UNIVERSIDADE DE BRASÍLIA
FACULDADE DE ADMINISTRAÇÃO, ECONOMIA, CONTABILIDADE E GESTÃO PÚBLICA – FACE
PROGRAMA DE PÓS-GRADUAÇÃO EM ADMINISTRAÇÃO – PPGA

JOSÉ RÔMULO DE CASTRO VIEIRA

PREDIÇÃO DO BOM E DO MAU PAGADOR NO PROGRAMA MINHA CASA, MINHA VIDA

Brasília – DF
2016

JOSÉ RÔMULO DE CASTRO VIEIRA

PREDIÇÃO DO BOM E DO MAU PAGADOR NO PROGRAMA MINHA CASA, MINHA VIDA

Dissertação de Mestrado submetida ao Programa de Pós-Graduação em Administração (PPGA) como requisito parcial para obtenção do grau de Mestre em Administração, na área de finanças, sob a orientação do Prof. Tit. Herbert Kimura.

Brasília – DF
2016

FOLHA DE APROVAÇÃO

Autor: JOSÉ RÔMULO DE CASTRO VIEIRA

Titulo: PREDIÇÃO DO BOM E DO MAU PAGADOR NO PROGRAMA MINHA CASA, MINHA VIDA

BANCA EXAMINADORA

Prof. Tit. Herbert Kimura _____
Orientador

Prof. Dr. Vinicius Amorim Sobreiro _____
Membro interno

Prof. Dr. Fabiano Guasti Lima _____
Membro externo

FICHA CATALOGRÁFICA

VIEIRA, JOSÉ RÔMULO DE CASTRO
PREDIÇÃO DO BOM E DO MAU PAGADOR NO PROGRAMA MINHA CASA, MINHA VIDA [Distrito Federal] 2016.
xvi, 88 p., 210 x 297 mm (PPGA/FACE/UnB, Mestre, Administração, 2016).
Dissertação de Mestrado - Universidade de Brasília, Faculdade de Administração, Economia, Contabilidade e Gestão Pública - FACE
Programa de Pós-Graduação em Administração - PPGA

1. Habitação pública.
2. Minha Casa, Minha Vida.
3. Risco de crédito.
4. Medidas prudenciais.

I. PPGA/FACE/UnB

REFERÊNCIA BIBLIOGRÁFICA

VIEIRA, J.R.C. (2016). *PREDIÇÃO DO BOM E DO MAU PAGADOR NO PROGRAMA MINHA CASA, MINHA VIDA*. Dissertação de Mestrado, Programa de Pós-Graduação em Administração - PPGA, Universidade de Brasília, Brasília, DF, 88 p.

CESSÃO DE DIREITOS

AUTOR: JOSÉ RÔMULO DE CASTRO VIEIRA
TÍTULO: PREDIÇÃO DO BOM E DO MAU PAGADOR NO PROGRAMA MINHA CASA, MINHA VIDA.
GRAU: Mestre em Administração ANO: 2016

JOSÉ RÔMULO DE CASTRO VIEIRA
Programa de Pós-Graduação em Administração - PPGA
Faculdade de Administração, Economia, Contabilidade e
Gestão Pública - FACE
Universidade de Brasília (UnB)
Campus Darcy Ribeiro
CEP 70919-970 - Brasília - DF - Brasil

DEDICATÓRIA

À minha família, amigos, namorada, colegas de trabalho e orientador pelo apoio, força, incentivo, companheirismo e amizade. Sem eles, nada disso seria possível. À minha filha que, inconscientemente, motiva-me a ser uma pessoa melhor todos os dias.

JOSÉ RÔMULO DE CASTRO VIEIRA

Agradecimentos

Agradeço a meu orientador, Herbert Kimura, pela sua oportuna orientação e por ter me guiado durante todo o período do mestrado, além de sua dedicação, competência e especial atenção nas revisões, fatores fundamentais para a conclusão deste trabalho.
Agradeço aos professores Vinicius Amorim Sobreiro e Flávio Luiz de Moraes Barboza, pela colaboração e sugestões para um melhor aperfeiçoamento desta dissertação e a todos os professores do mestrado que, de alguma forma, contribuíram com a minha formação.
Ao meu colega e amigo Leonardo Rangel por disponibilizar seu computador para o processamento dos dados.
Agradeço ainda à Caixa Econômica Federal e a meu gestor, Jucemar José Imperatori, pela compreensão necessária para realização deste trabalho.

JOSÉ RÔMULO DE CASTRO VIEIRA

RESUMO

Este trabalho tem como objetivo principal implementar diferentes modelos de previsão da inadimplência, a partir de métodos de *credit scoring* e técnicas computacionais com algoritmos de *Machine Learning* (Análise discriminante, regressão logística, *Decision Tree*, *Random Forest*, *Bootstrap Aggregating* e *Adaptive Boosting*) e comparar a adequação dos modelos de previsão da inadimplência que melhor identifiquem o bom e o mau pagador no Programa Minha Casa, Minha Vida. Para avaliar a adequação dos modelos de *Machine Learning*, foram realizados três testes com a obtenção dos índices *Area Under ROC Curve* (AUROC), *Kolmogorov–Smirnov* (KS) e *BRIER Score* com o intuito de validar os modelos em diferentes intervalos de tempo para variável dependente *default* (30, 60, 90, 120 dias), validar os modelos, considerando um número menor de observações (300.000) e validar os modelos sem o uso de variáveis discriminatórias (gênero, idade e estado civil). Verifica-se que a capacidade de predição dos modelos melhorou, à medida que o número de dias de atrasos utilizados para definir a variável *default*, aumentava. Os melhores resultados foram obtidos com *Bootstrap Aggregating* (*Bagging*), *Random Forest* (RF) e *Adaptive Boosting* (*AdaBoost*). Observa-se um impacto negativo considerável nos resultados quando utilizado um número menor de observações. Verificou-se também que a retirada de variáveis discriminatórias dos modelos preserva o poder discriminatório do sistema de classificação de risco de crédito. Aplicando o algoritmo *Bagging* no Programa Minha Casa, Minha Vida (PMCMV) a taxa de inadimplência que é de 11,80 % poderia ser reduzida para 2,95 %. Logo, 197.905 mil contratos inadimplentes deixariam de existir no PMCMV resultando em uma redução nas perdas com inadimplência de aproximadamente R$ 9,8 bilhões.

PALAVRAS-CHAVE: Habitação pública. Minha Casa, Minha Vida, Risco de crédito. Medidas prudenciais.

ABSTRACT

The main objective of this work is to implement different models of forecasting of default, from credit scoring methods and computational techniques with Machine Learning algorithms (discriminant analysis, logistic regression, decision tree, random forest, bootstrap aggregating and adaptive boosting) and compare The adequacy of the default models that best identify the good and the bad payer in the "Programa Minha Casa, Minha Vida"(PMCMV). In order to evaluate the suitability of the Machine Learning models, three tests were carried out to obtain the Area Under ROC curve (AUROC), Kolmogorov-Smirnov (KS) and BRIER Score indices with the aim of validating the models at different time intervals for variable (30, 60, 90, 120 days), validate the models, considering a smaller number of observations (300,000) and validate the models without the use of discriminatory variables (gender, age and marital status). It is verified that the prediction capacity of the models improved, as the number of days of delays used to define the default variable increased. The best results were obtained with bootstrap aggregating (Bagging), random forest (RF) and adaptive boosting (AdaBoost). A considerable negative impact on results is observed when a smaller number of observations are used. It was also found that the removal of discriminatory variables from the models preserves the discriminatory power of the credit risk classification system. Applying the Bagging algorithm in the "Programa Minha Casa, Minha Vida"(PMCMV) program, the default rate of 11.80% could be reduced to 2.95%. Therefore, 197,905 thousand defaulted contracts would cease to exist in the PMCMV resulting in a reduction in losses with delinquencies of approximately 9.8 billion of real.

KEY WORDS: Public housing. Minha Casa, Minha Vida, Credit risk. Prudential arrangements.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS

AdaBoost *Adaptive Boosting*.

Bagging *Bootstrap Aggregating*.

Boosting *Boosting*.

ANN *Artificial Neural Network*.

AUROC *Area Under ROC Curve*.

BACEN Banco Central do Brasil.

BCBS *Basel Committee on Banking Supervision*.

BNH Banco Nacional de Habitação.

BNN *Backpropagation Neural Network*.

CART *Classification and Regression Tree*.

CBR *Case-Based Reasoning*.

CNH Rede de Habitação Cleveland.

DT *Decision Tree*.

ECOA *Equal Credit Opportunity Act*.

FAR Fundo de Arrendamento Residencial.

FDS Fundo de Desenvolvimento Social.

FGHab Fundo Garantidor da Habitação Popular.

FGTS Fundo de Garantia por Tempo de Serviço.

FIESP Federação das Indústrias do Estado de São Paulo.

FJP Fundação João Pinheiro.

GP *Genetic Programming*.

HUD Departamento de Habitação e Desenvolvimento Urbano.

IBGE Instituto Brasileiro de Geografia e Estatística.

KS *Kolmogorov–Smirnov*.

LDA *Linear Discriminant Analysis*.

LIHTC *Low-Income Housing Tax Credit*.

LR *Logistic Regression*.

MBS Mortgage Títulos.

MDA Análise Discriminante Múltipla.

ML *Machine Learning*.

PMCMV Programa Minha Casa, Minha Vida.

PNAD Pesquisa Nacional por Amostra de Domicílios.

PNHR Programa Nacional de Habitação Rural.

RBF *Radial Basis Function kernel*.

RF *Random Forest*.

ROC *Receiver Operating Characteristi*c.

SBPE Sistema Brasileiro de Poupança e Empréstimo.

SFI Sistema Financeiro Imobiliário.

SVM *Support Vector Machines*.

UF Unidade da Federação.

INTRODUÇÃO

O crédito tem uma importância fundamental para o desenvolvimento da economia, uma vez que provê os recursos financeiros para que consumidores possam realizar seus projetos e adquirir bens com uma quantidade de dinheiro emprestado por uma instituição financeira e que deve ser reembolsado, com juros e em parcelas, proporcionando, além de um crescimento econômico, a melhoria na qualidade de vida das pessoas (GERTLER; KARADI, 2015). Conhecer e entender a dinâmica do crédito realizado por instituições públicas nos despertou o desejo de pesquisar e apontar os possíveis efeitos gerados na economia, dada a abrangência deste tema que implica uma problemática atual no processo de aquisição de bens, ou melhor, aquisição de uma habitação.

Esta dissertação é resultado de uma pesquisa no campo da administração voltada para finanças e risco de crédito na habitação pública, que nos permitiu coletar, organizar, selecionar e analisar um *corpus* composto de habitação pública, risco de crédito, métodos de *credit scoring* com técnicas tradicionais e com *Machine Learning*, abstraindo-se os autores, suas individualizações e resguardados os princípios éticos.

Para abordar o tema proposto, tendo como finalidade investigar, analisar e "decifrar", por meio da pesquisa de tipo documental, e dentro das limitações deste trabalho, os mecanismos associados à administração, na área de finanças, teve-se o cuidado de levantar, a partir do referencial teórico e análise dos modelos utilizados, as seguintes indagações:

1. Que medidas prudenciais poderiam ser adotadas para diminuir o número de inadimplentes no Programa Minha Casa, Minha Vida?
2. Qual o melhor algoritmo de *Machine Learning* para avaliar o risco de crédito no PMCMV?

Após o levantamento dessas indagações que nortearam a investigação, centrada em modelos *Machine Learning* , direciona-se o olhar para as categorias temáticas de análise pretendida, na seguinte ordem:

1. Habitação pública e risco de crédito;
2. Estudo teórico – risco de crédito e medidas prudenciais para o programa de habitação pública do Brasil; e
3. Estudo empírico – métodos de aprendizagem de máquina para avaliação de risco de crédito no programa de habitação pública do Brasil.

A caracterização deste texto dissertativo coloca como necessária a escolha de um referencial teórico-metodológico que possibilite a explicação satisfatória das características intrínsecas do

objeto em estudo – risco de crédito na habitação pública – a partir da análise e investigação dos *corpus* habitação pública, risco de crédito, métodos de *credit scoring* com técnicas tradicionais e com *Machine Learning*. Em função disso, opta-se pela preferência de autores da Administração, (WANG et al., 2012; TSUKAHARA et al., 2016; BROWN; MUES, 2012; ALTMAN; SAUNDERS, 1998) entre outros, que, sem dúvida, contribuíram, expressivamente, para a compreensão e análise crítica deste trabalho.

O objetivo geral da pesquisa foi implementar diferentes modelos de previsão da inadimplência, a partir de métodos de *credit scoring* e técnicas computacionais com algoritmos de *Machine Learning* (análise discriminante, regressão logística, *Decision Tree* (DT), RF, *Bagging* e *AdaBoost*) e comparar a adequação dos modelos de previsão da inadimplência que melhor identifiquem o bom e o mau pagador no PMCMV.

Dentre os objetivos específicos, destacam-se os seguintes: a) pesquisar sobre habitação popular no Brasil; b) focalizar e compreender as categorias capitais que constituem o sistema financeiro de habitação popular; c) analisar o resultado da pesquisa; d) avaliar a adequação dos modelos de *Machine Learning*.

Aborda-se, aqui, o tema "Predição do bom e do mau pagador no programa Minha casa, Minha vida", porque acredita-se que esse assunto é muito instigante para os estudos da Administração, mais precisamente para área de finanças, e por levar-se em consideração que no Brasil, as operações de crédito para pessoas físicas apresentaram um acelerado crescimento nos últimos anos. De acordo com dados do Banco Central do Brasil, o saldo das operações de crédito, no período de 2011 e 2014, tiveram um crescimento médio de 15,3 % totalizando em dezembro de 2014 um saldo de R$ 1.412 bilhão. Destacam-se, nessas operações de crédito, os recursos direcionados ao setor imobiliário, que tiveram um crescimento médio de 30% entre 2012 e 2014, com o saldo total indo de R$ 255 bilhões (2012) para R$ 431 bilhões (2014) (BACEN, 2014).

E ainda, porque a baixa renda ou ausência de renda foi um dos principais critérios para seleção dos beneficiários do PMCMV fazendo com que essas operações não fosse submetidas a qualquer tipo de análise de risco de crédito, conduzindo o programa para um elevado número de contratos inadimplentes. Segundo os dados de Brasil (2016), um quarto dos contratos do PMCMV faixa 1 está há mais de 90 dias em atraso. Avaliações e critérios subjetivos para conceder o crédito foram bastante utilizadas no passado, porém o risco de crédito pode ser melhor controlado com o uso de instrumentos estatísticos e de sistemas multivariados, possibilitando a mensuração do risco de forma mais objetiva e com uma abordagem empírica que enfatiza a predição (ALTMAN; SAUNDERS, 1998).

Assim sendo, é necessário explicar, neste ponto, os conceitos que orientam a nossa investigação, a análise e a estrutura do trabalho, que foi organizado em três capítulos. Cada um deles propõe apresentar assuntos relevantes à compreensão do tema em estudo, isto é, descrevendo sobre

habitação pública; estudo teórico: risco de crédito e medidas prudenciais para o programa de habitação pública do Brasil; estudo empírico: métodos de aprendizagem de máquina para avaliação de risco de crédito no programa de habitação pública do Brasil.

No capítulo 1 (um) – Descrevendo sobre habitação pública – destaca-se o algumas questões sobre o crédito para habitação popular, a habitação popular e o seu impacto social, a inadimplência no financiamento à habitação, o Programa Minha Casa, Minha Vida, a análises de risco de crédito e os métodos de aprendizagem de máquina.

No capítulo 2 (dois) – Estudo teórico: risco de crédito e medidas prudenciais para o programa de habitação pública do Brasil – faz-se um estudo com o objetivo de propor medidas prudenciais para concessão do crédito para aquisição de unidades habitacionais, como por exemplo, uma proposta de análise de risco de crédito e medidas que visam o aperfeiçoamento dos programas de habitação pública, com a implementação de outros incentivos, que não o fornecimento da propriedade da unidade habitacional, como evidenciado por Popkin (2004) sobre o HOPE VI, programa habitacional americano que combina subsídios para a revitalização física com o financiamento de melhorias de gestão, serviços de apoio e o uso de *voucher* de habitação para aluguel de moradias no mercado privado.

No capítulo 3 (três) – Estudo empírico: métodos de aprendizagem de máquina para avaliação de risco de crédito no programa de habitação pública do Brasil – elucida-se a nossa compreensão sobre a implementação de diferentes modelos de previsão da inadimplência, a partir de métodos de *credit scoring* e técnicas computacionais com algoritmos de aprendizagem de máquina (análise discriminante, regressão logística, DT, RF, *Bagging* e *AdaBoost*) com o intuito de comparar a adequação dos modelos de previsão da inadimplência que melhor identifiquem o bom e o mau pagador no contexto do PMCMV.

Já, nas Considerações finais, tenta-se elaborar um diagnóstico do presente e colaborar com a necessária compreensão dos interessados que estejam buscando uma interpretação dos conhecimentos da Administração com vistas na área de finanças, para que este estudo possa ser visto e analisado, também, como um lugar de troca de experiência, realização e crescimento pessoal/profissional.

Assim, espera-se que esta pesquisa possa, de algum modo, contribuir para o aprofundamento de temas que despertem, atualmente, grande interesse na área de administração voltada para finanças e risco de crédito, e que este estudo nos apresente como sendo um instrumento relevante de interpretação da realidade social e econômica que permeia as relações humanas.

1 HABITAÇÃO PÚBLICA E RISCO DE CRÉDITO

Neste capítulo, destaca-se a importância do crédito para habitação popular, visto que a habitação foi reconhecida como direito humano em 1948, com a Declaração Universal dos Direitos Humanos, tornando-se um direito inalienável, universal em todas as partes do mundo e essencial para a vida das pessoas. Portanto, será descrito sobre: crédito para habitação popular; habitação popular e o seu impacto social; inadimplência no financiamento à habitação; o PMCMV; análises de risco de crédito, análises multivariadas tradicionais, análise discriminante; regressão logística; e métodos de aprendizagem de máquina

1.1 CRÉDITO PARA HABITAÇÃO POPULAR

Cabe, neste ponto, abrir parênteses e explicar a definição de habitação que, conforme ABIKO (1995) "tem a função de abrigo". Pode ser definida como uma necessidade fundamental, básica do ser humano, uma vez que a moradia é vista como um dos principais investimentos para a constituição de um patrimônio. Outrossim, está ligada, convencionalmente, ao sucesso econômico e *status* do indivíduo.

Desse modo, a habitação é considerada um ativo caro e que pode levar diversos anos até que sua amortização seja realizada por completo. A disponibilidade e o custo do financiamento habitacional são, então, determinantes críticos da forma como os mercados de habitação funcionam e, possivelmente, os mecanismos de financiamento são fatores que explicam as mudanças no mercado. Logo, este tópico dedica-se aos estudos do crédito para a habitação popular e as consequências até então observadas.

Stegman (1991) abordou questões dos custos excessivos de alternativas de financiamento (empréstimos dos vendedores, empréstimos assumidos com todos os atos de confiança e empréstimos a amortizar no curto prazo), que surgiram na década de 1980 nos Estados Unidos em decorrência da retirada do governo federal de seu papel como credor de longo prazo e subsidiador de habitações popular. O estudo afirma que essas formas alternativas não podem ser uma estratégia nacional de longo prazo viável para a produção ou conservação de habitação de baixa renda, já que sem o governo federal em parcerias público-privadas, muitos estados optariam por ficar à margem, investindo pouco dos seus próprios recursos na produção de habitação, com isso a oferta permanente de habitação de baixa renda será, então, determinado mais por variações no custo de construção e distribuição de renda locais.

Balfour e Smith (1996), em um estudo de grupo focal na base comunitária da Rede de

Habitação Cleveland (CNH), evidenciaram que os benefícios da casa própria são atenuados para as famílias de baixa renda por parte das pressões financeiras de manutenção da casa e a falta de apoio da comunidade para seu novo empreendimento. Enquanto um valor considerável vai para a compra de propriedades, menores recursos ficam disponíveis para manutenção do lar, que é um fator chave para a autoestima dos moradores. O autor sugere enfatizar o desenvolvimento de práticas que fortaleçam a organização por meio do aumento da produção de habitação e reduzam a tensão financeira e psicológica que os problemas de habitação podem produzir para as famílias que vivem à margem da economia.

Com relação ao retorno financeiro dos projetos de habitação popular, Cummings e Di-Pasquale (1999), sobre *Low-Income Housing Tax Credit* (LIHTC), programa de produção de habitação Federal norte americano, buscaram analisar a viabilidade financeira e o tamanho dos subsídios que são prestados. Os teóricos observaram que as receitas apenas cobrem os custos de muitos dos projetos habitacionais do LIHTC e que o retorno ao patrimônio dos investidores caiu significativamente; talvez, refletido por uma maior compreensão dos riscos do projeto. O estudo também estima que os projetos LIHTC desenvolvidos por organizações sem fins lucrativos são 20,3% mais caro do que os desenvolvidos por organizações com a finalidade de lucros.

Já McClure (2000) estuda o LIHTC e a constante exigência de aportar maiores subsídios para alavancar o investimento e para a prestação de benefícios, o estudo apresenta o crédito fiscal como um mecanismo de entrega de subsídio muito ineficiente, principalmente quando o crédito fiscal é utilizado como um investimento.

Ainda sobre o LIHTC, Malpezzi e Vandell (2002) analisaram se a concessão de incentivos conseguiria aumentar a oferta de habitação ou se somente gerava a substituição das unidades não subsidiadas, que de outra forma teriam sido construídas. Seu estudo afirma que a urbanização e a estrutura etária do crescimento da população explicam grande parte da variação na oferta de habitação e afirmar que há uma taxa relativamente elevada de substituição.

Sinai e Waldfogel (2005) também estudaram o impacto dos subsídios oferecidos pelo governo ao estoque de habitação e verificaram que as unidades financiadas pelo governo elevam o número total de unidades, embora, a cada três unidades subsidiadas, duas unidades teriam sido fornecidas pelo mercado privado.

Já sobre o número de unidades habitacionais e o financiamento privado, Pillay e Naudé (2006) apresentam a dificuldade encontrada pela África do Sul, com um déficit de três milhões de unidades habitacionais, de prover habitação de baixa renda, dado que os bancos comerciais pouco conhecem sobre o comportamento, preferências e experiências de famílias de poder aquisito baixo para concessão de empréstimos nesse mercado. Há também, pouca ou nenhuma evidência para sugerir que as instituições bancárias tomem medidas inovadoras para desenvolver produtos de crédito à população de baixa renda, e, ainda, que o acesso ao crédito por parte das instituições

tradicionais é um fator limitante; dificultando, assim, o desenvolvimento de novas casas nos mercados imobiliários primários.

Ortalo-Magne e Rady (2006) apresentam importantes considerações a respeito da restrição de crédito. Uma delas é que ocorre um atraso na primeira compra da casa própria ou então as famílias são forçadas a adquirirem uma casa menor. A segunda é que mudanças no rendimento podem produzir reação exagerada no preço da habitação e, por fim, que há evidências de uma correlação forte e positiva entre os preços da habitação e o rendimento das famílias recém-criadas.

Alguns determinantes da restrição de crédito foram estudados por Warnock e Warnock (2008), os quais verificaram que países com maiores direitos legais para tomadores e credores (por meio de garantias e leis de falência), sistemas de informação de crédito mais profundos e um ambiente macroeconômico mais estável proporciona um sistemas de financiamento de habitação mais robustos. Portanto, isto é potencialmente positivo na disponibilidade de financiamento a longo prazo.

Há que se considerar, contudo, uma maior prudência com relação aos novos instrumentos que surgem. Green e Wachter (2007) evidenciaram a necessidade de se ter maior atenção e controle a gestão do risco e do uso de instrumentos complexos que vem surgindo no financiamento habitacional em países industrializados; incluindo, também, securitização e novos tipos de contratos de hipoteca e títulos lastreados em hipotecas Mortgage Títulos (MBS).

Isto posto, uma vasta expansão do crédito será necessária para atender a demanda habitacional da população de baixa renda, já que 4 (quatro) bilhões de pessoas, até 2050, são projetadas para residirem nas áreas urbanas dos países em desenvolvimento e a maior parte dessas pessoas terão uma renda baixa a moderada (FERGUSON; SMETS, 2010).

1.2 HABITAÇÃO POPULAR E O SEU IMPACTO SOCIAL

Conforme McCallum e Benjamin (1985), o conceito de habitação deve ser visto de forma mais ampla e com três importantes atributos: terra, moradia e serviços. Além disso, a habitação possui papel econômico e impacto em comunidades urbanas de baixa renda sob diversas formas, indo muito além da função de puro abrigo, podendo essas implicações serem categorizadas sob cinco formas: habitação de consumo (social); habitação como melhoria da saúde e bem-estar; habitação como macro setor econômico; habitação como estímulo para poupança e investimento; habitação como contributo indireto para renda e produção.

Entretanto, conforme esclarece Hoffman (1996), os programas de habitação não devem ser vistos como panaceia para os problemas sociais profundamente enraizados, mas sim, como elementos importantes na política de bem-estar social da qual o seu sucesso dependerá de uma boa

triagem de mutuários, da coordenação com agências de serviços sociais locais, escolas e segurança.

Nessa perspectiva de habitação como melhoria da saúde e do bem-estar, a pesquisa apresentada por Green e White (1997) buscou identificar se as crianças de proprietário permanecem mais tempo na escola do que as crianças de inquilinos. O estudo identificou que os proprietários de casas de baixa renda tinham maior probabilidade de permanecer na escola do que filhos de inquilino. Além disso, o estudo afirma que os filhos de proprietários de casas têm maior probabilidade de concluir o ensino médio do que filhos de inquilinos e possuem maiores oportunidades de terem rendimentos futuros maiores do que as crianças de locatários. O estudo corrobora com apoio às políticas do governo para incentivar as famílias de baixa renda, por meio de programas focalizados em conceder créditos tributários de uma só vez ou por meio de reduzidas taxas de financiamento.

Nessa mesma linha, Scanion (1998) faz uma extensa revisão da literatura, buscando identificar se a casa própria afeta as comunidades, por meio da promoção de uma maior acumulação de riqueza, da diminuição da mobilidade residencial e de uma maior participação da comunidade. As evidências teóricas sugerem que há efeitos positivos sobre o desenvolvimento da comunidade devido à acumulação de riqueza, maior cuidado da propriedade e o aumento da participação na comunidade. No entanto, as condições da vizinhança podem reduzir esses impactos positivos e a renda familiar deve ser suficiente e estável para tornar a casa própria viável e benéfica.

Por outro prisma, Laferrère e Blanc (2006) fizeram um contraste com os instrumentos de política habitacional utilizado na França (Construção de habitação pública, subsídios de aluguel e auxílios para famílias de baixa renda) em relação a seus homólogos norte-americanos e chamam atenção sobre a interação que os programas habitacionais podem ter gerado externalidades. Os subsídios de habitação empurram as rendas, de modo que o seu efeito global sobre o bem-estar não fica claro. A ajuda para a casa própria tem efeitos consideráveis, mas gera ganhos inesperados; sendo assim, a análise dos impactos das políticas de habitação deve considerar suas interações com todo o sistema de bem-estar.

Por outro lado, Shlay (2006) examina a viabilidade da casa própria como uma estratégia para famílias de baixa renda e argumenta que faltam evidências definitivas para substanciar as alegações de que a casa própria de baixa renda gera mudanças significativas nas famílias, bairros e mercados locais, uma vez que não está claro se a casa própria é uma causa ou consequência do ciclo de vida das famílias ou circunstâncias econômicas. Além disso, percebe-se que algumas políticas têm conduzido o preços para patamares inacessíveis.

Contudo, uma extensa revisão da literatura foi feita por Herbert e Belsky (2008) para avaliar a experiência das famílias de baixa renda com a casa própria e se elas eram capazes de se beneficiar com isso. As evidências mostram que os proprietários se beneficiam de melhores condições de saúde psicológica e física, os filhos de proprietários de baixa renda têm maior sucesso escolar, maior sucesso nos mercados de trabalho e são menos propensos a ter problemas comportamentais.

Entretanto, os proprietários enfrentam um maior risco de ser incapaz de sustentar a casa própria porque os benefícios na maior parte acumulam-se lentamente ao longo do tempo.

1.3 INADIMPLÊNCIA NO FINANCIAMENTO À HABITAÇÃO

Meltzer (1974), no tocante à disponibilidade de crédito e as decisões econômicas, identificou que o subsídio para empréstimos hipotecários incentiva a substituição da dívida hipotecária para outros tipos de dívidas e aumenta a quantidade de empréstimos e que a avaliação do risco de crédito habitacional aumenta à mesma medida que a quantidade de hipotecas de empréstimos aumentam.

Para Lawrence, Smith e Rhoades (1992), o tamanho da dívida e cobertura de pagamento, embora sejam fatores importantes a considerar na avaliação de risco de um empréstimo que leva vários anos para serem amortizado, tem o histórico de últimos pagamentos do mutuário como fator fundamental para estimar o risco de incumprimento. Outras informações como as taxas estaduais de desemprego, vendas no varejo e o valor de mercado da casa também são variáveis importantes para estimar o risco de crédito.

Elul et al. (2010) no que se referem aos gatilhos do comportamento de inadimplência, identificaram que patrimônio líquido negativo e a falta de liquidez, medida pela elevada taxa de utilização de cartão de crédito, como significativamente associados com o padrão de inadimplência; choques de desemprego estão associados com maior risco de inadimplência e uma segunda hipoteca implica risco significativamente mais alto de inadimplência.

Já Lambrecht, Perraudin e Satchell (1997), em sua análise dos impactos de variáveis conhecidas previamente à operação de crédito (*ex-ante*) sobre o comportamento da inadimplência do mutuário, concluem que o "fluxo de caixa", salário e juros pagos desempenham o maior papel sobre o comportamento de inadimplência do mutuário.

Sobre o comportamento do mutuário e os riscos do mercado imobiliário, Deng e Liu (2009) preconiza, que a maioria dos bancos chineses fazem empréstimos com base em apenas características dos mutuários, entretanto, os resultados evidenciam que as características dos mutuários e informações da garantia são importantes para determinar os riscos da operação.

Os programas habitacionais têm seus benefícios, conforme se evidência na revisão de literatura anterior, e entender as questões que levam à inadimplência nos permite conhecer melhor as variáveis que impactam sobre os diversos programas habitacionais. A seguir é apresentado PMCMV com alguns números relevantes do ponto de vista operacional e financeiro e, também, referentes as questões da inadimplência do programa.

1.4 O PROGRAMA MINHA CASA, MINHA VIDA - PMCMV

O PMCMV foi instituído, em 2009, com finalidade de criar mecanismos de incentivo à produção e aquisição de novas unidades habitacionais. É gerido, diretamente, pelo governo federal e conta com a participação das unidades da federação para implantação dos projetos. O programa ainda conta com a participação dos bancos estatais para administração dos recursos e para realização das operações de créditos, sendo que os recursos são originados do orçamento anual do governo e de outros fundos (Fundo de Arrendamento Residencial (FAR), Fundo de Desenvolvimento Social (FDS), Fundo de Garantia por Tempo de Serviço (FGTS)) (BRASIL, 2009).

O objetivo declarado do programa é a redução do déficit habitacional nacional, a inovação do programa situa-se na condição para o atendimento das famílias mais pobres, prevendo elevado subsídio para as famílias enquadradas na faixa 1 (entre 0 e 3 salários mínimos mensais de renda familiar), subsídio moderado para famílias da faixa 2 (entre 3 e 6 salários mínimos) e ausência de subsídio para as famílias da faixa 3 (entre 6 e 10 salários mínimos de renda). Além do mais, as três faixas têm acesso ao Fundo Garantidor da Habitação Popular (FGHab), uma espécie de seguro que viabiliza a compensação no caso de instabilidade de renda dos mutuários.

No período de 2009 a 2015, o programa financiou 4,2 milhões de unidades habitacionais para famílias que recebem entre 0 e 10 salários mínimos e chegou ao valor de 270 bilhões em operações de crédito, além de conceder, aproximadamente, 500 milhões em subsídios direto (BRASIL, 2016). Além do subsídio fornecido às pessoas físicas, há o incentivo à produção de unidades de habitação popular com a concessão de crédito subsidiado às pessoas jurídicas (no caso as construtoras) para construção de novas unidades de habitação popular.

Unidade Habitacionais Entregues	Valor Investido em Reais
2.335.850	R$ 158.447.080.939

Fonte: Brasil (2016).

Tabela 1.1: Quantidade de unidades entregues e valor financiado do PMCMV.

O aporte volumoso de recursos é um fator de destaque e os números do PMCMV ajudam a compreender a importância dele para a política habitacional brasileira e o impacto que proporcionou nas cidades ao logo dos sete anos de programa. De 2009 a 2015, aproximadamente 18 mil (dezoito mil) empreendimentos foram erguidos com os recursos do PMCMV e aproximadamente 160 bilhões (cento e sessenta bilhões de reais) foram destinados a financiar a construção desses empreendimentos. Com esses recursos mais de 2.35 milhões (dois milhões e trezentos mil) unidades habitacionais foram entregues e destinadas a famílias selecionadas pelo programa (BRASIL, 2016).

Em levantamento realizado pelo Departamento da Indústria da Construção da Federação

das Indústrias do Estado de São Paulo (FIESP), percebe-se que o PMCMV atuou na redução do déficit habitacional a uma taxa média anual de 2,8% entre 2010 e 2014. Conforme estudos da Fundação João Pinheiro, FJP (2010), em 2010 o déficit habitacional era de 6.941 milhões, já no levantamento realizado pela FIESP, seguindo o mesmo método da Fundação João Pinheiro (FJP), esse número passou para um déficit de 6,198 milhões de famílias em 2014.

Déficit Habitacional em 2010	Déficit Habitacional em 2014
6.940.691	6.198.294

Fonte: FJP (2010), Instituto Brasileiro de Geografia e Estatística (IBGE), FIESP.

Tabela 1.2: Déficit Habitacional 2010 *vs* 2014

É inegável os benefícios trazidos pelo PMCMV, porém outro fator que vem ganhando destaque no PMCMV é a crescente inadimplência. Conforme dados Brasil (2016), 11,8% dos contratos de financiamentos estão inadimplentes, considere inadimplente os contratos com atraso superior a 90 dias, e na faixa de renda mais baixa da população esse número chega a quase 20 %, são números preocupantes uma vez que a faixa de renda mais baixa, além de ter recebidos subsídios para a aquisição da moradia, se torna inadimplente perante seus credores, isso também pode caracterizar uma pequena falha na análise de risco de crédito antes do financiamento.

1.5 ANÁLISES DE RISCO DE CRÉDITO

1.5.1 Análises multivariadas tradicionais

1.5.1.1 Análise discriminante

A maioria das pesquisas iniciais sobre a avaliação do risco de crédito foram baseadas em análise discriminante. Com isto, a análise discriminante se tornou uma das técnicas mais populares para se avaliar o risco de crédito. O estudo mais conhecido, utilizando análise discriminante, foi desenvolvido por Altman e Saunders (1998), cuja pesquisa os teóricos apresentam o *Z-Score* como uma técnica de avaliação que prevê ou não se uma empresa é susceptível de entrar em falência no prazo de um ou dois anos.

A análise discriminante é relativamente fácil de implementar e gera resultados simples. No entanto, existem algumas limitações relacionadas com as suas aplicações na avaliação do risco de crédito. É necessário intensivos esforços de pré-processamento dos dados por meio de análise de seleção de variáveis, o que requer conhecimento especializado domínio e compreensão profunda dos dados. Ademais, este método não é eficaz para problemas com um pequeno tamanho de amostra (por exemplo, novos credores) e as suposições sobre os dados devem ser mantidas como sendo linearmente separáveis e devem seguir uma distribuição normal com a matriz de

covariância para cada grupo igual, isso torna o processo de modelagem e design de um fluxo de atualização contínua difícil de automatizar, já que, quando ocorrem mudanças na população, os modelos estáticos geralmente não conseguem se adaptar e podem precisar serem reconstruídos a partir do zero (YANG, 2007).

Recentemente, alguns estudos têm utilizado a análise discriminante como contraponto para avaliar a precisão da previsão de outros métodos de *credit scoring*. Lee (2007) estudou a aplicação de máquinas de vetor de suporte (SVM) para previsão de classificação de crédito com a Análise Discriminante como *Benchmarking* superado. Doumpos e Zopounidis (2007) faz uma comparação de vários métodos, dentre eles a Análise Discriminante, para concluir que os modelos combinados podem superar os modelos individuais para análise de risco de crédito com a redução da polarização e/ou a variância.

1.5.1.2 Regressão logística

A regressão logística é outro método bastante tradicional e popular para a análise de *credit scoring*, ela é semelhante à análise de regressão tradicional e seu uso deve estar de acordo com algumas dessa forma de análise, como, por exemplo, evitar a autocorrelação nos resíduos para evitar multicolinearidade nas variáveis independentes. A diferença entre a Regressão Logística e Análise Discriminante é que a análise discriminante deve satisfazer a suposição de distribuição normal e as matrizes iguais de covariância para descobrir o valor ideal. No entanto, a Regressão Logística não precisa destes pressupostos, mesmo que estes pressupostos não sejam satisfeitos a Regressão Logística ainda pode fornecer uma precisão relativamente alta de previsão (PRESS; WILSON, 1978).

Portanto, dado as condições restritivas da Análise Discriminantes, diversos modelos de *credit scoring* usam a regressão logística, embora a regressão logística possa ser executada bem em muitas aplicações, quando as relações do sistema não são lineares, a precisão da Regressão Logística é menor quando comparada com outros métodos. Isso tem levado ao surgimento de diversos estudos, (DOUMPOS; ZOPOUNIDIS, 2007; WU; HU; HUANG, 2014; TSAI, 2014), que também colocam a regressão logística como contraponto para avaliar a precisão da previsão de outros métodos de *credit scoring*.

Ong, Huang e Tzeng (2005) comparam a taxa de erro da regressão logística e de outros métodos com um método de *Genetic Programming (GP)*, método de aprendizado de máquina também utilizado para *credit scoring* e concluem que a GP, pode proporcionar um melhor desempenho. Já Yang (2007) também utiliza a regressão logística com um método de aprendizado de máquina, *Support Vector Machines (SVM)*, e conclui que nos novos métodos as características não lineares dos dados são automaticamente incluídas no modelo por meio de uma transformação do *kernel* que é ajustado on-line.

Apesar de bastante popular e tradicional na análise de *credit scoring*, tanto a regressão logística quanto a análise discriminante, começam a dar espaço para outros métodos mais precisos, uma vez que a melhoria na precisão de uma fração de um por cento pode traduzir-se em economias significativas.

1.5.2 Métodos de aprendizagem de máquina

Diante das limitações dos métodos paramétricos tradicionais (regressão logística e análise discriminante) os métodos de aprendizado de máquina começaram a serem utilizados na análise de *credit scoring*, uma vez que permitem que os algoritmos sejam constantemente aperfeiçoados, melhorando sua robustez e acurácia. Um olhar sobre a literatura publicada revela uma infinidade de artigos que comparam alguns métodos e concluem que um determinado método é melhor do que alguns outros métodos.

Feng et al. (1993) apresentam uma comparação de vários métodos de aprendizado de máquina, dentre eles árvores de decisão, redes neurais e classificadores estatísticos, e apresentam importantes conclusões: (1) nenhum método parece uniformemente superior aos outros, (2) métodos de aprendizado de máquina parecem ser superior para distribuições multimodais, e (3) métodos estatísticos são computacionalmente mais eficiente.

Isso posto, não pode-se selecionar um algoritmo e reivindicar sua superioridade sobre algoritmos concorrentes sem ter em conta os dados, as características do problema, bem como a adequação do algoritmo para esses dados. No entanto, pode-se possivelmente reivindicar a superioridade de um algoritmo para um conjunto de dados ou problema específico (PIRAMUTHU, 2006).

A seguir é feita uma revisão do estudos que utilizaram os métodos de aprendizado de máquina, *DT*, *SVM* e classificadores *ensemble* (*Bagging* e *AdaBoost*), que serão utilizados nesse trabalho.

1.5.2.1 Árvore de decisão

Árvore de decisão é formada por um conjunto de nós de decisão, perguntas, que permitem a classificação de cada caso e, também, proporcionam o aprendizado de forma indutiva: Cria-se uma hipótese baseada em instâncias particulares que gera conclusões gerais. Ben-David (1995) foi um dos pioneiros a utilizar árvore de decisão na avaliação do risco de crédito. Seu estudo aborda questões da monotonicidade de árvores de decisão e o quão está propriedade é desejável para a resolução de problemas como, por exemplo, na classificação de crédito e determinação do prêmio de seguro.

Galindo e Tamayo (2000) no seu trabalho de avaliação de risco de crédito usando aprendi-

zagem máquina e estatística, demostram que os modelos de árvore de decisão forneceram uma melhor estimativa com uma taxa de erro menor, quando comparado com Redes Neurais, algoritmo *K-Nearest Neighbor* e algoritmo PROBIT.

Mues et al. (2004), afirmam que árvores de decisão excessivamente grandes inibem a intuitividade e facilidade de utilização do conhecimento extraído. Sendo assim, o autores buscaram avaliar empiricamente até que ponto diagramas de decisão são capazes de fornecer uma descrição visual mais compacta do que a seu homologo em árvore de decisão. Verificou-se que o mecanismo de redução foi bastante eficaz, em que vários subgráficos foram partilhados, e que de outra forma seriam replicados quando usando uma representação de Árvore de Decisão.

Em 2007 Florez-Lopez (2007) propôs um modelo de três etapas para analisar os determinantes de *rating*, executando vários métodos multivariados (análise discriminante, logit e árvores de decisão C4.5, *Classification and Regression Tree (CART)*). E concluiu que o método de árvore de decisão oblíqua parece ser uma boa estratégia, proporcionando um equilíbrio bastante satisfatório entre precisão e inteligibilidade.

Já Wang et al. (2012), afirmaram que o desempenho do método de árvore de decisão é relativamente mais pobre, na análise de *credit scoring*, do que outras técnicas devido aos ruídos e atributos redundantes dos dados. Os autores propõem então árvore de decisão com classificadores *ensemble* para reduzir as influências dos dados sobre o ruído e os atributos redundantes de dados e para obter a precisão da classificação relativamente maior.

1.5.2.2 *Support Vector Machines* – SVM

Um dos estudos pioneiros sobre o uso do SVM em análise de risco de crédito foi o trabalho de análise de *rating* de crédito com SVM e redes neurais elaborado por Huang et al. (2004). No artigo o autor faz uma comparação entre *Backpropagation Neural Network (BNN)* e o SVM, tendo como resultados o SVM com uma precisão comparável à BNN.

Em 2006 Gestel et al. (2006) propuseram um modelo de *rating* com capacidade de aprendizagem *kernel* baseado em SVM e foi possível verificar que o SVM melhorou claramente o desempenho da classificação, embora a legibilidade do modelo diminuiu em certa medida.

Lee (2007), na tentativa de sugerir um modelo com melhor poder explicativo e estabilidade, utiliza uma técnica de validação cruzada para descobrir os valores dos parâmetros ideais de função do *kernel* RBF da SVM. Para avaliar a precisão da previsão da SVM, comparou-se o seu desempenho com os de Análise Discriminante Múltipla (MDA), *Case-Based Reasoning (CBR)* e BNN, tendo o SVM superado todos os outros métodos.

Kim e Ahn (2012) propuseram um novo tipo de classificador SVM que é projetado para estender os SVMs binários por meio da aplicação de uma estratégia de múltiplas classes ordinais.

Foram comparados os resultados do modelo com a Análise Discriminante, MLOGIT, CBR e redes neurais, tendo as técnicas de classificação de classe multi-típicos utilizando menos recursos computacionais.

Mais recentemente diversos autores têm produzido trabalhos utilizando SVM. Zhong et al. (2014) fizeram um trabalho comparando a eficácia do métodos aprendizagem de máquina e SVM para *ratings* de crédito corporativo. Shi, Zhang e Qiu (2013) compara o SVM tradicional com um modelo de ponderação RF-FWSVM que utiliza um *F-score* e RF . Niklis, Doumpos e Zopounidis (2014) criaram um modelo de classificação de risco de crédito baseada em opção de Black, Scholes e Merton utilizando modelo SVM adaptado aos requisitos de notação de crédito.

1.5.2.3 Classificadores *ensemble*

Durante os últimos anos, diferentes abordagens para classificadores *ensembles* foram aplicadas com êxito na análise de risco de crédito, demonstrando serem mais precisos do que os modelos de previsão individual. Twala (2010) foi um dos pioneiros no uso dos classificadores *ensemble* na avaliação de risco de crédito, seu estudo explorou o comportamento de cinco classificadores em diferentes tipos de ruído em termos de precisão da previsão de risco de crédito e observou que os classificadores levam a uma melhoria significativa no desempenho da análise de risco de crédito.

Wang et al. (2011) realizaram uma avaliação comparativa do desempenho de três classificadores *ensemble* (*Bagging, Boosting (Boosting) e Stacking*) com base em quatro algoritmos, (Regressão logística, DT, *Artificial Neural Network* (ANN) e SVM). Os resultados demostram que os classificadores *ensemble* podem melhorar substancialmente a classificação da análise de risco de crédito. *Bagging* teve melhor desempenho que *Boosting* em todos os conjuntos de dados de crédito. *Stacking e Bagging-DT* obtiveram o melhor desempenho em termos de precisão média, erro tipo I e erro tipo II.

Marqués, García e Sánchez (2012a) avaliando o desempenho de sete técnicas de previsão utilizado classificadores *ensemble* concluíram que a Árvore de Decisão (C4.5) constitui a melhor solução para classificadores *ensemble*, seguido de perto por rede neural e regressão logística, enquanto *nearest neighbour* e os classificadores *naive bayes* parecem ser significativamente pior.

1.5.2.4 *Bagging*

Bagging é um classificador projetado para melhorar a estabilidade e a precisão dos algoritmos de aprendizado de máquina, além de reduzir a variância. É uma técnica que combina os classificadores e fornecem resultados mais eficientes no que diz respeito a uma coleção (BREIMAN, 1996).

O algoritmo de *Bagging*, conforme descrito por Breiman (1996), segue as seguintes etapas:

1. Construção de uma amostra aleatória, t, selecionada do conjunto de dados;

2. Cálculo do estimador C_t no conjunto de dados do passo 1;

3. Repita os dois primeiros passos por $t = 1, ..., T$, em que T é o total de iterações definidas pelo executor; e

4. A partir daí, cada classificador determina um voto, em que x comporta os dados de cada elemento do conjunto de treinamento, conforme Equação 1.1.

$$C(x) = T^{-1} \sum_{t=1}^{T} C_t(x) \tag{1.1}$$

A classe com maior votação é escolhida como classificação para cada elemento do conjunto de dados.

1.5.2.5 *Boosting* e *AdaBoost*

O *Boosting* consiste no uso repetido de uma função de predição em diferente amostras do conjunto inicial, dessa forma, como no *Bagging*, cada classificador é treinado usando um conjunto de treinamento diferente. A diferença em relação ao *Bagging* é que a importância do voto é ponderado com base no desempenho de cada modelo, em vez de atribuição do mesmo peso para todos os votos, e os dados são re-amostrados (LANTZ, 2015).

O *AdaBoost* é um algoritmo derivado do *Boosting* e tem sido utilizado com êxito como classificador. A lógica central do algoritmo consiste em manter uma distribuição ou um conjunto de pesos sobre o conjunto base e esses pesos são reajustados nos próximos processamentos (TSAI; HSU; YEN, 2014).

2 ESTUDO TEÓRICO – RISCO DE CRÉDITO E MEDIDAS PRUDENCIAIS PARA O PROGRAMA DE HABITAÇÃO PÚBLICA DO BRASIL.

Aqui, neste capítulo, procura-se fazer uma abordagem dos temas que compõem o estudo teórico da pesquisa em pauta: habitação pública: identificação na literatura; benchmarking de programas de habitação pública; breve histórico de habitação social no Brasil; descrição do PMCMV; inadimplência do PMCMV e medidas prudenciais, com o objetivo de aperfeiçoar a análise dos dados coletados.

Políticas e programas de habitação são fundamentais para o bem-estar das pessoas de baixa renda e comunidades. Os desafios de fornecer assistência habitacional aos pobres são muitos e cercada por numerosas metas e prioridades conflitantes (COLLINS et al., 2005).

Nos últimos anos, o Brasil empreendeu maiores esforços para fornecer habitação pública de baixo custo com o objetivo de reduzir o número crescente do déficit de habitação evidenciado na Pesquisa Nacional por Amostra de Domicílios (PNAD) do IBGE (2008). Para isto, foi criado o PMCMV que objetiva incentivar a produção de novas unidades habitacionais e fornecer subsídios para o financiamento dessas unidades habitacionais.

A renda baixa ou inexistente é uma das principais referências para o critério de seleção dos beneficiários do PMCMV, uma vez que o público alvo dificilmente conseguirá ter o crédito de que necessita para adquirir uma casa com dinheiro de uma linha de crédito habitual, com a análise de risco de crédito convencional e com taxas relativamente altas, convertendo este tipo de empréstimo em algo impraticável para os pobres.

Ao utilizar a baixa renda como critério de seleção, o programa é tratado preponderantemente, pelos beneficiários, como um espécies de subsídio ou doação, não como uma modalidade de crédito ou financiamento, por isso, os dados de inadimplência são altamente desfavoráveis, situando-se num patamar superior a 10% da carteira (BRASIL, 2016). A alta inadimplência é potencialmente prejudicial para a continuidade do programa uma vez que são necessários infinitos aportes de recursos para que seja possível beneficiar outras pessoas, sendo assim, a incorporação de aspectos prudenciais ao programa como, por exemplo, a análise de risco e criação de outras alternativas de suporte para aqueles que não conseguem ter uma avaliação aprovada pode tonar o programa de habitação pública mais eficaz.

O objetivo deste capítulo centra-se em propor medidas prudenciais para concessão do crédito para aquisição de unidades habitacionais, com uma proposta de análise de risco de crédito e medidas que visam o aperfeiçoamento do programa de habitação pública, com a implementação de outros incentivos, que não o fornecimento da propriedade da unidade habitacional, como evidenciado por Popkin (2004) sobre o HOPE VI, programa habitacional americano que combina

subsídios para a revitalização física com o financiamento de melhorias de gestão, serviços de apoio e o uso de *voucher* de habitação para aluguel de moradias no mercado privado.

2.1 HABITAÇÃO PÚBLICA: IDENTIFICAÇÃO NA LITERATURA

O uso da terminologia de "habitação pública" varia entre os países. Alguns referem-se a habitação fornecida diretamente e financiado sob a égide do governo, para habitação dos proprietários, por exemplo, em Hong Kong e Singapura. Alguns se referem apenas a habitação para arrendamento fornecido diretamente para famílias de baixa renda ou grupos especiais da população e produzido com o subsídio do governo, por exemplo, na China, Coreia do Sul, Taiwan e Japão (CHIU, 2013).

Nas democracias europeias, a habitação social de arrendamento tem sido utilizada como um mecanismo de redistribuição de renda e historicamente como um meio para mobilizar as classes trabalhadoras. Suas políticas de habitação têm sido influenciadas pela história da habitação, fatores socioeconômicos e demográficos. São em grande parte três modelos de habitação social entre os estados-chave da União Europeia. O primeiro é o "modelo residual" utilizado no economicamente menos desenvolvidos do Sul da Europa e em partes dos ex-Estados comunistas da Europa Oriental, aonde a habitação social está reservada para os segmentos mais pobres da população. O segundo é conhecido como o "modelo generalista", onde a habitação social é feita acessível a um grupo mais amplo da população por estabelecimento de um limite para os níveis de renda, nesta categoria estão: França, Alemanha, Reino Unido e Bélgica. O "modelo universal" é a terceira categoria que não tem qualquer alvo social, mas tem como objetivo fornecer suporte subsidiado para arrendamento habitação. Este modelo ainda está em prática na Suécia e Holanda (WONG; GOLDBLUM, 2016).

Especificamente no contexto francês, habitação pública se caracteriza pelas ajudas diretas ou indiretas, dos poderes públicos, dos governos estaduais e locais e é explicitamente projetado para acomodar pessoas com rendimentos modestos (STEBE, 2013).

No Reino Unido e na Austrália o termo Habitação Social, é um termo genérico que inclui organizações de habitação pública e a comunidade de forma geral. Nos últimos anos, o termo habitação pública foi usado para denotar qualquer carcaça que foi subsidiada para permitir inquilinos com rendas abaixo do mercado de alugar um imóvel (JACOBS et al., 2010).

Para possibilitar uma discussão mais abrangente, este artigo adota uma definição mais ampla de habitação pública, ou seja, habitação para arrendamento ou propriedade da habitação fornecida pelo governo ou desenvolvedores que envolvem recursos públicos.

2.2 BENCHMARKING DE PROGRAMAS DE HABITAÇÃO PÚBLICA

A China tem uma experiência única de desenvolvimento de habitação pública. Antes de 1978, a oferta de habitação na China repousava sobre fortes ideologias socialistas: a habitação não era considerada como uma mercadoria com um valor de troca, mas sim como uma necessidade básica, como um direito. No entanto, o programa de habitação social, com base no empregador, foi formalmente abolido em 1998, e a esmagadora maioria do parque habitacional público foi rapidamente privatizada. Contudo, o governo chinês está novamente comprometido com o desenvolvimento de habitação pública em larga escala. Ao longo dos últimos anos, dados oficiais sugerem que a China começou a construir 16 (dezesseis) milhões de unidades de habitação pública e terminou 11 (onze) milhões de unidades durante o período do "Plano Quinquenal" (2006-2010). No início de 2011, o governo chinês anunciou o comprometimento de construir 36 (trinta e seis) milhões de unidades de habitação pública durante o período do "Plano Quinquenal" (2011-2015) (CHEN et al., 2013).

Entretanto, Huang (2012) considera que embora o governo chinês tem demonstrado um compromisso impressionante para habitação de baixa renda nos últimos anos, é justo dizer que o programa de habitação de baixa renda até agora falhou. A primeira razão é que o governo central não definiu objetivos claros para a habitação de baixa renda. Existem várias e, muitas vezes, conflitantes metas para o setor da habitação, em geral, e para a habitação de baixa renda, em particular. Isto resultou na falta de um plano estratégico para a habitação de baixa renda com mudanças políticas constantes.

A França também apresenta um importante caso de habitação pública, passando por um período maciço de construção de grandes conjuntos habitacionais. Mais de 1,4 milhões de unidades habitacionais foram construídas em toda a França nas Zups (*Zones à Urbaniser en Priorité*) durante o período 1965-1978, a França construiu quase meio milhão de novas unidades habitacionais para todos os grupos de renda na qual 80% receberam ajuda do governo. Destas unidades construídas, quatro foram construídas por empresas de economia mista, responsáveis pelos grandes conjuntos habitacionais (EISINGER, 1982).

O agrupamento do grande número de famílias de baixa renda em um habitat altamente subsidiado, onde os serviços de manutenção e qualidade foram difíceis de encontrar provou ser uma falha décadas posteriores na integração social e de justiça. Com uma baixa qualidade física dos grandes conjuntos habitacionais, uma elevada delinquência, baixa integração comunitária, marginalização e exclusão social (DRIANT, 2012). E, ainda, conforme relata Wong e Goldblum (2016), em 2000, a França enfrentava um déficit habitacional de 850.000 unidades. Enquanto dois milhões de unidades habitacionais foram desocupadas, cerca de quatro milhões de pessoas viviam em condições de habitação precária, em ruínas, abaixo do padrão ou inaceitavelmente pequenas.

Os EUA oferecem diversos subsídios à habitação. Um desses subsídios é administrado

pela Receita Federal e está disponível para todas as pessoas em qualquer nível de renda, e oferece incentivos fiscais para os proprietários, sendo o pagamento de juros e de impostos de propriedade local autorizados como despesas dedutíveis. O HUD é atualmente um dos principais apoiantes do governo federal dos EUA para o fornecimento de habitação acessível para as famílias de baixa renda. A Tabela 2.1 lista os programas habitacionais do HUD. O programa mais conhecido do HUD começou com a Lei de Habitação Pública de 1937, com o governo federal financiando a construção de conjuntos habitacionais públicos que são, então, propriedade e operados por autoridades locais de habitação. Há também outras formas de incentivos como os fornecimentos de *Vouchers* para famílias de baixa renda, idosos e pessoas com deficiência para obter uma habitação no mercado privado (MOTLEY; PERRY, 2013).

Programa	**População**	**Propósito**
Emergency solutions grants.	Desabrigados recente.	Ajudar os cidadãos rapidamente recuperar a habitação permanente estável após uma crise de habitação e falta de moradia.
Housing choice vouchers.	Idosos e deficientes de baixa renda.	Fornecer subsídios para essas pessoas para ajudá-los a pagar a habitação decente no mercado privado.
Privately owned subsidized housing.	Proprietários de apartamentos.	Ajuda proprietários de baixa renda com subsídios.
Public housing program.	Baixa renda, idosos e pessoas com deficiência.	Unidades habitacionais públicas subsidiadas.
Shelter plus care program.	Pessoas sem-abrigo e com deficiência.	Fornecer serviços de habitação e apoio a longo prazo.
Single room occupancy program.	Pessoas sem–teto.	Fornecer assistência de aluguel para pessoas sem abrigo e ajudar a reabilitar habitações individuais.
Supportive housing program.	Pessoas sem–teto.	Desenvolver habitação e serviços que permitem às pessoas sem–teto alcançar a estabilidade residencial, aumentar as habilidades e/ou renda, e obter uma maior auto–determinação de apoio.
Title V.	Pessoas sem–teto.	Fornecer propriedades que podem ser usadas para abrigo, serviços e outros usos.

Fonte: Housing (2016)

Tabela 2.1: Programas de habitação do HUD.

O que se observa nos programas de habitação pública é que existem inúmeras características próprias de cada país como, por exemplo, o grau em que casa própria deve ser incentivada; o equilíbrio entre a propriedade e o aluguel subsidiado; os critérios de seleção e acesso; o papel do Estado no financiamento habitacional; origem dos recursos financeiros; benefícios ofertados e a forma de gestão operacional, essas características dão identidade aos programas de habitação

pública e a forma como cada país vem tratando a habitação pública.

2.3 BREVE HISTÓRICO DE HABITAÇÃO SOCIAL NO BRASIL

Para compreender os programas de habitação pública no Brasil e em qual contexto se deu a criação do PMCMV, é interessante conhecer um pouco da história do déficit habitacional brasileiro e das políticas habitacionais anteriores. O déficit habitacional no Brasil é um problema social que vem, desde os tempos do Império, quando surgiram as primeiras favelas aos pés dos morros do Rio de Janeiro. Desde então, várias políticas públicas surgiram com o objetivo de minimiza o tamanho do problema (DAMICO, 2011).

Após 1964, a resposta do governo militar, à forte crise de moradia, foi a criação do Banco Nacional de Habitação (BNH), nesse período, embora a produção habitacional tenha sido significativa, ela esteve muito aquém das necessidades geradas pelo acelerado processo de urbanização que ocorreu no Brasil na segunda metade do século XX. Entre 1950 e 2000, a população urbana brasileira vivendo em cidades com mais de 20 mil habitantes cresceu de 11 milhões para 125 milhões. Na redemocratização, ao invés de uma transformação, ocorreu um esvaziamento e pode-se dizer que deixou propriamente de existir uma política nacional de habitação (BONDUKI, 2008).

Entre a extinção do BNH (1986) e a criação do Ministério das Cidades (2003), o setor do governo federal responsável pela gestão da política habitacional esteve subordinado a sete ministérios ou estruturas administrativas diferentes, caracterizando descontinuidade e ausência de estratégia para enfrentar o problema. A partir de 2005, alterações relevantes ocorreram na área de financiamento habitacional com a criação do Sistema Nacional de Habitação e de dois subsistemas – o de habitação de mercado e o de interesse social permitindo ampliar a aplicação de recursos do Sistema Brasileiro de Poupança e Empréstimo (SBPE) e Sistema Financeiro Imobiliário (SFI) em empreendimentos habitacionais, condição fundamental para que o FGTS fosse direcionado para a faixa de interesse social (BONDUKI, 2008).

Em 2009, os resultados da PNAD (IBGE, 2008), constatou um déficit de domicílios de 6.273 milhões de moradias no Brasil, das quais 82,6% estavam localizadas nas áreas urbanas. Diante deste cenário, o governo federal instituiu o PMCMV para criar mecanismos de incentivo à produção e aquisição de novas unidades habitacionais, visando superar o déficit habitacional urbano (BRASIL, 2009).

A produção social da moradia no Brasil passou a ganhar espaço e reconhecimento do Estado ao longo dos últimos anos, estando claros seu planejamento e institucionalização, em que diversos programas habitacionais vieram a financiar esta forma de produção. A análise da execução desses programas mostra situações de demanda reprimida, dificuldades operacionais

e restrições de fundos, em que os programas sucedem-se, passando o PMCMV a concentrar a produção habitacional, inclusive na modalidade voltada às entidades privadas sem fins lucrativos (cooperativas e associações) (BALBIM; KRAUSE, 2014).

2.4 DESCRIÇÃO DO PMCMV

Neste subtópico, é relevante fazer uma descrição, mesmo que seja sucinta, do PMCMV, com o intuito de entender melhor a finalidade deste programa habitacional implantado pelo governo federal.

O PMCMV foi instituído no Brasil em 2009 com finalidade de criar mecanismos de incentivo à produção e aquisição de novas unidades habitacionais. Para incentivo a aquisição de novas unidades, foi disponibilizado subsídios para as famílias de baixa renda e para o incentivo à produção de novas unidades foi instituído linhas de crédito subsidiadas às construtoras. No período de 2009 a 2015 o programa financiou 4,2 milhões de unidades habitacionais para famílias que recebem entre 0 e 10 salários mínimos e chegou ao valor de 160 bilhões em operações de crédito, além de conceder aproximadamente 500 milhões em subsídios direto (BRASIL, 2016).

O objetivo do PMCMV é a redução do déficit habitacional nacional, a inovação do programa situa-se na condição para o atendimento das famílias mais pobres, prevendo elevado subsídio para as famílias enquadradas na faixa 1 (entre 0 e 3 salários mínimos mensais de renda familiar), subsídio moderado para famílias da faixa 2 (entre 3 e 6 salários mínimos) e ausência de subsídio para as famílias da faixa 3 (entre 6 e 10 salários mínimos de renda). Além disto, as três faixas têm acesso ao FGHab, uma espécie de seguro que viabiliza a compensação no caso de instabilidade de renda dos mutuários. Desta forma, os valores da renda familiar definem os benefícios destinados aos selecionados do programa, conforme se verifica na Tabela 2.2.

Faixa de Renda	Valor de Referência	Descrição do Incentivo
Faixa 1	De 0 a 3 Salários Mínimos	Subsidio + Financiamento (Juros A)
Faixa 2	De 3 a 6 Salários Mínimos	Financiamento (Juros B)
Faixa 3	De 6 a 10 Salários Mínimos	Financiamento (Juros C)

Fonte: Brasil (2009).

Tabela 2.2: Faixa salarial e incentivos do PMCMV.

A seleção prioriza residentes em áreas de risco ou insalubres, que tenham sido desabrigadas e pessoas com deficiência. A renda do beneficiário também é critério para o acesso a outros benefícios como, por exemplo, o abatimento no valor dos emolumentos e descontos no seguro do imóvel.

Ao utilizar o critério de renda (ou ausência de renda) o risco de crédito, risco relacionado

à incerteza da qualidade de crédito do mutuário, passa por uma análise quase precária uma vez que a seleção dos mutuários são realizadas por Estados e Municípios e com um grande poder de discricionariedade, não levando e consideração características que poderiam distinguir um bom de um mau pagador.

Para operacionalizar o PMCMV, um conjunto de atores atuam direta ou indiretamente. O Ministério das Cidades é o responsável por estabelecer diretrizes, fixar regras, definir a distribuição de recursos entre as Unidades da Federação, além de acompanhar e avaliar o desempenho do programa. O Ministério da Fazenda e do Planejamento, Orçamento e Gestão em conjunto com o Ministério das Cidades decidem sobre os limites de renda familiar dos beneficiários. A Caixa Econômica Federal (banco estatal do Governo Federal) é a instituição financeira responsável pela administração dos recurso para o programa ela também atua juntamente com o Banco do Brasil (banco estatal do Governo Federal) realizando operações de crédito com os recursos do programa para pessoas físicas construtoras. Os Estados, Distrito Federal e Municípios atuam na indicação das áreas priorizadas para implantação dos projetos, isenção de tributos, aporte de recursos, indicação da demanda, indicação de solicitantes para a venda dos empreendimentos.

O PMCMV, então, se caracteriza como uma gestão centralizada no governo federal, seja por meio da administração direta (Ministérios), seja por meio do uso de bancos estatais. Aos estados e municípios resta a atuação operacional na indicação dos beneficiários e áreas para implantação dos projetos.

O PMCMV utiliza recursos de três fundos: FGTS, FAR e FDS. Cada fundo possui alguns critérios para permitir o uso dos seus recursos. Os recursos oriundos do FGTS, por exemplo, é destinado a famílias com renda até R$ 5.000,00 (cinco mil reais), já os recursos oriundos dos FAR é destinado a famílias com renda até R$1.600,00 (um mil e seiscentos reais). E há ainda recursos financeiros que são disponibilizados diretamente pelo Governo com base no orçamento federal para atender o Programa Nacional de Habitação Rural (PNHR). É importante frisar que, apesar de usar os fundos como fonte de recursos para o PMCMV o dinheiro é oriundo de fontes públicas, portanto, não há destinação de recursos privados. A Tabela 2.3 demostra, resumidamente, as características do PMCMV.

Características PMCMV	Descrição PMCMV
Objetivo.	Redução do déficit habitacional nacional.
Critério de seleção.	Renda como critério de seleção dos beneficiários.
Fonte de recursos.	Fonte de recursos exclusivamente pública.
Analise de risco de crédito.	Ausência de análise de risco de crédito dos beneficiários, dado a necessidade de contemplar a parcela mais pobre da população que não possui renda.
Subsídios.	Subsídio destinado à aquisição e construção de novas unidades habitacionais.
Gestão e operação.	Centralizada no governo federal, porém os municípios selecionam os beneficiários de acordo com o critério estabelecido.
Benefícios.	Taxa de juros subsidiada, isenção de taxas e emolumentos, seguro habitacional.

Fonte: Brasil (2009).

Tabela 2.3: Características PMCMV.

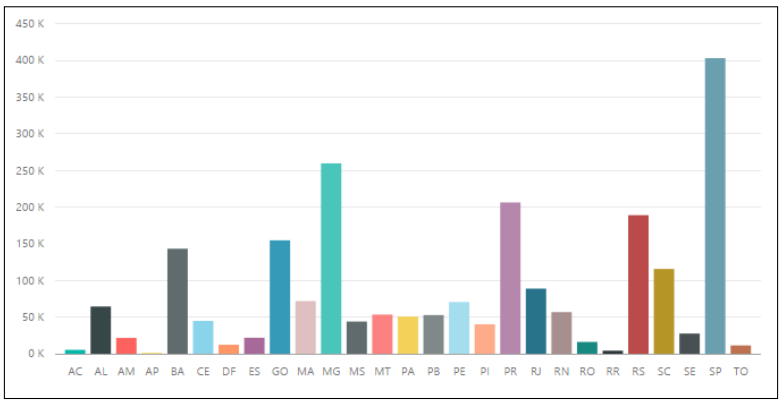
Os números do PMCMV ajudam a compreender a importância dele para a política habitacional brasileira e o impacto que proporcionou nas cidades ao logo dos sete anos de sua implementação. De 2009 a 2015, aproximadamente 18 mil (dezoito mil) empreendimento foram erguidos com os recursos do PMCMV e, aproximadamente, 160 (cento e sessenta bilhões) foram utilizados para financiar a construção destes empreendimentos. Com este recurso, mais de 2.300.000 (dois milhões e trezentas mil) unidades habitacionais foram produzidas.

Unidade Habitacionais Entregues	Valor Investido em Reais
2.335.850	R$ 158.447.080.939

Fonte: Brasil (2016).

Tabela 2.4: Unidades habitacionais entregues e valor financiado do PMCMV.

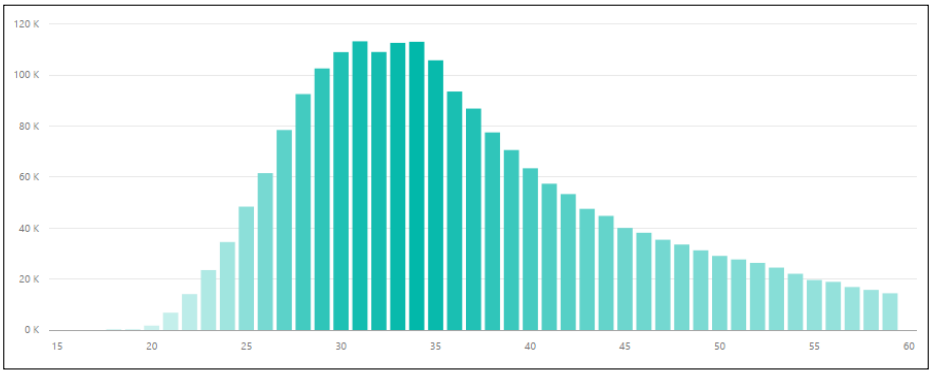
A Figura 2.1 demostra a distribuição do quantitativo de unidades habitacionais por Unidade da Federação (UF), mostrando que São Paulo, seguido de Minas Gerais, Paraná e Rio Grande do Sul foram os estados mais beneficiados pelo programa, levando-se em consideração o número de unidades habitacionais.



Fonte: Brasil (2016).

Figura 2.1: Contratos por UF.

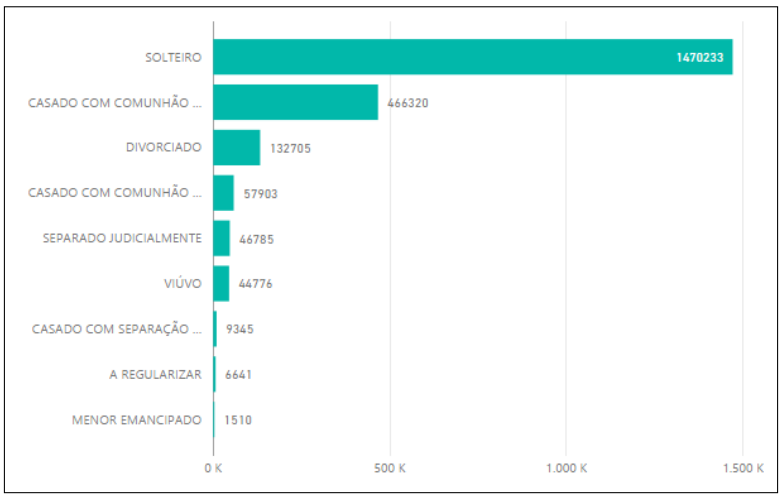
O perfil dos beneficiários do programa também apresentam características significativas. Os selecionados apresentam idade mínima de 18 anos e máxima de 110 anos já a idade média é de 38 anos. A Figura 2.2 evidência a distribuição da idade pelo quantitativo de contratos do programa.



Fonte: Brasil (2016).

Figura 2.2: Contratos por idade dos beneficiários.

Outra característica marcante é o estado civil dos beneficiários do programa, a grande maioria é solteiro (1,5 milhão), seguido por outros 467 mil que são casados com comunhão de bens, conforme pode ser visto na Figura 2.3. A renda média das famílias beneficiadas é de 2.000 reais e o gênero dos beneficiários está distribuído em 53% de mulheres e 47% de homens.



Fonte: Brasil (2016).

Figura 2.3: Contratos por estado civil.

Um resultado relevante do programa pode ser visto na Tabela: **??**, aonde em levantamento realizado pelo Departamento da Indústria da Construção da FIESP, percebe-se que o PMCMV atuou na redução do déficit habitacional a uma taxa média anual de 2,8% entre 2010 e 2014. Conforme estudos da FJP, FJP (2010), em 2010 o déficit habitacional era de 6.941 milhões, já no levantamento realizado pela FIESP, seguindo o mesmo método da FJP, esse número passou para um déficit de 6,198 milhões de famílias em 2014.

Déficit Habitacional em 2010	**Déficit Habitacional em 2014**
6.940.691	6.198.294

Fonte: FJP (2010), IBGE, FIESP.

Tabela 2.5: Déficit Habitacional 2010 *vs* 2014

2.5 INADIMPLÊNCIA DO PMCMV

O conceito de inadimplência ou *default* utilizado neste trabalho segue o *Basel Committee on Banking Supervision* (BCBS) que o define como atrasados em mais de 90 (noventa) dias em alguma obrigação material com o conglomerado financeiro (BASEL, 2007). Portanto, atrasos de pagamentos inferiores a 90 dias não serão contabilizados como inadimplentes.

O PMCMV conseguiu financiar um grande número de unidades habitacionais para a população de baixa renda, porém, atualmente a inadimplência dos financiamentos destinados a população de menor renda, naqueles em que o governo dá subvenções econômicas para reduzir as parcelas, ultrapassa os 18%. Isso certamente prejudica o programa que ficará condicionado

a novos aportes de recursos dado que a elevada inadimplência diminui a quantidade de recursos disponíveis para aquisição de novas unidades habitacionais PMCMV.

Verifica-se, conforme pode ser visto na Tabela 2.6, que a inadimplência é maior na faixa I do programa, ou seja, nos imóveis destinados às famílias com renda mensal bruta de até R$ 1.800,00 (um mil e oitocentos reais). A elevada inadimplência pode estar associada à percepção de que o imóvel está sendo doado, conforme abordou Gonzalez (2015), e ainda a baixa atenção dada aos devedores e a falta de instrumentos de cobrança e retomada do imóvel, que só foi criado com a publicação da lei 13.043/2014 (BRASIL, 2014) que possibilita aos fundos, na qualidade de credor fiduciário, realizar a retomada do imóvel e promover sua reinclusão no programa habitacional, destinando-o à aquisição por outro beneficiário.

Faixa de Renda	Contratos Inadimplentes	% em Relação ao Total
1	211538	9,45%
2	49695	2,2%
3	2700	0,12%

Fonte: Brasil (2016).

Tabela 2.6: Inadimplência por faixa de renda.

Avaliações e critérios subjetivos e a ausência de análise de risco de crédito para selecionar o beneficiário do programa é outro fator que pode ter contribuído para o índice de inadimplência. Os dados descritos a seguir sobre o perfis dos inadimplentes reforçam ainda que algumas diretrizes do programa contribuíram para o aumento da inadimplência, já que foram pautadas, por exemplo, em critérios de gênero e não na utilização de mecanismos estatístico para a seleção dos beneficiários.

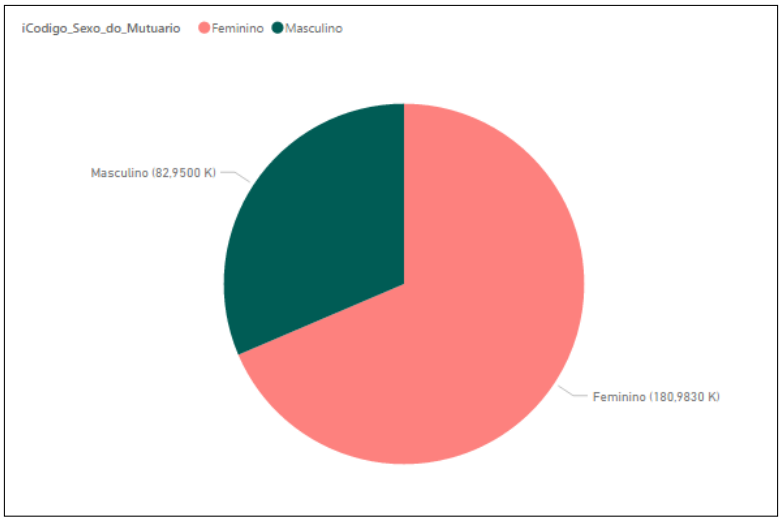
Movimentos sociais exigiram que os programas habitacionais públicos priorizassem mulheres, incluindo medidas que estipulassem preferência e que os contratos e acordos financeiros fossem feitos no nome das mulheres, não obstante a sua contribuição para a renda familiar ou o Estado civil. Em 2006, o Ministério das Cidades aprovou a Resolução 004/06, cujo Artigo 03 dá prioridade às famílias de baixa renda com mulheres chefes de família em seus programas habitacionais públicos (LEVY; LATENDRESSE; CARLE-MARSAN, 2016). O resultado desta estratégia de seleção pode ser visto na Tabela 2.7, na qual o número de beneficiários inadimplentes do sexo feminino é o dobro dos inadimplentes do sexo masculino.

Sexo	Contratos Inadimplentes	% em relação ao total
Homem	82.950	3,70%
Mulher	180.983	8,9%

Fonte: Brasil (2016).

Tabela 2.7: Sexos inadimplentes.

Embora alguns autores como, por exemplo, Levy, Latendresse e Carle-Marsan (2016), defendam que as mulheres devem ter um acesso diferente à habitação dos homens por causa de uma divisão do trabalho baseada no gênero, uma diferença de rendimento e uma política estatal de habitação historicamente dirigida de forma esmagadora aos homens. Os dados da inadimplência reforçam que um critério de seleção de beneficiários ancorado essencialmente em questões de gênero pode não ser a mais adequada em termos de sustentabilidade do programa. A Figura 2.4 demostra visualmente o resultado dos contratos inadimplentes por gênero.



Fonte: (BRASIL, 2016).

Figura 2.4: Contratos inadimplentes por gênero.

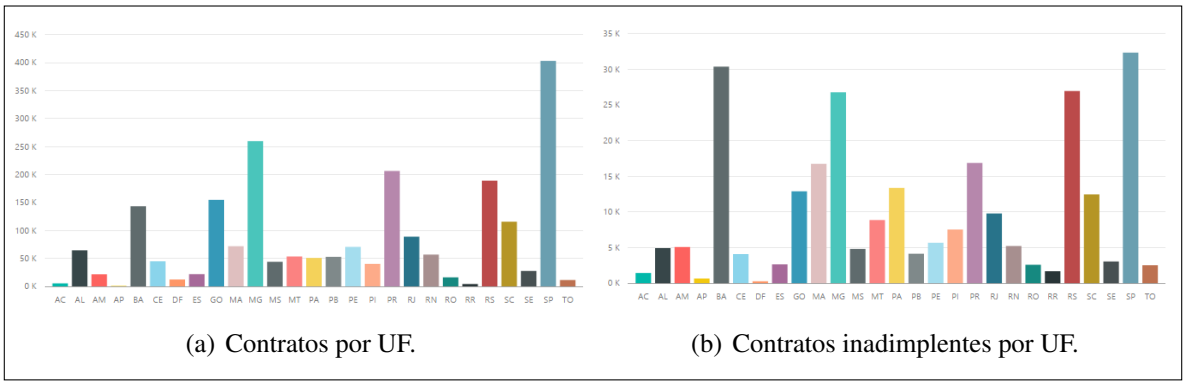
Apesar de priorizar o atendimento às famílias, com mulheres responsáveis pela unidade familiar, a legislação nada fala sobre o estado civil do beneficiário do programa. Do total de inadimplentes do programa verifica-se que 69 % são solteiros, conforme Tabela 2.8, ou seja, 12,42% dos contratos de beneficiários solteiros estão inadimplentes. Os casados com comunhão total de bens apresentam 13,02% dos contratos como inadimplentes. Já, os menores emancipados, apesar de ser um número reduzido de contratos, apresentam 21,78 % dos contratos como inadimplentes. Por outro lado, os contratos que possuem o estado civil por regularizar possuem 32,96 % dos contratos como inadimplente, o que reforça a necessidade de cadastrados qualificados.

Estado Civil	Contratos Inadimplentes
Solteiro	182625
Casado com comunhão parcial de bens	48459
Divorciado	11391
Casado com comunhão total de bens	7543
Viúvo	5339
Separado Judicialmente	5002
A regularizar	2189
Casado com separação de bens	1056
Menor emancipado	329

Fonte: Brasil (2016).

Tabela 2.8: Estado civil dos inadimplentes.

Ao analisar a inadimplência por Unidade da Federação, verifica-se que alguns estados seguem uma proporção no número de contratos e inadimplentes próximo aos níveis gerais (11,80%). São Paulo, por exemplo, tem um número de 403.169 mil contratos e 32.339 mil contratos inadimplentes, o que dá um total de 8,02 % de contratos inadimplentes. Entretanto, outros estados como, por exemplo, a Bahia, apresentam uma elevada proporção de inadimplentes, dos 143.396 mil contratos 30.372 estão inadimplentes o que representa 21,18 % dos contratos naquele estado. A Figura 2.5 traz um comparativo dos contratos por UF *vs* contratos inadimplentes por UF.



(a) Contratos por UF. (b) Contratos inadimplentes por UF.

Fonte: Brasil (2016).

Figura 2.5: Contratos inadimplentes por UF.

As características dos perfis dos inadimplentes torna ainda mais evidente que o estabelecimento de critérios com base na renda, no gênero, no estado civil ou no local de residência são insuficientes para garantir uma adequada seleção de beneficiários. O risco de crédito poderia ser melhor controlado com o uso de instrumentos estatísticos e de sistemas multivariados, possibilitando a mensuração do risco de forma mais objetiva e com uma abordagem empírica que enfatiza a predição, conforme já tratado por Altman e Saunders (1998). A seguir, são abordadas medidas prudenciais que poderiam ser adotadas nos programas habitacionais, de forma geral, visando assegura a sustentabilidade das fontes de financiamento para habitação pública.

2.6 MEDIDAS PRUDENCIAIS

2.6.1 Análise de risco de crédito para financiamentos de habitação pública

Entende-se, por análise de risco de crédito, um conjunto de modelos preditivos e suas técnicas subjacentes que ajudam as instituições financeiras na concessão de crédito. Estas técnicas decidem quem vai ter crédito e quanto de crédito que deve receber. Destarte, a análise de risco de crédito avalia o risco em emprestar para um cliente particular que é identificado como "bom pagador" e "mau pagador" (onde o comportamento negativo, por exemplo, *default* é esperado) (ŘEZÁČ; ŘEZÁČ et al., 2011).

A análise de risco de crédito, com o uso de instrumentos estatísticos e de sistemas multivariados para a mensuração do risco de forma mais objetiva, prever a probabilidade de que um candidato com qualquer pontuação, de acordo com o modelo estatístico, será classificado como "bom" ou "mau". Estas probabilidades ou pontuações, juntamente com outra consideração de negócios, tais como taxas esperadas de aprovação, lucros e perdas, são utilizadas como uma base para a tomada de decisão (ALTMAN; SAUNDERS, 1998).

Vários modelos para análise de crédito foram introduzidos nas últimas décadas. Os mais amplamente utilizados são regressão logística, análise discriminante, árvores de decisão, redes neurais e SVM (BROWN; MUES, 2012). Para a construção de um modelo preditivo é estabelecido a definição correta da variável dependente (que no caso é o *Default*), para que possa ser caracterizado com precisão o "bom" e o "mau" cliente. Esta definição baseia-se, geralmente, no número de dias de atraso após a data de vencimento.

Einav, Jenkins e Levin (2013), no seu estudo sobre o impacto da pontuação de crédito no consumo, identificaram dois benefícios distintos da classificação de risco: a capacidade de selecionar mutuários de alto risco (e aqui podemos incluir a população de baixa renda) e a capacidade de fazer empréstimos mais generosos para mutuários de baixo risco. Sendo assim, *credit scoring* permitiu um aumento nos lucros por empréstimo. Para a habitação pública, o importante não é o incremento no lucro decorrente da intermediação financeira, mas sim, a sustentabilidade das fontes de recursos, de forma que novos beneficiários possam ser atendidos pelos recursos do pagamento das prestações devidas.

Ao invés de rejeitar empréstimos de maior risco, a exemplo do PMCMV, o credor pode decidir o preço do risco, exigindo um prêmio de taxa de juros sobre os empréstimos com maior probabilidade prevista de inadimplência. O uso do *credit scoring* também pode ajudar com a recolha e perda mitigação de processo, por exemplo, permitindo que os credores concentrem os recursos humanos sobre os mutuários, cuja pontuação de crédito indica maior risco de inadimplência.

Avery et al. (1996) apresentam a relação entre a pontuação de crédito e a inadimplência. Eles mostram que mutuários inadimplentes desproporcionalmente têm pontuações baixa. Os mu-

tuários com baixa pontuação de crédito representaram apenas 1,5% de todos os recéns-originados empréstimos de taxa fixa, mas 17% daqueles que se tornaram inadimplente. Esta relação vale para outros tipos de produtos e empréstimos, por exemplo, os mutuários com baixa pontuação de crédito foram responsáveis por 2,1% das hipotecas de taxa fixa, mas eles representaram 32% das pessoas que se tornaram inadimplente. No entanto, eles também demostram que a maioria dos mutuários com pontuação de crédito baixa não são inadimplentes. Por exemplo, no caso dos empréstimos de taxa fixa convencional, apenas 4,4 % dos devedores com baixa pontuação de crédito se tornou inadimplente durante o período de desempenho. Assim, enquanto mutuários inadimplentes desproporcionalmente têm baixas pontuações, a maioria dos mutuários com baixa pontuação não são inadimplentes.

Portanto, a utilização de instrumentos estatísticos e *Machine Learning* (ML) permite criar um modelo de predição e mensuração do risco de forma mais objetiva, possibilitando diminuir o índice de inadimplência e ao mesmo tempo fornecer empréstimos para mutuários com uma renda baixa. Ao invés de rejeitar um empréstimo um sistema de predição pode permitir classificar corretamente o bom e o mau pagador, logo, o programa habitacional pode continuar beneficiando a população de baixa renda com empréstimos para aquisição de habitação, porém, de forma mais sustentável financeiramente.

Qual seria então o melhor modelo de *credit scoring* para predizer o bom e o mau pagador do PMCMV? Conforme Hand e Henley (1997), em geral não há melhor modelo global, já que ele dependerá dos detalhes do problema, da estrutura dos dados, das características usadas e da medida em que é possível separar as classes usando essas características. Esse ponto é objeto do estudo empírico apresentado no Capítulo 3.

2.6.2 Cadastro positivo

A lei 12.414, Brasil (2011), que dispõe sobre a formação de banco de dados com informações de adimplemento, o chamado cadastro positivo, tornando mais abrangentes, a partir do compartilhamento de informações positivas e negativas. Uma lista positiva inclui nomes, endereços ou outras informações de identificação sobre clientes em que um comerciante geralmente confia. Se um nome estiver na lista positiva, então ele pode substituir escores negativos de fraude. Por outro lado, uma lista negativa inclui nomes, endereços ou outras informações de identificação sobre clientes que um comerciante geralmente não confia. Se um nome estiver na lista negativa, uma transação proposta com o consumidor pode ser automaticamente negada, independentemente de uma pontuação de fraude.

Barren e Staten (2003) realizaram simulações para quantificar a vantagem de se adotar as informações positivas e o compartilhamento. Para eles, essa situação viabilizaria uma queda da inadimplência da ordem de 43% em comparação ao sistema exclusivo de informações negativas.

Dessa forma, a utilização deste instrumento poderia reduzir os índices de inadimplência dos programas de habitação pública, já que permitiria ter informação histórica, ou positiva, ou de adimplemento dos possíveis beneficiários e avaliar se o histórico da vida comercial pregressa do cadastrado é bom ou ruim. Ou seja, a combinação da análise de risco de crédito com o uso de outros instrumentos como, por exemplo, o cadastro positivo permite que aqueles usuários classificados como maus pagadores sejam avaliados mais de perto de acordo com um histórico de pagamentos já realizados em sua vida. Logo, um indivíduo poderia ser beneficiado pelo programa habitacional, mesmo apresentando baixa renda ou sendo classificado como um mau pagador, caso apresente um bom histórico de pagamento.

2.6.3 Ajuste do valor das parcelas em períodos de recessão

Se os possíveis beneficiários forem classificados como bons pagadores e ainda possuírem um histórico de bom pagador, porém, em um determinado momento de uma recessão esse beneficiário ou uma pessoa de sua família venha a perda seu emprego reduzindo, assim, a renda daquela família, isso poderá impactar no aumento da inadimplência, porém existem mecanismo que podem ser utilizados como uma medida prudencial.

O PMCMV, por exemplo, possui uma espécie de seguro, o FGHab, e em caso de desemprego ou perda de renda familiar, pode ser utilizado pelo prazo máximo de 36 meses. Esse prazo será calculado de acordo com a renda familiar bruta. As prestações pagas pelo fundo deverão ser pagas com juros e correção monetária no final do período de utilização ou após 12 meses contados da última prestação assumida.

Eberly e Krishnamurthy (2014), no seu artigo sobre políticas de crédito habitacional eficiente em momento de crise financeira, descrevem que um contrato de hipoteca bem projetado deve reduzir os pagamentos durante as recessões e reduzir a dívida, quando os preços das casas caem. Isso implica modificar os empréstimos para reduzir os pagamentos durante a crise, em vez de reduzir os pagamentos ao longo da vida do contrato de hipoteca, como por meio da redução da dívida.

Sendo assim, a utilização do mecanismo de ajuste do valor das parcelas em períodos de recessão ou perda de renda pode auxiliar na redução da inadimplência reduzindo a exposição dos credores ao risco.

2.6.4 *Voucher* para habitação pública

Caso a análise de risco de crédito e o conhecimento do histórico de pagamento do cliente não permitam quadramento de uma pessoa com baixa renda como um possível beneficiário do crédito para aquisição da habitação própria outras ações podem ser tomadas para amparada aquelas

pessoas que estão em situação de pobreza e que não possuem uma habitação como, por exemplo, o enquadramento de pessoas não selecionadas para o crédito imobiliário em outros tipos de benefícios, como o fornecimento de *vouchers* para aluguel de moradia no mercado privado.

O fornecimento de *vouchers* para o aluguel de moradia no mercado privado é uma prática usual em outros países. Motley e Perry (2013) demostram em seu artigo as diversas formas de incentivos à habitação nos Estados Unidos como, por exemplo, os fornecimentos de *vouchers* para famílias de baixa renda, idosos e pessoas com deficiência para obter uma habitação no mercado privado.

Jacobs (2008) também demostra que esse instrumento é utilizado na Austrália como o nome de Sistemas de Incentivos do inquilino (TIS) que é uma camada de benefícios, tais como aluguel subsidiado, disponibilizados especificamente para os inquilinos que satisfaçam as condições necessárias para um contrato de arrendamento.

O uso de *vouchers* poderia então reduzir o índice de inadimplência do programa uma vez que o atendimento e os benefícios fornecidos à população de baixa renda seria de acordo com o que sua análise de risco e histórico de pagamento permitisse. Sendo assim, caso uma pessoal não tenha condição de ser atendida com o crédito subsidiado, pois não foi aprovada na análise de risco e também não possui um bom histórico de pagamentos, ela poderia ser amparada pelo fornecimento de *vouchers* para aluguel de moradia no mercado privado até que seu nível de renda e as condições da operação de crédito permitam seu enquadramento como uma boa pagadora.

2.7 CONSIDERAÇÕES SOBRE O RISCO DE CRÉDITO E MEDIDAS PRUDENCIAIS NA HABITAÇÃO PÚBLICA

Conforme demostrado, anteriormente, políticas e programas de habitação são fundamentais para o bem-estar das pessoas de baixa renda e comunidades, entretanto é necessário estruturar o programa habitacional corretamente e implementar algumas medidas prudenciais na concessão do crédito. Ao utilizar apenas a baixa renda como critério de seleção verifica-se uma inadimplência superior a 10% da carteira do PMCMV (BRASIL, 2016). A alta inadimplência é, potencialmente, prejudicial para a continuidade do programa tendo em vista que é necessário infinito aportes de recursos para que seja possível beneficiar outras pessoas.

A incorporação de aspectos prudenciais pode tonar os programas de habitação pública mais eficaz. Dentre algumas medidas prudenciais a serem adotadas, a análise de risco de crédito com um conjunto de modelos preditivos e suas técnicas subjacentes podem ajudar as instituições financeiras na concessão do crédito habitacional uma vez que essas técnicas decidem quem vai ter crédito e quanto de crédito que deve receber.

Outra medida importante seria a criação e compartilhamento de um banco de dados com informações de adimplemento dos possíveis beneficiários, o chamado cadastro positivo, com informações positivas e negativas. Este mecanismo poderia reduzir os índices de inadimplência dos programas de habitação pública, já que permitiria ter informação histórica ou positiva, ou de adimplemento dos possíveis beneficiários e avaliar se o histórico da vida comercial pregressa do cadastrado é bom ou ruim.

Mesmo com a adoção das medidas anteriores, pode ser que em algum momento os beneficiários venham a passar por um momento de dificuldade financeira. Sendo assim, o ajuste no valor das parcelas, em períodos de recessão, poderia fazer com que o beneficiário não se tornasse inadimplente. Auxiliando, portanto, na diminuição do descumprimento do contrato e reduzindo a exposição dos credores ao risco.

E para aqueles que não tiveram o seu crédito aprovado outras medidas poderia ser adotadas visando o bem estar da população de forma geral. Por exemplo, o fornecimento de *vouchers* para o aluguel de moradia no mercado privado é uma prática usual em outros países. Esse instrumento poderia reduzir o índice de inadimplência do programa, uma vez que o fornecimento de crédito à população de baixa renda seria de acordo com a análise de risco e o histórico de pagamento permitissem e, por outro lado, a população seria amparada com essa modalidade de benefício.

As medidas citadas, anteriormente, não se enquadram em um rol taxativo, mas apenas exemplificativo; porém, a adoção dessas medidas seria suficiente para reduzir o índice de inadimplência do programa habitacional, garantindo maior sustentabilidade dos recursos financeiros para novas gerações.

No estudo empírico desta dissertação, que é apresentado a seguir, demostra-se como a aplicação do melhor instrumento de *Machine Learning* pode reduzir a inadimplência do PMCMV de 11,80% para menos de 3%. O que reforça a necessidade de uma adequada análise de risco de crédito para os financiamentos da habitação pública. Desta maneira é imperativo que medidas prudenciais sejam adotadas nos programas de habitação pública, para que novas gerações possam ser beneficiadas com o crédito à aquisição de sua casa própria.

3 ESTUDO EMPÍRICO – MÉTODOS DE APRENDIZAGEM DE MÁQUINA PARA AVALIAÇÃO DE RISCO DE CRÉDITO NO PROGRAMA DE HABITAÇÃO PÚBLICA DO BRASIL.

3.1 MÉTODO E DESIGN DE PESQUISA

Este trabalho se propôs a identificar modelos de previsão que possam ser gerados por algoritmos, a partir de dados reais, para classificação de crédito entre devedores, sem uma preocupação com uma eventual fundamentação teórica para a inclusão de uma variável explicativa sobre o potencial de pagamento.

O trabalho foi desenvolvido com o uso de técnicas estatísticas tradicionais de análise do risco de crédito (LEWIS, 1992): Análise discriminante (HAND; OLIVER; LUNN, 1998) e Regressão logística (JR; LEMESHOW, 2004). E métodos de aprendizagem de máquina (MITCHELL et al., 1997): Árvore de decisão (BREIMAN et al., 1984), *Random Forest* (ALI et al., 2012), *Support Vector Machines* (VAPNIK; VAPNIK, 1998) (HSU et al., 2003) e classificadores *Ensemble* (BREIMAN, 1996) e (MITCHELL et al., 1997) (*Bagging e Adaboost*) para comparar a adequação, robustez e acurácia dos modelos para a previsão de inadimplência na base de dados em estudo.

3.2 BASE DE DADOS

Neste estudo, foi utilizada uma base de dados com informações das operações de créditos realizadas no âmbito do PMCMV e que está disponível em Brasil (2016). Os dados correspondem somente a linha de crédito destinado à pessoas físicas do PMCMV com taxas prefixada, pagamentos de prestações mensais e subsídio governamental.

Os dados de crédito coletados referem-se à base histórica do período de março de 2009 a dezembro de 2015. Destaca-se que o foco do trabalho envolve o estudo da aplicabilidade dos modelos de aprendizagem de máquina na análise de risco de crédito, a partir de dados do tomador de recursos, sem a preocupação de se considerar situações conjunturais do mercado de crédito como um todo. Em particular, este estudo analisa uma carteira de crédito de elevado nível de risco de crédito, na qual mais de 10% dos clientes estão inadimplentes.

Os empréstimos contraídos, nesse produto, têm um valor médio de R$ 51.000 (cinquenta e um mil reais), sendo que o tomador possui uma idade média de 38 (trinta e oito) anos e renda bruta mensal média de R$ 2.000 (dois mil reais). Os tomadores estão espalhados em todas as 27 (vinte e sete) unidades da federação e, proporcionalmente, divididos em: 54% do sexo feminino; e 46% do

sexo masculino. Do total de 2,24 milhões de contratos, 311 mil não estão com o pagamento em dia.

3.2.1 Seleção das amostras

As informações foram extraídas da base de dados com 2,2 milhões de contratos, dentre estes: adimplentes e inadimplentes. Para o processamento dos modelos, foi necessário retirar todos os registros em branco da base restando 1,5 milhões de registros qualificados. Ou seja, quando um registro apresentou qualquer dado em branco para qualquer uma das variáveis o registro foi excluído do universo, uma vez que sua presença inviabilizaria o processamento dos modelos.

Após a extração dos dados e seleção das variáveis, o banco de dados foi dividido em duas amostras aleatórias, uma para o treinamento do modelo, com 70% do banco de dados (1.070.448 observações), e outra para validação do modelo, com 30% (458.763 observações), buscando preservar as mesmas características de inadimplente tanto para a base de desenvolvimento quanto para a base de avaliação.

Todos os modelos foram treinados em um total de 1.070.448 observações e validados em 458.763 observações. Entretanto, Para avaliar a eficiência dos modelos, diante do tamanho da amostra, também foram realizados testes, considerando uma amostra da base com 300.000 observações, o que representa 20% da base de dados qualificada, e que possibilita a convergência de todos os modelos. Um número maior ou menor de observações poderia ser utilizado, entretanto, como o objetivo era avaliar a eficiência dos modelos em um número menor de observações, entende-se que 300.000 registros são adequados para essa avaliação.

3.2.2 Tratamento das variáveis

Dentro do universo de contratos, foram calculadas 4 (quatro) variáveis de inadimplência que são: atrasos superiores 29 dias; atrasos superiores 59 dias; atrasos superiores 89 dias e atrasos superiores a 119 dias; caracterizando-os como mau pagadores caso o contrato se enquadre em uma dessas faixas, já os demais contratos foram caracterizados como bons pagadores. A criação destas variáveis foi feita para construir cenários e comparar qual algoritmo se comporta melhor para a predição da inadimplência no intervalo de dias de atraso proposto. O critério de bom e mau pagador é necessário para investigar a acuidade classificatória dos modelos e a segregação dos diferentes tipos de tomadores. Foram eliminadas da base objeto de análise deste estudo os contratos liquidados e os contratos cancelados.

Lembrando que, eventuais variações no cenário econômico do país e estratégias políticas de incentivo ao financiamento imobiliário, não foram objeto desse estudo e, portanto, não foram avaliados.

As Tabelas 3.2 e 3.3 apresentam os nomes das variáveis independentes da base de dados utilizadas neste estudo. Os autores, Lawrence, Smith e Rhoades (1992), Lambrecht, Perraudin e Satchell (1997), Deng e Liu (2009), Shi, Zhang e Qiu (2013), Saberi et al. (2013), utilizaram variáveis semelhantes para avaliar o risco de crédito utilizando aprendizado de máquina ou para avaliar a inadimplência, conforme descrito na Tabela 3.1.

Variáveis Estudadas	**Estudos Referenciais**
Gênero.	Shi, Zhang e Qiu (2013) e Saberi et al. (2013).
Renda mensal.	Lawrence, Smith e Rhoades (1992), Lambrecht, Perraudin e Satchell (1997) e Deng e Liu (2009).
Estado civil.	Lawrence, Smith e Rhoades (1992), Lambrecht, Perraudin e Satchell (1997), Deng e Liu (2009), Shi, Zhang e Qiu (2013) e Saberi et al. (2013).
UF.	Deng e Liu (2009).
Município.	Deng e Liu (2009).
Valor da renda familiar.	Deng e Liu (2009).
Idade.	Lawrence, Smith e Rhoades (1992), Deng e Liu (2009) e Shi, Zhang e Qiu (2013).
Valor total do imóvel.	Lawrence, Smith e Rhoades (1992).
Valor financiado.	Lawrence, Smith e Rhoades (1992) e Shi, Zhang e Qiu (2013).
Prazo do financiamento.	Lawrence, Smith e Rhoades (1992), Shi, Zhang e Qiu (2013) e Saberi et al. (2013).
Taxa de juros.	Shi, Zhang e Qiu (2013).

Tabela 3.1: Estudos referenciais.

3.2.3 Variáveis independentes do cadastro do cliente

As variáveis independentes do cadastro dos clientes disponíveis na base de dados são:

Variável de Cadastro	Descrição	Tipo
Id.	Número identificador do tomador de crédito.	Informado
Gênero.	Sexo do tomador do crédito.	Informado
Data de nascimento.	Data de nascimento do tomador do crédito.	Informado
Renda mensal.	Renda mensal recebida pelo tomador do crédito.	Informado
Estado civil.	Estado civil do tomador do crédito.	Informado
UF.	Unidade da federação onde reside o tomador de crédito.	Informado
Valor da renda familiar.	Valor da renda familiar do tomador em salários mínimos.	Informado
Idade.	Idade do tomador de crédito.	Calculado

Fonte: Brasil (2016).

Tabela 3.2: Variáveis independentes do cadastro do cliente.

3.2.4 Variáveis independentes da operação de crédito contratadas

As variáveis independentes das operações de crédito contratadas para o financiamento imobiliário são:

Variável de Crédito	Descrição	Tipo
Adimplente.	Marcação utilizada para definir se o contrato é adimplente ou inadimplente.	Calculado
Data da inadimplência.	Data do início da inadimplência da operação de crédito.	Informado
Valor total do imóvel.	Valor do imóvel objeto da contratação de crédito.	Informado
Valor financiado.	Valor em Reais que foi financiado na operação de crédito.	Informado
Prazo do financiamento.	Prazo discriminado em meses utilizado no financiamento.	Informado
Taxa de juros.	Taxa de juros pactuada na contratação da operação de crédito.	Informado
Entrada.	Valor em Reais que foi dado de entrada no financiamento do imóvel.	Informado
Valor do subsídio concedido.	Valor do subsídio concedido pelo governo para operação de crédito.	Informado
Sistema de amortização.	Sistema de amortização utilizado na operação de crédito.	Informado

Fonte: Brasil (2016).

Tabela 3.3: Variáveis independentes da operação de crédito realizada.

3.2.5 Variável dependente

A variável dependente deste estudo é o "*default*" ou "Inadimplente" que representa o evento de não pagamento ou de pagamento da obrigação contratual em dia. Conforme Resolução 2682/99 do Banco Central do Brasil, um tomador é considerado "*default*" se sua classificação "média" for igual ou pior que "E". A Resolução estabeleceu que as instituições financeiras deveriam classificar suas exposições de crédito em nove níveis de risco de acordo com o sistema de classificação da Tabela 3.4

	AA	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**
Provisão (%)	0	0,5	1	3	10	30	50	70	100
Níveis de atraso (dias)	-	-	15-30	31-60	61-90	91-120	121-150	151-180	>180

Fonte: Resolução BACEN 2682/99.

Tabela 3.4: Classificação da inadimplência.

Para esse estudo foi considerado quatro cenários diferentes para o cálculo da variável "*default*", conforme Tabela 3.5:

Cenário	**Variável**	**Descrição**
Cenário 1 para execução dos algoritmos.	*Default*	Parcelas vencidas a mais de 29 dias.
Cenário 2 para execução dos algoritmos.	*Default*	Parcelas vencidas a mais de 59 dias.
Cenário 3 para execução dos algoritmos.	*Default*	Parcelas vencidas a mais de 89 dias.
Cenário 4 para execução dos algoritmos.	*Default*	Parcelas vencidas a mais de 119 dias.

Fonte: Brasil (2016).

Tabela 3.5: Variável dependente.

Em resumo, o estudo consistiu na execução de 3 (três) testes cujo os objetivos eram: Avaliar o comportamento dos modelos considerando diferentes intervalos de tempo para variável dependente *default* (30, 60, 90 e 120 dias); avaliar o comportamento dos modelos considerando um número menor de observações (300 mil observações); e avaliar o comportamento dos modelos sem o uso de variáveis que podem ser consideradas discriminatórias (Gênero, idade e estado civil), e que estavam disponíveis na base de dados. A Tabela 3.6 resume o testes realizados nesse estudo.

Descrição do teste	Objetivo
Teste 1.	Avaliar os modelos em diferentes intervalos de tempo para variável dependente *default* (30, 60, 90, 120 dias).
Teste 2.	Avaliar os modelos considerando um número menor de observações (300.000).
Teste 3.	Avaliar os modelos sem o uso de variáveis discriminatórias (Gênero, idade e estado civil).

Tabela 3.6: Variável dependente.

3.3 TÉCNICAS ESTATÍSTICAS

3.3.1 Análise discriminante

O objetivo da análise discriminante é a alocação de indivíduos em grupos bem definidos, no caso adimplentes e inadimplentes, a partir de uma função discriminante cujo resultado almejado da função é a obtenção de coeficientes para cada uma das variáveis independentes e que permitem determinar em qual grupo o indivíduo será classificado (adimplente ou inadimplente). A função discriminante de classificação dos adimplentes e inadimplentes, explicada por n variáveis independentes têm a seguinte forma descrita na Equação 3.1:

$$Y = b_0 + b_1.X_1 + b_2.X_2 + ... + b_n.X_n \tag{3.1}$$

Aonde,

Y é o escore discriminante, variável dependente;

b_i, para $i = 0, 1, 2, ..., n$, é o intercepto e os coeficientes que ponderam as variáveis independentes X_i na função discriminante; e

X_i, para $i = 1, 2, ..., n$, representam os valores da $i - sima$ variável discriminatória independente;

A classificação de cada indivíduo é efetuada por meio do cálculo da função discriminante estimada. O cliente deverá ser classificado como bom pagador se estiver mais próximo deste grupo do que do grupo dos maus pagadores, isto é, se a distância entre o seu escore discriminante e o centroide do grupo 1 (um) for menor que a distância entre o seu escore e o centroide do grupo 2 (dois), e no grupo dos maus pagadores no caso contrário (HAND; OLIVER; LUNN, 1998).

3.3.2 Regressão logística

A regressão logística caracteriza-se por descrever a relação entre variáveis independentes X_i e uma variável dependente Y, representando a presença ou ausência de uma característica. A regressão logística descreve o comportamento matemático de Y em função dos valores de X_i utilizando o método de estimação da máxima verossimilhança (JR; LEMESHOW, 2004). Desta forma, ao invés do escore da análise discriminante, que representa a proximidade em relação a um grupo, na regressão logística estima-se uma relação linear entre variáveis e a probabilidade de pertencer a um ou outro grupo, no caso, adimplente ou inadimplente. A expressão do modelo logístico é apresentada na Equação 3.2:

$$P(Y = 1) = \frac{1}{1 + e^{-z}} \tag{3.2}$$

aonde:

$$z = b_0 + \sum_{i}^{n} b_i X_i \tag{3.3}$$

$P(Y = 1)$ representa a probabilidade de um indivíduo ser inadimplente ou adimplente, dado seus valores das variáveis X_i; e

X_i representa as variáveis independentes.

3.3.3 Árvore de decisão

O modelo árvore de decisão compreende uma série de decisões lógicas, semelhante a um fluxograma, com nós de decisão indicando uma decisão a ser tomada em um atributo. A decisões são similares a regras *if-then* em que tomam como entrada uma situação descrita por um conjunto de atributos e retorna uma decisão, que é o valor predito para o valor de entrada (BREIMAN et al., 1984).

A probabilidade de que uma amostra arbitrária pertença à uma classe C_j é estimada pela Equação 3.4:

$$P_i = \frac{freq(C_j, S)}{|S|} \tag{3.4}$$

Onde $|S|$ é o número de amostras no conjunto S.

Para lidar com a classificação no modelo de árvore de decisão são introduzidos dois novos conceitos, Entropia e o Ganho. A entropia é o grau de pureza do conjunto e define o número de bits

necessários, em média, para representar a informação em falta, usando codificação ótima (TSAI; HSU; YEN, 2014).

Dado um conjunto S, com instâncias pertencentes à classe i, com probabilidade pi, a entropia é calculada conforme Equação 3.5:

$$Entropia(S) = \sum p_i log_2 p_i \tag{3.5}$$

O ganho de informação representa a diferença entre a informação necessária para identificar um elemento de A e a informação necessária para identificar um elemento de A após o valor do atributo S ter sido avaliado. Desta forma, *Gain*(S, A) é o ganho de informação devido ao atributo S (TSAI; HSU; YEN, 2014). O ganho (*gain*) define a redução na entropia e Ganho(S, A) significa a redução esperada na entropia de S, ordenando pelo atributo A. O ganho é dado pela Equação 3.6:

$$Ganho(S, A) = Entropia(S) - \sum \frac{|S|_v}{|S|}.Entropia(S_v) \tag{3.6}$$

3.3.4 Random Forest

Segundo Lantz (2015), o método RF, que se baseia em conjuntos de DT, combina versatilidade e potência em uma abordagem de aprendizado de máquina única. O método utiliza apenas uma pequena parte aleatória do conjunto completo de observações, podendo lidar com grandes conjuntos de dados.

RF é um classificador que consiste em uma coleção de árvores classificadoras estruturadas $h(x, \ominus_K), k = 1...$; aonde os $\ominus_K$ são vetores aleatórios independentes e identicamente distribuídos e cada árvore lança um único voto para a classe mais popular a partir dos dados de entrada x (BREIMAN, 2001).

3.3.5 *Support Vector Machines* – SVM

O objetivo de um SVM é criar uma fronteira plana, chamado de hiperplano, que leva a partições dos dados em lados razoavelmente homogêneos. A construção de um hiperplano de separação é dado por uma função *Kernel* $K(x_i, x_j)$ que é o produto dos vetores de entrada x_i e x_j (TSAI; HSU; YEN, 2014), conforme Equação 3.7 :

$$K(x_i, x_j) = \Phi(x_i).\Phi(x_j) \tag{3.7}$$

A função $Kernel$ pode estar associada a funções lineares, funções de base radial – *Radial Basis Function kernel (RBF)*, polinômios ou sigmoides. Neste estudo foram utilizadas as funções

linear (Equação 3.8) e *Radial Basis Function kernel* (Equação 3.9) dado que os ganhos de performance dos diferentes tipos de $kernel$ são incrementais (REN; HU; MIAO, 2016) e o uso do $Kernel$ radial e linear permite avaliar o SVM:

$$kernel\ Linear\colon K(x_i, x_j) = x_i^T x_j \tag{3.8}$$

$$RBF\colon K(x_i, x_j) = e^{(-\gamma||x_i - x_j||^2)} \tag{3.9}$$

Entre as principais limitações das SVMs encontram-se a sua sensibilidade a escolhas de valores de parâmetros e a dificuldade de interpretação do modelo gerado. O $Kernel$ ***Linear*** não fornece grande previsibilidade em conjuntos de dados não separáveis, já o $Kernel$ RBF fornece previsões superiores em casos não separáveis (VAPNIK; VAPNIK, 1998).

3.3.6 *Bagging*

O algoritmo de *Bagging*, conforme descrito por Breiman (1996), segue as seguintes etapas:

1. Construção de uma amostra aleatória, t, selecionada do conjunto de dados;
2. Cálculo do estimador C_t no conjunto de dados do passo 1;
3. Repita os dois primeiros passos por $t = 1, ..., T$, em que T é o total de iterações definidas pelo executor; e
4. A partir daí, cada classificador determina um voto, em que x comporta os dados de cada elemento do conjunto de treinamento, conforme Equação 3.10.

$$C(x) = T^{-1} \sum_{t=1}^{T} C_t(x) \tag{3.10}$$

A classe com maior votação é escolhida como classificação para cada elemento do conjunto de dados.

3.3.7 *AdaBoost*

O algoritmo empregado neste estudo segue Wang et al. (2011) descrito a seguir:

1. Inicia-se a distribuição dos pesos $D_1(i) = \frac{1}{m}$, para $t = 1, 2, ...T$; sendo que $i = 1, 2, ...m$; e D_t é a ponderação interativa;
2. Treina a base de aprendizado h_t para D usando a distribuição D_t, conforme Equação 3.11 ;

$$\varepsilon_i = Pr_i \cong D_i[h_t(x_i \neq y_i)] \tag{3.11}$$

3. Mensura-se o erro h_t, conforme Equação 3.12;

$$\alpha_t = \frac{1}{2} ln \frac{1 - \varepsilon_t}{\varepsilon_t} \tag{3.12}$$

4. Determina o peso de h_t, conforme Equação 3.13;

$$D_{t+1}(i) = \frac{D_t(i)}{Z_t} \begin{cases} e^{-\alpha_t} \; se \; h_t(x_i) = y_i \\ e^{\alpha_t} \; se \; h_t(x_i) \neq y_i \end{cases} \tag{3.13}$$

5. Atualiza a distribuição; e

6. É calculada $H(x)$, conforme Equação 3.14.

$$H(x) = sign \sum_{t=1}^{T} \alpha_t h_t(x) \tag{3.14}$$

3.4 INSTRUMENTOS ESTATÍSTICOS – R

A aplicação das técnicas estatísticas foram feitas no R, versão 3.3.1, utilizando as bibliotecas descritas na Tabela 3.7. O processamento foi realizado em um servidor com a seguintes especificações: Processador i7-3770 3.40Ghz, com 4 núcleos e 8 *threads* e 16GB RAM.

Modelo	**Biblioteca**
RF.	randomForest e H2O.
DT.	C50.
AdaBoost.	C50.
Bagging.	ipred.
SVM linear.	parallelSVM e e1071.
SVM radial.	parallelSVM e e1071.
Logistic Regression (LR).	gmodels.
Linear Discriminant Analysis (LDA).	MASS.

Tabela 3.7: Bibliotecas R.

Adicionalmente, também foram utilizadas as bibliotecas: "ROCR" para geração das curvas *Receiver Operating Characteristi*c (ROC) e cálculo da AUROC; e também a biblioteca "verification" que calcula o índice de *BRIER score*.

As bibliotecas H2O e parallelSVM foram utilizado para implementar o processamento em paralelo (multi-core), visto que o número de 1,5 milhões de observações inviabiliza o processamento dos modelos SVM e RF em série, conforme constatado durante o processamento dos dados. A implementação original do randomForest é lento e ineficiente no uso de memória. Ele não pode lidar por padrão com um grande número de categorias daí a necessidade de utilizar a biblioteca H2O. H2O é rápido, eficiente de memória e usa todos os núcleos e ainda trata as variáveis categóricas automaticamente. Cabe ressaltar que a limitação nada tem a ver com o R *per se*, visto que a implementação RF pela biblioteca randomForest que é ineficiente. Da mesma forma, o parallelSVM implementa o princípio de ensacamento para o algoritmo SVM possibilitando o processamento de uma forma paralela utilizando todos os núcleos da máquina.

3.5 PROCEDIMENTOS DE AVALIAÇÃO E VALIDAÇÃO DOS MODELOS

Vários autores (TASCHE, 2006; SUN; WANG, 2005; TSUKAHARA et al., 2016) mencionam diversos métodos que podem ser utilizados para validar o poder discriminante de um sistema de classificação de risco. Porém, observa-se certa convergência na escolha desses métodos. Há uma maior incidência do uso da curva ROC, cujo índice relacionado é a AUROC e cujo os componentes da curva são a especificidade (probabilidade de que estes irão corretamente classificar os não *default*) e sensibilidade (probabilidade de classificar corretamente o *default*). Verifica-se também a predominância do uso do *BRIER score*, que avalia as previsões de probabilidade de *default* individuais e também do teste KS, que quantifica a maior distância entre a distribuição acumulada de *default* e não–*default*.

As ferramentas estatísticas de validação de sistemas internos de classificação de risco de crédito apresentadas neste trabalho têm o objetivo de medir o poder discriminatório de tais sistemas em relação a contrapartes boas e más. A seguir, são apresentados os procedimentos utilizados para a avaliar e validar os modelos de classificação.

3.5.1 Matriz de confusão

Para a avaliação dos modelos foram utilizadas medidas-padrão para classificação de crédito (WANG et al., 2011). Estas medidas são: a precisão média, o erro tipo I e o erro tipo II; que podem ser explicadas a partir de uma matriz de confusão, como mostrado na Tabela 3.8.

Valor Observado			
Valor Predito	Y = 1	Y = 1 VP (Verdadeiro Positivo)	Y = 0 FP (Falso Positivo)
	Y = 0	FN (Falso Negativo)	VN (Verdadeiro Negativo)

Tabela 3.8: Matriz de confusão.

A medidas são definidas da seguinte forma:

Precisão Média ou Acurácia: Porcentagem de amostras positivas e negativas classificadas corretamente sobre a soma de amostras positivas e negativas e pode ser calculado conforme Equação 3.15.

$$Preciso\ Mdia = \frac{VP + VN}{VP + FP + FN + VN} \tag{3.15}$$

Sensibilidade: Porcentagem de amostras positivas classificadas corretamente sobre o total de amostras positivas. É a capacidade do sistema em predizer corretamente a condição para casos que realmente a têm e pode ser calculado conforme Equação 3.16.

$$Sensibilidade = 1 - Erro\ Tipo\ I = 1 - \frac{FP}{VP + FP} \tag{3.16}$$

Especificidade: Porcentagem de amostras negativas identificadas corretamente sobre o total de amostras negativas. É a capacidade do sistema em predizer corretamente a ausência da condição para casos que realmente não a tem e pode ser calculado conforme Equação 3.17.

$$Especificidade = 1 - Erro\ Tipo\ II = 1 - \frac{FN}{FN + VN} \tag{3.17}$$

3.5.2 *Receiver Operating Characteristic* - ROC

ROC é um método visual que pode ser construído a partir de duas amostras de escores, uma para casos anormais, como devedores inadimplentes, e outra para casos normais. A curva ROC é um teste que busca mostrar a relação, normalmente antagônica, entre a sensibilidade e a especificidade. A área da curva está relacionada com a distribuição de frequência de eventos de inadimplência e não inadimplência e permitem quantificar a exatidão de um teste, pois, quanto maior a área sob a curva ROC, maior é a precisão (ENGELMANN; HAYDEN; TASCHE, 2003). Portanto, a curva ROC será útil na comparação dos testes entre os algoritmos, sendo um algoritmo mais preciso quanto maior for a área sob a curva ROC, conforme Figura 3.1.

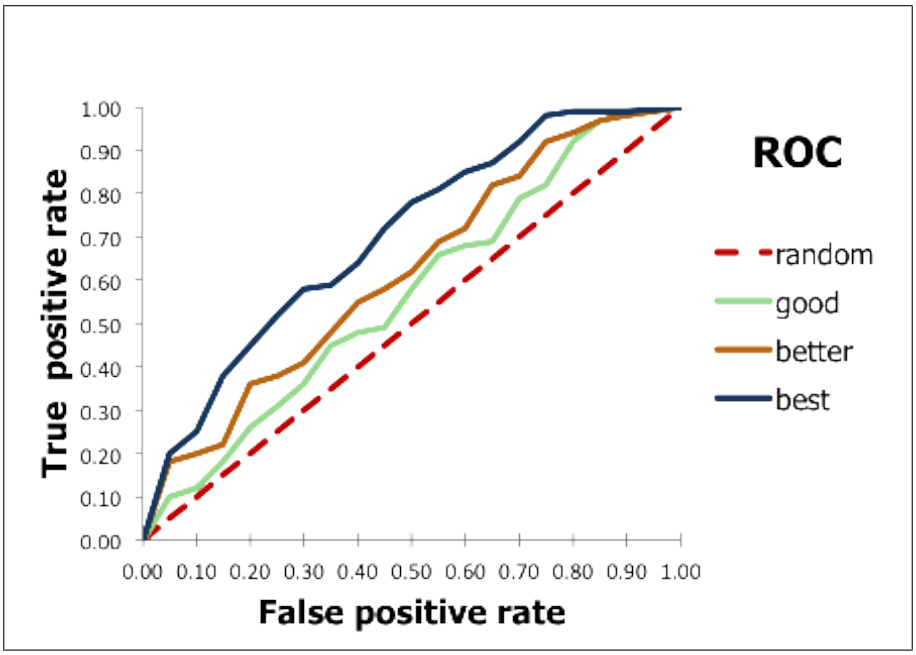


Figura 3.1: Curva ROC.

Para calcular a área da curva ROC é necessário conhecer o valor do score C, que separa os inferiores a C como potenciais inadimplentes e os superiores a C como potenciais não-inadimplentes.

$$HR(C) = \frac{H(C)}{N_D} \tag{3.18}$$

Aonde:

$H(C)$ é o número de inadimplentes com score inferior a C; e

N_D é o total de inadimplentes.

Isso feito é necessário conhecer taxa de falsos alarmes $FAR(C)$ conforme Equação 3.19:

$$FAR(C) = \frac{F(C)}{N_{ND}} \tag{3.19}$$

Aonde:

$F(C)$ corresponde ao número de não-inadimplentes com pontuações menores que o índice C; e

N_{ND} corresponde ao número total de não-inadimplentes.

A partir de agora é possível chegar ao valor da área da curva ROC, onde numa situação ideal a área da curva ROC seria igual a 1 e pode ser calculado utilizando a Equação 3.20:

$$AUROC = \int_0^1 HR(FAR)d(FAR) \tag{3.20}$$

A área tem variações de 0,5 a 1,0. O valor de 0,5 é para um modelo aleatório, sem poder discriminativo e 1.0 para um modelo perfeito. Ela situa-se entre 0,5 e 1,0 para qualquer modelo de avaliação razoável na prática. Um sistema de classificação plausível é acima de 0,5, senão é um sistema de seleção aleatória.

Tsukahara et al. (2016), no estudo sobre técnicas utilizadas para a validação de modelos de risco de crédito, demostrou que AUROC é a métrica mais sensíveis a mudanças nas condições de mercado. Ostrowski e Reichling (2011) descobriram que AUROC pode, em certas circunstâncias, levar a conclusões erradas e causar modelos de baixo desempenho bem classificado com este indicador. Sendo assim, resta a necessidade de utilizar outros instrumentos de validação, logo, foi realizado o teste KS e *BRIER Score* para todos os modelos.

3.5.3 *Kolmogorov–Smirnov – KS*

Os estudos de Lilliefors (1967) apresentam um processo para testar um conjunto de n observações a partir de uma distribuição normal. De forma simplificada, Lilliefors (1967) propõe um teste de hipótese para a medida D, que é a diferença absoluta entre a função de distribuição acumulada da amostra; e a função de distribuição acumulada normal com média e variância iguais aos da amostra. Para validar modelos de risco de crédito, vale a pena notar que o objetivo não é analisar a normalidade de uma distribuição, mas sim para verificar se o modelo pode distinguir os inadimplentes dos adimplentes. Para tal finalidade, o teste de KS pode ser utilizado para quantificar a maior distância entre a distribuição acumulada de inadimplentes e adimplentes, calculado conforme Equação 3.21:

$$kS = max|F_D(S) - F_{ND}(S)| \tag{3.21}$$

Aonde:

F_D corresponde à função de distribuição acumulada de casos *default*;

F_{ND} corresponde à função de distribuição acumulada de casos não–*default*; e

S corresponde à pontuação.

O teste KS é utilizado no mercado financeiro como um dos indicadores de eficiência de modelos de *credit scoring*, sendo que o mercado considera um bom modelo aquele que apresente

um valor de KS igual ou superior a 30. O teste KS busca encontrar a diferença máxima entre duas distribuições acumuladas, se houver uma diferença de pelo menos 30 entre a distribuição dos adimplentes e dos inadimplentes, o modelo consegue discriminar satisfatoriamente os dois grupos, pois a diferença é grande o suficiente para diferenciar os grupos, logo, foram aceitáveis todos os resultados gerados o que implica na viabilidade de implantação do modelo proposto para análise de crédito da instituição, podendo ter impactos positivos na estratégia de concessão de crédito.

3.5.4 *BRIER Score*

BRIER Score é uma função que mede a precisão das previsões probabilísticas. Quanto menor a pontuação *BRIER* é para um conjunto de previsões, melhores são as previsões. O *BRIER Score* pode ser calculada de acordo com a Equação 3.22.

$$BRIER = \frac{1}{N}\sum_{i=1}^{N}(Pi - Oi)^2 \tag{3.22}$$

Aonde:

Pi é a probabilidade de ocorrer o evento; e

Oi é uma variável binaria (1/0) e significa se o evento foi observado ou não.

3.6 RESULTADOS E DISCUSSÃO

3.6.1 Teste 1 – Avaliação dos modelos em diferentes intervalos de tempo para variável dependente *default* (30, 60, 90, 120 dias).

Para avaliar a performance dos modelos diante dos diferentes critérios utilizados para a variável dependente, *default* ou inadimplente, os modelos foram executados com quatro critérios diferentes de dias para caracterizar a operação como inadimplente, quais sejam: 30, 60, 90 e 120 dias. Será iniciado a apresentação dos resultados para a variável *Default* igual a 90 dias, uma vez que corresponde ao critério utilizado pelo Banco Central do Brasil (BACEN) para definir uma determinada operação como inadimplente.

3.6.1.1 Avaliação dos modelos para variável dependente *default* igual a 90 dias

Das 445.761 observações separadas para avaliação do RF, o modelo previu corretamente que 9.368 são inadimplentes e 427.539 são adimplentes, resultando em uma precisão média de 95,4%. O modelo DT foi capaz de prever corretamente 5.636 inadimplentes e 427.685 adimplentes, resultando em uma precisão média de 94,45%. O modelo *AdaBoost* foi capaz de prever corretamente 8.602 inadimplentes e 429.286 adimplentes, resultando em uma precisão média de 95,45%. O modelo *Bagging* foi capaz de prever corretamente 14.515 inadimplentes e 429.031 adimplentes, resultando em uma precisão média de 96,68%. O modelo LR foi capaz de prever corretamente 5.690 inadimplentes e 419.448 adimplentes, resultando em uma precisão média de 92,67%. O modelo LDA foi capaz de prever corretamente 834 inadimplentes e 429.486 adimplentes, resultando em uma precisão média de 93,80%.

Já o modelo SVM, *kernel* radial e linear, não foi capaz de classificar corretamente os adimplentes e inadimplentes para os dados da base que estava sendo testada. Neste modelo todas a 445.761 observações foram classificadas como adimplentes, resultando em uma precisão média de 93,99%, porém apresentando especificidade igual a 0. Ressalte-se que foram trabalhados diversos parâmetros do modelo SVM para que ele pudesse resultar em uma melhor classificação, entretanto o modelo SVM não conseguiu obter uma performance satisfatória para essa base de dados em específico. Os dados de precisão média, sensibilidade e especificidade estão resumidos na Tabela 3.9

A partir dos dados apresentados na Tabela 3.9 foi traçado um comparativo entre a especificidade e sensibilidade dos modelos em avaliação. A Figura 3.2 demostra que os modelos *Bagging* seguido por *AdaBoost* e RF apresentaram melhores resultados. Em contrapartida, os modelos SVM e LDA obtiveram piores resultados como classificadores. Por essa figura é possível notar a superioridade das técnicas de aprendizado de máquina no distanciamento entre os pontos no gráfico dos algoritmos de aprendizado de máquina e do modelo tradicional, Regressão Logística.

Modelo	Precisão Média	Sensibilidade	Especificidade
RF.	95,24%	95,91%	72,05%
DT.	94,45%	95,12%	61,77%
AdaBoost.	95,45%	95,76%	82,01%
***Bagging*.**	**96,68%**	**97,04%**	**87,14%**
LR.	92,67%	95,04%	32,67%
LDA.	93,80%	94,14%	33,08%
SVM linear.	93,99%	93,99%	0
SVM radial.	93,99%	93,99%	0

Tabela 3.9: Sensibilidade, especificidade e precisão média dos modelos para *default* superior a 90 dias.

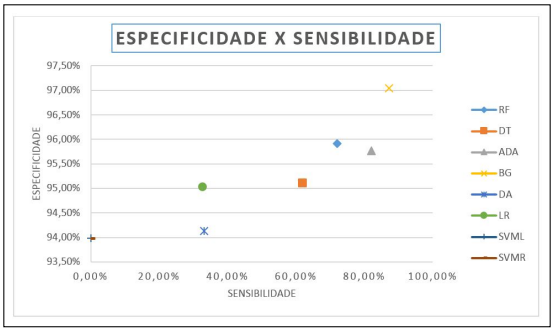
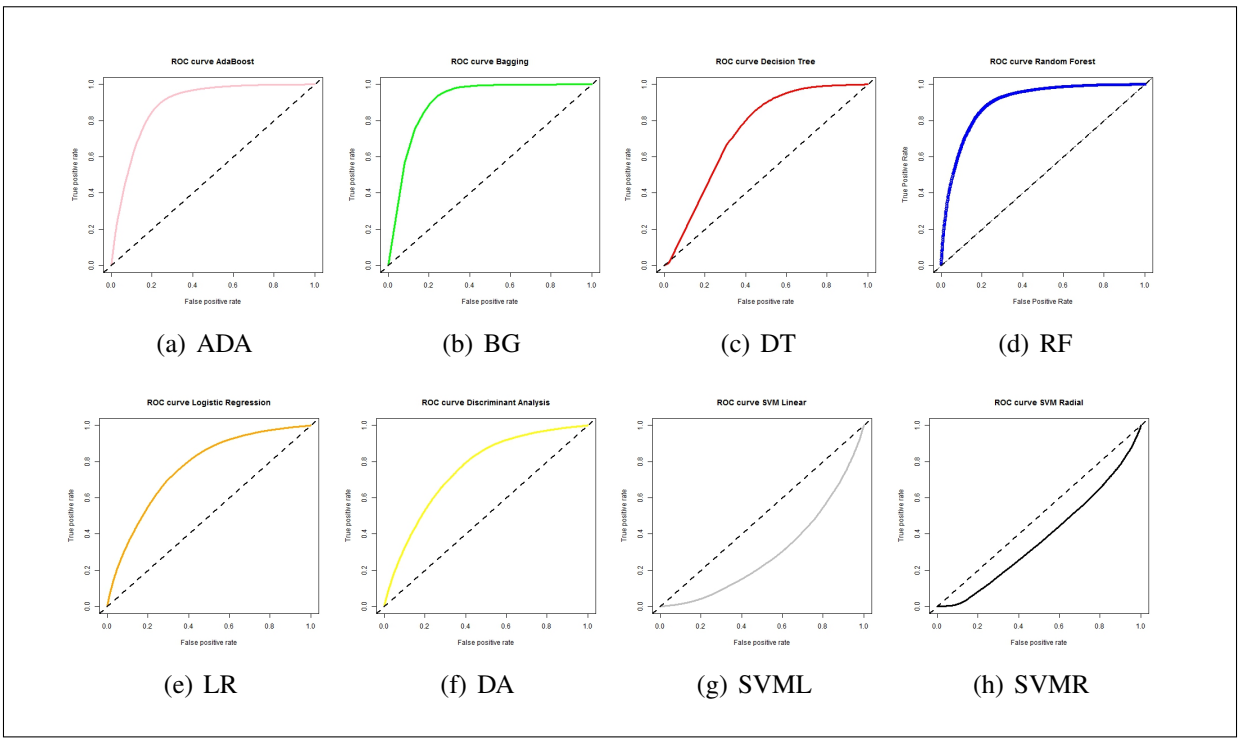


Figura 3.2: Sensibilidade *vs* Especificidade dos modelos para *default* superior a 90 dias.

A Figura 3.4 apresenta as curvas ROC para os modelos avaliados. A AUROC para o algoritmo *Bagging* foi a maior de todos os modelos, 90,43%, seguido por RF, 89,92%, e *AdaBoost* 88,85%.



(a) ADA (b) BG (c) DT (d) RF

(e) LR (f) DA (g) SVML (h) SVMR

Figura 3.3: Curva ROC para variável *default* superior a 90 dias.

A Figura 3.4 apresenta o *benchmarking* das curvas ROC para os modelos em avaliação, considerando a variável *default* superior a 90 dias. É interessante observar que os três melhores métodos (*Bagging*, *AdaBoost* e RF) são baseados em classificadores *ensemble* de estimadores.

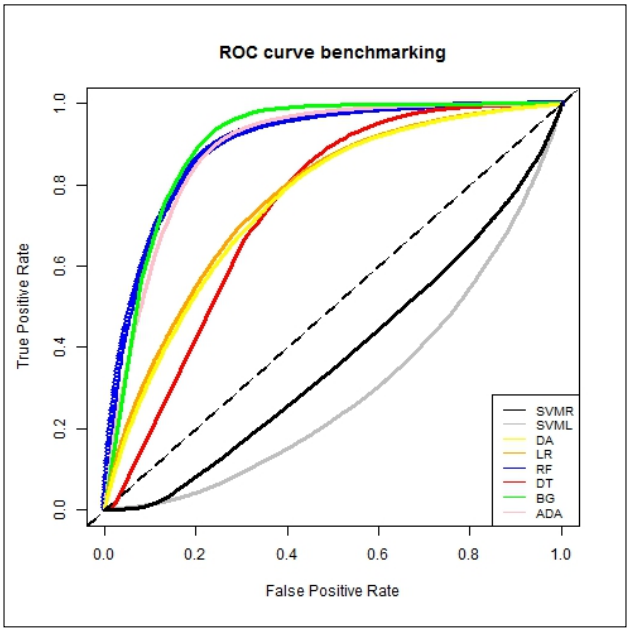


Figura 3.4: Benchmarking modelos *default* superior a 90 dias.

Observa-se relativa convergência dos resultados dos índices *BRIER score* e KS com os

resultados demostrados anteriormente para curva ROC e para a AUROC. *Bagging* obteve o melhor valor para o índice de *BRIER score*, 0,03032, seguido por *AdaBoost* com 0,03971 e RF com 0,04146. Já o teste KS apresentou melhor resultado para *Bagging*, 0,69, seguido por RF, 0,66, e *AdaBoost*, 0,65. A Tabela 3.10 apresenta os resultados dos índices *BRIER score* e KS para os algoritmos avaliados.

Modelo	**AUC**	**KS**	***BRIER Score***
RF.	0.8992894	0.6619846	0.04146
DT.	0.7382974	0.4071035	0.04840
AdaBoost.	0.8885443	0.6565465	0.03971
***Bagging*.**	**0.9043971**	**0.6925993**	**0.03032**
LR.	0.7657149	0.4053571	0.05270
LDA.	0.7568453	0.3948803	0.05830
SVM linear.	0.2973585	0	0.84830
SVM radial.	0.3762618	2.319255e-06	0.84700

Tabela 3.10: Resultados dos modelos com a variável *default* superior a 90 dias.

Para avaliar o custo de processamento de cada algoritmo foi monitorado o tempo gasto para treinar o modelo juntamente com o tempo necessário para realizar a predição na base de teste. Dentre os três algoritmos que melhor performaram (*Bagging*, *AdaBoost* e RF) RF foi o que teve menor custo computacional, levando aproximadamente 28 minutos para sua execução. Já o algoritmo *AdaBoost* teve o segundo menor custo computacional. Cabe ressaltar que a execução do RF foi realizada utilizando o processamento multi-core (processamento em todos os núcleos do chip da CPU) enquanto que o *AdaBoost* foi realizado o processamento de forma serial. A Tabela 3.11 demostra o tempo de execução gasto para todos os demais algoritmos.

Modelo	**Usuário (ms)**	**Sistema (ms)**	**Decorrido (ms)**
RF.	12.07	1.56	1681.69
DT.	67.34	0.75	68.11
AdaBoost.	3030.00	17.44	3048.15
Bagging.	13279.72	27.22	13229.65
LR.	10948.36	667.56	4399.77
LDA.	17.03	3.21	13.27
SVM linear.	3229.97	21.37	13585.91
SVM radial.	9400.05	21.46	25609.03

Tabela 3.11: Tempo decorrido em milissegundos para cada modelo com a variável *default* superior a 90 dias.

3.6.1.2 Avaliação dos modelos para variável dependente *default* igual a 30 dias

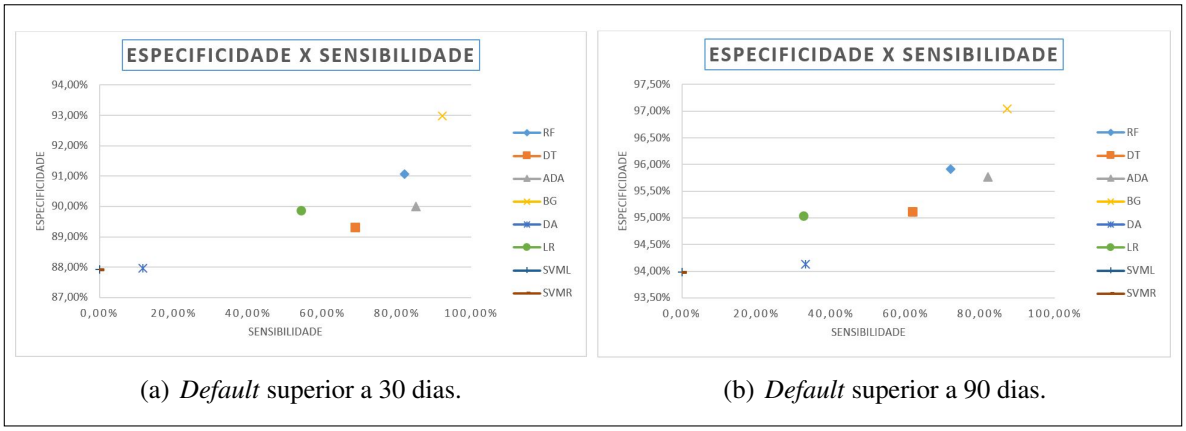
Para avaliar a performance dos modelos utilizando o critério de 30 dias para definir a variável dependente, *default* ou inadimplente, os modelos foram executados seus resultados comparados com os resultados obtidos para variável dependente, *default*, igual a 90 dias, uma vez que corresponde ao critério utilizado pelo BACEN para definir uma determinada operação como inadimplente.

Ao utilizar o critério de 30 dias para definir a variável *default* verifica-se que todos os modelos tiveram um redução significativa na precisão média, conforme evidencia-se no comparativo PM 90 *vs* PM 30 da Tabela 3.12. O modelo menos sensível a essa nova caracterização da variável *default* foi *Bagging*, com uma redução de 3,88% na precisão média, por outro lado, LDA foi o modelo que apresentou a maior perda na precisão média, 6,56%.

Modelo	PM 90	PM 30	Sens. 90	Sens. 30	Esp. 90	Esp. 30
RF.	95,24%	90,66%	95,91%	91,06%	72,05%	82,01%
DT.	94,45%	88,80%	95,12%	89,30%	61,77%	68,75%
AdaBoost.	95,45%	89,86%	95,76%	89,99%	82,01%	85,13%
***Bagging*.**	**96,68%**	**92,93%**	**97,04%**	**92,98%**	**87,14%**	**92,15%**
LR.	92,67%	87,75%	95,04%	89,85%	32,67%	54,22%
LDA.	93,80%	87,64%	94,14%	87,96%	33,08%	11,67%
SVM linear.	93,99%	87,93%	93,99%	87,93%	0	0
SVM radial.	93,99%	87,93%	93,99%	87,93%	0	0

Tabela 3.12: Sensibilidade, especificidade e precisão média dos modelos para *default* superior a 30 dias *vs* 90 dias.

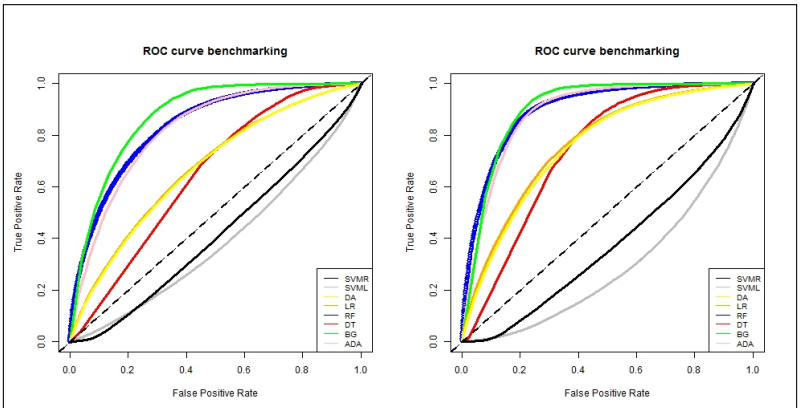
A Figura 3.5 demostra a relação de sensibilidade e especificidade dos modelos para a variável *default* superior a 30 dias em comparação com a relação de especificidade e sensibilidade para a variável *default* superior a 90 dias. Observa-se que apesar dos modelos terem uma redução geral na sua precisão média, quando comparado com a variável *default* superior a 90 dias, verifica-se que foi mantido a hierarquia dos resultado obtidos anteriormente para os modelos (*Bagging*, RF e *AdaBoost*).



(a) *Default* superior a 30 dias. (b) *Default* superior a 90 dias.

Figura 3.5: Sensibilidade *vs* Especificidade dos modelos para *default* superior a 30 dias *vs* 90 dias.

O comparativo das curvas ROC dos modelos com a variável *default* superior a 30 dias *vs* 90 dias, Figura 3.6, possibilita visualizar a redução da AUROC dos modelos quando a variável *default* é caracterizada como atrasos superiores a 30 dias. Este fato pode ser explicado porque com o passar do tempo tornar-se cada vez mais improvável que o devedor pague suas obrigações de crédito ao banco, ficando mais nítido os atributos dos devedores.



(a) *Default* superior a 30 dias. (b) *Default* superior a 90 dias.

Figura 3.6: Curva ROC comparativa com variável *default* superior a 30 dias *vs* 90 dias.

Quando se visualiza os índices AUROC, KS e *BRIER Score* dos modelos com variável *default* definida como superior a 30 dias de atraso *vs* 90 dias, também se evidencia a redução na capacidade de classificação dos modelos, conforme pode ser visto na Tabela 3.13.

Modelo	AUC 90	AUC 30	KS 90	KS 30	*BRIER* 90	*BRIER* 30
RF.	0.8992894	0.8355241	0.6619846	0.5094635	0.04146	0.08576
DT.	0.7382974	0.6461897	0.4071035	0.2418152	0.04840	0.09866
AdaBoost.	0.8885443	0.8214115	0.6565465	0.5014284	0.03971	0.08387
***Bagging*.**	**0.9043971**	**0.8756473**	**0.6925993**	**0.6000578**	**0.03032**	**0.06229**
LR.	0.7657149	0.6744533	0.4053571	0.2538376	0.05270	0.10210
LDA.	0.7568453	0.6711700	0.3948803	0.2480742	0.05830	0.11160
SVM linear.	0.2973585	0.3887173	0	0	0.84830	0.64050
SVM radial.	0.3762618	0.4152368	2.319255e-06	0	0.84700	0.65920

Tabela 3.13: Resultados dos modelos com a variável *default* superior a 30 dias *vs* 90 dias.

3.6.1.3 Avaliação dos modelos para variável dependente *default* igual a 60 dias

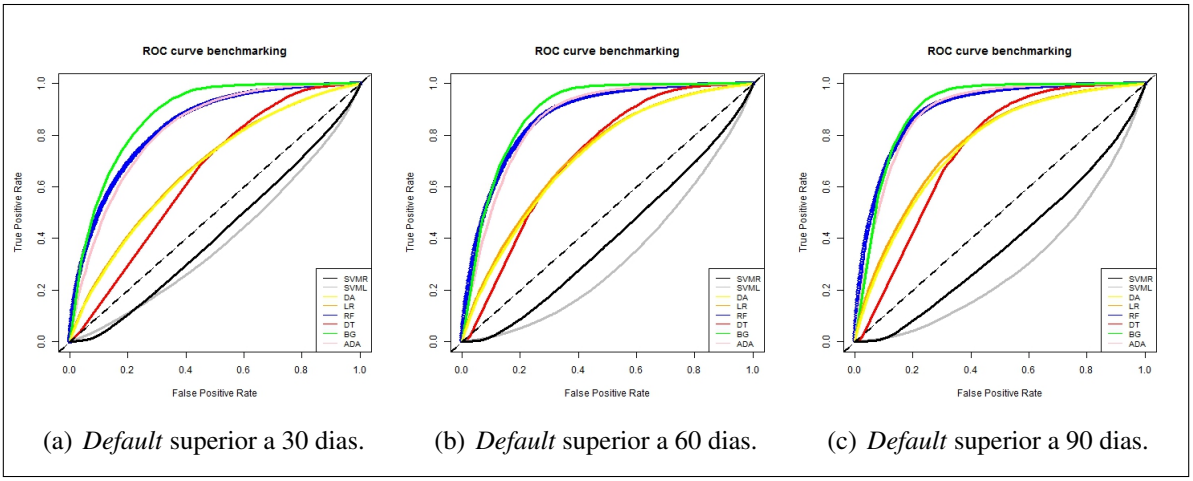
Ao utilizar a variável dependente *default* caracterizando-a como atrasos superiores a 60 dias, verificamos uma evolução na precisão média dos modelos quando comparado com o critério de 30 dias, conforme Tabela 3.14, entretanto, observa-se também que a precisão média dos modelos, de formal geral, foi inferior aos resultados obtidos quando caracterizado a variável *default* superior a 90 dias. O modelo de LDA foi o que obteve maior ganho de precisão com um aumento de 4,92%. O modelo que obteve o menor ganho foi o *Bagging* com uma elevação de 2,88% na precisão média. *Bagging* também foi o modelo com maior precisão média tanto para o critério de 30 quanto para o critério de 60 e 90 dias, o que pode demostrar efetiva capacidade do *Bagging* para classificação dos adimplentes e inadimplentes no banco de dados em estudo.

Modelo	PM 90	PM 30	PM 60
RF.	95,24%	90,66%	93,94%
DT.	94,45%	88,80%	92,96%
AdaBoost.	95,45%	89,86%	93,97%
***Bagging*.**	**96,68%**	**92,93%**	**95,68%**
LR.	92,67%	87,75%	92,18%
LDA.	93,80%	87,64%	91,24%
SVM linear.	93,99%	87,93%	92,43%
SVM radial.	93,99%	87,93%	92,43%

Tabela 3.14: Sensibilidade, especificidade e precisão média dos modelos para *default* superior a 60 dias *vs* 30 dias e 90 dias.

Visualmente, Figura 3.7, também pode-se verificar uma melhora geral das curvas ROC dos modelos com a variável *default* superior a 60 dias, quando comparado com o critério de 30 dias para a mesma. As curvas ROC dos modelos *AdaBoost* e RF se aproximaram do modelo *Bagging*, diferente do que se observa quando a variável *default* é superior a 30 dias.

Figura 3.7: Curva ROC comparativa com variável *default* superior a 60 dias *vs* 30 dias e 90 dias.



(a) *Default* superior a 30 dias. (b) *Default* superior a 60 dias. (c) *Default* superior a 90 dias.

Quando se observa os resultados da AUROC dos modelos utilizando o critério da variável *default* superior a 60 dias, Tabela 3.15, é possível identificar a melhora dos índices, de forma geral, quando comparado com os resultados obtidos com a variável *default* superior a 30 dias. O modelo *Bagging* foi o que apresentou melhor resultado (AUC igual a 0.88) seguido por RF e *AdaBoost*.

Modelo	AUC 90	AUC 30	AUC 60
RF.	0.8992894	0.8355241	0.8735597
DT.	0.7382974	0.6461897	0.717630
AdaBoost.	0.8885443	0.8214115	0.8621694
***Bagging*.**	**0.9043971**	**0.8756473**	**0.8895560**
LR.	0.7657149	0.6744533	0.7232733
LDA.	0.7568453	0.6711700	0.7164254
SVM linear.	0.2973585	0.3887173	0.3281406
SVM radial.	0.3762618	0.4152368	0.3997688

Tabela 3.15: Resultados AUC dos modelos com a variável *default* superior a 60 dias *vs* 30 dias e 90 dias.

Os índices KS também apresentaram relativa melhora se comparado à variável *default* superior a 30 dias, entretanto, não foram capazes de superar os modelos com a variável *default* superior a 90 dias. *Bagging* foi o que obteve o melhor resultado, seguido por RF e *AdaBoost*, conforme pode ser visto na Tabela 3.16.

Modelo	KS 90	KS 30	KS 60
RF.	0.6619846	0.5094635	0.6027268
DT.	0.4071035	0.2418152	0.3398986
AdaBoost.	0.6565465	0.5014284	0.6005207
***Bagging*.**	**0.6925993**	**0.6000578**	**0.6499971**
LR.	0.4053571	0.2538376	0.3332069
LDA.	0.3948803	0.2480742	0.3217469
SVM linear.	0	0	0
SVM radial.	2.319255e-06	0	0

Tabela 3.16: Resultados KS dos modelos com a variável *default* superior a 60 dias *vs* 30 dias e 90 dias.

Na mesma linha dos resultados obtidos anteriormente para variável *default* superior a 60 dias, o *BRIER Score* dos modelos teve uma relativa melhora, quando comparado com os resultados obtidos com a variável *default* superior a 30 dias. *Bagging* apresentou o melhor resultado e verifica-se que *AdaBoost* teve resultado melhor que RF, diferente dos resultados anteriores para AUROC e KS. O comparativo dos índices para os diversos critérios da variável *default* estão descritos na Tabela 3.17.

Modelo	*BRIER Score* 90	*BRIER Score* 30	*BRIER Score* 60
RF.	0.04146	0.08576	0.05330
DT.	0.04840	0.09866	0.06159
AdaBoost.	0.03971	0.08387	0.05153
***Bagging*.**	**0.03032**	**0.06229**	**0.03900**
LR.	0.05270	0.10210	0.06612
LDA.	0.05830	0.11160	0.07171
SVM linear.	0.84830	0.64050	0.83370
SVM radial.	0.84700	0.65920	0.82930

Tabela 3.17: Resultados *BRIER Score* dos modelos com a variável *default* superior a 60 dias *vs* 30 dias e 90 dias.

3.6.1.4 Avaliação dos modelos para variável dependente *default* igual a 120 dias

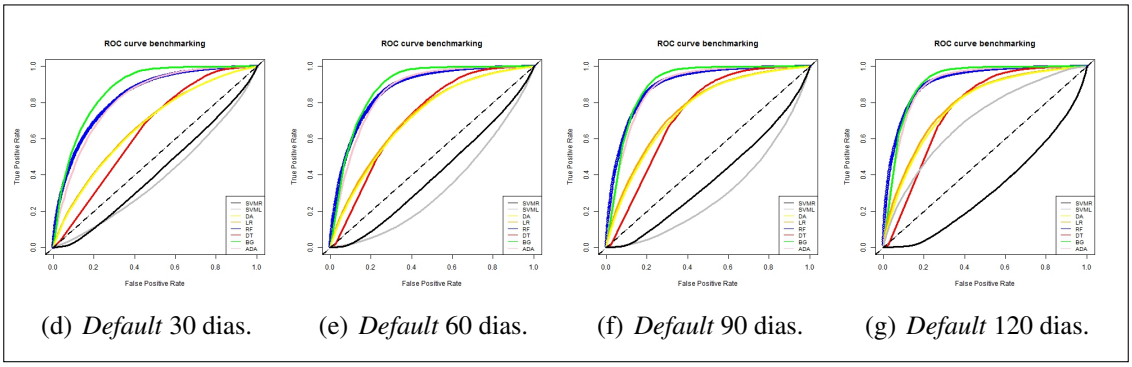
Por fim, para avaliar o impacto dos diferentes valores da variável *default* foi utilizado o parâmetro de 120 dias de atraso. Observa-se que, apesar dos modelos apresentarem uma melhor precisão média, percebe-se um menor incremento nos ganhos, quando comparado com os resultados obtidos utilizando outros critérios para a variável *default*. LR foi o modelo que teve melhor incremento na precisão média, aumento de 0,84%, e *Bagging* foi o modelo que teve menor ganho, aumento de 0,46%, entretanto *Bagging* também foi o modelo que apresentou melhor precisão média, 97,12%, conforme pode ser visto na Tabela 3.18.

A curva ROC dos modelos, Figura 3.8, demostra o incremento na AUROC à medida que o valores que caracterizam a variável *default* aumentam. Também é possível identificar que os

Modelo	PM 90	PM 30	PM 60	PM 120
RF.	95,24%	90,66%	93,94%	95,88%
DT.	94,45%	88,80%	92,96%	95,15%
AdaBoost.	95,45%	89,86%	93,97%	96,16%
***Bagging*.**	**96,68%**	**92,93%**	**95,68%**	**97,12%**
LR.	92,67%	87,75%	92,18%	93,46%
LDA.	93,80%	87,64%	91,24%	94,44%
SVM linear.	93,99%	87,93%	92,43%	93,42%
SVM radial.	93,99%	87,93%	92,43%	94,67%

Tabela 3.18: Sensibilidade, especificidade e precisão média dos modelos para *default* superior a 120 dias *vs* 30, 60 e 90 dias.

ganhos obtidos no tamanho da área sob a curva ROC são menores de 90 dias para 120 dias. Porém, há que se observar o incremento na curva ROC do modelo SVM linear. Tal fato pode ser justificado já que o aumento no número de dias de inadimplência para a caracterização da variável *default* evidencia melhor as características dos perfis dos inadimplentes permitindo que o modelo SVM linear obtenha melhores resultados na classificação dos inadimplentes.



(d) *Default* 30 dias. (e) *Default* 60 dias. (f) *Default* 90 dias. (g) *Default* 120 dias.

Figura 3.8: Curva ROC comparativa com variável *default* superior a 120 dias *vs* 30, 60 e 90 dias.

O comparativo dos índices da AUROC para os modelos diante dos diferentes valores para a variável dependente demostra claramente o incremento no aumento da área sob a curva à medida que os valores para a variável dependente aumenta. Interessante notar que a AUROC do modelo RF apresentou melhor resultado (0.9159156) do que o modelo *Bagging* (0.9159105), diferente do que vinha sendo apresentado nos índices anteriores para variável *default* igual a 30, 60 e 90 dias, conforme pode ser visto na Tabela 3.19.

Modelo	AUC 90	AUC 30	AUC 60	AUC 120
RF.	0.8992894	0.8355241	0.8735597	0.9159156
DT.	0.7382974	0.6461897	0.717630	0.7646174
AdaBoost.	0.8885443	0.8214115	0.8621694	0.9059096
***Bagging*.**	**0.9043971**	**0.8756473**	**0.8895560**	**0.9159105**
LR.	0.7657149	0.6744533	0.7232733	0.7933989
LDA.	0.7568453	0.6711700	0.7164254	0.7839488
SVM linear.	0.2973585	0.3887173	0.3281406	0.7007318
SVM radial.	0.3762618	0.4152368	0.3997688	0.3226471

Tabela 3.19: Resultados AUC dos modelos com a variável *default* superior a 120 dias *vs* 30, 60 e 90 dias.

A mesma tendência de incremento nos índices à medida que aumenta o valor da variável *default* também se aplica para o índice KS, conforme se evidencia na Tabela 3.20, entretanto, observa-se que o modelo *Bagging* foi o que apresentou melhor resultado para a variável *default* superior a 120 dias

Modelo	KS 90	KS 30	KS 60	KS 120
RF.	0.6619846	0.5094635	0.6027268	0.6971009
DT.	0.4071035	0.2418152	0.3398986	0.4485496
AdaBoost.	0.6565465	0.5014284	0.6005207	0.6933222
***Bagging*.**	**0.6925993**	**0.6000578**	**0.6499971**	**0.7186262**
LR.	0.4053571	0.2538376	0.3332069	0.4540041
LDA.	0.3948803	0.2480742	0.3217469	0.4441733
SVM linear.	0	0	0	0.2968042
SVM radial.	2.319255e-06	0	0	0

Tabela 3.20: Resultados KS dos modelos com a variável *default* superior a 120 dias *vs* 30, 60 e 90 dias.

O *BRIER Score* também ratifica a superioridade do modelo *Bagging* frente ao outros modelos, pois apresentou o melhor índice (0.02652) em contrapartida *AdaBoost* apresentou o segundo melhor índice (0.03412) seguido por RF (0.03611).

Modelo	*BRIER Score* 90	*BRIER Score* 30	*BRIER Score* 60	*BRIER Score* 120
RF.	0.04146	0.08576	0.05330	0.03611
DT.	0.04840	0.09866	0.06159	0.04231
AdaBoost.	0.03971	0.08387	0.05153	0.03412
***Bagging*.**	**0.03032**	**0.06229**	**0.03900**	**0.02652**
LR.	0.05270	0.10210	0.06612	0.04671
LDA.	0.05830	0.11160	0.07171	0.05249
SVM linear.	0.84830	0.64050	0.83370	0.05041
SVM radial.	0.84700	0.65920	0.82930	0.85320

Tabela 3.21: Resultados *BRIER Score* dos modelos com a variável *default* superior a 120 dias *vs* 30, 60 e 90 dias.

De forma geral, o que se observa nos testes é que a capacidade de predição dos modelos aumentaram quando o valor para caracterizar a variável *default* aumentou. Entretanto, os ganhos obtidos na precisão do modelo para variável *default* acima de 120 dias são relativamente menores, isto reforça o entendimento do BACEN de caracterizar a variável como atrasos superiores a 90 dias (ANNIBAL et al., 2009).

Bagging foi o modelo que apresentou melhores resultados para variável *default* superior 30, 60, 90 ou 120 dias. É interessante ressaltar que os três melhores métodos (*Bagging*, RF e *AdaBoost*) são baseados em classificadores *ensemble* de estimadores o que corrobora a superioridades dos modelos *ensemble* na análise de risco de crédito e reforça os resultados apresentados por Wang et al. (2012) e Lessmann et al. (2015).

Observa-se também uma diminuição no tempo de processamento dos modelos. Quando utilizado a variável *default* como superior a 30 dias o tempo de processamento foi de 44 horas, reduzindo para 28 horas quando a variável *default* era superior a 90 dias e para 26 horas quando a variável *default* era superior a 120 dias.

3.6.2 Teste 2 – Avaliação dos modelos com amostra de 300.000 observações.

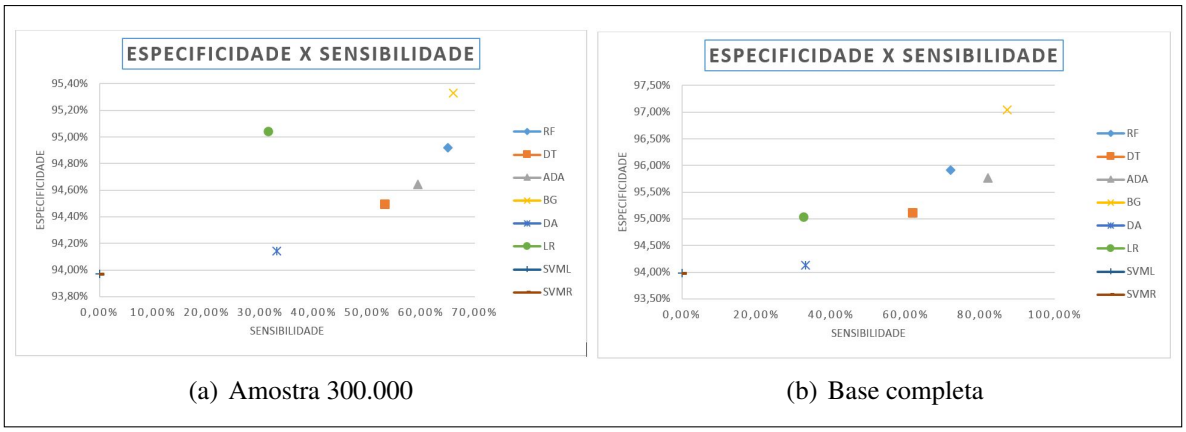
Para avaliar a performance dos algoritmos para classificação dos inadimplentes e adimplentes foi executado um teste com um número menor de observações, 300.000 (trezentas mil). Observa-se que houve uma redução na precisão média de todos os algoritmos, conforme pode ser visto na Tabela 3.22.

Modelo	PM	PM amostra	Sens.	Sens. amostra	Esp.	Esp. amostra
RF.	95,24%	94,44%	95,91%	94,92%	72,05%	65,03%
DT.	94,45%	94,04%	95,12%	94,49%	61,77%	53,29%
AdaBoost.	95,45%	94,20%	95,76%	94,64%	82,01%	59,36%
***Bagging*.**	**96,68%**	**94,68%**	**97,04%**	**95,33%**	**87,14%**	**66,00%**
LR.	92,67%	92,48%	95,04%	95,04%	32,67%	31,50%
LDA.	93,80%	93,75%	94,14%	94,14%	33,08%	33,04%
SVM linear.	93,99%	93,97%	93,99%	93,97%	0	0
SVM radial.	93,99%	93,97%	93,99%	93,97%	0	0

Tabela 3.22: Sensibilidade, especificidade e precisão média dos modelos para *default* superior a 90 dias e amostra de 300.000 observações.

A redução na precisão dos modelos está de acordo com o estudo de Marqués, García e Sánchez (2012b) que avaliaram as precisões de regressão logística, análise discriminante, árvores de decisão e redes neurais em dois conjuntos de dados através de 20 amostras de tamanho crescente e 29 distribuições amostrais e que fornecem evidência de que o uso de amostras maiores na análise de risco de crédito prevê um aumento significativo na precisão dos algoritmos.

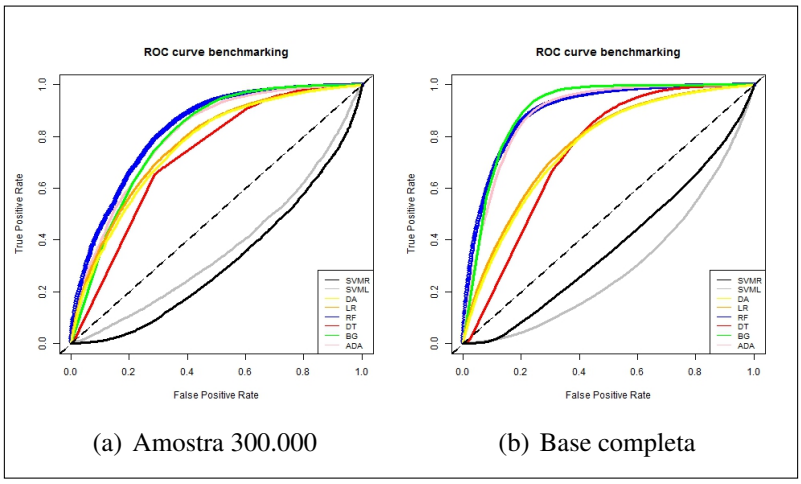
Conforme pode ser visto na Figura 3.9, o algoritmo RF obteve uma melhor performance que o algoritmo *AdaBoost* quando realizado os teste na amostra, diferente do que ocorreu com os algoritmos quando testados na base completa. Isto pode se justificar, pois conforme descreve Trevor, Robert e Jerome (2001), RF são candidato ideais para ensacamento (redução da variância pela média de muitos modelos ruidosos) e podem capturar a interação complexa, o que poderia explicar seu melhor poder preditivo numa amostra em contraposição ao *AdaBoost*.



(a) Amostra 300.000 (b) Base completa

Figura 3.9: Sensibilidade *vs* Especificidade dos modelos para *default* superior a 90 dias e amostra de 300.000 observações.

A curva ROC apresentada na Figura 3.10 e os dados demostrados na Tabela 3.23 corroboram os resultados encontrados por Jones, Johnstone e Wilson (2015) em que classificadores lineares simples, LDA e LR preveem com bastante precisão sobre as amostras, em alguns casos, realizando comparativamente bem para as estruturas modelo mais flexível. Pode-se verificar que todos os métodos de ML tiveram redução da AUROC, entretanto, LDA e LR apresentaram ligeiro aumento nos índices da AUROC. Já com relação aos métodos de ML, verifica-se a redução da AUROC para todos os modelos.



(a) Amostra 300.000 (b) Base completa

Figura 3.10: Curva ROC para variável *default* superior a 90 dias.

Os índices AUROC, *BRIER Score* e KS da amostra apresentaram significativa redução quando comparados com os resultados obtidos para os modelos em toda a base de dados. Ressalta-se, porém, o maior AUROC do algoritmo RF na amostra de dados, quando comparado com os outros algoritmos também na amostra de dados. Já o *BRIER Score* e o KS do algoritmo *Bagging* foi o que apresentou o melhor resultado, conforme pode ser visto na Tabela 3.23.

Modelo	AUC	AUC amostra	KS	KS amostra	*BRIER*	*BRIER* amostra
RF.	**0.8992894**	**0.8269959**	**0.6619846**	**0.5072386**	**0.04146**	**0.04691**
DT.	0.7382974	0.7250404	0.4071035	0.3674051	0.04840	0.05204
AdaBoost.	0.8885443	0.8012433	0.6565465	0.4685078	0.03971	0.04935
Bagging.	0.9043971	0.7987889	0.6925993	0.4717409	0.03032	0.04664
LR.	0.7657149	0.7707621	0.4053571	0.4085858	0.05270	0.05287
LDA.	0.7568453	0.7609289	0.3948803	0.3997563	0.05830	0.05871
SVM linear.	0.2973585	0.3692660	0	1.86905e-05	0.84830	0.84610
SVM radial.	0.3762618	0.3166604	2.319255e-06	0	0.84700	0.84520

Tabela 3.23: Resultados dos modelos com a variável *default* superior a 90 dias e amostra de 300.000 observações.

Dado que os modelos foram executados em uma amostra de 300.000 observações houve uma grande redução do custo computacional quando comparado com a execução na base completa. O tempo total gasto para executar os modelos na amostra de 300.000 observações foi de aproximadamente 1 hora, enquanto que para a base completa foi necessário 28 horas para executar os modelos com a variável *default* superior a 90 dias.

Diferente do que ocorreu com o tempo de processamento para a base completa, onde RF teve um menor tempo perante os três algoritmos que obtiveram uma melhor performance em classificar adimplentes e inadimplentes, no processamento realizado na amostra, dentre os três algoritmos que melhor performaram (RF, *AdaBoost* e *Bagging*), *AdaBoost* foi o que teve menor custo computacional, levando aproximadamente 5 minutos para sua execução. A Tabela 3.24 apresenta o custo computacional que foi gasto para treinar e realizar a predição em cada modelo.

Modelo	Usuário (ms)	Sistema (ms)	Decorrido (ms)
RF.	5.08	0.73	639.50
DT.	16.39	0.14	16.53
AdaBoost.	292.21	0.25	292.53
Bagging.	750.88	8.08	733.53
LR.	2013.50	207.22	1260.25
LDA.	3.84	0.87	3.42
SVM linear.	113.72	3.19	396.27
SVM radial.	301.39	2.97	864.05

Tabela 3.24: Tempo decorrido em milissegundos para cada modelo com a variável *default* superior a 90 dias e amostra de 300.000 observações.

3.6.3 Teste 3 – Avaliação dos modelos sem o uso de variáveis discriminatórias (gênero, idade e estado civil).

O *Equal Credit Opportunity Act* (ECOA) é uma lei dos Estados Unidos, promulgada em 28 de outubro de 1974 que torna ilegal qualquer credor discriminar qualquer requerente, com relação a qualquer aspecto de uma transação de crédito com base na raça, cor, religião, origem nacional, sexo, estado civil, idade, ou o fato de que a totalidade ou parte do rendimento provenha de um programa de assistência pública. A lei aplica-se a qualquer pessoa que, no curso normal dos negócios, participa regularmente em uma decisão de crédito, incluindo bancos, varejistas, empresas de cartões bancários, empresas financeiras e cooperativas de crédito (KOH; TAN; GOH, 2015).

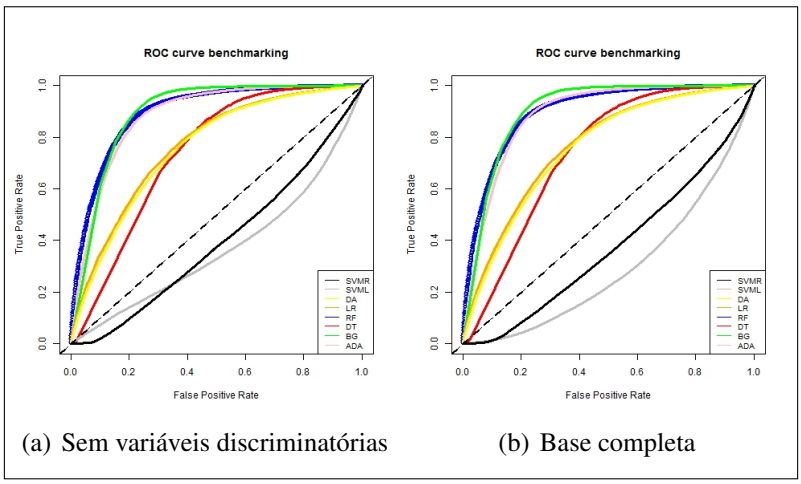
No Brasil, não há explicitamente uma proibição para o uso de variáveis discriminatórias nos modelos de risco, sendo exigência do órgão regulador (BACEN) a validação do poder discriminatório do sistema de classificação de risco de crédito. Porém este teste buscou identificar qual o impacto sobre o poder discriminatório do modelo com a retirada das variáveis gênero, idade e estado civil da base de treinamento.

Quando se analisa a precisão média dos modelos, em comparação com os resultados obtidos considerando o uso de todas a variáveis, verifica-se uma pequena redução na precisão média deles, entretanto, observa-se que as reduções foram muito baixas, conforme pode ser visto na Tabela: 3.25. Interessante notar ainda que para LR a precisão média do modelo chegou a aumentar de 92,67% para 92,68%.

Modelo	PM	PM SV	Sen.	Sens. SV.	Esp.	Esp. SV.
RF.	95,24%	95,18%	95,91%	95,96%	72,05%	69,95%
DT.	94,45%	94,37%	95,12%	94,99%	61,77%	60,79%
AdaBoost.	95,45%	95,22%	95,76%	95,56%	82,01%	79,57%
***Bagging*.**	**96,68%**	**96,54%**	**97,04%**	**97,00%**	**87,14%**	**84,58%**
LR.	92,67%	92,68%	95,04%	94,99%	32,67%	32,36%
LDA.	93,80%	93,78%	94,14%	94,13%	33,08%	32,31%
SVM linear.	93,99%	93,98%	93,99%	93,99%	0	0
SVM radial.	93,99%	93,98%	93,99%	93,99%	0	0

Tabela 3.25: Sensibilidade, especificidade e precisão média dos modelos para *default* superior a 90 dias *vs default* superior a 90 dias sem variáveis discriminatórias.

A Figura 3.11, com as curvas ROC dos resultados sem as variáveis discriminatórias e dos resultados com a base completa, reforça a evidência de que o impacto da retirada das variáveis discriminatórias foram baixos. Visualiza-se grande similaridades das curvas do quadrante "Sem variáveis discriminatórias" com o quadrante "Base completa" o que reforça a superioridade das variáveis da operação de crédito para a efetiva classificação dos inadimplentes como, por exemplo: valor do subsídio, valor do imóvel, renda mensal, valor financiado, etc.



(a) Sem variáveis discriminatórias (b) Base completa

Figura 3.11: Curva ROC para variável *default* superior a 90 dias *vs default* superior a 90 dias sem variáveis discriminatórias.

Na comparação da AUROC dos modelos, Tabela 3.26, é possível verificar como a perda no valor da área sob a curva ROC foi incremental. O modelo RF, por exemplo, passou de uma AUC de 0.8992894 para 0.8970298, redução de 0,25% na AUROC do modelo. Os demais índices, KS e *BRIER Score*, também seguem a mesma linha, apresentando perdas incrementais e reforçando a superioridade do algoritmo *Bagging*, RF e *AdaBoost* para a efetiva classificação dos inadimplentes.

Modelo	**AUC**	**AUC SV**	**KS**	**KS SV**	***BRIER Score***	***BRIER Score* SV**
RF.	0.8992894	0.8970298	0.6619846	0.6583683	0.04146	0.04121
DT.	0.7382974	0.7377127	0.4071035	0.4009901	0.04840	0.04852
AdaBoost.	0.8885443	0.8825037	0.6565465	0.6398294	0.03971	0.04108
***Bagging*.**	**0.9043971**	**0.8945465**	**0.6925993**	**0.6794313**	**0.03032**	**0.03128**
LR.	0.7657149	0.7650284	0.4053571	0.4065277	0.05270	0.05279
LDA.	0.7568453	0.7560313	0.3948803	0.3948835	0.05830	0.05842
SVM linear.	0.2973585	0.3710326	0	0	0.84830	0.84830
SVM radial.	0.3762618	0.3969884	0	0	0.84700	0.84790

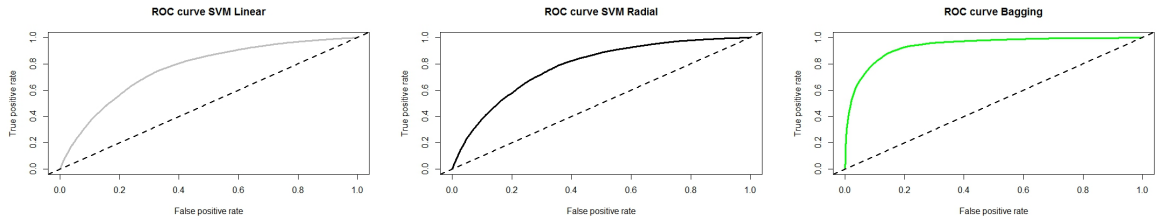
Tabela 3.26: Resultados dos modelos com a variável *default* superior a 90 dias *vs default* superior a 90 dias sem variáveis discriminatórias.

Por estes resultados pode-se inferir, para este caso em específico, que é viável retirar as variáveis discriminatórias dos modelos e preservar o poder discriminatório do sistema de classificação de risco de crédito. Observa-se que estes resultados estão de acordo com o trabalho de Palmuti e Pichiatti (2004) que identificaram que, apesar do seu modelo utilizar variáveis discriminatórias, como idade e sexo, as variáveis mais relevantes para a formação dos grupos (adimplente e inadimplente), foram: valor liberado, grau de formalidade, renda, valor da prestação, prazo de pagamento e taxa de juros. As demais variáveis mostraram-se pouco significativas para serem usadas enquanto medida de dissimilaridade entre os grupos.

3.6.4 Observações a respeito do modelo SVM

A implementação do modelo SVM no R está de acordo com Meyer e Wien (2015), entretanto, para averiguar se o SVM poderia obter uma melhor performance com a alteração de alguns parâmetros ou com a adequação da base de dados foram feitos o seguintes testes: alteração na semente de seleção dos dados de teste e treino; retirada dos parâmetros de "*Cost*" e "*Gamma*" obtidos com a função "*Tune*" e utilização dos valores padrão; a retirada das variáveis categóricas da base de dados; o processamento de índices obtidos a partir de duas variáveis; e o balanceamento do número de observações da classe adimplente com o mesmo número de observações da classe inadimplentes.

Após a realização destes procedimentos verificou-se que o balanceamento do número de observações da classe adimplente com o mesmo número de observações da classe inadimplentes melhorou a performance do modelo SVM fazendo com ele conseguisse classificar corretamente os inadimplentes e adimplentes. A Figura 3.12 demostra a curva ROC dos modelos SVM linear e radial e *Bagging*, após balancear igualmente as classes adimplentes e inadimplentes.



(a) SVM linear (b) SVM radial (c) *Bagging*

Figura 3.12: Curva ROC dos modelos com as classes (adimplente e inadimplente) igualmente balanceadas.

É possível verificar pela curva ROC do SVM que houve uma melhora considerável do modelo com o balanceamento das classes. Porém, ao executar o algoritmo *Bagging* na base com as classes igualmente balanceadas fica nítido a superioridade do *Bagging* em comparação com o algoritmo SVM. A Tabela 3.27 compara os índices AUC, KS e *BRIER Score* para os modelos com as classes balanceadas.

Modelo	AUC	KS	*BRIER Score*
***Bagging*.**	**0.936809**	**0.7363862**	**0.1009**
SVM linear.	0.7655737	0.4186507	0.1989
SVM radial.	0.7799931	0.4322487	0.1921

Tabela 3.27: Resultados dos modelos com as classes (adimplente e inadimplente) igualmente balanceadas.

Verifica-se que o *Bagging* apresenta os melhores índices de forma geral (AUC, KS, *BRIER Score*) e é possível verificar que o *kernel* radial obteve melhores resultados do que *kernel* linear.

Hsu et al. (2003) descrevem em seu artigo que os usuários do SVM se depararam frequentemente com problemas ao utilizar este modelo, sendo assim o autores propõem os seguintes procedimentos para melhorar a performance do SVM:

- Transformar dados para o formato de um pacote SVM;
- Realizar escalonamento simples nos dados;
- Usar validação cruzada para encontrar o melhor parâmetro C e γ;
- Usar o melhor parâmetro C e γ para treinar todo o conjunto de treinamento;
- Testar diferentes *kernel*;

Dessa forma, pode-se dizer que o SVM é sim capaz de predizer os inadimplentes e adimplentes na base de dados do PMCMV, contudo, é necessário um maior cuidado com o balanceamento adequado no número de observações. Como o objetivo deste trabalho era comparar a adequação dos modelos de previsão da inadimplência que melhor identificassem o bom e o mau pagador e considerando que os demais modelos não tiveram problema ao serem executados na base completa, pode-se dizer que os modelos *Bagging*, RF e *AdaBoost* apresentam maior praticidade e performance para a base de dados do PMCMV.

CONSIDERAÇÕES FINAIS

A disponibilidade do crédito habitacional é vital para o bom funcionamento do setor imobiliário, dado que a habitação é um dos principais preços de compra dos consumidores comuns e normalmente variam de 4 vezes a renda anual nos países desenvolvidos a 8 vezes a renda anual em economias emergentes (BALL, 2003).

Em consequência, a melhoria do setor imobiliário de um país pode melhorar a saúde pública reduzindo a probabilidade de surtos de doenças, estimular o crescimento econômico com a criação de empregos, e ter consequências sociais importantes, influenciando na redução do crime e aumentando a cidadania (WARNOCK; WARNOCK, 2008).

Por outro lado, a demanda habitacional é relativamente impactada pelo nível de renda, o que tem levado governos a atuarem com programa habitacionais e para o aumentado da renda dos indivíduos, seja minimizando as externalidades/desigualdades, seja fornecendo subsídios direto às famílias de baixa renda (OLSEN, 2003).

A adoção de algumas medidas prudenciais como, por exemplo: Análise de risco de crédito para financiamentos de habitação pública; utilização de um cadastro positivo; o ajuste do valor das parcelas em períodos de recessão e a utilização de *voucher* para habitação pública, poderiam reduzir drasticamente a inadimplência para o PMCMV.

Mais especificamente quanto a análise de risco de crédito, avaliações e critérios subjetivos para conceder o crédito foram bastante utilizadas no passado, porém o risco de crédito pode ser melhor controlado com o uso de instrumentos estatísticos e de sistemas multivariados, possibilitando a mensuração do risco de forma mais objetiva e com uma abordagem empírica que enfatiza a predição (ALTMAN; SAUNDERS, 1998). Esses modelos são conhecidos como *credit scoring* e buscam aceitar ou rejeitar uma determinada operação de crédito comparando a probabilidade estimada de incumprimento da obrigação contratual com um limiar adequado, utilizando métodos de análise discriminante, regressão linear, regressão logística e árvore de decisão, etc. (DOMINGOS, 2012).

Os resultados apresentados demostram que a utilização de instrumentos estatísticos e *Machine Learning* para a avaliação do risco de crédito no PMCMV podem trazer resultados significativos em termos de redução da inadimplência já que foi possível, com alguns algoritmos, obter razoável precisão na identificação do bom e do mau pagador na base de dados com 1,5 milhões de registros.

Os resultados também demostram a superioridade dos algoritmos *ensemble* bem como a adequação deles para estes dados em especifico. Como visto anteriormente, os três melhores

métodos (*Bagging*, RF e *AdaBoost*) são baseados em classificadores *ensemble* de estimadores o que corrobora a superioridades dos modelos *ensemble* na análise de risco de crédito e reforça as contribuições apresentados pelo autores Wang et al. (2012) e Lessmann et al. (2015).

Na avaliação da performance dos modelos, diante dos diferentes critérios utilizados para a variável dependente, *default*, verifica-se que a capacidade de predição dos modelos melhoraram à medida que o número de dias de atrasos utilizados para definir a variável *default* aumentavam. Observa-se, entretanto, que os ganhos nos resultados dos modelos de 90 para a 120 dias de atrasos foram menores, o que reforça que a utilização da variável *default* como 90 dias é adequada para se obter bons resultados de um modelo de classificação de inadimplentes e adimplentes.

Outro ponto importante observado, no processamento dos modelos, é a redução do custo computacional à medida que o tempo da variável *default* também aumentava. O tempo de processamento, que era de 44 horas para variável *default* superior a 30 dias de atraso, foi reduzindo gradativamente, até chegar em 26 horas quando a variável *default* era superior a 120 dias de atraso.

Já no teste executado com um número menor de observações, 300.000 observações, observa-se uma redução nos resultados dos algoritmos de forma geral. Porém, LDA e LR apresentaram ligeiro aumento nos índices da AUROC o que demostra a capacidade desses algoritmos com um número reduzido de observações ou com uma amostra pequena.

Verificou-se também que a retirada de variáveis discriminatórias do modelos preserva o poder discriminatório do sistema de classificação de risco de crédito, dessa forma, pode-se inferir, que as variáveis mais importante para a classificação estão associadas as características da operação crédito, como por exemplo: valor liberado, valor prestação, prazo de pagamento e taxa de juros, o que corrobora o estudo de Palmuti e Pichiatti (2004).

Numa pequena simulação, utilizando o algoritmo que obteve os melhores resultados nos testes (*Bagging*), fica ainda mais claro o ganho que poderia ser obtido ao utilizar o *Machine Learning* para análise de risco dos possíveis beneficiários. A taxa de inadimplência do programa que é de 11,80 % poderia ser reduzida para 2,95 % com a utilização do algoritmo *Bagging* para classificar possíveis beneficiário como bom ou mau pagador. Além do que, apenas 0,46% das análises de risco seriam equivocadas (classificar bons pagadores como maus). Logo, 197.905 mil contratos inadimplentes deixariam de existir para o programa, e, considerando uma média de R$ 50.000,00 por operação de crédito e descartando a hipótese de retomada do imóvel em caso de inadimplemento, pode se esperar uma redução nas perdas com inadimplência de 9,8 bilhões.

Sendo assim, dado a magnitude dos benefícios elencados com a utilização de um adequado modelo de *credit scoring* para predizer o bom e o mau pagador do PMCMV o seu uso torna-se imperioso para um eficaz controle do risco de crédito em futuras operações.

Ressalta-se que este estudo não fez qualquer adequação ou transformação dos dados como, por exemplo, categorização de variáveis e escalonamento simples nos dados, logo pode-se

considerar que está é uma limitação deste estudo uma vez que a adequação desses dados poderiam trazer ganhos de performance para os modelos, em especial para o SVM, já que ele requer que cada instância de dados sejam representadas como um vetor de números reais (HSU et al., 2003). Portanto, um trabalho com a adequação deste dados como, por exemplo, configurações de variáveis como valor da prestação dividido pela renda e uso de variáveis relativas poderiam ser também significativas no modelo e, portanto, pode ser explorado em estudos futuros.

Uma outra limitação deste estudo se deve ao fato de dois modelos terem sido executados utilizando o processamento em paralelo (multi-core) e os demais terem sido executados em serial. Isso acaba por limitar a comparação do custo computacional dos diferentes algoritmos, pois com o processamento em paralelo o uso do computador é melhor aproveitado. Como sugestão de estudos futuro, a comparação de todos os modelo em serial ou em paralelo pode dar a real dimensão dos custos computacionais de cada modelo.

Sugere-se ainda para estudo futuros ampliar o número de modelos a serem utilizados como, por exemplo, a utilização de *Machine Learning* com algoritmos genéticos, lógica fuzzy e redes neurais, que podem ser uma boa alternativa para predição de bons e maus pagadores. Ademais, sugere-se implementar a análise de diferentes custos de classificação errada. Afinal, classificar um mau tomador como bom é mais custoso do que classificar um bom tomador como ruim.

REFERÊNCIAS

ABIKO, A. *Introdução à gestão habitacional. São Paulo, EPUSP, 1995. Texto técnico da Escola Politécnica da USP, Departamento de Engenharia de Construção Civil*. [S.l.], 1995.

ALI, J.; KHAN, R.; AHMAD, N.; MAQSOOD, I. Random forests and decision trees. *International Journal of Computer Science Issues (IJCSI)*, Citeseer, v. 9, n. 5, 2012.

ALTMAN, E. I.; SAUNDERS, A. Credit risk measurement: Developments over the last 20 years. *Journal of Banking and Finance*, CFA Institute, v. 21, n. 11/12, 1998.

ANNIBAL, C. A. et al. *Inadimplência do Setor Bancário Brasileiro: Uma avaliação de suas medidas*. [S.l.], 2009.

AVERY, R. B.; BOSTIC, R. W.; CALEM, P. S.; CANNER, G. B. Credit risk, credit scoring, and the performance of home mortgages. *Fed. Res. Bull.*, HeinOnline, v. 82, p. 621, 1996.

BACEN, B. C. d. B. *Relatório de Economia Bancária e Crédito*. [S.l.], 2014.

BALBIM, R.; KRAUSE, C. Produção social da moradia: um olhar sobre o planejamento da habitação de interesse social no brasil. *Revista Brasileira de Estudos Urbanos e Regionais*, v. 16, n. 1, p. 189–201, 2014.

BALFOUR, D. L.; SMITH, J. L. Transforming lease-purchase housing programs for low income families: Towards empowerment and engagement. *Journal of Urban Affairs*, Wiley Online Library, v. 18, n. 2, p. 173–188, 1996.

BALL, M. *Improving Housing Markets. RICS Leading Edge Series*. 2003.

BARREN, J.; STATEN, M. The value of comprehensive credit reports: Lessons from the us experience. *Credit reporting systems and the international economy*, MIT Press Cambridge, MA, v. 8, p. 273–310, 2003.

BASEL, I. *An Explanatory Note on the Basel II IRB Risk Weight Functions*. 2007.

BEN-DAVID, A. Monotonicity maintenance in information-theoretic machine learning algorithms. *Machine Learning*, Springer, v. 19, n. 1, p. 29–43, 1995.

BONDUKI, N. Política habitacional e inclusão social no Brasil: Revisão histórica e novas perspectivas no governo Lula. *Revista eletrônica de Arquitetura e Urbanismo*, v. 1, p. 70–104, 2008.

BRASIL. Lei nº 11.977, de 7 de julho de 2009. 2009.

BRASIL. Lei nº. 12.414, de 9 de junho de 2011. 2011.

BRASIL. Lei nº 13.043, de 13 de novembro de 2014. 2014.

BRASIL, M. d. C. M. *Indicadores sobre Minha Casa Minha Vida*. 2016. Acessado em 2015 jan 15. Disponível em: <http://dados.gov.br/dataset/minha-casa-minha-vida>.

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

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

BREIMAN, L.; FRIEDMAN, J.; STONE, C. J.; OLSHEN, R. A. *Classification and regression trees*. [S.l.]: CRC press, 1984.

BROWN, I.; MUES, C. An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications*, Elsevier, v. 39, n. 3, p. 3446–3453, 2012.

CHEN, J.; JING, J.; MAN, Y.; YANG, Z. Public housing in mainland china: History, ongoing trends, and future perspectives. In: *The Future of Public Housing*. [S.l.]: Springer, 2013. p. 13–35.

CHIU, R. L. The transferability of public housing policy within Asia: Reflections from the Hong Kong-Mainland China case study. In: *The Future of Public Housing*. [S.l.]: Springer, 2013. p. 3–12.

COLLINS, M. E.; CURLEY, A. M.; CLAY, C.; LARA, R. Evaluation of social services in a hope vi housing development: resident and staff perceptions of successes and barriers. *Evaluation and Program Planning*, Elsevier, v. 28, n. 1, p. 47–59, 2005.

CUMMINGS, J. L.; DIPASQUALE, D. The low-income housing tax credit an analysis of the first ten years. *Housing Policy Debate*, Taylor & Francis, v. 10, n. 2, p. 251–307, 1999.

DAMICO, F. O programa minha casa, minha vida e a caixa econômica federal. *TRABALHOS PREMIADOS*, p. 33, 2011.

DENG, Y.; LIU, P. Mortgage prepayment and default behavior with embedded forward contract risks in china's housing market. *The Journal of Real Estate Finance and Economics*, Springer, v. 38, n. 3, p. 214–240, 2009.

DOMINGOS, P. A few useful things to know about machine learning. *Communications of the ACM*, ACM, v. 55, n. 10, p. 78–87, 2012.

DOUMPOS, M.; ZOPOUNIDIS, C. Model combination for credit risk assessment: A stacked generalization approach. *Annals of Operations Research*, Springer, v. 151, n. 1, p. 289–306, 2007.

DRIANT, J.-C. Défaire les grands ensembles. *La Ville en débat*, Presses Universitaires de France, p. 13–24, 2012.

EBERLY, J.; KRISHNAMURTHY, A. Efficient credit policies in a housing debt crisis. *Brookings Papers on Economic Activity*, Brookings Institution Press, v. 2014, n. 2, p. 73–136, 2014.

EINAV, L.; JENKINS, M.; LEVIN, J. The impact of credit scoring on consumer lending. *The RAND Journal of Economics*, Wiley Online Library, v. 44, n. 2, p. 249–274, 2013.

EISINGER, P. K. French urban housing and the mixed economy: The privatization of the public sector. *The ANNALS of the American Academy of Political and Social Science*, JSTOR, p. 134–147, 1982.

ELUL, R.; SOULELES, N. S.; CHOMSISENGPHET, S.; GLENNON, D.; HUNT, R. M. What'triggers' mortgage default? FRB of Philadelphia Working Paper, 2010.

ENGELMANN, B.; HAYDEN, E.; TASCHE, D. Testing rating accuracy. *Risk*, v. 16, n. 1, p. 82–86, 2003.

FENG, C.; SUTHERLAND, A.; KING, R.; MUGGLETON, S.; HENERY, R. Comparison of machine learning classifiers to statistics and neural networks. *AI&Statistics-93*, v. 6, p. 41, 1993.

FERGUSON, B.; SMETS, P. Finance for incremental housing; current status and prospects for expansion. *Habitat International*, Elsevier, v. 34, n. 3, p. 288–298, 2010.

FJP, F. J. P. Fundação João Pinheiro, centro de estatística e informações. *Belo Horizonte*, 2010.

FLOREZ-LOPEZ, R. Modelling of insurers' rating determinants. an application of machine learning techniques and statistical models. *European Journal of Operational Research*, Elsevier, v. 183, n. 3, p. 1488–1512, 2007.

GALINDO, J.; TAMAYO, P. Credit risk assessment using statistical and machine learning: Basic methodology and risk modeling applications. *Computational Economics*, Springer, v. 15, n. 1-2, p. 107–143, 2000.

GERTLER, M.; KARADI, P. Monetary policy surprises, credit costs, and economic activity. *American Economic Journal: Macroeconomics*, American Economic Association, v. 7, n. 1, p. 44–76, 2015.

GESTEL, T. V.; BAESENS, B.; DIJCKE, P. V.; GARCIA, J.; SUYKENS, J. A.; VANTHIENEN, J. A process model to develop an internal rating system: sovereign credit ratings. *Decision Support Systems*, Elsevier, v. 42, n. 2, p. 1131–1151, 2006.

GONZALEZ, L. E. Finanças e contabilidade-contribuições para melhorar o programa minha casa minha vida. *Anuário de Pesquisa GVPesquisa*, 2015.

GREEN, R. K.; WACHTER, S. M. The housing finance revolution. *U of Penn, Inst for Law & Econ Research Paper*, n. 09-37, 2007.

GREEN, R. K.; WHITE, M. J. Measuring the benefits of homeowning: Effects on children. *Journal of Urban Economics*, Elsevier, v. 41, n. 3, p. 441–461, 1997.

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

HAND, D. J.; OLIVER, J. J.; LUNN, A. D. Discriminant analysis when the classes arise from a continuum. *Pattern Recognition*, Elsevier, v. 31, n. 5, p. 641–650, 1998.

HERBERT, C. E.; BELSKY, E. S. The homeownership experience of low-income and minority households: A review and synthesis of the literature. *Cityscape*, JSTOR, v. 10, n. 2, p. 5–59, 2008.

HOFFMAN, A. V. High ambitions: The past and future of american low-income housing policy. *Housing Policy Debate*, Taylor & Francis, v. 7, n. 3, p. 423–446, 1996.

HOUSING, U. Departament of. *Programas de habitaçãdo HUD*. 2016. Disponível em: <https://portal.hud.gov/hudportal/HUD>.

HSU, C.-W.; CHANG, C.-C.; LIN, C.-J. et al. A practical guide to support vector classification. 2003.

HUANG, Y. Low-income housing in chinese cities: Policies and practices. *The China Quarterly*, Cambridge Univ Press, v. 212, p. 941–964, 2012.

HUANG, Z.; CHEN, H.; HSU, C.-J.; CHEN, W.-H.; WU, S. Credit rating analysis with support vector machines and neural networks: A market comparative study. *Decision support systems*, Elsevier, v. 37, n. 4, p. 543–558, 2004.

IBGE, I. B. d. E. *Pesquisa nacional por amostra de domicílios*. [S.l.: s.n.], 2008.

JACOBS, K. Contractual welfare ideology and housing management practice: The deployment of 'tenant incentive schemes' in Australia. *Urban Policy and Research*, Taylor & Francis, v. 26, n. 4, p. 467–479, 2008.

JACOBS, K.; ATKINSON, R.; SPINNEY, A.; PEISKER, V. C.; BERRY, M.; DALTON, T. What future for public housing? A critical analysis. Australian Housing and Urban Research Institute, Southern Research Centre, 2010.

JONES, S.; JOHNSTONE, D.; WILSON, R. An empirical evaluation of the performance of binary classifiers in the prediction of credit ratings changes. *Journal of Banking & Finance*, Elsevier, v. 56, p. 72–85, 2015.

JR, D. W. H.; LEMESHOW, S. *Applied logistic regression*. [S.l.]: John Wiley & Sons, 2004.

KIM, K.-j.; AHN, H. A corporate credit rating model using multi-class support vector machines with an ordinal pairwise partitioning approach. *Computers & Operations Research*, Elsevier, v. 39, n. 8, p. 1800–1811, 2012.

KOH, H. C.; TAN, W. C.; GOH, C. P. A two-step method to construct credit scoring models with data mining techniques. *International Journal of Business and Information*, v. 1, n. 1, 2015.

LAFERRÈRE, A.; BLANC, D. L. Housing policy: Low-income households in France. *A Companion to Urban Economics, Oxford: Blackwell Publishing*, 2006.

LAMBRECHT, B.; PERRAUDIN, W.; SATCHELL, S. Time to default in the uk mortgage market. *Economic Modelling*, Elsevier, v. 14, n. 4, p. 485–499, 1997.

LANTZ, B. *Machine learning with R*. [S.l.]: Packt Publishing Ltd, 2015.

LAWRENCE, E. C.; SMITH, L. D.; RHOADES, M. An analysis of default risk in mobile home credit. *Journal of Banking & Finance*, Elsevier, v. 16, n. 2, p. 299–312, 1992.

LEE, Y.-C. Application of support vector machines to corporate credit rating prediction. *Expert Systems with Applications*, Elsevier, v. 33, n. 1, p. 67–74, 2007.

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

LEVY, C.; LATENDRESSE, A.; CARLE-MARSAN, M. Gendering the urban social movement and public housing policy in são paulo. *Latin American Perspectives*, SAGE Publications, p. 0094582X16668317, 2016.

LEWIS, E. M. *An introduction to credit scoring*. [S.l.]: Fair, Isaac and Company, 1992.

LILLIEFORS, H. W. On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, Taylor & Francis Group, v. 62, n. 318, p. 399–402, 1967.

MALPEZZI, S.; VANDELL, K. Does the low-income housing tax credit increase the supply of housing? *Journal of Housing Economics*, Elsevier, v. 11, n. 4, p. 360–380, 2002.

MARQUÉS, A.; GARCÍA, V.; SÁNCHEZ, J. S. Exploring the behaviour of base classifiers in credit scoring ensembles. *Expert Systems with Applications*, Elsevier, v. 39, n. 11, p. 10244–10250, 2012.

MARQUÉS, A.; GARCÍA, V.; SÁNCHEZ, J. S. Two-level classifier ensembles for credit risk assessment. *Expert Systems with Applications*, Elsevier, v. 39, n. 12, p. 10916–10922, 2012.

MCCALLUM, D.; BENJAMIN, S. Low-income urban housing in the third world: Broadening the economic perspective. *Urban Studies*, SAGE Publications, v. 22, n. 4, p. 277–287, 1985.

MCCLURE, K. The low-income housing tax credit as an aid to housing finance: How well has it worked? *Housing Policy Debate*, Taylor & Francis, v. 11, n. 1, p. 91–114, 2000.

MELTZER, A. H. Credit availability and economic decisions: Some evidence from the mortgage and housing markets. *The Journal of Finance*, Wiley Online Library, v. 29, n. 3, p. 763–777, 1974.

MEYER, D.; WIEN, F. T. Support vector machines. *The Interface to libsvm in package e1071*, 2015.

MITCHELL, T. M. et al. *Machine learning*. [S.l.]: McGraw-Hill, Inc., New York, NY, 1997.

MOTLEY, C. M.; PERRY, V. G. Living on the other side of the tracks: An investigation of public housing stereotypes. *Journal of Public Policy & Marketing*, American Marketing Association, v. 32, n. special issue, p. 48–58, 2013.

MUES, C.; BAESENS, B.; FILES, C. M.; VANTHIENEN, J. Decision diagrams in machine learning: An empirical study on real-life credit-risk data. *Expert Systems with Applications*, Elsevier, v. 27, n. 2, p. 257–264, 2004.

NIKLIS, D.; DOUMPOS, M.; ZOPOUNIDIS, C. Combining market and accounting-based models for credit scoring using a classification scheme based on support vector machines. *Applied Mathematics and Computation*, Elsevier, v. 234, p. 69–81, 2014.

OLSEN, E. O. Housing programs for low-income households. In: *Means-tested transfer programs in the United States*. [S.l.]: University of Chicago Press, 2003. p. 365–442.

ONG, C.-S.; HUANG, J.-J.; TZENG, G.-H. Building credit scoring models using genetic programming. *Expert Systems with Applications*, Elsevier, v. 29, n. 1, p. 41–47, 2005.

ORTALO-MAGNE, F.; RADY, S. Housing market dynamics: On the contribution of income shocks and credit constraints. *The Review of Economic Studies*, Oxford University Press, v. 73, n. 2, p. 459–485, 2006.

OSTROWSKI, S.; REICHLING, P. Measures of predictive success for rating functions. *The Journal of Risk Model Validation*, Incisive Media Plc, v. 5, n. 2, p. 61, 2011.

PALMUTI, C.; PICHIATTI, D. Mensuração do risco de crédito por meio de análise estatística multivariada. *Revista Economia Ensaios*, v. 26, n. 2, p. 7–22, 2004.

PILLAY, A.; NAUDÉ, W. Financing low-income housing in south africa: Borrower experiences and perceptions of banks. *Habitat International*, Elsevier, v. 30, n. 4, p. 872–885, 2006.

PIRAMUTHU, S. On preprocessing data for financial credit risk evaluation. *Expert Systems with Applications*, Elsevier, v. 30, n. 3, p. 489–497, 2006.

POPKIN, S. J. A decade of HOPE VI: Research findings and policy challenges. The Urban Institute, 2004.

PRESS, S. J.; WILSON, S. Choosing between logistic regression and discriminant analysis. *Journal of the American Statistical Association*, Taylor & Francis, v. 73, n. 364, p. 699–705, 1978.

REN, Y.; HU, F.; MIAO, H. The optimization of kernel function and its parameters for SVM in well-logging. In: IEEE. *Service Systems and Service Management (ICSSSM), 2016 13th International Conference on*. [S.l.], 2016. p. 1–5.

ŘEZÁČ, M.; ŘEZÁČ, F. et al. How to measure the quality of credit scoring models. *Finance a úvěr: Czech Journal of Economics and Finance*, v. 61, n. 5, p. 486–507, 2011.

SABERI, M.; MIRTALAIE, M. S.; HUSSAIN, F. K.; AZADEH, A.; HUSSAIN, O. K.; ASHJARI, B. A granular computing-based approach to credit scoring modeling. *Neurocomputing*, Elsevier, v. 122, p. 100–115, 2013.

SCANION, E. Low-income homeownership policy as a community development strategy. *Journal of Community Practice*, Taylor & Francis, v. 5, n. 1-2, p. 137–154, 1998.

SHI, J.; ZHANG, S.-y.; QIU, L.-m. Credit scoring by feature-weighted support vector machines. *Journal of Zhejiang University SCIENCE C*, Springer, v. 14, n. 3, p. 197–204, 2013.

SHLAY, A. B. Low-income homeownership: American dream or delusion? *Urban Studies*, SAGE Publications, v. 43, n. 3, p. 511–531, 2006.

SINAI, T.; WALDFOGEL, J. Do low-income housing subsidies increase the occupied housing stock? *Journal of public Economics*, Elsevier, v. 89, n. 11, p. 2137–2164, 2005.

STEBE, J.-M. *Le logement social en France: Que sais-je? n 763*. [S.l.]: Presses Universitaires de France, 2013.

STEGMAN, M. A. The excessive costs of creative finance: Growing inefficiencies in the production of low-income housing. *Housing Policy Debate*, Taylor & Francis, v. 2, n. 2, p. 357–373, 1991.

SUN, M.-Y.; WANG, S.-F. Validation of credit rating models-a preliminary look at methodology and literature. *Review of Financial Risk Management*, v. 2, n. 94, p. 1–15, 2005.

TASCHE, D. Validation of internal rating systems and pd estimates. *The analytics of risk model validation*, v. 28, p. 169–196, 2006.

TREVOR, H.; ROBERT, T.; JEROME, F. The elements of statistical learning: Data mining, inference and prediction. *New York: Springer-Verlag*, v. 1, n. 8, p. 371–406, 2001.

TSAI, C.-F. Combining cluster analysis with classifier ensembles to predict financial distress. *Information Fusion*, Elsevier, v. 16, p. 46–58, 2014.

TSAI, C.-F.; HSU, Y.-F.; YEN, D. C. A comparative study of classifier ensembles for bankruptcy prediction. *Applied Soft Computing*, Elsevier, v. 24, p. 977–984, 2014.

TSUKAHARA, F. Y.; KIMURA, H.; SOBREIRO, V. A.; ZAMBRANO, J. C. A. Validation of default probability models: A stress testing approach. *International Review of Financial Analysis*, Elsevier, v. 47, p. 70–85, 2016.

TWALA, B. Multiple classifier application to credit risk assessment. *Expert Systems with Applications*, Elsevier, v. 37, n. 4, p. 3326–3336, 2010.

VAPNIK, V. N.; VAPNIK, V. *Statistical learning theory*. [S.l.]: Wiley New York, 1998. v. 1.

WANG, G.; HAO, J.; MA, J.; JIANG, H. A comparative assessment of ensemble learning for credit scoring. *Expert Systems with Applications*, Elsevier, v. 38, n. 1, p. 223–230, 2011.

WANG, G.; MA, J.; HUANG, L.; XU, K. Two credit scoring models based on dual strategy ensemble trees. *Knowledge-Based Systems*, Elsevier, v. 26, p. 61–68, 2012.

WARNOCK, V. C.; WARNOCK, F. E. Markets and housing finance. *Journal of Housing Economics*, Elsevier, v. 17, n. 3, p. 239–251, 2008.

WONG, T.-C.; GOLDBLUM, C. Social housing in France: A permanent and multifaceted challenge for public policies. *Land Use Policy*, Elsevier, v. 54, p. 95–102, 2016.

WU, H.-C.; HU, Y.-H.; HUANG, Y.-H. Two-stage credit rating prediction using machine learning techniques. *Kybernetes*, Emerald Group Publishing Limited, v. 43, n. 7, p. 1098–1113, 2014.

YANG, Y. Adaptive credit scoring with kernel learning methods. *European Journal of Operational Research*, Elsevier, v. 183, n. 3, p. 1521–1536, 2007.

ZHONG, H.; MIAO, C.; SHEN, Z.; FENG, Y. Comparing the learning effectiveness of BP, ELM, I-ELM, and SVM for corporate credit ratings. *Neurocomputing*, Elsevier, v. 128, p. 285–295, 2014.

APÊNDICES

APÊNDICE A – CÓDIGO FONTE DOS MODELOS NO R

```
rm(list = ls())
install.packages("randomForest")
install.packages("gmodels")
install.packages("C50")
install.packages("ipred")
install.packages("ROCR")
install.packages("MASS")
install.packages("tabplot")
install.packages("ffbase")
install.packages("Amelia")
install.packages("e1071")
install.packages("varhandle")
install.packages("ineq")
install.packages("caret")
install.packages("ff")
install.packages("doParallel")
install.packages("foreach")
install.packages("h2o")
install.packages("devtools")
install.packages("parallelSVM")
library(parallelSVM)
library(devtools)
library(randomForest)
library(gmodels)
library(ipred)
library(C50)
library(ROCR)
library(MASS)
library(tabplot)
library(ff)
library(ffbase)
library(Amelia)
library(e1071)
library(verification)
library(varhandle)
library(ineq)
library(doParallel)
library(foreach)
library(h2o)
localH2O <- h2o.init(nthreads = -1)
```

```
41 h2o.init()
42
43 base <- read.csv("C:\\R\\Planilha-Credito.txt", sep = ;, header = TRUE)
44
45 excluir <- c("ID_Tomador","Dias_atraso", "vNome_do_Produto", "mValor_da_
     ↪ Renda_Familiar_em_SM", "iFaixa_de_Renda")
46 base1<- base[,!(names(base)%in% excluir)]
47 #Transformando em fator
48 base1$Adimplente<-as.factor(base1$Adimplente)
49 base1$Entrada<-as.numeric(base1$Entrada)
50 base1$mValor_Subsidio_Total<-as.numeric(base1$mValor_Subsidio_Total)
51 base1$Valor.financiado<-as.numeric(base1$Valor.financiado)
52 base1$Valor.total.do.imovel<-as.numeric(base1$Valor.total.do.imovel)
53 base1$Renda.mensal<-as.numeric(base1$Renda.mensal)
54 base1$Prazo..meses<-as.numeric(base1$Prazo..meses)
55 base1$Idade<-as.numeric(base1$Idade)
56 base1$Taxa.de.juros<-as.numeric(base1$Taxa.de.juros)
57 set.seed(5489794)
58 amostra100 <- base1[runif(1529211, 1, nrow(base1)),]
59 set.seed(5489794)
60 index <- 1:nrow(amostra100)
61 testindex <- sample(index, trunc(length(index)*30/100))
62 test <- amostra100[testindex,]
63 train <- amostra100[-testindex,]
64 nrow(train)
65 nrow(test)
66 ########RandomForest----------------------------------------
67 train.h2o <- as.h2o(train)
68 frame0<- as.data.frame(train.h2o)
69 train.h2o<- frame0[-1,]
70 train.h2o <- as.h2o(train.h2o)
71 test.h2o <- as.h2o(test)
72 frame1<- as.data.frame(test.h2o)
73 test.h2o<- frame1[-1,]
74 test.h2o <- as.h2o(test.h2o)
75 colnames(train.h2o)
76 y.dep <- 11
77 x.indep <- c(1:10,12:13)
78 ptm <- proc.time()
79 rf <- h2o.randomForest(y = y.dep, x = x.indep, ntree=900, training_frame =
     ↪ train.h2o)
80 h2o.varimp(rf)
81 pred <- h2o.predict(rf, newdata = test.h2o)
82 labels <- as.data.frame(pred)[,1]
83 proc.time() - ptm
```

```
84  CrossTable(test$Adimplente, labels, prop.chisq = FALSE, prop.c = FALSE,
        ↪ prop.r = TRUE, dnn = c(actual default, predicted default))
85  # Plot ROC curve
86  perf = h2o.performance(rf, test.h2o)
87  jpeg("C:\\R\\RF.jpg", width = 480, height = 480, units = "px", pointsize =
        ↪ 12, quality = 100, bg = "white", res = NA)
88  plot(perf, main = "ROC curve Random Forest", col = "blue", lwd = 2)
89  abline(a = 0, b = 1, lwd = 2, lty = 2)
90  dev.off()
91  prob<- as.data.frame(h2o.predict(rf, newdata = test.h2o))[,3]
92  perf
93  #AUC
94  h2o.auc(perf , valid = TRUE)
95  #BRIER SCORE
96  obs <- unfactor(test$Adimplente)
97  forecast <- prob
98  a <- verify(obs, forecast)
99  summary(a)
100 c<-perf@metrics$thresholds_and_metric_scores$fpr
101 d<-perf@metrics$thresholds_and_metric_scores$tpr
102 #KS
103 max ((d) - (c))
104 h2o.shutdown(prompt = FALSE)
105 #Arvore de decisao-----------------------------------------
106 gc(reset = TRUE) # para o R liberar a memoria para o SO
107 test$Adimplente<-as.factor(test$Adimplente)
108 train$Adimplente<-as.factor(train$Adimplente)
109 ptm <- proc.time()
110 dt <- C5.0(train[-11], train$Adimplente)
111 #summary(dt)
112 #fazendo a predicao
113 pred1 <- predict(dt, test)
114 proc.time() - ptm
115 table(pred1,test$Adimplente)
116 #summary(pred)
117 CrossTable(test$Adimplente, pred1, prop.chisq = FALSE, prop.c = FALSE,
        ↪ prop.r = TRUE,dnn = c(actual default, predicted default))
118 #ROC
119 predDT <-predict(dt, test, type = "prob")[,2] #probabilidad de classe=yes
120 str(predDT)
121 pred1DT <- prediction(predictions = predDT, labels = test$Adimplente)
122 predict.rocr <- prediction(predDT, test$Adimplente) # valor real da classe
123 perf.DT <- performance(predict.rocr,"tpr", "fpr")
124 jpeg("C:\\R\\DT.jpg", width = 480, height = 480, units = "px", pointsize =
        ↪ 12, quality = 100, bg = "white", res = NA)
125 plot(perf.DT, main = "ROC curve Decision Tree", col = "red", lwd = 3)
```

```
126 abline(a = 0, b = 1, lwd = 2, lty = 2)
127 dev.off()
128 perf.auc <- performance(predict.rocr, measure = "auc")
129 str(perf.auc)
130 unlist(perf.auc@y.values)
131 obs <- unfactor(test$Adimplente)
132 forecast <- predDT
133 a <- verify(obs, forecast)
134 summary(a)
135 #KS
136 max (attr (perf.DT, y.values) [[1]] - attr (perf.DT, x.values) [[1]])
137 #ADABOOST - adaptive boosting-----------------------------------------
138 gc(reset = TRUE) # para o R liberar a memoria para o SO
139 test$Adimplente<-as.factor(test$Adimplente)
140 train$Adimplente<-as.factor(train$Adimplente)
141 ptm <- proc.time()
142 AdaB <- C5.0(train[-11], train$Adimplente, trials = 30)
143 #AdaB
144 #summary(AdaB)
145 #Compara os resultados testando na base teste
146 predAD <- predict(AdaB, test)
147 proc.time() - ptm
148 table(predAD,test$Adimplente)
149 CrossTable(test$Adimplente, predAD, prop.chisq = FALSE, prop.c = FALSE,
    ↪ prop.r = TRUE, dnn = c(actual default, predicted default))
150 #ROC
151 predAdaB <-predict(AdaB, test, type = "prob")[,2] #probabilidad de classe=
    ↪ yes
152 predict.rocr <- prediction(predAdaB, test$Adimplente) # valor real da
    ↪ classe
153 perf.AdaB <- performance(predict.rocr,"tpr", "fpr")
154 jpeg("C:\\R\\ADA.jpg", width = 480, height = 480, units = "px", pointsize
    ↪ = 12, quality = 100, bg = "white",  res = NA)
155 plot(perf.AdaB, main = "ROC curve AdaBoost", col = "pink", lwd = 3)
156 abline(a = 0, b = 1, lwd = 2, lty = 2)
157 dev.off()
158 perf.auc <- performance(predict.rocr, measure = "auc")
159 str(perf.auc)
160 unlist(perf.auc@y.values)
161 obs <- unfactor(test$Adimplente)
162 forecast <- predAdaB
163 a <- verify(obs, forecast)
164 summary(a)
165 #KS
166 max (attr (perf.AdaB, y.values) [[1]] - attr (perf.AdaB, x.values) [[1]])
167 ######BAGGING-----------------------------------------------------------
```

```
gc(reset = TRUE) # para o R liberar a memoria para o SO
test$Adimplente<-as.factor(test$Adimplente)
train$Adimplente<-as.factor(train$Adimplente)
#set.seed(12345)
ptm <- proc.time()
bag <- bagging(Adimplente ~ ., data = train, nbagg = 40)
#bag
bag_pred <- predict(bag, test)
proc.time() - ptm
table(bag_pred, test$Adimplente)
CrossTable(test$Adimplente, bag_pred, prop.chisq = FALSE, prop.c = FALSE,
    ↪ prop.r = TRUE, dnn = c(actual default, predicted default))
#ROC
predbag <-predict(bag, test, type = "prob")[,2] #probabilidad de classe=
    ↪ yes
pred <- prediction(predictions = predbag, labels = test$Adimplente)
predict.rocr <- prediction(predbag, test$Adimplente) # valor real da
    ↪ classe
perf.bag <- performance(predict.rocr,"tpr", "fpr")
jpeg("C:\\R\\BG.jpg", width = 480, height = 480, units = "px", pointsize =
    ↪  12, quality = 100, bg = "white",  res = NA)
plot(perf.bag, main = "ROC curve Bagging", col = "green", lwd = 3)
abline(a = 0, b = 1, lwd = 2, lty = 2)
dev.off()
perf.auc <- performance(predict.rocr, measure = "auc")
str(perf.auc)
unlist(perf.auc@y.values)
obs <- unfactor(test$Adimplente)
forecast <- predbag
a <- verify(obs, forecast)
summary(a)
#KS
max (attr (perf.bag, y.values) [[1]] - attr (perf.bag, x.values) [[1]])
#Analise Discriminante----------------------------------------------
gc(reset = TRUE) # para o R liberar a memoria para o SO
test$Adimplente<-as.factor(test$Adimplente)
train$Adimplente<-as.factor(train$Adimplente)
ptm <- proc.time()
AD<-lda(Adimplente~., data=train, type="prob")
#AD
#Faz a predicao para a base de validacao
predito<-predict(AD, test)
proc.time() - ptm
#Compara os resultados
CrossTable(test$Adimplente, predito$class, prop.chisq = FALSE, prop.c =
    ↪ FALSE, prop.r = TRUE, dnn = c(actual default, predicted default))
```

```
#ROC
pr <- prediction(predito$x, test$Adimplente)
prAD <- performance(pr, measure = "tpr", x.measure = "fpr")
jpeg("C:\\R\\DA.jpg", width = 480, height = 480, units = "px", pointsize =
    ↪ 12, quality = 100, bg = "white",  res = NA)
plot(prAD, main = "ROC curve Discriminant Analysis", col = "yellow", lwd =
    ↪ 3)
abline(a = 0, b = 1, lwd = 2, lty = 2)
dev.off()
perf.auc <- performance(pr, measure = "auc")
auc <- performance(pr, measure = "auc")
auc <- auc@y.values[[1]]
auc
str(auc)
obs <- unfactor(test$Adimplente)
forecast <- unfactor(predito$class)
a <- verify(obs, forecast)
summary(a)
#KS
max (attr (prAD, y.values) [[1]] - attr (prAD, x.values) [[1]])
#Regressao Logistica ----------------------------------------------
gc(reset = TRUE) # para o R liberar a memoria para o SO
test$Adimplente<-as.factor(test$Adimplente)
train$Adimplente<-as.factor(train$Adimplente)
ptm <- proc.time()
LR <- glm(Adimplente ~ . , family=binomial,data=train)
#Calcula a razao de chances
exp(cbind(OR = coef(LR), confint(LR)))
#Faz a previsao para a base de validacao (probabilidade)
predito<-predict(LR,test,type="response")
proc.time() - ptm
#Escolhe quem vai ser "1" e quem vai ser "0"
predito<-ifelse(predito>=0.8,1,0)
#Compara os resultados
table(predito,test$Adimplente)
#fazendo a predicao
CrossTable(test$Adimplente, predito,
prop.chisq = FALSE, prop.c = FALSE, prop.r = TRUE, dnn = c(actual default,
    ↪ predicted default))
#ROC
p <- predict(LR, test, type="response")
prLR <- prediction(p, test$Adimplente)
perf.LR<- performance(prLR, measure = "tpr", x.measure = "fpr")
jpeg("C:\\R\\LR.jpg", width = 480, height = 480, units = "px", pointsize =
    ↪ 12, quality = 100, bg = "white",  res = NA)
```

```
250 plot(perf.LR, main = "ROC curve Logistic Regression", col = "orange", lwd
    ↪ = 3)
251 abline(a = 0, b = 1, lwd = 2, lty = 2)
252 dev.off()
253 perf.auc <- performance(prLR, measure = "auc")
254 str(perf.auc)
255 unlist(perf.auc@y.values)
256 test$Adimplente<-as.factor(test$Adimplente)
257 obs <- unfactor(test$Adimplente)
258 forecast <- p
259 a <- verify(obs, forecast)
260 summary(a)
261 #KS
262 max (attr (perf.LR, y.values) [[1]] - attr (perf.LR, x.values) [[1]])
263 #######SVM ----------------------------------------
264 gc(reset = TRUE) # para o R liberar a memoria para o SO
265 test$Adimplente<-as.factor(test$Adimplente)
266 train$Adimplente<-as.factor(train$Adimplente)
267 #Kernel linear
268 tuned <- tune.svm(Adimplente~., data = test, gamma = 10^(-5:-2), cost =
    ↪ 10^(-4:-1))
269 tbper=tuned$best.performance
270 tbpar=tuned$best.parameter
271 tbper
272 tbpar
273 ptm <- proc.time()
274 svm.L = parallelSVM(Adimplente ~ ., data = train, kernel="linear", cost
    ↪ =0.1, gamma=0.01, probability=TRUE)
275 svm.L
276 svm.L.pred <- predict(svm.L, test,  probability=TRUE)
277 proc.time() - ptm
278 #Comparando os resultados
279 CrossTable(test$Adimplente, svm.L.pred, prop.chisq = FALSE, prop.c = FALSE
    ↪ , prop.r = TRUE, dnn = c(actual default, predicted default))
280 #ROC
281 predSVML <-predict(svm.L, test, probability=TRUE, decision.values=TRUE)
282 probSVM <- attr (predSVML, "probabilities")[,2]
283 pred <- prediction(predictions = probSVM, labels = test$Adimplente)
284 predict.rocr <- prediction(probSVM, test$Adimplente) # valor real da
    ↪ classe
285 perf.SVML <- performance(predict.rocr,"tpr", "fpr")
286 jpeg("C:\\R\\SVML.jpg", width = 480, height = 480, units = "px", pointsize
    ↪  = 12, quality = 100, bg = "white",  res = NA)
287 plot(perf.SVML, main = "ROC curve SVM Linear", col = "gray", lwd = 3)
288 abline(a = 0, b = 1, lwd = 2, lty = 2)
289 dev.off()
```

```
perf.auc <- performance(predict.rocr, measure = "auc")
str(perf.auc)
unlist(perf.auc@y.values)
obs <- unfactor(test$Adimplente)
forecast <- probSVM
a <- verify(obs, forecast)
summary(a)
#KS
max (attr (perf.SVML, y.values) [[1]] - attr (perf.SVML, x.values) [[1]])
#Kernel radial
ptm <- proc.time()
svm.R = parallelSVM(as.factor(Adimplente)~., data = train, kernel="radial"
    ↪ , probability=TRUE, cost=0.1, gamma=0.01)
#svm.R
svm.R.pred <- predict(svm.R, test)
proc.time() - ptm
#Comparando os resultados
CrossTable(test$Adimplente, svm.R.pred, prop.chisq = FALSE, prop.c = FALSE
    ↪ , prop.r = TRUE, dnn = c(actual default, predicted default))
#ROC
predSVMR <-predict(svm.R, test, probability=TRUE, decision.values=TRUE)
prob <- attr (predSVMR, "probabilities")[,2]
pred <- prediction(predictions = prob, labels = test$Adimplente)
predict.rocr <- prediction(prob, test$Adimplente) # valor real da classe
perf.SVMR <- performance(predict.rocr,"tpr", "fpr")
jpeg("C:\\R\\SVMR.jpg", width = 480, height = 480, units = "px", pointsize
    ↪ = 12, quality = 100, bg = "white", res = NA)
plot(perf.SVMR, main = "ROC curve SVM Radial", col = "black", lwd = 3)
abline(a = 0, b = 1, lwd = 2, lty = 2)
dev.off()
perf.auc <- performance(predict.rocr, measure = "auc")
str(perf.auc)
unlist(perf.auc@y.values)
obs <- unfactor(test$Adimplente)
forecast <- prob
a <- verify(obs, forecast)
summary(a)
#KS
max (attr (perf.SVMR, y.values) [[1]] - attr (perf.SVMR, x.values) [[1]])
#####Benchmarking
jpeg("C:\\R\\BM.jpg",
quality = 100,
bg = "white", res = NA)
plot(perf, main = "ROC curve benchmarking", col = "blue", lwd = 2)
plot(perf.DT, add = TRUE, col = "red", lwd = 3)
plot(perf.AdaB, add = TRUE, col = "pink", lwd = 3)
```

```
333 plot(perf.bag,  add = TRUE, col = "green", lwd = 3)
334 plot(perf.LR, add = TRUE, col = "orange", lwd = 3)
335 plot(prAD, add = TRUE, col = "yellow", lwd = 3)
336 plot(perf.SVML, add = TRUE, col = "gray", lwd = 3)
337 plot(perf.SVMR, add = TRUE,  col = "black", lwd = 3)
338 abline(a = 0, b = 1, lwd = 2, lty = 2)
339 legend("bottomright", legend=c("SVMR", "SVML", "DA", "LR", "RF", "DT", "BG
      ↪ ", "ADA"), col=c("black", "gray", "yellow", "orange", "blue", "red",
      ↪  "green", "pink"), lty=1:1, cex=0.8)
340 dev.off()
```

APÊNDICE B – MATRIZ DE CONFUSÃO PARA VARIÁVEL *DEFAULT* SUPERIOR A 90 DIAS

		Valor Observado	
		Y = Adimplente	Y = Inadimplente
Valor Predito	Y = Adimplente	427539	18222
	Y = Inadimplente	3634	9368

Matriz de confusão RF.

		Valor Observado	
		Y = Adimplente	Y = Inadimplente
Valor Predito	Y = Adimplente	427685	21954
	Y = Inadimplente	3488	5636

Matriz de confusão DT.

		Valor Observado	
		Y = Adimplente	Y = Inadimplente
Valor Predito	Y = Adimplente	429286	18988
	Y = Inadimplente	1887	8602

Matriz de confusão *AdaBoost*.

		Valor Observado	
		Y = Adimplente	Y = Inadimplente
Valor Predito	Y = Adimplente	429031	13075
	Y = Inadimplente	2142	14515

Matriz de confusão Bagging.

		Valor Observado	
		Y = Adimplente	Y = Inadimplente
Valor Predito	Y = Adimplente	429486	26756
	Y = Inadimplente	1687	834

Matriz de confusão LDA.

		Valor Observado	
		Y = Adimplente	Y = Inadimplente
Valor Predito	Y = Adimplente	419448	21900
	Y = Inadimplente	11725	5690

Matriz de confusão LR.

		Valor Observado	
		Y = Adimplente	Y = Inadimplente
Valor Predito	Y = Adimplente	431173	27590
	Y = Inadimplente	0	0

Matriz de confusão SVML.

		Valor Observado	
		Y = Adimplente	Y = Inadimplente
Valor Predito	Y = Adimplente	431173	27590
	Y = Inadimplente	0	0

Matriz de confusão SVMR.