Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Métodos de Estimação de Renda Presumida para Clientes Pessoa Física no ambiente de Open Banking.

Flávio Henrique de Souza Gonçalves

Dissertação apresentada como requisito parcial para conclusão do
Mestrado Profissional em Computação Aplicada

Orientador
Prof. Dr. João Carlos Félix Souza

Brasília
2022

Ficha Catalográfica de Teses e Dissertações

Está página existe apenas para indicar onde a ficha catalográfica gerada para dissertações de mestrado e teses de doutorado defendidas na UnB. A Biblioteca Central é responsável pela ficha, mais informações nos sítios:

http://www.bce.unb.br
http://www.bce.unb.br/elaboracao-de-fichas-catalograficas-de-teses-e-dissertacoes

Esta página não deve ser inclusa na versão final do texto.

Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Métodos de Estimação de Renda Presumida para Clientes Pessoa Física no ambiente de Open Banking.

Flávio Henrique de Souza Gonçalves

Dissertação apresentada como requisito parcial para conclusão do
Mestrado Profissional em Computação Aplicada

Prof. Dr. João Carlos Félix Souza (Orientador)
FT/UnB

Prof. Dr. Peng Yaohao
Ministério da Economia

Prof. Dr. João Gabriel de Moraes Souza
PPCA/UnB

Prof. Dr. Marcelo Ladeira
Coordenador do Programa de Pós-graduação em Computação Aplicada

Brasília, 10 de fevereiro de 2022

Dedicatória

Aos meus pais que sempre me incentivaram e patrocinaram meus estudos e pelo apoio e amor incondicional da minha esposa e meus filhos.

Agradecimentos

Primeiramente, a Deus, pela vida.

Aos meus pais, esposa e filhos, pelo apoio incondicional e paciência.

Ao Prof. Dr. João Carlos Félix Souza (Joca), pela compreensão, amizade e apoio na orientação, durante todos esses anos.

Ao Programa de Pós-Graduação em Computação Aplicada da Universidade de Brasília, seu corpo docente e discente, direção e administração pela oportunidade, em especial ao coordenador Prof. Dr. Marcelo Ladeira, pelas várias oportunidades de conclusão deste mestrado.

Aos colegas das turmas do curso PPCA, em especial a turma de 2015, por propiciar momentos de troca de experiência e amizade.

Aos amigos do BB, em especial ao Iram, Gerson, Varanda, Maranhão, Mietto e Diogo, por me ajudarem a realizar este sonho.

Ao meu grande amigo Raucélio, pela amizade, por ajudar a encarar esse desafio e pelo suporte no LaTeX.

A todos os amigos e familiares, pelo incansável apoio e por compreenderem minha ausência ao longo desses anos de estudo.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES), por meio do Acesso ao Portal de Periódicos.

Resumo

Uma das informações fundamentais no processo de concessão de crédito em uma instituição financeira é a capacidade de o indivíduo honrar os compromissos assumidos. Nesse contexto, a renda individual da pessoa física é de suma importância para determinar, de maneira adequada, a capacidade de pagamento e o volume de recursos a serem disponibilizados a cada cliente. Um modelo estatístico capaz de estimar a renda do indivíduo é deveras relevante, para além do processo de crédito, com influência, por exemplo, na exigência regulatória, na fidelização/prospecção de clientes, no combate à lavagem de dinheiro, na validação de informações prestadas sem a devida comprovação e sujeitas a risco operacional - no momento de internalização do dado -, entre outros. A posse de informações de renda fidedignas e atualizadas torna-se, portanto, grande vantagem competitiva para as instituições financeiras. Dado o início do *Open Banking* no Brasil (princípio que permite a abertura de dados no sistema financeiro), o desafio de estimar a renda presumida das pessoas físicas, por metodologia proprietária, pode ser o diferencial das instituições financeiras e *fintechs*, nesta nova arena do Sistema Financeiro Nacional (SFN). A partir deste ecossistema de compartilhamento de dados foi desenvolvido modelos estatísticos para presumir a renda de clientes de uma Instituição Financeira Brasileira com objetivo de aprimorar a consistência das informações cadastrais, mitigar os riscos de crédito e prospectar novos clientes. O presente trabalho objetiva descrever as etapas de elaboração do modelo preditivo de renda presumida de pessoa física, utilizando modelagem estatística - em especial a regressão quantílica -, aplicadas em base de dados de clientes em uma grande instituição financeira brasileira e comparar seus resultados com os modelos de renda presumida adquiridos de *Bureaus* de Crédito.

Palavras-chave: : Gestão de Riscos de Crédito, Renda Presumida, Regressão Quantílica, inteligência analítica, Open Banking.

Abstract

One of the fundamental pieces of information in the credit granting process to a financial institution is the individual's ability to honor commitments assumed. In this context, the individual income is of paramount importance to properly determine one's ability to pay and the volume of resources to be made available to each customer. A statistical model capable of estimating an individual's income is highly relevant, and goes beyond the credit process, as it influences, for instance, regulatory requirement, customer loyalty/prospecting, money laundering fighting, validation of information provided without proper evidence and subject to operational risk at the time of data internalization, among others. The possession of reliable and up-to-date income information becomes, therefore, a relevant competitive advantage for financial institutions. With the beginning of Open Banking in Brazil (a principle that allows data opening in the financial system), the challenge of estimating the presumed income of individuals, using a proprietary methodology, could be the differential for financial institutions and fintechs in this new arena of the National Financial System (SFN). From this data sharing ecosystem, statistical models were developed to estimate the income of clients of a Brazilian Financial Institution with the objective of improving the consistency of registration information, mitigating credit risks and prospecting new clients. The present work aims to describe the stages of elaboration of the predictive model of estimated income of individuals, using statistical modeling - in particular quantile regression -, applied to a database of clients in a large Brazilian financial institution and to compare its results with the models of presumed income acquired from Credit *Bureaus*.

Keywords: : Credit Risk Management, Presumed Income, Quantile Regression, analytical intelligence, Open Banking.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

Bacen Banco Central do Brasil.

CAPAG Capacidade de Pagamento.

CEP Código de Endereçamento Postal.

CMN Conselho Monetário Nacional.

CMP Capacidade Mensal de Pagamento.

CPF Cadastro de Pessoa Física.

EAD Exposição Dado o *Default*.

IF Instituição Financeira.

KS Kolmogorov–Smirnov.

LGD Perda Dado o *Default*.

LGPD Lei Geral de Proteção de Dados.

MAPE Erro Percentual Absoluto Médio.

PD Probabilidade de *Default*.

PF Pessoa Física.

RAIS Relação Anual de Informações Sociais.

ROC Receiver Operating Characteristic.

SFN Sistema Financeiro Nacional.

Capítulo 1

Introdução

Este trabalho tem por objetivo estimar a renda presumida das Pessoas Físicas de uma determinada Instituição Financeira (IF) Brasileira utilizando modelagem estatística em um ambiente de *data lake*. A renda do indivíduo é informação fundamental no processo de concessão de crédito. Calcular os rendimentos estimados dos clientes permite determinar, de maneira mais acurada, quais valores serão concedidos em operações a lhe serem ofertadas. Portanto, para as instituições financeiras, a posse da melhor informação de renda precisa e atualizada constitui vantagem competitiva, neste ambiente de *Open Banking.*

A presente pesquisa foi organizada da seguinte forma: nesta seção, foram abordados o problema da pesquisa - que envolve a utilização de técnicas de inteligência analítica para mensuração da renda presumida de clientes pessoa física, em uma instituição financeira - e seus objetivos geral e específicos - ; na seção seguinte, foram discutidos o referencial teórico - desenvolvendo-se abordagem do *Open Banking* no Brasil -, os conceitos e as principais características do risco de crédito, os estudos de renda presumida, os fundamentos e conceitos da técnica de regressão quantílica; na terceira seção, foi abordado o escopo de aplicação do estudo, ou seja, o portfólio em que serão implementados e desenvolvidos modelos de renda presumida; na quarta, serão detalhados os resultados obtidos com os modelos de renda presumida; por fim, tem-se a conclusão, na qual se apresentam os resultados finais da pesquisa, com síntese da aplicação da modelagem aqui discutida e sua utilização no ambiente de *Open Banking.*

1.1 Problema de Pesquisa

Uma falha de mercado ou assimetria de informação ocorre quando os mecanismos de mercado, não regulados pelo Estado e deixados livremente ao seu próprio funcionamento, originam resultados econômicos não eficientes ou indesejáveis do ponto de vista social.

As imperfeições (ou falhas) de mercado, porém, não se limitam aos casos de monopólio ou oligopólio, e outras situações vêm sendo estudadas, tais como: a indivisibilidade do produto, custos de transação elevados, externalidades, riscos e incertezas na oferta de bens e assimetria de informações entre os agentes econômicos, que são na verdade os fundamentos para a intervenção do Estado na economia ou para a regulação do mercado pelo Estado.

Em 2020, o Banco Central do Brasil (Bacen) e o Conselho Monetário Nacional (CMN) introduziram, na indústria financeira, o conceito de *Open Banking* - sistema financeiro aberto.

Nele, clientes de produtos e serviços financeiros podem permitir, de forma segura, ágil e conveniente, tanto o compartilhamento de suas informações entre diferentes instituições autorizadas pelo Banco Central quanto a movimentação de suas contas bancárias em diferentes plataformas - e não apenas pelo aplicativo ou site do banco.

Dentre as melhorias suscitadas pelo Bacen, o *Open Banking* trará mais competição ao (SFN), pois, com acesso ao histórico de dados dos usuários, as instituições participantes poderão fazer ofertas de produtos e serviços para clientes de seus concorrentes, trazendo benefícios ao consumidor, que poderá obter tarifas mais baixas e condições mais vantajosas.

A resolução Conjunta nº 3 do CMN e Bacen [1] alterou a Resolução Conjunta nº 1 [2], que dispõe sobre a implementação do Sistema Financeiro Aberto (*Open Banking*). O Art. 3º apresenta os objetivos do *Open Banking*, quais sejam, incentivar a inovação, promover a concorrência, aumentar a eficiência do Sistema Financeiro Nacional e do Sistema de Pagamentos Brasileiro e promover a cidadania financeira.

No ambiente *Open Banking*, para concessão de crédito a um cliente pessoa física, os órgãos reguladores das instituições financeira exigem a análise de crédito do cliente (*credit score)*, por meio de modelos estatísticos ou algoritmos matemáticos, nos quais, informações de renda ou da capacidade de pagamento dos clientes são variáveis essenciais. Se obtida de forma precisa, essa informação pode mitigar o risco de inadimplência em operações de crédito (*default*) e permitir que as IFs direcionem seus produtos e serviços a um público apto a consumi-los, dimensionado seu crédito, conforme sua capacidade de pagamento.

A renda presumida é, geralmente, uma das variáveis mais relevantes nos modelos estatísticos de credit score, Siddiqi [3], pois permite mensurar-se a capacidade de consumo ou pagamento de um consumidor. Entretanto, essa informação, caso não esteja corretamente calibrada, pode conduzir à concessão indevida de créditos, com grandes efeitos negativos.

A renda presumida pode ser estimada analisando-se diversas variáveis explicativas, como escolaridade, estado civil, profissão, ocupação, faixa etária, local de residência, situ-

ação na receita federal, dentre outras. Com base em todas essas informações, associadas a dados internos da Instituição Financeira, é possível estimar a renda presumida de uma pessoa física, visando a balizar - com segurança - os riscos envolvidos na concessão de um crédito, principalmente, aqueles sem vinculação de garantia. As instituições financeiras brasileiras, comumente, consomem informações de renda presumida produzidas por *Bureaus* de crédito.

Contudo, a necessidade de se aperfeiçoar os modelos de renda presumida e de se obter resultados personalizados e mais precisos sobre a capacidade de pagamentos dos seus clientes - além de reduzir custos -, tem levado as IFs a desenvolver modelos estatísticos proprietários para esta finalidade.

O início do *Open Banking* no Brasil, a necessidade de se obter modelos de renda presumida personalizados e mais precisos - o que permite às IFs oferecer crédito novo para clientes e não clientes, neste novo ecossistema de compartilhamento de dados - motivou a realização desta pesquisa, cujo objetivo é estimar a renda das Pessoas Físicas que já possuem relacionamento com uma determinada IF (Modelo *Behavior*) e daquelas sem relacionamento com a IF (Modelo *Bureau*).

1.1.1 Questões de Pesquisa

A seguir, destacam-se questões de pesquisa utilizadas para responder ao problema apresentado.

1. Os modelos de renda presumida mensuram, adequadamente, a capacidade de pagamento dos clientes das IFs?
2. Existem diferenças relevantes entre os modelos *behavior* (clientes com relacionamento com as IFs) e *Bureau* (clientes sem relacionamento com as IFs)?
3. As informações compartilhadas pelos clientes no ecossistema do *Open Banking* serão suficientes para estimar a renda presumida?

1.2 Justificativa

O aprimoramento contínuo dos modelos de risco de crédito traz maior segurança às instituições financeiras, gerando oportunidades para realização de negócios mais rentáveis e duradouros, beneficiando suas estruturas de capital e a sustentabilidade de seus negócios.

No Brasil, tanto para aplicação em seus modelos de *Credit Scoring* quanto nas regras de tomada de decisão que aferem limites de crédito, a maior parte das IFs adota modelos de renda presumida fornecidos por *Bureaus* de crédito sem nenhuma customização para seus clientes.

No entanto, com o início do *Open Banking* no Brasil, muitas IFs fizeram - e ainda fazem - investimentos consideráveis em inteligência analítica, visando ao uso de modelos de risco de crédito proprietários, com estimativas de renda presumida.

Nesse sentido, justifica-se o aprofundamento desta pesquisa na área para avaliar e melhor compreender modelos de renda presumida, em um cenário de compartilhamento de dados no *Open Banking*, e o seu poder de predição, na capacidade de pagamento dos clientes.

Mister se faz destacar que a gestão do risco de crédito é peça fundamental na definição da estratégia de negócios, no estabelecimento do apetite e tolerância a risco e na elaboração do orçamento das Instituições Financeiras que visam ao alinhamento do retorno esperado dos acionistas e à contribuição para o desenvolvimento do sistema financeiro nacional e da sociedade.

Por fim, vale ressaltar que o debate em torno do uso de inteligência analítica, na Gestão do Risco de Crédito, é crescente no meio acadêmico. Por isso, enfatizamos a importância do tema para um projeto de pesquisa e seu alinhamento com a busca constante das melhoras práticas adotadas na indústria financeira.

1.3 Objetivo Geral

Estimar modelos de renda presumida para Pessoas Físicas (PF) de uma determinada IF com fulcro em implementá-los tanto nos modelos de risco de crédito (*Credit Scoring*) quanto nas regras de tomada de decisão de uma IF que busca aferir limite de crédito, no ecossistema do *Open Banking*.

1.4 Objetivos Específicos

De forma mais específica, o estudo pretende contribuir em:

1. desenvolver modelos preditivos de renda presumida de Pessoas Físicas, para os clientes de uma grande IF;
2. analisar a utilização desses modelos de renda presumida em ambiente de compartilhamento de dados do *Open Banking*; e
3. aferir limites de crédito para clientes a partir da renda presumida estimada.

Capítulo 2

Referêncial Teórico

2.1 Open Banking

O *Open Banking* é um conjunto de regras e tecnologias que permite o compartilhamento de dados financeiros dos clientes entre instituições financeiras e de pagamentos, por meio da abertura e integração de seus respectivos sistemas.

Ele também é conhecido como "sistema bancário aberto", definido por Veiga et al. [4] como uma forma de compartilhamento de informações, produtos e serviços do sistema financeiro por instituições financeiras e pelas demais instituições autorizadas, a critério do usuário dos produtos e serviços.

O *Open Banking* já é prática adotada na União Europeia - precursora na regulamentação deste assunto e de meios de pagamento - que busca agregar inovação, como forma de eficiência no mercado financeiro, propiciar um ambiente mais competitivo e preservar os direitos consumeristas.

Hong Kong é outro exemplo de iniciativa de regulação em sistema de *Open Banking*. Com a edição, em 2017, de medidas destinadas à preparação de um ecossistema propício à implementação do *Open Banking*, garantindo competitividade no setor bancário, construindo ambiente seguro para promover as transações e, ainda, acompanhando tendências internacionais.

No Brasil, o projeto foi inspirado na experiência britânica, mas adaptado à realidade da indústria financeira brasileira, e propunha fornecer mais transparência ao sistema financeiro nacional.

O Banco Central do Brasil, nos últimos anos, vem introduzindo uma série de transformações no Sistema Financeiro Nacional, visando a adotar sua estratégia da Agenda BC+, promovendo mudanças e evoluções em termos de inclusão financeira, modernização de arcabouço legal, aumento de eficiência tecnológica e operacional e implantação de estímulos para promoção de créditos mais baratos. Nesse bojo, uma das medidas mais recentes é a

implantação do *Open Banking*, que estimula concorrência e redução no custo do crédito e que muda o paradigma tecnológico do mercado e dos serviços prestados pelo SFN, na medida em que oferta melhores produtos e serviços financeiros para o consumidor.

A Resolução Conjunta Nº 1, de 04 de maio de 2020 [2], é o marco do *Open Banking*, no Brasil, e dispõe sobre a implementação deste Sistema Financeiro Aberto por instituições financeiras, instituições de pagamento e demais instituições autorizadas a funcionar, pelo Banco Central do Brasil.

Constituem os principais objetivos do *Open Banking* no Brasil:

1. incentivar a inovação;
2. promover a concorrência;
3. aumentar a eficiência do Sistema Financeiro Nacional e do Sistema de Pagamentos Brasileiro;
4. promover a cidadania financeira.

O *Open Banking* abrange o compartilhamento de, no mínimo:

- I - dados sobre:
    a) canais de atendimento relacionados com: 1. dependências próprias; 2. correspondentes no País; 3. canais eletrônicos; e 4. demais canais disponíveis aos clientes;
    b) produtos e serviços relacionados com: 1. contas de depósito à vista; 2. contas de depósito de poupança; 3. contas de pagamento pré-pagas; 4. contas de pagamento pós-pagas; 5. operações de crédito; 6. operações de câmbio; 7. serviços de credenciamento em arranjos de pagamento; 8. contas de depósito a prazo e outros produtos com natureza de investimento; 9. seguros; e 10. previdência complementar aberta;
    c) cadastro de clientes e de seus representantes; e
    d) transações de clientes relacionadas com: 1. contas de depósito à vista; 2. contas de depósito de poupança; 3. contas de pagamento pré-pagas; 4. contas de pagamento pós-pagas; 5. operações de crédito; 6. conta de registro e controle de que trata a Resolução nº 3.402, de 6 de setembro de 2006; 7. operações de câmbio; 8. serviços de credenciamento em arranjos de pagamento; 9. contas de depósito a prazo e outros produtos com natureza de investimento; 10. seguros; 11. previdência complementar aberta; e
- II - serviços de:

a) iniciação de transação de pagamento; e

b) encaminhamento de proposta de operação de crédito

Goettenauer [5] define que o novo modelo de *Open Banking* pode ser compreendido com a abertura dos sistemas tecnológicos bancários e com a inclusão de Interfaces de Programação de Aplicações (*Application Programming Interface* - APIs) padronizadas, por meio das quais instituições externas passam a interagir, tecnologicamente, com bancos, no interesse e benefício dos clientes finais.

O *Open Banking* está calcado sob os pilares da eficiência do Sistema Financeiro Nacional e da oferta de crédito mais barato. Ele foi desenhado para propiciar o compartilhamento padronizado de dados e serviços por meio de APIs das instituições autorizadas pelo Banco Central do Brasil. No atinente aos dados de clientes (pessoa física - PF - ou jurídica - PJ), no escopo do *Open Banking*, é o cliente quem decidirá quando e com quem deseja compartilhá-los..

Uma das motivações para se implementar o *Open Banking*, de acordo com Cavalcanti [6], é o fato de o custo do crédito bancário no Brasil ser um dos maiores do mundo. Entre outros fatores, isso decorre da alta concentração bancária brasileira, da baixa mobilidade de clientes entre instituições financeiras e do alto custo do crédito embutido nas taxas de juros.

Outra motivação citada por Cavalcanti [6] reside na facilidade que o consumidor terá de gerenciar, unificadamente, suas contas e serviços, mesmo quando o produto pertencer à instituição diferente daquela na qual possui relacionamento.

No Brasil, a jornada de Open Banking foi iniciada em 2020, com a organização do Bacen, de IFs, de fintechs e de associações participantes do mercado financeiro e de pagamentos. Sua implementação foi dividida em quatro fases:

Fase 1 - disponibilização ao público, por parte das instituições participantes, de informações padronizadas sobre os seus canais de atendimento e suas características de produtos e serviços bancários tradicionais. Nessa fase, não foi compartilhado nenhum dado de cliente. Nela, puderam surgir soluções que comparavam diferentes ofertas de produtos e serviços financeiros, auxiliando as pessoas a escolherem a opção mais adequada ao seu perfil e necessidades. Entre as possíveis soluções, citam-se os comparadores de tarifas bancárias, de tipos de contas e de cartões de crédito.

Fase 2 - compartilhamento de dados cadastrais e transacionais sobre serviços bancários tradicionais (contas, crédito e pagamentos). A partir dessa fase, os clientes - caso queiram - poderão solicitar compartilhamento, entre instituições participantes, de seus dados cadastrais, de informações sobre transações em suas contas, cartão de crédito e produtos de crédito contratados. Vale reforçar que o compartilhamento

ocorre apenas se, expressamente, autorizado, sempre para finalidades determinadas e por um prazo específico. Será possível ao cliente cancelar a autorização, a qualquer momento, em quaisquer das instituições envolvidas. Como principal benefício, clientes terão a oportunidade de receber ofertas de produtos e serviços mais adequados ao seu perfil, a custos mais acessíveis e de forma mais ágil e segura. Poderão ainda surgir soluções mais personalizadas de gestão e de aconselhamento sobre finanças pessoais, por exemplo. O ecossistema financeiro como um todo também ganha com mais inovação, maior competitividade e racionalização de processos.

Fase 3 - introdução de transações de PIX por iniciadores de transação de pagamento, com a entrada gradual dos demais arranjos de pagamento. Nessa fase, surge a possibilidade de compartilhamento dos serviços de iniciação de transações de pagamento e de encaminhamento de proposta de operação de crédito. Isso abre caminho para o surgimento de novas soluções e ambientes para a realização de pagamentos e para a recepção de propostas de operações de crédito, possibilitando o acesso a serviços financeiros de forma mais fácil, célere e por meio de canais mais convenientes para o cliente, preservando a segurança do processo. Também nesses casos, o compartilhamento só acontece com a autorização prévia e específica do cliente.

Fase 4 - compartilhamento de informações sobre produtos de investimentos, previdência, seguros, câmbio, entre outros, ofertados e distribuídos no mercado. Dados sobre outros serviços financeiros passam a fazer parte do escopo do *Open Banking*. Clientes - sempre que quiserem e autorizarem - poderão compartilhar informações de operações de câmbio, investimentos, seguros, previdência complementar aberta e contas-salário, bem como acessar informações sobre as características dos produtos e serviços - com essa natureza - disponíveis para contratação no mercado.

Assim, amplia-se ainda mais a possibilidade de surgimento de novas soluções para oferta e contração de produtos e serviços financeiros mais integrados, personalizados e acessíveis, sempre com o consumidor no centro das decisões.

O principal desafio da regulação ou autorregulação bancária na implementação do *Open Banking*, conforme pontua Cavalcanti [6], está relacionado à tecnologia das interfaces. Esse ponto é central e exige um padrão mínimo de qualidade e operabilidade. Associadas a padrões tecnológicos, mencionam-se a segurança da informação e a proteção dos dados como o próximo desafio à regulamentação bancária, no que tange a *Open Banking*.

O *Open Banking* é um caminho sem volta e integra a transformação digital que as instituições financeiras estão vivenciando, mas também demanda regulação adequada para que o mercado, como um todo, possa funcionar de forma organizada e estável.

2.2 Risco de Crédito

Abordar-se-ão aqui os principais temas sobre conceitos de risco de crédito e utilização dos modelos de *Credit Scoring*, no apoio à tomada de decisão para concessão do crédito, nas instituições financeiras.

Os primeiros modelos de *Credit Scoring* foram desenvolvidos na década de 1960 e se referiam aos métodos de discriminação sugeridos por Fisher [7], nos quais os modelos eram baseados na sua função discriminante. Conforme assinala Thomas [8], David Durand, em 1941, foi o primeiro a reconhecer que a técnica de análise discriminante, desenvolvida por Fisher, poderia ser usada para separar bons e maus clientes.

De acordo com Kang e Shin [9], Durand apresentou um modelo que atribuía pesos para cada uma das variáveis, usando análise discriminante. Assim, a abordagem de Fisher pode ser vista como ponto inicial na evolução e modificação das metodologias utilizadas nesse tipo de problema, na gestão do risco de crédito até os dias atuais, em que técnicas como árvores de decisão, regressão logística, redes neurais, algoritmos genéticos e técnicas de *Machine Learning* são bastante utilizados pela Indústria Financeira.

Henry Markowitz [10] foi um dos pioneiros no desenvolvimento de modelos estatísticos para uso financeiro, o qual foi utilizado para medir o efeito da diversificação no risco total de uma carteira de ativos.

Fischer Black e Myron Scholes [11] desenvolveram um modelo para a precificação de opções, uma das mais importantes fórmulas usadas no mercado financeiro.

Diretores do Citicorp [12] lançaram o livro Risco e Recompensa: O Negócio de Crédito ao Consumidor, com as primeiras menções ao modelo de *Credit Scoring* (baseado em dados cadastrais dos clientes, utilizado nas decisões de aceitação de proponentes a créditos); ao modelo de *behaviour Scoring* (baseado em dados transacionais, utilizado nas decisões de manutenção ou renovação de linhas e produtos para os clientes). Estes e vários outros modelos são utilizados entre as principais ferramentas de suporte à concessão de crédito, por inúmeras instituições financeiras, no mundo.

O processo de avaliação e gerenciamento de risco de crédito em instituições financeiras vem passando por um movimento de revisão, ao longo dos últimos anos, com a proliferação das técnicas de inteligência analítica em grandes bancos de dados (*big data*), no ambiente *data lake*. O risco de crédito pode ser definido ou pelas perdas geradas por um evento de *default* do tomador ou pela deterioração da sua qualidade de crédito.

Diniz e Louzada [13] destacam que os modelos estatísticos passaram a integrar importante instrumento de auxílio aos gestores de risco, gestores de fundos, bancos de investimento, gestores de créditos e gestores de cobrança, levando-os à tomadas de decisão corretas e, por essa razão, as instituições financeiras passaram a aprimorá-los continuamente.

Conforme Resolução CMN No. 4557/17 [14], um *default* ou ativo problemático pode ser caracterizado pelo atraso no pagamento de uma obrigação por mais de 90 dias, pela honra antecipada de garantias vinculadas à operação, pela quebra de cláusula contratual restritiva (*covenant*), por pedidos de recuperação Judicial, concordata ou falência, abatimentos negociais e quando ocorre a restruturação das dívidas com vantagens para os clientes, em decorrência da deterioração creditícia, indicando que os fluxos de caixa esperados não serão suficientes para honrar as obrigações assumidas.

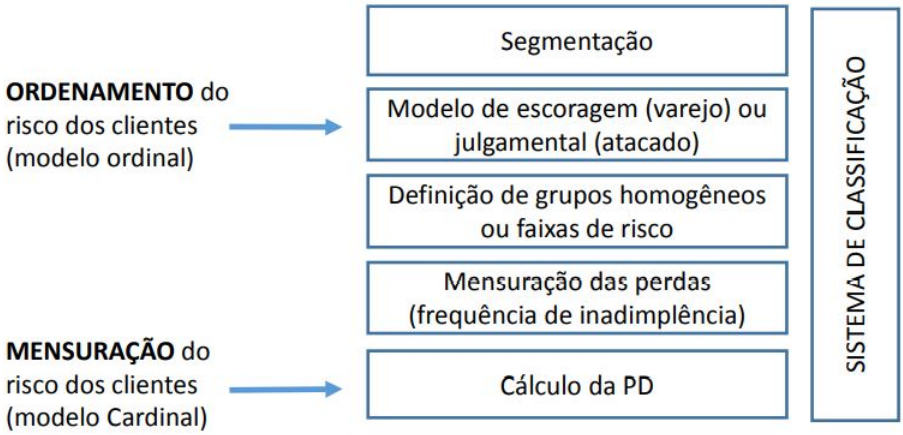
Neto e Brito [15] define que o conceito de risco de crédito pode ser analisado sob diversas perspectivas. Para uma instituição financeira, risco de crédito refere-se, principalmente, à atividade de colocar valor à disposição de um tomador de recursos sob a forma de um empréstimo ou financiamento, mediante compromisso de pagamento em data futura.

Por outro lado, para a Resolução CMN 3.721 [16], o risco de crédito é definido como a possibilidade de ocorrência de perdas associadas ao não cumprimento, pelo tomador ou pela contraparte, de suas respectivas obrigações financeiras nos termos pactuados; à desvalorização de contrato de crédito decorrente da deterioração na classificação de risco do tomador; à redução de ganhos ou remunerações; às vantagens concedidas na renegociação e aos custos de recuperação.

Os modelos de risco de crédito compõem ferramental técnico que supre de informações os gestores e contribuem para tomarem decisões que atendam às diretrizes estabelecidas nas políticas de crédito da instituição financeira.

Segundo Yanaka [17], para realizar a avaliação do risco de crédito, a Instituição Financeira necessita de estrutura para gerenciamento do risco de crédito que permita identificação, mensuração, controle e mitigação do risco de crédito, Figura 2.1 .

Figura 2.1: Sistema de Classificação de Risco de Crédito.



Fonte: Yanaka [17]

As instituições financeiras devem sempre avaliar o risco de crédito do tomador, antes de fazer a concessão do crédito. Os modelos para classificação do risco de crédito permitem classificar os clientes quanto à probabilidade de ocorrência de um evento de inadimplência *(default)*.

Na concessão do crédito, a classificação dos clientes por níveis de risco de crédito é informação primordial para a tomada de decisão e traz a percepção da qualidade da carteira originada.

Diniz e Louzada [13] citam diferentes tipos de modelos utilizados no problema de crédito, com o intuito de mitigar o risco crédito e aumentar a rentabilidade. Entre eles, podem-se citar, regressão logística e linear, análise de sobrevivência, redes probabilísticas, árvores de classificação, algoritmos genéticos, *machine learning* e redes neurais.

Para Sicsú [18], o objetivo dos modelos de *Credit Scoring* é prever, na data da decisão do crédito, a probabilidade de que o crédito, se concedido, incorra em perda para o credor. A probabilidade de isso ocorrer, ou seja, a probabilidade de perda em uma operação de crédito, é denominada risco de crédito.

Sicsú [18] define *Credit Scoring* como medida do risco de crédito e modelos de *Credit Scoring* como as fórmulas matemáticas de cálculo dos escores de crédito. Ele também distingue os conceitos de *Credit Scoring* e *rating*; enquanto o primeiro é, basicamente, um processo quantitativo, a determinação do *rating* de um cliente depende, em grande parte, de avaliações subjetivas, podendo ou não contemplar métodos quantitativos como partes do processo.

O risco de crédito pode ser avaliado a partir dos componentes que compreendem a probabilidade de *default* (PD), o risco de exposição do *defaut* (EAD) e o risco de recuperação do *default* (LGD). O risco de *default* está associado à probabilidade de ocorrer um evento de *default* com o tomador em um certo período; o risco de exposição decorre da incerteza em relação ao valor do crédito, no momento do *default* e, por fim, o risco de recuperação se refere à incerteza quanto ao valor que pode ser recuperado pelo credor, no caso de um *default* do tomador. O risco de recuperação depende do tipo do *default* ocorrido e das características da operação de crédito, como valor, prazo e garantias.

A probabilidade de *default* é classificada como 'risco cliente', pois está vinculado às características intrínsecas do tomador de crédito. Os riscos de exposição e de recuperação são classificados como 'risco operação', pois estão associados a fatores específicos da operação de crédito.

2.2.1 Modelos de Classificação do Risco de Crédito

Para Neto e Brito [15], os modelos na classificação de risco de crédito buscam avaliar o risco de um tomador ou uma operação, atribuindo uma medida que representa a ex-

pectativa de risco de *default*, geralmente expressa na forma de uma classificação de risco (*rating*) ou pontuação (escore).

Entre os modelos para classificação de risco de crédito, têm sido objeto de especial atenção dos pesquisadores os chamados modelos de previsão de *default* (*Credit Scoring*). São aqueles cujo objetivo principal é medir a probabilidade de um cliente incorrer em um evento de *default*, ao longo de um dado período.

Esses modelos de são desenvolvidos a partir de uma amostra de casos históricos de tomadoras de crédito, dividida em dois grupos: um que engloba os que incorreram em eventos de *default* - classificados como clientes ruins -; e outro que compreende os que não incorreram em *default* - classificados como clientes bons.

A partir das características dos clientes da amostra, são identificadas as variáveis que melhor discernem os clientes ruins dos clientes bons, no período analisado. Esse conjunto de variáveis selecionadas é utilizado para classificar novos clientes, de acordo com a probabilidade de se tornarem clientes ruins (*default*).

Os modelos *Credit Scoring*, geralmente, se baseiam em algoritmos estatísticos de análise multivariada, a exemplo de modelos de regressão lineares, análises discriminantes, regressões logísticas, árvores de decisão. Mais recentemente, outras técnicas têm sido utilizadas no desenvolvimento de modelos de risco de crédito, utilizando inteligência artificial ou machine learning, como redes neurais e algoritmos genéticos, dentre outros.

Sicsú [18] refere que os modelos de *Credit Scoring* podem ser classificados como *application Scoring*, utilizados para novos clientes solicitantes de crédito. Por outro lado, *Behavioral Scoring* são utilizados para clientes ou ex-clientes que já possuem histórico de crédito. A diferença dos dois modelos reside nas variáveis utilizadas para estimar o risco de crédito. O *Behavioral Scoring* utiliza, além das informações do a*pplication Scoring*, informações relativas ao histórico de bom pagador de créditos concedidos anteriormente. Os modelos *Behavioral Scoring*, por já incorporarem informações do comportamento passado dos clientes, tendem a fornecer modelos mais precisos que o *application Scoring*.

Para Ohtoshi [19], o principal motivo que diferencia as abordagens *Behavioral Scoring* e *application Scoring* é o fato de que, na segunda, há mais variáveis disponíveis para construção do modelo. Além de dados cadastrais e informação de restrições de crédito provenientes de *Bureau* de crédito, há características que descrevem o cliente e o comportamento de pagamento e utilização de produtos.

As etapas principais do desenvolvimento de modelos *Credit Scoring* descritas por Sicsú [18] são:

1. planejamento e definições do que se que modelar;
2. identificação das variáveis dependentes e independentes;

3. planejamento e seleção da amostra;
4. análise e tratamento de dados;
5. cálculo do algoritmo de escoragem (modelagem);
6. análise, comparação e validação do modelo desenvolvido; e
7. ajuste final do modelo e publicação para uso no negócio.

Diniz e Louzada [13] definem o processo para desenvolvimento de um modelo de crédito em várias etapas, entre as quais, planejamento amostral, determinação da pontuação de escore e validação e comparação de modelos.

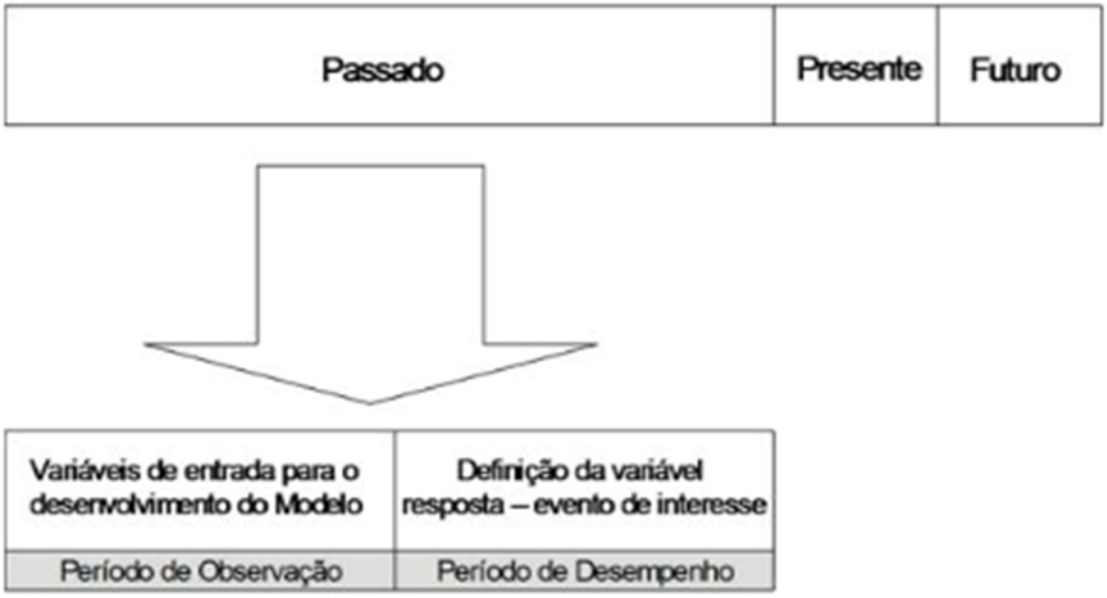
O planejamento amostral na construção de modelos *Credit Scoring* passa pela definição de para qual produto e para quais segmentos de mercados o modelo será desenvolvido. A base de dados utilizada para o desenvolvimento de um modelo é formada por clientes, cujos créditos foram concedidos e cujos desempenhos foram observados, durante um período no passado.

Um fator importante no desenvolvimento do modelo de *Credit Scoring* é o horizonte de previsão; é necessário estabelecer um espaço de tempo para a previsão. Esse será também o intervalo em que o modelo permite fazer previsões de quais clientes terão mais ou menos chances de se tornarem inadimplentes ou de serem menos rentáveis.

Thomas et al. [20] propuseram um período de 12 meses para modelos de *Credit Scoring*, sugerindo que a taxa de descumprimento dos clientes nas instituições financeiras, em função do tempo, aumenta no início, estabilizando-se somente após 12 meses.

Diniz e Louzada [13] tratam o fator tempo como primordial no desenvolvimento de modelos preditivos e, de forma geral, possui três importantes etapas. O passado é composto pelas operações para as quais já foram observados os desempenhos de crédito, durante um horizonte de previsão adotado. As informações cadastrais dos clientes, no momento da concessão do crédito, levantadas no passado mais distante, são utilizadas como variáveis de entrada para o desenvolvimento do modelo e os dados do passado mais recente, as observações dos desempenhos de crédito dos clientes, *default* ou não *default*, são utilizados para a determinação da variável resposta, como pode ser visto na Figura 2.2.

Figura 2.2: Estrutura temporal das informações para construção de modelos preditivos.



Fonte: Diniz e Louzada [13].

Os modelos preditivos construídos a partir de dados históricos podem se ajustar bem no passado, possuindo boa capacidade preditiva. Porém, o mesmo não ocorre, quando aplicados a dados mais recentes.

Diniz e Louzada [13] sugerem dividir a amostra em duas partes, quais sejam, desenvolvimento e validação. Para eles, esse procedimento é conveniente e resulta em benefícios técnicos. Isto é feito para que se verificar o desempenho e comparar os modelos. É importante que a amostra seja, suficientemente, grande para permitir esta divisão.

Lewis [21] acrescenta que, em geral, amostras com tamanhos menores de 1500 clientes bons e 1500 maus, podem inviabilizar a construção de modelos com capacidade preditiva aceitável para um modelo de *Credit Scoring*, além de não permitirem a sua divisão.

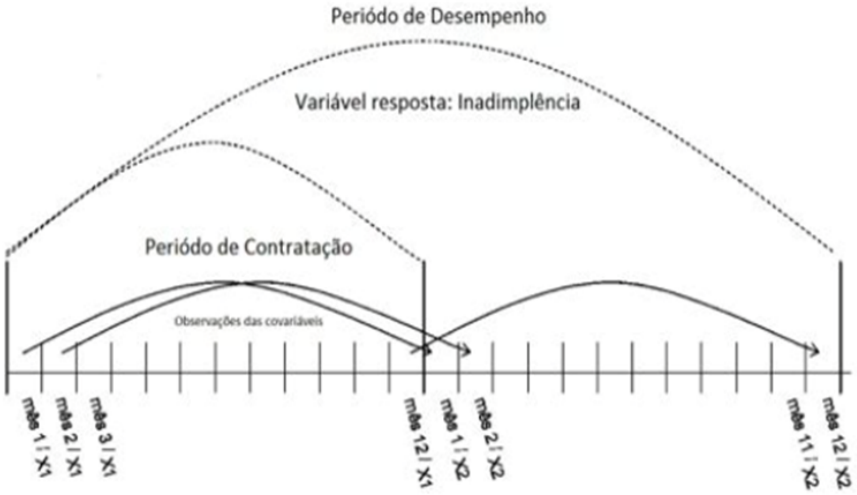
Em grande parte da modelagem com variável resposta binária, um desbalanceamento significativo da variável dependente é observado entre o número de bons e maus pagadores, nas bases de clientes das instituições financeiras. Essa situação pode prejudicar o desenvolvimento do modelo, vez que o número de maus pode ser muito pequeno e insuficiente para estabelecer perfis relacionados às variáveis explicativas e observar possíveis diferenças em relação aos bons clientes.

Para o desenvolvimento dos modelos *Credit Scoring* com problema de desbalanceamento, uma amostra de treinamento é selecionada utilizando a metodologia de *oversampling*. Para tal, considera-se uma amostra balanceada com 50% de bons clientes e 50% de maus clientes. A partir dessa amostra, busca-se atender as quantidades mínimas sugeridas por Lewis [21].

A sazonalidade na ocorrência do evento modelado também é fator a ser considerado, no planejamento amostral. Por exemplo, a seleção da amostra envolvendo momentos específicos no tempo, em que o comportamento do evento é atípico, pode afetar e comprometer, diretamente, o desempenho do modelo em vários pontos, ao longo do tempo. Tal fenômeno é, comumente, chamado de safras de clientes.

No contexto de *Credit Scoring*, Diniz e Louzada [13] mencionam a escolha de 12 safras ao longo de um ano, no intuito de minimizar, consideravelmente, a instabilidade do modelo provocada pelos fatores descritos anteriormente, conforme Figura 2.3.

Figura 2.3: Delineamento amostral com horizonte de previsão 12 meses e 12 safras de clientes.



Fonte: Diniz e Louzada [13].

Eles também indicam que a determinação da pontuação do escore permite melhor definição da técnica estatística a ser utilizada e, consequentemente, um aprimoramento do desenvolvimento do modelo. Essa análise inicial tem alguns objetivos, dentre os quais, destacam-se:

- identificação de eventuais inconsistências e presença de outliers;
- comparação dos comportamentos das covariáveis entre a amostra de bons e maus pagadores, identificando, assim, potenciais variáveis correlacionados com o evento modelado; e
- definição de possíveis transformações de variáveis e a criação de novas a serem utilizadas nos modelos.

A validação e a comparação dos modelos estão relacionadas ao quanto o escore produzido pelo modelo consegue distinguir os eventos bons e maus pagadores, uma vez que

se deseja identificar, previamente, esses grupos e tratá-los de forma distinta, através de diferentes políticas de crédito.

Seguindo o exposto por Diniz e Louzada [13], um aspecto importante na validação dos modelos é o temporal - nele, a situação ideal para se testar um modelo é a obtenção de amostras mais recentes. Isto permite que uma medida de desempenho mais próxima da real e atual utilização do modelo possa ser alcançada.

Em Estatística existem métodos padrões para descrever o quão duas populações são diferentes, quando relacionadas à alguma característica medida e observada.

Uma medida de separação muito utilizada na indústria financeira para avaliar um modelo de *Credit Scoring* é a estatística de Kolmogorov-Smirnov (KS). Os modelos podem também ser avaliados e comparados através da curva ROC (*Receiver Operating Characteristic*), que permite comparar o desempenho de modelos, elegendo-se critérios de classificação dos clientes em bons e maus pagadores, de acordo com a escolha de diferentes pontos de corte, ao longo das amplitudes dos escores observadas para os modelos obtidos.

Em um modelo *Credit Scoring*, o interesse é classificar os indivíduos em uma das duas categorias, bons ou maus clientes, e obter apropriado grau de acerto nestas classificações. Geralmente, nas amostras testes em que os modelos são avaliados, conhece-se a resposta dos clientes - em relação a sua condição de crédito. Assim, estabelecendo-se critérios para classificar estes clientes em bons e maus, torna-se possível comparar a classificação obtida com a verdadeira condição creditícia dos clientes.

Neste ponto, introduz-se a ideia de matriz de confusão. Trata-se de tabela na qual, facilmente, se identificam todos os quatro tipos de classificação do modelo. Com ela, pode-se calcular valores como acurácia, especificidade, sensibilidade etc.

Diniz e Louzada [13] indicam que estabelecer a matriz de confusão é determinar um ponto de corte (*cutoff)* no escore final dos modelos, indicando que indivíduos com pontuação acima desse *cutoff* seriam - por exemplo - classificados como bons e, abaixo desse valor, como maus clientes e, em seguida, comparar essa classificação com a situação real de cada indivíduo. Essa matriz descreve, portanto, uma tabulação cruzada entre a classificação predita através de ponto de corte único e a condição real e conhecida de cada indivíduo, em que a diagonal principal representa as classificações corretas e valores fora dessa diagonal correspondem a erros de classificação - Figura 2.4.

Onde,

- n : número total de clientes na amostra;
- b_B : número de bons clientes que foram classificados como Bons (acerto);
- m_M: número de maus clientes que foram classificados como maus (acerto);
- b_M : número de bons clientes que foram classificados como Maus (erro);

Figura 2.4: Matriz de confusão.

	Situação Real		
Previsão do Modelo	**Bom**	**Mau**	**Total**
Bom	b_B	b_M	b
Mau	m_B	m_M	m
Total	B	M	n

Fonte: Diniz e Louzada [13].

- m_B: número de maus clientes que foram classificados como Bons (erro);
- B : número total de bons clientes na amostra;
- M: número total de maus clientes na amostra;
- b: número total de clientes classificados como bons na amostra;
- m: número total de clientes classificados como maus na amostra.

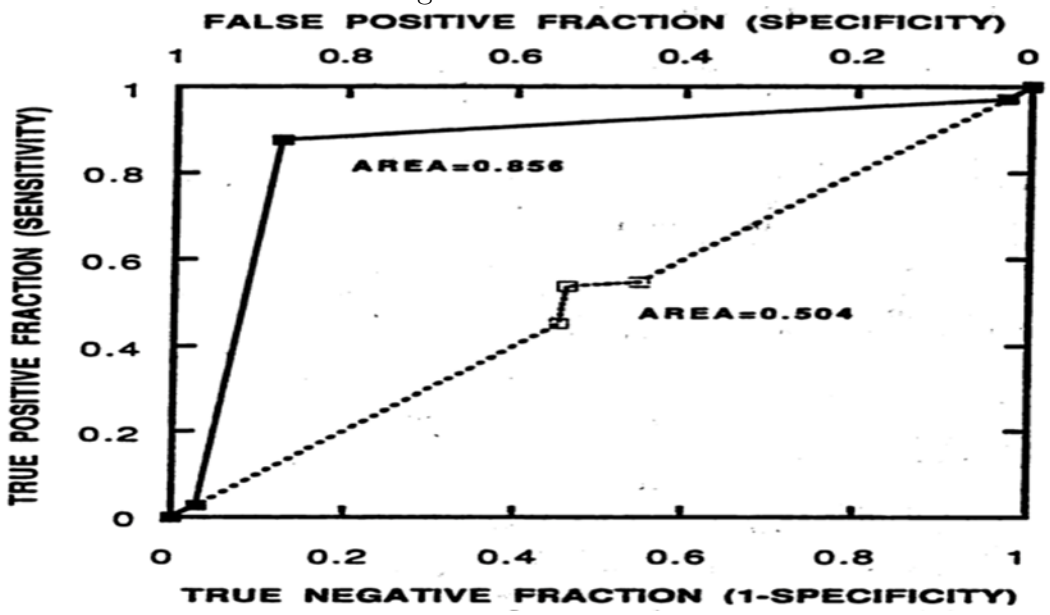
Na indústria financeira, duas medidas muito comuns e bastante utilizadas são a sensitividade e a especificidade. Sensitividade é a probabilidade de um indivíduo ser classificado como mau pagador, dado que realmente é mau, e especificidade é a probabilidade de um indivíduo ser classificado como bom pagador, dado que realmente é bom.

A curva ROC, proposta por Zweig e Campbell [22], é construída variando os pontos de corte, *cutoffs*, ao longo da amplitude dos escores fornecidos pelos modelos, a fim de se obter diferentes classificações dos indivíduos e, consequentemente, os respectivos valores para as medidas de sensitividade e especificidade de cada *cutoff* estabelecido.

Assim, a curva ROC, ilustrada na Figura 2.5, é obtida tendo, no seu eixo horizontal, os valores de (1- especificidade), ou seja, a proporção de bons clientes, classificados como maus clientes pelo modelo (falso negativo) e, no eixo vertical, a Sensibilidade, que é a proporção de maus clientes classificados realmente como maus. Consequentemente, a curva ROC deve ser interpretada de forma que, quanto mais a curva se distanciar da diagonal principal, melhor será o desempenho do modelo em questão.

Como já mencionado, este trabalho não se propõe ao desenvolvimento de modelos de *Credit Scoring* - assunto que possui vasta referência bibliográfica -, mas entendeu-se ser importante compreender seus objetivos e etapas para melhor aplicar a utilização da renda presumida nestes modelos.

Figura 2.5: Curva ROC.



Fonte: Zweig e Campbell [22].

2.3 Renda Presumida

Um dos problemas mais frequentes enfrentados por Instituições Financeiras, na hora de tomar decisões de crédito, é a vulnerabilidade em relação à veracidade das informações da renda mensal do cliente. Não é fácil confirmar esta renda, existe o risco de adulteração de documentos e a dificuldade de obtenção de confirmações junto aos empregadores.

A renda presumida é um modelo estatístico que estuda variáveis cadastrais e demográficas para estimar a faixa de renda dos clientes. Utiliza informações como região, profissão, idade, sexo, movimentações financeiras, dentre outras.

A renda presumida tem como objetivo estimar a faixa de renda de um indivíduo. Na análise e gestão do risco de crédito, é variável importante e, se bem utilizada, será bastante efetiva na análise do *Credit Scoring* e nas políticas de concessão de crédito. Uma das forma para se estimar a renda é a pesquisa de Frias-Martinez e Virseda [23] , na qual os autores apresentam modelo preditivo para dados socioeconômicos de clientes de telefone celulares em um país da América Latina. Os autores possuíam acesso aos dados das ligações, a geolocalização da torre de celular utilizada para cada ligação, o gênero, a idade e o endereço residencial de cada cliente. Foi construído, então, modelo de regressão linear multivariada, utilizando o *Ordinary Least Squares* (OLS) para cada uma das variáveis. A combinação dos três tipos de variáveis de telefonia resultou em modelo preditor de faixa de renda com R2 = 0,83, o que é consideravelmente alto, se comparada com o *benchmarking* do mercado.

Ele é bastante utilizado como variável na definição dos limites de crédito dos clientes,

relacionada ao poder de compra do consumidor, ajudando a estabelecer a capacidade de pagamento do indivíduo.

Mourão [24] dispõe sobre a análise de risco de crédito do cliente como sendo um cálculo efetuado a partir de diversos fatores, dos quais se destaca a comprovação de renda. Além de mitigar o risco de inadimplência em operações de crédito, a correta informação da renda permite a instituições financeiras direcionarem seus produtos e serviços a um dado público com o menor risco.

De maneira prática, ela permite ao gestor de crédito confirmar as informações apresentadas pelo consumidor, comparando o valor declarado com o valor da renda presumida. É possível aplicar uma série de lógicas para apurar a veracidade dos dados, definir limites e até dispensar a apresentação do documento de comprovação de renda. Também pode ser utilizada na revisão periódica dos limites de crédito, visando ao direcionamento e à personalização de novas ofertas aos clientes com perfil adequado e menor risco de inadimplência.

Valter Jr et al. [25] utilizaram, igualmente, modelos estatísticos para estimar a renda presumida. Com seu método, compararam o modelo lognormal em relação ao modelo linear generalizado gama. Eles utilizaram os dados da Pesquisa Nacional por Amostra de Domicílios de 2009 (PNAD), executada pelo Instituto Brasileiro de Geografia e Estatística (IBGE).

Kibekbaev e Duman [26] estimaram a renda dos clientes de bancos turcos. No estudo foi testado o desempenho de vários algoritmos de regressão (por exemplo, regressão de mínimos quadrados ordinários, regressão beta, regressão robusta, regressão ridge, MARS, ANN, LS-SVM e CART, bem como modelos de dois estágios que combinam várias técnicas) aplicados a cinco bancos de dados.

A motivação no artigo de Kibekbaev e Duman [26] é prever a renda dos clientes para os bancos turcos, considerando as novas regulamentações bancárias na Turquia. O regulador bancário da Turquia (BDDK) anunciou resoluções que trouxeram uma série de regras rígidas no uso de cartões de crédito. Uma das principais regras era lançar "limite único" para cartões de crédito; o limite do cartão de crédito do consumidor não poderia ultrapassar quatro vezes o valor de sua renda mensal, para compensar o uso descontrolado de empréstimos e cartões de crédito no mercado interno, além de reduzir os níveis de endividamento das famílias. Todos esses limites seriam aplicáveis a todos os bancos na Turquia, de modo que a soma dos limites dos diferentes bancos não poderia exceder esse "limite único".

Muitos estudos empíricos sobre modelagem de renda são encontrados na literatura. É assaz difícil obter informações exatas sobre a renda, riqueza e características dos indivíduos.

Swan [27] mostrou que as características observadas (como idade, ocupação, sexo, indústria e tempo de trabalho remunerado) explicam uma proporção relativamente pequena de variabilidade na renda. Carrier e Shand [28] argumentaram que os empregadores não se voluntariam facilmente para fornecer dados salariais. Bone e Mitchell [29] mostraram, por exemplo, que a obtenção de dados mais adequados e uma boa modelagem para o benefício da aposentadoria podem levar a melhores estimativas. Lazar [30] fez a previsão de renda utilizando métodos de análise de componentes principais e *support vector machine* aplicada em banco de dados na *Current Population Survey provided by the U.S. Census Bureau*. Esse estudo estatístico foi direcionado para a seleção de recursos relevantes visando a aumentar eficiência no algoritmo e até mesmo melhorar a precisão da classificação.

Para Yamnampet et al [31], a determinação de renda é uma aplicação importante do uso de inteligência analítica preditiva. Ali, a segmentação do cliente ocorre com base em diferentes dados demográficos. Eles apresentaram nova abordagem - com diferentes técnicas de classificação -, para minimizar o risco e o custo envolvidos na previsão de certos níveis de renda, demonstrando o desempenho de cada algoritmo, particularmente, na identificação de clientes.

Dentre os diversos benefícios que caracterizam o modelo da renda presumida, destacamos os citados abaixo, indicando a importância de uso para cada um:

- assertividade na tomada de decisões de crédito: permite conceder e ajustar os limites de crédito, aprimorando as políticas de decisão;.
- prevenção de fraudes: permite direcionar o nível de cuidado e atenção, na validação do comprovante de renda apresentado pelo consumidor; e
- inferência de renda para autônomos e profissionais liberais: permite que seja possível presumir a renda para.

A renda presumida é uma das variáveis de maior peso em modelos estatísticos de *Credit Score*, por possibilitar a mensuração da capacidade de consumo ou pagamento de um cliente. Entretanto, essa informação, caso não esteja corretamente calibrada, pode conduzir à concessão indevida de créditos, com grandes efeitos negativos.

O processo de análise de crédito evoluiu muito, ao longo dos anos. Até pouco tempo, era comum a instituições financeiras solicitar comprovante de renda do cliente, a fim de comprovar se ele tinha ou não atividade remunerada e qual o salário recebido. Atualmente, todo este processo ocorre de forma automatizada e em poucos segundos. A utilização de inteligências analíticas em *Big Data* possibilitou maior agilidade e segurança no processo de análise do perfil de crédito dos clientes, refletindo-se, diretamente, na redução de risco de fraudes e no risco de *default*.

Lessmann et al. [32] efetuaram pesquisa com bases de dados, avaliando, dentre uma série de algoritmos de classificação, algoritmos individuais e técnicas de *ensemble* homogêneo e heterogêneo. Dentre todos os algoritmos testados, três se destacaram por apresentar acurácia melhor do que a obtida na Regressão Logística, técnica, até então, mais utilizada no mercado, quais sejam, *Hill-climbing Ensemble Selection with Bootstrap Sampling (HCES-Bag)*, *Random Forest* (RF) e *Artificial Neural Networks* (ANN), respectivamente. Apesar de o algoritmo HCES-Bag ter sido melhor em acurácia, os autores verificaram que ANN e RF, respectivamente, obtiveram menor número de Falsos Negativos; isto é, deixavam passar menos casos de *default*.

Sundsøy et al. [33] demonstra como o status socioeconômico, em um grande conjunto de dados de telefones móveis, pode ser classificado com precisão, usando-se *deep learning*, o que evita o processo de engenharia de recursos manual e complicado dos modelos tradicionais de mineração de dados.

Uma das técnicas estatísticas mais utilizada pelas IFs e pelos Bureaus de Crédito na estimação da renda presumida é a regressão quantílica proposta por Koenker e Bassett [34] e Koenker [35]. Ela ganhou grande difusão em diferentes áreas cientificas.

2.4 Regressão Quantílica

Para Santos [36], os modelos de regressão são de extrema utilidade em estudos estatísticos, devido tanto à sua facilidade de interpretação quanto à grande diversidade de programas estatísticos hoje capazes de realizar esse tipo de análise. E dentre os métodos de estimação dos parâmetros do modelo, pode-se citar o método de minimização dos quadrados dos erros, como o mais utilizado.

Entre as vantagens da modelagem utilizando a regressão quantílica destacam-se a facilidade na interpretação dos resultados, diversidade de pacotes estatísticos que oferecem esse tipo de análise, utilização do método de minimização de erros absolutos ponderados, não pressupõe que os erros sigam a distribuição normal, as estimativas dos parâmetros não sofrem influência de valores extremos e permite a análise ao longo de toda a distribuição condicional da variável resposta nas covariáveis.

Montgomery et al. [37] descrevem a análise de regressão como técnica estatística utilizada para investigar e modelar o relacionamento entre variáveis.

O método clássico de minimização de mínimos quadrados para estimação de parâmetros, pode ser encontrado para consulta na obra-prima da estatística de Rao [38].

Entretanto, este trabalho está voltado para a técnica chamada de minimização de erros absolutos ponderados, que resulta nos modelos de regressão quantílica. Nas próximas

seções, será apresentada introdução ao uso da técnica de regressão quantílica linear, assim com sua definição e exemplos, para melhor elucidação do tema.

A regressão quantílica foi introduzida originalmente por Koenker e Bassett Jr [34]. Em seu artigo, os autores propõem a classe de modelos no contexto linear, apresentando-a como alternativa ao método de Estimação de Mínimos Quadrados. Destacam ainda sua eficiência na estimação dos parâmetros, sobretudo, para os casos em que os erros não seguem distribuição normal.

Minimização de Erros Absolutos

O método dos mínimos quadrados é, atualmente, a técnica mais utilizada. Entretanto, ela possui limitações que levaram à busca por outros métodos. A principal delas é que esta metodologia está fortemente associada à distribuição normal dos erros, quando essa hipótese não é provada, a performance do método na estimação de parâmetros fica prejudica. Nesse caso, na tentativa de satisfazer às suposições do modelo, Box e Cox [39] sugerem transformar a variável resposta, porém, esta alternativa pode dificultar a interpretação dos parâmetros do modelo ajustado.

Santos [36] indica outra limitação na aplicação do método de mínimos quadrados, qual seja, a influência que *outliers* exercem nas estimativas dos parâmetros do modelo. Faz-se, então, necessária análise da influência, no ajuste do modelo, vez que *outliers*, tanto na variável resposta quanto nas variáveis preditoras, podem atrapalhar a identificação da verdadeira relação entre as variáveis de interesse.

Por outro lado, o método de minimização dos erros absolutos é robusto na presença de *outliers*, na variável resposta. Além disso, quando a distribuição dos erros não é normal, ao estimar o valor mediano da distribuição, esse método se mostra melhor para descrever a posição central da distribuição condicional da variável resposta. A regressão quantílica se baseia no método dos erros absolutos, porém, para se estimar os diversos quantis de interesse, é feita ponderação na minimização desses erros.

Modelos de regressão quantílica apresentam os efeitos médios das covariáveis sobre a resposta, nos diferentes quantis da distribuição dela e, por isso, podem ser vistos como abordagem alternativa à metodologia de regressão usual. Ou seja, enquanto os modelos clássicos se limitam à análise das médias condicionais, a regressão quantílica permite a análise ao longo de toda a distribuição condicional da variável resposta, nas covariáveis.

A regressão quantílica esbarrou, durante muito tempo, na dificuldade de estimação dos parâmetros que, ao contrário dos modelos de regressão lineares usuais, não tem fórmula analítica. Porém, com o advento dos computadores e desenvolvimento das técnicas de programação linear, a metodologia vem ganhando espaço em estudos empíricos e pesquisas acadêmicas.

Definição de Quantis

Seja Y uma variável aleatória com função de distribuição acumulada dada por $F_Y(\cdot)$. O quantil de ordem τ para Y é definido como:

Definição 1: O quantil de ordem τ para Y, $\tau \in [0,1]$, é o menor y tal que $F_Y(y) = \tau$.

Em outras palavras, o quartil de ordem τ, que será denotado por $Q_Y(\tau)$, pode ser visto como resultado da função inversa $F_Y^{-1}(\tau)$, de modo que:

$$Q_Y = F^{-1}(\tau) = inf\{y : F_Y(y) \geq \tau\}, \tau \in [0,1]. \tag{2.1}$$

De acordo com as propriedades conhecidas da função de distribuição acumulada e de sua função inversa, se $F_Y(\cdot)$ é estritamente crescente, então existe um único número real y tal que $F_Y(y) = \tau$.

Em estudos empíricos, no entanto, a função $FY(\cdot)$, não é, em geral, conhecida. Dessa forma, considere uma amostra aleatória $\{y_1, \ldots, y_n\}$ de tamanho n da variável Y. Tem-se então a seguinte definição:

Definição 2: Uma estimativa para o quantil τ de Y é dada pelo menor valor y tal que:

$$\hat{F}_Y(y) = \left\{ \frac{1}{n} \sum_{i=1}^{n} \Im(y_i \leq y) \geq \tau, \quad \Im(y_i \leq y) = \begin{cases} 1, & y_i \geq y \\ 0, & y_i \geq y \end{cases} \right. \tag{2.2}$$

em que $\hat{F}_Y(y)$ é uma estimativa para $F_Y(y)$.

As definições apresentadas anteriormente se baseiam no conceito de ordenação dos dados. Apesar de serem as mais usuais, não são as únicas formas de definir, respectivamente, o quantil populacional e amostral de ordem τ de Y, mas também é possível apresentá-las à luz de um problema de otimização, conforme descrito a seguir. Considere a função de perda $\rho_\tau(Y - y)$, em que:

$$\rho\tau(u) = u\{\tau - \Im(u < 0)\}, \quad \Im(u < 0) = \begin{cases} 1, & u < 0 \\ 0, & u \geq 0 \end{cases} \tag{2.3}$$

sendo que $\rho\tau(u) \leq 0$, $\forall u$. Considerando Y variável aleatória contínua, observe que

$$E[\rho_\tau(Y - y)] = (\tau - 1) \int_{-\infty}^{y} (t - y)\, dF_Y(y) + \tau \int_{y}^{+\infty} (t - y)\, F_Y(y) \tag{2.4}$$

Como a função de distribuição acumulada é não decrescente, todo y em $y : F_Y(y) = \tau$ minimiza o valor esperado da função $\rho\tau(Y - y)$. Portanto, de acordo com a Definição 1, $y = E[\rho\tau(Y - y)]$ é quantil de ordem τ de Y. Assim, uma definição de quantil equivalente pode ser escrita como:

definição 3 O quantil de ordem τ para Y é dado por:

$$\arg\min_{y} E[\rho_\tau(Y-y)] \tag{2.5}$$

Considere, por exemplo, $\tau = \frac{1}{2}$, probabilidade associada ao quantil denominado mediana de Y. Observe que:

$$E\left[\rho_{\frac{1}{2}}(Y-y)\right] = \left(\frac{1}{2}-1\right)\int_{-\infty}^{y}(t-y)\,dF_Y(y) + \frac{1}{2}\int_{y}^{+\infty}(t-y)\,F_Y(y) = \frac{1}{2}E[Y-y] \tag{2.6}$$

Ou seja, minimizar $E\left[\rho_{\frac{1}{2}}(Y-y)\right]$ é equivalente a minimizar $E|Y-y|$. De um modo geral, o quantil definido de acordo com a Definição 3 pode ser visto como uma generalização do problema de minimizar a esperança dos resíduos absolutos resultantes ao usar y para predizer Y.

Observe que, de acordo com a Lei dos Grandes Números, para n suficientemente grande, tem-se que a média amostral da função $\rho_\tau(y_i - y)$ converge para o seu valor esperado. Dessa forma, pode-se escrever a seguinte definição:

Definição 4 Uma estimativa consistente para o quantil de ordem τ de Y é dada pelo valor y que minimiza a soma:

$$S_n(y) = \lim_{n\to\infty}\frac{1}{n}\sum_{1=1}^{n}\left[\rho_\tau(y_i - y)\right]. \tag{2.7}$$

Esta definição de quantil, que considera um problema de otimização, o objetivo é encontrar $y \in \{y_1, \ldots, y_n\}$ que minimiza a função

$$S_n(y) = \frac{1}{n}\sum_{1=1}^{n}\left[\rho_\tau(y_i - y)\right]. \tag{2.8}$$

Essa segunda definição de quantil será base para o entendimento de regressão quantílica, que é introduzida a seguir.

Uma forma intuitiva de entender a regressão quantílica é uma analogia aos modelos de regressão clássica que pode ser vista em Santos [36]. Neste caso, cada valor observado da variável resposta do estudo é dado pela soma de uma parte sistemática, que é quantil de ordem τ de Y_i , $f(x_i, \tau)$, e de um erro aleatório u_i. Isto é:

$$y_i = f(x_i, \tau) + u_i \tag{2.9}$$

com u_i independentes e identicamente distribuídas, $i = 1, \ldots, n$. Supondo-se que o quantil de ordem τ de u_i , condicional a x_i , é igual a zero, observe que a função a ser modelada pode ser expressa da seguinte forma:

$$Q_{y_i}|x_1\left(\tau\right) = f\left(x_i, \tau\right) \tag{2.10}$$

Essa forma de entender o modelo de regressão quantílica é importante para o desenvolvimento da teoria inferencial.

Koenker [35] define que a suposição de erros identicamente distribuídos não é condição necessária para ajuste da regressão quantílica. Ao contrário da metodologia clássica de regressão, os modelos de regressão quantílica são capazes de incorporar a informação de heterocedasticidade dos erros aleatórios independentes.

A interpretação dos parâmetros de seus modelos de regressão quantílica, ocorre da seguinte maneira, considere, que $f(x_i, \tau) = x_i^T \beta(\tau)$ para um τ fixado. Neste caso, a interpretação dos parâmetros $\beta(\tau)$ é essencialmente a mesma de qualquer outro modelo linear, no sentido de se dar em função da taxa de variação. Ou seja, o coeficiente $\beta_j(\tau)$, $j = 1, \ldots, p$, pode ser interpretado como a taxa de variação no τ-ésimo quantil da variável resposta Y ao variar-se em uma unidade o valor da j-ésima covariável mantendo-se os valores das demais variáveis fixos,

$$\beta_j\left(\tau\right) = \frac{\partial Q_Y|x\left(\tau\right)}{\partial x_j}. \tag{2.11}$$

Para estimação dos parâmetros $\beta(\tau)$, relembre inicialmente que, no caso univariado, de acordo com Definição 4, o quantil de ordem τ pode ser consistentemente estimado encontrando-se o valor y da amostra que minimiza a função $\sum_{i=1}^{n} \rho_\tau(y_i - y)$. Na presença das covariáveis, o valor y é modelado por $Q_{Y_i}(\tau|x_i)$, que no caso linear é dado por $x_i^T \beta(\tau)$. Então, o interesse é encontrar $b(\tau)$, estimativa de $\beta(\tau)$, que minimiza a função:

$$S_n\left[b\left(b\right)\right] = \frac{1}{n}\sum_{i=1}^{n} \rho_\tau\left(y_i - z_i^t \beta\left(\tau\right)\right). \tag{2.12}$$

Pelas propriedades bastante conhecidas de cálculo, tem-se que o valor $b(\tau)$ que minimiza a função acima é também raiz da seguinte função de estimação:

$$D_n\left[b\left(\tau\right)\right] = \frac{dS_n\left[b\left(tau\right)\right]}{db\left(\tau\right)} = \frac{1}{n}\sum_{i=1}^{n} x_i\{\Im\left(y_i x_i^T b\left(\tau\right) \leq 0\right)\}. \tag{2.13}$$

No entanto, não é trivial encontrar a raiz dessa equação que, por envolver função indicadora, não assume fórmula analítica. Como alternativa, a literatura sugere reformular-se esta função para equação equivalente, entendendo regressão quantílica como resultado de

um problema de programação linear, em que é possível encontrar a solução, usando-se métodos já conhecidos e consolidados.

O uso das ferramentas de programação linear permitiu o desenvolvimento da regressão quantílica. Entre as técnicas utilizadas para a resolução destes problemas, pode-se citar o método simplex, processo interativo que se inicia com solução que satisfaz as restrições lineares, e faz a busca pela solução que resulta no menor valor da função objetivo (ou maior, em problemas de maximização).

Uma interpretação geométrica do método simplex e maiores detalhes sobre a técnica podem ser encontrados em Koenker [35] e Davino et al. [40]. Em problemas de minimização dos erros absolutos do modelo, o primeiro algoritmo eficiente que fez uso de programação linear foi o proposto por Barrodale e Roberts [41].

Mais tarde, já no contexto de regressão quantílica, Koenker e d'Orey [42] propuseram uma adaptação do método simplex, que é bastante conveniente para um número moderado de observações.

Para grandes amostras, a literatura sugere o uso de outra técnica, computacionalmente, mais eficiente, a do ponto interior, proposta por Portnoy e Koenker [43]. Uma introdução a essa técnica e uma comparação com o método simplex podem ser encontradas em Chen e Wei [44].

A regressão quantílica é uma técnica que permite a modelagem de qualquer quantil de ordem τ de interesse, $\tau \in [0, 1]$. Em alguns casos, inclusive, tem-se interesse em estudar todos os quantis, de modo a compreender toda a distribuição da variável resposta em função das covariáveis.

O problema de estimação dos parâmetros $\beta(\tau)$ foi resolvido aplicando-se ferramenta de programação linear. Uma vez estimados os parâmetros, é importante conhecer suas propriedades e métodos inferenciais disponíveis.

2.4.1 Propriedades e Inferência

Observe que o vetor de parâmetros estimados $b(\tau)$ depende de τ, claramente de y e da matriz de covariáveis X observados na amostra, sendo $b(\tau)$, y e X conforme definidos anteriormente. Dessa forma, para enunciar as propriedades que seguem, denote $b(\tau) = b(\tau, y, X)$.

Teorema de Koenker e Bassett [42]. Seja A matriz não singular de dimensão $p \times p$, $y \in R^p$, e $a > 0$. Então, para qualquer $\tau \in [0, 1]$, pode-se mostrar que:

1. $b(\tau, ay, X) = ab(\tau, y, X)$
2. $b(\tau, -ay, X) = ab(1 - \tau, y, X)$

3. $b(\tau, y + Xy, X) = b(\tau, y, X) + y$

4. $b(\tau, y, XA) = A - 1b(\tau, y, X)$

As propriedades (1) e (2) tratam de equivariância de escala, enquanto que a propriedade (3) aborda o contexto conhecido como equivariância de regressão; e a (4) é chamada de equivariância da reparametrização da matriz de planejamento proposta por Koenker [35].

Outra importante propriedade da regressão quantílica é a equivariâcia sob transformações monótonas. Relembre que, nos problemas de regressão usuais, quando a transformação da variável resposta se faz necessária para obtenção de propriedades desejáveis dos estimadores, - como linearidade, por exemplo -, a interpretação dos parâmetros é comprometida, uma vez que deve ser feita em função da variável transformada. Isso pode ser demonstrado pela desigualdade de Jensen, em que:

$$E\left(g\left(Y\right)\right) \neq g\left(E\left(Y\right)\right) \tag{2.14}$$

Por outro lado, em regressão quantílica tem-se que:

$$Qg\left(Y\right)\left(\tau\right) = g\left(QY\left(\tau\right)\right) \tag{2.15}$$

que deriva diretamente do fato de que $P(Y \leq y) = P(g(Y) \leq g(y))$.

Considere agora o modelo linear definido conforme:

$$y_i = x_i^T \beta\left(\tau\right) + u_i i. \tag{2.16}$$

Suponha que os erros u_i são independentes e identicamente distribuídos com função de distribuição $F(\cdot)$, e que o quantil de ordem τ de u_i seja igual a zero. Considere ainda uma sequência $\tau_1, \ldots, \tau_m$ de probabilidades de interesse, ver Koenker [35]:

1. A função densidade $f(\cdot)$, associada à função de distribuição acumulada $F(\cdot)$, é tal que
$$f\left(F^{-1}\left(\tau\right)\right) > 0, \quad j = 1, \ldots, m.$$

2. O modelo é ajustado com intercepto.

3. Para Q sendo uma matriz positiva definida, ocorre
$$\lim_{n\to\infty} \sum x_i x_i^T = Q$$

Nessas condições, pode-se mostrar que:

$$\sqrt{n}\left(b(\tau_1)-\beta(\tau_1),\ldots,b(\tau_m)-\beta(\tau_m)\right) \xrightarrow{\mathrm{D}} \mathcal{N}\left(0, V\left(\tau_1, ldots, \tau_m\right)\right), \tag{2.17}$$

em que $V(\tau_1,\ldots,\tau_m) = \Omega(\tau_1,\ldots,\tau_m,F) \bigotimes Q^{-1}$, e $\Omega(\tau_1,\ldots,\tau_m,F)$ é a matriz de covariâncias entre os m quantis amostrais, e $\bigotimes$ representa o produto de Koenker. Ou seja, sob as suposições do modelo, os estimadores dos parâmetros da regressão quantílica são não viesados e seguem distribuição normal assintótica. Sob condições adicionais, é possível mostrar a consistência do estimador. Ver, por exemplo, Koenker [35].

Estimada a matriz de covariância, e sob a distribuição normal assintótica, é possível construir intervalos de confiança para os parâmetros a fim de avaliar se podem ser considerados diferente de zero. Neste caso, tem-se que:

$$IC\left(\beta_i, 1-\alpha\right) = b_i\left(\tau\right) \pm t_{\left(\frac{\alpha}{2},n-1\right)} \sqrt{\frac{\hat{V}}{n}}, \tag{2.18}$$

em que $\hat{V}(\tau)$ é uma estimativas de $V(\tau)$ e $t_{\left(\frac{\alpha}{2},n-1\right)}$ é o quantil de ordem $\frac{\alpha}{2}$ da distribuição t-Student com $n-1$ graus de liberdade.

Outra metodologia de construção de intervalos de confiança se baseia na técnica de *bootstrap*. Trata-se de um esquema de *reamostragem* que consiste em selecionar n pares (y_i, x_i) com reposição da amostra original de tamanho n, de modo que cada par tenha probabilidade de $\frac{1}{n}$ de ser sorteado. Esse procedimento é repetido B vezes e, para cada uma delas, o vetor de parâmetros $b(\tau)$ é calculado. Cada uma dessas B estimativas contribui para a estimação do erro padrão dos parâmetros.

Então, um intervalo de confiança para o parâmetro $b_i(\tau)$, $i = 1,\ldots,p$, com coeficiente de confiança $1-\alpha$ é dado por:

$$IC\left(\beta_i, 1-\alpha\right) = b_i\left(\tau\right) \pm t_{\left(\frac{\alpha}{2},n-1\right)} \widehat{EP\left[b_i\left(\tau\right)\right]}, \tag{2.19}$$

A desvantagem em adotar a metodologia de *bootstrap* é que, para grandes amostras, o custo operacional é bastante alto. Santos [36] apresenta outras formas de estimativa dos intervalos de confiança.

Em relação a testes de hipóteses, dois tipos são de interesse. Em primeiro lugar, quer-se testar se parâmetros dentro de um mesmo quantil são iguais a constantes conhecidas, como zero, por exemplo. Neste caso, a literatura sugere a aplicação do teste de Wald, que não apresenta grandes complicações, uma vez estimada a matriz de covariâncias. Outro teste de interesse é avaliar e comparar parâmetros de diferentes quantis.

Conforme apresentado em Koenker [35], é possível escrever teste único de hipóteses para avaliar essas duas situações.

Ele também aponta esse teste de hipóteses geral para regressão quantílica como uma alternativa aos convencionais testes na detecção de heterocedasticidade dos parâmetros, vez que a metodologia de regressão quantílica é a presença de valores discrepantes na variável resposta, Koenker [35].

2.4.2 Seleção de Variáveis nos Modelos de Regressão Quantílica

A seleção de variáveis é etapa crucial tanto nos modelos clássicos de regressão bem como nos modelos de regressão quantílica. A seleção de variáveis, no caso clássico, já apresenta longo desenvolvimento teórico e prático, com a utilização das medidas de regularização, desde o trabalho de Tibshirani [45].

Em um ambiente rico de dados, em que há grande número de covariáveis, é indesejável manter os preditores irrelevantes no modelo final, pois tal cenário dificulta a interpretação do modelo resultante e pode reduzir sua capacidade preditiva.

Na estrutura de regularização, muitos tipos diferentes de penalidades foram introduzidos para alcançar um conjunto de variáveis com propriedades ótimas de previsão (propriedades de oráculo). A penalidade L1 foi utilizada no LASSO proposto por Tibshirani [45] para seleção de variáveis.

Fan e Li [46] propuseram abordagem unificada, via regressão por mínimos quadrados penalizados não concêntricos, que realiza simultaneamente a seleção de variáveis e estimativas de coeficientes. Ao escolher uma função de penalidade não-côncava apropriada, esse método mantém consideráveis méritos na melhor seleção de subconjuntos e da regressão reestimada: produz solução esparsa, garante a estabilidade da seleção de modelos e fornece estimativas parciais para grandes coeficientes. Estas são as três propriedades desejáveis de uma boa penalidade.

Utilizando-se de pesos adaptativos para penalizar diferentes coeficientes na penalidade, Zou [47] introduziu o LASSO adaptativo, demonstrando suas propriedades oraculares. Resultados semelhantes também foram estabelecidos em Zou [48] e Zhang [49]. Eles estudaram o LASSO adaptativo, em modelos de risco proporcional. Huang [50] e Lin [51] estudaram a seleção de variáveis, no cenário de dimensionalidade maior que o tamanho da amostra.

Antes disso, porém, Koenker [52] aplicou a penalidade LASSO ao modelo de regressão quantílica de efeito misto para dados longitudinais, no intuito de encorajar encolhimento estimativo dos efeitos aleatórios.

Wu e Liu [53] desenvolveram o caminho da solução da regressão quantílica penalizada L1. Basicamente, a ideia é penalizar os coeficientes de diferentes covariáveis em um nível diferente, usando pesos adaptativos.

No caso da regressão dos mínimos quadrados, Zou e Yuan [47] propuseram usar, como pesos, as recíprocas das estimativas dos mínimos quadrados ordinários elevadas em alguma potência. A generalização direta, para o nosso caso de regressão quantílica, é usar as estimativas de regressão quantílica não penalizada como pesos. A regressão quantil penalizada adaptativa-LASSO resolve:

$$min_{\beta} Q(\tau) \tag{2.20}$$

onde

$$Q_{yL_1}(\tau) = \sum_{i=1}^{n} \rho_\tau \left(y_i - x_i^{'}\beta\right) + n\lambda_n + \sum_{j=1}^{d} \tilde{w}_j \beta_j \tag{2.21}$$

Considerando os dados $\{(x_i, y_i), \quad i = 1, \ldots, n\}$ n observações do modelo linear

$$y_i = x_i^{'}\beta + \epsilon_i = x_{i1}^{\beta_1} + x_{i2}^{\beta_2} + \epsilon_i, \quad i = 1, \ldots, n \tag{2.22}$$

Com $P(\epsilon_i < 0) = \tau$, $x_i = \left(x_{i1}^{'}, x_{i2}^{'}\right)^{'}$ onde $x_{1i} \in \mathbb{R}^s$, $x_{2i} \in \mathbb{R}^{d-s}$ onde os verdadeiros coeficientes da regressão são $\beta_1 = \beta_{10}$ componentes diferentes de zero e $\beta_2 = \beta_{20} = 0$, resultando então, $\beta_0 = \left(\beta_{10}^{'}, \beta_{20}^{'}\right)^{'}$. O que significa que os s primeiros regressores são estatisticamente relevantes e os $s-d$ são variáveis que apenas agregam ruídos. As hipóteses necessárias são:

- Os erros da regressão ϵ_i são independentes e identicamente distribuída, com τ-ésimo quantil contínua com densidade $f(\cdot)$ positiva na vizinhança de zero.
- O design x_i, $i = 1, \ldots, n$, é uma sequência determinística, para qual existe uma matriz positiva definida $\sum$ tal que

$$\lim_{n \longleftarrow \infty} \frac{\left(\sum_{i=1}^{n} x_i x_i^{'}\right)}{n} = \sum .$$

Os métodos tradicionais de estimação e seleção de variáveis em regressão quantílica envolvem alta intensidade de processamento computacional que, muitas vezes - nem mesmo no contexto amostral -, pode ser implementado para usos cotidianos de processamento. A alternativa para tais situações é reduzir o número de quantis a ser estimado. Essa alternativa adota todos os passos de ordenação de quantis desejados, conforme já descrito, e, em seguida, estima cada quantil (decis, quartis etc.) dos procedimentos clássicos de inferência linear e seleção de variáveis.

Capítulo 3

Aplicação

3.1 Escopo de Aplicação

Descrevemos aqui o experimento realizado com base nas premissas e conceitos estabelecidos, no projeto de pesquisa, para aprimoramento da área de gestão de riscos de crédito. O intuito desse estudo foi realizar, a partir de uma base de clientes de uma instituição financeira brasileira, as etapas para desenvolvimento de modelos preditivos de renda presumida de pessoa física por CPF.

3.2 Base de Dados

Nesta seção, é apresentada a construção da base de dados utilizada para a estimação de renda presumida nos modelos *Behavior* e *Bureau*, no período de dezembro de 2019 - para o processamento da primeira etapa -, e fevereiro de 2020, para a segunda etapa. Foi utilizado o histórico dos seis meses anteriores, para todos os casos em que foi possível obter essa informação.

Os registros dos clientes foram obtidos no cadastro de uma grande IF e em base de dados de órgãos governamentais, com data base de 2018 - respeitando todos os critérios de anonimização dos clientes, previsto na Lei Geral de Proteção de Dados (LGPD).

A base de dados possui total de 520 variáveis - qualitativas e quantitativas, cadastrais e comportamentais -, utilizadas para definir *clusters* e estimar parâmetros da regressão quantílica.

3.2.1 Caracterização da Base de Dados

Base de Dados *Behavior*: base de dados de clientes Pessoas Física com determinadas características comportamentais - excluídos os falecidos ou marcados com indícios de

fraude -, residentes no exterior, menores de 16 anos e situação inativa na receita federal. Para classificar o cliente como *Behavior*, ou seja, aquele com fatores comportamentais suficientes para o modelo, foram estabelecidos os seguintes critérios:

1. Clientes com investimentos abaixo de R$ 170,00 (cento e setenta reais) que:
   (a) não possuem saldo de créditos, nos últimos seis meses;
   (b) saldo no cartão, nos últimos seis meses, igual a zero;
   (c) item saldo médio - na conta corrente -, nos últimos seis meses, igual a zero;
   (d) quantidade de transferências igual a zero; e
   (e) nenhum DOC ou TED realizado nesse período.
2. clientes com saldo médio de conta corrente, nos últimos seis meses, abaixo de R$ 140,00 (cento e quarenta reais) e não possuem todos dos demais citados anteriormente.
3. clientes com saldo no cartão abaixo de R$ 150,00 (cento e cinquenta reais) e sem todos os demais critérios citados anteriormente..

Os casos aplicados aos critérios acima serão considerados clientes *Bureau* e todos os demais serão marcados com *Behavior*.

Após esse tratamento, foi obtida base de dados com 21 milhões de clientes, para estimativa do modelo *Behavior*.

Base de Dados *Bureau*: para o conjunto de informações *Bureau,* foram utilizadas bases de dados de clientes PF não *behavior* da IF, adicionadas de clientes de bases de dados externas - em órgãos parceiros da IF -, selecionando-se o tipo PF - excluídos clientes residentes no exterior, falecidos, marcados como fraude, menores que 16 anos e com CPF em situação inativa na Receita Federal. Desconsiderando os clientes que foram marcados como *Behavior*, chegou-se a uma base de dados com 164 milhões de clientes PF para estimar o modelo *Bureau*.

3.2.2 Tratamento da base de dados e análise descritiva

O público a ser estimado pelo modelo foi definido como pessoas físicas, maiores de 16 anos de idade, com situação do CPF regular na Receita Federal e não residentes no exterior.

Do total de 253,6 milhões de CPFs obtidos, foram excluídos 88,9 milhões de pessoas físicas, entre falecidos, fraudes, residentes no exterior, menores de 16 anos e com CPF irregular na Receita Federal, conforme Tabela 3.1.

Os clientes foram marcados em *Behavior* e *Bureau*, sendo que o modelo *Bureau* também utilizará os clientes marcados como comportamentais.

O processo de classificação dos clientes foi dividido em duas etapas de processamento; na primeira, foram classificados de acordo com variáveis do cadastro da IF, para definir os clientes *Behavio*r e, consequentemente, marcando todos os demais como *Bureau*.

Tabela 3.1: Quantidade de CPF e exclusões.

Menor / CPF Irregular Residente exterior?	Produtor Rural?	Falecido/Fraude?	Qtd. Clientes
Não	Não	Não	164.677.147
Não	Não	Sim	10.095.911
Sim	Não	Não	73.828.273
Sim	Sim	Sim	732.551
Não	Sim	Não	4.311.261
Total			253.645.143

e, para a etapa final, utilizou-se o aprendizado da primeira etapa, ajustando-se a marcação dos clientes, conforme números apresentados na tabela 3.2 abaixo:

Tabela 3.2: Classificação dos clientes.

Tipo do Cliente	Qtd. Clientes
Bureau	143.085.987
Behavior	21.591.160
Total	164.677.147

Entre os clientes utilizados para os modelos, 8,9% recebem proventos na Instituição Financeira, 51,9% são mulheres e 0,01% teve o sexo não informado, conforme tabela 3.3.

Tabela 3.3: Distribuição de Recebedores de Proventos por Sexo.

Recebedor / Sexo	Homem	Mulher	Não Informado	Total
Não	72.536.530	77.489.698	11.975	150.038.203
Sim	6.550.624	8.086.220	2.100	14.638.944
Total	79.087.154	85.575.918	14.075	164.677.147

Clientes com renda atualizada, nos últimos dois anos, representam apenas 22,9% do total de CPFs. Clientes com renda zerada (não informada), 119 milhões, ver tabela 3.4.

Das ocupações informadas pela IF ou pela base externa de fonte de dados, 44% não possuem ocupação ou não têm informação preenchida para o CPF e 5,1% estão cadastrados como aposentados, como pode ser visto na tabela 3.5.

Para determinar a região de localização do CEP foi utilizado o radical do CEP (seus três primeiros dígitos), com os três primeiros dígitos, gerando um total de 940 códigos

Tabela 3.4: Análise da qualidade da variável renda.

Renda Atualizada	Renda Zerada	Qtd. Clientes
Não	Não	7.295.940
Não	Sim	119.576.501
Sim	Não	37.647.247
Sim	Sim	157.459
Total		164.677.147

Tabela 3.5: Análise da Ocupação.

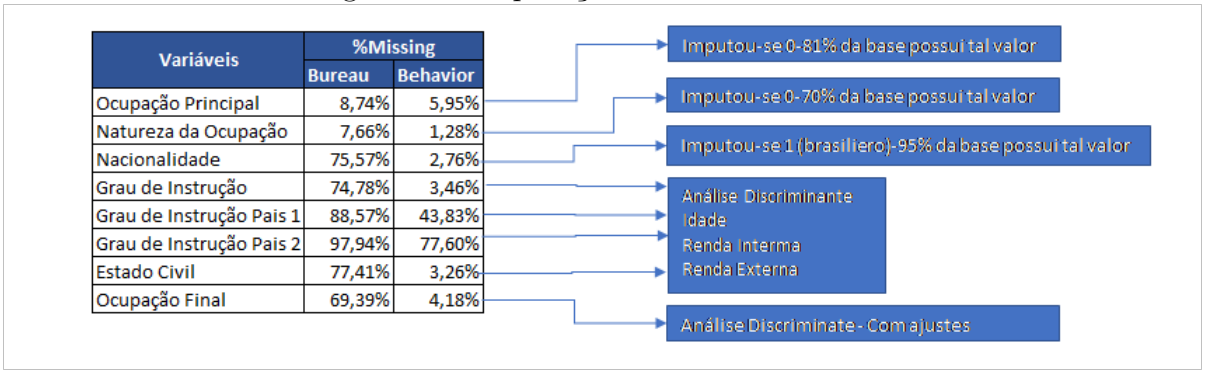
Ocupação	Qtd. Clientes
Sem ocupação	72.526.585
Aposentado	8.420.213
Demais	83.730.349
Total	164.677.147

distintos. O mais concentrado perfaz 0,9% do total de indivíduos e pertence à região de Manaus.

3.2.3 Imputação de Valores Faltantes

Foram imputados valores nas variáveis categóricas necessárias ao processo de clusterização, para que todos os clientes possam ser alocados, devidamente, em seus respectivos grupos semelhantes, Figura 3.1.

Figura 3.1: Imputação de valores faltantes



Fonte: Produção do próprio autor.

No caso do modelo *Behavior*, as variáveis que sofreram imputação de dados foram: estado civil, sexo, ocupação principal, grau de instrução e os três primeiros dígitos do CEP; o modelo *Bureau* usou as mesmas variáveis do modelo comportamental - com exceção do estado civil, pois possui grande quantidade de valores vazios, inviabilizando a imputação dos dados.

Para algumas variáveis imputadas, tais como ocupação na receita federal, código de natureza da ocupação e código de nacionalidade, aplicaram-se valores com grande concentração - acima de 70% - dos dados observados.

Já para as variáveis grau de instrução, grau de instrução do pai e da mãe e código do estado civil, foi aplicada análise de discriminante, utilizando-se, como fatores de imputação, idade, renda cadastrada na IF e renda informada na fonte de dados externa.

3.2.4 Definição da Variável Resposta

Foi definido que a variável resposta do modelo será a Renda Bruta dos clientes por CPF, considerando-se como critério a média dos últimos seis meses, excluídos o maior e o menor casos, adicionando-se a renda constante da fonte de informações em bases externas, no ano vigente; quando não havia essa informação, foram utilizados os dados da RAIS (Relação Anual de Informações Sociais). A renda utilizada segue a distribuição observada na Tabela 3.6 abaixo

Tabela 3.6: Estatísticas Descritivas Renda Bruta.

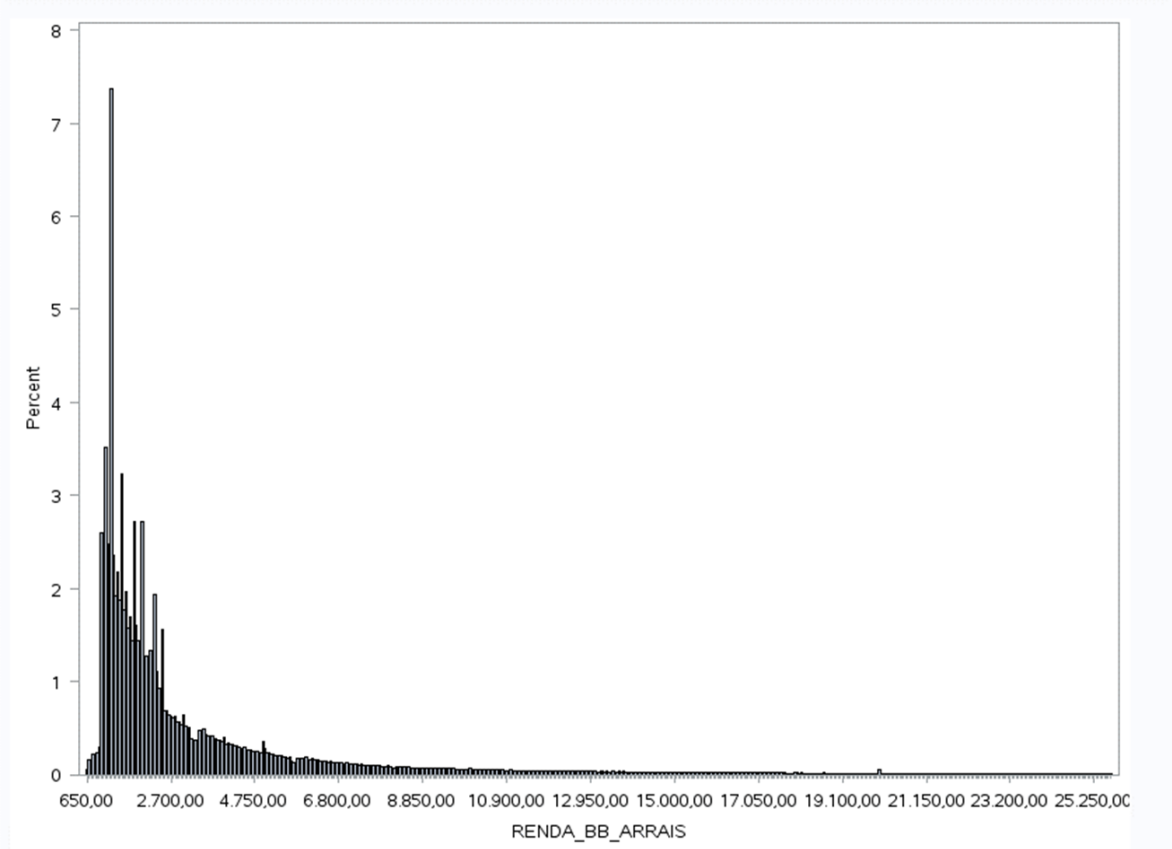
Medidas/Quantis	Renda Observada
Média	3.179
Desvio Padrão	9.226
Coef. Variação	2,902
Máximo	28.036.974
1%	505
5%	956
10%	1.010
25% Q1	1.216
50% Mediana	1.799
75% Q3	3.206
90%	6.745
95%	10.496
99%	19.945

Na distribuição da renda observada, os resultados indicam elevado grau de heterogeneidade e presença de valores extremos. A assimetria à esquerda é uma clara evidência de valores extremos, na distribuição de renda, como mostra a figura 3.2 na página seguinte.

Os resultados sugerem a necessidade de modelos que, primeiramente, possam segmentar essa distribuição ou que possam garantir uma estimação dentro do contexto de superdispersão de dados.

O coeficiente de variação nesse contexto torna-se a variável decisiva para a finalidade de previsão (2,90 ou 290% para a renda de maneira geral). Sem eliminação da superdis-

Figura 3.2: Distribuição da Variável Renda Bruta.



Fonte: Produção do próprio autor.

persão da renda observada, qualquer modelo apresentará viés preditivo, estatisticamente, significativo e não poderá apresentar acurácia aceitável.

Com intuito de evitar a superdispersão dos dados utilizados no modelo, a amostra para clusterização - e respectiva regressão quantílica aplicada - não considerou valores abaixo de R$300,00 (trezentos reais) e acima de R$33.763,00 (trinta e três mil e setecentos e sessenta e três reais - teto do funcionalismo público, aplicado no Brasil, na data em que os dados foram coletados).

Para eliminarmos a superdispersão da variável resposta, o modelo propõe a criação de *clusters* construídos a partir de dados cadastrais e, portanto, que possam ser aplicados tanto para clientes no modelo *Bureau* quanto para clientes no modelo *Behavior*.

Esses dados cadastrais estão divididos em variáveis categóricas e contínuas. Para obtermos *cluster* de variáveis categóricas, é necessária a utilização de método que represente uma distância numérica contínua. O método K-modes, disponível no software SAS *Miner*, utiliza a distância de Jaccard; contudo, tem limitação de uso em uma amostra dos dados, pois, mesmo para bases acima de um milhão de observações, a complexidade computacional de uma matriz com mais de um milhão de registros esgota os recursos de memória,

incorrendo em parada do procedimento.

Uma alternativa é a utilização de componentes principais esparsos para dados binários, Lee [54]. Nessa metodologia, as variáveis categóricas são transformadas em variáveis *dummy* e estimam-se componentes principais esparsos utilizados junto aos componentes principais de variáveis contínuas na estimação dos *clusters* em alta dimensão.

O procedimento de cluster em alta dimensão estende o método K-means para alta dimensão com método centroide mais próximo hierárquico (Anderberg [55]).

Diferentemente do que acontece no método de árvore de regressão, em ambientes ricos de dados, não há, implicitamente, teste estatístico para construção de clusters e, dessa forma, não há como incorrer em erro do tipo I. Esse fato, isoladamente, não evita a existência de missclassification (falsa-classificação). Contudo, existe a garantia de que - mesmo nesses casos -, os grupos de falsas-classificações estarão dentro de um grupo de médias próximas.

Capítulo 4

Solução Proposta

4.1 Resultados Obtidos

O desenvolvimento dos modelos foi dividido em duas etapas, na etapa inicial, os CPFs dos clientes foram classificados como *Behavior* e *Bureau*, considerando-se a classificação do tipo de cadastro da IF. Para marcar o cliente como *behavior*, foram selecionados aqueles que possuíam os seguintes tipos de cadastro: identificação, simplificação, básico, intermediário ou completo. Os demais CPFs foram classificados como *Bureau*.

Ainda na primeira etapa do desenvolvimento dos modelos, houve a necessidade de realizar tratamento na variável resposta renda bruta, considerando-se a renda bruta informada na IF ou nas bases de dados da fonte externa, independentemente, do tempo de atualização das rendas. Apesar de o MAPE dentro de cada cluster da primeira etapa ter sido satisfatório, verificou-se uma considerável quantidade de clientes indevidamente classificados como *Behavior* ou *Bureau*.

Para corrigir essa classificação incorreta, utilizou-se a estimação dos modelos de regressão quantílica nessa etapa inicial, possibilitando a identificação das principais variáveis associadas a diferentes distribuições de renda. Esses resultados preliminares permitiram a correta classificação dos clientes como *Behavior* e *Bureau*.

Na segunda etapa, foi utilizado procedimento de análise de *cluster* em alta dimensão, considerando variáveis categóricas, Lee [54]. A análise de *cluster* tem por objetivo principal a alocação de observações de uma quantidade relativamente pequena de agrupamentos homogêneos internamente e heterogêneos entre si.

Inicialmente, foi gerada amostra aleatória estratificada com um milhão de observações, analisando as variáveis: os três primeiros dígitos do CEP, sexo e ocupação. Uma vez estimados os *clusters* amostrais, os clientes foram escorados, gerando-se marcação única de *cluster* para cada CPF do modelo *Bureau*.

Com essa marcação, foi gerada nova amostra aleatória estratificada, agora considerando o *cluster* como variável de estratificação, em que foi estimada - para cada *cluster* - uma regressão quantílica, consideradas as demais covariáveis possíveis, no contexto ADA-LASSO. Dentro de cada cluster, os decis com os melhores resultados de MAPE são escorados, no cluster populacional. A programação e os resultados das regressões estão apresentados nos apêndices A e B.

Esse mesmo procedimento foi implementado para o CPFs classificados como *Behavior*, a programação e os resultados das regressões podem ser consultados nos apêndices C e D.

A última etapa do desenvolvimento do modelo ocorreu com ajustes na variável resposta, selecionando-se apenas as rendas da IF - atualizadas, há pelo menos 2 anos -, e as rendas da base externa, com posição de 2018. Também foram feitos ajustes no processo de clusterização, retirando-se a variável resposta tanto na geração do *cluster* quanto no ajuste da classificação de cliente *Bureau* e *Behavior*, usando-se todos os resultados dos modelos de regressão quantílica das etapas anteriores.

O objetivo principal de se agrupar os clientes em cluster é reduzir a superdispersão da população e alocar os clientes em blocos com características homogêneas, para, então - na etapa posterior -, utilizar os modelos de regressão quantílica para estimação da renda presumida. Portanto, para uma renda presumida de maior habilidade preditiva é fundamental que os *clusters* gerados tenham coeficiente de variação abaixo de um.

Para o modelo *Bureau*, foi utilizada amostra aleatória estratificada de um milhão de CPFs, considerando-se os o radical do CEP e selecionando-se somente clientes com a variável resposta atualizada e maior que zero.

Após todos os ajustes desta última etapa, novamente os *clusters* amostrais foram obtidos e, posteriormente, escorados na população. Uma nova amostra aleatória estratificada por *cluster* é apresentada e, considerando essa amostra, são estimados modelos de regressão quantílica com todas as demais covariáveis, em contexto ADA-LASSO. Os decis de melhor MAPE são utilizados para escorar o *cluster* populacional.

Neste processo, esses clusters foram escorados na população, gerando clientes marcados como *Bureau* na base total. A etapa final forneceu outra amostra estratificada - por *cluster*, radical do CEP e código da ocupação - com o tamanho de 1.290.597, para uso na etapa de regressão quantílica. Destacamos que essa dimensão da amostra foi necessária para garantir presença suficiente de clientes em todos os estratos amostrais, conforme a Tabela 4.1.

Tabela 4.1: Resultados do Modelo Bureau.

Cluster	MAPE	Clientes Modelagem	Renda Atual	Clientes Total	Renda Presumida					
					Média	Desv. Pad.	Coef. Var.	Min.	Max.	Limite Sup.
1	0,597	91.043	6134,572	289.018	1.833,730	16,633	0,907	1.466,048	2.356,899	2.929,380
2	0,251	7.552.065	1872,826	18.321.699	1.029,648	7,825	0,760	1.014,541	1.337,859	1.288,437
3	0,343	1.113.723	2356,073	11.049.193	1.592,340	3,281	0,206	1.553,897	1.632,881	2.138,611
4	0,534	319.018	5343,875	634.447	2.374,204	32,821	1,382	2.235,466	3.518,811	3.641,952
5	0,477	27.890	4955,669	58.579	1.613,346	15,063	0,934	1.551,161	2.358,503	2.382,275
6	0,494	49.426	6723,463	98.692	1.712,364	4,614	0,269	1.708,923	1.735,029	2.558,700
7	0,385	3.530.460	1739,433	9.355.554	1.014,068	7,351	0,725	966,919	1.073,751	1.404,361
8	0,465	53.655	4730,608	105.690	1.543,519	8,122	0,526	1.483,753	2.199,268	2.261,893
9	0,440	605.211	4288,511	1.449.946	2.354,202	51,579	2,191	2.115,944	3.522,285	3.390,231
10	0,772	139.295	9427,771	313.686	3.458,652	358,156	10,355	2.478,664	5.216,856	6.127,636
11	0,335	166.323	3187,215	413.823	1.455,249	4,851	0,333	1.431,734	1.810,271	1.942,166
12	0,484	1.542.450	5029,092	2.522.011	2.726,343	95,150	3,490	2.473,609	4.085,649	4.044,618
13	0,597	74.230	7379,866	195.794	1.551,894	0,991	0,064	1.549,233	1.559,078	2.478,776
14	0,719	57.464	12324,104	108.677	1.998,789	38,601	1,931	1.894,691	2.387,287	3.435,067
15	0,575	284.747	5388,744	729.966	2.366,077	79,402	3,356	2.111,344	3.534,332	3.726,599
16	0,767	62.113	15432,791	95.240	4.946,843	1.078,704	21,806	3.241,876	7.873,994	8.738,608
17	0,401	1.608.969	2682,802	6.251.691	1.448,665	5,390	0,372	1.386,804	2.120,742	2.029,854
18	0,575	500.891	5433,457	1.337.694	2.565,159	149,462	5,827	2.396,058	3.853,150	4.041,334
19	0,556	758.718	4372,853	1.662.339	3.482,596	55,935	1,606	3.411,858	5.207,860	5.419,921
20	0,626	90.081	7383,343	277.019	1.768,672	4,200	0,237	1.754,057	2.631,791	2.876,592
21	0,836	237	14247,835	747	2.507,560	588,705	23,477	2.093,068	3.885,853	4.603,434
22	0,658	72.166	7695,216	197.789	2.466,447	81,049	3,286	2.300,148	3.694,070	4.088,648
23	0,541	118.445	5420,776	425.943	2.455,530	153,617	6,256	1.354,356	3.698,456	3.783,342
24	0,624	222.567	6987,225	654.609	2.431,077	61,212	2,518	2.247,601	3.640,927	3.947,887
25	0,475	2.582.982	3793,531	6.064.710	2.966,818	57,622	1,942	2.885,982	4.450,047	4.375,939
26	0,561	71.897	6663,285	208.179	1.619,753	23,605	1,457	1.256,537	2.382,322	2.528,394
27	0,446	1.903.967	3089,910	5.305.675	2.525,018	34,304	1,359	2.385,244	3.778,138	3.650,931
28	0,266	7.927.986	1658,400	42.303.512	1.290,784	2,612	0,202	1.091,878	1.405,642	1.634,301
29	0,470	108.081	4456,426	347.770	1.536,138	6,049	0,394	1.483,338	1.580,501	2.258,773
30	0,593	650.280	5550,920	1.235.174	3.020,028	192,651	6,379	2.658,738	4.533,816	4.809,466
31	0,610	62.521	6712,680	119.464	2.208,327	58,330	2,641	1.113,580	3.226,296	3.556,009
32	0,601	54.303	8698,587	91.809	3.009,154	397,984	13,226	2.287,598	4.622,858	4.816,378
33	0,536	471.081	5402,587	1.177.726	2.223,411	65,348	2,939	1.155,262	3.335,311	3.415,953
34	0,595	68.844	6387,111	227.351	1.756,903	8,598	0,489	1.743,861	1.798,391	2.801,747
35	0,302	4.334.751	1962,827	22.594.335	1.486,979	3,687	0,248	1.467,506	2.206,011	1.936,014
36	0,362	3.065.523	2728,085	19.886.715	1.508,508	4,839	0,321	1.430,716	2.228,611	2.053,997
37	0,667	244.269	7762,170	530.113	2.816,596	175,390	6,227	2.529,845	4.236,407	4.694,870
38	0,575	1.507.623	5155,725	2.425.452	2.722,644	91,038	3,344	2.443,868	4.083,077	4.287,084
39	0,591	89.170	7245,669	271.899	1.605,590	21,203	1,321	1.529,938	2.355,449	2.554,247
40	0,588	223.232	6056,665	417.090	2.555,101	222,972	8,727	1.316,737	3.844,882	4.058,282
41	0,462	494.497	5125,111	1.205.449	3.072,380	94,586	3,079	2.904,253	4.605,535	4.493,204
42	0,653	367.595	8182,114	604.370	2.634,357	278,245	10,562	2.207,645	3.975,958	4.354,814
43	0,717	504.434	7135,203	755.117	2.956,192	355,045	12,010	2.173,254	4.440,300	5.074,978
44	0,718	462.225	8470,861	1.171.451	3.149,535	141,585	4,495	2.496,359	4.727,096	5.410,739
45	0,505	65.607	5134,152	209.860	1.368,889	12,231	0,894	1.330,565	1.442,352	2.060,629
46	0,571	330.833	6125,832	686.378	3.113,519	134,626	4,324	2.901,778	4.672,963	4.891,982
47	0,540	105.692	8148,883	204.471	1.946,873	0,000	0,000	1.946,873	1.946,873	2.998,474
Média	0,542	44.738.580	6003,933	164.593.916	2.250,861	112,453	3,817	1.935,385	3.208,818	3.531,862
Média Pond.	0,384									

Fonte: Produção do próprio autor.

Devido ao grande volume de dados na amostra, foi necessário adaptar a regressão quantílica para um contexto de alta dimensão, considerando-se decis. Novamente, foi selecionado o resultado de menor erro percentual médio (MAPE), entres os decis de cada *cluster* amostral e escorado o *cluster* populacional.

Considerando-se o fato de que, no modelo *Behavior*, a variável resposta tem quantidade razoável de informações preenchidas e atualizadas, a variável renda bruta foi incluída na análise de *cluster*, gerando, consequentemente, melhor resultado para os CPFs com renda preenchida e alocando os clientes sem renda informada ou desatualizada dentro de um mesmo *cluster*.

Para o modelo *Behavior*, foi utilizada amostra aleatória estratificada de um milhão de CPFs por radical CEP, com os três primeiros dígitos, selecionando-se somente clientes com a variável resposta maior que zero e atualizada.

O procedimento de *cluster* amostral elimina a superdispersão da renda, permitindo sua escoragem na população, marcando dessa forma os clientes *Behavior* na base total. Um novo procedimento de amostra estratificada foi implementado considerando uma vez mais: radical do CEP, código da ocupação e *cluster* com o tamanho de um milhão, para ser utilizado na etapa de regressão quantílica, Tabela 4.2.

O *cluster* número 64 do modelo *Behavior* foi caracterizado com os clientes comportamentais de renda não informada ou desatualizada e, para esse caso em específico, foi feito mais um processo de modelagem, tratando apenas os clientes alocados nesse *cluster* como uma nova população. O procedimento de *cluster* aplicado a esse *cluster* 64 gerou 33 novos clusters, salientando-se que, para geração desses agrupamentos, não se considerou a renda, haja vista seus valores não constarem na nova população e, portanto, não possuírem efeitos discriminatórios para o caso.

Dado que o tratamento do cluster número 64 é feito para um agrupamento de CPFs sem renda informada ou com renda desatualizada, o acompanhamento da medida do erro percentual médio (MAPE) não é aconselhado para certificar a qualidade da informação final, haja vista não haver renda observável.

Os modelos para estimação de renda presumida (*Bureau* e *Behavior*) partem, inicialmente, do tratamento da superdispersão, por meio de uma análise de *cluster* em alta dimensão. A partir de *cluster* homogêneos, em termos da variável resposta, utilizaram-se modelos de regressão quantílica - adaptados para decis com procedimento de ADA-LASSO de seleção de variáveis dentro em cada decil.

Algumas suposições estão implícitas nessa abordagem: assumiu-se a hipótese esparsa de dados em alta dimensão e, portanto, todas as propriedades de oráculo são admitidas, Tibshirani [45]. Contudo, podem existir outras funções de *linkage* capazes de oferecer maior habilidade preditiva. Aquela limitação poderia ser tratada com diferentes classes

de modelos estimadas, após a geração dos *clusters*, concomitantemente, com algoritmos de aprendizado de máquina para seleção, via *bootstrap*, de melhores previsões como, por exemplo, *LogitBoost* [56] ou AdaBoost [57].

Tabela 4.2: Resultados do Modelo Behavior.

Cluster	MAPE	Clientes	Renda			Clientes	Renda Presumida				
		Modelagem	Média	Desv. Pad.	Coef. Var.	Total	Média	Desv. Pad.	Coef. Var.	Min.	Max.
1	0,0124	149.660	7.971,800	119,039	1,493	149.660	7.972,630	0,000	0,000	7.972,630	7.972,630
2	0,0151	190.898	7.558,735	134,634	1,781	190.898	7.558,785	0,000	0,000	7.558,785	7.558,785
3	0,0075	7.076	32.020,747	285,691	0,892	7.077	31.994,222	0,000	0,000	31.994,222	31.994,222
4	0,0056	12.846	26.365,569	178,756	0,678	12.846	26.355,753	0,000	0,000	26.355,753	26.355,753
5	0,0052	12.107	25.355,652	224,119	0,884	12.107	25.337,538	0,000	0,000	25.337,538	25.337,538
6	0,0129	126.343	10.340,318	155,122	1,500	126.343	10.333,128	3,238	0,031	9.718,984	10.675,195
7	0,0012	33.759	29.907,929	167,126	0,559	33.759	29.940,000	0,000	0,000	29.940,000	29.940,000
8	0,0100	9.562	33.135,200	383,423	1,157	10.448	33.094,717	4,332	0,013	33.042,713	33.454,501
9	0,2593	234.436	748,828	197,363	26,356	240.760	608,622	1,664	0,273	607,997	778,454
10	0,0046	7.394	31.367,314	390,040	1,243	7.394	31.350,000	0,000	0,000	31.350,000	31.350,000
11	0,0459	753.578	1.439,437	79,716	5,538	753.578	1.454,624	0,085	0,006	1.454,601	1.461,372
12	0,0238	355.933	5.637,037	154,054	2,733	355.933	5.626,171	1,420	0,025	5.624,914	5.998,991
13	0,0053	12.825	27.374,855	182,099	0,665	12.825	27.373,948	5,776	0,021	27.335,805	27.594,187
14	0,0220	427.048	4.696,404	120,057	2,556	427.048	4.697,506	3,772	0,080	4.256,706	6.362,461
15	0,0189	287.561	6.643,450	145,234	2,186	287.561	6.645,360	0,000	0,000	6.645,360	6.645,360
16	0,0066	25.495	19.533,217	154,922	0,793	25.495	19.527,560	0,000	0,000	19.527,560	19.527,560
17	0,0082	44.983	17.515,562	168,318	0,961	44.983	17.518,930	1,801	0,010	17.467,313	17.711,945
18	0,0074	40.118	18.049,875	156,664	0,868	40.118	18.054,392	0,000	0,000	18.054,392	18.054,392
19	0,0054	11.715	26.858,324	178,664	0,665	11.715	26.865,095	0,000	0,000	26.865,095	26.865,095
20	0,0204	339.020	6.141,404	144,540	2,354	339.020	6.140,380	0,000	0,000	6.140,380	6.140,380
21	0,0189	361.473	3.704,719	85,835	2,317	361.473	3.708,908	17,323	0,467	3.167,383	5.497,140
22	0,0089	68.892	13.476,986	142,984	1,061	68.892	13.481,228	2,633	0,020	13.264,094	13.484,514
23	0,0326	732.430	2.992,629	114,064	3,811	732.430	2.994,841	2,119	0,071	2.594,046	3.957,651
24	0,0070	24.653	20.559,751	170,703	0,830	24.653	20.557,244	0,802	0,004	20.513,763	20.569,189
25	0,0122	139.639	9.303,130	136,154	1,464	139.639	9.300,002	0,755	0,008	9.201,132	9.300,403
26	0,0090	59.350	14.466,679	151,302	1,046	59.350	14.470,215	0,739	0,005	14.458,168	14.470,476
27	0,0391	860.138	2.591,812	117,241	4,524	860.138	2.593,271	0,000	0,000	2.593,271	2.593,271
28	0,0076	13.967	29.289,545	268,164	0,916	13.967	29.283,245	0,000	0,000	29.283,245	29.283,245
29	0,0061	17.115	23.234,763	169,786	0,731	17.115	23.229,325	1,089	0,005	23.195,719	23.231,492
30	0,0149	155.315	9.799,764	168,503	1,719	155.315	9.806,671	0,000	0,000	9.806,671	9.806,671
31	0,0083	56.589	14.963,670	144,946	0,969	56.589	14.974,336	2,782	0,019	14.902,696	15.045,792
32	0,0080	47.664	15.937,584	150,298	0,943	47.664	15.944,604	0,935	0,006	15.939,662	16.132,281
33	0,0097	76.012	13.005,107	149,036	1,146	76.012	13.000,769	0,000	0,000	13.000,769	13.000,769
34	0,0083	47.504	16.958,028	165,326	0,975	47.504	16.960,321	0,200	0,001	16.960,200	16.969,893
35	0,9219	1.073	30.562,153	8.278,731	27,088	118.079	3.332,277	0,000	0,000	3.332,277	3.332,277
36	0,0067	31.399	19.041,063	152,357	0,800	31.399	19.031,800	0,000	0,000	19.031,800	19.031,800
37	0,0273	504.492	3.415,453	110,602	3,238	504.492	3.424,384	2,248	0,066	3.254,788	4.088,978
38	0,0070	24.213	21.118,513	174,328	0,825	24.213	21.115,749	12,928	0,061	19.115,465	21.115,980
39	0,0542	1.251.243	2.183,896	126,105	5,774	1.251.243	2.228,918	0,000	0,000	2.228,918	2.228,918
40	0,0100	89.524	12.513,604	148,796	1,189	89.524	12.512,353	0,000	0,000	12.512,353	12.512,353
41	0,0056	13.824	28.608,821	279,290	0,976	13.824	28.620,000	0,000	0,000	28.620,000	28.620,000
42	0,0477	883.724	1.657,193	95,913	5,788	883.724	1.644,414	0,000	0,000	1.644,414	1.644,414
43	0,0064	25.343	20.033,314	154,779	0,773	25.343	20.023,398	0,000	0,000	20.023,398	20.023,398
44	0,0083	50.756	15.446,285	149,017	0,965	50.756	15.444,190	0,000	0,000	15.444,190	15.444,190
45	0,0122	150.433	8.390,924	119,849	1,428	150.433	8.393,187	0,000	0,000	8.393,187	8.393,187
46	0,0348	1.506.874	1.023,583	70,069	6,845	1.506.874	1.010,734	0,000	0,000	1.010,734	1.010,734
47	0,0057	16.553	23.738,338	166,684	0,702	16.553	23.742,813	4,864	0,020	23.599,686	24.296,039
48	0,0194	405.297	3.972,405	90,709	2,283	405.297	3.984,542	2,183	0,055	3.028,167	3.984,631
49	0,0064	20.005	22.221,079	169,158	0,761	20.005	22.230,468	2,807	0,013	22.186,727	22.314,261
50	0,0092	67.493	13.964,317	150,506	1,078	67.493	13.968,907	0,000	0,000	13.968,907	13.968,907
51	0,0133	123.189	10.884,449	168,506	1,548	123.189	10.884,698	19,013	0,175	6.518,859	10.885,156
52	0,0160	227.254	7.121,613	133,825	1,879	227.254	7.116,591	0,000	0,000	7.116,591	7.116,591
53	0,0117	105.589	11.985,068	164,098	1,369	105.589	11.988,164	0,000	0,000	11.988,164	11.988,164
54	0,0070	36.868	18.559,963	152,613	0,822	36.868	18.564,323	3,748	0,020	18.360,144	19.218,829
55	0,0068	23.331	21.680,362	174,473	0,805	23.331	21.691,531	0,000	0,000	21.691,531	21.691,531
56	0,0062	18.535	22.734,283	167,388	0,736	18.535	22.739,687	0,000	0,000	22.739,687	22.739,687
57	0,0061	15.569	30.115,926	248,778	0,826	15.569	29.980,534	2,717	0,009	29.873,028	29.989,740
58	0,0813	1.059.303	1.186,460	116,548	9,823	1.063.696	1.199,234	0,000	0,000	1.199,234	1.199,234
59	0,0119	105.369	11.433,510	158,147	1,383	105.369	11.436,840	0,000	0,000	11.436,840	11.436,840
60	0,0053	17.332	31.204,466	248,784	0,797	17.332	31.329,335	5,079	0,016	30.980,322	31.329,924
61	0,0082	46.945	16.433,491	157,124	0,956	46.945	16.432,333	0,000	0,000	16.432,333	16.432,333
62	0,0064	12.753	27.978,179	216,364	0,773	12.753	27.967,407	0,000	0,000	27.967,407	27.967,407
63	0,0243	466.709	5.144,863	144,225	2,803	466.709	5.152,906	5,089	0,099	3.755,125	5.876,067
64	0,3215	98.703	778,023	341,390	43,879	6.497.035	590,873	1,366	0,231	297,002	841,605
65	0,0068	19.150	24.846,180	199,770	0,804	19.150	24.862,366	0,000	0,000	24.862,366	24.862,366
66	0,0063	16.861	25.841,812	193,278	0,748	16.861	25.850,680	0,000	0,000	25.850,680	25.850,680
67	0,0232	451.238	4.296,147	115,669	2,692	451.238	4.297,894	1,408	0,033	3.852,930	4.481,421
68	0,0059	17.190	24.261,658	175,866	0,725	17.190	24.263,939	0,000	0,000	24.263,939	24.263,939
69	0,0387	1.263.178	1.849,423	100,740	5,447	1.263.181	1.841,084	7,577	0,412	1.205,208	2.749,017
70	0,0139	153.805	8.835,954	142,982	1,618	153.805	8.836,399	0,000	0,000	8.836,399	8.836,399
Média	0,0356	15.064.215	15.200,123	283,020	3,132	21.591.160	14.806,990	1,750	0,032		
Média Pond.	0,0395										

Fonte: Produção do próprio autor.

4.2 Comparação dos Resultados dos Modelos

Aqui realiza-se comparação dos resultados encontrados nos modelos de renda presumida desenvolvidos e customizados para a Instituição Financeira com aqueles externos, adquiridos de dois grandes *Bureaus* de crédito, no mercado brasileiro.

Para realizar essa comparação, foi utilizado o mesmo agrupamento encontrado nos *clusters* estimados, na etapa de modelagem, atribuída a renda informada por CPF pelos dois Bureaus externos e calculada a estatística MAPE, dentro de cada *cluster* por modelo desenvolvido.

Ao comparar o MAPE do modelo Bureau desenvolvido para os clientes sem relacionamento com a instituição financeira com aquele dos dois modelos externos adquirido no mercado, nota-se que, em média, o modelo Bureau da IF possui o menor erro percentual absoluto médio ponderado (0,384). Entretanto, o modelo externo 2 consegue presumir a renda com menor erro, na maioria dos *clusters* de clientes classificados como *Bureau* e renda superior a R$ 4 mil, como pode ser observado na Tabela 4.3.

Tabela 4.3: Comparação dos MAPE do Modelo Bureau com os modelos externos de mercado.

Cluster	**Bureau**	**Modelo 1**	**Modelo 2**	**Cluster**	**Bureau**	**Modelo 1**	**Modelo 2**
1	**0,597**	0,641	0,678	26	0,561	0,602	**0,551**
2	**0,251**	0,483	0,529	27	0,446	0,545	**0,438**
3	**0,343**	0,615	0,513	28	**0,266**	0,411	0,537
4	**0,534**	0,539	0,462	29	**0,470**	0,555	0,558
5	**0,477**	0,503	0,636	30	0,593	0,533	**0,474**
6	**0,494**	0,545	0,537	31	0,610	0,581	**0,495**
7	**0,385**	0,481	0,586	32	0,601	0,585	**0,473**
8	**0,465**	0,569	0,525	33	0,536	0,576	**0,476**
9	**0,440**	0,565	0,444	34	0,595	0,596	**0,520**
10	0,772	0,569	**0,533**	35	**0,302**	0,571	0,503
11	**0,335**	0,432	0,564	36	**0,362**	0,488	0,466
12	0,484	0,564	**0,451**	37	0,667	**0,597**	0,608
13	0,597	0,574	**0,573**	38	0,575	0,573	**0,534**
14	0,719	0,598	**0,587**	39	**0,591**	0,617	0,596
15	0,575	0,553	**0,498**	40	**0,588**	0,598	0,621
16	0,767	**0,495**	0,563	41	0,462	0,557	**0,450**
17	**0,401**	0,615	0,549	42	0,653	0,563	**0,544**
18	0,575	0,709	**0,505**	43	0,717	0,508	**0,480**
19	0,556	0,531	**0,437**	44	0,718	**0,626**	**0,626**
20	0,626	0,626	**0,560**	45	**0,505**	0,585	0,573
21	0,836	0,653	0,631	46	0,571	0,565	**0,496**
22	0,658	0,600	**0,540**	47	0,540	0,558	**0,520**
23	0,541	0,592	**0,467**				
24	**0,624**	0,652	0,672	**média**	**0,542**	**0,566**	**0,532**
25	0,475	0,513	**0,435**	**média pond.**	**0,384**	**0,513**	**0,513**

Fonte: Produção do próprio autor.

Ao comparar o MAPE do modelo *Behavior* desenvolvido para os clientes com relacionamento na instituição financeira com o MAPE dos dois modelos externos adquiridos no mercado, percebe-se, nitidamente, que a acurácia dos modelos desenvolvidos nos 70 *clusters* apresenta o menor erro percentual absoluto médio e, consequentemente, a menor média ponderada (Tabela 4.4).

Tabela 4.4: Comparação dos MAPE do Modelo Behavior com os modelos externos de mercado.

Cluster	Behavior	Modelo 1	Modelo 2	Cluster	Behavior	Modelo 1	Modelo 2
1	0,012	0,583	0,362	37	0,027	0,621	0,427
2	0,015	0,582	0,358	38	0,007	0,648	0,665
3	0,008	0,737	0,771	39	0,054	0,616	0,463
4	0,006	0,660	0,725	40	0,010	0,608	0,484
5	0,005	0,671	0,715	41	0,006	0,684	0,740
6	0,013	0,589	0,397	42	0,048	0,627	0,472
7	0,001	0,673	0,745	43	0,006	0,640	0,651
8	0,010	0,735	0,778	44	0,008	0,614	0,568
9	0,259	1,135	1,588	45	0,012	0,592	0,366
10	0,005	0,700	0,755	46	0,035	0,750	0,644
11	0,046	0,597	0,494	47	0,006	0,657	0,697
12	0,024	0,581	0,368	48	0,019	0,616	0,423
13	0,005	0,655	0,732	49	0,006	0,661	0,678
14	0,022	0,593	0,374	50	0,009	0,603	0,529
15	0,019	0,582	0,356	51	0,013	0,595	0,424
16	0,007	0,635	0,639	52	0,016	0,582	0,355
17	0,008	0,602	0,605	53	0,012	0,605	0,469
18	0,007	0,599	0,613	54	0,007	0,608	0,623
19	0,005	0,668	0,727	55	0,007	0,650	0,670
20	0,020	0,584	0,363	56	0,006	0,656	0,685
21	0,019	0,608	0,397	57	0,006	0,712	0,758
22	0,009	0,602	0,513	58	0,081	0,560	0,590
23	0,033	0,627	0,432	59	0,012	0,596	0,449
24	0,007	0,636	0,654	60	0,005	0,740	0,767
25	0,012	0,576	0,354	61	0,008	0,606	0,587
26	0,009	0,608	0,544	62	0,006	0,685	0,738
27	0,039	0,628	0,435	63	0,024	0,583	0,364
28	0,008	0,680	0,745	64	0,322	0,908	1,342
29	0,006	0,664	0,693	65	0,007	0,664	0,712
30	0,015	0,593	0,375	66	0,006	0,666	0,720
31	0,008	0,613	0,556	67	0,023	0,600	0,389
32	0,008	0,602	0,575	68	0,006	0,658	0,702
33	0,010	0,605	0,504	69	0,039	0,546	0,473
34	0,008	0,602	0,593	70	0,014	0,586	0,363
35	0,922	0,783	0,775	Média	0,036	0,642	0,589
36	0,007	0,622	0,636	Média Pond.	0,039	0,624	0,498

Fonte:Produção do próprio autor.

4.3 Uso da Renda Presumida no ecossistema do *Open Banking*

O uso da renda presumida no ecossistema do *Open Banking* se dá, principalmente, na gestão do risco de crédito e envolve os seguintes processos: atualização cadastral automatizada, análise do risco de crédito dos clientes, cálculo da capacidade de pagamento e aferição ou ajuste nos limites de crédito.

A renda presumida é utilizada como informação cadastral pela maioria das instituições financeiras e, em alguns casos, há a dispensa de comprovação de renda - o que traz eficiência operacional - e melhorias no processo de análise de crédito e, consequentemente, na experiência do cliente junto à instituição.

De maneira prática, ela permite confirmar as informações apresentadas pelo cliente, comparando-se o valor declarado com o da renda presumida. Ela permite utilizar-se de algoritmos automatizados para apurar a veracidade dos dados, definir limites e até dispensar a apresentação do documento de comprovação de renda.

Instituições financeiras também usam a renda presumida para validar a informação que o cliente compartilha, no ambiente do *Open Banking*, evitando fraudes cadastrais ou de subscrição.

Outra utilização da renda presumida no *Open Banking* é na análise do risco de crédito do cliente. Geralmente, essa informação é utilizada como covariável nos modelos de *application score*, utilizados para classificar o risco do cliente, no momento da contratação de uma operação de crédito.

Ela pode ser utilizada de forma direta ou combinada com outras variáveis cadastrais/comportamentais, com o propósito de aumentar o poder discriminatório de classificação dos clientes bons e ruins.

Com o avanço dos algoritmos de aprendizado de máquina, o processo de análise de crédito evoluiu muito, ao longo dos anos. Até pouco tempo, era comum a instituição financeira solicitar contracheque (*holerite*) do proponente ao crédito, a fim de comprovar se ele de fato possuía atividade remunerada e qual seria sua capacidade de pagamento. Além disso, a conferência da existência de restrições junto a órgãos de proteção de crédito era obrigatória. Atualmente, para que se possa dar respostas online ao cliente, todo este processo ocorre de forma automatizada e em poucos segundos. Isso é primordial no ambiente de *Open Banking*, em se é exigido dar ao cliente respostas online.

Outro processo que utiliza bastante a renda presumida é a análise da Capacidade de Pagamento dos Cientes - CAPAG ou, em caso de crédito parcelado, a Capacidade Mensal de Pagamento - CMP. Ocorrer com certa frequência, principalmente, para clientes desbancarizados, na informalidade ou muito jovens. Essas pessoas possuem um histórico

de informações muito "pobre", o que torna a aderência da Renda Presumida, no modelo estatístico, extremamente baixa.

Assim, buscar alternativas no mercado ou desenvolver modelos proprietários, customizados para os seus clientes, é uma das principais saídas para IFs garantir que a Renda Presumida gerada internamente e a real situação de um proponente ao crédito não sejam discrepantes. Do contrário, uma empresa pode, simplesmente, negar o crédito e perder, para a concorrência, um cliente com grande potencial; no competitivo ambiente do *Open Banking*, esse seria um prejuízo altamente significativo.

A renda presumida é bastante utilizada como variável para aferir limites de crédito. Ela está relacionada ao poder de compra do cliente, ajudando a estabelecer a capacidade de pagamento do indivíduo, no momento de compartilhamento de dados no *Open Banking*, e pode ser utilizada na revisão periódica dos limites de crédito, visando ao direcionamento e à personalização de novas ofertas a clientes com perfil adequado e ao menor risco de inadimplência.

Além das situações mencionadas, a renda presumida pode ser utilizada na recuperação do crédito, permitindo calcular a capacidade de pagamento dos devedores, otimizando esforços e direcionando, mais adequadamente, acordos de pagamento e negociações das dívidas vencidas.

Capítulo 5

Conclusão

A renda presumida constitui, no Brasil, modelo bastante comercializado por *Bureaus* de crédito. Ela é bastante utilizada por instituições financeiras e *fintechs*, na gestão do risco de crédito. Por isso, a área de pesquisa que une *big data* e modelagem estatística tem atraído a atenção da indústria financeira e da academia.

Em tal contexto, este trabalho fez proposta de modelagem da renda presumida, utilizando dados reais de instituição financeira brasileira, abordagem da modelagem estatística clássica, aplicação da análise de *clusters* e regressão quantílica, para fazer as estimativas das rendas dos clientes.

A modelagem estatística estende-se, da análise ao tratamento de dados, passando pela imputação de valores faltantes e delineamento de agrupamento por técnica de análise de *clusters*, concluindo pelas estimativas de parâmetros por meio da regressão quantílica.

Este estudo descreveu funcionalidades e propostas de soluções para estimativa de renda presumida, adicionando casos de uso prático.

Como resultado, foram desenvolvidos dois grupamentos de modelos denominados modelo *Behavior* (para os clientes que já se relacionavam com a IF) e modelo *Bureau* (para clientes sem relacionamento com a IF) que demonstraram seu poder preditivo de mensurar a capacidade de pagamento dos clientes, por meio da renda presumida.

As características dos modelos desenvolvidos, a partir da técnica proposta e aplicados em dados reais, demonstraram as diferenças relevantes entre o poder preditivo entre as estimativas da renda presumida dos clientes classificados como *behavior* e *bureau*, demonstrando a importância da segregação desses grupos antes do desenvolvimento da modelagem.

Os modelos desenvolvidos, *behavior* e *buereau* , utilizam-se de covariáveis que estão contempladas no *book* de dados a serem compartilhadas pelos clientes no ambiente do *Open Banking*, sendo suficiente para estimar a renda presumida.

Dado esse escopo, as contribuições desta dissertação são:

1. formalização de modelagem estatística, utilizando-se técnica de regressão quantílica, para presumir a renda dos clientes pessoa física;

2. utilização da renda presumida, na gestão do risco de crédito em IFs e *Fintechs*;

3. alternativas ao uso de modelos de renda presumida de mercado, comercializados sem qualquer customização paras IFs.

4. implementação dos modelos apresentados em plataforma analítica especializada em *Open Banking*.

5.1 Trabalhos Futuros

A área de gestão do risco de crédito e da modelagem estatística para presumir renda de clientes possui diversos desafios ainda abertos, identificados durante a revisão bibliográfica e o desenvolvimento da prova de conceito desta dissertação.

Vez que a proposta descrita nesta dissertação tem escopo e restrições bem definidos, não foram incluídas tantas outras possibilidades de aplicação da renda presumida, como, por exemplo, prevenção a fraudes, eficiência operacional e funcionalidades para tarefas mais específicas e regulatórias. Assim, incluem-se como possíveis evoluções da modelagem proposta neste trabalho:

1. existência de outras funções de *linkage* que podem oferecer maior habilidade preditiva dos modelos. Esse tema poderia ser tratado com outras classes de modelos, estimadas após a geração dos *clusters*, concomitantemente, a algoritmos de aprendizado de máquina para seleção, via *bootstrap*, de melhores previsões, como, por exemplo LogitBoost [56] ou AdaBoost [57];

2. desenvolvimento de modelos híbridos, combinando os modelos de renda presumida proprietários com os modelos de mercado dos *Bureaus* de crédito, objetivando-se alcançar um aperfeiçoamento nas estimativas dos clientes que não possuem relacionamento com a IFs.

3. estimativas de faturamento presumido para micro e pequenas empresas.

Ainda, como trabalho futuro, a própria modelagem proposta pode ter módulos automatizados na otimização dos *clusters* e das estimativas das regressões aplicados, em ambiente *data lake* de plataforma analítica. Neste caso, poder-se-ia lançar mão de técnicas de *machine learning* para se determinar a forma mais performática de criar serviços, combinando-se diversas tecnologias disponíveis neste ambiente.

Referências

[1] CMN: *Resolução Conjunta Nº 3, de 24 de junho de 2021 (que a altera a Resolução conjunta Bacen/CMN nº 1 de 04 de maio de 2020, que dispõe sobre a implementação do Sistema Financeiro Aberto (Open Banking).* 2

[2] CMN: *Resolução Conjunta Nº 1, de 04 de maio de 2020 (Dispõe sobre a implementação do Sistema Financeiro Aberto (Open Banking).* 2, 6

[3] Siddiqi, Naeem: *Credit risk scorecards: developing and implementing intelligent credit scoring*, volume 3. John Wiley & Sons, 2012. 2

[4] Veiga, Fábio Da Silva, Sandro Mansur Gibran e Silvana Fátima Mezaroba Bonsere: *Open banking: Expectativas e desafios para o mercado financeiro no brasil.* Administração de Empresas em Revista, 1(15):203–226, 2020. 5

[5] Goettenauer, Carlos: *Open banking e o modelo de banco em plataforma: a necessidade de reavaliação da definição jurídica de atividade bancária.* Revista da Procuradoria-Geral do Banco Central, 14(1):13–27, 2020. 7

[6] Cavalcante, Eric Jardim: *O novo paradigma tecnológico do setor financeiro nacional: a implantação do open banking no brasil.* 2021. 7, 8

[7] Fisher, Ronald A: *The use of multiple measurements in taxonomic problems.* Annals of eugenics, 7(2):179–188, 1936. 9

[8] Thomas, Lyn C: *A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers.* International journal of forecasting, 16(2):149–172, 2000. 9

[9] Kang, S e Kyung shik Shin: *Customer credit scoring model using analytic hierarchy process.* Informs & Korms, Seoul, páginas 2197–2204, 2000. 9

[10] Markowitz, Harry M: *Portfolio selection.* Yale university press, 1968. 9

[11] Black, Fischer e Myron Scholes: *The pricing of options and corporate liabilities.* Em *World Scientific Reference on Contingent Claims Analysis in Corporate Finance: Volume 1: Foundations of CCA and Equity Valuation*, páginas 3–21. World Scientific, 2019. 9

[12] LAWRENCE, David B: *Risco e recompensa: o negócio de crédito ao consumidor.* Traduçao de Debbie McKey. Nova York: Individual Bank, Citicorp, 1984. 9

[13] Diniz, Carlos e Francisco Louzada: *Métodos estatısticos para análise de dados de crédito*. Em *6th Brazilian Conference on Statistical Modeling in Insurance and Finance, Maresias-SP*, 2013. 9, 11, 13, 14, 15, 16, 17

[14] CMN: *Resolução 4.557 de 23 de fevereiro de 2017 (Dispõe sobre a estrutura de gerenciamento de riscos, a estrutura de gerenciamento de capital e a política de divulgação de informações*. 10

[15] Brito, Giovani Antonio Silva e Alexandre Assaf Neto: *Modelo de classificação de risco de crédito de empresas*. Revista Contabilidade & Finanças, 19:18–29, 2008. 10, 11

[16] CMN: *Resolução 3.721 de 2009 (Estrutura de Gerenciamento de Risco de Crédito)*. 10

[17] Yanaka, Guilherme M: *Ensaios em gestão de risco e regulação bancária*. Tese de Doutoramento, 2014. 10

[18] Sicsú, Abraham Laredo: *Credit Scoring: desenvolvimento, implantação, acompanhamento*. Blucher, 2010. 11, 12

[19] Ohtoshi, Claudia: *Uma comparação de regressão logística, árvores de classificação e redes neurais: analisando dados de crédito*. Tese de Doutoramento, Universidade de São Paulo, 2003. 12

[20] Thomas, Lyn C, Dabid B Edelman e Jonathan N Crook: *Credit scoring and its applications: Siam monographs on mathematical modeling and computation*. Philadelphia: University City Science Center, SIAM, 2002. 13

[21] Lewis, Edward M: *An introduction to credit scoring*. Fair, Isaac and Company, 1992. 14

[22] Zweig, Mark H e Gregory Campbell: *Receiver-operating characteristic (roc) plots: a fundamental evaluation tool in clinical medicine*. Clinical chemistry, 39(4):561–577, 1993. 17, 18

[23] Frias-Martinez, Vanessa e Jesus Virseda: *On the relationship between socio-economic factors and cell phone usage*. Em *Proceedings of the fifth international conference on information and communication technologies and development*, páginas 76–84, 2012. 18

[24] Mourão, Roberto Nunes: *Mineração de dados para previsão de renda de clientes com contas-correntes digitais*. 2018. 19

[25] Júnior, Valter E Silva, Renata MCR Souza, Getúlio JA Amaral e Hélio G Souza Júnior: *Estimation methods of presumed income*. Em *International Conference on Neural Information Processing*, páginas 235–241. Springer, 2013. 19

[26] Kibekbaev, Azamat e Ekrem Duman: *Benchmarking regression algorithms for income prediction modeling*. Information Systems, 61:40–52, 2016. 19

[27] Swan, Neil: *Problems in dynamic modeling of individual incomes.* Em *Swedish Conference on Microsimulation, Stockholm*, 2006. 20

[28] Carriere, Jacques F e Kevin J Shand: *New salary functions for pension valuations.* North American Actuarial Journal, 2(3):18–26, 1998. 20

[29] Bone, Christopher M e Olivia S Mitchell: *Building better retirement income models.* North American Actuarial Journal, 1(1):1–10, 1997. 20

[30] Lazar, Alina: *Income prediction via support vector machine.* Em *ICMLA*, páginas 143–149. Citeseer, 2004. 20

[31] Yamnampet, Ghatkesar: *Comparative analysis of classification models on income prediction.* International Journal on Recent and Innovation Trends in Computing and Communication, 5(4):451–455. 20

[32] Lessmann, Stefan, Bart Baesens, Hsin Vonn Seow e Lyn C Thomas: *Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research.* European Journal of Operational Research, 247(1):124–136, 2015. 21

[33] Sundsøy, Pål, Johannes Bjelland, Bjørn Atle Reme, Asif M Iqbal e Eaman Jahani: *Deep learning applied to mobile phone data for individual income classification.* Em *Proceedings of the 2016 International Conference on Artificial Intelligence: Technologies and Applications, Bangkok, Thailand*, páginas 24–25, 2016. 21

[34] Koenker, Roger e Gilbert Bassett Jr: *Regression quantiles.* Econometrica: journal of the Econometric Society, páginas 33–50, 1978. 21, 22

[35] Koenker: *Quantile Regression (Econometric Society Monographs; No. 38).* Cambridge university press, 2005. 21, 25, 26, 27, 28, 29

[36] Santos, Bruno Ramos dos: *Modelos de regressão quantílica.* Tese de Doutoramento, Universidade de São Paulo, 2012. 21, 22, 24, 28

[37] Montgomery, Douglas C, Elizabeth A Peck e G Geoffrey Vining: *Introduction to linear regression analysis.* John Wiley & Sons, 2021. 21

[38] Rao, Calyampudi Radhakrishna: *Linear statistical inference and its applications*, volume 2. Wiley New York, 2009. 21

[39] Box, George EP e David R Cox: *An analysis of transformations.* Journal of the Royal Statistical Society: Series B (Methodological), 26(2):211–243, 1964. 22

[40] Davino, Cristina, Marilena Furno e Domenico Vistocco: *Quantile regression: theory and applications*, volume 988. John Wiley & Sons, 2013. 26

[41] Barrodale, Ian e Frank DK Roberts: *An improved algorithm for discrete l_1 linear approximation.* SIAM Journal on Numerical Analysis, 10(5):839–848, 1973. 26

[42] Koenker, Roger W e Vasco d'Orey: *Algorithm as 229: Computing regression quantiles.* Applied statistics, páginas 383–393, 1987. 26

[43] Portnoy, Stephen e Roger Koenker: *The gaussian hare and the laplacian tortoise: computability of squared-error versus absolute-error estimators*. Statistical Science, 12(4):279–300, 1997. 26

[44] Chen, Colin e Ying Wei: *Computational issues for quantile regression*. Sankhyā: The Indian Journal of Statistics, páginas 399–417, 2005. 26

[45] Tibshirani, Robert: *Regression shrinkage and selection via the lasso*. Journal of the Royal Statistical Society: Series B (Methodological), 58(1):267–288, 1996. 29, 41

[46] Fan, Jianqing e Runze Li: *Variable selection via nonconcave penalized likelihood and its oracle properties*. Journal of the American statistical Association, 96(456):1348–1360, 2001. 29

[47] Zou, Hui: *The adaptive lasso and its oracle properties*. Journal of the American statistical association, 101(476):1418–1429, 2006. 29, 30

[48] Zou, Hui e Ming Yuan: *Composite quantile regression and the oracle model selection theory*. The Annals of Statistics, 36(3):1108–1126, 2008. 29

[49] Zhang, Hao Helen e Wenbin Lu: *Adaptive lasso for cox's proportional hazards model*. Biometrika, 94(3):691–703, 2007. 29

[50] Huang, Jian, Shuangge Ma e Cun Hui Zhang: *Adaptive lasso for sparse high-dimensional regression models*. Statistica Sinica, páginas 1603–1618, 2008. 29

[51] Lin, Zhengyan, Yanbiao Xiang e Caiya Zhang: *Adaptive lasso in high-dimensional settings*. Journal of Nonparametric Statistics, 21(6):683–696, 2009. 29

[52] Koenker, Roger: *Quantile regression for longitudinal data*. Journal of Multivariate Analysis, 91(1):74–89, 2004. 29

[53] Wu, Yichao e Yufeng Liu: *Variable selection in quantile regression*. Statistica Sinica, páginas 801–817, 2009. 29

[54] Lee, Seokho, Jianhua Z Huang e Jianhua Hu: *Sparse logistic principal components analysis for binary data*. The annals of applied statistics, 4(3):1579, 2010. 37, 38

[55] Anderberg, Michael R: *The broad view of cluster analysis*. Cluster analysis for applications, páginas 1–9, 1973. 37

[56] Friedman, Jerome, Trevor Hastie e Robert Tibshirani: *Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors)*. The annals of statistics, 28(2):337–407, 2000. 42, 49

[57] Freund, Yoav, Robert E Schapire *et al.*: *Experiments with a new boosting algorithm*. Em *icml*, volume 96, páginas 148–156. Citeseer, 1996. 42, 49

Apêndice A

Modelo *Bureau*

```
%MACRO Z;
%DO J=1 %TO 10;

DATA QUANTIL&J;
RETAIN RENDA_BB_ARRAIS;
SET MODEL_BUR_CLUS&i;
IF _N_ > (&TOT/10)*(&J-1) AND _N_ <= (&TOT/10)*&J;
RUN;

ODS OUTPUT PARAMETERESTIMATES = DEMAIS.PAR\_SEL\_CL&I.\_D&J;
PROC GLMSELECT DATA = QUANTIL&J TESTDATA = DEMAIS.VALID_BUR_CLUS&i;
     CLASS CD_SEXO CEP_SRF_3 COD_OCUP_FINAL;
     MODEL RENDA_BB_ARRAIS = IDADE_SRF
     ANOT_SERASA_EXNO ANOT_SCPC_EXNO ANOT_CADIN_EXNO RENDA_MEDIA_CONJUGE
     RENDA_MEDIA_FILHO RENDA_MEDIA_PAIS RENDA_MEDIA_AVO RENDA_MEDIA_IRMAOS
     RENDA_MEDIA_DEMAIS RESTITUICAO_ANT1 RESTITUICAO_ANT2 RESTITUICAO_ANT3
     RESTITUICAO_ANT4 RESTITUICAO_ANT5 RESTITUICAO_ANT6
     CD_SEXO CEP_SRF_3 COD_OCUP_FINAL FAIXA_RENDA
         / SELECTION = ELASTICNET(STEPS=120 CHOOSE=CVEX) CVMETHOD=SPLIT(4);
   SCORE DATA = DEMAIS.VALID_BUR_CLUS&i OUT=DEMAIS.SCORE_BUR_CLUS_&i._&J;
 STORE DEMAIS.RESULT&i._&J;
RUN;

DATA DEMAIS.MAPE_BUR_CLUS&i._&j;
    SET DEMAIS.SCORE_BUR_CLUS_&i._&j;
WHERE p_RENDA_BB_ARRAIS NE .;
```

```
DECIL = &J;
MAPE = (ABS(RENDA_BB_ARRAIS - p_RENDA_BB_ARRAIS) / RENDA_BB_ARRAIS);
KEEP NR_CPF_CGC CLUSTER_BUR  RENDA_BB_ARRAIS p_RENDA_BB_ARRAIS DECIL MAPE;
RUN;
%END; %MEND Z; %Z;
%END; %MEND X; %X;
```

Apêndice B

Exemplos dos Resultados Modelo *Bureau*

10:50 Wednesday, January

The GLMSELECT Procedure

Data Set	WORK.QUANTIL1
Test Data Set	DEMAIS.VALID_BUR_CLUS1
Dependent Variable	RENDA_BB_ARRAIS
Selection Method	ELASTICNET
Stop at Specified Number of Steps	120
Choose Criterion	External Cross Validation
External Cross Validation Method	Split
External Cross Validation Fold	4
Effect Hierarchy Enforced	None

Observation Profile for Analysis Data

Number of Observations Read	863
Number of Observations Used	757
Number of Observations Used for Training	757

Observation Profile for Test Data

Number of Observations Read	3699
Number of Observations Used	502

Class Level Information

Class	Levels	Values
CD_SEXO	2	1 2
CEP_SRF_3	69	10 11 13 14 15 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 46 47 48 49 50 51 52 53 54 55 56 57 58 60 61 62 63 64 65 66 68 69 70 71 72 74 76 77 78 79 80 81 82 83 84 85 86

		138 858	
COD_OCUP_FINAL	93	1 2 3 4 5 8 9 10 13 15 17 18 19 20 21 22 25 26 27	
		28 29 30 31 32 33 34 37 38 40 41 44 45 46 49 50 51	
		53 54 55 59 60 66 71 72 75 76 77 82 83 84 85 86 87	
		88 90 91 94 96 98 99 111 113 131 149 150 153 160	
		163 169 173 175 177 198 206 207 210 241 243 244	
		247 ...	

Dimensions

Number of Effects	21
Number of Effects after Splits	182
Number of Parameters	182

10:50 Wednesday, January

The GLMSELECT Procedure

Elastic Net Selection Summary

Step	Effect Entered	Effect Removed	Number Effects In	ASE	Test ASE	CVEX PRESS
0	Intercept		1	0	8763542.06	0*

Selection stopped because the selected model is a perfect fit.

10:50 Wednesday, January

The GLMSELECT Procedure
Selected Model

The selected model, based on External Cross Validation, is the model at Step 0.

Effects: Intercept

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value
Model	0	0	.	.
Error	756	0	0	
Corrected Total	756	0		

Root MSE	0
Dependent Mean	0
R-Square	.
Adj R-Sq	.
AIC	-Infty
AICC	-Infty
SBC	-Infty
ASE (Train)	0
ASE (Test)	8763542

CVEX PRESS	0

Parameter Estimates

Parameter	DF	Estimate
Intercept	1	0

Score Information

Input Data Set	DEMAIS.VALID_BUR_CLUS1
Output Data Set	DEMAIS.SCORE_BUR_CLUS_1_1
Number of Observations Read	3699
Number of Observations Scored	3699

Apêndice C

Modelo *Behavior*

```
\%MACRO X;
\%DO I= 1 \%TO 1;
DATA MODEL_BEH_CLUS&i;
SET DEMAIS.MODEL_BEH_CLUS&i; RUN;
PROC SORT DATA = MODEL_BEH_CLUS&i;
BY RENDA_BB_ARRAIS;
RUN;

PROC SQL;
SELECT
COUNT(NR_CPF_CGC)
  INTO :TOT
  FROM MODEL_BEH_CLUS&i;
QUIT;

%MACRO Z;

%DO J=1 %TO 10;
DATA QUANTIL&J;
RETAIN RENDA_BB_ARRAIS;
SET MODEL_BEH_CLUS&i;
IF _N_ > (&TOT/10)*(&J-1) AND _N_ <= (&TOT/10)*&J;
RUN;

ODS OUTPUT PARAMETERESTIMATES = DEMAIS.PAR_SEL_BEH_CL&I._D&J;
PROC GLMSELECT DATA = QUANTIL&J
```

```
    TESTDATA = DEMAIS.VALID_BEH_CLUS&i;
CLASS CD_SEXO CD_OCP_PPL_PF CD_NTZ_OCP
CEP_SRF_3 COD_OCUP_FINAL COD_ETDO_CVIL
COD_NCLD COD_GRAU_INST
GRAU_INST_PAIS1 GRAU_INST_PAIS2
    CARACT_ESP_1 CARACT_ESP_2 CARACT_ESP_3
    CARACT_ESP_4 CARACT_ESP_5 CARACT_ESP_6
    CARACT_ESP_7 CARACT_ESP_8    CARACT_ESP_9
    CARACT_ESP_10 CARACT_ESP_11 CARACT_ESP_12
    CARACT_ESP_13 CARACT_ESP_14 CARACT_ESP_15
    CARACT_ESP_16
    NM_MCA NM_MOD IN_UNES_N OAB MAIS_MEDICOS
    REC_SALARIO_BB PROFESSOR ADMIN_EMPRESARIO
    MEDICO ADVOGADO ENGENHEIRO ENFERMEIRO
    CONTADOR ODONTOLOGO FISIOTERAPEUTA PSICOLOGO
    VETERINARIO JORNALISTA NUTRICIONISTA
    PUBLICITARIO ECONOMISTA TECNOLOGO FONOAUDIOLOGO FOTOGRAFO;
    MODEL RENDA_BB_ARRAIS = IDADE_SRF ANOT_SERASA
    ANOT_SCPC ANOT_CADIN SALDO_LIVELO RENDA_MEDIA_DOC_TED
    RENDA_MEDIA_TRANS_INT VLR_IMOVEIS_URB VLR_IMOVEIS_RUR
    VLR_SEMOV VLR_MOVEIS VL_SDO_MEDI_CRDR VL_SDO_MEDI_DVDR
    VL_TTL_CRD_CT   VL_TTL_DEB_CT VL_VST_REAL_FAT VL_PCLD_REAL_FAT
    VL_CPR_USD_FAT  VL_TTL_PGTO_RLZD VL_TTL_RTV VL_SDO_MEDI_OPR
    VL_SDO_MEDI_APL VL_SDO_MEDI_SMT  VL_SDO_MEDI_CC_AA
    VL_SDO_RFCD_CRT VL_MEDI_APL_SMT VL_CC_APL_ULT_TIM
    VL_CC_APL_ANT_TIM VL_SDO_ULT_SEIS_MM VL_SDO_DVDR_CC_AA
    TRANS_TED_DOC QT_TED_DOC VALOR_ENVIADO QT_ENVIOS QT_TRAN_DGTL
    QT_TRAN_N_DGTL MARGEM VL_RC VL_DSP MFB CAPITAL_PRUDENCIAL
    INVEST_1 INVEST_2 INVEST_3 INVEST_4 INVEST_5
    INVEST_6 INVEST_7 INVEST_8 INVEST_9 INVEST_10
    INVEST_11 INVEST_12 INVEST_13 INVEST_14 INVEST_15
    INVEST_16 INVEST_17 INVEST_18 INVEST_19 INVEST_20
    INVEST_21 INVEST_22   INVEST_23
    SALDO_1 SALDO_2 SALDO_3  SALDO_4 SALDO_5  SALDO_6 SALDO_7
    SALDO_8 SALDO_9 SALDO_10 SALDO_11 SALDO_12 SALDO_13   SALDO_14
    SALDO_15 SALDO_16
    INAD90_1 INAD90_2 INAD90_3 INAD90_4 INAD90_5
```

```
 INAD90_6 INAD90_7 INAD90_8  INAD90_9 INAD90_10
 INAD90_11 INAD90_12 INAD90_13 INAD90_14 INAD90_15
 INAD90_16
 CANAIS_40 CANAIS_18 CANAIS_52  CANAIS_51 CANAIS_31
 CANAIS_55 CANAIS_24 CANAIS_16 CANAIS_23 CANAIS_7
 CANAIS_13 CANAIS_41 CANAIS_15 CANAIS_36   CANAIS_46
 CANAIS_27 CANAIS_44 CANAIS_25 CANAIS_4 CANAIS_32
 CANAIS_8 CANAIS_34 CANAIS_57
 SFN_4  SFN_15 SFN_11 SFN_12 SFN_5 SFN_1 SFN_3 SFN_14 SFN_6
 SFN_10 SFN_9 SFN_8
 INAD_SFN_4 INAD_SFN_15 INAD_SFN_11 INAD_SFN_12 INAD_SFN_5
 INAD_SFN_1 INAD_SFN_3 INAD_SFN_14 INAD_SFN_6 INAD_SFN_10
 INAD_SFN_9 INAD_SFN_8
 CARTAO_13 CARTAO_16 CARTAO_18 CARTAO_19
 CARTAO_20 CARTAO_21 CARTAO_22 CARTAO_7
 CARTAO_1 CARTAO_10 CARTAO_11 CARTAO_12
 CARTAO_2 CARTAO_23 CARTAO_24   CARTAO_25
 CARTAO_3 CARTAO_9 CARTAO_14 CARTAO_5
 CARTAO_6 CARTAO_8 CARTAO_15 CARTAO_17
 CARTAO_4
 QT_CHQ_DLVD_AA QT_CHQ_DLVD_SMT QT_DD_EXC_CHQ
 QT_VRC_CHQ_ANT_TIM QT_TTL_ADTT_SMT QT_DD_ADTT_AA
 QT_DD_EXC_CHQ_SMT VL_CONSORCIO_IMOVEIS
 VL_CONSORCIO_AUTO VL_CONSORCIO_MOTO VL_SEGURO_AUTO VL_SEGURO_PATR
 VL_SEGURO_PESSOAS VL_PREVIDENCIA PGTO_ENERGIA_ELETRICA
 PGTO_AGUA PGTO_SUPERMERCADO PGTO_COMBUSTIVEL PGTO_CELULAR
 PGTO_TV_CABO
 TMP_CTA TMP_CDTO
 ANOT_SERASA_EXNO ANOT_SCPC_EXNO ANOT_CADIN_EXNO
 RENDA_MEDIA_CONJUGE RENDA_MEDIA_FILHO RENDA_MEDIA_PAIS
 RENDA_MEDIA_AVO RENDA_MEDIA_IRMAOS RENDA_MEDIA_DEMAIS
RESTITUICAO_ANT1 RESTITUICAO_ANT2 RESTITUICAO_ANT3
RESTITUICAO_ANT4 RESTITUICAO_ANT5 RESTITUICAO_ANT6
FAIXA_RENDA CD_SEXO CD_OCP_PPL_PF CD_NTZ_OCP CEP_SRF_3
COD_OCUP_FINAL COD_ETDO_CVIL COD_NCLD
COD_GRAU_INST GRAU_INST_PAIS1 GRAU_INST_PAIS2
CARACT_ESP_1 CARACT_ESP_2 CARACT_ESP_3 CARACT_ESP_4
```

```
    CARACT_ESP_5 CARACT_ESP_6 CARACT_ESP_7 CARACT_ESP_8
 CARACT_ESP_9 CARACT_ESP_10 CARACT_ESP_11 CARACT_ESP_12
 CARACT_ESP_13 CARACT_ESP_14 CARACT_ESP_15 CARACT_ESP_16
    NM_MCA NM_MOD IN_UNES_N OAB MAIS_MEDICOS REC_SALARIO_BB
 PROFESSOR ADMIN_EMPRESARIO MEDICO ADVOGADO ENGENHEIRO
 ENFERMEIRO CONTADOR ODONTOLOGO FISIOTERAPEUTA PSICOLOGO
 VETERINARIO JORNALISTA NUTRICIONISTA PUBLICITARIO ECONOMISTA
 TECNOLOGO FONOAUDIOLOGO FOTOGRAFO /
 SELECTION = ELASTICNET(STEPS=120 CHOOSE=CVEX) CVMETHOD=SPLIT(4);
   SCORE DATA = DEMAIS.VALID_BEH_CLUS&i
   OUT=DEMAIS.SCORE_BEH_CLUS_&i._&J;
    STORE DEMAIS.RESULT_BEH&i._&J;
RUN;

DATA DEMAIS.MAPE_BEH_CLUS&i._&j;
SET DEMAIS.SCORE_BEH_CLUS_&i._&j;
WHERE p_RENDA_BB_ARRAIS NE .;
DECIL = &J;
MAPE = (ABS(RENDA_BB_ARRAIS - p_RENDA_BB_ARRAIS) / RENDA_BB_ARRAIS);
KEEP NR_CPF_CGC CLUSTER_BEH  RENDA_BB_ARRAIS p_RENDA_BB_ARRAIS DECIL MAPE;
RUN;
%END; %MEND Z; %Z;
%END; %MEND X; %X;
```

Apêndice D

Exemplo dos Resultados Modelo *Behavior*

15:09 Wednesday, January

127173

15:09 Wednesday, January

The GLMSELECT Procedure

Data Set	WORK.QUANTIL1
Test Data Set	DEMAIS.VALID_BEH_CLUS1
Dependent Variable	RENDA_BB_ARRAIS
Selection Method	ELASTICNET
Stop at Specified Number of Steps	120
Choose Criterion	External Cross Validation
External Cross Validation Method	Split
External Cross Validation Fold	4
Effect Hierarchy Enforced	None

Observation Profile for Analysis Data

Number of Observations Read	12717
Number of Observations Used	7827
Number of Observations Used for Training	7827

Observation Profile for Test Data

Number of Observations Read	54502
Number of Observations Used	8909

Class Level Information

Class	Levels	Values

```
CD_SEXO                    2    1 2
CD_OCP_PPL_PF            251    0 1 3 4 5 8 9 10 11 12 14 15 16 17 19 20 21 22 23
                                25 26 27 28 29 30 31 32 33 34 37 38 39 40 41 42 43
                                44 45 48 50 51 52 53 54 58 59 60 62 64 66 71 72 73
                                74 75 76 77 81 82 83 84 85 86 87 88 89 90 91 94 95
                                96 98 99 100 101 102 103 104 105 106 107 108 109
                                ...
CD_NTZ_OCP                28    0 1 2 3 4 5 6 7 8 10 11 12 13 14 15 16 17 18 19 20
                                21 23 31 33 41 42 61 91
CEP_SRF_3                145    10 11 12 13 14 15 20 21 22 23 24 25 26 27 28 29 30
                                31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47
                                48 49 50 51 52 53 54 55 56 57 58 60 61 62 63 64 65
                                66 67 68 69 70 71 72 74 75 76 77 78 79 80 81 82 83
                                84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99
                                ...
COD_OCUP_FINAL           265    0 1 2 3 4 5 8 9 10 11 12 13 14 15 16 17 18 19 20
                                21 22 23 25 26 27 28 29 30 31 32 33 34 37 38 39 40
                                41 42 43 44 45 46 48 49 50 51 52 53 54 55 58 59 60
                                62 64 66 67 69 71 72 73 74 75 76 77 78 81 82 83 84
                                85 86 87 88 89 90 91 94 95 96 97 98 99 100 101 102
                                ...
COD_ETDO_CVIL             10    0 1 2 3 4 5 6 7 8 99
COD_NCLD                   6    0 1 2 3 4 5
COD_GRAU_INST              9    0 1 2 3 5 6 7 8 9
GRAU_INST_PAIS1            9    0 1 2 3 5 6 7 8 9
GRAU_INST_PAIS2            9    0 1 2 3 5 6 7 8 9
CARACT_ESP_1               2    0 1
                                                        15:09 Wednesday, January

                          The GLMSELECT Procedure

                          Class Level Information

Class                 Levels    Values

CARACT_ESP_2               2    0 1
CARACT_ESP_3               2    0 1
CARACT_ESP_4               2    0 1
CARACT_ESP_5               2    0 1
CARACT_ESP_6               1    0
CARACT_ESP_7               2    0 1
CARACT_ESP_8               2    0 1
CARACT_ESP_9               2    0 1
CARACT_ESP_10              2    0 1
CARACT_ESP_11              2    0 1
CARACT_ESP_12              2    0 1
CARACT_ESP_13              2    0 1
CARACT_ESP_14              2    0 1
CARACT_ESP_15              2    0 1
CARACT_ESP_16              2    0 1
nm_mca                    30    Apple BLU Blackview CUBOT DOOGEE Google Guiabolso
                                HUAWEI LENOVO LG Electro LGE Lenovo Meizu Mirage
                                Motorola Multilaser Não encont OUKITEL OnePlus
                                Quantum Sony TCL TP-LINK Xiaomi asus motorola
                                positivo samsung ulefone vernee
nm_mod                   292    2014819 4034E 5010E 5026J 5051J 5085N 71S 8050E
                                9008J ASUS_A001D ASUS_A007 ASUS_I01WD ASUS_T00J
```

		ASUS_X008D ASUS_X00AD ASUS_X00DD ASUS_X00HD
		ASUS_X00ID ASUS_X00LD ASUS_X00QD ASUS_X00RD
		ASUS_X00TD ASUS_X013D ASUS_X017D ASUS_X018D
		ASUS_Z00AD ASUS_Z00LD ...
IN_UNES_N	2	0 1
OAB	2	0 1
MAIS_MEDICOS	1	0
REC_SALARIO_BB	2	0 1
PROFESSOR	2	0 1
ADMIN_EMPRESARIO	2	0 1
MEDICO	2	0 1
ADVOGADO	2	0 1
ENGENHEIRO	2	0 1
ENFERMEIRO	2	0 1
CONTADOR	2	0 1
ODONTOLOGO	2	0 1
FISIOTERAPEUTA	2	0 1
PSICOLOGO	2	0 1
VETERINARIO	2	0 1
JORNALISTA	2	0 1
NUTRICIONISTA	2	0 1
PUBLICITARIO	2	0 1
ECONOMISTA	2	0 1
TECNOLOGO	2	0 1
FONOAUDIOLOGO	2	0 1
FOTOGRAFO	2	0 1

15:09 Wednesday, January

The GLMSELECT Procedure

Dimensions

Number of Effects	257
Number of Effects after Splits	1337
Number of Parameters	1337

15:09 Wednesday, January

127173

15:09 Wednesday, January

The GLMSELECT Procedure

Data Set	WORK.QUANTIL1
Test Data Set	DEMAIS.VALID_BEH_CLUS1
Dependent Variable	RENDA_BB_ARRAIS
Selection Method	ELASTICNET
Stop at Specified Number of Steps	120
Choose Criterion	External Cross Validation
External Cross Validation Method	Split
External Cross Validation Fold	4
Effect Hierarchy Enforced	None

Observation Profile for Analysis Data

Number of Observations Read	12717
Number of Observations Used	7827
Number of Observations Used for Training	7827

Observation Profile for Test Data

Number of Observations Read	54502
Number of Observations Used	8909

Class Level Information

Class	Levels	Values
CD_SEXO	2	1 2
CD_OCP_PPL_PF	251	0 1 3 4 5 8 9 10 11 12 14 15 16 17 19 20 21 22 23 25 26 27 28 29 30 31 32 33 34 37 38 39 40 41 42 43 44 45 48 50 51 52 53 54 58 59 60 62 64 66 71 72 73 74 75 76 77 81 82 83 84 85 86 87 88 89 90 91 94 95 96 98 99 100 101 102 103 104 105 106 107 108 109 ...
CD_NTZ_OCP	28	0 1 2 3 4 5 6 7 8 10 11 12 13 14 15 16 17 18 19 20 21 23 31 33 41 42 61 91
CEP_SRF_3	145	10 11 12 13 14 15 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 60 61 62 63 64 65 66 67 68 69 70 71 72 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 ...
COD_OCUP_FINAL	265	0 1 2 3 4 5 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 25 26 27 28 29 30 31 32 33 34 37 38 39 40 41 42 43 44 45 46 48 49 50 51 52 53 54 55 58 59 60 62 64 66 67 69 71 72 73 74 75 76 77 78 81 82 83 84 85 86 87 88 89 90 91 94 95 96 97 98 99 100 101 102 ...
COD_ETDO_CVIL	10	0 1 2 3 4 5 6 7 8 99
COD_NCLD	6	0 1 2 3 4 5
COD_GRAU_INST	9	0 1 2 3 5 6 7 8 9
GRAU_INST_PAIS1	9	0 1 2 3 5 6 7 8 9
GRAU_INST_PAIS2	9	0 1 2 3 5 6 7 8 9
CARACT_ESP_1	2	0 1

The GLMSELECT Procedure

Class Level Information

Class	Levels	Values
CARACT_ESP_2	2	0 1
CARACT_ESP_3	2	0 1
CARACT_ESP_4	2	0 1
CARACT_ESP_5	2	0 1
CARACT_ESP_6	1	0
CARACT_ESP_7	2	0 1

CARACT_ESP_8	2	0 1
CARACT_ESP_9	2	0 1
CARACT_ESP_10	2	0 1
CARACT_ESP_11	2	0 1
CARACT_ESP_12	2	0 1
CARACT_ESP_13	2	0 1
CARACT_ESP_14	2	0 1
CARACT_ESP_15	2	0 1
CARACT_ESP_16	2	0 1
nm_mca	30	Apple BLU Blackview CUBOT DOOGEE Google Guiabolso HUAWEI LENOVO LG Electro LGE Lenovo Meizu Mirage Motorola Multilaser Não encont OUKITEL OnePlus Quantum Sony TCL TP-LINK Xiaomi asus motorola positivo samsung ulefone vernee
nm_mod	292	2014819 4034E 5010E 5026J 5051J 5085N 71S 8050E 9008J ASUS_A001D ASUS_A007 ASUS_I01WD ASUS_T00J ASUS_X008D ASUS_X00AD ASUS_X00DD ASUS_X00HD ASUS_X00ID ASUS_X00LD ASUS_X00QD ASUS_X00RD ASUS_X00TD ASUS_X013D ASUS_X017D ASUS_X018D ASUS_Z00AD ASUS_Z00LD ...
IN_UNES_N	2	0 1
OAB	2	0 1
MAIS_MEDICOS	1	0
REC_SALARIO_BB	2	0 1
PROFESSOR	2	0 1
ADMIN_EMPRESARIO	2	0 1
MEDICO	2	0 1
ADVOGADO	2	0 1
ENGENHEIRO	2	0 1
ENFERMEIRO	2	0 1
CONTADOR	2	0 1
ODONTOLOGO	2	0 1
FISIOTERAPEUTA	2	0 1
PSICOLOGO	2	0 1
VETERINARIO	2	0 1
JORNALISTA	2	0 1
NUTRICIONISTA	2	0 1
PUBLICITARIO	2	0 1
ECONOMISTA	2	0 1
TECNOLOGO	2	0 1
FONOAUDIOLOGO	2	0 1
FOTOGRAFO	2	0 1

15:09 Wednesday, January

The GLMSELECT Procedure

Dimensions

Number of Effects	257
Number of Effects after Splits	1337
Number of Parameters	1337

15:09 Wednesday, January

The GLMSELECT Procedure

Elastic Net Selection Summary

tep	Effect Entered	Effect Removed	Number Effects In	ASE	Test ASE	CVEX PRESS
0	Intercept		1	0	10562692.3	0*

Selection stopped because the selected model is a perfect fit.

15:09 Wednesday, January

The GLMSELECT Procedure

Selected Model

The selected model, based on External Cross Validation, is the model at Step 0.

Effects: Intercept

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value
Model	0	0	.	.
Error	7826	0	0	
Corrected Total	7826	0		

Root MSE	0
Dependent Mean	0
R-Square	.
Adj R-Sq	.
AIC	-Infty
AICC	-Infty
SBC	-Infty
ASE (Train)	0
ASE (Test)	10562692
CVEX PRESS	0

Parameter Estimates

Parameter	DF	Estimate
Intercept	1	0

Score Information

Input Data Set	DEMAIS.VALID_BEH_CLUS1
Output Data Set	DEMAIS.SCORE_BEH_CLUS_1_1
Number of Observations Read	54502
Number of Observations Scored	54502