

INSPER

PROGRAMA DE MESTRADO PROFISSIONAL EM ECONOMIA

DANIEL ALVES DE BRITO FILHO

**APLICAÇÃO DE ÁRVORES DE REGRESSÃO ADITIVAS
BAYESIANAS NO DESENVOLVIMENTO DE MODELOS DE ESCORE
DE CRÉDITO NO BRASIL**

SÃO PAULO

2016

DANIEL ALVES DE BRITO FILHO

**APLICAÇÃO DE ÁRVORES DE REGRESSÃO ADITIVAS
BAYESIANAS NO DESENVOLVIMENTO DE MODELOS DE ESCORE
DE CRÉDITO NO BRASIL**

Dissertação apresentada ao Programa de Mestrado Profissional em Economia do Insper, como parte dos requisitos para obtenção do título de Mestre em Economia.

Área de concentração: Finanças

Orientador: Prof. Dr. Rinaldo Artes

SÃO PAULO

2016

Alves de Brito Filho, Daniel

Aplicação de árvores de regressão aditivas bayesianas no desenvolvimento de modelos de escore de crédito no Brasil / Daniel Alves de Brito Filho – São Paulo, 2016

XX f.

Dissertação (Mestrado – Programa de Mestrado Profissional em Economia) – Insper, 2016

Orientador: Rinaldo Artes

1. Escore de Crédito 2. Aprendizado de Máquina 3. Regressão Logística 4. BART 5. *Random Forests*

DANIEL ALVES DE BRITO FILHO

**APLICAÇÃO DE ÁRVORES DE REGRESSÃO ADITIVAS
BAYESIANAS NO DESENVOLVIMENTO DE MODELOS DE ESCORE
DE CRÉDITO NO BRASIL**

Dissertação apresentada ao Programa de Mestrado Profissional em Economia do Insper, como parte dos requisitos para obtenção do título de Mestre em Economia.

Área de concentração: Finanças

Orientador: Prof. Dr. Rinaldo Artes

DATA DA APROVAÇÃO: __/__/__

BANCA EXAMINADORA

Profa. Dra. Andrea Minardi

Inper

Profa. Dra. Lúcia Barroso

Instituto de Matemática e Estatística

Prof. Dr. Rinaldo Artes

Inper

DEDICATÓRIA

Dedico este trabalho à minha esposa Kátia e meu filho Bruno que tiveram paciência e compreensão nos momentos em que tive que me ausentar para os estudos do mestrado e conclusão desta dissertação. Também dedico este trabalho aos meus familiares que sempre me apoiaram nos estudos.

AGRADECIMENTOS

Agradeço ao meu orientador, Rinaldo Artes, pela compreensão e paciência em sempre indicar o melhor caminho para o desenvolvimento desta dissertação. A sua parceria foi essencial no meu processo de aprendizado e crescimento.

À empresa Serasa Experian, por fornecer a base de dados para o desenvolvimento e conclusão desta pesquisa.

Aos colegas do Banco Rabobank International Brasil S/A, pelo apoio e compreensão. Além da tranquilidade nos momentos em que tive que me dedicar aos estudos do mestrado e desenvolvimento desta dissertação.

RESUMO

A análise de crédito é uma atividade fundamental para as instituições financeiras. Os modelos de escore de crédito tornaram-se uma ferramenta importante, devido à necessidade de padronização e agilidade nas análises de crédito, existindo situações em que a aprovação ou recusa do crédito é totalmente automatizada. Segundo Thomas (2009), a técnica mais utilizada na construção de modelos de escore de crédito é a regressão logística. Por outro lado, outras técnicas, reunidas sob o termo aprendizado de máquina, têm sido aplicadas em modelos de classificação. Como podemos observar em Kruppa et al. (2013) e Lessmann et al. (2015), esses modelos têm apresentado resultados superiores aos modelos de regressão logística. Este trabalho propõe uma comparação entre o modelo de regressão logística e os modelos de aprendizado de máquina BART e *Random Forests*. Para o desenvolvimento dos modelos foi utilizada uma base de dados fornecida pela empresa Serasa Experian contendo informações do *bureau* de crédito referente a clientes de operações de crédito direto ao consumidor no varejo. Para a avaliação da performance dos modelos foram utilizadas a estatística de Kolmogorov-Smirnov e o coeficiente de Gini. Também foi gerado um intervalo de confiança para a métrica área sob a curva (AUC) para testar a hipótese dos modelos possuírem a mesma performance. Como principal resultado, a análise realizada confirma a superioridade do modelo BART sobre o modelo de regressão logística no banco de dados analisado. Além disso, os resultados sugerem que o modelo *Random Forests* é superior ao modelo de regressão logística somente quando ajustado na amostra balanceada analisada, dado que a performance da regressão logística melhorou quanto ajustado na base de desenvolvimento desbalanceada. Os melhores modelos BART ajustados, tanto na amostra balanceada quanto na amostra desbalanceada, foram superiores ao modelo *Random Forests*, nos dados analisados. Porém, o modelo BART padrão e *Random Forests* apresentaram performance similar e não podemos afirmar que um modelo foi superior ao outro.

Palavras-Chaves: Escore de Crédito. Aprendizado de Máquina. Regressão Logística. BART. *Random Forests*.

ABSTRACT

The credit risk assessment is a vital activity for any financial institution. The credit scoring models become an important tool due to the standardization and speed necessities on the credit process, having situations where the credit approval or rejection is fully automated. According to Thomas (2009), the logistic regression has been the most used technique to build up credit scoring models. This paper proposes a comparison between the logistic regression model and models created using machine learning techniques BART and Random Forests. The database used to develop these models was provided by Serasa Experian, which was related to retail credit transactions for consumers. The performance of these models was assessed using the Kolmogorov-Smirnov statistic and the Gini coefficient. A confidence interval was also generated to the area under curve (AUC) metric also to support models performance comparison. The main result of this paper confirms the superiority of the machine learning model BART against the logistic regression. On the other hand, results suggest a superiority of the machine learning model Random Forests model against the logistic regression only when fitted in the under sampling data base, however, the logistic regression improved when fitted in the unbalanced development data base with bias correction and its performance was the same of the Random Forests model. The best chosen BART models, fitted in both the under sampling data base and the unbalanced data base, have had a better performance against the Random Forests model. However, the standard BART model presented similar results against Random Forests and we could not conclude which one was better than the other.

Keywords: Credit Scoring. Machine Learning. Logistic Regression. BART. Random Forests.

SUMÁRIO EXECUTIVO

A análise de crédito é uma atividade fundamental para os bancos de varejo, tanto para melhorar a assertividade na concessão de crédito classificando corretamente os “bons” e “maus” pagadores, quanto na melhor precificação dos empréstimos e custo do capital. Os modelos de escore de crédito são ferramentas que auxiliam no processo análise de crédito e surgiram da necessidade de padronizar e dar agilidade ao processo. Os primeiros modelos de escore de crédito foram introduzidos na década de 1950, inicialmente para decidir sobre a concessão de crédito para o ano seguinte. A ideia dos modelos era utilizar dados de tomadores de crédito de um ou dois anos anteriores e seus históricos de crédito para construir modelos de escore de crédito com o único objetivo de prever se o tomador de crédito se tornaria um “bom” ou “mau” pagador no ano seguinte. O modelo de escore de crédito que se consagrou foi a regressão logística, uma extensão da regressão linear, que se tornou o *benchmark* na construção desses modelos. A regressão logística relaciona as características de um cliente (por exemplo, idade, sexo, salário, estado civil, etc.) com a sua classificação entre “bom” ou “mau” pagador.

Muito foi feito nos sistemas de escore de crédito, incluindo o uso de ferramentas de mineração de dados e inteligência artificial. Algumas técnicas de aprendizado de máquina (do inglês *Machine Learning*) têm sido aplicadas na análise de crédito para classificar um cliente como “bom” ou “mau” pagador. Alguns estudos comparando modelos desenvolvidos utilizando a regressão logística e modelos desenvolvidos com técnicas de aprendizado de máquina sugerem uma melhor performance destes últimos. Os modelos de aprendizado de máquina ganharam maior atenção devido ao constante incremento na capacidade de processamento dos computadores.

Este trabalho desenvolve e compara o desempenho de modelos de escore de crédito aplicados a uma base de dados fornecida pela Serasa-Experian, um *bureau* de crédito brasileiro e uma das maiores empresas do mundo em análise e informações. Foram desenvolvidos modelos baseados em regressão logística e em duas técnicas de aprendizado de máquina: BART – *Bayesian Additive Regression Trees* – e *Random Forests*.

A base de dados era composta por 10.356 observações de clientes de operações de crédito direto ao consumidor no varejo, cada um classificado como “bom” ou “mau” pagador em 2014, e 198 variáveis explicativas. Para o desenvolvimento dos modelos, a base de dados foi dividida em duas partes: uma chamada de base de desenvolvimento, utilizada para estimar os modelos e uma chamada de base de validação *out-of-sample*, utilizada para testar os modelos estimados.

Os modelos estimados foram utilizados para prever os “bons” e “maus” pagadores da base de validação. Os resultados das previsões foram comparados através de indicadores de performance, como a estatística de Kolmogorov-Smirnov e o coeficiente de Gini. Também foi gerado um intervalo de confiança para a métrica área sob a curva (AUC) para testar a hipótese dos modelos possuírem a mesma performance.

Os resultados dessa comparação indicam uma superioridade do modelo de aprendizado de máquina BART quando comparado ao modelo de regressão logística. Porém, os resultados sugerem que o modelo de aprendizado de máquina *Random Forest* foi superior ao modelo de

regressão logística somente quando ajustado na amostra balanceada analisada, dado que a performance da regressão logística melhorou quanto ajustado na base de desenvolvimento desbalanceada. Os melhores modelos BART ajustados, tanto na amostra balanceada quanto na amostra desbalanceada, foram superiores ao modelo *Random Forests*, nos dados analisados. Porém, o modelo BART padrão e *Random Forests* apresentaram performance similar e não podemos afirmar que um modelo foi superior ao outro.

Os resultados encontrados vão ao encontro de diferentes trabalhos da literatura especializada e nos faz recomendar o uso de modelos de aprendizado de máquina na avaliação de concessão de crédito.

LISTA DE ILUSTRAÇÕES

Figura 1 – Exemplo de particionamento de árvore de decisão.....	21
Figura 2 – Curvas ROC – Regressão Logística	31
Figura 3 – Curva ROC – <i>Random Forests</i>	33
Figura 4 – Curva ROC – BART padrão	34
Figura 5 – Comparação curva ROC dos modelos balanceados	36
Figura 6 – Comparação curva ROC dos modelos desbalanceados.....	37
Figura 7 – Convergência do MCMC do modelo BART padrão balanceado	47
Figura 8 – Convergência do MCMC do modelo BART padrão desbalanceado.....	47

LISTA DE TABELAS

Tabela 1 – Distribuição priori do número de nós terminais.....	25
Tabela 2 – Base fornecida pelo <i>bureau</i> de crédito.....	28
Tabela 3 – Base de desenvolvimento desbalanceada	29
Tabela 4 – Base de validação.....	29
Tabela 5 – Base de desenvolvimento balanceada.....	29
Tabela 6 – Performance do modelo de Regressão Logística reduzido balanceado	30
Tabela 7 – Performance do modelo de Regressão Logística reduzido desbalanceado .	30
Tabela 8 – Performance dos modelos Random Forests balanceados.....	32
Tabela 9 – Performance dos modelos Random Forests desbalanceados	32
Tabela 10 – Performance dos modelos BART balanceados	34
Tabela 11 – Performance dos modelos BART desbalanceados.....	34
Tabela 12 – Comparação da performance dos modelos balanceados.....	35
Tabela 13 – Testes de hipótese dos modelos balanceados	36
Tabela 14 – Comparação da performance dos modelos desbalanceados.....	37
Tabela 15 – Testes de hipótese dos modelos desbalanceados.....	38
Tabela 16 – Poder de Predição.....	49
Tabela 17 – Exemplo de Cálculo do Peso da Evidência e Valor de Informação	50
Tabela 18 – Dicionário de Dados	52
Tabela 19 – Resultado dos Cálculos do Valor de Informação	57
Tabela 20 – Resultado dos Agrupamento das Variáveis em Categorias	61
Tabela 21 – Estimação do modelo de regressão linear reduzido balanceado	64
Tabela 22 – Estimação do modelo de regressão linear reduzido desbalanceado	64

SUMÁRIO

1	INTRODUÇÃO	14
2	REVISÃO BIBLIOGRÁFICA	16
2.1	UMA REVISÃO DO ESCORE DE CRÉDITO	16
2.2	TÉCNICAS UTILIZADAS NOS MODELOS DE ESCORE DE CRÉDITO.....	17
2.3	MODELOS BASEADOS EM ÁRVORES.....	17
3	METODOLOGIA	20
3.1	ÁRVORES DE CLASSIFICAÇÃO	20
3.2	REGRESSÃO LOGÍSTICA	22
3.3	RANDOM FORESTS	22
3.4	BART	23
3.4.1	<i>Modelo de Soma de Árvores</i>	23
3.4.2	<i>Priori de Regularização</i>	24
4	BASE DE DADOS	27
5	DESENVOLVIMENTO DOS MODELOS	28
5.1	PREPARAÇÃO DA BASE	28
5.2	BASE PARA DESENVOLVIMENTO E VALIDAÇÃO DO MODELO	28
5.3	APLICAÇÃO DOS MODELOS.....	29
5.3.1	<i>Regressão Logística</i>	30
5.3.2	<i>Random Forests</i>	31
5.3.3	<i>BART</i>	33
5.3.4	<i>Comparação dos Modelos</i>	35
6	CONCLUSÕES	39
7	REFERÊNCIAS BIBLIOGRÁFICAS	41
8	APÊNDICES	45
8.1	APÊNDICE 1 – GERAÇÃO DE AMOSTRAS POSTERIORI E INFERÊNCIA.....	45
8.2	APÊNDICE 2 – PESO DA EVIDÊNCIA E VALOR DA INFORMAÇÃO.....	48
8.3	APÊNDICE 3 – MÉTODOS DE COMPARAÇÃO DE ÁREAS SOB A CURVA	51
8.4	APÊNDICE 4 – DICIONÁRIO DE DADOS	52
8.5	APÊNDICE 5 – RESULTADO DOS CÁLCULOS DO VALOR DE INFORMAÇÃO	57
8.6	APÊNDICE 6 – RESULTADO DAS RECATEGORIZAÇÕES DAS VARIÁVEIS	61
8.7	APÊNDICE 7 – ESTIMAÇÃO DOS MODELOS DE REGRESSÃO LINEAR.....	64

1 INTRODUÇÃO

Uma das principais ferramentas utilizadas na análise de crédito atualmente é o escore de crédito. De acordo com Thomas, Edelman e Crook (2002), o escore de crédito tem sido fundamental para permitir o crescimento de consumo nas últimas quatro décadas.

Segundo Anderson (2007), o estudo do escore de crédito pode ser dividido em duas partes. A primeira é o significado da palavra crédito que deriva do latim *credo*, que significa “Eu acredito” ou “Eu confio” e também está associada ao sentido de “consumo imediato e pagamento no futuro”. Em segundo lugar, o escore é uma ferramenta para ordenar os tomadores de empréstimo de acordo com suas características.

A análise de crédito é uma atividade fundamental para os bancos de varejo. Os modelos de escore de crédito tornaram-se uma ferramenta importante devido à necessidade de padronização e agilidade nas análises, existindo, hoje, situações em que a aprovação ou recusa do crédito é totalmente automatizada. Dada a importância do processo de concessão de crédito para os bancos, ser mais assertivo na identificação dos clientes com maior probabilidade de inadimplência, significa reduzir custos (pela redução da perda por inadimplência) ou aumentar as receitas (pelo melhor aproveitamento dos bons clientes que poderiam não ser considerados por serem erroneamente classificados como maus pagadores).

Desde 2004, com o acordo de Basileia II (BCBS, 2006; BIS, 2004), os bancos foram incentivados a melhorar seus modelos internos de risco de crédito para conseguir a autorização para utilizá-los como base para alocação de capital ajustado a esse risco. Para conseguir a certificação para utilização de modelos avançados de escore de crédito, conhecidos como *Advanced Internal Rating Approach (A-IRB)*, os bancos precisam demonstrar sua capacidade de avaliar seus riscos de forma acurada. Bancos com a certificação para utilização do *A-IRB* possuem vantagens competitivas em relação a outros bancos, pois são autorizados pelos reguladores a alocar menor capital ao risco de crédito.

Muito foi feito no desenvolvimento de sistemas de escore de crédito, incluindo o uso de ferramentas de mineração de dados e inteligência artificial. Porém, novas abordagens ainda são restritas pela falta de aceitação de modelos menos intuitivos ou que vão além dos modelos padronizados implantados na indústria bancária, desenvolvidos pelos próprios bancos ou fornecidas por provedores de solução. Desenvolver e implantar um sistema de escore de crédito demanda tempo e investimento e pode levar de 12 a 18 meses, por isso, não é surpresa que bancos utilizam modelos de escore de crédito inalterados por vários anos.

Segundo Thomas (2009), a técnica mais utilizada na construção de modelos de escore de crédito é a regressão logística. Por ser a técnica mais utilizada, ela será utilizada como *benchmark* para comparação do desempenho dos modelos de escore de crédito desenvolvidos nesta dissertação.

Algumas técnicas de aprendizado de máquina têm sido utilizadas em diversas áreas do conhecimento como medicina, biologia e genética, principalmente em problemas de classificação. Estas técnicas podem ser aplicadas na análise de crédito para classificar um

tomador de empréstimo como “bom” ou “mau” pagador. Como podemos observar em Kruppa et al. (2013) e Lessmann et al. (2015), esses modelos têm apresentado resultados superiores aos modelos de regressão logística.

Um dos principais objetivos desta dissertação é verificar, em uma base de dados real de um *bureau* de crédito brasileiro, fornecida pela empresa Serasa Experian, a eficácia de dois modelos de aprendizado de máquina. A primeira técnica é chamada *Bayesian Additive Regression Trees* (BART), onde diversas árvores de classificação são combinadas de forma a aumentar poder preditivo da análise. Além disso, o BART propõe a definição de uma distribuição à priori para os parâmetros do modelo de forma a induzir uma distribuição posteriori conveniente gerando, supostamente, um resultado final mais promissor. A segunda técnica é chamada de *Random Forests*, que também combina árvores de classificação para aumentar o poder preditivo da análise.

Os modelos de escore de crédito encontrados na literatura normalmente são aplicados a bases de dados fornecidas por empresas privadas. Após revisão bibliográfica não foi encontrada nenhuma aplicação em uma base de dados obtida diretamente de um *bureau* de crédito. Além disso, também após a revisão bibliográfica, foi encontrada apenas uma aplicação do modelo BART em escore de crédito, os autores Zhang and Härdle (2010) aplicaram o modelo em um banco de dados de empresas Alemãs (“*German Creditreform database*”). Também não foi encontrado nenhum estudo de modelagem de escore de crédito que utilizou um intervalo de confiança para comparação da eficiência dos modelos.

Esta dissertação está dividida em oito capítulos: Capítulo 1, esta introdução. Capítulo 2 faz uma revisão bibliográfica da história do desenvolvimento dos modelos de escore de crédito, a evolução destes modelos e as principais técnicas utilizadas nos modelos escore de crédito, incluindo modelos baseados em aprendizado de máquina – como BART e *Random Forests*. O Capítulo 3 apresenta as principais técnicas utilizadas nesta dissertação para o desenvolvimento e comparação da performance dos modelos escores de crédito. O Capítulo 4 descreve a base de dados utilizada para o desenvolvimento dos modelos. O Capítulo 5 descreve o desenvolvimento do modelo, desde a preparação da base de dados até a aplicação dos modelos e a comparação dos resultados. O Capítulo 6 traz as principais conclusões encontradas pela aplicação das técnicas de aprendizado de máquina e sugestões para trabalhos futuros. Demais capítulos apresentam as referências bibliográficas e os apêndices.

2 REVISÃO BIBLIOGRÁFICA

Este capítulo está dividido em três partes, na primeira é feita uma revisão bibliográfica sobre o desenvolvimento e utilização do escore de crédito, a segunda sobre técnicas utilizadas no desenvolvimento de modelos de escore de crédito e a terceira sobre os modelos baseados em árvores.

2.1 UMA REVISÃO DO ESCORE DE CRÉDITO

A análise de crédito pode ser considerada como um problema de classificação de clientes entre maus e bons pagadores. Seu principal objetivo é prever se um indivíduo se tornará um mau pagador em um determinado horizonte de tempo, dado um conjunto de características.

Os métodos de classificação tentam determinar uma função que melhor segregue os indivíduos entre bons e maus pagadores. O primeiro método para segregar grupos foi proposto no trabalho original de Fisher (1936), aplicado a um problema geral de classificação de variedades de plantas, neste trabalho foram desenvolvidos os princípios da Análise Discriminante. Durand et al (1941) aplicou esta metodologia na área de finanças para distinguir entre bons e maus pagadores de um empréstimo. A análise discriminante foi o primeiro método utilizado para o desenvolvimento de um escore de crédito. Altman (1968) utilizou o método para previsão de falência de empresas corporativas.

De acordo com Sousa, Gama e Brandão (2016), as primeiras aplicações do modelo de escore de crédito foram focadas na concessão de dois tipos de empréstimo: crédito ao consumidor e crédito a empresas. Com o grande crescimento do segmento de cartões de crédito demandou-se que a atividade de concessão de crédito fosse automatizada, o que foi possível graças ao crescimento do poder computacional.

Chandler e Coffman (1979) apontam como os principais benefícios do escore de crédito o ganho de tempo de processamento das propostas de crédito, com consequência na maior agilidade na tomada de decisão; minimização dos custos e esforços do processo de crédito; e diminuição dos erros cometidos.

A pesquisa de Abdou e Pointon (2011) apontou que a coleta de informação é um problema crítico na construção dos modelos de escore de crédito, normalmente as informações das características dos indivíduos são utilizadas (como sexo, idade, salário, estado civil, etc.). A seleção de variáveis é baseada em uma análise estatística, porém de acordo com a pesquisa nenhum autor estabeleceu uma razão teórica para seleção das variáveis do modelo, que em geral depende do banco de dados disponibilizado pelas instituições privadas. Ainda, segundo a pesquisa de Abdou e Pointon (2011), as principais críticas aos modelos de escore de crédito são:

- algumas vezes fatores econômicos não são considerados; em geral os modelos não são padronizados e são diferentes de um mercado para o outro, pois não se busca estabelecer uma razão teórica para a seleção das variáveis do modelo, ou seja, a seleção de variáveis é, em geral, baseada em uma análise estatística;
- a utilização de um ponto de corte no escore para classificação entre “bons” e “maus” pagadores pode causar erros de classificação como rejeitar um bom pagador ou aceitar

um mau pagador; os modelos são normalmente baseados no passado, a não ser que sejam atualizados constantemente; o resultado dos modelos é geralmente dicotômico (“bom” ou “mau” pagador), enquanto pode haver outras saídas possíveis como atrasos de pagamentos, prorrogações e renegociações. Vale ressaltar a existência de modelos que permitem a previsão de múltiplos eventos.

Ainda, a pesquisa de Abdou e Pointon (2011) apontou que a coleta de informação é um problema crítico na construção dos modelos de escore de crédito. O tamanho da amostra também é outro problema bastante discutido. Em algumas aplicações para o crédito no varejo, utilizaram-se amostras menores do que 1100 observações.

2.2 TÉCNICAS UTILIZADAS NOS MODELOS DE ESCORE DE CRÉDITO

A regressão logística (Steenackers e Goovaerts, 1989; Hosmer Jr e Lemeshow, 2013) foi utilizada para o desenvolvimento de escore de crédito e tornou-se o método mais utilizado na indústria financeira (Anderson, 2007). O uso de técnicas de inteligência artificial, importadas da teoria de aprendizagem estatística, como árvores de decisão (Breiman, Friedman, Olshen e Stone 1984) e redes neurais (Desai, Crook e Overstreet, 1996; Malhotra e Malhotra, 2002) cresceram em sistemas de escore de crédito. Métodos de aprendizagem estatística receberam grande atenção na última década em pesquisas relacionadas à área de finanças, tanto para construção de escore de crédito quanto para previsão de falência (Li, Shiue e Huang, 2006), classificação de falência (Lensberg, Eilifsen e McKee, 2006), análise de estresse (Gestel, 2006) e decisão de aplicação e retorno financeiro (West, Dellana e Qian, 2005; Xia et al., 2000). Além disso, técnicas de regressões (Lee e Chen, 2005) e agrupamento (Wei, Yun-Zhong e Ming-Shu, 2014) também foram adaptadas para o problema de escore de crédito.

A escolha de algoritmos de aprendizagem é um problema, pois depende dos métodos disponíveis e da preferência do usuário (Jain, Duin e Mao, 2000). Como alternativa para utilização de um único método, a tendência está evoluindo para o uso de sistemas híbridos (Hsieh, 2005), por exemplo, o uso de técnicas de agrupamento para separar e isolar amostras não representativas e redes neurais para a modelagem do escore de crédito. Novos conceitos de adaptação às mudanças (Pavlidis, 2012) e modelagem dinâmica (Crook e Bellotti, 2010) estão começando a ser explorados na análise de risco de crédito.

Ferramentas padronizadas de classificação incluem análise discriminante linear e quadrática, além do modelo logístico. Por outro lado, outras técnicas, como modelos baseados em aprendizado de máquina, têm sido aplicadas em modelos de classificação. Como podemos observar em Kruppa et al. (2013) e Lessmann et al. (2015), esses modelos têm apresentado resultados superiores aos modelos de regressão logística.

2.3 MODELOS BASEADOS EM ÁRVORES

Outra técnica não paramétrica conhecida como partição recursiva ou modelo de classificação com regressão de árvore, também conhecida com *Classification Regression Tree* (CART), foi proposto por Breiman, Friedman, Olshen e Stone (1984). O modelo consiste em construir uma árvore baseada na subdivisão binária da amostra em sub árvores. Primeiramente,

seleciona-se a variável independente que segrega a amostra entre, por exemplo, bons e maus pagadores, da forma mais eficiente possível baseada no menor custo de erro de classificação (Zekic-Susac, Sarlija e Bencic, 2004); formam-se assim dois grupos (nós) de observações. Em cada um dos grupos, separadamente, identifica-se uma nova variável independente que permita, de maneira eficiente discriminar bons e maus pagadores – novos grupos de observações são criados. O processo é repetido de forma recursiva até que a amostra seja particionada adequadamente – atendendo a critérios de qualidade pré-definidos.

Outros autores também exploraram métodos alternativos baseados no CART. Breiman (1996) propôs o método chamado *Bagging*, onde várias árvores de decisão são criadas de forma aleatória para posteriormente serem combinadas, cada indivíduo será classificado como “bom” ou “mau” pagador por cada uma das árvores geradas pelo processo em um esquema de votação e a classificação final do cliente será obtido pela maioria da contagem das árvores. Breiman (2001) propôs uma alteração ao modelo *Bagging*, que chamou de *Random Forests*, onde algumas variáveis independentes também são selecionadas de forma aleatória na construção de cada uma das árvores de decisão. Podemos observar a técnica de *Random Forests* aplicada principalmente nas áreas da genética e medicina. Porém, como podemos observar em Kraus (2014), Zhou et al. (2012) e Malekipirbazari et al. (2015), alguns estudos obtiveram bons resultados aplicando a técnica de *Random Forests* à modelo de escore de crédito.

Chipman, George e McCulloch (1998) propuseram uma versão bayesiana para o modelo CART. Os dois pontos fundamentais desse trabalho foram definir uma priori para os parâmetros do modelo a fim de induzir a distribuição posteriori e através de um processo estocástico selecionar um modelo mais promissor. A vantagem desta abordagem está no fato de ao se especificar uma priori para os parâmetros, pode-se dar menor peso para modelos indesejados e expressar preferência para certos preditores. A consequência é que o modelo vai determinar uma distribuição posteriori com maior probabilidade para as melhores árvores. O algoritmo Metropolis-Hastings (Hastings, 1970; Peskun, 1973) pode ser utilizado neste processo estocástico para estimar as distribuições a posteriori.

Chipman, George e McCulloch (2010) estenderam seu trabalho e propuseram um modelo baseado na soma de árvores bayesianas, o modelo foi chamado de *Bayesian Additive Regression Trees* (BART). A ideia principal deste modelo é a imposição de uma priori que efetivamente regulariza o modelo, mantendo as árvores individuais simples. Sem tal regularização, grandes árvores individuais poderiam prejudicar o modelo e limitar a vantagem da soma de árvores. A soma de árvore permite que o ajuste do modelo aos dados aconteça de forma livre de uma árvore para outra, por isso o modelo BART possui uma flexibilidade para problemas complexos. Conceitualmente, o modelo BART pode ser classificado como um modelo Bayesiano não paramétrico. Os autores comparam o desempenho do modelo BART com a regressão logística e outros métodos baseados em aprendizagem estatística como LASSO (Efron et al., 2004), *Gradient Boosting* (Friedman, 2001), redes neurais (Desai, Crook e Overstreet, 1996; Malhotra e Malhotra, 2002) e o *Random Forests* (Breiman, 2001). A conclusão dos autores é que o modelo BART pode ser comparado favoravelmente com outros métodos de aprendizado de máquina.

O método BART foi aplicado para modelagem de risco de crédito por Zhang and Härdle (2010). Os autores aplicaram o método BART para o escore de crédito e o renomearam para *Bayesian Additive Classification Tree* (BACT). Os autores utilizaram um banco de dados de empresas Alemãs ("*German Creditreform database*"), que continha informações financeiras de 20.000 empresas solventes e 1.000 empresas insolventes no período de 1996 e 2002. Os autores também comparam o desempenho do método BACT com a regressão logística e outros métodos baseados em aprendizagem estatística como CART (Breiman, Friedman, Olshen e Stone, 1984), *Random Forest* (Breiman, 2001) e *Gradient Boosting* (Friedman, 2001). A conclusão dos autores é que o modelo BACT apresenta ótimo desempenho quando comparado com a regressão logística e outros métodos de aprendizado de máquina. O método BACT também é robusto para valores extremos das variáveis de entrada e pode ser ajustado para bases de dados com pouco número de observações.

3 METODOLOGIA

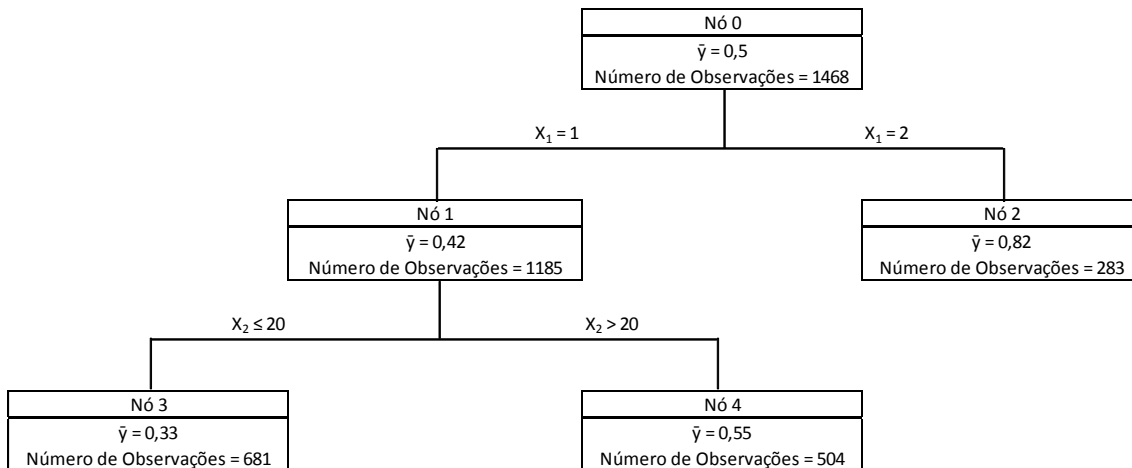
Neste capítulo serão revistas as metodologias utilizadas neste trabalho para a previsão do escore de crédito. A seção 3.1 introduz a ideia do modelo de árvore de classificação. Da seção 3.2 até a seção 3.4 são apresentados os modelos que serão desenvolvidos como a regressão logística, *Random Forests* e o BART.

3.1 ÁRVORES DE CLASSIFICAÇÃO

Para melhor entendimento dos modelos apresentados nesta dissertação é necessária uma breve explicação do modelo de árvores de classificação proposto por Breiman, Friedman, Olshen e Stone (1984). Neste contexto, há uma variável dependente e um conjunto de variáveis independentes. O objetivo é, tendo por base os valores observados das variáveis independentes (X_1, \dots, X_p), particionar a amostra em grupos homogêneos em relação à variável dependente (y).

A construção de uma árvore de classificação binária começa pela identificação da variável independente que melhor segregue a amostra em grupos distintos em relação à variável dependente (Zekic-Susac, Sarlija e Bencic, 2004). Formam-se assim dois grupos (nós) de observações. Na Figura 1 temos o “Nó 0” (nó pai) com toda a base de dados, contendo 1.468 observações e com valor médio da variável dependente de 0,5. O nó pai será dividido em dois novos nós denominados “Nó 1” e “Nó 2”, também chamados de nós filhos, para esta divisão será utilizada a variável que melhor segrega em “bons” e “maus” pagadores segundo algum critério de qualidade, no caso, busca-se a construção de grupos nos quais as diferenças entre as médias da variável dependente seja a maior possível. A variável X_1 , que assume os valores 1 ou 2, foi a escolhida, sendo que o “Nó 1” receberá todos as observações com $X_1 = 1$ e o “Nó 2” as observações com $X_1 = 2$. Em cada um dos grupos, separadamente, identifica-se uma nova variável independente que permita, de maneira eficiente segregar a amostra – novos grupos de observações são criados. O processo é repetido de forma recursiva até que algum critério de parada seja satisfeito, ou seja, a partir da aplicação do critério o nó não será mais dividido e será um nó terminal – atendendo a critérios de qualidade pré-definidos. Um exemplo de critério de parada seria o número mínimo de indivíduos em um nó terminal, ou seja, o nó não será mais dividido caso tenha um número de indivíduos menor ou igual a um parâmetro pré-estabelecido. Voltando à Figura 1, temos o “Nó 1” que será dividido em dois novos nós: 3 e 4, sendo que a variável X_2 traz a melhor divisão do “Nó 1”. Assim, o “Nó 3” receberá as observações com $X_1 = 1$ e $X_2 \leq 20$ e o “Nó 4” as observações com $X_1 = 1$ e $X_2 > 20$. Os nós “Nó 2” ($X_2 = 2$), “Nó 3” e “Nó 4” são chamados de nós folha ou nós terminais, pois a partir destes nós a amostra não será mais particionada.

Figura 1 – Exemplo de particionamento de árvore de decisão



Podemos aqui introduzir a notação do modelo com uma única árvore de decisão:

- $x_i = \{x_{1i}, \dots, x_{pi}\}$ são os valores observados para as variáveis independentes para o indivíduo i .
- T é uma árvore binária com um conjunto de nós interiores e um conjunto de nós terminais.
- L é o número de nós terminais da árvore T , sendo os nós terminais $l = 1, \dots, L$.
- $M = \{\mu_1, \mu_2, \dots, \mu_L\}$ são parâmetros associados a cada nó terminal l da árvore T .

Na árvore apresentada na Figura 1, $L = 3$ e $M = \{0,82; 0,33; 0,55\}$.

As regras de decisão são divisões binárias do espaço de predição na forma $\{x_i \in A\}$ e $\{x_i \notin A\}$, onde A é um subconjunto do domínio de x_i ;

Cada x_i está associado a um único nó terminal de T de acordo com uma sequência de regras de decisões do topo da árvore para baixo. Então, x_i está associado a um único parâmetro μ_l referente ao seu nó terminal.

Chipman, George e McCulloch (1998) definem o modelo de uma única árvore como:

$$y_i = g(x_i; T, M) + \varepsilon_i, \text{ sendo } \varepsilon_i \text{ um erro aleatório.} \tag{1}$$

Dados T e M , utilizamos a função $g(x_i; T, M)$ para associar o indivíduo i a um nó terminal $\mu_l \in M$. Abaixo está a representação da função para o caso da Figura 1. Por exemplo, dado um indivíduo representado por $x_1 = \{1, 10\}$, a função $g(x_1; T, M)$ o associaria ao parâmetro $\mu_3 = 0,33$.

$$g(x_i; T, M) = \begin{cases} 0,82, & \text{se } x_{1i} = 2 \\ 0,33 & \text{se } x_{1i} = 1 \text{ e } x_{2i} \leq 20 \\ 0,55 & \text{se } x_{1i} = 1 \text{ e } x_{2i} > 20 \end{cases} \tag{2}$$

3.2 REGRESSÃO LOGÍSTICA

A regressão logística deriva da regressão linear, sendo que a regressão linear é utilizada quando a variável dependente pode assumir qualquer valor. Porém, existem situações em que a variável dependente pode assumir apenas dois valores. Em escore de crédito a variável dependente é binária ($y = 0$ ou 1) e determina se um tomador de empréstimo é um “bom” ou “mau” pagador.. Seja $p_i = P(y = 1)$, probabilidade de um cliente ser um “mau” pagador, então o modelo de regressão logística é dado por:

$$p_i = \frac{e^{\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_p x_{pi}}}{1 + e^{\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_p x_{pi}}}. \quad (3)$$

O modelo pode ser linearizado por meio da transformação logito dada por:

$$\ln\left(\frac{p_i}{1 - p_i}\right) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_p x_{pi} \quad (4)$$

É comum que as amostras possuam baixo número de observações de maus pagadores quando comparado aos bons pagadores (Brown e Mues, 2012). Este desbalanceamento de classes está presente e a técnica de regressão logística pode não ser capaz de lidar bem com a situação, gerando modelos ruins (King e Zeng, 2001). Uma solução é aumentar a classe minoritária sinteticamente (Chawla, Bowyer, Hall e Kegelmeyer, 2002) ou reduzir a classe majoritária (Yap et al, 2014). Outra abordagem é fazer a correção do viés do modelo através da mudança no intercepto do modelo, para classificação dos clientes entre “bons” e “maus” pagadores (King e Zeng, 2001). Esse ajuste não altera a ordem de classificação dos escores dos clientes gerados por meio da regressão logística, alterando apenas os eventuais pontos de corte utilizados no processo de prever bons e maus clientes.

3.3 RANDOM FORESTS

O modelo *Random Forests* é uma evolução do modelo *Bagging* (Breiman, 1996), por isso faz sentido iniciar a explicação pelo modelo *Bagging*.

No *Bagging* várias árvores de decisão são criadas de forma aleatória para posteriormente serem combinadas. Cada árvore é formada a partir de uma amostra *bootstrap*, extraída com reposição, da amostra original de desenvolvimento. Este procedimento é repetido diversas vezes até o limite pré-estabelecido de números de árvores. Por último, cada indivíduo será classificado como “bom” ou “mau” pagador por cada uma das árvores geradas pelo processo anterior. A classificação final do cliente é obtida por meio de um esquema de votação levando-se em conta o número de árvores que classificam o cliente como “bom” ou “mau” – se a maioria das árvores o classifica como “bom” ele será considerado “bom”, o inverso se a maioria o classificar como “mau”. Segundo Breiman (1996), o *Bagging* representa uma evolução em relação aos modelos de uma única árvore, pois é mais estável na previsão de novos clientes.

Porém, o modelo *Bagging* sempre utiliza as mesmas variáveis na montagem das árvores, o que pode tornar os resultados observados em diferentes árvores, semelhantes.

O *Random Forests* sugere uma alteração em relação ao *Bagging*, onde algumas, ou todas, variáveis independentes também são selecionadas de forma aleatória na construção de cada uma das árvores de decisão. O percentual de variáveis independentes que serão selecionadas em cada sorteio aleatório é um parâmetro do modelo. Segundo Breiman (2001), essa seleção aleatória de variáveis tem o potencial de utilizar toda a informação disponível no conjunto de variáveis.

3.4 BART

O modelo *Bayesian Additive Regression Trees* (BART) pode ser apresentado em três partes, primeiro o modelo aditivo de árvores de decisão, segundo a especificação da priori para os parâmetros do modelo de forma a induzir a distribuição posteriori e por último o processo estocástico para geração da distribuição posteriori.

3.4.1 MODELO DE SOMA DE ÁRVORES

Alguns modelos de classificação assumem a existência de uma variável contínua y^* , não observável, que determina no valor de y , sendo:

$$\begin{cases} y = 1; \text{ se } y^* \geq 0 \\ y = 0; \text{ se } y^* < 0 \end{cases} \quad (6)$$

A ideia do modelo BART proposta por Chipman, George e McCulloch (2010) é relacionar y^* com as variáveis independentes por meio de um modelo probito aditivo, em que cada termo da soma é representado por uma árvore baseada nos preditores x_i . Podemos então escrever o modelo aditivo de árvores como sendo:

$$y_i^* = g(x_i; T_1; M_1) + \dots + g(x_i; T_m; M_m) + \varepsilon = \sum_{k=1}^m g(x_i; T_k; M_k) + \varepsilon = G(x_i) + \varepsilon, \quad (7)$$

em que:

$\varepsilon \sim N(0,1)$ é o erro aleatório do modelo.

m é o número de árvores que serão utilizadas no modelo.

T_k é a árvore de decisão k com um conjunto de nós interiores de decisão e um conjunto de nós terminais.

L_k é o número de nós terminais da árvore T_k , sendo os nós terminais $l = 1, \dots, L_k$.

$M_k = \{\mu_{1k}, \mu_{2k}, \dots, \mu_{L_k k}\}$ são os parâmetros associados à árvore T_k e

$$G(x) = \sum_{k=1}^m g(x; T_k; M_k).$$

cada x_i está associado a um único nó terminal de T_k . Então, o indivíduo i está associado a um único μ_{lk} referente a um nó terminal da árvore T_k .

dados T_k e M_k , utilizamos a função $g(x_i; T_k, M_k)$ para associar um $\mu_{lk} \in M_k$ a um indivíduo. Na aplicação de escore de crédito seria associar a predição μ_{lk} referente ao nó terminal da árvore T_k aos valores dos preditores de entrada x_i .

Note que,

$$P(y = 1|x_i) = P(y^* \geq 0|x_i) = P(G(x_i) + \varepsilon \geq 0) = P(G(x_i) \geq \varepsilon) = \Phi[G(x_i)], \quad (8)$$

onde Φ é a função distribuição acumulada da normal padrão.

3.4.2 PRIORI DE REGULARIZAÇÃO

Com o objetivo de fazer inferência utilizando o modelo de soma de árvores de uma forma Bayesiana, precisamos especificar uma distribuição priori conjunta para a estrutura da árvore T_i e os parâmetros μ_{lk} associados aos nós terminais da árvore T_k . Assumimos que a priori da estrutura de árvores dos parâmetros associados aos nós terminais possuem distribuições independentes. A distribuição priori conjunta pode ser definida como:

$$p\{(T_1, M_1), (T_2, M_2), \dots, (T_m, M_m)\} = \prod_{k=1}^m p(T_k, M_k) = \prod_{k=1}^m p(M_k|T_k)p(T_k),$$

definindo $p(M_k|T_k) = \prod_{l=1}^L p(\mu_{lk}|T_k)$, temos:

$$p\{(T_1, M_1), (T_2, M_2), \dots, (T_m, M_m)\} = \prod_{k=1}^m p(T_k) \prod_{l=1}^L p(\mu_{lk}|T_k). \quad (9)$$

PRIORI PARA T_k

Utilizando a sugestão dada por Chipman, George e McCulloch (2010), a construção da distribuição priori para T_k passa por três etapas:

1) Modelagem da probabilidade de um nó não ser terminal

Esta probabilidade está diretamente associada ao tamanho da árvore. Os autores sugerem que a probabilidade de uma divisão do nó η seja dada por:

$$p_{DIVISAO}(\eta|T_k) = \frac{\alpha}{(1 + d_{\eta k})^\beta}, \quad (10)$$

sendo:

$d_{\eta k}$ é a profundidade (nível) do nó n na árvore,

$\alpha \in (0,1)$ e $\beta \in [0, \infty)$ são hiperparâmetros positivos que controlam a probabilidade de uma nova partição, quanto maior a profundidade do nó η na árvore T , menor será a probabilidade de uma nova partição e maior a probabilidade do nó se tornar um nó terminal. Os autores Chipman, George e McCulloch (2010) recomendam a utilização dos valores padrões: $\alpha = 0,95$ e $\beta = 2$, com estes parâmetros maior probabilidade será dada para a geração de árvores individuais de tamanho 2 ou 3. O pacote *bartMachine* do software R apresenta todos os parâmetros citados que devem ser fornecidos pelo usuário. A Tabela 1 mostra a distribuição priori do número de nós terminais de uma árvore para diferentes valores de α e β .

Tabela 1 – Distribuição priori do número de nós terminais

	Configuração 1	Configuração 2	Configuração 3
α	0,50	0,95	0,95
β	2,00	2,00	0,10
Probabilidade priori de árvores com 1 nó terminal	0,500	0,050	0,050
Probabilidade priori de árvores com 2 nós terminais	0,383	0,552	0,012
Probabilidade priori de árvores com 3 nós terminais	0,098	0,275	0,004
Probabilidade priori de árvores com 4 nós terminais	0,017	0,092	0,002
Probabilidade priori de árvores com ≥ 5 nós terminais	0,003	0,031	0,932

Fonte: Zhang and Härdle (2010)

2) Modelagem da variável que dividirá cada nó

Os autores recomendam o uso de uma distribuição uniforme para a escolha da variável que será utilizada na divisão do nó (Chipman, George e McCulloch, 1998). Assim, se existirem a_η variáveis disponíveis no nó η , a probabilidade de uma dessas variáveis ser escolhida é $1/a_\eta$.

3) Modelagem da regra de decisão em cada nó, dada a variável que definirá a divisão

A sugestão, para a escolha do ponto de corte, é novamente o uso de uma especificação uniforme (termo sugerido por Chipman, George e McCulloch, 1998), agora sobre os diferentes valores assumidos pela variável, que orientará a formação de uma nova divisão para a definição do ponto de corte:

- se a variável for quantitativa, deve-se escolher o valor do ponto de corte aleatoriamente entre os valores assumidos pela variável do nó η ;
- se a variável for qualitativa a escolha se dará de modo aleatório entre todas as possibilidades de partição.

PRIORI PARA $\mu_{lk}|T_k$

A distribuição priori dos parâmetros $\mu_{lk}|T_k$ associados aos nós terminais da árvore T_k é definida como $\mu_{lk} \sim N(0; \sigma_u^2)$. Para simplificação do modelo e facilitação da geração da

distribuição posteriori, assume-se que μ_{lk} segue uma distribuição normal e a média dos parâmetros μ_{lk} é zero, ou seja, $E[\mu] = 0$.

Do modelo de soma de árvores temos que y^* associa-se à soma dos m diferentes parâmetros μ_{il} , ou seja, para um vetor de preditores x um parâmetro μ_{il} é associado pela função $g(x; T_k; M_k)$. Dado que os m diferentes parâmetros μ_{lk} têm distribuições priori independentes temos que:

$$G(x) \sim N(0; m\sigma_u^2) \tag{11}$$

Assim, podemos definir um intervalo de confiança para $G(x)$ como sendo:

$$\begin{aligned} G_{min} &= -z\sigma_u\sqrt{m} \\ G_{max} &= z\sigma_u\sqrt{m}, \end{aligned} \tag{12}$$

onde z determina o nível de confiança, por exemplo, se $z = 2$, então a probabilidade dos valores de y^* estarem entre o intervalo G_{min} e G_{max} é que 95,45%.

A partir de (7), temos $P(y = 1|x) = \Phi[G(x)]$. Desse modo, Chipman, George e McCulloch (2010) sugerem que seria razoável supor, por exemplo, que $-3 \leq G(x) \leq 3$. Assim, um modo de determinar o valor de σ_u , seria a partir de uma escolha conveniente de z fazer $G_{max} = 3$ (consequentemente, $G_{min} = -3$). Desse modo,

$$\sigma_u = \frac{3}{z\sqrt{m}} \tag{13}$$

Os autores recomendam o uso de z entre 1 e 3, tendo obtido bons resultados com $z = 2$.

Para computação do problema Bayesiano formulado foi utilizada a técnica de MCMC (do inglês *Monte Carlo Markov Chain*) resumida no Apêndice 1.

Segundo Chipman, George e McCulloch (2010), o BART possui uma característica interessante, pois consegue facilmente detectar e inferir modelos reduzidos em problemas com grande número de variáveis e amostras com pequena quantidade de observações. Este método de seleção de variáveis é menos efetivo quando a quantidade de árvores é muito alta, pois tende a misturar variáveis predictoras importantes com aquelas não relevantes. Bleich et al. (2014) propôs um modelo de seleção de variáveis baseado no BART, que utiliza como critério a proporção de vezes que a variável foi utilizada como regra de segregação de um novo ramo das diversas árvores dividido pelo total de ramos do modelo, ou seja, seleciona aquelas variáveis que aparecem com mais frequência no modelo ajustado de soma de árvores. O pacote *bartMachine* do software R implementa o procedimento sugerido por Bleich et al. (2014).

4 BASE DE DADOS

A base de dados utilizada neste trabalho foi fornecida pela empresa Serasa Experian (que será referenciada neste trabalho como *bureau* de crédito), sendo formada por clientes de operações de crédito direto ao consumidor no varejo. Esta base se diferencia dos estudos convencionais de escore de crédito no varejo, que normalmente utilizam variáveis como as características dos indivíduos, pois traz variáveis específicas de um *bureau* de crédito. O banco de dados fornecido pelo *bureau* de crédito possui 10.356 observações de clientes de operações de crédito direto ao consumidor no varejo e 198 variáveis do ano de 2014. O dicionário de dados é apresentado no Apêndice 4.

O primeiro grupo de variáveis preditoras inclui a quantidade de demanda por crédito de um específico tomador de empréstimo, em diversos segmentos diferentes e em diversos períodos de tempo. Os segmentos são cheques, imóveis, bancos, financeiras, indústrias, seguros, serviços, telefonia, varejista, utilidades e outros. Os períodos de demanda por crédito são até 30 dias, de 31 dias a 60 dias, de 61 dias a 90 dias, de 91 dias a 180 dias e de 181 dias até 360 dias. Das variáveis disponíveis o *bureau* recomendou desconsiderar as quantidades de demandas por crédito do segmento de telecomunicações e em todos os períodos, isso excluiu cinco variáveis do banco de dados. Após as exclusões este grupo totalizou 68 variáveis independentes.

O segundo grupo de variáveis preditoras se relaciona ao primeiro grupo e medem o tempo em dias desde a primeira demanda e desde a última demanda por crédito de um específico tomador de empréstimo por diversos segmentos. Os segmentos são cheques, bancos, financeira, seguros, telecomunicação e varejo. Este grupo totalizou 12 variáveis independentes.

O terceiro grupo de variáveis preditoras está relacionado à quantidade de eventos do tomador de empréstimo registrados no *bureau* de crédito em determinados períodos de tempo. Os eventos registrados no *bureau* são dívidas ativas ou resolvidas, protestos, cheques devolvidos, negativas por bancos ou financeiras ativas ou resolvidas, negativas por empresas que não são bancos ou financeiras ativas ou resolvidas e credores ativos. Os períodos de tempo variam de 1 mês, 2 meses, 3 meses, 6 meses, 12 meses, 2 anos e 5 anos. Das variáveis disponíveis o *bureau* de crédito recomendou desconsiderar seis variáveis do banco de dados. Após as exclusões este grupo totalizou 60 variáveis independentes.

Por fim, o quarto grupo de variáveis preditoras se relaciona ao terceiro grupo de variáveis e mede o valor financeiro registrado no *bureau* de crédito relacionado aos eventos descritos. Este grupo totalizou 40 variáveis independentes.

Também se verificou que algumas variáveis possuíam um grande número de observações com valores em branco, por desconhecimento do significado de observações com valores em branco as variáveis foram desconsideradas. Foram excluídas 13 variáveis após esta análise.

Além das variáveis descritas, o *bureau* de crédito também forneceu a identificação se o tomador de empréstimo era “bom” ou “mau” pagador, variável dependente utilizada na calibração do modelo de escore de crédito baseado em dados passados. Porém, não foram informados pelo *bureau* de crédito os critérios que classificaram o tomador de empréstimo como “bom” ou “mau” pagador.

5 DESENVOLVIMENTO DOS MODELOS

Todos os modelos de regressão foram estimados pelo aplicativo R (R Core Team, 2016), utilizando os pacotes *bartMachine* (Kapelner e Bleich, 2013), *randomForest* (Liaw e Wiener, 2002) e *Zelig* (Imai, King e Lau, 2009) para regressão logística com correção de viés.

5.1 PREPARAÇÃO DA BASE

Após uma análise descritiva das variáveis preditoras disponíveis na base de dados, verificou-se que as variáveis de quantidade tinham uma grande concentração de observações sem nenhuma (zero) demanda por crédito ou nenhum (zero) evento registrado do tomador de empréstimo, existindo casos em que algumas variáveis assumiram o valor zero em todas as observações da amostra. O mesmo se concluiu para as variáveis de tempo em dias, onde uma grande quantidade de observações assumiu o valor zero. Por isso, foram utilizadas as técnicas de valor de informação para desconsiderar variáveis do modelo caso seu poder de predição fosse nenhum, fraco ou suspeito e o peso da evidência para recategorização das variáveis, as técnicas estão resumidas no Apêndice 2.

Após a análise pelo valor de informação foram excluídas 141 variáveis do modelo, restando 31 variáveis. Os resultados dos cálculos do valor de informação são apresentados no Apêndice 5.

Após a eliminação das variáveis pelo método do valor de informação, foram executados os cálculos do peso da evidência nas 31 variáveis restantes. Os resultados das recategorizações das variáveis utilizando o método WOE estão demonstrados no Apêndice 6.

5.2 BASE PARA DESENVOLVIMENTO E VALIDAÇÃO DO MODELO

Para a construção dos modelos de score de crédito normalmente é utilizada uma amostra, coletada a partir do histórico de crédito. Conforme citado anteriormente o *bureau* de crédito forneceu para este trabalho uma amostra de dados composta por operações de crédito direto ao consumidor no varejo no ano de 2014, sendo dividida entre clientes “bons” e “maus” pagadores. A Tabela 2 apresenta a proporção de “bons” e “maus” pagadores da base de dados.

Tabela 2 – Base fornecida pelo *bureau* de crédito

Classe Clientes	Quantidade	%
BOM	9.319	90%
MAU	1.037	10%
Total	10.356	100%

Esta amostra foi dividida aleatoriamente em um subconjunto para desenvolvimento ou calibragem do modelo e outro subconjunto para teste da acurácia dos modelos, que é tipicamente conhecida como dados fora da amostra (*out-of-sample*), ou base de validação. Por possuir diferentes quantidades de “bons” e “maus” pagadores esta base de desenvolvimento será mencionada neste trabalho como base de desenvolvimento desbalanceada.

Segundo Abdou e Pointon (2011), alguns estudos de modelos de escore de crédito segregaram a amostra em 50% para treinamento, desenvolvimento ou calibragem do modelo e 50% para validação do modelo (*out-of-sample*). Enquanto outros estudos utilizaram 70% para treinamento ou calibragem e 30% para teste do modelo. Para a base de desenvolvimento foram selecionadas aleatoriamente 70% da amostra original fornecida pelo *bureau* de crédito. A Tabela 3 apresenta a proporção de “bons” e “maus” pagadores da base de desenvolvimento.

Tabela 3 – Base de desenvolvimento desbalanceada

Classe Clientes	Quantidade	%
BOM	6.522	90%
MAU	727	10%
Total	7.249	100%

Para a base de validação foram utilizados os demais 30% da amostra original não selecionada na base de desenvolvimento. A Tabela 4 apresenta a proporção de “bons” e “maus” pagadores da base de validação.

Tabela 4 – Base de validação

Classe Clientes	Quantidade	%
BOM	2.797	90%
MAU	310	10%
Total	3.107	100%

Este trabalho, num primeiro momento, ajusta os modelos com a redução da classe majoritária, neste caso os “bons” pagadores. A Tabela 5 apresenta a proporção de “bons” e “maus” pagadores da base de desenvolvimento com a correção de balanceamento ou base de desenvolvimento balanceada.

Tabela 5 – Base de desenvolvimento balanceada

Classe Clientes	Quantidade	%
BOM	727	50%
MAU	727	50%
Total	1.454	100%

Este trabalho também ajusta os modelos com a base de desenvolvimento desbalanceada, mostrada na Tabela 3. Para a regressão logística é aplicou-se uma correção do viés conforme King e Zeng (2001), mudando o intercepto para classificação dos clientes entre “bons” e “maus” pagadores.

5.3 APLICAÇÃO DOS MODELOS

Nesta seção serão apresentados os modelos desenvolvidos utilizando as técnicas de regressão logística, *Random Forests* e *BART*. A seção 5.3.1 descreve os resultados dos modelos

de regressão logística, a seção 5.3.2 descreve dos modelos *Random Forests*, a seção 5.3.3 descreve os resultados dos modelos BART ajustados com diversas prioris e na seção 5.3.4 compara-se a performance dos modelos. Os modelos foram ajustados e testados na base de desenvolvimento balanceada e na base de desenvolvimento desbalanceada.

Para avaliação dos modelos de escore de crédito desenvolvidos nesta dissertação foram utilizadas as técnicas estatísticas de Kolmogorov-Smirnov e o coeficiente de Gini. Para comparação de performance dos modelos foi utilizado o teste estatístico proposto por DeLong, DeLong e Clarke-Pearson (1988), descrito de forma resumida no Apêndice 3. Além disso, foi utilizada a geração de intervalo de confiança pelo método de *bootstrap*, proposta por Carpenter e Bithell (2000), para comparação das curvas ROC.

5.3.1 REGRESSÃO LOGÍSTICA

Primeiramente, foi ajustado um modelo de regressão logística reduzido na base de desenvolvimento balanceada, onde as variáveis preditoras foram selecionadas através do método *Stepwise Forward*, utilizando as variáveis cujo efeito apresentasse p-valor inferior a 10%. De forma similar foi ajustado um modelo de regressão logística reduzido na base de desenvolvimento desbalanceada, as estimações dos modelos reduzidos estão detalhados no Apêndice 7. A Tabela 6 apresenta a performance do modelo reduzido balanceado e a Tabela 7 a performance do modelo reduzido desbalanceado. A Figura 2 demonstra a curva ROC do modelo reduzido balanceado e do modelo reduzido desbalanceado.

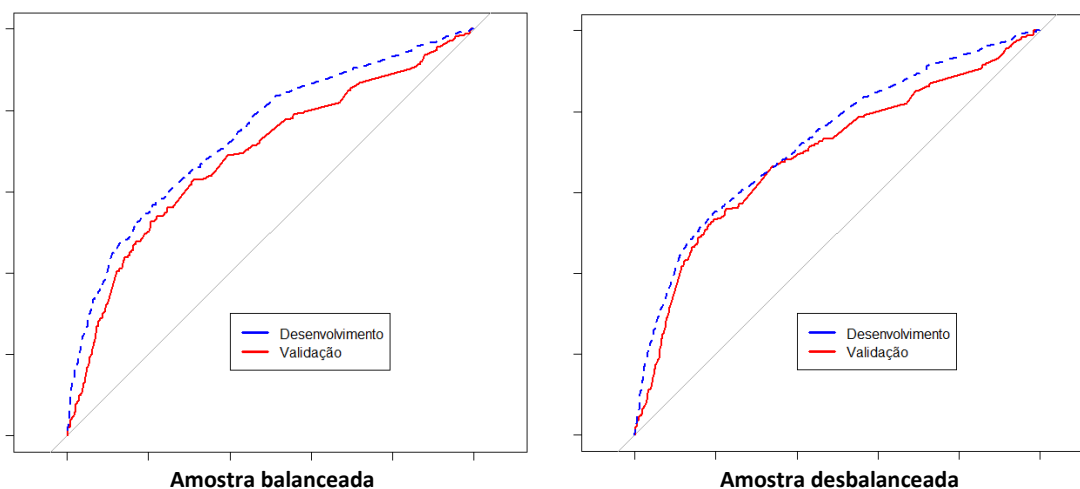
Tabela 6 – Performance do modelo de Regressão Logística reduzido balanceado

Base	KS	AUC	Gini
Desenvolvimento	35,90%	73,96%	47,91%
Validação	31,99%	68,85%	37,69%

Tabela 7 – Performance do modelo de Regressão Logística reduzido desbalanceado

Base	KS	AUC	Gini
Desenvolvimento	35,56%	73,03%	46,07%
Validação	33,84%	69,45%	38,90%

Figura 2 – Curvas ROC – Regressão Logística



Observamos uma melhora de performance do modelo de regressão logística reduzido quando ajustado na base de desenvolvimento desbalanceada com correção de viés. Pela análise da curva ROC, aparentemente o modelo desbalanceado melhorou para classificação nas faixas de escore mais baixas.

5.3.2 *RANDOM FORESTS*

Inicialmente foram ajustados modelos *Random Forests* na base de desenvolvimento balanceada com 500 árvores e 10%, 25%, 50% e 100% como o percentual de variáveis utilizadas a cada sorteio aleatório, eles serão mencionados nesta seção como modelos balanceados. Também foram ajustados modelos na base de desenvolvimento desbalanceada utilizando os mesmos parâmetros, eles serão mencionados neste trabalho como modelos desbalanceados. Ao todo foram gerados oito modelos. Os parâmetros aplicados nos modelos foram baseados no estudo de Chipman, George e McCulloch (2010). Também foi controlada a quantidade mínima de indivíduos que um nó terminal pode conter, ou seja, caso um nó filho apresente uma quantidade de indivíduos menor que o parâmetro, um novo ramo da árvore não pode ser criado. Este parâmetro é importante, pois se não for especificado, o crescimento da árvore pode ocorrer até o extremo em que nós terminais possuam apenas um indivíduo, ocorrendo o fenômeno de *overfitting*. Os modelos foram ajustados com quantidade mínima de indivíduos em nós terminais igual a 30. O pacote *randomForest* do software R apresenta os três parâmetros citados que devem ser fornecidos pelo usuário.

A Tabela 8 mostra o resultado dos modelos *Random Forests* balanceados para os vários percentuais de variáveis utilizadas a cada sorteio aleatório, 500 árvores e quantidade mínima indivíduos em nós terminais igual a 30. A Tabela 9 mostra o resultado dos modelos *Random Forests* desbalanceados para os mesmos parâmetros.

Tabela 8 – Performance dos modelos Random Forests balanceados

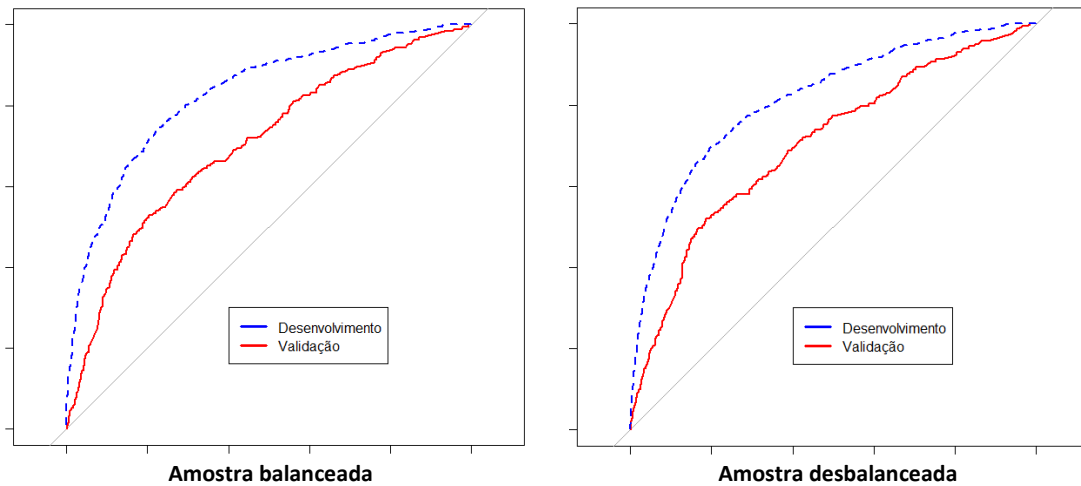
Base	% de variáveis utilizadas no sorteio			
	aleatório	KS	AUC	Gini
Desenvolvimento	10%	46,63%	80,43%	60,86%
	25%	52,27%	82,98%	65,95%
	50%	54,88%	84,16%	68,32%
	100%	56,40%	84,77%	69,54%
Validação	10%	32,14%	69,99%	39,99%
	25%	32,99%	70,35%	40,71%
	50%	32,33%	70,36%	40,71%
	100%	31,94%	69,93%	39,85%

Tabela 9 – Performance dos modelos Random Forests desbalanceados

Base	% de variáveis utilizadas no sorteio			
	aleatório	KS	AUC	Gini
Desenvolvimento	10%	49,90%	81,54%	63,08%
	25%	59,62%	86,45%	72,89%
	50%	60,21%	87,05%	74,10%
	100%	60,82%	87,42%	74,83%
Validação	10%	33,24%	70,31%	40,61%
	25%	32,39%	69,79%	39,58%
	50%	32,07%	69,40%	38,80%
	100%	31,03%	68,81%	37,61%

Os modelos que apresentaram melhor performance na base de desenvolvimento, como seria esperado, foram aqueles com maior quantidade de variáveis, tanto os modelos balanceados quanto os desbalanceados. Enquanto que na base de validação os modelos balanceados que apresentaram melhor performance foram aqueles que utilizaram 25% e 50% das variáveis. Já os modelos desbalanceados que apresentaram melhor performance na base de validação foram aqueles que utilizaram 10% e 25% das variáveis. A Figura 3 mostra as curvas ROC do modelo balanceado – que utilizou 25% das variáveis e apresentou melhor performance para previsão entre “maus” e “bons” pagadores na base de validação – e do modelo desbalanceado – que utilizou 10% das variáveis e apresentou melhor performance para previsão entre “maus” e “bons” pagadores na base de validação.

Figura 3 – Curva ROC – *Random Forests*



Todos os oito modelos testados apresentaram uma boa capacidade de separar “bons” e “maus” pagadores. Também apresentaram performance superior aos da regressão logística para previsão na base de desenvolvimento. Porém, a performance superior também nas previsões na base de validação sugere a superioridade do método *Random Forests* sobre a regressão logística.

5.3.3 BART

Foram ajustados modelos BART completos, ou seja, com todas as variáveis disponíveis após o tratamento do banco de dados, tanto na base de desenvolvimento balanceada quanto na base de desenvolvimento desbalanceada, eles serão mencionados nesta seção como modelos balanceados ou modelos desbalanceados. Primeiramente, foram ajustados modelos utilizando os parâmetros padrão para as prioris sugeridos por Chipman, George e McCulloch (2010), com número de árvores, m , igual a 50 e $z = 2$. Lembrando que para aplicações em classificação, $z = 2$ determina que a probabilidade priori de $E(y^*|x)$ esteja entre um intervalo entre G_{min} e G_{max} é de 95,45%. Além dos modelos padrão, também foram ajustados modelos com árvores de tamanho 200 e foram testados 4 valores para z (1, 2, 3 e 5), que determina que a probabilidade a priori de $E(y^*|x)$ estar entre um intervalo entre G_{min} e G_{max} é de 68,27%, 95,45%, 99,73% e cerca de 100% respectivamente. Estes parâmetros combinados entre si e aplicados nas duas bases geraram dezesseis modelos. A Figura 4 traz as curvas ROC do modelo BART padrão balanceado e a do modelo BART padrão desbalanceado. A Tabela 10 mostra a performance dos modelos balanceados aplicados à base de validação. A Tabela 11 mostra a performance dos modelos desbalanceados aplicados a base de validação.

Figura 4 – Curva ROC – BART padrão

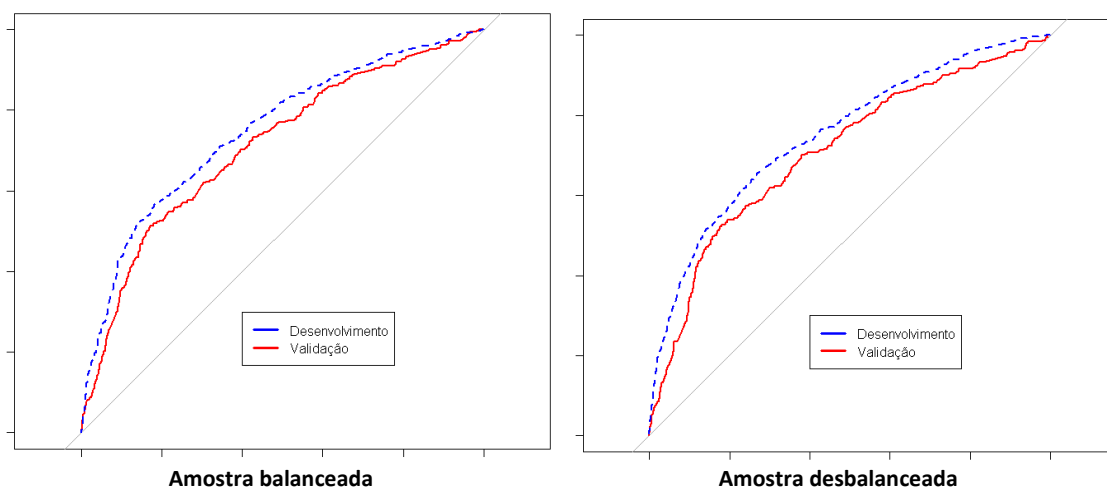


Tabela 10 – Performance dos modelos BART balanceados

Base	(m) Quantidade de árvores	(z) Priori do intervalo de $E(y^* x)$	KS	AUC	Gini
Validação	50	1	32,99%	70,91%	41,81%
	50	2	34,24%	71,02%	42,04%
	50	3	33,81%	70,92%	41,84%
	50	5	30,50%	69,44%	38,89%
	200	1	33,71%	70,88%	41,76%
	200	2	33,96%	70,73%	41,47%
	200	3	33,31%	70,10%	40,21%
	200	5	28,64%	68,34%	36,67%

Tabela 11 – Performance dos modelos BART desbalanceados

Base	(m) Quantidade de árvores	(z) Priori do intervalo de $E(Y x)$	KS	AUC	Gini
Validação	50	1	32,41%	70,68%	41,37%
	50	2	34,70%	70,91%	41,82%
	50	3	34,89%	71,33%	42,67%
	50	5	34,81%	71,60%	43,20%
	200	1	34,45%	70,55%	41,11%
	200	2	34,99%	71,09%	42,17%
	200	3	35,06%	71,25%	42,50%
	200	5	34,63%	71,52%	43,04%

Todos os dezesseis modelos testados apresentaram uma grande capacidade de separar “bons” e “maus” pagadores. Apresentaram performance superior a dos modelos de regressão logística na base de validação, e superior a maioria dos modelos obtidos via *Random Forests*, exceto para $z = 5$, no modelo balanceado, sugerindo uma superioridade do método BART sobre os modelos de regressão logística e *Random Forests*.

Os modelos BART com maior quantidade de árvores e diferentes priori para o intervalo de $E(y^*|x)$ não apresentaram performance superior ao modelo BART padrão com número de árvores igual a 50 e $z = 2$, sugeridos por Chipman, George e McCulloch (2010). Os resultados sugerem que a utilização do número de árvore padrão igual a 50 seria preferida devido ao menor tempo de processamento quando comparado com modelos que utilizaram quantidade de árvores igual a 200.

5.3.4 COMPARAÇÃO DOS MODELOS

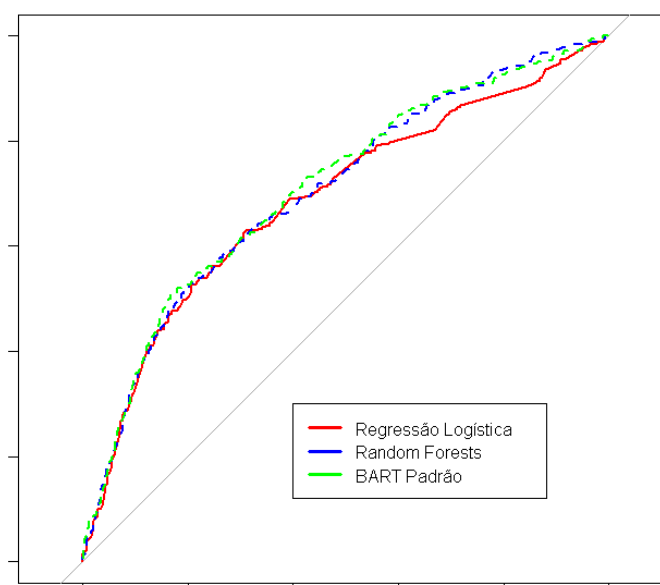
Nesta seção será apresentada a comparação dos modelos. Para esta comparação foi selecionado o melhor modelo ajustado na base balanceada de cada uma das técnicas apresentadas, assim como o melhor modelo ajustado na base desbalanceada de cada uma das técnicas. A seleção dos modelos foi baseada na estatística KS e no coeficiente de Gini calculados na base de validação, os modelos selecionados foram aqueles que apresentaram estatística KS e um coeficiente de Gini superiores.

Primeiramente foram selecionados os melhores modelos ajustados na base balanceada. Para a regressão logística foi utilizado o modelo reduzido. No caso do *Random Forests* foi selecionado o modelo que utilizou 25% das variáveis. Já para o modelo BART foi selecionado o modelo ajustado com a priori padrão. Foram comparados os modelos balanceados aplicados à base de validação, que mostra a capacidade de generalização do modelo uma vez que os clientes desta base não foram utilizados no processo de estimação do modelo. A Tabela 12 mostra a comparação dos modelos balanceados aplicados na base de validação e na Figura 5 temos a curva ROC dos modelos.

Tabela 12 – Comparação da performance dos modelos balanceados

Modelo	KS	AUC	Gini
Regressão Logística	31,99%	68,85%	37,69%
<i>Random Forests</i>	32,99%	70,35%	40,71%
BART Padrão	34,24%	71,02%	42,04%

Figura 5 – Comparação curva ROC dos modelos balanceados



A comparação das curvas ROC sugere que os modelos de aprendizado de máquina BART e *Random Forests* balanceados possuem uma performance similar, mesma capacidade de classificar “bons” e “maus” pagadores, ao modelo de regressão logística balanceado para faixas de escore mais baixas. Ainda, a comparação sugere que os modelos BART e *Random Forests* possuem uma performance superior de classificação para faixas de escore mais altas.

Para confirmação dos resultados da comparação dos modelos balanceados foram realizados testes de hipótese para verificar se as áreas sob a curva (AUC) dos métodos são iguais, para isso foram utilizadas as técnicas de Delong e *bootstrap*. A Tabela 13 mostra os resultados dos testes de hipótese.

Tabela 13 – Testes de hipótese dos modelos balanceados

Método	Teste de Hipótese	z	p-Valor
Delong	H ₀ : AUC Regressão Logística = AUC <i>Random Forests</i>	-1,9579	0,05025
	H ₀ : AUC BART Padrão = AUC <i>Random Forests</i>	-1,2463	0,21270
	H ₀ : AUC BART Padrão = AUC Regressão Logística	-3,3224	0,00089
Bootstrap	H ₀ : AUC Regressão Logística = AUC <i>Random Forests</i>	-1,9337	0,05315
	H ₀ : AUC BART Padrão = AUC <i>Random Forests</i>	-1,2462	0,21270
	H ₀ : AUC BART Padrão = AUC Regressão Logística	-3,3322	0,00086

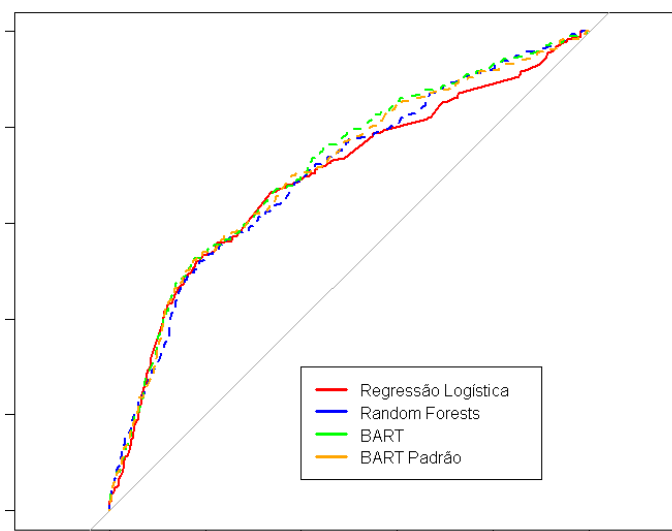
Com os resultados dos testes de hipótese obtidos para os modelos balanceados, podemos dizer que o modelo BART foi superior ao modelo de regressão logística ($p < 0,01$). Não há evidências para rejeitar a hipótese de que o modelo BART tem a mesma performance do modelo *Random Forests* (método *Delong* com $p = 0,21270$ e *bootstrap* com $p = 0,21270$). A comparação entre os modelos *Random Forests* e a regressão logística ficou no limite da significância de 5%, sugerindo uma superioridade do modelo BART.

De forma similar foram selecionados os melhores modelos ajustados na base desbalanceada. Para a regressão logística foi utilizado o modelo reduzido com correção de viés. No caso do *Random Forests* foi selecionado o modelo que utilizou 10% das variáveis. Já para o modelo BART foi selecionado que o modelo com 50 árvores e $z = 5$, pois possui a maior área sob a curva. Para fins de comparação, também foi selecionado o modelo BART ajustado com a priori padrão. A Tabela 14 mostra a comparação dos modelos desbalanceados aplicados a base de validação e na Figura 6 temos a curva ROC dos modelos.

Tabela 14 – Comparação da performance dos modelos desbalanceados

Modelo	KS	AUC	Gini
Regressão Logística	33,84%	69,45%	38,90%
<i>Random Forests</i>	33,24%	70,31%	40,61%
BART	34,81%	71,60%	43,20%
BART Padrão	34,70%	70,91%	41,82%

Figura 6 – Comparação curva ROC dos modelos desbalanceados



Apesar da melhora de performance observada no modelo de regressão logística desbalanceado, a comparação das curvas ROC ainda sugere que sua performance é inferior aos modelos BART e *Random Forests* para classificação de “bons” e “maus” pagadores nas faixas de escore mais altas.

Para confirmação dos resultados da comparação dos modelos desbalanceados também foram realizados testes de hipótese para verificar se as áreas sob a curva (AUC) dos métodos são iguais. A Tabela 15 mostra os resultados dos testes de hipótese.

Tabela 15 – Testes de hipótese dos modelos desbalanceados

Método	Teste de Hipótese	z	p-Valor
DeLong	H ₀ : AUC Regressão Logística = AUC <i>Random Forests</i>	-0,9222	0,35640
	H ₀ : AUC BART = AUC <i>Random Forests</i>	-2,0282	0,04254
	H ₀ : AUC BART = AUC Regressão Logística	-2,8691	0,00412
	H ₀ : AUC BART Padrão = AUC <i>Random Forests</i>	-0,8845	0,37640
	H ₀ : AUC BART Padrão = AUC Regressão Logística	-1,9769	0,04806
Bootstrap	H ₀ : AUC Regressão Logística = AUC <i>Random Forests</i>	-0,9667	0,33370
	H ₀ : AUC BART = AUC <i>Random Forests</i>	-2,0040	0,04507
	H ₀ : AUC BART = AUC Regressão Logística	-2,7881	0,00530
	H ₀ : AUC BART Padrão = AUC <i>Random Forests</i>	-0,9067	0,36460
	H ₀ : AUC BART Padrão = AUC Regressão Logística	-1,9581	0,05022

Com os resultados dos testes de hipótese obtidos para os modelos desbalanceados, podemos dizer que o melhor modelo BART foi superior aos modelos de regressão logística ($p < 0,01$) e *Random Forests* ($p < 0,05$). O modelo BART padrão também foi superior ao modelo de regressão logística ($p < 0,05$), porém com 5% de significância não podemos rejeitar a hipótese que o modelo BART padrão tem mesma performance do modelo *Random Forests*. Além disso, com 5% de significância não podemos rejeitar a hipótese que os modelos de regressão logística e *Random Forests* têm a mesma performance.

Resultado semelhante foi observado por Zhang e Härdle (2010), onde o modelo BART aplicado ao escore de crédito de uma base de dados de empresas Alemãs apresentou resultado superior à regressão logística.

6 CONCLUSÕES

Esta dissertação avaliou empiricamente o desempenho de dois modelos de aprendizado de máquina, o BART e o *Random Forests*, aplicados ao escore de crédito para previsão de “bom” ou “mau” pagador. A performance dos modelos foi comparada a da regressão logística, pois atualmente é o modelo mais utilizado no mercado para o escore de crédito.

Para o desenvolvimento dos modelos foi utilizada a base de dados fornecida pela Serasa Experian (*bureau* de crédito Brasileiro), sendo formada por clientes de operações de crédito direto ao consumidor no varejo. Diferentemente das bases de dados normalmente utilizadas para o desenvolvimento do escore de crédito, que utilizam variáveis como as características dos indivíduos, a base de dados utilizada nesta dissertação traz variáveis com a visão de um *bureau* de crédito. Esta base de dados foi dividida em duas partes, sendo uma parte para o desenvolvimento do modelo e a outra utilizada para validação do modelo desenvolvido (chamada *out-of-sample* ou base de validação).

Os resultados da AUC ou coeficiente de Gini sugerem que os modelos de aprendizado de máquina BART e *Random Forests* foram superiores a regressão logística tanto na amostra balanceada quanto na amostra desbalanceada. Já os resultados da estatística KS sugerem que o modelo BART foi superior ao modelo de regressão logística tanto na amostra balanceada quanto na amostra desbalanceada, sendo que o modelo *Random Forests* foi superior a regressão logística somente na amostra balanceada. Além disso, foram realizados testes de hipótese para comparação da métrica AUC dos modelos. Os resultados da comparação sugerem que o melhor modelo BART é superior aos modelos *Random Forests* e regressão logística tanto na amostra balanceada quanto na amostra desbalanceada. O modelo BART padrão foi superior ao modelo de regressão logística, porém sugere uma performance similar ao modelo *Random Forests*. Já os resultados da comparação do modelo *Random Forests* sugerem que ele é superior ao modelo de regressão logística somente na amostra balanceada.

Os resultados empíricos sugerem que os modelos de aprendizado de máquina BART e *Random Forests* apresentam uma boa capacidade de separar “bons” e “maus” pagadores, tanto na amostra balanceada quanto na amostra desbalanceada. As comparações das curvas ROC dos modelos também sugerem que os modelos BART e *Random Forests* possuem melhor capacidade de separar “bons” e “maus” pagadores para faixas de escore mais altas.

Os resultados encontrados confirmam a superioridade do modelo BART em relação ao modelo de regressão logística. Apesar das diferenças de desempenho dos modelos BART e *Random Forest* não serem muito maiores que o modelo de regressão logística e considerando sua aplicação em classificação de “bons” ou “maus” pagadores de um empréstimo, a melhor performance dos modelos de aprendizado de máquina pode representar ganhos significativos para as instituições financeiras pela redução da perda de conceder um empréstimo para um “mau” pagador ou pelo aumento receita de não negar um empréstimo a um “bom” pagador.

Como futuras pesquisas, poder-se-ia investigar outras técnicas de aprendizado de máquina, como Redes Neurais, *Gradient Boosting* ou LASSO, que também tem se mostrado promissoras para aplicação em problemas de classificação, como podemos observar em Chipman, George e McCulloch (2010) e Zhang and Härdle (2010). Uma outra extensão deste

trabalho seria aumentar a classe minoritária sinteticamente (Chawla, Bowyer, Hall e Kegelmeyer, 2002) para criar amostras balanceadas. Mais estudos são necessários para avaliar a performance do modelo de regressão logística com correção de viés (King e Zeng, 2001) em outras aplicação de escore de crédito.

Outras pesquisas importantes seriam a aplicação do uso de sistemas híbridos (Hsieh, 2005), novos conceitos de adaptação às mudanças (Pavlidis, 2012) e modelagem dinâmica (Crook e Bellotti, 2010) a uma base de dados de um *bureau* de crédito.

7 REFERÊNCIAS BIBLIOGRÁFICAS

ABDOU, Hussein A.; POINTON, John. Credit scoring, statistical techniques and evaluation criteria: A review of the literature. **Intelligent Systems in Accounting, Finance and Management**, v. 18, n. 2-3, p. 59-88, 2011.

ALTMAN, Edward I. Financial ratios, discriminant analysis and the prediction of corporate bankrupt. **The Journal of Finance**, v. 23, p. 589-609, 1968.

ANDERSON, Raymond. **The credit scoring toolkit: Theory and practice for retail credit risk management and decision automation**. Oxford University Press, 2007.

ANDRIEU, Christophe; FREITAS, Nando de; DOUCET, Arnaud; JORDAN, Michel I. An Introduction to MCMC for Machine Learning. **Machine Learning**, 50, 5-43, 2003.

BCBS (2006). International convergence of capital measurement and capital standards: A revised framework - comprehensive version. **Bank for International Settlements**.

BIS (2004). Implementation of Basel II: Practical considerations. **Bank for International Settlements**.

BLEICH, Justin, et al. Variable selection for BART: an application to gene regulation. **The Annals of Applied Statistics**, v. 8, n. 3, p. 1750-1781, 2014.

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

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

BREIMAN, Leo; FRIEDMAN, J.; OLSHEN, R.; STONE, C. **Classification and regression trees**. CRC press, 1984.

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

CARPENTER, James; BITHELL, John. Bootstrap confidence intervals: when, which, what? **A practical guide for medical statisticians**. 2000.

CHANDLER, Gary G.; COFFMAN, John Y. A comparative analysis of empirical vs. judgmental credit evaluation. **Financial Review**, v. 14, n. 4, p. 23-23, 1979.

CHAWLA, N. V.; BOWYER, K. W.; HALL, L. O.; KEGELMEYER, W. P. SMOTE: synthetic minority over-sampling technique. **Journal of artificial intelligence research**, v. 16, p. 321-357, 2002.

CHIPMAN, Hugh A.; GEORGE, Edward I.; MCCULLOCH, Robert E. Bayesian CART model search. **Journal of the American Statistical Association**, v. 93, n. 443, p. 935-948, 1998.

CHIPMAN, Hugh A.; GEORGE, Edward I. e MCCULLOCH, Robert E. BART: Bayesian Additive and Regression Trees. **The Annals of Applied Statistics**, 2010, Vol. 4, No. 1, 266-298.

CROOK, Jonathan; BELLOTTI, Tony. Time varying and dynamic models for default risk in consumer loans. **Journal of the Royal Statistical Society: Series A (Statistics in Society)**, v. 173, n. 2, p. 283-305, 2010.

DELONG, Elizabeth R.; DELONG, David M.; CLARKE-PEARSON, Daniel L. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. **Biometrics**, p. 837-845, 1988.

DESAI, Vijay S.; CROOK, Jonathan N.; OVERSTREET, George A. A comparison of neural networks and linear scoring models in the credit union environment. **European Journal of Operational Research**, v. 95, n. 1, p. 24-37, 1996.

DURAND, David, et al. Risk elements in consumer instalment financing. **NBER Books**, 1941.

EFRON, Bradley, et al. Least angle regression. **The Annals of statistics**, v. 32, n.2, p. 407-499, 2004.

FISHER, Ronald A. The use of multiple measurements in taxonomic problems. **Annals of eugenics**, v. 7, n. 2, p. 179-188, 1936.

FRIEDMAN, Jerome H. Greedy function approximation: a gradient boosting machine. **Annals of statistics**, p. 1189-1232, 2001.

GESTEL, Tony Van, et al. Bayesian kernel based classification for financial distress detection. **European journal of operational research**, v. 172, n. 3, p. 979-1003, 2006.

HANLEY, James A.; HAJIAN-TILAKI, Karim O. Sampling variability of nonparametric estimates of the areas under receiver operating characteristic curves: an update. **Academic radiology**, n. 4, v. 1, p. 49-58, 1997.

HANLEY, James A.; MCNEIL, Barbara J. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. **Radiology**, n. 148, v. 3, p. 839-843, 1983.

HASTINGS, W. Keith. Monte Carlo sampling methods using Markov chains and their applications. **Biometrika**, v. 57, n. 1, p. 97-109, 1970.

HOSMER JR, David W.; LEMESHOW, Stanley; STURDIVANT, Rodney X. **Applied logistic regression**. John Wiley & Sons, 2013.

HSIEH, Nan-Chen. Hybrid mining approach in the design of credit scoring models. **Expert Systems with Applications**, v. 28, n. 4, p. 655-665, 2005.

IMAI, Kosuke; KING, Gary; LAU, Olivia. Zelig: Everyone's statistical software. **R package version**, v. 3, n. 5, 2009.

JAIN, Anil K.; DUIN, Robert P. W.; MAO, Jianchang. Statistical pattern recognition: A review. **IEEE Transactions on pattern analysis and machine intelligence**, v. 22, n. 1, p. 4-37, 2000.

KAPELNER, Adam; BLEICH, Justin. bartmachine: Machine learning with bayesian additive regression trees. **arXiv preprint arXiv**, p. 1312-2171, 2013.

KING, Gary; ZENG, Langche. Logistic regression in rare events data. **Political analysis**, v. 9, n. 2: p. 137-163, 2001.

KRAUS, Anne. **Recent Methods from Statistics and Machine Learning for Credit Scoring**. 2014. 144 f. Tese (Doutorado) – Faculdade de Matemática, Informática e estatística, Universidade München, Alemanha

KRUPPA, Jochen, et al. Consumer credit risk: Individual probability estimates using machine learning. **Expert Systems with Applications**, v. 40, n. 13, p. 5125-5131, 2013.

LEE, Tian-Shyug; CHEN, I. F. A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines. **Expert Systems with Applications**, v. 28, n. 4, p. 743-752, 2005.

LENSBERG, Terje; EILIFSEN, Aasmund; MCKEE, Thomas E. Bankruptcy theory development and classification via genetic programming. **European Journal of Operational Research**, v. 169, n. 2, p. 677-697, 2006.

LESSMANN, Stefan, et al. Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. **European Journal of Operational Research**, v. 247, n. 1, p. 124-136, 2015.

LI, Sheng-Tun; SHIUE, Weissor; HUANG, Meng-Huah. The evaluation of consumer loans using support vector machines. **Expert Systems with Applications**, v. 30, n. 4, p. 772-782, 2006.

LIAW, Andy; WIENER, Matthew. Classification and regression by randomForest. **R news**, n. 2, v.3, p. 18-22, 2002.

MALHOTRA, Rashmi; MALHOTRA, Davinder K. Differentiating between good credits and bad credits using neuro-fuzzy systems. **European journal of operational research**, v. 136, n. 1, p. 190-211, 2002.

MALEKIPIRBAZARI, Milad; AKSAKALLI, Vural. Risk assessment in social lending via random forests. **Expert Systems with Applications**, v. 42, n. 10, p. 4621-4631, 2015.

PAVLIDIS, Nicos G., et al. Adaptive consumer credit classification. **Journal of the Operational Research Society**, v. 63, n. 12, p. 1645-1654, 2012.

PESKUN, Peter H. Optimum monte-carlo sampling using markov chains. **Biometrika**, v. 60, n. 3, p. 607-612, 1973.

R Core Team (2016). **R: A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria. URL: <https://www.R-project.org/>.

SIDDIQI, Naeem. **Credit risk scorecards: developing and implementing intelligent credit scoring**. John Wiley & Sons, 2012.

SOUSA, Maria Rocha; GAMA, João; BRANDÃO, Elísio. A new dynamic modeling framework for credit risk assessment. **Expert Systems With Applications**, v. 45, p. 341-351, 2016.

STEENACKERS, Ann; GOOVAERTS, Marc J. A credit scoring model for personal loans. **Insurance: Mathematics and Economics**, v. 8, n. 1, p. 31-34, 1989.

THOMAS, Lyn C. **Consumer Credit Models: Pricing, Profit and Portfolios: Pricing, Profit and Portfolios**. OUP Oxford, 2009.

THOMAS, Lyn C.; EDELMAN, David B.; CROOK, Jonathan N. **Credit scoring and its applications**. Siam, 2002.

THOPPAY, Sudarson Mothilal. Computes Weight of Evidence and Information Values. **Changes on CRAN - Porting R to Darwin/X11 and Mac OS X**, v. 38, 2013.

WEI, Gao; YUN-ZHONG, Cao; MING-SHU, Cheng. **A new dynamic credit scoring model based on the objective cluster analysis**. In: Practical Applications of Intelligent Systems. Springer Berlin Heidelberg, p. 579-589, 2014.

WEST, David; DELLANA, Scott; QIAN, Jingxia. Neural network ensemble strategies for financial decision applications. **Computers & Operations research**, v. 32, n. 10, p. 2543-2559, 2005.

XIA, Yusen, et al. A model for portfolio selection with order of expected returns. **Computers & Operations Research**, v. 27, n. 5, p. 409-422, 2000.

YAP, Bee Wah, et al. **An application of oversampling, undersampling, bagging and boosting in handling imbalanced datasets**. In: Proceedings of the First International Conference on Advanced Data and Information Engineering (DaEng-2013). Springer Singapore, p. 13-22, 2014.

ZEKIC-SUSAC, Marijana; SARLIJA, Natasa; BENSIC, Mirta. **Small business credit scoring: a comparison of logistic regression, neural network, and decision tree models**. In: Information Technology Interfaces. 26th International Conference on IEEE, p. 265-270, 2004.

ZHANG, Junni L.; and HÄRDLE, Wolfgang K. The Bayesian Additive Classification Tree applied to credit risk modelling. **Computational Statistics and Data Analysis**, v. 54, n. 5, p. 1197-1205, 2010.

ZHOU, Lifeng; WANG, Hong. Loan default prediction on large imbalanced data using random forests. **TELKOMNIKA Indonesian Journal of Electrical Engineering**, v. 10, n. 6, p. 1519-1525, 2012

8 APÊNDICES

8.1 APÊNDICE 1 – GERAÇÃO DE AMOSTRAS POSTERIORI E INFERÊNCIA

Para computação do problema Bayesiano formulado, técnicas de MCMC (do inglês *Monte Carlo Markov Chain*) são frequentemente utilizadas (Andrieu, Freitas, Doucet e Jordan, 2003).

A partir da base de dados de desenvolvimento \mathcal{D} , a configuração Bayesiana formulada no modelo de árvores nos induz a uma distribuição posteriori para todos os parâmetros desconhecidos:

$$p((T_1, M_1), \dots, (T_m, M_m), y^* | \mathcal{D}), \quad (14)$$

onde:

$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ é a base de dados de desenvolvimento com N observações.

$G = \{G(x_i)\}_{i=1}^N$ é a estimativa da probabilidade de um cliente ser um “mau” pagador, tratada no problema como uma variável desconhecida.

m é o número de árvores de decisão que serão utilizadas no modelo.

T_k é a árvore de decisão k com um conjunto de nós interiores de decisão e um conjunto de nós terminais.

L_k é o número de nós terminais da árvore T_k , sendo os nós terminais $l = 1, \dots, L_k$.

$M_k = \{\mu_{1k}, \mu_{2k}, \dots, \mu_{L_k k}\}$ são as predições associadas à cada nó terminal l da árvore T_k .

Vamos iniciar pela geração da distribuição posteriori de (T_k, M_k) . Definimos $T_{(k)}$ como sendo o conjunto de tamanho $m - 1$ de árvores excluindo a árvore T_k . Conforme mostrado em Chipman, George e McCulloch (2010), o algoritmo MCMC é utilizado para geração de \mathcal{K} sucessivas amostras de (T_j, M_j) da distribuição posteriori definida por:

$$p\{(T_j, M_j) | T_{(j)}, M_{(j)}, y^*, \mathcal{D}\} \quad (15)$$

sendo $j = 1, \dots, m$. Detalhes sobre essa distribuição podem ser encontrados em Chipman, George e McCulloch (2010, pg 274).

Em seguida são gerados valores de y^* da distribuição:

$$p(y^* | (T_1, M_1), \dots, (T_m, M_m), \mathcal{D}). \quad (16)$$

O algoritmo MCMC gera uma sequência de amostras para $(T_1, M_1), \dots, (T_m, M_m), y^*$, no qual converge em probabilidade para $p((T_1, M_1), \dots, (T_m, M_m), y^* | \mathcal{D})$. Sendo que, $G(x_i) = g(x_i; T_1; M_1) + \dots + g(x_i; T_m; M_m)$ representa o valor ajustado do preditor x_i calculados pelas m árvores. Podemos então gerar amostras de y_i^* ($i = 1, \dots, N$) de forma independente seguindo a distribuição normal:

$$\begin{cases} y_i^* \sim N(G(x_i), 1) e y_i^* \geq 0, \text{ se } y_i = 1 \\ y_i^* \sim N(G(x_i), 1) e y_i^* < 0, \text{ se } y_i = 0 \end{cases} \quad (17)$$

Iniciamos a cadeia de cálculo com m árvores de um único nó e então repetimos as \mathcal{K} iterações até que uma convergência satisfatória é obtida. A cada iteração, os nós terminais de cada árvore poderão aumentar ou diminuir. Cada parâmetro μ_{il} associado aos nós terminais também poderão mudar e por consequência o valor estimado de y^* . O modelo de soma de árvore, com seu grande número de parâmetros, permite que o ajuste do modelo aos dados \mathcal{D} aconteça de forma livre de uma árvore para outra, por isso o modelo BART possui uma flexibilidade para problemas complexos (Chipman, George e McCulloch, 2010).

Ignoramos as primeiras \mathcal{B} amostras posteriori, utilizadas para a convergência do modelo e utilizamos as S amostras subsequentes ($\mathcal{K} - \mathcal{B}$) da distribuição posteriori para fazer inferência. Sendo as amostras S representada por $\{(T_1^{(s)}, M_1^{(s)}), \dots, (T_m^{(s)}, M_m^{(s)})\}_{s=1}^S$, a probabilidade de um cliente ser um “mau” pagador será:

$$\Phi \left\{ g \left(x, T_1^{(s)}, M_1^{(s)} \right) + \dots + g \left(x, T_m^{(s)}, M_m^{(s)} \right) \right\} \quad (18)$$

onde Φ é a função da distribuição normal acumulada.

Para a estimativa de $P(y = 1)$ para um x qualquer, a escolha natural é a média das amostras posteriori S , dado por :

$$\frac{1}{S} * \sum_{s=1}^S \Phi \left\{ g \left(x, T_1^{(s)}, M_1^{(s)} \right) + \dots + g \left(x, T_m^{(s)}, M_m^{(s)} \right) \right\} \quad (19)$$

Finalmente, utilizamos o resultado da equação para classificar um indivíduo particular. Se a probabilidade média for maior que 0,5, então o cliente será classificado como “mau” pagador, senão será classificado como um “bom” pagador.

Figura 7 – Convergência do MCMC do modelo BART padrão balanceado

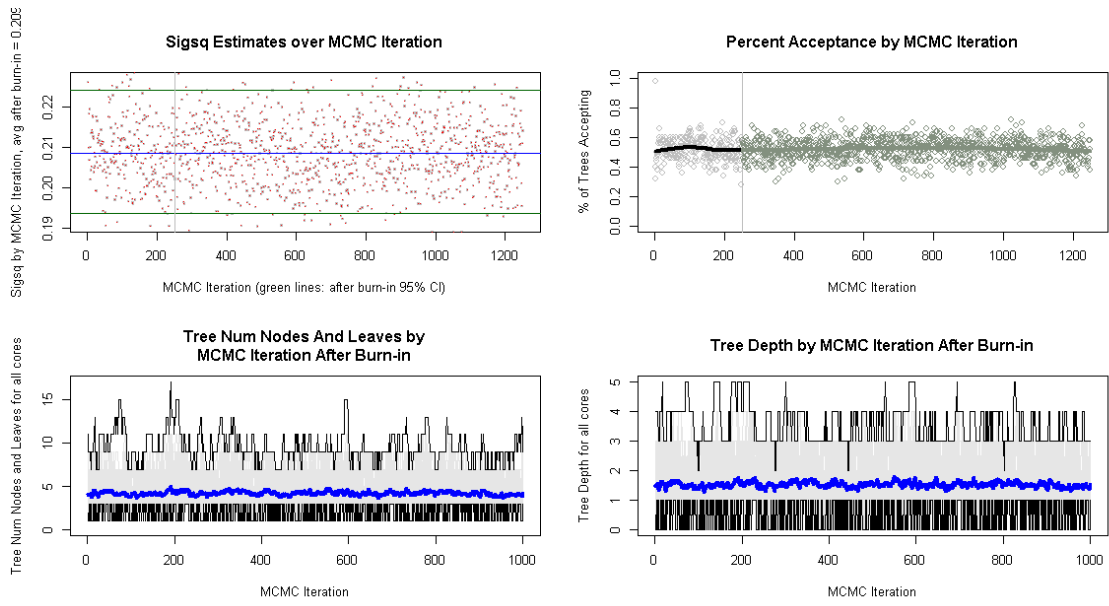
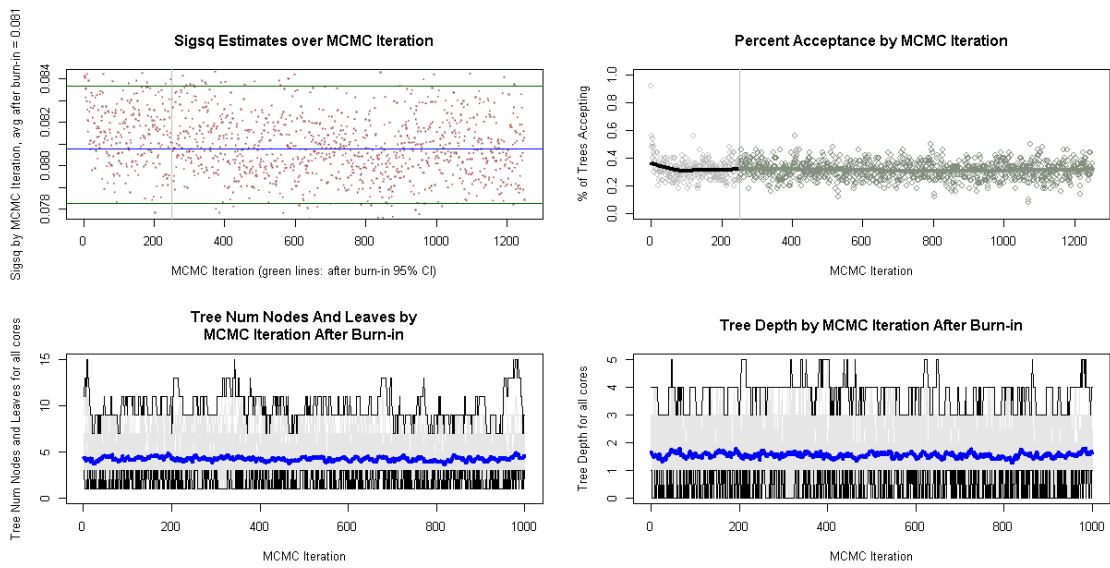


Figura 8 – Convergência do MCMC do modelo BART padrão desbalanceado



8.2 APÊNDICE 2 – PESO DA EVIDÊNCIA E VALOR DA INFORMAÇÃO

Siddiqi (2012) recomenda que a análise das características iniciais de uma variável envolve primeiramente analisar seu poder de predição e em seguida agrupar seus valores em classes, esta é uma forma para lidar com *outliers* e classes raras, gerando assim uma nova variável categorizada e também com a não linearidade, no caso de variáveis quantitativas.

A razão de proporções entre “bons” e “maus” pagadores na classe de valores i da variável independente ou preditor, $\frac{DistribBom_i}{DistribMau_i}$, sendo $DistribBom$ a proporção de “bons” pagadores na classe de valores i da variável independente e $DistribMau$ a proporção de “maus” pagadores na classe de valores i da variável independente, é chamada de chance de informação (do inglês *Information Odds*). Chances de informação sejam maiores que um sugerem uma maior probabilidade dos clientes serem “bons” pagadores na classe de valores i da variável independente ou preditor. Chance de informação menor que um sugere que existe uma maior probabilidade dos clientes serem “maus” pagadores. Aplicar a função logarítmica sobre a chance de informação é uma forma útil de analisar a probabilidade de um “bom” pagador na classe de valores i da variável independente ou preditor, métrica conhecida como peso da evidência (do inglês *Weight of Evidence* ou WOE). O peso da evidência é medido pela seguinte expressão:

$$WOE_i = \ln\left(\frac{DistribBom_i}{DistribMau_i}\right) \quad (20)$$

Para o agrupamento dos valores de uma variável em classes podemos utilizar a técnica WOE. Quanto maior a diferença entre WOE nas classes de valores da variável independente maior o poder de predição do grupo. O objetivo é identificar grupos de valores com WOE similar, para isso é realizado um teste de hipóteses do tipo $H_0: WOE_i = WOE_{i+1}$, calculado pela estatística qui-quadrado. Não é o objetivo descrever o cálculo da estatística qui-quadrado, porém utilizar o teste de hipótese para confirmar se uma nova classe de valores deverá ser criada ou não.

Por outro lado, para determinar as variáveis mais importantes foi utilizada a técnica estatística chamada de valor de informação (do inglês *Information Value* ou IV). O valor de informação vem da teoria da informação e consiste em medir o poder preditivo de uma variável independente. Se estivermos interessados em medir o quão bem um preditor está separando os “bons” e “maus” pagadores, seria razoável pensar na diferença das proporções de “bons” e “maus” pagadores ponderada pelo peso da informação. Assim o valor de informação é medido pela seguinte expressão:

$$IV = \sum_{i=1}^n (DistribBom_i - DistribMau_i) * \ln\left(\frac{DistribBom_i}{DistribMau_i}\right) \quad (21)$$

O valor de informação pode também ser expresso como:

$$IV = \sum_{i=1}^n (DistribBom_i - DistribMau_i) * WOE_i \tag{22}$$

Variáveis com valores de informação próximos de zero possuem baixo poder de predição. O poder de predição das variáveis vai aumentando conforme seus valores de informação vão aumentando. Siddiqi (2012) sugere uma classificação para o poder de predição de uma variável entre nenhum, fraco, médio ou forte, como mostrado na Tabela 16. Sendo variáveis com valor de informação acima de 0,5 consideradas como suspeitas ou com “excesso de predição”, devendo estas ser excluídas do modelo ou utilizadas com precaução. Esta é uma sugestão da autora baseado em práticas da indústria, diferentes praticantes podem optar por utilizar outra regra para classificação do poder de predição das variáveis baseado no valor de informação.

Tabela 16 – Poder de Predição

Valor de Informação	Poder de Predição
< 0,02	Nenhum
de 0,02 até 0,1	Fraco
de 0,1 até 0,3	Médio
de 0,3 até 0,5	Forte
> 0,5	Excesso de predição

Fonte: Siddiqi (2012)

O pacote *woe* (Thoppay, 2013) do software R implementa os cálculos do WOE e IV, assim como sugere um poder de predição para cada uma das variáveis.

A Tabela 17 mostra um exemplo de cálculo do peso da evidência e valor de informação aplicado à variável quantidade de demanda por crédito em até 180 dias. Podemos observar que não há uma grande demanda por crédito em até 180 dias, com uma proporção total de 53,29% de “bons” e 35,68% de “maus” pagadores, o WOE_0 foi calculado para as proporções e obtivemos $WOE_0 = 0,4011$. Já uma única demanda por crédito em 180 dias possui uma proporção de 24,66% de “bons” e 22,95% de “maus” pagadores e o peso da evidência de $WOE_1 = 0,0718$. Como mencionado, podemos realizar um teste de hipótese para verificar se o WOE_0 da classe de valores sem nenhuma demanda por crédito é igual ao WOE_1 da classe de valores com uma única demanda por crédito, por limitação do pacote *woe* os resultados do p-Valor não são apresentados na Tabela 17. Devemos realizar o teste de hipóteses sucessivamente aos pares para todos os valores i da variável independente ou preditor, ao final do processo definir os agrupamentos dos valores que serão utilizados para criação de categorias. O resultado foi que para este preditor foram criadas quatro categorias de valores: nenhuma demanda por crédito, com apenas uma demanda por crédito, com duas demandas e acima de 3 demandas.

Tabela 17 – Exemplo de Cálculo do Peso da Evidência e Valor de Informação

Domínio da Variável "SQtCons180d"	Qtde "bom" pagador	Qtde "mau" pagador	Proporção "bom" pagador	Proporção "mau" pagador	WOE	IV
0	4.966	370	53,29%	35,68%	0,4011	0,0706
1	2.298	238	24,66%	22,95%	0,0718	0,0012
2	968	147	10,39%	14,18%	-0,3109	0,0118
3	468	99	5,02%	9,55%	-0,6424	0,0291
4	251	58	2,69%	5,59%	-0,7307	0,0212
5	133	34	1,43%	3,28%	-0,8317	0,0154
6	98	26	1,05%	2,51%	-0,8689	0,0126
7	52	15	0,56%	1,45%	-0,9525	0,0085
8	27	9	0,29%	0,87%	-1,0971	0,0063
9	17	12	0,18%	1,16%	-1,8474	0,0180
10	15	6	0,16%	0,58%	-1,2794	0,0053
11	7	7	0,08%	0,68%	-2,1957	0,0132
12	6	4	0,06%	0,39%	-1,7903	0,0058
13	5	2	0,05%	0,19%	-1,2794	0,0018
14	2	2	0,02%	0,19%	-2,1957	0,0038
15	1	1	0,01%	0,10%	-2,1957	0,0019
16	1	1	0,01%	0,10%	-2,1957	0,0019
17	1	0	0,01%	0,00%	N/A	N/A
18	0	3	0,00%	0,29%	N/A	N/A
19	0	1	0,00%	0,10%	N/A	N/A
21	0	1	0,00%	0,10%	N/A	N/A
22	1	0	0,01%	0,00%	N/A	N/A
23	0	1	0,00%	0,10%	N/A	N/A
27	1	0	0,01%	0,00%	N/A	N/A
38	1	0	0,01%	0,00%	N/A	N/A

8.3 APÊNDICE 3 – MÉTODOS DE COMPARAÇÃO DE ÁREAS SOB A CURVA

A comparação de duas curvas ROC produzidas a partir de dois modelos de escore de crédito diferentes utilizando a mesma base de clientes requer um teste estatístico. Hanley e McNeil (1983) propuseram um método paramétrico baseado nas métricas AUC das duas curvas, assumindo que a métrica AUC segue uma distribuição normal e também supostamente existe uma correlação positiva entre as métricas AUC das duas curvas ROC, que melhoraria a sensibilidade estatística da comparação entre as curvas AUC. O escore z da distribuição normal padronizada é calculado pela equação:

$$z = \frac{AUC_1 - AUC_2}{\sigma_{AUC_1}^2 + \sigma_{AUC_2}^2 - 2r\sigma_{AUC_1}\sigma_{AUC_2}}, \quad (23)$$

onde:

AUC_1 é a área sob a curva ROC 1

AUC_2 é a área sob a curva ROC 2

σ_{AUC_1} é o desvio padrão da curva ROC 1

σ_{AUC_2} é o desvio padrão da curva ROC 2

r é a correlação entre as curva ROC 1 e curva ROC 2

De acordo com Hanley e Hajian-Tilaki (1997), pelo menos quatro métodos foram propostos para calcular a variância da estimativa métrica AUC e dois métodos são extensíveis para calcular a covariância das estimativas da métrica AUC. Nesta dissertação foi utilizado o método não paramétrico DeLong (DeLong, DeLong e Clarke-Pearson, 1988) para estimar o desvio-padrão e a covariância das métricas AUC de duas curvas ROC, para depois ser utilizado para o cálculo do escore z da distribuição normal.

8.4 APÊNDICE 4 – DICIONÁRIO DE DADOS

Tabela 18 – Dicionário de Dados

#	Variável	Tipo	Descrição
1	dt_contrato	Num	Data de concessão
2	PERF60_M12	Num	Variável resposta
3	UF	Char	UF
4	SQtCons30d	Num	Demanda por crédito nos últimos 30 dias
5	SQtCons60d	Num	Quantidade Consulta 60 dias
6	SQtCons90d	Num	Demanda por crédito nos últimos 90 dias
7	SQtCons180d	Num	Demanda por crédito nos últimos 180 dias
8	SQtCons360d	Num	Demanda por crédito nos últimos 360 dias
9	SQtConsOrigBCO30d	Num	Demanda por crédito em bancos 30 dias
10	SQtConsOrigBCO60d	Num	Demanda por crédito em bancos 60 dias
11	SQtConsOrigBCO90d	Num	Demanda por crédito em bancos 90 dias
12	SQtConsOrigBCO180d	Num	Demanda por crédito em bancos 180 dias
13	SQtConsOrigBCO360d	Num	Demanda por crédito em bancos 360 dias
14	SQtConsOrigVAR30d	Num	Demanda por crédito em varejista nos últimos 30 dias
15	SQtConsOrigVAR60d	Num	Demanda por crédito em varejista nos últimos 60 dias
16	SQtConsOrigVAR90d	Num	Demanda por crédito em varejista nos últimos 90 dias
17	SQtConsOrigVAR180d	Num	Demanda por crédito em varejista nos últimos 180 dias
18	SQtConsOrigVAR360d	Num	Demanda por crédito em varejista nos 360 dias
19	SQtConsOrigFIN30d	Num	Demanda por crédito em financeira nos últimos 30 dias
20	SQtConsOrigFIN60d	Num	Demanda por crédito em financeiras nos últimos 60 dias
21	SQtConsOrigFIN90d	Num	Demanda por crédito em financeira nos últimos 90 dias
22	SQtConsOrigFIN180d	Num	Demanda por crédito em financeira nos últimos 180 dias
23	SQtConsOrigFIN360d	Num	Demanda por crédito em financeira nos últimos 360 dias
24	SQtConsOrigTEL30d	Num	Demanda por crédito em telefonia nos últimos 30 dias
25	SQtConsOrigTEL60d	Num	Demanda por crédito em telefonia nos últimos 60 dias
26	SQtConsOrigTEL90d	Num	Demanda por crédito em telefonia nos últimos 90 dias
27	SQtConsOrigTEL180d	Num	Demanda por crédito em telefonia nos últimos 180 dias
28	SQtConsOrigTEL360d	Num	Demanda por crédito em telefonia nos últimos 360 dias
29	SQtConsCheq30d	Num	Demanda por cheque nos últimos 30 dias
30	SQtConsCheq60d	Num	Demanda por cheque nos últimos 60 dias
31	SQtConsCheq90d	Num	Demanda por cheque nos últimos 90 dias
32	SQtConsCheq180d	Num	Demanda por cheque nos últimos 180 dias
33	SQtConsCheq360d	Num	Demanda por cheque nos últimos 360 dias
34	SQtConsOrigIND30d	Num	Demanda por crédito em indústria nos últimos 30 dias
35	SQtConsOrigIND60d	Num	Demanda por crédito em indústria nos últimos 60 dias
36	SQtConsOrigIND90d	Num	Demanda por crédito em indústria nos últimos 90 dias
37	SQtConsOrigIND180d	Num	Demanda por crédito em indústria nos últimos 180 dias
38	SQtConsOrigIND360d	Num	Demanda por crédito em indústria nos últimos 360 dias
39	SQtConsOrigSER30d	Num	Demanda por crédito em serviço nos últimos 30 dias
40	SQtConsOrigSER60d	Num	Demanda por crédito em serviço nos últimos 60 dias
41	SQtConsOrigSER90d	Num	Demanda por crédito em serviço nos últimos 90 dias

#	Variável	Tipo	Descrição
42	SQtConsOrigSER180d	Num	Demanda por crédito em serviço nos últimos 180 dias
43	SQtConsOrigSER360d	Num	Demanda por crédito em serviço nos últimos 360 dias
44	SQtConsOrigTELECOM30d	Num	Desconsiderar
45	SQtConsOrigTELECOM60d	Num	Desconsiderar
46	SQtConsOrigTELECOM90d	Num	Desconsiderar
47	SQtConsOrigTELECOM180d	Num	Desconsiderar
48	SQtConsOrigTELECOM360d	Num	Desconsiderar
49	SQtConsOrigSEGUROS30d	Num	Demanda por seguros nos últimos 30 dias
50	SQtConsOrigSEGUROS60d	Num	Demanda por seguros nos últimos 60 dias
51	SQtConsOrigSEGUROS90d	Num	Demanda por seguros nos últimos 90 dias
52	SQtConsOrigSEGUROS180d	Num	Demanda por seguros nos últimos 180 dias
53	SQtConsOrigSEGUROS360d	Num	Demanda por seguros nos últimos 360 dias
54	SQtConsUtil30d	Num	Demanda por crédito em utilities nos últimos 30 dias
55	SQtConsUtil90d	Num	Demanda por crédito em utilities nos últimos 90 dias
56	SQtConsUtil180d	Num	Demanda por crédito em utilities nos últimos 180 dias
57	SQtConsUtil360d	Num	Demanda por crédito em utilities nos últimos 360 dias
58	SQtConslmov30d	Num	Quantidade Consulta Imóveis em 30 dias
59	SQtConslmov90d	Num	Quantidade Consulta Imóveis em 90 dias
60	SQtConslmov180d	Num	Quantidade Consulta Imóveis em 180 dias
61	SQtConslmov360d	Num	Quantidade Consulta Imóveis em 360 dias
62	STpPriConsOrigBCO	Num	Tempo desde Primeira demanda banco
63	STpPriConsOrigVAR	Num	Tempo desde Primeira demanda varejista
64	STpPriConsOrigFIN	Num	Tempo desde Primeira demanda financeira
65	STpPriConsOrigTEL	Num	Tempo desde Primeira demanda telecom
66	STpPriConsCheq	Num	Tempo desde Primeira demanda a cheque
67	STpUltConsOrigBCO	Num	Tempo desde Última demanda banco
68	STpUltConsOrigVAR	Num	Tempo desde Última demanda varejista
69	STpUltConsOrigFIN	Num	Tempo desde Última demanda financeira
70	STpUltConsOrigTEL	Num	Tempo desde Última demanda telecom
71	STpUltConsCheq	Num	Tempo desde Última demanda a cheque
72	STpPriConsOrigSEGUROS	Num	Tempo desde Primeira demanda seguros
73	STpUltConsOrigSEGUROS	Num	Tempo desde Última demanda seguros
74	SQtEmpConsOrigSEGUROS30d	Num	Quantidade de empresas diferentes que demandou crédito de seguradoras nos últimos 30 dias
75	SQtEmpConsOrigSEGUROS60d	Num	Quantidade de empresas diferentes que demandou em de seguradoras nos últimos 60 dias
76	SQtEmpConsOrigSEGUROS90d	Num	Quantidade de empresas diferentes que demandou em de seguradoras nos últimos 90 dias
77	SQtEmpConsOrigSEGUROS180d	Num	Quantidade de empresas diferentes que demandou em de seguradoras nos últimos 180 dias
78	SQtEmpConsOrigSEGUROS360d	Num	Quantidade de empresas diferentes que demandou em de seguradoras nos últimos 360 dias
79	SQtEmpCons30d	Num	Quantidade de empresas diferentes que demandou crédito nos últimos 30 dias
80	SQtEmpCons60d	Num	Quantidade de empresas diferentes que demandou crédito nos últimos 60 dias
81	SQtEmpCons90d	Num	Quantidade de empresas diferentes que demandou crédito nos últimos 90 dias
82	SQtEmpCons180d	Num	Quantidade de empresas diferentes que demandou crédito nos últimos 180 dias
83	SQtEmpCons360d	Num	Quantidade de empresas diferentes que demandou crédito nos últimos 360 dias
84	SQtEmpConsCheq30d	Num	Quantidade de empresas diferentes que demandou cheque nos últimos 30 dias
85	SQtEmpConsCheq60d	Num	Quantidade de empresas diferentes que demandou cheque nos últimos 60 dias
86	SQtEmpConsCheq90d	Num	Quantidade de empresas diferentes que demandou cheque nos últimos 90 dias

#	Variável	Tipo	Descrição
87	SQtEmpConsCheq180d	Num	Quantidade de empresas diferentes que demandou cheque nos últimos 180 dias
88	SQtEmpConsCheq360d	Num	Quantidade de empresas diferentes que demandou cheque nos últimos 360 dias
89	XQtRestrTot	Num	Quantidade de dívidas Totais
90	XQtRestrInc030d	Num	Quantidade de dívidas incluídos nos últimos 30 dias
91	XQtRestrInc090d	Num	Quantidade de dívidas incluídos nos últimos 90 dias
92	XQtRestrInc180d	Num	Quantidade de dívidas incluídos nos últimos 180 dias
93	XQtRestrInc181d	Num	Quantidade de dívidas incluídos acima de 180 dias
94	XQtRestrIncU30d	Num	Desconsiderar
95	XQtRestrIncU60d	Num	Desconsiderar
96	XQtRestrIncU90d	Num	Desconsiderar
97	XQtRestrIncU180d	Num	Desconsiderar
98	XQtRestrInc1A	Num	Quantidade de dívidas incluídos nos últimos 1 ano
99	XQtRestrInc2A	Num	Quantidade de dívidas incluídos nos últimos 2 anos
100	XQtRestrInc5A	Num	Quantidade de dívidas incluídos nos últimos 5 anos
101	XVIRestrInc030d	Num	Valor
102	XVIRestrInc090d	Num	Valor
103	XVIRestrInc180d	Num	Valor
104	XVIRestrInc181d	Num	Valor
105	XVIRestrIncU30d	Num	Valor
106	XVIRestrIncU60d	Num	Valor
107	XVIRestrIncU90d	Num	Valor
108	XVIRestrIncU180d	Num	Valor
109	XVIRestrInc1A	Num	Valor
110	XVIRestrInc2A	Num	Valor
111	XVIRestrInc5A	Num	Valor
112	XQtRestrAti	Num	Quantidade de dividas ativas
113	XQtRestrAti030DPI	Num	Desconsiderar
114	XVIRestrAti	Num	Valor
115	XVIRestrAti030DPI	Num	Valor
116	XQtRestrRes030d	Num	Quantidade de dívidas resolvidas nos últimos 30 dias
117	XQtRestrRes060d	Num	Quantidade de dívidas resolvidas nos últimos 60 dias
118	XQtRestrRes090d	Num	Quantidade de dívidas resolvidas nos últimos 90 dias
119	XQtRestrRes180d	Num	Quantidade de dívidas resolvidas nos últimos 180 dias
120	XQtRestrRes360d	Num	Quantidade de dívidas resolvidas nos últimos 360 dias
121	XQtRestrResU6m	Num	Desconsiderar
122	XQtRestrResU6m030DPI	Num	Quantidade de dívidas resolvidas nos últimos 6 meses com mais de 30 dias de atraso
123	XQtRestrResU6m060DPI	Num	Quantidade de dívidas resolvidas nos últimos 6 meses com mais de 60 dias de atraso
124	XQtRestrResU6m090DPI	Num	Quantidade de dívidas resolvidas nos últimos 6 meses com mais de 90 dias de atraso
125	XQtRestrResU6m180DPI	Num	Quantidade de dívidas resolvidas nos últimos 6 meses com mais de 180 dias de atraso
126	XQtRestrResU6m360DPI	Num	Quantidade de dívidas resolvidas nos últimos 6 meses com mais de 360 dias de atraso
127	XQtRestrAtiAtraso30d	Num	Quantidade de dívidas ativo atraso>30 dias
128	XQtRestrAtiAtraso60d	Num	Quantidade de dívidas ativo atraso>60 dias
129	XQtRestrAtiAtraso90d	Num	Quantidade de dívidas ativo atraso>90 dias
130	XQtRestrAtiAtraso180d	Num	Quantidade de dívidas ativo atraso>180 dias
131	XQtRestrAtiAtraso720d	Num	Quantidade de dívidas ativo atraso>720 dias

#	Variável	Tipo	Descrição
132	XQtProt	Num	Quantidade Protestos
133	XQtProtInc30d	Num	Quantidade Protestos incluídos nos últimos 30 dias
134	XQtProtInc60d	Num	Quantidade Protestos incluídos nos últimos 60 dias
135	XQtProtInc90d	Num	Quantidade Protestos incluídos nos últimos 90 dias
136	XQtProtInc180d	Num	Quantidade Protestos incluídos nos últimos 180 dias
137	XQtProtInc1A	Num	Quantidade Protestos incluídos no último 1 ano
138	XQtProtInc2A	Num	Quantidade Protestos incluídos nos últimos 1 anos
139	XQtProtInc5A	Num	Quantidade Protestos incluídos nos últimos 5 anos
140	XVIProt	Num	Valor
141	XQtProtAti	Num	Quantidade Protestos ativas
142	XVIProtAti	Num	Valor
143	XQtCCFs	Num	Quantidade Cheques devolvidos
144	XQtCCFsInc30d	Num	Quantidade Cheques devolvidos nos últimos 30 dias
145	XQtCCFsInc60d	Num	Quantidade Cheques devolvidos nos últimos 60 dias
146	XQtCCFsInc90d	Num	Quantidade Cheques devolvidos nos últimos 90 dias
147	XQtCCFsInc180d	Num	Quantidade Cheques devolvidos nos últimos 180 dias
148	XQtCCFsInc1A	Num	Quantidade Cheques devolvidos no último ano
149	XQtCCFsInc2A	Num	Quantidade Cheques devolvidos nos últimos 2 anos
150	XQtCCFsInc5A	Num	Quantidade Cheques devolvidos nos últimos 5 anos
151	XVICCFs	Num	Valor
152	XVICCFsInc30d	Num	Valor
153	XVICCFsInc60d	Num	Valor
154	XVICCFsInc90d	Num	Valor
155	XVICCFsInc180d	Num	Valor
156	XVICCFsInc1A	Num	Valor
157	XVICCFsInc2A	Num	Valor
158	XVICCFsInc5A	Num	Valor
159	XQtCCFsResU6m	Num	Quantidade Cheques devolvidos pagos nos últimos 6 meses
160	XQtCCFsResU12m	Num	Quantidade Cheques devolvidos pagos nos últimos 12 meses
161	XQtCCFsResU24m	Num	Quantidade Cheques devolvidos pagos nos últimos 24 meses
162	XVICCFsResU6m	Num	Valor
163	XVICCFsResU12m	Num	Valor
164	XVICCFsResU24m	Num	Valor
165	XQtRefins	Num	Quantidade de dívidas de bancos/financeiras
166	XVIRefins	Num	Valor
167	XQtRefinsATI	Num	Quantidade Negativações por bancos/financeiras ativas
168	XVIRefinsATI	Num	Valor
169	XQtRefinsResU6m	Num	Quantidade de dívidas de bancos/financeiras resolvidas nos últimos 6 meses
170	XQtRefinsResU12m	Num	Quantidade de dívidas de bancos/financeiras resolvidas nos últimos 12 meses
171	XQtRefinsResU24m	Num	Quantidade de dívidas de bancos/financeiras resolvidas nos últimos 24 meses
172	XVIRefinsResU6m	Num	Valor
173	XVIRefinsResU12m	Num	Valor
174	XVIRefinsResU24m	Num	Valor
175	XQtPefins	Num	Quantidade Negativações por não bancos/financeiras
176	XVIPefins	Num	Valor

#	Variável	Tipo	Descrição
177	XQtPefinsATI	Num	Quantidade Negativações por não bancos/financeiras ativas
178	XVIPefinsATI	Num	Valor
179	XQtPefinsResU6m	Num	Quantidade de dívidas de não bancos/financeiras resolvidas nos últimos 6 meses
180	XQtPefinsResU12m	Num	Quantidade de dívidas de não bancos/financeiras resolvidas nos últimos 12 meses
181	XQtPefinsResU24m	Num	Quantidade de dívidas de não bancos/financeiras resolvidas nos últimos 24 meses
182	XVIPefinsResU6m	Num	Valor
183	XVIPefinsResU12m	Num	Valor
184	XVIPefinsResU24m	Num	Valor
185	XQtRestrOrigBCO	Num	Quantidade de dívidas Bancos
186	XVIREstrOrigBCO	Num	Valor
187	XQtRestrOrigFIN	Num	Quantidade de dívidas financeiras
188	XVIREstrOrigFIN	Num	Valor
189	XQtRestrOrigTEL	Num	Quantidade de dívidas telecom
190	XVIREstrOrigTEL	Num	Valor
191	XQtCredoresAti	Num	Quantidade Credores ativos
192	XQtCredoresTot	Num	Quantidade Credores totais
193	XQtRestrLDep	Num	Quantidade de dívidas Lojas de departamentos
194	XVIREstrLDep	Num	Valor
195	SCORE_2BUREAUX	Num	Score crédito
196	SEG_UF	Char	UF
197	CONCEITO	Num	Variável resposta
198	id	Num	

8.5 APÊNDICE 5 – RESULTADO DOS CÁLCULOS DO VALOR DE INFORMAÇÃO

Tabela 19 – Resultado dos Cálculos do Valor de Informação

#	Variável	Valor de Informação	Número de Categorias	Poder de Predição
1	STpUltConsOrigVAR	1,8231	4	Suspeito
2	XQtRestrInc5A	1,0499	2	Suspeito
3	XQtRestrInc2A	0,3935	3	Forte
4	STpUltConsOrigBCO	0,3098	5	Forte
5	SQtEmpCons180d	0,2407	3	Médio
6	SQtCons180d	0,2151	4	Médio
7	SQtEmpCons360d	0,2145	3	Médio
8	SQtConsCheq60d	0,2075	3	Médio
9	SQtCons90d	0,1820	3	Médio
10	SQtEmpCons90d	0,1797	2	Médio
11	SQtConsOrigSER180d	0,1759	3	Médio
12	SQtConsOrigIND60d	0,1722	3	Médio
13	SQtCons360d	0,1628	3	Médio
14	SQtConsCheq90d	0,1503	3	Médio
15	SQtConsOrigSER360d	0,1503	3	Médio
16	SQtConsOrigIND90d	0,1500	3	Médio
17	SQtConsOrigIND180d	0,1490	3	Médio
18	SQtEmpCons60d	0,1472	2	Médio
19	STpPriConsCheq	0,1456	3	Médio
20	STpPriConsOrigSEGUROS	0,1454	2	Médio
21	STpUltConsOrigSEGUROS	0,1454	2	Médio
22	SQtConsCheq180d	0,1411	3	Médio
23	STpPriConsOrigFIN	0,1411	3	Médio
24	STpPriConsOrigVAR	0,1314	3	Médio
25	STpPriConsOrigBCO	0,1245	3	Médio
26	STpPriConsOrigTEL	0,1198	2	Médio
27	STpUltConsOrigTEL	0,1198	2	Médio
28	SQtCons60d	0,1104	2	Médio
29	SQtCons30d	0,1097	2	Médio
30	SQtConsOrigSER90d	0,1054	3	Médio
31	SQtConsOrigSER30d	0,1053	3	Médio
32	SQtConsCheq360d	0,1047	2	Médio
33	SQtConsOrigIND360d	0,1047	2	Médio
34	SQtConsOrigSER60d	0,0916	2	Frac
35	SQtConsOrigIND30d	0,0793	2	Frac
36	SQtConsOrigFIN180d	0,0757	2	Frac
37	SQtConsOrigFIN360d	0,0568	2	Frac
38	SQtConsOrigVAR180d	0,0461	2	Frac
39	SQtConsOrigBCO180d	0,0400	2	Frac
40	SQtConsOrigVAR360d	0,0381	2	Frac
41	SQtConsOrigBCO360d	0,0379	2	Frac
42	SQtEmpCons30d	0,0356	2	Frac
43	XQtRestrAtiAtraso720d	0,0270	2	Frac
44	XQtRestrAtiAtraso60d	0,0183	2	Nenhum
45	XVICCFsResU12m	0,0179	2	Nenhum
46	SQtConsImov360d	0,0176	2	Nenhum
47	XQtRestrAti	0,0149	2	Nenhum
48	SQtConsOrigBCO30d	0,0147	2	Nenhum

#	Variável	Valor de Informação	Número de Categorias	Poder de Predição
49	SQtConsCheq30d	0	1	Nenhum
50	SQtConsImov180d	0	1	Nenhum
51	SQtConsImov30d	0	1	Nenhum
52	SQtConsImov90d	0	1	Nenhum
53	SQtConsOrigBCO60d	0	1	Nenhum
54	SQtConsOrigBCO90d	0	1	Nenhum
55	SQtConsOrigFIN30d	0	1	Nenhum
56	SQtConsOrigFIN60d	0	1	Nenhum
57	SQtConsOrigFIN90d	0	1	Nenhum
58	SQtConsOrigSEGUROS180d	0	1	Nenhum
59	SQtConsOrigSEGUROS30d	0	1	Nenhum
60	SQtConsOrigSEGUROS360d	0	1	Nenhum
61	SQtConsOrigSEGUROS60d	0	1	Nenhum
62	SQtConsOrigSEGUROS90d	0	1	Nenhum
63	SQtConsOrigTEL180d	0	1	Nenhum
64	SQtConsOrigTEL30d	0	1	Nenhum
65	SQtConsOrigTEL360d	0	1	Nenhum
66	SQtConsOrigTEL60d	0	1	Nenhum
67	SQtConsOrigTEL90d	0	1	Nenhum
68	SQtConsOrigVAR30d	0	1	Nenhum
69	SQtConsOrigVAR60d	0	1	Nenhum
70	SQtConsOrigVAR90d	0	1	Nenhum
71	SQtConsUtil180d	0	1	Nenhum
72	SQtConsUtil30d	0	1	Nenhum
73	SQtConsUtil360d	0	1	Nenhum
74	SQtConsUtil90d	0	1	Nenhum
75	SQtEmpConsCheq180d	0	1	Nenhum
76	SQtEmpConsCheq30d	0	1	Nenhum
77	SQtEmpConsCheq360d	0	1	Nenhum
78	SQtEmpConsCheq60d	0	1	Nenhum
79	SQtEmpConsCheq90d	0	1	Nenhum
80	SQtEmpConsOrigSEGUROS180d	0	1	Nenhum
81	SQtEmpConsOrigSEGUROS30d	0	1	Nenhum
82	SQtEmpConsOrigSEGUROS360d	0	1	Nenhum
83	SQtEmpConsOrigSEGUROS60d	0	1	Nenhum
84	SQtEmpConsOrigSEGUROS90d	0	1	Nenhum
85	XQtCCFs	0	1	Nenhum
86	XQtCCFsInc180d	0	1	Nenhum
87	XQtCCFsInc1A	0	1	Nenhum
88	XQtCCFsInc2A	0	1	Nenhum
89	XQtCCFsInc30d	0	1	Nenhum
90	XQtCCFsInc5A	0	1	Nenhum
91	XQtCCFsInc60d	0	1	Nenhum
92	XQtCCFsInc90d	0	1	Nenhum
93	XQtCCFsResU12m	0	1	Nenhum
94	XQtCCFsResU24m	0	1	Nenhum
95	XQtCCFsResU6m	0	1	Nenhum
96	XQtCredoresAti	0	1	Nenhum
97	XQtCredoresTot	0	1	Nenhum
98	XQtPefins	0	1	Nenhum
99	XQtPefinsATI	0	1	Nenhum
100	XQtPefinsResU12m	0	1	Nenhum
101	XQtPefinsResU24m	0	1	Nenhum
102	XQtPefinsResU6m	0	1	Nenhum
103	XQtProt	0	1	Nenhum
104	XQtProtAti	0	1	Nenhum

#	Variável	Valor de Informação	Número de Categorias	Poder de Predição
105	XQtProtInc180d	0	1	Nenhum
106	XQtProtInc1A	0	1	Nenhum
107	XQtProtInc2A	0	1	Nenhum
108	XQtProtInc5A	0	1	Nenhum
109	XQtProtInc60d	0	1	Nenhum
110	XQtRefinsResU12m	0	1	Nenhum
111	XQtRefinsResU24m	0	1	Nenhum
112	XQtRefinsResU6m	0	1	Nenhum
113	XQtRestrAtiAtraso180d	0	1	Nenhum
114	XQtRestrAtiAtraso30d	0	1	Nenhum
115	XQtRestrAtiAtraso90d	0	1	Nenhum
116	XQtRestrInc030d	0	1	Nenhum
117	XQtRestrInc090d	0	1	Nenhum
118	XQtRestrInc180d	0	1	Nenhum
119	XQtRestrInc181d	0	1	Nenhum
120	XQtRestrInc1A	0	1	Nenhum
121	XQtRestrLDep	0	1	Nenhum
122	XQtRestrOrigBCO	0	1	Nenhum
123	XQtRestrOrigFIN	0	1	Nenhum
124	XQtRestrOrigTEL	0	1	Nenhum
125	XQtRestrRes030d	0	1	Nenhum
126	XQtRestrRes060d	0	1	Nenhum
127	XQtRestrRes090d	0	1	Nenhum
128	XQtRestrRes180d	0	1	Nenhum
129	XQtRestrRes360d	0	1	Nenhum
130	XQtRestrResU6m030DPI	0	1	Nenhum
131	XQtRestrResU6m060DPI	0	1	Nenhum
132	XQtRestrResU6m090DPI	0	1	Nenhum
133	XQtRestrResU6m180DPI	0	1	Nenhum
134	XQtRestrResU6m360DPI	0	1	Nenhum
135	XQtRestrTot	0	1	Nenhum
136	XVICCFs	0	1	Nenhum
137	XVICCFsInc180d	0	1	Nenhum
138	XVICCFsInc1A	0	1	Nenhum
139	XVICCFsInc2A	0	1	Nenhum
140	XVICCFsInc30d	0	1	Nenhum
141	XVICCFsInc5A	0	1	Nenhum
142	XVICCFsInc60d	0	1	Nenhum
143	XVICCFsInc90d	0	1	Nenhum
144	XVICCFsResU24m	0	1	Nenhum
145	XVICCFsResU6m	0	1	Nenhum
146	XVIPefins	0	1	Nenhum
147	XVIPefinsATI	0	1	Nenhum
148	XVIPefinsResU12m	0	1	Nenhum
149	XVIPefinsResU24m	0	1	Nenhum
150	XVIPefinsResU6m	0	1	Nenhum
151	XVIProt	0	1	Nenhum
152	XVIProtAti	0	1	Nenhum
153	XVIRefins	0	1	Nenhum
154	XVIRefinsATI	0	1	Nenhum
155	XVIRefinsResU12m	0	1	Nenhum
156	XVIRefinsResU24m	0	1	Nenhum
157	XVIRefinsResU6m	0	1	Nenhum
158	XVIRestrAti	0	1	Nenhum
159	XVIRestrAti030DPI	0	1	Nenhum
160	XVIRestrInc090d	0	1	Nenhum

#	Variável	Valor de Informação	Número de Categorias	Poder de Predição
161	XVIRestrInc181d	0	1	Nenhum
162	XVIRestrInc1A	0	1	Nenhum
163	XVIRestrInc2A	0	1	Nenhum
164	XVIRestrInc5A	0	1	Nenhum
165	XVIRestrIncU180d	0	1	Nenhum
166	XVIRestrIncU30d	0	1	Nenhum
167	XVIRestrIncU60d	0	1	Nenhum
168	XVIRestrIncU90d	0	1	Nenhum
169	XVIRestrLDep	0	1	Nenhum
170	XVIRestrOrigBCO	0	1	Nenhum
171	XVIRestrOrigFIN	0	1	Nenhum
172	XVIRestrOrigTEL	0	1	Nenhum

8.6 APÊNDICE 6 – RESULTADO DAS RECATEGORIZAÇÕES DAS VARIÁVEIS

Tabela 20 – Resultado dos Agrupamento das Variáveis em Categorias

#	Variável	Classe	Qtde "bom" pagador	Qtde "mau" pagador	Proporção "bom" pagador	Proporção "mau" pagador	Odds	WOE	IV
1	SQtCons30d	$x \leq 0,5$	8050	760	0,8638	0,7329	1,1787	0,1644	0,0215
		$x > 0,5$	1269	277	0,1362	0,2671	0,5098	-0,6738	0,0882
2	SQtCons60d	$x \leq 0,5$	7207	645	0,7734	0,6220	1,2434	0,2178	0,0330
		$x > 0,5$	2112	392	0,2266	0,3780	0,5995	-0,5116	0,0774
3	SQtCons90d	$x \leq 0,5$	6608	548	0,7091	0,5284	1,3418	0,2940	0,0531
		$0,5 < x \leq 1,5$	1675	222	0,1797	0,2141	0,8396	-0,1748	0,0060
		$> 1,5$	1036	267	0,1112	0,2575	0,4318	-0,8398	0,1229
4	SQtCons180d	$x \leq 0,5$	4966	370	0,5329	0,3568	1,4935	0,4011	0,0706
		$0,5 < x \leq 1,5$	2298	238	0,2466	0,2295	1,0744	0,0718	0,0012
		$1,5 < x \leq 2,5$	968	147	0,1039	0,1418	0,7328	-0,3109	0,0118
		$> 2,5$	1087	282	0,1166	0,2719	0,4289	-0,8465	0,1314
5	SQtCons360d	$x \leq 0,5$	3322	229	0,3565	0,2208	1,6143	0,4789	0,0650
		$0,5 < x \leq 3,5$	4590	501	0,4925	0,4831	1,0195	0,0193	0,0002
		$> 3,5$	1407	307	0,1510	0,2960	0,5100	-0,6734	0,0977
6	SQtConsCheq60d	$x \leq -0,5$	2007	203	0,2154	0,1958	1,1002	0,0955	0,0019
		$-0,5 < x \leq 693,5$	4487	697	0,4815	0,6721	0,7164	-0,3336	0,0636
		$> 693,5$	2825	137	0,3031	0,1321	2,2946	0,8306	0,1421
7	SQtConsCheq90d	$x \leq -0,5$	4122	309	0,4423	0,2980	1,4844	0,3950	0,0570
		$-0,5 < x \leq 406,5$	4131	658	0,4433	0,6345	0,6986	-0,3587	0,0686
		$> 406,5$	1066	70	0,1144	0,0675	1,6946	0,5275	0,0247
8	SQtConsCheq180d	$x \leq -0,5$	3734	302	0,4007	0,2912	1,3759	0,3191	0,0349
		$-0,5 < x \leq 9,5$	3924	627	0,4211	0,6046	0,6964	-0,3618	0,0664
		$> 9,5$	1661	108	0,1782	0,1041	1,7114	0,5373	0,0398
9	SQtConsCheq360d	$x \leq -0,5$	5074	398	0,5445	0,3838	1,4187	0,3497	0,0562
		$x > -0,5$	4245	639	0,4555	0,6162	0,7392	-0,3021	0,0485
10	SQtConsOrigIND60d	$x \leq -0,5$	2007	203	0,2154	0,1958	1,1002	0,0955	0,0019
		$-0,5 < x \leq 1,5$	3947	632	0,4235	0,6095	0,6950	-0,3639	0,0677
		$> 1,5$	3365	202	0,3611	0,1948	1,8537	0,6172	0,1026
11	SQtConsOrigIND90d	$x \leq -0,5$	4122	309	0,4423	0,2980	1,4844	0,3950	0,0570
		$-0,5 < x \leq 155,5$	4094	654	0,4393	0,6307	0,6966	-0,3616	0,0692
		$> 155,5$	1103	74	0,1184	0,0714	1,6586	0,5060	0,0238
12	SQtConsOrigIND180d	$x \leq -0,5$	3734	302	0,4007	0,2912	1,3759	0,3191	0,0349
		$-0,5 < x \leq 116,5$	4065	645	0,4362	0,6220	0,7013	-0,3548	0,0659
		$> 116,5$	1520	90	0,1631	0,0868	1,8794	0,6309	0,0482
13	SQtConsOrigIND360d	$x \leq -0,5$	5074	398	0,5445	0,3838	1,4187	0,3497	0,0562
		$x > -0,5$	4245	639	0,4555	0,6162	0,7392	-0,3021	0,0485
14	SQtConsOrigSER30d	$x \leq -0,5$	4477	355	0,4804	0,3423	1,4034	0,3389	0,0468
		$-0,5 < x \leq 2,5$	3856	596	0,4138	0,5747	0,7199	-0,3286	0,0529
		$> 2,5$	986	86	0,1058	0,0829	1,2758	0,2436	0,0056
15	SQtConsOrigSER90d	$x \leq -0,5$	4995	400	0,5360	0,3857	1,3896	0,3290	0,0494

#	Variável	Classe	Qtde "bom" pagador	Qtde "mau" pagador	Proporção "bom" pagador	Proporção "mau" pagador	Odds	WOE	IV
		-0,5 < x ≤ 0,5	1837	221	0,1971	0,2131	0,9250	-0,0780	0,0012
		> 0,5	2487	416	0,2669	0,4012	0,6653	-0,4076	0,0547
16	SQtConsOrigSER180d	x ≤ 0,5	6218	498	0,6672	0,4802	1,3894	0,3289	0,0615
		0,5 < x ≤ 1,5	1475	185	0,1583	0,1784	0,8872	-0,1197	0,0024
		> 1,5	1626	354	0,1745	0,3414	0,5111	-0,6711	0,1120
17	SQtConsOrigSER360d	x ≤ 0,5	5504	423	0,5906	0,4079	1,4479	0,3701	0,0676
		0,5 < x ≤ 1,5	1232	150	0,1322	0,1446	0,9140	-0,0900	0,0011
		> 1,5	2583	464	0,2772	0,4474	0,6195	-0,4789	0,0815
18	SQtEmpCons60d	x ≤ 0,5	7696	685	0,8258	0,6606	1,2502	0,2233	0,0369
		x > 0,5	1623	352	0,1742	0,3394	0,5131	-0,6673	0,1103
19	SQtEmpCons90d	x ≤ 0,5	7248	604	0,7778	0,5824	1,3353	0,2892	0,0565
		x > 0,5	2071	433	0,2222	0,4176	0,5322	-0,6307	0,1232
20	SQtEmpCons180d	x ≤ 0,5	6236	481	0,6692	0,4638	1,4427	0,3665	0,0753
		0,5 < x ≤ 1,5	1712	201	0,1837	0,1938	0,9478	-0,0536	0,0005
		> 1,5	1371	355	0,1471	0,3423	0,4298	-0,8445	0,1649
21	SQtEmpCons360d	x ≤ 0,5	5515	415	0,5918	0,4002	1,4788	0,3912	0,0750
		0,5 < x ≤ 2,5	2533	304	0,2718	0,2932	0,9272	-0,0756	0,0016
		> 2,5	1271	318	0,1364	0,3067	0,4448	-0,8102	0,1380
22	STpPriConsOrigBCO	x ≤ 1,5	6226	598	0,6681	0,5767	1,1586	0,1472	0,0135
		1,5 < x ≤ 865,5	1057	255	0,1134	0,2459	0,4613	-0,7738	0,1025
		> 865,5	2036	184	0,2185	0,1774	1,2313	0,2081	0,0085
23	STpPriConsOrigVAR	x ≤ -0,5	2224	376	0,2387	0,3626	0,6582	-0,4182	0,0518
		-0,5 < x ≤ 0,5	5417	418	0,5813	0,4031	1,4421	0,3661	0,0652
		> 0,5	1678	243	0,1801	0,2343	0,7684	-0,2634	0,0143
24	STpPriConsOrigFIN	x ≤ -0,5	1970	343	0,2114	0,3308	0,6391	-0,4477	0,0534
		-0,5 < x ≤ 1,5	5427	413	0,5824	0,3983	1,4622	0,3800	0,0700
		> 1,5	1922	281	0,2062	0,2710	0,7611	-0,2730	0,0177
25	STpPriConsOrigTEL	x ≤ -0,5	3487	566	0,3742	0,5458	0,6856	-0,3775	0,0648
		x > -0,5	5832	471	0,6258	0,4542	1,3779	0,3205	0,0550
26	STpPriConsCheq	x ≤ -0,5	2445	429	0,2624	0,4137	0,6342	-0,4554	0,0689
		-0,5 < x ≤ 0,5	5411	411	0,5806	0,3963	1,4650	0,3819	0,0704
		> 0,5	1463	197	0,1570	0,1900	0,8264	-0,1907	0,0063
27	STpUltConsOrigBCO	x ≤ 0,5	2383	323	0,2557	0,3115	0,8210	-0,1973	0,0110
		0,5 < x ≤ 103,5	851	189	0,0913	0,1823	0,5010	-0,6911	0,0628
		103,5 < x ≤ 617,5	1775	291	0,1905	0,2806	0,6788	-0,3875	0,0349
		617,5 < x ≤ 774,5	991	86	0,1063	0,0829	1,2823	0,2486	0,0058
		> 774,5	3319	148	0,3562	0,1427	2,4955	0,9145	0,1952
28	STpUltConsOrigTEL	x ≤ -0,5	3487	566	0,3742	0,5458	0,6856	-0,3775	0,0648
		x > -0,5	5832	471	0,6258	0,4542	1,3779	0,3205	0,0550
29	STpPriConsOrigSEGUROS	x ≤ -0,5	3166	547	0,3397	0,5275	0,6441	-0,4399	0,0826
		x > -0,5	6153	490	0,6603	0,4725	1,3973	0,3346	0,0628
30	STpUltConsOrigSEGUROS	x ≤ -0,5	3166	547	0,3397	0,5275	0,6441	-0,4399	0,0826
		x > -0,5	6153	490	0,6603	0,4725	1,3973	0,3346	0,0628

#	Variável	Classe	Qtde "bom" pagador	Qtde "mau" pagador	Proporção "bom" pagador	Proporção "mau" pagador	Odds	WOE	IV
31	XQtRestrInc2A	x ≤ 246,5	6673	537	0,7161	0,5178	1,3828	0,3241	0,0642
		246,5 < x ≤ 609	733	324	0,0787	0,3124	0,2517	-1,3793	0,3225
		> 609	1913	176	0,2053	0,1697	1,2095	0,1902	0,0068

8.7 APÊNDICE 7 – ESTIMAÇÃO DOS MODELOS DE REGRESSÃO LINEAR

Tabela 21 – Estimação do modelo de regressão linear reduzido balanceado

Variável	Coefficiente	Erro Padrão	z	p-Valor
Intercepto	-0,5984	0,3266	-1,832	0,06692
XQtRestrInc2A_Cat(246;609]	1,0872	0,2419	4,495	0,00001
XQtRestrInc2A_Cat(609;985]	0,1779	0,2044	0,871	0,38391
STpUltConsOrigBCO_Cat(0,5;104]	0,8777	0,2992	2,933	0,00335
STpUltConsOrigBCO_Cat(104;618]	0,8083	0,2348	3,443	0,00058
STpUltConsOrigBCO_Cat(618;774]	0,2554	0,2341	1,091	0,27531
STpUltConsOrigBCO_Cat(774;1820]	-0,5545	0,1905	-2,911	0,00360
STpPriConsOrigBCO_Cat(1,5;866]	-0,9954	0,3143	-3,167	0,00154
STpPriConsOrigBCO_Cat(866;1820]	-1,0400	0,3005	-3,461	0,00054
SQtEmpCons180d_Cat(0,5;1,5]	0,6299	0,2582	2,439	0,01472
SQtEmpCons180d_Cat(1,5;16]	0,7914	0,3013	2,626	0,00863
SQtConsCheq60d_Cat(-0,5;406]	0,4825	0,3166	1,524	0,12748
SQtConsCheq60d_Cat(406;1820]	-0,2905	0,1774	-1,637	0,10153
SQtConsOrigSER90d_Cat(-0,5;0,5]	-1,1751	0,4724	-2,487	0,01287
SQtConsOrigSER90d_Cat(0,5;1810]	-0,9954	0,4378	-2,274	0,02299
SQtCons30d_Cat(0,5;9]	0,4868	0,2242	2,171	0,02993
SQtCons60d_Cat(0,5;17]	-0,3503	0,2183	-1,605	0,10849
SQtEmpCons360d_Cat(0,5;2,5]	2,1988	0,7951	2,765	0,00569
SQtEmpCons360d_Cat(2,5;16]	2,4208	0,8168	2,964	0,00304
SQtConsOrigSER30d_Cat(-0,5;2,5]	-1,5114	0,6789	-2,226	0,02601
SQtConsOrigSER30d_Cat(2,5;1820]	-0,3230	0,2302	-1,403	0,16061
STpPriConsOrigTEL_Cat(-0,5;1820]	0,4344	0,2643	1,643	0,10032

Tabela 22 – Estimação do modelo de regressão linear reduzido desbalanceado

Variável	Coefficiente	Erro Padrão	z	p-Valor
Intercepto	-2,4672	0,2235	-11,038	0,00000
XQtRestrInc2A_Cat(246;609]	1,2782	0,1585	8,064	0,00000
XQtRestrInc2A_Cat(609;985]	0,1311	0,1493	0,878	0,38010
STpUltConsOrigBCO_Cat(0,5;104]	0,7453	0,2137	3,487	0,00049
STpUltConsOrigBCO_Cat(104;618]	0,7174	0,1756	4,085	0,00004
STpUltConsOrigBCO_Cat(618;774]	0,0033	0,1792	0,019	0,98513
STpUltConsOrigBCO_Cat(774;1820]	-0,5318	0,1549	-3,433	0,00060
STpPriConsOrigBCO_Cat(1,5;866]	-0,9564	0,2207	-4,334	0,00001
STpPriConsOrigBCO_Cat(866;1820]	-1,0251	0,2210	-4,638	0,00000
SQtEmpCons180d_Cat(0,5;1,5]	0,2227	0,1897	1,174	0,24047
SQtEmpCons180d_Cat(1,5;16]	0,5443	0,2100	2,592	0,00955
SQtConsCheq60d_Cat(-0,5;406]	0,1460	0,2183	0,669	0,50354
SQtConsCheq60d_Cat(406;1820]	-0,4245	0,1402	-3,028	0,00246
SQtConsOrigSER90d_Cat(-0,5;0,5]	-1,1041	0,4119	-2,680	0,00736
SQtConsOrigSER90d_Cat(0,5;1810]	-0,9935	0,3942	-2,520	0,01172

Variável	Coefficiente	Erro Padrão	z	p-Valor
SQtCons30d_Cat(0,5;9]	0,3119	0,1498	2,081	0,03739
SQtCons60d_Cat(0,5;17]	-0,0868	0,1504	-0,577	0,56409
SQtEmpCons360d_Cat(0,5;2,5]	1,8277	0,5445	3,357	0,00079
SQtEmpCons360d_Cat(2,5;16]	2,0156	0,5559	3,626	0,00029
SQtConsOrigSER30d_Cat(-0,5;2,5]	-0,9235	0,3461	-2,669	0,00762
SQtConsOrigSER30d_Cat(2,5;1820]	-0,2022	0,1903	-1,062	0,28813
STpPriConsOrigTEL_Cat(-0,5;1820]	0,2383	0,1713	1,391	0,16419