Departamento de Ciências e Tecnologias de Informação

Optimização da Gestão de Contactos via Técnicas de *Business Intelligence*: aplicação na banca

Sérgio Miguel Carneiro Moro

Dissertação submetida como requisito parcial para obtenção do grau de

Mestre em Gestão de Sistemas de Informação

Orientador:
Doutor Paulo Cortez, Professor Associado,
Universidade do Minho

Co-orientador:
Doutor Raul M. S. Laureano, Professor Auxiliar,
ISCTE – Instituto Universitário de Lisboa

Setembro, 2011

Resumo

Com a massificação de campanhas publicitárias é cada vez mais reduzido o efeito que as mesmas têm sobre o público-alvo, pelo que os responsáveis de *Marketing* têm cada vez mais apostado em campanhas direccionadas. Desta forma, a área de *Business Intelligence* reveste-se de um enorme potencial com vista à melhoria da selecção de contactos a efectuar. Em particular, realçam-se as técnicas de *Data Mining*, a partir das quais se pode extrair conhecimento útil a partir de dados não tratados. Devido a pressões externas, como a crise financeira internacional, e à concorrência interna, as instituições financeiras portuguesas têm apostado no *Marketing* direccionado, optimizando as suas campanhas de forma a aumentar a sua eficiência.

Assim, esta dissertação irá focar-se num estudo de caso de uma instituição bancária nacional, tendo em conta dados de campanhas de subscrição de depósitos a prazo efectuadas entre 2008 e 2010, os quais serão utilizados na aplicação de técnicas de *Data Mining* para a optimização dessas campanhas. Da investigação resulta um modelo explicativo da evolução das campanhas, nomeadamente na sua capacidade de previsão do sucesso dos contactos. Desta forma, é possível extrair conhecimento útil e que poderá suportar decisões de negócio pelos gestores, que poderão assim conceber campanhas que beneficiem das características identificadas pelo modelo.

Palavras-Chave: *Marketing* Direccionado, *Business Intelligence*, *Data Mining*, Gestão de Contactos, Gestão de Campanhas, CRISP-DM.

Abstract

The increasingly vast number of marketing campaigns over time has reduced its effect on the general public. That led the marketing managers to invest on directed campaigns, which have been enhanced strongly through Business Intelligence techniques to help select the best set of available contacts for each campaign. In particular, by using Data Mining techniques which allow to extract knowledge from raw data.

Furthermore, economical pressures and competition has led marketing managers to invest on directed campaigns and its optimization to increase efficiency. Having this information into account, the present dissertation will focus on a case-study about a Portuguese financial institution, by using its data about directed marketing campaigns of long-term deposits subscription (which were executed between 2008 and 2010) and applying Data Mining techniques with a goal on the optimization of future similar campaigns. From this research, an explanatory model is conceived that can, with a good precision, predict success in contacts (subscription of the deposit). Valuable knowledge can be extracted from this model in the form of characteristics that can be used to benefit future similar campaigns.

Keywords: Directed Marketing, Business Intelligence, Data Mining, Contact Management, Campaign Management, CRISP-DM.

Agradecimentos

Findo este trabalho, importa, acima de tudo, agradecer ao Professor Paulo Cortez, por todo o apoio e disponibilidade prestados ao longo do ano lectivo, tendo inclusive permitido que assistisse novamente às suas aulas de *Business Intelligence* e observasse o entusiasmo dos colegas do ano seguinte com o método activo de ensino. É de realçar o empenho demonstrado ao responder sempre às minhas solicitações inclusive numa altura em que a prioridade da sua vida é a alegria de voltar a ser pai, facto pelo qual lhe dou oficialmente os parabéns.

Para o Professor Raul Laureano fica também uma palavra de apreço para toda a ajuda e conselhos preciosos, muitos deles discutidos em saudáveis conversas através de *Internet Messenger*. O seu entusiasmo contagiante, sempre com novas ideias, foi determinante como incentivo a prosseguir na investigação.

Adicionalmente, importa referir todo o apoio da parte da instituição cujos quadros de empregados integro, em especial o responsável da equipa a que pertenço, o André Padrão, e os responsáveis da Direcção de Sistemas de Informação, Dr. José Freitas, Dr. José Cabeças e Eng.º Mário Rodrigues, que consideraram este mestrado como sendo relevante e em linha com os objectivos da instituição, e ainda dos colegas Hugo Carrilho, Nuno Pais, Natércia Rodrigues e Ricardo Viana, da Direcção de *Marketing*, que se disponibilizaram prontamente para me fornecer os dados à medida que iam sendo necessários.

A nível pessoal, agradeço aos meus pais e avó, pelo incentivo, e ao meu irmão Artur que, como doutorado e investigador na Universidade Nova de Lisboa (ainda que noutra área), me foi transmitindo alguma da sua experiência académica. Peço ainda desculpas às duas mulheres da minha vida, a minha querida esposa Tânia, e a minha filha Marta, por alguma indisponibilidade da minha parte nesta altura – obrigado pela vossa paciência.

Por último, um agradecimento a uma senhora muito especial que, sem ser directamente da minha família, tem acompanhado a minha vida desde o ano 2000. O seu apoio incansável foi fundamental para a conclusão deste projecto. A sua presença constituiu um pilar sólido que manteve a minha família unida durante esta fase em que não consegui estar tão disponível. E o único presente que lhe ofereci foi uma neta! Muito obrigado, Hermínia Claudina Silva, por tudo.

Índice

Índice de Figuras

Índice de Tabelas

Lista de Abreviaturas

BI – *Business Intelligence*

CRISP-DM – *CRoss-Industry Standard Process for Data Mining*

DM – *Data Mining*

DT – *Decision Trees* / Árvores de Decisão

e-mail – *Electronic Mail* / Correio Electrónico

IMC - *Integrated Marketing Communications*

IVR – *Interactive Voice Response*

MSE - *Mean Squared Error*

NB – *Naïve Bayes*

OLAP – *On-Line Analytical Processing*

ROC – *Receiver Operating Characteristic*

SMS – *Short Message Service*

SQL– *Structured Query Language*

SVM – *Support Vector Machine*

1. Introdução

O trabalho exposto neste documento resulta da investigação decorrida entre Setembro de 2010 e Agosto de 2011. Com uma base de relatórios contendo resultados de campanhas de *Marketing* direccionado de um banco, a ideia era trazer valor acrescentado para o negócio. Para tal, recorreu-se a uma abordagem de *Data Mining*. Neste capítulo são descritos os *drivers*[1] e objectivos iniciais que desencadearam a investigação.

1.1. Enquadramento e Motivação

A necessidade de cativar público, no sentido de captar novos clientes, tem sido desde sempre o principal catalisador de acções publicitárias cada vez mais arrojadas e inovadoras. Contudo, as acções publicitárias em massa com que o público é bombardeado têm reduzido de sobremaneira a sua eficácia. Daí que a aposta tem sido no sentido do *Marketing* direccionado. Este não é mais do que a canalização dos esforços da comunicação no sentido de cada cliente específico, ou seja, contactar cada cliente especificamente de entre uma lista pré--seleccionada, de forma a atingir o objectivo final da campanha de *Marketing*, conforme indicado por Tapp (2008).

No entanto, mais recentemente, o próprio *Marketing* direccionado tem sido fortemente criticado, o que tem conduzido ao surgimento de legislação específica, com algumas medidas que têm reduzido o seu impacto do ponto de vista do público-alvo. Uma dessas medidas foi a obrigatoriedade de disponibilizar a opção de *opt-out*, ou seja, de permitir ao cliente a possibilidade de indicar explicitamente que quer ficar excluído de qualquer campanha promocional. Segundo Tapp (2008), no Reino Unido o público tem criticado fortemente a utilização de campanhas telefónicas desencadeadas pelas empresas, o que tem conduzido a que seja cada vez mais utilizada a opção *opt-out* no sentido de deixar de todo de ser contactado.

Posto isto, torna-se cada vez mais importante que seja efectuada uma selecção criteriosa do público-alvo a atingir, escolhendo cada indivíduo de acordo com as suas características específicas, de modo a apenas contactar aqueles que, à partida e de acordo com a informação disponível, são bons candidatos a responderem positivamente ao objectivo da campanha de *Marketing*, isto é, têm o perfil adequado à campanha. Tal premissa impõe novas exigências ao nível da gestão de contactos e, consequentemente, ao nível das ferramentas informáticas de suporte.

Os factores atrás descritos motivaram o autor, que integra uma equipa de desenvolvimento de *software* direccionada para responder a solicitações provenientes da área de *Marketing* de uma instituição bancária portuguesa, para esta investigação Nesse contexto, esteve envolvido directamente num projecto de grande dimensão temporal e complexidade

[1] *Drivers* – factores impulsionadores do negócio.

cujo objectivo era o desenvolvimento de uma plataforma para campanhas direccionadas através de diversos canais não presenciais disponíveis na instituição. Tal projecto suscitou interesse em constatar qual o resultado da execução das diversas campanhas de *Marketing*. Esse interesse não se foca naquilo que a solução informática implementada permite, mas sim na área cuja responsabilidade é directamente da equipa de *Marketing*: a selecção dos clientes a contactar e a eficiência traduzida nos resultados de cada campanha.

De acordo com a investigação de Page e Luding (2003), que teve por base uma amostra aleatória de 153 clientes duma instituição bancária, o público-alvo de campanhas evidencia uma atitude negativa no que diz respeito a estratégias de *Marketing* direccionado, ou seja, a taxa de intenções de aquisição como resultado de uma campanha direccionada é muito baixa. Tal resultado sugere que é uma área onde claramente há uma grande margem de evolução.

Assim, ganha cada vez mais importância o conhecimento que se tem do público-alvo das campanhas de *Marketing*. Para optimizar o *Marketing* é importante agilizar as campanhas em torno de dois vectores:

- **eficácia** – aumentando os resultados positivos aquando das interacções com os clientes;
- **eficiência** – optimizando a gestão dos meios de comunicação, no sentido de utilizar os meios mais onerosos apenas para clientes com potencial elevado de responder positivamente à campanha.

Tendo estas necessidades presentes, a introdução da temática associada ao *Business Intelligence* afigura-se como apropriada, uma vez que o mesmo pode ser definido como um conceito que engloba metodologias com o objectivo de fornecer aos gestores de organizações a capacidade de analisar dados para permitir suportar as decisões de negócio (Turban *et al.*, 2010). Neste contexto, uma das suas vertentes é o *Data Mining*, ou seja, a descoberta de conhecimento a partir de dados já existentes, pelo que é com naturalidade que se proponha a sua aplicação a uma área cujos alicerces são construídos com base em informação como é o caso do *Marketing*.

As expectativas apontam para que a investigação corrente conduza a benefícios directos a aplicar na gestão de contactos de *Marketing* através da utilização de técnicas de *Business Intelligence*. Ou seja, o resultado deste trabalho poderá interessar a responsáveis e gestores de *Marketing*, uma vez que se espera que evidencie o potencial do *Business Intelligence* quando aplicado ao *Marketing*. Por outro lado, será interessante para académicos com enfoque no *Business Intelligence* uma vez que demonstrará uma aplicação prática das diversas técnicas.

Desta forma, a dissertação permitirá expor a utilização de técnicas de *Business Intelligence* como forma de potenciar o negócio das organizações. Sendo um caso assente na utilização de dados de uma instituição financeira, o estudo poderá abrir novas perspectivas a

organizações nesta área de negócio para a optimização dos seus recursos. Logo, é possível que este estudo seja utilizado como base para novas implementações e soluções para optimização da comunicação com os clientes, em benefício tanto das instituições, como dos clientes. Da mesma forma, poderá conduzir a novas investigações nesta área, contribuindo para um incremento do conhecimento.

1.2. Objectivos

Pretende-se efectuar um estudo de caso assente nos dados de uma solução aplicacional de campanhas de *Marketing* de uma instituição bancária portuguesa, na tentativa de identificar padrões comportamentais de clientes que permitam optimizar o funcionamento das próprias campanhas. Para tal, utilizar-se-ão técnicas e metodologias na área de *Business Intelligence* que permitam extrair conhecimento útil. Os dados a analisar correspondem a registos criados entre 2008 e 2010 e têm origem em várias campanhas, efectuadas por diversos canais (de *inbound* e *outbound*[2]).

A solução aplicacional enquadra-se na área de CRM[3] Operacional[4], uma vez que os contactos são seleccionados externamente para carregamento nas campanhas, cingindo-se a aplicação à gestão desses contactos face aos diversos canais, com algumas regras passíveis de serem parametrizadas. Assim, estando o enfoque do *Business Intelligence* no apoio à tomada de decisão, o trabalho de investigação centrar-se-á em *Data Mining*, ou seja, na descoberta de padrões que se traduzam em conhecimento útil e aplicável no caso em estudo.

Sendo actualmente a selecção de contactos a realizar feita de forma manual, pela própria instituição, pretende-se que a investigação conduza a um melhoramento do processo de selecção dos clientes e que se traduza na optimização dos resultados das campanhas, em consonância com as orientações estratégicas (directamente influenciadas por factores externos como a crise financeira internacional e nacional que criam uma maior pressão na racionalização dos recursos ao dispor da instituição).

[2] *Inbound* – a iniciativa do contacto é do cliente, e *Outbound* – quem despoleta o contacto é a instituição.
[3] CRM (Customer Relationship Management) – uma "abordagem" que coloca o cliente no centro do desenho dos processos do negócio, sendo desenhado para perceber e antecipar as necessidades dos clientes actuais e potenciais, de forma a suprimi-las adequadamente (Anderson e Kerr, 2001).
[4] CRM Operacional – visa melhorar a eficiência da relação entre uma empresa e os seus clientes, através da integração de ferramentas tecnológicas para proporcionar um melhor serviço ao cliente.

1.3. Organização

Para além desta introdução, a investigação está documentada em mais cinco capítulos.

No Capítulo 2, apresenta-se o enquadramento teórico associado à temática da gestão de campanhas de *Marketing* e de canais de comunicação pelos quais as campanhas são usualmente realizadas. Posteriormente, são apresentados os conceitos associados ao *Business Intelligence* e, mais concretamente, às técnicas de *Data Mining*. Finalmente, é feita uma exposição de vários estudos de aplicações práticas de *Data Mining* na gestão de campanhas.

No Capítulo 3, é descrito o problema, bem como todo o contexto envolvente associado ao caso em estudo e, em particular, a metodologia utilizada para o abordar. De seguida (Capítulo 4), são explanados os trabalhos e experiências práticas decorrentes da investigação. Este capítulo consubstancia a aplicação das técnicas descritas no Capítulo 2 ao caso prático mencionado no Capítulo 3. Assim, inclui a indicação das ferramentas utilizadas, e desenrola-se com a estrutura da metodologia base utilizada, o CRISP-DM (a qual será descrita mais à frente).

Após a componente prática de *Data Mining* sobre o caso em estudo, são discutidos (Capítulo 5) os resultados finais e extraídas conclusões no que se refere à utilidade do ponto de vista do negócio. Adicionalmente, são ainda mencionadas algumas possibilidades de aprofundamento da investigação em trabalhos futuros.

Importa ainda referir que alguns dos tópicos considerados não essenciais ou demasiado exaustivos são apresentados em anexos próprios, os quais são oportunamente referenciados nos Capítulos adequados.

2. *Marketing* e *Business Intelligence*

Segundo Kotler (2002), o *Marketing* pode ser definido do ponto de vista social e do ponto de vista da gestão empresarial. No primeiro caso, é considerado um processo pelo qual indivíduos e grupos obtêm o que necessitam através da criação, oferta e troca de produtos e serviços livremente entre si. Já do ponto de vista da gestão, pode ser descrito simplesmente como a arte de vender produtos. Mais precisamente, o objectivo é conhecer e compreender o cliente de uma forma tal que se lhe possa oferecer o produto ou serviço mais adequado e do qual ele mais necessite.

Neste contexto, a gestão de contactos com os clientes necessita de ser eficiente e eficaz de modo a poder, por um lado, responder às solicitações dos clientes e, por outro, potenciar eventuais acções de *Marketing* para promover produtos ou serviços. Após esta introdução, este capítulo estender-se-á em vários subcapítulos focando áreas mais específicas e relacionadas directamente com a investigação a levar a cabo.

Assim, começam por ser focados aspectos relacionados com o *Marketing* direccionado (2.1), canais de comunicação utilizados (2.2) e os centros de gestão de contactos (2.3), de forma a enquadrar o caso. Posteriormente, são enumeradas as características e aspectos relacionados com *Business Intelligence* e *Data Mining* (2.4) e descritos alguns estudos da sua aplicação em *Marketing* (2.5).

2.1. *Marketing* Direccionado

No sentido de promoverem os seus produtos e serviços, as empresas têm à sua disposição duas formas distintas de *Marketing* (Ling e Li, 1998): em massa e direccionado.

O *Marketing* **em massa** utiliza diversos meios de comunicação para promover produtos e serviços sem discriminar o público-alvo. Era tido como uma forma bastante eficaz de promoção quando a procura por produtos era elevada como, por exemplo, a procura por dispositivos electrónicos após a Segunda Guerra Mundial. Contudo, actualmente, num mundo global, cada vez mais competitivo, e com uma grande pressão da oferta sobre a procura, fruto de uma concorrência mais aguerrida, o *Marketing* de massas é cada vez menos eficaz. A resposta positiva a cada contacto é cada vez menor, em muitos casos, uma taxa de resposta de 1% é um valor usual (Ling e Li, 1998).

Alternativamente, ao invés de tentar promover produtos/serviços por um público indiscriminado, o *Marketing* **direccionado** baseia-se em estudos e análises para seleccionar os clientes mais apropriados para um determinado produto/serviço. É expectável que, desta forma, a taxa de resposta possa ser melhorada significativamente. Assim, o *Marketing* direccionado é cada vez mais popular devido à pressão da concorrência e ao custo de cada promoção (Ou *et al.*, 2003). Com taxas cada vez mais reduzidas nas próprias acções direccionadas, importa optimizar a escolha do público-alvo, de forma a melhorar a eficiência

das campanhas promocionais, reduzindo o custo associado às mesmas face ao retorno, reflectido directamente na taxa de resposta. Para tal, são aplicados modelos preditivos para determinar a probabilidade de cada cliente responder positivamente à promoção (Cohen, 2004).

Contudo, no estudo de Page e Luding (2003) é evidenciado o impacto negativo que o *Marketing* direccionado tem junto do público-alvo (no caso vertente, foi analisado o sector bancário). A mesma investigação sugere que a intenção de aquisição é significativamente influenciada pela atitude de cada indivíduo face à publicidade direccionada, sendo a influência do canal utilizado mais reduzida.

2.2. Canais de Comunicação

A utilização da tecnologia como facilitador da comunicação é um facto. Ao longo dos tempos têm surgido diversas tecnologias para ultrapassar as limitações que a simples comunicação verbal *in-loco* tem associadas (distância, persistência no tempo, etc.). Assim, surgiu (Whittaker, 2003): **tecnologia síncrona** (e.g. telefone, videoconferência) e **assíncrona** (e.g. correspondência/cartas, telégrafo, FAX, *e-mail*, SMS). No estudo de Whittaker (2003) é dado um grande enfoque a todos os aspectos da comunicação que um determinado canal permite. Por exemplo, um canal visual como videoconferência transmite uma mensagem subjacente às expressões faciais do interlocutor, algo que não está disponível num simples telefonema. Assim, a tecnologia pode permitir que um canal de comunicação seja linguístico ou visual (telefone *versus* videoconferência).

Uma outra classificação dos canais de comunicação é a relacionada com a interactividade: alguns canais são síncronos e bidireccionais, ou seja, **interactivos** (e.g. telefone ou *chat*), enquanto que outros não o são (*e-mail*, SMS). Não são, no entanto, analisados nesse estudo canais em massa, como a televisão, a rádio ou a imprensa escrita.

Tendo em conta a abordagem apresentada por Whittaker (2003), pode-se definir uma tabela classificativa dos diversos canais de comunicação usuais, com algumas ressalvas que importa mencionar. Assim, considera-se que um canal é bidireccional desde que se possa responder a uma comunicação exactamente pelo mesmo canal; por exemplo, num determinado *site* na *Internet* é possível disponibilizar um formulário para obter *feedback* (ou ter um sistema de mensagens associado ao próprio *site*), logo considera-se que a *Internet* é um canal bidireccional. Já no caso da imprensa escrita, é impossível ao público responder pelo mesmo canal (quanto muito, pode-se responder a anúncios por outros canais como o telefone ou o correio tradicional), pelo que se considera que é exclusivamente unidireccional. A Tabela 1 apresenta as características dos principais canais de comunicação.

Tabela 1 - Classificação de canais de comunicação

Canal	Forma de comunicação			Sincronismo		Tipo (direcção)	
	Escrita	Oral	Visual	Síncrono	Assíncrono	Unidir.	Bidir.
Telefone		X		X			X
E-mail	X				X		X
SMS	X				X		X
Correio	X				X		X
Televisão			X	X		X	
Rádio		X		X		X	
Internet			X		X		X
Imprensa			X		X	X	
Chat	X			X			X

Fonte: adaptado de Whittaker (2003: 246-247)

Pode ainda ser adicionada a classificação relativamente a quem toma a iniciativa de estabelecer inicialmente o contacto, ou seja, se o contacto é ***inbound*** (o cliente/público-alvo inicia o contacto) ou ***outbound*** (caso contrário). Ainda que esta classificação seja principalmente utilizada para o canal telefónico – como, por exemplo, no estudo de Evenson *et al.* (1999) – a mesma não está associada de forma directa ao canal, uma vez que qualquer canal bidireccional pode ser de *inbound* ou de *outbound*. Todos os canais unidireccionais – como, de resto, se pode depreender directamente da Tabela 1 – são de *outbound*, ou seja, a iniciativa é da instituição que pretende transmitir a mensagem ao público-alvo.

Nos anos recentes têm surgido novos canais de comunicação alternativos, tipicamente ligados ao aparecimento de novas tecnologias (*Internet*, tecnologia móvel).

Danaher e Rossiter (2006) realizaram um estudo, com abrangência a todo o território australiano, para compreender, quer do ponto de vista dos emissores de *Marketing*, quer dos indivíduos receptores, a eficácia relativa dos novos e velhos canais de comunicação no que diz respeito a acções de *Marketing*. A selecção do canal adequado a uma determinada acção promocional é um tópico muito complexo. Uma das conclusões do estudo é a de que o canal de comunicação mais adequado depende das preferências de cada indivíduo.

A *Internet* providenciou novos canais de comunicação. No entanto, vários problemas que lhe estão associados como vírus informático, fraude, *spam* e invasão de privacidade causam um impacto negativo na sua utilização para acções de *Marketing*. Como resultado, o investimento é ainda superior nos canais tradicionais, ainda de acordo com o estudo de Danaher e Rossiter (2006).

No que diz respeito às intenções de aquisição como resultado da campanha, as conclusões do estudo apontam para uma convergência de opinião entre emissores e receptores, sendo que o telefone, *e-mail*, SMS e a venda porta-a-porta têm resultados inferiores (menos intenções de aquisição) do que a correspondência escrita, televisão, catálogos, rádio e imprensa escrita. No entanto, os custos dos canais electrónicos são muito

inferiores, o que resulta numa maior utilização destes canais e que, como tal, acaba por saturar os receptores das campanhas, afectando a sua eficácia.

O estudo de Owen e Humphrey (2009) aborda a forma como o *Marketing online* está a utilizar as novas redes sociais, com interacções entre vários indivíduos, ao invés de ser apenas uma interacção entre o emissor da campanha de *Marketing* e o público-alvo. Dessa investigação resulta uma proposta de infra-estrutura para estudar a forma como os novos canais de comunicação *online* estão a emergir e como poderão evoluir no futuro.

Os elementos desta infra-estrutura são: 1) *core*/tecnológicos, ou seja, associados a toda a base tecnológica que dá suporte à comunicação em rede; 2) competitivos/comerciais, isto é, tem de existir uma motivação comercial para potenciar o crescimento de um novo canal; 3) políticos/regulatórios, uma vez que a regulação jurídica é fundamental para credibilizar as mensagens divulgadas pelo respectivo canal; e 4) sociais, dado que é necessário que o canal tenha uma boa aceitação da parte do público e seja efectivamente utilizado, para que possa ser atractivo do ponto de vista do *Marketing*.

Este estudo identifica ainda alguns riscos potenciados pelos canais *online* associados a ataques com publicidade ou boatos negativos e personificação de figuras públicas com o objectivo de divulgar uma imagem negativa sobre a pessoa, muitas vezes com efeitos nefastos no impacto público. Um dos casos típicos referido é o de indivíduos em campanha eleitoral política, cuja necessidade de angariação de empatia pública é fundamental para obter os resultados desejados.

2.3. Centros de Gestão de Contactos (*Contact-Centers*)

As instituições financeiras estão no cerne das mudanças mais radicais na escolha e investimento no que diz respeito a canais de comunicação (Evenson *et al.*, 1999). Inicialmente, os *call-centers*[5] surgiram para reduzir os custos no tratamento de problemas ou questões de clientes. No entanto, já no estudo indicado, datado de 1999, se referia que os *call-centers* se estavam rapidamente a tornar numa forma de elevado potencial de disponibilizar serviços e outras mais-valias que se pudessem traduzir num retorno financeiro para as instituições. Da investigação supracitada conclui-se que são factores preponderantes os recursos humanos e a tecnologia na disponibilização de *call-centers* eficientes.

Actualmente, os *call-centers,* na verdadeira acepção da palavra, isto é, que apenas disponibilizam o tradicional serviço telefónico, estão a transformar-se em *contact-centers*, ou centros de gestão de contacto, disponibilizando outros canais que não apenas atendimento

[5] C*all-center*: centro equipado para receber e efectuar um grande volume de chamadas telefónicas (retirado de http://www.thefreedictionary.com/), acedido em Agosto de 2011.

telefónico por equipas de assistentes (Koole e Mandelbaum, 2002): atendimento telefónico automático (ou IVR[6]), *e-mail*, FAX, *Internet* ou *chat*.

Conforme referido por Danaher e Rossiter (2006), os gestores de *Marketing* possuem na actualidade uma miríade de canais de comunicação para a divulgação de produtos/serviços das suas empresas. A escolha do canal apropriado para cada campanha e indivíduo é complexa. Daí que tenha surgido um novo conceito, ***Integrated Marketing Communications* (IMC)**, ou a comunicação de *Marketing* integrada, o qual é explanado por Owen e Humphrey (2009).

A ideia em geral é a de que existe uma variedade de métodos, campanhas e canais de comunicação à disposição de uma organização, sendo que, para se tirar um melhor proveito da panóplia indicada, é necessário haver coordenação e centralização das actividades que envolvem cada uma das componentes referidas no estudo, sejam elas *core*/tecnológicas, competitivas/comerciais, políticas/regulatórias ou sociais. Neste contexto, é usual que uma organização defina as diversas estratégias de *Marketing* de uma forma global, servindo-se, posteriormente, de uma série de ferramentas tecnológicas para que sejam seleccionados os produtos/serviços mais apropriados a promover para cada cliente. Para a divulgação, utilizará os canais de comunicação à disposição do centro de gestão de contactos, bem como outros que, por razões físicas ou outras, não possam ser geridos a partir do centro (por exemplo, a difusão de anúncios publicitários em *mass-media*, como televisão, rádio ou imprensa, ou o contacto pessoal por parte de promotores humanos).

2.4. *Business Intelligence* e *Data Mining*

As organizações estão a apostar cada vez mais em tecnologia de suporte à decisão, incluíndo o *Business Intelligence*. Turban *et al.* (2010) desenvolveram um modelo para ajudar a compreender essa aposta, sendo que, nesse modelo, os factores ambientais e exteriores criam pressões e/ou oportunidades para as organizações, as quais necessitam de tomar decisões para a definição da sua estratégia. É precisamente neste contexto que surge a área de *Business Intelligence*, como facilitador da tomada de decisão, através do fornecimento de conhecimento que sirva de suporte à decisão.

De acordo com Turban *et al.* (2010), o termo *Business Intelligence* é um chapéu debaixo do qual se enquadram arquitecturas, ferramentas, bases de dados, aplicações e metodologias. O principal objectivo de *Business Intelligence* é permitir o acesso interactivo a dados e a sua manipulação, e fornecer aos gestores de negócio e analistas a possibilidade de efectuar uma análise consubstanciada.

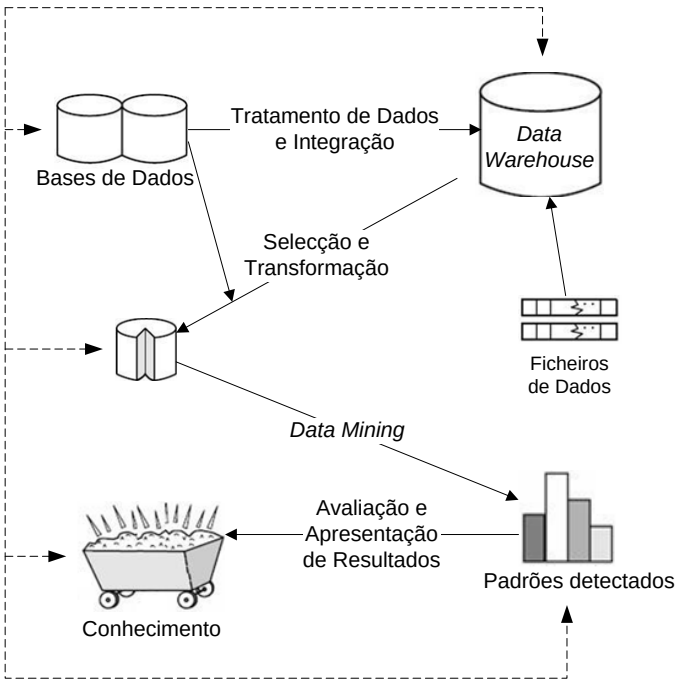
[6] *Interactive Voice Response* (IVR): sistema de atendimento automático telefónico em que o utilizador interage através de DTMF (*dual-tone, multi-frequency*), ou da própria voz (*speech-to-text*).

Associada ao *Business Intelligence* surge a necessidade de descoberta de conhecimento em Bases de Dados. Tal conceito enquadra-se perfeitamente no contexto de *Business Intelligence* e é especialmente relevante tendo em conta o manancial de dados de que as organizações dispõem, mas com o problema de, tipicamente, os dados se encontrarem dispersos, uma vez que respondem a diferentes tipos de necessidades e funções.

Assim, uma das formas de organizar os dados para responder a solicitações no âmbito de *Business Intelligence* é agrupá-los em *Data Warehouses* (armazéns de dados), que consistem em repositórios de dados com diversas origens, organizados sob um esquema único e num único local de forma a facilitar a obtenção de conhecimento para apoio à gestão (Han e Kamber, 2006). Tal implica processos de limpeza e integração de dados e ferramentas OLAP (*On-Line Analytical Processing*), ou seja, técnicas de análise com funcionalidades como sumarização, consolidação e agregação, para além da capacidade de ver a informação de diferentes perspectivas. Apesar de este tipo de ferramentas suportarem a tomada de decisão, são necessárias outras ferramentas adicionais para uma análise de dados mais aprofundada.

No seguimento do exposto e, entrando numa área mais específica debaixo do enorme chapéu que é o *Business Intelligence*, *Data Mining* é um termo usado para descrever a descoberta de conhecimento a partir de grandes quantidades de dados (Turban *et al.*, 2010). Por outras palavras e, segundo Witten e Frank (2005), trata-se de um processo de descoberta de padrões em dados. Esses padrões têm de ter utilidade, de modo a que permitam alguma vantagem no conhecimento que transmitem. De acordo com Han e Kamber (2006), *Data Mining* é um passo essencial da descoberta de conhecimento em que métodos inteligentes são aplicados com o objectivo de extrair padrões de dados (Figura 1).

Figura 1 - *Data Mining* e a descoberta de conhecimento

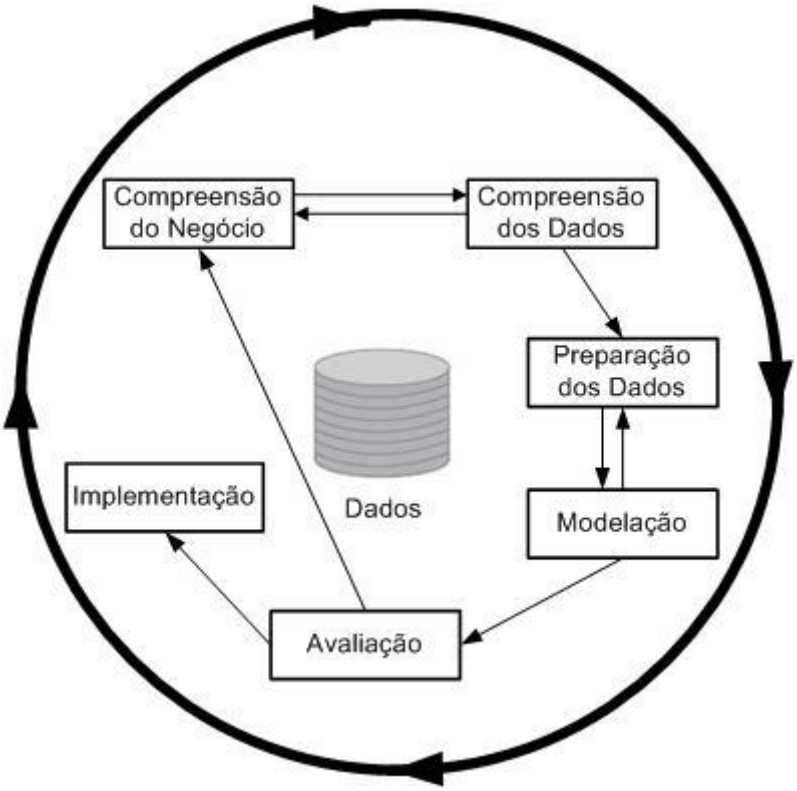


Fonte: adaptado de Han e Kamber (2006: 6)

Assim, a descoberta de conhecimento engloba *à priori* etapas para tratamento de dados, e *à posteriori* passos para transformar os padrões obtidos em conhecimento. As setas a tracejado na Figura 1 representam a natureza iterativa e flexível do processo. Contudo, dado o uso actual mais lato do termo *Data Mining* e, tal como utilizado por Han e Kamber (2006), o termo *Data Mining* irá doravante ser usado como um sinónimo do processo da descoberta de conhecimento.

Não existe uma metodologia universalmente aceite para a condução de projectos de descoberta de conhecimento em Bases de Dados, apesar de alguns esforços nesse sentido. Neste contexto, uma das metodologias que surgiu com grande aceitação foi a CRISP-DM (*CRoss-Industry Standard Process for Data Mining*). Esta surgiu precisamente na tentativa de definir um modelo processual *standard*, não-proprietário e gratuito para sistematizar a descoberta de conhecimento (Chapman *et al.*, 2000). De acordo com esta metodologia, o ciclo de vida de um projecto de *Data Mining* consiste em seis fases, conforme apresentado na Figura 2.

Figura 2 - Fases do modelo CRISP-DM



Fonte: adaptado de Chapman *et al.* (2000: 10)

A sequência de fases do ciclo de vida não é rígida, até porque é sempre necessário recuar para fases anteriores. O resultado de cada fase determina qual a próxima fase a executar. As setas indicadas na Figura 2 referem-se às dependências mais importantes e frequentes entre fases. O círculo exterior simboliza a natureza cíclica do processo de *Data Mining*, uma vez que não termina assim que a solução é disponibilizada para ser usada.

Cada uma das fases pode-se decompor em tarefas genéricas, sendo que, para cada uma destas, existirão tarefas específicas. Por seu turno, cada tarefa específica decompõe-se em instâncias de processos. Assim, o modelo CRISP-DM disponibiliza quatro níveis de abstracção, sendo que, para além das fases, as tarefas genéricas são suficientemente abrangentes para cobrir as diversas actividades associadas a *Data Mining*. Porém, as tarefas específicas são especializadas no sentido em que cada uma permite lidar apenas com um dado tipo de problema. Já as instâncias de processos representam as actividades executadas para um determinado problema específico de *Data Mining*.

Relativamente às seis fases macro apresentadas na Figura 2 as mesmas podem ser sucintamente descritas do seguinte modo:

1. Compreensão do negócio – nesta fase inicial é necessário entender os requisitos do negócio e converter esse conhecimento numa definição clara do problema e desenhar um plano preliminar para atingir os objectivos;
2. Compreensão dos dados – inclui a colecta inicial de dados e actividades para permitir ganhar alguma familiaridade com os dados de forma a poder identificar problemas relacionados com a sua qualidade, ou detectar subconjuntos de interesse acerca dos quais se possam formular hipóteses;
3. Preparação dos dados – tarefas de preparação de dados com a finalidade de construir um conjunto de dados final, o qual possa servir para alimentar directamente os modelos de *Data Mining*;
4. Modelação – aplicação de diversas técnicas aos dados, de forma a obter um (ou mais) modelo(s) afinados através de parametrizações que permitam representar o conhecimento;
5. Avaliação – análise dos modelos obtidos para verificar se, de facto, atingem os objectivos de negócio; e
6. Implementação – disponibilização do modelo para que possa ser utilizado directamente pelas equipas de gestão do negócio.

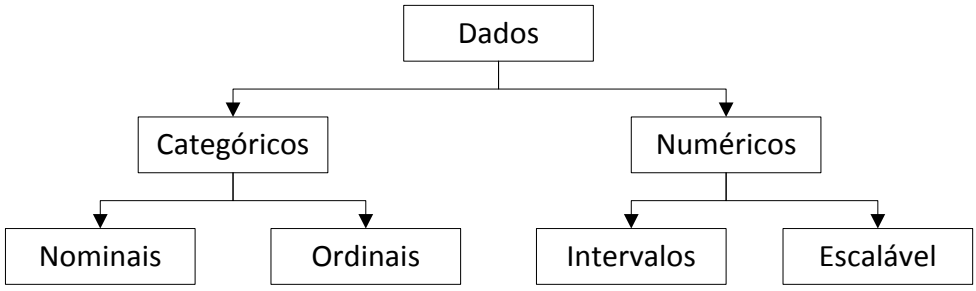
Entrando em maior detalhe no conceito de *Data Mining*, este implica por si só um processo com vários passos iterativos, e ainda diversos tipos de experiências para verificar se os padrões descobertos nos dados existentes permanecem válidos face a novos dados, pelo menos com um certo grau de certeza (Turban *et al.*, 2010). A necessidade de experiências e validações justificam que se afirme que o processo não é trivial no sentido em que não existe uma simples receita a seguir, mas antes vários caminhos alternativos a explorar e que dependem, entre outros factores, dos dados e do conhecimento que se pretende extrair.

Conforme se pode constatar pela Figura 1, é necessário fornecer *inputs* para se efectuar *Data Mining*. Assim, de acordo com Witten e Frank (2005), esses *inputs* tomam a forma de conceitos, instâncias e atributos. Os conceitos representam uma descrição do que se pretende aprender. A informação fornecida ao processo de *Data Mining* é representada por um conjunto de instâncias, em que cada uma representa uma ocorrência independente do conceito

a aprender. Por exemplo, se o conceito a aprender for a previsão de respostas dadas a uma determinada prova de avaliação por alunos de um curso, então cada instância corresponderá ao conjunto de respostas dadas na prova por cada um dos alunos. Cada instância é caracterizada por um conjunto de valores referentes a atributos em que cada um mede um aspecto diferente da instância. No exemplo anterior, cada atributo corresponde a cada uma das perguntas da prova de avaliação, e as respostas correspondem aos valores caracterizadores da instância, ou seja, da prova de avaliação de determinado aluno.

Podem existir diferentes tipos de atributos, contudo as técnicas típicas de *Data Mining* só conseguem funcionar com tipos categóricos (também conhecidos como discretos, enumerados, qualitativos ou categoriais) ou numéricos (também designados como métricos ou quantitativos). Conforme a Figura 3, os dados categóricos podem ser nominais (quando não existe nenhuma relação de ordem entre cada categoria (valor) – por exemplo, o estado civil: solteiro, casado/junto, divorciado/separado ou viúvo), ou ordinais (quando pode ser definida uma ordem – por exemplo, habilitações: nenhuma, básico, secundário e superior. Já os dados numéricos são escaláveis, isto é, podem representar um qualquer valor numérico numa escala (contínua – por exemplo, idade – ou discreta – por exemplo, o número de filhos).

Figura 3 - Taxonomia simples para os tipos de atributos em *Data Mining*



Fonte: adaptado de Turban *et al.* (2010: 197)

Muitas vezes, os conjuntos de dados a analisar pelas técnicas de *Data Mining* contêm *missing values*[7]. Estes valores são frequentemente indicados através de valores sem nexo para o atributo em causa, por exemplo, no caso de atributos numéricos usando um valor negativo para um atributo cujo leque de valores seja o conjunto dos números naturais (ex: um atributo que indica o número de filhos, e que poderá ter o valor -1 se esse número for desconhecido) e, no caso de atributos nominais, usando um espaço ou uma barra para representar os casos desconhecidos (Witten e Frank, 2005).

É importante haver algum cuidado na escolha de valores para representar os *missing values* – pode haver necessidade de distinguir diferentes tipos de *missing values* – por exemplo, num questionário (uma instância do conjunto de dados), face a uma pergunta sobre rendimento financeiro, alguém pode-se recusar a responder a uma determinada questão mas,

[7] A tradução de *missing values* seria "valores em falta", "valores omissos" ou "não-respostas", no entanto, como o termo *missing values* é sobejamente utilizado quer pelas ferramentas de *Data Mining*, quer inclusive na literatura portuguesa, no decorrer do presente documento será utilizado o termo em inglês.

noutro caso, pode ter simplesmente havido um esquecimento de preencher a resposta (nesta situação poder-se-ia adicionar uma opção de resposta "não respondo", a qual teria de ter um tratamento específico, uma vez que o valor apropriado e usual seria numérico, por se tratar do rendimento).

Alternativamente, pode-se optar por tratar os *missing values* substituindo-os por valores reais, ao invés de os utilizar tal como estão para alimentar o processo de *Data Mining*. Assim, de acordo com Brown e Kros (2003), esse tratamento pode ser efectuado de diversas formas:

- Ignorar os registos (atributos/exemplos) com valores omissos;
- Substituir (Imputação) cada valor omisso por:
    - Valor dado por um perito (*case substitution*);
    - Valor médio, mediana ou mais comum (moda) do atributo;
    - Valor retirado de outra base de dados (*cold deck*);
    - Valor do exemplo mais semelhante/próximo (*hot deck*);
    - Valor estimado por regressão linear;
    - Combinação dos métodos anteriores (*multiple imputation*).

Importa ainda ter em conta que, usualmente, os dados utilizados para *Data Mining* não são especificamente obtidos com esse fim. Assim, é muito provável que existam valores erróneos, os quais depois poderão afectar o desempenho das técnicas de *Data Mining*. Por exemplo, pode haver erros tipográficos para a introdução do valor num atributo nominal o que resultará (provavelmente) no facto da técnica assumir o valor diferente como uma enumeração completamente distinta da pretendida. Adicionalmente, certos dados mudam com o tempo – por exemplo, a morada de um cliente pode mudar.

Poderão ainda existir alguns valores atípicos (ou *outliers*), os quais poderão estar de todo errados ou ser apenas alguns casos raros. Num ou noutro caso estes valores poderão influenciar a técnica de *Data Mining* na sua aprendizagem ao analisar os dados. Caso seja necessário, as instâncias com valores deste tipo poderão ser descartadas.

Os diversos modelos de *Data Mining* (concretizados na implementação das respectivas técnicas) permitirão, depois do processo de aprendizagem com base nos dados e instâncias já existentes, para determinados dados de *input* de uma nova instância, obter os valores estimados para os dados de *output*. Esses dados de *output* poderão corresponder a apenas um atributo ou a vários. Se o atributo a prever for nominal, então trata-se de um problema de classificação, isto é, pretende-se prever a probabilidade de uma nova instância pertencer a uma determinada classe para o atributo. Caso o atributo seja numérico, então é um problema de regressão, sendo que o atributo a prever terá um valor numérico que o modelo considere o mais provável face ao que "aprendeu" através dos dados iniciais. Pode suceder ainda que o objectivo não seja prever o valor de um atributo, mas antes agrupar instâncias com características semelhantes, ou seja, em que os valores dos seus atributos sejam aproximados o suficiente para que a técnica de *Data Mining* considere que ambas as instâncias pertencem

ao mesmo grupo. Neste caso, trata-se de um problema de segmentação (ou *clustering*), e como exemplos dessas técnicas tem-se a rede neuronal SOM (*Self-Organization Map*) e o algoritmo *k-Means* (Mingoti e Lima, 2005).

De acordo com Hand *et al.* (2001), a regressão pode ser linear, no caso de o atributo de *output* ter uma relação linear (aproximada) com apenas um atributo de *input*, ou múltipla no caso de o atributo de *output* ter uma relação linear com dois ou mais atributos. Pode ainda ser uma regressão não linear quando a relação do *output* (variável dependente) com os dados de *input* (variáveis independentes) não for linear (nos parâmetros). Nestes casos, em Estatística, é necessário conhecer o modelo matemático *à priori*. Existem, no entanto, modelos de *Data Mining* adequados a situações em que não se conhece a função matemática.

Importa fazer neste ponto uma breve nota referente à relação entre *Data Mining* e outras ciências, nomeadamente a Estatística. Assim, segundo Kuonen (2004), a Estatística é a ciência que permite aprender a partir de dados, e inclui todo o processo de colecta, tratamento e gestão de dados, bem como a sumarização e a apresentação dos resultados. No entanto, em *Data Mining*, o objectivo não é só modelar ou prever, mas também definir processos para resolver problemas relacionados com a obtenção de conhecimento. Adicionalmente, *Data Mining* utiliza técnicas não só do domínio da Estatística, mas também da Inteligência Artificial na procura de soluções para a extracção de conhecimento. Cada técnica poderá ser adequada a apenas um tipo de problema (classificação ou regressão). A Tabela 2 apresenta algumas tipologias de modelos de *Data Mining* e o fim a que se destinam.

Tabela 2 - Técnicas de *Data Mining* e tipos de problemas a que se adequam

Técnicas mais usuais	**Pode ser usada em problemas de:**		
	Classificação	Regressão	Segmentação
Regras de classificação	X		
Árvores de decisão	X	X	
Naive Bayes	X		
Random forests	X	X	
Linear discriminant analysis (LDA)	X		
Regressão linear múltipla		X	
Regressão logística	X		
K-Vizinhos mais próximos	X	X	
Redes neuronais artificiais	X	X	
Máquinas de vectores de suporte	X	X	
Self-Organization Map			X
K-Means			X

Após se obter um modelo através da aprendizagem recorrendo a dados iniciais, importa obter métricas que permitam avaliar a capacidade de previsão desse modelo, isto é,

medir a qualidade do modelo. Existem vários métodos, dependendo do tipo de *output* que se pretende prever.

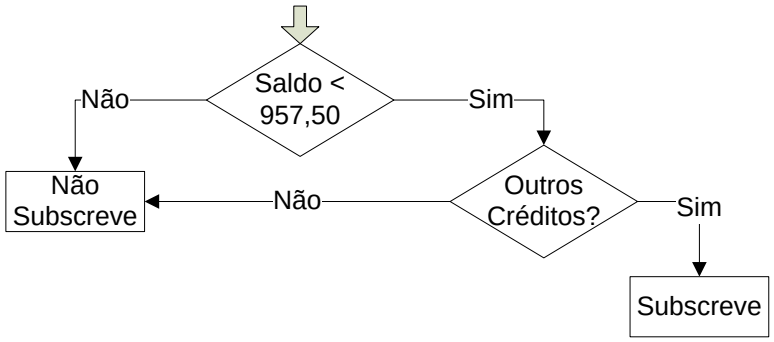
Uma vez que existe uma miríade de modelos, o enfoque será dado em especial àqueles utilizados no âmbito desta investigação.

De forma a facilitar a compreensão da restante revisão literária e enquadramento, será utilizado um exemplo de um problema de classificação que foi desenvolvido unicamente com este propósito. Assim, foi criado um ficheiro com 98 instâncias de uma hipotética campanha de subscrição de um cartão de crédito bancário, com quatro atributos caracterizadores: se a instituição é o banco principal do cliente, o saldo médio da conta à ordem principal, a indicação da existência de outros créditos na instituição e a indicação se o cliente detém um cartão de débito.

No que diz respeito a problemas de classificação, uma das técnicas mais utilizadas é a das Árvores de Decisão (Apte e Weiss, 1997). Uma das suas principais vantagens assenta na sua facilidade de compreensão, dado que são modelos fáceis de interpretar por qualquer ser humano. Por esta mesma razão, um bom exercício para analisar os dados poderá ser a contrução de uma árvore de forma manual, aplicando algum bom senso e, principalmente, conhecimento do "negócio" a analisar (Witten e Frank, 2005).

Na Figura 4 pode-se visualizar a árvore gerada para o exemplo indicado acima.

Figura 4 - Árvore de Decisão para a subscrição de um cartão de crédito



Cada nó da árvore apresenta uma decisão, a qual é baseada num ou mais atributos. Neste caso, é possível inferir a seguinte regra: um cliente subscreve o cartão se tiver outros créditos e um saldo inferior a 957,50 euros. Naturalmente que este se trata de um exemplo muito simples. Em casos reais, a árvore pode ter inúmeros níveis, tornando-a mais complexa.

Outra técnica muito simples é a *Naive Bayes*, na qual se assume que nenhuma das variáveis de entrada influencia qualquer das restantes, situação que raramente ocorre num problema real (Zhang, 2004). Portanto, para a probabilidade de ocorrer um determinado resultado contribuem de modo independente várias variáveis. Apesar desta limitação, é uma técnica que tem sido utilizada com sucesso para resolução de problemas, o que se deve em parte ao facto de, em determinadas condições, a dependência entre alguns atributos pode ser anulada pela dependência entre outros.

Contudo, algumas técnicas mais elaboradas têm surgido e que permitem obter modelos com melhores performances. Uma das mais recentes é a de Máquinas de Vectores de Suporte, que se baseia na definição e utilização de vectores de suporte que contenham apenas os exemplos mais representativos do universo de treino (Cortes e Vapnik, 1995). Tendo os mesmos definidos, é aplicada uma transformação não linear aos atributos de entrada através de uma função de *kernel* que permite definir o hiperplano óptimo de separação entre as possíveis classes de saída (Hearst *et al*., 1998). Uma das funções de *kernel* mais populares é o *gaussiano*, que apresenta menos parâmetros do que outros *kernels* (e.g. polinominal).

Após se obter um modelo através da aprendizagem recorrendo a dados iniciais, importa obter métricas que permitam avaliar a capacidade de previsão desse modelo, isto é, medir a qualidade do modelo. Existem vários métodos, dependendo do tipo de *output* que se pretende prever. Adicionalmente, importa referir que a aprendizagem pode ser supervisionada, se se baseia em instâncias exemplificativas do problema para "ensinar" a técnica de *Data Mining* numa fase dita de treino (adequada aos problemas de classificação e regressão); se, por outro lado, a aprendizagem se baseia apenas na descoberta de regularidades (semelhanças) nos dados de entrada, procurando agrupamentos (segmentação/*clustering*) a partir das instâncias fornecidas para a fase de treino, então é uma aprendizagem não supervisionada.

Caso se pretenda prever uma classe, ou seja, qual o valor de um atributo enumerado e, portanto, referente a um problema de classificação, pode-se utilizar uma matriz de confusão para avaliar a qualidade do modelo. Esta permite mapear os valores previstos com os valores observados de acordo com Kohavi e Provost (1998). Na Tabela 3 é possível verificar a matriz de confusão gerada para o exemplo da subscrição do cartão de crédito já mencionado atrás. A mesma corresponde aos resultados da aplicação do modelo de árvores de decisão ao conjunto de dados de teste, considerando uma resposta positiva (subscrição do cartão) sempre que o modelo assim o indicar com uma probabilidade superior a 0,5.

Tabela 3 - Matriz de confusão - previsão de subscrição de cartão de crédito

↓Observado\Previsto→	Não	Sim	Total
Não	18	5	23 (69,7%)
Sim	4	6	0 (30,3%)
Total	22 (66,7%)	11 (33,3%)	

Pela matriz pode-se constatar que se previram correctamente 6 "subscrições" e 18 "não subscrições". No entanto, verifica-se que o modelo previu a "não subscrição" para 4 casos em que o cliente acabaria por subscrever o cartão. Para, do ponto de vista do negócio, se avaliar correctamente o modelo seria necessário atribuir valores financeiros, quer para a subscrição do cartão, quer para o custo de cada contacto, e só assim se poderia avaliar melhor a eficiência do modelo. Tendo em conta a matriz de confusão, podem-se determinar algumas medidas de avaliação do modelo (Fawcett, 2005), sendo também apresentados os resultados, muitas vezes expressos em percentagem, para o exemplo indicado na Tabela 3:

- *Accuracy* (ACC) ou exactidão:

$$\frac{Positivos\ e\ Negativos\ correctamente\ classificados}{Total\ de\ positivos + Total\ de\ negativos} = \frac{18+6}{33} = 0{,}727\ (72{,}7\%) \qquad (1)$$

- *True Positive Rate* (TPR) ou sensibilidade:

$$\frac{Positivos\ correctamente\ classificados}{Total\ de\ positivos} = \frac{6}{10} = 0{,}600\ (60\%) \qquad (2)$$

- *False Positive Rate* (FPR) ou taxa de falsos positivos:

$$\frac{Negativos\ incorrectamente\ classificados}{Total\ de\ negativos} = \frac{5}{23} = 0{,}217\ (21{,}7\%) \qquad (3)$$

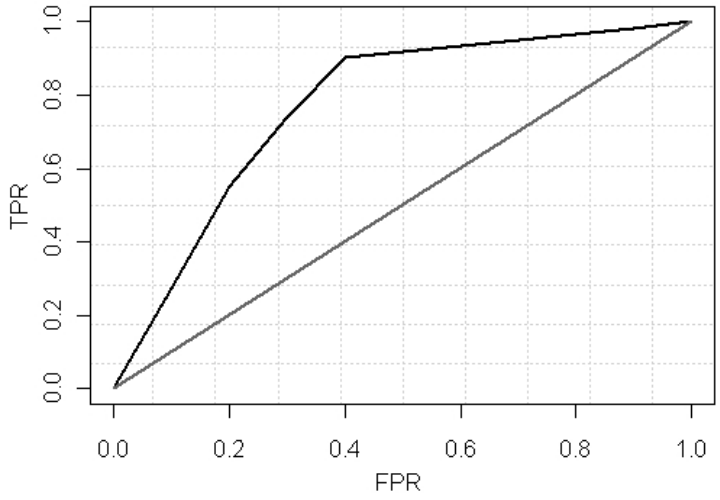
- Especificidade:

$$\frac{Negativos\ correctamente\ classificados}{Total\ de\ negativos} = 1 - \mathrm{FPR} = 0{,}783\ (78{,}3\%) \qquad (4)$$

Assim, tem-se que a Taxa de Falsos Positivos é igual a 1 – especificidade.

Conforme descrito por Fawcett (2005), a taxa de falsos positivos (eixo das coordenadas) e a sensibilidade (eixo das ordenadas) permitem traçar a curva ROC (*Receiver Operating Characteristic*). Como ambas as medidas se tratam de proporções, tal significa que ambos os valores variam entre 0 e 1. Tendo em conta que um modelo define uma probabilidade entre 0 e 1 para um resultado correspondente à classe objectivo, cada ponto da curva representa um determinado valor a partir do qual se considera que essa probabilidade é, de facto, a classe objectivo. Para o exemplo da subscrição de cartão de crédito já referido anteriormente, a curva ROC resultante é a traçada na Figura 5.

Figura 5 - Curva ROC - previsão de subscrição de cartão de crédito



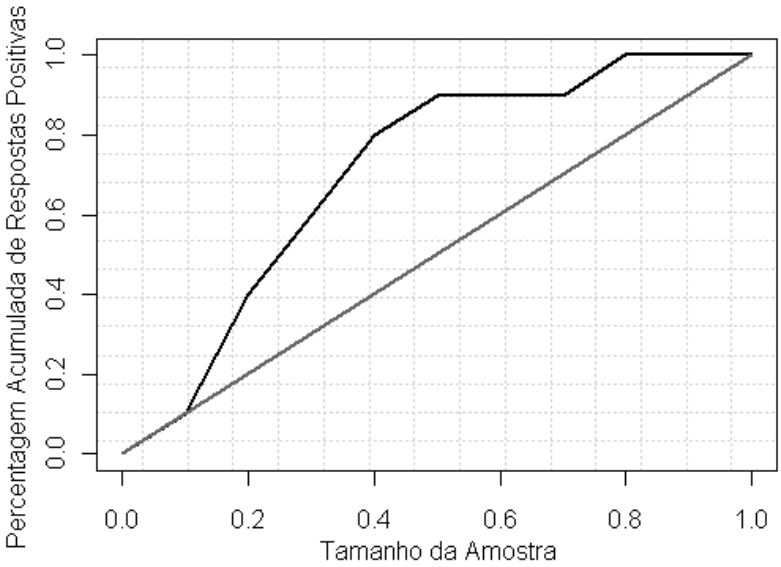
Tendo em conta as taxas (2) e (3) apresentadas acima, conclui-se que a previsão perfeita é representada pelo ponto (0, 1), ou seja, foram previstos correctamente todos os casos positivos, e não houve nenhum caso negativo incorrectamente classificado como positivo. O ponto (0, 0) representa a estratégia de nunca prever uma classificação positiva. A estratégia oposta, de prever sempre uma classificação positiva, corresponde ao ponto (1, 1). A diagonal traçada entre o ponto (0, 0) e o ponto (1, 1) corresponde a um classificador aleatório que, em média, preveja correctamente metade dos casos.

Efectuando uma mera análise visual ao gráfico da Figura 5, pode-se dizer que um ponto no "espaço ROC" corresponde a uma melhor previsão do que outro ponto nesse espaço caso esteja mais para "noroeste", ou mais próximo do ponto (0, 1) (Fawcett, 2005). De uma forma mais formal, o desempenho global do modelo é dado pela "área debaixo da curva", ou seja, por:

$$AUC = \int_0^1 ROC\ dD = 0{,}774 \qquad (5)$$

O método provavelmente mais utilizado para medir a performance de modelos em que se pretende obter respostas para campanhas de *Marketing* é o *Lift* (Coppock, 2002). Tipicamente, quantifica-se através da divisão da população em decis (dez grupos com o mesmo número de instâncias cada), ordenados pela previsão que o modelo dá para a probabilidade de responder positivamente à campanha de *Marketing,* de forma decrescente, ou seja, colocando as instâncias correspondentes aos indivíduos cuja probabilidade de resposta positiva é maior no primeiro decil, e assim por diante. Na Figura 6 é apresentada a curva do *Lift* cumulativo para o exemplo da subscrição do cartão de crédito, ou seja, o número de respostas positivas previstas dividido pelo número total de respostas positivas para cada decil. A *baseline* corresponde à diagonal traçada entre os pontos (0, 0) e (1, 1), podendo ser interpretada da seguinte forma: num determinado decil N, tem-se que N% da população total capturaria N% de respostas positivas.

Figura 6 - Curva *Lift* - previsão de subscrição de cartão de crédito



Fonte: adaptado de Coppock (2002: 1)

Quanto maior for a área entre a *baseline* e a curva *Lift* cumulativa, melhor é o modelo, uma vez que conseguiu concentrar as respostas positivas nos primeiros decis. No exemplo, verifica-se que, se se seleccionar a metade das instâncias de teste com maior probabilidade de sucesso, se obtêm 90% do total de subscrições que se obteriam com o total do conjunto de instâncias.

De acordo com Coppock (2002), o *Lift* é especialmente útil para avaliar a performance relativa de vários modelos alternativos. Por exemplo, se um modelo de árvores de decisão

providencia um *Lift* mais elevado do que um modelo de redes neuronais artificiais para os mesmos dados, então a análise *Lift* torna-se um factor de decisão chave na escolha do modelo com melhor capacidade de previsão.

Anteriormente foram referidas medidas para avaliar a performance dos modelos; estas medidas pretendem avaliar a capacidade de previsão de um modelo obtido através de "aprendizagem" para prever o *output* para novas instâncias. No entanto, o que acontece tipicamente é que, para a definição de um modelo, é apenas fornecido um grande conjunto de instâncias, o qual tem de servir, quer para a aprendizagem, quer para a validação do modelo obtido. Uma técnica simples de valiação de modelos supervisionados é o *holdout*, que consiste na divisão do conjunto inicial de instâncias num conjunto de treino, para a construção do modelo, e noutro para teste, para avaliar o desempenho do modelo. Naturalmente que, quanto mais dados se tiverem nesta fase, melhor será o modelo (mais dados de treino para aprendizagem) e melhor será avaliado o seu desempenho (mais dados para teste), embora a partir de uma determinada quantidade de dados, esse efeito diminua (Witten e Frank, 2005).

No entanto, alguns problemas poderão advir desta validação. Se o conjunto de instâncias inicial for reduzido, então pode suceder que o subconjunto definido para treino não tenha uma dimensão significativa, o que irá resultar numa má definição do modelo. Uma estratégia simples para reduzir esta possibilidade é efectuar uma validação cruzada (*cross--validation*): o treino é feito em várias iterações sobre o conjunto de dados, sendo que, a cada iteração, são definidos dois subconjuntos diferentes de treino e de testes. A técnica é executada sempre sobre o subconjunto de treino mas, como este varia a cada execução, consegue-se, com as várias iterações, obter resultados que reflictam todo o conjunto de dados inicial, tirando proveito da diversidade do mesmo na sua totalidade.

Após se obter um bom modelo na sua capacidade de previsão, devidamente validado através de um conjunto de dados de teste, importa aferir da sua utilidade prática com dados reais. O mesmo pode servir como modelo de previsão a aplicar em dados futuros, para que seja possível utilizar os resultados previstos pelo mesmo em benefício do negócio.

No entanto, o modelo pode também ser útil para explicar a influência que os atributos de entrada têm na sua definição, por exemplo, através de um método de análise da sensibilidade, conforme proposto por Cortez e Embrechts (2011). Nesse estudo, os autores propõem a utilização da curva VEC (*Variable Effect Characteristic*) para uma interpretação gráfica da influência de um atributo de *input* no atributo de *ouput*. Assim, o conhecimento obtido por um modelo explicativo pode também ser utilizado pelos gestores em benefício do negócio.

2.5. Aplicações Práticas de *Data Mining* em contextos associados ao *Marketing*

Data Mining não é uma área nova, mas sim um conceito que engloba a utilização de diversas disciplinas (Turban *et al.*, 2010). Ou seja, ainda que várias das técnicas utilizadas em *Data Mining* já tenham algumas décadas de existência e tenham vindo a ser aplicadas esporadicamente e em determinadas situações, foi no decorrer da década de 1990 que se tornou uma área específica, muito por força de responder a necessidades e pressões colocadas pelo negócio.

Desta forma, os primeiros artigos científicos que referem *Data Mining* surgem na segunda metade da última década do século passado, sendo que uma boa parte envolve a aplicação de diversas técnicas a casos de estudo contextualizados dentro das áreas científicas de *Marketing* e Finanças.

As pressões competitivas levaram ao surgimento de novas metodologias e técnicas de *Marketing* e a tecnologia potenciou o surgimento de novos canais de comunicação que foram desde logo inundados com promoções de *Marketing* cada vez mais agressivas. Todo este contexto envolvente transpôs-se de uma forma ampliada para o início do século XXI.

Ling e Li (1998) apoiam-se na experiência adquirida em projectos de *Marketing* direccionado utilizando *Data Mining*. Um dos problemas que os levou a enveredar por *Data Mining* e a explorar esta área em maior detalhe foi o facto de a taxa de respostas positivas ser muito baixa (aproximadamente 1% para os casos que estudaram), tratando-se de uma característica muito própria do *Marketing* direccionado, uma vez que é cada vez mais difícil cativar o público-alvo para promoções e campanhas, dada a quantidade de publicidade com que este é inundado.

Outro dos problemas identificados é também caracterizador de *Marketing*: os erros de classificação não têm todos a mesma importância. Enquanto que reconhecer compradores entre não compradores (falsos positivos) significa apenas que são efectuados contactos desnecessários (no entanto, é o que acontece usualmente, com as taxas de respostas positivas baixas), já o facto de não se reconhecer compradores, ou seja, de se classificar como não compradores indivíduos que acabariam por responder positivamente à campanha (falsos negativos) é um erro a evitar de todo, uma vez que são oportunidades de negócio que se perdem.

Para resolver este problema, é necessário escolher técnicas de aprendizagem que classifiquem cada instância com uma determinada probabilidade de ocorrência. Se todas as instâncias forem ordenadas do indivíduo mais provável comprador para o menos provável, então basta que os gestores de negócio seleccionem os N mais prováveis compradores como alvos da campanha, sendo N uma variável que podem usar quer para controlar os custos, quer para controlar a probabilidade mínima que a técnica indicou para uma resposta positiva. Desta

forma, para conseguir avaliar a eficácia da técnica pode ser usado o método *Lift*, tal como o fizeram Ling e Li (1998).

O principal ponto referido na secção de conclusões e discussão final desse estudo pode ser generalizado a todos os casos de aplicação prática de métodos de *Data Mining*: o objectivo é sempre obter o melhor ROI[8]. O que interessa acima de tudo ao negócio é que seja lucrativo. Após se ter os indivíduos tidos como potenciais alvos da campanha, ordenados por ordem decrescente da probabilidade de aquisição, é necessário estimar custos da própria campanha (por exemplo, o custo do envio de cartas); tendo também as probabilidades de aquisição devidamente estimadas e, supondo que é possível estimar o resultado financeiro de uma resposta positiva (há que ter em atenção que nem sempre as campanhas têm como objectivo a aquisição de produtos), pode-se então escolher finalmente o conjunto de indivíduos que tornarão a campanha mais eficaz.

Ainda que a tarefa de encontrar padrões comportamentais nos dados de negócio não seja nova, tradicionalmente era realizada por analistas de negócio recorrendo a técnicas estatísticas. Com o advento e proliferação de diversas ferramentas informáticas de *Data Mining*, os gestores e analistas de negócio têm um natural constrangimento que se prende com a dificuldade de entender toda esta panóplia de novos termos e conceitos.

Esta dificuldade foi identificada por Bose e Mahapatra (2001) que publicaram um artigo com o objectivo de facilitar a compreensão de alguns dos conceitos actuais à data. Assim, foram analisadas algumas das técnicas mais conhecidas e utilizadas e verificadas as vantagens e desvantagens de cada uma quando aplicadas a cinco diferentes categorias de problemas: Finanças, telecomunicações, *Marketing*, análise *Web* e outras.

Conforme já foi mencionado, uma das áreas de grande interesse pela utilização de *Data Mining* dentro da banca é a detecção de fraude. O artigo publicado por Brause *et al.* (1999) demonstra a investigação levada a cabo para a prevenção de transacções fraudulentas com cartões de crédito. Um dos problemas identificados é a taxa de incidência de fraudes muito baixa, na ordem de 0,1% face ao total de transacções (inferior cerca de 10 vezes face à encontrada por Ling e Li (1998) para os casos de *Marketing* direccionado). Foi desenvolvido com sucesso um modelo baseado em redes neuronais artificiais que conseguia lidar com este problema específico.

Outra área no sector financeiro em que *Business Intelligence* em geral pode ser aplicado para extraír informação útil para o negócio é a melhoria de eficiência na gestão operacional de balcões de atendimento ao público, conforme foi investigado por Leite *et al.* (2009). Ao contrário da maioria das publicações referidas anteriormente nesta secção, no caso deste artigo, o enfoque não é técnico mas, fundamentalmente, no impacto a nível da gestão da organização. A solução de *Business Intelligence* implementada no estudo de caso é muito abrangente e foi aplicada com sucesso a um dos maiores bancos brasileiros. As principais

[8] *Return on investment* - é a relação entre o dinheiro ganho ou perdido através de um investimento, e o montante de dinheiro investido (Keen e Digrius, 2003).

conclusões apontam como factores críticos de sucesso para a implementação de projectos de *Business Intelligence* de grande envergadura a necessidade de um forte patrocinador dentro da organização, ou seja, de um responsável com um elevado poder de decisão, e de uma orientação focada no alinhamento com o negócio, de forma a cativar todos os intervenientes no processo.

Os dois artigos mencionados anteriormente demonstram como *Business Intelligence* tem sido utilizado para que a área financeira beneficiasse do conhecimento adicional adquirido com as novas técnicas e metodologias.

A publicação de Kim e Street (2004) apresenta aos gestores de *Marketing* uma abordagem de *Data Mining* para selecção de clientes num contexto de *Marketing* direccionado baseada em redes neuronais artificiais guiadas por algoritmos genéticos. Apesar das inúmeras publicações sobre *Data Mining* aplicado ao contexto de *Marketing* direccionado, os autores argumentam que, no geral, não têm sido abordados dois aspectos fundamentais do ponto de vista dos gestores de negócio.

Por um lado, estes pretendem conhecer o processo (no fundo, as regras) que suportaram uma determinada decisão. É muito mais fácil deste modo aceitarem indicações de um "computador" que irão fundamentar todo o seu negócio do que se forem apenas *outputs* de uma caixa negra, sobre a qual desconhecem o seu funcionamento, tal como o são muitas das soluções de *Business Intelligence* e *Data Mining* comercializadas no mercado.

Por outro, é crítico para os gestores que possam conhecer o número de indivíduos que devam ser incluídos numa determinada campanha para que possam maximizar o lucro ou aumentar a sua quota de mercado ou, pelo menos, que permitam cobrir os custos operacionais de execução da campanha (esta visão está, de resto, em linha com a necessidade identificada também por Ling e Li (1998) de ter sempre em mente o principal objectivo: obter o máximo retorno do investimento). De forma a atingir este objectivo, os gestores necessitam de uma análise com sensibilidade suficiente para que demonstre como o valor resultante do objectivo da campanha (por exemplo, o lucro) varie em função dos parâmetros da campanha (por exemplo, o número de indivíduos abrangidos).

Como conclusões, os autores indicam que a utilização de um algoritmo genético adequado a um modelo de redes neuronais artificiais permite a construção de modelos de previsão que reflictam directamente a decisão dos gestores de *Marketing*.

No que diz respeito à qualidade de serviço de um *call*-center, um dos factores que influencia o resultado de cada contacto é a qualidade do atendimento. Assim, é de realçar a versatilidade na aplicação de técnicas de *Data Mining*, que podem ser também utilizadas para melhorar a qualidade do serviço, conforme investigado por Paprzycki *et al.* (2004). Nesta investigação conclui-se que a gestão do c*all-center* deve-se focar no treino dos assistentes de atendimento e na melhoria dos produtos oferecidos ao invés de analisar o tempo dispendido nas chamadas.

No estudo de Hu (2005) foi abordado o problema da retenção de clientes bancários através de *Data Mining*, ou seja, um problema de classificação. No projecto conduzido pela investigação, foram utilizadas árvores de decisão, duas técnicas de *Naïve Bayes* e redes neuronais. Adicionalmente, foi ainda utilizada uma técnica híbrida que consiste em utilizar as classificações obtidas pelas quatro técnicas anteriores, em que o resultado previsto para uma determinada instância consiste no resultado obtido pela maioria das técnicas isoladas. O modelo obtido com esta solução revelou-se melhor na capacidade de previsão do que qualquer um dos restantes modelos isoladamente.

Li *et al*. (2010) analisaram dados referentes a clientes detentores de cartões de crédito de um banco chinês, de modo a aplicar técnicas de *Data Mining* para a segmentação dos clientes e eventuais acções de *Marketing* direccionado. Os dados iniciais constituiam 72544 instâncias e 27 atributos. Para a modelação no que diz respeito ao *Marketing* direccionado, foram usadas quatro técnicas: redes neuronais, árvores de classificação e regressão (*C & R Tree*), árvores *Chi-Square Automatic Interaction Detection* (*CHAID*) e árvores C5.0. Apesar de as redes neuronais proporcionarem o modelo com melhor capacidade de previsão, o mesmo foi rejeitado por ser de difícil explicação. Assim, o modelo adoptado foi o das árvores C5.0, o segundo melhor modelo nas suas previsões. Com base nesse modelo, foram definidas várias regras de classificação, concluíndo o estudo pela análise de algumas delas, com percentagens de casos a satisfazerem-nas superior a 80%.

Claramente que existe a necessidade nos bancos de organizar os seus clientes por perfis, de forma a facilitar futuras acções de *Marketing*. Taghva *et al*. (2011) propoem efectuar a segmentação de clientes através de *Self-Organization Map* (*SOM*), organizando os dados de entrada em três grupos: pessoais, financeiros e de utilização de serviços *online*. Após elegerem um *cluster* com 981 clientes, identificado através de *SOM*, foram analisadas as regras de associação, através das quais foram apontados 12 serviços de entre 21 como sendo utilizados mais regularmente pelos clientes identificados.

O estudo de Javaheri (2008) foca a selecção de alvos de *Marketing* direccionado na banca através de *Data Mining*. O objectivo é implementar um modelo preditivo com base em dados históricos (30000 clientes de um banco e 85 atributos de entrada) para a selecção de clientes com maior probabilidade de responderem positivamente a ofertas de produtos bancários constituindo, portanto, um problema de classificação. O modelo é construído recorrendo a Máquinas de Vectores de Suporte, o qual tem uma boa capacidade de previsão, avaliada com recurso a uma análise *Lift*, permitindo, por exemplo, que os 40% clientes com maior probabilidade de resposta positiva contenham, de facto, 80% de respostas positivas. No entanto, o autor aponta a limitação de não dispor de dados caracterizadores do perfil pessoal do cliente para além da idade, apesar da extensa lista de atributos, que acaba por estar completamente relacionada com o perfil histórico enquanto cliente bancário. Adicionalmente, os atributos não incluem os resultados de anteriores contactos no contexto de *Marketing* direccionado.

Para terminar este capítulo de revisão da literatura, é de referir a publicação de Ngai *et al.* (2009), em que é efectuado um levantamento de publicações relacionadas com a aplicação de técnicas de *Data Mining* a contextos de CRM, tendo sido revistos artigos publicados entre 2000 e 2006. É de realçar que o *Marketing* direccionado se enquadra no contexto muito geral de CRM, de acordo com o indicado no referido artigo. Em Ngai *et al.* (2009), pode-se verificar a forma como o *Marketing* direccionado (enquadrado na dimensão de "Atracção de Clientes") se integra com os restantes conceitos associados a CRM (Tabela 4).

Tabela 4 - Distribuição de artigos de acordo com o modelo de classificação proposto

Dimensões CRM	**Elementos CRM**	**Tipo de Problema de *Data Mining***	**Modelos de *Data Mining***	**Autores**
Atracção de clientes	*Marketing* direccionado	Classificação	Regressão logística	(Prinzie e Poel, 2005)
			Rede de classificação ingénua de *Bayes*	(Baesens *et al.*, 2002)
			Árvore de decisão	(Buckinx *et al.*, 2004)
			Algoritmos genéticos	(Ahn *et al.*, 2006) (Chiu, 2002)
			Redes neuronais artificiais e algoritmos genéticos	(Kim e Street, 2004)
		Segmentação	Detecção de *Outlier*	(He *et al.*, 2004)

Fonte: adaptado de Ngai *et al*. (2009: 2597)

2.6. Sumário

O *Marketing* é a ciência que estuda a divulgação e promoção de produtos e serviços. Quando a divulgação é dirigida a determinados indivíduos, em função das suas características, trata-se de *Marketing* direccionado. O surgimento de centros de contacto tem permitido a proliferação de campanhas direccionadas, conduzindo, entre outros factores, a uma saturação do público-alvo e consequente diminuição dos resultados atingidos pelas acções promocionais. Nesse sentido, importa encontrar formas de optimizar os contactos a efectuar, nomeadamente, através da análise da informação decorrente de campanhas anteriores de forma a extraír conhecimento, o que pode ser feito com recurso a técnicas de *Data Mining*.

Na revisão da literatura efectuada neste capítulo foram apresentados vários estudos bem sucedidos referentes a aplicações de *Data Mining* a contextos de *Marketing*. As técnicas de análise e tratamento de dados, modelação e avaliação dos modelos obtidos permitem suportar um projecto de *Data Mining*. No entanto, é de realçar que o enfoque deve ser sempre

direccionado para a optimização do objectivo de negócio, só assim se consegue cativar os gestores para patrocinarem investimentos em *Data Mining*.

3. Metodologia

Este capítulo inicia-se com uma descrição do contexto envolvente do caso em estudo (3.1), sendo depois explanadas as características do próprio caso (3.2), nomeadamente no que se refere à informação resultante da execução das campanhas de *Marketing* direccionado e que constituirá o conjunto de dados de suporte à investigação.

Posteriormente, é referido no secção 3.3 o planeamento inicialmente efectuado aquando do arranque da investigação e a forma como o problema é abordado com recurso à metodologia CRISP-DM, sendo ainda indicadas as ferramentas e técnicas de *Data Mining* utilizadas, respectivamente, nas secções 3.4 e 3.5.

3.1. Contextualização

No âmbito do estabelecimento de contactos com clientes por canais alternativos (não presenciais), a área de *Marketing* de uma instituição financeira portuguesa possui à sua disposição uma solução tecnológica desenvolvida pela área de desenvolvimento informático da mesma instituição, à medida das necessidades e requisitos expostos. Essa solução enquadra-se na plataforma multicanal da organização, que expõe canais alternativos para usufruto dos clientes (*homebanking*[9], banca telefónica, SMS, serviço para dispositivos móveis, etc.), constituindo, no entanto, um módulo autónomo da referida plataforma.

Esta solução gira em torno do conceito de campanha. Assim, uma campanha é a noção de contexto único da comunicação com um mesmo objectivo e que pode ser efectuada de diversas maneiras, por diversos canais e em fases distintas do tempo. O objectivo de campanha pode ser qualquer um que faça sentido do ponto de vista do negócio da instituição, conforme definido pela área de *Marketing*. Por exemplo, pode ser um inquérito de satisfação, uma campanha de angariação de novos clientes (porque não é obrigatório que as campanhas sejam direccionadas apenas para clientes), a venda de um produto financeiro, entre outros.

Os canais de comunicação a serem utilizados podem ser quaisquer de que a instituição disponha e que estejam integrados com a solução. Até ao final de 2011, os canais disponíveis eram os seguintes: telefónico na vertente de *inbound* e *outbound*, SMS, *e-mail*, mensagens enviadas e recebidas através do serviço de *homebanking* e páginas *Web* específicas no serviço de *homebanking*.

Toda a solução informática é gerida a partir do centro de gestão de contactos (*contact-center*) da instituição, sendo as campanhas geridas através de uma aplicação de administração, onde se seleccionam quer as características, quer o segmento de clientes alvo

[9] *Homebanking* – banco *online*, ou electrónico, permite aos clientes de um banco acederem de uma forma autenticada e segura a serviços bancários (execução de transacções financeiras) através de um portal disponível na *Internet* (retirado de http://www.portal-financeiro.com/homebanking.html), acedido em Agosto de 2011.

da campanha. A operacionalização da campanha é feita também através do centro de gestão de contactos, quer com o atendimento telefónico, quer respondendo às solicitações de clientes através dos restantes canais não presenciais (SMS, *e-mail*, entre outros).

No entanto, importa referir que nem todos os canais são geridos através do centro de gestão de contactos da instituição. Por exemplo, publicidade nos meios de comunicação social não é tratada pelo *contact-center*, ou abordagens comerciais aos clientes quando estes se dirigem aos balcões de atendimento, efectuadas pelos colaboradores da instituição. Naturalmente que é mais difícil, senão mesmo impossível, aferir com clareza qual o impacto que o investimento em publicidade fora dos canais de comunicação geridos pela instituição tiveram no sucesso ou não da campanha para um determinado cliente (como são meios de comunicação fora do controle directo, a sua eficácia apenas pode ser medida de um modo global aos objectivos da campanha).

Tal facto é importante, uma vez que, para o caso em estudo, não será possível avaliar o impacto de publicidade em meios externos à organização na execução da campanha de subscrição telefónica do produto. Ou seja, o estudo será apenas com base na comunicação através dos canais disponíveis na solução descrita atrás, ainda que possam existir outras variáveis que poderão influenciar essa mesma comunicação.

3.2. Descrição do Caso

Conforme mencionado, a solução tecnológica associada ao canal escolhido tem de ser alimentada com uma lista de clientes para cada campanha, sendo esta uma tarefa manual. Para a selecção dos mais adequados como alvo para atingir o objectivo de uma determinada campanha, é solicitada à área de informática uma lista de clientes com determinadas características, sendo depois a mesma afinada casuisticamente. Assim, ainda que, para a selecção dos clientes pela área de *Marketing* estejam definidas algumas condições (por exemplo, seleccionam apenas os clientes com mais de X euros de saldo médio anual), não existe nenhuma forma sistematizada de incorporar o conhecimento adquirido durante as campanhas e em campanhas anteriores no processo de selecção de clientes para novas campanhas.

No entanto, a solução permite exportar relatórios de campanhas com os resultados da execução das mesmas, para que a área de *Marketing* possa efectuar análises de negócio e tratamento estatístico. É através destes relatórios que serão obtidos os dados a analisar. Contudo, como esses dados são apenas referentes à campanha, foram obtidos adicionalmente alguns dados socioeconómicos caracterizadores de cada cliente alvo da campanha.

As campanhas a analisar ocorreram temporalmente entre Maio de 2008 e Novembro de 2010. Durante este período foram criadas e executadas 72 campanhas. Todas as campanhas são segmentadas, isto é, correspondem a *Marketing* direccionado. Por outras palavras, são especificamente configuradas para surgirem apenas aos clientes que foram

devidamente seleccionados. Apesar da diversidade de canais de comunicação disponíveis, para as campanhas em análise apenas foram utilizados os seguintes canais: telefónico e através de páginas específicas no *homebanking*.

No caso de o cliente telefonar para o *Contact-Center* e ser abrangido pela campanha, o mesmo poderá ser abordado pelo agente nesse âmbito, tratando-se de *inbound* telefónico. Assim, conforme definido na secção 2.2, este caso corresponde ao *cross-selling*[10] telefónico, ou seja, aproveitar o facto de o cliente estar a telefonar para tentar promover uma das campanhas das quais o cliente é alvo, poupando assim à instituição o custo da chamada.

Desta forma, um dos dados que será contabilizado será o número de vezes que o assistente de atendimento telefónico viu uma dada campanha (bem como a última vez que viu), ou seja, teve oportunidade de abordar o cliente no âmbito dessa campanha (não significa que o tenha feito). Caso o cliente seja abordado, então terá de ser obtido um resultado, tal como para o *telemarketing* e conforme será descrito mais à frente.

No caso mais usual, o *telemarketing* é utilizado, ou seja, as chamadas são efectuadas pela instituição para os clientes alvos da campanha. Trata-se de um meio de comunicação intrusivo, pelo que existe legislação específica para permitir ao cliente inviabilizar que seja contactado por uma organização. Assim, é de realçar que alguns clientes poderão ficar, à partida, excluídos desta forma de contacto se o indicarem explicitamente, não sendo possível contabilizá-los no presente estudo.

Doravante, para facilitar a leitura, os canais utilizados para as campanhas em estudo serão apenas designados por *inbound* telefónico, *telemarketing* e *homebanking*.

Por cada chamada executada por ambos os canais telefónicos, por serem objectivamente síncronos e bidireccionais, existe sempre um resultado associado (no caso do *inbound* telefónico, tal é válido apenas quando o assistente tem oportunidade de abordar o cliente no âmbito da campanha), o qual pode ser um dos enumerados na Tabela 5. Esta tabela é assim constituída por três colunas, a primeira com o resultado tal e qual com o nome que surgirá nos dados importados, a segunda com a descrição desse resultado do ponto de vista do negócio, e a terceira onde se indica se corresponde a um estado terminal.

Objectivamente, um resultado só não é terminal se daí resultar um novo agendamento telefónico no âmbito do mesmo contacto. No entanto, interessa, do ponto de vista do negócio, conhecer as razões desse agendamento, daí que existam vários resultados possíveis respeitantes a agendamentos, conforme está patente na Tabela 5.

[10] *Cross-selling* – venda cruzada: corresponde a aproveitar um contacto da iniciativa do cliente para vender outro produto não relacionado directamente com o serviço que está, no momento, a ser prestado ao cliente.

Tabela 5 - Resultados possíveis para um contacto

Resultado	**Descrição**	**Estado terminal?**
Sucesso do contacto	Atingido o objectivo da campanha	Sim
Insucesso	Não atingido o objectivo da campanha	Sim
Agendamento Decisor	Chamada atendida pelo cliente alvo da campanha que, no entanto, solicitou ser contactado mais tarde. **Descontinuado a partir de 2008-07-10**	Não
Agendamento outros	Chamada telefónica atendida por outro indivíduo que não o cliente alvo da campanha, tendo sido possível agendar novo contacto para outro telefone (supostamente) pertencente ao cliente – ou indicado pelo indivíduo que atendeu, ou outro disponível na lista de telefones da base de dados (neste caso, pode ter sucedido que a chamada nem tenha sido atendida, mas o assistente, por ter números de telefone disponíveis do cliente, decidiu reagendar a mesma – tal é analisado casuisticamente)	Não
Agendamento Decisor - Prod. Apresentado	Chamada atendida pelo cliente alvo da campanha que, no entanto, apesar de ter permitido ao assistente a apresentação do produto/serviço em campanha (ou seja, houve interesse em ouvir a proposta), solicitou ser contactado mais tarde. **Novo a partir de 2008-07-11**	Não
Agendamento Decisor - Prod. Não Apresentado	Chamada atendida pelo cliente alvo da campanha que, no entanto, solicitou ser contactado mais tarde, sem ter permitido ao assistente efectuar qualquer introdução à campanha no diálogo. **Novo a partir de 2008-07-11**	Não
Agendamento gravador	A chamada foi atendida por um gravador de chamadas, tendo, no entanto, o assistente tido a possibilidade de agendar novo contacto. **Novo a partir de 2009-10-27**	Não
Abortado porque o agente efectuou um *cleanup*	Contacto terminado pelo *software* que gera automaticamente as chamadas – operação incorrecta do assistente (permite assim alertar os assistentes para o comportamento incorrecto na utilização da aplicação)	Sim
Não atendeu a chamada	Chamada não atendida e não houve a possibilidade de agendar novo contacto (ou o assistente assim decidiu)	Sim
Não é o dono	Chamada atendida por outro indivíduo que indica que o telefone já não pertence ao cliente alvo da campanha	Sim
FAX em vez de telefone	Chamada atendida por um FAX	Sim
Contacto de *outbound* abandonado	Contacto terminado pelo *software* que gera automaticamente as chamadas (não foi possível por razões técnicas despoletar a chamada para nenhum dos telefones do contacto)	Sim

É de realçar que a própria área de negócio sentiu necessidade de discriminar mais detalhadamente alguns dos estados, daí que o resultado de “Agendamento Decisor” tenha sido desdobrado em dois, o “Agendamento Decisor - Produto Apresentado” e o “Agendamento Decisor - Produto Não Apresentado”. Tal permite que seja possível distinguir as situações em

que o cliente esteve receptivo para ouvir o assistente enquanto este lhe apresentava o produto/serviço em campanha. Assim, no âmbito do pedido à área técnica para acrescentarem estes estados, foi ainda solicitado que as chamadas correspondentes a agendamentos em que o cliente já tenha sido introduzido à campanha tivessem prioridade face às demais, canalizando assim os esforços para os contactos com maior potencial de sucesso.

Tipicamente, para ambos os canais telefónicos existe um guião para gerir o fluxo do diálogo, com várias perguntas opcionais, sendo que, em cada opção, poderá ser colocada uma questão diferente ao cliente, tentando assim adequar o diálogo do assistente de atendimento às respostas do cliente. Caso existam outros resultados que não apenas a conclusão/fecho de campanha para esse contacto, então poderá existir uma ficha de venda a preencher no final do contacto. Alguns dos dados relevantes dessa ficha poderão ser uma conta à ordem e um montante para subscrição de um determinado produto, caso o objectivo da campanha seja a venda desse produto.

Já o *homebanking* corresponde a algumas páginas específicas que surgem, quer ao entrar no serviço, quer em certos menus intermédios acessíveis a partir da árvore de opções. No entanto, não é possível com a solução actual determinar onde é que surgiu a campanha, pelo que se contabiliza como um acesso à campanha (um contacto). Como a solução não permite criar páginas interactivas (ou seja, apenas são páginas *Web* de visualização), no caso deste canal apenas são contabilizados o número de visualizações que cada cliente viu a "promoção", bem como a data e hora da última visualização.

Conforme referido no início desta secção, os dados resultantes dos contactos da campanha são exportados através de relatórios. Cada relatório é um ficheiro de texto no formato CSV[11]. No entanto, existem dois tipos de relatórios, um para registar as visualizações, e outro para registar os resultados.

Assim, no relatório de visualizações é obtida informação dos seguintes dados:

- canal de comunicação;
- número de visualizações da campanha; e
- data da última visualização da campanha.

O atributo referente ao número de visualizações é um número natural (incluindo o zero). Já o atributo para a data da última visualização da campanha é a data e hora da última visualização, no seguinte formato: "aaaa-mm-dd hh:mm:ss.ms"[12].

[11] CSV – *comma-separated values* ou *character-separated values*, corresponde a um formato de ficheiro de texto em que o separador é uma vírgula (",") ou um ponto-e-vírgula (";"), dependendo do facto de o separador numérico decimal ser um ponto (".") ou uma vírgula (","), respectivamente.

[12] Formato de data e hora utilizado nos dados: "aaaa-mm-dd hh:mm:ss.ms", em que "aaaa" é o ano, "mm" o mês, "dd" o dia, "hh" a hora, "mm" os minutos, "ss" os segundos e "ms" os milissegundos – por exemplo: 2008-05-27 21:22:32.66.
Para dados referentes apenas à hora, o formato é o seguinte "hh:mm:ss", com o mesmo significado mencionado para a data e hora.

Este relatório de visualizações apenas serve para os canais de *inbound* (quer para o *homebanking*, quer para o *inbound* telefónico), uma vez que, no caso do *telemarketing*, o contacto é despoletado pela instituição. No entanto, os atributos têm um significado um pouco diferente para os dois canais de *inbound* em análise.

No caso do *homebanking*, uma visualização contabilizada no relatório significa que, de facto, a página *Web* foi disponibilizada no *browser* do cliente (em princípio, o cliente efectivamente viu a campanha). Por outro lado, no caso do *inbound* telefónico significa apenas que, durante uma chamada efectuada pelo cliente (por iniciativa deste e com interesses muito provavelmente diferentes ao objectivo da campanha), o assistente verificou na aplicação de apoio ao atendimento que o cliente era alvo da campanha de *inbound* telefónico e, eventualmente, pode ter tentado efectuar uma abordagem comercial, a qual pode ou não ter resultado num contacto efectivo no âmbito da campanha.

Assim, enquanto que o relatório de visualizações da campanha no *homebanking* para o negócio é um indicador relativamente fiel que o cliente teve conhecimento da campanha, já no caso do *inbound* telefónico esse indicador permite verificar que houve a possibilidade de falar da campanha ao cliente e analisar este dado em conjunto com a efectivação do contacto ou não .

O relatório de resultados permite obter informação dos seguintes dados:

- canal de comunicação;
- resultado do contacto, de entre os indicados na Tabela 5;
- número de telefone;
- data e hora de contacto;
- duração do contacto;
- data e hora de agendamento; e
- montante de subscrição (no caso do objectivo da campanha ser a subscrição de um produto financeiro).

Este relatório serve apenas para os canais em que existe um contacto síncrono com o cliente, ou seja, para este estudo tratam-se dos canais telefónicos, o *telemarketing* e o *inbound* telefónico. Isto porque é nestes canais que é possível estabelecer um diálogo com o cliente e obter uma resposta à campanha (o resultado do contacto). Assim, os dados caracterizadores do contacto são o número de telefone, a data/hora do contacto e a duração do mesmo. No caso do resultado ser um agendamento é registada também a data/hora para a qual foi agendada novo contacto. Se o resultado for um sucesso e se tratar de uma campanha de venda de produto financeiro, então têm-se também o montante subscrito.

Importa voltar a frisar que se pretende optimizar as campanhas em geral por via de uma selecção mais rigorosa e restritiva de contactos. Tal permitirá, por um lado, obter ganhos ao nível dos custos operacionais (agentes e comunicações, entre outros), por outro, melhorar a

própria relação com os clientes pelo facto de se ser mais selectivo na análise à receptividade da campanha por parte do cliente.

Nesta fase, dada a diversidade de campanhas (já referida em 3.1 e que será analisada em maior detalhe no decorrer da investigação), não é ainda possível ser preciso na definição da expectativa dos resultados desta investigação uma vez que a mesma contempla todo um processo de *Data Mining* que engloba uma clarificação ao nível da própria definição de objectivos.

3.3. Planeamento

Exposta a base da qual se pretendia que a investigação evoluisse, na secção 3.2, tornou-se crítico definir, numa primeira fase, um planeamento adequado ao intervalo temporal em que o projecto decorreu. Desta forma, a metodologia adoptada, CRISP-DM (Chapman *et al*., 2000), tinha de ser passível de se ajustar dinamicamente à medida que o projecto de investigação evoluisse e tendo em conta o tempo disponível em cada fase.

Uma das grandes vantagens da metodologia CRISP-DM é o facto de a mesma ser simultaneamente iterativa e cíclica, permitindo retroceder nas diversas fases e tarefas, de forma a enriquecer os modelos obtidos ou mesmo rever totalmente as condições que os originaram. Assim, o CRISP-DM pretende ser apenas um guia do processo de procura de conhecimento, sendo totalmente flexível. O objectivo a cumprir com este dinamismo é o de permitir a liberdade de explorar novos caminhos que inclusive podem conduzir à revisão dos objectivos de negócio.

Para iniciar o projecto de investigação propriamente dito, era necessário obter o acesso aos dados. Assim, foi garantida previamente a disponibilidade dos dados e autorizado o seu uso no âmbito desta investigação (preservando dados de carácter identificador de clientes e outros considerados confidenciais).

O projecto iniciou-se com um objectivo lato, de âmbito abrangente, com vista à optimização em geral das campanhas. No entanto, a metodologia CRISP-DM encarregou-se de direccionar os esforços da investigação para um objectivo mais concreto, conforme se verá no capítulo 4. É devido ao facto de o próprio CRISP-DM englobar uma fase de definição de objectivos que pode ser revisitada inúmeras vezes que se optou por não se fechar à partida um objectivo e trabalhar nesse sentido apenas. A desvantagem é que as iterações sucessivas podem inviabilizar algum do trabalho efectuado antes. Desta forma, assumiu-se que o planeamento inicial serviria apenas como uma *baseline* inicial que poderia ter de ser refeita. É por esta razão que não são descritos já à partida os dados a analisar (os mesmos começarão por ser analisados na primeira fase da primeira iteração da metodologia CRISP-DM, na secção 4.1.1), uma vez que os mesmos acabaram por ser alterados (quer a nível de instâncias, quer a nível de atributos) no decorrer da investigação.

A próxima secção, 3.4, refere todas as ferramentas utilizadas no âmbito da investigação, e a 3.5 aborda as técnicas de *Data Mining* adoptadas para a fase de modelação. Nos capítulos subsequentes são, em primeiro lugar, descritos os passos da investigação com o evoluir do CRISP-DM (capítulo 4) e os resultados para o modelo final e respectivas análise e conclusões (capítulo 5).

3.4. Ferramentas Utilizadas

Contendo esta investigação diversas etapas de carácter prático, foi necessário recorrer a diferentes tecnologias para servirem de suporte a cada fase. A escolha das diversas soluções tecnológicas adoptadas teve como base o objectivo específico a que se propunha, dependendo da situação concreta, conforme adiante se explica.

No entanto, dada a miríade de soluções que existem actualmente para resolver os mesmos problemas, o enveredar por uma ou outra opção acaba por também ter algumas condicionantes do contexto específico do ambiente em que a investigação decorre (escolha de soluções mais baratas e mais fáceis de ser obtidas) e também de algum cariz pessoal (optar por tecnologias já conhecidas do investigador ou dos seus orientadores), de forma a facilitar a a obtenção de resultados no decurso da investigação, para a qual a tecnologia é apenas um meio.

A principal opção que teria de ser tomada seria a da escolha da tecnologia a utilizar para a extracção de conhecimento, ou seja, para aplicar os métodos de *Data Mining* propriamente ditos, de forma a obter modelos. Essa solução seria também utilizada para avaliar os modelos e validá-los.

Assim, tendo como base as premissas enumeradas atrás, a opção recaiu na biblioteca *rminer*[13] (Cortez, 2010). Esta biblioteca é um pacote *open source*[14] desenvolvido especificamente para *Data Mining* e é instalada no ambiente de programação estatístico R[15], que também é *open source*, e está disponível em diversas versões para vários sistemas operativos. O *rminer* já foi utilizado em vários casos práticos com resultados positivos, conforme referenciado por Cortez e Silva (2008) e por Silva *et al*. (2010).

Como reforço da opção pela ferramenta R, pode-se constatar que a mesma tem aumentado os seus indíces de popularidade e satisfação no que diz respeito à sua utilização para *Data Mining*, nomeadamente comparativamente às ferramentas BMDP, JMP, Minitab, R, R-PLUS, Revolution R, S-PLUS, SAS, SPSS, Stata, Statistica, e Systat, de acordo com o estudo de Muenchen (2010).

[13] Biblioteca de *Data Mining* assente no ambiente R e disponível em http://www3.dsi.uminho.pt/pcortez/rminer.html, acedido em Maio de 2011.

[14] O *software open-source* trata-se de *software* de código aberto, o qual está disponível gratuitamente para qualquer pessoa, podendo não só aceder ao *software*, mas também ao seu código fonte.

[15] O R é uma linguagem e ambiente de programação *open-source* com fins estatísticos, e está disponível em http://www.r-project.org/.

O estudo de Javaheri (2008) reforça a escolha da ferramenta R, uma vez que também foi a escolha para um contexto semelhante, de selecção de clientes para *Marketing* direccionado na banca, tendo o resultado sido bastante positivo, conforme já referido em 2.5.

Adicionalmente, outras tecnologias foram utilizadas. Numa fase inicial de recolha e normalização de dados, quer com origem nos diversos relatórios, quer com origem no ficheiro caracterizador de clientes (que será explicado na secção 4.1.1), foi desenvolvida uma aplicação *Web* em *Java*[16], a qual foi instalada num servidor *Tomcat*[17]. Essa aplicação obtinha a informação das diferentes fontes e aglutinava a mesma numa base de dados relacional, instalada num servidor *mySQL*[18]. Tal permitia a que fosse possível, posteriormente, obter um *flat-file*[19] tal como é requerido pelas aplicações de *Data Mining* (caso do *rminer*), bastando para tal efectuar *queries* à base de dados.

Para apoio à análise de atributos, foi usada a ferramenta *rattle*[20] (Williams, 2009; 2011), que é *open-source* e também está disponível para várias plataformas e sistemas operativos. Esta ferramenta assenta também no ambiente estatístico R e permite a execução de diversas técnicas de *Data Mining*. No entanto, foi usada, sobretudo, pela sua capacidade de sintetizar em gráficos simples a forma como os atributos afectam o resultado das campanhas.

Por último, importa referir que foi utilizada em diversas fases da investigação a ferramenta *Microsoft Excel*[21] quer para visualização do conjunto de dados, quer para pequenos tratamentos de dados (por exemplo, eliminação de caracteres portugueses, não compatíveis com o *rminer*).

3.5. Técnicas de *Data Mining*

Para a fase de Modelação, foram escolhidas três técnicas de *Data Mining* já referidas em 2.4, de crescente nível de complexidade algorítmica: *Naïve Bayes*, Árvores de Decisão (binárias), e Máquinas de Vectores de Suporte. Tal opção foi tomada tendo em conta que a investigação se desenvolveu em torno da metododologia CRISP-DM, com uma redução significativa da quantidade dos dados de entrada e uma melhoria substancial da sua qualidade à medida que se progredia no processo iterativo inerente à própria metodologia. Ou seja, a expectativa inicial era a de que a complexidade dos dados fosse demasiada para avançar logo para uma modelação recorrendo a Máquinas de Vectores de Suporte, daí a opção de executar técnicas mais simples, mas mais rápidas devido a exigências computacionais menores.

[16] *Java* é uma marca registada da *Oracle Corp.* (http://www.oracle.com/technetwork/java/index.html).
[17] *Tomcat* é uma marca registada da *Apache Software Foundation* (http://tomcat.apache.org/).
[18] *MySQL* é uma marca registada da *Oracle Corp.* (http://www.mysql.com/).
[19] Um *flat-file* é um documento estático de texto, contendo vários registos de conjuntos de dados não relacionados entre si.
[20] O *rattle* é uma ferramenta *open-source* assente no ambiente R a qual permite visualizar e transformar dados de forma a facilitar a modelação, e está disponível em http://rattle.togaware.com/.
[21] O *Microsoft Excel* é uma ferramenta de folhas de cálculo, disponível no pacote *Microsoft Office*, sendo ambos marcas registadas da *Microsoft Corp*.

Relativamente à aposta em Máquinas de Vectores de Suporte face a outras alternativas (por exemplo, redes neuronais), a justificação prende-se com os resultados de alguns estudos mais recentes que apontam esta técnica como muito eficiente em casos com algumas características similares. Um dos que mais se assemelha no seu contexto é o de Javaheri (2008), já referido quer em 2.4, quer em 3.4, e cujo modelo foi gerado precisamente recorrendo a esta técnica, com resultados muito satisfatórios (de acordo com a análise *Lift*, 40% das instâncias originais aglutinariam, pela aplicação do modelo, 80% dos resultados positivos). A função de *kernel* utilizada será a gaussiana.

Para todos os modelos, a divisão entre dados de treino e dados de teste foi efectuada dividindo o conjunto em 2/3 para treino e 1/3 para teste. Essa divisão foi efectuada de forma aleatória, não contemplando uma eventual ordenação. No decorrer das iterações da metodologia CRISP-DM verificou-se que se dispunha de um elevado número de instâncias para treino e teste (da ordem das várias dezenas de milhar), pelo que se considerou que 2/3 eram suficientes para a modelação.

Como métricas de avaliação da qualidade dos modelos foram utilizadas a matriz de confusão e medidas associadas, como a exactidão e a sensibilidade, e ainda a curva ROC e a respectiva área debaixo da mesma que, quanto maior, melhor é o modelo na sua capacidade de previsão. Para a matriz de confusão foi assumido que se consideraria como sendo um sucesso, caso o modelo assim o previsse com uma probabilidade acima de 0,5. Para uma comparação entre modelos e uma análise mais exaustiva foi utilizada a curva *Lift*.

Uma vez que se pretende um modelo explicativo, no final foi avaliada a importância de cada atributo de entrada no melhor modelo obtido, isto é, como é que cada um contribuia para a definição do modelo.

4. Trabalho Realizado

A investigação desenrolou-se em torno da metodologia CRISP-DM. Dada a natureza iterativa e cíclica da mesma, optou-se por se definir três iterações para afinar melhor os modelos e resultados obtidos. Considerou-se que cada iteração em si estaria fechada e que, consequentemente, teria de se avançar para uma nova, sempre que se chegava a uma fase em que, para melhorar os resultados obtidos, teria de se retroceder algumas das fases desta metodologia.

As secções subsequentes descrevem os ciclos da metodologia CRISP-DM que foram executados, de acordo com a Figura 2. Desta forma, cada ciclo terá a sua própria secção, devidamente numerada (e. g. 4.X), sendo que cada uma das fases percorridas nessa iteração terá a sua própria sub-secção (e. g. 4.X.Y). Os nomes das seis fases principais, recorde-se, são os seguintes:

1. Compreensão do Negócio;
2. Compreensão dos Dados;
3. Preparação dos Dados;
4. Modelação;
5. Avaliação;
6. Implementação.

Tal como indicado na secção 2.4, o CRISP-DM não é uma metodologia rígida, pelo que em algumas iterações poderão não ser executadas todas as fases. Em cada iteração subsequente pretende-se um maior afunilamento no que se refere a uma definição clara do que se pretende atingir e com que tipo de modelo. Assim, a cada iteração foi dado um nome que sugerisse por si só qual o seu principal enfoque.

4.1. CRISP-DM – Iteração 1: Dados vs. Objectivos

4.1.1. Compreensão do Negócio

Numa fase preliminar torna-se premente garantir a viabilidade do projecto. Desta forma, a metodologia CRISP-DM arrancou com duas tarefas base, as quais estão intimamente interligadas: a determinação dos objectivos e a avaliação de recursos e constrangimentos (da qual deverá resultar um ponto de situação). Caso seja de todo impossível justificar os objectivos de negócio com os recursos disponíveis, o projecto não é viável.

Considerando-se que esta fase é fulcral para direccionar a investigação para o rumo adequado, optou-se por se dividir a mesma, não tanto de acordo com as tarefas do CRISP-DM, mas mais pelas diferentes etapas que decorreram para melhor apurar objectivos tangíveis na distância temporal pré-definida para o projecto. Segue-se assim uma descrição das etapas, devidamente discriminadas em sub-secções diferentes.

4.1.1.1. Dados de Execução do Contacto

Conforme mencionado na secção 3.2, a base de suporte ao estudo são relatórios de execução de contactos no âmbito de campanhas. Tratam-se, portanto, de dados que consubstanciam desde logo o que se pretende analisar, ou seja, o resultado das campanhas, conforme se pode constatar na Tabela 6.

Tabela 6 - Atributos de contacto e os seus tipos

Nome	Descrição e valores possíveis	Tipo
Agente	Identificador do agente que atendeu a chamada	Nominal
Telefone do cliente	Número de telefone do cliente	Nominal
Data e hora do contacto	O *Contact-Center* só funciona entre as 7h00 e a 1h00 do dia seguinte, e nos dias úteis, sendo os contactos de *oubound* despoletados entre as 10h00 e as 23h00	Nominal
Data e hora para agendamento		Nominal
Duração do contacto	Duração do contacto no formato “hh:mm:ss”, em que hh=hora, mm=minuto, e ss=segundos	Numérico

Adicionalmente, algumas campanhas também funcionaram em modo de visualização, isto é, contabilizando quer as vezes que um cliente alvo viu a campanha no canal de *homebanking*, quer as vezes que um agente, que recebendo uma chamada telefónica do cliente alvo, teve a oportunidade de abordar o cliente no contexto da mesma (podendo, no entanto, não o ter feito). Ainda que com âmbitos distintos, conforme explicado na secção 3.2, em ambos os casos os atributos registados são os mesmos e podem ser constatados na Tabela 7.

Tabela 7 - Atributos de visualizações e os seus tipos

Nome	Descrição e valores possíveis	Tipo
Número de visualizações	Contabilização das vezes que o cliente (ou o agente, no caso do canal telefónico e, aquando de uma chamada com o cliente) viu a campanha - [0; ∞[	Numérico
Data e hora da última visualização	Data e hora da última visualização (em que visualização tem o mesmo significado que no caso do atributo “Número de visualizações”)	Nominal

Todos os atributos caracterizadores de contactos (incluindo as visualizações) só são conhecidos após o arranque de cada campanha, o que inviabiliza o seu uso para a implementação de um modelo preditivo de forma a seleccionar à partida os clientes a serem alvos de cada campanha.

No entanto, o conhecimento da influência que cada atributo tem no resultado final pode ser utilizado pelos gestores de *Marketing* para condicionar o decurso da campanha em

direcção aos objectivos do negócio. Desta forma, fica assumido já à partida que o modelo a obter será um modelo explicativo do funcionamento das campanhas, devendo o conhecimento obtido ser transmitido aos gestores na forma da influência que os atributos têm para o resultado a atingir.

4.1.1.2. Dados Caracterizadores de Clientes

Foi identificada uma lacuna referente aos dados disponíveis: os mesmos diziam apenas respeito a informação de contacto (com excepção do número de telefone, que é caracterizador do indivíduo – o cliente não muda de número de telefone a cada contacto). Desta forma, tornava-se necessário enriquecer a qualidade e quantidade de dados, de modo a poder alimentar as técnicas de *Data Mining* com informações relevantes acerca dos clientes. Foi solicitada essa informação à Direcção de *Marketing*, tendo sido fornecidas as características do cliente indicadas na Tabela 8.

Tabela 8 - Atributos de clientes e os seus tipos

Nome	Descrição e valores possíveis	Tipo
Data de nascimento	Data no formato aaa/mm/dd (aaaa=ano, mm=mês e dd=dia)	Nominal
Profissão	São 1726 os valores possíveis	Nominal
Situação/categoria profissional	Situação laboral (são 27 os valores possíveis)	Nominal
Estado civil	Casado, Divorciado, Separado, Solteiro e Viuvo	Nominal
Morada	Dados referentes à morada de residência	Nominal
Morada complementar		Nominal
Localidade		Nominal
Código postal		Nominal
Código postal local		Nominal
Freguesia		Nominal

Estes dados foram fornecidos para 220.989 clientes através de um ficheiro CSV com uma linha por cliente. Uma das verificações residiu em saber se os dados abrangiam todos os clientes referidos nos relatórios de campanhas, algo que se confirmou, ou seja, para todos os clientes alvo de campanhas havia dados caracterizadores do cliente enquanto indivíduo.

Ainda que alguns desses atributos pudessem ter um valor vazio, era importante que existissem tantos registos nesse ficheiro de dados de clientes quantos os clientes abrangidos pelas campanhas. Considera-se que esta verificação é fundamental para garantir a obtenção de modelos adequados de *Data Mining* e, consequentemente, enquadra-se na tarefa do CRISP-DM de avaliação de recursos, pelo que foi efectuada nesta fase e não na fase de compreensão dos dados.

4.1.1.3. Os Diversos Objectivos de Negócio

De acordo com a metodologia CRISP-DM (Chapman *et al.*, 2000), a primeira fase é a compreensão do negócio. Só terminada esta se pode transitar para a fase de compreensão dos dados, a qual providenciará o suporte para uma melhor identificação dos objectivos.

No que diz respeito à definição dos objectivos de negócio, havia que efectuar uma análise das campanhas correspondentes aos relatórios obtidos. Assim, procedeu-se a um levantamento inicial das campanhas a analisar, elaborando uma listagem de todas no sentido de identificar similaridades entre elas. A listagem total das 72 campanhas é apresentada em detalhe na Tabela 25 do anexo A.

Após analisar as características das campanhas, em especial os objectivos de negócio a atingir por cada uma, conclui-se que uma campanha tem um intervalo temporal perfeitamente definido, ainda que a campanha possa ser replicada/clonada para ser reutilizada no futuro, com o mesmo objectivo. Por exemplo, uma campanha para venda de um depósito a prazo para uma selecção específica de clientes pode ter decorrido no mês de Junho de 2008 e depois ser criada uma nova campanha exactamente nos mesmos moldes, mas com outra lista de clientes (eventualmente com alguns repetidos) para ser executada noutro mês, eventualmente através de outro canal. No entanto, o objectivo de negócio é o mesmo, pelo que, por exemplo, um cliente que tenha visto um produto na página do *homebanking* na primeira campanha pode-se lembrar do mesmo na campanha seguinte que incidiu sobre o mesmo produto.

Adicionalmente, os fluxos de diálogo para os canais telefónicos são repetidos, com poucas alterações. Desta forma, pode-se extrapolar, dos resultados de campanhas anteriores para um determinado produto, de que forma uma nova campanha se comportará face a um conjunto diferente de clientes (este e o período em que a campanha decorre são as grandes variáveis entre uma campanha e outra). O *feedback* obtido da área de negócio permite suportar esta análise. Como resultado desta primeira fase de compreensão do negócio, as campanhas foram agrupadas por grupo e por produto/serviço, caso aplicável, conforme se pode constatar na Tabela 9.

Como os dados são apenas os provenientes dos relatórios (com excepção dos caracterizadores de clientes) e o objectivo de cada contacto é atingir o sucesso, medido a partir do resultado indicado no relatório respectivo, só será possível avaliar dados para campanhas em que tenha sido emitido o relatório de resultados. Ou seja, ainda que, por exemplo, uma campanha tenha sido criada especificamente apenas para divulgar um produto através do *homebanking*, o seu resultado não pode ser calculado apenas com os dados disponíveis nos relatórios referidos, uma vez que apenas são contabilizadas as vezes que cada cliente viu a página *Web*. Desta forma, todas as campanhas para as quais apenas existisse relatório de visualizações tiveram de ser excluídas deste estudo.

Tabela 9 - Objectivos de negócio das campanhas

Grupo	**Produto ou Serviço**
Aplicação Mista	DM X
Apoio Domiciliário	CR
Associados	Validação de Dados
Cartão de Crédito	CC X
Cartão de Crédito	CC Y
Cartão de Crédito	CC Z
Cheques	Aviso Cheques por levantar
Crédito Individual Pré Aprovado	
Depósito a Prazo	DP U
Depósito a Prazo	DP W
Depósito a Prazo	DP X
Depósito a Prazo	DP Y
Depósito a Prazo	DP Z
Gestores de Cliente	Inquérito de Satisfação
Informação sobre clausulado	particulares
Obrigações de Caixa	OC X
PPR	Reforço
Reactivação de Clientes	
Recuperação de Crédito	
Seguro Específico	SE X
Seguro Pessoal	SP X
Seguro Viagem	Inquérito de Satisfação

Nota: para confidencialidade, os nomes dos produtos e serviços foram mascarado

Posto isto, das 72 campanhas inicialmente em estudo ficaram excluídas sete campanhas. Os objectivos de negócio distintos que se pretendiam atingir com esta investigação eram 22, havendo várias campanhas que se repetem no tempo para o mesmo objectivo.

Uma vez que o resultado de cada contacto é um de entre os enumerados na Tabela 5, todos os problemas a analisar são de classificação. Adicionalmente, nos casos de venda de produtos e em que haja uma resposta positiva ao *output* de classificação (ou seja, se o resultado do contacto for um sucesso), trata-se também de um problema de regressão: analisar quais os montantes subscritos e definir modelos que os consigam prever.

4.1.1.4. O Objectivo de Negócio Alvo

Considerando que a existência de vários objectivos de negócio implicaria aplicar a metodologia CRISP-DM tantas vezes quantos os objectivos (que são completamente distintos entre si), torna-se natural a escolha de um único objectivo para este estudo. Tanto mais que tal permite um maior aprofundamento na aplicação da metodologia CRISP-DM para obtenção de conhecimento, aumentando sobejamente a probabilidade de conduzir a resultados mais apurados para esse objectivo.

Um dos critérios óbvios a ter em conta para a escolha desse objectivo seria a quantidade de dados disponíveis nos relatórios: quanto maior, maior a probabilidade de se obter bons modelos (Witten e Frank, 2005). Tendo em conta tal facto, o grupo de campanhas predominante no conjunto de dados disponível diz respeito a depósitos a prazo, quer em quantidade de registos, quer em campanhas e sua dispersão temporal. No entanto, vários produtos de aforro foram alvo das diferentes campanhas de depósitos a prazo. Tendo em conta que os depósitos a prazo têm características muito similares (capital garantido, taxa e prazo pré-definidos[22]) e, dada a concorrência elevada que tem conduzido a taxas similares, a opção para os clientes bancários põe-se, acima de tudo, entre a escolha de um depósito a prazo ou outro produto de aforro com maior risco e, eventualmente, maior retorno.

Desta forma, considerou-se que o objectivo principal seria a subscrição de um depósito a prazo, independentemente de qual o produto. Naturalmente que o tipo de produto poderia também ter alguma influência no resultado. No entanto, para a investigação subsequente e, dado o exposto atrás, assume-se que essa influência é negligenciável face a outras condicionantes. Por conseguinte, o atributo de *output* que representava o valor a ser previsto era o resultado do último contacto, o qual poderia ser um dos definidos na Tabela 5. Assim, os dados a servirem de base são constituídos por dezassete campanhas, correspondentes a cinco depósitos a prazo distintos, sendo o número de instâncias independentes igual a 79354.

4.1.1.5. Novas Considerações sobre os Dados de Clientes

Após obter os dados caracterizadores de cliente indicados acima, considerou-se que os mesmos poderiam ser insuficientes para auxiliar na construção de modelos adequados. Mais importante ainda, faltavam dados caracterizadores do cliente não tanto enquanto indivíduo, mas como cliente bancário.

Assim, foram solicitados novos dados à Direcção de *Marketing*. Depois de verificar o que era possível ser fornecido, foram obtidos os dados adicionais (face aos já indicados na Tabela 8) referidos na Tabela 10.

[22] De acordo com http://economiafinancas.com/2009/03/o-que-e-um-deposito-a-prazo/ (acedido em Agosto de 2011).

Verificou-se para este novo conjunto de dados que abrangia todos os clientes alvo das campanhas, ainda que alguns dados específicos pudessem não vir preenchidos para alguns clientes.

Tabela 10 - Atributos de cliente adicionais e os seus tipos

Nome	Descrição e valores possíveis	Tipo
Título honorífico	22 valores possíveis	Nominal
Bloqueios gerais	Indicador da existência de bloqueios bancários gerais sobre o cliente bancário - (S)im ou (N)ão	Nominal (binário)
Bloqueios informativos	Indicador da existência de bloqueios informativos - (S)im ou (N)ão	Nominal (binário)
Bloqueio de cheques	Indicador da existência de bloqueios para a utilização de cheques - (S)im ou (N)ão	Nominal (binário)
Inibição de cheques	Indicador da inibição total na utilização de cheques - (S)im ou (N)ão	Nominal (binário)
Associado	O cliente é associado da associação mutualista da instituição? - (S)im ou (N)ão	Nominal (binário)
Sexo	(M)asculino ou (F)eminino	Nominal (binário)
Habilitações literárias	Inferior ao 4º Ano, 4º Ano, Ciclo, 9º Ano, 12º Ano, Curso Médio, Curso Superior, Outro, Desconhecidas	Nominal
Créditos em mora	Indicador da existência de prestações de créditos em mora - (S)im ou (N)ão	Nominal (binário)
Saldo médio anual	Saldo médio anual das contas à ordem das quais o cliente é titular	Numérico
Cartão débito	Indicador da existência de bloqueios para a utilização de cheques - (S)im ou (N)ão	Nominal (binário)
Conta ordenado	O cliente tem conta ordenado? - (S)im ou (N)ão	Nominal (binário)
Cartão crédito	O cliente tem cartão de crédito? - (S)im ou (N)ão	Nominal (binário)
Crédito habitação	O cliente tem crédito habitação? - (S)im ou (N)ão	Nominal (binário)
Crédito individual	O cliente tem crédito individual? - (S)im ou (N)ão	Nominal (binário)
Domiciliações	O cliente tem domiciliações? - (S)im ou (N)ão	Nominal (binário)

4.1.1.6. Alguns Resultados Obtidos nesta Fase

Nesta fase crucial de compreensão do negócio com vista à definição de objectivos foram obtidos dois *outputs* chave para todo o restante estudo: por um lado, foi fixado o objectivo em definir um modelo explicativo do comportamento das campanhas para a subscrição de produtos a prazo; por outro e, decorrente do objectivo, concluiu-se que o atributo a prever seria o resultado do último contacto, o qual traduziria a eficácia do contacto.

4.1.2. Compreensão dos Dados

Na fase anterior verificou-se que existem os recursos que, à primeira vista, adequam-se ao objectivo definido: obtenção de um modelo de optimização de contactos. Assim, de acordo com a metodologia CRISP-DM, segue-se a fase de analisar exaustivamente todos os dados disponíveis, dados de contacto no âmbito da campanha e dados de cliente, para poder identificar eventuais necessidades ou lacunas.

4.1.2.1. Dados de Contacto

Relativamente aos dados de contacto, os mesmos foram extraídos directamente dos relatórios de resultados. No entanto, cada contacto propriamente dito correspondia a um registo no ficheiro de relatórios. Assim, se um mesmo cliente, no âmbito da mesma campanha, fosse contactado, por exemplo, quatro vezes e, admitindo que o resultado de cada contacto fosse um novo agendamento (o qual pode ter diversas razões), ter-se-iam quatro registos com informação distinta, ainda que o contexto fosse o mesmo (mesmo cliente e campanha).

Com o exposto e, tendo em conta que as ferramentas de *Data Mining* habituais recebem como *input* um ficheiro com um conjunto de dados constituído por registos (linhas) independentes entre si, era necessário condensar numa única linha toda a informação dos diversos contactos realizados para o mesmo cliente e campanha. Para tal, optou-se por guardar a informação do primeiro e último contacto, descartando toda a informação de contactos intermédios. Adicionalmente, criou-se um atributo para registar o número total de contactos efectuados com o cliente.

Considerou-se que a informação perdida poderia ser descartada com base em algumas premissas. Em primeiro lugar, o contacto final, o qual seria o último, acabava por ser o mais relevante, uma vez que traduzia todo o esforço de todos os contactos num resultado final. Para os restantes contactos, a importância relativa de cada um já era alvo de uma maior discussão. No entanto, o primeiro contacto acabava por ser a introdução que o cliente tinha à campanha. Para além de que só existiriam contactos subsequentes, caso existisse o primeiro contacto, pelo que optou-se por se manter também a informação deste contacto inicial.

Adicionalmente, dos 79354 registos finais (pares cliente/campanha) que resultaram da conversão efectuada de acordo com a Tabela 11, apenas 27870 tinham mais de dois contactos, pelo que não se perde nenhuma informação para cerca de 65% dos casos.

Desta forma, os dados de relatórios de contactos foram desdobrados do modo que se pode constatar na Tabela 11.

Tabela 11 - Conversão de registos de relatórios de contactos

Cliente / Campanha	**N.º Contacto**
X / Depósito A	x1
X / Depósito A	x2
X / Depósito A	x3
X / Depósito A	x4
X / Depósito A	x5
Y / Depósito A	y1
Y / Depósito A	y2

Cliente / Campanha	**Contacto**		**N.º de contactos**
	1º	**Último**	
X / Depósito A	x1	x5	5
Y / Depósito A	y1	y2	2

Refira-se ainda que os relatórios de resultados contêm elementos associados ao próprio resultado da operação (*outputs*), nomeadamente o montante subscrito para o depósito a prazo, caso o resultado seja um sucesso, e ainda a duração do último contacto. Obviamente que o montante subscrito (atributo numérico com o valor em euros) não pode ser usado para alimentar as técnicas de *Data Mining*, pois tal só é conhecido após o contacto ser efectuado, ou seja, o seu estudo constituiria um problema de regressão, pelo que foi excluído.

Relativamente aos dados dos relatórios de visualizações, os atributos que os mesmos disponibilizavam já se resumiam a um registo por cliente e campanha, pelo que puderam ser transpostos directamente para o ficheiro de *input* a ser usado nas técnicas de *Data Mining*.

4.1.2.2. Dados de Cliente

Nesta fase, urgia trabalhar sobre o ficheiro de dados caracterizadores de clientes no sentido de aferir a qualidade dos dados e a sua adequação para serem adicionados aos dados dos contactos de campanhas.

Uma boa parte dos dados (13 de entre 21) obtidos eram nominais binários, tipicamente com um de dois valores: (S)im ou (N)ão, e não existiam *missing values*. Assim, a expectativa é a de que estes dados possam enriquecer os modelos criados sem aumentar em demasia a sua complexidade, uma vez que têm apenas duas categorias.

Conforme já foi indicado atrás, todos os clientes foram caracterizados por dados do ficheiro de clientes, ainda que existissem alguns *missing values* para alguns dos atributos – caso do saldo médio anual que tem 14746 valores desconhecidos – havendo outros que, não tendo um valor omisso implícito, têm um valor omisso explícito, ou seja, foi-lhes atribuído um valor pela entidade bancária para representar o desconhecimento – casos do estado civil (valor “OMISSO NO DOCUMENTO”) ou do título honorífico (valor “SEM TITULO HONORIFICO”).

Alguns dos atributos nominais não binários têm demasiados valores possíveis, o que resulta numa dispersão elevada e que poderá dificultar a utilização destes atributos por parte

das técnicas de *Data Mining* aquando da construção de modelos – por exemplo, existem 1726 profissões diferentes e 2377 freguesias de residência.

Considera-se ainda que a qualidade dos dados é suficientemente boa no que diz respeito a *missing values* para que não seja necessário considerar técnicas de substituição como as referidas em 2.4.

4.1.2.3. Histórico de Contactos

Uma vez que se dispõe de informação de vários contactos realizados ao longo do tempo para cada campanha e, podendo o mesmo cliente ter sido abrangido por várias campanhas, poderá existir informação válida de contactos passados (caso o cliente tenha sido contactado anteriormente) e que poderá ser muito útil para alimentar as técnicas de *Data Mining* e assim obter modelos mais fiáveis e robustos na sua capacidade de previsão.

Desta forma, afigurava-se como adequada a criação de informação de "memória" de modo a manter este histórico, contendo o mesmo um resumo dos contactos efectuados no passado com o mesmo contexto (mesmo cliente e campanhas efectuadas anteriormente).

Com base na informação descrita na Tabela 6, foram criados vários atributos para resumir contactos anteriores, condensando o resultado dos mesmos num único registo de vários atributos. Conforme está patente na Tabela 12, a maioria trata-se de contadores como, por exemplo, os números de sucessos e insucessos anteriores, ou de somatórios como, por exemplo, o montante total subscrito anteriormente em depósitos a prazo no âmbito de campanhas.

Tabela 12 - Atributos de histórico de contactos e os seus tipos

Nome e descrição	**Tipo**
N.º de dias desde o último contacto para qualquer campanha	Numérico
N.º de dias desde o primeiro contacto para qualquer campanha	Numérico
N.º total de contactos em campanhas anteriores	Numérico
N.º total de sucessos anteriores	Numérico
N.º total de insucessos anteriores	Numérico
Resultado do último contacto para a última campanha – Um dos valores apresentados na Tabela 5	Nominal
Montante subscrito (em euros) na última campanha	Numérico
Montante total subscrito para todas as campanhas	Numérico
N.º total de visualizações para todas as campanhas no *homebanking*	Numérico
N.º total de visualizações para todas as campanhas na banca telefónica	Numérico

No entanto, a maioria dos registos (65339) correspondem a contactos para clientes que nunca haviam sido contactados no passado no âmbito de campanhas de depósitos a prazo.

Assim, apenas se possui histórico para 14015 dos registos de contactos de clientes (17,7%), o que poderá limitar a sua utilidade na modelação.

4.1.2.4. Agrupamento dos Atributos

Nesta fase foram analisados os dados disponíveis e verificada a sua adequação para servirem de *input* às técnicas de *Data Mining*. De um modo geral, considera-se que os dados têm uma boa qualidade – foram todos fornecidos pela mesma Direcção da instituição bancária, havendo poucos casos de *missing values*. Toda a informação que irá servir como *input* às técnicas de *Data Mining* organiza-se em vários registos, sendo os atributos agrupados consoante o tipo de informação que contêm, conforme pode ser verificado na Tabela 13.

Tabela 13 - Agrupamento dos atributos

Grupo	**Tipo**	**Detalhe na:**	**Total de atributos**
Identificadores de Cliente, Campanha e Objectivo			3
Informação de Cliente	Pessoal	Tabela 8 e Tabela 10	13
	Bancária		13
Informação de Execução da Campanha	Primeiro Contacto	Tabela 6	5
	Número de Contactos		1
	Último Contacto		5
	Informação de Visualização	Tabela 7	4
Informação de Histórico de Contactos		Tabela 12	10
			54

4.1.3. Preparação dos Dados

Uma vez que as ferramentas que irão ser usadas para construir os modelos de *Data Mining* necessitam de um *input* único constituído por um ficheiro de texto em que cada linha corresponde a um registo independente de uma ocorrência (cada linha, no caso vertente, é um resumo dos contactos efectuados no âmbito de um mesmo contexto de campanha e para o mesmo cliente), toda a informação disponível e analisada na fase anterior foi transferida para um flat-file com cada registo independente dos demais, tendo cada registo como chave um par <cliente, campanha>.

Adicionalmente, era necessário extrapolar alguns dados mediante os já existentes. Mais especificamente, tendo a data de nascimento, fazia sentido obter a idade do cliente (em anos) à data do contacto efectuado. A data de nascimento por si só pode não ser uma grande ajuda para a obtenção de bons modelos na sua capacidade de previsão, mas a idade possibilita às técnicas a tentativa de agrupar os clientes de acordo com a sua faixa etária de modo a prever o seu comportamento.

Relativamente ao atributo da data e hora de contacto (e também do agendamento) referido na Tabela 6 que, como campo único, não teria muita utilidade, dado que, provavelmente, é diferente para todas as instâncias, foi desdobrado em vários atributos, que se encontram descritos na Tabela 14.

Tabela 14 - Atributos extraídos da data e hora de contacto e os seus tipos

Atributo	Descrição e valores possíveis	Tipo
Dia da semana	Dia da semana em que o contacto foi feito – *Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, Sunday*	Nominal
Dia do mês	1 a 31, consoante o mês	Nominal
Mês	De 1 a 12 (Janeiro a Dezembro)	Nominal
Hora	De 7 a 22 (o *Contact-Center* só funciona entre as 7h00 e a 1h00 da manhã, sendo os contactos de *oubound* despoletados entre as 10h00 e as 23h00) – valor inteiro, ou seja, se o contacto foi feito às 9h49, o valor para este atributo é 9	Nominal

O identificador do cliente, bem como a sua morada (atributos morada e morada complementar) foram descartados por serem características específicas do indíviduo. Da mesma forma, o número do telefone do cliente para onde foi efectuada a chamada também acaba por não ter muito valor para as técnicas de *Data Mining*, uma vez que é praticamente um identificador do cliente (a não ser no caso de telefones partilhados, por exemplo, o número de casa, que é utilizado por todos os elementos da família). No entanto, neste caso, extraíu-se o tipo de telefone através do prefixo do número, ou seja, se começasse por "9" então era um telefone móvel, caso contrário, um telefone fixo. A duração do contacto que, pelo seu formato, descrito na Tabela 6, não seria devidamente utilizada pelas técnicas de *Data Mining*, dado ser uma cadeia de caracteres de formatação complexa, foi também convertida num atributo numérico simples, a duração do contacto em segundos.

Para levar a cabo todas as acções de processamento mencionadas, foi desenvolvida uma aplicação específica, a qual se encontra descrita no anexo B.

Assim, a base de informação para modelação nesta fase tem 79354 instâncias, 6499 das quais correspondentes a sucessos, 60 atributos de entrada, e 12 categorias possíveis para resultados de contacto[23].

4.1.4. Modelação

Nesta fase procede-se à execução das diversas técnicas de *Data Mining* implementadas pela biblioteca que irá ser utilizada, o *rminer*. Tratando-se de uma primeira

[23] A Tabela 26 do Anexo C apresenta uma síntese da informação obtida nesta fase e condensada no ficheiro de dados. O mesmo apresenta uma estrutura conceptual de acordo com o já descrito em 4.1.2.4.

iteração no método CRISP-DM, ir-se-ia agora efectuar os primeiros ensaios com a ferramenta R e o *rminer*.

As técnicas seleccionadas foram as seguintes:

- NB – *Naïve Bayes*;
- DT – Árvores de Decisão;
- SVM – Máquina de Vectores de Suporte.

Uma vez que se trata da primeira iteração, em que não foi efectuado um tratamento exaustivo dos dados e se pretende efectuar mais iterações para melhorar os resultados, foi efectuada apenas uma tentativa de execução de cada uma das técnicas acima indicadas.

Segue-se abaixo um excerto do código executado no ambiente estatístico R para a obtenção (treino) e teste do modelo NB. As funções específicas do *rminer* estão sublinhadas.

```
DF<-read.table("DADOS_OBJECTIVO_DP_iter1.csv",sep=";",header=TRUE) # leitura de dados
AT=c(2,3,6,7,8,9,10:25,27,29:64,5) # lista de índices de atributos a seleccionar
DF=DF[,AT] # ficar apenas com os dados referentes aos atributos seleccionados

# execução de modelo e teste do mesmo
MNB=mining(FINAL_RESULTADO~.,DF,method=c("holdout",2/3),model="naivebayes",Runs=1)
# execução de métricas
# o parâmetro TC é a Target Class (Objectivo) – 12=Sucesso
print(mmetric(MNB,metric="CONF",TC=12)) # obter a matriz de confusão
print(mmetric(MNB,metric="AUC",TC=12)) # a área debaixo da curva ROC
print(mmetric(MNB,metric="ACC",TC=12)) # exactidão referente à matriz de confusão
print(mmetric(MNB,metric="TPR",TC=12)) # sensibilidade referente à matriz de confusão

print(summary(DF)) # sumário de dados, com métricas para cada atributo
cat("linhas:",nrow(DF),"colunas:",ncol(DF),"\n") # dimensão da tabela de dados
savemining(MNB, "mining1_nb.output",ascii=TRUE) # guardar o resultado da execução
```

A função *mining* é a mais relevante, uma vez que executa a técnica referida no parâmetro *model*, obtendo um modelo com base nos dados indicados no parâmetro *method*, e efectuando testes de validação desse modelo. Este processamento pode ser executado tantas vezes quantas as indicadas no parâmetro *Runs*. Com uma selecção aleatória de dados para treino e para teste, é assim possível efectuar uma validação cruzada (embora não tenha sido o caso nesta fase, uma vez que apenas foi efectuada uma execução – *Runs*=1). O resultado é uma estrutura de dados complexa com informação sobre os resultados da execução. Como se trata de um processamento moroso e complexo, é possível guardar os resultados para posterior análise através da função *savemining*.

Relativamente à execução nesta fase de modelação, a única técnica para a qual foi possível obter um modelo foi a NB. Quer para DT, quer para SVM, verificou-se que o seu processamento não terminava ao fim de várias horas (mais de 12, para ambos os casos), isto

efectuando a execução num computador relativamente recente, com processador de dois núcleos e dois *gigabytes* de memória física, com o sistema operativo *Microsoft Windows XP*©.

4.1.5. Avaliação

As métricas para os testes de validação efectuados com o único modelo obtido, o NB, são as indicadas na Tabela 15.

Tabela 15 - Matriz de confusão e métricas (NB - 1ª iteração) – amostra de teste

Real \ Previsto	Insucesso	Sucesso
Insucesso	24019	270
Sucesso	1751	416

AUC = 0,7759665
ACC = 0,9236090
TPR = 0,1919705
FPR = 0,0111161

Legenda: AUC = área da curva ROC (*Area Under the ROC Curve*), ACC = exactidão (*Accuracy*), TPR = sensibilidade (*True Positive Rate*), FPR = taxa de falsos positivos (*False Positive Rate*)

Nota: uma vez que haviam várias categorias de resultados distintos, considerou-se para a obtenção da matriz de confusão e respectivas métricas aqui apresentadas que o resultado a atingir era um sucesso e todos os restantes seriam fracassos (ou insucessos).

Este primeiro modelo, que não foi sujeito a uma validação cruzada no sentido de minimizar uma eventual má selecção de dados para treino, apresenta um bom desempenho, conseguindo um valor para a AUC de 0,776, bastante acima dos 0,5 de um modelo aleatório.

4.1.6. Sumário

Nesta iteração foi dado um especial enfoque na definição concreta de objectivos, de forma a alavancar o resto do estudo em torno de um propósito conciso. Assim, não é de estranhar que a fase mais trabalhosa tenha sido a primeira, precisamente a compreensão do negócio e a consequente definição de objectivos. Posteriormente, verificou-se a adequação dos dados disponíveis para suportar este estudo de uma forma sólida, nas fases de análise e tratamento de dados.

Ao se avançar para a execução de modelos, verificou-se que o volume e qualidade de dados não permitiam, nesta iteração, obter ainda modelos com técnicas diferentes (e, eventualmente, com melhor qualidade). Concluiu-se que, efectivamente, a informação que estava a ser fornecida como *input* continha vários dados que dificultavam a construção de modelos (provavelmente porque nem todos esses dados serão úteis, levando a que as diversas técnicas “dispersassem” na procura de padrões que pudessem ser traduzidos em conhecimento), pelo que os dados necessitariam de ser trabalhados, nomeadamente, descartando quer na vertical (atributos) quer na horizontal (registos). Foi com esta premissa que se avançou para uma nova iteração, de acordo com a metodologia CRISP-DM, retrocedendo nas diversas fases cíclicas.

4.2. CRISP-DM – Iteração 2: Objectivo "Subscrição do Depósito"

4.2.1. Compreensão do Negócio

Decorrente da iteração anterior, urgia direccionar os esforços do estudo numa convergência maior em torno do objectivo de subscrição de depósitos por parte dos clientes contactados. A constatação de que a obtenção de modelos era difícil recorrendo a tecnologia que, embora recente, já deu provas da sua eficácia noutros projectos de investigação, conduziu a uma reflexão sobre a utilidade da informação com vista ao objectivo proposto.

Com os pressupostos expostos no parágrafo anterior em mente prosseguiu-se a uma reavaliação dos objectivos de negócio.

Na iteração anterior, o objectivo era prever qual o resultado final decorrente dos contactos a um cliente no âmbito de uma só campanha. No entanto, uma análise mais detalhada ao resultado do último contacto para campanha e cliente conduz a alguns números inesperados: dos 79354 registos, existem 23537 que não são resultados terminais sequer (caso dos agendamentos de chamadas que acabaram por não ser efectuados por diversas razões – decisão de gestão ou, tipicamente, a campanha terminou sem que a chamada pudesse ser realizada), ou que correspondem a erros que impossibilitaram de todo que o contacto fosse concretizado. Ora em termos de negócio, o objectivo é a subscrição do produto (resultado = sucesso); só se pode assumir que esse objectivo não é, de todo, atingível quando o cliente diz explicitamente que não pretende subscrever o produto (resultado = insucesso). Desta forma, considera-se que os resultados de sucesso e insucesso são resultados conclusivos; já os restantes resultados terminais considera-se que são cancelados. Os totais agrupados de acordo com o tipo de resultado para o último contacto (conclusivo, cancelado ou agendamento) podem ser consultados na Tabela 16.

Tabela 16 - Total de resultados por grupo

Resultado	Total	Total por Grupo
Sucesso do Contacto	6499	55817
Insucesso	49318	(conclusivos)
Não é o dono	1011	8365
Não Atendeu Chamada	5091	(cancelados)
Fax em vez de telefone	151	
Contacto de *outbound* abandonado	2059	
Abortado porque o agente efectuou um *cleanup*	53	
Agendamento Outros	9640	15172
Agendamento Gravador	231	(agendamentos)
Agendamento Decisor - Prod. Não Apresentado	1763	
Agendamento Decisor - Prod. Apresentado	2916	
Agendamento Decisor	622	

Uma vez que, do ponto de vista do negócio, é difícil definir em que resultariam os contactos que, por uma ou outra razão, não puderam ser concluídos, pode ser reformulado o objectivo de modo a considerar apenas a previsão de resultados conclusivos.

Estes constituem um total de 55817 registos, um número ainda assim considerável para ser utilizado como *input* das técnicas de *Data Mining*. O número de atributos foi também reduzido uma vez que, ao se excluir os agendamentos como resultados finais de contactos, deixou de fazer sentido considerar os atributos relacionados com a informação sobre a data e hora do agendamento (o valor destes atributos era sempre vazio).

4.2.2. Compreensão dos Dados

A expectativa é a de que o facto de se ter reduzido o objectivo à previsão de um atributo nominal binário (ou é sucesso ou é insucesso), juntamente com a redução do conjunto de dados, possam contribuir para a obtenção de melhores modelos, facilitando assim a execução das técnicas de *Data Mining* que, na iteração anterior, se revelou algo limitativa.

Assim, nesta iteração a opção foi a de não efectuar nenhuma análise e tratamento de dados adicional e avançou-se para a modelação, de forma a verificar o impacto desta redefinição de objectivos, a qual, por si só, já resultou numa alteração radical do conjunto de dados a analisar.

Esta decisão de omitir fases é suportada na definição da metodologia CRISP-DM (Chapman *et al*., 2000), em que é consubstanciado que, na Figura 2, as setas interiores representam as dependências mais usuais, mas não obrigatórias, entre fases.

No entanto, dois dos atributos utilizados na iteração anterior surgem desde logo como bons candidatos a serem excluídos para a modelação, uma vez que são identificadores com pouca variação: a identificação do objectivo e a identificação da campanha.

Desta forma, avança-se para a modelação com um conjunto de dados de 55817 registos, 6499 dos quais correspondentes a sucessos, e 54 atributos, com dois valores possíveis para o atributo referente ao resultado do contacto.

4.2.3. Modelação

Os modelos foram obtidos recorrendo à função *mining* e efectuando 20 execuções da técnica com uma selecção aleatória dos dados de 2/3 para treino e 1/3 teste, conforme já referido também na secção 3.5. A execução de inúmeras modelações e respectivos testes permitem efectuar uma validação cruzada, minimizando o impacto de uma selecção de dados não exemplificativos do universo.

Para esta segunda iteração, apesar de se ter diminuído a complexidade dos dados (menos registos e o atributo correspondente ao resultado a prever reduzido a nominal binário), verificou-se que, para a ferramenta e biblioteca utilizada (*rminer*), os dados continuavam a ser

demasiado complexos para que se pudesse obter um modelo com recurso a Máquinas de Vectores de Suporte (SVM). Ainda assim e, como complemento ao modelo que se obteve através de *Naïve Bayes* (NB), já foi possível obter um modelo através de árvores de decisão (DT), pelo que os resultados de ambos serão analisados na próxima fase, de acordo com a metodologia CRISP-DM.

4.2.4. Avaliação

Os resultados das validações dos modelos NB e DT podem ser verificados na Tabela 17 e Tabela 18, respectivamente.

Tabela 17 - Matriz de confusão e métricas (NB - 2ª iteração) – amostra de teste

Real \ Previsto	Insucesso	Sucesso
Insucesso	315061	13739
Sucesso	27506	15834

AUC = 0,8298096
ACC = 0,8891681
TPR = 0,3653438
FPR = 0,0417853

Tabela 18 - Matriz de confusão e métricas (DT - 2ª iteração) – amostra de teste

Real \ Previsto	Insucesso	Sucesso
Insucesso	311623	17177
Sucesso	26913	16427

AUC = 0,7635523
ACC = 0,8815231
TPR = 0,3790263
FPR = 0,0522415

Legenda: AUC = área da curva ROC (*Area Under the ROC Curve*), ACC = exactidão (*Accuracy*), TPR = sensibilidade (*True Positive Rate*), FPR = taxa de falsos positivos (*False Positive Rate*)

É de realçar que os números apresentados nas matrizes de confusão são o somatório do resultado dos testes para as 20 execuções. Já as métricas AUC, ACC, TPR e FPR são as médias de cada um dos tipos de métricas para as 20 execuções.

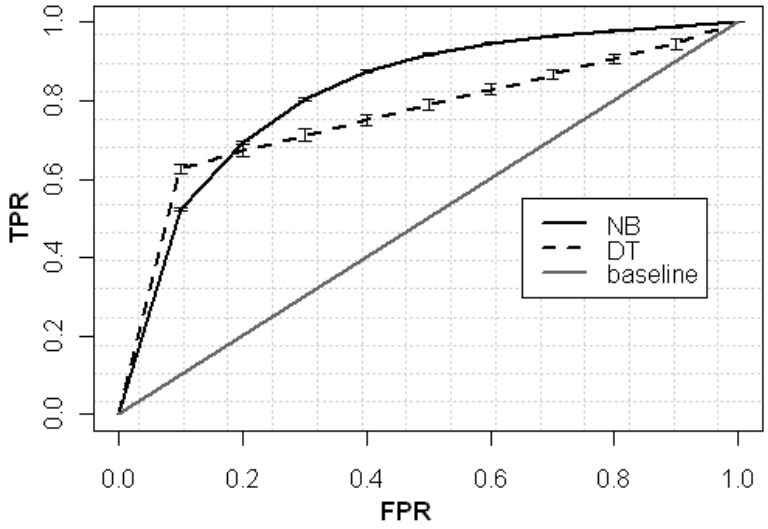
À partida e, a julgar pelos valores das métricas, o modelo NB obteve ligeiramente melhores resultados do que o DT. A AUC é claramente superior no caso do NB (0,83 face a 0,76 para o DT). Para uma comparação mais clara entre ambos os modelos, pode ser executada a curva ROC recorrendo-se à função *mgraph* da biblioteca *rminer*. Segue-se abaixo o excerto de código R executado para a obtenção das curvas ROC num mesmo gráfico a partir das execuções (a sublinhado estão as funções específicas do *rminer*).

```
# executar modelação e testes, com cross-validation (Runs=20)
MNB2=mining(FINAL_RESULTADO~.,DF,method=c("holdout",2/3),model="naivebayes",Runs=20)
MDT2=mining(FINAL_RESULTADO~.,DF,method=c("holdout",2/3),model="dt",Runs=20)

# desenhar as curvas ROC num mesmo gráfico
L=vector("list",2); L[[1]]=MNB2; L[[2]]=MDT2
mgraph(L,graph="ROC",TC=2,leg=list(pos=c(0.65,0.55),leg=c("NB","DT")),
       baseline=TRUE,Grid=15,main="ROC para o resultado final")
```

As curvas ROC podem ser observadas na Figura 7.

Figura 7 - Curvas ROC para os modelos NB e DT (2ª iteração)



Ambos os modelos têm uma capacidade de previsão bastante superior ao modelo aleatório, representado pela recta diagonal que divide a área do gráfico ao meio.

No caso do modelo DT, a taxa de negativos classificados incorrectamente, ou seja, os insucessos que o modelo previu como sucessos (FPR), constituindo custos desnecessários, aumenta de uma forma mais acentuada a partir de um valor de sensibilidade (TPR) de, aproximadamente, 0,65. Tal significa que, a partir deste valor, o melhoramento na capacidade de detecção de sucessos poderá ser demasiado oneroso no que diz respeito aos insucessos incorrectamente classificados, ou seja, só pode ser obtido aumentando drasticamente a detecção errónea de insucessos, os quais constituem custos operacionais a evitar.

Já para o modelo NB, a curva é mais regular, não havendo propriamente um ponto de mudança acentuada de comportamento. A uma taxa de sensibilidade de 0,80 a taxa de negativos classificados incorrectamente tem o valor aproximado de 0,28, o que significa que se consegue capturar uma grande parte dos sucessos sem um prejuízo elevado ao considerar alguns dos insucessos como sucessos.

Comparando ambos os modelos no mesmo gráfico e, considerando que o que interessa é obter um valor de TPR o mais elevado possível e um valor de FPR o mais baixo possível, ou seja, maximizar a AUC (área debaixo da curva ROC), verifica-se que o DT é ligeiramente melhor do que o NB para uma sensibilidade de até 0,70. A partir deste valor, o modelo NB é substancialmente melhor.

Relativamente à primeira iteração, o modelo NB obtido nesta (iteração) conseguiu um valor de AUC de 0,83, superior, portanto, ao valor de 0,78 obtido anteriormente, o que representa uma melhoria gradual.

4.2.5. Sumário

Nesta iteração, o grande foco esteve na redefinição do objectivo de negócio a atingir: a subscrição ou não do depósito a prazo. Tal conduziu a uma revisão do atributo a prever – ao se eliminarem os registos com resultados inconclusivos, reduziu-se os valores possíveis a dois ("Sucesso" e "Insucesso"), o que beneficia a execução das técnicas para obtenção de modelos de previsão, permitindo um modelo com melhor capacidade de prever o resultado do que aquele obtido aquando da primeira iteração. Além disso já foi possível obter um modelo com árvores de decisão (DT).

Havendo nesta iteração uma redefinição muito importante do objectivo e, consequentemente, do atributo a prever, considerou-se que deveria ser avaliado o impacto na obtenção de modelos decorrente desta grande reformulação. Assim, optou-se por não se efectuar uma análise mais detalhada dos diversos atributos no sentido de compreender a importância que os mesmos teriam na execução das diversas técnicas para a definição de modelos.

Para a próxima iteração, retroceder-se-á à fase de compreensão de dados, havendo um grande enfoque nesta e na fase de preparação de dados. Com 54 atributos para analisar, continua a ser complexo para a ferramenta validar quais os atributos realmente relevantes na previsão do resultado, pelo que há que reduzir este número.

4.3. CRISP-DM – Iteração 3: Utilidade dos dados

4.3.1. Compreensão dos Dados

Nesta iteração, o enfoque serão os dados de *input*, nomeadamente a sua análise e tratamento com vista a suportar a obtenção de modelos de *Data Mining* com melhor capacidade de previsão, permitindo que seja utilizada a técnica de Máquina de Vectores de Suporte (SVM) disponível na biblioteca *rminer*.

Assim, estando bem definido o objectivo a atingir, traduzido no respectivo atributo a prever, esta iteração inicia-se pela fase de compreensão dos dados.

Aquando da iteração anterior, verificou-se que o conjunto de dados era ainda demasiado complexo. Desta forma, é necessário analisar cada atributo e verificar o quão influi na variável a prever, o resultado do último contacto. Para tal, irá ser utilizada a ferramenta *rattle* que permite obter gráficos relacionando o resultado final com o atributo a analisar.

Os gráficos gerados e utilizados na análise subsequente relacionam sempre duas características, um atributo de *input* e o atributo de *output* a prever. O tipo de gráfico escolhido dependeu do tipo de atributos a analisar e do seu leque de valores possíveis. Para tipos nominais binários ou com poucos valores enumerados foi escolhido o gráfico em forma de mosaico, uma vez que permitia ter uma noção visual das proporções de dados (Friendly, 1999).

Já no caso de haver vários valores possíveis para um tipo nominal, optou-se, em vários casos, por um gráfico de barras. No caso de tipos de dados numéricos, o gráfico mais utilizado foi o do diagrama de extremos e quartis (*boxplot*) podendo, em alguns casos, ser utilizado um histograma, conforme disponibilizado pela ferramenta *rattle*.

Uma vez que existem 54 atributos a analisar, nesta secção opta-se por se detalhar apenas a análise de quatro gráficos, um de cada tipo. Em três desses casos verificou-se que os atributos influenciavam nitidamente o resultado, e no outro caso que não existia influência directa.

Um dos atributos a analisar é o indicador da existência do crédito habitação ou seja, se o cliente tem contraído um empréstimo para habitação na instituição bancária em causa. Este atributo é nominal binário e pode assumir os valores: SIM ou NÃO.

Na Figura 8 é possível verificar o gráfico gerado pelo *rattle* para a relação entre possuir crédito habitação e o sucesso numa campanha de depósito a prazo. Assim, sendo o gráfico do tipo mosaico, pode-se verificar que, no conjunto de registos, existem quase tantos clientes sem crédito habitação (44,8%), como clientes com este tipo de empréstimo (55,2%). Proporcionalmente, já é do conhecimento de análises em iterações anteriores que existem bastante mais registos de insucessos do que de sucessos.

No entanto, pelo gráfico, pode-se constatar que, em proporção, o número de sucessos para clientes sem crédito habitação é bastante maior que para clientes com crédito habitação (16,29% face a 7,97%).

Pode-se concluir que a detenção de um empréstimo para habitação influencia o resultado da campanha de subscrição de depósito a prazo. A conclusão de que os clientes sem crédito habitação estarão mais aptos (ou disponíveis) para subscrição de depósitos a prazo pode ter várias interpretações, sendo talvez a mais consensual a que atribui a quem não tem crédito habitação uma maior disponibilidade financeira para investir em produtos de aforro, como é o caso dos depósitos a prazo.

Assim, se um atributo influencia o resultado, isto é, se, pelo facto de o atributo tomar um valor diferente passa a existir uma proporcionalidade diferente no que diz respeito a sucessos e insucessos na campanha, então o mesmo não pode ser de todo excluído e as técnicas de *Data Mining* deverão ser alimentadas com mais este atributo uma vez que o mesmo poderá contribuir para a construção de melhores modelos de previsão.

É possível constatar na Figura 9 o diagrama de extremos e quartis gerado pelo *rattle* para o montante total de subscrições anteriores. O que se verifica, quer para os insucessos, quer para os sucessos, é que existem inúmeros *outliers*, concentrando-se a grande maioria dos casos em torno do valor zero (a mediana, o primeiro e o terceiro quartis são iguais a zero, o que significa que, ou os clientes foram contactados anteriormente e não subscreveram o depósito, ou então nem sequer foram contactados), apesar de a média geral ser de 435,2 euros.

Figura 8 - Gráfico mosaico para a influência do atributo crédito habitação no resultado

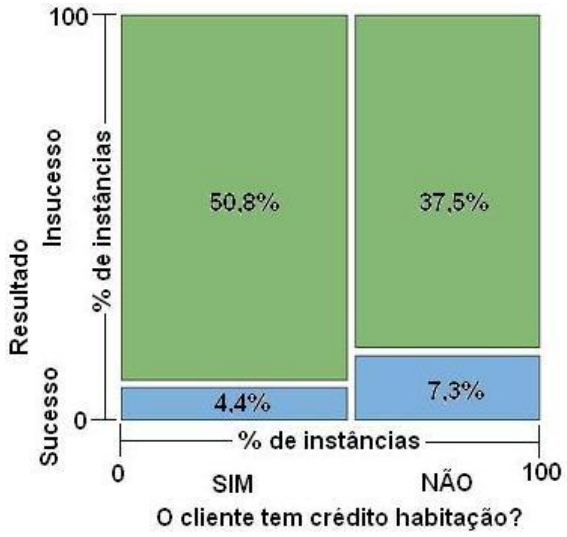
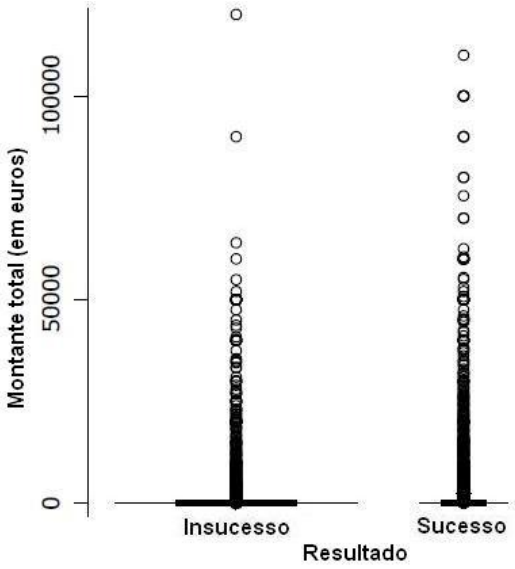


Figura 9 - *Boxplot* para a influência do montante total de subscrições anteriores no resultado



Desta forma, pode-se descartar o montante total de subscrições anteriores do conjunto de atributos de entrada uma vez que, aparentemente, o comportamento do resultado final não é influenciado pelo valor desse atributo.

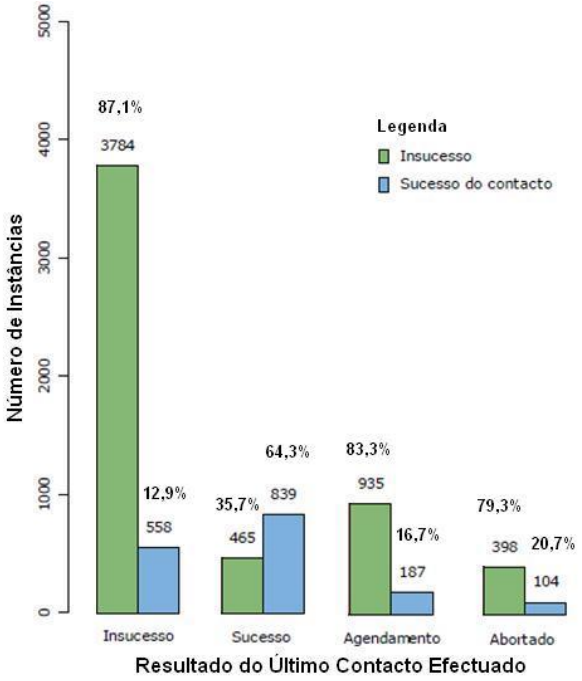
No caso de se ter informação do resultado do último contacto, tal é claramente útil para prever o resultado final. Na Figura 10 é visível que, se o último contacto tiver tido sucesso, é muito mais provável que o próximo contacto também resulte num sucesso. Um resultado anterior de sucesso significa que a probabilidade de o próximo também o ser é de cerca de 0,643, ao passo que um insucesso reduz a probabilidade para 0,129.

Para alguns atributos numéricos optou-se por se gerar um histograma, tal como o da Figura 11 para o caso da idade. Verifica-se que a curva de distribuição de sucessos é inferior à de insucessos entre os 40 e os 60 anos de idade, e que é claramente superior para idades superiores a 60 anos. Tal indicia uma maior probabilidade de sucesso para clientes possivelmente já reformados, provavelmente devido a terem uma maior disponibilidade financeira (em muitos casos, o crédito para habitação é concedido por um prazo limitado pela idade da reforma). Assim, trata-se de um atributo que não pode ser, de todo, descartado.

Com esta análise através de gráficos gerados pelo *rattle*, descartou-se o montante total de subscrições anteriores e consolidou-se a ideia de manter a idade, a existência de crédito para habitação e o resultado do último contacto como atributos relevantes para a construção de modelos.

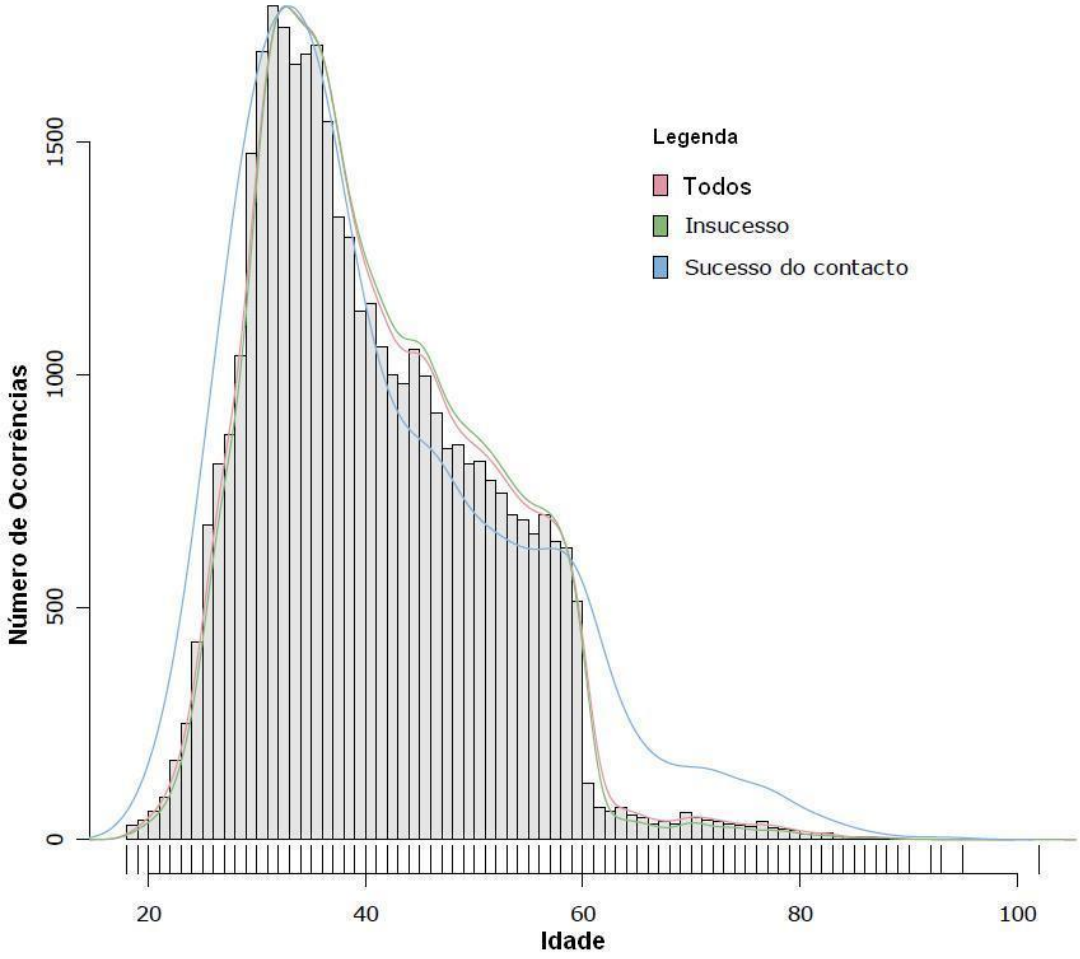
No entanto, existem outros casos em que uma simples análise aos atributos permite concluir que os mesmos não serão úteis para *Data Mining*, dada a pouca variedade de valores (ou categorias, para os tipos nominais).

Figura 10 - Gráfico de barras para a influência do último contacto no resultado



Nota: fez-se uma simplificação para este gráfico, condensando todos os agendamentos num único valor, bem como todos os resultados terminais diferentes de insucesso ou sucesso (contactos abortados)

Figura 11 - Histograma para a influência da idade no resultado



Por exemplo, para o atributo da existência de um indicador de bloqueio geral de cliente (neste caso, o bloqueio é impeditivo de prosseguir com algumas transacções sem que um supervisor da instituição autorize as mesmas), só 0,07% dos registos é que têm o valor S (Sim). Trata-se de um dos casos em que é muito pouco provável que as técnicas de *Data Mining* consigam extrair informação útil referente a uma previsão de um eventual resultado do contacto no âmbito de uma campanha (Sengupta e Sil, 2010). Assim, é um atributo que pode ser excluído dos dados de *input*.

Uma vez que são inúmeros os atributos a analisar bem como os gráficos que suportam essa análise, opta-se por não se apresentar mais gráficos, disponibilizando, no entanto, na Tabela 19, o conjunto de atributos eliminados de cada grupo, bem como o número de atributos antes e após a eliminação. A decisão definitiva de excluir atributos será reforçada na fase de modelação (secção 4.3.3).

Tabela 19 - Atributos excluídos na 3ª iteração do CRISP-DM

Grupo	Antes	Excluídos	Depois
Informação Pessoal de Cliente	11	Profissão Localidade Código postal Código postal local Freguesia Sexo	5
Informação Bancária de Cliente	13	Bloqueios gerais Bloqueios informativos Bloqueio de cheques Inibição de cheques Associado Conta ordenado	7
Informação do Último Contacto	7	Dia da semana de contacto	6
Informação do Primeiro Contacto	8	Dia da semana de contacto	7
Informação de Visualizações	4	Data da última visualização pelo agente Número de visualizações pelo agente Data da última visualização no *homebanking* Número de visualizações no *homebanking*	0
Informação de Histórico	10	N.º de dias desde o primeiro contacto N.º total de sucessos anteriores N.º total de insucessos anteriores Montante subscrito na última campanha Montante total subscrito para todas as campanhas Número total de visualizações pelo agente Número total de visualizações no *homebanking*	3

No caso dos atributos referentes às visualizações, quer para informação de visualizações no âmbito do decurso da campanha, quer para os totais da informação de histórico, verificou-se que o número de registos com estes atributos preenchidos era mínimo: 0,995% para o caso das visualizações por agentes para a banca telefónica e 0,136% para as visualizações no *site* do *homebanking*. Desta forma, estes atributos foram descartados.

Após esta fase, restaram 29 atributos para caracterizar cada contacto e respectiva análise pelas técnicas de *Data Mining*. Na Tabela 26 do anexo C é apresentado o dicionário de dados completo onde são indicados todos os atributos utilizados em cada iteração.

4.3.2. Preparação dos Dados

No que se refere aos dados e, decorrente da análise efectuada na fase anterior com recurso à ferramenta *rattle*, verificou-se que a mesma classificava erradamente alguns atributos como numéricos uma vez que os mesmos eram efectivamente apenas constituídos por códigos numéricos, apesar de serem atributos nominais.

Existia, desta forma, a possibilidade de a própria biblioteca *rminer* considerar esses atributos como numéricos e assim não os tratar adequadamente para a construção dos modelos.

De forma a dissipar qualquer dúvida sobre os tipos destes atributos e a forma como estes se apresentavam perante as técnicas, foi acrescentada uma *string* (cadeia de caracteres alfabéticos) a cada um dos valores possíveis.

Os campos alterados foram os indicados na Tabela 20.

Tabela 20 - Atributos numéricos mal classificados e alteração aos mesmos

Grupo	Atributo	Alteração
Último contacto	Agente	Adicionada a *string* “agente” antes do valor
	Dia	Adicionada a *string* “dia” após o valor
	Mês	Adicionada a *string* “mes” após o valor
	Hora	Adicionada a *string* “hora” após o valor
Primeiro contacto	Agente	Adicionada a *string* “agente” antes do valor
	Dia	Adicionada a *string* “dia” após o valor
	Mês	Adicionada a *string* “mes” após o valor
	Hora	Adicionada a *string* “hora” após o valor

Um outro caso em que os dados estavam incorrectos era para o atributo da duração do primeiro contacto: no caso de não ter havido um contacto anterior ao final (ou seja, a interacção no âmbito da campanha para o cliente teve apenas um contacto), então a duração tinha o valor “-1”. Ora, sendo um campo numérico positivo, ou este valor teria de ser considerado um NA, ou

então, como não houve contacto, a duração teria de ser igual a zero. Optou-se por esta última solução.

Para validar se as alterações teriam impacto na definição dos modelos, foram executadas as técnicas *Naïve Bayes* (NB) e árvores de decisão (DT). Naturalmente que fica implícito um avanço para as fases seguintes de modelação e avaliação, e um retrocesso para esta, o que é perfeitamente válido perante a flexibilidade da metodologia CRISP-DM. Se, para o caso do NB, houve apenas uma ligeira melhoria, considerando que a *Area-Under-the-Curve* (AUC) respeitante à curva ROC passou de 0,83 para 0,87, já no caso da DT, a melhoria foi muito significativa, passando de 0,75 para 0,87. Adicionalmente, a utilização de Máquinas de Vectores de Suporte (SVM) implica que seja necessário que o conjunto de dados não disponha de *missing values*, ao contrário do que sucedia para as técnicas NB e DT.

Analisando o conjunto, verificou-se que existiam 10606 registos que continham um ou mais atributos com *missing values*, face aos dados utilizados na iteração anterior, ou seja, a sua omissão significaria que permaneciam 45211 registos no conjunto de dados para treino e testes. Desta forma, optou-se por esta solução uma vez que se considerou que os dados que restaram eram perfeitamente suficientes para a modelação e validação dos modelos (testes).

Assim, o total de dados a utilizar era constituído por 45211 registos, dos quais 5289 correspondiam a sucessos e 29 atributos explicativos.

4.3.3. Modelação

Ainda no que se refere à eliminação de atributos, é importante referir que a exclusão de atributos somente devido à análise da influência do mesmo sobre o resultado não é trivial, uma vez que um atributo pode por si só não influenciar mas, em conjunto com outros atributos, ter uma influência importante no resultado a prever.

Foi com esta premissa em mente que se fizeram diversas experiências antes de se decidir definitivamente eliminar cada atributo previamente excluído na secção 4.3.1.

Assim, uma validação óbvia efectuada foi simplesmente remover o atributo e obter um modelo para validar se as previsões se mantinham nos mesmos valores percentuais face ao modelo com o atributo. Os resultados práticos de tal processo reforçaram a análise efectuada, permanecendo o conjunto de dados com 29 atributos. Tais resultados não são apresentados por serem demasiado extensos e terem servido apenas para confirmar a análise anterior.

Para a modelação no âmbito desta iteração foram executadas as três técnicas NB, DT e SVM através da função *mining* do *rminer*, com o parâmetro *Runs*=20, ou seja, cada modelação e teste respectivo foi executada 20 vezes para cada uma das técnicas, de forma a efectuar uma validação cruzada com robustez, em termos de confiança dos valores de erro.

Como novidade nesta fase face às iterações anteriores está o facto de se ter utilizado com sucesso a técnica SVM.

4.3.4. Avaliação

O resultado da validação dos três modelos, NB, DT e SVM, é apresentado nas Tabelas Tabela 21, Tabela 22 e Tabela 23, respectivamente.

Tabela 21 - Matriz de confusão e métricas (NB - 3ª iteração) – amostra de teste

Real \ Previsto	Insucesso	Sucesso
Insucesso	235740	30420
Sucesso	10512	24748

AUC = 0,8702377
ACC = 0,8642028
TPR = 0,7018718
FPR = 0,1142922

Tabela 22 - Matriz de confusão e métricas (DT - 3ª iteração) – amostra de teste

Real \ Previsto	Insucesso	Sucesso
Insucesso	256783	9377
Sucesso	19136	16124

AUC = 0,8679249
ACC = 0,9054044
TPR = 0,4572887
FPR = 0,0352307

Tabela 23 - Matriz de confusão e métricas (SVM - 3ª iteração) – amostra de teste

Real \ Previsto	Insucesso	Sucesso
Insucesso	258242	7918
Sucesso	19233	16027

AUC = 0,9379262
ACC = 0,9099230
TPR = 0,4545377
FPR = 0,0297490

Legenda: AUC = área da curva ROC (*Area Under the ROC Curve*), ACC = exactidão (*Accuracy*), TPR = sensibilidade (*True Positive Rate*), FPR = taxa de falsos positivos (*False Positive Rate*)

A AUC sintetiza num único valor o desenho da curva ROC e, quanto maior este, melhor o modelo na sua capacidade de previsão. Assim, tem-se que o NB é ligeiramente melhor do que o DT. No entanto, o modelo que aparenta ter melhores resultados é o SVM: com uma AUC igual a 0,938, trata-se de um modelo de elevada qualidade, explicativo de uma grande percentagem da subscrição de depósitos a prazo no âmbito de campanhas direccionadas, de acordo com os testes efectuados.

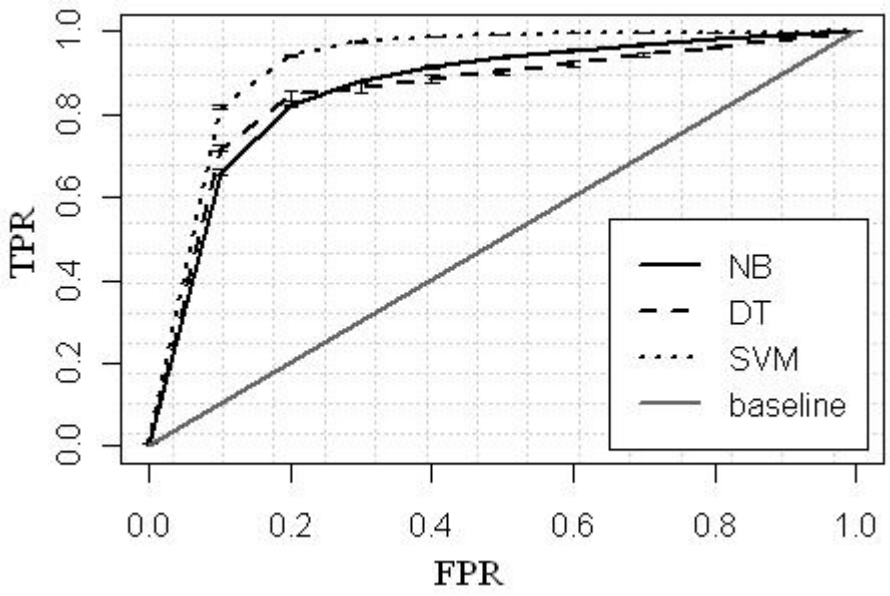
Há, no entanto, uma taxa, a sensibilidade (TPR) que se pretende bastante elevada, uma vez que resume os acertos na previsão de sucessos: caso seja baixa, significa que há vários contactos para os quais o modelo prevê insucesso quando o resultado real seria um sucesso, perdendo assim oportunidades de negócio. Tendo em conta esta taxa, o modelo NB é o que apresenta um valor superior (0,7 contra aproximadamente 0,45 dos outros modelos).

Apesar disso, há que relembrar que as métricas apresentadas para TPR, ACC e FPR, baseadas nos resultados da matriz de confusão, têm em conta que consideram como sendo sucesso sempre que o modelo assim o preveja com uma probabilidade superior a 0,5. Desta forma, se se pretender melhorar a TPR, pode-se considerar uma probabilidade menor para ser assumido um sucesso, fazendo assim deslizar o ponto considerado na curva ROC para a

direita, e aumentando a FPR, que se pretende o mais baixa possível, dado que traduz os insucessos incorrectamente classificados como sucessos pelo modelo.

Posto isto, é importante obter também as curvas ROC, as quais podem ser visualizadas na Figura 12 para os três modelos.

Figura 12 - Curvas ROC para os modelos NB, DT e SVM (3ª iteração)



Tendo em conta a análise ROC, fica claro que o melhor modelo é o SVM, com uma curva superior à dos restantes modelos. É possível conseguir um valor próximo de 1,0 para a TPR com uma taxa de 0,2 para a FPR, optimizando assim as oportunidades de negócio.

Comparando os modelos NB e DT com os obtidos com recurso às mesmas técnicas na 2ª iteração, verifica-se uma melhoria substancial na AUC nos dois casos: o NB passou de 0,83 para 0,87, enquanto que o DT passou de 0,76 para 0,87.

No capítulo seguinte (5 – conclusões) será realizada uma análise mais detalhada relativamente ao impacto que o conhecimento adquirido através do melhor modelo obtido poderá ter em benefício do negócio.

4.3.5. Implementação

A obtenção de um bom modelo explicativo traduz-se em conhecimento que se pretende que venha a beneficiar futuras campanhas. Desta forma, a fase de implementação permitirá concretizar os benefícios em termos de mais-valias para o negócio. No entanto, os limites temporais associados a esta investigação não permitiram avançar para essa fase, ficando a realização da mesma para trabalho futuro, conforme se descreve no Capítulo 5.

4.3.6. Sumário

Nesta 3ª iteração foi dado especial enfoque aos dados e à forma como os mesmos podem ser trabalhados para terem maior utilidade na definição de bons modelos explicativos.

Tal como se previa, o facto de se excluir vários dos atributos permitiu optimizar a execução das técnicas: estas passaram a demorar muito menos tempo a gerar os respectivos modelos. A dúvida, no entanto, seria saber como iria afectar a capacidade de previsão desses modelos. O que se verificou com os dados de teste foi que os modelos obtinham curvas ROC e respectivas AUC com melhor capacidade de previsão.

Para tal, também, terá certamente contribuído o tratamento de dados, em que se forçou a que alguns atributos, que estavam a ser erradamente considerados como numéricos, passassem a ser analisados como nominais pelas técnicas.

Por vezes, uma pequena diferença na forma como os dados são considerados pela técnica pode fazer uma diferença significativa na definição do modelo, como o comprovou a melhoria na AUC detectada para o caso das árvores de decisão. Tal significa que o tratamento de dados é uma fase sobre a qual se deve trabalhar em maior detalhe em iterações subsequentes.

A exclusão de atributos de *input*, que resultou em bons modelos explicativos do sucesso das campanhas, traduz-se nalguns benefícios: por um lado, não é necessário efectuar mais uma recolha de dados adicionais (que, em alguns casos, pode ser complexa ou computacionalmente, ou burocraticamente) e, por outro, as próprias técnicas para gerar os modelos e, em especial, a execução dos mesmos, fica simplificada e mais eficiente.

5. Conclusões

5.1. Síntese e Análise de Resultados Relevantes

Numa época de crise financeira mundial o nosso país, em particular, devido a inúmeras razões económicas (externas e internas), tem dificuldades acrescidas de financiamento. Neste contexto, é fulcral para as instituições bancárias a retenção de capital por meios próprios – no caso vertente, cativando os clientes a aforrar dinheiro na instituição – de forma a poderem continuar a dispor de capital para investir e financiar a própria economia nacional.

A prova disto é o facto de todas as instituições bancárias estarem a remunerar os seus depósitos a taxas cada vez mais elevadas, algo que preocupa o Banco de Portugal que, em Junho de 2011 anunciou uma iniciativa para obrigar os bancos a comunicar taxas que ultrapassem em três pontos percentuais o valor da taxa Euribor aplicável na operação respectiva[24].

Tendo esse cenário presente, o qual pressiona também a concorrência entre instituições, não basta esperar apenas que os clientes surjam no balcão de atendimento para subscrever produtos de aforro; tornam-se, por isso, cada vez mais importantes todas as acções proactivas de contactar os clientes e ir de encontro às suas expectativas e necessidades, cativando assim a subscrição dos produtos de poupança.

No entanto, é precisamente o contexto económico que pressiona a uma redução global de custos e que, talvez até contraditoriamente, acaba por afectar também o orçamento disponível para efectuar os contactos com os clientes, levando a reduções quer no custo de comunicações, quer no custo de contratação e formação dos operadores humanos (sendo operações algo intrusivas, de venda directa, descarta-se a possibilidade de utilização de chamadas automáticas, tratadas por um sistema de IVR).

Assim a redução de custos e consequente aumento da eficiência dos contactos pode ser conseguida através do conhecimento adquirido sobre o funcionamento das campanhas, o qual pode ser usado para a optimização das mesmas. Tal implica que a selecção dos contactos tenha por base um modelo com uma excelente capacidade de previsão, para que a sua análise no que diz respeito à explicação do sucesso do contacto possa traduzir-se em conhecimento útil, a ser utilizado em campanhas futuras.

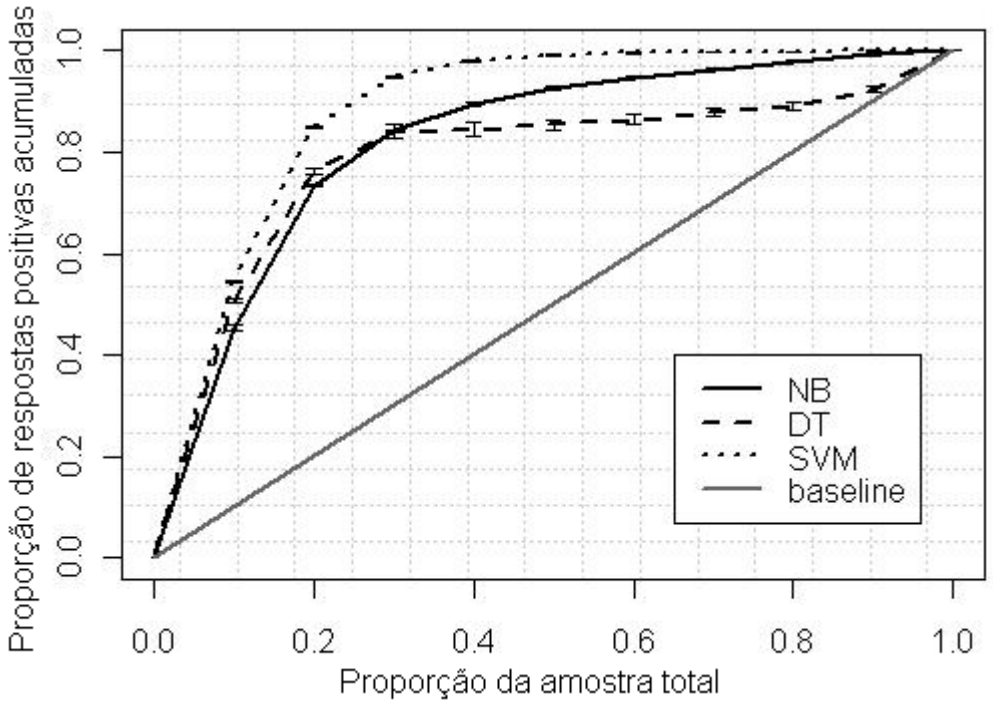
O cálculo do benefício resultante de uma subscrição de um depósito a prazo é demasiado complexo dada a entropia[25] que lhe está associada. São inúmeras as variáveis que teriam de ser consideradas, por exemplo, se o cliente que subscreveria o produto não fosse contactado e se se dirigisse ao balcão para o subscrever por sua própria iniciativa, então o benefício seria nulo, apesar de o contacto, a ser feito, ter resultado num sucesso.

[24] Retirado do *Semanário Sol*, "Taxa de juro nos depósitos a prazo mais do que duplica num ano", edição *online* de 11 de Julho de 2011 – http://sol.sapo.pt/inicio/Economia/Interior.aspx?content_id=23861.
[25] Medida de desordem ou imprevisibilidade – retirado de *Dicionário Priberam da Língua Portuguesa, Edição Online,* http://www.priberam.pt/DLPO/, Agosto de 2011.

Adicionalmente, definiu-se como objectivo principal deste estudo a definição de um modelo explicativo da subscrição de um depósito a prazo no âmbito de uma campanha, não se equacionando o valor investido por cada subscrição. Ou seja, o modelo a obter com esta investigação torna-se extremamente útil, no sentido em que permite explicar o sucesso dos contactos com base em informação distinta, possibilitando assim uma reconfiguração de campanhas futuras tendo em conta o conhecimento adquirido.

Com todas as considerações já expostas relativamente à quantificação de benefícios e, tendo obtido na 3ª iteração do CRISP-DM três modelos com boa capacidade de explicar o sucesso (ainda que um deles sobressaísse), é relevante justificar perante a equipa de gestão da área de negócio (no caso, a Direcção de *Marketing*) a escolha e interpretação de um modelo. Assim, poderá ser analisada na Figura 13 a curva *Lift* para os três modelos: *Naïve Bayes* (NB), Àrvores de Decisão (DT) e Máquinas de Vectores de Suporte (SVM).

Figura 13 - *Lift* acumulativo dos modelos obtidos na 3ª iteração do CRISP-DM



A figura permite, de uma forma simples, visual, ter a percepção de como seria afectada uma campanha no que diz respeito ao número de sucessos em função do conjunto de contactos seleccionados por cada modelo.

Quanto maior o valor de respostas positivas acumuladas, mais contactos referentes a sucessos são abrangidos no lote seleccionado da amostra total. A *baseline* pode ser interpretada da seguinte forma: a utilização de N% dos contactos resultaria em N% do total de sucessos. Logo, quanto maior a diferença entre a *baseline* e a curva *Lift* para o modelo e,

consequentemente, a área debaixo da curva *Lift* (ALIFT), melhor o modelo na capacidade de concentrar os sucessos nos primeiros decis.

Por exemplo, pode-se constatar para os modelos NB e DT que, com uma selecção de apenas 20% dos contactos (os primeiros dois decis, que contêm os clientes com maior probabilidade de sucesso), se irão obter mais de 75% do total de sucessos.

Desta forma, é possível comparar os modelos e considerar que o SVM é substancialmente melhor do que os restantes: uma selecção de 30% do universo de contactos disponíveis para serem efectuados selecciona a quase totalidade dos contactos que resultariam em sucesso. Confirma-se assim os resultados da análise efectuada com a curva ROC (Figura 12), ou seja, o modelo SVM é de elevada qualidade.

Definido o melhor modelo (SVM), resta então analisar o mesmo no que diz respeito à importância dos atributos (Figura 14) para a sua definição e assim poder extrair o conhecimento sobre o funcionamento dos contactos na subscrição de depósitos a prazo.

Claramente sobressai na Figura 14 a importância da duração da chamada no sucesso do contacto. A duração por si só tem uma influência superior a 20% na definição do modelo obtido. No entanto, dois outros atributos surgem também com alguma importância no sucesso. Um deles é o resultado obtido no âmbito da última campanha de depósitos a prazo de que o cliente foi alvo. O outro é o mês em que o contacto foi efectuado, significando que alguns meses são melhores para este tipo de campanhas.

Através de uma curva VEC (*Variable Effect Characteristic*), é possível verificar como é que os atributos mais importantes na definição do modelo SVM influenciaram o resultado. Assim, na Figura 15 pode-se verificar que, a uma duração de chamada entre 15 e 50 minutos (entre 1000 e 3000 segundos), aproximadamente, corresponde uma probabilidade de quase 1,0 de o contacto resultar num sucesso. A tal não será certamente alheio o facto de haver uma necessidade de descrever detalhadamente as características do depósito aos clientes que se mostrem interessados.

No entanto, eventualmente, um cliente com mais disponibilidade para ouvir o agente poderá ficar mais receptivo ao produto, devendo o agente investir no prolongamento do diálogo. O conhecimento deste facto poderá conduzir a um investimento na formação dos agentes para treinar a sua capacidade de cativar através do diálogo. Apesar de tudo, a partir de 50 minutos a probabilidade de sucesso diminui, o que poderá ser devido a uma saturação do cliente face a um diálogo demasiado prolongado.

Figura 14 - Importância dos atributos para a explicação do resultado (modelo SVM)

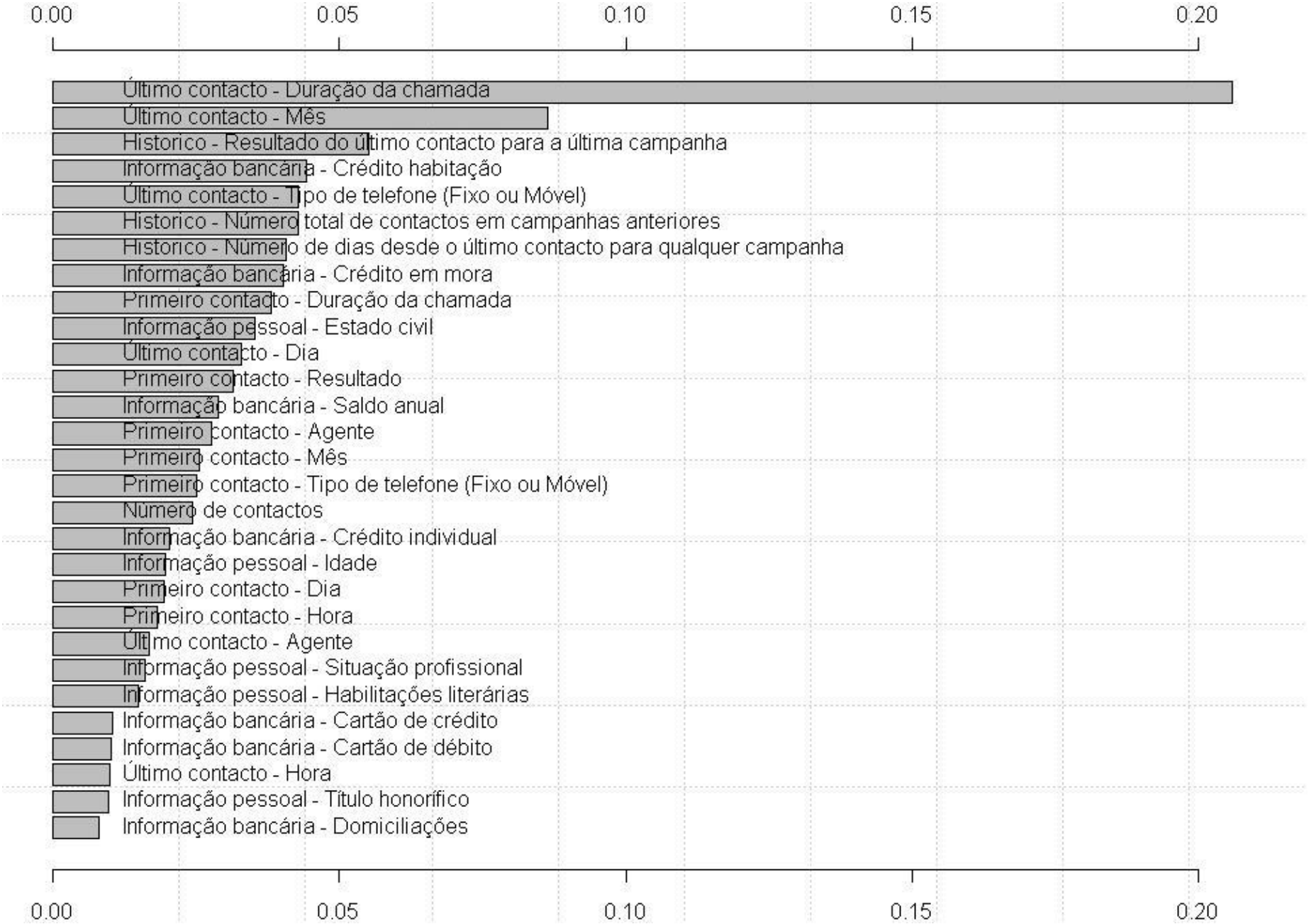
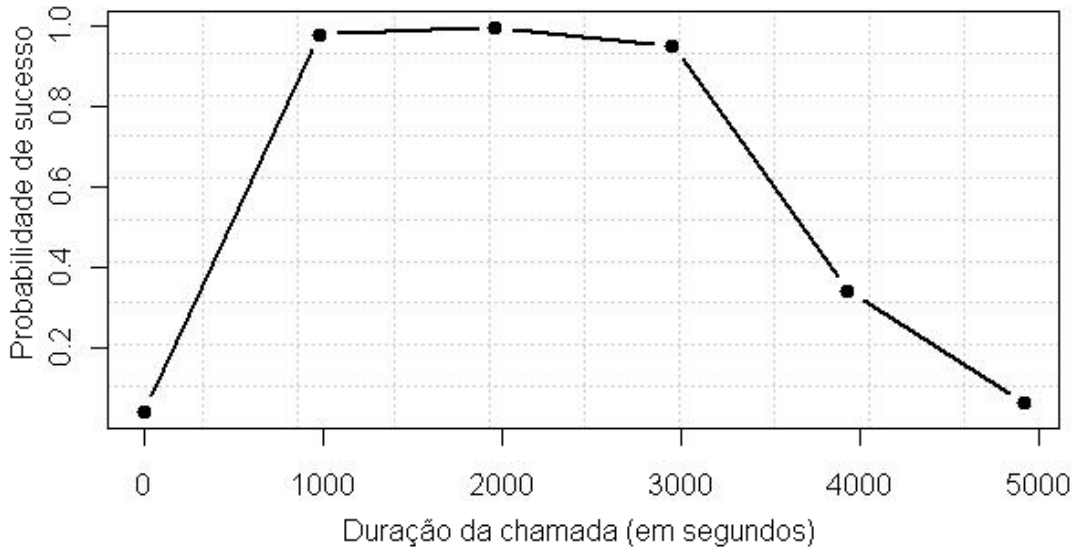


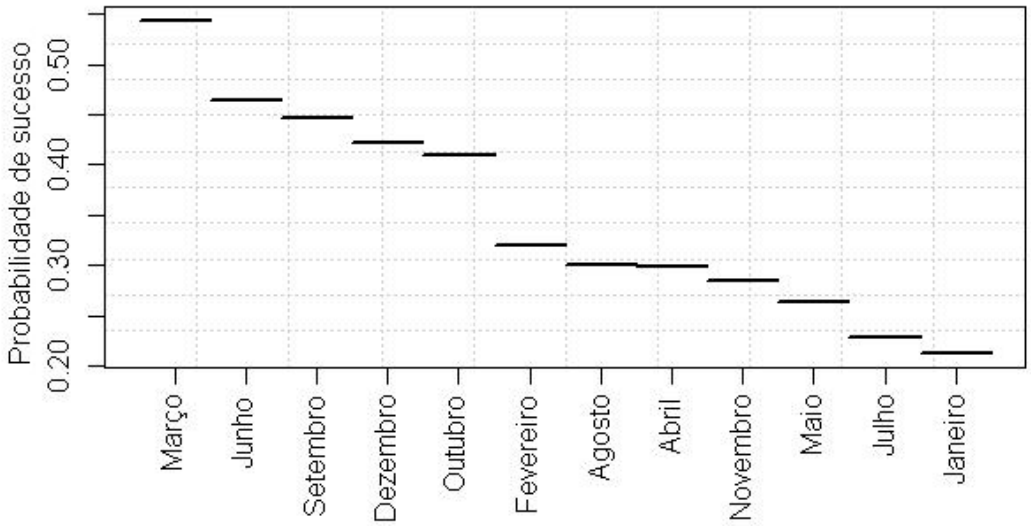
Figura 15 - Influência da duração da chamada no resultado



Relativamente ao mês de contacto, pode-se observar na Figura 16 que existe uma maior probabilidade de sucesso se o contacto for efectuado no último mês de cada trimestre (Março, Junho, Setembro e Dezembro). Para estes meses e ainda para Outubro, a probabilidade de sucesso traduzida no modelo corresponde a mais de 0,4.

No caso do resultado do último contacto para a última campanha, a influência corresponde ao que já havia sido identificado aquando da análise da Figura 10: a ocorrência de um sucesso anterior aumenta fortemente as probabilidades de um sucesso no futuro.

Figura 16 - Influência do mês de contacto no resultado



Desta forma e, através da análise da curva VEC, é possível perceber como é que cada atributo utilizado no modelo influencia o sucesso na subscrição de depósitos a prazo.

5.2. Discussão

A redução da eficácia das acções de *marketing* direccionado, proactivo, e intrusivo para as pessoas que são alvo das respectivas campanhas é uma realidade cada vez mais presente – existe uma saturação generalizada da parte dos consumidores a qual tem vindo inclusive a ser suportada por legislação apropriada para defender o público em geral.

Quando o contexto económico-social se degrada abruptamente (como é o caso da actualidade, onde uma crise financeira à escala mundial tem levado a uma reformulação dos paradigmas da gestão empresarial e governamental), maior é a ineficácia das acções publicitárias em geral. Trata-se, portanto, de um novo factor de pressão e que leva as organizações a apostarem numa melhoria da eficiência interna dos seus processos relacionados com *Marketing*. Quando estes estão dependentes de grandes volumes de dados, sendo os mesmos dinâmicos, a aposta tem vindo a ser no sentido de melhorar a utilidade da informação da organização.

No caso das acções de *Marketing* direccionado e, uma vez que é previsível que o seu retorno (ROI) seja cada vez menor (em função do contexto já descrito), é importante aumentar a sua eficiência, optimizando os recursos disponíveis no sentido de tirar maior proveito de cada contacto.

É nestas situações que a aposta em soluções de *Data Mining* se tem verificado como adequada para a melhoria da eficiência, nomeadamente, através da definição e implementação de modelos explicativos que, com o recurso aos dados já existentes na organização, permitam dar resposta à necessidade de informação actualizada para suporte a decisões de negócio. No caso do *Marketing* direccionado, tal traduz-se em várias acções de melhoria dos recursos existentes, por exemplo, seleccionando melhor os contactos a efectuar, realizando acções de formação para educar os agentes a tirarem proveito do conhecimento adquirido ou executando as campanhas em alturas em que seja expectável um maior ROI.

Para a investigação que culminou no trabalho exposto neste documento, a base partiu de um caso a estudar, referente a contactos efectuados no âmbito de campanhas para subscrição de depósitos a prazo. O objectivo principal era a definição de um modelo que tivesse uma capacidade de previsão suficientemente boa no que diz respeito aos contactos de forma a que pudesse ser extraído conhecimento, utilizando-o como ferramenta de suporte à definição das campanhas.

Tendo em conta as pressões externas para a redução de custos, a opção lógica seria a utilização de ferramentas *freeware* (idealmente *open-source*, para que pudessem ser alteradas se necessário) de forma a construir o modelo adequado.

No final deste trabalho, pode-se concluir que as ferramentas utilizadas, nomeadamente a biblioteca *rminer* suportada sobre a plataforma R (ambas soluções *open-source*), são suficientemente robustas e maduras para serem usadas num ambiente real. Têm a vantagem de disponibilizar o código fonte, para alterações que os utilizadores considerem adequadas, e

de uma boa comunidade de apoio e interessada no seu desenvolvimento. Adicionalmente, em época de contenção de custos, o facto de serem gratuitas assume maior relevância, dispensando o licenciamento dispendioso das soluções mais usuais disponibilizadas pelas grandes corporações.

Algumas das desvantagens verificadas prendem-se com o volume dos dados de *input* que alimentam as técnicas, o qual é algo limitado (foi necessário haver uma redução substancial da quantidade de dados e melhoria da sua qualidade para se obter o modelo mais complexo de Máquinas de Vectores de Suporte), e ainda com o suporte técnico fornecido, que é praticamente inexistente em Portugal, para além da comunidade científica.

Tendo sido utilizada a metodologia CRISP-DM, importa avaliar os seus resultados. Assim, na Tabela 24 pode-se verificar como evoluíram os modelos gerados com o decorrer das iterações.

Tabela 24 - Evolução dos modelos

Iteração no CRISP-DM	**1ª**	**2ª**		**3ª**		
Instâncias × Atributos (Categorias Possíveis)	79354×59 (12)	55817 × 53 (2)		45211 × 29 (2)		
Técnica	NB	NB	DT	**NB**	**DT**	**SVM**
Número de Execuções (*Runs*)	1	20	20	20	20	20
AUC	0.776	0.823	0.764	**0.870**	**0.868**	**0.938**
ALIFT	0.687	0.790	0.591	**0.827**	**0.790**	**0.887**

Legenda: AUC – *Area Under the ROC Curve*, ALIFT – *Area Under the LIFT Curve*, *Runs* – Número de execuções na função *mining* do *rminer*

Constata-se que a sua aplicação resultou numa redução gradual do conjunto de dados a utilizar e, inversamente, num aumento gradual da qualidade dos modelos obtidos no que diz respeito à sua capacidade de previsão. Os valores para a análise através da curva *Lift* (ALIFT) acompanham os respectivos para a curva ROC (AUC): embora com interpretações distintas, em ambos os casos quanto maior o valor, melhor.

A utilização da metodologia CRISP-DM, flexível, dinâmica, incremental, e iterativa trabalhando sucessivamente sobre os resultados da iteração anterior, revelou-se adequada aos objectivos propostos. Tal permitiu trabalhar objectivos, dados que serviram de suporte, e efectuar diversos ensaios que possibilitaram avançar a cada iteração para uma refinação de todo o projecto de *Data Mining*, com objectivos cada vez mais concretos e dados cada vez mais adequados.

A evolução nos modelos, suportada por um trabalho exaustivo de tratamento de dados, permite ainda concluir que as ferramentas actuais de *Data Mining* por si só não podem ainda ter todo o ónus na definição de modelos: é fulcral haver previamente um bom e exaustivo trabalho na definição de objectivos e, em especial, na selecção criteriosa e tratamento dos

atributos de dados que, no fundo, irão caracterizar o registo perante a técnica de *Data Mining*. Só desta forma é possível obter um modelo de previsão adequado a um objectivo complexo.

Esta necessidade identificada de um trabalho nitidamente humano a nível de decisão de qual a informação adequada à técnica tem vários paralelismos noutros problemas computacionais, um dos quais é a utilização de técnicas de inteligência artificial para automatização na resolução de problemas: quando são utilizadas soluções generalizadas, com pouca ou nenhuma informação específica sobre o problema – soluções tipicamente independentes do domínio do problema – a performance obtida é muito inferior àquela obtida por soluções alimentadas por informação específica do domínio (Negnevitsky, 2005). Nestes casos, o facto de se transmitir conhecimento humano sobre o problema aos algoritmos de inteligência artificial constitui um guia para a procura da solução.

O melhor modelo foi obtido com recurso a Máquinas de Vectores de Suporte (SVM), confirmando assim o bom desempenho desta técnica mais recente face às restantes, mais antigas, *Naïve Bayes* (NB) e Árvores de Decisão (DT). Este resultado é equivalente ao publicado por Hearst *et al*. (1998), em que foram comparadas estas três técnicas, com uma clara vantagem para a SVM. Através deste modelo, foi possível extraír conhecimento na forma da definição da influência que alguns atributos têm para o próprio modelo. Três dos cinco atributos mais relevantes estão relacionados com o próprio decorrer dos contactos. O mais importante é a duração da chamada: o conhecimento adquirido permite condicionar o diálogo dos agentes no sentido de prolongar a chamada até uma duração que permita persuadir o cliente a adquirir o produto.

Esta investigação resultou num modelo explicativo através do qual se adquiriu conhecimento sobre o funcionamento das campanhas de subscrição de depósitos a prazo. Tal deverá ser transmitido à área de negócio gestora do *contact-center* de forma a que seja utilizado para a optimização da execução dos contactos.

Por contraponto, refere-se o trabalho de Javaheri (2008), em que foi definido um modelo preditivo para *Marketing* direccionado aplicado a um caso de estudo de um banco no Irão, com 30000 clientes e 85 atributos, todos com excepção da idade referentes a informação bancária conhecida antes do arranque das acções de *Marketing*. Também foi utilizado o ambiente R e a técnica SVM para a definição do modelo. De acordo com a análise *Lift* efectuada pelo autor, 20% do total de clientes consolidariam 60% das respostas positivas. Por comparação, no caso do modelo obtido na presente investigação, 20% dos contactos resultariam em cerca de 85% dos sucessos. Não sendo os estudos directamente comparáveis, os resultados de ambos indiciam o quão importantes são os atributos referentes à execução dos contactos, confirmando os resultados já expostos na Figura 14.

A extensão do trabalho efectuado aos restantes tipos de campanhas obrigaria a um projecto semelhante para cada tipo, dada a disparidade de objectivos a que cada tipo de

campanha se propõe. No entanto, as conclusões apresentadas permitem encarar com optimismo essa extensão.

Sendo este um estudo que compara diferentes técnicas de *Data Mining* numa única organização, mas com base numa amostra de grande dimensão (mais de 70 mil casos) é importante partilhar os resultados obtidos, quer com a comunidade científica, quer com os profissionais do *Marketing*. Desta forma, submeteu-se um resumo deste trabalho a uma das principais conferências, a nível mundial, sobre Modelação, a ESM – *European Simulation and Modelling Conference* -, 2011, com *proceedings* indexados no *ISI web of knowledge*, o qual foi aceite para apresentação e publicação (Moro *et al.*, 2011).

5.3. Limitações e Trabalho Futuro

No trabalho exposto foi possível percepcionar como é que determinados atributos influenciaram o resultado das campanhas de subscrição de depósitos a prazo e qual a sua importância na definição de um modelo explicativo. Para confirmar os resultados, seria interessante compará-los com os obtidos através de técnicas clássicas de estatística, como por exemplo, a regressão logística. Por um lado, poder-se-ia averiguar se as descobertas em ambos os casos eram similares e, por outro, permitiria verificar até que ponto as mais recentes técnicas de *Data Mining* estão em condições de substituir uma abordagem clássica.

A utilidade de um modelo explicativo vai no sentido de fornecer conhecimento às equipas de gestão do negócio. No entanto, a criação de um processo automático de utilização desse conhecimento pressuporia a implementação de um modelo preditivo. Tal permitiria uma selecção automática dos contactos mais adequados a uma campanha com base nas características de cada cliente, definidas em função dos valores dos seus atributos.

O facto de se ter usado informação sobre a execução de contacto, obtida somente após o início da campanha, relevou-se extremamente importante na definição do modelo: dos cinco atributos mais importantes para o modelo, três são referentes a dados de contacto, sendo dois deles os mais importantes. Esta revelação permite pressupor que seria difícil obter um bom modelo preditivo somente com os dados utilizados, removendo a informação de contacto.

Apesar de tudo, no estudo de Javaheri (2008) ficou patente que, apenas com atributos de informação bancária, era possível obter um modelo preditivo com uma capacidade de previsão bastante aceitável. Nesse trabalho foram usados 84 atributos relacionados todos com informação bancária, muitos deles quantitativos. Por comparação, no caso da presente investigação, só foram usados 13 atributos, e todos nominais binários (com excepção do saldo bancário). Assim, deveria ser obtida informação bancária mais detalhada para auxiliar na definição eventual de um modelo preditivo.

Adicionalmente, alguma da informação pessoal do cliente acabou por não ser utilizada. Por exemplo, o local de residência: como existiam demasiadas freguesias e códigos postais,

não foram utilizados para a definição dos modelos. No entanto, pode-se extrapolar o concelho e distrito de residência tendo em conta o código postal, com recurso a uma base de dados pública disponibilizada pelos CTT[26]. Tais dados poderiam ser úteis na previsão de sucessos, uma vez que é um facto que há distritos onde o rendimento *per capita* é superior a outros[27].

Existe ainda informação que poderia ser relevante para um modelo preditivo: em qualquer produto que se pretenda vender, existem duas características chave – o preço e a qualidade. No caso de depósitos a prazo, a qualidade não é relevante, uma vez que as instituições possuem, em geral, uma imagem sólida e suportada por fundos de garantia do Banco de Portugal. Assim, o preço pode ser um factor decisivo e, no caso dos depósitos, é consubstanciado pela taxa de juro oferecida. Desta forma, o conjunto de dados poderia ser enriquecido pela adição de, pelo menos, mais três atributos: a taxa oferecida no depósito, a taxa média dos depósitos a prazo de todos os bancos, e uma taxa de referência para a generalidade das operações bancárias (eventualmente, ou a taxa Euribor, ou a taxa de referência do Banco Central Europeu - BCE).

[26] CTT (Correios Telégrafos e Telefones) – Empresa de Correios de Portugal.

[27] Retirado do *Semanário Expresso*, "Mapa dos salários em Portugal", edição *online* de 31 de Março de 2011 – http://aeiou.expresso.pt/mapa-dos-salarios-em-portugal=f574045.

Referências Bibliográficas

Ahn, H., Kim, K., and Han, I. (2006). Hybrid genetic algorithms and case-based reasoning systems for customer classification, *Expert Systems with Applications*, 23(3), 127–144.

Anderson, K. and Kerr, C. (2001). *Customer Relationship Management – 1st edition*. McGraw--Hill, USA.

Apte, C. and Weiss, S. (1997). Data mining with decision trees and decision rules, *Future Generation Computer Systems*, 13(2-3), 197–210.

Baesens, B., Viaene, S., Poel, D., Vanthienen, J., and Dedene, G. (2002). Bayesian neural network learning for repeat purchase modelling in direct marketing, *European Journal of Operational Research*, 138(1), 191–211.

Bose, I. and Mahapatra, R. (2001). Business data mining – a machine learning perspective, *Information & Management*, Elsevier, 39(3), 211–225.

Brause, R., Langsdorf, T. and Hepp, M. (1999). Neural Data Mining for Credit Card Fraud Detection, *Proceedings 11th IEEE International Conference on Tools with Artificial Intelligence*, pp. 103–106, IEEE.

Brown, M. and Kros J. (2003). Data mining and the impact of missing data, *Industrial Management & Data Systems*, 103(8), 611–621.

Buckinx, W., Moons, E., Poel, D. and Wets, G. (2004). Customer-adapted coupon targeting using feature selection, *Expert Systems with Applications*, 26(4), 509–518.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. and Wirth, R. (2000). *CRISP-DM 1.0 - Step-by-step data mining guide*, CRISP-DM Consortium. URL: http://www.crisp-dm.org, acedido em Outubro de 2010.

Chiu, C. (2002). A case-based customer classification approach for direct marketing, *Expert Systems with Applications*, 22(2), 163–168.

Cohen, M. (2004). Exploiting response models: optimizing cross-sell and up-sell opportunities in banking, *Journal Information Systems - Knowledge discovery and data mining* (KDD 2002), Elsevier, 29(4), 327-341.

Coppock, D. (2002). Why Lift? – Data Modeling and Mining, *Information Management Online*, June 21, 2002. URL: http://www.information-management.com/news/5329-1.html, acedido em Outubro de 2010.

Cortes, C. and Vapnik, V. (1995). Support Vector Networks, *Machine Learning*, 20(3), 273–297.

Cortez, P. (2010). Data Mining with Neural Networks and Support Vector Machines using the R/rminer Tool. In P. Perner (Ed.), Advances in Data Mining - Applications and Theoretical Aspects, *Proceedings of the 10th Industrial Conference on Data Mining*, LNAI 6171, pp. 572–583, Berlin, Germany, July, 2010. Springer.

Cortez, P. and Embrechts, M. (2011). Opening Black Box Data Mining Models Using Sensitivity Analysis, *Proceedings of IEEE Symposium on Computational Intelligence and Data Mining (Paris, France)*, pp. 341–348.

Cortez, P. and Silva, A. (2008), Using Data Mining to Predict Secondary School Student Performance, *Proceedings of the 5th Future Business Technology Conference*, pp. 5–12.

Danaher, P. and Rossiter, J. (2006). *A Comparison of the Effectiveness of Marketing Communication Channels: Perspectives from Both Receivers and Senders*, In Occasional Paper, University of Auckland, New Zealand. URL: http://www.mailmarketing.com.au/files/AuspostDanahersFullReport_1_.pdf, acedido em Março de 2011 .

Evenson, A., Harker, P. and Frances F. (1999). Effective Call Center Management: Evidence from Financial Services, *The Working Paper Series, The Wharton School, University of Pennsylvania*, Philadelphia, USA.

Fawcett, T. (2005). An introduction to ROC analysis, Pattern Recognition Letters, Elsevier, 27(8), 861–874.

Friendly, M. (1999). Extending Mosaic Displays: Marginal, Partial, and Conditional Views of Categorical Data, *Journal of Computational and Graphical Statistics*, 8, 373–395.

Han, J. and Kamber, M. (2006). Data Mining – Concepts and Techniques, 2nd edition, Elsevier, USA.

Hand, D., Mannila, H., and Smyth, P. (2001). *Principles of Data Mining*, MIT Press, Cambridge, MA.

He, Z., Xu, X., Huang, J. and Deng, S. (2004). Mining class outliers: Concepts, algorithms and applications in CRM, *Expert Systems with Applications*, 27(4), 681–697.

Hearst, M., Dumais, S., Osman, E., Platt, J., and Scholkopf, B. (1998). Support Vector Machines, *IEEE Intelligent Systems*, 13(4), 18–28.

Hu, X. (2005). A data mining approach for retailing bank customer attrition analysis, *Applied Intelligence*, 22(1), 47–60.

Javaheri, H. (2008). *Response modeling in direct marketing: a data mining based approach for target selection*, Master Thesis, Tarbiat Modares University, Iran and Luleå University of Technology, Sweden.

Keen, J. and Digrius, B. (2003). *Making Technology Investments Profitable: ROI Roadmap to Better Business Cases*, John Wiley & Sons, USA.

Kim, Y. e Street, W. (2004). An Intelligent System for Customer Targeting: A Data Mining Approach, *Decision Support Systems*, Elsevier, 37(2), 215–228.

Kohavi, R. and Provost, F. (1998). Glossary of Terms, *Machine Learning*, 30(2–3), 271–274.

Koole. G. and Mandelbaum, A. (2002). Queueing Models of Call Centers: An Introduction, *Annals of Operations Research*, 113, 41–59.

Kotler, P. (2002). *Marketing Management, Millenium Edition – Customer Edition for University of Phoenix*, Pearson Custom Publishing, USA.

Kuonen, D. (2004). *Data Mining and Statistics: What is the Connection?*, In TDAN.com, URL: http://www.tdan.com/view-articles/5226/, acedido em Janeiro de 2011.

Leite, F., Diniz, E. and Jayo, M. (2009). Utilização de Business Intelligence para gestão operacional de agências bancárias: um estudo de caso, *Revista Eletrônica de Sistemas de Informação*, 2(8). URL: http://revistas.facecla.com.br/index.php/reinfo/article/view/576/446, acedido em Janeiro de 2011.

Li, W., Wu, X., Sun, Y. and Zhang, Q. (2010). Credit Card Customer Segmentation and Target Marketing Based on Data Mining, *Proceedings of International Conference on Computational Intelligence and Security*, pp. 73-76, Lecture Notes in Computer Science, Springer.

Ling, C. X. and Li, C. (1998). Data Mining for Direct Marketing: Problems and Solutions, *Proceedings of the 4th International Conference on Knowlege Discovery and Data Mining (KDD'98)*, pp. 73–79, Association for the Advancement of Artificial Intelligence, AAAI Press.

Mingoti, S. and Lima, J. (2005). Comparing SOM neural network with Fuzzy c-means, K-means and traditional hierarchical clustering algorithms, *European Journal of Operational Research*, Elsevier, 174(3), 1742-1759.

Moro, S., Laureano, R. and Cortez, P. (2011). Using Data Mining for Bank Direct Marketing: an Application of the CRISP-DM Methodology, *Proceedings of European Simulation and Modelling Conference* (ESM 2011), October 24–26, Guimarães, Portugal (aceite para publicação e apresentação).

Muenchen, R. (2010). The Popularity of Data Analysis Software. URL: http://r4stats.com/popularity, acedido em Fevereiro de 2011.

Negnevitsky, M. (2005). *Artificial Intelligence: A Guide to Intelligent Systems, 2nd edition*, Pearson, USA.

Ngai, E., Xiu, L. and Chau, D. (2009). Application of data mining techniques in customer relationship management: A literature review and classification, *Expert Systems with Applications*, Elsevier, 36(2), 2592–2602,.

Ou, C., Liu, C., Huang, J. and Zhong, N. (2003). On Data Mining for Direct Marketing, *Proceedings of the 9th international conference on Rough sets, fuzzy sets, data mining, and granular computing (RSFDGrC 2003), 2639*, pp. 491–498.

Owen, R. and Humphrey, P. (2009). The Structure of Online Marketing Communication Channels, *Journal of Management and Marketing Research*, Academic and Business Research Institute, 2, 54–62.

Page, C. and Luding, Y. (2003). Bank manager's direct marketing dilemmas – customer's attitudes and purchase intention, *International Journal of Bank Marketing*, Emerald Insight, 21(3), 147–163.

Paprzycki, M., Abraham, A., Guo, R. and Mukkamala, S. (2004). Data Mining Approach for Analyzing Call Center Performance, *Proceedings of the 17th international conference on Innovations in applied artificial intelligence* (IEA/AIE'2004), Springer, pp. 1092–1101.

Prinzie, A., and Poel, D. (2005). Constrained optimization of data-mining problems to improve model performance: A direct-marketing application, *Expert Systems with Applications*, 29(3), 630–640.

Sengupta, N. and Sil, J. (2010), Dimension Reduction Using Rough Set Theory For Intrusion Detection System, Proceedings of the 4th National Conference on Computing For Nation Development, pp. 410–415.

Silva, F., Cortez, P. and Cadavez, V. (2010), Using Multiple Regression, Neural Networks and Support Vector Machines to Predict Lamb Carcasses Composition, *Proceedings of the 6th International Conference on Simulation and Modelling in the Food and Bio-Industry*, pp. 41–45.

Taghva, M., Bamakan, S. and Toufani, S. (2011). A data mining method for service marketing: A case study of banking industry, *Management Science Letters*, Growing Science, 1, 253–262.

Tapp, A. (2008). Introducing direct marketing. In *Principles of direct and database marketing – a digital orientation, Fourth Edition*, Prentice Hall, pp. 1–52.

Turban, E., Sharda, R. and Delen, D. (2010). *Decision Support and Business Intelligence Systems – 9th edition*, Prentice Hall Press, USA.

Whittaker, S. (2003). Theories and Methods in Mediated Communication. In *The Handbook of Discourse Processes*, Lawrence Erlbaum Associates, pp. 243–286.

Williams, G. (2009). Rattle: a data mining GUI for R, *The R Journal*, 1(2), 45–55.

Williams, G. (2011). *Data Mining with Rattle and R: The Art of Excavating Data for Knowledge Discovery*, Springer, USA.

Witten, I. and Frank, E. (2005). *Data Mining – Pratical Machine Learning Tools and Techniques, 2nd edition*, Elsevier, USA.

Zhang, H. (2004). The Optimality of Naïve Bayes, *Proceedings of the 17th FLAIRS conference*, AAAI Press.

Anexos

A. Campanhas em Estudo

A Tabela 25 contém a listagem de todas as campanhas que seriam alvo da investigação, constituindo, como tal, o caso em estudo. A cada linha corresponde uma campanha comercializada pela área de *Marketing*, com um objectivo de negócio bem definido, uma selecção de clientes própria, e um intervalo temporal em que decorre.

Tabela 25 - Lista inicial de campanhas a analisar

Produto ou Serviço		Objectivo	Relatório		Campo	Banca Telefónica		*Home*
Grupo	Nome		Visualizações	Contactos	Montante	*Inbound*	*Outbound*	*banking*
Crédito Individual Pré Aprovado		Informar	X					X
Crédito Individual Pré Aprovado		Informar	X					X
Crédito Individual Pré Aprovado		Informar	X					X
Crédito Individual Pré Aprovado		Informar	X					X
Depósito a Prazo	DP X	Actuar		X	X		X	
Depósito a Prazo	DP X	Actuar		X	X	X		
Depósito a Prazo	DP Y	Actuar		X	X		X	
Seguro Pessoal	SP X	Actuar		X	X		X	
Obrigações de Caixa	OC X	Actuar		X	X	X	X	
Obrigações de Caixa	OC X	Actuar		X	X		X	
Reactivação de clientes		Actuar		X			X	
Crédito Individual Pré Aprovado		Actuar		X	X	X	X	
Reactivação de clientes		Actuar		X			X	
Recuperação de Crédito		Actuar		X			X	
Depósito a Prazo	DP Y	Actuar	X	X	X	X		
Seguro Pessoal	SP X	Actuar		X	X		X	
Cartão de Crédito	CC X	Actuar		X	X		X	
Depósito a Prazo	DP Z	Actuar	X	X	X	X		
Depósito a Prazo	DP W	Actuar		X	X	X		
Informação sobre clausulado	particulares	Actuar		X		X		
Informação sobre clausulado	empresas	Informar	X			X		X
Aplicação Mista	DM X	Actuar		X	X		X	
Cartão de Crédito	CC Y	Actuar		X			X	
PPR	Reforço	Actuar		X	X		X	
PPR	Reforço	Actuar		X	X		X	
Cartão de Crédito	CC X	Actuar		X			X	
Seguro Pessoal	SP X	Actuar		X	X		X	
Obrigações de Caixa	OC X	Actuar		X	X		X	

Produto ou Serviço		**Objectivo**	**Relatório**		**Campo**	**Banca Telefónica**		***Home***
Grupo	**Nome**		**Visualizações**	**Contactos**	**Montante**	***Inbound***	***Outbound***	***banking***
Obrigações de Caixa	OC X	Actuar		**X**	**X**		**X**	
Depósito a Prazo	DP W	Actuar		**X**	**X**		**X**	
Depósito a Prazo	DP W	Actuar		**X**	**X**		**X**	
Depósito a Prazo	DP Y	Actuar	**X**	**X**	**X**		**X**	**X**
Depósito a Prazo	DP Y	Informar	**X**					**X**
Obrigações de Caixa	OC X	Actuar		**X**	**X**		**X**	
Obrigações de Caixa	OC X	Actuar		**X**	**X**		**X**	
Seguro Pessoal	SP X	Actuar		**X**	**X**		**X**	
Apoio domiciliário	CR	Actuar		**X**			**X**	
Obrigações de Caixa	OC X	Actuar		**X**	**X**		**X**	
Cartão de Crédito	CC X	Actuar		**X**	**X**		**X**	
Crédito Individual Pré Aprovado		Informar	**X**					**X**
Crédito Individual Pré Aprovado		Informar	**X**					**X**
Crédito Individual Pré Aprovado		Informar	**X**					**X**
Crédito Individual Pré Aprovado		Informar	**X**					**X**
Crédito Individual Pré Aprovado		Actuar		**X**	**X**		**X**	
Depósito a Prazo	DP U	Actuar		**X**	**X**		**X**	
Seguro viagem	Inquérito Satisfação	Actuar		**X**			**X**	
Obrigações de Caixa	OC X	Actuar		**X**	**X**		**X**	
Obrigações de Caixa	OC X	Actuar		**X**	**X**		**X**	
Depósito a Prazo	DP Y	Actuar		**X**	**X**		**X**	
Cartão de Crédito	CC X	Actuar		**X**			**X**	
Obrigações de Caixa	OC X	Actuar		**X**	**X**		**X**	
Seguro Pessoal	SP X	Actuar		**X**	**X**		**X**	
Obrigações de Caixa	OC X	Actuar		**X**	**X**		**X**	
Seguro específico	SE X	Actuar		**X**			**X**	
Obrigações de Caixa	OC X	Actuar		**X**	**X**		**X**	
Depósito a Prazo	DP Y	Actuar		**X**	**X**		**X**	
Depósito a Prazo	DP Y	Actuar	**X**	**X**	**X**	**X**	**X**	**X**
Crédito Individual Pré Aprovado		Informar	**X**					**X**
Associados	Validação de Dados	Actuar	**X**	**X**		**X**		**X**
Obrigações de Caixa	OC Y	Informar	**X**	**X**				**X**
Depósito a Prazo	DP Y	Actuar	**X**	**X**	**X**	**X**	**X**	**X**
Cheques	Aviso Cheques por levantar	Actuar	**X**	**X**		**X**	**X**	
Cartão de Crédito	CC Z	Actuar	**X**	**X**		**X**	**X**	
Cartão de Crédito	CC U	Informar	**X**					**X**

Produto ou Serviço		Objectivo	Relatório		Campo	Banca Telefónica		*Home*
Grupo	**Nome**		**Visualizações**	**Contactos**	**Montante**	***Inbound***	***Outbound***	***banking***
Gestores de Cliente	Inquérito de Satisfação	Actuar		**X**			**X**	
Depósito a Prazo	DP Y	Actuar	**X**	**X**	**X**	**X**	**X**	**X**
Crédito Automóvel		Informar	**X**					**X**
Depósito a Prazo	DP V	Informar	**X**					**X**
Depósito a Prazo	DP M	Informar	**X**					**X**
Depósito a Prazo	DP E	Informar	**X**					**X**
Depósito a Prazo	DP Y	Actuar	**X**	**X**	**X**	**X**	**X**	**X**
Depósito a Prazo	DP Y	Actuar	**X**	**X**	**X**	**X**	**X**	**X**

As campanhas com o mesmo objectivo de negócio possuem o mesmo grupo e produto/serviço, conforme analisado e descrito em 4.1.1.3. Para cada campanha específica, pode existir o objectivo de "Informar", ou seja, apenas se pretende informar o cliente, pelo que é apenas produzido o relatório de visualizações, ou o objectivo de "Actuar", sendo este termo suficientemente abrangente quer para indicar que se pretende que os clientes subscrevam um produto ou serviço no âmbito da campanha, quer para outros mais específicos.

Nas campanhas de "Reactivação de clientes" o sucesso é medido pelo assistente quando determina um resultado correspondente, o qual se traduz na retenção do cliente ou nas campanhas de "Recuperação de crédito" em que o sucesso é a recuperação de – parte – do crédito em dívida.

De um modo ou de outro, o resultado é sempre derivado da formação específica transmitida aos assistentes no âmbito de cada campanha e pode carecer de alguma subjectividade inerente a uma avaliação casuística, em especial para objectivos mais complexos (por exemplo, como traduzir o sucesso para uma campanha de "Associados/Validação de Dados"? Será que é o facto de o cliente ter permitido essa validação? Ou será o facto de os dados *à priori* já estarem correctos?).

Adicionalmente, importa informar que, devido a questões de confidencialidade, os nomes reais identificativos dos diversos produtos e serviços foram mascarados.

As campanhas com linhas a cinzento significam que foram excluídas por não ser possível analisar se o objectivo das mesmas foi atingidas – na prática, significa que apenas se dispõe dos dados de visualização, ou seja, são campanhas informativas, não existindo nenhuma campanha subsequente com o mesmo "Grupo" e "Produto ou Serviço", mas com o objectivo de "Actuar".

Um valor de "X" na coluna "Campo Montante" indica que existe uma quantificação do objectivo a atingir traduzido num valor financeiro em €. Assim, nestas campanhas, se o resultado for um sucesso, existirá um montante associado a esse sucesso. As três últimas colunas indicam os canais de comunicação em que a campanha respectiva decorreu.

B. Aplicação de Carregamento de Dados

Conforme referido em 3.2, os dados para o estudo são obtidos através de vários relatórios, de dois tipos, um com o registo de visualizações, e outro com o registo dos resultados de contactos através de canais síncronos bidireccionais. Cada campanha pode ter apenas um relatório de visualizações e apenas um relatório de resultados, podendo, no entanto, ter um de cada tipo.

Por outro lado, a informação caracterizadora de cada campanha (a qual se encontra espelhada na Tabela 25) apenas estava disponível de forma visual, através de uma aplicação *Web*. Assim, uma das acções óbvias seria transpor esta informação para um formato electrónico que permitisse a sua utilização para uma análise geral das campanhas. Para tal, foi definido um ficheiro de propriedades para cada campanha para onde foram copiados os dados extraídos da aplicação *Web*. Esses dados são um reflexo directo das colunas da Tabela 25.

Com o intuito de conglomerar quer os dados dos vários relatórios, quer os ficheiros de propriedades caracterizadoras das campanhas, foi implementada uma aplicação *Web* em tecnologia *Java*[28], assente no servidor aplicacional *Tomcat*[29] para importar os diversos dados para uma Base de Dados relacional implementada em *MySQL*[30]. Todas as tecnologias utilizadas e referidas atrás têm a vantagem de serem *freeware*[31].

Para o armazenamento dos dados foram criadas as seguintes tabelas:

- DADOS_VISUALIZACAO – dados de relatórios de visualizações;
- DADOS_CONTACTO – dados de relatórios de resultados de contactos;
- CLIENTES – dados de clientes fornecidos;
- CONTACTOS – dados dos contactos dos clientes (números de telefones);
- CAMPANHAS – dados caracterizadores das campanhas;
- CANAIS – canais de comunicação e suas características.

Não se irá entrar em detalhe relativamente ao formato de cada tabela, por se considerar que foge completamente do âmbito deste documento.

Esta aplicação implementada é referida sucintamente neste documento em anexo uma vez que, embora não se enquadre directamente em nenhum passo do processo de descoberta de conhecimento, permitirá obter de uma forma integrada os ficheiros de dados exactamente com os atributos pretendidos e no formato adequado à sua utilização como *input* de modelos de *Data Mining* implementados pelas ferramentas a utilizar. A vantagem para o autor foi o facto de o mesmo dispor de bons conhecimentos de SQL e conseguir facilmente consolidar os dados a analisar num único ficheiro.

[28] *Java* é uma marca registada da *Oracle Corp.* (http://www.oracle.com/technetwork/java/index.html).
[29] *Tomcat* é uma marca registada da *Apache Software Foundation* (http://tomcat.apache.org/).
[30] *MySQL* é uma marca registada da *Oracle Corp.* (http://www.mysql.com/).
[31] *Freeware – software* de acesso gratuito.

C. Resumo de Dados

Neste anexo apresenta-se na Tabela 26 o dicionário de dados envolvidos na investigação. Os dados encontram-se agrupados de acordo com a sua origem. Nem todos os atributos são utilizados em todas as iterações do CRISP-DM, pelo que as últimas colunas são indicadores da utilização ou não do respectivo atributo na iteração correspondente à coluna (se o valor for X, então o atributo foi usado na iteração correspondente).

Para alguns dos tipos nominais, nos próprios dados da origem e que foram fornecidos para esta investigação já existiam valores que representavam o desconhecimento, ou seja, os *missing values*. Assim, quando, na coluna respectiva, forem contabilizados os *missing values*, para além dos valores NA (*not available*, ou seja, não está disponível), são indicados também aqueles para os quais existe um valor explícito.

Relativamente aos valores possíveis e *missing values* apresentados, importa reter que os mesmos foram recolhidos para os dados utilizados a partir da segunda iteração da metodologia CRISP-DM – da primeira para a segunda iteração, o objectivo foi revisto (ver 4.2.1), resultando daí que vários registos foram eliminados do conjunto de dados por não corresponderem aos novos objectivos.

Tabela 26 - Dicionário de Dados

Grupo de Dados 1 - Identificadores do Registo (3)

Nome	Descrição e valores possíveis	Tipo	*Missing values*	Iterações		
				1	2	3
ID Cliente	Identificador mascarado de cliente (>0)	Numérico	- - -			
ID Campanha	Identificador mascarado de campanha (>0)	Numérico	- - -	X		
ID Objectivo	Identificador do objectivo (ver Tabela 9) (>0)	Numérico	- - -	X		

Grupo de Dados 2 - Informação Pessoal de Cliente (13 atributos)

Nome	Descrição e valores possíveis	Tipo	*Missing values*	Iterações		
				1	2	3
Idade	Em anos, e à data do último contado para o par <cliente, campanha> (>0)	Numérico	- - -	X	X	X
Profissão	São 1726 os valores possíveis	Nominal	11470 *desconhecidos* e 5597 NA's	X	X	
Situação profissional	Situação laboral (são 27 os valores possíveis)	Nominal	9144 *desconhecidos* e 2 NA's	X	X	X
Estado civil	Casado, Divorciado, Separado, Solteiro e Viuvo	Nominal	10 *desconhecidos* e 55 *omissos no documento*	X	X	X
Morada	Dados referentes à morada de residência	Nominal	247 NA's			
Morada complementar		Nominal	247 NA's			
Localidade		Nominal	4210 NA's	X	X	
Código postal		Nominal	247 NA's	X	X	
Código postal local		Nominal	247 NA's	X	X	
Freguesia		Nominal	1923 NA's	X	X	
Título honorífico	22 valores possíveis	Nominal	23 NA's	X	X	X
Sexo	(M)asculino ou (F)eminino	Nominal (binário)	- - -	X	X	
Habilitações Literárias	12 ANO, 4 ANO, 9 ANO, CICLO, CURSO MEDIO, CURSO SUPERIOR, DESCONHECIDAS, INFERIOR AO 4 ANO, OUTRO	Nominal	1319 *desconhecidas* e 27 NA's	X	X	X

Grupo de Dados 3 - Informação Bancária de Cliente (13 atributos)

Nome	Descrição e valores possíveis	Tipo	*Missing values*	Iterações		
				1	2	3
Bloqueios gerais	Indicador da existência de bloqueios gerais sobre o cliente bancário - (S)im ou (N)ão	Nominal (binário)	- - -	X	X	
Bloqueios informativos	Indicador da existência de bloqueios informativos - (S)im ou (N)ão	Nominal (binário)	- - -	X	X	
Bloqueios de cheques	Indicador da existência de bloqueios para a utilização de cheques - (S)im ou (N)ão	Nominal (binário)	- - -	X	X	
Inibição de cheques	Indicador da inibição total na utilização de cheques - (S)im ou (N)ão	Nominal (binário)	- - -	X	X	
Associado	O cliente é associado da associação mutualista da instituição? - (S)im ou (N)ão	Nominal (binário)	- - -	X	X	
Créditos em mora	Indicador da existência de prestações de créditos em mora - (S)im ou (N)ão	Nominal (binário)	- - -	X	X	X
Saldo médio anual	Saldo médio anual das contas à ordem das quais o cliente é titular	Numérico	7537 NA's	X	X	X
Cartão de débito	Indicador da existência de bloqueios para a utilização de cheques - (S)im ou (N)ão	Nominal (binário)	- - -	X	X	X
Conta ordenado	O cliente tem conta ordenado? - (S)im ou (N)ão	Nominal (binário)	- - -	X	X	
Cartão de crédito	O cliente tem cartão de crédito? - (S)im ou (N)ão	Nominal (binário)	- - -	X	X	X
Crédito habitação	O cliente tem crédito habitação? - (S)im ou (N)ão	Nominal (binário)	- - -	X	X	X
Crédito individual	O cliente tem crédito individual? - (S)im ou (N)ão	Nominal (binário)	- - -	X	X	X
Domiciliações	O cliente tem domiciliações subscritas? - (S)im ou (N)ão	Nominal (binário)	- - -	X	X	X

Grupo de Dados 4 - Informação Geral de Contacto (1 atributo)

Nome	Descrição e valores possíveis	Tipo	*Missing values*	Iterações		
				1	2	3
Número de contactos	Número total de contactos efectuados no contexto de uma campanha, incluíndo o último	Numérico	- - -	X	X	X

Grupo de Dados 5 - Informação do Último Contacto (11 atributos)

Nome		Descrição e valores possíveis	Tipo	*Missing values*	Iterações		
					1	2	3
Agente		Identificador mascarado do agente que atendeu a chamada	Nominal	- - -	X	X	X
Tipo de telefone do cliente		Tipo de telefone utilizado – (F)ixo ou (M)óvel	Nominal	10783 desconhecidos	X	X	X
Contacto	Dia da semana	Dia da semana em que o contacto foi feito – *Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, Sunday*	Nominal	- - -	X	X	
	Dia do mês	1 a 31, consoante o mês	Nominal	- - -	X	X	X
	Mês	De 1 a 12 (Janeiro a Dezembro)	Nominal	- - -	X	X	X
	Hora	De 7 a 22 (o *Contact-Center* só funciona entre as 7h00 e a 1h00 da manhã, sendo os contactos de *oubound* despoletados entre as 10h00 e as 23h00) – valor inteiro, ou seja, se o contacto foi feito às 9h49, o valor para este atributo é 9	Nominal	- - -	X	X	X
Agendamento	Dia da semana	Estes atributos têm significado igual aos equivalentes com o mesmo nome para o contacto	Nominal	- - -	X		
	Dia do mês		Nominal	- - -	X		
	Mês		Nominal	- - -	X		
	Hora		Nominal	- - -	X		
Duração		Duração do contacto em segundos	Numérico	- - -	X	X	X

Grupo de Dados 6 - Informação do Primeiro Contacto (8 atributos)

Nome		Descrição e valores possíveis	Tipo	*Missing values*	Iterações		
					1	2	3
Resultado		Um dos valores apresentados na Tabela 5	Nominal	17069 NA's (os mesmos dizem respeito a todos os casos em que houve apenas um contacto, o último – ou seja, nestes casos, a informação de primeiro contacto não foi preenchida, por ser igual à do último e único contacto)	X	X	X
Agente		Identificador mascarado do agente que atendeu a chamada	Nominal		X	X	X
Tipo de telefone do cliente		(F)ixo ou (M)óvel (e 7033 desconhecidos, para além dos NA's)	Nominal		X	X	X
Contacto	Dia da semana	Dia da semana em que o contacto foi feito – *Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, Sunday*	Nominal		X	X	
	Dia do mês	1 a 31, consoante o mês	Nominal		X	X	X
	Mês	De 1 a 12 (Janeiro a Dezembro)	Nominal		X	X	X
	Hora	De 7 a 22 (o *Contact-Center* só funciona entre as 7h00 e a 1h00 da manhã, sendo os contactos de *oubound* despoletados entre as 10h00 e as 23h00) – valor inteiro, ou seja, se o contacto foi feito às 9h49, o valor para este atributo é 9	Nominal		X	X	X
Duração		Duração do contacto em segundos	Numérico		X	X	X

Grupo de Dados 7 - Informação de Visualizações (4 atributos)

Nome	Descrição e valores possíveis	Tipo	*Missing values*	Iterações		
				1	2	3
Última visualização do agente	Data e hora da última visualização pelo agente da campanha (pode ou não ter abordado o cliente!)	Numérico	37819 NA's	X	X	
Número de visualizações do agente	Número de visualizações da campanha pelo agente aquando de contactos do cliente	Numérico		X	X	
Última visualização no *homebanking*	Data e hora da última visualização do cliente da campanha no *homebanking*	Numérico	39029 NA's	X	X	
Número de visualizações no *homebanking*	Número de visualizações da campanha pelo agente aquando de contactos do cliente	Numérico		X	X	

Grupo de Dados 8 - Informação de Histórico (10 atributos)

Descrição e valores possíveis	Tipo	*Missing values*	Iterações		
			1	2	3
N.º de dias desde o último contacto para qualquer campanha	Numérico	- - -	X	X	
N.º de dias desde o primeiro contacto para qualquer campanha	Numérico	- - -	X	X	X
N.º total de contactos efectuados anteriormente	Numérico	- - -	X	X	X
N.º total de sucessos anteriores	Numérico	- - -	X	X	
N.º total de insucessos anteriores	Numérico	- - -	X	X	
Resultado do último contacto para a última campanha – Um dos valores apresentados na Tabela 5 (estes valores foram recolhidos antes da 2ª iteração, daí poderem surgir valores diferentes de "Sucesso" ou "Insucesso")	Nominal (binário)	45462 NA's	X	X	X
Montante subscrito (em euros) na última campanha	Numérico	- - -	X	X	
Montante total subscrito para todas as campanhas	Numérico	- - -	X	X	
N.º total de visualizações para todas as campanhas no *homebanking*	Numérico	- - -	X	X	
N.º total de visualizações para todas as campanhas na banca telefónica	Numérico	- - -	X	X	