NOVA IMS Information Management School

MAA

Mestrado em Métodos Analíticos Avançados
Master Program in Advanced Analytics

Métodos Analíticos Avançados Aplicados ao Sector Bancário em Portugal

Relatório de Estágio

Joana Machado Nunes Sassetti

Relatório de Estágio apresentado como requisito parcial para obtenção do grau de Mestre em Métodos Analíticos Avançados

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

MÉTODOS ANALÍTICOS AVANÇADOS APLICADOS AO SECTOR BANCÁRIO EM PORTUGAL

por

Joana Machado Nunes Sassetti

Relatório de Estágio apresentado como requisito parcial para a obtenção do grau de Mestre em Métodos Analíticos Avançados

Orientador: Leonardo Vanneschi

Novembro 2018

AGRADECIMENTOS

"Faço colecção de pessoas maravilhosas."

À minha mãe e ao Pedro por serem um exemplo de trabalho e dedicação. Ao meu pai e à Suzi por me ensinarem que em todo o lado existem pessoas espectaculares. Aos quatro por me apoiarem incondicionalmente.

Aos meus avós, em particular à minha avó Lena por estar sempre disponível para me ouvir e fazer uns miminhos que me enchem sempre o estômago e o coração.

Aos meus irmãos Maria, Francisca, Nadinho, Manel, Gui e Nuno por serem verdadeiramente os meus melhores amigos.

Ao meu orientador Leonardo por todo o apoio e por conseguir encontrar sempre as palavras certas de motivação.

Ao Millennium, mas em particular a toda a equipa do CRM, que me aturou ao longo deste ano (+€), mais concretamente ao Tiago, por ter apostado e acreditado em mim, à Amélia, por tudo o que me ensinou e pelo exemplo de paciência e confiança, ao Miguel, ao Miguel, ao Fonfas e ao Ritchie, por todas as brincadeiras e cumplicidade, mas também nos momentos sérios por tudo o que me ensinaram e ainda ao Ziad por estar a fazer este caminho comigo desde o início do mestrado.

Ao meu namorado Tomaz por toda a ajuda e apoio, mas principalmente por me fazer sentir que "Home is wherever I'm with" him.

À Alexandra, à Clara e à Maria por embarcarem nas minhas loucuras e me ajudarem sempre a descomplicar todos os problemas.

Ao meu grande amigo Eduardo que me mostrou o que é o mundo dos dados, e partilhou comigo a sua paixão por esta área.

A todas as pessoas que se cruzaram comigo e que de uma maneira ou doutra me inspiraram, ensinaram, marcaram ou ajudaram.

RESUMO

Actualmente, com um mercado cada vez mais competitivo, os bancos têm de fazer melhor, diferente, inovar e marcar a diferença junto de quem interessa: os clientes. Neste sentido, a gestão da relação dos clientes desempenha um papel fundamental na criação de serviços e produtos com valor acrescentado para os clientes.

Neste relatório são apresentados três dos projectos desenvolvidos ao longo do estágio no departamento de Marketing, na equipa de CRM no Millennium BCP: desenvolvimento de um modelo de propensão à compra de uma solução integrada, através da combinação dos modelos resultantes da aplicação de redes neuronais e de regressões logísticas; criação da segmentação comportamental, onde foram definidos dois perfis diferentes mas complementares – o perfil de utilização de serviços bancários e o perfil de compras em POS; e análise da utilização de serviços incluídos numa solução integrada.

PALAVRAS-CHAVE

CRM; Banco; Modelo de Propensão à Compra; Segmentação Comportamental

ABSTRACT

Nowadays, with an increasingly competitive market, banks have to do better, different, innovate and make a difference with those who matters: customers. In this sense, customer relationship management plays a key role in creating products and services with added value for clients.

This report presents three of the projects developed during the internship in the Marketing department, in the CRM team at Millennium BCP: development of a propensity model for the purchase of a pack of services, through the combination of the models resulting from the application of neural networks and logistic regressions; creation of the behavioral segmentation, where two different but complementary profiles were defined - the use of banking services profile and the purchases in POS profile; and analysis of the use of services included in a pack of services.

KEYWORDS

CRM; Bank; Propensity Model; Behavioral Segmentation

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABELAS

LISTA DE SIGLAS E ABREVIATURAS

ABT	Analytical Base Table
ANOMES	Ano e mês considerados (exemplo: Novembro de 1995 = 199511)
ATM	Automated Teller Machine (caixa multibanco)
BCP	Banco Comercial Português
CF	Solução Integrada Cliente Frequente
CFN	Solução Integrada Cliente Frequente Negócios
Ciclo	O ano encontra-se dividido em 4 ciclos comerciais, onde 201802 representa o segundo ciclo de 2018
CRM	Customer Relantionship Management – Gestão da Relação com o Cliente
MCC	Merchant Category Code – código usado para diferenciar tipos de negócios e indústrias, onde cada um tem um código único
MS	Segmento de clientes Mass Market
OIC	Outra(s) Instituição(ões) de Crédito
PP	Segmento de clientes Prestige
POS	Point of Service – Máquina de cartões de débito ou crédito, ou outro terminal electrónico de venda
UU	Segmento de Cliente Plus
SI	Solução Integrada

1. INTRODUÇÃO

O sector financeiro tem sofrido muitas alterações nos últimos anos, seja pelas novas regulações, pelo aumento exponencial de informação, pela alteração do comportamento dos consumidores, ou pela competição cada vez mais intensa. [1] Os consumidores têm-se vindo a tornar mais exigentes, com melhor acesso a informação, consciencializados da existência de alternativas, e por isso esperam que as suas expectativas e necessidades sejam compreendidas e satisfeitas. [2]

Para os bancos conseguirem acompanhar estas mudanças e terem a capacidade de desenvolver relações duradouras com os seus clientes, têm investido na transformação da banca tradicional, deixando de serem vistos apenas como uma instituição financeira que recebe depósitos e concede créditos [3], para um ecossistema focado no cliente, de maneira a conseguirem entregar produtos ou serviços com mais valor, através da gestão de relacionamento com cada cliente.

A gestão da relação com clientes (Customer Relationship Management – CRM) representa o sistema que gere todos os aspectos de comunicação e interacção com os clientes. Os projectos deste âmbito têm como objectivo conhecer melhor os clientes, bem como perceber a quem melhor se destinam os produtos existentes no banco, de maneira a conseguir-se direccionar e personalizar mais as campanhas de marketing.

O objectivo deste relatório é expor as metodologias e abordagens utilizadas em três dos projectos desenvolvidos no decorrer do estágio no departamento Marketing de Retalho, na equipa de CRM, do maior banco privado português, o Millennium BCP, com dados provenientes de clientes reais.

O primeiro projecto desenvolvido foi a criação de um modelo de propensão à compra da solução integrada Cliente Frequente de Negócios. Uma solução integrada é um conjunto de produtos para um segmento específico, com uma mensalidade fixa associada, não dependendo da utilização dos serviços e produtos incluídos.

Este modelo tem como objectivo compreender e traçar o perfil de clientes que adquiriram esta solução integrada no passado e a mantiveram, através da aplicação de técnicas e algoritmos habitualmente utilizados na literatura científica, como redes neuronais, árvores de decisão ou regressões logísticas, para depois conseguir prever a aquisição deste produto por parte de clientes que potencialmente não irão cancelar a subscrição desta solução.

O Millennium actualmente segmenta dos clientes ,entre empresas e particulares de acordo com a **Figura 1**. Esta segmentação divide os clientes de acordo com o nível de envolvimento com o banco, como por exemplo o património financeiro ou o valor do ordenado ou pensão domiciliados, no caso de serem clientes particulares. O departamento de marketing de retalho gere a relação com os clientes particulares dos macro-segmentos Mass Market (MS), Plus (UU) e Prestige (PP) e de clientes empresas do macro-segmento de Negócios.

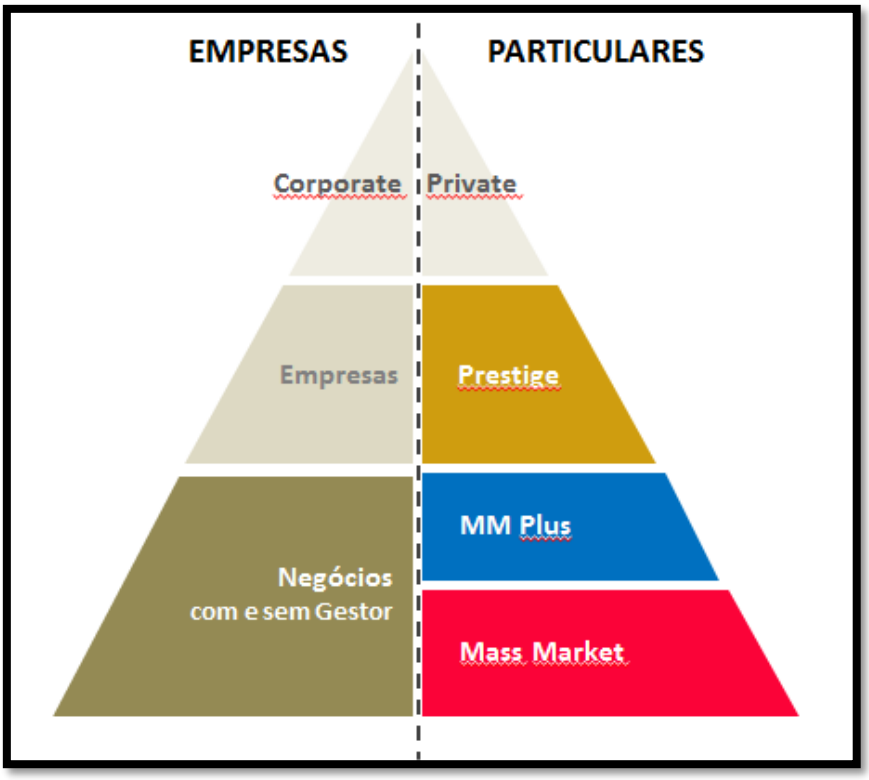


Figura 1- Segmentação dos clientes no Millennium BCP

Com o objectivo de contribuir para uma melhor compreensão dos clientes existentes no banco foi desenvolvido o projecto da Segmentação Comportamental, levando à criação de novos segmentos para clientes particulares, com outros factores em consideração para além das informações financeiras, que são os critérios actuais da segmentação de clientes.

Com este estudo espera-se conseguir perceber quem são os clientes do Millennium BCP dentro e fora do banco, através da análise de clusters com a aplicação do algoritmo K-Means, com base em técnicas que se considerem enriquecer o processo de extracção de conhecimento da informação existente.

O último projecto exposto neste relatório consiste numa análise de utilização dos serviços disponibilizados pela solução integrada (SI) Cliente Frequente. Esta solução integrada já existe no banco desde 2004, e apesar de ao longo do tempo se actualizarem as ofertas incluídas, muitas vezes existem serviços que são utilizados por uma minoria de clientes, ou por outro lado, outros que a maioria dos clientes aproveita.

Com o objectivo de perceber a utilização de todos os serviços incluídos, foi realizada esta análise, onde para cada serviço se comparam os valores que cada cliente pagaria caso não possuísse a solução. Para além disto, esta análise tem também como objectivo comunicar a cada cliente quanto poupou por deter a SI, e fomentar a utilização em pleno de todas as vantagens incluídas.

Este relatório está organizado da seguinte forma: na Secção 2 é dada a visão de investigadores e cientistas sobre o que é o CRM, o que é Data Mining e que tipos de tarefas adereçam. Na Secção 3 é dada uma contextualização ao leitor sobre a organização do CRM no Millennium BCP, e que tipo de informação existe relativa aos clientes. Na Secção 4 é exposta a metodologia utilizada nos projectos do modelo de propensão à compra e da segmentação comportamental. A Secção 5 contempla a apresentação de resultados da criação do modelo de propensão à compra, analisando a variação da taxa de conversão. Na Secção 6 são apresentados e analisados os segmentos obtidos através da análise de *clusters*, para a nova segmentação comportamental. Na Secção 7 é realizada a análise de utilização dos serviços da SI Cliente Frequente. A Secção 8 apresenta as conclusões e a Secção 9 as limitações e recomendações para futuros projectos.

2. REVISÃO DE LITERATURA

O sector bancário tem vindo a sofrer muitas alterações, sejam derivadas da globalização ou do aparecimento de fintechs, e cada vez mais os bancos para conseguirem manter ou aumentar a sua quota de mercado têm de ser mais proactivos e criativos quando lidam com clientes. Apesar destas instituições possuírem muita informação relativa aos seus clientes, esta é habitualmente utilizada para dar apoio à rede ou satisfazer requisitos de auditoria internos ou externos, mas na opinião de Rene Domingo, esta informação deveria ser principalmente utilizada para criar e gerar conhecimento sobre os clientes: o que são os clientes no banco, o que esperam, o que querem e como querem. [4]

A gestão da relação com os clientes ou CRM (Customer Relationship Management) tem-se tornado numa das estratégias de negócio mais utilizadas neste novo milénio e para Kim, Suh e Hwang pode ser visto como um esforço para gerir as interacções com os clientes, combinando regras de negócio e tecnologias que procuram perceber melhor os clientes de cada empresa. [5]

Um CRM eficaz consegue adquirir, analisar e partilhar conhecimento sobre e com os seus clientes, permitindo à empresa focar-se mais no cliente. [6] Consequentemente alguns dos potenciais benefícios que podem advir desta gestão de relacionamento com clientes são: o aumento da retenção e da lealdade de clientes, maior rentabilidade de clientes e criação de valor para o cliente através da personalização de produtos e serviços. [7]

> "We are drowning in information, but starved for knowledge" [1]
>
> John Naisbitt, 1982 [8]

A investigação e a tomada de decisão com base em dados passaram a ser mais do que uma tendência. A prospecção de dados ou data mining é o processo de extrair conhecimento interessante, não trivial, desconhecido e potencialmente útil de um grande conjunto de dados. [9]

"Data Mining is a knowledge discovery process"[2] - é o processo de analisar dados e sumarizá-los em informação útil. [10] Segundo Adeniyi, Wei e Yongquan, data mining pode também ser visto como a extracção de conhecimento de uma base de dados com o objectivo de descobrir relações e padrões escondidos nos dados, de maneira a torná-los compreensíveis para todos os utilizadores. [11]

[1] Em português: Estamos a afogar-nos em informação, mas famintos por conhecimento.
[2] Em português: A prospecção de dados é um processo de descoberta de conhecimento.

A prospecção de dados pode dividir-se em duas tarefas: modelação preditiva e modelação descritiva. A modelação preditiva é um tipo de aprendizagem supervisionada, isto é, existe um conhecimento prévio em relação ao resultados que se deveriam obter, e por isso o modelo tem de encontrar os melhores critérios de decisão para conseguir classificar novos casos desconhecidos com o menor erro possível. O segundo tipo tem como objectivo encontrar a estrutura natural dos dados e por isso descobrir padrões nestes onde não existe nenhum resultado definido *apriori*, sendo por isso um tipo de aprendizagem não supervisionada. [12]

Um dos exemplos de aplicação da modelação preditiva no sector bancário é a criação de modelos de propensão à compra. Estes modelos têm como objectivo separar os clientes que adquiriram um determinado produto dos que não o fizeram, através da aplicação de vários algoritmos e técnicas, percebendo as características diferenciadoras de cada um dos grupos, e no futuro conseguir prever com uma determinada probabilidade se cada cliente vai ou não adquirir esse produto. Quanto melhor um modelo preditivo conseguir separar os dois grupos de clientes, melhor é o poder discriminatório do modelo. [13]

A modelação descritiva é utilizada habitualmente em modelos de Clustering, Regras de Associação ou *Link Analysis*. Uma vez que as empresas estão cada vez mais interessadas em conhecer os seus clientes, são muitas vezes aplicados modelos de *clustering* para se encontrarem segmentos que possam caracterizar os clientes existentes. [14] De acordo com Kotler e Armstrong não existe uma maneira única ou certa para as empresas segmentarem os seus clientes, visto esta segmentação depender das variáveis que se utilizarem, mas estes autores sugerem utilizar variáveis geográficas, demográficas e comportamentais. [15]

3. ESTRUTURA DE DADOS NO CRM DO MILLENNIUM BCP

O CRM, no Millennium BCP, está integrado na área de Marketing do Retalho, e divide-se em três equipas: a equipa de Gestão de Dados, responsável pela manutenção dos dados do *datamart* do marketing, bem como garantir a qualidade destes; a equipa de Campanhas Comerciais que é responsável pela gestão de campanhas comerciais e integração destas nos canais existentes no banco; e a equipa de Análise e Modelos, onde se desenvolvem modelos de propensão à aquisição ou retenção de clientes, de produtos ou serviços, analisam-se metodologias para perceber e melhorar a relação com estes, e consequentemente melhorar a performance de campanhas ou futuras abordagens.

A equipa de Análise e Modelos desenvolveu a *Analytical Base Table* (ABT), grande parte com base nas tabelas existentes no *datamart*, para conseguir concentrar e consolidar toda a informação actual, histórica, variações e rácios sobre todos os clientes numa única tabela, com informação desde Janeiro de 2016, servindo como *input* para os modelos e análises realizados. Esta tabela é constituída por 6.661 variáveis de 9 blocos de informação diferentes: Demográficas, Segmentação, Relação, Posse de Produtos e Serviços, *Pen*, Rentabilidade, Responsabilidades, Transacionalidade, Seguros e *InformaDB*[3].

1. Variáveis Demográficas – informação básica do cliente e enquadramento geográfico, como por exemplo idade e morada (concelho, distrito, região, país);

2. Variáveis de Segmentação – informação comercial (tipo de cliente, segmento de marketing) e de risco (grau de risco, *triad*, scores de risco, *flags* de risco[4]);

3. Variáveis de Relação – antiguidade no banco, datas de abertura e/ou fecho de contas, comunicações comerciais e não comerciais (por canal e produto), reclamações, *clicks* no site e *cross-networking;*

[3] Base de dados fornecida pela empresa com o mesmo nome - InformaDB, relativa a entidades empresariais.

[4] Contencioso, Risco Ruptura, Insolvência, Moras, Crédito Vencido, Níveis de Alerta, Negativos, Cartão Vigiado.

4. Variáveis de Posse de Produtos e Serviços – indicadores de posse de cada família de produtos e serviços;

5. Variáveis da *Pen* – informação baseada nos censos de 2011, incluindo gastos em várias categorias, como hospitais ou comida;

6. Variáveis de Rentabilidade – informação relativa ao património financeiro, recursos à ordem ou prazo, títulos, valores de rentabilidade de cada cliente no banco em cada mês, valores totais de crédito em dívida, entre outros;

7. Variáveis de Responsabilidades – informação relativa a todas as responsabilidades que cada cliente possui de forma global, interna e noutras instituições de crédito, por situação de crédito e/ou por produto financeiro;

8. Variáveis de Transacionalidade – informação relativa à caracterização da utilização de cartões (de débito e crédito), empréstimos (créditos, desconto comercial, *factoring*, *confirming*, *trade finance*), poupanças e investimentos, bem como aos saldos e perfilagem da movimentação financeira;

9. Variáveis de Seguros – informação relativa à posse de seguros, sinistros e prémios;

10. Variáveis da *InformaDB* – apenas para Empresas, incluí informação financeira, como indicadores de balanço, rácios da empresa e liquidez, e informação não financeira, como nº de colaboradores, incidentes judiciais, caracterização jurídica e relações com o exterior.

Tabela 1- Peso de cada bloco de informação constituinte da ABT

Bloco de Informação	Nº de Colunas	% do Total
Demográficas	62	0.9%
Segmentação	51	0.8%
Relação	698	10.5%
Posse de Produtos e Serviços	611	9.2%
Pen	189	2.8%
Rentabilidade	134	2.0%
Responsabilidades	1.077	16.2%
Transacionalidade	3.031	45.5%
Seguros	31	0.5%
InformaDB	777	11.7%
Total	**6.661**	**100%**

4. METODOLOGIA

Esta secção contém a metodologia utilizada em cada um dos projectos, nomeadamente descrição das fontes de informação e a apresentação das técnicas utilizadas, bem como as fórmulas e significados das estatísticas usadas. Para além disto possui também uma descrição sobre as ferramentas utilizadas para o desenvolvimento dos projectos

4.1. CLIENTE FREQUENTE DE NEGÓCIOS (CFN)

Na criação de um modelo de propensão à compra podem ser consideradas 3 etapas: **Pré-Modelação** – onde é definida a target, e realizada a redução de dimensionalidade; **Modelação** – onde se seleccionam as variáveis mais relevantes e comparam-se os vários modelos criados, escolhendo o melhor modelo – modelo campeão; e **Pós-Modelação** – onde se realiza o *backtesting*, lança-se a campanha piloto, para os clientes indicados pelo modelo e acompanham-se os resultados obtidos.

4.1.1. Pré-Modelação

Antes de criar o modelo, é importante perceber que clientes são relevantes estudar e modelar, bem como que variáveis são estatisticamente mais interessantes analisar, de acordo com o objectivo que estiver estabelecido na target, seguindo o processo representado pela **Figura 2**.

Figura 2 - Processo da Pré-Modelação

Primeiramente, define-se o universo de clientes que queremos analisar, ou seja, no âmbito do produto que será o alvo de modelação, neste caso o CFN, que tipos de clientes interessam incluir na análise, ou quais são os critérios que os clientes têm de ter para poderem adquirir este produto.

Neste caso falamos de um produto que é destinado a pequenas e médias empresas e por isso faz sentido incluir na análise apenas clientes de negócios ou empresas / clientes não particulares. Esta solução integrada (SI) é destinada a todos os clientes de negócios, que pertençam à rede de retalho,

contudo, para a criação do modelo foram considerados apenas os clientes Activos[5] ou Adormecidos[6], que nunca possuíram esta SI.

4.1.1.1. Definição da Target

Definindo o universo de clientes que queremos analisar, criam-se depois as regras de definição da target que neste caso será binária, explicitando claramente o que é o evento 0 e o evento 1, que serão os alvos de predição do modelo criado. Esta atribuição dependerá do objectivo do modelo, pois se o objectivo for criar um modelo de propensão à compra o valor 1 representará os clientes que adquiriram o produto (entre outros critérios), mas se por outro lado o objectivo for prever o abandono (modelo de churn) o evento 1 será um cliente abandonar o banco. A definição clara e objectiva da target é crucial para todo o desenvolvimento do projecto.

Com o modelo, espera-se conseguir prever em cada mês se cada cliente de negócios vai adquirir o CFN e manter a SI durante pelo menos 6 meses. Para construir a target da base de clientes a modelar, definem-se as janelas de previsão, onde para cada mês – **M** - se faz a seguinte verificação:

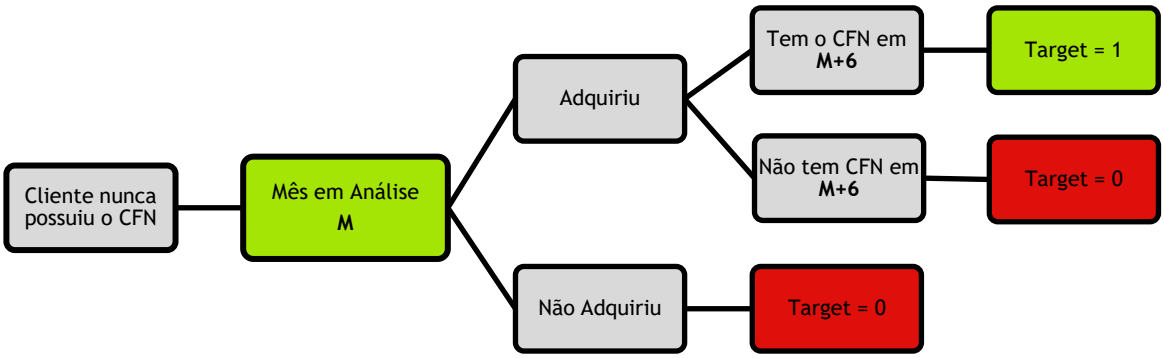


Figura 3 – Processo representativo do CFN na atribuição da target para cada mês em análise

Desta maneira consegue-se perceber quais os alvos que o modelo deve procurar, que neste caso são os clientes que nunca possuíram o CFN, e que num determinado mês adquiriram-no e mantiveram-no durante pelo menos 6 meses. As janelas consideradas foram desde Junho de 2016 (ANOMES = 201606) a Janeiro de 2017 (ANOMES = 201701), e se o mesmo cliente estiver em várias janelas terá

[5] Clientes que realizaram pelo menos uma transacção por iniciativa própria nos últimos 6 meses

[6] Clientes sem transacções por iniciativa própria nos últimos 6 meses, mas com saldo superior a 25€ ou inferior a -25€ ou com produtos associados à conta

primazia a possua a target com o valor 1, e caso não exista, a que for mais recente, atribuindo desta maneira para cada cliente um ANOMES de referência único.

4.1.1.2. Análise de Variáveis e Observações

Posteriormente analisam-se todas as variáveis da ABT individualmente, considerando o ANOMES de referência de cada cliente, mantendo cada variável ou registo se a métrica em causa estiver abaixo do nível definido como *default*, de acordo com a tabela seguinte:

Tabela 2 – Análises de linhas e colunas na Pré-Modelação

Ordem	Nível de Operação	Tipo de Análise	*Default*	Descrição
1º	Variáveis	Identificação de IDs	-	Exclusão de variáveis que sejam consideradas Ids
2º	Análise Colunas - Variáveis	Missings por variável	50%	Cálculo da percentagem de missings e exclusão de variáveis cujo valor seja superior ao definido
2º	Análise Colunas - Variáveis	Moda por variável	100%	Identificação da moda de cada variável, cálculo da percentagem de registos na moda e exclusão das variáveis cujo valor seja 100%.
2º	Análise Colunas - Variáveis	Missings + Moda por variável	100%	Cálculo da percentagem de missings e de registos na moda e exclusão das variáveis cuja soma das percentagens seja 100%
3º	Análise Linhas - Clientes	Missings por registo	75%	Cálculo da percentagem de missings e exclusão dos registos cujo valor seja superior a 75%
3º	Análise Linhas - Clientes	Zero por registo	90%	Cálculo da percentagem de entradas por registos com valor igual a 0 e exclusão dos que tenham um valor superior a 90%

O processo de análise de linhas e colunas na pré-modelação levou a que vários registos (clientes) fossem retirados da análise, visto possuírem muitos valores omissos ou muitas entradas com o valor zero. Para além disto, 2.331 variáveis foram excluídas desta análise, pelos critérios explicitados no gráfico abaixo representado.

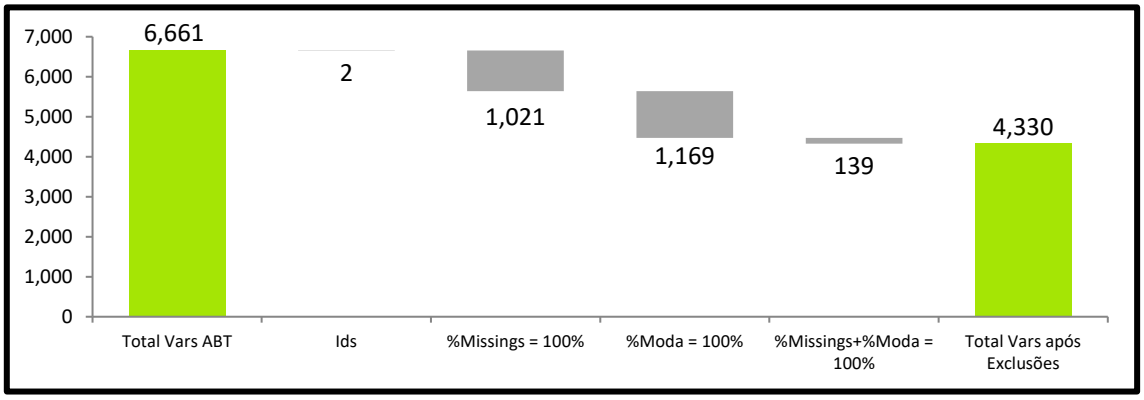


Figura 4 - Exclusão de Variáveis existentes na ABT

4.1.1.3. Ranking de Variáveis

Durante uma análise de dados para modelação é importante perceber a relação estatística entre variáveis, porque para além de poder trazer inputs importantes para o desenvolvimento do modelo, os resultados obtidos podem ser enviesados se uma ou mais variáveis estiverem muito relacionadas entre si ou não relacionadas com a target. [16] Neste sentido, com o objectivo de construir um ranking com todas as variáveis, foram determinados os seguintes indicadores estatísticos:

- *Worth* por Variável

  Para calcular o *worth* de cada variável foi construída uma árvore de decisão com apenas um nó (raíz) e de acordo com o critério especificado, dependendo da natureza de cada atributo.

  Para variáveis Intervalares foram medidas duas estatísticas diferentes de acordo com a divisão sugerida para cada variável: o p-value do teste de Fisher e a Variância. Para testar a diferença significativa entre os diferentes ramos definidos é calculado o p-value do teste de Fisher, onde quanto maior este valor mais significativa é a variável. Por outro lado, a variância mede o erro quadrático médio e é dado pela seguinte fórmula, onde N é o número de observações e Y_i a observação i:

$$Variância = \frac{1}{N}\sum_{i}(Y_i - \bar{Y})^2$$

Para variáveis nominais e categóricas foram calculadas a entropia e o índex de Gini. A entropia de uma variável mede o grau de pureza de um atributo. Dada uma variável, com observações pertencentes à classe i, com uma probabilidade da target possuir valor 1 p_i, temos:

$$Entropia = \sum_{i} p_i \times log_2(p_i)$$

O índex de Gini mede o índice de dispersão de cada variável. Dado um atributo com observações pertencentes à classe i, com probabilidade p_i da target possuir o valor 1:

$$Gini = 1 - \sum_{i} {p_i}^2$$

Estas estatísticas são calculadas através do procedimento PROC ARBOR no SAS Base.

Depois de serem calculadas as métricas aplicáveis para cada variável, todas as estatísticas são *standardizadas*[7] entre 0 e 1, e o *Worth* de cada variável será a média das métricas já normalizadas calculadas para cada uma.

- R^2 por variável

  Esta estatística mede o efeito que cada variável tem na target. Podem ser considerados quatro efeitos: *Class*, *Group*, Var e AOV16, que podem ser aplicáveis dependendo da natureza de cada variável.

  O efeito *Class* estima todas as interacções possíveis entre cada par de variáveis categóricas ou nominais, combinando todas as classes de ambas as variáveis.

  O efeito *Group* analisa o efeito de juntar classes de cada variável categórica ou nominal em grupos maiores, medindo a diminuição do efeito *Class*.

  O efeito VAR é estimado através da construção de uma regressão linear para cada variável intervalar e a target, medindo a percentagem da variância da target explicada pela regressão construída.

  O efeito AOV16 é calculado para cada variável numérica através da criação de intervalos igualmente espaçados, num máximo de 16, e testa a relação entre cada variável e a target.

[7] *Standardizar* uma métrica é transformar os valores desta de maneira a estarem compreendidos entre 0 e 1.

Estas estatísticas foram calculadas através do procedimento DMINE no SAS Base.

Para cada variável o valor de R^2 representa a média entre os efeitos acima descritos que sejam aplicáveis, sendo que estes são *standardizados* primeiro.

- Qui-quadrado por variável

  O teste do Qui-Quadrado mede a relação entre duas variáveis, neste caso entre cada uma das variáveis e a target. Tem como hipótese nula a independência das variáveis, ou seja a não existência de relação entre as variáveis e a target. Através do p-value pode-se inferir sobre esta relação, nomeadamente se possuir um valor superior a 10% (nível de significância máximo habitualmente utilizado) então pode-se concluir que existe evidência para afirmar que a variável e a target não são independentes.

  Este teste foi realizado no SAS Base através do procedimento PROC FREQ.

  Os valores de Qui-Quadrado resultantes de todas as variáveis são *standardizados* entre 0 e 1, sendo este valor normalizado que serve como métrica para o ranking das variáveis.

- Information Value (IV) por variável

  O IV tem como objectivo medir a capacidade de cada variável distinguir os valores da target. Depois de esta métrica ser calculada é *standardizada*, ficando com um valor entre 0 e 1, sendo este o valor do IV de cada variável. O IV pode ser calculado através da seguinte fórmula:

$$IV = \sum_i \left(Distribuição_{Target=1} - Distribuição_{Target=0}\right) \times ln\left(\frac{Distribuição_{Target=1}}{Distribuição_{Target=0}}\right)$$

  O Information Value de cada variável foi calculado através do PROC Tabulate no SAS Base.

- Correlação por variável

  Esta métrica mede a relação estatística entre duas variáveis (causal ou não causal), e entre cada tipo de variáveis existe uma fórmula para calcular esta estatística, como descrito pela **Tabela 3**.

A correlação de Pearson avalia a relação linear entre duas variáveis intervalares ou binárias. A correlação de Spearman mede a relação entre variáveis ordinais ou entre variáveis ordinais e intervalares ou binárias, onde avalia a tendência para as variáveis mudarem juntas mas não necessariamente com uma relação linear. O coeficiente de Cramer é uma medida de associação entre variáveis categóricas e quanto maior o valor do coeficiente maior a associação entre variáveis. [17]

Estes coeficientes foram calculados através do PROC FREQ no SAS Base.

Tabela 3 – Correlação entre cada tipo de variáveis

Correlação	Intervalar	Nominal	Ordinal	Binária
Intervalar	**Pearson**	**Cramer**	**Spearman**	**Pearson**
Nominal		**Cramer**	**Cramer**	**Cramer**
Ordinal			**Spearman**	**Spearman**
Binária				**Pearson**

O Ranking das variáveis consiste na combinação linear dos indicadores considerados, levando à eliminação de variáveis que não estejam relacionadas com a target ou que estejam muito correlacionadas entre si.

Primeiramente é calculado o ranking para cada variável de acordo com a relação com a target, que resulta da soma dos indicadores estatísticos acima mencionados – *Worth*, Qui-Quadrado, R^2, Correlação e *Information Value* – já *standardizados*, resultando num valor entre 0 e 5. Neste ranking foram dados a todos indicadores o mesmo peso, visto se considerar que são igualmente relevantes para a selecção de variáveis.

Posteriormente, tendo em consideração a correlação calculada entre variáveis de acordo com a **Tabela 3**, entre duas variáveis que possuíssem um coeficiente de correlação superior a 0,7, foi eliminada a que possuísse um ranking inferior.

4.1.1.4. Balanceamento da Amostra

Antes de começarmos a modelar os dados, como estamos perante um evento que podemos considerar raro, visto que a target possui o valor 1 em 10% dos casos, é importante balancear os alvos que temos, porque quando estamos perante dados muito desequilibrados, os algoritmos tendem a resultar em atribuir a todas as observações o caso mais comum [18], resultando em falsas classificações [19].

Para balancear os alvos de modelação existem dois métodos que são usualmente utilizados: *oversampling* - copiar observações do caso menos comum, e *undersampling* – eliminar observações do caso mais comum.

O método de amostragem utilizado neste projecto foi o *undersampling*, que consiste em manter todos os registos da classe tem menos observações, que é a representada pelo valor 1, e aleatoriamente escolher o mesmo número de observações da outra classe, de maneira a obter o mesmo número de clientes em cada uma das classes, como se pode observar pela **Figura 5**. Desta maneira obtemos 50% de clientes cuja target é 1 e 50% cuja target é 0, tendo assim uma amostra balanceada.

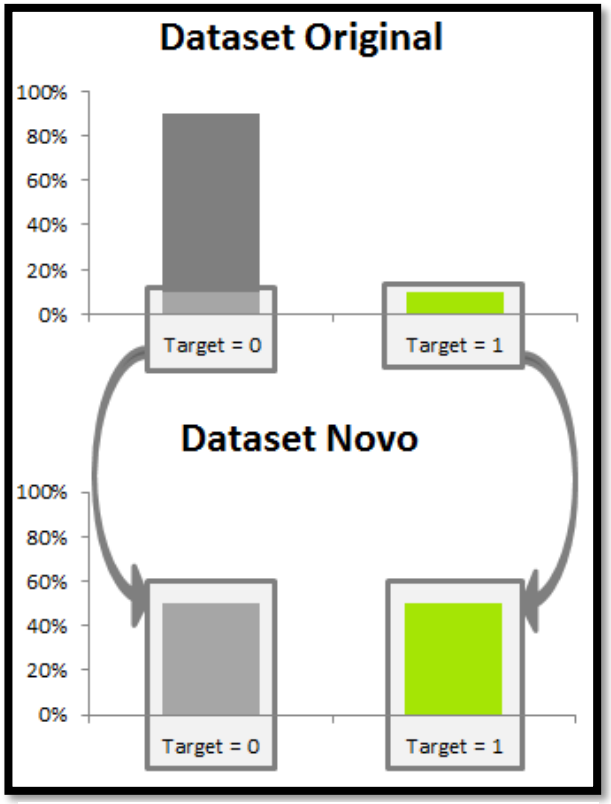


Figura 5 - Balanceamento da amostra com UnderSampling

4.1.2. Modelação

4.1.2.1. Partição de dados

Com os dados já balanceados, sendo o objectivo resolver um problema preditivo, é fundamental perceber como o modelo que criamos se comporta com dados desconhecidos, e por isto na fase da modelação é muito importante dividir as observações existentes em pelo menos duas amostras distintas: amostra de treino – utilizada para construir o modelo e estimar os parâmetros -, amostra

de validação – utilizada para validar os resultados e melhorar a precisão do modelo construído – e a amostra de teste – utilizada para testar o poder preditivo do modelo numa amostra não utilizada antes, sendo que este grupo é adequado apenas para grandes bases de dados e por isso é opcional. A partição habitual, entre grupo de treino e validação, quando não existe amostra de teste, é 70-30 ou 60-40 [20]. No desenvolvimento deste projecto foi criada apenas a amostra de treino e validação e utilizada a partição de 70% para a amostra de treino e 30% para a amostra de validação, através do nó *Data Partition* no SAS Enterprise Miner.

4.1.2.2. Selecção de Variáveis

A complexidade de um modelo pode ser um agente crucial para o sucesso deste, e um dos factores que pode contribuir para esta complexidade é a quantidade de variáveis que os algoritmos utilizam, que podem ser pouco relevantes para o objectivo da modelação. Neste sentido, as variáveis que se revelem correlacionadas com a target devem ser incluídas, mas variáveis que se revelem muito correlacionadas entre si devem ser eliminadas, já que se deve excluir toda a informação que se mostre redundante, levando a uma diminuição na variância e uma melhor precisão dos resultados obtidos [21]. Esta selecção de variáveis foi realizada através do nó StatExplore, onde para além de se poder analisar as estatísticas e distribuições de cada variável pode-se também seleccionar os atributos com maior correlação com a target.

Outro método utilizado para seleccionar variáveis foi através do R^2 (para variáveis numéricas) e Qui-Quadrado (para variáveis categóricas ou nominais) no nó Variable Selection, onde se determina a dependência entre cada variável e a target, e observando a sua relação: se a target for dependente de um atributo então este será mantido, caso sejam considerados independentes então a variável será eliminada, visto o atributo não ter impacto na target.

4.1.2.3. Transformação de Variáveis

Muitas vezes o poder preditivo das variáveis escolhidas pode ser amplificado através de uma transformação adequada, sejam as variáveis numéricas, nominais ou categóricas. Através do nó Transform Variables, pode-se escolher a melhor transformação matemática para cada atributo, maximizando a sua normalidade ou a correlação com a target. [22] Para variáveis intervalares existem dois tipos de transformações possíveis: transformações simples – aplicação de uma função matemática como a logaritmica, raiz quadrada, inversa, exponencial -, e transformações de *binning* – criar classes a partir da variável intervalar: através da criação de *buckets*, onde se divide a variável em

intervalos iguais; através de quantis, onde as observações são distribuídas em grupos com o mesmo número de observações em cada; ou através de optimal binning em relação à target, onde as observações são divididas em grupos com o intuito da distribuição da target em cada grupo ser significativamente diferente dos outros grupos. Relativamente a variáveis de classe existem também duas transformações possíveis: agrupar níveis raros num único ou criar indicadores binários para cada classe da variável.

De acordo com cada tipo de variável foram realizadas transformações diferentes. Em variáveis intervalares foi realizada a transformação que maximizava a correlação com a target, e em variáveis categóricas ou nominais foi realizado o agrupamento das classes raras em cada variável, onde o nível de corte definido foi de 5%, ou seja, classes que possuíssem menos de 5% de observações eram agrupadas numa única.

4.1.2.4. Algoritmos Utilizados

Sendo o objectivo neste projecto prever a compra do CFN por parte de cada cliente, temos uma tarefa de modelação preditiva. Para resolver este problema são habitualmente usados diversos algoritmos e modelos, e por isso neste projecto foram utlizadas árvores de decisão, regressões, redes neuronais e modelos ensemble.

Árvores de Decisão

Os modelos com base em algoritmos de árvores de decisão são muito populares em problemas de classificação e de regressão, visto serem baseados em regras facilmente percebidas. As árvores de decisão são estruturas utilizadas para dividir um grande número de observações em subgrupos através da aplicação de regras promovendo a homogeneidade de cada um, de acordo com a target definida. [23]

Ao contrário das árvores na natureza, as representações destas árvores têm uma lógica de cima para baixo, onde a raiz está no topo. Possui também nós, incluindo a raiz, que representam testes às variáveis e em “ramos” que representam as respostas aos testes. Relativamente aos nós, este são designados por nós internos ou de decisão se possuírem ramos de saída, caso contrário são designados por nós terminais ou “folhas”. [24]

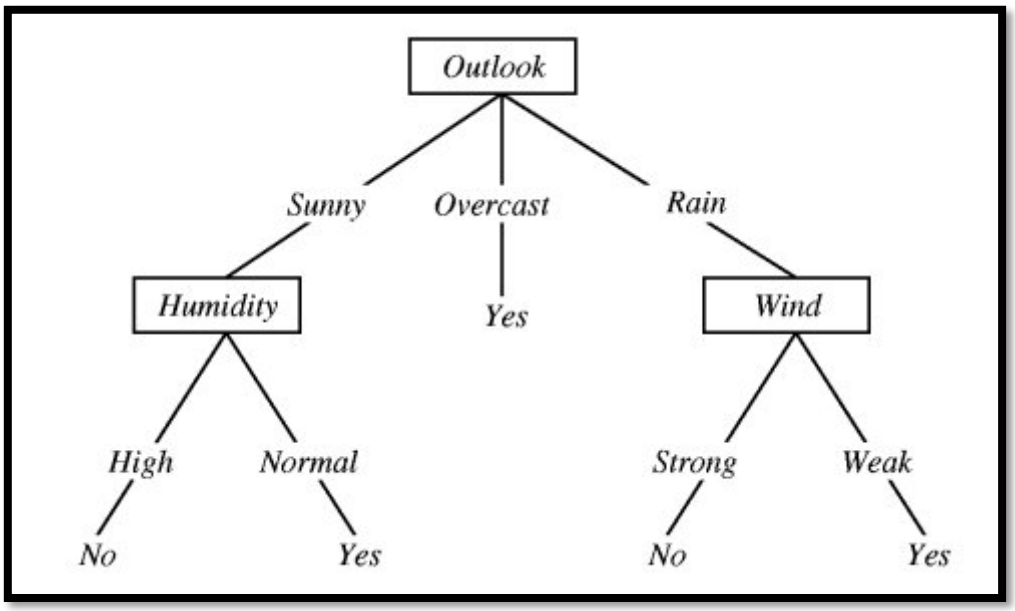


Figura 6 - Exemplo de uma Árvore de Decisão para escolher jogar ténis [25]

A cada iteração encontramos a melhor variável a utilizar de acordo com o critério utilizado, ou seja, encontra-se a variável que tem maior poder de diferenciar as classes da target e para isso são aplicadas diversas medidas de selecção de atributos, resultando no critério de divisão para cada nó. Este processo termina quando as folhas ou nós terminais possuírem distribuições da target o mais homogéneos possíveis.

No SAS Enterprise Miner foi aplicado este algoritmo através do nó Decision Tree, em que se podem escolher os critérios de selecção de variáveis, definir o tamanho máximo da árvore, bem como o número de divisões permitido em cada variável.

Regressões Logísticas

Um método estatístico que analisa a relação entre várias variáveis e a target é a regressão logística. Este modelo paramétrico combina as variáveis de input de maneira a minimizar a diferença entre o número obtido e o esperado. Como é um algoritmo que só funciona com variáveis numéricas, as variáveis categóricas são transformadas em dummies no algoritmo (onde cada classe passa a ser representada por uma variável binária).

Nas regressões logísticas existem três tipos de selecção de variáveis: *forward, backward* e *stepwise*. Com a selecção *forward* o modelo começa com apenas a ordenada na origem (*intercept*), e vão

sendo testadas e seleccionadas variáveis de acordo com o seu p-value, até nenhuma nova variável se mostrar significante no modelo. Na selecção *backward* o processo é o inverso: o modelo começa com todos os atributos e este vão sendo retirados do modelo iterativamente de acordo com o seu p-value, até todas as variáveis serem relevantes no modelo. Por último o *stepwise* é uma mistura dos dois primeiros, visto começar como o *forward*, ou seja sem variáveis, e a cada iteração é avaliada a inclusão de novas variáveis ou a exclusão de variáveis existentes.

No SAS Enterprise Miner as regressões foram aplicadas através do nó Regression, onde se pode parametrizar o modelo de selecção de variáveis, bem como os valores de p-value de entrada e de saída de cada variável no modelo.

Redes Neuronais

Os algoritmos de redes neuronais são baseados na estrutura existente no cérebro humano, que simulam o seu processo de aprendizagem. Os resultados de uma rede neural são pesos atribuídos internamente, distribuídos pela rede construída. Cada um dos neurónios recebe todos os atributos balanceados com um peso característico de cada variável. Os pesos atribuídos a cada conexão são ajustados de maneira a minimizar a diferença entre o objectivo e o output. Em cada rede pode existir uma camada intermédia, denominada por *hidden layer* com várias camadas escondidas -*hidden units* - que permite ao algoritmo aumentar o seu poder em reconhecer padrões e por isso melhorar os resultados finais, apesar de aumentar a complexidade do próprio modelo.

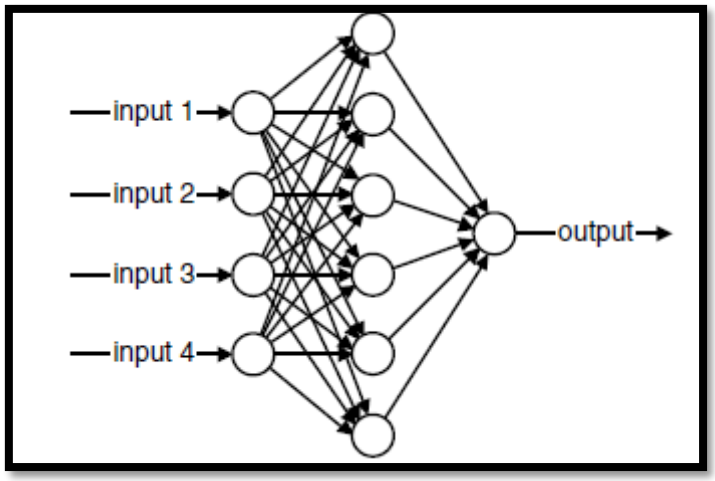


Figura 7 - Exemplo do funcionamento de uma rede neuronal [23]

No SAS Enterprise Miner este algoritmo foi aplicado através do nó Neural Networks, onde foram experimentados vários de parâmetros, nomeadamente o número de camadas escondidas, o que permite à rede desenvolver uma aprendizagem automática.

Modelo Ensemble

O modelo gerado pelo ensemble é baseado na combinação dos resultados dos modelos previamente construídos e daí originar um único score. Este modelo pode combinar os resultados dos outros modelos através da média dos resultados ou da votação, e esta combinação pode originar um modelo com maior poder preditivo e mais estável que os anteriores, visto não depender apenas de um algoritmo.

4.1.2.5. Comparação de modelos

Neste projecto o objectivo é conseguir encontrar o melhor modelo, e por isso depois de se aplicarem vários algoritmos para gerar diversos modelos é fundamental definir métricas para se encontrar o "tal", o melhor modelo. De acordo com o projecto em estudo considerou-se relevante analisar as seguintes métricas: ROC *Index*, *Missclassification Rate*, *Cumulative Percent Captured Response*, *Cumulative Lift*.

- ROC *Index* – permite quantificar a capacidade do modelo em detectar os resultados falsos positivos (observações que foram classificados como 1 mas na realidade são 0) e os falsos negativos (observações que foram classificados como 0 mas na realidade são 1);
- *Missclassification Rate* – representa a percentagem de observações que foram mal classificadas através do modelo em causa;
- *Cumulative Percent Captured Response* – representa o acumulado das respostas positivas previstas até ao decil em análise face ao total das respostas positivas previstas na amostra;
- *Cumulative Lift* – mostra o factor de elevação de quantas vezes é melhor usar um modelo preditivo do que não utilizá-lo.

4.1.3. Pós-Modelação

Nesta última fase, depois de termos encontrado o modelo campeão, para testar o seu poder preditivo realiza-se o *backtesting*, ou seja, faz-se um teste com os dados reais e não necessariamente

balanceados de um ANOMES que não foi utilizado para a modelação, para avaliar a discrepância entre os resultados obtidos através do modelo e as observações reais. [26]

Como resultado do modelo preditivo construído aplicado a este novo ANOMES teremos um score, ou seja, valores entre 0 e 1 para cada cliente/observação que representa a propensão de cada cliente a adquirir o produto segundo o modelo que criámos. Sabendo isto, os valores de output do modelo são organizados de forma crescente e agrupados em 5 escalões, de acordo com o percentil a que pertencerem:

- Escalão Baixo: Percentil de 0 a 25 exclusive;
- Escalão Médio-Baixo: Percentil de 25 a 50 exclusive;
- Escalão Médio Alto: Percentil de 50 a 75 exclusive;
- Escalão Alto: Percentil de 75 a 90 exclusive;
- Escalão Superior: Percentil de 90 a 100.

Seguidamente é avaliado o poder discriminatório do modelo, ou seja, é analisada a distribuição real da target do ANOMES em causa pelos 5 escalões, para perceber se o modelo consegue prever com precisão os resultados verdadeiros, ou seja, se a percentagem de clientes com target igual a 1 é superior nas classes superiores e nas classes inferiores o contrário: maior percentagem de clientes com valor 0 face ao valor da percentagem de clientes com target 1 naquele escalão. Se o modelo não possuir um bom poder discriminatório, então o modelo tem de ser refeito ou alguma alteração tem de ser feita, como por exemplo alterar a combinação de variáveis ou de modelos, visto o criado não conseguir prever com precisão o comportamento dos clientes.

Os escalões que cumpram o critério de no *backtesting* possuírem uma percentagem maior de clientes com o valor da target 1 no escalão face à percentagem com valor 0 são seleccionados para fazerem a triagem de clientes em campanhas futuras, segundo o score que possuam em cada ANOMES, e consequentemente o escalão a que pertençam, sendo que só serão escolhidos os clientes que pertençam a escalões que cumpram o critério acima descrito.

4.2. SEGMENTAÇÃO

Qualquer segmentação, mesmo lidando com poucas variáveis, depende maioritariamente do analista que a executa, e é mais desafiante quanto maior o número de variáveis. Contudo, para conseguirmos obter resultados mais precisos têm de ser os dados a orientar-nos, e para isso foi usada a análise de *clusters*, que é uma técnica estatística que separa observações em grupos (denominados por *clusters)*. Este tipo de análise tem como objectivo agrupar observações de maneira a que as que pertençam ao mesmo grupo sejam semelhantes (homogéneas), mas observações de grupos diferentes sejam distintas (heterogéneas) [27].

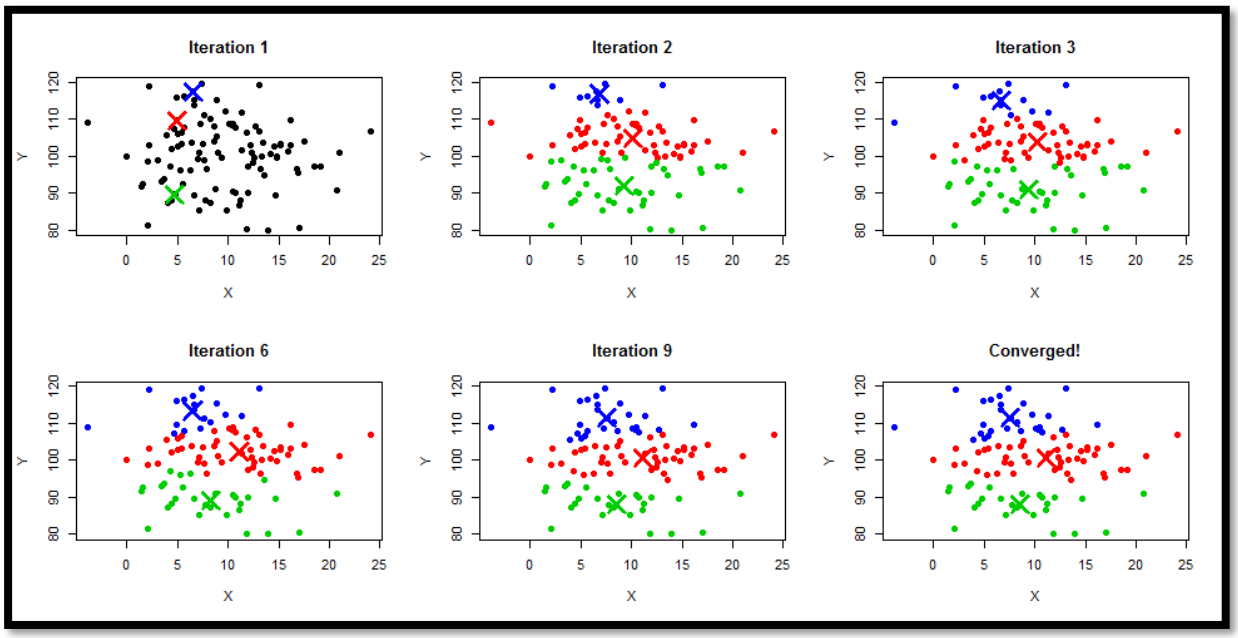
Neste projecto, para desenvolver a análise foi utilizado o K-*Means* - um dos métodos mais antigos e usados para análise de *clusters*, onde k representa o número de *clusters* escolhido. Este método consiste em escolher K *centroids*[8] iniciais, onde K é definido *a-priori* e cada ponto é assignado ao *centroid* que estiver mais próximo, de maneira a que cada ponto pertença a um único *cluster*. Depois, o *centroid* de cada cluster é actualizado de maneira a representar o ponto médio do grupo. Repete-se a atribuição de cada ponto a um *cluster*, bem como a actualização dos *centroids* até estes convergirem e se manterem na mesma posição. Muitas vezes esta condição de convergência é substituída por uma mais fraca, como por exemplo repetir os passos descritos até apenas 1% das observações se alterarem. [28]

Para cada ponto pertencer a um *cluster*, em cada uma das iterações, é tido em consideração o centroid mais próximo de cada observação. Ora para esta atribuição precisamos de definir primeiro o que é "estar mais próximo". A noção de proximidade usualmente utilizada, quando se trabalha com pontos num espaço euclidiano (n-dimensional), é a distância euclidiana, onde considerando A e B, dois pontos n-dimensionais, tem a seguinte fórmula:

$$dist(A,B) = \sqrt{\sum_{i=1}^{n} (A_i - B_i)^2}$$

Utilizando a distância euclidiana, podemos ver a figura seguinte como exemplo da aplicação do algoritmo K-*Means*, num espaço bi-dimensional:

[8] Um *centroid* é um ponto representativo de um grupo de observações – *cluster* -, habitualmente é o ponto médio.



9

Figura 8 - Exemplo do algoritmo K-*Means*

Tendo já a noção de proximidade em mente, é importante definir a função objectivo a utilizar, isto é, como podemos comparar resultados de várias aplicações do algoritmo K-*Means* (com inicializações diferentes ou com números de clusters diferentes). Para se poder avaliar a qualidade de um processo de *clustering* utiliza-se habitualmente a soma do erro quadrático (denominado por SSE – *Sum of Squared Error*), que calcula o quadrado da distância entre cada ponto (x) e o centroid do cluster (C_i) a que ficou afectado no final do algoritmo, através da fórmula:

$$SSE = \sum_{i=1}^{K} \sum_{x \epsilon C_i} dist(C_i, x)^2$$

Quanto menor o valor do SSE, melhor os *centroids* representam as observações. No limite, se utilizarmos o mesmo número de *clusters* e de observações, teremos o valor de SSE com o valor 0, o que seria um caso óbvio de *overfitting*[10], e por isso tem de se encontrar o equilíbrio entre o número de clusters e o valor de SSE. Para encontrar este equilíbrio pode-se utilizar o *elbow graphic* ("gráfico

[9] http://www.learnbymarketing.com/methods/k-means-clustering/ no dia 27 de Outubro de 2018

[10] Overfitting ou sobreajuste é um conceito estatístico utilizado para descrever um modelo que não possui uma boa capacidade de generalização e consequentemente se ajustou demasiado aos dados.

do cotovelo"), que é um gráfico que relaciona o número de *clusters* com o SSE de cada número definido, onde o ponto de cruzamento das linhas de tendência, de cada um dos lados, indica o número de *clusters* sugerido para as variáveis utilizadas no processo de *clustering*, e desta maneira consegue-se escolher o K – número de *clusters* - mais adequado. (Na secção 6 poderá ver-se como o *elbow graphic* foi aplicado.)

Contudo, num processo de *clustering* não apenas a informação de cada variável é relevante, como também a escala desta. Como a medida de distância utilizada - distância Euclideana - é sensível a diferenças de magnitude ou escalas das variáveis utilizadas, a uniformização da escala é uma peça fulcral desta análise, levando a que cada variável tenha a mesma magnitude e variabilidade. [29] Para além disso, esta normalização contribuirá também para que todas as variáveis tenham o mesmo peso na análise [30]. Por isso, a cada variável utilizada na análise, foi subtraída a média a cada observação e dividiu-se pelo desvio padrão, para todas as variáveis uma distribuição normal padrão – média igual a zero e desvio padrão igual a 1 – e consequentemente a mesma escala e peso no processo de *clustering*.

4.2.1. Fontes de Informação

Esta análise consiste em criar grupos de clientes que tenham comportamentos semelhantes com base nas variáveis definidas. O resultado final depende maioritariamente das variáveis que se escolher, visto que se se quiser uma segmentação relacionada com o valor do cliente, teremos de escolher como *inputs* variáveis como, por exemplo, a rentabilidade. Se por outro lado o objectivo for compreender o perfil de risco do cliente terá de se utilizar informação relacionada com, por exemplo, a posse de produtos de risco.

Para este estudo foram consideradas relevantes analisar seis fontes de informação diferentes: Demografia, Relação, Segmentação, Posse Actual, Transacionalidade e Compras em POS.

1. **Demografia**: idade, sexo e localidade, que foi agrupada em três grupos: Lisboa e Porto, Resto do País e Fora de Portugal;

Tabela 4 - Estatísticas da variável demográfica intervalar

Variável	Mínimo	Mediana	Máximo	Média	Desvio Padrão
IDADE_CLI	2	60	214	60	13,82

2. **Relação**: o nº de anos de relação com o banco, o crédito vivo (em euros) e o nº de comunicações recebidas no último mês;

Tabela 5 - Estatísticas das variáveis de relação intervalares

Variável	Mínimo	Mediana	Máximo	Média	Desvio Padrão
CR_VIVO	0,00	9,12	3.241.633,00	12.134,27	43.546,63
NUM_COMUNIC_3M	0,00	6,00	371,00	28,39	43,83
NUM_VENDAS_3M	0,00	0,00	16,00	0,20	0,63
RECENS_CLI	24,00	27,00	109,00	27,63	2,94

3. **Segmentação**: estado de actividade do cliente no banco, o património financeiro, *flags* de risco e o grau de risco de cada cliente (agrupado em Superior, Alto, Médio Alto, Médio Baixo, Baixo e Inferior);

Tabela 6 - Estatísticas da variável de segmentação intervalar

Variável	Mínimo	Mediana	Máximo	Média	Desvio Padrão
PATRI_FIN	0,00	5.044,91	12.500.000,00	39.324,67	159.884,90

4. **Posse Actual**: produtos e serviços que cada cliente detinha no mês em análise (Abril), incluindo produtos das três famílias existentes: Crédito (por exemplo: Cartão de Crédito, Crédito Imobiliário, Crédito ao Consumo); Recursos, Seguros e Cartões de Débito (por exemplo: Produtos de Investimento e Poupança) e Serviços (por exemplo: canais de *Self-Banking*, domiciliações de vencimento ou pagamentos). Com base nestas variáveis criaram-se mais duas: o nº de cartões (nº de cartões de crédito + nº de cartões de débito) e o nº total de produtos.

Tabela 7 - Estatísticas das variáveis de posse intervalares

Variável	Mínimo	Mediana	Máximo	Média	Desvio Padrão
CART_DB	0	1	8	117.893,00	0,74
CR_CART	0	0	7	0,59	0,74
CR_CURTO_PRAZO	0	0	3	0,10	0,30
CR_IMOBILIARIO	0	0	3	0,19	0,51
CR_PESSOAL_AUTOMOVEL	0	0	3	0,10	0,30
INVEST_POUP	0	1	41	1,31	2,23
INVEST_POUP_CTAS_POUP	0	0	5	0,23	0,46
MILLENNIUM_PT_SITE	0	0	2	0,37	0,48
MOBILE_BANKING_APP	0	0	1	0,14	0,35
NR_CART	0	2	11	1,77	1,15
NUM_PROD	1	9	54	10,03	4,60
SEGURO	0	0	11	0,83	1,24
SERV_DOMICILIACOES_PAGAMENTOS	0	1	1	0,74	0,44
SERV_DOMICILIACOES_VENCIMENTO	0	1	2	0,59	0,55
SOLUCOES_INTEGRADAS	0	1	4	0,78	0,69

5. **Transacionalidade**: Logins em canais digitais (aplicação e site), valor e montante de levantamentos e depósitos, nº de requisição de cheques, total de créditos e débitos efectuado por iniciativa do cliente[11], nº de pagamentos de serviços, nº de domiciliações de pagamentos e nº de transferências discriminadas por possuírem um âmbito eventual ou permanente, terem sido realizadas em balcão, ATM ou digital, e o destinatário ser cliente BCP ou cliente OIC. A cada uma destas variáveis foi analisada a variação mensal, trimestral e semestral, com o objectivo de perceber se existia alguma que fosse constante, ou muito pouco variável.

[11] Consideram-se débitos por iniciativa do cliente todas as transacções que impliquem diminuir o saldo da conta ordenadas pelo cliente (levantamentos, transferências do cliente) e créditos as que impliquem aumentar o saldo (depósitos, transferências para o cliente).

Tabela 8 - Estatísticas das variáveis de transacionalidade intervalares

Variável	Mínimo	Mediana	Máximo	Média	Desvio Padrão
N_DEP_3M	0,00	0,00	412,00	1,53	4,21
N_DOMPAG_3M	0,00	6,00	233,00	8,96	9,97
N_LEV_3M	0,00	7,00	416,00	11,52	14,11
N_LOGS_DIG_3M	0,00	0,00	3.071,00	34,18	106,05
N_PAGSERV_3M	0,00	3,00	678,00	5,60	9,15
N_REQ_CHEQ_6M	0,00	0,00	57,00	0,24	0,81
N_TOT_CR_INICLI_6M	0,00	14,00	3.082,00	19,73	31,10
N_TOT_DB_INICLI_6M	-3,00	94,00	2.090,00	126,85	12,43
N_TOT_TRF_6M	0,00	3,00	884,00	10,08	21,82
N_TRF_ATM_3M	0,00	0,00	181,00	1,26	32,35
V_DEP_3M	0,00	0,00	3.232.157,00	2.547,60	27.119,62
V_LEV_3M	0,00	500,00	201.550,00	929,11	1.685,86
V_TOT_CR_INICLI_6M	0,00	9.480,00	28.133.832,00	25.559,47	135.865,60
V_TOT_DB_INICLI_6M	-43,60	6.911,08	14.235.790,00	17.013,77	73.946,67
V_TOT_TRF_6M	0,00	555,00	13.540.765,00	7.197,24	60.379,37
V_TRF_DIG_6M	0,00	0,00	2.000.000.000,00	47.652,50	7.757.286,00
N_TRF_OIC_3M	0,00	0,00	239,00	2,13	4,31
N_TRF_OIC_6M	0,00	1,00	440,00	4,30	8,41
V_TRF_BCP_6M	0,00	0,00	1.126.460,00	2.113,81	14.840,48
V_TRF_OIC_3M	0,00	0,00	1.116.719,00	967,31	6.593,37
V_TRF_OIC_6M	0,00	51,29	1.716.649,00	1.959,48	11.335,81
V_TOT_TRF_EVEN_DB_3M	0,00	85,60	5.300.027,00	3.011,42	27.898,48
V_TOT_TRF_EVEN_DB_6M	0,00	400,00	13.540.765,00	6.280,56	59.008,94

6. **Compras em POS**: Com base num histórico de 6 meses de compras, obteve-se o número e montante total de compras em POS. Como a cada POS está associado o MCC da empresa, consegue-se perceber onde cada cliente gasta em maior montante e com maior frequência. Ao todo existem cerca de 500 MCC's diferentes, e por isso foram agrupados em 23 categorias diferentes: Animais, Arte e Cultura, Beleza e Jóias, Casa, Combustíveis, Desporto, Ensino, Entretenimento, Governo, Justiça, Lojas,

Material Escritório / Escolar, Moda, Plantas, Saúde, Seguros, Serviços, Supermercados, Transportes, Tecnologia, Veículos e Serviços, Viagens e outros (todos os MCC que não se enquadravam em nenhuma outra categoria). Foram criadas mais duas categorias, que incluem MCC das várias categorias anteriormente definidas, mas que se consideraram relevantes: Crianças (inclui roupas infantis, escolas, creches, entre outros) e Entretenimento Adultos (inclui bares, casinos, entre outros). Como se verificou que o número e montante total de compras e o número e montante de compras em supermercados estavam muito relacionados, visto ser o lugar mais frequente de utilização de cartões em POS, decidiu-se criar duas novas variáveis o novo nº de compras e o novo montante de compras (€), ou seja o nº e montante totais sem os valores referentes às compras realizadas em supermercados. Estas variáveis são fundamentais para se poder calcular o peso de cada categoria no montante gasto, excluindo supermercados.

Na tabela abaixo apresentada podem-se encontrar as estatísticas das variáveis relativas ao número (N_'CATEGORIA') e montante de compras (V_'CATEGORIA') de cada categoria, bem como do número (N_TOT_COMPR) e montante (V_TOT_COMPR) totais de compras. Na tabela os números positivos representam compras efectuadas e os negativos devoluções efectuadas no período em análise.

Tabela 9 - Estatísticas das variáveis de compras em POS intervalares

Variável	Mínimo	Mediana	Máximo	Média	Desvio Padrão
N_ANIMAIS	-1	0	219	1,24	3,67
N_ARTE_CULTURA	-1	0	440	2,19	6,98
N_BELEZA_JOIAS	-1	0	410	2,68	6,04
N_CASA	-3	1	292	3,00	5,74
N_COMBUSTIVEIS	0	4	452	11,06	1,87
N_COMPR_TOT	-2	84	2.862	123,94	1,32
N_CRIANCAS	-2	0	152	1,37	4,46
N_DESPORTO	-1	0	121	1,11	2,68
N_ENSINO	-4	0	150	0,22	1,52
N_ENTRETENIMENTO	-1	0	891	2,25	8,78
N_ENTRE_A	0	0	536	0,81	5,97
N_GOVERNO	-4	0	735	1,28	0,53

N_JUSTICA	0	0	38	0,07	0,58
N_LOJAS	-16	3,00	643,00	7,04	12,00
N_MATERIAL	-2	0	165	0,39	1,74
N_MODA	-2	3	203	7,71	1,26
N_OUTROS	-60	1	658	4,11	9,73
N_PLANTAS	-1	0	136	0,91	2,52
N_SAUDE	0	3	163	6,74	9,55
N_SEGUROS	0	0	33	0,06	0,42
N_SERVICOS	-28	0	1.617	2,11	8,94
N_SUPERMERCADOS	0	26	1.624	44,45	53,24
N_TECNOLOGIA	-5	0	159	0,95	2,86
N_TRANSPORTES	-97	0	939	3,02	17,16
N_VEICULOS_SERV	-2	0	301	1,09	3,28
N_VIAGENS	-5	0	2.323	1,98	11,61
V_ANIMAIS	-12,39	0,00	109.500,00	62,03	473,02
V_ARTE_CULTURA	-319,47	0,00	44.141,13	111,94	485,10
V_BELEZA_JOIAS	-27,99	0,00	89.832,00	164,58	708,43
V_CASA	-398,00	32,39	32.456,16	260,77	789,66
V_COMBUSTIVEIS	0,00	112,04	53.971,71	37,22	748,17
V_COMPR_TOT	-2.433,94	3.260,92	10.372.330,00	6.234,48	48.036,25
V_CRIANCAS	-855,80	0,00	62.005,65	98,88	604,51
V_DESPORTO	-364,61	0,00	18.047,13	59,18	217,91
V_ENSINO	-4.703,30	0,00	44.002,00	44,03	419,18
V_ENTRETENIMENTO	-718,90	0,00	754.550,00	172,27	3.331,17
V_ENTRE_A	-387,86	0,00	754.550,00	71,94	3.242,41
V_GOVERNO	-448,32	0,00	10.225.266,00	551,12	43.406,52
V_JUSTICA	0,00	0,00	11.899,02	5,24	87,90
V_LOJAS	-3.050,41	99,99	139.761,40	310,74	954,78
V_MATERIAL	-28,82	0,00	21.028,83	26,03	155,17
V_MODA	-110,44	116,72	589.140,10	462,55	2.455,79
V_OUTROS	-4.147,38	43,00	176.936,20	319,45	1.555,34
V_PLANTAS	-44,15	0,00	127.813,40	39,41	488,52
V_SAUDE	-964,79	95,57	113.062,40	338,58	962,35
V_SEGUROS	0,00	0,00	10.000,00	9,90	105,99
V_SERVICOS	-3.701,01	0,00	3.528.520,00	174,30	12.371,70

V_SUPERMERCADOS	0,00	796,95	1.255.900,00	1.431,87	5.633,01
V_TECNOLOGIA	-1.400,00	0,00	48.237,78	62,01	385,17
V_TRANSPORTES	-823,33	0,00	171.684,90	46,19	680,79
V_VEICULOS_SERV	-276,86	0,00	197.446,10	256,99	1.817,97
V_VIAGENS	-2.855,73	0,00	519.621,40	469,81	3.321,57

4.3. FERRAMENTAS UTILIZADAS

Nos projectos apresentados neste relatório, para além do Excel, foram usadas as ferramentas SAS Base e SAS Enterprise Miner.

O SAS Base é um programa que permite manipular dados e bases de dados, armazenar informação e permite realizar várias análises estatísticas. A sintaxe simples e a facilidade em corrigir erros (graças à janela de log muito descritiva e compreensiva) desta ferramenta permite criar todos os programas necessários para desenvolver os projectos nas fases de pré e pós modelação. Por outro lado, o SAS Base é um software pago, que possui muitas limitações a nível de representação gráfica.

O SAS Enterprise Miner foi utlizado no desenvolvimento dos projectos do modelo de propensão à compra e da segmentação comportamental. É uma ferramenta de data mining, que possui várias "caixas" com operações e análises que se arrastam para o diagrama principal, construindo o nosso projecto, mas também permite escrever o próprio código, fornecendo insights importantes para conseguirmos tomar decisões mais apoiadas em estatísticas de cada variável ou do próprio modelo. Esta ferramenta tem uma grande limitação que é a falta de capacidade de processamento de grandes volumes de dados, o que levou a que tanto no desenvolvimento tanto do modelo de propensão à compra como na criação da segmentação comportamental se tivesse de utilizar uma amostra com uma dimensão muito inferior à da população.

5. MODELOS DE PROPENSÃO À COMPRA

Com o objectivo de tornar as campanhas de venda da solução integrada Cliente Frequente de Negócios (CFN) mais eficazes, aumentando a taxa de conversão de contactos em vendas, foi criado um modelo de propensão à compra deste produto, ou seja, um modelo que atribui a cada cliente considerado elegível um score, i.e. uma probabilidade de aquisição do produto alvo de modelação.

Depois da fase de pré-modelação e posterior tratamento de dados já no SAS Enterprise Miner, foram utilizadas abordagens diferentes com o objectivo de perceber quais os modelos que produziam os melhores resultados de acordo com as métricas estabelecidas na metodologia. Em cada um dos algoritmos foram testados vários parâmetros para se poder inferir sobre que valores conduziam a melhores resultados.

Na tabela abaixo apresentada, para cada modelo testado, estão os valores das métricas do ROC *Index, Missclassification Rate, Cumulative Lift* e *Cumulative Percentual Captured Response,* para as amostras de validação no SAS Enterprise Miner.

Tabela 10 - Modelo de Propensão: Comparação de modelos

Modelo	ROC Index	Missclassification Rate	Cumulative Lift	Cumulative % Captured Response
Regressão - Backwards	0.95	0.11	1.95	47.9%
Regressão - Stepwise	0.95	0.10	1.91	46.82%
Árvore de Decisão	0.93	0.10	1.87	46.82%
Rede Neuronal	0.95	0.10	1.93	48.35%
Ensemble	0.96	0.10	1.97	49.4%

De acordo com todas as métricas consideradas, nomeadamente *ROC Index, Missclassification Rate, Cumulative Percent Captured Response* e *Cumulative Lift*, pode-se considerar que o modelo

Ensemble é o que produz os melhores resultados na amostra de validação e por isso é o modelo campeão, que junta os resultados das regressões e da rede neuronal. O modelo construído com as árvores de decisão não foi incluído nesta combinação porque piorava as estatísticas do modelo ensemble.

O gráfico abaixo representa a curva de ROC do modelo Ensemble, que tem em conta dois aspectos a *Sensitivity* e a *Specificity*. A *Sensitivity* mede a rapidez com que o modelo captura os eventos 1, e a *Specificity* que mede a rapidez com que o modelo captura os eventos 0. Quanto mais a curva estiver afastada da *Baseline* (linha azul nos gráficos) e perto do ponto (0,1) mais preciso será o modelo, conduzindo a melhores resultados.

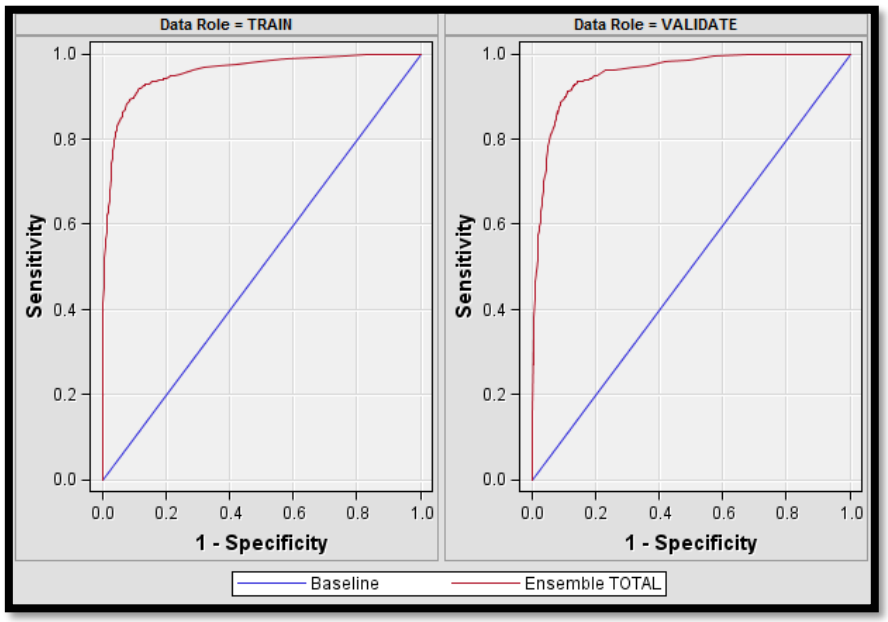


Figura 9 - Modelo de Propensão: ROC Index

O modelo ensemble possui a *cumulative lift* com o valor de 1.97 no percentil 25, o que revela que, ao contactar apenas os 25% dos clientes melhor classificados pelo modelo iremos apurar 1.97 vezes mais respostas positivas do que não usando modelo ou usando uma selecção aleatória dos alvos. Para além disso este modelo possuir um valor de 48.54% de respostas capturadas acumuladas percentuais no percentil 25 significa que aos contactar 25% dos clientes iremos captar aproximadamente 48.5% dos eventos classificados como 1.

Depois de escolher o modelo campeão foi realizado o *backtesting*, cujos resultados se apresentam em baixo.

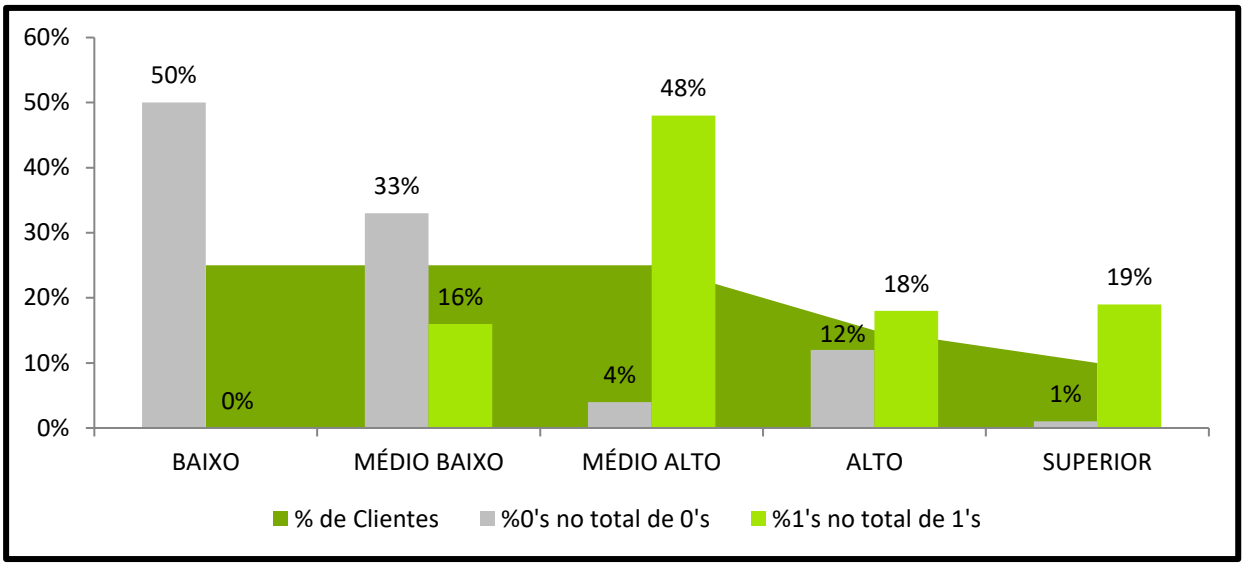


Figura 10 - Modelo de Propensão: Resultados do Backtesting

Através da distribuição dos escalões apresentada pode-se perceber como o modelo classificou os clientes de um ANOMES que não tinha sido utilizado para a modelação - Fevereiro de 2017. Neste caso o modelo parece estar a separar os eventos 0 e 1 relativamente bem, pois até ao percentil 50 a probabilidade do evento 1 acontecer face ao evento 0 é inferior, enquanto que a partir desse percentil a tendência inverte.

Nos primeiros dois escalões captura-se cerca de 83% dos clientes que foram considerados como 0, ou seja que não compraram o CFN ou que não o mantiveram durante pelo menos 6 meses e captura somente 16% dos eventos contrários. Por outro lado nos escalões Médio Alto, Alto e Superior capturaram-se cerca de 85% dos eventos considerados como 1, ou seja os clientes que foram contactados e que mantiveram a solução durante pelo menos 6 meses, e apenas 17% dos eventos classificados como 0.

Apesar do escalão Médio Alto apresentar uma percentagem de eventos 1 capturados mais elevada do que o pensado como normal, considerou-se que como possuía uma grande diferença de proporções dos eventos capturados distinguia bem os valores alvos.

Como o modelo demonstra ter um poder discriminatório considerado bom nos três escalões superiores, o modelo seguiu para produção onde os alvos foram escolhidos com base nessa classificação.

O modelo entrou em produção no ANOMES 201802 e desde essa altura, comparando com períodos homólogos, veio a melhorar os resultados, visto a taxa de venda ter aumentado de 1,7% em média para 5,4% no segundo ciclo e no terceiro ciclo de 2,2% em média para 7,7%. Em ambos os casos apesar do número de listados para a campanha se ter mantido semelhante, o número de contactados diminuiu resultando numa diminuição na taxa de contacto.

Tabela 11 - Modelo de Propensão: Taxas de Venda antes de depois do modelo

Ciclo	Taxa de Contacto	Taxa de Intenção de Compra	Taxa de Venda
201602	41%	9%	1,6%
201702	28%	7%	1,8%
201802	19%	12%	5,4%
201603	15%	5%	2,5%
201703	22%	6%	1,9%
201803	14%	15%	7,7%

6. SEGMENTAÇÃO COMPORTAMENTAL

Actualmente os clientes particulares no BCP são segmentados tendo em conta o seu património financeiro, vencimento e idade. Ora, com todos os dados que o banco possui dos seus clientes, ainda não existe nenhuma informação trabalhada relativa ao seu comportamento: quem é o cliente no banco? O que faz? O que compra? Onde gasta? E na óptica do marketing, ainda faz sentido oferecer o mesmo produto/serviço com a mesma abordagem a clientes diferentes só porque têm idades iguais? Para responder a estas perguntas surgiu o estudo da segmentação comportamental, cujo objectivo é a criação de novos segmentos, com base no comportamento destes.

Neste capítulo estão descritas as etapas do projecto pela seguinte ordem: escolha das fontes de informação, clientes a estudar, definição dos *clusters* de segmentação e resultados por perfil de clientes.

6.1. CLIENTES ANALISADOS

Os clientes alvos desta análise são clientes particulares do banco, e por isso que pertençam aos segmentos Mass Market (MS), Plus (UU) ou Prestige (PP). Para além disso, como se pretende estudar o comportamento actual dos clientes, estabeleceu-se que se iriam analisar apenas os clientes cujo estado de actividade fosse A1, i.e., clientes primeiros titulares de pelo menos uma conta com transacções realizadas por iniciativa própria nos últimos 6 meses, retirando à população a analisar 19% dos clientes inicialmente considerados.

Depois, com base nas variáveis definidas anteriormente, e nos clientes acima descritos, foi feito o processo de identificação de *outliers*[12] no SAS Enterprise Miner para cada uma das variáveis, onde se utilizou a lógica, na qual, a partir do primeiro espaçamento significativo entre intervalos, as observações à sua direita deveriam ser consideradas outliers e, como tal, excluídas, excepto em situações nas quais o número de observações fosse significativo, podendo afectar a qualidade dos segmentos criados.

[12] Valor que apresenta um grande afastamento em relação aos demais da amostra [1]

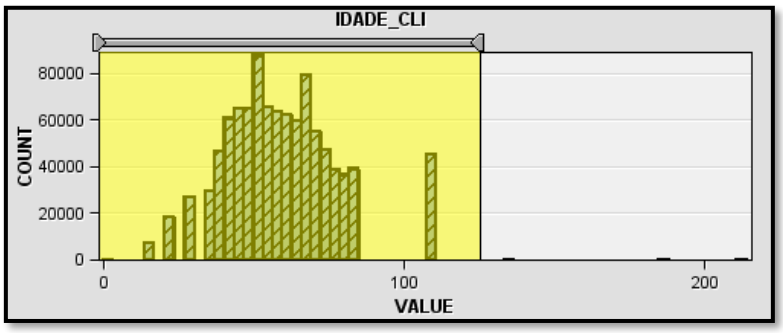


Figura 11 - Processo de identificação de outliers na Idade do Cliente

Podemos ver o exemplo da distribuição da idade dos clientes, onde a partir do valor 110 se consideraram que os clientes eram outliers, por a partir desse valor existir um espaçamento muito significativo e à direita deste valor não existirem observações em número considerável.

Ao todo foram considerados como outliers, seguindo a lógica anterior, 0.71% dos clientes activos (A1) e por isso foram excluídos da análise. No final da identificação de clientes com valores extremos, a proporção de segmentos presente na população a estudar era de 55.63% de clientes MS, 26.74% de clientes UU e 17.63% de clientes PP.

6.2. Definição dos Clusters de Segmentação

De acordo com as fontes de informação descritas acima, considerou-se relevante definir dois perfis, e por isso duas análises de *clusters*, dando desta forma uma visão mais alargada do que o cliente é dentro e fora do banco. O primeiro é baseado na informação relativa às compras feitas com cartões em POS, que tem o objectivo de encontrar perfis que expliquem onde é que o cliente gasta mais dinheiro e que categorias têm mais peso no orçamento de cada um – **Perfil de Compras em POS**. O segundo pretende explicar o que o cliente é e o que faz no banco, com informação relativa à demografia, posse e transacionalidade – **Perfil de Utilização de Serviços Bancários**.

Para cada perfil definido foi desenvolvida uma análise de *clusters* no *SAS Enterprise Miner*, através da técnica K-*Means*, onde K representa o número de clusters definido, sendo que para definir o número de clusters foi usado o *Elbow Graphic*.

6.2.1. Perfil de Compras em POS

Como referido anteriormente, com este perfil pretende-se perceber quais os interesses e obrigações (traduzidos em gastos) de cada cliente, e compreender quem é o cliente fora do banco. Para isso foram usadas as informações relativas aos gastos em cada uma das categorias de MCCs.

Através do Elbow Graphic tentou-se perceber qual o número de *clusters* a utilizar para este perfil:

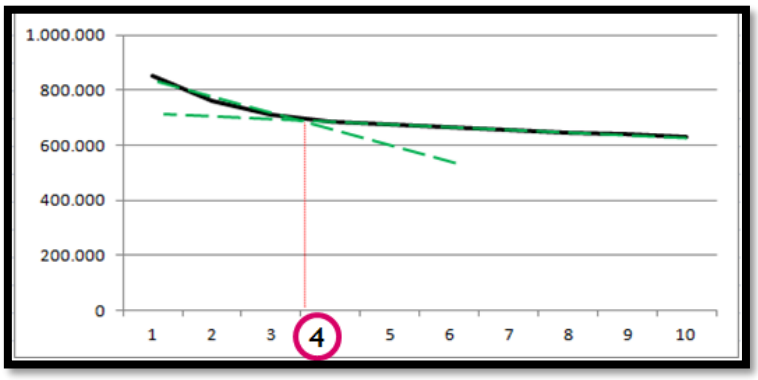


Figura 12 - Elbow Graphic para o Perfil de Compras

Neste gráfico pode-se observar cada número de *clusters* definido e o respectivo SSE, que tem por objectivo encontrar o ponto em que aumentar o número de clusters a usar não diminui significativamente o erro quadrático acumulado. Neste caso o número de *clusters* sugerido pelas variáveis utilizadas nesta segmentação é quatro, visto ser o ponto onde as duas rectas de tendência se encontram, aproximadamente.

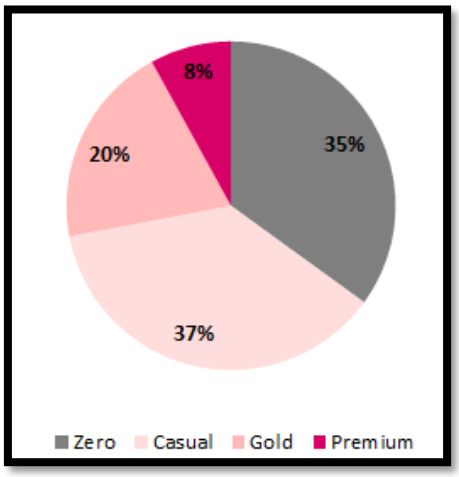


Figura 13 - Distribuição dos segmentos do perfil de compras em POS

Após a aplicação do processo de *clustering*, tendo por base as variáveis consideradas para a segmentação de compras em POS, obtiveram-se quatro segmentos: Zero, Casual, Gold e Premium. Passemos à análise de cada um dos perfis obtidos para a segmentação dos clientes por Compras em POS, onde nos gráficos apresentados, para cada variável normalizada, a vermelho está a distribuição da população e a azul a distribuição do *cluster*:

- **Zero – *Cluster* 1 – 35%**

Este *cluster* caracteriza-se por ter clientes que, na sua maioria, não realizaram qualquer compra em POS nos últimos 6 meses, incluindo em supermercados, que é a categoria mais frequente em toda a população.

Aproximadamente 98% destes clientes não compraram nada através de POS em supermercados, sendo que na população total apenas 58% não o haviam feito. Relativamente ao número de compras no resto das categorias, onde apenas 31% da população não tinha efectuado compras em qualquer categoria excepto supermercados, neste segmento cerca de 85% não fez nenhuma compra. Estes clientes destacam-se dos outros segmentos, também, por gastarem menos em despesas do Governo.

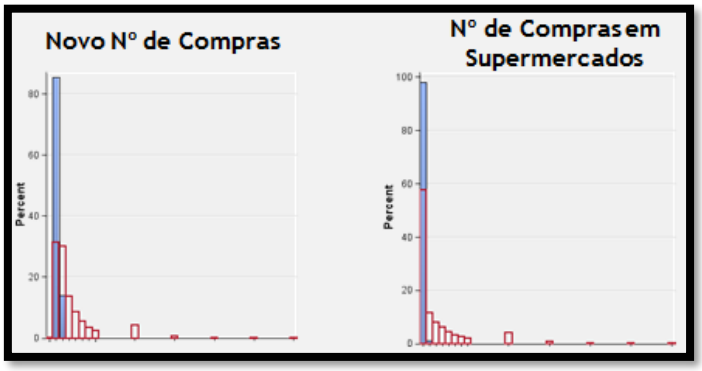


Figura 14 - Zero: Variáveis Características

- **Casual – *Cluster* 2 – 37%**

Este *cluster* é representado por clientes que efectuam poucas compras. Apesar de não gastarem tanto como a média, as compras em moda e em restaurantes têm algum peso no montante gasto.

Neste segmento apenas 42% dos clientes não fizeram qualquer compra, sendo que cerca de 50% da população total está nessa situação. As compras em moda foram realizadas por 65% do segmento, valor superior ao da população. Para além disto, este perfil caracteriza-se também pelos gastos em

restaurantes terem um peso no total gasto superior ao do resto da população (em todas as categorias excepto supermercados).

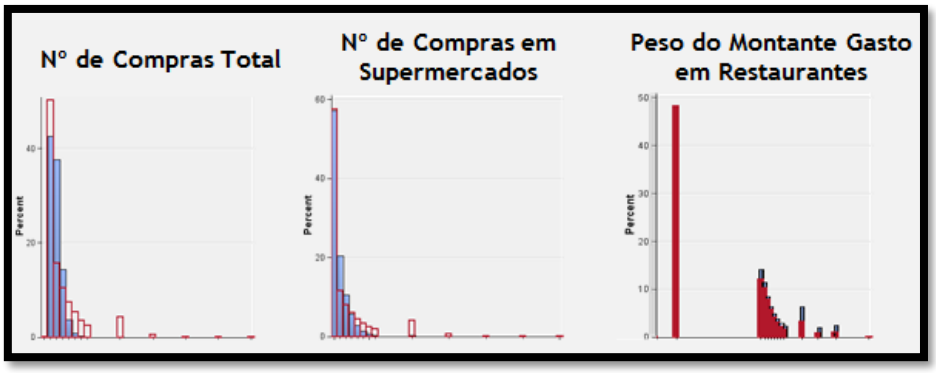


Figura 15 - Casual: Variáveis Características

- **Gold – *Cluster* 3 – 20%**

Este *cluster* caracteriza-se por ter clientes que compram mais do que a média, em número e montante. Gastam sobretudo em Combustíveis, Lojas, Supermercados e Moda. A saúde é a categoria que assume um peso superior no montante gasto total relativamente ao resto da população.

Todos os clientes neste segmento fizeram pelo menos uma compra nos últimos 6 meses, através de POS, como se pode verificar pela **Figura 16**, e em número superior ao do resto da população. Relativamente às categorias, pode-se observar que apenas 19% dos clientes Gold não realizaram compras de combustíveis, ao contrário do total da população, onde este número ascende a 69%. Para além disto, o montante gasto em Saúde faz com que esta categoria tenha um peso maior no total gasto por estes clientes face ao resto da população.

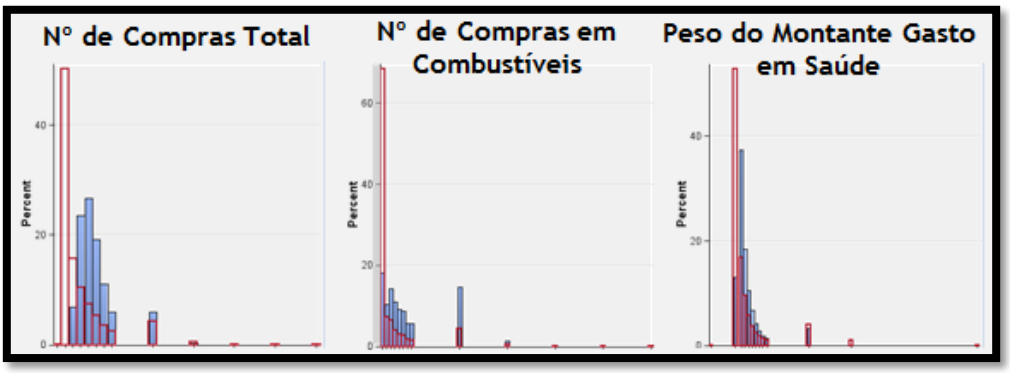


Figura 16 - Gold: Variáveis Características

- **Premium – *Cluster* 4 – 8%**

Este *cluster* tem clientes que efectuam muitas compras, de elevado valor, em várias categorias. A destacar o nº e montante gasto em moda, beleza e jóias, crianças, casa, lojas, saúde e desporto.

Em relação ao número de compras, 99% dos clientes deste segmento fizeram compras em Moda, valor muito superior à distribuição da população. Tanto em Beleza e Jóias, como em Crianças este perfil possui clientes que despendem muito mais do que o resto da população.

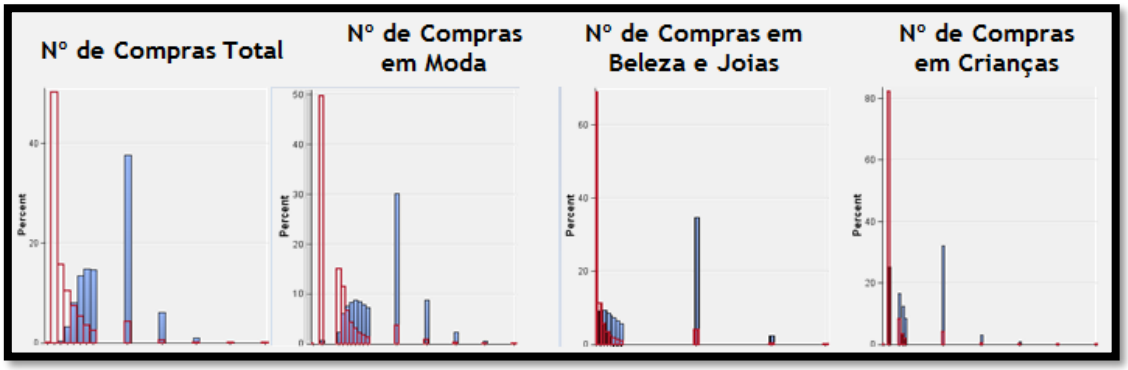


Figura 17 - Premium: Variáveis Características

6.2.2. Perfil de Utilização de Serviços Bancários

A criação deste perfil pretende perceber quem são os clientes dentro do banco, o que fazem, o que possuem? Quem são eles? Para isso utilizou-se informação relativa à demografia, posse de produtos bancários e transacionalidade do cliente.

Através do *Elbow Graphic* abaixo apresentado, consegue-se perceber que o número de *clusters* a utilizar de acordo com as variáveis de input é quatro, visto este ser o número em que aumentar o número de *clusters* não favorece substancialmente o SSE, observável através das linhas de tendência sugeridas pelo gráfico.

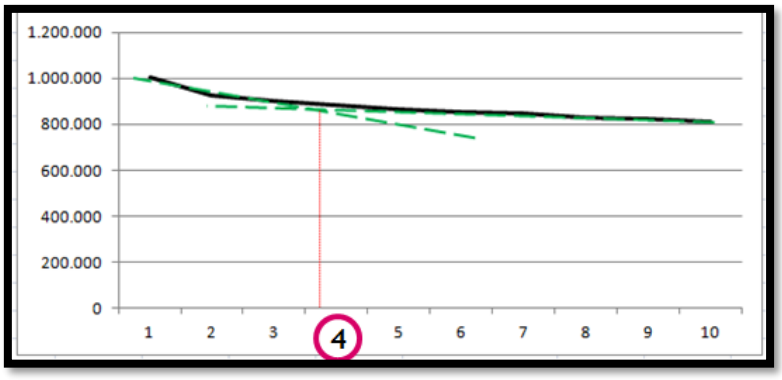


Figura 18 - *Elbow Graphic* para o Perfil de Utilização de Serviços Bancários

Tendo em consideração o resultado do *Elbow Graphic*, aplicou-se o algoritmo K-*Means* com as variáveis acima definidas (demográficas, transaccionais, de relação, de segmentação e de posse) para criar os quatro segmentos para o perfil de utilização de serviços bancários - Pouco Activos, Pensionistas, Produtos e Transferências - onde se obteve a seguinte distribuição:

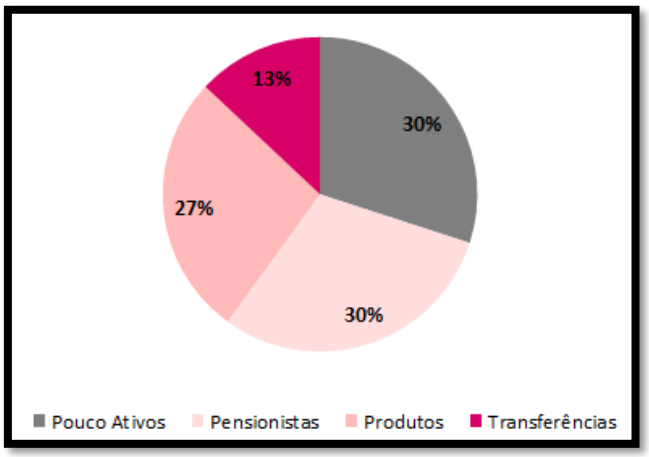


Figura 19 - Distribuição dos segmentos do perfil de utilização de serviços bancários

Passemos à análise de cada um dos perfis obtidos para a segmentação dos clientes por Utilização de Serviços Bancários, onde nos gráficos apresentados a vermelho está a distribuição da população e a azul a distribuição do *cluster*:

- **Pouco Activos – *Cluster* 1 – 30%**

Este *cluster* caracteriza-se por ter clientes que não têm muito envolvimento com o banco.

Relativamente a transacções a débito por iniciativa do cliente (como levantamentos ou transferências para outra contas), cerca de 71% dos clientes não efectuou nenhuma transacção deste tipo nos últimos 6 meses, mais especificamente 80% não fez nenhum levantamento. Para além disto, estes clientes possuem menos produtos do que o resto da população, como se pode verificar pelo segundo gráfico, onde 39% destes clientes não possuem qualquer cartão.

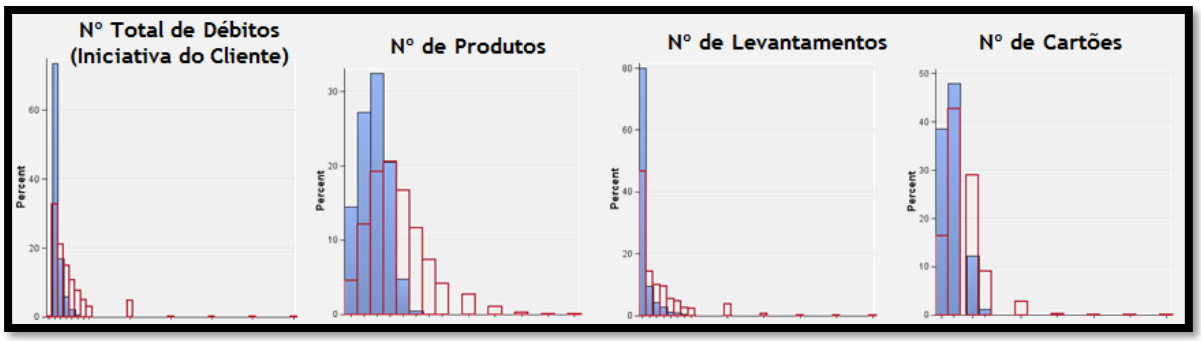


Figura 20 - Pouco Activos: Variáveis Características

- **Pensionistas – *Cluster* 2 – 30%**

Este *cluster* é caracterizado por ter clientes maioritariamente mais velhos, não digitais visto que 98% destes clientes não fez nenhum login na aplicação ou no site, e onde a maioria dos clientes – 73% - possui têm o vencimento domiciliado, nomeadamente a reforma. Apesar de não realizarem muitos levantamentos nem depósitos, o valor destas operações por transacção é muito superior ao do resto da população.

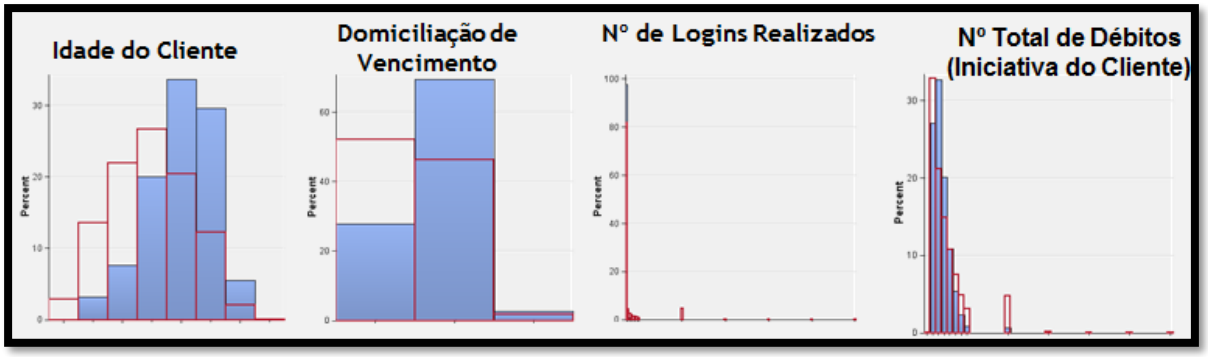


Figura 21 - Pensionistas: Variáveis Características

- **Produtos – *Cluster* 3 – 27%**

Os clientes que pertencem a este *cluster* são os que detêm mais produtos. Têm no banco o crédito imobiliário, vários cartões (possuem mais cartões de débito e/ou crédito do que o resto da população) e/ou seguros. Para além disto, cerca de 84% destes clientes possuem pelo menos uma domiciliação de pagamentos no banco.

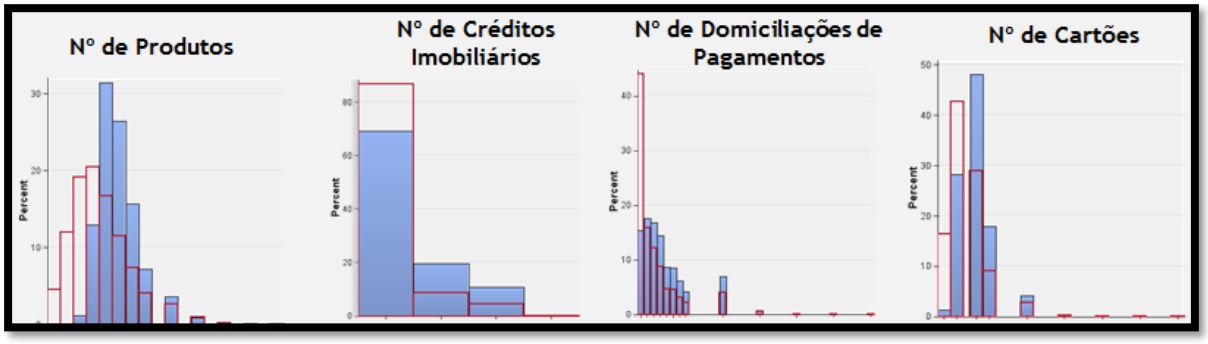


Figura 22 - Produtos: Variáveis Características

- **Transferências – *Cluster* 4 – 13%**

Este *cluster* é representado por clientes que têm mais débitos (transferências para outras contas ou levantamentos) e créditos (depósitos, recebimento de transferências ou ordenado) do que o resto da população, onde todos os clientes deste cluster têm mais do que uma transacção de cada tipo realizada nos últimos 6 meses. Fazem muitas transferências, principalmente para OIC e através de canais digitais, nomeadamente aplicação ou site.

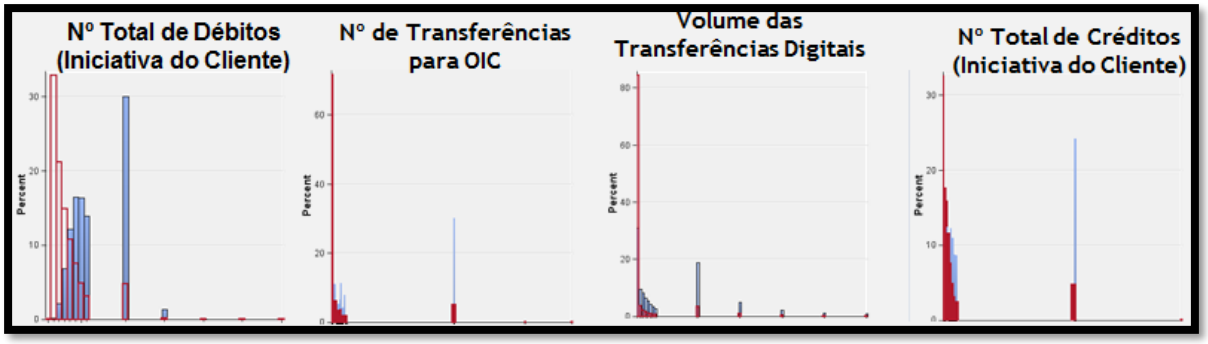


Figura 23 - Transferências: Variáveis Características

6.3. Resultados por Perfil de Cliente

Depois de definidos os dois perfis, considerou-se interessante cruzá-los, tornando o perfil de utilização de serviços bancários no principal perfil e o de compras em POS no secundário, com o objectivo de perceber quais os hábitos de compras (de acordo com os clusters criados) dos clientes de cada um dos segmentos do perfil de utilização de serviços bancários. Para cada um dos segmentos deste perfil foi analisada a distribuição dos clientes pelos macro segmentos existentes no banco (MS, UU e PP) e pela posse de produtos.

6.3.1. Clientes Pouco Activos

Estes clientes, como visto anteriormente, não possuem um grande envolvimento com o banco, e quando cruzado com o perfil de Compras em POS, é maioritariamente constituído (95%) pelos segmentos Zero - clientes que não fazem compras - e Casual - clientes que fazem poucas compras. Este segmento é também formado por 75% de clientes Mass Market, 15% de Plus e apenas 10% de Prestige. Para além disto, é o único segmento que possui clientes menores (cerca de 15%).

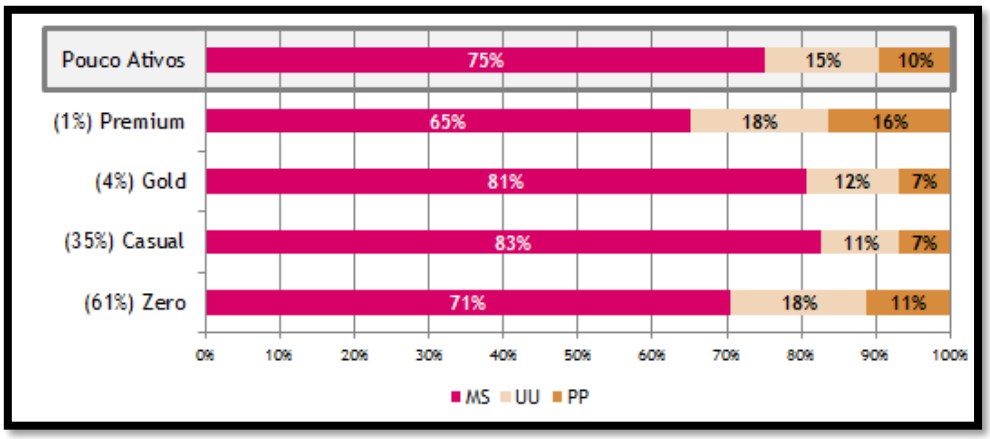


Figura 24 - Pouco Activos: Distribuição por Macro Segmentos

Estes clientes descrevem - se também por possuírem menos produtos comparativamente com o resto dos segmentos, como se pode verificar pela posse de cartões de débito que são o meio de pagamento mais utilizado em Portugal [31], apenas 60% destes clientes têm cartão de débito. Para além disto, dentro deste segmento, pode observar-se que os clientes Zero são os que possuem mais crédito habitação e produtos de poupança e investimento, em relação aos outros perfis de compras.

	Crédito	Crédito ao Consumo	Crédito Habitação	Cartões	Cartão Débito	Cartões de Crédito	Seguros	Poupança	Investimento	Site *
(61%) Zero	12%	3%	6%	41%	37%	9%	12%	36%	17%	26%
(35%) Casual	12%	2%	1%	97%	95%	16%	7%	19%	8%	47%
(4%) Gold	13%	1%	1%	99%	98%	18%	5%	19%	7%	56%
(1%) Premium	24%	2%	1%	100%	95%	30%	4%	22%	10%	69%
Pouco Ativos	**12%**	**2%**	**4%**	**63%**	**60%**	**12%**	**10%**	**29%**	**14%**	**34%**

*Corresponde a ter acedido ao site nos últimos 6 meses.

Figura 25 - Pouco Activos: Posse de Produtos Bancários

6.3.2. Clientes Pensionistas

Os clientes assinalados como Pensionistas possuem maioritariamente idades superiores a 60 anos, e são principalmente dos segmentos Casual e Zero (83%), que não fazem muitas compras, mas o segmento Gold tem mais expressão do que no grupo de clientes anterior (clientes pouco activos). Para além disto, a distribuição dos macro segmentos é mais equilibrada, onde os clientes Plus e Prestige ganham maior representação.

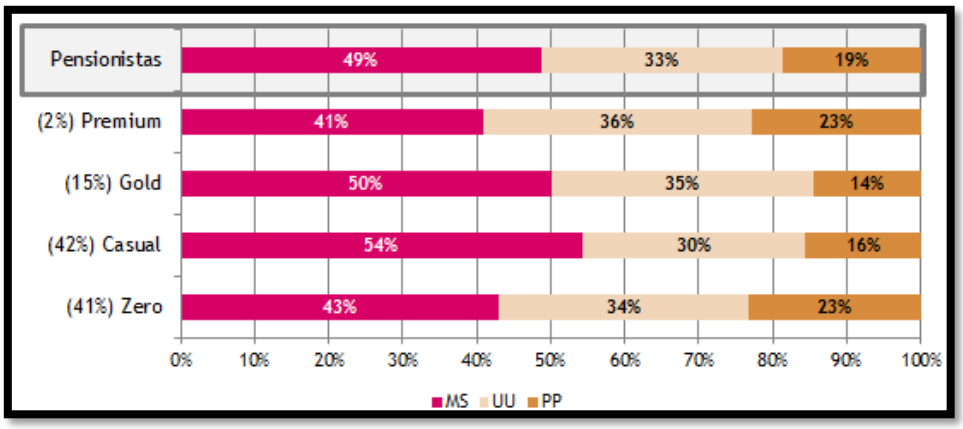


Figura 26 - Pensionistas: Distribuição por Macro Segmentos

Relativamente à posse de produtos, este segmento caracteriza-se por ser o que possui menos produtos de crédito, o que pode derivar da idade avançada destes clientes (menos necessidade, mas também menos cedência de crédito por parte do banco), e ainda por ser o menos digital, visto ter poucos clientes que acedem ao site e aplicação. Por outro lado, este segmento é o que possui mais produtos de poupança.

	Crédito	Crédito ao Consumo	Crédito Habitação	Cartões	Cartão Débito	Cartões de Crédito	Seguros	Poupança	Investimento	Site *
(41%) Zero	7%	1%	2%	64%	57%	20%	23%	47%	30%	15%
(42%) Casual	12%	1%	1%	99%	97%	22%	20%	36%	23%	18%
(15%) Gold	10%	1%	1%	100%	99%	18%	17%	33%	21%	22%
(2%) Premium	15%	0%	1%	100%	99%	23%	14%	32%	20%	32%
Pensionistas	**10%**	**1%**	**2%**	**85%**	**81%**	**21%**	**21%**	**40%**	**26%**	**18%**

*Corresponde a ter acedido ao site nos últimos 6 meses.

Figura 27 - Pensionistas: Posse de Produtos Bancários

6.3.3. Clientes Produtos

Estes clientes provêm principalmente dos segmentos Gold e Casual (77%), que são na maioria Mass Market e Plus, o que mostra que realizam mais compras do que os dois segmentos apresentados anteriormente. Para além disto, têm a idade especialmente compreendida entre 40 e 60 anos.

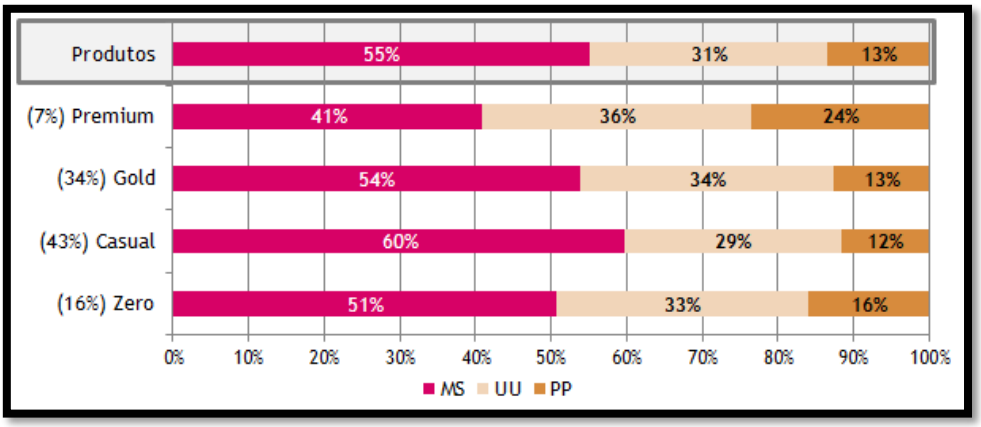


Figura 28 - Produtos - Distribuição por Macro Segmentos

De todos os segmentos do perfil de utilização de serviços bancários, este é o que possui mais crédito na forma de crédito habitação ou crédito ao consumo, onde 67% dos clientes possui pelo menos um produto deste tipo de crédito, destacando de todos os subsegmentos os clientes classificados como Zero, já que 53% possuem crédito habitação.

Dos clientes classificados como Produtos, 66% possui pelo menos um seguro, o que pode derivar da exigência de possuir um produto deste tipo quando se adquire um crédito habitação ou pessoal de maior montante.

	Crédito	Crédito ao Consumo	Crédito Habitação	Cartões	Cartão Débito	Cartões de Crédito	Seguros	Poupança	Investimento	Site *
(16%) Zero	77%	22%	53%	87%	77%	53%	83%	23%	23%	42%
(43%) Casual	69%	25%	31%	99%	97%	60%	68%	22%	18%	52%
(34%) Gold	61%	21%	23%	100%	99%	59%	58%	24%	17%	54%
(7%) Premium	59%	11%	19%	100%	99%	62%	48%	25%	19%	58%
Produtos	**67%**	**22%**	**31%**	**98%**	**95%**	**58%**	**66%**	**23%**	**19%**	**52%**

*Corresponde a ter acedido ao site nos últimos 6 meses.

Figura 29 - Produtos: Posse de Produtos Bancários

6.3.4. Clientes Transferências

Os clientes classificados neste segmento possuem uma distribuição demográfica semelhante aos clientes pertencentes ao *cluster* Produtos, com idades maioritariamente compreendidas entre 40 e 60. Para além disto, é o único segmento que a percentagem de clientes Premium ascende a 43%

(comparando com os outros segmentos, onde o máximo era 7%), levando a que este segmento seja composto principalmente por clientes que utilizam o BCP como primeiro banco, já que fazem muitas compras e transferências – como o próprio nome sugere.

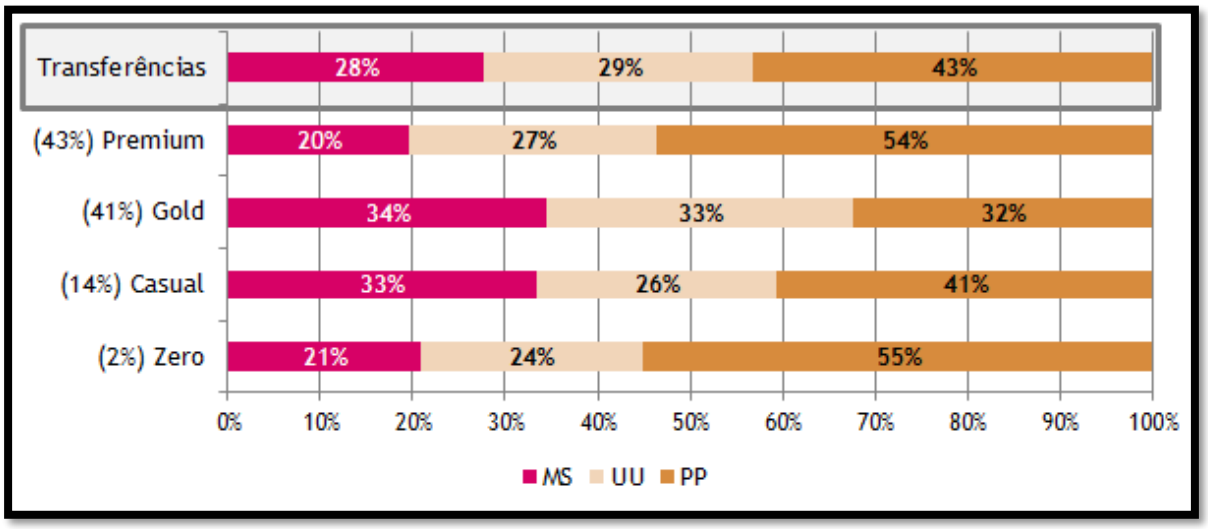


Figura 30 - Transferências: Distribuição por Macro Segmentos

Uma das características deste segmento é ser constituído por clientes que dão primazia ao canal digital, particularidade que se comprova através da **Figura 31**, que mostra que 92% destes clientes utilizaram o *site* nos últimos 6 meses. Para além disto, são o segmento que tem clientes com mais cartões de crédito, onde 74% destes clientes possuem pelo menos um produto deste tipo (factor que pode estar relacionado com a tendência destes clientes para compram mais em número e montante). Por fim, estes clientes são os que mostram maior interesse por produtos de investimento, como acções, certificados ou obrigações, visto que 32% destes clientes possuem pelo menos um produto deste tipo.

	Crédito	Crédito ao Consumo	Crédito Habitação	Cartões	Cartão Débito	Cartões de Crédito	Seguros	Poupança	Investimento	Site *
(2%) Zero	53%	13%	28%	81%	63%	60%	59%	41%	46%	91%
(14%) Casual	68%	18%	23%	99%	95%	72%	57%	40%	37%	94%
(41%) Gold	70%	22%	26%	100%	99%	70%	57%	38%	28%	92%
(43%) Premium	77%	15%	27%	100%	99%	78%	57%	40%	34%	91%
Transferências	**72%**	**18%**	**26%**	**99%**	**98%**	**74%**	**57%**	**39%**	**32%**	**92%**

*Corresponde a ter acedido ao site nos últimos 6 meses.

Figura 31 - Transferências: Posse de Produtos Bancários

7. ANÁLISE DE UTILIZAÇÃO DE SERVIÇOS

Uma das dificuldades que existe quando se gere uma solução integrada é a de perceber se os clientes utilizam as vantagens que são dadas, e consequentemente se existe alguma ou algumas alterações a fazer em termos de produtos incluídos na solução, ou mesmo nos preços praticados. Para além disso, pensou-se ser interessante comunicar aos clientes, por *e-mail*, quanto haviam poupando por deter uma solução integrada, detalhando a utilização de cada individuo nos vários serviços/produtos incluídos. Neste sentido foi elaborada uma análise ao tipo de utilização das vantagens incluídas no Cliente Frequente (solução integrada existente desde 2004), tendo em consideração o valor pago pelo cliente, bem como o valor que seria pago se não detivesse a solução.

O universo considerado inicialmente era constituído pelos clientes cujas contas, em Dezembro de 2017, possuíam a solução integrada Cliente Frequente (CF). Destes clientes analisaram-se apenas os clientes considerados elegíveis, i.e., clientes que não tinham nenhuma *flag* de risco activa, com *e-mail* registado, e que aceitavam e-mails externos, reduzindo o universo de clientes a analisar para 37,7% do inicialmente considerado, como se pode verificar pela **Figura 32**.

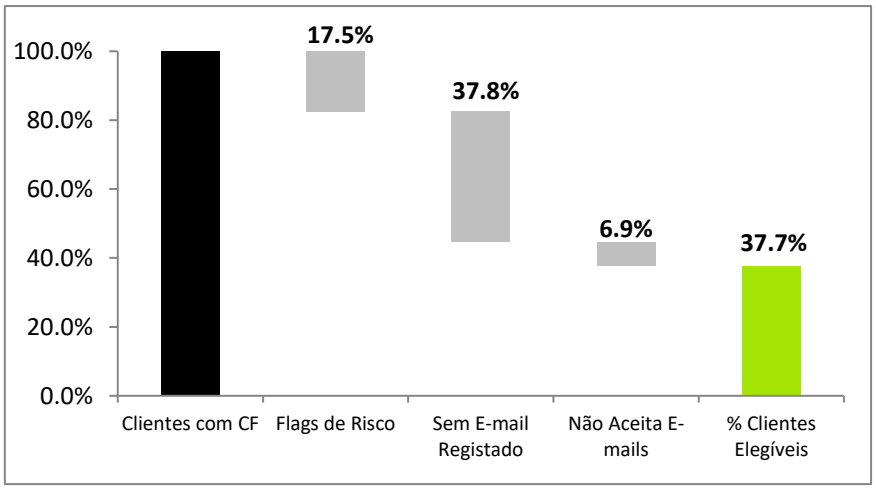
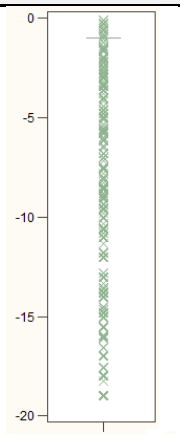
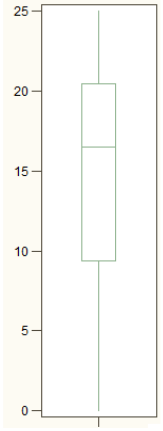
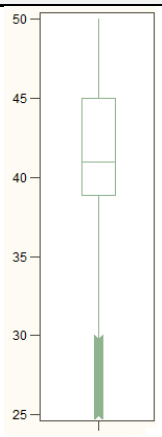
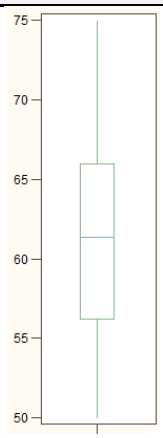
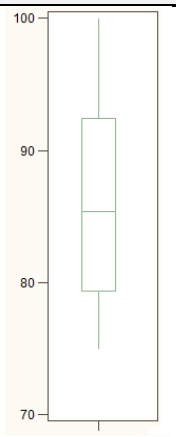
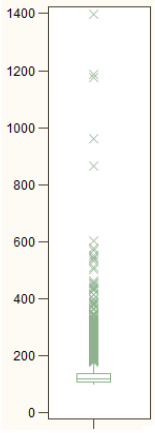


Figura 32 - Universo de clientes com Cliente Frequente em Dezembro de 2017

As contas dos clientes elegíveis foram distribuídas por 6 grupos diferentes, de acordo com o valor absoluto que haviam poupado com a solução integrada no decorrer do ano de 2017, onde o grupo 1 representa os clientes que não pouparam por ter a solução, e restantes 5 grupos representam os clientes que o fizeram, tendo valores de poupança diferentes, com um intervalo de 25€. O objectivo desta divisão foi tentar potenciar uma diferenciação nos *e-mails* enviados aos clientes após a análise,

onde um cliente que tenha poupado 100€ teria um corpo de texto (ou outro aspecto) diferente de outro cliente que tenha poupado apenas 10€, por exemplo.

Tabela 12 - Grupos de clientes por valor poupado com a SI

Grupos	Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5	Grupo 6
Intervalos da poupança	[-19€; 0€]	]0€; 25€[	[25€; 50€[	[50€;75€[	[75€; 100€[	[100€;1398,10€]
Total de contas (%)	1.9%	5.7%	25.3%	38.7%	18.5%	9.8%
Distribuição do valor poupado (€)						

Através da Tabela 12, pode perceber-se a distribuição de cada grupo pelo intervalo a que pertence, onde no grupo 1, a maioria dos indivíduos perdeu 1.5€ por possuir a solução integrada, enquanto nos grupos 2, 4 e 5 houve uma distribuição mais equilibrada entre os extremos do intervalo a que pertencem. No grupo 3, houve uma maior acumulação de clientes entre os 39€ e os 45€, e por último no grupo 6 (representativo das pessoas que pouparam valores superiores a 100€ por possuírem a solução), a globalidade dos indivíduos poupou entre 103€ e 130€.

7.1. Análise por Produto

Nesta solução, ao se pagar a comissão mensal, existem vários serviços a que o cliente tem acesso. Estes serviços são: a comissão de manutenção da conta, bem como a de, no máximo, três contas filhas; a anuidade de um cartão de débito e um de crédito por titular de conta (no máximo dois

titulares); transferências cujo valor da comissão é no máximo 1€[13]; um módulo de cinco cheques mensais e dois seguros - seguro de assistência na urgência médica ao domicílio e seguro de responsabilidade civil familiar.

7.1.1. Comissão da Solução Integrada

Neste produto, em cada mês podem ser pagas duas comissões diferentes: 4.5€ no caso do cliente domiciliar o ordenado e 8€ caso contrário. Habitualmente, nos primeiros 3 ou 6 meses de aquisição desta solução, os clientes estão isentos desta comissão, como oferta de boas-vindas. Considerando um ano completo, se um cliente tiver o ordenado domiciliado, ao longo dos 12 meses pagou 54€, e por outro lado um que não o tenha pagou 96€.

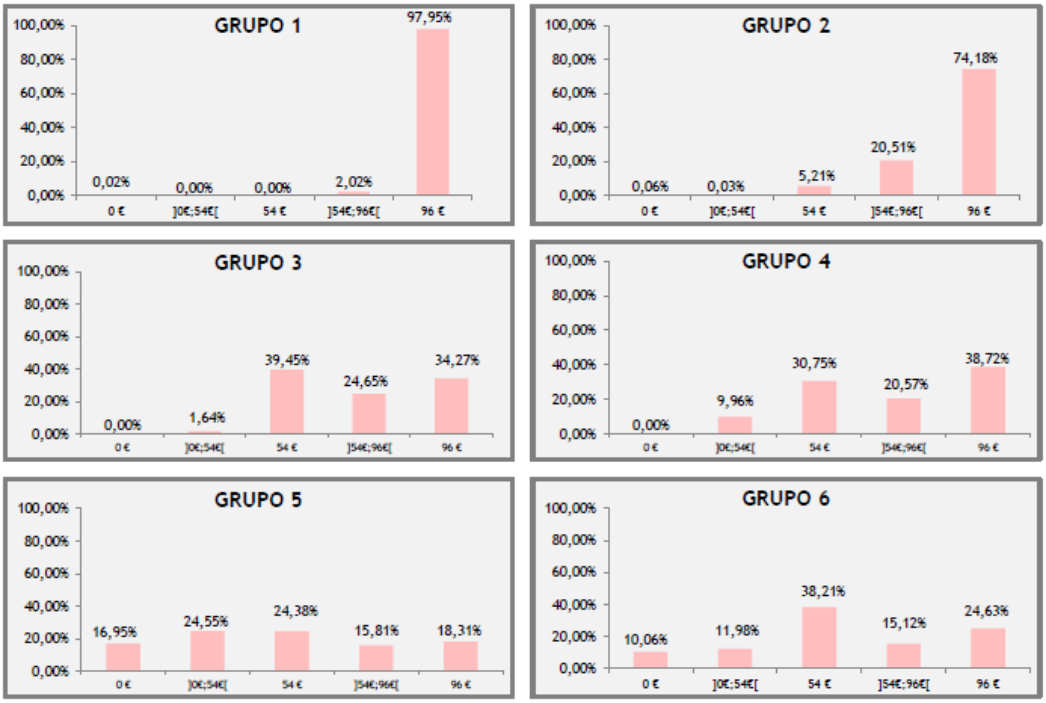


Figura 33 – Distribuição do Valor Acumulado da Comissão da Solução no ano de 2017

Através da Figura 33, pode-se concluir que os clientes que menos pouparam em 2017, pagaram a totalidade da solução, na sua maioria. Por outro lado, alguns clientes pertencentes aos grupos 5 e 6 estiveram isentos de pagar a comissão e por isto pode-se concluir que são clientes que estão no máximo há 6 meses no banco, o que contribuiu para terem poupado aparentemente mais.

[13] Ver no anexo 11.1 - Excerto do preçário das comissões de transferências em vigor, desde 5 de Março de 2018

7.1.2. Comissão de Manutenção da Conta e das Contas Filhas

A comissão de manutenção de conta deveria ser recebida mensalmente por conta existente no banco, mas existem 24 razões que levam a que este valor não seja cobrado, para além de possuir uma solução integrada, como por exemplo possuir património em Macau, ou ter a reforma domiciliada.

Todos os meses são verificadas as condições que poderiam isentar cada cliente desta comissão, incluindo a posse de uma solução integrada, neste caso o Cliente Frequente. Na figura seguinte pode-se observar a distribuição do valor que deveria ter sido cobrado em 2017, por grupo, caso cada cliente não possuísse a solução em causa.

Figura 34 - Distribuição da comissão de manutenção que seria cobrado em 2017

Pela figura acima representada, pode-se concluir que a maioria dos clientes não pagaria a comissão, mesmo que não possuísse esta solução, principalmente os clientes do grupo 1. Isto pode significar que se se quiser valorizar mais esta solução, as condições de isenção devem ser revistas.

Esta solução, para além de isentar a conta associada, possibilita a isenção da comissão de manutenção de até 3 contas filhas. Porém quando a solução foi criada não existia um número limite de contas filhas que se podia isentar, o que permitiu a que clientes que adquiriram o CF antes da

limitação de associações de contas filhas usufruíssem e continuem a usufruir da isenção de todas as contas.

Através da figura em baixo, pode-se perceber que este factor pode ter um peso muito grande no total poupado, visto que o valor desta comissão pode ascender a 411,60€ (valor que não cobrado por possuir a solução).

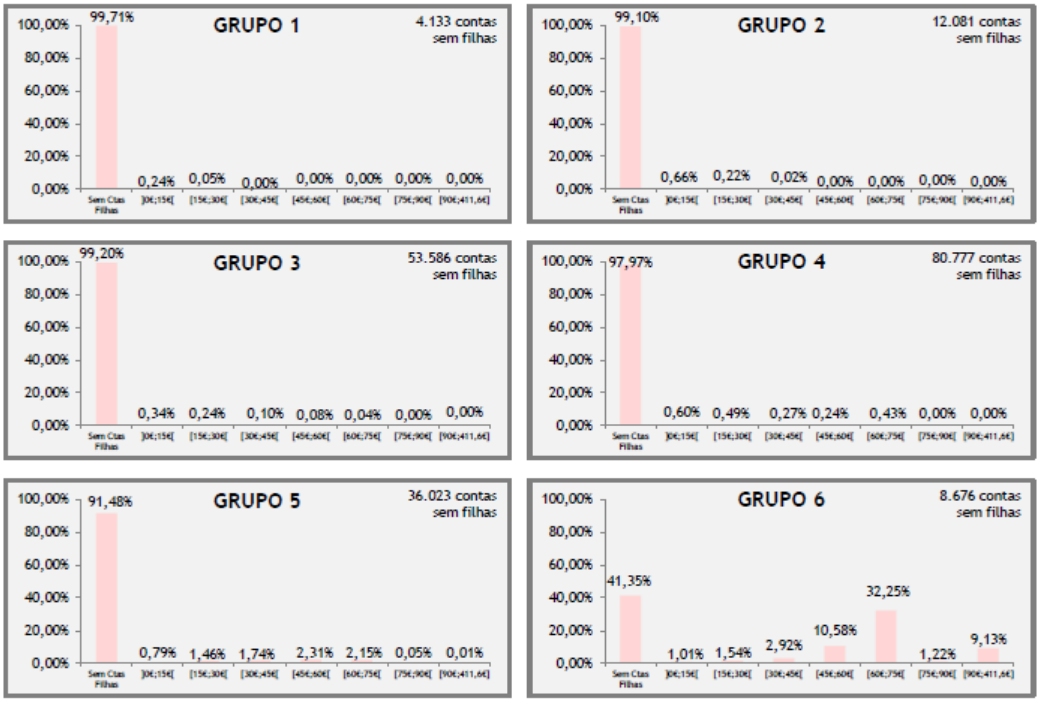


Figura 35 - Distribuição da comissão de manutenção das contas filhas que seria cobrado em 2017

Apenas 8% dos clientes com Cliente Frequente possuem pelo menos uma conta filha associada, e destes, somente 0.5% possuem mais do que três contas filhas associadas (clientes que pertencem todos ao grupo 6).

7.1.3. Requisição de Cheques

O pagamento através de cheques está a cair cada vez mais em desuso, já que em 2017 representou apenas 1.3% do volume de pagamentos [31]. A solução inclui uma requisição de um módulo de 5 cheques mensais, que teria um custo de 6.25€ por cada módulo requisitado. Apesar desta oferta, somente 0.4% dos clientes com CF requisitaram este meio de pagamento pelo menos uma vez, e destes apenas 2.2% requisitaram em dois meses diferentes, como se pode verificar pela figura

seguinte. Percebemos também que não existe nenhuma diferença significativa entre os vários grupos estabelecidos.

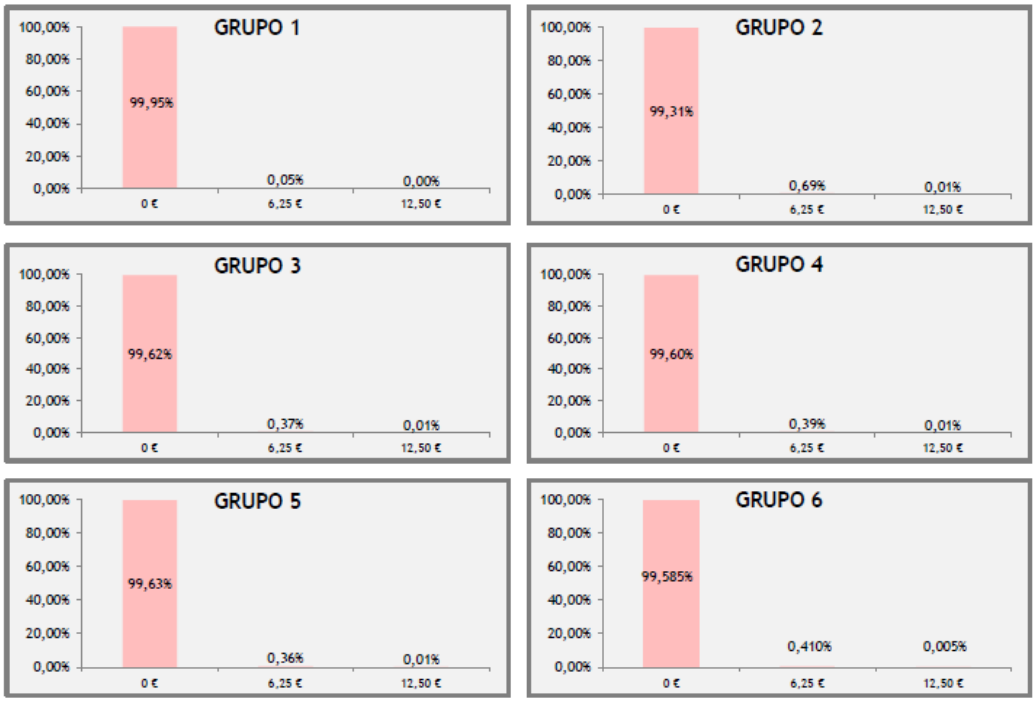


Figura 36 - Distribuição da comissão da requisição de cheques que seria cobrado em 2017

7.1.4. Anuidade de Cartões de Débito

Os cartões de débito são o meio de pagamento mais utilizado atualmente em Portugal [31], com 14.6 milhões de cartões de débito registados na rede Multibanco. A solução integrada em estudo isenta até dois cartões de débito - um cartão por cada titular da conta - onde cada cartão tem uma anuidade associada de 18€. De todos os clientes com CF, existem 14% que não possuem nenhum cartão de débito, e destes apenas 30% são elegíveis para contactar - e por isso representados na análise, e na figura seguinte:

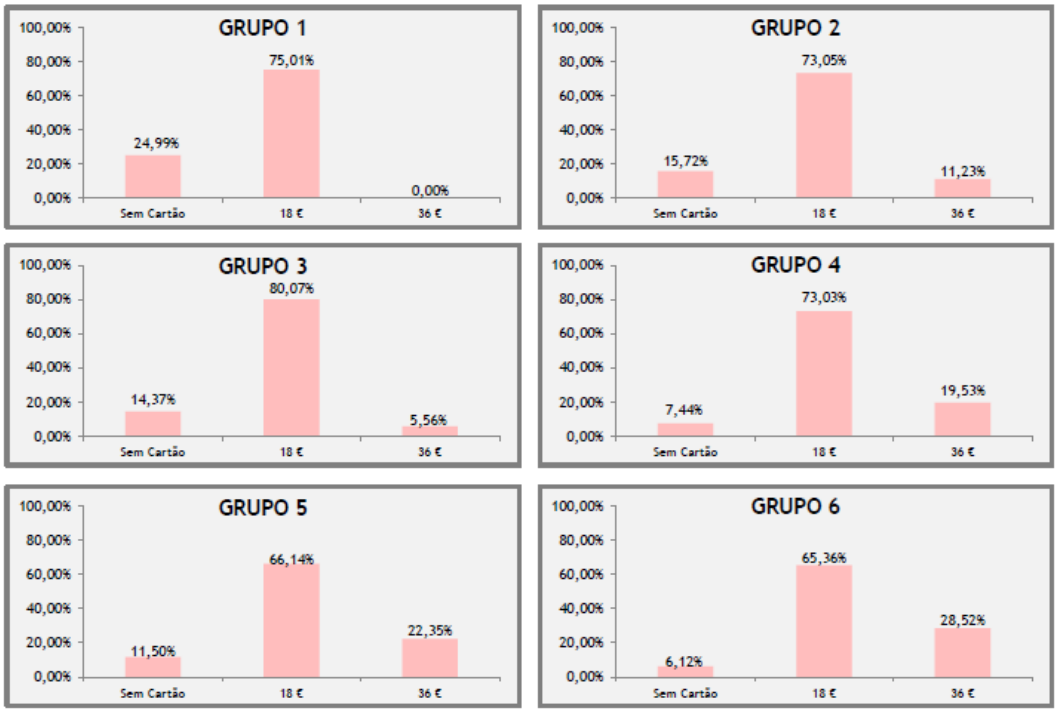


Figura 37 - Distribuição da anuidade dos cartões de débito que seria cobrada em 2017

Através dos gráficos acima representados, observa-se que 10.6% dos clientes com a solução (e elegíveis para contactar) não possuem cartão de débito, no entanto, a maioria destes clientes utilizam o cartão de crédito ou a sucursal como meios de pagamento. Para além disto, nota-se uma tendência crescente para cada conta ter dois cartões isentos (e por isso dois titulares), passando de 0.00% no grupo 1 para 28.52% no grupo 6.

7.1.5. Anuidade de Cartões de Crédito

O Cliente Frequente inclui a anuidade de dois cartões de crédito – um por titular- que têm valores diferentes, de acordo com o cartão escolhido (dentro de três possibilidades), variando entre 0€ e 25€. No decorrer desta análise, foi notado que estavam a ser isentos da anuidade, todos os cartões de crédito que os clientes com a solução possuíam, levando a uma perda que ascendeu a cerca de 650.000€, em 2017.

Observando a **Figura 38**, conclui-se que 55% das contas com Cliente Frequente não possuem cartão de crédito, muitas vezes derivado da falta de conhecimento do lado do consumidor. Para além disto, nota-se que existem muitos clientes que não pagariam a anuidade, mesmo que não possuíssem a

solução, que pode derivar de três situações: a anuidade do cartão ser 0€, realizar compras para obter descontos na anuidade até alcançar o valor desta, ou possuir um crédito especial, sendo que a maioria dos clientes derivam da última razão enunciada.

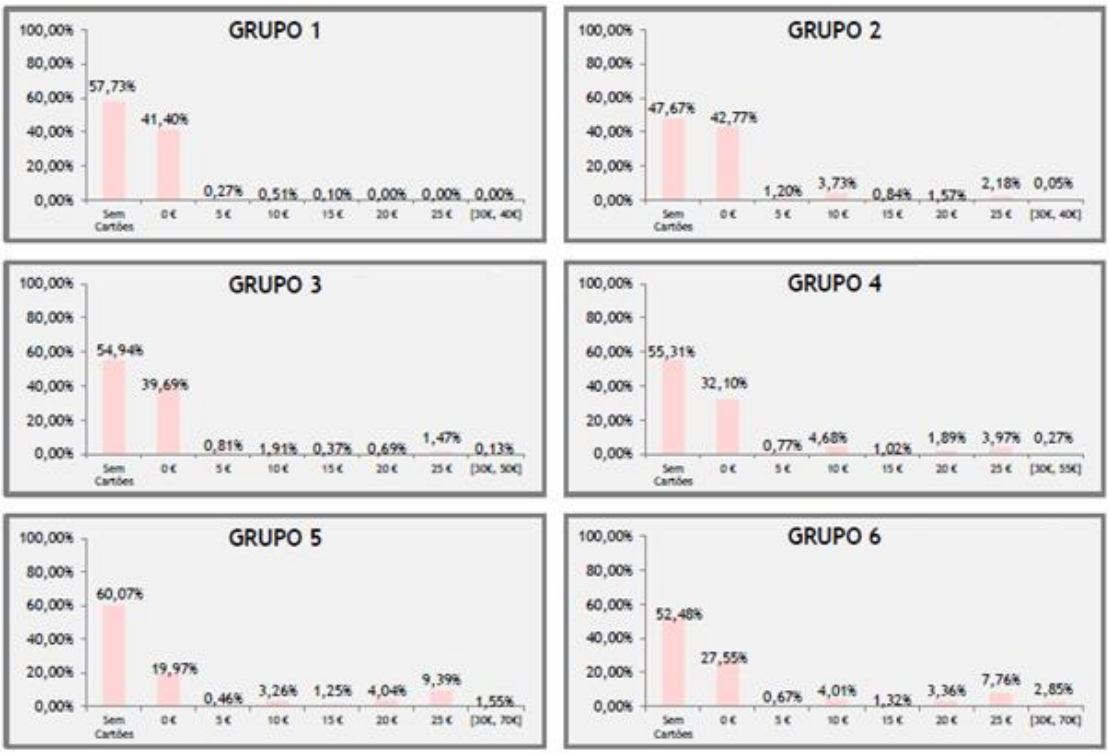


Figura 38 - Distribuição da anuidade de cartões de crédito que seria cobrado em 2017

7.1.6. Comissão de Transferências

Todas as transferências realizadas em ATM estão isentas de comissão, porém existem vários canais onde esta comissão existe quando se realiza uma transferência, nomeadamente Internet, Mobile, Balcão e Telefone, como se pode verificar pelo anexo **11.1**. O CF isenta todas as transferências cuja comissão associada seja até 1€. Cerca de 61% das contas em análise não realizaram qualquer transferência em 2017, porém cerca de 2% dos clientes com CF, com as operações realizadas em 2017, chegaram a poupar entre 500€ e 1355€ e pertencem ao grupo 6, da figura abaixo.

Figura 39 - Distribuição da comissão de transferências que seriam cobradas em 2017

7.2. VISÃO TRANSVERSAL

Relativamente à comissão de manutenção, a maioria do Grupo 1 seria isento por outros motivos que não a posse da SI, e por isso levou a que poupassem menos, por outro lado os clientes dos grupos 4, 5 e 6 pouparam em média 26,81€ nesta comissão por estarem isentos pela solução.

Os clientes do grupo 6 pouparam em média 40.49€ na comissão de manutenção das contas filhas, ao contrário de todos os outros que possuíam valores próximos de 0€, (pois alguns destes clientes deste grupo possuem mais do que três contas filhas).

Por fim, em relação à comissão de transferências, os clientes do grupo 6 são os que mais beneficiaram desta oferta, poupando em média 13.82€, ao longo de 2017, o que indicia que são clientes que utilizam o banco nas suas operações do dia-a-dia.

8. CONCLUSÕES

Este relatório tinha como objectivo expor alguns dos projectos realizados durante o decorrer do estágio no CRM do Millennium BCP. Sendo a gestão da relação com os clientes uma área que está a ganhar muita importância em qualquer empresa por poder dar insights relevantes ao negócio, foi muito interessante perceber o que um banco como o Millennium está a desenvolver nesta área.

Todos os projectos desenvolvidos tiveram como foco os clientes e como objectivo perceber o que os clientes querem, através da criação do modelo de propensão à compra do CFN, quem os clientes são e o que fazem dentro e fora do banco através da segmentação comportamental, e ainda perceber como utilizam os serviços disponibilizados pelo banco através da análise da utilização de serviços.

O modelo de propensão criado tinha como objectivo aumentar a taxa de conversão dos contactos efectuados, através a aplicação de métodos analíticos avançados em dados históricos, para se conseguir prever a propensão de cada cliente adquirir o CFN no futuro.

No desenvolvimento do modelo foram experimentados os algoritmos sugeridos pela literatura científica, nomeadamente árvores de decisão, redes neuronais e regressões lineares. As árvores de decisão não revelaram produzir tão bons resultados quanto os outros dois algoritmos. A combinação dos resultados da rede neuronal e das regressões logísticas num único modelo – modelo ensemble - produziu os resultados mais precisos, com melhor poder preditivo.

O modelo ensemble, que foi considerado como o melhor modelo, conseguiu aumentar o número de vendas, tornando os contactos efectuados mais eficazes, através da selecção dos clientes melhor classificados pelo modelo, ou seja, os que estavam classificados acima do percentil 50.

Para a criação da segmentação comportamental, através do estudo do comportamento dos clientes, criaram-se dois perfis diferentes: Compras em POS e o de Utilização de Produtos Bancários. Estes dois perfis permitem ao banco perceber que tipos de clientes existem na sua base dando duas visões diferentes, mas que podem ser complementares.

No perfil de compras em POS, destacam-se os segmentos Gold e Premium como os que mais gastam, e em diversas categorias. Os clientes Zero no princípio da análise pareciam ser clientes com grandes probabilidades de abandonar o banco (clientes Churn), porque não fazem compras, mas com o cruzamento dos dois perfis deu a entender que afinal são clientes que ou são menores e têm as suas

contas poupança no banco, e daí não fazerem compras, ou então são clientes que possuem pelo menos um crédito habitação no banco, sem mais nenhum envolvimento e consequentemente não realizam nenhuma compra. Com a criação deste perfil o banco pode conseguir criar parcerias mais direccionadas e oferecer aos clientes vantagens e promoções nas categorias que cada perfil gasta mais (como por exemplo oferecer aos clientes Casual promoções em supermercados para aumentar a utilização dos cartões do banco) com o objectivo de elevar o envolvimento com o banco.

No perfil de utilização de serviços bancários existem 4 tipos de clientes, que foram classificados como Pouco Activos, Pensionistas, Produtos e Transferências, de acordo com a sua actividade/ comportamento no banco, bem como os produtos detidos por cada cliente. Com este perfil, o banco consegue direccionar melhor as campanhas de marketing existentes, oferecendo a cada perfil de clientes uma proposta com mais valor e personalizada, levando a uma fidelização maior.

Através da análise da utilização de serviços bancários, conseguiu-se perceber melhor os padrões de utilização dos serviços incluídos no CF. De todas as vantagens incluídas, a mais residual é a Requisição de Cheques, já que é utilizada por apenas 0.4% dos clientes considerados na análise. Para além disto, as características que se revelaram mais impactantes nas diferenças do valor total poupado em 2017 foram: a comissão de manutenção da conta e das contas filhas, e a comissão das transferências.

9. LIMITAÇÕES E RECOMENDAÇÕES PARA TRABALHOS FUTUROS

Qualquer um dos projectos desenvolvidos pode ser melhorado e aperfeiçoado, desde o desenvolvimento de cada um à implementação e acompanhamento dos resultados obtidos.

Tanto no desenvolvimento do projecto do modelo de propensão como no projecto da segmentação comportamental, uma das grandes limitações foi a pouca capacidade da versão do SAS Enterprise Miner usada em lidar com grandes volumes de dados, levando à diminuição substancial do número de clientes utilizados para treinar o modelo de propensão como também levando à possível diminuição da representatividade dos segmentos dos perfis obtidos.

Em relação ao modelo criado, apesar de estar a gerar bons resultados, é importante monitorizar o desempenho do modelo. Ao longo do tempo é normal que o poder preditivo do modelo diminua, e por isso é importante avaliar periodicamente várias dimensões, nomeadamente a estabilidade da população e das variáveis, o valor da informação do modelo e das variáveis, a performance do modelo através, por exemplo, das métricas ROC e Gini e validar o Lift do modelo existente.

No desenvolvimento da segmentação comportamental foram consideradas seis fontes de informação diferentes (Demografia, Relação, Segmentação, Posse Actual, Transacionalidade e Compras em POS). Ao considerar apenas as compras realizadas em POS corre-se o risco de estas não serem representativas do total de compras efectuadas pelos clientes, já que os canais digitais têm cada vez mais expressão. Para além disto, com um mundo cada vez mais digital, incluir informações de redes sociais ou pesquisas na internet seriam bons *inputs* para incluir numa futura segmentação.

Para a criação dos segmentos foi utilizado o K-Means, e devido à limitação de tempo, técnicas como a análise de componentes principais (PCA) ou SOM Kohonen não foram aplicadas, por não se ter investigado o suficiente sobre estes algoritmos e por isso deixo aqui como sugestão de trabalho futuro.

Em relação à análise de utilização de serviços incluídos no CF seria interessante desenvolver também uma análise de *clusters* para se conseguir perceber que perfis-tipo de utilização existem.

Contudo, com a análise realizada, conclui—se que existem clientes que não pouparam nada ou muito pouco, e tendo esta análise o objectivo de comunicar a estes clientes quanto pouparam por deter a solução, na abordagem através de *e-mail* para mostrar a poupança de cada cliente por possuir o

Cliente Frequente, para além de incluir os valores actuais de cada vantagem usufruída, também se poderia incluir, para os grupos que menos pouparam, uma previsão destes valores no caso da utilização ser superior ou alguns incentivos para aproveitar melhor a solução e consequentemente aumentar o envolvimento com o banco.

10.BIBLIOGRAFIA

[1] INDRA, “From the Tradiotional Banking System to the customer-centric financial ecosystem,” 2014.

[2] K. Heinonen, “Multiple perspectives on customer relantionships,” *International Journal of Bank Marketing,* vol. 32, n.º 6, pp. 450-456, 2014.

[3] Dicionário infopédia da Língua Portuguesa, Porto: Porto Editora, 2003 - 2018.

[4] R. T. Domingo, “Apllying Data Mining to Banking - Business Management Articles,” 2003. [Online]. Available: http://www.rtdonline.com/BMA/BSM/4.html. [Acedido em Novembro 2018].

[5] J. Kim, E. Suh e H. Hwang, “A model for evaluating the effectiveness of CRM using the balanced scorecard,” *Journal of Interactive Marketing,* vol. 17, n.º 2, pp. 5-19, 2003.

[6] D. Jutla, J. Craig e P. Bodorik, “Enabling and Measuring Electronic Customer Relantionship Management Readiness,” em *Proceedings of the 34th Hawaii International Conference on System Sciences*, 2001.

[7] M. Stone, N. Woodcock e M. Wilson, “Managing the Change from MArketing Planning to Customer Relationship Management,” *Long Rnage Planning,* vol. 29, n.º 5, pp. 675-683, 1996.

[8] J. Naisbitt, Megatrends, Warner Books, 1982.

[9] J. Han, Data Mining: Concepts and Techniques, Illinois: University of Illinois at Urbana - Champaign, 2006.

[10] E. Jyoti e E. A. S. Walia, “A Review on Recommendatyion System and Web Usage Data Mining using K-Nearest Neighbor(KNN) method,” *International Research Journal of Engineering and Technology,* vol. 5, n.º 4, pp. 2931-2934, 2017.

[11] D. A. Adeniyi, Z. Wei e Y. Yongquan, “Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method,” *Applied COmputing and Informatics,* vol. 12, pp. 90-108, 2014.

[12] D. J. Hand, “Principles of Data Mining,” *Drug Safety,* vol. 30, n.º 7, pp. 621-622, 2007.

[13] L. J. Mester, “What's the point of credit scoring?,” *Business Review,* vol. 3, pp. 3-16, 1997.

[14] S. McKechnie, “Integrating Intelligent Systems into MArketing to Support Market Segmentation Decisions,” *Intelligent Systems in Accounting, Finance and Management,* vol. 14, n.º 3, pp. 117-127, 2006.

[15] P. Kotler e G. Armstrong, Princípios do Marketing, Harlow: Pearson, 2012.

[16] J. Brownlee, “How to Use Correlation to Understand the Relationship Between Variables,” 27 Abril 2018. [Online]. Available: https://machinelearningmastery.com/how-to-use-correlation-to-understand-the-relationship-between-variables/. [Acedido em 7 Outubro 2018].

[17] S. Siegel e N. Castellan, Non Parametric statistics for the behavioral sciences, McGraw Hill, 1988.

[18] D. L. Olson, *Data Set Balancing,* University of Nebraska, USA.

[19] T. Hoens e N. Chawla, Imbalanced Datasets: From Sampling to Classifiers, Haibo He & Yunqian Ma, 2013.

[20] T. H. Davenport e D. McNeill, Analytics in Healthcare and the Life of Sciences: Stratagies, Implementation Methods, and Best Practices, International Institute for Analytics, 2017.

[21] M. A. Brookhart, S. Scnneeweiss, K. J. Rothman, R. J. Glynn, J. Avorn e T. Sturmer, “Variable Selection for Propensity Score Models,” *American Journal of Epidemiology,* vol. 163, n.º 12, pp. 1149-1156, 2006.

[22] K. S. Sarma, Variable Selection and Tranaformation of Variables in SAS Enterprise Miner 5.2, Ecostat Research Corp. , White Plains NY, 2007.

[23] M. J. A. Berry e G. S. Linoff, Data Mining Techniques - For MArketing, Sales and Customer Relantionship Management, Indiana: Wiley Publishing, Inc., 2004.

[24] P. L. Galinha, “Data Mining no Turismo em Portugal,” Nova IMS, Lisboa, 2017.

[25] T. M. Mitchell, “Chapter 3,” em *Machine Learning*, McGraw-Hill, 1997.

[26] P. D. Santomil, L. O. González, O. M. Cunill e J. M. M. Lindahl, “Backtesting an equity risk model under Solvency II,” *Journal of Business Research,* vol. 89, pp. 216-222, 2018.

[27] L. Kaufman e P. J. Rousseeuw, Fiding groups in data: an introduction to cluster analysis, Hoboken: John Wiley & Sons, 2005.

[28] P.-N. Tan, M. Steinbach, A. Karpatne e V. Kumar, “Cluster Analysis: Basic Concepts and Algorithms,” em *Introduction to Data Mining*, 2018, pp. 487 - 500.

[29] P. Sneath e R. Sokal, Numerical Taxonomy, San Francisco: Freeman, 1973.

[30] H. Romerburg, Cluster Analysis for Researchers, Belmont: CA: Lifetime Learnings Publications, 1984.

[31] B. d. Portugal, “Relatório dos Sistemas de Pagamentos,” 2017.

11.ANEXOS

11.1. EXCERTO DO PREÇÁRIO DAS COMISSÕES DE TRANSFERÊNCIAS EM VIGOR, DESDE 5 DE MARÇO DE 2018

Disponível em https://ind.millenniumbcp.pt/pt/info/Pages/precario.aspx

BANCO COMERCIAL PORTUGUÊS, S.A. Entrada em vigor: 05-mar-2018

5. TRANSFERÊNCIAS (PARTICULARES) (ÍNDICE)

5.1. Ordens de transferência

Escalões		Canal de recepção da ordem de Transferência						Outras condições
		Balcão	Telefone C/ operador	Telefone S/ Operador	Internet e Mobile	ATM	Maq. Rede Interna	
1. Transferências Internas / Nacionais - Emitidas em euros								
1.1 - Para conta domiciliada na própria Instituição de Crédito								
- com o mesmo ordenante e beneficiário								
Pontuais ou Data Futura	Qualquer montante	1,70 €	1,70 €	Grátis	Grátis	Grátis	Grátis	
Permanentes		1,00 €	1,00 €	Grátis	Grátis	n/a	n/a	
- com ordenante e beneficiário distintos								Nota (8)
Pontuais ou Data Futura	Qualquer montante	1,70 €	1,70 €	0,35 €	Grátis	Grátis	Grátis	
Permanentes		1,00 €	1,00 €	0,30 €	Grátis	n/a	n/a	
1.2 - Para conta domiciliada noutra Instituição de Crédito								
- Normais								
- Com indicação de IBAN / NIB								
Pontuais ou Data Futura	Até 1.000 Euros	5,30 €	5,30 €	1,00 €	1,00 €	Grátis	Grátis	
	De 1.000,01 Euros a 50.000 Euros	5,75 €	5,75 €	1,45 €	1,45 €	Grátis	Grátis	
	De 50.000,01 Euros a 99.999,99 Euros	7,00 €	7,00 €	1,70 €	1,70 €	n/a	n/a	
	Igual ou Superior a 100.000 Euros	22,50 €	22,50 €	19,50 €	19,50 €	n/a	n/a	
Permanentes	Até 1.000 Euros	4,70 €	4,70 €	0,75 €	0,75 €	n/a	n/a	
	De 1.000,01 Euros a 50.000 Euros	5,20 €	5,20 €	1,00 €	1,00 €	n/a	n/a	
	De 50.000,01 Euros a 99.999,99 Euros	6,70 €	6,70 €	1,45 €	1,45 €	n/a	n/a	
	Igual ou Superior a 100.000 Euros	22,50 €	22,50 €	19,50 €	19,50 €	n/a	n/a	
- Sem indicação de IBAN								
Pontuais ou Data Futura / Permanentes	Qualquer montante	31,25 €				n/a	n/a	
- Urgentes								
Com indicação de IBAN	Qualquer montante	Acresce 19,00 € ao preço da ordem				n/a	n/a	
Sem indicação de IBAN	Qualquer montante	Acresce 19,00 € ao preço da ordem				n/a	n/a	
1.3 - Transferências MB WAY								
- Para conta domiciliada na própria Instituição de Crédito	Até 750,00 €	n/a	n/a	n/a	Grátis	n/a	n/a	Nota (9)
- Para conta domiciliada noutra Instituição de Crédito	Até 750,00 €	n/a	n/a	n/a	1,30 €	n/a	n/a	
Acresce Imposto	Acresce Imposto do Selo à taxa de 4%							

11.2. PROJECTOS DESENVOLVIDOS AO LONGO DO ESTÁGIO