os agregados patrimoniais dos demais participantes do segmento bancário cresceram abaixo de dois dígitos no mesmo período, exceto em ativos e operações de crédito que cresceram com taxas de 13% e 16% respectivamente. Como consequência dessa expansão, a representatividade das cooperativas de crédito no segmento bancário teve um crescimento relevante saltando de 4,1% em 2011 para 6,2% em 2015 a participação no patrimônio líquido do setor.

Cuevas e Fischer (2006) observam que as cooperativas de crédito, embora estejam difundidas pela maior parte dos países no mundo, são entidades pouco

estudadas quando comparadas com as demais instituições que realizam intermediação financeira.

Giarola et al. (2009) argumentam que as cooperativas de crédito possuem duas vantagens para os associados, comparativamente às demais instituições financeiras: o *spread* bancário e o custo do empréstimo são menores; e as sobras líquidas, o resultado financeiro dos serviços prestados pelas cooperativas menos as despesas administrativas, retornam para os cooperados de acordo com a decisão da assembleia geral. No entanto, para manter-se viável com taxas de juros mais baixas e clientes com maior risco de crédito em um setor altamente lucrativo e dominado por instituições de grande porte, a cooperativa deve manter um nível adequado de retenção de sobras para realizar os investimentos necessários nos ativos operacionais e na gestão de riscos, liquidez e alavancagem.

Giarola et al. (2009) também argumenta que uma característica particular das cooperativas de crédito é o fato de que o seu insumo produtivo, ou seja, o dinheiro, é gerado majoritariamente pelos depósitos e aquisição de quotas-partes dos associados e então repassado como empréstimos a outros associados. Essa característica particular das cooperativas de crédito exige eficiência na gestão das atividades da cooperativa, uma vez que a má gestão administrativa é revertida negativamente para todos os associados.

Lima, Araújo e Amaral (2008) argumentam que a mais importante das vantagens das cooperativas de crédito é permitir aos associados o acesso aos serviços financeiros, mesmo em tempos em que ocorra racionamento de crédito no mercado financeiro tradicional. A carga tributária reduzida é uma importante vantagem competitiva que permite às cooperativas se diferenciarem das demais instituições financeiras do setor de varejo bancário por possuir custos potencialmente mais baixos.

Além das vantagens, as cooperativas de crédito possuem fraquezas quando comparadas com outras instituições do setor bancário. A vulnerabilidade para problemas de agência é um fator de risco para sustentabilidade das cooperativas de crédito no longo prazo (CUEVAS; FISCHER, 2006).

Segundo Eisenhardt (1989), a teoria da agência tem interesse em resolver dois problemas que podem ocorrer em relações de agência. Estes problemas são: a dificuldade tática e onerosa para o principal verificar o comportamento do agente; e desejos e objetivos conflitantes entre o principal e o agente. O que caracteriza o primeiro problema é o fato de que o principal não tem meios para se certificar que o

agente se comportou de maneira adequada. O segundo problema é o compartilhamento de risco que emerge quando principal e agente possuem diferentes atitudes em relação ao risco. Nesta situação, o principal e o agente podem preferir diferentes ações devido as diferentes preferencias por risco.

Segundo Cornforth (2004) a teoria da agência atribui que o principal interesse dos acionistas é maximizar o lucro e que características do mercado de controle corporativo, como a pressão de acionistas majoritários, a ameaça de aquisição e o monitoramento do conselho de administração irão ajudar os gestores a se manterem alinhados em torno deste objetivo. Em sociedades cooperativas a situação é diferente, as cooperativas são estabelecidas para servir aos interesses dos seus associados e, portanto, o lucro é um meio para se atingir uma finalidade e não uma finalidade em si.

O conflito de agência em cooperativas de crédito possui características particulares quando comparado com as demais instituições financeiras. Primeiro, os proprietários das cooperativas são ao mesmo tempo clientes. Por haver heterogeneidade na preferência dos clientes por aplicações rentáveis ou por empréstimos com taxas atrativas, ocorre o conflito de interesses entre aplicadores e tomadores. Segundo, as decisões tomadas em assembleia geral, incluindo a eleição do quadro administrativo, são baseadas no voto de cada associado, não havendo distinção pelo volume de participação no capital social. Por fim, os associados eleitos para a diretoria geralmente possuem menos qualificação técnica que a mão de obra especializada existente na indústria financeira (LIMA; ARAÚJO; AMARAL, 2008).

Souza (2017) argumenta que quanto mais homogêneo for o quadro de membros com relação a interesses e visão para o longo prazo, menores tendem a ser os custos de agência da cooperativa de crédito, pois nessa situação o voto expressará a vontade da maioria. Nas cooperativas de crédito muito heterogêneas, ao contrário, as decisões se afastarão muito da vontade de cada associado e, então, organizações de *shareholders* , cujo objetivo de maximizar o lucro é mais bem definido e alcançável, tendem a ter mais vantagens (HART; MOORE, 1998).

O estudo empírico realizado por Cuevas e Fischer (2006) sugere que quanto maior o tamanho de uma cooperativa de crédito, menor tende a ser o seu nível de eficiência. O aumento do número de associados contribui para enfraquecer a coesão dos objetivos traçados para a cooperativa e, por consequência, a governança, podendo até mesmo, em casos extremos, levar a instituição à falência.

A identificação das variáveis que possuem relação com o risco de crédito dos sócios permite o controle estatístico e objetivo do processo de concessão de empréstimos e financiamentos. Uma vez que esse processo é objetivo e padronizado, é possível obter escalabilidade e maior imparcialidade na tomada de decisão, havendo, portanto, maior probabilidade de redução de custos de agência para a cooperativa.

Modelos de *credit scoring e profit scoring* são ferramentas que potencialmente podem ajudar a reduzir conflitos de agência entre sócios e agentes na etapa de concessão de crédito. Com a utilização dos modelos na tomada de decisão, espera-se que os recursos cedidos pela cooperativa para empréstimos retornem multiplicados pela taxa de retorno necessária para a sustentação e cumprimento dos objetivos da instituição, eliminando-se assim, razões divergentes do interesse geral dos sócios para a concessão de empréstimos.

2.1. MODELOS DE *CREDIT SCORING* E *PROFIT SCORING*

Durante a Segunda Guerra Mundial e, de maneira mais acentuada, a partir da segunda metade do século XX, foram introduzidos e aprimorados os modelos de *credit scoring* que se estabeleceram como referência para a indústria financeira. A primeira técnica de partição de dados em categorias foi introduzida por Fisher (1936) e é conhecida como análise discriminante. O primeiro modelo de *credit scoring* que se tem registro está no trabalho publicado por Durand (1941).

Dentre as abordagens utilizadas para a classificação estatística de qualidade creditícia, destacam-se como referenciais para a indústria de crédito de varejo as técnicas de análise discriminante e de regressão logística (CROOK; EDELMAN; THOMAS, 2007; LEWIS, 1992).

De acordo com Sanchez-Barrios, Andreeva e Ansell (2016), a modelagem de previsão de lucro de clientes tem emergido como ferramenta para decisão de concessão de crédito. Tomadores de crédito são agrupados de acordo com a sua lucratividade para o ofertante de crédito baseando-se em índices de previsão de lucratividade no lugar das probabilidades de insolvência e as respectivas perdas financeiras associadas.

Com a evolução da computação e o desenvolvimento de ferramentas para análise de grandes bancos de dados, as instituições financeiras começaram a ampliar o escopo do processo de análise de crédito ponderando a redução do risco de perdas

com a maximização dos lucros. Além da previsão do risco de insolvência de uma operação de crédito, passou-se a ser estimada a lucratividade que um cliente pode propiciar ao longo do tempo com a utilização de crédito (SERRANO-CINCA; GUTIÉRREZ-NIETO, 2016).

Segundo Yap, Ong e Husain (2011), o aumento da capacidade de processamento computacional permitiu que modelos mais eficientes baseados em técnicas de aprendizado de máquina (*machine learning*) fossem introduzidos nas instituições financeiras.

Segundo Forsyth (1984), o termo *machine learning* refere-se ao conjunto de técnicas aplicadas em algoritmos computacionais para a melhoria de sistemas de tomada de decisão. Os dados de entrada são processados pelos algoritmos e os resultados produzidos são utilizados para analisar futuras novas entradas, havendo, portanto, a retenção de conhecimento sobre o comportamento dos dados. Técnicas estatísticas são utilizadas para a avaliação da qualidade dos resultados produzidos e estabelecer critérios para a utilização dos dados durante o processo de aprendizagem.

Os algoritmos de métodos de *machine learning* podem ser caracterizados em algoritmos de aprendizado supervisionado e aprendizado não supervisionado. Os algoritmos de aprendizado supervisionado recebem a informação sobre a classe dos dados durante a etapa de desenvolvimento do modelo, para então, posteriormente realizar a classificação de novos conjuntos de entrada sem indicação de classes. Os algoritmos de aprendizado não supervisionado não recebem a informação sobre as classes dos dados, sendo o próprio algoritmo responsável por identificar as classes existentes nos dados (SILVERIO, 2015).

A modelagem de *profit scoring* permite, por exemplo, a previsão da taxa interna de retorno das operações de empréstimo que cada cliente pode gerar de acordo com características comportamentais e demográficas. Diversos autores têm recomendado a utilização de técnicas avançadas de *machine learning* para modelagem de *profit scoring* devido à maior precisão e maior robustez para acomodar relações não-lineares (SERRANO-CINCA; GUTIÉRREZ-NIETO, 2016).

Na tabela a seguir, estão listados alguns dos principais métodos de classificação em *machine learning* e publicações de relevância relacionadas a aplicações para *credit scoring e profit scoring*.

Tabela 2: Métodos de *machine learning* e aplicações

	Técnica	Propósito	Autores
1	Árvores de Decisão	*Credit Scoring*	Srinivisan e Kim (1987)
3	Redes Neurais	*Credit Scoring*	Malhorta (2003)
6	knn (*k-Nearest Neighbors*)	*Credit Scoring*	West (2000)
7	Máquinas de Vetores de Suporte	*Credit Scoring*	Schebesch e Stecking (2005)
8	Regressão Logística	*Profit Scoring*	So et al. (2014)
9	Redes Neurais	*Profit Scoring*	Verbraken et al. (2014)
8	Análise de Sobrevivência	*Profit Scoring*	Sanchez-Barrios, Adreeva e Ansell (2016)
9	CHAID	*Profit Scoring*	Serrano-Cinca e Gutiérrez-Nieto (2016)

Fonte: elaborada pelo autor

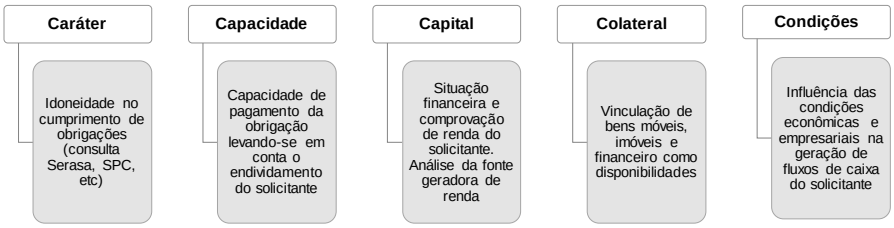
2.2. ANÁLISE DE CRÉDITO EM INSTITUIÇÕES FINANCEIRAS

Usualmente, os profissionais da área de crédito de bancos comerciais utilizam duas modalidades de procedimentos para realizar a análise de crédito de pessoas físicas para concessão de crédito: a análise subjetiva e a análise objetiva (LEWIS, 1992) .

Segundo Securato (2002), a análise subjetiva de crédito assenta-se sobre um conjunto de informações armazenadas pela área de análise de crédito. Essas informações são tipicamente dados cadastrais, financeiros, de caráter e de relacionamento. A subjetividade advém da interpretação e julgamento de cada analista de crédito para identificar fatores de risco que evidenciem a capacidade de pagamento comprometida de pessoas físicas.

Segundo Gitman (1997) e Lewis (1992), a análise subjetiva de credito possui cinco variáveis de análise: caráter, capacidade, colateral e condições. Essas variáveis são descritas na Figura 1.

Figura 1: Cs da análise de crédito



Fonte: Elaborada pelo autor a partir de figura semelhante publicada em Santos e Famá (2007)

A análise objetiva de crédito utiliza métodos estatísticos como critérios para tomada de decisão de concessão de crédito (SANTOS; FAMÁ, 2007).

A tomada de decisão na análise objetiva de crédito é baseada na probabilidade de um tomador de crédito vir a se tornar inadimplente no futuro, tendo como base os seus atributos de risco (THOMAS; EDELMAN; CROOK, 2002).

A redução de perdas financeiras por incumprimento e de custos de aprendizado sobre o comportamento dos clientes estão entre os principais argumentos favoráveis à modelagem de *credit scoring*. Esta abordagem tende a apresentar desempenho superior quando comparada com a análise subjetiva para a determinação da probabilidade de insolvência para pedidos de crédito e a renovação de limites de crédito dos clientes (LEWIS, 1992).

2.3. VARIÁVEIS UTILIZADAS NA ANÁLISE OBJETIVA DE CRÉDITO

A seguir, são elencadas as variáveis utilizadas nesse estudo para análise objetiva de crédito, assim como a relação esperada destas variáveis com a inadimplência segundo estudos na área de risco de crédito.

Modalidade da operação de crédito

É a forma de operação de crédito escolhida pelo requerente de acordo com o seu propósito financeiro, por exemplo, crédito pessoal, crédito para aquisição de bens, crédito consignado, entre outras.

Segundo Lawrence e Elliehausen (2008), as operações de crédito pessoal e crédito consignado possuem risco maior que outras modalidades de crédito direto ao consumidor, uma vez que as conveniências para contratação da dívida, tais quais

limites pré-aprovados para pequenos volumes e prazos curtos, atraem tomadores com maior propensão ao incumprimento de obrigações.

Informações de agências de classificação de crédito

São informações relacionadas com o histórico de pagamento dos clientes no mercado de crédito e são obtidas por meio de consultas de cadastro negativo em empresas que fornecem estes dados tais como SPC e Serasa. Internamente as instituições financeiras podem monitorar o comportamento histórico de pagamento dos seus clientes por meio da análise de relatórios gerenciais que informam a pontualidade no pagamento de parcelas.

Muitas instituições financeiras utilizam estas informações como principal critério para decisão de concessão de crédito, sem levar em consideração a influência de outros fatores que não são identificados no histórico de pagamento dos clientes no mercado de crédito (AVERY; BOSTIC; CALEM, 2000).

Avery et al. (2000) argumentam que decisões de concessão de crédito baseadas preponderantemente em informações de relatórios de crédito fornecidos por *bureaux* de crédito possuem menor eficiência que modelos de *credit scoring*, que agregam outras variáveis do processo de análise de crédito tais como: renda, tempo no emprego e situação de saúde.

Rendimentos

São informações relacionadas à renda mensal líquida dos tomadores de crédito e são obtidas mediantes a apresentação de comprovantes de rendimento emitidos pelas fontes pagadoras.

Black e Morgan (1998) sugerem que existe uma relação direta entre a perda de renda e a taxa de inadimplência de pessoas físicas em empréstimos bancários, ressaltando a importância da análise da capacidade financeira durante o processo de análise de concessão de crédito.

Garantias

São informações relacionadas à vinculação de bens móveis e imóveis como garantia para a contração de empréstimos. Tais dados são obtidos preferencialmente da declaração de imposto de renda do tomador ou documentos registrados em

cartório que atestem a posse da totalidade ou predominância dos bens dados como garantia.

Segundo Hynes (1998) , a colocação de bens patrimoniais como garantias em contratos de empréstimos aumenta a probabilidade de pontualidade na amortização dos saldos devedores, uma vez que os tomadores possuem o risco de ter o arresto dos bens exigidos pela instituição financeira como garantia.

Tempo de relacionamento bancário

É calculada como o tempo decorrido entre a abertura da conta corrente e a realização da operação de crédito.

Segundo Chakravarty e Scott (1998), quanto maior for o tempo de existência da conta corrente, maior será a possibilidade do banco de adquirir conhecimento sobre a idoneidade do cliente, o que leva a uma maior proporção de correntistas com histórico positivo de pagamento para contas com maior tempo de duração.

Taxa de juros

É a taxa de juros nominal da operação.

De acordo com Serrano-Cinca e Gutiérrez-Nieto (2016) e Dinh e Kleimeier (2007), a taxa de juros de uma operação de crédito deve ser maior tanto quanto for a sua probabilidade de *default* para compensar as perdas oriundas por incumprimento dos empréstimos inadimplentes.

Valor Financiado

É a valor de empréstimo requerido pelo associado para a cooperativa de crédito.

O estudo realizado por Araújo et al (2007) em uma instituição de microcrédito sugere que o valor financiado é inversamente proporcional à probabilidade de *default* de uma operação de crédito, uma vez que os empréstimos com valores maiores são fornecidos preferencialmente para clientes com histórico positivo de pagamento.

Endividamento

É a razão entre o valor da parcela do empréstimo e a renda mensal do tomador.

Serrano-Cinca e Gutiérrez-Nieto (2016) argumentam que quanto maior for o endividamento de um tomador de crédito, maior será a sua probabilidade de *default*,

uma vez que o tomador terá menor disponibilidade de renda para o pagamento das despesas domésticas e não recorrentes.

Histórico de pontualidade no pagamento de obrigações

É o número de meses decorridos desde a última ocorrência de um atraso maior que 30 dias para o pagamento de uma parcela em aberto de um determinado cliente nos últimos dois anos. No caso de não haver eventos observados para um dado cliente, esta variável assume valor zero.

Segundo os estudos empíricos de Desai et. al (1997), clientes com histórico de inadimplência possuem maior probabilidade de *default.*

Prazo da operação

É o prazo em dias entre a data de contratação da operação e a data prevista contratualmente para a liquidação do empréstimo.

Segundo Dinh e Kleimeier (2007), a preferência dos clientes por maiores prazos para a contratação de crédito está relacionada diretamente com o risco de crédito da operação. Os clientes com incumprimento de obrigações possuem maior representatividade em operações de prazo mais longo comparativamente a clientes adimplentes.

2.4. COMPORTAMENTO ESPERADO DAS VARIÁVEIS

A utilização de modelos estatísticos para análise objetiva de crédito tem como pressuposto o uso de variáveis que estão relacionadas significativamente com a capacidade de pagamento dos tomadores de crédito (LEWIS, 1992).

As variáveis selecionadas foram contextualizadas com estudos na área de risco de crédito apresentados na seção de revisão bibliográfica. É importante destacar a definição da variável RATING para este estudo; trata-se da nota de qualidade de crédito concedida, de forma subjetiva, pela cooperativa para novos contratos de empréstimo com base em informações cadastrais e dados fornecidos por agências de classificação de crédito (ou *bureaux* de crédito)

A seguir, as variáveis utilizadas nos modelos estão relacionadas com o efeito esperado na determinação da probabilidade de *default*. Adotou-se como convenção o efeito negativo elevando a probabilidade de *default,* aumentando, portanto, o risco da operação e o efeito positivo reduzindo o risco da operação e a probabilidade de *default*.

Tabela 3: Expectativas para variáveis independentes categorizadas

Variável	Níveis da variável	Efeito esperado na PD	Referência
MOD (Modalidade da operação)	outros empréstimos	positivo	Lawrence and Elliehausen (2008)
	crédito pessoal - sem consignação em folha de pagam.	negativo	
	crédito pessoal - com consignação em folha de pagam.	negativo	
	aquisição de bens – veículos automotores	positivo	
RATING (Análise Subjetiva)	Rating não A	negativo	Avery et al (2010)
	Rating A	positivo	
GARANTIA	Sem Garantia	negativo	Hynes (1988)
	Possui Garantia	positivo	
RENDA (Rendimentos Mensais)	até R$ 1.000	negativo	Black e Morgan (1998)
	de R$ 1.000 até R$3.000	positivo	
	de R$ 3.000 até R$6.000	positivo	
	mais de R$ 6.000	positivo	

Fonte: elaborado pelo autor

Tabela 4: Expectativas para variáveis independentes contínuas

Variável	Definição	Efeito esperado na PD	Referência
TEMPOREL (Tempo de Relacionamento)	tempo decorrido entre a abertura de conta na instituição e a operação contratada.	positivo	Chakravarty e Scott (1998)
TAXA (Taxa de Juros)	taxa de juros da operação	negativo	Serrano-Cinca e Gutiérrez-Nieto (2016)
ENDIV (Endividamento)	razão do valor de parcela sobre renda mensal	negativo	Serrano-Cinca e Gutiérrez-Nieto (2016)
DLQANT (Histórico de Atrasos)	quantidade de eventos anteriores de atraso com duração maior que 30 dias.	negativo	Chakravarty e Scott (1998) Desay et al (1997)
PRAZO (Prazo do Empréstimo)	prazo da operação contando da contratação ao vencimento em dias	negativo	Dinh et al (2007)
VALOR (valor emprestado)	valor solicitado para empréstimo	positivo	Araújo et al (2007)

Fonte: elaborado pelo autor

2.5. COMPARAÇÃO DE TÉCNICAS DE MODELAGEM

Técnicas emergentes de *machine learning* para a modelagem de *credit scoring,* como *random forests,* têm apresentado desempenho superior aos modelos tradicionais baseados nas técnicas de análise discriminante e regressão logística (LESSMANN; BAESENS; THOMAS, 2015). Deseja-se verificar se o mesmo ocorre com os dados da cooperativa de crédito.

Para realizar as modelagens de *credit scoring* serão utilizados o algoritmo *random forests* para classificação preditiva com *machine learning* e a técnica de regressão logística. A estatística de Kolmogorov-Smirnov, o Coeficiente de Gini e a área sobre a curva ROC (*receiver operating characteristic*) são os indicadores de qualidade que serão utilizados para comparar o desempenho das técnicas.

3. METODOLOGIA

Nesta seção serão apresentadas as diferentes metodologias utilizadas no desenvolvimento dos modelos de *credit scoring e profit scoring*.

3.1. REGRESSÃO LOGÍSTICA

A regressão logística é um modelo que permite prever a probabilidade de ocorrência de uma possível realização de uma variável nominal, frequentemente binária ou dicotômica. Os benefícios deste método para modelagem de *credit scoring* são: o fato da regressão logística não assumir que a associação entre variáveis dependentes e independentes seja linear, tampouco que o conjunto de variáveis tenha distribuição normal. A forma da equação de regressão logística é definida pela transformação logito:

$$\ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_m x_{mi} = x_i^T \beta, \quad (1)$$

sendo o termo p_i a probabilidade de um cliente ser "bom", ou seja, não entrar em default; β_j são os parâmetros do modelo, e x_{ij} é o valor da variável preditora j, $j = 1, \ldots, m$, para o indivíduo i, com i variando de 1 até n, $x_i^T = (1, x_{1i}, \ldots x_{mi})$ e $\beta = (\beta_0, \beta_1, \ldots \beta_m)^T$.

Os parâmetros da equação (1) podem ser estimados por meio do método da máxima verossimilhança (HOSMER et al. 2013)

A fração $\frac{p_i}{1-p_i}$ determina a razão entre "bons" e "maus" clientes e é chamada em inglês de *odds* ou chance. O logaritmo da chance é chamado de logito. Aplicando a exponenciação nos dois lados da equação (1), obtemos a seguinte transformação:

$$p_i = \frac{e^{x_i^T \beta}}{1 + e^{x_i^T \beta}} = \frac{1}{1 + e^{-x_i^T \beta}} \quad (2)$$

3.2. RANDOM FORESTS

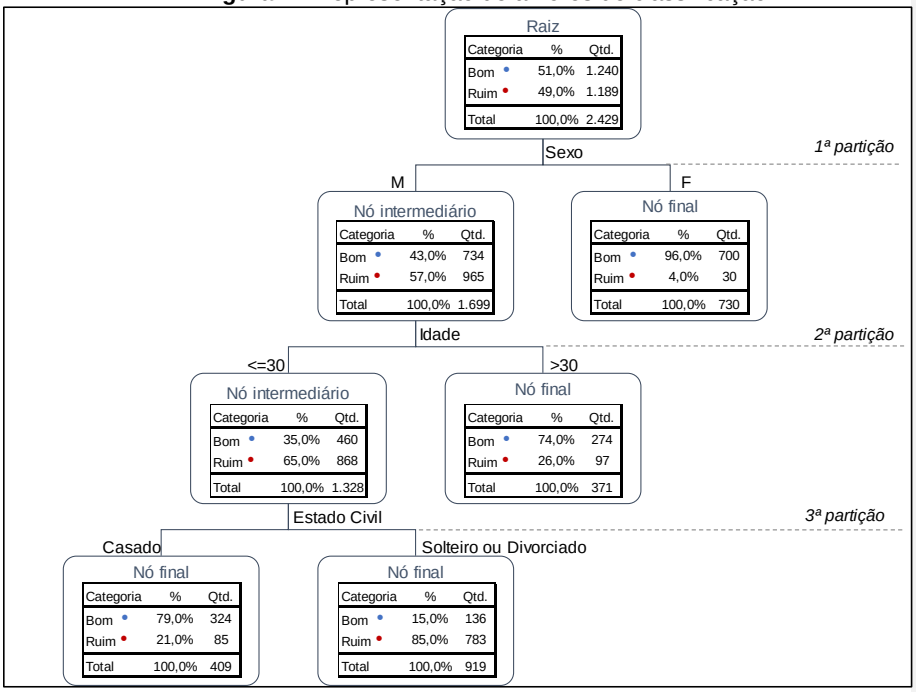
Random Forests (BREIMAN, 2001) é um algoritmo de *machine learning* utilizado para aprendizado supervisionado a partir da combinação de árvores de decisão. A técnica utilizada nesta abordagem é uma evolução do algoritmo *bootstrap aggregating*, também conhecido como *bagging* (BREIMAN, 1996). Portanto, para

compreender o *Random Forest*, é necessário abordar os conceitos dos algoritmos de árvores de decisão e *bagging*.

Classification and Regression Trees (CART) é uma técnica não-paramétrica proposta por Breiman et al. (1984) para o desenvolvimento de modelos de classificação e regressão preditiva, por meio de árvores de decisão. O conjunto de dados para modelagem é definido por uma variável dependente contínua ou nominal e uma lista de variáveis independentes. Este método executa partições binárias sucessivas nos dados buscando a formação de subconjuntos homogêneos até que sejam atendidos critérios de qualidade (ou pureza) pré-definidos. Uma técnica usual para avaliar a qualidade dos subconjuntos, ou nós das árvores, é o cálculo do coeficiente de Gini.

A previsão de valores ou categorias para a variável dependente é realizada com base na estrutura de classificação presente nos nós terminais da árvore de decisão. A Figura 2 ilustra a estrutura simplificada de uma árvore de regressão e classificação:

Figura 2: Representação de árvores de classificação



Fonte: (SILVERIO, 2015)

No exemplo da Figura 2 podem-se visualizar os processos sucessivos de divisão binária dos dados em classes. A partir do nó raiz, é selecionada a variável (ou partição) que permite uma melhor separação de grupos segundo critérios de qualidade como, por exemplo, o Coeficiente de Gini. O processo se desenvolve com o particionamento posterior dos novos nós gerados.

O critério de parada do algoritmo para as partições sucessivas é, por exemplo, a variação de qualidade do último particionamento comparado com o particionamento imediatamente anterior. Se essa variação não for relevante, essa partição será descartada e o nó anterior da árvore será registrado como nó terminal. O processo estará concluído quando não houver mais nós para serem particionados.

No método *bagging*, o conjunto de dados é dividido aleatoriamente em um grande número de subamostras, sorteadas na amostra original, com reposição, e para cada subamostra é gerada uma árvore de decisão. A previsão é calculada como a média (para regressão) ou a maioria de votos (para classificação) das previsões obtidas de cada árvore desenvolvida para as subamostras. Essa técnica permite o aumento da estabilidade e da precisão dos resultados, uma vez que os efeitos de ruídos e *outliers* são atenuados com as diversas amostragens.

Uma importante limitação da técnica *bagging* é a possibilidade de geração de árvores muito parecidas, o que por sua vez eleva a taxa de erro de previsão do modelo, uma vez que as variáveis independentes são sempre as mesmas.

A técnica *random forest* possui dois passos adicionais à seleção de subamostras realizada na técnica bagging. Havendo M variáveis preditoras no conjunto de dados, para cada subamostra serão selecionadas aleatoriamente $m < M$ variáveis na construção das árvores individuais. O valor m é mantido constante durante o processo de aprendizagem do modelo. Esta característica permite a redução dos erros de previsão por problemas de multicolinearidade, por exemplo.

4. RESULTADOS

A base de dados utilizada para a criação dos modelos é composta por variáveis comportamentais e demográficas de empréstimos observados em um período de 24 meses (de janeiro de 2015 até dezembro de 2016), tendo sido fornecida por uma cooperativa de crédito habilitada pelo Banco Central do Brasil. Para atender aos objetivos do estudo foram consideradas apenas as operações com clientes do tipo Pessoa Física. A cooperativa não realizou o armazenamento em computadores dos registros das propostas de crédito rejeitadas durante o mesmo período.

Cada registro da base de dados representa uma operação de crédito realizada no período, e contém informações relativas às características da operação contratada, informações do tomador de crédito e do desenvolvimento da operação. Este último conjunto abrange o histórico de pagamento, a marcação de atrasos e a evolução do saldo devedor até a liquidação da operação por quitação ou baixa para prejuízo.

O critério escolhido para classificação de clientes “maus” pagadores para a modelagem *de credit scoring* é a ocorrência de atraso de pagamento igual ou superior a noventa dias após a data de vencimento.

O critério escolhido para a definição de lucratividade das operações na modelagem de *profit scoring* é o cálculo da taxa interna de retorno (TIR) sobre o valor dos empréstimos, o mesmo utilizado por Serrano-Cinca e Gutiérrez-Nieto (2016). Os parâmetros de cálculo da taxa interna de retorno são os pagamentos recebidos e os custos envolvidos em cada operação contratada.

Segundo Brealey e Myers (2012), a taxa interna de retorno de uma aplicação ou investimento é definida como a taxa de desconto que faz com que o valor presente líquido de um fluxo de caixa com duração de T períodos seja igual a zero, conforme a fórmula seguir:

$$VPL = C_0 + \sum_{I=1}^{T} \frac{C_i}{(1+TIR)^i} = 0 \quad (3)$$

A Tabela 5 contém os eventos de fluxo de caixa dos empréstimos contidos na base de dados que serão usados para o cálculo da variável dependente TIR utilizada

no modelo de *profit scoring*. O anexo I contém o gráfico com o exemplo de um fluxo de caixa típico de uma operação de crédito realizada na cooperativa.

Tabela 5: Eventos de fluxo de caixa para cálculo da TIR de um empréstimo

Sinal	Evento	Momento	Definição	Cálculo
Negativo	Liberação de financiamento	Início da operação	Saída do caixa da cooperativa para concessão de empréstimo	Valor da operação
	Custos operacionais líquidos	Início da operação	Custos operacionais, descontando-se a tarifa de confecção de cadastro	1% sobre o valor financiado, segundo a cooperativa.
	Custo de alocação de capital	Final de cada mês e a liquidação do contrato	Custo apurado aplicando-se a taxa de referência sobre o saldo médio devedor do mês	CDI do mês sobre o saldo devedor
Positivo	Pagamentos diversos	Data de pagamento	Recebimento de parcelas vencidas ou antecipadas e encargos de atraso	Valor da parcela

Fonte: elaborado pelo autor

4.1. DESENVOLVIMENTO DOS MODELOS

Todos os modelos foram ajustados com o aplicativo R (R CORE TEAM, 2000) O método *random forest* foi implementado com o pacote *randomForest* (SVETNIK; LIAW; TONG, 2003).

A variável dependente do modelo de *credit scoring* para classificação de risco de crédito é identificada pela letra Y e é categorizada com os valores de reposta 0 para clientes “maus pagadores” e 1 para clientes “bons pagadores”. A Tabela 6 contém o resumo das ocorrências da variável Y na base de dados.

Tabela 6: Variável dependente de classificação de risco de crédito - Y

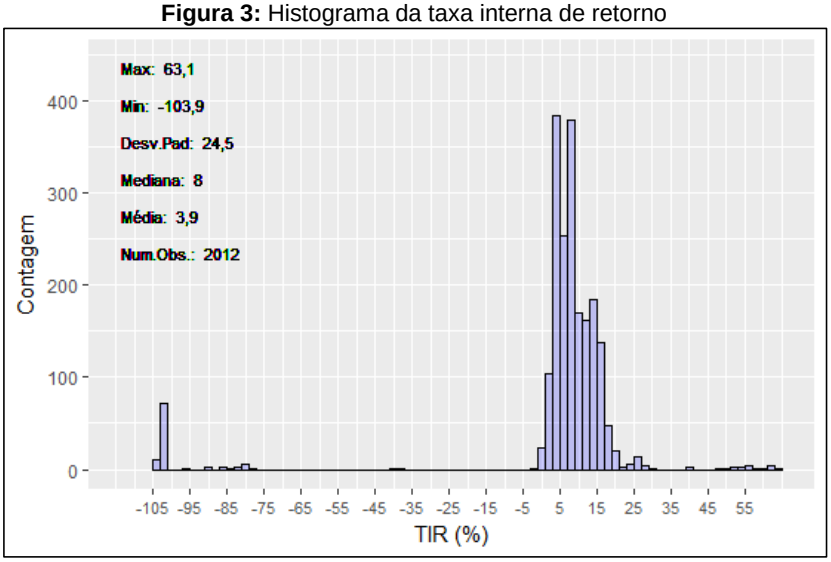
Eventos da variável dependente Y	Quantidade	%
Operações com atraso maior ou igual a 90 dias (MAUS)	174	9%
Operações com atraso menor que 90 dias (BONS)	1838	91%
Quantidade de registro da base de dados	2012	100%

Fonte: elaborado pelo autor

A variável dependente do modelo de *profit score* é a taxa interna de retorno das operações, TIR.

A Figura 3 contém a análise descritiva da variável dependente TIR. Estão representados o histograma com frequências das taxas e as medidas de dispersão. Pode-se observar que, devido aos eventos de *default,* a distribuição é assimétrica. Os eventos de *default* deslocam acentuadamente a cauda esquerda da distribuição, uma vez que esses eventos estão associados a perdas expressivas. Para os valores de TIR positivos pode-se verificar que a distribuição é concentrada no intervalo que compreende as taxas de retorno de 5% até 15%.

Figura 3: Histograma da taxa interna de retorno



Fonte: elaborado pelo autor

Para as variáveis independentes categóricas serão apresentados os cruzamentos destas variáveis com Y e os respectivos testes de homogeneidade.

Para as variáveis independentes contínuas serão apresentadas as medidas de dispersão destas variáveis, a correlação com a variável dependente TIR e os valores médios das variáveis independentes nos níveis da variável dependente Y (bons e maus pagadores).

4.2. ANÁLISE DESCRITIVA DOS DADOS

A Tabela 7 contém o resumo da análise descritiva das variáveis independentes categorizadas. Foram aplicados os testes de homogeneidade qui-quadrado para verificar se as proporções de bons e maus pagadores diferem entre as categorias das variáveis independentes e o teste de comparação de médias (ANOVA) entre os níveis

das variáveis independentes para a TIR. Todos os testes, exceto para a variável GARANTIA, identificaram diferenças significativas com o valor p < 0,01.

Tabela 7: Análise descritiva das variáveis independentes nominais

		Variável dependente Classificação de Risco (Y) *		Variável dependente Taxa Interna de Retorno (TIR) **		
Variável	**Níveis da Variável**	**Qtd Y=0**	**Qtd Y=1**	**Media**	**Mediana**	**Desv. Padrão**
MOD	Outros empréstimos	39	43	-9,44	2,64	56,46
	Crédito pessoal - sem consignação em folha de pagam.	104	1380	4,09	7,64	21,95
	Crédito pessoal - com consignação em folha de pagam.	11	399	8,88	10,26	13,20
	Aquisição de bens – veículos automotores	20	16	-30,56	4,55	52,08
RATING	Rating não A	27	4	-23,82	5,54	63,02
	Rating A	147	1834	4,33	8,00	23,17
GARANTIA	Sem Garantia	157	1733	3,95	7,89	24,06
	Possui Garantia	17	105	3,07	10,83	30,55
RENDA	Até R$ 1.000	10	32	-4,02	8,00	39,91
	De R$ 1.000 até R$3.000	107	1121	4,93	8,25	23,32
	De R$ 3.000 até R$6.000	26	487	4,84	7,16	19,48
	Mais de R$ 6.000	31	198	-2,30	7,73	34,17

** significativa (p<0,01) para o teste qui-quadrado para todas as variáveis*
*** significativa (p<0,01) para o teste F-ANOVA para todas as variáveis exceto GARANTIA*
Fonte: elaborado pelo autor

A variável GARANTIA não apresentou significância estatística para o teste F-ANOVA (p>0,1) que pode ser um indicativo que a sua contribuição para os modelos de *profit scoring* pode não ser significativa.

A Tabela 8 contém o resumo a análise descritiva das variáveis independentes contínuas. Foi aplicado o teste t para a comparação de médias das variáveis independentes entre os níveis da variável Y com diferença significativa (p<0,01). Ao observar-se a correlação da TIR com as variáveis contínuas, verifica-se correlação

com magnitude importante com todas as variáveis exceto endividamento e tempo de relacionamento.

Tabela 8: Análise descritiva das variáveis contínuas

		Medidas de dispersão				Associação com as variáveis dependentes		
Variável	**Definição**	**Media**	**Max**	**Min**	**Desv. Padrão**	**Media (Y=0)**	**Media (Y=1)**	**Correl (TIR)**
TEMPOREL	Tempo de relacionamento em anos	6,25	17,00	0,10	5,27	2,74*	6,58*	0,10**
TAXA	Taxa de juros percentual da operação	28,14	241,66	12,68	12,08	42,71*	26,76*	-0,18**
VALOR	Valor contratado em reais	3912	106500	100	5252	4386	3867	0,01
ENDIV	Razão percentual do valor de parcela sobre a renda mensal	12,48	156,28	0,01	13,24	16,70*	12,08*	-0,03
DLQANT	Quantidade de eventos anteriores de atraso com mais de 30 dias	0,14	16,00	0,00	0,98	0,64*	0,09*	-0,15**
PRAZO	Prazo da operação em dias	515	1837	14	309	545	512	0,02

** significativa (p<0,01) para o teste t de Student*
*** significativa (p<0,01) para a correlação de Pearson*

Fonte: elaborado pelo autor

É importante observar que as variáveis VALOR e PRAZO não apresentaram significância para nenhum dos testes. Para melhorar o resultado do teste t, foram aplicadas transformações nas variáveis conforme descrito na Tabela 9.

Tabela 9: Variáveis transformadas para *credit scoring*

Variável	**Descrição**
LnTEMPOREL	Log do tempo de relacionamento em anos
TAXA	Taxa de juros da operação
lnVALOR	Log do valor contratado
lnENDIV	Log da razão do valor de parcela sobre renda mensal
sqrtDLQANT	Raiz quadrada da quantidade de eventos anteriores de atraso
sqrtPRAZO	Raiz quadrada do prazo da operação em dias

Fonte: elaborado pelo autor

A multicolinearidade ocorre quando as variáveis independentes de um modelo de regressão são altamente correlacionadas. Hair et al. (1998) argumentam que a multicolinearidade pode aumentar sensivelmente o erro tanto na explicação quanto na estimação dos coeficientes do modelo de regressão.

A presença do efeito de multicolinearidade pode ser detectada com a estatística Fator de Inflação da Variância (VIF) que é calculada para cada variável independente.

Na Tabela 10 pode-se observar que a variável lnVALOR possui o valor da estatística VIF elevado, o que pode levar a distorções na explicação do efeito das variáveis independentes nos modelos. Ao se retirar essa variável e realizar um novo teste VIF (coluna “VIF depois”), percebe-se que a mesma possui multicolinearidade com as variáveis sqrtPRAZO e lnENDIV. Os valores da estatística VIF são pequenos para as demais variáveis, não havendo, portanto, outras evidências de multicolinearidade nas variáveis independentes contínuas. Para eliminar os problemas decorrentes deste efeito, decidiu-se retirar a variável ln(VALOR) do conjunto de dados para modelagem.

Tabela 10 : Valores da estatística VIF

Variável	VIF antes	VIF depois
lnTEMPOREL	1,14	1,14
TAXA	1,15	1,15
lnENDIV	2,64	1,16
ln(VALOR)	4,30	
sqrtDLQANT	1,37	1,37
sqrtPRAZO	3,40	1,24

Fonte: elaborado pelo autor

4.3. MODELOS LOGÍSTICOS

Foram gerados 3 modelos com diferentes cenários de análise para a variável dependente Y. O modelo completo contém todas as variáveis selecionadas para o estudo. Os modelos 2 e 3 foram gerados para avaliar se os modelos de *credit scoring* possuem desempenho superior à análise subjetiva baseada em informações de idoneidade fornecidas por *bureaux* de crédito e critérios subjetivos para concessão de empréstimos. Os índices AUROC e Gini foram utilizados para comparar o desempenho dos modelos.

Tabela 11: Modelos de regressão logística

	Modelo 1 (completo)		Modelo 2 (sem a análise subjetiva)		Modelo 3 (contém apenas a análise subjetiva)	
	coef.	*P*	*coef.*	*p*	*coef.*	*p*
(Intercepto)	-1.03	.353	2.61	.006	-1.91	<.001
MOD						
Crédito pessoal - com consignação em folha de pagam.	2.98	<.001	3.00	<.001		
Crédito pessoal - sem consignação em folha de pagam.	2.38	<.001	2.28	<.001		
Outros empréstimos	0.13	.839	-0.06	.927		
RATING	3.88	<.001			4.43	<.001
RENDA						
De R$ 1.000 até R$3.000	1.01	.047	1.01	.039		
De R$ 3.000 até R$6.000	1.39	.015	1.32	.015		
Mais de R$ 6.000	0.12	.837	0.28	.613		
lnTEMPOREL	0.50	<.001	0.48	<.001		
GARANTIA	-0.22	.554	-0.15	.680		
TAXA	-0.07	<.001	-0.07	<.001		
lnENDIV	-0.01	.964	-0.02	.889		
sqrtDLQANT	-1.55	<.001	-1.76	<.001		
sqrtPRAZO	-0.06	<.001	-0.05	<.001		
Observações	2012		2012		2012	
AUROC	0,94		0,93		0,58	
GINI	0,88		0,86		0,15	

Fonte: elaborado pelo autor

A partir dos dados da Tabela 9 observa-se que o modelo completo possui desempenho superior ao modelo 3 e ligeiramente superior, mas muito próximo, ao modelo 2. O modelo 2 possui desempenho superior ao modelo 3, o que sugere que a inclusão de outras variáveis traz benefícios para a classificação de risco em relação ao uso isolado da variável RATING (classificação subjetiva de risco de crédito).

As variáveis GARANTIA, ENDIV, RENDA (acima de R$ 6000) e MOD (outros empréstimos) não apresentaram significância estatística. A variável GARANTIA possui também sinal diferente do esperado. Na análise descritiva, observando-se as operações “ruins” com e sem garantia, nota-se com que a proporção de operações “ruins” com garantia (17 para 105) é maior que a proporção de operações “ruins” sem garantia (157 para 1733). Essa evidência sugere a existência de fragilidades no processo de avaliação de garantias na formalização dos contratos de crédito. Embora a variável ENDIV não tenha apresentado significância estatística, o seu sinal está de

acordo com o efeito esperado descrito na seção de comportamentos esperados. As variáveis RENDA (acima de R$ 6000) e MOD (outros empréstimos) possuem poucos exemplares na amostra (229 e 82 casos respectivamente para um total de 2012 observações), o que que pode ter contribuído para o valor-p não significante. O anexo II contém o gráfico de diagrama de caixa (ou *boxplot*) do valor dos rendimentos dos associados agrupados por categorias de classificação de risco de crédito e renda.

Observando-se os sinais da variável MOD, verifica-se que o produto “Aquisição de veículos” possui risco de crédito maior que os produtos de crédito pessoal e crédito não consignado, comportando-se, portanto, de maneira diferente do previsto na seção de comportamentos esperados.

4.4. COMPARAÇÃO DE MODELAGENS DE *CREDIT SCORING*

Após o desenvolvimento do modelo de regressão logística para avaliação da probabilidade de *default* e avaliação dos fatores relevantes para avaliação de risco crédito, deseja-se verificar se o modelo de *credit scoring* construído com a técnica *random forest* possui desempenho superior à técnica de regressão logística.

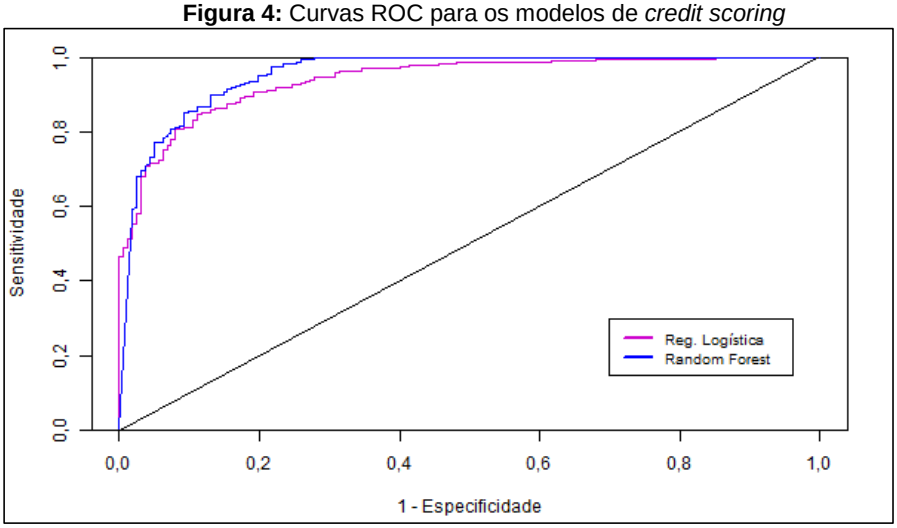
A comparação dos modelos de *credit scoring* desenvolvidos por meio de índices de qualidade empregados para modelos de classificação preditivos calculados na Tabela 12 e as curvas ROC na Figura 4 sugerem que a modelagem *random forest* possui desempenho superior à regressão logística.

Tabela 12: Comparação do desempenho dos modelos de *credit scoring*

Modelos	**AUROC**	**KS**	**Gini**
Regressão Logística (Modelo 1)	0,94	0,69	0,88
Random Forest	0,96	0,90	0,92

Fonte: elaborado pelo autor

Figura 4: Curvas ROC para os modelos de *credit scoring*



Fonte: elaborado pelo autor

4.5. VALIDAÇÃO CRUZADA DOS MODELOS DE *CREDIT SCORING*

Para avaliar a capacidade de generalização dos modelos foi utilizado o método de validação cruzada *k-fold cross validation* (KOHAVI, 1995). Neste método os dados foram aleatoriamente particionados em 15 grupos (k=15) mutualmente exclusivos. Para cada partição foi gerado um modelo a partir dos dados obtidos com a exclusão desta partição, e o modelo foi subsequentemente avaliado usando a partição removida.

Pode-se verificar que os resultados obtidos com a validação cruzada são muito próximos daquele obtido com o modelo desenvolvido com todas as observações da base de dados. De forma geral, o modelo desenvolvido com a técnica *random forest* possui desempenho superior ao modelo desenvolvido com a técnica de regressão logística.

Tabela 13 : Validação cruzada com o método *k-fold* para *credit scoring*

Partição	KS *Random Forest*	KS Reg Logística	Gini *Random Forest*	Gini Reg Logística
1	0,91	0,69	0,91	0,87
2	0,92	0,68	0,90	0,87
3	0,90	0,69	0,92	0,88
4	0,90	0,69	0,91	0,87
5	0,90	0,68	0,92	0,87
6	0,91	0,67	0,91	0,86
7	0,91	0,70	0,91	0,88
8	0,91	0,66	0,91	0,87
9	0,91	0,70	0,92	0,88
10	0,91	0,69	0,92	0,88
11	0,90	0,68	0,91	0,88
12	0,91	0,68	0,91	0,88
13	0,92	0,68	0,91	0,88
14	0,91	0,68	0,93	0,87
15	0,91	0,68	0,92	0,88

Fonte: elaborado pelo autor

4.6. DESENVOLVIMENTO DO MODELO DE *PROFIT SCORE*

O modelo desenvolvido com a técnica de *random forest* foi comparado com a técnica de regressão pelo método dos mínimos quarados ordinários com erro padrão robusto.

Para amenizar problemas decorrentes da falta de linearidade da variável dependente e melhorar o desempenho do modelo de regressão com mínimos quadrados ordinários, serão adicionadas variáveis quadráticas ao conjunto de variáveis contínuas. Estas variáveis estão explicitadas na Tabela 14.

Tabela 14: Variáveis quadráticas adicionais para o modelo de *profit scoring*

Variável	Descrição
TEMPOREL_QUAD	Quadrado do tempo de relacionamento em anos
TAXA_QUAD	Quadrado da taxa de juros da operação
ENDIV_QUAD	Quadrado da razão do valor de parcela sobre renda mensal
sqrtDLQANT	Raiz quadrada da quantidade de eventos anteriores de atraso
sqrtPRAZO	Raiz quadrada do prazo da operação em dias

Fonte: elaborado pelo autor

O conjunto de variáveis categorizadas não sofreu nenhuma alteração quando comparado com o conjunto de variáveis categóricas utilizadas na modelagem de *credit scoring*. Conforme observado na análise descritiva das variáveis categóricas, a

variável GARANTIA não atingiu significância estatística no teste F-ANOVA com a variável dependente TIR, logo, espera-se que essa variável também não tenha significância estatística no modelo de regressão pelo método dos mínimos quadrados ordinários.

A Tabela 15 contém os resultados da modelagem de regressão pelo método dos mínimos quadrados ordinários com as estimativas dos coeficientes e a utilização de erro padrão robusto para controle da heterocedasticidade presente nos dados.

Pode-se verificar que a adição de variáveis quadráticas produziu o ajuste não-linear desejado. Os resultados sugerem que para as variáveis com a forma quadrática adicional, mantido tudo o mais constante, existe um valor crítico para o qual o efeito sobre a TIR passa a ter o sinal alterado, havendo um ponto de TIR mínima ou TIR máxima segundo uma curva quadrática. Para as variáveis categóricas com significância estatística o efeito na variável dependente tem o sinal igual ao que foi observado que no modelo de *credit scoring.*

As variáveis RATING, GARANTIA e RENDA (acima de R$ 6000) não atingiram significância estatística. Esse resultado já era esperado para variável GARANTIA na análise descritiva. A variável RENDA (acima de R$ 6000) possui poucos exemplares na amostra (229 casos), o que que pode ter contribuído para o valor-p não significante. Os resultados da variável RATING sugerem que a avaliação subjetiva de crédito não influencia significativamente a taxa interna de retorno das operações. O sinal da variável PRAZO é diferente do sinal encontrado para esta variável no modelo de *credit scoring*, o que sugere que as operações contratadas com prazo maior oferecem retornos maiores que as operações contratadas com vencimento em curto prazo.

Tabela 15: Regressão linear com erro padrão robusto

		erro padrão robusto		
Termo	**Coeficiente**	**Erro Padrão**	**Z**	**P**
(Intercepto)	-60,230	14,385	-4,187	3,E-05 ***
MOD				
Crédito pessoal - com consignação em folha de pagam.	28,018	6,595	4,248	2,E-05 ***
Crédito pessoal - sem consignação em folha de pagam.	25,976	6,614	3,928	9,E-05 ***
Outros empréstimos	25,734	8,186	3,144	2,E-03 ***
RATING	12,492	10,538	1,185	0,24
RENDA				
De R$ 1.000 até R$3.000	9,560	5,490	1,741	0,08 *
De R$ 3.000 até R$6.000	9,373	5,524	1,697	0,09 *
Mais de R$ 6.000	5,462	5,728	0,953	0,34
lnTEMPOREL	-1,224	0,492	-2,486	0,01 ***
TEMPOREL_QUAD	0,021	0,004	5,394	7,E-08 ***
GARANTIA	3,049	2,898	1,052	0,29
TAXA	0,358	0,166	2,157	0,03 **
TAXA_QUAD	-0,003	0,001	-4,315	2,E-05 ***
lnENDIV	3,471	0,774	4,486	7,E-06 ***
ENDIV_QUAD	-0,002	0,001	-2,021	0,04 **
sqrtDLQANT	-27,718	5,624	-4,929	8,E-07 ***
sqrtPRAZO	0,242	0,085	2,829	5,E-03 ***

** significativa (p<0,1)*
*** significativa (p<0,05)*
**** significativa (p<0,01)*
Fonte: elaborado pelo autor

A técnica *random forest* foi implementada para a previsão da TIR com as mesmas variáveis da regressão linear. A partir da análise dos índices de desempenho dos modelos na Tabela 16 pode-se verificar que o erro quadrático médio (MSE) do modelo *random forest* (MSE=183,29), é consideralvemente menor que obtido pelo modelo com regressão linear (MSE=427,61). O mesmo ocorre observando-se o coeficiente de determinação, o modelo *random forest* possui maior explicação sobre os valores observados que o modelo de regressão linear. A contribuição das variáveis para o modelo *random forest*, calculadas no programa R com base no pacote *randomForest* (SVETNIK; LIAW; TONG, 2003) são apresentadas no anexo III.

Tabela 16: Resultados da modelagem de profit scoring

MSE Random Forest	**MSE Regressão Linear**	**R^2 Random Forest**	**R^2 Regressão Linear**
183,29	427,61	0,70	0,29

Fonte: elaborada pelo autor

4.7. VALIDAÇÃO CRUZADA DOS MODELOS DE *PROFIT SCORING*

Assim como ocorreu com os modelos de *credit scoring*, para avaliar a capacidade de generalização dos modelos de *profit scoring* foi utilizado o método de validação cruzada *k-fold cross validation* (KOHAVI, 1995). Neste método os dados foram aleatoriamente particionados em 15 grupos (k=15) mutualmente exclusivos.

Pode-se verificar na Tabela 17 que os a técnicas possuem pouca oscilação nos resultados. As estatísticas indicam, de um modo geral, a superioridade do método *random forest.*

Tabela 17: Validação cruzada com o método *k-fold* para *profit scoring*

Partição	**MSE *Random Forest***	**MSE *Regressão Linear***	**R^2 *Random Forest***	**R^2 *Regressão Linear***
1	185,00	431,13	0,70	0,29
2	202,41	439,32	0,67	0,28
3	202,17	422,83	0,67	0,30
4	194,33	436,47	0,68	0,28
5	181,62	409,68	0,69	0,30
6	188,47	428,78	0,69	0,29
7	194,52	441,40	0,68	0,28
8	196,77	430,06	0,67	0,28
9	200,33	419,59	0,66	0,28
10	181,78	416,78	0,68	0,27
11	175,70	435,44	0,71	0,29
12	179,35	420,27	0,70	0,29
13	182,83	428,79	0,70	0,29
14	196,09	422,75	0,68	0,30
15	186,52	418,40	0,69	0,30

Fonte: elaborado pelo autor

5. CONCLUSÕES

A análise de crédito é uma atividade de fundamental importância para o objetivo das cooperativas de crédito de fornecer empréstimos com taxas acessíveis para um público que possui dificuldade de acesso ao crédito e se associa às cooperativas para obter e prover ajuda financeira mútua. Um dos requisitos para essa estrutura operar de forma sustentável é a eficiência na decisão de aprovação de crédito, reduzindo a exposição dessas sociedades ao risco de perdas financeiras severas por incumprimento no pagamento de empréstimos.

A análise subjetiva de crédito, prática ainda muito comum nas cooperativas de crédito brasileiras, possui riscos tanto de natureza operacional quanto de natureza moral para estas sociedades, uma vez que concessões de empréstimos podem ser facilitadas para clientes com risco de crédito elevado. Conflitos de agência podem surgir durante a análise subjetiva de crédito entre o solicitante do empréstimo, que também é sócio da cooperativa, e o agente contratado pela diretoria eleita pelos sócios para analisar a viabilidade das solicitações de crédito.

A substituição da análise subjetiva pela análise objetiva de crédito com modelos determinísticos pode ser benéfica para as cooperativas de crédito, pois permite que os resultados da análise de crédito sejam sempre consistentes com as variáveis de entrada, tornando o processo imparcial e com critérios claros para a decisão de aprovação de crédito. Uma outra característica da análise objetiva de crédito desejável para cooperativas de crédito é a gestão das perdas financeiras com o estabelecimento de uma taxa de corte para o nível de risco máximo aceitável para novas operações. A escolha das variáveis independentes e relacionadas com o risco de crédito é condição básica para que a análise objetiva tenha resultados satisfatórios.

Neste estudo verificou-se se as variáveis presentes na base de dados de uma cooperativa de crédito possuem efeito positivo ou negativo sobre a probabilidade de *default* de acordo com estudos anteriores na área de risco de crédito. A partir da análise dos coeficientes do modelo de regressão logística construído com a base de dados, verificou-se que as seguintes variáveis apresentaram significância estatística ($p<0,05$): modalidade da operação, *rating* de julgamento subjetivo para risco de crédito, renda, tempo de relacionamento, taxa de juros da operação, histórico de inadimplência e o prazo da operação. Estas variáveis possuem o sinal de acordo com o efeito esperado, com exceção da variável modalidade da operação onde o

financiamento de compra de veículos demonstrou-se mais arriscado que operações de empréstimo pessoal e empréstimo consignado. Uma possível causa para esse efeito é uma restrição da própria base de dados. Os dados coletados refletem operações contratadas e finalizadas no intervalo de dois anos (indo de 2015 até 2016). Como o produto de financiamento de veículos possui prazos mais longos, seria necessário utilizar uma base de dados com maior intervalo de duração para confirmar o efeito desta variável.

Modelos de *credit scoring* e *profit scoring* permitem à gestão das cooperativas selecionar os clientes com maior potencial para pagamento das operações, assim como verificar se a lucratividade das operações está adequada aos objetivos de sustentabilidade da instituição e oferta de empréstimos e aplicações financeiras com taxas atrativas para os sócios.

Para a modelagem de *credit scoring* foram utilizadas a técnica referencial de regressão logística e a técnica *random forests* pertencente à família de algoritmos de *machine learning*. Para a modelagem de *profit scoring*, foi construído um modelo baseado na técnica de regressão com *random forests* que foi comparado com um modelo de regressão linear com erro padrão robusto.

Ao comparar-se os valores dos índices Gini e AUROC, os resultados sugerem que os modelos de *credit scoring* possuem qualidade preditiva muito superior à análise subjetiva. Esta evidência favorece o argumento presente na literatura que a análise subjetiva de crédito possui desempenho inferior à análise objetiva realizada com modelos de *credit scoring*.

Comparando-se o desempenho de técnicas de modelagem de *credit scoring*, o modelo de desenvolvido com a técnica *random forests* possui desempenho superior ao modelo desenvolvido à técnica de regressão logística. Com o emprego desta técnica espera-se uma melhoria importante nos resultados financeiros da cooperativa, dado que as perdas financeiras evitadas não se transformarão em custos adicionais para a organização. As economias geradas ajudam a reduzir as taxas de financiamento, melhorando a oferta de crédito contribuindo para os objetivos da cooperativa de crédito.

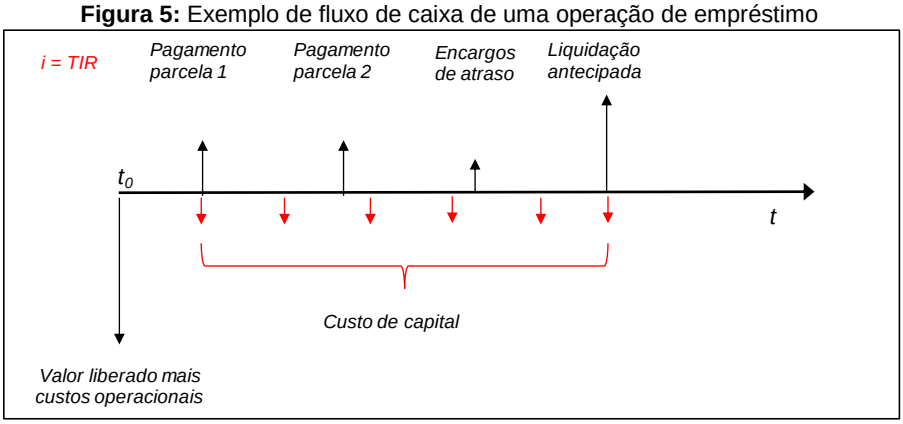
Para futuras pesquisas, sugere-se a utilização de outras técnicas de *machine learning* populares tais quais redes neurais e máquinas de vetores de suporte para modelagem de *credit scoring* em cooperativas de crédito. Para *profit scoring* sugere-se análise de lucratividade com utilizando como variável dependente o retorno

ajustado ao risco no capital (RAROC). A técnica de análise de sobrevivência pode ser utilizada para avaliar se é possível prever lucros ou perdas em eventos de atraso de pagamento e liquidação antecipada de contratos de crédito.

6. ANEXOS

6.1. ANEXO I - FLUXO DE CAIXA DE UMA OPERAÇÃO DE EMPRÉSTIMO

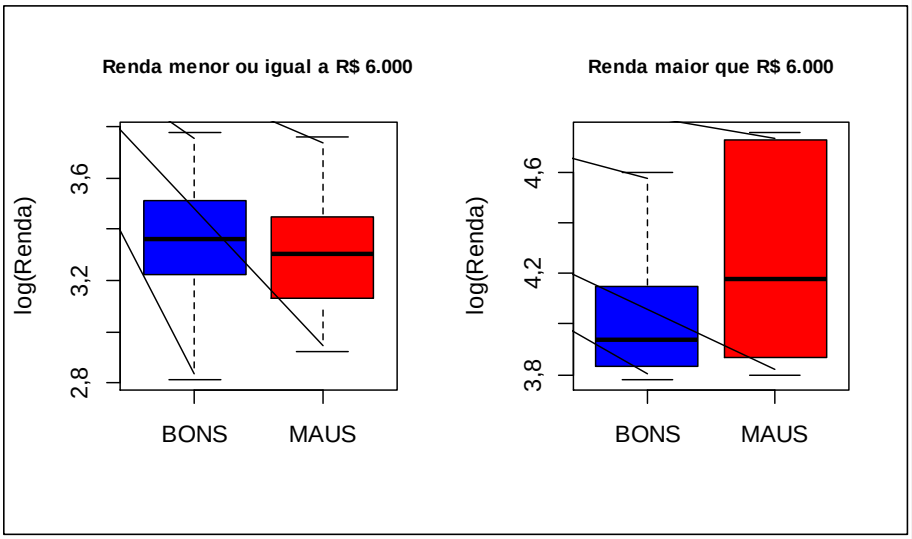
Figura 5: Exemplo de fluxo de caixa de uma operação de empréstimo



Fonte: elaborado pelo autor

6.2. ANEXO II – BOXPLOT DA VARIÁVEL RENDA

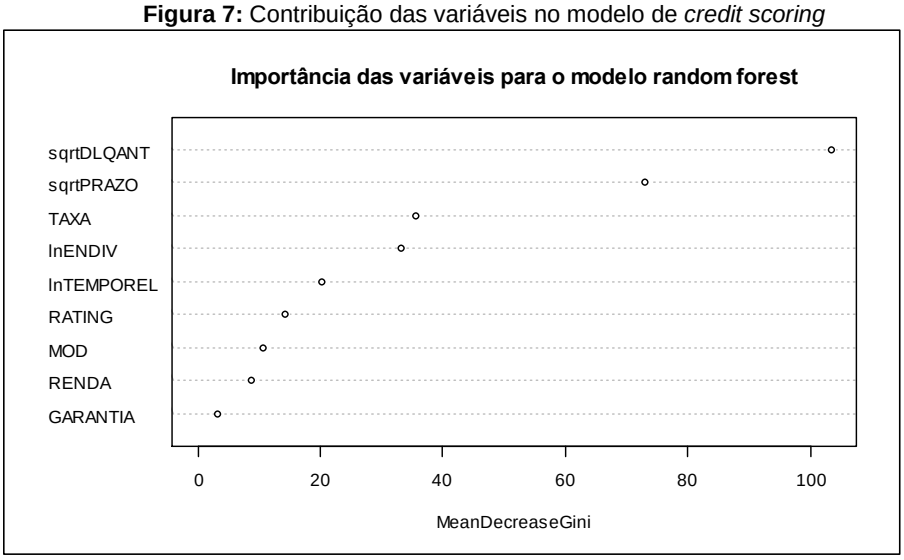
Figura 6: Qualidade do crédito cedido em função da renda



Fonte: elaborado pelo autor

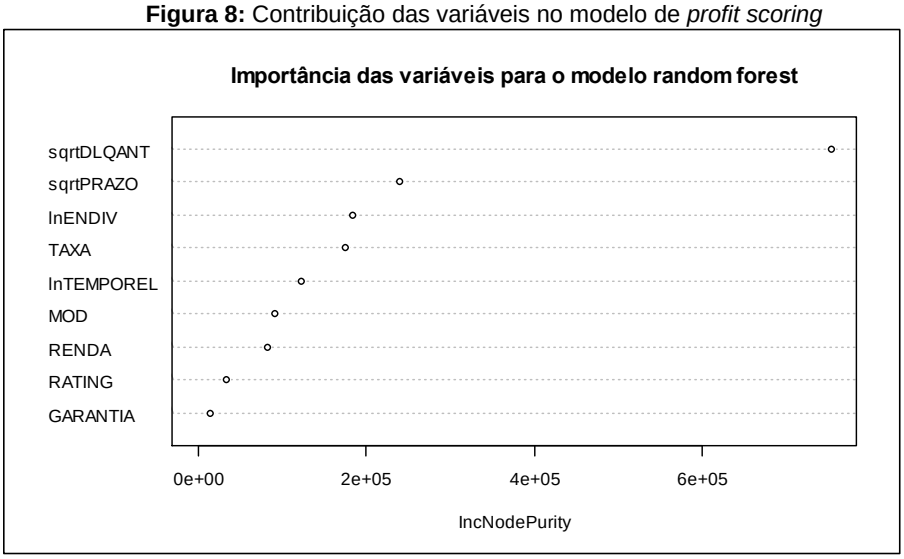
6.3. ANEXO III – CONTRIBUIÇÃO DAS VARIÁVEIS PARA OS MODELOS DE*CREDIT SCORING* E *PROFIT SCORING*

Figura 7: Contribuição das variáveis no modelo de *credit scoring*



Fonte: elaborado pelo autor

Figura 8: Contribuição das variáveis no modelo de *profit scoring*



Fonte: elaborado pelo autor

7. REFERÊNCIAS

ARAÚJO, E. A.; MONTREUIL CARMONA, C. U. DE. Desenvolvimento de Modelos Credit Scoring com Abordagem de Regressão Logística para a Gestão da Inadimplência de uma Instituição de Microcrédito. **Contabilidade Vista & Revista - UFMG**, 2007.

AVERY, R. B.; BOSTIC, R. W.; CALEM, P. S. Credit Scoring: Statistical Issues and Evidence from Credit-Bureau Files. **Real Estate Economics**, v. 28, n. 3, p. 523–547, set. 2000.

BLACK, S. E.; MORGAN, D. P. **Risk and the democratization of credit cards**. [s.l.] Federal Reserve Bank of New York New York, 1998.

BREALEY, R. A.; MYERS, S. C.; ALLEN, F. **Principles of corporate finance**. [s.l.] Tata McGraw-Hill Education, 2012.

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

BREIMAN, L. Random forests. **Machine learning**, v. 45, n. 1, p. 5–32, 2001.

BREIMAN, L.; FRIEDMAN, J.; STONE, C. J. **Classification and regression trees**. [s.l.] CRC press, 1984.

CHAKRAVARTY, S.; SCOTT, J. S. Relationship and rationing in the consumer loan market. **Relationship and rationing in the consumer loan market**, 1998.

CORNFORTH, C. The Governance of cooperatives and mutual associations: a paradox perspective. **Annals of Public and Cooperative Economics**, v. 75, n. 1, p. 11–32, mar. 2004.

CROOK, J. N.; EDELMAN, D. B.; THOMAS, L. C. Recent developments in consumer credit risk assessment. **European Journal of Operational Research**, v. 183, n. 3, p. 1447–1465, 2007.

CUEVAS, C. E.; FISCHER, K. P. **Cooperative financial institutions: issues in governance, regulation, and supervision**. Washington: The World Bank, 2006.

DINH, T. H. T.; KLEIMEIER, S. A credit scoring model for Vietnam's retail banking

market. **International Review of Financial Analysis**, v. 16, n. 5, p. 471–495, 2007.

DURAND, D. Risk Elements in Consumer Instalment Financing. **National Bureau of Economic Research**, 1941.

EISENHARDT, K. M. Agency Theory: An Assessment and Review. **The Academy of Management Review**, v. 14, n. 1, p. 57–74, 1989.

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

FORSYTH, R. **Machine learning systems**. Aslib proceedings. **Anais**...MCB UP Ltd, 1984

GERIZ, S. D. As cooperativas de crédito no arcabouço institucional do sistema financeiro nacional. **Prima Facie-Direito, História e Política**, v. Vol 3, n. N. 4, p. 82–110, 2010.

GIAROLA, E.; SANTOS, A. C. DOS; FERREIRA, RO. D. N. Conflitos de interesses em cooperativas de crédito: uma análise sob a ótica da Social Network Analysis. **XXXIII Encontro da ANPAD**, p. 1–16, 2009.

GITMAN, L. J. **Princípios de administração financeira**. [s.l.] Harbra, 1997.

HAIR, J. F.; BLACK, W. C.; ANDERSON, R. E. **Multivariate data analysis**. [s.l.] Prentice hall Upper Saddle River, NJ, 1998. v. 5

HART, O.; MOORE, J. **Cooperatives vs. outside ownershipNational Bureau of Economic Research**, 1998.

HOSMER JR, DAVID W LEMESHOW, S.; STURDIVANT, R. X. **Applied logistic regression**. [s.l.] John Wiley & Sons, 2013.

HYNES, R. M. Three essays on consumer bankruptcy and exemptions. 1998.

KOHAVI, R. **A study of cross-validation and bootstrap for accuracy estimation and model selection**. Ijcai. **Anais**...Stanford, CA, 1995

LAWRENCE, E. C.; ELLIEHAUSEN, G. A comparative analysis of payday loan customers. **Contemporary Economic Policy**, v. 26, n. 2, p. 299–316, 2008.

LESSMANN, S.; BAESENS, B.; THOMAS, L. C. Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. **European Journal of Operational Research**, v. 247, n. 1, p. 124–136, 2015.

LEWIS, E. **Introduction to Credit Scoring**. San Rafael: The Athena Press, 1992.

LIMA, R. E. DE; ARAÚJO, M. B. V. DE; AMARAL, H. F. Conflitos de Agência: um etudo comparativo dos aspectos inerentes a empresas tradicionais e cooperativas de crédito. **RCO - Revista de Contabilidade e Organizações**, v. 2, n. 4, p. 148–157, 2008.

MALHOTRA, R.; MALHOTRA, D. K. Evaluating consumer loans using neural networks. **Omega**, v. 31, p. 83–96, 2003.

PINHEIRO, M. A. H. **Cooperativas de Crédito História da evolução normativa no Brasil**. Brasília: BCB, 2008.

POLÔNIO, W. A. **Manual das Sociedades Cooperativas**. São Paulo: Atlas, 1999.

SANCHEZ-BARRIOS, L. J.; ANDREEVA, G.; ANSELL, J. “Time-To-Profit Scorecards for Revolving Credit”. **European Journal of Operational Research**, v. 249, n. 2, p. 397–406, 2016.

SANTOS, J. O. DOS; FAMÁ, R. Avaliação da aplicabilidade de um modelo de credit scoring com varíaveis sistêmicas e não-sistêmicas em carteiras de crédito bancário rotativo de pessoas físicas. **Revista Contabilidade & Finanças**, v. 18, p. 105–117, 2007.

SCHEBESCH, K. B.; STECKING, R. Support vector machines for classifying and describing credit applicants: Detecting typical and critical regions. **Journal of the Operational Research Society**, v. 56, p. 1082–1088, 2005.

SECURATO, J. R. **Crédito: análise e avaliação do risco**. São Paulo: Saint Paul, 2002.

SERRANO-CINCA, C.; GUTIÉRREZ-NIETO, B. The use of profit scoring as an alternative to credit scoring systems in peer-to-peer (P2P) lending. **Decision Support Systems**, v. 89, p. 113–122, set. 2016.

SILVA FILHO, G. T. DA. Avaliação de desempenho em cooperativas de crédito: uma aplicação do modelo de gestão econômico -GECON. **Organizações Rurais & Agroindustriais**, v. 4, n. 1, 2002.

SILVERIO, M. **APLICAÇÃO DE ALGORITMOS DE APRENDIZADO DE MÁQUINA NO DESENVOLVIMENTO DE MODELOS DE ESCORE DE CRÉDITO**. [s.l: s.n.].

SO, M. C.; THOMAS, L. C.; SEOW, H.-V. Using a transactor/revolver scorecard to make credit and pricing decisions. **Decision Support Systems**, v. 59, p. 143–151, 2014.

SOARES, M. M. S.; SOBRINHO, A. D. DE M. **Microfinanças O Papel do Banco Central do Brasil e a Importância do Cooperativismo de Crédito**. Brasília: Banco Central do Brasil, 2008.

SOUZA, F. A. P. DE. **Competição entre cooperativas de crédito e bancos em mercados locais**, 2017.

SRINIVISAN, V.; KIM, Y. H. Credit granting a comparative analysis of classificatory procedures. **Journal of Finance**, v. 42, p. 655–683, 1987.

SVETNIK, V.; LIAW, A.; TONG, C. Random forest: a classification and regression tool for compound classification and QSAR modeling. **Journal of chemical information and computer sciences**, v. 43, n. 6, p. 1947–1958, 2003.

TEAM, R. C. R language definition. **Vienna, Austria: R foundation for statistical computing**, 2000.

TECH, V. Credit-scoring models in the credit-onion environment using neural networks and genetic algorithms. **IMA Journal of Management Mathematics**, p. 323–346, 1997.

THOMAS, L. C.; EDELMAN, D. B.; CROOK, J. N. **Credit Scoring and its Applications**. Philadelphia: SIAM, 2002.

VERBRAKEN, T.; BRAVO, C.; BAESENS, B. Development and application of consumer credit scoring models using profit-based classification measures. **European Journal of Operational Research**, v. 238, n. 2, p. 505–513, 2014.

WEST, D. Neural network credit scoring models. **Computers and Operational Research**, v. 27, p. 1131–1152, 2000.

YAP, B. W.; ONG, S. H.; HUSAIN, N. H. M. Using data mining to improve assessment of credit worthiness via credit scoring models. **Expert Systems with Applications**, v. 38, n. 10, p. 13274–13283, 2011.