UNIVERSIDADE FEDERAL DE JUIZ DE FORA
DEPARTAMENTO DE ESTATÍSTICA
CURSO DE ESTATÍSTICA

ISADORA CONSOLI SILVA LUPCHINSKI

ANÁLISE DE CLASSES LATENTES: CLASSIFICAÇÃO DE PERFIS NA CONCESSÃO DE CRÉDITO

JUIZ DE FORA
2018

ISADORA CONSOLI SILVA LUPCHINSKI

ANÁLISE DE CLASSES LATENTES: CLASSIFICAÇÃO DE PERFIS NA CONCESSÃO DE CRÉDITO

Trabalho de Conclusão de Curso apresentado ao Curso de Estatística da Universidade Federal de Juiz de Fora, como requisito parcial para obtenção do grau de Bacharel em Estatística.
Orientador: Ronaldo Rocha Bastos

JUIZ DE FORA
2018

ANÁLISE DE CLASSES LATENTES: CLASSIFICAÇÃO DE PERFIS NA CONCESSÃO DE CRÉDITO

ISADORA CONSOLI SILVA LUPCHINSKI

Trabalho de Conclusão de Curso apresentado ao Curso de Estatística da Universidade Federal de Juiz de Fora, como requisito parcial para obtenção do grau de Bacharel de Estatística.

Aprovado em 06 de julho de 2018.

BANCA EXAMINADORA

_____

PhD - Ronaldo Rocha Bastos - Orientador

Universidade Federal de Juiz de Fora

_____

Doutor - Augusto Carvalho Souza

Universidade Federal de Juiz de Fora

_____

Phd - Marcel de Toledo Vieira

Universidade Federal de Juiz de Fora

“Ce quon voit et ce qu’on ne voit pas”

Frédéric Bastiat

AGRADECIMENTOS

À Deus, por me permitir concluir mais essa jornada em uma universidade e curso muito bons.

Aos meus pais, Rogério e Isabel, minha inspiração para muitos momentos, por todo apoio, valores, conhecimentos e ensinamentos que me transmitiram.

Ao meu marido, Lucas, meu calmante emocional e fonte de segurança para minhas ansiedades, que sempre me incentivou a dar o meu melhor, a não desistir, a honrar os princípios que recebi e me deu força nos momentos mais difíceis desta caminhada.

Aos meus irmãos, Kássio e Maria Teresa, que mesmo não entendendo bem o que um Estatístico faz e achando que eu deveria fazer um curso na área da saúde, sempre torceram pelo meu sucesso e respeitaram a minha escolha.

Aos demais familiares, pelo carinho, orações e preocupação. Saibam que vocês me ajudaram muito.

Ao meu orientador, Ronaldo, por todo conhecimento, amparo e direcionamento.

Aos demais professores do Departamento de Estatística, pelas gratas contribuições e aprendizados.

Aos amigos de faculdade, saibam que vocês tornaram essa trajetória muito mais agradável, divertida e proveitosa. Valorizem-se! Sejam justos. Ser um estatístico não é pra qualquer um. Essa é a profissão que escolhemos para nossas vidas, é o nosso trabalho.

RESUMO

A análise de concessão de crédito é um ramo da área financeira muito estudado e importante. Isso porque deixou de utilizar critérios meramente subjetivos como o julgamento humano, e passou a empregar técnicas estatísticas e computacionais mais rápidas, precisas e confiáveis. E como forma principal e preliminar para conceder o crédito, é necessário uma análise criteriosa do tipo de cliente que esse mais se enquadra, tal como adimplente ou inadimplente. Diante disso, a análise de classes latentes (LCA) surge como uma nova proposta para a avaliação do perfil dos clientes sujeitos a concessão do crédito bancário. Neste contexto, este trabalho objetivou apresentar quais variáveis determinam a conjectura das características dos bons e dos maus pagadores, através da LCA, como forma de ajudar as instituições credoras na tomada de decisão. Além disso, também foi alvo deste trabalho estudar técnicas e procedimentos sobre a análise de risco de crédito, bem como conhecer o banco de dados e analisar suas características de forma exploratória. Para tanto, foram utilizadas estatísticas descritivas, regressões logística simples, múltipla e o comando stepwise do R, com intuito de fazer uma pré-seleção das variáveis do LCA que, em seguida, foram utilizadas no pacote poLCA do mesmo software. A partir da análise de dados foi possível perceber, mesmo que com poucas variáveis manifestas, uma distinção entre duas classes latentes: os bons e os maus pagadores. No entanto, a partir desse trabalho sugere-se futuros estudos associados a outras metodologias, tais como: validação cruzada, análise discriminante e análise de *cluster*.

Palavras-Chaves: Análise Classe Latente. Regressão Logística. Crédito. Perfis de Clientes.

ABSTRACT

Credit granting analysis is a very studied and important branch of finance. This is because it no longer uses purely subjective criteria such as human judgment, but employs faster, more accurate and reliable statistical and computational techniques. As a primary and preliminary way to grant credit, a careful analysis of the type of client (good and bad payers is necessary. Therefore, the latent class analysis (LCA) emerges as a new proposal for the evaluation of clients' profile subject to the granting of bank credit. In this context, this paper aims to present which variables determine the conjecture of the characteristics of good and bad payers, through LCA, as a way of helping creditor institutions in decision-making. In addition, it was also the objective of this paper to study techniques and procedures on credit risk analysis, as well as to know the database and analyze its characteristics in an exploratory manner. For that, descriptive statistics, simple and multiple logistic regressions, and stepwise regression were used, in order to pre-select the LCA variables, which were then used in the poLCA package of the R software. From the analysis of data, it was possible to perceive a distinction between two latent classes, even with only a few observed variables: the good and the bad payers. However, this work suggests future studies associated with other methodologies, such as cross-validation, discriminant analysis and cluster analysis.

Keywords: Latent Class Analysis. Logistic Regression. Credit. Customer Profiles.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

O crédito é um tema muito interessante e que existe desde os primórdios da humanidade, facilitando o comércio e possibilitando a sobrevivência de muitos fatos históricos. Tem suma importância no desenvolvimento econômico mundial, capaz de financiar a evolução das nações e também de diversas guerras. Quando bem empregado, faz empresas crescerem, ou falirem se caso contrário, auxilia pessoas na tomada de decisões, à novas ideias, projetos, desenvolvimento de tecnologias e muito mais. Por isso a análise de concessão de crédito é uma área muito estudada e que necessita de tantos aprimoramentos.

Conceder crédito a alguém deixou de ser um método tradicional em que prevalecia a confiança e as experiências do julgamento humano. Hoje, devido as pressões de oferta e demanda, dos aspectos econômicos e das disputas de mercado, surgiram, a partir do século XX, formas mais precisas e eficazes de se avaliar e “julgar” a concessão de crédito (NUNES, 2011).

A concessão de crédito está ligada com o provimento antecipados de recursos que, em geral, é feita de forma monetária (dinheiro em espécie). Com esses recursos, o indivíduo contemplado consegue acesso a bens e serviços que, de outro modo, não seria obtido tão facilmente, ou que tardaria um pouco mais. No entanto, esta operação financeira causa um impacto na instituição credora, já que fornecer crédito a um indivíduo que não possui condições de efetuar seu pagamento - num futuro acordado entre as partes (instituição credora e cliente) - pode exigir renegociações, obrigar a empresa a arcar com o crédito cedido até que o cliente faça o pagamento e, em alguns casos, até fazer com que a empresa se responsabilize completamente pela dívida, gerando desta forma, custos adicionais e até mesmo prejuízos para a instituição.

Através da análise de concessão de crédito é possível detectar fraudadores, eliminando, na cadeia de créditos, o nome das pessoas inidôneas. Em outras palavras, a análise de crédito é um importante processo para manter o equilíbrio entre clientes e empresas, diminuindo os riscos de consumidores inadimplentes, e

consequentemente, o risco das dívidas não serem pagas e fazendo, portanto, com que as empresas possam trabalhar com cliente mais comprometidos.

Diante deste cenário, a importância desse trabalho está ligado à necessidade em criar um perfil dos clientes e classificá-los em bons ou maus pagadores, visando ser um procedimento de concessão de crédito mais seguro e inteligente, para assim obter uma análise mais precisa durante a concessão ou não do crédito.

O objetivo geral deste trabalho é realizar uma avaliação dos clientes, caracterizando-os em bons ou maus pagadores. Ou seja, trata-se de um problema de classificação, pois busca-se encontrar um agrupamento mínimo de características, de um conjunto de dados disponíveis, para predizer características de dados futuros.

Portanto, este trabalho buscou reunir dados/informações com o propósito de responder ao seguinte problema de pesquisa: quais fatores determinam e contribuem para a avaliação do perfil dos clientes bancários para classificar os indivíduos em bons ou maus pagadores, utilizando análise de classes latentes?

Em outras palavras, é possível dizer que o interesse prioritário é o de classificar o perfil destes indivíduos para verificar se os mesmos se enquadram como sendo bom pagador ou não, utilizando a análise de classes latentes (LCA) como metodologia estatística principal. Isso, porque à medida que esta classificação é bem feita, pode gerar ganhos (tanto financeiros quanto em termos de tempo e agilidade de processamento) para a instituição credora.

Diante do exposto e de um mercado financeiro altamente competitivo, é perceptível a necessidade de modelos e metodologias rápidas, precisas e eficientes, no sentido de minimizar os erros e perdas, de modo a tentar diminuir a probabilidade de inadimplência, e consequentemente, reduzir as chances de custos adicionais e, em pior caso, prejuízos de não recebimento. Para tanto, vale a pena ressaltar a importância da veracidade das informações prestadas pelos clientes e a necessidade de se levar em conta a conjuntura da instituição credora e do meio em que ela está inserida.

Desta forma, optou-se por estudar a análise de classes latentes como uma proposta diferenciada para abordar a concessão de crédito. Já que esta envolve a identificação de relações entre variáveis, utilizando tanto as variáveis observadas, como aquelas não observadas ou latentes, como é o caso do recebimento da concessão de crédito. Além disso, segundo Silva (2011, p.49), “o propósito de uma LCA é a verificação dos padrões de variação em indicadores dependentes e a identificação de grupos ou classes com comportamento relativamente homogêneo”, o que vai ao encontro da aplicação deste trabalho, já que o objetivo principal é tentar prever em qual perfil um determinado cliente sujeito à análise de concessão de crédito se enquadra.

Além disso, pode-se dizer que o modelo de classes latentes permite: descrever o perfil de cada classe latente encontrada e calcular a probabilidade de cada indivíduo pertencer a uma dada classe, identificando a classe a que o indivíduo pertence. Esta metodologia pode ter diversas aplicações, que incluem desde a análise de pesquisas de opinião, estilo de vida e escolha do consumidor até outros fenômenos sociais e comportamentais.

O presente trabalho está estruturado de forma a permitir um ganho de entendimento das motivações, do embasamento teórico e chegando na proposição de solução.

No capítulo 1, são apresentadas a introdução e a relevância do tema no contexto da área financeira, porém voltado para uma aplicação estatística, além dos objetivos do trabalho.

O capítulo 2, apresenta uma revisão de literatura acerca da análise de crédito, área esta que é o foco da aplicação deste trabalho.

O capítulo 3, por sua vez, apresenta todo o embasamento conceitual e teórico das metodologias utilizadas. Nesse capítulo são apresentados os fundamentos, noções preliminares, principalmente quanto ao modelo de mistura, modelo de regressão logística e da análise de classes latentes.

É no capítulo 4 que são apresentados a aplicação da LCA em dados reais de uma instituição financeira, o detalhamento exploratório do mesmo, a construção de

modelos para estimar a probabilidade de classificação dos clientes, bem como sua descrição e interpretação; e por fim os resultados.

O capítulo 5 é o de conclusão. Nele são discutidos os resultados da aplicação da proposta do trabalho.

O capítulo 6 é o de considerações finais. Nele são apresentadas algumas possibilidades de diferentes metodologias que também poderiam ser usadas e dos novos temas que se abrem a partir desse trabalho.

Por fim, as referências bibliográficas registram as várias fontes utilizadas para formar a base conceitual de todo o trabalho e que contribuem como caminhos para o leitor seguir na busca por mais informações e dados referentes ao tema que o embasam.

1.1 JUSTIFICATIVA

Desta forma, devido à necessidade em classificar os clientes em bons ou maus pagadores, essa pesquisa se justifica através da aplicação de técnicas estatísticas e computacionais, em contribuição para o seu público alvo (instituições financeiras), em obter um método de avaliação do perfil dos clientes sujeitos à análise de concessão de crédito bancário, já que o principal objetivo dessas transações é evitar conceder crédito aos maus pagadores (que pode ser evadida com a identificação desses, com dados históricos, comportamentais, indicadores financeiros e registros de negativação, por exemplo). Desta forma, este trabalho está estruturado do ponto de vista da instituição financeira.

1.2 OBJETIVO GERAL

O presente trabalho tem como objetivo geral apresentar quais fatores determinam a avaliação do perfil dos clientes bancários para classificar os indivíduos em bons ou

maus pagadores como forma de ajudar na tomada de decisão, via análise de classes latentes, com a finalidade de apontar benefícios para instituições financeiras, tendo como base o banco de dados reais de uma instituição financeira do sul da Alemanha.

1.3 OBJETIVOS ESPECÍFICOS

Estudar técnicas e procedimentos sobre a análise de risco de crédito;

Conhecer o banco de dados que será utilizado neste trabalho e analisar as características dele de forma exploratória;

Usar técnicas e metodologias estatísticas, tais como: regressão logística e análise de classes latentes;

Identificar e classificar o perfil dos clientes sujeitos à análise de risco de crédito;

Descrever os resultados.

2 REVISÃO DE LITERATURA

2.1 CRÉDITO

O conceito de crédito pode ser visto sob diversas perspectivas, porém, mediante a ótica bancária, o crédito pode ser interpretado como sendo um empréstimo de uma determinada quantia monetária mediante a promessa de pagamentos futuros. Ele envolve a criação de expectativas de recebimento/pagamento num futuro combinado e acordado por ambas as partes (instituição credora e cliente). Nesse sentido, Caouette, Altman e Narayanan (1998) afirmam que o risco de crédito é a chance de que essa expectativa não se cumpra.

É interessante, aliás, entender o funcionamento da concessão de crédito bancário. Mas há um fato que se sobrepõe a isso, que é a necessidade de conhecer e classificar o cliente por meio de técnicas estatísticas.

Indivíduos comuns quando emprestam dinheiro, por exemplo, às pessoas mais próximas, fazem uma análise de risco pessoal, analisando a possibilidade de receber ou não desse conhecido em um futuro próximo. O que acaba sendo um critério particular, subjetivo e de julgamento humano. E não obstante, acontece com as instituições financeiras, que, no entanto, necessitam de uma busca mais aprofundada e objetiva do perfil dos clientes sujeitos a análise de crédito, já que se tratam de pessoas desconhecidas, e que cada inadimplência resulta em prejuízos para a instituição credora. É sinal de que há, enfim, uma inevitabilidade para ser estudada e gerida por parte dessas instituições.

Ainda nesse contexto, pode-se notar o quão importante e necessária é a análise de risco de crédito e seu funcionamento:

> [...]Quando esses clientes necessitam de recursos, eles recorrem ao banco, que tem como norma elaborar uma análise minuciosa para a concessão do crédito pretendido, baseados primordialmente em critérios pessoais e financeiros. O banco busca com isso colher indícios de insolvência de clientes, pois a preocupação é que a quantia emprestada não retorne mais com os respectivos encargos financeiros, que são juros e correção monetária (NETO e SÉRGIO, 2009, p.31).

Embora os autores citados acima tenham mais de 10 anos de diferença nas publicações, é possível perceber ideias semelhantes. Caso contrário, teríamos uma regressão na busca por melhoramentos no setor e, consequentemente, não teríamos um sistema bancário de análise de risco de crédito tão eficiente como temos no Brasil, por exemplo. Além disso, conforme citado acima, não se trata de querer ganhar vantagem em cima do cliente, mas sim de resguardar a instituição financeira de possíveis prejuízos e calotes. Dessa forma, é importante considerar que uma análise de risco de crédito bem feita, traz resultados não só para o cliente, mas para a instituição credora como um todo, seja porque diminui o risco de fraudes, seja nesse caso uma precaução para não sobre cair sob os demais clientes o prejuízo (NETO e SERGIO, 2009).

É preciso, porém, ir mais além do que tais conceitos. Uma vez que, a grosso modo, trata-se de um caso de classificação e agrupamento estatístico, já que envolve o fato de conceder ou não o crédito ao cliente, analisando pois, inicialmente, suas características demográficas, financeiras e socioeconômicas. Como bem nos assegura Nunes (2011, p. 14): "A classificação,[...] consiste na predição de um valor categórico como, por exemplo, predizer se o cliente é digno de crédito (bom pagador) ou se ele não é digno (mau pagador)". Desse modo, a classificação (predição) tem papel importante nesse cenário de concessão de crédito, pois, se bem feita, ampara o processo de concessão de crédito.

A análise de crédito envolve a habilidade de fazer uma decisão de crédito, dentro de um cenário de incertezas e constantes mutações e informações incompletas. Esta habilidade depende da capacidade de analisar logicamente situações, não raro, complexas, e chegar a uma conclusão clara, prática e factível de ser implementada (SCHRICKEL, 1994).

Sendo assim, investir em conhecimentos de técnicas estatísticas e computacionais aprimoradas nesta área, pode resultar em avanços no processo de classificação dos clientes em bons ou maus pagadores, e, consequentemente, na análise de concessão de crédito. Podemos perceber, que esse quadro remete a uma tarefa que muitos profissionais da área de exatas e de gerenciamento se debruçam a tentar resolver. Além do mais, não é exagero afirmar que esse tema necessita de modelos que

estejam em constante modificação, visando melhorias, adequações às circunstâncias presentes (históricas e políticas do país e da instituição credora) e, principalmente, na boa capacidade de generalização dos modelos, tencionando o aumento de especificidade dos mesmos, de modo que o número de acertos de classificação desses, erre o quanto menos.

2.2 CREDIT SCORING

Um dos inconvenientes da concessão de crédito é o risco do não recebimento. Se a análise de crédito não for bem feita, o lucro obtido com uma transação financeira pode ser total ou parcialmente afetado com as despesas de cobrança de um mau pagador ou com a perda do crédito. Neste cenário, uma das técnicas usualmente empregadas na análise de crédito é justamente o credit scoring.

O credit scoring é, a grosso modo, o resultado do ajuste de um modelo que se faz ca partir de algumas informações importantes, definidas pelo credor, que irão determinar se o cliente terá o crédito aprovado ou não. O credor colhe informações, basicamente, sobre a base de clientes dele e verifica quais são os pontos em comum entre os bons e os maus pagadores. Dessa forma, criam-se vários perfis com os clientes já cadastrados e compara-se com as informações prestadas do novo cliente, de modo a verificar em qual perfil este último se enquadra.

Não se trata de uma modelagem tão trivial, seja porque depende do modelo estar bem ajustado e ter uma boa capacidade de previsão e adequação, seja porque depende das informações prestadas pelos clientes, o que nem sempre é possível verificar a veracidade. Em suma, é preciso trazer à tona a importância da qualidade das informações, pois quanto mais amplas e precisas, melhores tendem a ser as condições de avaliação do risco de cada cliente e operação.

O desenvolvimento de um modelo de credit scoring consiste, de forma geral, em determinar uma função das variáveis cadastrais dos clientes que possa auxiliar na tomada de decisão para aprovação de crédito, envolvendo cartões de créditos, cheque especial, atribuição de limite, financiamento de veículo, imobiliário e varejo.

Normalmente esses modelos são desenvolvidos a partir de bases históricas de performance de crédito dos clientes e também de informações pertinentes ao produto (LOUZADA e DINIZ, 2012).

Conforme citado acima, pode-se dizer que o modelo de credit scoring é um modelo que fornece uma “régua”, no sentido de ter uma escala ordinal, ou seja, cada valor tem um peso, que serve como parâmetro de classificação. Por exemplo, clientes com score baixo, ou seja, que possuem baixa probabilidade de pagar (consequentemente, é um perfil com alto risco de não pagar) pode ser automaticamente reprovado, enquanto que o cliente com pontuação intermediária, poderá ir para uma segunda etapa como usuário e assim, aprofundar no quesito de concessão ou não do crédito. Além disso, o autor deixa claro, que o credit scoring pretende classificar se um determinado cliente é ou não merecedor do crédito a partir de certas informações cadastrais.

Existem vários fatores que se associam à probabilidade de não pagamento (inadimplência). Nesse contexto, um modelo de credit scoring deve combinar os fatores mais importantes associados à inadimplência, inter-relacioná-los e atribuir um valor para tornar possível uma análise numérica. Dessa forma, na prática, quanto menor for o escore, maior será a possibilidade de não pagamento e portanto de perdas associadas à clientes inadimplentes (GHERARDI e GHIELMETTI, 1997).

O credit scoring é uma parte do processo de análise de concessão de crédito, porém de grande relevância, pois é uma forma confiável que o comércio, financeiras, bancos e prestadoras de serviços em geral têm de obter dados estatísticos sobre os clientes, pois permitem análises com agilidade, faz comparação entre perfis de clientes (e com certo grau de acerto expressivo), faz a análise de diversas propostas para aquele perfil de consumidor (seu investimento) e, consequentemente, auxilia na tomada de decisões.

Além disso, pode-se citar como vantagem do uso do credit scoring a diminuição dos custos, pois clientes que são mau avaliados, ou seja, com score muito baixo, por exemplo, podem ser automaticamente reprovados.

Dessa forma, tais vantagens mostradas acima podem possibilitar o crescimento controlado da carteira de clientes das instituições credoras, mas mantendo baixas as taxas de perdas, e consequentemente, aumentando a rentabilidade do negócio. Além de melhorar a organização das informações, de modo a contribuir para a sistematização e administração das mesmas, pois contribuem para o desenvolvimento do processo de concessão de crédito (BESS, 2007).

Conforme explicado acima, nota-se que os resultados desse quadro de vantagens do uso do credit scoring são devidos à expansão do mercado de crédito massificado, centrado na avaliação do risco de crédito e em especial, na inadimplência, que por sua vez, passou a demandar análises mais rápidas e homogêneas na avaliação da concessão do crédito.

Além disso, fica evidente que o desenvolvimento de sistemas computacionais possibilitou o tratamento estatístico adequado desses conjuntos de dados. Conforme o exposto acima, espera-se, dessa forma, contribuir com os resultados da análise de concessão de crédito.

3 METODOLOGIA

3.1 INTRODUÇÃO

Neste capítulo serão abordadas os principais métodos e técnicas de modelagem para a análise de classes latentes, LCA (do inglês *Latent Class Analysis*), que é aplicado em muitos domínios. Os exemplos podem ser encontrados em diversas áreas, desde a psicologia e sociologia, passando pelo marketing (para avaliar os gostos e preferências dos clientes) até o setor financeiro, que é o foco desde trabalho.

As ciências comportamentais e sociais (por considerar muitos construtos como variáveis latentes) usam da LCA para avaliar diferentes contextos e interesses, tais como, a depressão em adolescentes (Lanza, Flaherty, e Collins, 2003); a influência do estilo de vida nas escolhas de transporte (Silva, 2013); os temperamentos de bebês, de modo a analisar os perfis comportamentais de crianças aos 4 meses e o grau de sinais comportamentais de medo aos 14 meses (Stern, Arcus, et al., 1995); ou uma LCA para avaliar agências de viagem e grupos estratégicos (Fernandes, 2013), e ainda, para expressar a pobreza como um construto multidimensional (Dewilde, 2004); além de analisar o uso de maconha e atitudes entre os alunos do ensino médio americano de 1977 a 2001 (Chung, 2006).

A análise de classe latente - também conhecida como análise de estrutura latente - pode ser usada para identificar agrupamentos, *clusters,* de "tipos" semelhantes de indivíduos ou observações de dados categóricos multivariados, estimar as características desses grupos latentes e retornar a probabilidade de cada observação pertencer a cada grupo.

A LCA envolve a identificação das relações entre variáveis, utilizando tanto as variáveis observadas, bem como aquelas não observadas ou latentes, comumente utilizadas em equações estruturais. “O propósito de uma LCA é a verificação dos padrões de variação em indicadores dependentes e a identificação de grupos ou classes com comportamento relativamente homogêneo” (SILVA, 2013, p.49).

Na LCA a classificação de cada indivíduo em uma classe é feita com base na probabilidade do mesmo pertencer a essa classe (likelihood of class membership).

Isto é feito assumindo-se que existe uma variável latente (não observada) que pode ser deduzida a partir dos dados em análise, e esta variável latente é usada para explicar a variância dos dados.

Pode ser o caso de o pesquisador ter algumas ideias sobre quantos e quais tipos de classes existem na população, mas estas ideias não estão tão bem estabelecidas para serem tratadas de maneira confirmatória. Cabe ao pesquisador gerar vários modelos para os dados, cada um sendo diferenciado dos outros pelo número de classes latentes. O ajuste relativo desses modelos é então comparado para identificar o número ótimo de classes. Tal ajuste é avaliado por meio de índices projetados para determinar qual modelo pode reproduzir melhor os dados observados. Dentre essas estatísticas para avaliar o ajuste do modelo estão a AIC e BIC. Além de confiar em tais avaliações estatísticas de ajuste de modelo, o pesquisador também deve considerar a coerência substantiva das classes latentes ao selecionar um modelo. Esta questão é importante da mesma forma que a interpretabilidade é um critério chave na determinação do número ideal de fatores, isto é, a solução final de um LCA deve ser defensável com base nos tipos de indivíduos que agrupamos, no que diz respeito às suas respostas sobre os indicadores utilizados na análise, bem como outras variáveis potencialmente pertinentes, tais como características demográficas (FINCH e FRENCH, 2015)

Resumidamente, pode-se dizer que a análise fornece informações sobre padrões de comportamento de risco e a prevalência desses padrões. E a modelagem de classes latentes refere-se a um grupo de técnicas para identificar subgrupos inobserváveis, ou latentes, dentro de uma população. Desta forma, a análise de classe latente é um método estatístico usado para identificar um conjunto de classes latentes discretas e mutuamente exclusivas de indivíduos com base em suas respostas a um conjunto de variáveis categóricas observadas.

3.2 NOÇÕES PRELIMINARES

Neste capítulo introduzimos algumas noções preliminares para que o leitor consiga acompanhar o desenvolvimento das análises seguintes, já que estas dependem de alguns conhecimentos prévios.

3.2.1 Modelo de Mistura

A LCA pode ser enxergada como um caso especial de agrupamento. Nele, se pressupõe que as observações se originem de uma quantidade de grupos (classes) e modelos, cada um com a sua própria distribuição de probabilidade. Como a variável latente é nominal, o modelo de classe latente é tido como um tipo de modelo de mistura finita (GENGE, 2014).

Ainda segundo Genge (2014), modelos de mistura finita são uma técnica muito utilizada para modelar a heterogeneidade não observada ou aproximar funções de distribuição geral. Eles são usados em muitas áreas diferentes, como agronomia, agricultura, astronomia, biologia, economia, genética, marketing e medicina. Uma visão geral dos modelos de mistura é dada em McLachlan e Peel (2000).

A seguir, algumas noções preliminares a respeito de misturas finitas a fim de aproximar o leitor da nomenclatura utilizada.

Segundo Faria (2006), seja X uma variável aleatória com valores num espaço R e cuja função de densidade de probabilidade é dada por:

$$f(x) = \sum_{j=1}^{g} \pi_j \, f_j(x) \tag{1}$$

Tal que $f_j(x)$ são funções de densidade de probabilidade, ou seja, são densidades componentes da mistura e as quantidades π_j são designadas por proporções ou pesos de mistura, que precisam respeitar os critérios abaixo:

$$0 \leq \pi_j \leq 1 \qquad e \qquad \sum_{j=1}^{g} \pi_j = 1 \tag{2}$$

O número de componentes *g* pode ser um valor conhecido ou um parâmetro a estimar a partir de uma amostra.

A função de distribuição de X é uma mistura finita de g distribuições e a densidade de probabilidade dada na equação (1) é uma mistura finita de *g* funções de densidade de probabilidade.

Em muitas aplicações as densidades componentes da mistura pertencem a uma família paramétrica, e passam a ser representadas por $f_j(x, \theta_j)$ onde θ_j é o vetor dos parâmetros desconhecidos da j-ésima densidade componente da mistura. Neste caso, a função densidade de probabilidade dada em (1) pode ser representada da seguinte forma:

$$f(x; \varphi) = \sum_{j=1}^{g} \pi_j \, f_j(x; \theta_j) \tag{3}$$

Sendo que φ é o vetor que contém todos os parâmetros desconhecidos do modelo de mistura e que pode ser definido de acordo com a expressão abaixo:

$$\varphi = (\pi_1, \pi_2, \dots, \pi_{g-1}, \xi^T)^T \tag{4}$$

Onde ξ é o vetor que contém os parâmetros $\theta_1, \dots, \theta_g$.

3.2.2 Método da Máxima Verossimilhança

Para fazer a estimação dos parâmetros dos modelos de mistura, foi utilizado o método da máxima verossimilhança, aplicado inicialmente por Rao em 1948.

Seja $x = (x_1^T, \dots, x_n^T)^T$ uma amostra aleatória de n realizações independentes da variável aleatória de *g* distribuições cuja função densidade de probabilidade é definida em (3).

Uma vez que os $\boldsymbol{x}$ são independentes, a função de verossimilhança é obtida pelo produto dos termos dados na expressão (3) e é definida por:

$$L(\varphi) = \prod_{i=1}^{n} f(x;\varphi) = \prod_{i=1}^{n} \left(\sum_{j=1}^{g} \pi_j \, f_j(x;\theta_j) \right) \tag{5}$$

Onde denota-se por φ o vetor de parâmetros desconhecidos, ou seja, o vetor a estimar usando o método da máxima verossimilhança. Logo, a função de log verossimilhança é dada por:

$$l(\varphi) = ln[L(\varphi)] = \sum_{i=1}^{n} log \left(\sum_{j=1}^{g} \pi_j \, f_j(x;\theta_j) \right) \tag{6}$$

O princípio da máxima verossimilhança é obter o valor de φ que maximize $l(\varphi)$, ou equivalentemente $L(\varphi)$, que pode ser obtido derivando a equação acima e igualando a zero em relação a cada parâmetro, da seguinte forma:

$$\frac{\partial l(\varphi)}{\partial \varphi} = 0 \tag{7}$$

3.2.3 Algoritmo EM

Segundo Faria (2011), o algoritmo EM é uma ferramenta computacional utilizada para o cálculo do estimador de máxima verossimilhança (EMV) de forma iterativa, principalmente em problemas envolvendo dados incompletos (estes problemas caracterizam-se pela inexistência de alguma informação dos dados). Este algoritmo converge para o EMV e tem como base a ideia de substituir uma difícil maximização por uma sequência de maximizações mais fáceis, envolvendo dois passos: o passo “E” (esperança) que calcula o valor esperado do logaritmo da verossimilhança completa; e o passo “M”, que encontra seu máximo. Os passos são repetidos até se atingir a convergência. Esta ferramenta é utilizada para fazer a estimação do modelo de classes latentes (LCM).

3.2.4 Regressão Logística

Muitas vezes o pesquisador está interessado em usar técnicas estatísticas de análise de regressão, através de modelos que descrevam a relação entre uma variável resposta a partir de uma ou mais variáveis explicativas. No caso de a variável resposta Y_i ser binária (dicotômica), ela só poderá assumir dois possíveis estados (Y_i =1 ou Y_i =0), ou seja, ou o “sucesso” ou o “fracasso”, respectivamente.

O modelo de regressão logística binária, em que a variável resposta do modelo tem distribuição Bernoulli (ou Binomial) e a função de ligação é a função logística é um dos modelo lineares generalizados mais utilizados, tanto na área da saúde quanto na financeira.

O modelo de regressão logística com apenas uma variável explicativa é chamado de modelo de regressão logística simples. No entanto, assim como no modelo de regressão linear, é possível ajustar o modelo para a variável resposta levando em consideração mais de uma variável explicativa, que é chamado de modelo de regressão logística múltipla. Desta forma, este trabalho utilizou destes modelos, tanto o de regressão simples, quanto o de regressão múltipla para verificar a relação entre as variáveis explicativas com a variável resposta, individualmente e posteriormente em conjunto, respectivamente.

De acordo com Hosmer e Lemeshow (1989) considere um conjunto de p variáveis independentes denotadas como um vetor $\boldsymbol{x}' = (x_1, x_2, \dots, x_p)$. E as probabilidades de sucesso e fracasso são dadas respectivamente por $P(Y = 1|\boldsymbol{x}) = \pi(\boldsymbol{x})$ e $P(Y = 0|\boldsymbol{x}) = 1 - \pi(\boldsymbol{x})$. Neste caso, a função de ligação é dada pelo logit do modelo de regressão múltipla, dado pela equação:

$$g(\boldsymbol{x}) = \ln\left(\frac{\pi(\boldsymbol{x})}{1-\pi(\boldsymbol{x})}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \tag{8}$$

E neste caso, o modelo de regressão logística é dado por

$$\pi(\boldsymbol{x}) = \frac{e^{g(\boldsymbol{x})}}{1 + e^{g(\boldsymbol{x})}} = \frac{1}{1 + e^{-g(\boldsymbol{x})}} \tag{9}$$

Para ajustar o modelo de regressão logística apresentado acima, é necessário estimar os parâmetros $\beta_0 \; e \; \beta_j$, $j = 1,2, \dots, p$ via método da máxima verossimilhança.

Considerando n pares de observações de uma amostra (x_i, y_i) independente, tal que x_i é o valor esperado da variável explicativa da i-ésima observação, $i = 1,2, \dots, n$ e y_i representa o valor esperado da variável binária.

A função de distribuição da probabilidade de $Y_i \sim Ber(\pi(\boldsymbol{x}))$ para o modelo de regressão logística que é dado por:

$$f(y_i, \pi_i) = \pi(\boldsymbol{x})^{y_i}\big(1 - \pi(\boldsymbol{x})\big)^{1-y_i} \text{ de modo que } \begin{cases} y_i = 0 \\ y_i = 1 \end{cases} \tag{10}$$

A função de verossimilhança é dada pelo produtos dos termos da equação acima, em consequência de que $Y_1, Y_2, \dots, Y_n$ serem independentes:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi(x_{\boldsymbol{i}})^{y_i}\big(1 - \pi(x_{\boldsymbol{i}})\big)^{1-y_i} \tag{11}$$

onde $\boldsymbol{\beta} = \big(\beta_0, \beta_1, \dots, \beta_p\big)^T$ é o vetor de parâmetros desconhecidos. E a função log verossimilhança é dada pelo $ln[L(\boldsymbol{\beta})]$, da seguinte forma:

$$l(\boldsymbol{\beta}) = l\,n[L(\boldsymbol{\beta})] = \sum_{i=1}^{n} y_i ln\,[\pi(x_{\boldsymbol{i}})] + (1 - y_i) ln[1 - \pi(x_{\boldsymbol{i}})] \tag{12}$$

Pelo princípio da máxima verossimilhança, como visto anteriormente, é necessário utilizar a derivação da equação acima em relação a cada parâmetro e igualá-las a zero afim de obter o valor de $\boldsymbol{\beta}$ que maximize $l(\boldsymbol{\beta})$. No entanto, tais derivações e consequentes igualdades a zero são não lineares em $\boldsymbol{\beta}$ de forma que são necessários métodos iterativos, como por exemplo o método de Newton Raphson para a resolução do sistema de equações. Que podem ser encontrados com mais detalhes, por exemplo, em Louzada e Diniz (2012), Nunes (2011) e Dantas e DeSouza (2008).

3.2.4 Teorema de Bayes

O teorema de Bayes (também conhecido por lei de Bayes ou regra de Bayes) é um corolário da lei da probabilidade total, que descreve a probabilidade de um evento, baseado em um conhecimento a priori que pode estar relacionado ao evento. Além disso, ele mostra como alterar as probabilidades a priori tendo em vista novas evidências para obter probabilidades a posteriori (BUSSAB e MORETTIN, 2006) e pode ser expresso matematicamente da seguinte forma:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{13}$$

em que A e B são eventos e $P(B) \neq 0$.

Onde:

. $P(A)$ *e* $P(B)$ são as probabilidades a priori de A e B;

. $P(A|B)$ é a probabilidade a posteriori (probabilidade condicional de A dado B);

. $P(B|A)$ é a probabilidade a posteriori (probabilidade condicional de B dado A);

Este teorema calcula a probabilidade a posteriori $P(A|B)$ através das probabilidades a priori das hipóteses $P(A)$, juntamente com $P(B)$ e $P(B|A)$, que representam respectivamente a probabilidade a priori do conjunto de dados B ser observado e a chance do mesmo ser observado dada a hipótese A.

Este teorema é utilizado para calcular as estimativas das probabilidades a posteriori do modelo de classes latentes.

3.2.5 Matriz de confusão

Uma matriz de confusão mostra o número de previsões corretas e incorretas feitas pelo modelo de classificação em comparação com os resultados reais (valor alvo) nos dados. A matriz é *N* x *N* , onde *N* é o número de valores alvo (classes). O desempenho

de tais modelos é comumente avaliado usando os dados da matriz. A tabela a seguir exibe uma matriz de confusão 2x2 para duas classes (geralmente uma positiva e uma negativa).

Tabela 1: Matriz de confusão

		Previsto		
		Mau Pagador	Bom pagador	Total
Observado	Mau Pagador	a	b	a+b
	Bom pagador	c	d	c+d
	Total	a+c	b+d	a+b+c+d

Fonte: Autoria própria (2018).

- A letra **a** na tabela acima representa a quantidade de clientes classificados como maus pagadores dado que eles são realmente maus pagadores, ou seja, o modelo fez uma previsão correta, acertou;
- A letra **b** representa a quantidade de clientes classificados como bons pagadores quando na verdade não os são (ou seja, o modelo previu errado, equivalente ao erro do tipo 2, utilizado em estatística);
- A letra **c** é a quantidade de clientes classificados como maus pagadores quando na verdade são bons pagadores, ou seja, seria equivalente ao erro do tipo 1, já que o modelo previu erroneamente;
- A letra **d** é a de clientes classificados como bons pagadores dado que eles realmente são, ou seja, o modelo fez uma boa previsão dos adimplentes, ele acertou.

É possível perceber que, neste caso, o erro do tipo 2 é o mais grave para a instituição credora, já que seu interesse é evitar errar na previsão dos maus pagadores (classificação em que pode levar a prejuízos quando classificada erroneamente), ou seja, quanto menor o valor de **b** – e maior os valores de **a** e **d** - melhor tende a ser a classificação.

A partir desta tabela é possível observar medidas comumente utilizadas em problemas de classificação binária, tais como:

- Acurácia (AC): corresponde à proporção total de predições acertadas (corretas).

$$AC = \frac{a + d}{a + b + c + d} \tag{14}$$

- Sensibilidade: corresponde à proporção de clientes classificados corretamente como inadimplentes quando na verdade os são.

$$Sensibilidade = \frac{a}{a + b} \tag{15}$$

- Especificidade: corresponde à proporção de clientes classificados corretamente como adimplentes quando na verdade os são.

$$Especificidade = \frac{d}{c + d} \tag{16}$$

- Falsos positivos (FP): corresponde à proporção de clientes serem classificados pelo modelo como bons pagadores dado que eles são, na verdade, maus pagadores.

$$FP = \frac{b}{a + b} \tag{17}$$

- Falsos negativo (FN): corresponde à proporção de clientes serem classificados pelo modelo como maus pagadores dado que eles são bons pagadores.

$$FN = \frac{c}{c + d} \tag{18}$$

3.2.6 poLCA

O poLCA é um pacote implementado no software livre R utilizado para a estimação de modelos de classes latentes, também chamados por alguns autores de modelos de regressão de classes latentes para variáveis de resultados politômicos; ele foi

utilizado neste trabalho para realizar a estimação do LCM e apresentado no capítulo de aplicação:

> O modelo de classe latente básica é um modelo de mistura finita no qual as distribuições de componentes são consideradas tabelas de classificação cruzada de múltiplas vias com todas as variáveis mutuamente independentes. O modelo de regressão de classe latente permite ainda ao pesquisador estimar os efeitos das covariáveis na previsão da participação na classe latente. O poLCA usa algoritmos de maximização de expectativas e de Newton-Raphson para encontrar estimativas de máxima verossimilhança dos parâmetros do modelo. (LINZER e LEWIS, 2011, p.1)

Este pacote é o mais usado na área de estudo de análises de classes latentes pois é considerado o mais amigável e completo, e ainda,

> O poLCA também inclui ferramentas para visualizar resultados de modelos, criar conjuntos de dados categóricos multivariados simulados com uma estrutura categórica latente não observada e pós processá-los para produzir várias outras quantidades de interesse, incluindo percentuais de células esperados e probabilidades posteriores de associação de classe latente (LINZER e LEWIS, 2011, p.26).

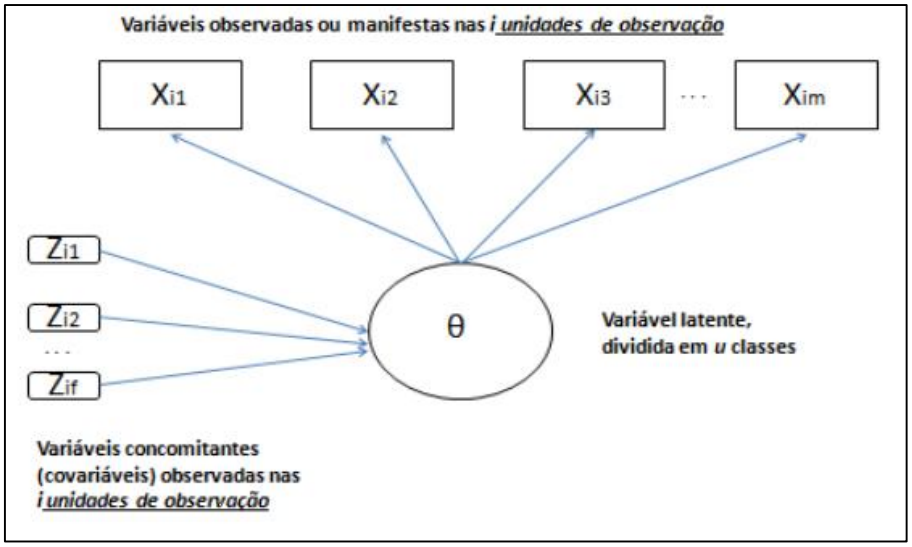
No entanto, existem outros pacotes no R que desempenham papeis semelhantes, como por exemplo: e1071, gllm e randomLCA. No entanto, tais pacotes só podem estimar o modelo básico para variáveis de resultado dicotômico.

O poLCA pode ser baixado no link: http://CRAN.R-project.org/package=poLCA.

3.3 MODELO DE CLASSES LATENTES

O diagrama a seguir esquematiza de forma mais clara o modo como o modelo de classes latentes com covariáveis são criados:

Figura 1: Esquema de Modelo de Classes Latentes (MCL)



Fonte: (BASTOS, 2016, p. 148)

Observe na figura acima que as informações sobre as características socioeconômicas dos clientes sujeitos a análise de crédito se encontram representadas pelas variáveis observadas ou manifestas, X_{im} tal que i representa cada indivíduo e m representa cada pergunta. A variável latente θ, que divide os i indivíduos em u classes, é o traço não observado que define os grupos ou classes em que o universo se divide, determinando, por sua vez, os valores das variáveis manifestas. As características sociais, familiares, de trabalho, etc., podem ser incorporadas ao modelo como variáveis concomitantes (covariáveis) Z_{if}, tal que f é o número de covariáveis observadas. Dessa forma, pode-se avaliar a influência de tais variáveis sobre o traço latente que se deseja estudar, que neste caso é o recebimento do crédito.

É possível afirmar que o modelo de classe latente se enquadra como modelo de mistura com u componentes, em que cada um destes segue uma distribuição paramétrica. Além disso, cada um desses componentes possui um peso π_s associado, que representa a probabilidade a priori para uma observação ter sido originada deste componente e a distribuição da mistura é dada pela soma ponderada sobre os componentes u, conforme notação adotada por Genge (2014).

3.3.1 Generalização do modelo

O objetivo do modelo tradicional de classe latente é determinar o menor número de classes latentes s que é suficiente para explicar as associações observadas entre as variáveis manifestas.

De acordo com Genge (2014), os modelos de classe latente aproximam a distribuição conjunta observada das variáveis manifestas como a soma ponderada do número finito, u das tabelas de classificação cruzada.

O modelo de classe latente pode ser escrito como:

$$f(\boldsymbol{x}_i|\boldsymbol{\theta}) = \sum_{s=1}^{u} \pi_s \, f_s(\boldsymbol{x}_i, \boldsymbol{\theta}_s) \tag{19}$$

em que:

f_s: distribuição de probabilidade da classe latente s;

$\boldsymbol{x}_i$: vetor de variáveis observadas ou manifestas [$x = (x_{i1}, \dots, x_{im})$];

π_s: probabilidade a priori da classe s, tal que $\pi_j \in (0,1); \sum_{s=1}^{u} \pi_s = 1$;

$\boldsymbol{\theta}_s$: vetor de parâmetros específicos à classe latente s;

$\boldsymbol{\theta}$: vetor de todos os parâmetros para o modelo de classes latentes (mistura).

A probabilidade em cada célula da tabela de componentes é simplesmente o produto das respectivas probabilidades marginais condicionais da classe. Uma soma ponderada dessas tabelas de componentes forma uma aproximação (estimativa de densidade) da distribuição de casos através das células da tabela observada. Observações com conjuntos semelhantes de respostas sobre as variáveis manifestas x_i tenderão a se agrupar dentro das mesmas classes latentes.

Cada componente s da mistura representada pela fórmula acima, pode ser entendida como o produto de distribuições multinomiais condicionalmente independentes dos parâmetros θ_{sj} e que pode ser escrito da seguinte maneira:

$$f_s(\boldsymbol{x_i}|\theta_s) = \prod_{j=1}^{m}\prod_{h=1}^{l_j}\left(\theta_{sjh}\right)^{\boldsymbol{x_{ijh}}} \tag{20}$$

onde

. $\boldsymbol{x_i} = (x_{ijh}; j = 1, \dots, m; h = 1, \dots, l; i = 1, \dots, n)$;

. $l = \sum_{j=1}^{m} l_j$ ou seja, o número total de categorias de todas as m variáveis manifestas;

. $\theta_s = \left(\theta_{ijh}; j = 1, \dots, m; h = 1, \dots, l; i = 1, \dots, n\right)$.

A distribuição conjunta de probabilidade é expressa por:

$$P(\boldsymbol{x_i}|\theta_s) = \sum_{s=1}^{u}\pi_s\prod_{j=1}^{m}\prod_{h=1}^{l_j}\left(\theta_{sjh}\right)^{\boldsymbol{x_{ijh}}} \tag{21}$$

Os parâmetros a serem estimados pelo LCM são θ_{sjh} e π_s

A partir das estimativas de $\hat{\pi}_s$ e $\hat{\theta}_{sjh}$e utilizando o teorema de Bayes é possível chegar à estimativa da probabilidade a posteriori (de cada indivíduo pertencer a cada classe s condicionalmente aos valores das variáveis observadas $\boldsymbol{x_i}$):

$$\hat{P}(s|\boldsymbol{x_i}) = \frac{\pi_s f\left(\boldsymbol{x_i}|\widehat{\boldsymbol{\theta}}_s\right)}{\sum_{q=1}^{u}\pi_q f\left(\boldsymbol{x_i}|\widehat{\boldsymbol{\theta}}_s\right)} \tag{22}$$

Os parâmetros dos LCM são estimados via método de máxima verossimilhança através do algoritmo EM.

3.3.2 Modelo de classes latentes com covariáveis

A inclusão de uma ou mais covariáveis pode auxiliar o pesquisador em melhorar compreensão da natureza da solução da classe latente (ou seja, quais os significados das classes), e melhorar a capacidade da própria solução da classe latente para recuperar as classes reais na população. Tais covariáveis devem ser selecionadas com cuidado, focando em variáveis que tenham relação teórica com as classes que se espera que existam na população.

O modelo de mistura levando em consideração as covariáveis $\boldsymbol{z_i}$ é um modelo de classes latentes estendido do modelo de classes latentes básico. As covariáveis permitem predizer a participação na classe latente, ao passo que o modelo básico, cada resposta tem a mesma probabilidade de pertencer a cada classe latente. Desta forma, o modelo de classes latentes com covariáveis pode ser expresso da seguinte maneira:

$$f(\boldsymbol{x}_i, \boldsymbol{z_i}|\boldsymbol{\theta}) = \sum_{s=1}^{u} \pi_{si}(\boldsymbol{z_i}, \alpha_s)\, f_s(\boldsymbol{x}_i, \boldsymbol{\theta}_s) \tag{23}$$

tal que:

. $\boldsymbol{z_i}$:é o vetor de covariáveis $\boldsymbol{z_i} = [\boldsymbol{z_{i1}}, \dots, \boldsymbol{z_{if}}]$**;**

. $\pi_{si}(\boldsymbol{z_i}, \alpha_s)$**:**é a probabilidade a priori da classe s, $\pi_{si}(\boldsymbol{z_i}, \alpha_s) \in (0,1)$;

$\sum_{s=1}^{u} \pi_s\,(\boldsymbol{z_i}, \alpha_s) = 1$;

. α_s: é o vetor de coeficientes para as covariáveis, de tamanho f+1, correspondente à classe s.

Foi utilizado o pacote poLCA do software R para modelar a análise de classes latentes e levar em consideração o efeito das covariáveis através da função de ligação logit multinomial.

Além do modelo básico, existem $u-1$ vetores de parâmetros α_s a serem estimados. Desta forma, a probabilidade a posteriori pode ser expressa da seguinte forma:

$$\hat{P}(s|\boldsymbol{x_i}, \boldsymbol{z_i}) = \frac{\pi_{si}(\boldsymbol{z_i}, \hat{\alpha}_s) f\left(\boldsymbol{x_i}|\hat{\theta}_s\right)}{\sum_{q=1}^{u} \pi_q(\boldsymbol{z_i}, \hat{\alpha}_s) f\left(\boldsymbol{x_i}|\hat{\theta}_s\right)} \quad (24)$$

Tais modelos possibilitam que os efeitos das covariáveis e de prováveis diante da probabilidade a posteriori apresentada acima.

3.3.3 Seleção de Modelo

Unindo traços das abordagens de Lanza, Collins, et al. (2007) e Linzer e Lewis (2011) um dos benefícios da LCA, em contraste com outras técnicas estatísticas para dados agrupados, é a variedade de ferramentas disponíveis para avaliar o ajuste do modelo e determinar um número apropriado de classes latentes para um dado conjunto de dados. Em algumas aplicações, o número de classes latentes será selecionado por razões principalmente teóricas. Em outros casos, entretanto, a análise pode ser de natureza mais exploratória, com o objetivo de localizar o modelo mais adequado ou mais parcimonioso.

Uma variedade de ferramentas pode ser usada em conjunto para a seleção de modelos quanto a qualidade de ajuste, incluindo a estatística de razão de verossimilhança G^2 e as duas medidas de ajuste mais utilizadas: o AIC - Critério de Informação de Akaike (AKAIKE, 1974) e o BIC - Critério de Informação Bayesiano (SCHWARZ, 1978).

Segundo Fernandes (2013) modelos com valores menores para a razão G^2 apresentam maior evidência a favor de sua adequação aos dados da amostra. Embora modelos com diferentes números de classes latentes sejam tecnicamente aninhados, a distribuição da estatística de razão de verossimilhança comparando dois modelos não deve ser comparada a um teste qui-quadrado; a diferença estatística G^2 pode ser usada apenas de maneira aproximada para comparar o ajuste do modelo.

O AIC e o BIC são critérios de informações penalizados do modelo de log-verossimilhança que podem ser usados para comparar modelos concorrentes (por exemplo, modelos com diferentes números de classes latentes) ajustados aos mesmos dados. Um menor AIC e BIC para um determinado modelo sugere que o trade-off entre ajuste e parcimônia é preferível. No entanto, em geral, o modelo que minimiza o valor desses critério é o escolhido.

Os critérios de parcimônia procuram encontrar um equilíbrio entre o ajuste excessivo e insuficiente do modelo aos dados, penalizando a log-verossimilhança por uma função do número de parâmetros que estão sendo estimados. Além disso, a interpretabilidade do modelo deve ser considerada. Por exemplo, cada classe deve ser distinguível das outras com base nas probabilidades item-resposta, nenhuma classe deve ser trivial em tamanho (ou seja, com uma probabilidade de adesão próxima de zero), e deve ser possível atribuir um significado coerente, no sentido de possuir uma descrição mínima de rótulo para cada classe (DZIAK, COFFMAN, *et al.*, 2012).

Os critérios AIC e BIC são os mais populares critérios de informação de análise de classes latentes disponíveis no pacote poLCA do R e que serão utilizados em análises posteriores. Eles podem ser expressos de acordo com as equações abaixo:

$$AIC = -2\log f\left(\boldsymbol{x}_i\middle|\hat{\theta}_s, M_s\right) + \vartheta_s \log(n) \tag{25}$$

$$BIC = -2\log f\left(\boldsymbol{x}_i\middle|\hat{\theta}_s, M_s\right) + 2\vartheta_s \tag{26}$$

em que :

. $f\left(\boldsymbol{x}_i\middle|\hat{\theta}_s, M_s\right)$ é a máxima log-verossimilhança para o modelo M_s;

. ϑ_s é o número de parâmetros a serem estimados no modelo;

. n é o número de observações

Além disso, o primeiro termo de cada um dos critérios calcula a qualidade do ajuste, enquanto que o segundo termo é a penalização pela complexidade do modelo.

4 APLICAÇÃO

4.1 DESCRIÇÃO DOS DADOS

Para ilustrar o desenvolvimento de um modelo de análise de classes latentes, foi utilizado um conjunto de dados reais de um banco do sul da Alemanha, que consiste de 1000 observações de clientes requerentes de crédito bancário. São consideradas 20 variáveis explicativas, todas categóricas e a variável resposta dicotômica confiabilidade de crédito (pagou ou não pagou) como sendo uma variável importante de comparação, pois o interesse é justamente tentar prevê-la através da análise de classes latentes. Os dados estão divididos em 700 requerentes adimplentes, ou seja, que pagaram o empréstimo e 300 inadimplentes, isto é, não pagaram, representados pela variável “kredit”. Os dados brutos podem ser obtidos através do endereço eletrônico: https://archive.ics.uci.edu/ml/datasets.html.

As variáveis utilizadas estão divididas em três tabelas mostradas a seguir, tal que a primeira coluna contém o nome das variáveis seguido de suas respectivas categorias e por fim, a descrição com o significado de cada variável (rótulo). Além disso, é possível perceber que elas estão divididas quanto ao caráter das variáveis, no sentido de que as duas primeiras tabelas abordam 14 variáveis de perfil socioeconômico e a terceira com 7 variáveis de perfil demográfico.

A sigla DM contida na tabela 2 representa unidades monetárias (Deutsche Mark) que corresponde ao marco alemão, que foi a moeda oficial na República Federal da Alemanha de 1949 a 2002.

Tabela 2: Variáveis de caráter socioeconômico

Nome da Variável	Categorias	Descrição
kredit	0 = "Crédito não reembolsado" 1= "Crédito reembolsado"	Variável resposta para cada requerente
status_conta_corrente	1= "Nenhuma conta atual" 2= "Nenhum saldo da conta ou saldo devedor" 3= "[0;200) DM" ano"	Balanço do saldo em conta.
duração_meses	1= "Até 21 meses"; 2= "Mais que 21 meses";	Prazo em que o cliente obteve ao receber o empréstimo.
emprestimo_anterior	1= "Muito ruim"; 2= "Ruim - Conta crítica"; 3= "Nenhum empréstimo até agora / todos os empréstimos anteriores pagos"; 4= "Bom - Empréstimos ainda existentes com o banco até agora perfeitamente"; 5= "Muito bom - Empréstimos anteriores liquidados no banco corretamente"	Histórico de empréstimos anteriores.
motivo_emprestimo	1= "Carro"; 2= "Carro usado"; 3= "Móveis/equipamentos"; 4= "Radio/televisão"; 5= "Eletrodomésticos"; 6= "Consertos"; 7= "Treinamento"; 8= "Férias"; 9= "Requalificação profissional"; 10= "Negócios"; 11= "Outros".	Variável que demostra qual o propósito do empréstimo.
poupança_cat	1= "Sem poupança"; 2= "<100 DM"; 3= "[100;500) DM"; 4= "[500;1000) DM"; 5= "[1000;+] DM"	Valor depositado em poupança ou de ações, em Deutsche Marks (DM).

Fonte: Adaptado de Haun (2014)

Tabela 3: Variáveis de caráter socioeconômico

Nome da Variável	Categorias	Descrição
porcentagem_emprestimo	1= "[35;+]", 2= "[25;35)", 3= "[20,25)", 4= "<20".	Equivale às parcelas, em % da renda disponível.
outros_credores	1= "Não", 2="Candidato", 3= "Fiador".	Indica se o cliente possui algum fiador ou co-pretendente.
bens_solicitante	1= "Não detectável / não possui ativos", 2= "Carro / outro", 3= "Seguro de vida", 4= "Propriedade de casas e terras".	Equivale aos recursos ou bens de maior valor disponíveis dados como garantia para o empréstimo.
emprestimos_add	1= "Em outro banco", 2= "Em loja de departamentos", 3= "Nenhum".	Se o cliente possui empréstimos solicitados em outra instituição.
tipo_apartamento	1= "Deixado de graça", 2= "Apartamento alugado", 3= "Próprio"	Qual o tipo de moradia do cliente.
n_emprestimos_anteriores	1= "1" , 2= "2 ou 3"; 3= "4 ou 5"; 4= "6 ou +".	Quantidade de empréstimos anteriores, incluindo o atual.
linha_telefone_fixo	1= "Não" 2= "Sim (em meu nome)"	Se o cliente possui linha telefônica fixa.
montante_emprestimo	1= "[15001;20000]"; 2= "[10001;15000]"; 3= "[7501;10000]"; 4= "[5001;7500]"; 5= "[2501;5000]"; 6= "[1501;2500]"; 7= "[1001;1500]"; 8= "[501;1000]"; 9= "<=500".	Montante do empréstimo. Valor do empréstimo

Fonte: Adaptado de Haun (2014)

Tabela 4: Variáveis de caráter demográfico

Nome da Variável	Categorias	Descrição
sexo	1= "Homem divorciado/separado"; 2= "Mulher div./sep/casada"; 3= "Homem solteiro"; 4= "Homem casado"; 5= "Mulher solteira".	Sexo e situação conjugal do cliente.
mora_endereço	1= "Menos de 1 ano", 2= "1 a 4 anos", 3= "4 a 7 anos", 4= "Mais de 7 anos",	Corresponde ao tempo que o cliente mora no atual endereço.
ocupaçao	1= "Desempregado (buscando)", 2= "Desempregado (não buscando)", 3= "Funcionários qualificados/funcionários públicos para o serviço de nível médio", 4= "Gestor/executivo/autônomo/funcionário público sênior".	Ocupação do cliente.
n_dependentes	1= "3 ou mais" 2= "0 a 2"	Número de dependentes que o cliente detém.
imigrante	1= "Sim" 2= "Não"	Se o cliente é imigrante.
idade_cat	1= "[0;25]", 2= "[26;39]", 2= "[40;59]", 2= "[65;+]", 2= "[60;64]",	Idade do cliente na data do pedido do empréstimo
tempo_trabalho	1= "Desempregado"; 2= "Até 1 ano" 3= "1 a 4 anos"; 4= "4 a 7 anos"; 5= "7 ou mais anos".	Há quanto tempo o cliente trabalha com o atual empregador?

Fonte: Adaptado de Haun (2014)

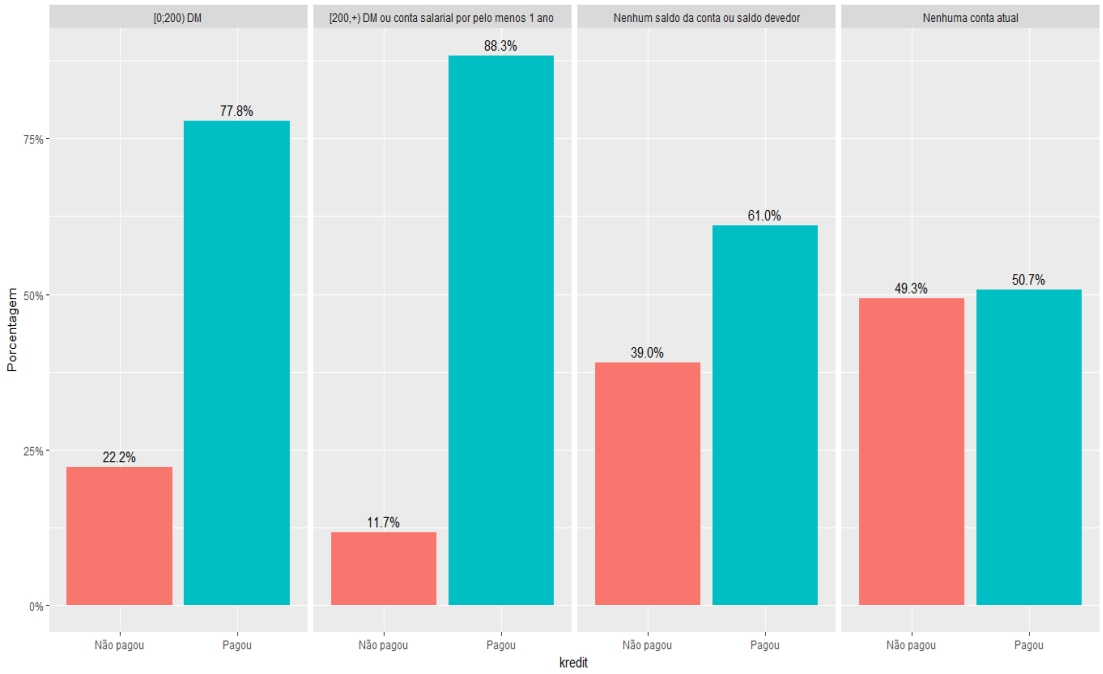
4.2 DESCRIÇÃO DAS VARIÁVEIS

Inicialmente, será mostrado a análise univariada de cada uma das variáveis, afim de conhecê-las com mais intensidade. Em seguida uma análise bivariada entre cada variável explicativa e a variável dependente “kredit”.

- Variável: **status_conta_corrente**

Dos 1000 clientes em análise 27,4% não possuem conta bancária; 26,9% não possuem saldo em conta; 6,3% detém entre zero e duzentos DM em conta enquanto que 39,4% dos demais clientes possuem mais que 200 DM ou conta salário por um ano ao menos.

Figura 2: Gráfico bivariado entre status_conta_corrente x kredit



Fonte: Autoria própria (2018).

Tabela 5: Tabela cruzada entre status_conta_corrente x kredit

		kredit		
		Crédito não foi reembolsado	Crédito reembolsado	Total
Conta corrente em banco	Nenhuma conta atual	135	139	274
	Nenhum saldo da conta ou saldo devedor	105	164	269
	[0;200) DM	14	49	63
	[200,+) DM ou conta salarial por pelo menos 1 ano	46	348	394
Total		300	700	1000

Fonte: Autoria própria (2018).

- Varável: **duraçao_meses**

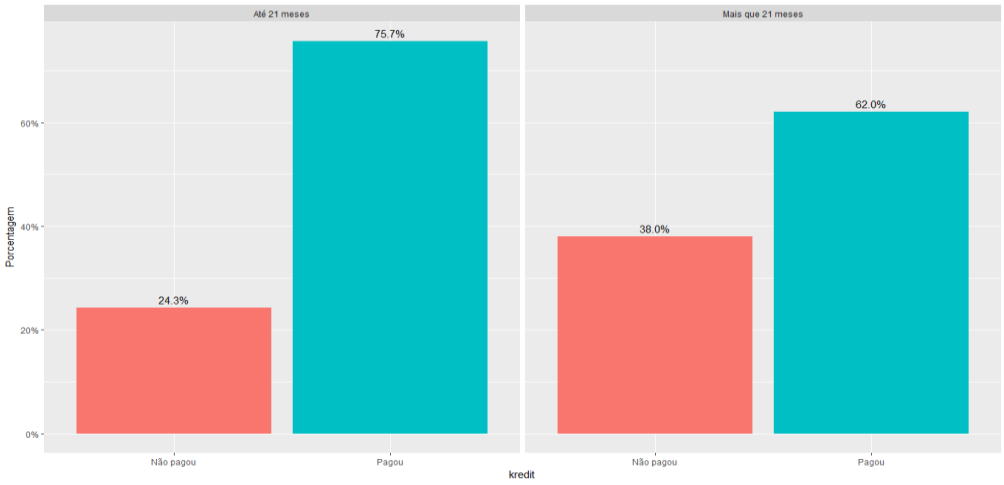
58,4% dos clientes obtiveram prazos de até 21 meses, enquanto que os demais (41,6%) mais que de 21.

Tabela 6: Tabela cruzada entre duraçao_meses x kredit

		kredit		
		Crédito não foi reembolsado	Crédito reembolsado	Total
Duração do emprestimo em meses	Até 21 meses	142	442	584
	Mais que 21 meses	158	258	416
Total		300	700	1000

Fonte: Autoria própria (2018).

Figura 3: Gráfico bivariado entre duraçao_meses x kredit



Fonte: Autoria própria (2018).

- Variável: **emprestimo_anterior**

Esta variável remete ao histórico do cliente em empréstimo anteriores.

Tabela 7: Tabelas de frequências da variável emprestimo_anterior

		Frequência	Porcentual	Porcentagem válida	Porcentagem acumulativa
Válido	Muito ruim	40	4,0	4,0	4,0
	Ruim - Conta crítica	49	4,9	4,9	8,9
	Nenhum empréstimo até agora / todos os empréstimos anteriores pagos	530	53,0	53,0	61,9
	Bom - Empréstimos ainda existentes com o banco até agora perfeitamente	88	8,8	8,8	70,7
	Muito bom - Empréstimos anteriores liquidados no banco corretamente	293	29,3	29,3	100,0
	Total	1000	100,0	100,0	

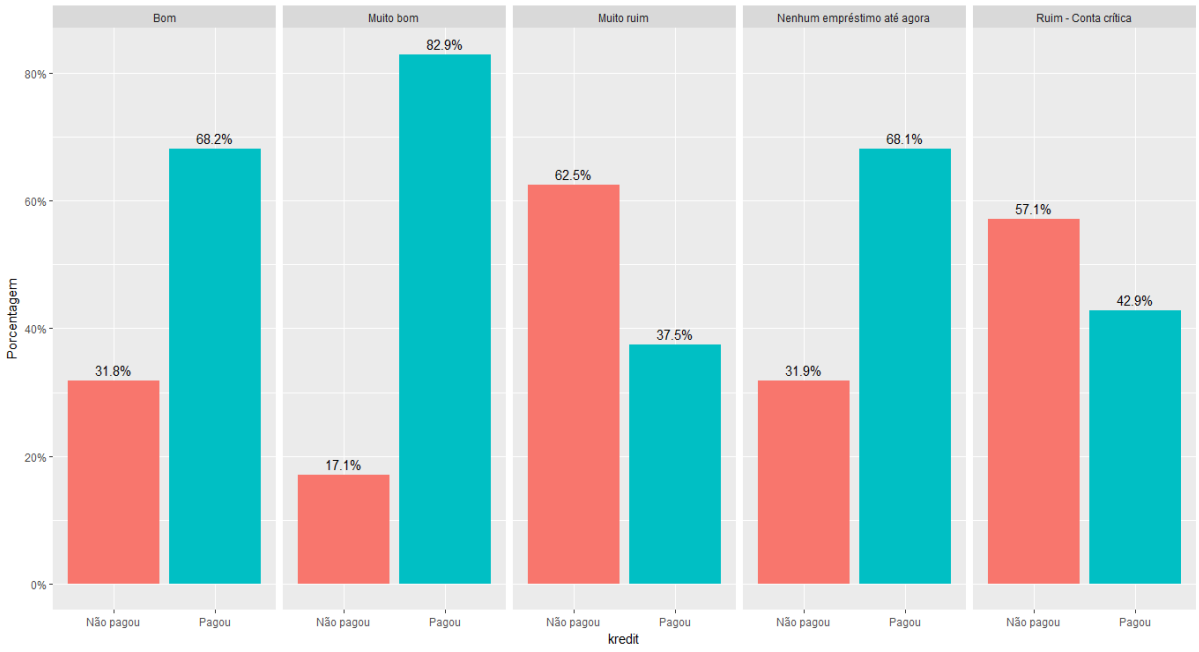
Fonte: Autoria própria (2018).

Tabela 8: Tabela cruzada entre empréstimo_anterior x kredit

		kredit		
		Crédito não foi reembolsado	Crédito reembolsado	Total
Histórico de pagamento anterior (de ruim a muito bom)	Muito ruim	25	15	40
	Ruim - Conta crítica	28	21	49
	Nenhum empréstimo até agora / todos os empréstimos anteriores pagos	169	361	530
	Bom - Empréstimos ainda existentes com o banco até agora perfeitamente	28	60	88
	Muito bom - Empréstimos anteriores liquidados no banco corretamente	50	243	293
Total		300	700	1000

Fonte: Autoria própria (2018).

Figura 4: Gráfico bivariado entre emprestimo_anterior x kredit



Fonte: Autoria própria (2018).

- Variável: **motivo_empréstimo**

Nesta variável o cliente relata qual o motivo de solicitar o empréstimo. E a categoria que mais se destacou foi a de “Móveis/equipamentos”, enquanto que a menos requisitado foi a categoria “Férias”.

Tabela 9: Tabela de frequências da variável motivo_emprestimo

		Frequência	Porcentual	Porcentagem válida	Porcentagem acumulativa
Válido	Carro novo	103	10,3	10,3	10,3
	Carro usado	181	18,1	18,1	28,4
	Móveis/equipamentos	280	28,0	28,0	56,4
	Radio/televisão	12	1,2	1,2	57,6
	Eletrodoméstico	22	2,2	2,2	59,8
	Consertos	50	5,0	5,0	64,8
	Férias	9	,9	,9	65,7
	Requalificação profissional	97	9,7	9,7	75,4
	Negócio	12	1,2	1,2	76,6
	Outros	234	23,4	23,4	100,0
	Total	1000	100,0	100,0	

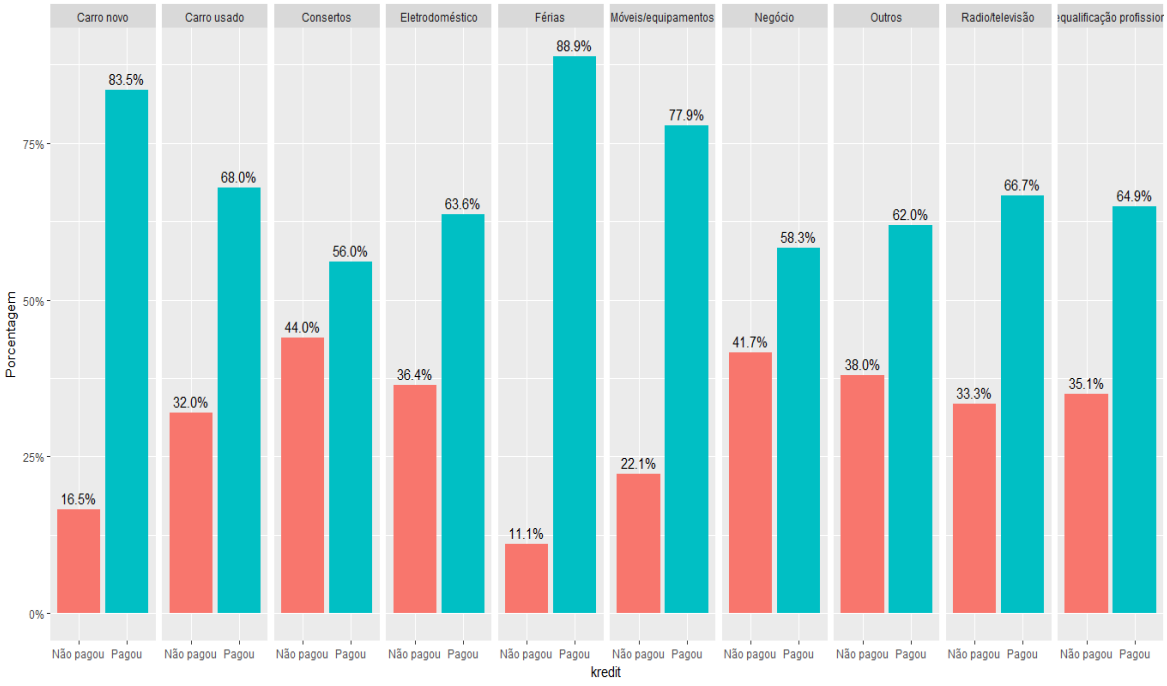
Fonte: Autoria própria (2018).

Tabela 10: Tabela cruzada entre motivo_emprestimo x kredit

		kredit		
		Crédito não foi reembolsado	Crédito reembolsado	Total
Finalidade do empréstimo	Carro novo	17	86	103
	Carro usado	58	123	181
	Móveis/equipamentos	62	218	280
	Radio/televisão	4	8	12
	Eletrodoméstico	8	14	22
	Consertos	22	28	50
	Férias	1	8	9
	Requalificação profissional	34	63	97
	Negócio	5	7	12
	Outros	89	145	234
Total		300	700	1000

Fonte: Autoria própria (2018).

Figura 5: Gráfico bivariado entre motivo_emprestimo x kredit



Fonte: Autoria própria (2018).

- Variável: **montante_cat**

Esta variável mostra o quantitativo dos montantes das categorias de empréstimo solicitada pelos clientes, de modo que de 1501 DM a 5000 DM representa pouco mais de 50% dos valores solicitados.

Tabela 11: Tabela de frequências da variável montante_cat

		Frequência	Porcentual	Porcentagem válida	Porcentagem acumulativa
Válido	[15001;20000]	5	,5	,5	,5
	[10001;15000]	35	3,5	3,5	4,0
	[7501;10000]	46	4,6	4,6	8,6
	[5001;7500]	102	10,2	10,2	18,8
	[2501;5000]	275	27,5	27,5	46,3
	[1501;2500]	231	23,1	23,1	69,4
	[1001;1500]	190	19,0	19,0	88,4
	[501;1000]	98	9,8	9,8	98,2
	<=500	18	1,8	1,8	100,0
	Total	1000	100,0	100,0	

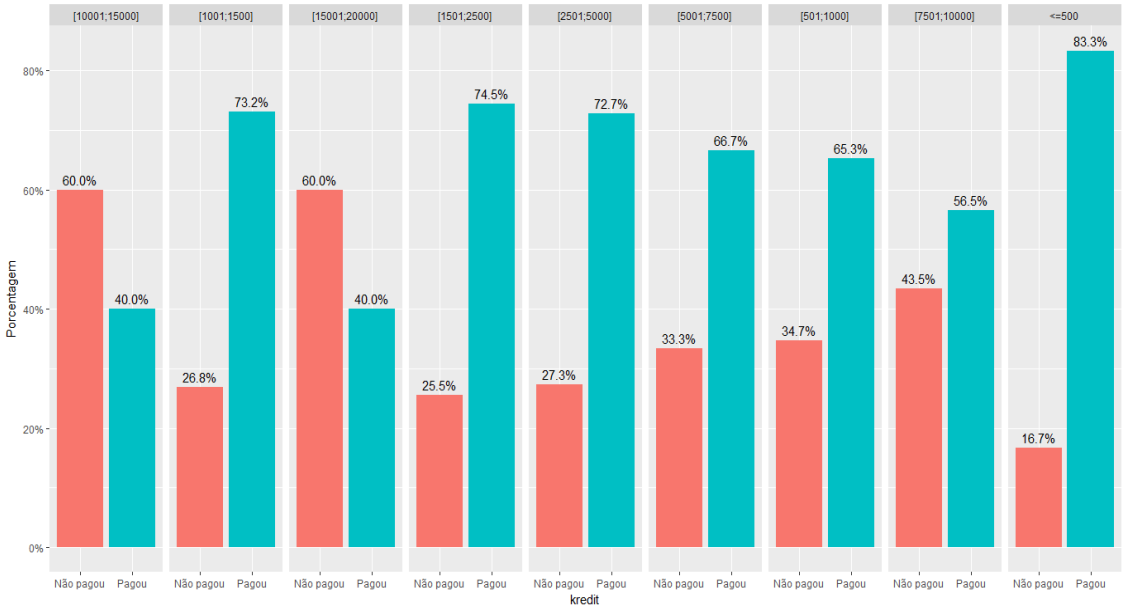
Fonte: Autoria própria (2018).

Tabela 12: Tabela cruzada entre montante_cat x kredit

		kredit		
		Crédito não foi reembolsado	Crédito reembolsado	Total
Valor do empréstimo em € (discretizada)	[15001;20000]	3	2	5
	[10001;15000]	21	14	35
	[7501;10000]	20	26	46
	[5001;7500]	34	68	102
	[2501;5000]	75	200	275
	[1501;2500]	59	172	231
	[1001;1500]	51	139	190
	[501;1000]	34	64	98
	<=500	3	15	18
Total		300	700	1000

Fonte: Autoria própria (2018).

Figura 6: Gráfico bivariado entre montante_cat x kredit



Fonte: Autoria própria (2018).

- Variável: **poupança**

Mais da metade dos clientes, 78,6%, possuem menos que 100 DM em conta na data da aplicação do questionário ou não detinham conta poupança. Enquanto que os 21,4% demais apresentam uma reserva maior ou igual a 100 DM.

Tabela 13: Tabela de frequências da variável poupança

		Frequência	Porcentual	Porcentagem válida	Porcentagem acumulativa
Válido	Sem poupança	603	60,3	60,3	60,3
	<100 DM	103	10,3	10,3	70,6
	[100;500) DM	63	6,3	6,3	76,9
	[500;1000) DM	48	4,8	4,8	81,7
	[1000;+] DM	183	18,3	18,3	100,0
	Total	1000	100,0	100,0	

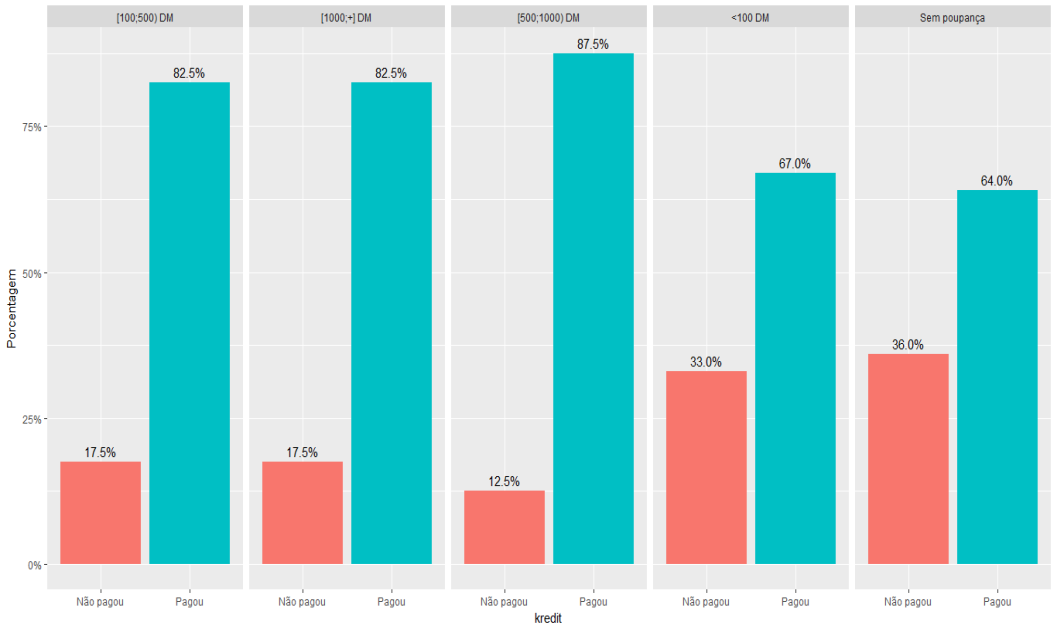
Fonte: Autoria própria (2018).

Tabela 14: Tabela cruzada entre poupança x kredit

		kredit		
		Crédito não foi reembolsado	Crédito reembolsado	Total
Conta poupança ou títulos	Sem poupança	217	386	603
	<100 DM	34	69	103
	[100;500) DM	11	52	63
	[500;1000) DM	6	42	48
	[1000;+] DM	32	151	183
Total		300	700	1000

Fonte: Autoria própria (2018).

Figura 7: Gráfico bivariado entre poupança x kredit



Fonte: Autoria própria (2018).

- Variável: **n_empréstimos_anteriores**

Aqui o cliente relata quantos empréstimos anteriores já fez, incluindo o atual.

Tabela 15: Tabela de frequências da variável n_emprestimos_anteriores

		Frequência	Porcentual	Porcentagem válida	Porcentagem acumulativa
Válido	1	633	63,3	63,3	63,3
	2 ou 3	333	33,3	33,3	96,6
	4 ou 5	28	2,8	2,8	99,4
	6 ou +	6	,6	,6	100,0
	Total	1000	100,0	100,0	

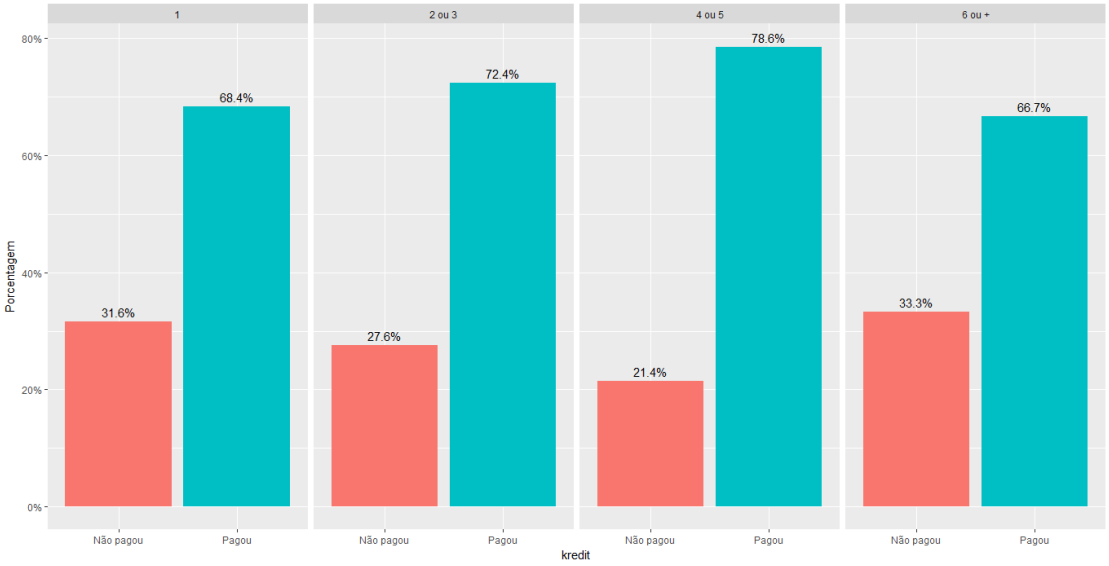
Fonte: Autoria própria (2018).

Tabela 16: Tabela cruzada entre n_emprestimos_anteriores x kredit

		kredit		
		Crédito não foi reembolsado	Crédito reembolsado	Total
Número de empréstimos parcelados anteriores no banco (incluindo o atual)	1	200	433	633
	2 ou 3	92	241	333
	4 ou 5	6	22	28
	6 ou +	2	4	6
Total		300	700	1000

Fonte: Autoria própria (2018).

Figura 8: Gráfico bivariado entre n_emprestimos_anteriores x kredit



Fonte: Autoria própria (2018).

- Variável: **porcentagem_emprestimo**

Esta variável representa o valor do empréstimo em termos do percentual da renda total do cliente.

Tabela 17: Tabela de frequências da variável porcentagem_emprestimo

Taxa em % de rendimento disponível

		Frequência	Porcentual	Porcentagem válida	Porcentagem acumulativa
Válido	[35;+]	136	13,6	13,6	13,6
	[25;35)	231	23,1	23,1	36,7
	[20;25)	157	15,7	15,7	52,4
	<20	476	47,6	47,6	100,0
	Total	1000	100,0	100,0	

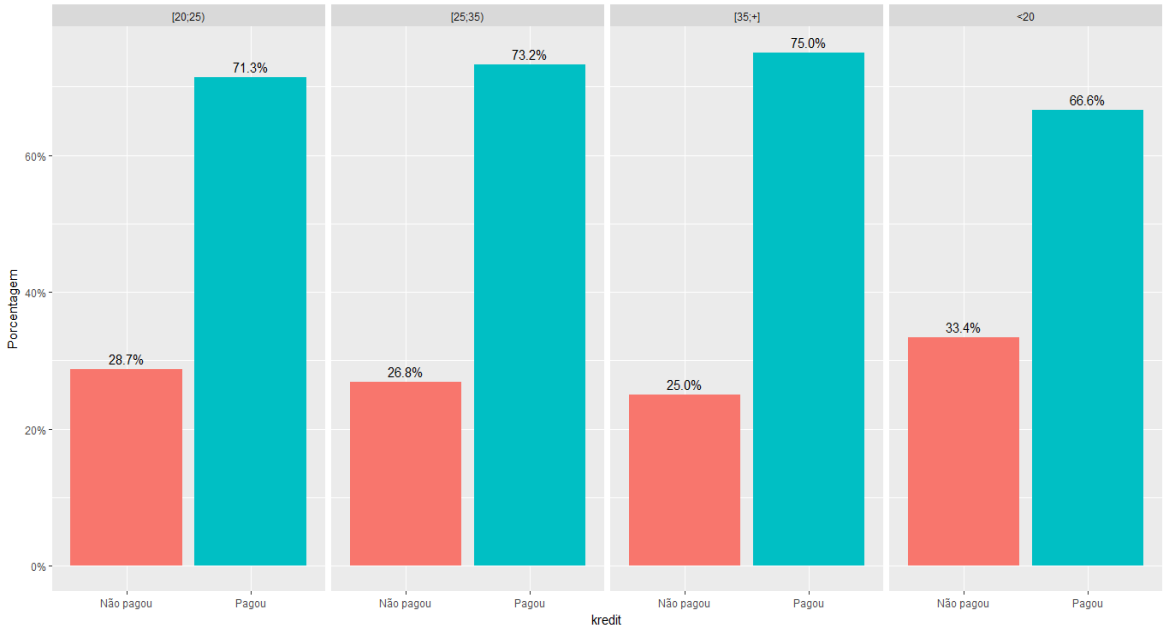
Fonte: Autoria própria (2018).

Tabela 18: Tabela cruzada entre porcentagem_emprestimo x kredit

		kredit		
		Crédito não foi reembolsado	Crédito reembolsado	Total
Taxa em % de rendimento disponível	[35;+]	34	102	136
	[25;35)	62	169	231
	[20;25)	45	112	157
	<20	159	317	476
Total		300	700	1000

Fonte: Autoria própria (2018).

Figura 9: Gráfico bivariado entre porcentagem_emprestimo x kredit



Fonte: Autoria própria (2018).

- Variável: **outros_credores**

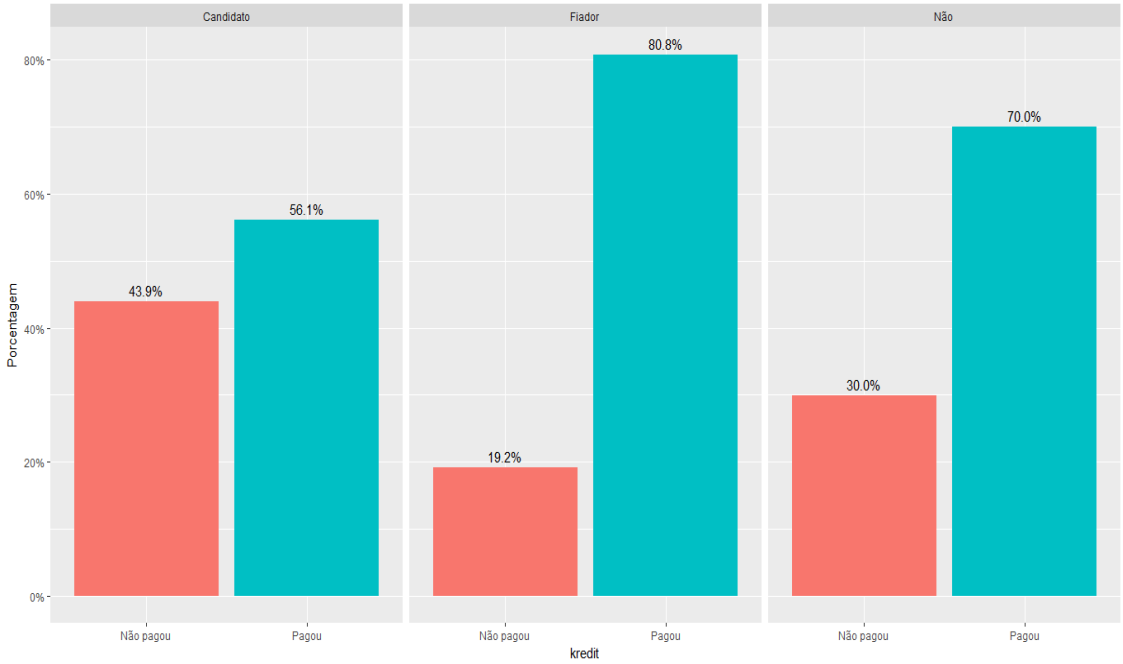
Mais de 90% dos clientes não possuem nenhum outro credor, enquanto que 4,1% relatam ter pelo menos um candidato a avalista e os 5,2% demais um fiador.

Tabela 19: Tabela cruzada entre outros_credores x kredit

		kredit		
		Crédito não foi reembolsado	Crédito reembolsado	Total
Outros garantes envolvidos	Não	272	635	907
	Candidato	18	23	41
	Fiador	10	42	52
Total		300	700	1000

Fonte: Autoria própria (2018).

Figura 10: Gráfico bivariado entre outros_credores x kredit



Fonte: Autoria própria (2018).

- Variável: **bens_solicitante**

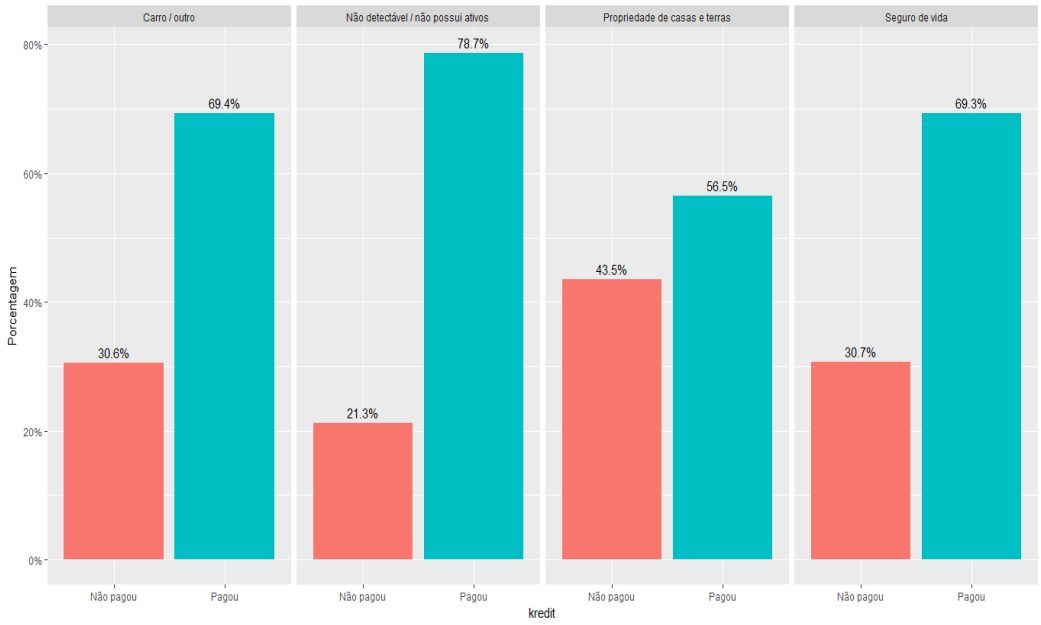
Aqui o cliente declara o bem de maior valor que possui como forma de garantia para obter o empréstimo, no entanto, 28,2% deles relatam não possuir nenhum bem e 23,2% colocam o carro, 15,4% a própria casa ou terras e os outros 33,2% o seguro de vida.

Tabela 20: Tabela cruzada entre bens_solicitante x kredit

		kredit		
		Crédito não foi reembolsado	Crédito reembolsado	Total
Ativos disponíveis	Não detectável / não possui ativos	60	222	282
	Carro / outro	71	161	232
	Seguro de vida	102	230	332
	Propriedade de casas e terras	67	87	154
Total		300	700	1000

Fonte: Autoria própria (2018).

Figura 11: Gráfico bivariado entre bens_solicitante x kredit



Fonte: Autoria própria (2018).

- Variável: **tempo_trabalho**

Nesta variável, o cliente responde há quanto tempo trabalha com o empregador atual ou se está desempregado.

Tabela 21: Tabela de frequências da variável tempo_trabalho

		Frequência	Porcentual	Porcentagem válida	Porcentagem acumulativa
Válido	Desempregado	62	6,2	6,2	6,2
	Até 1 ano	172	17,2	17,2	23,4
	1 a 4 anos	339	33,9	33,9	57,3
	4 a 7 anos	174	17,4	17,4	74,7
	7 ou mais anos	253	25,3	25,3	100,0
	Total	1000	100,0	100,0	

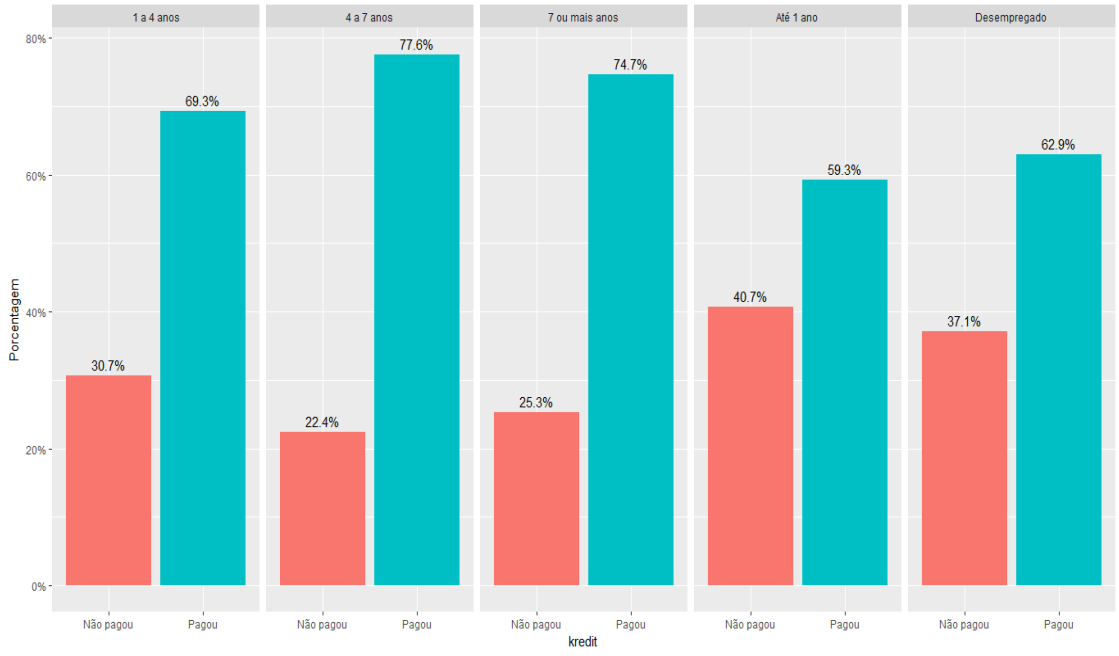
Fonte: Autoria própria (2018).

Tabela 22: Tabela cruzada entre tempo_trabalho x kredit

		kredit		
		Crédito não foi reembolsado	Crédito reembolsado	Total
Duração do emprego com o empregador atual	Desempregado	23	39	62
	Até 1 ano	70	102	172
	1 a 4 anos	104	235	339
	4 a 7 anos	39	135	174
	7 ou mais anos	64	189	253
Total		300	700	1000

Fonte: Autoria própria (2018).

Figura 12: Gráfico bivariado entre tempo_trabalho x kredit



Fonte: Autoria própria (2018).

- Variável: **tipo_apartamento**

O cliente relata qual é o tipo da atual moradia, se herdada/deixado de graça ou apartamento alugado ou próprio.

Tabela 23: Tabela de frequências da variável tipo_apartamento

		Frequência	Porcentual	Porcentagem válida	Porcentagem acumulativa
Válido	Deixado de graça	179	17,9	17,9	17,9
	Apartamento alugado	714	71,4	71,4	89,3
	Próprio	107	10,7	10,7	100,0
	Total	1000	100,0	100,0	

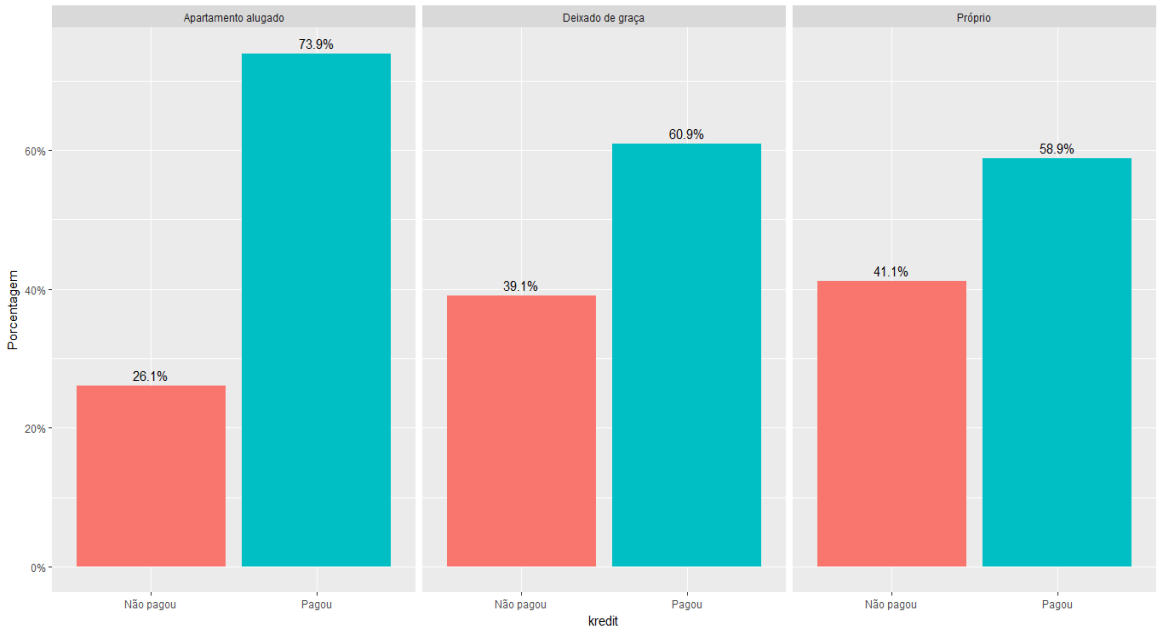
Fonte: Autoria própria (2018).

Tabela 24: Tabela cruzada entre tipo_apartamento x kredit

		kredit		
		Crédito não foi reembolsado	Crédito reembolsado	Total
Tipo de apartamento	Deixado de graça	70	109	179
	Apartamento alugado	186	528	714
	Próprio	44	63	107
Total		300	700	1000

Fonte: Autoria própria (2018).

Figura 13: Gráfico bivariado entre tipo_apartamento x kredit



Fonte: Autoria própria (2018).

- Variável: **emprestimos_add**

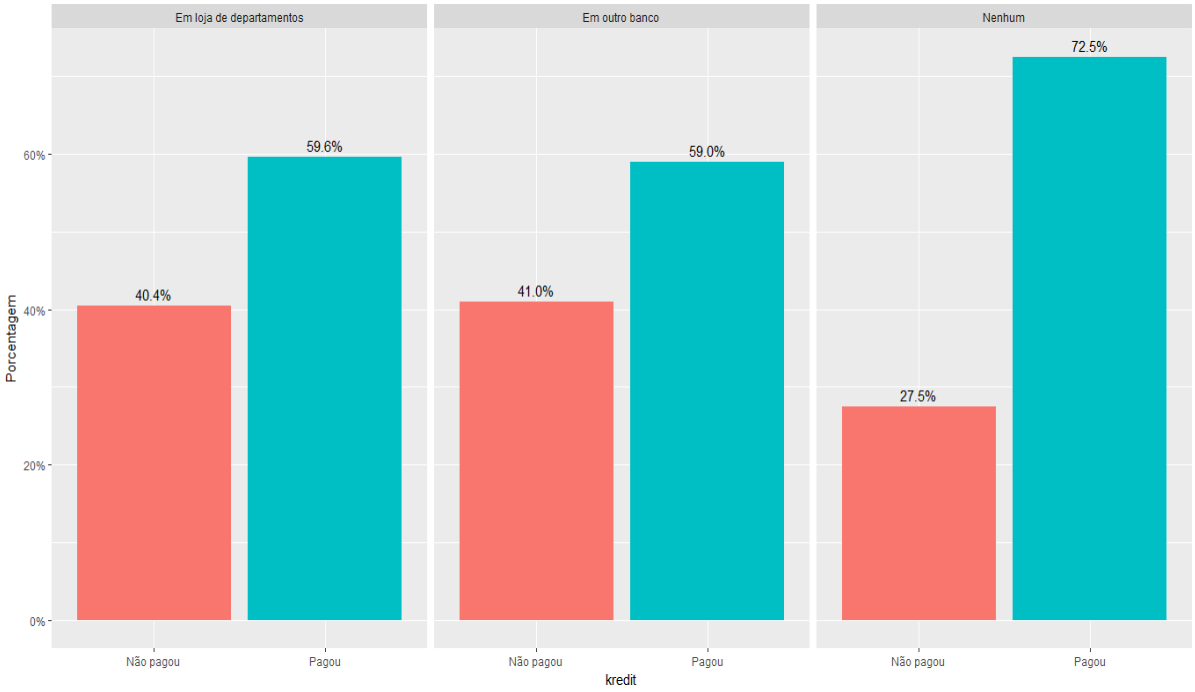
Nesta variável, 81,4% dos clientes relatam não possuir nenhum outro empréstimo adicional em outro banco e 13,9% já dizem o contrário, enquanto que 4,7% dos demais informam que possuem crédito imobiliário/em loja de departamento.

Tabela 25: Tabela cruzada entre emprestimos_add x kredit

		kredit		
		Crédito não foi reembolsado	Crédito reembolsado	Total
Empréstimos adicionais a prestações	Em outro banco	57	82	139
	Em loja de departamentos	19	28	47
	Nenhum	224	590	814
Total		300	700	1000

Fonte: Autoria própria (2018).

Figura 14: Gráfico bivariado entre emprestimos_add x kredit



Fonte: Autoria própria (2018).

- Variável: **mora_endereço**

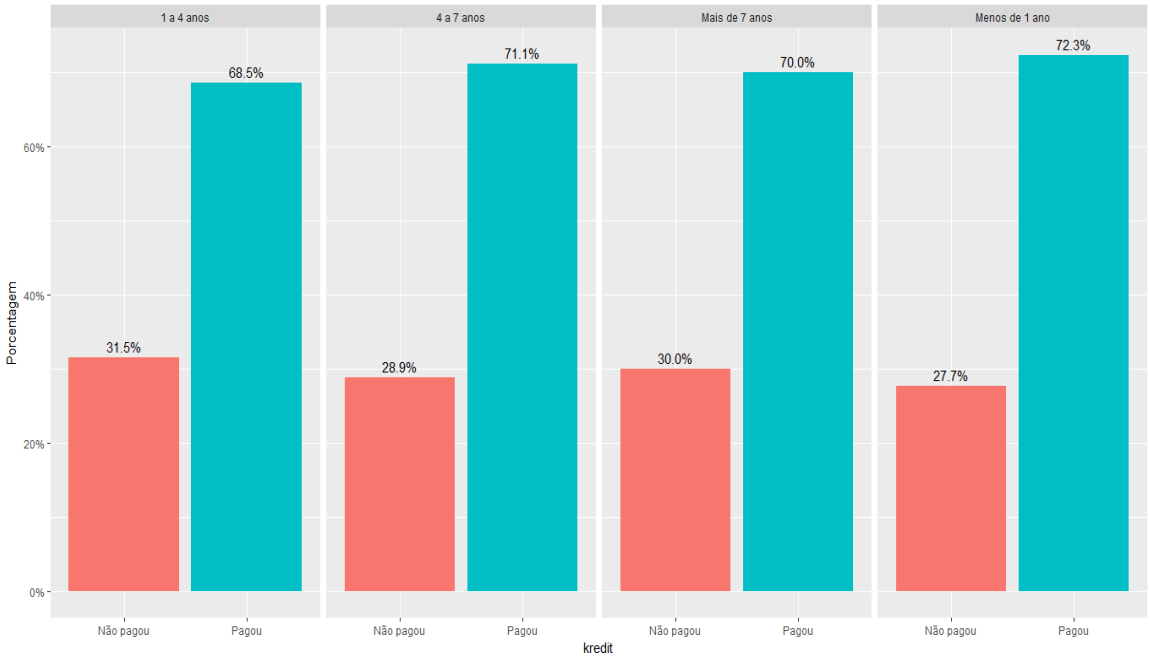
Esta variável busca informações sobre o tempo em que o cliente mora no atual endereço.

Tabela 26: Tabela cruzada entre mora_endereço x kredit

		kredit		
		Crédito não foi reembolsado	Crédito reembolsado	Total
Mora no atual endereço há quanto tempo	Menos de 1 ano	36	94	130
	1 a 4 anos	97	211	308
	4 a 7 anos	43	106	149
	Mais de 7 anos	124	289	413
Total		300	700	1000

Fonte: Autoria própria (2018).

Figura 15: Gráfico bivariado entre mora_endereço x kredit



Fonte: Autoria própria (2018).

- Variável: **sexo**

Aqui o cliente informa o sexo e sua situação conjugal. É possível perceber, no entanto, que não foi encontrado nenhum mulher solteira neste conjunto de dados e que, no entanto, a grande maioria é de homens solteiros.

Tabela 27: Tabela de frequências da variável sexo

		Frequência	Porcentual	Porcentagem válida	Porcentagem acumulativa
Válido	Homem divorciado/separado	50	5,0	5,0	5,0
	Mulher div./sep/casada	310	31,0	31,0	36,0
	Homem solteiro	548	54,8	54,8	90,8
	Homem casado	92	9,2	9,2	100,0
	Total	1000	100,0	100,0	

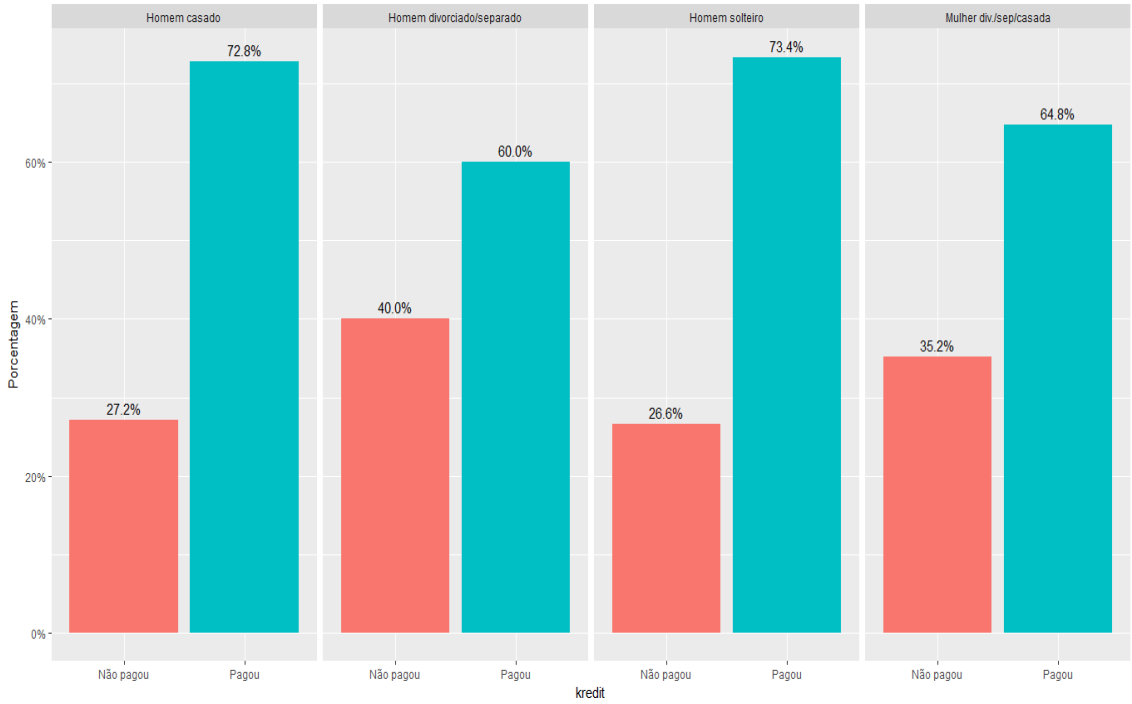
Fonte: Autoria própria (2018).

Tabela 28: Tabela cruzada entre sexo x kredit

		kredit		
		Crédito não foi reembolsado	Crédito reembolsado	Total
Estado civil e gênero	Homem divorciado/separado	20	30	50
	Mulher div./sep/casada	109	201	310
	Homem solteiro	146	402	548
	Homem casado	25	67	92
Total		300	700	1000

Fonte: Autoria própria (2018).

Figura 16: Gráfico bivariado entre sexo x kredit



Fonte: Autoria própria (2018).

- Variável**: idade_cat**

Representa a idade do cliente na data do pedido do crédito, categorizada.

Tabela 29: Tabela de frequências da variável idade_cat

		Frequência	Porcentual	Porcentagem válida	Porcentagem acumulativa
Válido	[0;25]	190	19,0	19,0	19,0
	[26;39]	511	51,1	51,1	70,1
	[40;59]	248	24,8	24,8	94,9
	[65;+]	23	2,3	2,3	97,2
	[60;64]	28	2,8	2,8	100,0
	Total	1000	100,0	100,0	

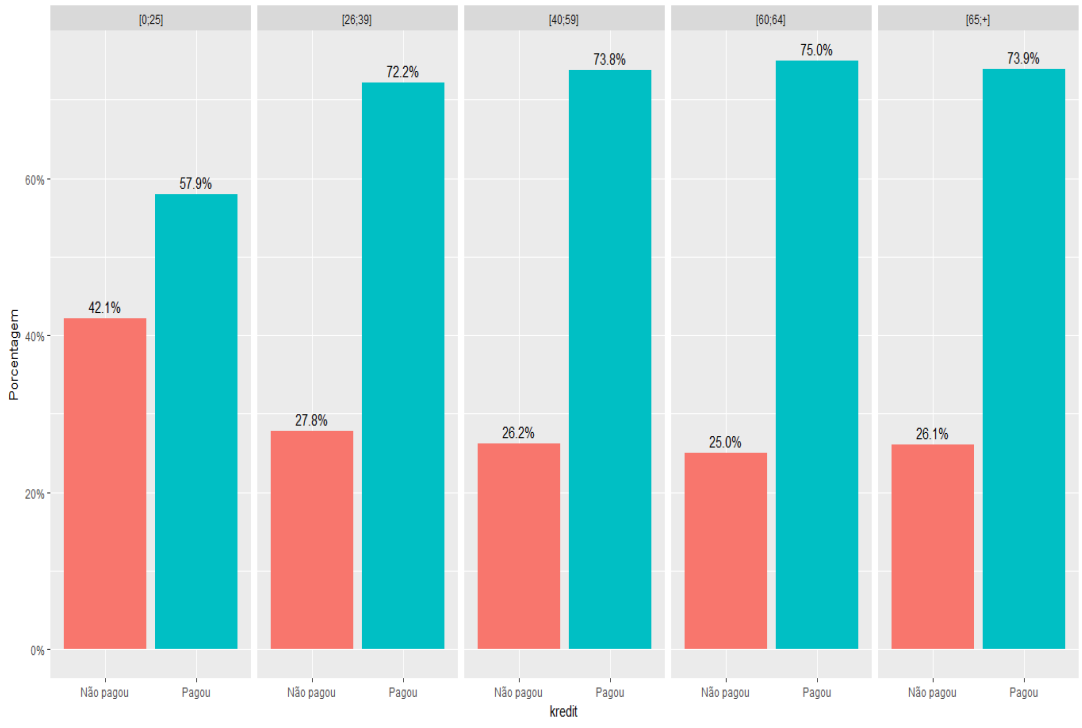
Fonte: Autoria própria (2018).

Tabela 30: Tabela cruzada entre idade_cat x kredit

		kredit		
		Crédito não foi reembolsado	Crédito reembolsado	Total
idade em anos	[0;25]	80	110	190
	[26;39]	142	369	511
	[40;59]	65	183	248
	[65;+]	6	17	23
	[60;64]	7	21	28
Total		300	700	1000

Fonte: Autoria própria (2018).

Figura 17: Gráfico bivariado entre idade_cat x kredit



Fonte: Autoria própria (2018).

- Variável: **ocupaçao**

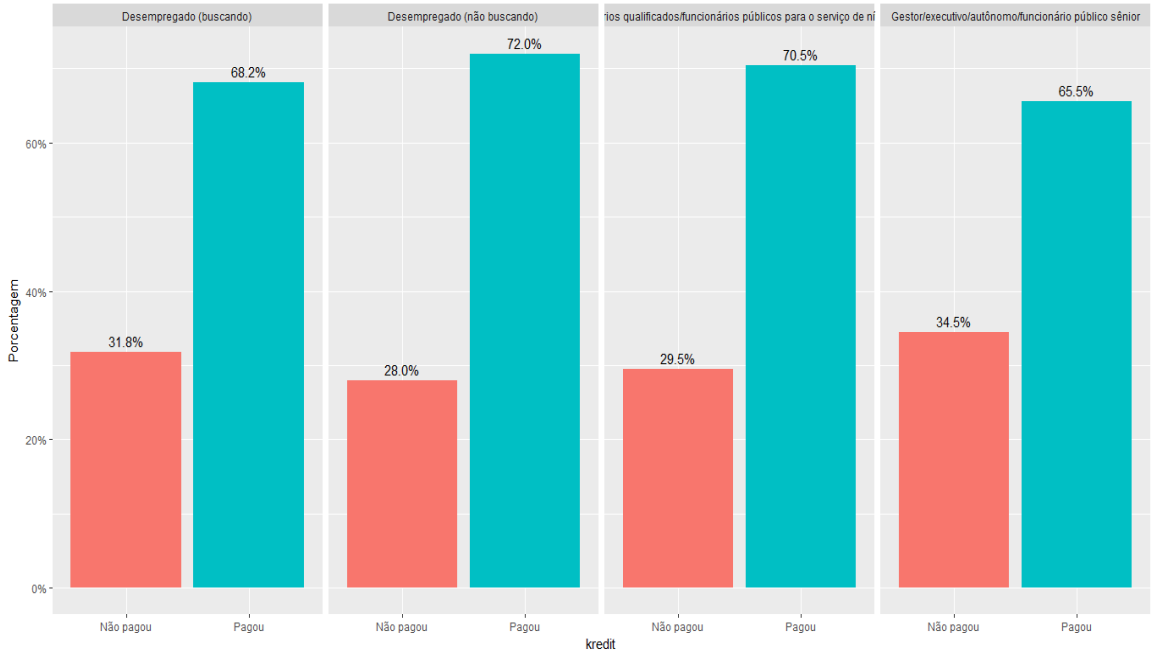
Mais da metade dos clientes, 77,8% relatam ser gestores (executivos) ou serem qualificados. Ao passo que 22,2% dos demais estão desempregados, sendo que destes, apenas 2,2% estão buscando por emprego.

Tabela 31: Tabela cruzada entre ocupaçao x kredit

		kredit		
		Crédito não foi reembolsado	Crédito reembolsado	Total
Ocupação atual do cliente	Desempregado (buscando)	7	15	22
	Desempregado (não buscando)	56	144	200
	Funcionários qualificados/funcionários públicos para o serviço de nível médio	186	444	630
	Gestor/executivo/autônomo/funcionário público sênior	51	97	148
Total		300	700	1000

Fonte: Autoria própria (2018).

Figura 18: Gráfico bivariado entre ocupaçao x kredit



Fonte: Autoria própria (2018).

- Variável: **numero_dependentes**

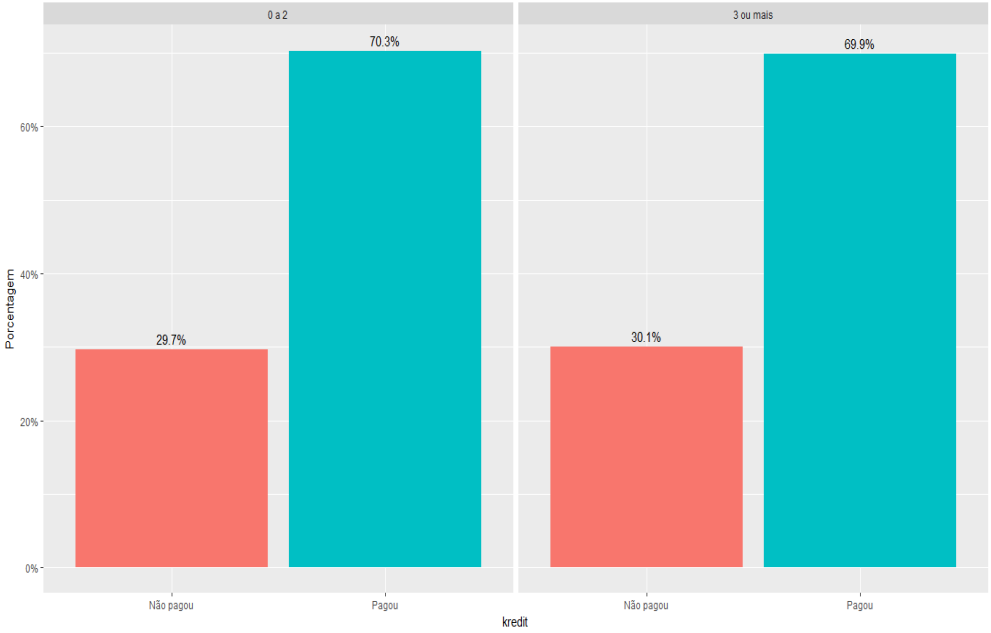
84,5% dos clientes afirmam ter 3 ou mais dependentes, enquanto que os 15,5% dizem ter dois ou menos.

Tabela 32: Tabela cruzada entre n_dependentes x kredit

		kredit		
		Crédito não foi reembolsado	Crédito reembolsado	Total
Número de pessoas dependentes	3 ou mais	254	591	845
	0 a 2	46	109	155
Total		300	700	1000

Fonte: Autoria própria (2018).

Figura 19: Gráfico bivariado entre n_dependentes x kredit



Fonte: Autoria própria (2018).

- Variável: **linha_telefone_fixa**

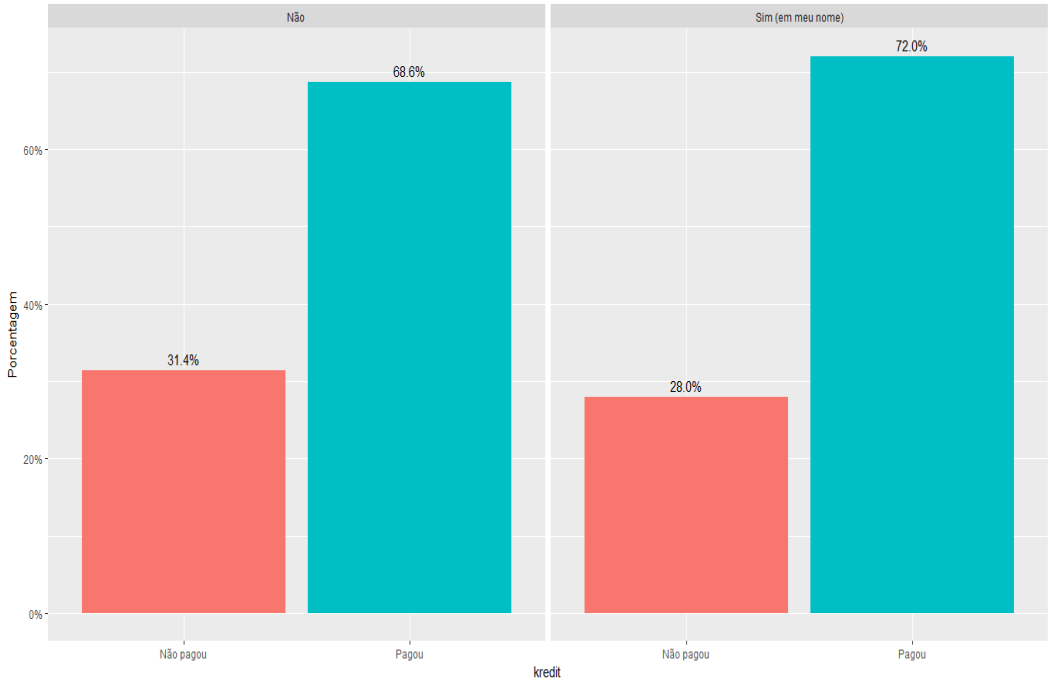
Mais da metade, 59,6% dos clientes declaram não possuir telefone fixo.

Tabela 33: Tabela cruzada entre linha_telefone_fixa x kredit

		kredit		
		Crédito não foi reembolsado	Crédito reembolsado	Total
linha telefonica fixa?	Não	187	409	596
	Sim (em meu nome)	113	291	404
Total		300	700	1000

Fonte: Autoria própria (2018).

Figura 20: Gráfico bivariado entre linha_telefone_fixa x kredit



Fonte: Autoria própria (2018).

- Variável**: imigrante**

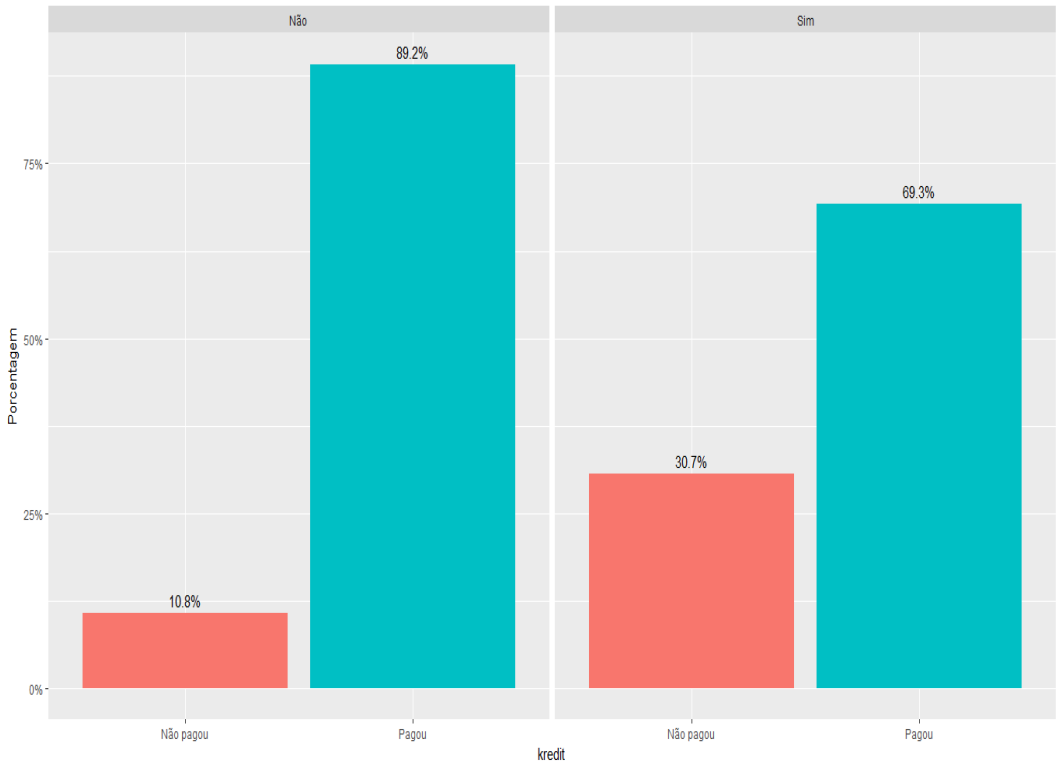
A grande maioria dos clientes dizem ser imigrantes, 96,3%.

Tabela 34: Tabela cruzada entre imigrante x kredit

		kredit		
		Crédito não foi reembolsado	Crédito reembolsado	Total
Trabalhador estrangeiro	Sim	296	667	963
	Não	4	33	37
Total		300	700	1000

Fonte: Autoria própria (2018).

Figura 21: Gráfico bivariado entre imigrante x kredit



Fonte: Autoria própria (2018).

4.3 CONSTRUÇÃO DOS MODELOS

Antes da aplicação direta da análise de classes latentes foi adotado como ponto de partida o modelo de regressão logística para a seleção das variáveis importantes a serem consideradas na LCA, discutido na subseção 3.2.4. Para tanto, foi ajustado um modelo de regressão logística de cada variável explicativa, separadamente, com a variável depende “kredit”, ou seja, foram realizados modelos de regressão logística simples para verificar a significância individual de cada variável. Para o desenvolvimento dos modelos aqui expostos, foi utilizado o software R Studio versão 1.1.453.

Dentre os modelos de regressão logística simples, apenas sete variáveis não apresentaram significância estatística ao nível de 5%, são elas: “porcentagem_emprestimo”, “linha_telefone_fixa”, “mora_endereço”, “ocupaçao”, “n_dependentes”, “sexo” e “n_emprestimos_anteriores”.

Em seguida, também foi realizado o modelo de regressão logística múltipla, de modo a considerar todas as variáveis explicativas, representado por “~.”, com intuito de verificar as variáveis com influências significativas sobre a variável dependente, levando em consideração até mesmo aquelas que não foram significativas no modelo de regressão logística simples, para verificar se essas ainda permanecem sem relevância.

A seguir o código no R e a matriz de confusão do modelo:

```
> modelo_logistico <- glm(credito ~., family = "binomial", data =
credito_tcc)
```

Tabela 35: Matriz de confusão do modelo de regressão logística

		Predito		
		Credito não reembolsado	Credito reembolsado	Total
Real	Credito não reembolsado	168	132	300
	Credito reembolsado	74	626	700
	Total	242	758	1000

Fonte: Autoria própria (2018).

Acurácia=79,40%, Sensibilidade= 56%, Especificidade= 89,43%, FP=44% e FN=10,57%.

É possível verificar que este modelo possui um valor expressivo quanto a sua especificidade, já que representa cerca de 90%, neste sentido, pode-se dizer que o modelo está prevendo bem os bons pagadores.

Para a seleção das variáveis a serem incluídas no modelo, foi utilizado o procedimento estatístico chamado *stepwise*, que consiste em adicionar e remover iterativamente preditores, no modelo preditivo, para encontrar o subconjunto de variáveis no conjunto de dados que resulta no modelo de melhor desempenho, ou seja, um modelo que reduz o erro de previsão. Além disso, foi escolhido o passo backward, como direção para a escolha das variáveis, que começa com todos os preditores do modelo (modelo completo), remove iterativamente os preditores que menos contribuem e para quando todos os preditores são estatisticamente significativos. Este procedimento usa o BIC como critério de ajuste para a seleção das variáveis.

Este é o comando do modelo final no R Studio:

```
> stepwise(modelo_logistico, direction = "backward", criterion = "BIC")
Step:  AIC=1099.38
credito ~ as.factor(imigrante) + as.factor(status_conta_corrente) +
    as.factor(duraçao_meses) + as.factor(emprestimo_anterior)
                                  Df Deviance    AIC
<none>                                 1030.3 1099.4
- as.factor(imigrante)              1   1037.5 1099.7
- as.factor(emprestimo_anterior)    4   1061.8 1103.2
- as.factor(duraçao_meses)          1   1044.0 1106.1
- as.factor(status_conta_corrente)  3   1138.7 1187.0

Call:  glm(formula = credito ~ as.factor(imigrante) +
as.factor(status_conta_corrente) + as.factor(duraçao_meses) +
as.factor(emprestimo_anterior),
    family = "binomial", data = credito_tcc)
```

Como resultado é possível verificar apenas quatro variáveis foram signicativas ao nível de 5%, são elas: “status_conta_corrente”, “duração_meses”, “empréstimo_anterior e “imigrante”, sendo que apenas a última delas é de caráter demográfico, o que nos indica ser uma possível covariável a ser testada no LCM.

Desta forma, estas quatro variáveis foram levadas para a análise de classes latentes via pacote poLCA no R, sendo que a variável “imigrante” foi inserida como covariável

no modelo, com intuito de verificar a influência desta variável de caráter demográfico. Sendo assim, o modelo a ser testado no poLCA será composto de 3 variáveis (“status_conta_corrente”, “duração_meses”, “empréstimo_anterior) e uma covariável: imigrante. Para isso, é necessário, inicialmente, testar o modelo sem a presença da covariável e posteriormente com ela, afim de verificar sua influência.

Além disso, foram utilizadas 2 classes latentes, tendo em vista que o interesse desta pesquisa é de distinguir e classificar dois perfis diferentes, que são: os pagadores e os não pagadores. Entretanto, a título de curiosidade também foram realizadas as análises para 3 classes latentes, porém a interpretação ficou confusa com o acréscimo desta classe, ou seja, não conseguiu garantir interpretações, impossibilitando conclusões mais claras e objetivas.

Três variáveis manifestas foram levadas para esta análise e a variável que descreve o status de aplicação ou seja, a variável “kredit” que representa bom ou mal pagador, além da covariável “imigrante”.

Desta forma, foram feitos 4 modelos, são eles:

```
1)  lca1 <- poLCA(cbind(emprestimo_anterior, duração_meses,
status_conta_corrente, kredit)~1, credito_tcc, nclass = 2,graphs = T)
2)  lca2 <- poLCA(cbind(emprestimo_anterior, duração_meses,
status_conta_corrente)~1, credito_tcc, nclass = 2,graphs = T)
3)  lca3 <- poLCA(cbind(emprestimo_anterior, duração_meses,
status_conta_corrente, kredit)~imigrante,credito_tcc, nclass =
2,graphs = T)
4)  lca4 <- poLCA(cbind(emprestimo_anterior, duração_meses,
status_conta_corrente)~imigrante, credito_tcc, nclass = 2,graphs = T)
```

Sendo que os dois primeiros modelos estão sem a covariável, representados no modelo por “~1” e os dois seguintes, descritos com “~imigrante”. Além disso, os modelos 1 e 3 contém a variável dependente “kredit”, por motivos de visualização gráfica, afim de verificar se a classificação esperada está de acordo com a classificação real. Enquanto que os modelos 2 e 4 não contém a presença desta.

4.4 RESULTADOS

Segue abaixo a lista com as variáveis que foram levadas em consideração para a criação do modelo no poLCA:

Tabela 36: Lista de variáveis consideradas no poLCA

Nome da Variável	Categorias
status_conta_corrente	1= "Nenhuma conta atual" 2= "Nenhum saldo da conta ou saldo devedor" 3= "[0;200) DM" 4= "[200,+) DM ou conta salarial por pelo menos 1 ano"
duração_meses	1= "Até 21 meses"; 2= "Mais que 21 meses";
emprestimo_anterior	1= "Muito ruim"; 2= "Ruim - Conta crítica"; 3= "Nenhum empréstimo até agora / todos os empréstimos anteriores pagos"; 4= "Bom - Empréstimos ainda existentes com o banco até agora perfeitamente"; 5= "Muito bom - Empréstimos anteriores liquidados no banco corretamente"
kredit	0 = "Crédito não reembolsado" 1= "Crédito reembolsado"
Covariável: imigrante	1= "Sim" 2= "Não"

Fonte: Autoria própria (2018).

Vale a pena destacar que a variável “kredit” foi inserida no modelo poLCA a título de comparação gráfica, ou seja, para possibilitar melhor visualização da distinção entre as classes no gráfico de probabilidades de resposta de item condicional por variável manifesta.

No pacote poLCA existem dois critérios de informação principais: critério de informação de Akaike (AIC) e critério de informação bayesiano (BIC), discutidos na subseção 3.3.3, que são especialmente úteis na comparação de modelos. O BIC é mais amplamente utilizado na LCA. Além disso, um modelo com um valor BIC mais

baixo é preferido do que um modelo com um valor BIC mais alto. Uma definição mais geral do BIC baseia-se o log-likelihood e o número de parâmetros.

O modelo 1 a seguir é composto das 3 variáveis escolhidas no stepwise, porém acrescido da variável dependente “kredit” e sem a covariável “imigrante”, a fim de verificar a qualidade da classificação e posteriormente a relevância desta covariável.

```
> f1 <- cbind(emprestimo_anterior, duração_meses, status_conta_corrente,kredit)~1
> lca1 <- poLCA(f1, credito_tcc, nclass = 2,graphs = T)
Conditional item response (column) probabilities,
 by outcome variable, for each class (row)
```

$emprestimo_anterior

	Bom - Empréstimos ainda existentes com o banco até agora perfeitamente	Muito bom - Empréstimos anteriores liquidados no banco corretamente	Muito ruim	Nenhum empréstimo até agora / todos os empréstimos anteriores pagos	Ruim - Conta crítica
class 1:	0.0859	0.1241	0.0979	**0.5784**	0.1138
class 2:	0.0890	0.3701	0.0136	**0.5079**	0.0194

$duração_meses

	Até 21 meses	Mais que 21 meses
class 1:	0.4408	**0.5592**
class 2:	**0.6493**	0.3507

$status_conta_corrente

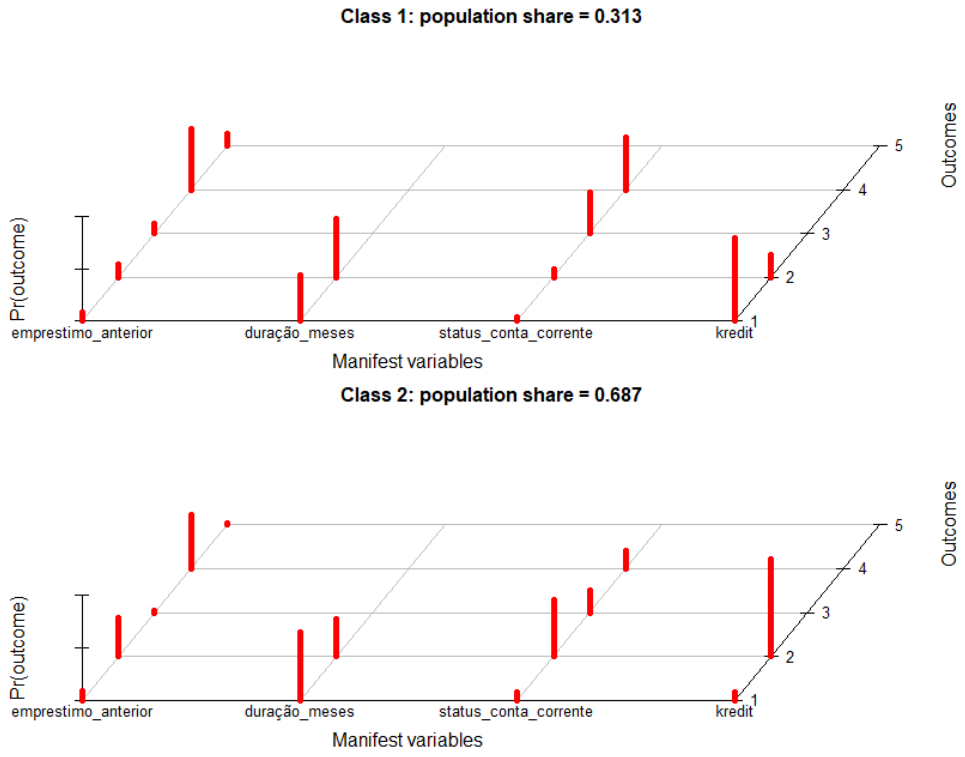
	[0;200) DM	[200,+) DM ou conta salarial por pelo menos 1 ano	Nenhum saldo da conta ou saldo devedor	Nenhuma conta atual
class 1:	0.0372	0.0739	0.3910	**0.4979**
class 2:	0.0748	**0.5400**	0.2133	0.1719

$kredit

	Crédito não foi reembolsado (não pagou)	Crédito reembolsado (pagou)
class 1:	**0.7900**	0.2100
class 2:	0.0765	**0.9235**

```
Estimated class population shares
 0.3133   0.6867

Predicted class memberships (by modal posterior prob.)
 0.282    0.718

=========================================================
Fit for 2 latent classes:
=========================================================
number of observations: 1000
number of estimated parameters: 19
residual degrees of freedom: 60
maximum log-likelihood: -3612.087

AIC(2): 7262.174
BIC(2): 7355.421
G^2(2): 107.1432 (Likelihood ratio/deviance statistic)
```

Figura 22: Gráfico de classificação com 2 classes, com a variável “kredit” e sem a covariável “imigrante”



Fonte: Autoria própria (2018).

A figura anterior representa a estimativa do modelo de classe latente com duas classes, obtido através do comando graphs = TRUE na função do pacote poLCA do R. Cada grupo de barras vermelhas representa as probabilidades condicionais, por classe latente, de serem classificadas por cada uma das categorias de variáveis. Barras mais altas correspondem a probabilidades condicionais mais próximas a 1 de uma determinada classificação.

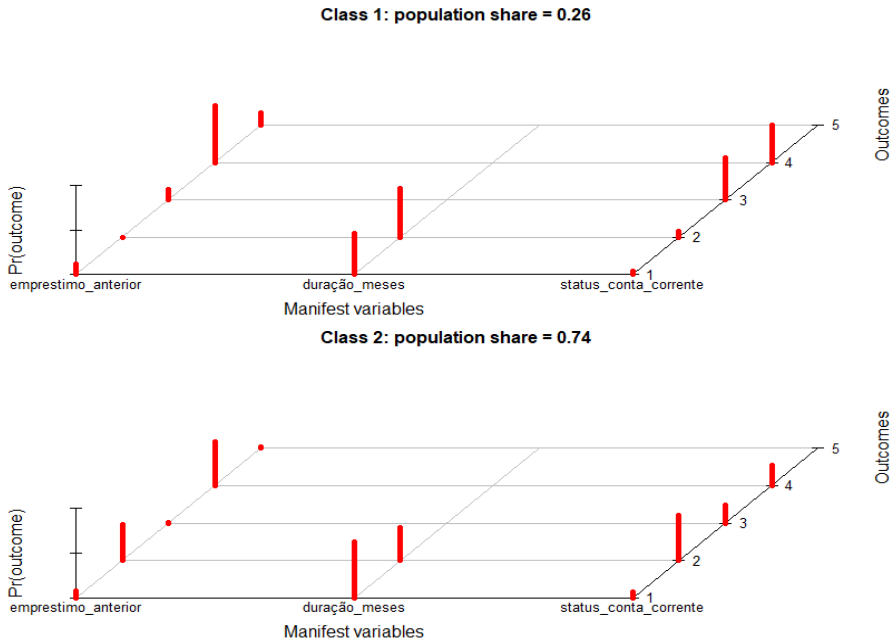
Esta figura é uma representação gráfica das probabilidades mostradas no modelo lca1. A partir dela é possível visualizar algumas semelhanças e, principalmente diferenciações. Como por exemplo, que a classe 2 representa os bons pagadores e a classe 1, os maus.

Como dito anteriormente, a variável “kredit” não faz parte do modelo selecionado por critério de ajuste de qualidade e foi incluída apenas com a finalidade visual/gráfica de verificar interpretação das classes. Logo, o modelo estimado sem esta é indicado a seguir.

```
> f2 <- cbind(emprestimo_anterior, duração_meses, status_conta_corrente)~1
> lca2 <- poLCA(f2, credito_tcc, nclass = 2,graphs = T)
Conditional item response (column) probabilities,
 by outcome variable, for each class (row)

$emprestimo_anterior
         Bom - Empréstimos ainda existentes com o banco até agora perfeitamente  Muito bom - Empréstimos anteriores liquidados no banco corretamente  Muito ruim  Nenhum empréstimo até agora / todos os empréstimos anteriores pagos  Ruim - Conta crítica
class 1:                0.1162                                                  0.0000      0.1180                                                 0.6307                0.1351
class 2:                0.0781                                                  0.3959      0.0126                                                 0.4946                0.0188

$duração_meses
         Até 21 meses  Mais que 21 meses
class 1:       0.4600            0.5400
class 2:       0.6275            0.3725

$status_conta_corrente
         [0;200) DM  [200,+) DM ou conta salarial por pelo menos 1 ano  Nenhum saldo da conta ou saldo devedor  Nenhuma conta atual
class 1:     0.0421                                             0.0676                                0.4694              0.4209
class 2:     0.0703                                             0.5086                                0.1986              0.2224
```

```
Estimated class population shares

 0.2599 0.7401


Predicted class memberships (by modal posterior prob.)
 0.228 0.772

=========================================================
Fit for 2 latent classes:
=========================================================
number of observations: 1000
number of estimated parameters: 17
residual degrees of freedom: 22
maximum log-likelihood: -3089.122

AIC(2): 6212.244
BIC(2): 6295.676
G^2(2): 49.29355 (Likelihood ratio/deviance statistic)
```

Figura 23: Gráfico de classificação com 2 classes, sem a variável “kredit” e sem a covariável “imigrante”



Fonte: Autoria própria (2018).

É possível notar que mesmo sem a variável “kredit” o perfil se mantém, no sentido de que os bons pagadores pagam com prazos menores e possuem mais de 200 DM em

conta, enquanto que os maus pagadores demoram mais tempo para pagar e estão sem saldo em conta ou com o saldo devedor.

A seguir é mostrado o modelo lca3, em que contém a variável dependente “kredit” (com intuito de verificar se acontece alguma mudança de perfil quando inserida a covariável) e a covariável “imigrante” (selecionada no stepwise).

```
> f3 <- cbind(emprestimo_anterior, duração_meses, status_conta_corrente,kredit)~imigrante
> lca3 <- poLCA(f1, credito_tcc, nclass = 2,graphs = T)
Conditional item response (column) probabilities,
 by outcome variable, for each class (row)
```

$emprestimo_anterior

	Bom - Empréstimos ainda existentes com o banco até agora perfeitamente	Muito bom - Empréstimos anteriores liquidados no banco corretamente	Muito ruim	Nenhum empréstimo até agora / todos os empréstimos anteriores pagos	Ruim - Conta crítica
class 1:	0.0893	0.1231	0.0975	**0.5786**	0.1116
class 2:	0.0874	0.3719	0.0133	**0.5074**	0.0199

$duração_meses

	Até 21 meses	Mais que 21 meses
class 1:	0.4394	**0.5606**
class 2:	**0.6512**	0.3488

$status_conta_corrente

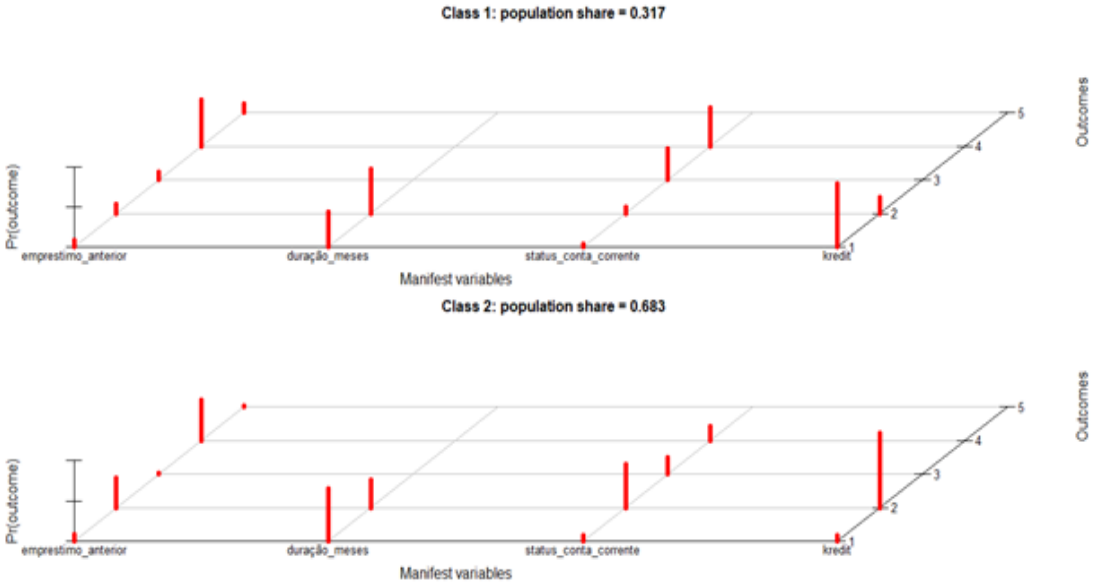
	[0;200) DM	[200,+) DM ou conta salarial por pelo menos 1 ano	Nenhum saldo da conta ou saldo devedor	Nenhuma conta atual
class 1:	0.0385	0.084	0.3922	**0.4854**
class 2:	0.0744	**0.538**	0.2118	0.1758

$kredit

	Crédito não foi reembolsado (não pagou)	Crédito reembolsado (pagou)
class 1:	**0.7939**	0.2061
class 2:	0.0706	**0.9294**

```
Estimated class population shares
 0.3172 0.6828

Predicted class memberships (by modal posterior prob.)
 0.284 0.716

=========================================================
Fit for 2 latent classes:
=========================================================
02/01/
             Coefficient  Std. error  t value  Pr(>|t|)
(Intercept)      2.26879     0.84482    2.686     0.009
imigranteSim    -1.54128     0.80944   -1.904     0.062
=========================================================
number of observations: 1000
number of estimated parameters: 20
residual degrees of freedom: 59
maximum log-likelihood: -3608.855

AIC(2): 7257.71
BIC(2): 7355.865
G^2(2): 104.6502 (Chi-square goodness of fit)
```

Figura 24: Gráfico de classificação com 2 classes, com a variável “kredit” e com a covariável “imigrante”

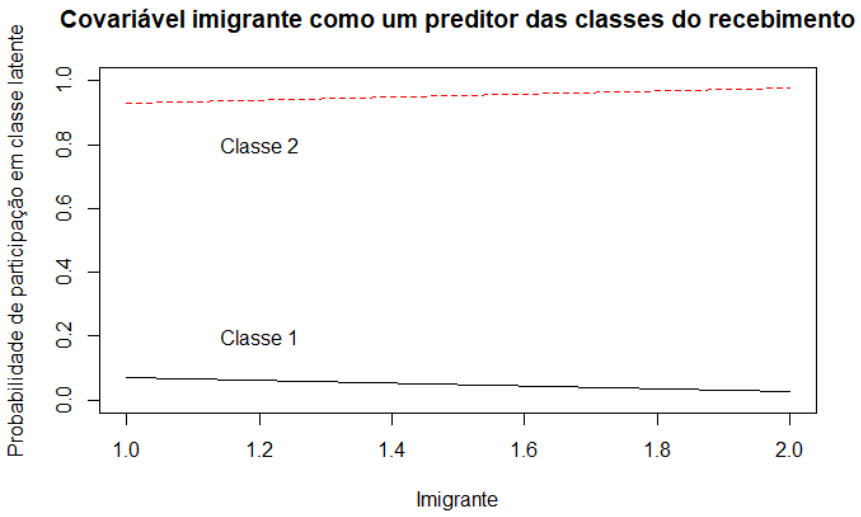


Fonte: Autoria própria (2018).

Além da informação para o modelo básico, a saída poLCA também inclui o coeficiente estimado nas covariáveis no modelo de regressão via LCA, mas que neste trabalho é a covariável “imigrante” e seus erros padrão.

A classe 2 ainda permanece sendo a classe dos bons pagadores e a classe 1 a dos maus pagadores. No entanto, a presença da covariável não foi significativa ao nível de 5%, ou seja, a presença da covariável “imigrante” parece não ter melhorado a interpretação e contribuição para o modelo. Fato verificado no gráfico abaixo de perfis de classes por categoria de covariável, pois as retas estão paralelas, indicando não apresentar discriminação entre as classes. Além disso, vale a pena ressaltar que mais de 90% dos clientes são imigrantes, ou seja, apesar de esta variável ter sido significativa no modelo de regressão logística, ela não foi significativo na LCA, no sentido de não conseguir ajudar na compreensão e interpretação do modelo.

Figura 25: Probabilidade de pertencer a classe latente com a covariável imigrante



Fonte: Autoria própria (2018).

E finalmente, o modelo lca4 sem a variável “kredit” e com a covariável “imigrante”:

```
> f4 <- cbind(emprestimo_anterior, duração_meses, status_conta_corrente)~imigrante
> lca4 <- poLCA(f1, credito_tcc, nclass = 2,graphs = T)
Conditional item response (column) probabilities,
 by outcome variable, for each class (row)
```

$emprestimo_anterior

	Bom - Empréstimos ainda existentes com o banco até agora perfeitamente	Muito bom - Empréstimos anteriores liquidados no banco corretamente	Muito ruim	Nenhum empréstimo até agora / todos os empréstimos anteriores pagos	Ruim - Conta crítica
class 1:	0.1313	0.0000	0.1099	**0.6256**	0.1332
class 2:	0.0692	0.4201	0.0097	**0.4885**	0.0125

$duração_meses

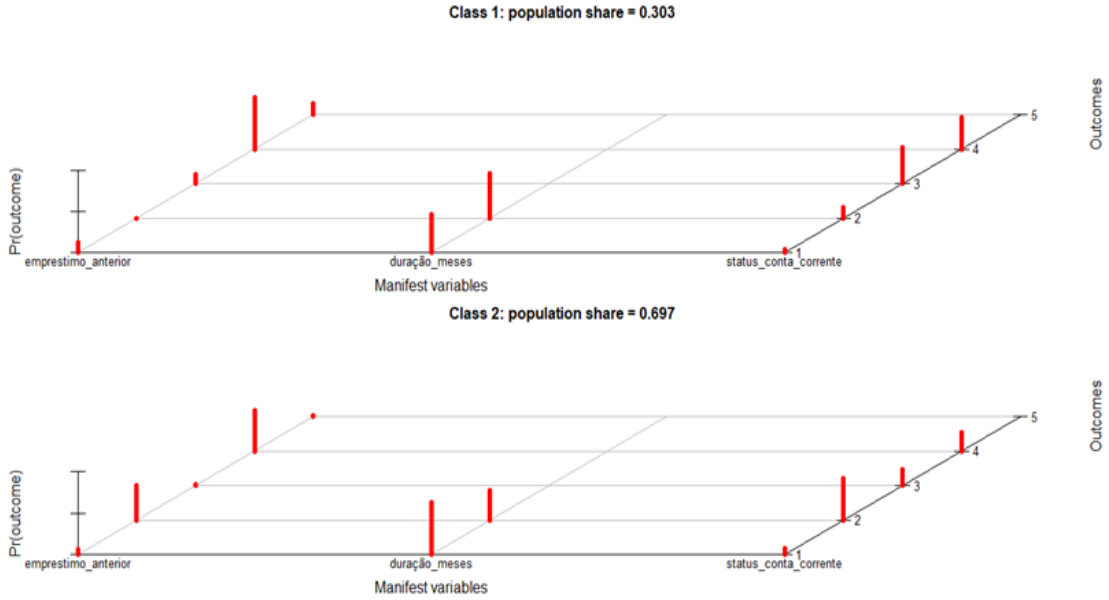
	Até 21 meses	Mais que 21 meses
class 1:	0.4608	**0.5392**
class 2:	**0.6374**	0.3626

$status_conta_corrente

	[0;200) DM	[200,+) DM ou conta salarial por pelo menos 1 ano	Nenhum saldo da conta ou saldo devedor	Nenhuma conta atual
class 1:	0.0452	0.1299	**0.4388**	0.3862
class 2:	0.0707	**0.5086**	0.1953	0.2253

```
Estimated class population shares

 0.3026 0.6974


Predicted class memberships (by modal posterior prob.)
 0.248 0.752

=========================================================
Fit for 2 latent classes:
=========================================================
02/01/
             Coefficient  Std. error  t value  Pr(>|t|)
(Intercept)      2.05227     1.08723    1.888     0.073
imigranteSim    -1.25141     0.91038   -1.375     0.184
=========================================================
number of observations: 1000
number of estimated parameters: 18
residual degrees of freedom: 21
maximum log-likelihood: -3087.568

AIC(2): 6211.136
BIC(2): 6299.476
G^2(2): 49.52707 (Chi-square goodness of fit)
```

Figura 26: Gráfico de classificação com 2 classes, sem a variável “kredit” e com a covariável “imigrante”



Fonte: Autoria própria (2018).

É possível notar que o perfil dos clientes não foi alterado com a presença da covariável “imigrante”, no sentido de que ela não foi significativa ao nível de 5% (p-valor= 0,184), ou seja, não foi relevante para o modelo inserir esta covariável. Porém vale ressaltar que o número de imigrantes neste conjunto de dados (963) era muito superior aos dos não imigrantes (37). Além disso, 29,6% dos imigrantes não pagaram o crédito recebido pela instituição credora, enquanto que 66,7% dos imigrantes pagaram e apenas 4% dos nativos (não imigrantes) deram calote (não pagaram).

À vista disso, segue abaixo uma tabela resumo com os AIC’s e BIC’s dos 4 modelos gerados no poLCA.

Tabela 37: Seleção do modelo LCA

Modelo	OBSERVAÇÃO	AIC	BIC
lca.1	Com kredit e sem covariável	AIC(2): 7262.174	BIC(2): 7355.421
lca.2	Sem kredit e sem covariável	AIC(2): 6212.244	BIC(2): 6295.676
lca.3	Com kredit e com covariável	AIC(2): 7257.71	BIC(2): 7355.865
lca.4	Sem kredit e com covariável	AIC(2): 6211.136	BIC(2): 6299.476

Fonte: Autoria própria (2018).

De acordo com a tabela acima, nota-se que o modelo que apresentou o menor BIC foi o lca.2, enquanto o modelo lca.4 foi o que apresentou o menor AIC. Porém, pelo princípio da parcimônia, o modelo selecionado foi o lca.2, porque não contém a variável dependente “kredit” e a covariável “imigrante”, visto que essa não foi significativa ao nível de 10%.

Desta forma, é possível descrever com mais profundidade o modelo lca.2:

- A primeira e segunda classes são descritas pelos tomadores de empréstimos, maus pagadores e bons pagadores, respectivamente.

- A primeira classe (25,99%) representa os tomadores de empréstimos que não possuem saldo em conta corrente no banco onde estão tentando obter o crédito ou estão com saldo devedor (46,94%) e a duração (prazo) superior a 21 meses (54,00%), ou seja, são pessoas que já possuem um histórico devedor no banco e que demoram a pagar. Além disso, 63,07% desses clientes não receberam nenhum empréstimo até agora.

- Até 74,01% dos pedidos de crédito pertencem à segunda classe e esta, geralmente caracteriza os bons pagadores, cujo status de conta corrente existente é de conta salarial ou está acima de 200 DM (50,86%). Além disso, o tempo de duração do empréstimo é inferior a 21 meses (62,75%). Vale ressaltar que essas pessoas também não tiveram nenhum empréstimo recebido até agora ou então pagaram todos os empréstimos anteriores (49,46%), não tendo havido diferenciação de classes nesta variável (empréstimo anterior). Mas

resumidamente, pode-se dizer que foram clientes que honraram com o prazo de devolução do empréstimo, possuem uma maior quantia de dinheiro em conta corrente e que devolveram outros créditos recebidos, quando requerido.

- Portanto, há duas classes e seus perfis são distintos.

Apesar de serem apenas 3 variáveis manifestas para descrever o perfil dos clientes em análise de concessão de crédito, vale a pena ressaltar que o modelo conseguiu distinguir com clareza as duas classes latentes (bom e mau pagador).

5 CONCLUSÃO

Com o avanço das tecnologias e o rápido processamento das informações, a análise de concessão de crédito deixou de ser um mero julgamento humano, baseados em critérios subjetivos e passou a ser um meio obrigatório para os analistas, de modo mais objetivo, rápido e confiável, visando reduzir as perdas, que neste caso, são os prejuízos com a inadimplência.

Neste sentido, o desenvolvimento do presente estudo possibilitou uma análise de como o conteúdo de análise de classes latentes pode ser usado, em especial para a concessão de crédito, em que foi possível realizar uma aplicação acerca do uso dos recursos do LCA para classificar o perfil dos bons e maus pagadores. Além disso, também permitiu utilizar outras metodologias estatísticas, como a análise de regressão logística simples e múltipla para auxiliar na escolha das variáveis para a construção do modelo final.

Diante disso, determinar a inadimplência e o perfil dos maus e bons pagadores é de suma importância para qualquer instituição credora. No entanto, ter acesso a esses dados nem sempre é possível, tendo em vista que são dados confidenciais. Neste sentido, o banco de dados apresentado neste trabalho é uma das pouquíssimas bases de dados disponíveis, que é de um banco do Sul da Alemanha, em que possui um contexto e realidade diferente da nossa aqui no Brasil. Além disso, esses dados são desbalanceados, ou seja, possuem proporções diferentes de adimplentes e inadimplentes além de possuírem poucas variáveis explicativas, relativamente.

Via análise descritiva, foi possível verificar o perfil dos maus pagadores como sendo homens imigrantes solteiros, com idade entre 26 e 39 anos, que solicitaram entre 2501 e 5000 DM para outras finalidades (diferentes das categorias descritas na variável "motivo_emprestimo"), sem conta poupança, mas que colocaram o seguro de vida como garantia para obter o crédito, além disso, a grande maioria se diz profissionais qualificados e com 3 ou mais dependentes. No entanto, vale destacar que este perfil é baseado nas frequências observadas na base de dados, e ainda, possui um caráter exploratório.

Usando a mesma analogia, os bons pagadores são homens imigrantes solteiros, qualificados, com idade entre 26 e 39 anos, que trabalham com o mesmo empregador há pelo menos um ano e no máximo 4, que solicitaram entre 2501 e 5000 DM, para comprarem móveis ou equipamentos, mas que tinham mais de 200 DM em conta poupança e pagaram o empréstimo num prazo menor, e também colocaram o seguro de vida como garantia para obter o empréstimo.

De modo geral, via análise de classes latentes, é possível dizer que o perfil dos maus pagadores é caracterizado por ter um status de conta corrente devedor ou sem saldo, além de demorarem mais de 21 meses para pagar o crédito recebido.

Enquanto que os bons pagadores apresentam um perfil oposto ao apresentado acima, já que os clientes desta classe são pessoas que devolveram todos os créditos recebidos e no prazo inferior a 21 meses e com uma conta salarial há pelo menos um ano ou acima de 200 DM por ano.

Desta forma, é possível verificar o ganho que a análise de classes latentes trouxe para a classificação do perfil dos clientes sujeitos a concessão de crédito, visto que conseguiu discriminar de forma mais clara cada uma das classes, apesar de poucas variáveis manifestas (explicativas). E mais, a LCA possui um caráter confirmatório, ou seja, busca confirmar as hipóteses e fazer inferências, ao passo que a análise descritiva detém um perfil exploratório, e neste caso, essencialmente, bivariado, e apenas lança suposições sobre as hipóteses.

6 CONSIDERAÇÕES FINAIS

O objetivo deste trabalho foi usar a análise de classes latentes para verificar o grau de discriminação desta metodologia, apesar de possuir alguns critérios subjetivos, como por exemplo a escolha do número de classes.

No entanto, dada a importância do tema, torna-se necessário destacar que outras metodologias podem ser utilizadas para a exploração deste banco de dados, como por exemplo a validação cruzada, a análise discriminante, a análise de *cluster*, dentre outros. Aumentando, assim, as possibilidades de aplicação da proposta do trabalho e dos novos temas que se abrem a partir desse.

REFERÊNCIAS BIBLIOGRÁFICAS

AKAIKE, H. A new look at the statistical model identification. **IEEE Transactions on Automatic Control**, p. 716–723, 1974.

ANDERSON, T. W.; CARLETON, R. O. Sampling theory and sampling experiments in latent structure analysis. **Journal of the American Statistical Association**, January 1957.

BANFIELD, J. D.; RAFTERY, A. E. Model-based Gaussian and non-Gaussian clustering. **Biometrics**, p. 803-821, 1993.

BASTOS, R. R. Processo de elaboração de uma investigação quantitativa sobre o conhecimento financeiro de estudantes do ensino fundamental de escolas públicas. **Revista de Educação, Ciências e Matemática**, v. 6, n. 3, set/dez 2016.

BESS, A. L. **Risco operacional: análise de sobrevivência aplicada a dados de risco operacional.** Universidade de São Paulo. São Paulo. 2007.

BUSSAB, W. D. O.; MORETTIN, P. A. **Estatística básica.** 5ª. ed. São Paulo: Saraiva, 2006.

CAOUETTE, J. B. et al. **Gestão do Risco de Crédito:** O próximo grande desafio financeiro. São Paulo: Qualitymark, 1999.

CAOUETTE, J. B.; ALTMAN, E. I.; NARAYANAN, P. **Managing Credit Risk:** The Next Great Financial Challenge. New York: John Wiley & Sons, 1998.

CHUNG, H.; FLAHERTY, B. P.; SCHAFER, J. L. Latent class logistic regression: application to marijuana use and attitudes among high school seniors. **Journal of the Royal Statistical Society: Series A**, v. 169, p. 723–743, April 2006.

CHUNG, H.; PARK, Y.; LANZA,. Latent transition analysis with covariates: pubertal timing and substance use behaviours in adolescent females. **Statistics in Medicine**, v. 24, p. 2895–2910, August 2005.

COLLINS, L. M.; LANZA, S. T. **Latent class and latent transition analysis:** With applications in the social, behavioral, and health sciences. Hoboken: NJ: John Wiley & Sons, Inc., v. 718, 2010.

DANTAS, R. ; DESOUZA, S. A. **Modelo de risco e decisão de crédito baseado em estrutura de capital de informação assimétrica**. Pesquisa Operacional. Ceará, p. 263-284. 2008.

DEWILDE, C. The Multidimensional Measurement of Poverty in Belgium and Britain: A Categorical Approach. **Social Indicators Research**, v. 68, p. 331–369, September 2004.

DZIAK, J. J. et al. **Sensitivity and Specificity of Information Criteria**. Pennsylvania State University. Pennsylvania, p. 31. 2012.

FARIA, S. M. F. D. S. **Modelos de Mistura: Aplicações em Análise de Regressão**. Universidade do Porto. Porto, p. 291. 2006.

FARIA, V. **Estimação de Máxima Verossimilhança via Algoritmo EM**. Universidade Federal de Juiz de Fora. Juiz de Fora, p. 44. 2011. (CDU /NA).

FERNANDES, P. C. **Agências de viagem e grupos estratégicos: uma análise de classes latentes**. Insper Instituto de Ensino e Pesquisa. São Paulo, p. 49. 2013.

FINCH, W. H.; FRENCH, B. F. **Latent variable modeling with R.** New York: Routledge, 2015.

GENGE, E. A. A Latent Class Analysis of the Public Attitude Towards the Euro Adoption in Poland, v. 8, p. 427-442, 2014.

GHERARDI, C.; GHIELMETTI, S. Escoragem de Crédito: Metodologia que Identifica Estatisticamente o Risco de Crédito. **Tecnologia do Crédito**, São Paulo, v. 01, n. 02, Setembro 1997.

GIBSON, W. A. **"Applications of the mathematics of multiple-factor analysis to problems of latent structure analysis**. Ph.D. dissertation, Department of Psychology, University of Chicago. Chicago. 1951.

GIBSON, W. A. Three multivariate models: Factor analysis, latent structure analysis, and latent profile analysis. **Psychometrika**, 1959.

GOODMAN, L. A. The Multivariate Analysis of Qualitative Data: Interactions among Multiple Classifications. **Journal of the American Statistical Association**, v. 65, p. 226–256, 1970.

GOODMAN, L. A. **Exploratory latent structure analysis using both identifiable and unidentifiable models.** Biometrika: 61, v. 215–231, 1974.

GREEN, B. F. J. A general solution for the latent class model of latent structure analysis. **Psychometrika**, v. 2, n. 16, p. 151-166, June 1951.

HAGENAARS, J. A.; MCCUTCHEON, A. L. **Applied Latent Class Analysis**. Cambridge: Cambridge University Press, 2002.

HAUN, M. **Cognitive Computing Steigerung des systemishchen Intelligenzprofils**. [S.l.]: Springer Vieweg, 2014.

HENRY, N. W. Latent structure analysis. **In S. Kotz & N. L. Johnson (eds.), Encyclopedia of Statistical Sciences**, New York: Wiley, v. 4, p. 497–504, 1983.

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

KOOPMANS, T. C. **Identification problems in latent structure analysis**. University of Chicago. Cowles Commission Discussion Paper: Statistics, No. 360. Unpublished manuscript. 1951.

LANZA, S. T.; COLLINS, M. Pubertal timing and the stages of substance use in females during early adolescence. **Prevention Science**, v. 3, p. 69-82, Março 2002.

LANZA, S. T.; FLAHERTY, B. P.; COLLINS, L. M. Latent Class and Latent Transition Analysis. **In Schinka, J. A., Velicer, W. F. (Eds.), Handbook of psychology**, v. 2, n. research methods in psychology, p. 663-685, april 2003.

LANZA, T. et al. PROC LCA: A SAS Procedure for Latent Class Analysis. **Struct Equ Modeling**, v. 4, n. 14, p. 671–694, 2007.

LAZARSFELD, P. F.; DUDMAN, J. The general solution of the latent class case. The use of mathematical models in the measurement of attitudes. **Santa Monica: RAND Corporation**, 1951.

LAZARSFELD, P. F.; HENRY, N. W. **Latent structure analysis**. Boston: Houghton-Mifflin, 1968.

LINZER, D. A.; LEWIS, J. B. poLCA: An R Package for Polytomous Variable Latent Class Analysis. **Journal of Statistical Software**, v. 42, n. 10, June 2011.

LOUZADA, F.; DINIZ, C. **Modelagem Estatística Para Risco de Crédito**. Simpósio Nacional de Probabilidade e Estatística. João Pessoa, p. 178. 2012.

MADANSKY, A. Partitioning methods in latent class analysis. **Santa Monica, CA: RAND Corporation**, Paper P-1644, March 1959.

MADANSKY, A. Latent structures. **In D. L. Sills (ed.), International Encyclopedia of the Social Sciences**, New York: Macmillan and Free Press, v. 9, p. 33-38, 1968.

MCHUGH, R. B. Efficient estimation and local identification in latent class analysis. **Psychometrika**, n. 21, p. 331–347, 1956.

MCHUGH, R. B. Note on Efficient estimation and local identification in latent class analysis. **Psychometrika**, n. 23, p. 273–274, 1958.

MCLACHLAN, G.; PEEL, D. **Finite mixtre models**. Wiley. New York. 2000.

NETO, J. L. D. C.; SÉRGIO, R. S. G. **Análise de Risco e Crédito**. Curitiba: IESDE Brasil S.A., 2009.

NUNES, L. **Aplicação do modelo de Regressão Logística para o apoio à decisão de crédito**. Universidade Federal de Juiz de Fora. Juiz de Fora, p. 57. 2011. (CDU N/A).

SANTOS, J. O. D.; FAMA, R. Avaliação da aplicabilidade de um modelo de credit scoring com varíaveis sistêmicas e não-sistêmicas em carteiras de crédito bancário rotativo de pessoas físicas. **Contabilidade & Finanças**, [online], v. 18, n. 44, p. 105-117, 2007. ISSN ISSN 1519-7077.

SCHRICKEL, W. K. **Análise de Crédito:** concessão e gerência de empréstimos. 4ª. ed. São Paulo: Atlas, 1994.

SCHWARZ, G. Estimating the dimension of a model. **Annals of Statistics**, v. 6, n. 2, p. 461–464, 1978.

SENGER, L. J.; CALDAS JUNIOR, J. Análise de risco de crédito utilizando redes neurais artificiais. **Revista do Centro de Ciências da Economia e Informática da Universidade da Região da Campanha (URCAMP)**, Bagé, v. 5, n. 8, Agosto 2001. ISSN ISSN 1415-2061.

SILVA, A. H. **A Influência do Estilo de Vida nas Escolhas de Transporte: Uma Análise de Classes Latentes**. Universidade de Brasília. Brasília, p. 216. 2013. (ENC/FT/UnB).

STERN, H. S. et al. Using Mixture Models in Temperament Research. **International Journal of Behavioral Development**, v. 18, p. 407–423, September 1995.

TITTERINGTON, D.; SMITH, A.; MAKOV, U. Statistical Analysis of Finite Mixture Distributions. **John Wiley & Sons**, 1985.

WALKER, J. L. **Extended Discrete Choice Models: Integrated Framework, Flexible Error Structures, and Latent Variables**. Massachusetts Institute of Technology. [S.l.], p. 208. 2001.

WOLFE, J. H. **Object cluster analysis of social areas**. Tese de Doutorado. University of California. 1963.

WOLFINGER, R. Covariance structure selection in general mixed models. **Communications in Statistics - Simulation**, Ontario, v. 22, n. 4, p. 1079-1106, 1993.

APÊNDICE – ACRÔNIMOS

AIC Akaike Information Criterion (Critério de Informação de Akaike)

BIC Bayesian Information Criterion (Critério de Informação Bayesiano)

EM Expectation Maximization (Maximização da esperança)

LCA Latent Class Analysis (Análise de Classes Latentes)

LCM Latent Class Model (Modelo de Classes Latentes)

LCR Latent Class Regression (Regresssão de Classes Latentes)