Departamento de Matemática

Mestrado em Modelação Estatística e Análise de Dados

O endividamento das famílias: estudo de caso de uma empresa de recuperação de créditos

Dissertação de Mestrado sob orientação da Professora Doutora Andreia Dionisio
Universidade de Évora

Célia João da Cruz Marmelo

Évora, 2011

Departamento de Matemática

Mestrado em Modelação Estatística e Análise de Dados

O endividamento das famílias: estudo de caso de uma empresa de recuperação de créditos

Dissertação de Mestrado sob orientação da Professora Doutora Andreia Dionisio
Universidade de Évora

Célia João da Cruz Marmelo

Évora, 2011

AGRADECIMENTOS

À professora Andreia Dionisio, pela disponibilidade em transmitir os seus conhecimentos e pelas criticas construtivas desde os primeiros esboços desta dissertação. Agradeço também por todo o apoio, paciência e disponibilidade demonstrados durante a orientação deste trabalho.

À professora Ana Sampaio, por todas as sugestões de alterações com vista a melhorar este trabalho.

Aos meus pais, um especial obrigado, pois sem eles ao meu lado jamais teria tido força para continuar e ultrapassar todas as contrariedades. Foram eles a razão desta minha luta para atingir mais este objectivo. Obrigado por serem quem são e fazerem com que tudo na minha vida faça sentido.

À minha amiga Marilia, pois acredito que se não fossem as palavras de estímulo dela, ao longo deste trabalho, mesmo a centenas de quilómetros de distância, jamais o teria acabado.

Aos meus amigos, pelo grande apoio e motivação e às minhas colegas de trabalho, por serem muito mais que isso.

Índice

Índice de Tabelas

Índice de Figuras

O endívidamento das famílias: estudo de caso de uma empresa de recuperação de créditos

Resumo

Com a abertura do mercado de crédito aos consumidores em Portugal, multiplicaram-se as formas de crédito, as instituições que o concedem e os bens e serviços que podem ser adquiridos. Contudo, o constante recurso ao crédito ganhou contornos mais graves originando situações de sobreendívidamento nos portugueses.

O presente estudo analisa e define o perfil dos devedores de clientes de uma empresa de recuperação de crédito, tornando possível que esta defina estratégias de gestão das carteiras dos diversos clientes, de forma a maximizar os seus resultados.

Para o efeito são aplicadas várias técnicas: análise de clusters, análise discriminante e regressão logística.

Este estudo revelou, com base num conjunto de variáveis previamente seleccionadas, e após a aplicação na análise de clusters, que existem para o cliente A três grupos de devedores com características muito próprias - *Devedores jovens, cautelosos e mais avessos ao risco; Devedores atrevidos e conhecedores das ofertas do crédito* e *Devedores mais velhos e mais propensos ao risco.* Para o cliente B existem dois grupos de devedores - *Devedores mais velhos, menos cautelosos e mais propensos ao risco* e *Devedores jovens e mais cautelosos.*

Após a análise de clusters, recorreu-se à análise discriminante e à regressão logística, de forma a identificar quais as variáveis que contribuíam de forma significativa para definir o perfil do devedor. Concluiu-se que os devedores com maior número de dívidas, com maior antiguidade e com montantes mais elevados tendem a ser devedores menos racionais aumentando a probabilidade de incumprimento.

The families' indebtedness: case study about a recovery credits enterprise

Abstract

With the opening of the credit market to Portuguese consumers, the forms of obtaining credit, the institutions that grant that credit and the goods and services that can be purchased have multiplied. Nevertheless, the constant need of acquiring a credit originated serious situations of indebtedness.

This study analyses and defines the profile of client's debtors from the enterprise of credit recovery, promoting the opportunity to define strategies to manage portfolios of various clients in order to maximize their results.

The cluster analysis applied to Client A, and based on a set of variables previously selected, resulted in three groups of debtors - *young people, more cautious and risk averse; fearless debtors that are aware of credit offers and Older and more risk prone debtors*. For Client B, two debtors groups were determined - *older people that are less cautious and more risk prone; and young and more cautious debtors*.

In order to identify which are the variables that contribute to the definition of the debtor's profile, the discriminant analysis and logistic regression were applied. The results showed that the debtors with more debts, higher amounts and more antiquity of debts tend to be the less rational debtors, wich will increase the likelihood of default.

1. Introdução

O crédito aos consumidores é um fenómeno recente no nosso país, que data da década de noventa, sendo mais recente que na generalidade dos países europeus, particularmente os do norte (Marques e Frade, 2003).

A procura de crédito ao consumo é um fenómeno das sociedades modernas que está relacionado com aspectos económicos, sociais e culturais. Assim, os indivíduos que recorrem ao crédito de consumo caracterizam essa procura mediante a taxa de desemprego, rendimento das famílias, taxas de juro, poupança, taxa de inflação, quadro jurídico e fiscal, preferências, estilos de vida e valores culturais privilegiados (Marques *et al*, 2000).

O perfil do utilizador do crédito ao consumo verifica-se maioritariamente em indivíduos do sexo masculino, e o grupo etário que mais recorre ao crédito situa-se entre os 35-44 anos, seguindo-se o grupo dos 25-34 anos e finalmente o grupo dos 45-54 anos de idade. Relativamente às classes sociais, verifica-se um menor recurso ao crédito nos indivíduos pertencentes a classes sociais com menores rendimentos, em particular, domésticas, desempregados, reformados e estudantes. Por outro lado, os indivíduos pertencentes as classes médias/superiores são os que acedem com maior frequência ao crédito para aquisição de bens e serviços. No entanto, são os de meios mais desfavorecidos os que têm mais dificuldade em pagar as dívidas (Marques *et al*, 2000).

Os consumidores de créditos dirigem a sua procura para determinado tipo de bens e serviços como automóveis/motas, obras em casa, electrodomésticos, mobiliário e equipamentos informáticos, sendo os grupos etários mais jovens ou os mais velhos os que normalmente recorrem ao crédito automóvel e as mulheres recorrem mais ao crédito para obras em casa, mobiliário e electrodomésticos (Marques *et al*, 2000).

Nos últimos anos verificou-se em Portugal um aumento do crédito ao consumo, e para tal situação muito têm contribuído as instituições financeiras com bastante publicidade, com o objectivo óbvio de angariar clientes, dando-lhes assim a possibilidade de acesso ao crédito com mais frequência e também para a aquisição de bens cada vez mais diversificados (Marques e Frade, 2003).

Pode-se referir que o acréscimo de créditos ao consumo que existiram, estavam relacionados com o lazer, férias e despesas de saúde das famílias, o que reflecte a mudança do estilo de vida por parte dos indivíduos.

Este constante recurso ao crédito ganhou contornos mais graves quando as famílias se começaram a ver impossibilitadas de fazer face aos encargos dos créditos contraídos, gerando situações de endívidamento e até mesmo de sobreendívidamento. Em muitos casos, as pessoas acumularam três, quatro, cinco ou mais créditos em simultâneo (crédito à habitação, crédito automóvel, crédito férias, crédito para aquisição de electrodomésticos, etc.), e muitas vezes recorreram a créditos para pagar juros de créditos anteriormente contraídos, entrando numa verdadeira "espiral de endívidamento".

Este aumento de recurso ao crédito que se verificou nos últimos anos, tem vindo a diminuir nos últimos meses, no entanto o tema do endívidamento dos portugueses continua a ser bastante actual, pois esta crise que despoletou no ano de 2008, não foi de forma alguma passageira, e muitos são os efeitos que se estão a sentir, e que ainda se irão fazer sentir. O aumento das taxas de juro, número crescente de falências e consequente aumento do desemprego, são alguns dos efeitos que já se fazem sentir actualmente.

Apesar de este ser um tema bastante actual e que a todos deve preocupar, em Portugal, e nos outros Estados-membros, não há um sistema centralizado de recolha e tratamento de estatísticas regulares e desagregadas sobre o endívidamento e sobreendívidamento das famílias. A escassez de estatísticas pormenorizadas, sobre o endívidamento e o sobreendívidamento limita a investigação neste campo, quer para efeitos de diagnóstico da situação em cada país, quer para estabelecer comparações entre países e avaliar a dimensão destas questões a nível europeu ou mundial.

Também a existência de estudos nesta área é escassa, bem como no que diz respeito à recuperação do crédito, no entanto algumas referências dos autores anteriormente referidos, terão um contributo positivo na realização deste estudo, como é o caso de Marques *et al*. (2000), Santos (2007) e Frade (2003).

Neste estudo pretende-se analisar o perfil dos indivíduos que recorreram ao crédito e a outros produtos financeiros e que não conseguiram cumprir com os seus compromissos. Serão analisadas características como a idade, sexo, localidade, profissão, valor das dívidas, número de dias de incumprimento, número de dívidas.

Todos os casos em estudo, após o incumprimento foram encaminhados para uma empresa especializada em recuperação de crédito, a qual forneceu os dados para este estudo.

A escolha deste tema está relacionada com o facto de me encontrar de momento a trabalhar numa empresa de recuperação de crédito. Nesta empresa, existe um departamento de reporting e estatísticas, do qual faço parte, e que tem um papel importantíssimo na tomada de decisões de gestão. Este departamento tem como um dos principais objectivos, fornecer

regularmente informação aos diversos departamentos, sobre a gestão das carteiras de vários clientes, para que sejam adoptadas as melhores estratégias na recuperação das respectivas dívidas.

Numa primeira fase, pretende-se segmentar os devedores em grupos com características homogéneas recorrendo à análise de clusters. Seguidamente, serão utilizadas técnicas que irão permitir identificar quais as variáveis que melhor diferenciam os grupos criados - análise discriminante e regressão logística. Também será utilizada a análise de componentes principais que tem como objectivo representar as variáveis originais num número mais reduzido de componentes, não correlacionadas entre si.

Para a análise dos dados será utilizado o programa SPSS (versão 16.0).

O presente trabalho de investigação está organizado da seguinte forma:

No primeiro capítulo, apresenta-se um enquadramento geral do estudo onde se expõe o problema e os objectivos de investigação e uma breve descrição dos métodos de análise a utilizar.

Seguidamente, o Capítulo II revê a literatura existente, passando pelo impacto que a crise económica tem no recurso ao crédito, pela evolução do recurso ao crédito e pelas aplicações existentes das diversas técnicas estatísticas neste tipo de estudos, nomeadamente, análise de clusters, análise discriminante, análise factorial e regressão logística. São apresentados alguns estudos relevantes na área.

O terceiro capítulo apresenta a metodologia de trabalho, as várias técnicas a utilizar, apresentação das técnicas, conceitos, regras de decisão, interpretação de resultados e a validação dos mesmos.

O quarto capítulo apresenta e discute os dados em estudo, nomeadamente as amostras, as variáveis seleccionadas e os resultados obtidos e validados de acordo com os critérios de validação definidos e sua interpretação.

A dissertação termina com a apresentação das conclusões obtidas e das limitações sentidas ao longo da elaboração deste estudo, assim como propostas de investigação futuras.

2. Revisão bibliográfica

2.1. Evolução do recurso ao crédito e do endívidamento

Segundo Lewis (1992) o crédito ao consumo já existe há mais de 3000 anos, desde a época dos Babilónios. No entanto, a concessão de crédito aos consumidores particulares de uma forma massificada é um fenómeno dos últimos 50 anos. O aparecimento dos cartões de crédito na década de 60 permitiu aos consumidores comprarem praticamente tudo a crédito, desde um jantar, passando pela compra de um livro ou de um computador e acabando nas viagens.

Na Europa o crédito ao consumo, teve um crescimento rápido na segunda metade da década de 80, tendo-se mantido essa tendência em muitos países Europeus na década de 90 e nos primeiros anos do século XXI. Portugal não ficou imune a esta tendência, apesar de mais tardiamente, na década de 90 do século XX, provocando uma dinâmica significativa no consumo privado e um exponencial crescimento do endívidamento das famílias.

Num ambiente de optimismo generalizado, quanto ao desempenho da economia nacional, os consumidores portugueses procuraram recuperar de um atraso de décadas em relação aos restantes países da UE, intensificando o acesso a determinados bens e serviços sendo que a prioridade foi dada à aquisição da habitação.

Ao longo do século XX, multiplicaram-se as formas de crédito, as instituições que o concedem e os produtos que podem ser por ele adquiridos. Para alguns, o crédito passou a constituir uma forma de gestão corrente do orçamento familiar, sobretudo, através dos cartões de crédito, cujos riscos são conhecidos. Na sociedade actual, os indivíduos sempre tentaram, e cada vez mais, adoptar um estilo de vida característico de uma classe superior à sua, e o crédito ao consumo pode assumir essa função, conferindo algum status (Marques *et al.,* 2000).

Segundo o projecto de investigação dirigido por Catarina Frade – Projecto desemprego e endívidamento das famílias (2003), a procura por novos estilos de vida, associada a novos valores, em que as famílias privilegiam as despesas com habitação, transporte e comunicações, serviços de saúde, serviços culturais, cuidados de beleza, viagens e serviços de hotelaria e restauração, todas elas, aquisições frequentemente realizadas a crédito, em detrimento das despesas com bens alimentares, bebidas, vestuário e calçado e outros bens de primeira necessidade, tiveram um papel determinante na alteração dos padrões do consumo e endívidamento.

Outra das causas do aumento do endividamento das famílias, prende-se com o facto de terem surgido alterações recentes no mercado de trabalho, como o trabalho temporário, o teletrabalho e a diminuição das horas de trabalho, que consequentemente vêm alargar os tempos de consumo. Estes tempos de consumo estão muitas vezes associados a tempos de lazer, o que levou a haver adaptações da oferta comercial, de forma a combinar num mesmo espaço as duas vertentes, consumo e lazer.

Cada vez mais os espaços de consumo se tornaram espaços de lazer. Veja-se o exemplo dos centros comerciais, locais de eleição dos consumidores para passarem algum do seu tempo de lazer, aproveitando para fazer algumas compras. A expansão destes novos formatos comerciais tem uma forte influência no consumo. Assim, a oferta comercial teve que se ir modernizando e adaptando aos novos hábitos dos novos consumidores, mais atrevidos e exigentes. A este tipo de oferta, há que acrescentar outra mais recente – o comércio electrónico, cujo peso é difícil de medir, mas cujo impacto no crédito ao consumo poderá ser considerável, uma vez que as compras são feitas através do cartão de crédito (Frade *et al.*, 2003).

2.2. Sobreendividamento das famílias

Nos últimos anos tudo se tem conjugado, na sociedade portuguesa, para o crescente endividamento das famílias. No entanto, um novo risco espreita: o da ruptura financeira de consumidores que se endividaram excessivamente ou que viram os seus rendimentos diminuir por alguma razão.

Com a crise financeira em 2008, verificou-se uma tendência decrescente do acesso ao crédito ao consumo em quase todos os países. Também em Portugal, entre 2007 e 2008, o crédito à habitação sofreu uma queda, e os restantes créditos (crédito para a educação, viagens, crédito consolidado e equipamento de escritório) viram o seu peso aumentar na distribuição do crédito total das famílias. No entanto o crédito habitação ainda representava mais de dois terços do crédito das famílias.

Apesar de antecipar o rendimento, o crédito não o aumenta, e o prazer de consumir que proporciona no presente, implica quase sempre a restrição de consumos futuros (Marques *et al*., 2000). Numa situação de estabilidade financeira e laboral, o crédito permite melhorar a acessibilidade a determinados bens e serviços, contribuindo para o conforto das famílias. No entanto, há sempre o risco de algo correr mal, de existir um acontecimento na vida de um

devedor que o impeça de cumprir com os seus compromissos financeiros, originando em alguns casos, situações de sobreendívidamento ou insolvência. Algumas situações que podem originar estes casos, são por exemplo o desemprego, um divórcio, um acidente ou até mesmo uma doença, pois podem determinar a perda de rendimento, e/ou o aumento das despesas do indivíduo e do seu agregado familiar, no entanto é o desemprego que é referido como umas das causas mais salientes (Frade *et al*., 2003). O projecto coordenado por Catarina Frade teve como objectivo estudar os contornos da relação entre o desemprego e o sobreendívidamento, procurando determinar de que forma a perda do emprego e/ou precarização das condições laborais esteve na origem de situações de ruptura financeira das famílias portuguesas, que haviam recorrido ao crédito para habitação e consumo. Neste projecto, foi formulada a hipótese, de que em Portugal haveria uma emergência de casos de sobreendívidamento motivados pelo aumento de desempregados. Começou por ser feita uma caracterização base do problema, isto é, do endívidamento dos consumidores e da evolução da taxa de desemprego.

Posteriormente foram definidas duas amostras. Uma das amostras com indivíduos desempregados que, no momento em que perderam o emprego possuíam uma ou mais dívidas, bem como desempregados fabris com dívidas de crédito e a outra amostra compreendia consumidores sobreendívidados que solicitaram apoio à DECO, tendo em vista a renegociação das dívidas com os credores. Para esta segunda amostra pretendia-se perceber, qual a origem do sobreendívidamento, uma situação de desemprego actual ou anterior, do próprio ou de outro membro do agregado familiar, ou se estava relacionado com a deterioração das condições laborais.

Entre os dois grupos observou-se um perfil distinto quanto aos hábitos de consumo e aos padrões de endívidamento.

Os desempregados fabris apresentaram um consumo pouco intenso e diversificado, antes e durante a fase de desemprego e mais direccionado para responder às necessidades básicas. Manifestaram também uma grande aversão ao crédito e só consideravam socialmente aceitável o crédito para aquisição de habitação. Este perfil de consumo e endívidamento é típico de um contexto rural ou semi-urbano e com uma escala de valores mais conservadores, mais favorável à poupança e ao consumo moderado.

Quanto aos sobreendívidados da DECO afectados por uma situação de desemprego, constatou-se que os hábitos de consumo, pelo seu estilo de vida marcadamente urbano, se revelavam mais complexos e diversificados. O recurso ao crédito para aquisição de casa, carro e outros bens de consumo, fazia parte dos seus hábitos de gestão financeira, sendo frequentes

situações de multiendividamento. O crédito pessoal e o cartão de crédito serviam para fazer face a despesas correntes.

O "triângulo de risco de sobreendividamento" (vide Figura 1) é uma representação gráfica resultante da reflexão realizada no projecto desemprego e endividamento das famílias (Frade *et al*., 2003). Num dos vértices encontra-se o multiendividamento, noutro, uma perturbação grave da situação laboral e no terceiro, a fragilidade das poupanças e das redes sociais. É desta combinação destes três factores que pode imergir o fenómeno do sobreendividamento, surgindo uma "relação perigosa" entre desemprego e sobreendividamento e sobre a qual se deve reflectir, famílias, mercado e regulador público, de forma a prevenir e a saber lidar de forma consciente com este risco.

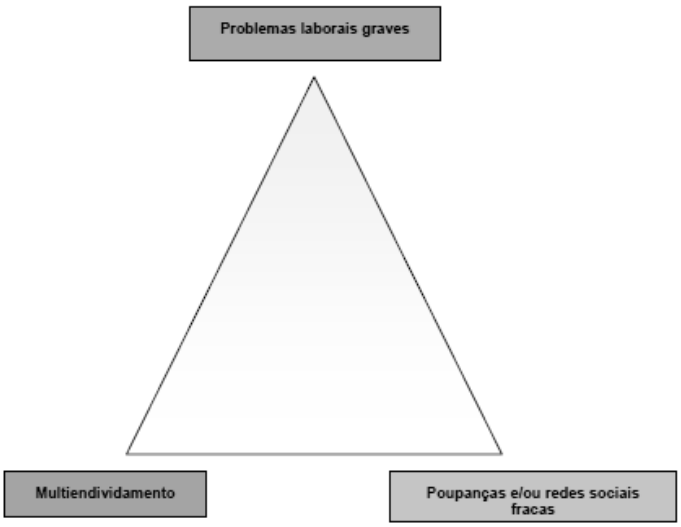

Figura 1: Triângulo de risco de sobreendividamento

Fonte: (Frade, 2003)

O Consumismo em Portugal tem-se vindo a verificar como uma atitude de consumir produtos ou serviços de forma pouco ponderada, devido à influência da publicidade exercida nos indivíduos;

- Actualmente os consumidores são avaliados tendo em consideração mudanças de âmbito socioeconómico, aumento do nível de escolaridade, a independência da mulher tanto a nível profissional como económico, maior acesso à informação, aumento da esperança de vida e a existência de mais tempo livre que facilita o acesso a um vasto leque de opções. Estas mudanças contribuíram para o aparecimento de um novo conceito de consumidor,

nomeadamente, consumidores mais exigentes, empreendedores, informados e com maior capacidade de decisão em tempo limite;

- No que respeita a Portugal, dispomos de estilos e de uma cultura de consumo que divide o indivíduo entre a fragmentação e a globalização. Assim, tem-se um país com menos poder de compra, relativamente aos respectivos parceiros europeus, uma economia de serviços com uma crescente dependência da globalização e das preocupações locais (Santos, 2007).
- A poupança não é uma prática recorrente para a maioria dos indivíduos, isto porque, a aquisição de créditos ao consumo está relacionada com a diminuição dos níveis de poupança, não havendo expectativas favoráveis, com uma evolução futura, para os rendimentos das famílias. Esta situação leva a que as famílias tenham um endividamento, que está acima dos seus rendimentos não dando lugar a perspectivas que tendam para a poupança (Marques *et al*, 2000).
- O fenómeno do crédito ao consumo é independente do sexo ou idade e actualmente existe um aumento de créditos ao consumo relacionados com o lazer, férias e despesas de saúde das famílias, que reflecte uma mudança do estilo de vida por parte dos indivíduos (Marques *et al*, 2000).
- O endividamento das famílias portuguesas está sobretudo relacionado com o crédito à habitação, verificado nas últimas décadas, com a liberalização do mercado, e com o surgimento de novas instituições bancárias, trazendo benefícios em termos de concorrência a quem quiser beneficiar de empréstimos. Contudo, os indivíduos estão mais prudentes pois existe mais acesso à informação sobre estes (Marques *et al*, 2000).

Tal como já foi referido inicialmente, o endividamento não é só um problema de Portugal, são vários os países, que se deparam com as mesmas dificuldades. A OCR Macro for DG Health & Consumer Protection (Directorate-General for Health & Consumer Protection) é uma comissão da União Europeia (UE) cujo trabalho consiste em garantir que todos os alimentos e bens de consumo que sejam vendidos na UE, são seguros e que o mercado interno da UE trabalha para o benefício dos consumidores e também para proteger e melhorar a saúde dos seus cidadãos. Em 2001 foi apresentado um relatório sobre o endividamento dos consumidores e algumas medidas propostas para avaliar o endividamento e o eventual sobreendividamento.

Numa primeira fase deste relatório (Rossi *et al*., 2001) são apresentadas as fontes de informação sobre o endividamento dos consumidores, tendo sido classificadas em três tipos: macro, micro e legais. As *Macro* foram obtidas através do sistema bancário, em inquéritos que são uma exigência quando se recorre ao crédito, e onde consta informação como os valores totais do empréstimos, o tipo de instituição de empréstimo, o tipo de empréstimo, o período de empréstimo, entre outras informações.

Micro refere-se a dados de inquéritos obtidos por amostragens periódicas que são realizadas em cada país da UE e também dados obtidos no ECHP (European Communities Household Panel) e EU HBS (EU harmonised Household Budget Survey). Segundo os autores do relatório estes dados fornecem informações qualitativas e quantitativas sobre uma amostra representativa da população de cada Estado-Membro da UE e encontram-se disponíveis a nível europeu.

As *Legais* consistem na recolha de informação nas várias fases do empréstimo ou após o incumprimento, para se obter informação sobre o número de processos que existem em diferentes fases do processo judicial, juntamente com o eventual resultado.

Posteriormente, são analisadas as informações básicas necessárias à construção de modelos adequados de endividamento do consumidor e sobreendividamento. Segundo os autores do relatório essas informações dizem respeito ao rendimento do consumidor, idade, as despesas de consumo, dívidas, bens (se disponível), os pagamentos efectuados e situação social.

Tal como tinha sido referido em estudos anteriores, também os autores deste relatório, definiram que tanto o endividamento, como o sobreendividamento são fenómenos naturais, que inevitavelmente afectam uma parte da população em qualquer altura e em qualquer circunstância económica.

O nível de endividamento em que se diz que uma família se torna excessivamente endividada depende de muitos fatores, entre os quais, o valor da dívida, a estrutura da dívida, o regime em que foi adquirido o crédito, a existência de bens e outras características pessoais e económicas da família, bem como factores externos, como o estado da economia do país (Rossi *et al*., 2001). Assim sendo, o conceito de endividamento, torna-se subjectivo, variando de família para família e de país para país, no entanto é importante ter em conta o rendimento do consumidor, idade e futuros rendimentos possíveis.

Foi então proposta a seguinte definição para sobreendívidamento:

"A person is over-indebted if he or she considers that he or she has difficulties in repaying debts, whether consumer debt or a mortgage."

Fonte: Rossi *et al* (2001)., "The problem of Consumer Indebtedness: Statistical Aspects \Consumer Indebtedness\ ORC Macro International Social Research", pag. 5.

Segundo esta definição uma família está sobreendívidada, se esta considera que tem dificuldade em pagar o valor das dívidas, ou seja, existe um relato da incapacidade em realizar o pagamento das dívidas, uma vez que estas têm um peso muito elevado nas suas despesas, e a família não encontra forma de cumprir com o acordado. Foi usada esta classificação, porque a maior parte das informações resultaram de pesquisas domiciliares, mas também porque ia ao encontro da literatura consultada aquando deste estudo.

Para analisar a relação entre o consumo e algumas variáveis explicativas, como o rendimento numa primeira fase e de seguida, as variáveis demográficas seleccionadas, foi proposto usar os dados Household Budget Surveys (HBS), cujos resultados combinados deram origem ao seguinte modelo de regressão:

$$\log(\mathrm{CON}_u) = \beta_0 + \beta_1(\mathrm{INC}_u) + \beta_2[\log(\mathrm{INC}_u)]^2 + \beta_3\mathrm{CH}_u + \beta_4\mathrm{AD}_u + \sum_{k=1}^{s}\beta_{k+4}\mathrm{AGE}_{ku} + \varepsilon_u \qquad (2.1.)$$

Onde:

CON - despesa total;

INC - é o rendimento líquido;

CH, AD e AGEk - variáveis demográficas que representam, respectivamente, o número de crianças, o número de adultos e 8 variáveis dummy que representam o género e a faixa etária do elemento do agregado familiar que foi tido como referência (Rossi et al., 2001).

Para a realização deste estudo, a amostra foi estratificada em seis categorias de idade (18-24, 25-34, 35-44, 45-54, 55-64 e 65 +), dentro de cada categoria calculou-se a percentagem de familias sobreendívidadas. Esta percentagem foi aplicada para 13 países, para a sua população real.

Os indicadores considerados foram os seguintes:

1. Rácio da dívida vs rendimento;
2. Rácio da dívida vs ativos;
3. Taxa de inadimplência de crédito
4. Média dos passivos por falência
5. Proporção de domicílios que se identificam como endívidados.

As várias tentativas para analisar a questão do endívidamento das famílias foram realizadas através de quatro medidas

- Famílias que têm empréstimos, com excepção de hipotecas;
- Familias que estão sobreendívidadas;
-Proporção de famílias com outros empréstimos de hipotecas que estão sobreendívidados;
- Individuos sobreendívidados

Segundo Rossi *et al.* (2001) estimou-se que em 1996 existiam aproximadamente 53 milhões de pessoas sobreendívidadas na UE, 18% da população total com mais de 18 anos, 16% das famílias da UE. Em todos os países, a percentagem de famílias com problemas de endívidamento é elevada, sendo que esta situação afecta todos os grupos etários e qualquer grupo de rendimento.

No entanto foram encontradas bastantes diferenças entre os vários estados membros da UE devido às diferenças existentes no desenvolvimento do crédito ao consumo, aos quadros jurídicos de cada país e também aos níveis de rendimento.

Neste estudo, onde foi estimado um modelo de regressão linear através do método dos mínimos quadrados e não foram encontradas evidências da verificação dos pressupostos Gauss-Markov deste método.

Num outro contexto Johansson *et al*. (2006) estudou o endívidamento das famílias suecas e a sua capacidade individual de pagamento das dívidas, de forma a analisar o risco de sobreendívidamento e as perdas no sector bancário. Além disso, analisou ainda o efeito dos choques macro-económicos, ou seja, a subida das taxas de juro e o aumento dos níveis de desemprego sobre o endívidamento das famílias.

Neste estudo (Johansson *et al*., 2006) também é realizada uma exposição detalhada de como o Riskbank[1] através dos Testes de Stress, utiliza dados micro para analisar a capacidade de endívidamento das famílias.

Na Suécia, a situação nos últimos anos tem levantado questões não somente sobre o impacto que a expansão acentuada do crédito poderá acarretar para a vulnerabilidade dos sectores doméstico e serviços bancários, mas também como o ambiente macroeconómico interno poderia ser afectado se este desenvolvimento cessasse.

A análise foi feita sobre os dados de riqueza e rendimento das famílias suecas em 2004. Para analisar a distribuição da dívida, rendimento, riqueza e capacidade de pagamento, as famílias foram divididas em 5 grandes categorias, de acordo com o seu rendimento, de forma a encontrar a vunerabilidade dos grupos, que sob stress podem originar perdas no sector bancário

As famílias que não possuíam nenhuma dívida, e portanto, são incapazes de provocar perdas com empréstimos, foram excluídas da análise.

Segundo Johansson *et al*. (2006) 18% das familias possuíam títulos e tinham rendimentos disponivel. Esta percentagem sobe em todas as categorias do rendimento, sendo que na maior 93% das familias possuíam títulos.

Enquanto o rácio da dívida total em 2004, se encontra um pouco acima de 120%, a categoria de maior rendimento tem um rácio superior a 190%

Todas as categorias de rendimento têm, em média, os activos a valer mais do que o dobro do valor dos seus passivos.

Segundo a investigação de Johansson *et al.* (2006) verificou-se que as diferenças também pode ser muito grandes dentro das categorias do rendimento. O grupo mais heterogêneo é o da categoria 1, pois trata-se de um grupo composto por indivíduos com situações de vida muito diferentes

As estatísticas mostraram que grande parte das famílias da 1ª categoria não têm emprego, rendimentos, activos ou passivos, e em média o rendimento desta categoria é muito baixo, sendo que muitas famílias têm dificuldade em pagar as contas.

Outra conclusão importante é que a maioria dos empréstimos são realizados pelas famílias de elevados rendimentos, bem como a maior parte dos reais e activos financeiros.

[1] Riskbank - Banco Central da Suécia, responsável pela política monetária com o objetivo de manter a estabilidade de preços e ao qual foi atribuido a tarefa de promover um sistema de pagamentos seguros e eficientes

Segundo os autores deste artigo (Johansson *et al*, 2006), os 20% das famílias de elevado rendimento representam 57% das dívidas e 44% do sector doméstico activo. Apenas 0,1% destas famílias foram consideradas vulneráveis.

A grande lição deste Teste de Stress é que o sector doméstico é muito mais sensível a aumentos na taxa de juros do que a mudanças nos níveis de desemprego. No entanto, nem mesmo um aumento acentuado das taxas de juros, combinado com as grandes quedas no valor dos activos reais, no sector doméstico, foi considerada capaz de gerar crédito de liquidação duvidosa no sector bancário (Johansson *et al*, 2006), no entanto o elevado endívidamento poderá dar origem a problemas individuais.

Nos artigos e estudos anteriores são analisadas as situações das famílias quanto ao endívidamento, situações que tendem a agravar-se devido a diversos factores externos, no entanto, este factores podem não afectar só as famílias mas também o governo desses países. Num artigo do International Journal of Mathematical Models and Methods in Applied Sciences, Smrčka (2011) explora as relações entre certos aspectos da dívida pública em vários países e o endívidamento das famílias nas modernas e desenvolvidas economias. Apesar das diferenças entre as economias nacionais dos estados modernos e a situação das famílias dos países desenvolvidos, existe uma série de padrões de comportamentos semelhantes que podem ser identificados entre os dois grupos.

É importante referir que as dívidas do governo são um importante fenómeno que tem crescido em grande escala somente nas últimas décadas, em particular na última metade do século. No entanto o endívidamento das famílias é um fenómeno mais recente, que têem vindo a ter um crescimento considerável nos últimos 25 anos (Smrčka, 2011).

Segundo Smrčka (2011) existem evidências significativas que comprovam que o comportamento das famílias, resulta do facto destas considerarem o endívidamento como aceitável e como um comportamento natural.

A análise de correlação dos rácios da dívida realizada nos vários países seleccionados e suas famílias, demonstrou forte depedência entre a evolução da dívida pública e suas famílias na Hungria e na República Checa. No entanto, não há evidências significativas que demonstrem a existência de correlação entre o rácio da dívida das famílias e a dívida pública nos restantes países da zona Euro.

2.3. Análise de clusters no recurso ao crédito

A Análise de Cluster é uma técnica de excelência em toda a Gestão, com uma ênfase especial no Marketing e na segmentação de mercados com base em características demográficas, geográficas, psicológicas, etc.

Um exemplo de aplicação desta técnica é o estudo de Lourosa (2009) cujo objectivo é segmentar uma base de dados com clientes que recorreram ao crédito ao consumo. Numa primeira fase, o autor fixou como objectivo identificar características diferenciadoras entre os diversos clusters e seguidamente analisar o comportamento dos clientes face ao crédito.

Através da aplicação do Método Não-Hierárquico K- Means identificaram-se 5 grupos (Lourosa, 2009). Os resultados finais permitiram, numa perspectiva de marketing e enquadrados no mercado do crédito ao consumo, concluir da adequabilidade dos 5 clusters.

De forma resumida e de acordo com o autor citado, pode caracterizar-se o perfil dos clientes nos 5 clusters da seguinte forma:

Cluster 1: Forte predominância dos homens, com residência concentrada no Porto e Aveiro, com uma idade média de 42 anos, casados com comunhão de adquiridos, com 2 ou mais dependentes e com casa com hipoteca.

Cluster 2: Maior presença de homens com idades no intervalo dos 26 a 35 anos, com residência entre Lisboa e Porto, casados com comunhão de adquiridos mas sem dependentes. Vivem em casas sem hipoteca e alugadas e trabalham em actividades como a segurança e serviços domésticos.

Cluster 3: Existe um equilibro entre mulheres e homens, com uma idade média elevada (60 anos), vivem em Lisboa em casas sem hipoteca e são funcionários públicos ou reformados.

Cluster 4: Predominam os solteiros com uma idade entre os 26 a 35 anos, vivem em casa dos familiares, são trabalhadores não especializados e com contratos a termo, com um baixo nível de rendimento.

Cluster 5: Mulher com uma idade média de 48 anos, a viver em Lisboa/Porto/Aveiro, casada com comunhão de adquiridos, trabalhando na função pública e vivendo em casa com hipoteca. Tem um nível de rendimento acima dos 2.000€.

Em termos de comportamento verificam-se diferenças importantes com o segmento 3 e 5 a apresentarem níveis de cumprimento muito bons e o segmento 2 e 4 com indicadores muito

fracos. Dos resultados obtidos importa referir as seguintes ilações: observa-se uma correlação negativa entre o rendimento auferido e o incumprimento; confirma-se uma relação positiva entre a idade e o cumprimento; não se consegue inferir sobre uma tendência clara em relação à estabilidade do emprego medido pelo número de anos no emprego; os segmentos onde predominam as denominadas operações sem juros têm um melhor comportamento de pagamento; os segmentos com os maiores valores de mensalidade do crédito à habitação são os que apresentam taxas de incumprimento mais baixas.

Enquanto no estudo anterior o objectivo foi caracterizar o perfil dos indivíduos que recorreram ao crédito, Frade *et al*. (2008), estudaram o perfil dos indivíduos sobreendívidados em Portugal.

O sobreendívidamento é um risco que está cada vez mais associado a famílias das economias de mercado mais desenvolvidas, e que resulta do facto de terem sido ampliadas as ofertas de crédido aos consumidores, que lhe permitiram financiar uma vasta gama de bens e serviços. O sobreendívidamento é cada vez mais uma realidade para um maior número de indivíduos, que se encontram expostos a um contexto macroeconómico adverso, onde se combina o aumento das despesas relacionadas com bens essenciais e os combustíveis, com a subida excessiva das taxas de juro e a crise global.

Neste cenário, com a expectativa de um agravamento das condições de solvabilidade das famílias, prevenir é urgente. É esta realidade que tem ocupado esta equipa nas diversas análises e iniciativas, sempre acompanhados, de forma privilegiada, pelo Observatório do Endívidamento dos Consumidores (OEC) e o Centro de Estudos Sociais da Faculdade de Economia da Universidade de Coimbra.

Tal como neste estudo, em que se pretende traçar o perfil dos vários devedores da empresa de recuperação de crédito XPTO, para adoptar estratégias adequadas à recuperação das dívidas, também no estudo de Frade *et al.*, (2008), houve a consciência que para definir estratégias preventivas se deve partir do conhecimento da realidade social e económica do fenómeno em estudo. Assim, o que se pretendeu neste estudo foi traçar um perfil dos sobreendívidados portugueses, partindo da análise dos casos apoiados pela Associação para a Defesa dos Consumidores – DECO – entre Janeiro de 2005 e Outubro de 2008, que oferece-se um retrato dos contornos pessoais e financeiros que o fenómeno do sobreendívidamento tem vindo a evidenciar. Para o efeito, foi elaborado um questionário – Questionário OEC-DECO – cujo preenchimento *online* foi da responsabilidade dos técnicos da DECO de cada uma das delegações da Associação.

O questionário era constituído por 47 questões e oito secções distintas que organizavam os conteúdos como se apresenta na Tabela 1.

Secção	Conteúdos	Nº questões
Identificação	Delegação; Nº de processo; data de preenchimento	4
Caracterisiticas sociodemográficas	Sexo; idade; estado civil; composição do agregado familiar; residência; escolaridade; profissão; situação na profissão	17
Caracteristicas financeiras	Rendimento; dividas de crédito contraidas; dividas de crédito em atraso; outras dividas em atraso	11
Motivações do consumidor	Recurso ao crédito; entidades de crédito; incumprimento	5
Contactos com a entidade de crédito	Tentatiava de renegociação das dividas, resultado do pedido de renegociação das dividas; cálculo das dividas em atraso	4
Gestão das dificuldades financeiras	Redes formais e informais de solidariedade	3
Consequências do sobreendividamento	Exclusão de sobreendividamentos do mercado de trabalho	2
Percepção do risco de crédito	Intenção de recorrer ao crédito num futuro próximo	1

Tabela 1- Conteúdos do questionário OEC-DECO.

Fonte: (Frade *et al*., 2008)

Na tabela anterior, pode observar-se as variáveis que foram escolhidas para este estudo, e que permitiram construir um retrato pessoal e financeiro dos indivíduos em estudo.

Após a recolha da amostra, foi feita uma análise exaustiva da distribuição geográfica da amostra, do perfil sócio demográfico dos entrevistados, das características do endívidamento e das características do incumprimento, por último foi analisado o risco do sobreendívidamento.

A distribuição geográfica obtida, embora distorcida pela própria divisão territorial determinada pela DECO para estabelecer as suas delegações (todo o interior Norte está afecto às delegações do Porto e Viana do Castelo, e todo o interior Centro, às de Coimbra e Santarém, o que pode, por questões de acessibilidade gerar a sua subrepresentação na amostra), parece apontar para a tendência marcadamente urbana deste fenómeno, sobretudo para a sua localização junto das duas grandes áreas metropolitanas do país, Lisboa e Porto.

Do total dos inquiridos, 60,5% situa-se na faixa etária dos 30 aos 49 anos, 18,7% situa-se na faixa etária dos 50 aos 59 anos e 12,6% entre os 20 e os 29 anos.

Quanto à condição dos inquiridos perante o trabalho, dos 2079 sujeitos que responderam, a maioria (62,6%) exerce uma profissão, 20,2% estão desempregados, 10,2% são reformados e 3,7% estão desempregados, mas realizam pequenos trabalhos informais.

Constatou-se que os sobreendívidados são maioritariamente indivíduos casados e com filhos, com um nível de instrução médio (3º Ciclo do ensino básico e ensino secundário),

empregados por conta de outrem, cujo agregado dispõe entre 500 e 1500 euros mensais e que estão multiendívidados, acumulando crédito à habitação e automóvel com pelo menos um crédito pessoal e um cartão de crédito.

Na sua maioria são influenciados pela publicidade na escolha da entidade a que vão pedir crédito e pela acessibilidade que é ter o crédito oferecido no ponto de venda do bem ou serviço. Embora tenham contraído o crédito também para aceder a bens essenciais, é sobretudo para fazer face às dificuldades financeiras que estão a atravessar que o contrataram. Não surpreende, por isso, que o desemprego seja a principal razão pela qual deixam de cumprir, tal como já se tinha analisado no estudo de Frade *et al*. (2003). Mas também é reconhecido por alguns técnicos da DECO que os apoiam, que há má gestão do orçamento familiar, o que também contribui em parte para a ruptura financeira destes agregados multiendívidados, com valores que variam entre os 2 e os 13 créditos.

Do total dos sobreendívidados em estudo, apenas um terço fez esforços para renegociar directamente com os credores, e dos fizeram esse esforço, a esmagadora maioria não teve qualquer sucesso.

A grave situação financeira em que estão e que tem uma ligação muito forte com a presença de dívidas de crédito torna-os mais avessos ao crédito. A esmagadora maioria, independentemente de nível de rendimento, faixa etária, habilitações literárias ou estado civil, afirmou que não voltaria a contrair crédito face à experiência por que estavam a passar.

Outra conclusão significativa deste estudo é a perda de importância relativa do crédito habitação e crédito automóvel, e o aumento que aconteceu no crédito pessoal. No período em análise houve um aumento do recurso ao crédito pessoal, sendo que, e segundo alguma informação adicional do questionário, mostrou que muitos destes créditos pessoais são do tipo "crédito por telefone" ou "crédito fácil". O risco é por isso evidente e muito elevado, pois trata-se de créditos com taxas de juro muito elevadas, de grande acessibilidade, que parecem ser pedidos para fazer face a dificuldades financeiras correntes e ao incumprimento de outras dívidas de crédito, mas que acabam por resultar num agravamento do multiendívidamento e da espiral de incumprimento.

2.4. Regressão logística no recurso ao crédito

De acordo com Hair *et al.* (1998), a logit é útil para situações nas quais se deseja predizer a presença ou ausência de uma característica, ou resultado, baseado em valores das variáveis independentes. Pode ser utilizada, por exemplo, para se mensurar a probabilidade do risco de crédito em situações de operação de vendas a prazo, empréstimos ou financiamentos. A probabilidade máxima pode ser estimada pela logit, após a transformação da variável dependente em variável de base logarítmica, permitindo que seja calculada a probabilidade de um certo evento acontecer.

Num artigo de Minussi *et al*. (2002), foi construído um modelo de previsão de solvência utilizando a regressão logística. A capacidade de prever o que pode acontecer no futuro e de optar entre as várias alternativas é fundamental nos dias de hoje. A forma de administrar o risco e a vontade de com ele fazer ou não determinadas opções são elementos-chave que impulsionam o sistema económico. No caso de uma instituição financeira, esta questão assume maior relevância, pois ao se considerar correctamente os vários factores que permitem mensurar os níveis de risco sobre uma carteira de activos, pode representar uma grande capacidade de competitividade.

Este artigo, demonstra a importância da aplicação de um modelo econométrico no processo de deferimento de crédito para uma instituição financeira e representa também, uma importante contribuição ao apresentar os resultados do teste de uma nova técnica (regressão logística) para avaliar o risco de crédito.

3. Metodologia

A revisão da literatura, realizada em livros, artigos publicados e estudos realizados, permitiu a identificação de alguns dos métodos utilizados com objectivos similares ao deste trabalho (vide por exemplo Morais (2008), Minussi *et al.* (2002), Ramos (2007), Frade *et al*. (2003), Rossi *et al*. (2001), Johansson *et al*. (2006), Smrčka (2011)), assim como algumas variáveis utilizadas, interpretações e conclusões mais significativas.

De salientar, que na maioria dos artigos publicados em revistas internacionais e em cadernos de relatórios governamentais (e outros) sobre o tema em epígrafe, os modelos usados cingiram-se essencialmente a regressões lineares, cujos pressupostos nem sempre foram devidamente verificados. Tal deve-se ao facto de grande parte dos estudos sobre esta temática serem realizados numa abordagem essencialmente governamental e económica.

Neste capítulo serão alvo de análise as principais técnicas estatísticas exploradas neste trabalho de investigação, dando especial importância à estatística multivariada, nomeadamente: análise clusters, análise discriminante e análise factorial. Com o intuito de explicar o eventual incumprimento do pagamento das dívidas por parte das famílias, é utilizada a regressão logística. Ao longo da análise descritiva dos métodos serão feitos alguns apontamentos relativos à comparação de métodos, limitações das diversas técnicas e identificação dos métodos mais adequados para os dados em estudo.

As técnicas utilizadas são largamente conhecidas e utilizadas na comunidade académica, como tal, a apresentação e discussão neste trabalho será sucinta.

3.1. Análise de Clusters

A análise de clusters é uma técnica de análise multivariada que permite agrupar sujeitos ou variáveis em grupos homogéneos, relativamente a uma ou mais características comuns entre eles (Hair *et al*. (1998); Maroco (2010); Lourosa (2009)).

A aplicação deste tipo de análise acontece nas mais diversas áreas, como é o caso da Demografia, para se efectuar o agrupamento dos concelhos em níveis similares de desenvolvimento sócio-económico (Coutinho *et al.*, 2000), em Psicologia, no agrupamento das técnicas pediátricas utilizadas nos hospitais universitários Brasileiros (Doca, 2009), na área Financeira para identificar grupos de investidores com características homogéneas, em termos de preferências, atitudes e comportamentos (Morais, 2008), estudar o perfil de grupos

de clientes que recorrem ao crédito de forma a identificar o risco de cada grupo (Lourosa, 2009) ou identificar as caracteristicas que definem o perfil dos indivíduos sobreendívidados (Frade *et al*., 2008),

Neste trabalho, pretende-se perante um grupo de indivíduos devedores de algumas carteiras de clientes e algumas características, classifica-los em grupos distintos entre si, de forma a possibilitar a classificação de cada um desses grupos quanto ao risco que apresentam na recuperação de dívidas. A análise de clusters permite fazer a classificação de objectos e pessoas sem pressupostos demasiado rígidos, observando-se apenas as semelhanças ou dissemelhanças entre elas, sem definir previamente critérios de inclusão em qualquer agrupamento.

Nesta análise podem-se definir algumas etapas (Reis, 2001):

1. A selecção de indivíduos ou de uma amostra de indivíduos a serem agrupados;
2. A definição de um conjunto de variáveis a partir das quais será obtida a informação necessária ao agrupamento dos indivíduos;
3. A definição de uma medida de semelhança ou distância entre cada dois indivíduos;
4. A escolha de um critério de agregação ou desagregação dos indivíduos;
5. A validação dos resultados.

3.1.1. Métodos da Análise de Clusters

Nesta análise podem ser utilizados três tipos de abordagens: *Métodos Hierárquicos, Métodos não Hierárquicos (K-Means)* e o Método *Two Step Clusters*.

Nos *Métodos Não Hierárquicos*, fixa-se inicialmente um número *K* de clusters que se pretende constituir e faz-se uma classificação inicial dos *n* indivíduos em *K* clusters. Realizando-se transferências de indivíduos de cluster para cluster, até se encontrar a melhor classificação, ou seja, até obtermos os clusters mais homogéneos possíveis.

No contexto deste método, a ANOVA fornece os testes F para cada variável, identificando a contribuição de cada variável na descriminação entre clusters. As variáveis não significativas podem ser retiradas, uma vez que não contribuem para a diferenciação dos clusters. Este método exige que se saiba à priori, o número de clusters desejado, sendo esta uma limitação deste método.

Os *Métodos Hierárquicos* podem dividir-se em técnicas aglomerativas e divisivas, ambas partindo de uma matriz de semelhanças ou de dissemelhanças entre os *n* indivíduos e consideram-se os *n* indivíduos como *n* clusters.

O processo efectua-se por etapas onde se funde um par de clusters em cada uma das etapas. A fusão a efectuar numa dada etapa é a fusão dos dois subgrupos (clusters) considerados mais "semelhantes". A forma usual de representar graficamente as sucessivas fusões de subgrupos num método classificativo é através de um dendograma, isto é, de uma representação em forma de árvore, como se pode ver na Figura 2.

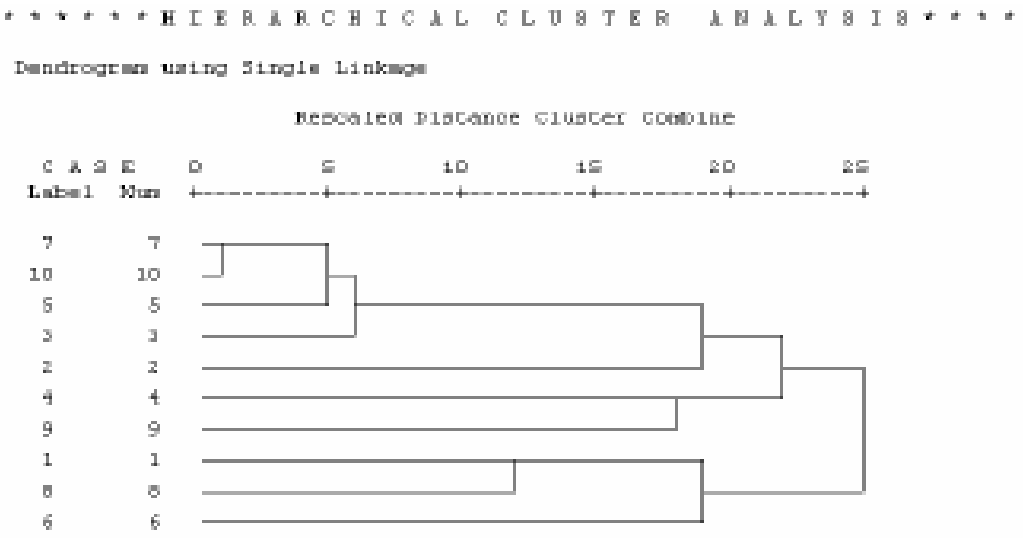

Figura 2- Exemplo de um dendograma

Um corte no dendograma a qualquer nível de aglomeração produz uma classificação em k subgrupos ($1 \leq k \leq n$).

Os métodos hierárquicos são mais adequados para amostras de tamanho mais reduzido.

Os métodos anteriormente referidos são utilizados em situações em que apenas existe um tipo de variável.

Para situações em que a amostra tem um tamanho mais elevado e existem simultaneamente variáveis categóricas (nominais ou ordinais) e contínuas, o Método Two Step Clusters, será o mais adequado (Morais, 2008) e por isso utilizado neste estudo.

Neste método as medidas de distância entre os grupos são estimadas a partir de métodos de máxima - verosimilhança quando as variáveis são categóricas, para as quais se assume uma

Distribuição Multinomial. Apesar destes pressupostos serem dificilmente verificados em dados reais, o algoritmo do modelo encontra uma solução razoável mesmo quando os pressupostos não são verificados.

O procedimento utilizado é feito em duas etapas, uma primeira onde se encontram vários pequenos clusters e a segunda, onde a partir desses clusters se encontra uma resposta óptima para o melhor número de agrupamentos, e o qual tem o objectivo de manter a maior homogeneidade interna em cada grupo e maior heterogeneidade entre os grupos.

A primeira etapa da estimação é feita por aproximação, e para cada registo o algoritmo, baseado na medida de distância, verifica se este deve ser agrupado a algum cluster previamente formado ou se começa um novo. O objectivo é reduzir o tamanho da matriz que contém as distâncias entre os possíveis pares de clusters. A segunda etapa, parte dos clusters formados na etapa anterior e cria o número de agrupamentos desejados, através de método hierárquico aglomerativo. O número de clusters pode ser previamente fixado ou pode ser calculado a partir de dois critérios - Critério de Informação de Akaike – AIC e Critério Bayesiano de Schwarz – BIC.

Critério de informação de Akaike – AIC

$$AIC = -2LogL(\hat{\theta}) + 2(p) \quad \text{Proposto por Akaike (1974)} \tag{3.1.}$$

Onde:

p - número de parâmetros a serem estimados no modelo

$\hat{\theta}$ - estimativa de máxima verosimilhança.

Critério de informação bayesiano – BIC

$$BIC = -2Logf(x_n \left| \hat{\theta}) + pLogn \quad \text{Proposto por Schwarz (1978)} \right. \tag{3.2.}$$

Onde:

p - número de parâmetros a serem estimados

n - número de observações da amostra (Emiliano *et al.*, 2002).

Estes critérios apesar de não constituirem estatísticas de testes, permitem analisar os vários modelos, ajudando na decisão sobre o número de clusters, pois medem o grau de ajuste de um modelo estatístico estimado. Os critérios são calculados para cada potencial solução, e para valores mais reduzidos de ambos os critérios, melhores são os ajustes dos modelos.

No entanto pode acontecer que o valor dos critérios BIC e AIC continue a diminuir com o aumento do número de clusters, o que aumenta a complexidade na caracterização da estrutura. Quando esta situação acontecer, os critérios sugerem que se analise em simultâneo o "ratio of BIC/AIC changes" (rácio que mede o aumento percentual do parâmetro BIC/AIC de uma determinada solução com outra imediatamente anterior) e o "ratio of distance measures" (Morais, 2008).

De acordo com Emiliano *et al*. (2002), ao seleccionamos modelos é preciso ter em mente, que não existem modelos verdadeiros. Há modelos que se aproximam da realidade, o que causa perda de informação. Deste modo, é necessário fazer a selecção do "melhor" modelo, de entre aqueles que foram ajustados aos dados em estudo.

3.1.2. Validação dos resultados

Na análise Two Step Clusters a caracterização dos clusters é feita através da interpretação de gráficos onde se pode observar o ranking de importância de cada variável na formação de cada grupo (variable importance plots). O eixo dos "y" apresenta as variáveis, por ordem decrescente de importância, para a formação de cada cluster.

Esta análise permite a selecção do número de clusters através de critérios estatísticos, incorporando tanto variáveis categóricas quanto continuas no algoritmo de agrupamentos, no entanto esses critérios diferem entre o tipo de variáveis (Nunes, 2009).

A análise das variáveis, categóricas e contínuas, realiza-se de forma distinta, como se pode ver de seguida:

- **Variáveis categóricas** – Teste Qui-quadrado Pearson para a frequência esperada. A estatística é calculada com base na comparação da distribuição de valores numa variável no total da amostra e no cluster encontrado. A hipótese nula é a de distribuição de probabilidades da variável no cluster é igual à distribuição na amostra. Se a hipótese nula é rejeitada a variável em causa é relevante na diferenciação do cluster.

- **Variáveis contínuas** – Teste T para a igualdade entre médias. A estatística compara a média da variável no cluster e na amostra total. Estatísticas negativas no t-test significam que a variável toma valores mais pequenos no cluster. A hipótese nula é que a média da variável

no cluster é igual á média na amostra. Se a hipótese nula é rejeitada a variável em causa é relevante na diferenciação do cluster (Morais, 2008).

Os métodos hierárquicos e o K-means, não são apresentados de forma mais detalhada, pois não serão utilizados neste trabalho.

3.2.Análise discriminante

Após a classificação dos dados em grupos relativamente homogéneos, utilizando a análise de clusters, pretende-se identificar variáveis que melhor diferenciam os grupos criados, através da análise discriminante.

A análise discriminante é uma técnica multivariada, cujo objectivo é descobrir as características que distinguem vários grupos, de tal forma que conhecendo uma característica de um novo indivíduo, seja fácil prever o grupo a que pertence (Maroco, 2010).

Esta técnica foi proposta, na primeira metade do séc. XX por Fisher, enquanto investigador da Estação Agrónomica de Rothamsted, como critério para a classificação de novas espécies vegetais, segundo as suas características biométricas (Maroco, 2010). A análise discriminante pode ainda ter aplicações na área do Marketing, quando uma empresa pretende identificar quais são as características que levam um grupo de consumidores a comprar determinado produto ou a não comprarem, na área da saúde, quando se pretende definir quais são as variáveis que diferenciam um grupo de indivíduos com determinado tipo de doença.

De acordo com Reis (2001) a análise discriminante é levada a cabo através de uma ou mais combinações lineares das variáveis independentes utilizadas (X_j), sendo que, cada combinação linear (Y_i) constitui uma função discriminante.

Um exemplo de função discriminante é o que se apresenta de seguida:

$$y_i = a_{i0} + a_{i1}X_1 + a_{i2}X_2 + \ldots + a_{ip}X_p \qquad (3.3.)$$

Onde:

a_{ip} - coeficientes de ponderação

X_j - variáveis discriminantes não normalizadas

n - número total de indivíduos nos k grupos

p - número de variáveis discriminantes

Esta função é conhecida como função discriminante linear de Fisher. Após a dedução da primeira função discriminante, os pesos das funções seguintes são obtidos sobre a restrição adicional de que os *scores* das funções não estejam correlacionados, isto é, $Cov(Y_i, Y_j)=0$ (Maroco, 2010).

Nesta função os pesos discriminantes são estimados de modo que a variabilidade dos *scores* da função discriminante seja máxima entre os grupos e mínima dentro dos grupos.

Para que esta análise possa ser aplicada, devem ser verificados os seguintes pressupostos (Hair *et al*, 1998):

- Normalidade multivariada das variáveis independentes;
- Homogeneidade das matrizes de variância e co-variância;
- Ausência de multicolinearidade entre variáveis independentes.

A análise discriminante pode também ser entendida como um sistema de pontuações, que a cada indivíduo faz corresponder uma pontuação, resultante de uma média ponderada dos valores que, para ele, assumem as variáveis independentes.

Ao considerar as funções discriminantes como eixos, definindo um espaço *p-dimensional* (*p* variáveis), cada indivíduo poderá ser representado nesse espaço por um ponto, cujas coordenadas são dadas pelos valores das *p variáveis* para esse indivíduo. Se se obtiverem grupos com um grande aglomerado de pontos, com contornos bem definidos e bem separados uns dos outros, isso significa, que os vários grupos apresentam um comportamento bastante diferenciado em relação às variáveis. Mesmo que se encontrem alguns pontos de vários grupos sobrepostos, tem de ser possível definir os seus territórios e posicioná-los a partir de uma medida da sua posição, o *centróide do grupo*.

3.2.1. Funções discriminantes

Como referido anteriormente, na interpretação espacial, cada grupo (representado pelo seu centróide) vai ser tratado como um ponto, e cada função discriminante como um eixo, mas por vezes pode acontecer dois, três ou mais pontos serem colineares[2], não criando qualquer nova dimensão. Assim, para determinar o número de funções discriminantes, recorre-se ao teste de Wilks.

[2] Colineares – dois ou mais pontos dizem-se colineares se incidem todos sobre a mesma recta, i.é., $\exists L : A, B, C, ... \in L$

Segundo o teste de Wilks, o nº máximo de funções discriminantes é igual ao nº de grupos menos um, ou ao nº de variáveis discriminantes, sendo o critério de escolha baseado no menor destes dois valores.

Uma vez estimadas as funções discriminantes e as respectivas médias dos grupos, é necessário testar se as médias dos grupos são significativamente diferentes, para tal, recorre-se ao teste Λ de Wilks para a igualdade de médias dos *k* grupos, onde as hipóteses a testar são as seguintes:

$$\begin{cases} H_0 : \mu_1 = \mu_2 = \ldots = \mu_K \\ H_1 : \mu_i \neq \mu_j ; com\, i \neq j \end{cases} \tag{3.4.}$$

A estatística de teste, para testar a igualdade das médias dos k grupos, é dada por:

$$V = -\left[(n-1) - \frac{1}{2}(p+k) \right] \cdot \ln \Lambda \text{, onde V}\sim \chi^2_{p\times(k-1)} \tag{3.5.}$$

Onde:

k - número de grupos

p - número de variáveis discriminantes;

n - número total de indivíduos nos k grupos

Se o valor de V for superior a um valor crítico retirado da distribuição do χ^2 com $p\times(k-1)$ graus de liberdade, pode-se então concluir que a solução discriminante é estatisticamente significativa, ou seja, rejeita-se H_0 pois a média entre os grupos é estatisticamente diferente.

Segundo Reis (2001), uma medida de avaliação da importância de cada função discriminante é, utilizar a percentagem do valor próprio que lhe está associado, uma vez que a soma dos valores próprios é a medida da variância total explicada pelas variáveis de partida.

Para testar o poder discriminante de cada função discriminante, recorre-se então ao teste Λ de Wilks, onde as hipóteses são:

$$\begin{cases} H_0 : \lambda_1 = \lambda_2 = \ldots = \lambda_K = 0 \\ H_1 : \exists \lambda_r \neq 0; r = 1,2,\ldots,s \end{cases} \tag{3.6.}$$

Onde:

λ_k - valor próprio associado a cada função discriminante.

A estatística de teste, para testar o poder discriminatório da *j-ésima* função discriminante, é dada por:

$$V_j = \left[(n-1) - \frac{1}{2}(p+k) \right] \cdot \ln(1+\lambda_j) \quad \text{, onde V}\sim \chi^2_{(p+k-2j)} \tag{3.7.}$$

São então efectuados testes sucessivos aos valores V_j, com *j=1,2,...,s-1,s*, onde V_j é o poder discriminatório residual depois de retirar o efeito da *j-ésima* função, ou seja, V_2 corresponde ao poder discriminatório, depois de retirados os efeitos discriminatórios da 1ª e 2ª função.

À medida que o termo residual, após remoção das *j* primeiras funções discriminantes, se torna menor que o percentil de ordem 100 (1- α) da distribuição de χ^2 com (p-j)(k-j-1), poder-se-á concluir que apenas as j funções discriminatórias têm poder para a discriminação entre os k grupos (Reis, 2001).

Outra medida importante da função discriminante é a correlação canónica. Trata-se de uma medida de associação entre a função discriminante e um conjunto de (k-1) variáveis que definem a pertença aos grupos. A correlação canónica, é outra forma de testar a validade da função para distinguir os grupos.

Os coeficientes de correlação canónica podem ser calculados a partir dos valores próprios associados a cada função discriminante, utilizando a seguinte equação (Reis, 2010):

$$r_j^* = \sqrt{\frac{\lambda_j}{1+\lambda_j}} \tag{3.8.}$$

3.2.2. Interpretação dos coeficientes estandardizados, não estandardizados e estruturais

Os coeficientes da função discriminante são utilizados para o cálculo de um *score* para cada caso, a partir das variáveis explicativas não estandardizadas.

O score é calculado como o produto dos valores das variáveis explicativas pelo respectivo coeficiente. Após determinadas as funções discriminantes, os seus coeficientes não estandardizados podem ser utilizados para a classificação dos indivíduos. Os coeficientes não estandardizados dão a contribuição absoluta de cada variável para a formação dos *scores* individuais, informação que não se pode comparar.

A interpretação e comparação de coeficientes só pode ser feita quando se tem os coeficientes estandardizados, e neste caso, se se ignorar o sinal obtém-se a contribuição relativa das variáveis em relação à função discriminante.

Para determinar a semelhança entre uma variável e a função discriminante, devem calcular-se os coeficientes estruturais. Quando o valor dos coeficientes estruturais se aproxima de 1 em valor absoluto, pode concluir-se que a função detém grande parte da informação contida da variável, caso o valor se aproxime de 0, conclui-se que não existe uma relação significativa entre as variáveis e a função (Reis, 2001).

3.2.3. Classificação dos indivíduos

A análise discriminante, permite identificar qual o grupo mais provável de um indivíduo pertencer, conhecendo as suas características. É portanto uma técnica que permite classificar e realizar previsões dos indivíduos em determinados grupos pré-definidos. Para ser possível a classificação dos indivíduos, existem dois critérios:

- **Critério de Fisher** – Este critério é utilizando quando existem apenas dois grupos

Para classificar os indivíduos segundo o critério de Fisher é calculado o ponto médio, ou o ponto crítico Y_C definido a partir das médias dos dois grupos e a classificação do indivíduo U com a característica X_U (Y_U).

$$Y_C = \frac{\overline{Y}_1 + \overline{Y}_2}{2} = \frac{\hat{a}'\overline{X}_1 + \hat{a}'\overline{X}_2}{2} \tag{3.9}$$

$$Y_U = \hat{a}X_U \tag{3.10}$$

O individuo classifica-se no grupo 1 se $Y_U > Y_C$ e no grupo 2 se $Y_U \leq Y_C$

Onde Y_C é o ponto médio dos dois grupos e Y_U é a classificação do indivíduo U com a característica X_U.

Para que este critério se possa aplicar é necessário que se verifique os pressupostos de normalidade e heterogeneidade.

- **Critério de Bayes** – Este critério é utilizado quando existem mais de dois grupos

Quando se está perante uma população normal, pode-se calcular a probabilidade de um indivíduo pertencer a um grupo pelas probabilidades condicionadas e o teorema de Bayes.

Seja $P[G_i]$ a probabilidade à priori do indivíduo pertencer ao grupo i, sem haver informação adicional, e $P[Y|G_i]$ a probabilidade de um indivíduo ter determinado *score* dado que pertence ao grupo i, a probabilidade de um indivíduo pertencer a um grupo dado o seu *score* é dado pela seguinte fórmula (Reis, 2001):

$$P[G_i|Y] = \frac{P[Y|G_i].P[G_i]}{\sum P[Y|G_i].P[G_i]} \tag{3.11}$$

Depois de determinar as classificações dos indivíduos nos respectivos grupos, pode avaliar-se a eficácia classificativa da análise discriminante, construindo uma tabela (ver Tabela 2), para se poder comparar as classificações iniciais com as classificações à posteriori, onde n_{ij} é o número de indivíduos classificados inicialmente no grupo i e posteriormente no grupo j.

Grupo original	#	Grupo previsto			
		1	2	...	k
1	n_1	n_{11}	n_{12}		n_{1k}
2	n_2	n_{21}	n_{22}		n_{2k}
...					
k	n_k	n_{k1}	n_{k2}		n_{kk}
Desconhecido	n_0	n_{01}	n_{02}		n_{0k}

Tabela 2- Matriz de classificações

Fonte: (Reis, 2001)

A partir desta matriz é possível calcular:

- Percentagem de casos correctamente classificados:

$$PCC = \frac{\sum_{i=1}^{k} n_{ij}}{n} \times 100 \tag{3.12}$$

- Percentagem de casos incorrectamente classificados:

$$PIC = \frac{\sum_{i=1}^{k}\sum_{\substack{i=1 \\ j\neq i}}^{k} n_{ij}}{n} \times 100 \tag{3.13}$$

Para que esta análise esteja concluída, resta saber a partir de que valores de percentagens de casos correctamente classificados, se aceita os resultados como aceitáveis. Para tal existem os critérios do acaso máximo (Equação 3.14) e do acaso proporcional (Equação 3.15):

$$C_{MAX} = \max_j \frac{n_j}{n} \tag{3.14}$$

$$C_{PRO} = \sum_{j=1}^{k}\left(\frac{n_j}{n}\right)^2 \tag{3.15}$$

A percentagem de casos correctamente classificados é aceitável, se ultrapassar qualquer dos dois valores (Reis, 2001).

3.3. Análise de componentes principais

A Análise de componentes principais é um método estatístico multivariado, que permite transformar um conjunto de variáveis correlacionadas entre si, num conjunto de variáveis não correlacionadas, as quais são chamadas de *Componentes Principais*, e que resultam de combinações lineares do conjunto inicial.

Para tal, consideram-se as *p variáveis* $x_1, x_2, ..., x_p$ de forma a encontrar as combinações lineares dessas variáveis iniciais, para criar indíces $CP_1, CP_2, ..., CP_m$, os quais vão ter variância máxima e que são denominados de *Componentes Principais*. Para *m* componentes e *p* variáveis, as componentes principais são expressas da seguinte forma:

$$\begin{aligned} CP_1 &= a_{11}x_1 + a_{12}x_2 + \cdots + a_{1p}x_p \\ CP_2 &= a_{21}x_1 + a_{22}x_2 \cdots + a_{2p}x_p \\ &\cdots \\ CP_m &= a_{1m}x_1 + a_{2m} + \cdots + a_{pm}x_p \end{aligned} \tag{3.16}$$

Estas componentes são geralmente ordenadas de modo a que, a CP_1 explique a maior variação de todas as variáveis iniciais, a CP_2 explique a segunda maior variação das mesmas,

e assim por diante. Logo, $Var(CP_1) \geq Var(CP_2) \geq \ldots \geq Var(CP_m)$, onde $Var(CP_j)$ $(j = 1, \ldots, m)$ representa a variância de cada CP_j no conjunto dos dados iniciais.

A redução da dimensão dos dados consiste em considerar um número *m* muito menor do que as *p* componentes principais, desprezando deste modo algumas das componentes, as quais deverão apresentar informação pouco relevante do ponto de vista estatístico.

Há que ter em conta que, se as variáveis iniciais forem não correlacionadas, então, o estudo em causa não faz qualquer sentido lógico, e que os melhores resultados são obtidos quando, as ditas variáveis, estão altamente correlacionadas, positiva ou negativamente.

3.3.1. Etapas da ACP/Pressupostos base

Antes de iniciar a análise de componentes principais, há que verificar para cada variável a evidência (ou não) de alguns pressupostos base:

1. A normalidade da distribuição;
2. Existência de *outliers*;
3. Achatamento da distribuição;
4. Simetria da distribuição.

Na aplicação da análise em componentes principais existem 4 etapas a seguir:

1) Estimar a matriz de correlações ou matriz de variâncias-covariâncias, e testar se é possível a realização deste tipo de análise. Existem alguns testes para esse efeito:

- *Teste de Esfericidade de Bartlett:* Testa a hipótese de a matriz de correlações ser uma matriz identidade, ou seja, de as variáveis não estarem correlacionadas entre si, donde se pressupõem a rejeição da Hipótese nula.

$$\begin{cases} H_O : R = I \\ H_1 : R \neq I \end{cases} \tag{3.17}$$

A estatística de teste a usar é a seguinte:

$$\chi^2 = -\left[n - 1 - \frac{1}{6}(2p + 5) \right] \ln|R| \sim \chi^2_{\frac{1}{2}P(P-1)} \tag{3.18}$$

- *Estatística de Kaiser-Meyer-Olkin (KMO):* Esta estatística compara as correlações entra as variáveis.

$$KMO = \frac{\sum_i \sum_j r_{ij}^2}{\sum_i \sum_j r_{ij}^2 + \sum_i \sum_j a_{ij}^2} \tag{3.19}$$

Onde:

r_{ij} - coeficiente de correlação observado entre as variáveis *i* e *j*

a_{ij} -coeficiente de correlação parcial entre as variáveis *i* e *j*.

A partir desta estatística, obteve-se a Tabela 3 que permite avaliar a fiabilidade desta análise:

KMO	*Análise das Componentes Principais*
1-0,90	*Muito Boa*
0,80-0,90	*Boa*
0,70-0,80	*Média*
0,60-0,70	*Razoável*
0,50-0,60	*Má*
<0,50	*Inaceitável*

Tabela 3 - Tabela de comparação da estatistica KMO

Fonte: (Reis, 2001)

2) Cálculo do número de componentes principais

Se algumas variáveis iniciais forem linearmente dependentes, alguns valores próprios serão nulos, logo $m \prec p$ e assim a variação total poderá ser completamente explicada pelas primeiras *m* componentes principais. Por isso, deve retirar-se algumas componentes com valores próximos de zero desta análise e não irá implicar uma perda significativa de informação.

Para determinar o número de componentes a excluir da análise deve proceder-se da seguinte forma:

- Representar gráficamente a percentagem de variância explicada por cada componente. Quando esta percentagem se reduz a curva passa a ser quase paralela ao eixo das abssicas, as componentes correspondentes devem ser excluidas;
- Incluir as componentes suficientes para explicar mais de 70% da variância total;
- Excluir as componentes cujos valores próprios são inferiores à média, menores que 1 se a análise for feita a partir de uma matriz de correlações (Critério de Kaiser).

3) Interpretação das componentes principais

Para tornar as C.P. mais facilmente interpretáveis, deve proceder-se à sua rotação[3]

A interpretação será tanto mais fácil quanto a contribuição de uma variável se aproximar de 100% num factor e apenas 0% nos restantes.

Os principais métodos de rotação das C.P. são:

Varimax - Método onde se pretende maximizar a variação entre os pesos de cada componente principal, ou seja, que para cada componente principal existam apenas alguns pesos significativos e todos os outros estejam próximos de zero, havendo uma clara falta de associação;

Quartimax - Método onde se pretende simplificar as linhas de uma matriz de pesos, ou seja, tornar os pesos de cada variável elevados para um número reduzido de componentes principais e próximo de zero para as restantes;

Equimax - Este método pretende ser uma espécie de combinação entre os métodos anteriores, uma vez que se concentra em simplificar não só as linhas e as colunas, mas sim a combinação de linhas e colunas.

4) Determinar o valor das componentes principais

Determinar o valor que cada factor têm para cada indivíduo, contruindo a matriz de *scores.*

[3] Rotação – A aplicação de um método de rotação tem como objectivo principal a transformação dos coeficientes das componentes principais numa estrutura simplificada.

3.3.2. Análise de componentes principais vs Análise factorial

Outra técnica exploratória multivariada que existe muito idêntica à análise de componentes principais é a análise factorial.

Ambas as técnicas permitem a representação das variáveis originais num número mais reduzido de componentes/factores, no entanto existem algumas diferenças entre estas. No caso da análise de componentes principais, como já se pode analisar, a informação contida nas variáveis originais é resumida a um número reduzido de componentes ortogonais entre si, que explicam o máximo da variância das variáveis originais. Na análise factorial, são identificados factores latentes (variáveis não observadas e subjacentes aos dados) que explicam as intercorrelações observadas nas variáveis originais.

Outra diferença entre estas duas técnicas tem a ver com o facto de a análise factorial ter como principio que a variável pode ser decomposta em duas partes, uma parte comum e uma parte única. A primeira parte da sua variação partilhada com as outras variáveis, enquanto a segunda parte é específica da sua própria variação. Assim, pode-se distinguir estes dois métodos pelo facto de na análise de componentes principais se considerar a variação total, e na análise factorial só é retida a variação comum, partilhada por todas as variáveis.

Também as equações mostram outra diferença, pois na análise factorial existe uma parcela adicional de erro, que explica uma parcela da variância não explicada pelos factores comuns.

As componentes principais são expressas como combinações lineares das variáveis ortogonais, como vimos anteriormente.

$$\begin{aligned} CP_1 &= a_{11}x_1 + a_{12}x_2 + \cdots + a_{1p}x_p \\ CP_2 &= a_{21}x_1 + a_{22}x_2 \cdots + a_{2p}x_p \\ &\cdots \\ CP_m &= a_{1m}x_1 + a_{2m} + \cdots + a_{pm}x_p \end{aligned} \tag{3.20}$$

Na análise factorial, cada variável observada é descrita como uma função dos factores comuns:

$$\begin{aligned} x_1 &= a_{11}FC_1 + a_{21}FC_2 + \ldots + a_{m1}FC_m + e_1 \\ x_2 &= a_{12}FC_1 + a_{22}FC_2 + \ldots + a_{m2}FC_m + e_2 \\ &\ldots \\ x_p &= a_{1p}FC_1 + a_{2p}FC_2 + \ldots + a_{mp}FC_m + e_p \end{aligned} \tag{3.21}$$

(Reis *et al.*, 2001)

A análise factorial é mais apropriada em estudos teóricos que não sejam influenciados por erros de variância, como por exemplo, encontrar dimensões associadas aos itens de uma escala de ansiedade.

A análise de componentes principais é aconselhada para tratamento empírico dos dados quando se pretende reduzir um número elevado de variáveis a um número limitado de componentes, como por exemplo, um estudo onde os consultores de uma empresa tenham que reduzir 60 indicadores de produtividade num conjunto sumário de índices. (Martinez *et al.*, 2008).

3.4.Regressão logística

Muitas são as situações reais em que o pesquisador sente necessidade de construir um modelo matemático, para estudar alguns fenómenos de observação.

A regressão logística tem sido referida por diversos autores, entre os quais Cox (1989), Hosmer e Lemeshow (2000), Hair *et al*. (1998), Rossi et al. (2001), Minussi *et al*. (2002), como uma ferramenta bastante poderosa na modelação estatística, adequada a variáveis categóricas. No estudo de Minussi *et al*. (2002) é possível analisar a importância desta ferramenta na construção de um modelo econométrico que é utilizado no processo de deferimento de crédito numa instituição bancária, utilizando vários factores que permitem medir o risco de crédito de cada cliente.

A regressão logística consiste em relacionar, através de um modelo, a variável resposta categórica, com as variáveis explicativas que influenciam a ocorrência de determinado fenómeno. A variável resposta é geralmente dicotómica, enquanto as variáveis explicativas podem ser categóricas ou contínuas.

Pretende-se, com a regressão logística, avaliar o risco de incumprimento dos devedores, o que em parte será feito na análise discriminante.

Tanto a análise discriminante como a regressão logística enquadram-se num conjunto de métodos estatísticos multivariados de dependência, pois relacionam variáveis independentes com uma variável dependente categórica (Hair *et al.,* 1998).

No entanto, segundo alguns autores a técnica de regressão logística tornou-se um método padrão de análise de regressão para variáveis medidas de forma dicotómica. Além disso, a regressão logística pode ser utilizada de forma bem mais geral, pois não faz suposições quanto à distribuição das variáveis independentes.

3.4.1. Função logística

A função usada na regressão logística para estimar a probabilidade de uma determinada realização j(j=1,…,n) da variável dependente ser 1, P[Y_j=1]= $\hat{\pi}_j$, é a função logística cuja forma genérica, para uma única variável independente X é:

$$\hat{\pi}_j = \frac{e^{\beta_0+\beta_1 X_j}}{1+e^{\beta_0+\beta_1 X_j}} \tag{3.22}$$

Onde:

X_j -variáveis independentes

β_i - Coeficientes de *logit*

Caso exista mais de uma variável independente o modelo é:

$$\hat{\pi}_j = \frac{e^{\beta_0+\beta_1 X_{1j}+\ldots+\beta_p X_{pj}}}{1+e^{\beta_0+\beta_1 X_{1j}+\ldots+\beta_p X_{pj}}} \tag{3.23}$$

Este modelo pode ser ajustado à regressão não linear, a solução tradicional consiste em linearizar esta função com a transformação Logit ($\hat{\pi}$):

$$Logit(\hat{\pi}) = \ln\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) \tag{3.24}$$

O modelo de regressão logística com mais de uma variável independente é (Maroco, 2010):

$$Logit(\hat{\pi}) = \beta_0 + \beta_1 X_{1j} + \ldots + \beta_p X_{pj} \tag{3.25}$$

3.4.2. Validação dos resultados

Para validar o modelo de regressão obtido, deverão ser realizados testes quanto à significância do modelo, ao seu ajustamento aos dados e também à significância dos coeficientes do modelo.

Para testar a significância do modelo ajustado, testam-se as seguintes hipóteses:

$$H_0 : \beta_1 = \beta_2 = ... = \beta_p = 0$$

$$H_1 : \exists_i : \beta_i \neq 0 (i = 1, ..., p) \quad (3.26)$$

A estatística de teste G^2 para testar a significância do modelo de regressão logística é dada pela seguinte equação:

$$G^2 = -2Ln\left[\frac{L_0}{L_c}\right] \sim \chi^2_{(p)} \quad (3.27)$$

Onde:

L_0 - verosimilhança do modelo nulo

L_C - verosimilhança do modelo completo.

Caso a hipótese nula não seja rejeitada, $p - value \geq \alpha$, o modelo não é estatisticamente significativo, ou seja, não se pode predizer a probabilidade do sucesso a partir das variáveis independentes do modelo.

Depois de testada a significância do modelo, as hipóteses a testar relativamente ao ajustamento do modelo aos dados, são as seguintes:

H_o: O modelo ajusta-se aos dados
H_1: O modelo não se ajusta aos dados (3.28)

A estatística de teste a usar é:

$$X^2_{HL} = \sum_{i=1}^{g} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{(g-2)} \quad \text{(para amostras grandes)} \quad (3,29)$$

Onde:

g – grupos definidos pelos decis de probabilidade de "sucesso" onde se classificam as duas classes da variável dependente dicotómica, geralmente g=10.

A hipótese nula não se rejeita se $p\text{-}value \geq \alpha$, ou seja, o modelo ajusta-se aos dados, se os valores observados (O_i) são suficientemente próximos dos valores esperados (E_i).

Após se chegar à conclusão que o modelo ajustado é significativo, há que identificar quais as variáveis que influenciam significativamente o modelo, para tal é usual recorrer ao teste de Wald, cujo objectivo é testar a significância dos coeficientes do modelo.

As hipóteses são as seguintes:

H_0: $\beta_i = 0 \left| \beta_0, \beta_1, \beta_{i-1}; \beta_{i+1}; \beta_p \right.$

H_1: $\beta_i \neq 0 \left| \beta_0, \beta_1, \beta_{i-1}; \beta_{i+1}; \beta_p (i = 1, ..., p) \right.$ (3.30)

A estatística de teste é:

$$Twald_i = \frac{\hat{\beta}_i}{S\hat{E}(\hat{\beta}_i)} \sim N(0,1)$$ (Para amostras grandes) (3.31)

Rejeita-se H_0 se $p\text{-}value \leq \alpha$, para cada um dos β_i. Esta distribuição só é válida para amostras de grande dimensão.

Outra forma de testar a qualidade do modelo é avaliar a força da associação da variável dependente com as variáveis independentes.

Na regressão logística as medidas da força de associação são aproximações ao coeficiente de determinação R^2. O SPSS calcula o R^2 de Cox & Snell e R^2 de Nagelkerke (Pestana *et al.,* 2009).

O R^2 de Cox & Snell (1998) é calculado como:

$$R^2_{CS} = 1 - e^{\frac{2(LL_C - LL_0)}{n}}$$, onde R^2_{CS} é a medida de associação (3.32)

Esta estatística nunca atinge o valor 1, mesmo quando o ajustamento é perfeito.

Em 1991 Nagelkerke, propôs uma correcção ao R^2_{CS} de modo a que este varie entre [0, 1]

$$R^2_N = \frac{R^2_{CS}}{1 - e^{\frac{2LL_0}{n}}}$$, onde R^2_N é a medida de associação (3.33)

3.4.3. Classificação de casos

Para avaliar a qualidade da classificação obtida pelo modelo, compara-se a percentagem global de classificações correctas obtidas no modelo, com a percentagem proporcional de classificações correctas por acaso. Esta percentagem é calculada a partir do número de indivíduos observados em cada uma das k classes (C_i) da variável dependente.

$$\text{Classificação correcta proporcional por acaso (\%)} = 100 \times \sum_{i=1}^{k} \left(\frac{C_i}{N} \right)^2 \quad (3.34)$$

"*Se a percentagem de casos classificados correctamente pelo modelo for superior em pelo menos 25% à percentagem de classificação proporcional por acaso, considera-se que o modelo tem boas propriedades classificativas*" (Maroco, 2010).

Outra medida que permite avaliar a capacidade discriminatória do modelo em relação aos individuos que tem determinada caracterisitca vs os indivíduos que não têm determinada característica é a área sob a curva ROC. Hosmer & Lemeshow (2000) apresentaram um conjunto de valores indicativos da área ROC, que podem ser utilizados para classificar o poder discriminatório do modelo de regressão obtido, e que se encontram na tabela seguinte:

Àrea ROC	Poder discriminante do modelo
0,5	Sem poder discriminatório
]0,5; 0,7[	Discriminação fraca
]0,7; 0,8[	Discriminação aceitável
]0,8; 0,9[	Discriminação boa
>=0,9	Discriminação excepcional

Tabela 4 - Tabela do poder discriminate do modelo

Fonte: (Maroco, 2010)

4. Dados e resultados

Neste capítulo são apresentados os dados e as variáveis que foram definidas para este estudo. É realizada uma análise descritiva da amostra e posteriormente são aplicadas as várias técnicas descritas no Capitulo 3, como a análise de clusters, análise discriminante, análise de componentes principais e regressão logística.

No final do capítulo são alvo de discussão e análise todos os resultados das diversas técnicas.

4.1.Apresentação dos dados

A metodologia desenvolvida no âmbito deste trabalho de investigação, visa analisar o comportamento dos devedores de determinados clientes da empresa XPTO, para tal, foi definido um universo de estudo, recolhidas amostras e seleccionadas variáveis.

O universo de estudo são todos os devedores de clientes da empresa XPTO, sendo que para este estudo foram apenas escolhidos dois clientes e desses dois clientes foram seleccionados os últimos lotes que foram entregues à empresa para fazer a gestão das suas dívidas.

O critério de selecção das amostras não foi aleatório, foram sugestões da empresa, devido ao volume de negócios que cada um dos clientes representa. O cliente A é um dos clientes mais antigos e que sempre foi dos mais representativos em termos de resultados de recuperação e de número de dossiers por lote, enquanto o cliente B é um dos clientes mais recentes, mas que apresentou desde o inicio, resultados muito elevados em termos de recuperação e onde houve uma resposta logo imediata ao tipo de gestão que a empresa XPTO estava a realizar.

O cliente A é representado por uma amostra com 2192 indivíduos e o cliente B por uma amostra com 13188 indivíduos.

A recolha das amostras realizou-se da seguinte forma:

- Numa primeira fase da recolha de dados, recorreu-se a documentos fornecidos pelos clientes, aquando da entrega das carteiras para gestão, onde se encontra informação como: sexo, idade, profissão, rendimento, valor da dívida, antiguidade da dívida, local de residência,

produto, último pagamento (data e valor), nº de dívidas (o documento fornecido pelo cliente, pode ter mais de uma dívida do mesmo devedor).

- Na segunda fase, recorreu-se à base de dados da empresa para realizar cruzamentos com a informação anterior e os dados existentes na base dados, de forma a obter variáveis como: outras dívidas (verificar se o devedor tem outras dívidas de outros clientes existentes na base de dados), chamadas telefónicas realizadas (durante o período de gestão, quantas chamadas foram realizadas com sucesso para este devedor) e número de acordos em ruptura (durante o período de gestão, são feitos acordos de pagamento com o devedor, acordos esses que podem entrar em ruptura, por incumprimento do devedor, e que posteriormente são renegociações com o mesmo, havendo assim a possibilidade de um devedor ter vários acordos ao longo do período de gestão e também várias rupturas).

As variáveis consideradas neste estudo tiveram como critério de selecção os resultados dos estudos referidos no Capitulo 2, mas também o facto de serem informação que os clientes e a empresa disponibilizam.

As variáveis seleccionadas para este estudo são as seguintes:

- **Sexo – variável categórica**
  - i Feminino - 1
  - ii Masculino - 2
- **Profissão – variável categórica –** esta variável teve que ser redefinida, pois a tabela das profissões do cliente, não permitia uma fácil interpretação dos resultados.
  - i Profissões liberais, técnicas/artísticas e literárias e qualificados -1
  - ii Membros das forças armadas e corporações militarizadas - 2
  - iii Pessoal dirigente, técnico e administrative - 3
  - iv Pessoal dos serviços, comércio, transportes e comunicações - 4
  - v Outros (desempregados, domésticos, temporários, estudantes, sazonais, desconhecidos, etc.) - 5
- **Região – variável categórica** – esta variável foi obtida a partir das moradas dos devedores, que se encontravam no documento do cliente.
  - i Norte -1
  - ii Centro -2
  - iii Lisboa e Vale do Tejo -3
  - iv Alentejo - 4

v Algarve - 5

vi Madeira - 6

vii Açores -7

- **Idade - variável contínua –** idade do devedor;
- **Dívida – variável contínua –** valor da dívida do devedor ao cliente (A ou B);
- **Tranches – variável categórica –** valor da dívida, agrupada em intervalos. A introdução desta variável deveu-se ao facto de a empresa XPTO, realizar algumas análise dos valores em dívida por tranches de montante, de forma a analisar qual a tranche onde há maior encaixe e em que altura do período de gestão esse encaixe acontece, com o objectivo de adoptar estratégias adequadas aos diversos valores de dívida e na altura mais correcta.

i [0, 500[- 1

ii [500, 1000[- 2

iii [1000, 1500[- 3

iv [1500, 2000[- 4

v [2000, +∞[- 5

- **Produto – variável categórica** – Tipo de crédito adquirido pelo devedor;

i Revolving - 1

ii Clássico - 2

- **Rendimento – variável contínua –** rendimento do devedor (cliente A) ou do agregado familiar (cliente B);
- **Nº de dívidas – variável contínua** – Número de dívidas do devedor;
- **Outras dívidas – variável contínua** - Número de dívidas do devedor de outros clientes;
- **Último pagamento – variável contínua** - Número de dias desde o último pagamento (cliente B);
- **Incumprimento/antiguidade da dívida – variável contínua** - Número de dias desde o incumprimento. (esta variável difere da anterior, pois o cliente após o incumprimento, pode ainda realizar algum pagamento);
- **Último valor – variável contínua –** valor do último pagamento;
- **Acordos_ruptura – variável contínua** – Número de acordos que entraram em ruptura por incumprimento do devedor;

- **Cham_sucesso – variável contínua** - Número de chamadas de sucesso realizadas para o devedor.

4.2.Análise descritiva

Nesta secção é apresentada a análise descritiva dos dados, com vista a uma análise preliminar dos mesmos.

4.2.1. Cliente A

Para o cliente A foram consideradas variáveis como: região, idade, sexo, dívida, tranches, produto, profissão, antiguidade, nº de dívidas, antiguidade da dívida, outras dívidas e rendimento.

Esta escolha das variáveis, como foi referido anteriormente, está relacionada com a informação que o cliente disponibiliza, sendo consideradas as mais importantes na empresa.

Inicialmente foram calculadas algumas medidas de estatística descritiva, como a média, mediana, moda, valor máximo, valor mínimo e desvio padrão.

Observe-se, na Tabela 5, os resultados obtidos na análise descritiva desta amostra.

		Divida	Antiguidade_divida	Idade	Nº de dividas	Outras_dividas	Rendimento
N	Válidos	2192	2192	2192	2192	2192	2192
	Em falta	0	0	0	0	0	0
Média		5602	398	44	1	0	715
Mediana		4326	424	42	1	0	624
Moda		2743	442	47	1	0	0
Desvio padrão		4886	96	12	1	0	467
Minimo		2	143	21	1	0	0
Máximo		24940	1384	74	6	1	5000

Tabela 5 – Análise descritiva da amostra – Cliente A

A amostra caracteriza-se por indivíduos com idades compreendidas entre o 21 e 75 anos, sendo a média das idades de 44 anos, com um desvio padrão de 12 anos, havendo assim uma grande variabilidade nas idades.

O rendimento médio dos devedores ronda os 715€, sendo o rendimento mais alto registado de 5000€.

Na tabela anterior pode constatar-se que os valores em dívida desta amostra são em média de 5.602€, sendo o valor mais alto em dívida de aproximadamente 25.000€. A antiguidade da dívida é em média de 398 dias sendo que o valor mais elevado de atraso de cumprimento é de 1384 dias. Pode também verificar-se que o número de dívidas varia entre 1 e 6, ou seja, existem clientes com 6 dívidas. O que confirma o que já foi referido anteriormente, que os indivíduos recorrem a sucessivos créditos, crédito habitação, crédito automóvel, crédito pessoal, cartão de crédito e por vezes a créditos para fazer face aos juros dos outros créditos contraídos, dando origem ao multiendívidamento.

No gráfico da Figura 3, pode observar-se que mais de 50% dos devedores é do sexo masculino, tal como se verificou na maioria dos estudos analisados. No estudo de Frade *et al*. (2008) constatou-se que a grande maioria dos devedores era casado, o que pode justificar a percentagem de indivíduos do sexo masculino, pois em muitos casos, quem representa a dívida é o chefe de familia.

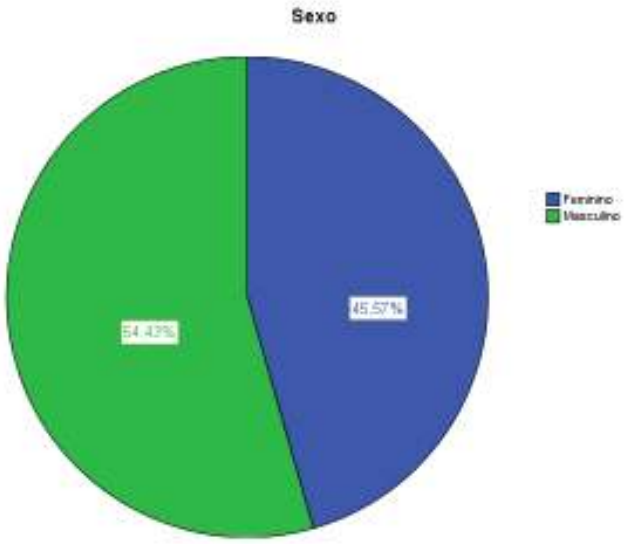

Figura 3 - Caracterização da amostra por género

Verificou-se também que (Figura 4) aproximadamente 50% dos indivíduos residem na região de Lisboa e Vale do Tejo e aproximadamente 23% residem na região norte. Estas

percentagens estão relacionadas com o efeito combinado de uma maior oferta com uma maior procura. Deve também referir-se que se tratam, de acordo com alguns autores Marques *et al*. (2000), Frade *et al. (*2008), das zonas de maior risco de crédito.

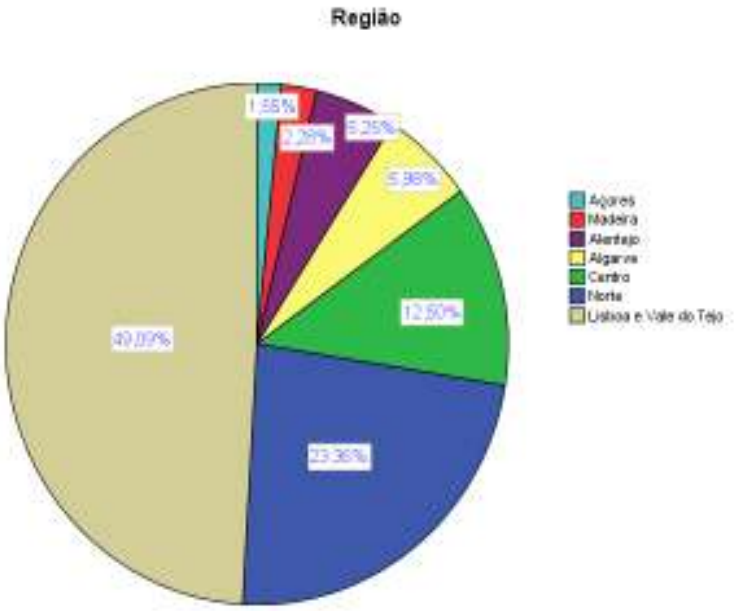

Figura 4 - Distribuição da amostra por zona geográfica

Relativamente à condição dos devedores perante o trabalho (Figura 5), 58,4% pertencem à categoria de pessoal dos serviços, comércio, transportes e comunicações, portanto devedores não qualificados, e aproximadamente 17% são pessoal dirigente, técnico e administrativo. Será ainda importante referir que 13% se encontram na categoria outros, onde estão considerados os desempregados, indivíduos que incorrem num incumprimento devido a uma situação de desemprego.

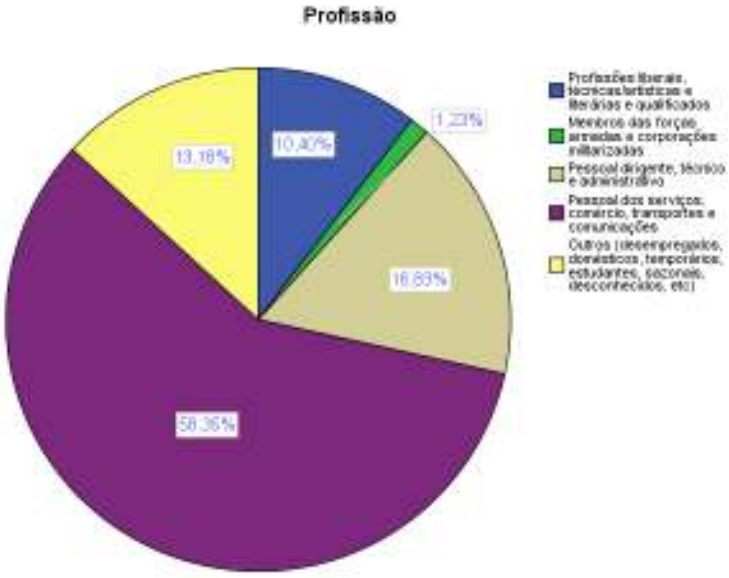

Figura 5 - Caracterização da amostra por situação profissional

Nesta primeira fase do estudo também se optou por analisar as associações entre algumas variáveis, de forma a complementar a análise descritiva dos dados e também para melhor conhecer a amostra e as variáveis utilizadas neste estudo.

Assim, para diversas variáveis realizou-se o seguinte teste de hipóteses.

$$\begin{cases} H_O: \text{ As variáveis são independentes} \\ H_1: \text{ As variáveis não são independentes, ou seja, as variáveis estão associadas.} \end{cases}$$

Para testar a hipótese nula, de que não existe relação entre as duas variáveis, usou-se a estatística designada de ***Qui-Quadrado (χ^2) à Independência***:

$\chi^2 = \sum_i \sum_j \frac{(o_{ij} - \varepsilon_{ij})^2}{\varepsilon_{ij}}$, onde o_{ij} são os valores observados, e ε_{ij} são os valores esperados

O teste do *Qui-Quadrado* não é mais do que uma comparação dos valores observados na tabela com os valores esperados, e permite averiguar se duas variáveis estão relacionadas.

Regra de Decisão:

- Se $\chi^2_{(q-1);(k-1)} < \chi^2$ não se rejeita a hipótese nula de as variáveis serem independentes
- Se $\chi^2_{(q-1);(k-1)} \geq \chi^2$ rejeita a hipótese de independência, ou seja, as variáveis estão associadas

Na Tabela 6 pode-se verificar que existe uma associação entre o sexo e as tranches de montante, pois no sexo masculino os valores das dívidas encontram-se mais concentrados nos intervalos mais altos, ou seja, os homens têm dívidas com montantes mais elevados. Segundo os estudos analisados, os créditos a que os indivíduos do sexo masculino recorrem são o crédito habitação e crédito automóvel, que são créditos com montantes superiores. Enquanto as mulheres recorrem mais ao crédito para obras em casa, compra de mobiliário e electrodomésticos, créditos com montantes mais reduzidos (Ramos, 2007).

Sexo*Tranches Cross tabulation

			Tranches					Total
			[0,500[	[500, 1000[	[1000, 1500[	1500, 2000	>2000	
Sexo	Feminino	Contagem	63	110	96	67	665	1001
		% total	2,9%	5,0%	4,4%	3,1,%	30,3%	45,8%
	Masculino	Contagem	56	100	104	62	869	1191
		% total	2,6%	4,6%	4,7%	2,8%	39,6%	54,3%
Total		Contagem	119	210	200	129	1534	2192
		% total	5,4%	9,6%	9,1%	5,9%	70,0%	100,0%

Tabela 6- Associação entre as variáveis Sexo e Tranches

Ao analisar o resultado do teste de *Qui-Quadrado*, na Tabela 7 para estas duas variáveis, pode-se concluir que a hipótese nula de indepedência é rejeitada, havendo uma associação estatisticamente significativa entre as duas variáveis.

Teste Qui-Quadrado

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	11,821 (a)	4	0,019
Likelihood Ratio	11,786	4	0,019
Linear-by-linear Association	9,609	1	0,002
N of valid cases	2197		

(a) 0 cells have expected count less than 5. The minimum expected count is 54,22

Tabela 7 - Aplicação do *Teste Qui-Quadrado* às variáveis Sexo e Tranches

Relativamente ao sexo e à antiguidade da dívida, pode verificar-se na Tabela 8 que os indivíduos do sexo masculino têm dívidas com antiguidade superior, ou seja, o intervalo de tempo desde o último dia do pagamento é superior para os homens, o que também pode estar

relacionado com o tipo de crédito a que recorrem, que têm montantes mais elevados e que se pode tornar mais difícil de efectuar os pagamentos.

Sexo*Antiguidade Cross tabulation

			Antiguidade				
			<200	[200, 300[	[300, 400[	>400	
Sexo	Feminino	Contagem	58	113	152	678	1001
		% total	2,6%	5,2%	6,9%	30,9%	45,7%
	Masculino	Contagem	63	105	140	883	1191
		% total	2,9%	4,8%	6,4%	40,3%	54,3%
Total		Contagem	121	218	292	1561	2192
		% total	5,5%	9,9%	13,3%	71,2%	100,0%

Tabela 8 – Associação entre as variáveis Sexo e Antiguidade

Analisando o teste do *Qui-Quadrado,* na Tabela 9 pode-se concluir que existe uma associação estatisticamente significativa entre as duas variáveis.

Teste Qui-Quadrado

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	10,361 (a)	3	0,016
Likelihood Ratio	10,335	3	0,016
Linear-by-linear Association	6,525	1	0,011
N of valid cases	2197		

(a) 0 cells have expected count less than 5. The minimum expected count is 55,13

Tabela 9 - Aplicação do *Teste Qui-Quadrado* às variáveis Sexo e Antiguidade

Também foi analisada a associação entre as variáveis "Tranches" e "Profissão".

Tranches*Profissão Cross tabulation

			Profissão					
			Profissões liberais, técnicas/artisticas e literárias e qualificados	Membros das forças armadas e corporações militarizadas	Pessoal dirigente, técnico e administrativo	Pessoal dos serviços, comércio, transportes e comunicações	Outros (desempregados, domésticos, temporários, estudantes, sazonais,	Total
Tranches	[0, 500[	Contagem	12	2	17	78	10	119
		% total	0,5%	0,1%	0,8%	3,6%	0,5%	5,4%
	[500, 1000[	Contagem	17	4	32	133	24	210
		% total	0,8%	0,2%	1,5%	6,1%	1,1%	9,6%
	[1000, 1500[	Contagem	23	1	22	128	31	205
		% total	1,0%	0,0%	1,0%	5,8%	1,4%	9,4%
	[1500, 2000[	Contagem	12	1	19	81	16	129
		% total	0,5%	0,0%	0,9%	3,7%	0,7%	5,9%
	>2000	Contagem	164	19	275	863	208	1529
		% total	7,5%	0,9%	12,5%	39,4%	9,5%	69,8%
Total		Contagem	228	27	365	1283	289	2192
		% total	10,4%	1,2%	16,7%	58,5%	13,2%	100,0%

Tabela 10 – Associação entre as variáveis Tranches e Profissão

De acordo com a Tabela 10 verifica-se que a maior frequência relativa de cada profissão está associada aos maiores valores de "Tranche". No entanto, ao observar o valor do teste do *Qui-Quadrado,* na Tabela 11, pode-se verificar que não existe evidência estatistica para rejeitar a hipotese de independência entre as variáveis, aparentemente não existe dependência entre as variáveis em causa. No entanto é interessante verificar que em praticamente todas as profissões, os indivíduos alvo de estudo apresentam dívidas que pertencem às tranches superioes a 2000 euros, o que pode estar relacionado com o tipo de dívida contraida, uma vez que segundo alguns autores, a maior percentagem das dívidas é do crédito habitação, o que justifica os valores mais elevados.

Teste Qui-Quadrado

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	18,574 (a)	16	0,291
Likelihood Ratio	19,74	16	0,232
Linear-by-linear Association	0,873	1	0,35
N of valid cases	2197		

(a) 4 cells have expected count less than 5. The minimum expected count is 1,46

Tabela 11 - Aplicação do *Teste Qui-Quadrado* às variáveis Tranches e Profissão

Por último, foi ainda analisada a associação entre as variáveis "Profissão" e "Antiguidade".

Na Tabela 12, verifica-se que a maior frequência relativa de cada profissão está associada a antiguidades superiores a 400 dias.

Antiguidade*Profissão Cross tabulation

			Profissão					
			Profissões liberais, técnicas/artisticas e literárias e qualificados	Membros das forças armadas e corporações militarizadas	Pessoal dirigente, técnico e administrativo	Pessoal dos serviços, comércio, transportes e comunicações	Outros (desempregados, domésticos, temporários, estudantes, sazonais, desconhecidos,	Total
Antiguidade	<200	Contagem	12	1	15	77	16	121
		% total	0,5%	0,0%	0,7%	3,5%	0,7%	5,5%
	[200, 300[	Contagem	24	0	25	142	27	218
		% total	1,1%	0,0%	1,1%	6,5%	1,2%	9,9%
	[300, 400[	Contagem	41	5	54	154	43	297
		% total	1,9%	0,2%	2,5%	7,0%	2,0%	13,5%
	>400	Contagem	151	21	271	910	203	1556
		% total	6,9%	1,0%	12,4%	41,5%	9,3%	71,0%
Total		Contagem	228	27	365	1283	289	2192
		% total	10,4%	1,2%	16,7%	58,5%	13,2%	100,0%

Tabela 12 – Associação entre as variáveis Antiguidade e Profissão

No entanto, ao observar o valor do teste do *Qui-Quadrado* na Tabela 13, pode-se verificar que não existe evidência estatistica para rejeitar a hipotese de independência entre as variáveis, assim, aparentemente não existe dependência entre as variáveis em causa. No entanto é interessante verificar que em todas as profissões, as antiguidades são superiores a 400 dias, o que também pode estar relacionado com o tipo de dívida contraida, pois tal como se verificou na análise anterior da associação entre as profissões e a tranche a que pertence o valor da dívida, os valores são superiores a 2000 euros, o que pode dar origem a uma maior antiguidade, pois existe mais dificuldade em pagar montantes superiores.

Teste Qui-Quadrado

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	18,820 (a)	12	0,093
Likelihood Ratio	21,76	12	0,04
Linear-by-linear Association	0,094	1	0,759
N of valid cases	2197		

(a) 4 cells have expected count less than 5. The minimum expected count is 1,49

Tabela 13 - Aplicação do Teste Qui-Quadrado às variáveis Antiguidade e Profissão

4.2.2. Cliente B

Da mesma forma que se procedeu para o cliente A, também foi realizada uma análise descritiva da amostra B.

Na Tabela 14, podem-se analisar algumas medidas estatísticas para as variáveis, sexo, idade, região, incumprimento, último pagamento, último valor, dívida e número de dívidas.

		idade	divida	último_pagamento	incumpriment	último_valor	n_dividas
N	Válidos	13188	13188	13188	13188	13188	13188
	Em falta	0	0	0	0	0	0
Média		41	1670	477	644	203	1
Mediana		40	886	376	543	80	1
Moda		35	60	0	337	100	1
Desvio padrão		11	2936	343	351	480	1
Minimo		16	1,05	0	0	0	1
Máximo		91	105424	2055	2010	19971	9

Tabela 14 - Análise descritiva da amostra – Cliente B

A idade média é 41 anos, a idade máxima é 91 anos e a idade mínima 16 anos, valores que não diferem muito do cliente A, nem dos estudos analisados e referidos no Capitulo 2. O número médio de dias de incumprimento é 477, sendo que o último pagamento foi em média de 100€. O valor em dívida varia entre 1,05€ e 105.423,76€, com um desvio padrão de 2.926,88€, o que mostra a grande variabilidade nos valores da dívida. O valor modal do número de dívidas é 1, havendo indivíduos com 9 dívidas. O número de dias desde o incumprimento é em média de 644 dias, enquanto no caso do cliente A, se verifica um

número médio de dias de incumprimento superior, o que pode estar relacionado com o tipo de instituição com que se estabeleceu o contrato, instituição bancária ou instituição de crédito.

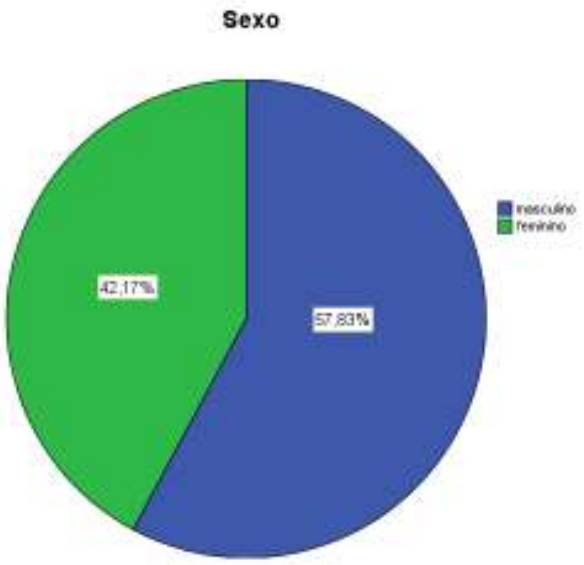

Figura 6- Caracterização da amostra por género

Nesta amostra mais de 50% dos indivíduos são do sexo masculino, como se tinha verificado na amostra relativa ao cliente A (Figura 6).

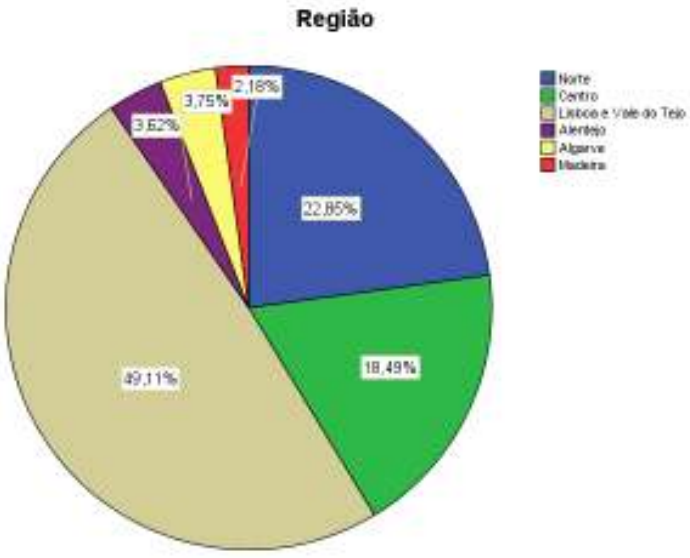

Figura 7- Distribuição da amostra por zona geográfica

Também nesta amostra se verifica uma maior tendência para devedores que residem na região de Lisboa e Vale do Tejo (49,11%), 24% dos indivíduos residem na região norte e apenas 2,18% residem na Madeira.

Foram também analisadas para o cliente B associações entre algumas variáveis.

Na Tabela 15, pode analisar-se a associação entre as variáveis "região" e "n dívidas", onde se conclui que a região com mais dívidas é Lisboa e Vale do Tejo, seguida da região Norte, o que se pode justificar pela oferta dessas regiões e também pela informação a que têm acesso os devedores dessas regiões, que em muito difere das regiões do interior, por exemplo. A região com menor percentagem de dívidas é a Madeira seguida do Algarve.

Região*n_dividas Cross tabulation

			n_dividas							
			1	2	3	4	5	6	9	Total
Região	Norte	Contagem	2677	299	27	4	0	6	0	3013
		% total	20,3%	2,3%	0,2%	0,0%	0,0%	0,0%	0,0%	22,8%
	Centro	Contagem	2162	212	42	8	15	0	0	2439
		% total	16,4%	1,6%	0,3%	0,1%	0,1%	0,0%	0,0%	18,5%
	Lisboa e Vale do Tejo	Contagem	5601	666	162	28	10	0	9	6476
		% total	42,5%	5,1%	1,2%	0,2%	0,1%	0,0%	0,1%	49,1%
	Alentejo	Contagem	434	32	12	0	0	0	0	478
		% total	3,3%	0,2%	0,1%	0,0%	0,0%	0,0%	0,0%	3,6%
	Algarve	Contagem	440	51	3	0	0	0	0	494
		% total	3,3%	0,4%	0,0%	0,0%	0,0%	0,0%	0,0%	3,7%
	Madeira	Contagem	259	26	3	0	0	0	0	288
		% total	2,0%	0,2%	0,0%	0,0%	0,0%	0,0%	0,0%	2,2%
Total		Contagem	11573	1286	249	40	25	6	9	13188
		% total	87,8%	9,8%	1,9%	0,3%	0,2%	0,0%	0,1%	100,0%

Tabela 15- Associação entre as variáveis Região e N_dívidas

Na Tabela 16, pode observar-se o valor do teste do *Qui-Quadrado*, que permite concluir que a associação entre as variáveis é significativa, para um nível de significância de 0,05

Teste Qui-Quadrado

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	119,302 (a)	30	0,093
Likelihood Ratio	128,758	30	0,04
Linear-by-linear Association	1,32	1	0,251
N of valid cases	13188		

(a) 19 cells have expected count less than 5. The minimum expected count is 0,13

Tabela 16 – Aplicação do Teste Qui-Quadrado às variáveis Região e N_dívidas.

Na Tabela 17 analisa-se a associação entre as variáveis "sexo" e "nº de dívidas".

sexo*n_dividas Cross tabulation

			n_dividas							Total
			1	2	3	4	5	6	9	
Sexo	Masculino	Contagem	6690	756	126	28	20	6	0	7626
		% total	50,7%	5,7%	1,0%	0,2%	0,2%	0,0%	0,0%	57,8%
	Feminino	Contagem	4883	530	123	12	5	0	9	5562
		% total	37,0%	4,0%	0,9%	0,1%	0,0%	0,0%	0,1%	42,2%
Total		Contagem	11573	1286	249	40	25	6	9	13188
		% total	87,8%	9,8%	1,9%	0,3%	0,2%	0,0%	0,1%	100,0%

Tabela 17 – Associação entre as variáveis Sexo e N_dívidas

A maior percentagem de dívidas são de devedores do sexo masculino, no entanto é no sexo feminino que se registam por devedor um número de dívidas superior a 6, esta situação pode ser justificada pelo que já foi referido anteriormente, o tipo de crédito a que as mulheres recorrem têm como objectivo a realização de obras, compra de imobiliário e electrodomésticos, que têm valores mais baixos, mas que podem dar origem a um maior número de dívidas.

Ao observar o valor do teste do *Qui-Quadrado* na Tabela 18, pode concluir-se que a associação entre as variáveis é estatisticamente significativa, para um nível de significância de 0,05.

Teste Qui-Quadrado

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	30,003 (a)	6	0,000
Likelihood Ratio	35,916	6	0,000
Linear-by-linear Association	0,298	1	0,585
N of valid cases	13188		

(a) 3 cells have expected count less than 5. The minimum expected count is 2,53

Tabela 18 – Aplicação do Teste Qui-Quadrado às variáveis Sexo e N_dívidas.

4.3. Análise de clusters

Nesta secção são apresentados os resultados da aplicação da análise de clusters, onde serão identificados os clusters obtidos para cada cliente. No final será feita a discussão dos resultados obtidos.

4.3.1. Cliente A

Como foi referido no Capitulo 3, antes de serem caracterizados os clusters, há que decidir qual o método mais adequado, e posteriormente qual a melhor solução para a caracterização do perfil dos devedores.

O método mais adequado para a amostra em estudo é o método *Two Step*, uma vez que existem variáveis categóricas e contínuas.

No primeiro passo determina-se o número de clusters óptimo, ou seja, aqueles que melhor caracterizam o conjunto de indivíduos, para tal serão usados os critérios BIC e AIC.

Da análise dos critérios BIC e AIC, verifica-se o seguinte:

- O valor mínimo de BIC surge em k=15 e o mesmo acontece para o valor mínimo de AIC, no entanto onde se verifica um maior decréscimo dos "*ratios of BIC changes"* e "*ratios of distance measures*" é em k=3, o que sugere que o número óptimo de clusters é k=3.

Auto-Clustering

Number of Clusters	Schwarz's Bayesian Criterion (BIC)	BIC Change[a]	Ratio of BIC Changes[b]	Ratio of Distance Measures[c]
1	15269,953			
2	12195,256	-3074,698	1,000	1,784
3	10529,249	-1666,007	,542	2,153
4	9825,414	-703,835	,229	1,205
5	9263,560	-561,854	,183	1,112
6	8771,659	-491,901	,160	1,238
7	8399,566	-372,093	,121	1,023
8	8038,926	-360,640	,117	1,144
9	7740,027	-298,899	,097	1,051
10	7461,860	-278,167	,090	1,448
11	7310,201	-151,659	,049	1,139
12	7193,034	-117,167	,038	1,357
13	7141,110	-51,924	,017	1,042
14	7096,469	-44,641	,015	1,085
15	7065,510	-30,958	,010	1,424

a. The changes are from the previous number of clusters in the table.

b. The ratios of changes are relative to the change for the two cluster solution.

c. The ratios of distance measures are based on the current number of clusters against the previous number of clusters.

Tabela 19 – Critério de informação Bayesiano - BIC

Auto-Clustering

Number of Clusters	Akaike's Information Criterion (AIC)	AIC Change[a]	Ratio of AIC Changes[b]	Ratio of Distance Measures[c]
1	15173,180			
2	12001,708	-3171,472	1,000	1,784
3	10238,928	-1762,780	,556	2,153
4	9438,319	-800,609	,252	1,205
5	8779,691	-658,628	,208	1,112
6	8191,017	-588,675	,186	1,238
7	7722,150	-468,866	,148	1,023
8	7264,737	-457,413	,144	1,144
9	6869,064	-395,673	,125	1,051
10	6494,123	-374,941	,118	1,448
11	6245,690	-248,433	,078	1,139
12	6031,749	-213,941	,067	1,357
13	5883,052	-148,697	,047	1,042
14	5741,637	-141,415	,045	1,085
15	5613,905	-127,732	,040	1,424

a. The changes are from the previous number of clusters in the table.

b. The ratios of changes are relative to the change for the two cluster solution.

c. The ratios of distance measures are based on the current number of clusters against the previous number of clusters.

Tabela 20 - Critério de informação de Akaike – AIC

4.3.2. Análise das variáveis após construção dos clusters

Após a formação dos clusters realiza-se a análise das variáveis, isto é, verifica-se para as diversas variáveis se o comportamento da amostra não difere em muito, através da contribuição das variáveis, para a segmentação da amostra.

Após a análise *two step cluster* com algumas variáveis, foi possível notar que existem variáveis com uma maior contribuição na diferenciação dos clusters. De seguida analisam-se essas variáveis.

Produto

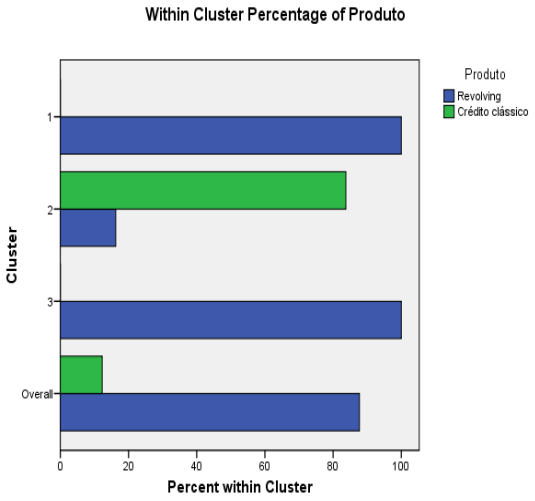

Figura 8 – Percentagem da variável "Produto" em cada cluster

Tranches

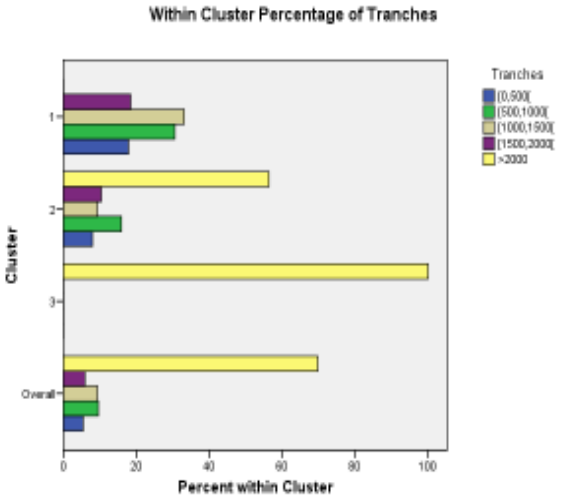

Figura 9 - Percentagem da variável "Tranches" em cada cluster

A variável "produto" é significativamente distinta apenas para o cluster 2, pois cerca de 80% dos devedores recorrem a outros produtos além do revolving, como é o caso do produto clássico, nos outros dois clusters os devedores apenas recorrem a produtos de revolving.

A variável "tranche" é significativa em todos os clusters, pois verificam-se comportamentos distintos nos três clusters. No cluster 3 é possível identificar os devedores com valores de dívida mais elevados, superiores a 2000€ e no cluster 1 os devedores com dívidas inferiores a 2000€.

Nº dívidas

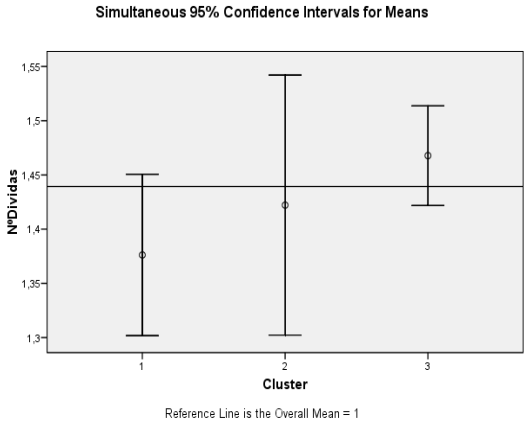

Figura 10 – Intervalo de confiança da variável "Nº Dívidas" por cluster

Rendimento

Figura 11 - Intervalo de confiança da variável "Rendimento" por cluster

Ambas as variáveis são notoriamente distintivas, sendo de realçar que o cluster 1 distingue-se pelos rendimentos médios inferiores à média da amostra, ao contrário do que acontece nos clusters 2 e 3, sendo que no cluster 3, existem uma maior variabilidade. A média dos rendimentos do cluster 2 é aproximadamente 800€, enquanto no cluster 1 é 600€. Quanto ao número de dívidas, o cluster 3 distingue-se por possuírem um número médio de dívidas superiores à média, no entanto nos clusters 1 e 2 existe uma maior variabilidade.

Dívidas

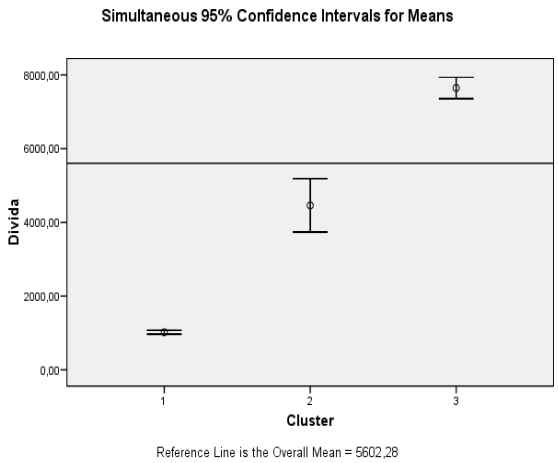

Figura 12 - Intervalo de confiança da variável "Dívida" por cluster

Antiguidade da dívida

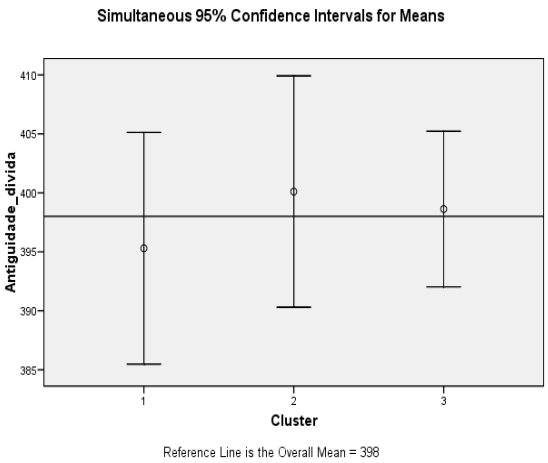

Figura 13 - Intervalo de confiança da variável "Antiguidade_dívida" por cluster

As variáveis "Dívida" e "Antiguidade da dívida" também são variáveis que distinguem os clusters. No cluster 3 verifica-se a existência de valores médios de dívida mais elevados e superiores à média da amostra, de aproximadamente 8000€, enquanto o cluster 1 tem os valores médios mais baixos, de aproximadamente 1000€. Em relação à variável "Antiguidade da dívida", o cluster 1 tem antiguidades médias inferiores à média da amostra, mas com grande variabilidade, enquanto os clusters 2 e 3 apresentam antiguidades médias superiores. No cluster 2 registam-se antiguidades de aproximadamente 400 dias.

Deve salientar-se o facto de a variável "Antiguidade da dívida" não ser uma variável tão distintiva quanto as restantes.

Idade

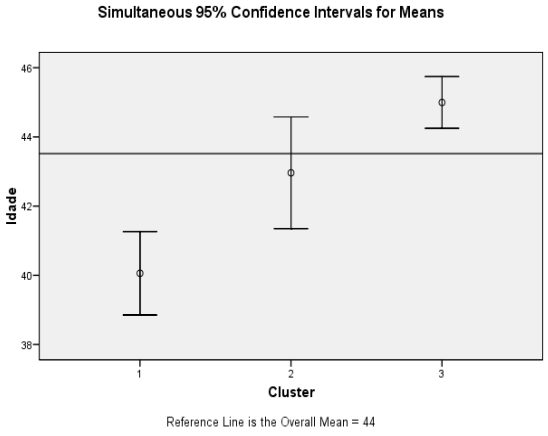

Figura 14 - Intervalo de confiança da variável "Idade" por cluster

Outras dívidas

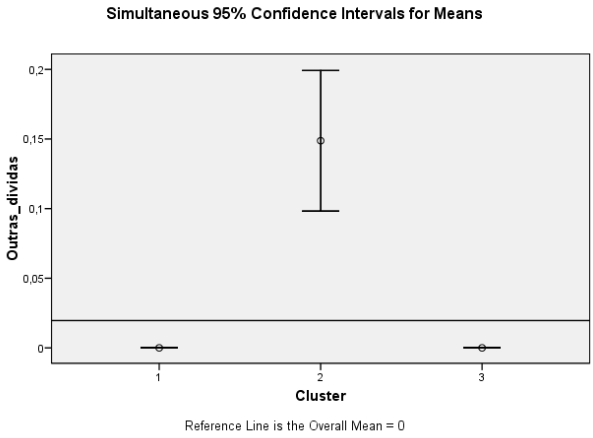

Figura 15 - Intervalo de confiança da variável "Outras_dívidas" por cluster

As variáveis "Idade" e "Outras dívidas" são variáveis bastante distintivas. O cluster 3 é o grupo com idades médias mais altas e superiores à média, sendo a média das idades próximas dos 45 anos. No cluster 1 as idades são mais baixas, e abaixo da média. No cluster 2, verifica-se uma maior variabilidade, e com média de idade abaixo da média da amostra.

O cluster 2, distingue-se dos restantes por existirem devedores com dívidas a outros clientes, ao contrário do que acontece no cluster 1 e 3, que não existem mais dívidas.

4.3.2.1. Caracterização dos segmentos de devedores

Após análise das variáveis mais significativas na construção dos clusters, construiu-se a Tabela 21, que irá ajudar na caracterização dos três segmentos de devedores identificados pela análise de cluster, atribuindo nomes a cada um dos segmentos que irão reflectir os aspectos mais importantes de cada cluster.

Variáveis	Cluster 1	Cluster 2	Cluster 3
Produto	100% dos individuos previlegiam produtos de revolving	Mais de 80% dos individuos previlegiam créditos clássicos	100% dos individuos previlegiam produtos de revolving
Tranches	Grupo de individuos com valores médios de dividas pertencentes às tranches de valores inferiores a 2000€		Grupo de individuos com valores médios de dividas pertencentes às tranches de valores superiores a 2000€
Nº de dividas	Devedores com número médio de dividas inferiores à média da amostra	Devedores com número médio de dividas inferiores à média da amostra. Clusters onde se encontram o menor e maior número de dividas por devedores. Maior variabilidade.	Devedores com número médio de dividas superiores à média da amostra
Rendimento	Individuos com rendimentos médios inferiores à média da amostra, de aproximadamente 600€.	Individuos com rendimentos médios superiores à média da amostra, de aproximadamente 800€. Maior variabilidade.	Individuos com rendimentos médios superiores à média da amostra, de aproximadamente 750€.
Dividas	Grupo com valores médios de dividas mais baixos, de aproximadamente 1000€		Grupo com valores médios de dividas mais altas, de aproximadamente 8000€
Antiguidade da divida	Individuos com dividas de antiguidade média inferior à média da amostra, que rondam aproximadamente os 395€. Maior variabilidade.	Individuos com dividas de antiguidade média superior à média da amostra, que rondam aproximadamente os 400€. Grande variabilidade.	Individuos com dividas de antiguidade média superior à média da amostra, que rondam aproximadamente os 398€.
Idade	Grupo de individuos com idades médias inferiores à média da amostra, aproximadamente 40 anos		Grupo de individuos com idades médias superiores à média da amostra, aproximadamente 45 anos
Outras dividas		Grupo de individuos que têm dividas em outros clientes	

Tabela 21 – Resumo da caracterização dos clusters

Nota: Nas células em branco considerou-se que a informação respectiva, não era significativa para a caracterização do respectivo cluster.

Cluster 1

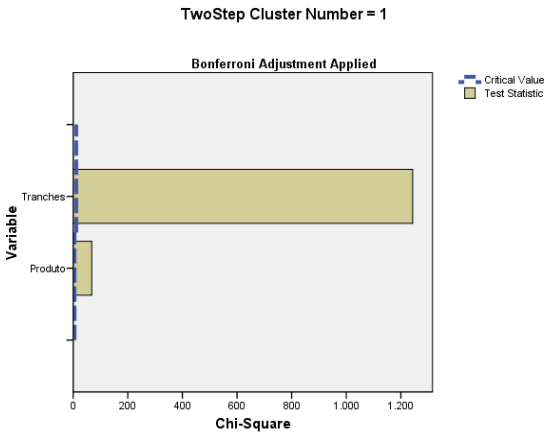

Figura 16 – Niveis de significância das variáveis categóricas, na formação do cluster 1

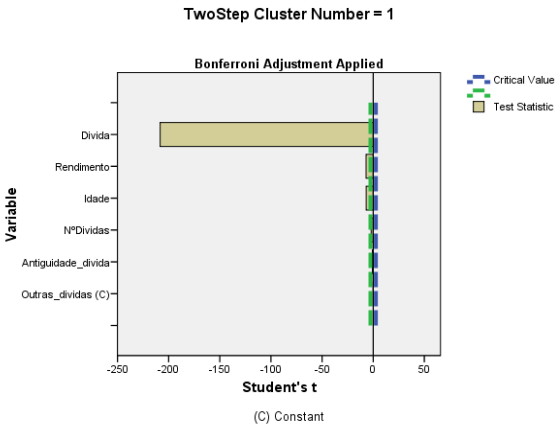

Figura 17 - Niveis de significância das variáveis numéricas, na formação do cluster 1

As variáveis com maior importância no cluster 1 são as variáveis "tranches", "produto", "rendimento", "dívida" e "idade"

1º cluster: Devedores jovens, cautelosos e mais avessos ao risco

Grupo de devedores jovens, conhecedores dos produtos de crédito, pois recorrem a crédito de revolving para montantes reduzidos. São indivíduos com rendimentos reduzidos sendo por isso cautelosos, contraindo dívidas com montantes mais baixos e também com um número de dívidas baixo.

Cluster 2

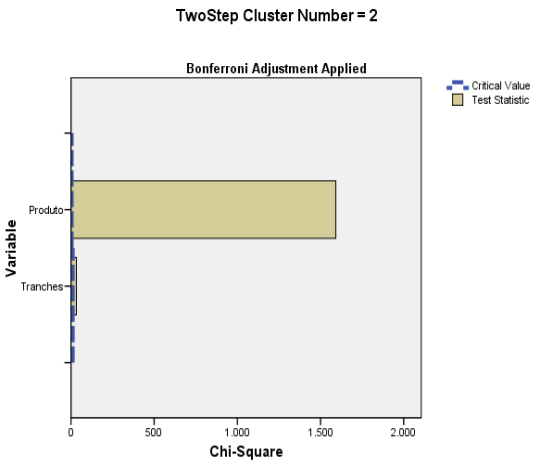

Figura 18 - Niveis de significância das variáveis categóricas, na formação do cluster 2

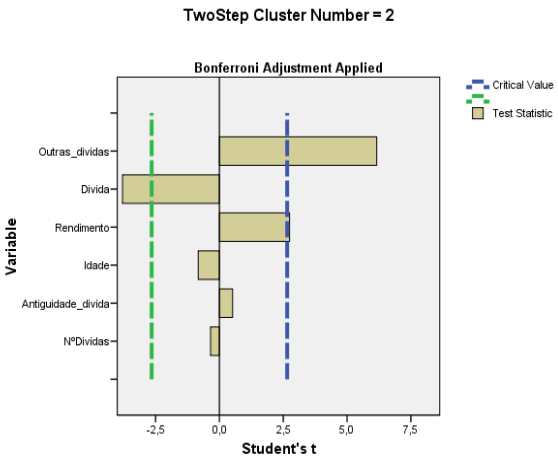

Figura 19 - Niveis de significância das variáveis numéricas, na formação do cluster 2

As variáveis com maior importância são as variáveis "produto", "outras dívidas", "dívida" e "rendimento".

2º cluster: Devedores atrevidos e conhecedores das ofertas do crédito

Grupo de devedores atrevidos, pois contraem várias dívidas, com montantes mais elevados, recorrendo a diferentes tipos de créditos, sendo maus pagadores das suas dívidas, apesar de serem aqueles com rendimentos médios mais elevados.

Cluster 3

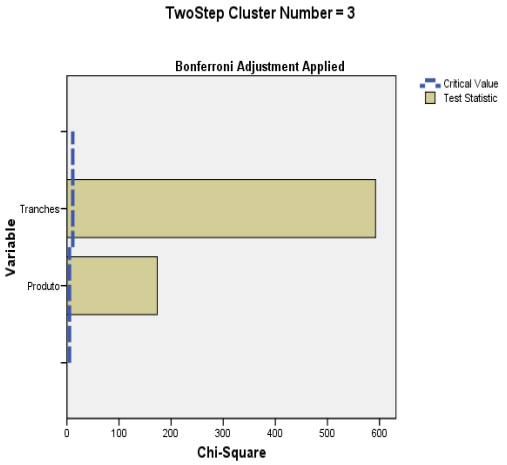

Figura 20 - Niveis de significância das variáveis categóricas, na formação do cluster 3

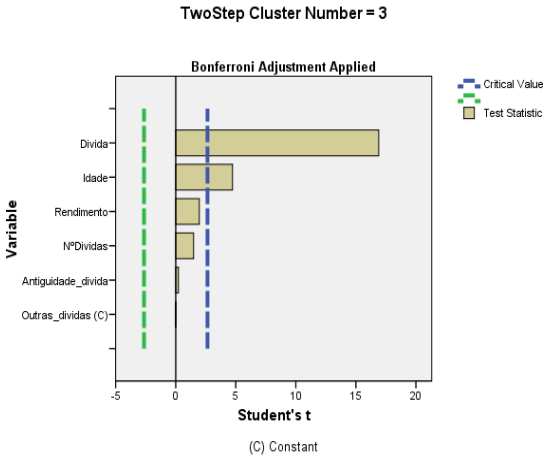

Figura 21 - Niveis de significância das variáveis numéricas, na formação do cluster 3

No cluster 3 as variáveis com maior importância relativas são as variáveis "tranche", "produto", "dívida", "idade".

3º cluster: Devedores mais velhos e mais propensos ao risco

Grupo de devedores mais velhos, menos racionais, pois contraem dívidas com montantes mais elevados recorrendo a produtos de revolving, possuindo várias dívidas com as mesmas características, sendo possivelmente devedores pouco conhecedores das ofertas de produtos de crédito.

Segundo esta análise realizada para o cliente A da empresa XPTO, pode-se concluir que os indivíduos pertencentes aos clusters 2 e 3 são aqueles que apresentam maior risco para empresa quanto à recuperação dos valores em dívida, devido ao seu perfil, enquanto os

indivíduos do cluster 1 apresentam menor risco quanto à recuperação dos valores em dívida, pois são indivíduos mais racionais e cautelosos, também os seus valores são mais reduzidos.

4.3.3. Cliente B

No primeiro passo através da análise dos critérios BIC e AIC, verifica-se o seguinte:

- O valor mínimo de BIC acontece em k=15 no entanto onde se verifica um maior decréscimo dos "*ratios of BIC changes"* e "*ratios of distance measures*" é em k=2, o que sugere que o número óptimo de clusters é k=2.

Auto-Clustering

Number of Clusters	Schwarz's Bayesian Criterion (BIC)	BIC Change [a]	Ratio of BIC Changes [b]	Ratio of Distance Measures [c]
1	55985,662			
2	39850,991	-16134,671	1,000	2,202
3	32585,364	-7265,626	,450	1,797
4	28591,799	-3993,566	,248	1,571
5	26091,497	-2500,302	,155	1,201
6	24029,271	-2062,225	,128	1,615
7	22795,580	-1233,692	,076	1,121
8	21707,680	-1087,899	,067	1,151
9	20777,067	-930,613	,058	1,243
10	20050,567	-726,501	,045	1,031
11	19349,497	-701,070	,043	1,002
12	18650,385	-699,111	,043	1,360
13	18166,421	-483,965	,030	1,119
14	17746,176	-420,245	,026	1,109
15	17378,420	-367,755	,023	1,095

a. The changes are from the previous number of clusters in the table.

b. The ratios of changes are relative to the change for the two cluster solution.

c. The ratios of distance measures are based on the current number of clusters against the previous number of clusters.

Tabela 22- Critério de informação bayesiano – BIC

Auto-Clustering

Number of Clusters	Akaike's Information Criterion (AIC)	AIC Change [a]	Ratio of AIC Changes [b]	Ratio of Distance Measures [c]
1	55895,594			
2	39670,856	-16224,738	1,000	2,202
3	32315,162	-7355,694	,453	1,797
4	28231,529	-4083,633	,252	1,571
5	25641,159	-2590,370	,160	1,201
6	23488,867	-2152,293	,133	1,615
7	22165,108	-1323,759	,082	1,121
8	20987,141	-1177,967	,073	1,151
9	19966,460	-1020,680	,063	1,243
10	19149,893	-816,568	,050	1,031
11	18358,755	-791,138	,049	1,002
12	17569,576	-789,179	,049	1,360
13	16995,544	-574,032	,035	1,119
14	16485,232	-510,313	,031	1,109
15	16027,409	-457,823	,028	1,095

a. The changes are from the previous number of clusters in the table.

b. The ratios of changes are relative to the change for the two cluster solution.

c. The ratios of distance measures are based on the current number of clusters against the previous number of clusters.

Tabela 23 - Critério de informação de Akaike - AIC

4.3.3.1. Análise das variáveis após construção dos clusters

Idade

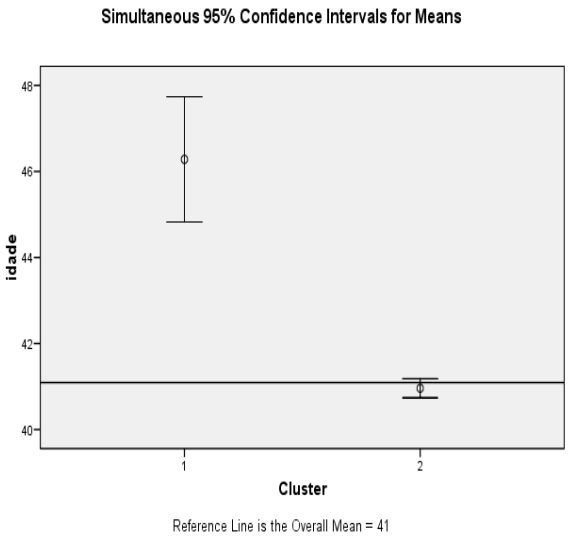

Figura 22 – Intervalos de confiança da variável "Idade" em cada cluster

Dívida

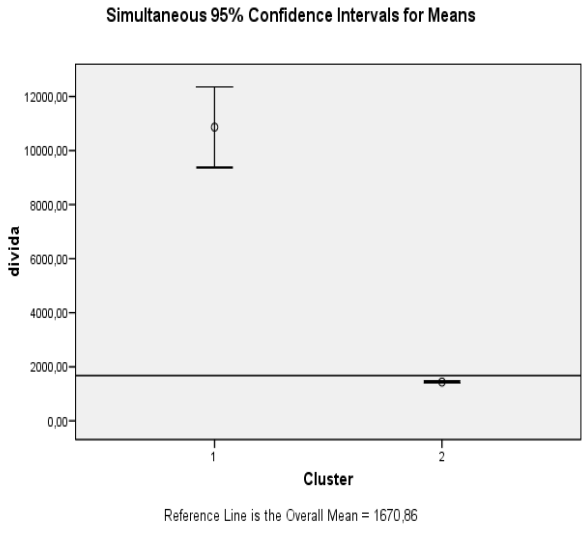

Figura 23 - Intervalos de confiança da variável "dívida" em cada cluster

As variáveis "idade" e "dívida" distinguem os dois clusters, sendo que no primeiro cluster os indivíduos têm idades e valores de dívida mais diversificados. O primeiro cluster

caracteriza-se por indivíduos com idades médias superiores à média da amostra, bem como montantes de dívidas superiores à média, a rondar os 11.000€.

Último_pagamento

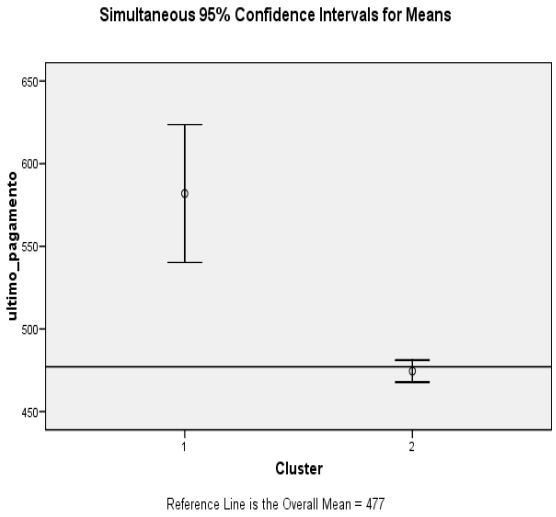

Figura 24 - Intervalos de confiança da variável "ultimo_pagamento" em cada cluster

Incumprimento

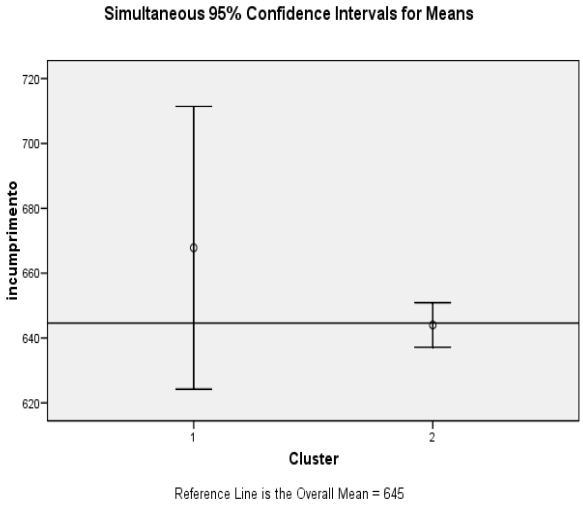

Figura 25 - Intervalos de confiança da variável "Incumprimento" em cada cluster

As variáveis "ultimo pagamento" e " incumprimento" são bastante distintivas. O cluster 1 caracteriza-se por periodo de dias desde o último pagamento bastante elevado, superior à média da amostra e com bastante variabilidade, também no cluster 1 o periodo médio de dias desde incumprimento é bastante elevado, sendo a média de aproximadamente 670 dias. O cluster 2 apresenta periodos de incumprimento mais reduzidos e com menor variabilidade.

Último_valor

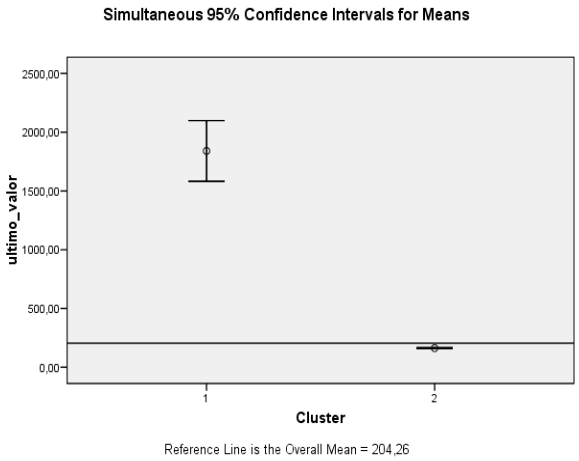

Figura 26 - Intervalos de confiança da variável "ultimo_valor" em cada cluster

n_dívidas

Figura 27 - Intervalos de confiança da variável "n_dívidas" em cada cluster

As variáveis "último valor" e "número de dívidas" são variáveis que distinguem os clusters de forma significativa. O primeiro cluster distingue-se pelos valores dos últimos pagamentos serem superiores, de aproximadamente 1.800€, bem como o número de dívidas, com uma média de aproximadamente 8 dívidas.

4.3.3.2. Caracterização dos segmentos devedores

Da mesma forma que se procedeu para o cliente A e com base na Tabela 24, são caracterizados os dois segmentos de devedores identificados pela análise de cluster.

Variáveis	Cluster 1	Cluster 2
Idade	Grupo de devedores com idades médias superiores à média da amostra, de aproximadamente 46 anos. Maior variabilidade.	Grupo de devedores com idades médias inferiores à média da amostra, de aproximadamente 41 anos.
Dividas	Valores médios das dividas mais elevados, de aproximadamente 11000€ e maior variabilidade.	Valores médios das dividas mais baixos e inferiores à média da amostra. Valores médios de aproximadamente 2000€.
Último_pagamento	Grupo de devedores que realizaram o último pagamento há mais tempo, tendo sido realizado em média há mais 550 dias.	Grupo de devedores que realizaram o último pagamento em média há menos de 500 dias.
Incumprimento	Periodo médio de incumprimento superior à média e com maior variabilidade.	Periodo médio de incumprimento superior à média, menor variabilidade e de aproximadamente 640 dias.
Último_valor	Grupo de devedores, que realizaram um último pagamento superior à média da amostra, sendo de aproximadamente 1750€.	Grupo de devedores, que realizaram um último pagamento inferior à média da amostra, de aproximadamente 200€.
N_dividas	Devedores com mais de uma divida, tendo em média 8 dividas	Devedores com apenas uma divida

Tabela 24 - Resumo da caracterização dos clusters

Cluster 1

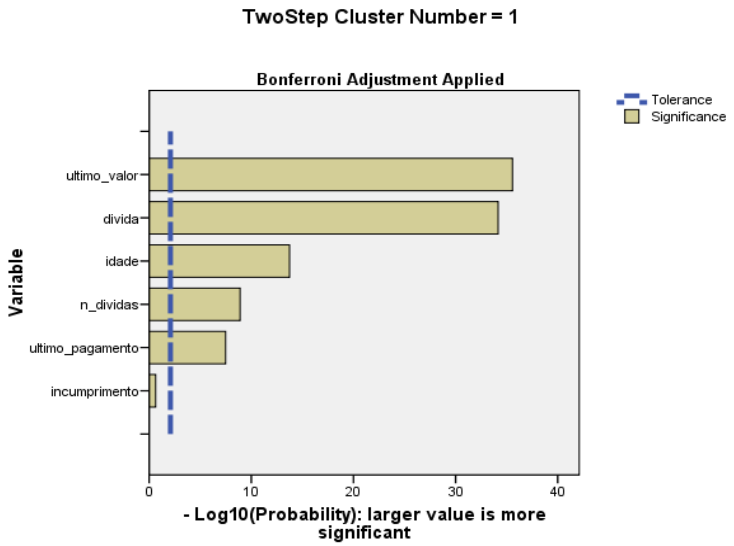

Figura 28 – Niveis de significância das variáveis numéricas, na formação do cluster 1.

As variáveis que mais contribuem para a construção do cluster 1 são o "último valor", "dívidas", "idade", "n_dívidas" e "último_pagamento" (Figura 28).

1º cluster: Devedores mais velhos, menos cautelosos e mais propensos ao risco.

Grupo que se caracteriza por indivíduos mais velhos, com dívidas de montantes superiores, com número de dias de incumprimento bastantes elevados, bem como o número de dias desde o último pagamento. São também indivíduos pouco cautelosos pois têm várias dívidas.

Cluster 2

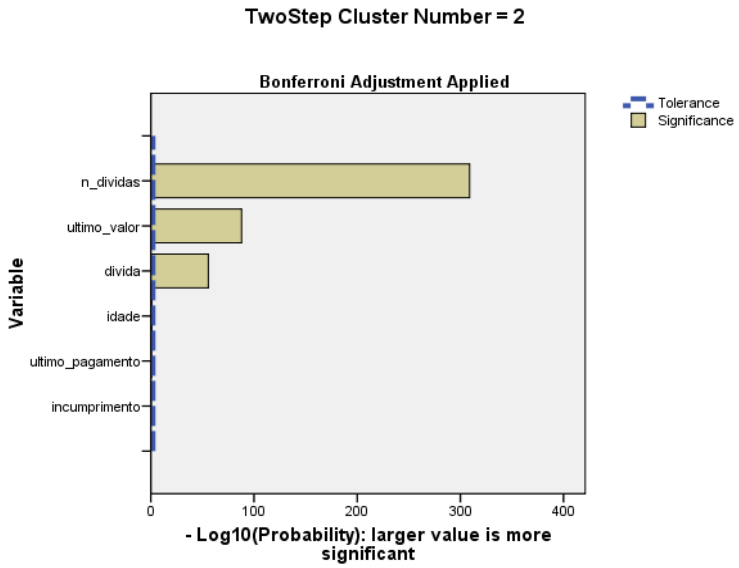

Figura 29 - Niveis de significância das variáveis numéricas, na formação do cluster 2.

As variáveis que mais contribuem para a construção do cluster 2 são o "último valor", "dívidas", "n_dívidas" e "último_pagamento" (Figura 29).

2º cluster: Devedores jovens e mais cautelosos

Grupo que se caracteriza por devedores mais jovens, mais cautelosos, pois o montante da dívida e o número de dias desde o último pagamento são mais baixos, e apenas têm em média uma dívida.

Assim, segundo esta análise realizada para o cliente B da empresa XPTO, pode-se concluir que os indivíduos pertencentes ao cluster 1 são aqueles que apresentam maior risco para empresa quanto à recuperação dos valores em dívida, pois apresentam divídas mais elevadas e com períodos de incumprimento bastante superiores à média.

Da análise realizada aos dois clientes, pode-se concluir, que apesar de se tratar de clientes distintos, com serviços distintos, existem algumas particularidades entre eles, como é o facto de os grupos com devedores mais jovens, serem mais cautelosos e apresentarem um risco de recuperação mais reduzido. Os grupos com devedores mais velhos, são menos racionais, mais atrevidos e mais propensos ao risco.[4]

4.4.Análise discriminante

Após a análise de clusters e conforme referido no Capitulo 3, é realizada a análise discriminante, utilizada para identificar qual ou quais das variáveis sob estudo permitem discriminar significativamente os grupos formados durante a análise de clusters.

4.4.1. Cliente A

4.4.1.1. Selecção de variáveis

Um dos primeiros passos da análise discriminante é a selecção das variáveis discriminantes, recorrendo-se à estatística de Λ Wilks para cada variável.

Teste de igualdade de médias

	Wilks' Lambda	F	df1	df2	Sig.
Tranches	0,314	2387,464	2	2189	0,000
Produto	0,164	5588,068	2	2189	0,000
Antiguidade_divida	1,000	0,309	2	2189	0,734
Idade	0,968	35,628	2	2189	0,000
NºDividas	0,997	3,128	2	2189	0,044
Outras_dividas	0,868	166,091	2	2189	0,000
Rendimento	0,979	24,046	2	2189	0,000
Divida	0,669	544,119	2	2189	0,000

Tabela 25 – Teste da igualdade de médias

Por observação da tabela pode-se concluir que a hipótese nula é rejeitada para a maioria das variáveis, com excepção da variável Antiguidade_dívida, ou seja, para as restantes

[4] Para ambos os clientes procedeu-se à mesma análise com as variáveis estandardizadas, no entanto obtiveram-se os mesmos resultados

variáveis, as médias são diferentes em pelo menos um grupo. Com um nível de significância de 1%, todas as variáveis tem médias diferentes para pelo menos um grupo.

Na Tabela 26 podem-se analisar as correlações existentes entre as diferentes variáveis, onde se observa que as correlações entre as variáveis são muito baixas pelo que faz sentido a interpretação dos seus coeficientes na matriz de estrutura. No entanto pode verificar-se que as variáveis "produto" e "outras_dívidas" têm uma alta correlação negativa, o que pode significar que com determinado produto, o número de dívidas por devedor diminui. Neste caso, se tivermos em conta a codificação, 1-Produtos de revolving e 2-Produtos clássicos, pode-se concluir que com os produtos clássicos os devedores têm menos dívidas.

Matriz de correlações e covariâncias (a)

		Tranches	Produto	Antiguidade_divida	Idade	NºDividas	Outras_dividas	Rendimento	Divida
Covariância	Tranches	0,504	-0,014	0,869	0,406	0,033	0,015	-8,819	641,102
	Produto	-0,014	0,016	0,065	0,019	-0,006	-0,016	-0,500	-16,135
	Antiguidade_divida	0,869	0,065	9283,064	10,944	6,115	-0,022	177,406	-2495,585
	Idade	0,406	0,019	10,944	133,427	1,080	-0,018	747,659	3807,281
	NºDividas	0,033	-0,006	6,115	1,080	0,533	0,006	24,232	62,707
	Outras_dividas	0,015	-0,016	-0,022	-0,018	0,006	0,017	-0,672	17,792
	Rendimento	-8,819	-0,500	177,406	747,659	24,232	-0,672	213576,655	190953,570
	Divida	641,102	-16,135	-2495,585	3807,281	62,707	17,792	190953,570	15960732,186
Correlação	Tranches	1,00	-0,15	0,01	0,05	0,06	0,17	-0,03	0,23
	Produto	-0,15	1,00	0,01	0,01	-0,06	-0,96	-0,01	-0,03
	Antiguidade_divida	0,01	0,01	1,00	0,01	0,09	0,00	0,00	-0,01
	Idade	0,05	0,01	0,01	1,00	0,13	-0,01	0,14	0,08
	NºDividas	0,06	-0,06	0,09	0,13	1,00	0,07	0,07	0,02
	Outras_dividas	0,17	-0,96	0,00	-0,01	0,07	1,00	-0,01	0,03
	Rendimento	-0,03	-0,01	0,00	0,14	0,07	-0,01	1,00	0,10
	Divida	0,23	-0,03	-0,01	0,08	0,02	0,03	0,10	1,00

(a) The covariance matrix has 2189 degrees of freedom.

Tabela 26 – Matriz de correlações e covariâncias

4.4.1.2. Funções Discriminantes

Segundo o teste de Wilks, o nº máximo de funções discriminantes é igual ao nº de grupos menos um, ou ao nº de variáveis discriminantes, sendo o critério de escolha baseado no menor destes dois valores, no caso em estudo, como se obteve 3 clusters, tem-se no máximo 2 funções discriminantes.

A Tabela 27 evidencia os valores próprios, que segundo Maroco (2010) são uma medida relativa do quão diferentes são os grupos na função discriminante. Neste caso obtêm-se duas funções discriminantes, onde na primeira função discriminante o valor próprio é 86,631 e explica 97,3% das diferenças entre os grupos, em contrapartida a segunda função tem de valor próprio 2,360 e explica 2,7% das diferenças entre grupos. Além dos valores próprios é também apresentada a correlação canónica, que demonstra o nivel de associação entre os

scores discriminantes e os grupos. Para utilizar estes valores como percentagem da variável dependente explicada pelo modelo, de acordo com Hair *et al.* (1998), deve-se elevar o resultado da correlação ao quadrado, $R_1^2 = (0{,}994)^2 = 99\%$, donde se conclui que a função explica 99% da discriminação entre os grupos, procedendo da mesma forma para a segunda função, tem-se que esta explica 70% da discriminação entre os grupos.

Valores próprios

Funções	Valores próprios	% de variância	%Acumulada	Correlação Canónica
1	86,631 (a)	97,3	97,3	0,994
2	2,360 (a)	2,7	100,0	0,838

(a) First 2 canonical discriminant functions were used in the analysis.

Tabela 27 – Importância das funções discriminantes

Ao observar a tabela pode-se concluir que o valor próprio da segunda função discriminante é bastante reduzido, com um poder discriminatório de apenas 2,7%, será que esta ainda é significativa?

Na Tabela 28, a estatística de Wilks permite testar a significância das funções discriminantes. O valor da estatística de Wilks, na presença das duas funções aproxima-se bastante de zero (0,003), e como o valor de Λ é uma medida inversa do poder discriminatório, significa que as funções iniciais têm elevado poder discriminatório para discriminar os 3 clusters. Ao remover a segunda função, o poder discriminatório diminui, aumentando o valor de Λ para 0,298, No entanto, continua a ser possível rejeitar a hipótese nula, pelo que são significativas as duas funções.

Wilks' Lambda

Teste da funções	Wilks' Lambda	Chi-square	df	Sig.
1 through 2	0,003	12433,0	10	0,000
2	0,298	2650,2	4	0,000

Tabela 28 – Poder discriminatório residual

Na Tabela 29 são apresentados os coeficientes das funções discriminantes estandardizados, onde é possível analisar quais as variáveis com maior poder de separar os grupos. Na primeira função, as variáveis com maior contribuição são produto, rendimento e outras_dívidas, enquanto na segunda função são as variáveis tranches, produto, rendimento e dívida.

Coeficientes estandardizados

	Função	
	1	2
Tranches	-0,048	0,927
Produto	3,521	0,101
Outras_dividas	3,427	-0,064
Rendimento	0,077	0,082
Divida	-0,017	0,240

Tabela 29 – Coeficientes estandardizados

Como anteriormente se verificou que a correlação entre as variáveis é muito fraca, alguns autores sugerem que a importância relativa das variáveis nas funções pondere sobre a correlação existente entre estas e a função discriminante.

Assim as variáveis que estão mais correlacionadas com as funções, poderão contribuir para a interpretação das funções discriminantes e dos grupos que as discriminam (Maroco, 2010).

Na função 1 as variáveis que mais contribuem são o produto, outras_dívidas e rendimento e na função 2 são as tranches e dívida.

Uma forma de ver qual a contribuição de cada variável para a discriminação entre os grupos, é analisando a correlação entre os valores de cada variável explicativa com a função discriminante através da matriz de estrutura.

Matriz estrutura

	Função	
	1	2
Produto	0,243*	0,011
Outras_dividas	0,042*	0,002
Antiguidade_divida(a)	0,012*	0,011
Tranches	-0,021	0,953*
Divida	-0,013	0,453*
Rendimento	0,008	0,082*
Idade(a)	0,011	0,079*
NºDividas(a)	0,003	0,059*

Pooled within-groups correlations between discriminating variables and standardized canonical
* Largest absolute correlation between each variable and any discriminant function
(a) This variable not used in the analysis.

Tabela 30 – Matriz estrutura

Nesta matriz (Tabela 30) é evidenciada a contribuição de cada variável para a função discriminante, sendo as mais importantes as que têm asterisco e as que têm maior valor absoluto, pois serão aquelas que mais contribuem para a função discriminante.

A Tabela 31 apresenta as variáveis seleccionadas para compor as funções e seus respectivos coeficientes não-estandardizados.

Coeficientes não-estandardizados

	Função	
	1	2
Tranches	-0,068	1,306
Produto	27,494	0,789
Outras_dividas	26,501	-0,496
Rendimento	0,000	0,000
Divida	0,000	0,000
(Constant)	-30,919	-6,883

Tabela 31 – Coeficientes não-estandardizados

Pode-se então escrever as duas funções discriminantes, da seguinte forma:

$$Z_1 = -30,919 + 27,494 \times \text{produto} - 0,068 \times \text{tranches} + 26,501 \times \text{outras\_dividas}$$
$$Z_2 = -6,883 + 0,789 \times \text{produto} + 1,306 \times \text{tranches} - 0,496 \times \text{outras\_dividas}$$

A interpretação das funções discriminantes também pode ser feita graficamente, onde cada grupo é representado pelo seu centróide e vai ser tratado como um ponto, e cada função discriminante como um eixo.

No gráfico seguinte (Figura 30) podem-se visualizar os clusters considerados e a distribuição dos indivíduos em cada cluster. Os indivíduos do cluster 3 são os mais bem classificados, em contrapartida os indivíduos do cluster 2 são os mais mal classificados, pois pode ver-se que existem indivíduos que se afastam muito do centróide, ou seja, há uma maior heterogeneidade no grupo.

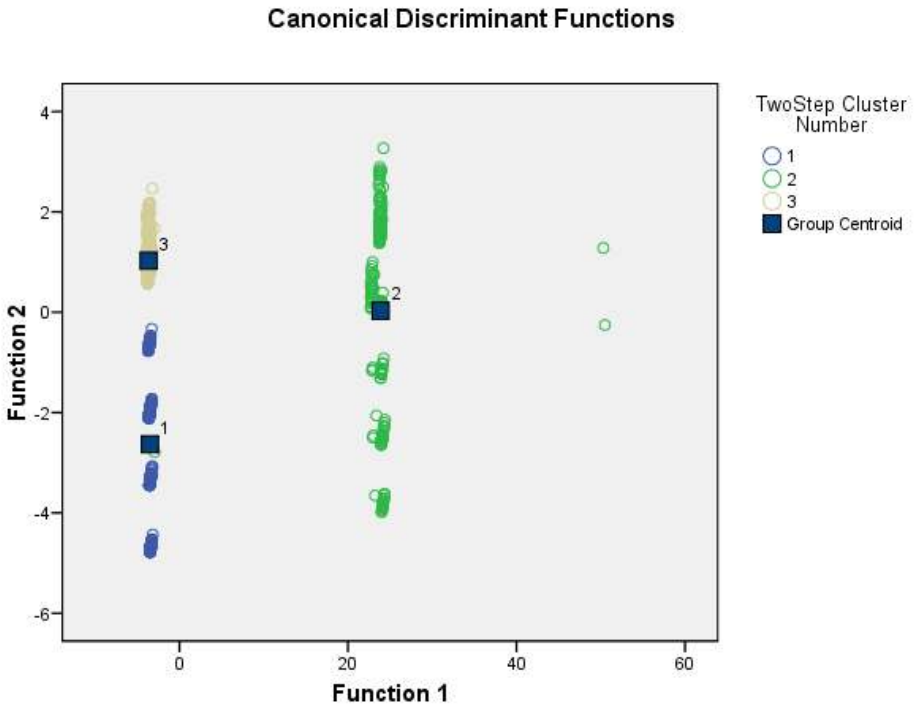

Figura 30- Representação gráfica dos centróides de cada cluster nas funções discriminantes

No gráfico anterior pode observar-se que o grupo 2 tem o valor máximo da função 1 e o grupo 1 o valor mínimo da função 2. A função 1 está associada essencialmente a outras dívidas e produtos, pode dizer-se que os devedores com mais dívidas em outros clientes e que recorrem a diversos produtos, tendem a ser mais atrevidos e mais conhecedores das ofertas de crédito. A função 2, tende a separar o grupo 1 (valores mais baixos) do grupo 3 (valores mais altos). Dadas as correlações positivas de variáveis como o rendimento, valor da dívida, tranche de montante e número de dívidas com a função 2, pode concluir-se que para valores mais baixos do rendimento, dos valores em dívida e do número de dívidas, os devedores tendem a ser mais cautelosos, mais avessos ao risco e contraem menos dívidas. Inversamente, para valores do rendimento mais altos, valores de dívida mais elevados e um maior número de dívidas, os devedores tendem a ser mais propensos ao risco, contraindo mais dívidas e com valores mais elevados.

4.4.1.3. Classificação dos indivíduos

Após se obterem as funções discriminantes calculam-se os coeficientes que irão permitir realizar a classificação dos indivíduos em estudo, nos respectivos grupos.

Para se afectar um indivíduo a um determinado grupo, existem vários procedimentos possíveis. Como se está a trabalhar em SPSS, será então necessário encontrar uma equação de classificação para cada grupo. A equação básica de classificação de um indivíduo no *j-ésimo* grupo tem a seguinte forma:

$$C_j = c_{jo} + c_{j1}X_1 + c_{j2}X_2 + ... + c_{jp}X_P$$

O indivíduo j será afectado ao grupo no qual se obtenha um maior resultado de classificação.

Na Tabela 32 têm-se os coeficientes de classificação.

Equações classificatórias

	TwoStep Cluster Number		
	1	2	3
Tranches	4,600	6,216	9,394
Produto	758,246	1512,786	756,373
Outras_dividas	716,206	1440,146	709,801
Rendimento	0,007	0,012	0,008
Divida	0,000	0,000	0,000
(Constant)	-388,390	-1528,735	-404,983

Fisher's linear discriminant functions

Tabela 32- Equações classificatórias

Por último é apresentado o resultado da classificação dos indivíduos em estudo, que permite analisar a eficácia classificativa do modelo.

Matriz de classificações (a)

			Grupo previsto			
	TwoStep Cluster Number		1	2	3	Total
Original	Contagem	1	438	0	99	537
		2	1	288	0	289
		3	0	0	1366	1366
	%	1	81,6	0,0	18,4	100,0
		2	0,3	99,7	0,0	100,0
		3	0,0	0,0	100,0	100,0

(a) 95,4% of original grouped cases correctly classified.

Tabela 33 – Matriz de classificações

Ao observar a Tabela 33 pode-se verificar que 99 indivíduos foram inicialmente classificados no cluster 1, no entanto pelas suas características aproximam-se mais do perfil de devedores mais velhos e mais propensos ao risco. Enquanto apenas um indivíduo inicialmente classificado no cluster 2, foi classificado no cluster 1 por se aproximar mais do perfil dos devedores mais jovens, cautelosos e mais avessos ao risco.

Verifique-se a eficácia classificativa do modelo calculando a probabilidade de classificação correcta:

$PCC = 95,4\%$

Determina-se de seguida os critérios do acaso máximo e do acaso proporcional

$$C_{MAX} = \max_j \frac{n_j}{n} = 0,623$$

$$C_{PRO} = \sum_{j=1}^{k} \left(\frac{n_j}{n} \right)^2 = 0,4656$$

Como o valor dos casos correctamente classificados é superior a estes dois critérios, pode-se aceitar a eficácia classificativa da análise realizada.

4.4.1.4. Validação dos resultados

Para que esta análise fique completa é necessária a verificação dos pressupostos: normalidade multivariada das variáveis independentes e homogeneidade das matrizes de variância e co-variância (Hair *et a.l*, 1998).

Normalidade

Teste de Normalidade

	Kolmogorov-Smirnov(a)			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Região	0,262	2192	0,000	0,849	2192	0,000
Divida	0,132	2192	0,000	0,871	2192	0,000
Tranches	0,421	2192	0,000	0,631	2192	0,000
Produto	0,526	2192	0,000	0,368	2192	0,000
Profissão	0,352	2192	0,000	0,753	2192	0,000
Antiguidade_divida	0,236	2192	0,000	0,796	2192	0,000
Idade	0,068	2192	0,000	0,972	2192	0,000
Nº Dividas	0,400	2192	0,000	0,634	2192	0,000
Sexo	0,364	2192	0,000	0,634	2192	0,000
Outras_dividas	0,537	2192	0,000	0,118	2192	0,000
Rendimento	0,137	2192	0,000	0,860	2192	0,000

(a) Lilliefors Significance Correction

Tabela 34 – Testes da normalidade das variáveis

O teste de Kolmogorov-Smirnov (Tabela 34) indica que as variáveis não seguem uma distribuição normal, no entanto, e dada a dimensão da amostra, pelo Teorema do Limite Central pode dizer-se que a distribuição das variáveis se aproxima de uma distribuição normal.

O pressuposto da homogeneidade das matrizes de variância e co-variância, é analisado através do teste Box's Muller. No entanto, e dada a existência de uma matriz singular nos clusters 1 e 3, não foi possível a realização do teste.

A decisão de aceitar os resultados depende da verificação dos pressupostos e caso estes não se verifiquem, coloca-se a questão: Podem considerar-se válidos os resultados da análise discriminante?

De acordo com alguns autores, a análise discriminante é robusta face a violações (Maroco, 2010) e como o tamanho da amostra é bastante grande, a análise poderá ser válida. Quanto à não existência de heterogeneidade, segundo alguns autores, se a correlação entre as médias e as variâncias é reduzida, a validade da análise discriminante pode ser assegurada (Friedman, 1989; McLachlan, 2004). Apesar de tudo, há que analisar cautelosamente os resultados obtidos.

4.4.2. Cliente B

4.4.2.1. Selecção das variáveis

Na Tabela 35, pode-se verificar que a hipótese nula, de igualdade de médias, é rejeitada para todas as variáveis à excepção da variável incumprimento, ou seja, para as variáveis escolhidas algumas têm médias diferentes em pelo menos um grupo, sendo significativas na diferenciação dos grupos.

Teste de igualdade de médias

	Wilks' Lambda	F	df1	df2	Sig.
idade	0,994	79,030	1	13186	0,000
divida	0,717	5215,283	1	13186	0,000
ultimo_pagamento	0,997	40,935	1	13186	0,000
incumprimento	1,000	2,161	1	13186	0,142
ultimo_valor	0,673	6399,169	1	13186	0,000
n_dividas	0,997	39,168	1	13186	0,000

Tabela 35 - Teste da igualdade de médias

Na Tabela 36 podem-se analisar as correlações existentes entre as diferentes variáveis, onde se observa que as correlações entre as variáveis são muito baixas pelo que faz sentido a interpretação dos seus coeficientes na matriz de estrutura.

Pooled Within-Groups Matrices(a)

		idade	divida	ultimo_pagamento	incumprimento	ultimo_valor	n_dividas
Covariância	idade	129,355	3205,646	225,547	160,559	129,360	0,305
	divida	3205,646	6177549,958	96047,221	109099,718	-125337,517	173,880
	ultimo_pagamento	225,547	96047,221	117230,651	66373,165	3691,717	5,814
	incumprimento	160,559	109099,718	66373,165	123531,605	-3052,924	-8,146
	ultimo_valor	129,360	-125337,517	3691,717	-3052,924	155270,015	3,826
	n_dividas	0,305	173,880	5,814	-8,146	3,826	0,259
Correlação	idade	1,000	0,113	0,058	0,040	0,029	0,053
	divida	0,113	1,000	0,113	0,125	-0,128	0,137
	ultimo_pagamento	0,058	0,113	1,000	0,552	0,027	0,033
	incumprimento	0,040	0,125	0,552	1,000	-0,022	-0,045
	ultimo_valor	0,029	-0,128	0,027	-0,022	1,000	0,019
	n_dividas	0,053	0,137	0,033	-0,045	0,019	1,000

(a) The covariance matrix has 13186 degrees of freedom.

Tabela 36 – Matriz de correlações e covariâncias

4.4.2.2. Funções discriminantes

De seguida determina-se as funções discriminantes para o conjunto de dados.

Na função discriminante obtida (Tabela 37), o valor próprio é igual a 1,018, explicando 100% das diferenças entre os grupos.

Valores próprios

Funções	Valores próprios	% de variância	%Acumulada	Correlação Canónica
1	1,018 (a)	100,0	100,0	0,710

(a) First 1 canonical discriminant functions were used in the analysis.

Tabela 37 - Importância das funções discriminantes

Na Tabela 38 pode-se ver que a função discriminante obtida tem um poder discriminativo significativo, com o valor da estatística de Wilks igual a 0,496.

Wilks' Lambda

Teste da funções	Wilks' Lambda	Chi-square	df	Sig.
1	0,496	9257,129	5	0,000

Tabela 38 - Poder discriminatório residual

Na Tabela 39 são apresentados os coeficientes das funções discriminantes estandartizados, onde se pode analisar que as variáveis com maior poder para separar os grupos são a idade e o último valor.

Coeficientes estandardizados

	Function
	1
idade	-0,024
divida	0,744
incumprimento	-0,065
ultimo_valor	0,786
n_dividas	-0,065

Tabela 39 - Coeficientes estandardizados

Como tinha sido referido anteriormente para avaliar a contribuição de cada variável para a função discriminante, deve-se analisar a matriz de estrutura (Tabela 40).

Matriz estrutura

	Function
	1
ultimo_valor	0,690
divida	0,623
idade	0,077
ultimo_pagamento(a)	0,066
n_dividas	0,054
incumprimento	0,013

Pooled within-groups correlations between discriminating variables and standardized canonical

(a) This variable not used in the analysis.

Tabela 40 - Matriz estrutura

Pode-se concluir que as variáveis que mais contribuem para a função discriminante são: "ultimo_valor", "dívida" e "n_dívidas", no entanto as restantes também contribuem mas de forma menos significativa. Além disso, estas variáveis, dão uma contribuição bastante significativa para a classificação dos devedores, como mais cautelosos e menos cautelosos.

Na Tabela 41 têm-se os coeficientes da função discriminante:

Coeficientes não-estandardizados

	Function
	1
idade	-0,002
divida	0,000
incumprimento	0,000
ultimo_valor	0,002
n_dividas	-0,127
(Constant)	-0,551

Unstandardized coefficients

Tabela 41 - Coeficientes não-estandardizados

Pode-se então escrever a função discriminante, da seguinte forma:

$$Z_1 = -0,551 - 0,002 \times \text{idade} + 0,002 \times \text{ultimo\_valor} - 0,127 \times \text{n\_dividas}$$

4.4.2.3. Classificação dos indivíduos

Tal como se procedeu para o cliente A, de seguida é apresentada a classificação dos indivíduos em estudo, de forma a analisar a eficácia classificativa do modelo.

Na Tabela 42 têm-se os coeficientes de classificação

Classification Results(a)

			Grupo previsto		
	TwoStep Cluster Number		1	2	Total
Original	Contagem	1	262	28	290
		2	40	12858	12898
	%	1	90,3	9,7	100,0
		2	0,3	99,7	100,0

(a) 99,5% of original grouped cases correctly classified.

Tabela 42 - Matriz de classificações

Ao observar a tabela pode-se verificar que 28 indivíduos foram inicialmente classificados no cluster 1, no entanto pelas suas características aproximam-se mais do perfil dos devedores do cluster 2. No cluster 2 inicialmente foram classificados 40 indivíduos, que posteriormente, pelas suas características serem mais semelhantes às dos devedores mais velhos, menos cautelosos e mais propensos ao risco, foram classificados no cluster 1.

A eficácia classificativa do modelo obtida pelo cálculo da probabilidade de classificação correcta é:

$$PCC = 99,5\%$$

Determina-se de seguida os critérios do acaso máximo e do acaso proporcional

$$C_{MAX} = \max_j \frac{n_j}{n} = 0,978$$

$$C_{PRO} = \sum_{j=1}^{k} \left(\frac{n_j}{n} \right)^2 = 0,957$$

Como o valor dos casos correctamente classificados é superior a estes dois critérios, pode-se aceitar a eficácia classificativa da análise realizada.

4.4.2.4. Validação dos resultados

De seguida são verificados os pressupostos: normalidade multivariada das variáveis independentes e homogeneidade das matrizes de variância e co-variância (Hair *et a.l*, 1998).

Na Tabela 43 encontra-se o resultado do teste de Kolmogorov-Smirnov.

Normalidade

Teste de Normalidade

	Kolmogorov-Smirnov(a)		
	Statistic	df	Sig.
idade	0,073	13188	0,000
divida	0,285	13188	0,000
ultimo_pagamento	0,131	13188	0,000
incumprimento	0,133	13188	0,000
ultimo_valor	0,336	13188	0,000
n_dividas	0,500	13188	0,000

(a) Lilliefors Significance Correction

Tabela 43 – Teste de Kolmogorov-Smirnov

O teste de Kolmogorov-Smirnov indica que as variáveis não seguem uma distribuição normal, no entanto, e dada a dimensão da amostra, pelo Teorema do Limite Central pode dizer-se que a distribuição das variáveis se aproxima de uma distribuição normal.

O pressuposto da homogeneidade das matrizes de variância e co-variância, é analisado através do teste Box's Muller, cujo resultado se encontra na Tabela 44.

Teste de Box's Muller

Box's M		21789,23
F	Approx.	114,652
	df1	15
	df2	944575,6
	Sig.	0,000

Testes null hypothesis of equal population covariance matrices

Tabela 44 – Teste de Box's Muller

Segundo o resultado do teste Box's Muller o pressuposto da heterogeneidade não é verificado. Novamente, se coloca a mesma questão, será que se podem considerar válidos os resultados da análise discriminante?

Da mesma forma que anteriormente, e segundo alguns autores, a análise discriminante é robusta face a violações (Maroco, 2010) e como o tamanho da amostra é bastante grande, pode considerar-se análise válida. Quanto à não existência de heterogeneidade, a validade da análise discriminante pode ser assegurada (Friedman, 1989; McLachlan, 2004). Apesar de tudo, há que analisar cautelosamente os resultados obtidos.

4.5.Análise de componentes principais

4.5.1. Cliente A

4.5.1.1. Validação dos resultados

Antes de se iniciar a análise ACP devem ser verificados os pressupostos, bem como a validade da análise.

Como se viu anteriormente o pressuposto da normalidade falha em todas as variáveis com o teste de Kolmogorov-Smirnov. De seguida analisou-se a normalidade usando o teste da assimetria, ou seja, consultando os valores na tabela das estatísticas descritivas para cada variável e concluiu-se que a seguinte relação $-1{,}96 < \frac{skewness}{std.skewness} < 1{,}96$, nunca se verificou, obtendo-se novamente que nenhuma das variáveis segue uma distribuição normal.

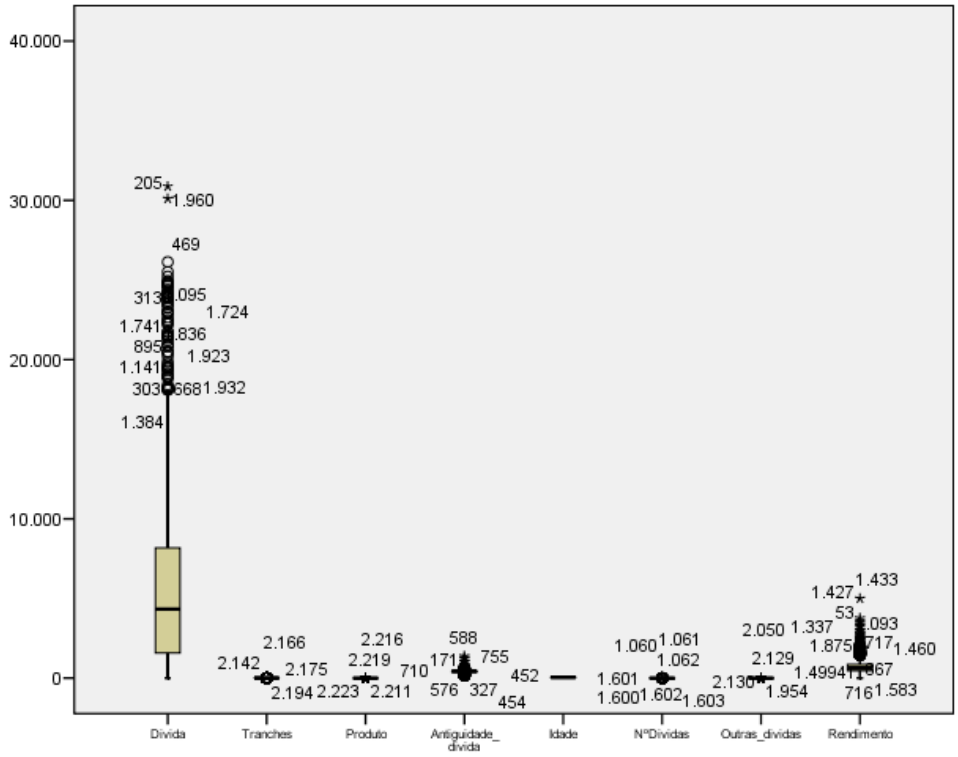

Figura 31 – Diagrama de extremos e quartis das variáveis

Pode-se concluir por observação do gráfico anterior (Figura 31), que para todas as variáveis existem outliers. Na análise descritiva da amostra pode também concluir-se que nenhuma das variáveis verifica o pressuposto de simetria e achatamento.

Na Tabela 45 – pode analisar-se a matriz de correlações:

Matriz de correlações

		Tranches	Produto	Divida	Antiguidade _divida	Idade	NºDividas	Outras_dividas	Rendimento
Correlação	Tranches	1,000	-0,114	0,542	0,018	0,175	0,078	0,049	0,079
	Produto	-0,114	1,000	-0,003	0,196	-0,016	-0,049	-0,033	0,078
	Divida	0,542	-0,003	1,000	-0,003	0,155	0,035	-0,011	0,151
	Antiguidade_divida	0,018	0,196	-0,003	1,000	-0,019	-0,045	-0,013	0,022
	Idade	0,175	-0,016	0,155	-0,019	1,000	0,133	-0,017	0,154
	NºDividas	0,078	-0,049	0,035	-0,045	0,133	1,000	0,059	0,070
	Outras_dividas	0,049	-0,033	-0,011	-0,013	-0,017	0,059	1,000	0,017
	Rendimento	0,079	0,078	0,151	0,022	0,154	0,070	0,017	1,000
Sig. (1-tailed)	Tranches		0,000	0,000	0,198	0,000	0,000	0,010	0,000
	Produto	0,000		0,445	0,000	0,221	0,011	0,062	0,000
	Divida	0,000	0,445		0,451	0,000	0,050	0,307	0,000
	Antiguidade_divida	0,198	0,000	0,451		0,191	0,017	0,275	0,152
	Idade	0,000	0,221	0,000	0,191		0,000	0,205	0,000
	NºDividas	0,000	0,011	0,050	0,017	0,000		0,003	0,000
	Outras_dividas	0,010	0,062	0,307	0,275	0,205	0,003		0,212
	Rendimento	0,000	0,000	0,000	0,152	0,000	0,000	0,212	

Tabela 45 – Matriz de correlações

É possivel observar que existem muitas variáveis que estão correlacionadas com valor absoluto baixo, pelo que se conclui que é pouco provável que partilhem factores comuns.

Através da estatística de KMO e o teste de Esfericidade de Bartlett's averigua-se se a aplicação da análise em componentes principais tem validade para as variáveis escolhidas.

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		0,538
Bartlett's Test of Sphericity	Approx. Chi-Square	1172,820
	df	28
	Sig.	0,000

Tabela 46 – Estatística de KMO e teste de Bartlett's

Como se pode verificar na Tabela 46, a estatística de KMO apresenta um valor de 0,538, o que significa que a análise factorial a realizar terá resultado mau. O teste de Bartlett's testa a hipótese da matriz de correlações ser uma matriz de identidade e como se pode verificar pelo *p-value*=0,00, rejeita-se a hipótese nula, concluindo-se com base nestes dados amostrais, que a matriz de correlações é diferente da matriz identidade.

Dado que não existem condições estatísticas para uma análise de componentes principais com qualidade, esta não será desenvolvida neste trabalho.

4.5.2. Cliente B

4.5.2.1. Validação dos resultados

Antes de iniciar a análise de componentes principais, verificam-se os pressupostos, e a validade da análise.

Como se viu anteriormente o pressuposto da normalidade falha em todas as variáveis com o teste de Kolmogorov-Smirnov. De seguida analisou-se a normalidade usando o teste da assimetria, ou seja, consultando os valores na tabela das estatísticas descritivas para cada variável e concluiu-se que a seguinte relação $-1{,}96 < \frac{skewness}{std.skewness} < 1{,}96$, nunca se verificou, obtendo-se novamente que nenhuma das variáveis segue uma distribuição normal.

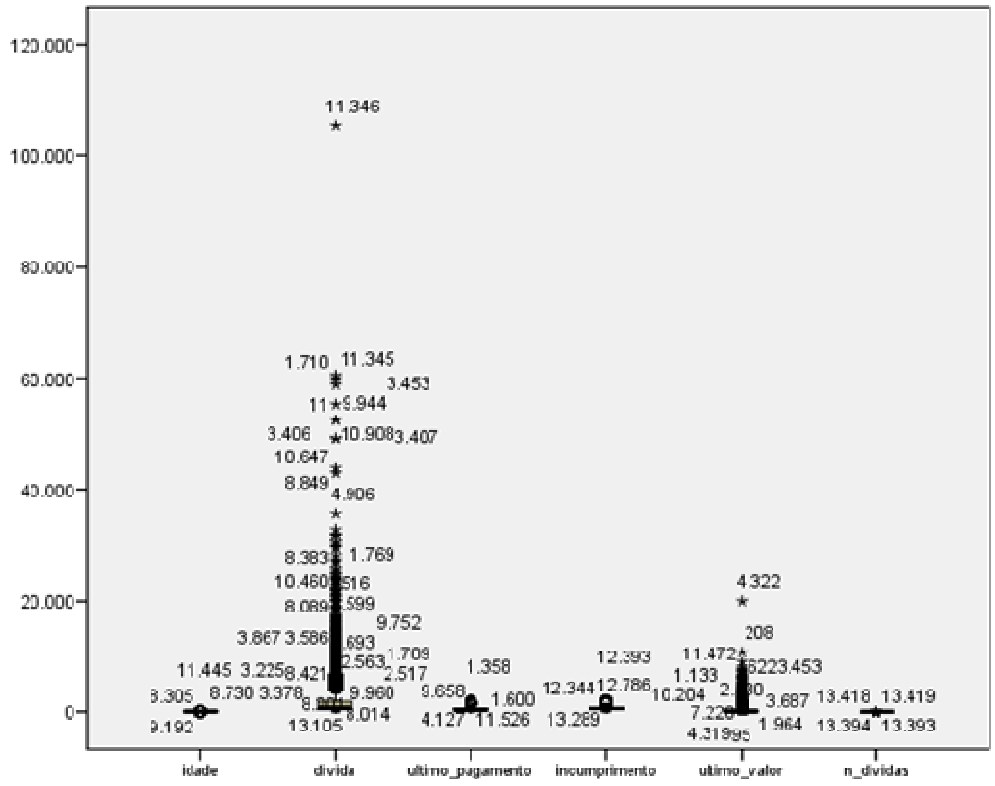

Figura 32 – Diagrama de extremos e quartis das variáveis

Pode-se concluir por observação do gráfico anterior (Figura 32), que para todas as variáveis existem outliers. Na análise descritiva da amostra pode também concluir-se que nenhuma das variáveis verifica o pressuposto de simetria e achatamento.

Na Tabela 47 pode analisar-se a matriz de correlações:

Matriz de correlações

		idade	divida	ultimo_pagamento	incumprimento	ultimo_valor	n_dividas
Correlação	idade	1,000	0,139	0,061	0,039	0,068	-0,001
	divida	0,139	1,000	0,126	0,113	0,216	0,015
	ultimo_pagamento	0,061	0,126	1,000	0,551	0,053	-0,001
	incumprimento	0,039	0,113	0,551	1,000	-0,011	-0,009
	ultimo_valor	0,068	0,216	0,053	-0,011	1,000	0,004
	n_dividas	-0,001	0,015	-0,001	-0,009	0,004	1,000
Sig. (1-tailed)	idade		0,000	0,000	0,000	0,000	0,473
	divida	0,000		0,000	0,000	0,000	0,040
	ultimo_pagamento	0,000	0,000		0,000	0,000	0,475
	incumprimento	0,000	0,000	0,000		0,104	0,138
	ultimo_valor	0,000	0,000	0,000	0,104		0,337
	n_dividas	0,473	0,040	0,475	0,138	0,337	

Tabela 47-Matriz de correlações

É possivel observar que existem muitas variáveis que estão correlacionadas com valor absoluto baixo, pelo que se conclui que é pouco provável que partilhem factores comuns.

Através da estatística de KMO e o teste de Esfericidade de Bartlett's averigua-se se a aplicação da análise em componentes principais tem validade para as variáveis escolhidas.

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		0,524
Bartlett's Test of Sphericity	Approx. Chi-Square	6117,596
	df	15
	Sig.	0

Tabela 48 – Estatistica de KMO e teste de Bartlett's

Como se pode verificar a estatística de KMO (Tabela 48) apresenta um valor de 0,524, o que significa que a análise factorial a realizar terá resultado mau. O teste de Bartlett's testa a hipótese da matriz de correlações ser uma matriz de identidade e como se pode verificar pelo o valor de *p-value*=0,00, rejeita-se a hipótese nula, concluindo-se com base nestes dados amostrais, que a matriz de correlações é diferente da matriz identidade.

Dado que não existem condições estatísticas para uma análise de componentes principais com qualidade, esta não será desenvolvida neste trabalho.

4.6.Regressão logística

De seguida irá ser avaliada a significância de determinadas variáveis sobre a probabilidade de os devedores serem cumpridores, recorrendo-se à regressão logistica.

Com base nos pagamentos efectuados pelo devedor durante o período de gestão, ou seja, 90 dias, foi definida a variável cumpridor, de acordo com a seguinte classificação:

- Devedores cumpridores - devedores com pagamentos durante os 90 dias do período de gestão da carteira;

- Devedores não cumpridores – devedores sem pagamentos durante os 90 dias do período de gestão da carteira.

Foi realizado este corte, a 90 dias, pois normalmente a partir deste período os processos são devolvidos ao cliente, para que este possa terminar a gestão dessas dívidas, que na maioria dos casos, são encaminhados para processos judiciais.

Para esta análise foram consideradas as seguintes variáveis:

- sexo;
- dívida;
- número de dívidas;
- dias de incumprimento
- número de acordos em ruptura;
- número de chamadas realizadas com sucesso;
- antiguidade

A escolha das variáveis teve como critério de selecção alguns dos estudos referidos no Capitulo 2, mas também o facto de a empresa disponibilizar esta informação mensalmente aos seus clientes, para os informar sobre a gestão das dívidas. Como tal, seria interessante, além de fornecer este tipo de informação, enriquece-la com o comportamento que cada variável tem sobre o facto de os devedores serem ou não cumpridores. Foram utilizadas variáveis diferentes entre os vários clientes pois a informação existente para os dois clientes é diferente.

De acordo com Hair (1998), o tamanho da amostra deve seguir os seguintes padrões:

- Proporção Mínima: 5 observações para cada variável independente. Como no modelo foram utilizadas 8 variáveis independentes, deveria ter-se no mínimo 40 observações;

- Mínimo de 20 observações por grupo.

4.6.1. Cliente A

Para este estudo foram considerados 2197 casos, não existindo missing cases, nem casos não seleccionados (vide Tabela 49).

Resumo dos casos seleccionados

Unweighted Cases(a)		N	Percentagem
Casos seleccionados	Incluidos na análise	2197	100
	Casos em falta	0	0
	Total	2197	100
Casos não seleccionados		0	0
Total		2197	100

(a) If weight is in effect, see classification table for the total number of cases.

Tabela 49 – Resumo dos casos seleccionados, em falta e não seleccionados.

Na Tabela 50 pode observar-se a primeira etapa desta análise, onde se considera o modelo nulo, ou seja, o modelo só com a constante, a classificação dos indivíduos dentro dos grupos, cumpridor e não cumpridor, as variáveis consideradas e as não consideradas na análise.

Tabela de classificação (a,b)

Observados			Previstos		Percentagem correcta
			Incumprimento		
			Não cumpridor	Cumpridor	
Step 0	Incumprimento	Não cumpridor	2057	0	100
		Cumpridor	140	0	0
	Overall Percentage				93,6

a Constant is included in the model.

b The cut value is ,500

Tabela 50 – Tabela de classificações

Na Tabela 50, pode-se observar que com o modelo nulo a percentagem de individuos bem classificados é de 93,6%, sendo que no grupo 1 todos os individuos estão bem classificados e no grupo 2 estão todos mal classificados.

Variáveis na equação

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	-2,687	0,087	946,639	1	0,000	0,068

Tabela 51 – Variáveis consideradas na equação

Na Tabela 51 tem-se o valor da constante na equação e o teste de Wald, onde se observa que a constante influencia o modelo de forma significativa.

Na Tabela 52, observam-se as variáveis que não são consideradas no modelo nulo e o nível de significância de cada uma nesse modelo.

Variáveis não consideradas na equação (a)

			Score	df	Sig.
Step 0	Variáveis	Antiguidade_divida	75,095	1	0,000
		NºDividas	27,011	1	0,000
		Acordos_ruptura	403,136	1	0,000
		Chamadas_sucesso	96,973	1	0,000
	Overall Statistics		499,633	4	0,000

(a) Residual Chi-Squares are not computed because of redundancies.

Tabela 52 – Variáveis não consideradas na equação

De seguida pode analisar-se o modelo completo, onde se observa na Tabela 53 o teste do rácio de verosimilhança entre o modelo nulo e os modelos em cada etapa (step), bloco (block) e modelo final (model).

No caso em estudo foi utilizado o método Enter, e como só existe uma etapa, os valores para o teste são todos iguais, sendo o valor da estatística de teste $G^2(5)=305,383$, com *p-value<0,01*, a hipótese nula de ausência de variáveis significativas é rejeitada, pelo que se conclui que existe pelo menos uma variável independente no modelo com poder predictivo sobre a variável dependente.

Block 1: Method = Enter

Teste de rácio verosimilhança

		Chi-square	df	Sig.
Step 1	Step	305,383	4	0
	Block	305,383	4	0
	Model	305,383	4	0

Tabela 53 – Teste de rácio verosimilhança

Na Tabela 54 pode observar-se o resultado do teste de Hosmer e Lemeshow, que tem distribuição aproximadamente ao qui-quadrado e que permite testar o ajustamento do modelo, comparando as frequências observadas com as esperadas na regressão logística. Os resultados apontam para um razoável ajustamento do modelo dos dados.

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	8,021	8	0,431

Tabela 54 – Teste de Hosmer and Lemeshow

O valor do teste da bondade do ajustamento de Hosmer e Lemeshow para o incumprimento é de 8,021 com sig. = 0,431, o que mostra alguma segurança num ajustamento razoável entre as frequências observadas e as esperadas. Com este teste, sobre o ajustamento, apenas se pode concluir que as variáveis independentes influenciam Y, sem medir a força dessa associação.

Na regressão logística as medidas da força de associação são aproximações ao coeficiente de determinação R^2, o qual procura explicar a percentagem devida às variáveis independentes.

Modelo Sumário

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	736,398 (a)	0,130	0,344

(a) Estimation terminated at iteration number 8 because parameter estimates changed by less than ,001.

Tabela 55 – Teste de associação da variável dependente e das variáveis independentes

Por observação da Tabela 55 pode-se concluir novamente que o grau da associação das variáveis independentes com a variável dependente é significativo.

Na Tabela 56 é apresentada a classificação dos indivíduos observada e prevista pelo modelo ajustado. Da leitura desta tabela, pode-se concluir que dos 2057 indivíduos classificados como não cumpridores, 2046 estão bem classificados (99,5%) e que dos 140 classificados como cumpridores, apenas 37 estão bem classificados (26,4%). A classificação global do modelo ajustado é de 94,8%.

Tabela de classificação(a)

Observados			Previstos		Percentagem correcta
			Incumprimento		
			Não cumpridor	Cumpridor	
Step 1	Incumprimento	Não cumpridor	2046	11	99,5
		Cumpridor	103	37	26,4
Overall Percentage					94,8

(a) The cut value is ,500

Tabela 56 – Classificação dos individuos

Para avaliar a qualidade da classificação feita pelo modelo, como referido no Capitulo 2 pode comparar-se a percentagem global de classificações correctas obtidas com o modelo, com a percentagem proporcional de classificações correctas por acaso que é

$[(2057/2197)^2+(149/2197)^2]*100\%=89\%$. A percentagem de casos classificados correctamente pelo modelo é superior à percentagem de classificação proporcional por acaso, no entanto apenas é superior em aproximadamente 6%, não se podendo concluir que o modelo tem boas propriedades classificativas. Esta conclusão, já era de esperar, se se observar a percentagem dos cumpridores bem classificados, que é de apenas 26,4%.

Perante estes resultados, e na existência de algumas dúvidas sobre a capacidade discriminativa do modelo, recorreu-se à construção da curva de ROC.

A tabela 57 dá a área sob a curva ROC (*c*=0,866), que é significativamente superior a 0,5 (p<0,001).

Area Under the Curve

Test Result Variable(s): Predicted probability

Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
,866	,014	,000	,838	,894

The test result variable(s): Predicted probability has at least one tie between the positive actual state group and the negative actual state group. Statistics may be biased.

a. Under the nonparametric assumption

b. Null hypothesis: true area = 0.5

Tabela 57 - Área abaixo da curva ROC

O modelo ajustado apresenta uma capacidade discriminativa boa (Hosmer & Lemeshow, 2000).

Perante os resultados da tabela 56, onde se verificou uma grande diferença entre a sensibilidade e especificidade do modelo, e com vista a diminuir a especificidade de forma a tomar valores mais próximos dos valores da sensibilidade, foram analisados as várias coordenadas da curva ROC. Pôde verificar-se que com um ponto de corte de 0,0667, os valores da sensibilidade e especificidade eram mais próximos, foi então testado novamente o modelo, tendo-se verificado que para estes valores, 75% dos incumpridores se encontram bem classificados e 76,4% dos cumpridores também se encontram bem classificados. Posteriormente a esta análise, foi feita uma análise dos resíduos, não se verificando a existência de observações influentes ou outliers.

A Tabela 58 resume toda a informação sobre as variáveis independentes no modelo completo. Pode também observar-se o teste de Wald, que permite concluir que todas as variáveis influenciam de forma significativa o modelo.

Variáveis consideradas na equação

		B	S.E.	Wald	df	Sig.	Exp(B)	95,0% C.I.for EXP(B)	
								Lower	Upper
Step 1(a)	Antiguidade_divida	-0,006	0,001	36,843	1	0,000	0,994	0,993	0,996
	NºDividas	-0,622	0,236	6,926	1	0,008	0,537	0,338	0,853
	Acordos_ruptura	1,940	0,234	68,979	1	0,000	6,962	4,404	11,005
	Chamadas_sucesso	2,593	0,444	34,048	1	0,000	13,370	5,596	31,945
	Constant	-2,210	0,613	13,014	1	0,000	0,110		

(a) Variable(s) entered on step 1: Antiguidade_divida, NºDividas, Acordos_ruptura, Chamadas_sucesso.

Tabela 58 – Resumo das variáveis consideradas na equação

De acordo com o output obtido, o modelo pode escrever-se da seguinte forma:

$$\text{Logit}(\hat{\pi}) = -2,210 - 0,006 \times \text{Antiguidade\_divida} - 0,622 \times \text{Nº dividas} + 1,984 \times \text{acordos\_ruptura} + 2,593 \times \text{cham\_sucesso}$$

A coluna Exp(B) é a exponencial dos coeficientes do modelo que estimam o rácio das hipóteses da variável dependente por unidade da variável independente.

Assim, pode concluir-se que:

- Por cada 30 dias a possibilidade de se tornar cumpridor diminui 82%, o que seria de esperar, pois se à medida que os meses de incumprimento passam e o cliente não realiza nenhum pagamento, a probabilidade de ele vir a pagar é bastante reduzido;

- A possibilidade se tornar cumpridor diminui 62,2% com o número de dívidas, pois com o acumular de dívidas, a capacidade do devedor de cumprir com os seus compromissos diminui;

- A possibilidade de ser cumpridor aumenta aproximadamente 7 vezes com o número de acordos em ruptura e aumenta 13 vezes com o número de chamadas com sucesso. O gestor realiza várias tentativas para negociar com o devedor, e consegue um acordo de pagamento, no entanto este não cumpre com o acordado e o gestor volta a entrar em contacto com o devedor. Apesar das sucessivas rupturas, pode significar que existe intenção de pagar, e apesar das dificuldades do devedor, este consegue realizar pagamentos.

4.6.2. Cliente B

Para este estudo foram considerados 13188 casos, não existindo missing cases, nem casos não seleccionados (vide Tabela 59)

Resumo dos casos seleccionados

Unweighted Cases(a)		N	Percentagem
Casos seleccionados	Incluidos na análise	13188	100
	Casos em falta	0	0
	Total	13188	100
Casos não seleccionados		0	0
Total		13188	100

(a) If weight is in effect, see classification table for the total number of cases.

Tabela 59 - Resumo dos casos seleccionados, em falta e não seleccionados

Nas Tabela 60 pode observar-se a primeira etapa desta análise, onde se considera o modelo nulo, ou seja, o modelo só com a constante, a classificação dos indivíduos dentro dos grupos, cumpridor e não cumpridor.

Tabela de classificação (a,b)

Observados			Previstos		Percentagem correcta
			Incumprimento		
			Não cumpridor	Cumpridor	
Step 0	Incumprimento	Não cumpridor	8118	0	100
		Cumpridor	5070	0	0
	Overall Percentage				61,6

a Constant is included in the model.

b The cut value is ,500

Tabela 60 – Tabela de classificações

No modelo nulo a percentagem de individuos bem classificados é de 61,6%, sendo que no grupo 1 todos os individuos estão bem classificados e no grupo 2 estão todos mal classificados.

Variáveis na equação

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	-0,471	0,018	691,585	1	0,000	0,625

Tabela 61 – Variáveis consideradas na equação

Na Tabela 61 tem-se o valor da constante na equação e o teste de Wald, onde se observa que a constante influencia o modelo de forma significativa.

Na Tabela 62, observam-se as variáveis que não são consideradas no modelo nulo e o nível de significância de cada uma modelo nulo.

Variáveis não consideradas na equação (a)

			Score	df	Sig.
Step 0	Variáveis	divida	149,596	1	0,000
		n_dividas	123,453	1	0,000
		Acordos_ruptura	723,929	1	0,000
		Cham_sucesso	133,487	1	0,000
		sexo	71,171	1	0,000
		dias_incum	415,299	1	0,000
	Overall Statistics		1317,951	6	0,000

Tabela 62 – Variáveis não consideradas na equação

Na próxima etapa pode-se analisar o modelo completo, na Tabela 63 observa-se o teste do rácio de verosimilhança entre o modelo nulo e os modelos em cada etapa (step), bloco (block) e modelo final (model).

No caso em estudo foi utilizado o método Enter, e como só existe uma etapa, os valores para o teste são todos iguais, sendo o valor da estatística de teste $G^2(7)=1433,396$, com *p-value*<0,01, a hipótese nula de ausência de variáveis significativas é rejeitada, pelo que se conclui que existe pelo menos uma variável independente no modelo com poder predictivo sobre a variável dependente.

Block 1: Method = Enter

Teste de rácio verosimilhança

		Chi-square	df	Sig.
Step 1	Step	1433,396	6	0
	Block	1433,396	6	0
	Model	1433,396	6	0

Tabela 63 - Teste de rácio verosimilhança

Na Tabela 64 pode-se observar o resultado do teste de Hosmer e Lemeshow, que tem distribuição aproximadamente ao qui-quadrado e que permite testar o ajustamento do modelo,

comparando as frequências observadas com as esperadas na regressão logística. Os resultados apontam para um razoável ajustamento do modelo dos dados.

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	11,472	8	0,176

Tabela 64 – Teste de Hosmer and Lemeshow

Concretizando, o valor do teste da bondade do ajustamento de Hosmer e Lemeshow para o incumprimento é de 11,472 com sig.=0,176, o que mostra alguma segurança num ajustamento razoável entre as frequências observadas e as esperadas. Com este teste, sobre o ajustamento, apenas se pode concluir que as variáveis independentes influenciam Y, sem medir a força dessa associação.

Na regressão logística as medidas da força de associação são aproximações ao coeficiente de determinação R^2, o qual procura explicar a percentagem devida às variáveis independentes.

Modelo Sumário

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	16138,193 (a)	0,130	0,140

(a) Estimation terminated at iteration number 8 because parameter estimates changed by less than ,001.

Tabela 65 - Teste de associação da variável dependente e das variáveis independentes

Por observação da Tabela 65 pode-se concluir novamente que o grau da associação das variáveis independentes com a variável dependente é significativo.

Na Tabela 66 é apresentada a classificação dos indivíduos observada e prevista pelo modelo ajustado. Da leitura desta tabela, pode-se concluir que dos 7449 indivíduos classificados como não cumpridores, 6849 estão bem classificados (91,8%) e que dos 5070 classificados como cumpridores, apenas 1284 estão bem classificados (25,3%). A classificação global do modelo ajustado é de 66,2%.

Tabela de classificação(a)

Observados			Previstos		Percentagem correcta
			Incumprimento		
			Não cumpridor	Cumpridor	
Step 1	Incumprimento	Não cumpridor	7449	669	91,8
		Cumpridor	3786	1284	25,3
		Overall Percentage			66,2

(a) The cut value is ,500

Tabela 66 – Tabela de classificações

Para avaliar a qualidade da classificação feita pelo modelo, tal como se procedeu no cliente A, pode-se comparar a percentagem global de classificações correctas obtidas com o modelo, com a percentagem proporcional de classificações correctas por acaso que é $[(8118/13188)^2+(5070/13188)^2]\times 100\%=53\%$.

A percentagem de casos classificados correctamente pelo modelo é superior à percentagem de classificação proporcional por acaso, no entanto apenas é superior em 13,2%, não se podendo concluir que o modelo tem boas propriedades classificativas. Tal como aconteceu no cliente A, e se se observar a percentagem dos cumpridores bem classificados, esta é de apenas 25,3%.

Das mesmas forma que se procedeu para o cliente A, recorreu-se à construção da curva de ROC.

A tabela 67 dá a área sob a curva ROC (c=0,781), que é superior a 0,5 (p<0,001).

Area Under the Curve

Test Result Variable(s): Predicted probability

Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
,781	,004	,000	,773	,789

The test result variable(s): Predicted probability has at least one tie between the positive actual state group and the negative actual state group. Statistics may be biased.

a. Under the nonparametric assumption

b. Null hypothesis: true area = 0.5

Tabela 67 - Área abaixo da curva ROC

O modelo ajustado apresenta uma capacidade de discriminação aceitável (Hosmer & Lemeshow, 2000).

Também para o clinte B, na tabela 66 se pôde verificar uma grande diferença entre a sensibilidade e especificidade do modelo, e com vista a diminuir essa diferença, foram analisados as várias coordenadas da curva ROC. Pôde verificar-se que com um ponto de corte de 0,385, os valores da sensibilidade e especificidade eram mais próximos, foi então testado novamente o modelo, tendo-se verificado que para estes valores, 63,3% dos incumpridores se encontram bem classificados e 61,9% dos cumpridores também se encontram bem classificados. Posteriormente a esta análise, foi feita uma análise dos resíduos, não se verificando a existência de observações influentes ou outliers.

A Tabela 68 resume toda a informação sobre as variáveis independentes no modelo completo. Note-se que relativamente à variável categórica utilizada, apenas são considerados os indivíduos do sexo feminino.

Pode também observar-se o teste de Wald, que permite concluir que todas as variáveis influenciam de forma significativa o modelo.

Variáveis consideradas na equação

		B	S.E.	Wald	df	Sig.	Exp(B)	95,0% C.I.for EXP(B)	
								Lower	Upper
Step 1(a)	sexo (1)	-0,276	0,038	53,319	1	0,000	0,759	0,705	0,817
	divida	0,000	0,000	80,556	1	0,000	1,000	1,000	1,000
	dias_incum	-0,001	0,000	260,996	1	0,000	0,999	0,999	0,999
	Acordos_ruptura	0,993	0,045	491,169	1	0,000	2,473	2,473	2,948
	n_dividas	-0,427	0,048	79,794	1	0,000	0,594	0,594	0,716
	Cham_sucesso	0,082	0,021	15,773	1	0,000	1,042	1,042	1,130
	Constant	0,744	0,073	104,572	1	0,000			

(a) Variable(s) entered on step 1: sexo, divida, dias_incum, acordos_ruptura, n_dividas, cham_sucesso

Tabela 68 – Variáveis consideradas na equação

De acordo com o output obtido, o modelo pode-se escrever da seguinte forma:

$$\text{Logit}(\hat{\pi}) = 0,744 - 0,276 \times \text{sexo(F)} - 0,001 \times \text{dias\_incumprimento} - 0,427 \times \text{n\_dividas} + 0,993 \times \text{acordos\_ruptura} + 0,082 \times \text{cham\_sucesso}$$

A coluna Exp(B) é a exponencial dos coeficientes do modelo que estima o rácio das hipóteses da variável dependente por unidade da variável independente.

Assim, pode concluir-se que:

-A possibilidade de ser cumpridor diminui 27,6% com o sexo feminino, situação que pode estar relacionada com o facto de os valores em dívidas associados ao sexo feminino

serem mais reduzidos, o que irá contribuir para uma maior facilidade de cumprimento dos pagamentos;

- Por cada 30 dias a possibilidade de se tornar cumpridor diminui 97%, tal como se verificou no cliente A;

- A possibilidade se tornar cumpridor diminui 42,7% com o número de dívidas. Com número de dívidas acumuladas, torna-se mais complicado para os devedores cumprirem com os seus compromissos, deixando de pagar os valores que se encontram em dívida;

- A possibilidade de ser cumpridor aumenta aproximadamente 2,5 vezes com o número de acordos em ruptura e aumenta 1 vez com o número de chamadas com sucesso.

4.7. Síntese do Capitulo

No presente Capítulo foram analisados os perfis dos devedores de dois clientes da empresa XPTO, realizando-se análise descritiva dos dados e recorrendo-se a técnicas estatísticas como a análise de cluster, análise discriminante, análise de componentes principais e regressão logística.

Da análise descritiva realizada para o cliente A, verificou-se que a média das idades é de 44 anos, mais de 50% dos indivíduos são do sexo masculino e aproximadamente 50% dos indivíduos residem na região de Lisboa e Vale do Tejo. Da análise de associações entre variáveis, verificou-se que existe uma associação entre o sexo e as tranches de montante, sendo que o sexo masculino se encontra associado a tranches de montante mais elevadas. Também se verificou que o sexo masculino se encontra associado a dívidas de antiguidade superior.

Para o cliente B, a média das idades é de 41 anos, e tal como se verificou na amostra A, mais de 50% dos indivíduos são do sexo masculino e aproximadamente 50% dos indivíduos residem na região de Lisboa e Vale do Tejo. Verificou-se também que a região se encontra associada ao número de dívidas, a região com mais dívidas é Lisboa e Vale do Tejo, seguida da região Norte, o que se pode justificar pelo acesso à informação que existente nestas regiões. Também no cliente B se verificou que o sexo se encontra associado ao número de dívidas, sendo que o sexo feminino se encontra associado a número de dívidas mais elevado.

Na análise de clusters para o cliente A, com uma amostra de 2192 devedores foram obtidos 3 clusters:

- *1º cluster: Devedores jovens, cautelosos e mais avessos ao risco*

Grupo de devedores jovens, conhecedores dos produtos de crédito, pois recorrem a crédito de revolving para montantes reduzidos. São indivíduos com rendimentos reduzidos sendo por isso cautelosos, contraindo dívidas com montantes mais baixos e também com um número de dívidas baixo.

- *2º cluster: Devedores atrevidos e conhecedores das ofertas do crédito*

Grupo de devedores atrevidos, pois contraem várias dívidas, com montantes mais elevados, recorrendo a diferentes tipos de créditos, sendo maus pagadores das suas dívidas, apesar de serem aqueles com rendimentos médios mais elevados.

- *3º cluster: Devedores mais velhos e mais propensos ao risco*

Grupo de devedores mais velhos, menos racionais, pois contraem dívidas com montantes mais elevados recorrendo a produtos de revolving, possuindo várias dívidas com as mesmas características, sendo possivelmente devedores pouco conhecedores das ofertas de produtos de crédito.

Destes resultados pode concluir-se que os devedores mais novos são mais cautelosos quando recorrem ao crédito, pois os valores das suas dívidas são mais baixos e recorrem a produtos de revolving, o que pode estar relacionado com o facto de estarem no inicio de vida, os rendimentos serem mais baixos o que não lhes permite fazer grandes investimentos. Por outro lado, os indivíduos mais velhos, são menos cautelosos, recorrem com mais frequência ao crédito, com valores mais altos e o risco de incumprimento é superior. Esta situação, justifica-se pelo facto de serem pessoas que em determinada altura tinham alguma estabilidade, fizeram vários investimentos, recorrendo a diversos créditos, como o crédito habitação, crédito automóvel, entre outros créditos, mas por alguma razão, como por exemplo, desemprego, doença, divórcio, ou outra situação, em determinada altura ficaram impossibilitados de cumprir com os seus compromissos, sendo um grupo de grande risco.

Ainda foi encontrado outro tipo de devedores, que são indivíduos mais atrevidos na altura de recorrer ao crédito, com rendimentos mais elevados, mostrando conhecimentos dos produtos existentes no mercado, recorrem a vários créditos os quais apresentam um incumprimento bastante elevado, tal como acontece com os montantes das dívidas.

Em seguida, aplicou-se a técnica estatística análise discriminante a cada um dos clientes, que permitiu identificar as variáveis que contribuíam de forma significativa para caracterizar o perfil dos devedores.

Considerando os resultados, concluiu-se que as variáveis que melhor representam o perfil dos devedores do cliente A são:

$$Z_1 = -30,919 + 27,494 \times \text{produto} - 0,068 \times \text{tranches} + 26,501 \times \text{outras\_dividas}$$
$$Z_2 = -6,883 + 0,789 \times \text{produto} + 1,306 \times \text{tranches} - 0,496 \times \text{outras\_dividas}$$

Onde:
Produto - tipo de produto a que o devedor recorreu para adquirir um crédito;
Tranches – Intervalos definidos para os valores das dívidas;
Outras_dívidas – variável que permite identificar outras dívidas dos devedores, além da dívida com o cliente A.

A função 1 está associada essencialmente a outras dívidas e produtos, pode dizer-se que os devedores com mais dívidas em outros clientes e que recorrem a diversos produtos, tendem a ser mais atrevidos e mais conhecedores das ofertas de crédito. A função 2, tende a separar os devedores do grupo 1, dos devedores do grupo 3. Dadas as correlações positivas de variáveis como os rendimentos, valor da dívida, tranche de montante e número de dívidas com a função 2, pode concluir-se que para valores mais baixos do rendimento, dos valores em dívida e do número de dívidas, os devedores tendem a ser mais cautelosos, mais avessos ao risco e contraem menos dívidas. Inversamente, para valores do rendimento mais altos, valores de dívida mais elevados e um maior número de dívidas, os devedores tendem a ser mais propensos ao risco, contraindo mais dívidas e com valores mais elevados.

As outras variáveis analisadas foram descartadas na análise discriminante por não discriminarem tão bem quanto as escolhidas. Dessa forma, demonstrou-se que o perfil dos devedores não se explica por um único indicador, é necessário um conjunto de indicadores reunidos estatisticamente, e este é variável de cliente para cliente.

Nesta análise os pressupostos de normalidade e homogeneidade não se verificaram, no entanto, de acordo com alguns autores (Hair *et al. (*1998), Maroco (2010)) a análise discriminante é robusta face a violações e como o tamanho das amostras é bastante grande, a

análise poderá ser válida. Apesar de tudo, há que analisar cautelosamente os resultados obtidos.

Para este estudo também foi sugerida a aplicação da análise de componente principais, no entanto após a análise dos pressupostos e das estatísticas que permitiam testar a validade desta análise, concluiu-se que estes não se verificavam e que o resultado seria muito mau, pelo que esta técnica não foi desenvolvida neste trabalho.

Por último, recorreu-se à regressão logística, técnica estatística cujos objectivos são idênticos à análise discriminante, no entanto de aplicação mais extensa. Apesar de serem técnicas semelhantes, as variáveis escolhidas para a regressão logística não foram as mesmas da análise discriminante, opção que se deveu ao facto de os pressupostos não terem sido verificados na análise discriminante. Assim, foram escolhidas novas variáveis que permitiram complementar o estudo realizado pela análise discriminante, pois foram analisadas variáveis obtidas durante o período de gestão. A regressão logística foi então utilizada para obter um modelo que relaciona a variável resposta categórica, que no caso deste estudo, é o cumprimento do pagamento dos valores em dívida, com as variáveis explicativas que influenciam a ocorrência de determinado fenómeno, ou seja, permitiu analisar o risco de cumprimento dos devedores segundo determinados factores.

Para o cliente A, obteve-se o seguinte modelo:

$$\text{Logit}(\hat{\pi}) = -2,210 - 0,006 \times \text{Antiguidade\_divida} - 0,622 \times \text{Nº dividas} + 1,984 \times \text{acordos\_ruptura} + 2,593 \times \text{cham\_sucesso}$$

Onde:

Antiguidade_dívida – variável que representa o número de dias desde o incumprimento, pagamento em falta;

Nºdívidas – número de dívidas que o devedor contraiu;

Acordos_ruptura – número de acordos realizados entre a empresa XPTO e o devedor que entraram em incumprimento;

Cham_sucesso – número de chamadas realizadas com sucesso para o devedor.

Desta análise pode concluir-se que as variáveis que contribuem de forma significativa para definir o risco de incumprimento são a antiguidade da dívida, número de dívidas, número de acordos em ruptura e chamadas telefónicas realizadas com sucesso. Verificando-se que:

- Por cada 30 dias a possibilidade de se tornar cumpridor diminui 82%, o que seria de esperar, pois se à medida que os meses de incumprimento passam e o cliente não realiza nenhum pagamento, a probabilidade de ele vir a pagar é bastante reduzido;

- A possibilidade se tornar cumpridor diminui 62,2% com o número de dívidas, pois com o acumular de dívidas, a capacidade do devedor de cumprir com os seus compromissos diminui;

- A possibilidade de ser cumpridor aumenta aproximadamente 7 vezes com o número de acordos em ruptura e aumenta 13 vezes com o número de chamadas com sucesso. O gestor realiza várias tentativas para negociar com o devedor, e consegue um acordo de pagamento, no entanto este não cumpre com o acordado e o gestor volta a entrar em contacto com o devedor. Apesar das sucessivas rupturas, pode significar que existe intenção de pagar, e apesar das dificuldades do devedor, este consegue realizar pagamentos.

Na análise de clusters para o cliente B, com uma amostra de 13188 devedores foram obtidos 2 clusters:

1º cluster: Devedores mais velhos, menos cautelosos e mais propensos ao risco.

Grupo que se caracteriza por indivíduos mais velhos, com dívidas de montantes superiores, com número de dias de incumprimento bastantes elevados, bem como o número de dias desde o último pagamento. São também indivíduos pouco cautelosos pois tem várias dívidas.

2º cluster: Devedores jovens e mais cautelosos

Grupo que se caracteriza por devedores mais jovens, mais cautelosos, pois o montante da dívida e o número de dias desde o último pagamento são mais baixos, e apenas têm em média uma dívida.

Neste cliente, diferenciam-se os devedores mais novos e mais cautelosos, dos devedores mais velhos, menos cautelosos e mais propensos ao risco, situação que se deve ao referido na análise do cliente A. São indivíduos em fases diferentes da vida, com estabilidades diferentes, que lhes permitem fazer mais ou menos investimentos, aumentando ou diminuindo, respectivamente, o risco de incumprimento.

Após a análise de clusters, foi também aplicada a análise discriminante ao cliente B e considerando os resultados obtidos, concluiu-se que as variáveis que melhor representam o perfil dos devedores do cliente B são:

$$Z_1 = -0,551 - 0,002 \times \text{idade} + 0,002 \times \text{ultimo\_valor} - 0,127 \times \text{n\_dividas}$$

Onde:
Idade – idade do devedor;
Ultimo_valor – valor do último pagamento realizado após o incumprimento;
N_dívidas – número de dívidas do devedor no cliente B

Pode-se concluir que as variáveis que mais contribuem para a função discriminante são: "ultimo_valor", "dívida" e "n_dívidas". São estas variáveis, que dão uma contribuição mais significativa para a classificação dos devedores como mais cautelosos e menos cautelosos. Com efeito, a equação obtida aponta para uma forte importância da variável que representa o número de dívidas dos indivíduos. Não será portanto de estranhar que o número de dívidas que um indivíduo ou agregado familiar possua ajude a discriminar entre a maior ou menor probabilidade para eventual sobreendívidamento.

Também para o cliente B, foram analisadas outras variáveis, que posteriormente foram descartadas na análise discriminante por não discriminarem tão bem quanto as escolhidas. Dessa forma, demonstrou-se que também o perfil dos devedores do cliente B não se explica por um único indicador, é necessário um conjunto de indicadores reunidos estatisticamente, e este é variável de cliente para cliente.

Nesta análise os pressupostos não se verificaram, no entanto, de acordo com alguns autores (Hair *et al. (*1998), Maroco (2010)), a análise discriminante é robusta face a violações e como o tamanho das amostras é bastante grande, a análise poderá ser válida. No entanto deverão ser analisados os resultados de forma cautelosa.

Por último, recorreu-se à regressão logística com o objectivo de criar um modelo que relacionasse a variável cumprimento com outras variáveis obtidas durante o período de getsão do cliente B, obtendo-se o seguinte resultado:

$$\text{Logit}(\hat{\pi}) = 0,744 - 0,276 \times \text{sexo(F)} - 0,001 \times \text{dias\_incumprimento} - 0,427 \times \text{n\_dividas}$$
$$+0,993 \times \text{acordos\_ruptura} + 0,082 \times \text{cham\_sucesso}$$

Onde:

Sexo – sexo do inquirido;

Dias_incumprimento – variável que representa o número de dias desde o incumprimento, pagamento em falta;

N_dívidas – número de dívidas que o devedor contraiu;

Acordos_ruptura – número de acordos realizados entre a empresa XPTO e o devedor que entraram em incumprimento;

Cham_sucesso – número de chamadas realizadas com sucesso para o devedor.

Desta análise pode concluir-se que as variáveis que contribuem de forma significativa para definir o risco de incumprimento são o sexo, dias de incumprimento, número de dívidas, número de acordos em ruptura e chamadas telefónicas realizadas com sucesso. Verificando-se que:

A possibilidade de ser cumpridor diminui 27,6% com o sexo feminino, situação que pode estar relacionada com o facto de os valores em dívidas associados ao sexo feminino serem mais reduzidos, o que irá contribuir para uma maior facilidade de cumprimento dos pagamentos;

- Por cada 30 dias a possibilidade de se tornar cumpridor diminui 97%, tal como se verificou no cliente A;

- A possibilidade se tornar cumpridor diminui 42,7% com o número de dívidas. Com número de dívidas acumuladas, torna-se mais complicado para os devedores cumprirem com os seus compromissos, deixando de pagar os valores que se encontram em dívida;

- A possibilidade de ser cumpridor aumenta aproximadamente 2,5 vezes com o número de acordos em ruptura e aumenta 1 vez com o número de chamadas com sucesso.

Da análise dos devedores dos dois clientes, apesar de serem utilizadas variáveis diferentes, constatam-se algumas semelhanças. Os devedores mais novos são mais cautelosos e mais avessos ao risco, enquanto os devedores mais velhos são menos cautelosos e mais propensos ao risco. Da análise discriminante concluiu-se que as variáveis que são estatisticamente significativas para descriminar os devedores do cliente A são diferentes do cliente B. Enquanto no cliente A as variáveis que mais contribuem para a discriminação dos indivíduos são o "produto", "tranches"e "outras_dívidas", para o cliente B são a "idade", "ultimo_valor" e "n_dívidas". No entanto em ambos os clientes verifica-se que os devedores se tornam menos racionais quando possuem mais dívidas, pois no caso do cliente A os

devedores tendem a ser mais atrevidos tornando-se maus pagadores, enquanto no cliente B, concluiu-se que existem uma tendência para ficarem menos cautelosos. Apesar de comportamentos aparentemente diferentes, pode ver-se um factor comum, que é a falta de controlo sobre a situação de aumento de dívidas.

Da regressão logística, para o cliente A obteve-se que as variáveis que melhor descrevem o fenómeno do cumprimento são "antiguidade_dívida", "nºdívidas", "acordos_ruptura" e "cham_sucesso" enquanto para o cliente B foram as variáveis "sexo(F)", "dias_incumprimento", "n_dívidas", "acordos_ruptura" e "cham_sucesso". Nesta última análise, pode constatar-se que a antiguidade da dívida, número de dívidas, os acordos em ruptura e as chamadas com sucesso são as variáveis que mais contribuem para a ocorrência do cumprimento dos compromissos dos devedores, nos dois clientes. Em ambos os casos, verifica-se que a probabilidade de cumprimento por parte do devedor diminui, com o aumento da antiguidade da dívida, situação bastante previsível, pois se o devedor até determinada data não pagou, demonstra que não tem capacidade financeira para o fazer, pois possivelmente a situação que originou este incumprimento deixa-o impossibilitado de cumprir com os seus compromissos por um longo período, como por exemplo, uma situação de desemprego. O mesmo sucede com o número de dívidas, o devedor fica incapacitado de pagar qualquer dívida, diminuindo assim a probabilidade de ocorrer um pagamento. Também se verificou em ambos os casos, que o facto de um devedor realizar vários acordos com a empresa de recuperação de crédito, apesar do número de rupturas dos acordos que ocorrem, demonstra que há intenção do devedor realizar um pagamento e que este acontece durante o período de gestão. Esta última situação, pode justificar-se pelo facto de o devedor terem incorrido num incumprimento por uma situação temporária e que após a sua resolução o devedor consegue cumprir com os seus compromissos.

5. Conclusão

O crédito aos consumidores constituiu, em Portugal, nos últimos anos, a forma de muitas famílias poderem melhorar a sua qualidade de vida e de terem acesso a determinados bens e serviços. Contudo, o sobreendívidamento é a outra face da democratização do crédito, pois ao alargar o acesso ao crédito, está a potenciar-se tal fenómeno. Para tal situação muito têm contribuído as instituições financeiras com bastante publicidade, com o objectivo de angariar clientes, dando-lhes assim a possibilidade de acesso ao crédito com mais frequência. No entanto este constante recurso ao crédito ganhou contornos mais graves quando as famílias se começaram a ver impossibilitadas de fazer face aos encargos dos créditos contraídos, gerando situações de endívidamento e até mesmo de sobreendívidamento. Em muitos casos, esta situação agravou-se ainda mais, quando as pessoas acumularam mais de um crédito, dando origem ao multiendívidamento (Frade *et al*., 2003).

Assim, credores, devedores e a sociedade em geral, têm a ganhar com a prevenção e com o tratamento do sobreendívidamento, seja do ponto de vista económico, seja de um ponto de vista social.

Neste estudo pretendeu-se analisar e definir o perfil dos devedores de dois clientes da empresa de recuperação de crédito XPTO. Nesta análise consideraram-se variáveis como a idade, sexo, localidade, profissão, valor das dívidas, número de dias de incumprimento, número de dívidas, número de chamadas telefónicas efectuadas com sucesso e número de acordos que entraram em ruptura.

Dados os objectivos fixados, foram seleccionadas várias ferramentas estatísticas, entre as quais, a análise de clusters, análise discriminante, regressão logística e análise de componentes principais. Esta selecção teve como base a literatura consultada, nomeadamente o estudo de Lourosa (2009), cujo objectivo era segmentar uma base de dados de clientes que recorreram ao crédito, recorrendo à análise de clusters. Outro estudo analisado foi o de Frade *et al*. (2008), onde se pretendia estudar o perfil dos indivíduos sobreendívidados em Portugal, também recorrendo à análise de clusters. No artigo de Minussi *et al*. (2002), é explorada a regressão logistica com o objectivo de construir um modelo de previsão de solvência.

De um modo geral, os resultados revelaram, que os devedores têm em média 42 anos e são na sua maioria do sexo masculino, resultado que em certa medida é o esperado, uma vez que, por motivos de natureza cultural, a situação mais usual é aquela em que o chefe de família lidera todo o processo de negociação no recurso ao crédito. Relativamente à

localização, aproximadamente 50% dos devedores residem na região de Lisboa e Vale do Tejo e 20% na região Norte. Quanto aos valores em dívida, verificou-se que os valores mais elevados e as dívidas com maior antiguidade, são referentes a indivíduos do sexo masculino, no entanto foi nos indivíduos do sexo feminino que se registou um maior número de dívidas por devedor.

Com base num conjunto de variáveis previamente seleccionadas, e após a aplicação da análise de clusters, este estudo revelou também que de uma forma geral existem grupos de devedores com características bastante próprias. Nos dois clientes analisados verificou-se a existência de grupos de devedores bastante distintos, os mais jovens, que são mais cautelosos e mais avessos ao risco e por outro lado, os mais velhos que são menos cautelosos e mais propensos ao risco. Num dos clientes ainda foi possível distinguir um outro grupo, onde se enquadram os devedores mais atrevidos e conhecedores das ofertas de crédito.

Desta análise pode concluir-se que o facto de os devedores mais novos serem mais cautelosos quando recorrem ao crédito, os valores das suas dívidas mais baixos e os produtos mais recorrentes serem de *revolving,* pode estar relacionado com o facto de estarem no início de vida, os rendimentos serem mais baixos o que não lhes permite fazer grandes investimentos. Por outro lado, os indivíduos mais velhos, são menos cautelosos, recorrem com mais frequência ao crédito, com valores mais altos e o risco de incumprimento é superior. Esta situação pode justificar-se pelo facto de serem indivíduos que em determinada altura tinham alguma estabilidade, fizeram vários investimentos, recorrendo a diversos créditos, como o crédito habitação, crédito automóvel, entre outros créditos, mas por alguma razão, como por exemplo, desemprego, doença, divórcio, ou outra situação, ficaram impossibilitados de cumprir com os seus compromissos, tornando-se um grupo de grande risco. No cliente A, ainda foi possível distinguir um terceiro grupo, onde foram encontrados outro tipo de devedores, que se pode caracterizar como mais atrevidos na altura de recorrer ao crédito, com rendimentos mais elevados, mostrando conhecimentos dos produtos existentes no mercado, recorrem a vários créditos os quais apresentam um incumprimento bastante elevado, tal como acontece com os montantes das dívidas.

Da aplicação da análise discriminante, obteve-se que as variáveis seleccionadas discriminam significativamente os grupos formados na análise de clusters e que em ambos os clientes as funções discriminantes obtidas explicam mais de 95% da variância dos grupos formados.

Nas duas funções obtidas para o grupo de devedores do cliente A, verificou-se que uma das funções estava associada essencialmente a outras dívidas e produtos, podendo concluir-se

que os devedores com mais dívidas em outros clientes e que recorrem a diversos produtos, tendem a ser mais atrevidos e mais conhecedores das ofertas de crédito. A outra função, tende a separar os devedores do grupo 1, dos devedores do grupo 3. Dadas as correlações positivas de variáveis como o rendimento, valor da dívida, tranche de montante e número de dívidas com essa função, pode concluir-se que para valores mais baixos do rendimento, dos valores em dívida e do número de dívidas, os devedores tendem a ser mais cautelosos, mais avessos ao risco e contraem menos dívidas. Inversamente, para valores do rendimento mais altos, valores de dívida mais elevados e um maior número de dívidas, os devedores tendem a ser mais propensos ao risco, contraindo mais dívidas e com valores mais elevados.

Para o cliente B pode-se concluir que as variáveis que mais contribuem para a função discriminante são: "ultimo_valor", "dívida" e "n_dívidas". São estas variáveis, que dão uma contribuição mais significativa para a classificação dos devedores como mais cautelosos e menos cautelosos. Com efeito, a equação obtida aponta para uma forte importância da variável que representa o número de dívidas dos indivíduos. Não será portanto de estranhar que o número de dívidas que um indivíduo ou agregado familiar possua ajude a discriminar entre a maior ou menor probabilidade para eventual sobreendívidamento.

Também neste estudo foi utilizada a regressão logística de forma a obter um modelo que permitisse caracterizar o perfil dos devedores dos dois clientes. Para o cliente A foram escolhidas as variáveis "antiguidade_dívida", "nºdívidas", "acordos_ruptura" e "cham_sucesso" e no cliente B foram as variáveis sexo, dias de incumprimento, número de dívidas, número de acordos em ruptura e chamadas telefónicas realizadas com sucesso para descrever o fenómeno do cumprimento.

Nesta análise verificou-se que a possibilidade de cumprimento por parte do devedor diminui, com o aumento da antiguidade da dívida, situação bastante previsível, pois se o devedor até determinada data não pagou, demonstra que não tem capacidade financeira para o fazer. Possivelmente a situação que originou este incumprimento deixa o devedor impossibilitado de cumprir com os seus compromissos por um longo período, como por exemplo, uma situação de desemprego. O mesmo sucede com o número de dívidas, o devedor fica incapacitado de pagar qualquer dívida, diminuindo assim a probabilidade de ocorrer um pagamento. Também se verificou que o facto de um devedor realizar vários acordos com a empresa de recuperação de crédito, apesar do número de rupturas dos acordos que ocorrem, demonstra que há intenção do devedor realizar um pagamento e que este acontece durante o período de gestão. Esta última situação, pode justificar-se pelo facto de o devedor ter

incorrido num incumprimento por uma situação temporária e que após a sua resolução o devedor consegue cumprir com os seus compromissos.

Também com o aumento do número de chamadas telefónicas efectuadas com sucesso a possibilidade de cumprimento aumenta, o que pode significar que o devedor demonstra intenção de realizar pagamentos, mas tem dificuldades, havendo necessidade de fazer várias negociações, mas acabam por acontecer pagamentos por parte do devedor.

No cliente B pôde ainda concluir-se que a possibilidade de ser cumpridor diminui 27,6% com o sexo feminino, situação que pode estar relacionada com o facto de os valores em dívidas associados ao sexo feminino serem mais reduzidos, o que irá contribuir para uma maior facilidade de cumprimento dos pagamentos.

Para o cliente A, concluiu-se que o modelo tem uma capacidade discriminante boa enquanto para o cliente B o modelo tem capacidade de discriminação aceitável.

Conhecendo as características dos devedores, é possível para empresa XPTO definir estratégias de gestão das carteiras dos diversos clientes, como por exemplo definir as campanhas de lançamento de chamadas telefónicas por intervalos do montante da dívida, por valores de rendimento, por número de dívidas do devedor, por antiguidade da dívida, sexo do devedor, etc. Desta forma, a empresa irá incidir sobre os devedores onde existe uma maior possibilidade de recuperar os valores em dívida, evitando em algumas situações, custos com grupos de devedores onde a probabilidade de recuperação é muito reduzida. Estes resultados, também poderão ser uma ajuda para os gestores definirem qual o melhor discurso quando contactam com determinado grupo de devedores.

Este estudo deparou-se com algumas dificuldades e limitações relacionadas com o tipo de dados, que não satisfizeram os pressupostos das várias técnicas adoptadas para esta análise, mas também com a informação que é disponibilizada pelos clientes e que permite a caracterização dos devedores. As variáveis disponibilizadas não permitiram fazer uma caracterização mais completa do perfil dos devedores, pois para tal, seria necessário ter acesso a mais informação sobre os devedores e de forma mais organizada e perceptível, como é o caso do tipo de produto, que surge de forma codificada, tendo sido impossível num dos clientes utilizar essa informação.

Outra limitação que se encontrou na aplicação do método de regressão logística, prendeu-se essencialmente com a composição da amostra, por não representar proporcionalmente os devedores cumpridores e incumpridores, dando origem a um enviesamento nos resultados.

Apesar de as conclusões não serem as desejadas, pode-se verificar, que algumas vão ao encontro dos resultados dos muitos estudos analisados na revisão bibliográfica, como é o caso

do estudo de Frade *et al.* (2008), onde foi analisado o perfil do sobreendívidados em Portugal, ou no relatório de Marques e Frade (2003) sobre o desemprego e o sobreendívidamento dos consumidores, o que fortalece as conclusões obtidas, mas aponta para a necessidade de complementar o trabalho de investigação realizado com mais informação sobre os devedores. Este trabalho sugere que algumas das variáveis escolhidas para definir o fenómeno do cumprimento sejam substituídas por outras variáveis, como por exemplo: Valor do empréstimo, taxa de juro, capital, prazo do empréstimo, número de dependentes, estado civil e nível académico.

Para que todas as técnicas estatísticas utilizadas neste estudo tenham uma melhor aplicabilidade é fundamental que em futuras investigações se obtenham bases de dados tão completas quanto possível e que retratem fidedignamente todas as características da população.

No entanto, apesar de tudo o que foi referido, um outro aspecto que não se pode deixar de salientar é a falta de estudos sobre o risco de incumprimento de indivíduos após uma situação de endívidamento. Os estudos existentes têm como objectivo conhecer o perfil dos indivíduos que recorrem ao crédito ou determinar o risco de endívidamento dos indivíduos que recorrem ao crédito.

A par disso os resultados obtidos sugerem para futuras investigações, estudos mais detalhados das características dos devedores e uma investigação na qual seja possível constatar o impacto de outras variáveis na determinação do risco de incumprimento. Uma proposta seria a elaboração de modelos de optimização da relação risco/retorno de uma carteira, aplicando instrumentos de optimização, como a programação linear, quadrática, entre outras, e utilizando variáveis, como o montante médio de pagamento, nº de chamadas com sucesso, taxa de transformação (nº de chamadas necessárias para obter um acordo). Com estes modelos pretende-se encontrar o valor óptimo para um conjunto de variáveis fundamentais, com vista à minimização do risco de crédito. Neste contexto, seria tido em conta o Acordo de Basileia e as respectivas directrizes.

6. Referências bibliográficas

Akaike, H. (1974), *A new look at the statistical model identification*. IEEE Transactions on Automatic Control., Boston, v.19, n.6, p.716-723.

Betti, G., Dourmashkin, N., Cristina Rossi, M., Verma, V., Yin, Y. (2001), *Study of the problem of Consumer Indebtedness: Statistical Aspects*, produced by OCR Macro for DG Health & Consumer Protection, European Commission.
http://ec.europa.eu/consumers/cons_int/fina_serv/cons_directive/fina_serv06_en.pdf (consulta em Fevereiro de 2010)

Branco, J. (2004), *Uma Introdução à Análise de Clusters,* Sociedade Portuguesa de estatística.

Coutinho, M.; Marquês, Maria; Soares, J. (2000), *Desigualdade regionais em Portugal: Uma análise estatística multivariada,* Trabalho da Area Departamental de Engenharia de Electrónica e Telecomunicações e de Computadores do ISEL.

Cox, D.R. & Snell, E. J. (1989), *The Analysis of Binary Data,* 2ª Edição, Edições Chapman and Hall.

Doca, F. (2009), *A Psicologia Pediátrica em Hospitais Universitários Brasileiros,* Tese de Mestrado da Universidade de Brasilia.

Emiliano, P. Veiga, E. Vivanco, M.; Menezes, F. (2002), *Critérios de informação de Akaike versus Bayesiano,* Trabalho do Instituto de Matemática, Estatística e Computação Cientifica de Campinas.

Frade, C.; Lopes, C.; Jesus, F.; Ferreira, T.; Marques, M. (2008), *Um Perfil dos Sobreendívidados em Portugal,* Relatório do Centro de Estudos Sociais.

Frade, C. (2003), *Desemprego e sobreendívidamento dos consumidores*, Relatório final do projecto desemprego e endívidamento das famílias (PIQS/ECO/50119/2003).

Friedman, J.H. (1989), *Regularized Discriminant Analysis.* Journal of the American Statistical Association.

Hair, A.; Black, T. (1998), *Multivariate Data Analysis*, 5ª Edição, Edições Bookman.

Hosmer, D. W. & Lemeshow, S. (2000), *Applied Logistic Regression*, 2º Edição, Edições John Wiley & Sons

Johansson, M., Persson, M., (2006), *Swedish households' indebtedness and ability to pay – a household level study*, Penning – Och Valutapolitik 3/2006, Sveriges Riksbank Economic Review nº 3 2006

Johnston, J.; Dinardo, J. (2003), *Método Econométrico,* 4ª Edição, Edições McGrawHill.

Lourosa, P. (2009), Segmentação de Mercado: Um Caso Específico do Crédito ao Consumo Afecto, Tese de Mestrado da Universidade do Porto

Malhotra, N. (2004), *Pesquisa de Marketing - Uma Orientação Aplicada*, 4ª Edição, Edições Bookman.

McLachlan, G. J. (2004), *Discriminant Analysis and Statistical Pattern Recognition*. Wiley Interscience.

Maroco, J. (2010), *Análise Estatística – Com Utilização do SPSS*, 3ª Edição, Edições Sílabo.

Martinez, L.; Ferreira, A. (2008), *Análise de Dados com SPSS*, 2ª Edição, Edições Escolar Editora.

Marques, M.; Neves, V.; Frade, C.; Lobo, F.; Pinto, P.; Cruz, C. (2000), *O Endívidamento dos Consumidores,* Edições Almedina.

Morais, I. (2008), *Segmentação dos Investidores Individuais Portugueses,* Tese de Mestrado da Universidade de Évora.

Minussi, J.; Damacena, C.; Ness, W. (2002), *Um Modelo de Previsão de Solvência Utilizando Regressão Logistica,* Artigo da Revista de Administração Contemporanêa do Brasil.

Nunes, M.(2009), *Perfil Profissional de Cirugiões-Dentistas e seu Desempenho Académico Durante a Graduação: Um Estudo com Egressos da Universidade Federal de Góias*, Pós-Graduação da Universidade Federal de Góias

Observatório do Endívidamento dos Portugueses, (2002), *Endívidamento e Sobreendívidamento das Famílias – Conceitos e Estatísticas para a sua Avaliação.*

Pereira, A. (2003), *Guia Prático de Utilização do SPSS,* Edições Sílabo.

Pestana, M.; Gageiro, J. (2009), *Análise Categórica, Árvores de Decisão e Análise Conteúdo – Em Ciências Sociais e da Saúde com o SPSS*, Edições Lidel.

Ramos, C. (2007), *Crédito ao Consumo em Portugal,* Trabalho da Universidade de Coimbra.

Reis, E. (2001), *Estatística Multivariada Aplicada*, 2ª Edição, Edições Sílabo.

Sarmento, A. (2005*), Experimentação e Avaliação de Modelos para um Problema de Atribuição de Crédito,* Tese de Mestrado da Universidade do Porto

Schwarz, G. (1978), *Estimating the dimensional of a model*. Annals of Statistics, Hayward, v.6, n.2, p.461-464.

Smrčka, L., (2011), *Government Indebtedness and Family Indebtedness as an Inseparable Twins in the Modern World*, publicado no "International journal of mathematical models and methods in applied sciences", vol. 5.

Wooldridge, J. (2006), *Introductory Econometrics – A Modern Approach*, 2ª Edição, International Student Edition.

7. Glossário

Neste glossário são apresentados alguns conceitos analisados ao longo da tese, que dada a sua especificidade no âmbito da empresa de recuperação de créditos, considerou-se pertinente a sua explanação.

Carteira – Grupo de devedores de um cliente da empresa XPTO, com características específicas, que se distinguem dos outros devedores do cliente, por exemplo, pelo produto que se encontra associado ao crédito, pela antiguidade da dívida, etc.

Por exemplo: Carteira 210 do cliente A - grupo de devedores do cliente A cujas dívidas têm antiguidade superior a 210 dias;

Crédito Clássico - Financiamentos de bens ou de serviços em que a aquisição é efectuada por um consumidor final e cujo crédito tem um plano de amortização rígido e pré-definido, nele se inclui o crédito concedido a particulares – crédito ao consumo – e o crédito concedido a empresas;

Crédito Revolving - Caracteriza-se pela existência de planos flexíveis de amortização da dívida, bem como pela existência de um "plafond" de crédito, que poderá estar, ou não, totalmente utilizado (atribuído antes da aquisição do bem ou serviço). Ex: cartões de crédito e abertura de crédito em conta corrente**;**

Endívidamento – Saldo devedor de um agregado familiar. Pode resultar apenas de uma dívida ou de mais do que uma em simultâneo, utilizando-se, neste último caso, a expressão multiendívidamento (Marques *et al*, 2000).

Incumprimento – Não pagamento atempado das prestações em dívida pelo devedor. Segundo as instituições bancárias, considera-se que há incumprimento definitivo quando se esgotam as possibilidades de renegociação e se inicia a acção judicial (Marques *et al*, 2000).

Lote – Grupo de devedores de um cliente da empresa XPTO, cujas dívidas deram entrada na empresa XPTO para sua gestão, num determinado periodo. Por exemplo: Carteira A, lote

de Março/2010 – grupo de devedores cujas dívidas tiveram o seu inicio de gestão na empresa XPTO em Março de 2010;

Sobreendívidamento - Situação em que o devedor se ache impossibilitado de cumprir com os seus compromissos financeiros, sem pôr em risco a subsistência do agregado familiar. São contempladas situações de insolvência das famílias com origem em causas diversas, que deverão ser claramente identificadas e onde a boa fé é factor crucial (OEC- Observatório do Endívidamento dos consumidores, 2002).