Modelos de Propensão ao Consumo baseados em Redes Neuronais Artificiais, o caso particular do Crédito Pessoal

por

Verónica Filipa dos Santos Silva

Dissertação apresentada como requisito
parcial para obtenção do grau

Mestre em Estatística e Gestão da Informação

pelo

Instituto Superior de Estatística e Gestão da Informação
da
Universidade Nova de Lisboa

Agradecimentos

Os meus agradecimentos dirigem-se, em primeiro lugar, ao meu orientador, Prof. Fernando Bação, pela sua disponibilidade, aos seus contributos bastante positivos e a paciência prestados durante a realização deste trabalho. Agradeço, ainda, à Dra. Sandra Catarino pelo apoio e disponibilidade na parte prática, em especial, sobre a utilização do SAS.

Gostaria, ainda, de expressar a minha gratidão à instituição financeira que me possibilitou o acesso às suas bases de dados de clientes tornando possível o desejo de realizar um projecto que possa ter utilidade no futuro. Em especial, aos meus directores que mais apoio me deram em diferentes momentos deste trabalho, o Paulo e a Isabel.

Por fim, ao João pelo apoio e compreensão infindáveis e, juntamente, a toda a minha família pelo apoio que transmitiu nos últimos meses de finalização deste trabalho.

Resumo

O objectivo do presente trabalho consiste na criação de um modelo de propensão ao consumo aplicado ao caso particular do Crédito Pessoal através da aplicação de uma das técnicas mais populares de *data mining*: as redes neuronais. O tema reside na selecção dos clientes mais propensos para constituição de um alvo a submeter a uma campanha reduzindo os custos de *mailing* e, em simultâneo, aumentando a taxa de resposta. Os dados utilizados baseiam-se nos clientes alvo de uma campanha que ocorreu no ano anterior à realização deste trabalho numa instituição financeira portuguesa, e que inclui variáveis como as características pessoais destes clientes e outras relacionadas com o envolvimento do cliente com a instituição. O desempenho dos modelos baseados nas redes neuronais, mais concretamente no algoritmo de Retropropagação, foi comparado com o dos modelos de regressão logística, sendo esta uma técnica mais tradicional e normalmente utilizada neste tipo de problemas de previsão. Este trabalho apresenta também as várias etapas do processo de construção dos modelos, desde a preparação dos dados em que, perante uma amostra desproporcional em termos de dimensão das duas classes da *target* se recorre às técnicas de *oversampling*, até à análise dos resultados. Dado o número elevado de modelos testados recorre-se ao critério do menor *RMSE* e *misclassification rate* para restringir o conjunto de modelos em análise. Após selecção dos melhores modelos, a sua performance foi avaliada com base em estatísticas como o *lift*, a curva *ROC* e matriz de custos/benefícios. Em conclusão, as redes neuronais originaram os melhores modelos, no entanto, os resultados obtidos não apresentaram diferenças muito expressivas comparativamente com a regressão logística.

Palavras-chave: modelos de propensão, redes neuronais, algoritmo de Retropropagação, regressão logística, *oversampling*, curva ROC, *Lift*.

Abstract

The aim of the present work consists in creating a propensity model applied to a particular case of the Personal Loan applying one of the most popular data mining techniques: the neural networks. The problem resides in the selection of the most disposed clients to constitute a target for a campaign reducing the mailing costs and, simultaneously, increasing the success rate. The data used in this work is based on a target of a campaign that occurred on a Portuguese financial institution on the year before the realization of this work and this database includes variables such as personal features and others related to the relation with the bank. The performance of the models created from neural networks, more precisely, based on BackPropagation was compared with the logistic regression models performance. The logistic regression is a more traditional technique and usually considered as appropriated to this prevision problems. This work presents the several steps of the process of models construction, since the data preparation where unbalanced sample needed to be oversampled, to the results analysis. Since it was obtained a great number of models to be compared, the minor RMSE and the minor misclassification rate were the selection criteria's to restrict the models set in analysis. After selecting the best models, these performances were evaluated based on different statistics: lift, ROC curve and matrix of costs/profits. In conclusion, neural networks derived the best models, however, the achieved results did not present great differences comparing to logistic regression.

Keywords: propensity models, neural networks, BackPropagation algoritm, logistisc regression, oversampling, ROC curve, Lift.

Índice

Lista de Figuras

Lista de Tabelas

Glossário

Dados/vectores de *input* – valores que as variáveis explicativas/dependentes tomam para cada indivíduo da amostra ou população.

Dados/vectores de *output* – valores estimados/previstos da variável em estudo em função das variáveis explicativas.

Data Mining – análise de dados com o objectivo de descobrir relações ou padrões não conhecidos, normalmente, em grandes volumes de dados.

Score – pontuação.

***Scoring* de Crédito** – método de atribuição de pontuações de risco de crédito com o objectivo de apoiar a decisão da concessão de crédito, baseado num valor de corte.

Target – variável em estudo; variável dependente ou de resposta.

Threshold – limiar ou ponto de quebra.

RMSE (Root Mean Square Error) – a raíz quadrada do erro quadrático médio.

Missclassification Rate – Taxa de classificações incorrectas.

1 Introdução

1.1 Tema

Nas sociedades contemporâneas, o consumo é um dos factores que mais tem contribuído e estimulado as actividades e o crescimento económicos. A denominação de 'sociedades de consumo', que é atribuída sobretudo às sociedades dos países ocidentais, advém da melhoria geral das condições de vida que se verificou nestes países a partir dos anos 60. Por outro lado, o progresso tecnológico e empresarial tem incidido cada vez mais, e com sucesso, em estratégias de aliciamento ao aumento do consumo, seja por via da publicidade seja pela criação constante de produtos e serviços que acrescentam algo aos já existentes.

No entanto, a conjuntura económica mundial que se tem feito sentir desde o início deste século, tem dificultado o acesso aos bens e serviços referidos, na medida em que a desaceleração das economias resultou numa subida do desemprego e na redução do poder de compra (sobretudo nas designadas classes médias).

Paralelamente, as instituições financeiras aproveitaram o decréscimo e estabilidade das taxas de juro, durante aquele período, para actuar cada vez mais como grandes financiadoras de indivíduos e empresas, tendo-se tornado também comum o recurso a soluções de financiamento para fazer face aos hábitos e necessidades de consumo. Nessa perspectiva, é cada vez mais frequente encontrar-se instituições financeiras com uma vasta panóplia de produtos de crédito para o consumo, convergindo para a satisfação das necessidades específicas e possibilidades de endividamento dos seus clientes.

Neste contexto, o Crédito Pessoal tem tido um grande desenvolvimento e aderência pelos consumidores sendo um produto muito direccionado para o financiamento rápido e fácil e que, como tal, tem servido como um grande estímulo ao consumo.

O aumento da concorrência e o valor que este negócio gera para as entidades financeiras tem impulsionado a ocorrência frequente de campanhas publicitárias, anunciando as melhores condições de financiamento para a compra de automóveis, material informático ou pacotes de férias, e tirando partido de factores como a sazonalidade na procura e o facto de existir cada vez menos apetência para a poupança. São também comuns as campanhas destinadas

exclusivamente aos clientes das instituições, aos quais são oferecidas condições vantajosas, com vista a aumentar a captação nestes produtos.

Devido ao crescimento considerável do peso e do potencial que representa este produto e, simultaneamente, sendo a concorrência cada vez maior e mais competitiva, existe, do ponto de vista das entidades financiadoras, uma crescente preocupação em optimizar o produto a as actividades subjacentes em termos da rendibilidade gerada e do risco inerente.

Uma das formas através das quais estas entidades têm tentado atingir esse objectivo, consiste na concepção de produtos e campanhas de divulgação que dêem resposta às necessidades dos seus clientes, e do mercado em geral. Mais concretamente, o estudo prévio do perfil dos clientes permite conhecer quais os propensos e os não propensos à aquisição de produtos como o Crédito Pessoal num determinado contexto e período de tempo.

O desenvolvimento de modelos de propensão (à aquisição de Crédito Pessoal) surge, assim, como resposta às questões inerentes à caracterização dos grupos de clientes que poderão ser alvo de campanhas ou condições pré-aprovadas e à realização de previsões sobre o sucesso das mesmas. Para tal, torna-se fundamental utilizar a informação disponível sobre os seus clientes, registada na forma de dados sobre as suas características pessoais e resultantes das transacções que envolvem a instituição, para extracção de conhecimento sobre o seu perfil e sobre as suas necessidades.

Por outro lado, devido à dimensão e ao elevado ritmo no aumento dos conjuntos de dados, os métodos de análise tradicionais têm vindo, progressivamente, a tornar-se ineficazes ou insuficientes para a extracção de informação e conhecimento úteis às organizações. O enorme número de dados e o número de variáveis susceptíveis de se relacionar tornam as tarefas computacionais pesadas e dispendiosas, verificando-se simultaneamente a inadequação dos modelos conhecidos.

Assim, as técnicas de *data mining*, como processo de extracção de informação, que se encontra oculta em bases de dados e que representa realidades que podem servir como mais-valia no apoio à decisão das organizações, vem de uma certa forma dar resposta a estes problemas.

A diversidade de ramos de negócio que uma instituição financeira pode comportar, implica a existência de vários tipos de problemas cuja resolução pode ser atingida através da utilização de *data mining*. Segundo Berry e Linoff (2000), existem seis fins ou actividades para os quais

o conhecimento extraído do processo de *data mining* pode ser utilizado: classificação, estimação, previsão, agrupamento ou criação de regras de associação, *clustering* e descrição.

Classificação consiste no processo de descoberta de um conjunto de modelos ou funções que descrevam ou distingam classes de dados (Han e Kamber, 2001). Através deste processo será possível classificar um determinado objecto de estudo e associá-lo a classes pré-definidas. Esta tarefa é muito utilizada na análise de risco, sobretudo nos casos de *scoring* de crédito em que, após realizada uma análise ao perfil do cliente, é-lhe associado um grau de risco e atribuída uma decisão de aprovação ou reprovação do crédito solicitado; e do *scoring* comportamental em que o recurso a abordagens de classificação pode ajudar na obtenção de uma melhor performance e estimação dos modelos (Hosemann e Fritz, 1999, citados por Klösgen e Zytkow, 2002, p.775), atribuindo a probabilidade de um cliente ou um produto alterar o seu nível de risco num determinado período de tempo.

No processo de classificação podem-se incluir também as tarefas de estimação dado que a resolução da maior parte dos problemas de classificação passa pela estimação de uma variável ou até mesmo pela atribuição de uma probabilidade.

Frequentemente, ao se fazer referência às tarefas de *data mining* o conceito de previsão confunde-se com o de classificação e de estimação uma vez que os dados são classificados de acordo com a previsão de comportamentos futuros ou a estimação de valores esperados (Berry e Linoff, 2000), como se poderá verificar nos exemplos dados anteriormente. No contexto de campanhas de crédito pessoal ou de cartões de crédito, é muito recorrente a aplicação da previsão para a modelação das respostas de um determinado conjunto de clientes a um *mailing* oferecendo condições vantajosas para aquisição de crédito ou para a activação de um cartão.

A identificação de associações consiste em encontrar tendências em conjuntos de dados no que diz respeito à sua relação de dependência, permitindo a exploração eficaz de padrões de comportamento. Com base nos registos de aquisição de vários produtos financeiros, poder-se-á verificar a existência da forte correlação entre a aquisição de dois ou mais produtos. Aproveitando esta relação entre os produtos/serviços, a instituição poderá agrupá-los em *packages* atractivos aos seus clientes.

Uma das técnicas mais populares é o *clustering* que, ao contrário da classificação e da previsão, em que se analisam dados através da associação a classes pré-definidas, analisa dados sem conhecimento prévio das classes. Os *clusters* dos objectos são formados para que

os objectos dentro do mesmo *cluster* tenham elevada similaridade mas muito distintos relativamente aos objectos pertencentes a outros *clusters* (Han e Kamber, 2001). No campo da análise de risco do crédito, esta técnica é utilizada numa fase inicial em que se definem as classes de clientes de baixo, médio e alto risco, e sempre que, por algum motivo, seja necessário averiguar a existência de uma nova classe de risco ou redefinir as classes existentes.

Por fim, a descrição permite apresentar a dependência dominante entre algumas variáveis. Nesta actividade, o conhecimento não é utilizado para realizar previsões (e.g., determinar quais os comportamentos futuros dos clientes), mas sim descrições de relações entre variáveis, podendo desta forma explicar alguns desses comportamentos (Klosgen e Zytkow, 2002).

O enquadramento do processo de *data mining* em qualquer empresa deverá ser analisado à luz dos objectivos e das áreas organizacionais em que é aplicado.

Requisitos comuns para a sua boa utilização e manutenção passam pelo bom conhecimento da área de negócio e do problema a resolver, pela capacidade de recolha e armazenamento de dados que devem representar o máximo da realidade inerente ao problema em questão, por um conhecimento tão amplo quanto possível das ferramentas e métodos existentes no motor de *data mining* e na capacidade de escolha dos mais adequados a cada problema, e, por fim, o conhecimento necessário para relacionar os resultados obtidos no processo de *data mining* com o contexto organizacional para atingir tomadas de decisão que resultem em mais-valias.

Como exemplo de sucesso, as redes neuronais representam um método poderoso para resolução de problemas de classificação e/ou previsão. Segundo Nürnberger, Pedrycz e Kruse citados por Klosgen e Zytkow, 2002, p. 304, "quando comparadas com os métodos estatísticos, as redes neuronais são úteis se não existir conhecimento prévio sobre a relação entre variáveis independentes e variáveis dependentes. Para além disso, as redes neuronais são facilmente extensíveis a várias variáveis independentes e dependentes sem que ocorra um aumento exponencial nos seus parâmetros".

Face às grandes potencialidades das redes neuronais artificiais, neste estudo, a estimação do modelo de propensão ao consumo de Crédito Pessoal será realizada através deste método. Realizar-se-á também a estimação de um modelo de propensão com base num modelo de regressão logística, sendo que *a posteriori* serão comparados os resultados da utilização dos dois métodos.

1.2 *Background* Científico

A pesquisa realizada para suporte deste trabalho incidiu, em primeira instância, em autores que pudessem contribuir para uma melhor compreensão dos problemas de propensão e das técnicas de modelação actualmente existentes para resolução dos mesmos.

A oferta de um determinado produto e consequentemente a sua aceitação no mercado tem de ser sempre equacionada face à propensão que os consumidores têm para o consumo desse produto. Existem vários factores que podem condicionar a propensão de um indivíduo ao consumo de um determinado bem ou serviço. O mais significativo será a necessidade que o indivíduo tem para o seu consumo. Se num negócio de venda de produtos de primeira necessidade a propensão de qualquer indivíduo é praticamente absoluta, quando se é confrontado com o problema da propensão para consumo de produtos financeiros, a estimação desse valor torna-se muito mais complexa, e muito menos atingível através de análises directas. Nesse caso, terão de ser levados em consideração factores que permitam diferenciar os consumidores e as respectivas decisões face ao consumo de um produto, que podem variar consoante a personalidade e preferências de cada indivíduo ou mesmo com o contexto no qual é realizada a compra.

Assim, para atingir conclusões sobre a propensão dos consumidores à aquisição de um bem ou serviço, pode ser necessária a criação de um modelo que através do *input* da realidade do(s) consumidor(es) (características pessoais, histórico de transacções, contexto socio-económico) possa responder com precisão a questões como irá, ou não, o indivíduo consumir determinado produto?

Desta forma, os modelos de propensão podem ser uma ferramenta de suporte à decisão em termos da concepção de um negócio ou produto, na medida em que disponibilizam conhecimento sobre os potenciais consumidores e, consequentemente, sobre o sucesso de um determinado produto ou campanha.

Dada a importância destas ferramentas e uma vez que existe uma vasta panóplia de modelos de propensão, a escolha do modelo que mais se adequa a cada problema específico torna-se numa questão da maior importância, sendo necessário considerar as diferenças existentes nos consumidores e em tudo o que os envolve. Neste contexto, os autores Roberts e Lilien, citados por Eliashberg e Lilien, 1993, p. 28, apresentam algumas justificações para a variedade de modelos que existem sobre o comportamento do consumidor, e que serão explanadas adiante.

No que diz respeito à personalidade dos consumidores, aos seus valores ou preferências, a diferença entre cada indivíduo pode significar que um modelo apropriado para descrever o comportamento de um consumidor, em particular, pode ser desadequado para outros consumidores ainda que se esteja a estudar a propensão a um mesmo produto ou campanha de vendas.

Da mesma forma, um modelo de propensão ao consumo deve ter em consideração que um mesmo indivíduo toma diferentes decisões mediante o produto que lhe é proposto, pelo que o modelo também se deve adequar ao produto em causa. A este respeito, e tendo em conta que neste trabalho se irá a abordar a propensão ao consumo de um serviço financeiro, ao qual poderá estar associado um investimento substancial e algum risco na adesão ao serviço, será de supor que, neste caso, a decisão de cada consumidor terá em consideração uma série de factores, que também terão necessariamente que ser levados em conta na construção do modelo de propensão.

Outro factor importante é o contexto no qual é tomada a decisão de compra de um produto, uma vez que a decisão de compra pode variar consoante a finalidade e circunstâncias da compra ou factores como o preço e a qualidade do produto a adquirir. No âmbito do produto de Crédito Pessoal, estes factores poderão resumir-se ao facto do cliente estar a responder ou não a uma campanha (com um preço vantajoso), ou se apresenta um quadro financeiro não muito favorável para adquirir algum bem/serviço que necessite em determinado momento.

Como já referido anteriormente, o conhecimento sobre o perfil e o comportamento do cliente é fundamental para a tomada de decisão das instituições, nomeadamente, na definição de produtos adequados aos seus clientes e de estratégias de *marketing*. Os modelos de propensão ao consumo estabelecem uma relação entre as características dos indivíduos e a probabilidade de aquisição de um dado produto, traduzindo-se assim no conhecimento de que a instituição necessita.

Roberts e Lilien, citados em Eliashberg e Lilien, 1993, p. 30, defendem que a necessidade, que motiva a compra, é activada por um estímulo interno ou externo. No primeiro caso, este estímulo poderá representar o sentimento de privação ou de desejo do indivíduo em adquirir um bem, seja este essencial ou supérfluo. No segundo, esta necessidade poderá ser impulsionada por um anúncio ou uma acção de promoção de um produto. É assim que os consumidores decidem se adquirem determinado bem ou não, ou qual a marca que devem escolher. Os modelos para prever uma ou outra das situações mencionadas (a decisão, ou não,

de compra de determinado produto ou a escolha de uma marca) recaem na teoria da escolha discreta.

"Quando existem exactamente duas escolhas (comprar o produto/não comprar o produto) os modelos de escolha discreta, também designados por modelos de escolha binária, têm sido aplicados a uma vasta variedade de problemas de classificação na área do marketing e noutras áreas" (Bem-Akiva e Lerman, 1985, citado por Eliashberg e Lilien, 1993, p. 32).

Lilien e Kotler (1992) sugerem vários modelos de comportamento do consumidor, concentrando-se no seu processo de decisão. O modelo *logit* foi o modelo apresentado como o mais apropriado para problemas de escolha discreta. Segundo estes autores, Gensch determinou através deste modelo a probabilidade de compra de determinada marca de um produto entre um conjunto de j marcas. No seguimento deste raciocínio e adequando este exemplo ao caso em estudo, pode-se definir $j=1$, considerando apenas o Crédito Pessoal do banco X e obtendo a probabilidade de o cliente adquirir este produto ou não.

Hoetker (2007) defende que no caso em que se pretende criar um modelo para saber qual de duas alternativas ocorre, o modelo *logit* é apropriado. É mencionado o caso de uma empresa que perante a necessidade de uma componente, terá de decidir se a fabrica ou, em alternativa, se a compra sendo que neste caso, o modelo *logit* determinará se a empresa tem, ou não, propensão ao fabrico da componente.

Por outro lado, e como objectivo deste trabalho, abordam-se as redes neuronais artificiais na perspectiva de servirem como ferramenta para resolver o problema de propensão ao consumo já referido, servindo assim de alternativa ao modelo *logit* tão referido na bibliografia consultada relativa ao comportamento do consumidor.

"Os modelos baseados em redes neuronais artificiais fornecem uma alternativa viável aos modelos clássicos de regressão. Estes modelos podem aprender através da experiência, podem generalizar e ver para além do ruído e distorções e podem extrair as principais características, mesmo na presença de dados irrelevantes. Podem também fornecer um elevado nível de robustez e tolerância à falha. Para além disso, estes modelos podem descobrir as transformações correctas para variáveis, detectar relações lineares fracas e lidar com *outliers*" (Hill, Marquez, Remus e Worthley citados em Trippi e Turban, 1993, pp 435-436).

Desta forma, a utilização das redes neuronais e do modelo *logit*, surgem como duas técnicas para resolução do problema deste trabalho, que se traduz em encontrar um modelo de

propensão ao consumo do produto de Crédito Pessoal de uma determinada instituição financeira.

Goss e Ramchandani (1998) comparam estes dois métodos ao caso real em que se pretende determinar quais os pacientes de um serviço de cuidados intensivos hospitalares que têm uma probabilidade razoável de recuperação. Neste estudo, os autores desenvolvem e propõem um modelo não paramétrico, baseado em redes neuronais artificiais e demonstram que os resultados obtidos superam os obtidos pela utilização do modelo *logit*.

No negócio do crédito pessoal existem outros exemplos nos quais se pode proceder à comparação das duas metodologias, nomeadamente na área do *scoring* de crédito. Os modelos de *scoring* de crédito têm como objectivo classificar os clientes como bons ou maus pagadores do crédito concedido, sendo utilizados na decisão sobre a concessão do crédito.

Liu e Schumann (2005) desenvolveram um modelo de *scoring* com base em diferentes algoritmos de classificação sendo as redes neuronais e a regressão logística duas delas. Estes autores concentraram-se na performance dos modelos obtidos, considerando três aspectos importantes: a simplicidade do modelo, a rapidez de modelação e a precisão do modelo.

Os estudos em que a aplicação das redes neuronais são casos de sucesso para *scoring* de crédito são inúmeros. Citam-se como exemplos Baesens, Gestel, Stepanova, Poel e Vanthienen (2005); Huang, Hung e Jiau (2006); Atiya (2001).

Através dos casos expostos é possível aferir da variedade de problemas aos quais os modelos de propensão podem dar resposta, através da utilização das redes neuronais.

No estudo aqui proposto, utilizar-se-ão o modelo baseado em redes neuronais e o modelo de regressão logística, com vista a apurar qual o que apresenta a melhor performance para o problema de propensão ao consumo do produto de Crédito Pessoal.

Apesar de toda a pesquisa realizada para a justificação da escolha das referidas técnicas, torna-se importante salientar o facto de existir escassa bibliografia sobre a modelação da propensão e, nomeadamente, *papers* com exposição de problemas desta natureza.

1.3 Objectivos

O objectivo deste trabalho consiste assim em desenvolver ferramentas que possibilitem a melhoria da performance do processo de atribuição de Crédito Pessoal, passando pela correcta definição do perfil de clientes que têm propensão à aquisição de crédito e, mais

concretamente, pela definição do perfil dos clientes aos quais determinadas campanhas devem ser destinadas. Assim, procurar-se-á atingir as seguintes metas:

- Compreender e aplicar os processos de selecção e redução de variáveis;
- Desenvolver um modelo de propensão ao consumo baseado em redes neuronais artificiais para o produto de Crédito Pessoal;
- Compreender as diferenças entre os indivíduos mais propensos e os menos propensos à aquisição do Crédito Pessoal;
- Proceder à comparação dos resultados entre a performance dos modelos baseados em redes neuronais artificiais e em modelos de regressão logística (*logit*).

Para alcançar estas respostas serão utilizados dados reais fornecidos por uma instituição financeira, e serão estimados modelos de propensão ao consumo de forma a reconhecer padrões de comportamento dos clientes que possam auxiliar as instituições a optimizar o processo de atribuição de Crédito Pessoal, e elaborar previsões sobre as taxas de sucesso de determinados produtos/campanhas.

1.4 Relevância

A relevância deste trabalho resulta da necessidade de melhoria nas campanhas de marketing directo que são realizadas por empresas do sector financeiro, com o objectivo de promover um produto concreto.

Actualmente, existe uma imensidão de publicidade que, diariamente, invade as caixas de correio dos consumidores, gerando alguma saturação e aumentando a sua relutância em analisar qualquer uma das ofertas que lhe são propostas. Esta realidade torna a preocupação de direccionar as campanhas para os clientes "certos" (i.e., os mais propensos) mais pertinente pois, se por um lado, o contacto indiscriminado aos clientes proporciona a mencionada saturação e desperdiça recursos da empresa, por outro, o não contacto a clientes que se encontrem em contexto propício à aquisição do produto poderá significar uma perda financeira significativa para a empresa.

Com o objectivo de aumentar a capacidade resposta das empresas a este tipo de questões, a proposta deste trabalho consiste na obtenção do conhecimento das características dos clientes propensos a um determinado produto de natureza financeira comercializado por uma dada instituição bancária.

Neste contexto e adicionalmente aos argumentos já apresentados, o desenvolvimento de um modelo de propensão poderá representar uma vantagem competitiva na medida em que habilitará esta instituição a optimizar as suas campanhas, orientando-as apenas para os clientes que apresentam uma maior probabilidade de aderir ao produto. Desta forma, pretende-se garantir o aumento das taxas de resposta de cada campanha, beneficiando-se, simultaneamente, de um processo mais eficiente de selecção e comunicação aos clientes alvo da campanha, nomeadamente, através da diminuição do volume de *mailings* enviados ou contactos efectuados.

Apesar do problema deste trabalho ter sido formulado na perspectiva de produtos financeiros, em particular do Crédito Pessoal, e de uma instituição em particular, o conhecimento dos clientes propensos a um determinado produto constitui uma preocupação constante e transversal aos mais diversos sectores empresariais. Assim, pode afirmar-se que a perspectiva deste estudo assenta na aplicabilidade a outras áreas económicas ou sociais.

Outro aspecto a realçar está relacionado com a amostra escolhida, designadamente com a desproporção existente entre as dimensões das classes da variável *target* – a resposta afirmativa *versus* a não resposta a uma campanha de crédito pessoal. Esta é uma característica comum em amostras cuja variável *target* espelha a taxa de sucesso de campanhas e este trabalho descreve a forma como o problema foi ultrapassado através da aplicação do método de *oversampling* e o sucesso dos resultados obtidos.

O trabalho desenvolvido aborda o problema da propensão na perspectiva da modelação preditiva, nomeadamente, através da concepção de modelos que possam prever com a maior exactidão possível quais os alvos certos para uma determinada campanha. A sua relevância prende-se também com as técnicas de modelação utilizadas para determinar estes modelos, como mostram ser adequadas ou não ao problema em estudo e como é avaliada a sua performance.

As redes neuronais constituem o foco principal da modelação uma vez que esta tem mostrado ser uma ferramenta de previsão com sucesso apresentando resultados similares ou melhores que os de técnicas estatísticas como a regressão logística conforme mencionado na secção 1.2.. A este respeito deverá realçar-se que embora a escolha da melhor técnica não seja linear, a exposição dos benefícios *versus* restrições que cada técnica revela na resolução do problema de propensão, poderão fornecer informação útil no desenvolvimento de trabalhos futuros.

A importância deste trabalho traduz-se num contributo para a literatura existente sobre a propensão ao consumo, os métodos de previsão utilizados na resolução de um problema neste âmbito.

1.5 Metodologia

A metodologia adoptada neste estudo é quantitativa e a mais adequada para adquirir conhecimento sobre o perfil dos clientes e elaborar previsões sobre a propensão dos clientes a um determinado produto. Neste seguimento, descrever-se-á as linhas de orientação fundamentais nesta abordagem.

Na fase de formulação do problema, debruçou-se sobre o tema da propensão consultando bibliografia sobre o comportamento do consumidor e pesquisando *case studies* centrados na resolução de problemas análogos e nos métodos estatísticos utilizados. Tendo como objectivo aprofundar mais os conceitos sobre *data mining*, em particular sobre as redes neuronais, realizou-se uma revisão dos conceitos e levantamento de procedimentos/métodos inerentes à construção das redes.

Paralelamente, com o objectivo de dar robustez ao trabalho foram solicitados dados reais a uma instituição de crédito, que reflectissem o comportamento de uma carteira de clientes em circunstâncias análogas às do problema formulado. A instituição em causa facultou os dados sob um compromisso de sigilo sobre informação que pudesse, de alguma forma, revelar o seu negócio ou a identidade e/ou características dos seus clientes.

Contextualizado o problema, a base de dados fornecida por esta instituição consiste numa campanha ocorrida em 2007 e que contém as respostas de cada cliente alvo de mailing, nomeadamente, se aderiu ou não ao produto em causa, o Crédito Pessoal. Relativamente às variáveis explicativas, definiu-se previamente um conjunto de características pessoais e outro conjunto de variáveis que reflectissem o seu envolvimento com o banco num espaço temporal anterior à campanha. No segundo conjunto, as variáveis foram definidas de acordo com os dados que a instituição possuía e como objectivo patente a obtenção de toda a informação que pudesse denunciar o comportamento do cliente propenso a este produto. As variáveis seleccionadas assim como o período de referência da amostra encontram-se descritos e explicados na secção 5.1.

O facto de se ter tido acesso a dados de clientes de instituições bancárias, sobre as quais existe uma grande exigência de actualização e fiabilização deste tipo de dados, constituiu uma vantagem ao nível da consistência e validade dos mesmos. Porém, a exploração dos dados, sobretudo por se tratarem de dados reais, implicou a aplicação de algumas técnicas de tratamento e também de transformação dos dados.

Como é normal neste género de amostras que consistem em sucessos/insucessos de campanhas, a base de dados revela uma dimensão muito reduzida de indivíduos que responderam afirmativamente à campanha comparativamente aos que não responderam, o que implicou a introdução de uma técnica que melhora a representatividade da amostra conferindo maior equilíbrio entre as duas classes – o *oversampling*.

As redes neuronais, designadamente, as multicamada foram escolhidas como a técnica de modelação principal, estando esta escolha fundamentada na bibliografia consultada desde o início deste trabalho. Dois motivos conduziram a que esta técnica tenha surgido como a mais interessante: o facto de ser adaptada do funcionamento do cérebro humano com as capacidades de aprendizagem e generalização e, por outro lado, a sua grande aplicabilidade. Em contrapartida, a regressão logística, o segundo técnica de modelação escolhida, surgiu como uma técnica concorrente às redes neuronais por ser mais conhecida e tradicional, sendo muito citada nos artigos consultados e que serão referenciados adiante com o devido enquadramento.

Por fim, a validação dos modelos obtidos assim como a comparação dos seus resultados basearam-se em testes e medidas de qualidade que pudessem ser aplicados a ambas as técnicas utilizadas – a regressão logística e a aplicação das redes neuronais. Alguns exemplos da avaliação da performance dos modelos e determinantes para a escolha do melhor modelo são o menor RMSE, o *lift*, a matriz custo/benefício e a curva ROC.

Como a comparação é um dos principais objectivos deste trabalho, decidiu-se não abordar mais testes específicas do *logit* que não teriam aplicação nas redes.

Na fase de tratamento dos dados, nomeadamente na aplicação de *oversampling* assim como todas as estimações e cálculos inerentes à estimação dos modelos e à validação foram realizados utilizando, essencialmente, o *software* considerado mais adequado para essas tarefas – o SAS V8 da SAS System.

1.6 Organização

Os capítulos que se seguem descrevem as várias temáticas que foram alvo de estudo e desenvolvimento para a conclusão deste trabalho.

Conforme já foi referido, um dos grandes objectivos deste trabalho consiste em desenvolver um modelo baseado em redes neuronais artificiais. Assim, o capítulo 2 aborda os conceitos essenciais associados a este tema, iniciando-se com o fundamental enquadramento histórico. Em seguida, apresentam-se os conceitos numa lógica de complexidade crescente, ou seja, introduz-se o conceito de neurónio e, a partir desse ponto, desenvolve-se a teoria inerente às redes neuronais artificiais multicamada. Ainda neste capítulo, são descritos a metodologia de construção de uma rede neuronal artificial e o funcionamento das mesmas, nomeadamente, como se processa a aprendizagem. No contexto da aprendizagem, apresenta-se o algoritmo de aprendizagem mais conhecido – o de Retropropagação – e algumas das suas variantes, e aborda-se um dos problemas mais comuns na utilização de redes neuronais, a *sobreaprendizagem*. Por fim, são enumeradas algumas das vantagens associadas à utilização das redes neuronais artificiais.

O capítulo 3 descreve os modelos baseados na regressão logística, introduzindo os seus pressupostos e explicando como esta técnica se adequa ao problema em estudo, esclarecendo que o grande objectivo da sua utilização neste trabalho se prende com a necessidade de comparação com as redes neuronais. Após esta introdução, a especificação destes modelos e a forma de estimação dos seus parâmetros são descritos assim como as suas propriedades. Para finalizar, são enunciados alguns testes, disponíveis na diversa literatura consultada, que podem ser utilizados para avaliar a performance do modelo e que terão tradução prática no trabalho de modelação realizado sobre a amostra em estudo.

No capítulo 4 são desenvolvidos os vários passos de preparação dos dados para que possam ser modelados, nomeadamente, a introdução de pontos essenciais sobre o tema da amostragem e transformação dos dados. A secção referente à amostragem inclui questões relacionadas com a própria selecção dos dados, o *oversampling* e a divisão dos dados nos conjuntos de treino, validação, teste e *score*. Este capítulo faz, ainda, referência às técnicas de modelação utilizadas. Por fim, são abordadas várias medidas de avaliação da qualidade da modelação, as quais se tornarão fundamentais para a análise dos modelos obtidos através das redes neuronais e da regressão logística permitindo a comparação das suas performances.

Contextualizados os suportes teóricos para este problema, segue-se o capítulo 5, que se refere à análise dos resultados extraídos do processo de modelação com recurso às técnicas descritas nos capítulos anteriores. Num primeiro momento, é feita uma referência à constituição da amostra e à forma como os dados foram preparados para poderem ser submetidos aos processos de modelação. Posteriormente, são apresentados os resultados derivados de cada técnica e, no caso particular das redes neuronais, de cada algoritmo de aprendizagem, complementando-os com conclusões elucidativas sobre as suas performances.

Por fim, no capítulo 6, são apresentadas as principais conclusões do trabalho desenvolvido tendo em consideração os objectivos propostos, apontando-se para um cenário sobre o trabalho que, no futuro, poderá ser realizado em continuação e em complementaridade ao presente projecto.

2 Redes Neuronais

Neste capítulo serão apresentados os principais conceitos associados às redes neuronais artificiais e os passos essenciais para a sua compreensão e utilização.

Na sua essência, as redes neuronais artificiais "são modelos simplificados do sistema nervoso central do ser humano", (Cortez e Neves, 2000), nos quais é possível replicar alguns dos componentes deste sistema e o seu funcionamento, como se verá no decurso deste capítulo.

Com base neste pressuposto, e em primeira instância, é importante referir que dada a abrangência deste tema, existem várias definições possíveis para as redes neuronais artificiais, e que, simultaneamente, não existe uma definição que seja universalmente aceite. No entanto, para os propósitos deste estudo, pode-se utilizar uma definição que sendo bastante genérica não deixa de ser útil e explicativa. Assim, uma rede neuronal artificial pode ser descrita como uma rede composta por vários (em certos casos muitos) processadores (também designados neurónios ou unidades) simples, sendo que cada um destes processadores possui uma pequena quantidade de memória local. Estas unidades encontram-se ligadas (conectadas) entre si e partilham informação numérica, e cada uma delas processa apenas os dados que recebe por via destas mesmas ligações.

Um dos aspectos centrais das redes neuronais artificiais está relacionado com a "regra de aprendizagem", pela qual a rede tem a capacidade de "ajustar" as ponderações associadas às ligações entre os neurónios, com base nos dados que lhe são disponibilizados e que esta usa no processamento.

É por via desta regra de aprendizagem que a rede, com base nos dados disponíveis, "aprende", sendo que o conceito de aprendizagem está intimamente relacionado com o conceito de "generalização", ou seja, a capacidade de produzir resultados correctos em previsões feitas com base em dados que não fazem parte do conjunto onde a aprendizagem foi desenvolvida.

Reflectindo sobre esta explicação, é possível estabelecer um paralelo entre a forma "indutiva" como uma rede "aprende" e o processo de aprendizagem de uma criança, cujo cérebro à nascença já comporta "uma estrutura fortemente conexionista com capacidade de aprender através da *experiência*", (Cortez e Neves, 2000).

Haykin (1994) defende que a *aprendizagem* consiste no seguinte processo: a rede é estimulada por um determinado ambiente, que provoca o ajustamento dos pesos das ligações entre nós e que, na sequência desta alteração, será obtida uma resposta.

Perante problemas de classificação e regressão, Bishop (1997) afirma que as redes neuronais disponibilizam uma plataforma potente para representar soluções não lineares a partir de várias variáveis de *input*, convertendo-as noutras de *output*. A forma de conseguir essas soluções baseia-se no ajustamento de determinados parâmetros, sendo a *aprendizagem* o processo de optimização desses parâmetros com base nos dados de *input*.

Um item importante a considerar no processo de aprendizagem é optar pelo conjunto de regras base ou algoritmo (de aprendizagem ou treino) mais apropriado consoante o problema em estudo. Existem dois tipos distintos de aprendizagem: a supervisionada e a não supervisionada.

Na aprendizagem supervisionada, são fornecidos à rede tanto os dados de *input* como os valores esperados de *output*, aos quais se espera que os valores de *output* da rede adiram por completo ou quase.

Conforme o citado em Cortez e Neves (2000), " a rede aprende a partir de um conjunto de padrões (P), onde cada *padrão* (*p*), também chamado de *exemplo* ou *caso de treino*, é composto por um vector de entrada (x^p) e por um vector de resposta ou saída (s^p). Durante o processo de aprendizagem é efectuada uma comparação entre o valor desejado (t^p) com o valor de saída da rede, originando um *erro* ($e^p = t^p \Theta y^p$, sendo Θ a função de erro."

O erro, ou seja, a diferença entre o valor de *output* obtido e o esperado será utilizado através de um algoritmo de aprendizagem para efectuar correcções aos pesos. Estes ajustes consistem, à partida, numa melhoria progressiva da performance da rede, traduzindo-se assim na minimização do erro.

Os algoritmos de aprendizagem supervisionada mais utilizados são diferentes versões do conhecido como erro de retropropagação, que será abordado mais adiante.

A grande diferença entre a aprendizagem não supervisionada e a anterior é a não indicação à rede de qualquer informação sobre a resposta correcta ou o valor esperado de *output*. Apenas são dados à rede os dados de *input*. Os algoritmos deste tipo de aprendizagem descobrem as relações entre as características dos dados de entrada ou encontram padrões de comportamento, agrupando os vectores de entrada similares nos mesmos grupos. Por outras

palavras, a rede altera os pesos sinápticos por forma a que os vectores de *input* mais semelhantes sejam associados ao mesmo *output*. Os exemplos mais conhecidos deste tipo de aprendizagem são os mapas *Kohonen* que são bastante utilizados, sobretudo, em tarefas de *clustering* mas estes não serão objecto neste estudo.

Usualmente, a aplicação das redes neuronais artificiais incide sobre problemas de classificação ou previsão, sendo "frequentemente utilizadas para análise estatística e modelação de dados, e o seu papel entendido como uma alternativa à regressão não-linear *standard* ou às técnicas de análise de *clusters*" (Cheng & Titterington, 1994, citado por Gurney, 2001, p. 5). Contudo, a sua aplicabilidade estende-se a diversos sectores, desde as áreas de negócio às áreas de pesquisa e desenvolvimento científico.

De acordo com os exemplos de aplicação das redes neuronais dados por Fausett (1994), o ruído existente nas linhas telefónicas, sobretudo em chamadas de longa distância, foi reduzido pela aplicação das redes neuronais. O reconhecimento automático de caracteres que são escritos manualmente, também é possível devido à sua utilização. Na medicina, efectuou-se o treino de uma rede, para que através da apresentação de alguns sintomas e características do paciente, a rede devolvesse informação sobre o diagnóstico e o medicamento que melhor poderia tratar o paciente. Como último exemplo, no sector financeiro, as redes neuronais são um instrumento utilizado na decisão de atribuição, ou não, de um empréstimo a um cliente. Este apoio à decisão é feito com base na experiência passada, ou seja, são comparadas as características do cliente em causa com a de outros clientes que, nalgum momento, incumpriram no pagamento do crédito. Quanto mais se assemelharem as características do cliente em causa, às características dos clientes "fraudulentos", maior será a probabilidade de a instituição renunciar a concessão do empréstimo.

Com esta breve abordagem fica aqui patente a enorme abrangência, aplicabilidade e utilidade deste tema.

Quanto à organização do capítulo, depois de um curto resumo sobre a história das redes neuronais, onde serão referenciados os principais marcos e investigadores que contribuíram para o desenvolvimento desta área, serão expostos os principais conceitos associados às redes neuronais e à sua constituição. Neste ponto, e com base nos conceitos referidos, explicar-se-ão as formas como podem ser construídas as redes neuronais artificiais e as potencialidades que resultam de uma escolha correcta da sua arquitectura, fazendo-se, sempre que considerado útil, a analogia com o paradigma humano.

Nos últimos pontos, expõe-se o factor primordial para a existência das redes neuronais artificiais, que consiste na sua capacidade de aprendizagem. Explicar-se-á a forma como se desenvolve e funciona este processo de aprendizagem artificial, introduzindo os métodos de aprendizagem mais comuns e algumas técnicas comummente utilizadas e que permitem melhorar o processo e evitar alguns dos obstáculos com que se pode deparar.

Por fim, e após estarem devidamente descritos os passos essenciais para criar e manusear uma rede neuronal artificial, serão apontadas algumas das potencialidades das redes neuronais artificiais.

2.1 Breve Introdução Histórica

O conhecimento do cérebro humano é um assunto que desde há milhares de anos ocupa a ciência. Este esforço e parte do seu sucesso em muito se fica dever ao trabalho pioneiro de Ramón e Cajál em 1911, que introduziram o conceito de neurónios como sendo parte estrutural do cérebro humano.

Com base no conceito de neurónio humano, o desenvolvimento das redes neuronais artificiais iniciou-se há cerca de 60 anos com a motivação de, simultaneamente, entender o cérebro humano e aproveitar algumas das suas potencialidades.

O primeiro grande passo em direcção às redes neuronais artificiais foi dado em 1943, quando McCulloch, um neurofisiologista, e um jovem matemático, Walter Pitts, elaboraram um artigo que descrevia o funcionamento dos neurónios. Nesse mesmo artigo, McCulloch e Pitts, modelaram uma rede neuronal utilizando circuitos eléctricos.

Em 1949, Donal Hebb, publicou o livro "The Organization of Behavior" do qual se destaca a referência ao facto de as sinapses se fortalecerem cada vez que são usadas, conceito este que se viria a tornar fundamental para entender a aprendizagem do cérebro humano – regra de Hebb.

Frank Rosenblatt, em 1958, introduziu o perceptrão, demonstrando a sua utilidade na classificação de um conjunto contínuo de *inputs* numa de duas classes linearmente separáveis. O perceptrão é a rede neuronal mais antiga que ainda é utilizada actualmente.

Com os avanços tecnológicos alcançados durante os anos 50, foi finalmente possível simular redes neuronais, tendo sido feitos alguns avanços nesse campo. Em 1959 e 1962, Widrow e Hoff, da Universidade de Stanford, desenvolveram alguns modelos que se tornaram nas

primeiras aplicações das redes neuronais artificiais a problemas reais. Neste período, mais concretamente em 1960, estes autores desenvolveram uma regra de aprendizagem, ainda hoje, bastante popular que é denominada de regra Delta ou Widrow-Hoff . Esta regra, criada para "uma rede neuronal com apenas uma camada foi a percursora da regra do algoritmo de Retropropagação para as redes multicamada" (Fausett, 1994).

Em 1969, Minsky e Papert, apontaram as limitações teóricas dos modelos de redes neuronais com uma camada, no seu livro "Perceptrons". Devido a esta visão pessimista e à falta de um método de aprendizagem para a rede multicamada, a investigação sobre as redes neuronais ficou praticamente eclipsada durante uma década, pelo que apenas em 1982 com o trabalho do físico John Hopfield, onde se desenvolveu as redes neuronais baseadas em ponderadores fixos e activações adaptativas e que são utilizadas como redes de memória associativa, se voltou a dar a devida atenção a esta área.

Rumelhart, Hinton e Williams, em 1986, baseando-se e melhorando a descoberta anteriormente realizada por Paul Werbos (1974), e, posteriormente, por David Parker (1985) e LeCun (1986) – um método que propagava os erros da camada de saída para as camadas intermédias precedentes – publicaram a solução poderosa para o treino de uma rede de várias camadas que ficou conhecida como o algoritmo de aprendizagem de Retropropagação (*"Back-Propagation")*. O sucesso desta abordagem ficou demonstrado pelo sistema "NETtalk" desenvolvido por Sejnowski e Rosenberg, em 1987 que convertia um texto em Inglês para voz, de uma forma bastante perceptível.

2.2 Estrutura da Rede Neuronal

De modo geral, uma rede neuronal é um modelo concebido para realizar algumas tarefas que o cérebro humano também é capaz de realizar, processando a informação de forma análoga.

O sistema neuronal humano está continuamente a receber informação que, após ser compreendida, possibilita a tomada de decisões.

"A rede neuronal é um processador distribuído paralela e massivamente que tem uma propensão natural para armazenar conhecimento experimental e torná-lo disponível para ser utilizado. Esta definição sugere a semelhança ao cérebro humano em dois aspectos: o conhecimento é adquirido pela rede através do processo de aprendizagem e as forças das

ligações entre os neurónios, conhecidas por pesos sinápticos, são utilizadas para o armazenamento do conhecimento" (Haykin, 1994).

2.2.1 O neurónio artificial

O neurónio artificial, ou também designado por nó, é a unidade de processamento fulcral para o processamento da rede neuronal. O modelo do nó tem três componentes principais:

- *Um conjunto de sinapses*, em que a cada uma está associada um peso w_{ij} (o i refere-se ao nó em questão e o j ao nó de onde partiu o estímulo ou sinal (x_j). Este peso exerce um determinado efeito sobre as sinapses: excitatório para valores positivos e inibitório para valores negativos.
- *Um integrador ou função de combinação (g)*, que soma os n estímulos de entrada ponderados pelos pesos associados às sinapses, ou seja, uma combinação linear.
- *Uma função de activação (f)*, que limita a amplitude do sinal de saída a um número mais restrito de valores.

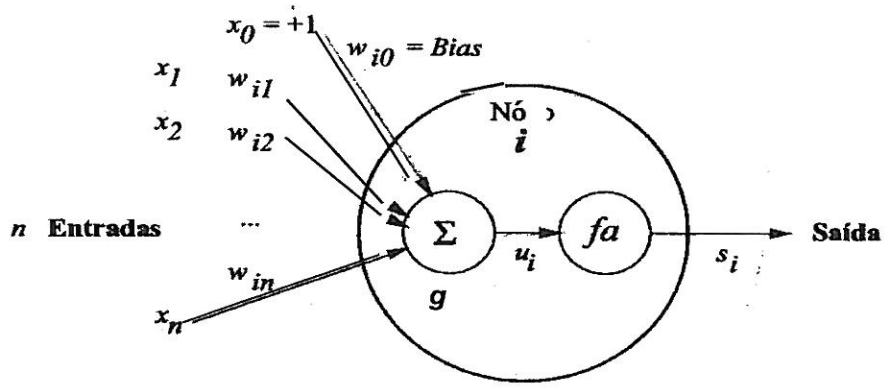

Figura 1 – Estrutura geral de um nó, adaptado de Cortez, P. e Neves, J. (2000)

Como mostra a figura 1, poderá existir uma sinapse extra, *bias*, cujo estímulo de entrada será x_0 que normalmente toma o valor +1 e multiplicado pelo peso w_{i0}. Este incremento irá influenciar o processamento computacional no sentido de equilibrar os resultados finais, podendo ser visto como mais um parâmetro de ajustamento e criando, assim, maior flexibilidade no ajustamento da função de interesse.

Os estímulos $x_1, x_2, ..., x_n$ são colocados como entrada das sinapses e estes são multiplicados pelo respectivo peso $w_{i1}, w_{i2}, ..., w_{in}$, sendo u_i a combinação linear de saída:

$$u_i = g(1 \times w_{i0}, x_1 \times w_{i1}, x_2 \times w_{i2}, ..., x_n \times w_{in}) = \sum_{j=1}^{n} w_{ij}\, x_j \,.$$

De um modo geral, para o nó i com uma determinada função de activação f contínua, diferenciável e monótona não-decrescente, a sua saída representar-se-á como $s_i = f(u_i)$.

Cada neurónio tem um nível de actividade que corresponde a uma função dos *inputs* que recebe. Por sua vez, cada neurónio transmitirá o resultado dessa função para os demais como se de um sinal se tratasse. No entanto, cada neurónio tem capacidade para transmitir apenas um sinal de cada vez, apesar de o poder fazer para vários neurónios simultaneamente.

McCulloch e Pitts apresentaram em 1943 um desenho muito simplificado da maioria das redes neuronais, onde definem a variável denominada *threshold*, θ_i , valor ao qual será comparado o resultado da combinação linear u_i. Nesta representação, estes autores utilizaram como função de activação:

$$s_i = \begin{cases} +1 & se\ u_i \geq \theta_i \\ -1 & se\ u_i < \theta_i \end{cases},$$

conhecida como *função sinal*. Esta função assim como a função limiar são utilizadas sobretudo para tomadas de decisão dos nós para classificação e no reconhecimento de padrões (Negnevitsky, 2002). Contudo, a função de activação pode tomar diferentes formas de acordo com os objectivos do analista. A figura 2 apresenta apenas algumas das opções mais comuns, como exemplo, a função linear que fornece um *output* igual ao *input* e é usada quando se pretende especificamente uma aproximação linear.

Nome	Função f	Contradomínio
limiar	$\begin{cases} 1, se\ u_i \geq 0 \\ 0, se\ u_i < 0 \end{cases}$	{0,1}
linear	u_i	]-∞,+∞[
por troços	$\begin{cases} 1, se\ u_i \geq 0.5 \\ ku_i, se -0.5 < u_i < 0.5 \\ 0, se\ u_i \leq -0.5 \end{cases}$	[0,1]
logística	$\frac{1}{1+\exp(-ku_i)}$	[0,1]
tangente hiperbólica	$\tanh(ku_i)$	[-1,1]
sin	$\sin(u_i \bmod 2\Pi)$	[-1,1]
cos	$\cos(u_i \bmod 2\Pi)$	[-1,1]
gaussiana	$\exp\left(\frac{-u_i 2}{2k^2}\right)$	[-1,1]
quadrada	$-sign(u_i)u_i^2$	]-∞,+∞[

Tabela 1 – Funções de activação típicas, adaptado de Cortez, P. e Neves, J. (2000)

Como refere Haykin (1994), a função sigmoidal é a função de activação mais utilizada na construção das redes neuronais artificiais, definida como estritamente crescente, suave e com propriedades assintóticas. Um exemplo da sigmoidal é a *função logística*:

$$f(u_i) = \frac{1}{1+\exp(-av)},$$

onde a representa o declive. Normalmente utilizada quando os valores de *output* esperados são binários ou estão contidos no intervalo [0;1].

A caracterização de uma rede neuronal consiste nos padrões de conexão entre os neurónios (arquitectura), a forma de determinar os pesos sinápticos (algoritmo de aprendizagem) e a função de activação.

2.2.2 O Perceptrão Multicamada

Quando se faz uma referência à arquitectura de redes neuronais, está-se a falar da forma como os nós se interligam na estrutura da rede. As redes multicamada encontram-se organizadas por camadas: uma camada de entrada, uma camada de saída e por uma ou mais camadas intermédias. Note-se que a camada de entrada não é referida como uma camada da rede pelo facto de nesta não serem efectuados quaisquer cálculos (Cortez e Neves, 2000).

"Os nós intermédios (das camadas intermédias) são simples unidades de processamento que combinam múltiplos dados de *input* resultando num único *output*. O Perceptrão Multicamada é uma extensão do perceptrão simples introduzido por Rosenblatt em 1958, que consiste numa única camada de entrada e outra de saída, sem camadas intermédias.

A introdução de unidades intermédias resolveu as restrições do perceptrão simples, dado que este não tem a capacidade de resolução de determinados problemas de classificação, nomeadamente, dos problemas não separáveis linearmente" (Kjosgen e Zytkow, 2002).

De um modo geral, a principal função dos nós das camadas intermédias (circunscrita na figura 2 por primeiro quadrado a tracejado, da esquerda para a direita) consiste em descobrir características que se encontram escondidas nos dados de *input* e que são representadas pelos pesos desses nós. Esta informação será utilizada pela camada de *output*, de forma determinante, para a obtenção de um padrão final.

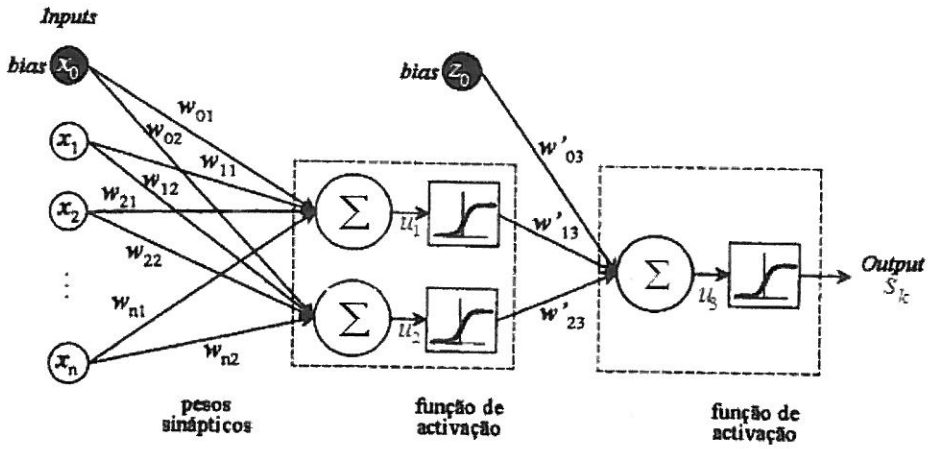

Figura 2 – Estrutura da rede multicamada com uma camada intermédia, adaptado de Chorão, L.A. (2005)

Quanto maior for o número de camadas intermédias, maior será a capacidade da rede para resolver problemas que possuem uma elevada complexidade. Está, no entanto, provado que duas camadas escondidas são suficientes para aproximar qualquer função, independentemente da sua complexidade (Bishop 1997).

Relativamente ao número de unidades de cada camada intermédia, os autores Freeman e Skapura (1992) aconselham a que se adicionem unidades caso a rede falhe na convergência para uma solução mas que, caso contrário, se tente reduzir o seu número optimizando a performance da rede.

O aumento de camadas intermédias implica necessariamente um aumento exponencial do tempo de aprendizagem da rede e introduz o risco de *sobreaprendizagem*. Este problema consiste na tendência que a rede tem em assumir que toda a realidade é exactamente igual aos dados de treino, verificando-se que quanto mais ajustada aos dados de treino estiver a rede, menor é a sua capacidade de generalização. Assim, quando existe um elevado número de parâmetros disponíveis (muitas unidades e camadas) a rede rapidamente "memoriza" os dados de treino perdendo, assim, capacidade de generalização.

O perceptrão multicamada apresenta algumas características distintivas, como a não linearidade dos *outputs* de todos os neurónios da rede (não se perdendo, no entanto, a diferenciabilidade), a existência de camadas intermédias, que permitem que a rede aprenda tarefas complexas, e por fim, o elevado grau de conexão assente nas sinapses da rede.

2.2.3 Metodologia de Construção da Rede

A construção de uma rede neuronal é um processo algo complexo em que se torna imprescindível uma boa definição dos objectivos da rede e o conhecimento razoável da natureza do problema ao qual se irá aplicar a rede.

Segundo Swingler (2001), o projecto do ciclo de vida de uma rede neuronal consiste na sequência de passos:

1. Definição da tarefa a desempenhar e desenho do plano: traçar todos os requisitos para o sistema final.
2. Viabilidade: verificar se o problema se adequa a uma solução de redes neuronais.
3. Definição da amostra: definir quais os dados a apresentar à rede.
4. Desenho da rede: garantindo a capacidade de generalização e a precisão do modelo.
5. Recolha de dados: medindo a quantidade de dados seleccionados.
6. Verificação dos dados: determinar quais os dados seleccionados e se permitirão a resolução do problema.
7. Treino e teste: construir a melhor rede a partir da amostra de dados.
8. Análise do erro: determinar quais os tipos e origem dos erros
9. Análise da rede: A partir da rede final obter conclusões e regras.
10. Implementação do sistema: a utilização e a monitorização da rede final.

Saliente-se que este plano não refere que deve existir um método de recolha da amostra, o qual é exigência para qualquer trabalho, ou o acesso aos dados, temas que serão abordados no capítulo 4.

2.2.4 A Regra Delta

A eficácia de uma rede está associada à sua capacidade de aprendizagem, sendo que o seu sucesso depende essencialmente de dois aspectos: a arquitectura da rede e o método de aprendizagem.

A escolha do método de aprendizagem é totalmente influenciada pela tarefa/objectivo da questão em estudo. Mais concretamente e no que respeita ao tema do presente trabalho, a regressão e a previsão são tarefas para uma aprendizagem supervisionada.

Como referido anteriormente, o objectivo da aprendizagem supervisionada é minimizar o erro que reside entre o valor esperado e o valor de saída da rede e, por conseguinte, como regra de

aprendizagem muito utilizada nas redes multicamada apresenta-se o gradiente descendente, também conhecido por *regra delta*. A aplicação desta regra justifica-se também pelo seu sucesso na convergência para as melhores soluções em problemas não linearmente separáveis. De acordo com Mitchell (1997), "a ideia chave por detràs da regra delta é usar o gradiente descendente para encontrar, no espaço com todas as hipóteses possíveis de vectores de pesos, os pesos que melhor se ajustam aos exemplos de treino".

De uma forma concisa, a aprendizagem de uma rede inicia com determinados valores aleatórios nos pesos sinápticos e, de iteração para iteração, estes valores são alterados tantas vezes quanto necessário para a obtenção de um valor aceitável. Para cada peso w_i associado ao vector de entrada x_i aplica-se a seguinte fórmula:

$$w_i \leftarrow w_i + \Delta w_i .$$

Definindo η como a *taxa de aprendizagem* e $\nabla\xi$ o gradiente da função de erro, esta regra sugere qual o ajustamento a efectuar aos pesos sinápticos de forma a reduzir a função de erro, traduzindo-se na resolução da equação:

$$\Delta w = \eta \nabla \xi \ .$$

A taxa de aprendizagem é utilizada para moderar a velocidade de alteração dos pesos sinápticos em cada passo e o gradiente de uma função, que neste caso trata-se da função do erro em que as suas variáveis são os pesos, indica a direcção na qual a função poderá crescer ou decrescer mais rapidamente conforme o seu sinal seja, respectivamente, positivo ou negativo.

No contexto das redes multicamada, apesar da existência de um considerável número de algoritmos de aprendizagem, a regra *delta* é muito importante pois o gradiente descendente serve de base a um dos mais populares algoritmos: o algoritmo de *Retropropagação*. Este algoritmo permite a aprendizagem através de redes com um elevado número de ligações sinápticas.

2.2.5 O Algoritmo de Retropropagação

De acordo com Cortez e Neves em 2000, o algoritmo de Retropropagação "prevalece como um marco para a comunidade das redes neuronais artificiais, procurando o mínimo da função de erro no espaço de procura dos pesos, baseando-se em métodos de gradiente descendente. A

combinação de pesos que minimiza a função de erro é considerada a solução para o problema de aprendizagem".

O sucesso deste algoritmo deve-se ao facto de obter uma boa *generalização*, o que significa que consegue extrair, através da aprendizagem, as semelhanças mais significativas entre os diferentes vectores de *input* e, simultaneamente, ignorar os dados irrelevantes.

O primeiro passo neste algoritmo é a inicialização da rede que consiste na escolha dos valores dos pesos sinápticos. Uma boa escolha dos valores iniciais destes parâmetros poderá contribuir para o sucesso do processo de aprendizagem. Em contrapartida, "uma má escolha poderá conduzir a um fenómeno conhecido por "saturação prematura" (Lee & tal., 1991 citado por Haykin em 1994, p. 156). Este fenómeno ocorre quando o erro se mantém quase constante durante um determinado período de tempo do processo de aprendizagem, e após esse tempo continua a decrescer, pelo que esta solução não pode ser considerada como um mínimo local". Cortez e Neves (2000) aconselham a evitar valores muito pequenos ou muito elevados.

Para prosseguir com a explanação deste algoritmo, torna-se essencial definir algumas notações adicionais que serão usadas nesta secção:

x_j - estímulo apresentado ao nó j

s_j - valor de *output* gerado pelo nó j

t_j - valor esperado de *output* gerado pelo nó j

E - número de nós de entrada

$succ(i)$ – conjunto dos nós j da camada com que o nó i estabelece ligações.

Importa, ainda, referir que a demonstração que se segue e que explica os passos do algoritmo, assim como o desenvolvimento de equações se baseiam, essencialmente, em Mitchelle (1997) e Cortez e Neves (2000).

O algoritmo baseia-se em duas fases, a etapa inicial, designada por passo *em frente*, traduz-se na apresentação aos nós de entrada os estímulos (x_i) propagando-se de camada em camada e sendo convertidos num único valor através do integrador:

$$u_i = \sum_j x_j w_{ij},$$

em que $x_0 = 1$. O valor deste somatório será transformado pela função de activação f gerando assim o valor de saída, s_j. Seguidamente, é calculado o erro com base na função de custo:

$$\xi = \frac{1}{2}\sum_j (t_j - s_j)^2 .$$

O segundo passo consiste na chamada *retropropagação do erro* desde os nós de saída até aos de entrada, ajustando os pesos sinápticos iterativamente. Especificando para uma dada iteração t e uma taxa de aprendizagem η, o ajustamento aplicado ao peso sináptico w_{ij} (que estabelece a ligação entre o nó j ao nó i) através da regra delta que se define como:

$$\Delta w_{ij}(t) = -\eta * \frac{\partial \xi}{\partial w_{ij}}. \qquad \textbf{(Equação 1)}$$

A derivada do erro em ordem ao peso sináptico traduz-se na seguinte forma:

$$\frac{\partial \xi}{\partial w_{ij}} = \frac{\partial \xi}{\partial u_i}\frac{\partial u_i}{\partial wij}$$
$$= \frac{\partial \xi}{\partial u_i} x_i =$$
$$= \frac{\partial \xi}{\partial s_i}\frac{\partial s_i}{\partial u_i} x_i .$$

Para o ajustamento dos pesos, torna-se necessário resolver esta derivada tendo em consideração os dois tipos de pesos existentes nas redes multicamada: os associados aos nós de saída e os associados aos nós pertencentes às camadas intermédias.

No primeiro caso, o primeiro termo expressa-se através da equação:

$$\frac{\partial \xi}{\partial s_i} = \frac{\partial}{\partial s_i}\frac{1}{2}\sum_{j \in outputs}(t_j - s_j)^2 ,$$

onde as derivadas $\frac{\partial}{\partial s_i}(t_j - s_j)^2$ serão igualadas a zero excepto quando $j=i$:

$$\frac{\partial \xi}{\partial s_i} = \frac{\partial}{\partial s_i}\frac{1}{2}(t_i - s_i)^2$$
$$= \frac{1}{2}2(t_i - s_i)\frac{\partial (t_i - s_i)}{\partial s_i}$$
$$= -(t_i - s_i) \qquad \textbf{(Equação 2)}$$

Relativamente ao segundo termo, requer recordar que a função de activação f é a função sigmoidal e a sua derivada de primeira ordem define-se como $f(x)(1 - f(x))$. Aplicando esta última fórmula, o segundo termo traduz-se na seguinte igualdade:

$$\frac{\partial s_i}{\partial u_i} = s_i (1 - s_i) \qquad \textbf{(Equação 3)}$$

Após substituição das equações 2 e 3 na igualdade 1, obtém-se como resultado final a regra de variação dos pesos para as unidades de saída:

$$\Delta w_{ij}(t) = -\eta * \frac{\partial \xi}{\partial w_{ij}} = \eta (t_i - s_i) s_i (1 - s_i) x_i .$$

Para o caso em que o nó *i* pertence às camadas intermédias, a derivação deverá reflectir as diferentes formas em que w_{ij} poderá influenciar o *output* da rede e, por conseguinte, o erro global ξ. Para tal, justifica-se a referência ao conjunto $succ(i)$, composto por todos os nós cujos *inputs* incluem o valor de *output* do nó *i*:

$$\begin{aligned}
\frac{\partial \xi}{\partial u_i} &= \sum_{j \in succ(i)} \frac{\partial \xi}{\partial u_j} \frac{\partial u_j}{\partial u_i} \\
&= \sum_{j \in succ(i)} \frac{\partial \xi}{\partial u_j} \frac{\partial u_j}{\partial s_i} \frac{\partial s_i}{\partial u_i} \\
&= \sum_{j \in succ(i)} \frac{\partial \xi}{\partial u_j} w_{ji} s_i (1 - s_i) \\
&= s_i (1 - s_i) \sum_{j \in succ(i)} \frac{\partial \xi}{\partial u_j} w_{ji}
\end{aligned}$$

De forma a simplificar a expressão, denote-se a derivada parcial $\frac{\partial \xi}{\partial u_j}$ como α_j e, obtém-se:

$$\alpha_i = s_i (1 - s_i) \sum_{j \in succ(i)} \alpha_j w_{ji} .$$

Por fim, consegue-se exprimir a variação dos pesos para as camadas intermédias:

$$\Delta w_{ij} = \eta \alpha_i x_i .$$

Após o cálculo da variação $\Delta w_i(t)$, é efectuada a actualização dos pesos:

$$w_i(t+1) = w_i(t) + \Delta w_i(t) .$$

A taxa de aprendizagem desempenha um papel importante na performance da rede e no ajustamento dos pesos, como se poderá deduzir da equação 1. Quando menor (maior) for o seu valor, menores (maiores) serão as variações nos pesos de uma iteração para a seguinte.
A velocidade da aprendizagem depende do valor de η, sendo mais lenta se o valor de η for pequeno devido à necessidade que a rede tem em executar um maior número de iterações. Se

o valor da taxa de aprendizagem for grande, poderá provocar uma grande instabilidade no treino e a sua não convergência. Freeman e Skapura (1992) consideram que η deverá tomar valores pequenos, mais concretamente, entre 0.05 e 0.25 para garantir a convergência para uma solução.

Para evitar que a rede tenha grandes oscilações na proximidade dos melhores valores (mínimos locais) sem os alcançar, o método Retropropagação incrementa o termo *momentum* na regra delta:

$$\Delta w_{ij}(t) = -\eta * \frac{\delta \xi}{\delta w_{ij}} + \alpha \Delta w_{ij}(t-1)$$

onde α representa a constante *momentum* , normalmente é um número positivo e pertence ao intervalo [0;1]. O papel deste termo é garantir que as alterações dos pesos seguem na mesma direcção.

Desta forma, a variação dos pesos será influenciada pelas alterações efectuadas anteriormente em que a equação anterior pode ser reescrita da seguinte forma:

$$\Delta w_{ij}(t) = -\eta \sum_{k=o}^{t} \alpha^{t-k} \frac{\delta \xi(k)}{\delta w_{ij}(k)}.$$

Note-se que o uso do sinal negativo em todas as equações anteriores está relacionado com o facto de o gradiente ser *descendente* no espaço dos pesos.

Em suma, o algoritmo de Retropropagação consiste no "treino da rede em que os nós pertencentes às camadas intermédias organizam-se de tal forma que nós diferentes "aprendem" a reconhecer características diferentes do espaço total de *inputs*. Após o treino, quando apresentado um padrão de *input* arbitrário que contenha ruído ou esteja incompleto, as unidades das camadas intermédias responderão com um *output* activo se o novo *input* contiver um padrão que se assemelhe com as características que as unidades individuais aprenderam a reconhecer durante o treino. Inversamente, as unidades das camadas intermédias têm a tendência de inibir os seus *outputs* se os padrões de *input* não contiverem as características para cujo reconhecimento foram treinados". (Freeman e Skapura, 1992)

2.2.6 Variantes do Algoritmo de Retropropagação

Neste trabalho, a aplicação das redes neuronais não assentará apenas no algoritmo de Retropropagação básico mas também nalgumas das suas variantes. O objectivo será comparar

a performance dos diferentes métodos e concluir qual o que melhor se adequa a este caso prático. Segundo Bishop (1995), " diferentes algoritmos irão possuir melhor performance em diferentes problemas e, como tal, não é possível recomendar um algoritmo de optimização universal".

Enquanto que o algoritmo de Retropropagação descrito na secção anterior ajusta os pesos com base na direcção descendente, a direcção correspondente ao negativo do gradiente tendo presente que, no algoritmo do gradiente conjugado a procura do mínimo da função do erro realiza-se na direcção do gradiente (positivo). Esta pequena diferença na direcção torna a convergência mais rápida no caso do gradiente conjugado.

Nos algoritmos de gradiente conjugado, a cada iteração é efectuado o ajustamento ao tamanho do passo. Este processo decorre de uma pesquisa realizada na direcção do gradiente conjugado, para encontrar o tamanho que minimiza a performance da função naquela direcção. De acordo com este método e sob a forma concebida por Polak e Ribiere, 1969, citados em Bishop, 1997, p.280, a actualização dos pesos é determinada por:

$$\Delta w_{t+1} = w_t + \alpha d_t,$$

onde

$$d_t = -g_t + \beta_t d_{t-1} \text{ e } \beta_t = \frac{(g_t^T - g_{t-1}^T) g_t}{g_{k-1}^T \cdot g_{k-1}},$$

sendo α a taxa de aprendizagem e d_t a direcção da pesquisa no espaço de todos os pesos possíveis na iteração t, e g_t o gradiente do erro relativamente aos pesos na iteração t. Bishop (1997) refere que este método pode ser visto como uma forma do gradiente descendente em que o *momentum* e a taxa de aprendizagem são determinados automaticamente em cada iteração.

O algoritmo Quasi-Newton baseia-se no conhecido método de Newton e é utilizado como alternativa ao algoritmo do gradiente conjugado para uma mais rápida convergência.

O método de Newton assume que a função do erro poderá ser aproximada por uma função quadrática na região próxima do valor óptimo e usa a 1ª e 2ª derivadas para determinar o ponto (mínimo) cujo valor do gradiente é zero:

$$\Delta w_{t+1} = -H_t^{-1} g_t,$$

onde H é a matriz Hessiana e g_t é o gradiente do erro relativamente aos pesos na iteração t.

A aplicação deste método torna-se complexa devido ao cálculo da matriz Hessiana, mais concretamente, as derivadas de segunda ordem da função erro, ξ, enquanto que no algoritmo Quasi-Newton, a matriz inversa, a realmente utilizada, é aproximada a cada iteração.

"Ao contrário dos algoritmos de Retropropagação baseados no método dos gradientes conjugados, a precisão da localização do mínimo no algoritmo Quasi-Newton não necessita de ser tão eficaz, o que se traduz numa menor exigência computacional" (Bishop, 1995; Reed e Marks, 1999, citados em Ayala Calvo, 2007, p. 2102.

O algoritmo de Levenberg-Marquadt é uma variante do método de Newton, "concebido especificamente para a soma dos erros quadráticos" (Bishop, 1997, p. 290. No entanto, como se baseia no método de Gauss-Newton, não requer a determinação da matriz Hessiana bastando uma aproximação tal como no método Quasi-Newton. Esta aproximação será efectuada através da matriz Jacobiana (Hagan, 1996, citado em Ayala Calvo,2007, p. 2103., sendo obtida a partir das primeiras derivadas parciais dos erros em relação aos vários pesos sinápticos. Esta técnica é considerada mais potente que o gradiente descendente por Andrade, Neto e Rosa (2001), tendo a seguinte regra de actualização dos pesos:

$$\Delta w_{t+1} = \left(J_t^T J_t + \mu_t I\right)^{-1} J_t^T e_t,$$

onde J é a matriz Jacobiana das derivadas de cada erro para cada peso, I é a matriz Identidade, μ é um escalar, e e é o vector dos erros na iteração t. À medida que μ tende para zero, a expressão aproximar-se-á do método de Gauss-Newton com a aproximação à matriz Hessiana, enquanto que para valores elevados de μ tem-se a fórmula do gradiente descendente, conforme referido por Andrade, Neto e Rosa (2001).

O comprimento do passo é dado por μ^{-1}, pelo que neste último cenário, o erro decrescerá uma vez que o comprimento do passo na direcção do gradiente descendente será muito pequeno.

2.2.7 Modo de Treino

Na aplicação do algoritmo de *Retropropagação*, pode-se optar por um dos dois tipos de treino da rede: modo incremental ou modo não-incremental (*batch*).

Esta diferenciação está relacionada com a forma através da qual são calculados os ajustamentos aos pesos, que por sua vez, terão influência na redução do erro médio do conjunto de treino.

No modo *incremental*, o momento de actualização dos pesos ocorre após a apresentação de cada um dos vários casos de treino. Utilizando o modo *não-incremental*, os ajustamentos são realizados após o fim de cada época, ou seja, após terem sido processados todos os casos de treino à rede neuronal. Apenas nesta fase, e após apurado o erro médio, é efectuada a actualização da rede com base nesse valor.

A regra do gradiente descendente, utilizada no algoritmo de retropropagação, apenas actualiza os pesos depois de cada época, ou seja, aplica o método não incremental. Consequentemente, a convergência pode tornar-se mais lenta, não sendo, igualmente, garantido que a solução obtida corresponda a um mínimo global, dada a possibilidade de existência de vários mínimos locais.

2.2.8 O Problema da Sobreaprendizagem

A *sobreaprendizagem* é um problema comum para qualquer generalização ocorre quando a rede neuronal "aprende" todo e qualquer padrão existente nos casos de treino, melhorando a sua performance para estes casos em detrimento da performance no conjunto de teste.

Existem algumas condições que podem contribuir para uma boa generalização, mitigando o efeito da *sobreaprendizagem*. "Em princípio, quanto maior for o conjunto de casos de treino melhor se comporta a rede, dado que assimila mais informação. Mas, por outro lado, um conjunto de casos de validação demasiado pequeno poderá parar o treino numa altura indesejada, ou seja, a rede não será convenientemente testada", Cortez (1997). Torna-se, assim, necessário impor algumas regras para que o treino da rede não termine cedo demais.

Como em qualquer processo iterativo, também na utilização de redes neuronais, nomeadamente, no treino de um conjunto de dados através do algoritmo de *Retropropagação*, podem definir-se regras para determinar o momento em que o processo deve terminar, ou seja, o ponto no qual se assume que os resultados não são passíveis de ser melhorados.

Tendo em conta que na submissão de um conjunto de treino, o que se pretende é a redução do erro médio da rede, os vários critérios que podem ser definidos deverão incidir sobre a evolução desse valor.

Se o treino da rede cessa sem atingir uma solução considerada razoável, a alteração do número de nós pertencentes às camadas intermédias ou dos parâmetros de aprendizagem ou, ainda, a alteração dos valores iniciais dos pesos são algumas das medidas que poderão servir para contornar o problema. Por outro lado, quando a rede obtém uma solução aceitável não há

garantia de que essa solução corresponde a um mínimo global. Na realidade, na maior parte dos casos trata-se de um mínimo local e não global, ambos representados na figura 3, o que não constitui necessariamente um problema desde que a solução encontrada esteja próxima do expectável, na perspectiva do erro.

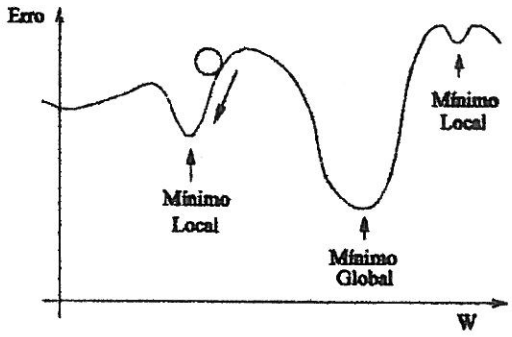

Figura 3 – Mínimos locais e globais

Para o treino de redes neuronais, existem várias funções de erro possíveis mas a mais utilizada traduz-se na equação:

$$E = \sum_{i} (t_i - s_i)^2 ,$$

para t_i o valor esperado do *output* e s_i o valor (real) de *output* do neurónio, medindo assim o quadrado dos erros.

Deve realçar-se que a questão da boa generalização não se prende apenas com o tamanho dos conjuntos de dados mas também com a sua representatividade, assim como o facto de que os dados de *input* deverão conter a informação necessária à determinação do *output*. Estas variantes deverão ser consideradas no momento da recolha de dados e da definição do problema.

2.3 Os Benefícios das Redes Neuronais

As redes neuronais têm uma grande capacidade de aprendizagem e proporcionam muitas vantagens para os analistas/gestores quando conjugadas com o bom conhecimento do seu funcionamento/topologia e do domínio da aplicação. Medsker, Turban e Trippi, citados em Turban e Trippi, 1993, p.11, nomeiam alguns benefícios:

Generalização e aprendizagem. A capacidade para generalizar a partir de exemplos e não requerem base de conhecimento prévio para "aprender"de forma eficiente, pois podem fazê-lo através da experiência. Mesmo perante ruído, correlação entre variáveis ou informação incompleta, estas conseguem processar uma resposta razoável.

Tolerância ao erro. As redes mantêm o seu desempenho ainda que algumas das suas conexões ou nós sejam danificados. Como a informação está distribuída ao longo da rede neuronal, apenas um dano muito forte na sua estrutura pode fazer com que a degradação da sua performance seja muito acentuada.

Adaptabilidade. Intrinsecamente, as redes neuronais têm a capacidade de adaptar os pesos das suas sinapses a alterações no meio envolvente, sem necessidade de alterar a sua estrutura.

Não linearidade. As redes possuem a aptidão para resolver problemas reais de natureza não linear.

Rapidez. O processamento de tarefas complexas de forma mais rápida que os sistemas convencionais.

3 Regressão Logística

A análise da regressão consiste no conhecimento das relações entre as variáveis, da sua relevância na explicação de determinados fenómenos, e, ainda, na previsão de eventos futuros que servirá de apoio à tomada de decisão.

Os modelos de regressão logística têm tido uma grande aplicabilidade nas mais diversas áreas, desde as ciências naturais às ciências económicas ou financeiras e, no seguimento do referido no capítulo anterior, estes modelos são comummente aplicados a problemas de marketing, nomeadamente, os relacionados com propensão ao consumo nos quais o processo de ajustamento considera cada observação da amostra como uma decisão de Bernoulli simples, em que a ocorrência de uma das classes se traduz num *sucesso*.

De um modo geral, a regressão logística assenta no pressuposto de que a variável de interesse (dependente) é uma função probabilística de determinadas variáveis (explicativas e independentes) e de um erro, determinando assim a probabilidade de ocorrência de um evento sob a influência desse conjunto de variáveis. A particularidade do modelo *logit* consiste na natureza da variável de interesse, ou seja, do facto de esta ser binária, por exemplo a resposta do mercado ou de um cliente a uma campanha.

Esta técnica estatística surge neste trabalho como o método tradicional e de comparação com as redes neuronais por forma a avaliar o desempenho e eficácia de uma técnica em que se pressupõe um determinado comportamento dos dados e outra em que se desconhece qualquer dado ou informação sobre os mesmos.

Este capítulo divide-se em duas partes principais: uma primeira em que há uma descrição do modelo de regressão logística apresentando a sua especificação e os seus pressupostos e, uma segunda em que se apresentam alguns testes à performance do modelo que são específicos e, normalmente, utilizados neste tipo de regressão. Contudo, esta não será uma abordagem aprofundada uma vez que o objectivo consiste na comparação entre as duas técnicas de modelação e, portanto, atribuir-se-á mais importância a métodos que permitam a aplicação à regressão e às redes, nomeadamente, as expostas na secção 4.6.

3.1 Especificação do modelo

A modelação de dados por um modelo *logit* justifica-se pela necessidade em explicar uma variável de natureza qualitativa, a variável endógena (objecto de estudo) Y que tem as características de uma variável binária, tomando valor '1' em caso de ocorrência do fenómeno em estudo, e '0', caso contrário.

"Juntamente com outros modelos probabilísticos para variáveis dependentes qualitativas, o modelo *logit* teve origem na análise de experiências biológicas. Se amostras de insectos são expostas a um insecticida com vários níveis de concentração, a proporção morta varia consoante a dosagem aplicada. Para um simples animal esta é uma experiência com uma determinada variável de estímulo contínua ou uma resposta aleatória discreta, sendo "sobrevivente" ou "morto". Este mesmo esquema aplica-se a pacientes com um determinado tratamento que consoante a sua intensidade recuperará ou não da doença, ou ao consumo dos agregados familiares com diferentes níveis de rendimento, que responderão a este incentivo comprando ou não um carro ou um outro bem de longa duração" (Cramer, 2003). Este último exemplo refere-se a uma variável discreta (de opção: aquisição ou não aquisição) no comportamento do consumidor, tal como o caso em estudo: cliente adere, ou não, ao Crédito Pessoal.

Perante a variável dependente Y e k variáveis explicativas, $X_1, X_2, ..., X_k$.

Defina-se a variável latente, não observável, Y^*, como:

$$Y_i = \begin{cases} 1, \text{se } Y_i^* > 0 \\ 0, \text{caso contrário} \end{cases},$$

sendo esta variável expressa da seguinte forma:

$$Y_i^* = X_i \beta_i + u_i,$$

e admitindo que u_i tem uma distribuição logística.

A resolução do problema centra-se na estimação da probabilidade condicionada de ocorrer um evento de sucesso (cliente adere ao produto, i.e., Y=1) como uma função de um conjunto de variáveis explicativas. Então,

$$P(Y=1 \mid X) = P(Y^* > 0 \mid X) = P(u > -(\beta_0 + x\beta \mid x)$$
$$= 1 - G\left[-\beta_0 + x^t\beta\right] = G(\beta_0 + x^t\beta) \quad [1],$$
$$= G(z) = \exp(z)/[1+\exp(z)]$$

onde $z = X^T\beta = \beta_0 + x^t\beta = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k$ e $G(z)$ é a função densidade acumulada de z tomando valores entre 0 e 1 e a forma da função logística.

Em muitos casos, a finalidade é estimar o efeito da variável explicativa X_i sobre a probabilidade de sucesso P(Y=1|X), o que é um pouco complexo pela natureza não linear de G(.). Então, supondo que X_i é uma variável contínua, o pretendido é obtido pela derivada parcial:

$$\frac{\partial P(y \mid X)}{\partial X_i} = G(X^T\beta)\left[1 - G(X^T\beta)\right]\beta_i .$$

O modelo *logit* aplicado ao Crédito Pessoal conforme explicado no capítulo anterior tem como finalidade calcular a probabilidade de um cliente aderir a este produto financeiro em função de um conjunto de variáveis relacionadas com as características pessoais e de envolvimento do cliente com a instituição, tomando a forma:

$$P(individuo\ adquire\ crédito\ pessoal) = P(y=1 \mid x) = \frac{e^{\beta_0+\beta_1 x_1+...+\beta_k x_k}}{1+e^{\beta_0+\beta_1 x_1+...+\beta_k x_k}} ,$$

em que x_i, i = 1, ...,K, são as designadas variáveis explicativas e β_i, i =0,..., k, são os coeficientes de regressão, parâmetros desconhecidos a estimar.

A particularidade dos coeficientes de regressão do modelo *logit* é a sua interpretação, pois estes traduzem a probabilidade de um cliente, com um determinado perfil e com um determinado envolvimento com o Banco, possuir propensão à compra deste produto financeiro.

3.2 Estimação dos parâmetros e as suas propriedades

"Devido à natureza não linear de E(y|x), o método dos mínimos quadrados não é aplicável na estimação dos parâmetros. Para a estimação de modelos com variáveis dependentes limitadas, os métodos da máxima verosimilhança são indispensáveis dado que a estimação através deste

[1] Esta igualdade deriva da assumpção de u ser independente de x e de possuir uma distribuição logística, que resulta numa distribuição simétrica relativamente a zero. Isto é, $1 - G(-z) = G(z)$.

método baseia-se na distribuição de Y dado X sendo a heteroscedasticidade na Var(y|x) automaticamente tida em consideração" (Wooldridge, 2003).

Em particular, o modelo *logit* é, normalmente, estimado através do método da máxima verosimilhança devido às suas boas propriedades assintóticas, nomeadamente, à obtenção de estimadores mais precisos.

" A especificação *logit* tornou-se muito atractiva nos anos 70 e 80 na área computacional porque a estimação de máxima verosimilhança do modelo logit era considerada fiável através da tecnologia computacional disponível nessa altura." (Horowitz e Savin, 2001)

Neste contexto, segue-se a explanação do método de estimação dos coeficientes de regressão e as suas propriedades.

Se a amostra em estudo é composta por *n* indivíduos independentes entre si e identicamente distribuídos, a função verosimilhança é o produto da probabilidade de cada observação. Então, supondo a existência de *m* observações tais que $y_i = 0$ e *n-m* observações tais que $y_i = 1$, tem-se:

$$\begin{aligned} L(\beta) &= P(y_1 = 0).P(y_2 = 0)...P(y_m = 0).P(y_{m+1} = 1)...P(y_n = 1) \\ &= \prod_{i=1}^{m}[1 - G(x_i\beta)]\prod_{i=m+1}^{n} G(x_i\beta) \\ &= \prod_{i=1}^{n} G(x_i\beta)^{y_i}\,[1 - G(x_i\beta)]^{1-y_i} \end{aligned}$$

Normalmente, utiliza-se a função logaritmo da verosimilhança, a log-verosimilhança, que para uma determinada observação *i* (*i*=1,...,n), expressa-se da seguinte forma:

$$l_i(\beta) = y_i \log[G(x_i\beta)] + (1 - y_i)\log[1 - G(x_i\beta)],$$

pode, assim, deduzir a expressão da função verosimilhança final, para os *n* indivíduos:

$$L(\beta) = \sum_{i=1}^{n} l_i(\beta).$$

Pelo método da máxima verosimilhança, os estimadores de máxima verosimilhança obtêm-se igualando a 1ª derivada de $L(\beta)$, em ordem a cada β_k, a zero.

As funções verosimilhança do modelo *logit* são globalmente côncavas, o que representa uma vantagem a nível computacional pois na implementação do método de optimização não será necessário salvaguardar a distinção entre máximos locais e máximos globais uma vez que estes são iguais.

Os estimadores de máxima verosimilhança de β, $\hat{\beta}$, máximos da função (1) têm propriedades assimptóticas bastante importantes tais como:

i. Consistência: $p\lim(\hat{\beta}) = \beta$

ii. Eficiência: $\sqrt{n}\left(\hat{\beta}-\beta\right) \xrightarrow{n} N(0,V)$, para dada matriz V definida positiva.

iii. Normalidade: $\hat{\beta} \overset{n}{\sim} N\left(\beta, I^{-1}(\beta)\right)$, onde $I(\beta)$ é a matriz de Informação de Fisher,

onde a matriz variância-covariância define-se por $\left[\sum_{i=1}^{n} \frac{\left[g\left(X_i^t\hat{\beta}\right)\right]^2 X_i X_i^t}{G\left(X_i^t\hat{\beta}\right)\left(1-G\left(X_i^t\hat{\beta}\right)\right)}\right]^{-1}$

No presente estudo, ao referir-se a este estimador estar-se-á a admitir estas três propriedades dada a dimensão da amostra em estudo.

3.3 Medidas de Qualidade de Ajustamento

Nas secções seguintes serão descritos alguns dos testes mais conhecidos para a aferição da qualidade dos modelos de regressão logística.

3.3.1 Teste de Wald

O teste de Wald pressupõe a estimação do modelo apenas sem restrições, comparando as estimativas obtidas de β com os valores dos parâmetros do modelo com restrições, possuindo uma distribuição Normal.

Considerem-se as seguintes restrições $R\beta = r$, onde R é uma matriz conhecida do tipo q x k (q <k) e r é um vector conhecido do tipo q x *1*. As hipóteses em teste são:

$$H_0 : R\beta = r \ \ versus \ \ H_1 : R\beta \neq r \ .$$

O vector $\left(R\hat{\beta}-r\right)$ revela o quão próximo se está da verificação da hipótese nula, sendo que o óptimo será a situação em que se encontra perto do valor nulo.

Sob H_0, $\left(R\hat{\beta}-r\right)$ tem uma distribuição assimptoticamente normal com média *zero* e matriz variância-covariância $R\,I^{-1}(\beta)\,R^T$ [2], donde resulta a estatística de teste

[2] Tal deve-se à propriedade de normalidade dos estimadores de máxima verosimilhança: $\hat{\beta} \overset{a}{\sim} N\left[\beta; I^{-1}(\beta)\right]$

$$W = \left(R\hat{\beta} - r\right)^T \left[R\, I^{-1}(\beta)\, R^T\right]^{-1} \left(R\hat{\beta} - r\right) \overset{a}{\sim} \chi^2_q,$$

representando q o número de restrições a testar.

Para testar a significância individual dos parâmetros, isto é, conhecer quais os coeficientes estatisticamente nulos e, por conseguinte, as variáveis (individualmente) relevantes para a explicação do modelo em estudo, basta reduzir o número de restrições a um, $q =$, e substituir r por 0.

A regra de decisão consiste em rejeitar a hipótese nula ao nível de significância de 5% se *p-value* < 0.05 e não rejeitar caso contrário.

3.3.2 Teste da Razão da Verosimilhança

O teste da razão da verosimilhança permite avaliar a significância global do modelo, determinando se as variáveis são, ou não, conjuntamente relevantes.

Para a realização deste teste será necessário maximizar a função log-verosimilhança admitindo o modelo sem restrições e com restrições ($\beta_1 = \beta_2 = ... = \beta_k = 0$), representando-se a estatística do teste da seguinte forma:

$$LR = -2\left(\ln \hat{L}_r - \ln \hat{L}\right),$$

onde $\ln L$ representa o modelo sem restrições e $\ln L_r$ o modelo com restrições.

Sob a hipótese nula, $H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$ e supondo que se está a testar k restrições, tem-se $LR \overset{a}{\sim} \chi^2_q$. Com um nível de significância de 5%, não se rejeita H_0 se *p-value*> 0.05 e rejeita-se, caso contrário.

3.3.3 Teste de McFadden

McFadden em 1974 sugere um coeficiente equivalente ao coeficiente de determinação R^2 utilizado na Regressão Linear Múltipla, embora com um significado diferente, o Pseudo-R^2:

$$\theta = 1 - \frac{L_\beta}{L_0}$$

onde $\theta \in [0;1]$ $L_\beta = L(\hat{\beta})$ e $L_0 = L(\hat{\beta}_0)$.

Esta medida é muito popular e útil na análise da bondade de ajustamento do modelo dado tomar valores entre 0 e 1.

Se L_β for igual a zero então θ tomará o valor 1, o que significa que o modelo possui um ajustamento perfeito. Caso este coeficiente se aproxime de 0, pode-se concluir que o modelo em questão tem pouco poder explicativo.

3.3.4 Teste de Hosmer-Lemeshow

O teste de Hosmer-Lemeshow permite avaliar a bondade de ajustamento do modelo obtido. Porém, para a construção da estatística de teste torna-se necessário tomar alguns procedimentos.

Em primeiro lugar, as observações são ordenadas em G grupos pela sua probabilidade (estimada) de sucesso $\hat{P}_i$. A primeira questão que surge está relacionada com o número de grupos a considerar e a forma de agrupamento das observações: "na implementação do teste de Hosmer-Lemeshow, estes autores usaram 10 grupos, definidos pelos decis de $\hat{P}_i$ ou pelos decis das observações ordenadas" (Cramer, 2003).

O segundo passo consiste no cálculo da frequência esperada de sucessos, ou seja, a soma das probabilidades esperadas ($\overline{m}_g$, soma das probabilidades estimadas) e a frequência observada (m_g). Em seguida, determina-se a probabilidade de sucesso estimada para cada grupo g, sendo esta a média de $\hat{P}_i$: $\overline{P}_g = \dfrac{\overline{m}_g}{n_g}$ e admitindo n_g a dimensão do grupo g.

Finalmente, obtém-se a estatística de teste através da fórmula:

$$HL = \sum_g \frac{\left(m_g - \overline{m}_g\right)^2}{n_g \overline{P}_g \left(1 - \overline{P}_g\right)},$$

convergindo assimptoticamente para uma distribuição χ^2 com G-2 graus de liberdade.

A hipótese nula deste teste considera que os valores observados são iguais aos esperados. Admitindo um nível de significância $\alpha = 5\%$, estar-se-á perante à aceitação da hipótese nula se *p-value*> 0.05 e da sua rejeição, caso contrário.

Uma das grandes vantagens deste teste consiste na possibilidade de observar onde residem as discrepâncias que levam à rejeição do modelo.

4 Modelação Preditiva

"O marketing está na linha da frente das aplicações, com mais sucesso, da Descoberta de Conhecimento em Bases de Dados[3] com o objectivo de prever e analisar o comportamento do cliente. Tipicamente, os principais interesses de negócio consistem em conhecer os perfis dos clientes e do produto (qual o cliente que compraria determinado produto), prever a fidelização e a retenção do cliente, e avaliar a eficiência de campanhas de vendas, do marketing directo ou de respostas a mailings" (Klösgen e Zytkow, 2002).

Neste capítulo abordar-se-á a metodologia seguida no processo de modelação e serão desenvolvidas e aprofundadas as técnicas inerentes às várias etapas do processo tais como a preparação e tratamento dos dados, a modelação e a avaliação da qualidade dos modelos obtidos.

A incidência sobre as fases de selecção e tratamento dos dados ganha relevância neste capítulo considerando que, segundo Andrienko e Andrienko, 1998, citados por Ramos e Santos, 2003, p. 5, estas fases, adicionalmente à interpretação dos resultados, ocupam a maior parte do tempo do processo e "constituem mais uma questão de "arte" do que uma rotina que possa ser automatizada". A aplicação das técnicas de *data mining* representa, normalmente, 20% do tempo utilizado em todo o processo e, por outro lado, é melhor suportada automaticamente (*software*).

De forma resumida, serão referenciadas os métodos estatísticos mais apropriadas para a resolução de problemas de propensão – as redes neuronais artificiais e uma técnica mais tradicional, o modelo *logit*, para efeitos de comparação.

Por fim, a análise da qualidade da modelação, intrinsecamente ligada à qualidade dos dados, permitirá confirmar, ou não, o sucesso do projecto.

4.1 Metodologia

A metodologia que será aplicada neste estudo será semelhante à utilizada pelo SAS Institute, normalmente através da ferramenta de *data mining* que esta empresa disponibiliza e escolhida

[3] Para alguns autores, a Descoberta de Conhecimento em Bases de Dados (Knowledge Discovery in Databases) e Data Mining têm exactamente o mesmo significado. Para outros, estes são dois conceitos diferentes considerando o Data Mining apenas como uma fase do processo de Descoberta de Conhecimento em Bases de Dados.

para a concretização deste trabalho: o Enterprise Miner. Esta metodologia é conhecida pelo acrónimo SEMMA. Groth (2000) tenta traduzir este processo de modelação num conjunto de passos metodológicos, a saber: *Sample (amostragem)*, *Explore (exploração)*, *Modify (modificação)*, *Model (modelação)* e *Assess (avaliação)*.

A amostragem, em termos muito gerais, passa por constituir um conjunto de dados que deverá ter uma dimensão suficientemente grande de forma a conter informação relevante. Por outro lado, este conjunto de dados deve ser suficientemente pequeno para viabilizar alguma eficiência computacional agilizando o processo de criação de vários modelos em tempo útil.

Numa segunda fase, explorar os dados com o objectivo de conhecer a sua distribuição, confirmar a existência relações entre variáveis, antecipar as suas tendências e até detectar anomalias/ incoerências nos dados.

Após esta análise que fará com que sobressaiam algumas conclusões prévias sobre a amostra em estudo, poderá ser necessário modificá-la, seleccionando, criando e transformando as variáveis convergindo para a construção do modelo.

Em seguida, através das técnicas/ferramentas de previsão apropriadas procede-se à modelação dos dados e à interpretação dos resultados.

Numa última etapa, avaliam-se os modelos segundo critérios e medidas de qualidade de ajustamento e de previsão. Tipicamente, para além das métricas utilizadas para avaliar a qualidade intrínseca do modelo existem também procedimento para realizar avaliações comparativas entre modelos concorrentes.

Um ponto essencial a incluir neste processo é a definição do problema. Este item antecede todos os passos, refinando-se nas três primeiras fases acima referidas (amostragem, exploração e modificação dos dados), uma vez que a formulação do problema estará sempre dependente do acesso dos dados, de quais os dados que se encontram disponíveis, da sua qualidade e de que variáveis poderão ser criadas de forma a obter o modelo que melhor espelhe a realidade (actual e futura).

4.2 Amostragem

Esta secção expõe todas as questões que deverão ser consideradas na constituição da amostra, especialmente, quando se confronta com a particularidade dos dados de campanhas de marketing, em que são apresentadas geralmente proporções bem distintas na variável *target*.

Aspectos como a dimensão e a representatividade serão abordados constantemente na selecção de dados, na aplicação do método de *oversampling* (conceito que será explicado mais adiante) e na divisão da amostra em vários subconjuntos necessários para a construção de um bom modelo de previsão.

4.2.1 Selecção dos dados

Este passo permite a selecção de dados relevantes para os objectivos do modelo. Nesta fase, o subconjunto de dados a seleccionar, o tamanho necessário da amostra e o período de tempo a considerar são tópicos que devem ser levados em conta.

Os critérios de selecção deverão contemplar os objectivos do estudo assim como a disponibilidade, qualidade e volume dos dados. Também se torna necessário ressalvar que muitas vezes a selecção dos dados não é realizada de acordo com o que o analista considera como variáveis essenciais e relevantes, mas sim de acordo com os dados disponíveis.

Tendo em conta que o presente trabalho se centra na criação de um modelo para definição de um conjunto de clientes alvo para uma campanha com base em dados históricos, ocorrerão dois momentos de selecção: inicialmente, a selecção será feita sobre os dados relativos aos clientes alvo de uma campanha anterior e, posteriormente, será feita uma segunda escolha aleatória sobre os restantes clientes da instituição, donde resultará o conjunto de clientes a quem será dirigido um *mailing*.

Esta sucessão de passos assenta no facto de se admitir "que todos os métodos de previsão partem do princípio de que as experiências do passado serão usadas no futuro, assumindo-se que as experiências que condicionaram o passado serão válidas no futuro" (Cortez, 1997).

Assim, num primeiro momento, ponderar-se-á sobre quais são as variáveis chave para a descrição do cliente e subsequente aplicação no modelo, como as características pessoais (por exemplo, o sexo e a idade), o envolvimento do cliente com a instituição financeira com variáveis como a antiguidade enquanto cliente e o valor do investimento em aplicações financeiras – variáveis de *input* do modelo. Da mesma forma, será necessário recolher informação sobre variáveis a utilizar para o *output* do modelo, que neste caso poderá ser a resposta de clientes a campanhas realizadas no passado.

Esta recolha de informação é bastante importante pelo que a janela temporal dos dados deve ser coerente e respeitar algumas condicionantes. Segundo Berry e Linoff, 2000, se não se garantir que todos os dados de *input* dizem respeito a um período temporal anterior ao dos

dados de *output*, corre-se o sério risco de se chegar a um modelo que fornece previsões inválidas.

Em consonância com o referido por Weiss e Indurkhya (1998), apesar de a previsão mais comum consistir em prever na unidade de tempo imediatamente a seguir ao período de latência[4] esta poderá incidir em várias unidades do tempo futuro. Saliente-se que a duração das campanhas de Crédito Pessoal, normalmente, é superior a um mês e, como tal, fará sentido que a previsão incida sobre a duração da campanha e não sobre a unidade de tempo considerada – o mês.

A dimensão da amostra assim como o método de amostragem também são factores fundamentais para a correcta selecção dos dados e para a representatividade da amostra. Estes factores encontram-se intrinsecamente interligados.

A dimensão da amostra tem repercussão no sucesso do processo de modelação dado que influencia a qualidade da solução final. Quanto mais pequena for a amostra, mais pobre será no que respeita à diversidade dos elementos constituintes, pelo que as estimações obtidas não serão as melhores. Por outro lado, uma amostra com muitos indivíduos tende a proporcionar melhores resultados, no entanto, requer um maior esforço computacional.

No caso em estudo, sobre uma campanha de Crédito Pessoal, o facto de a amostra estar dividida em dois grupos de indivíduos com proporções bem distintas (o grupo dos indivíduos que responderam positivamente à última campanha *versus* o grupo que não respondeu à mesma campanha), obrigará a que se aplique uma técnica de selecção/redução de dados com o objectivo de os tornar mais representativos e, consequentemente, mais úteis para uma modelação mais precisa: o *oversampling*.

4.2.2 *Oversampling*

De acordo com a definição apresentada por Berry e Linoff (2000), *oversampling* é o processo de criação de um conjunto de dados para modelação através da recolha de maior número de eventos raros e um menor número de eventos comuns, ajustando assim o rácio de eventos da amostra. Estes autores afirmam que, no caso em que amostra apresenta apenas dois tipos de eventos, o pressuposto usual para um modelo produzir bons resultados consiste em conseguir

[4] Tempo presente, isto é, período de tempo em que é construído o modelo de previsão.

que o evento raro suceda em percentagens compreendidas entre 10 e 40% do total de ocorrências da amostra e o rácio se estabeleça entre 20 e 30%.

O cenário do tema em estudo recai na situação anterior, ou seja, o objectivo consiste, em traços muito gerais, conhecer quais os indivíduos propensos à aquisição de um determinado produto em que a base de estudo é uma amostra de clientes seleccionados para o envio de um *mailing* no âmbito de uma campanha de Crédito Pessoal e em que a percentagem de aquisições é muito pequena. Face ao exposto, é fácil antever um modelo que só preverá as não aquisições e, como o interesse da instituição recai sobre os indivíduos propensos, a aplicação deste modelo não serviria. O processo de *oversampling* apresenta-se com uma possível resolução para este problema.

Na análise de campanhas de marketing realizadas nas mais diversas áreas de negócio é frequente e normal a constatação de taxas de resposta muito baixas relativamente ao número de contactos realizados ou ao número de *mailings* enviados, podendo atingir percentagens entre 1 e 10%. Perante uma amostra deste género, em que os eventos (resposta positiva à campanha) são considerados raros, o analista é confrontado com uma preocupação essencial, que consiste em distinguir qual o método de amostragem adequado para atingir representatividade q.b. de forma a obter estimadores precisos e, consequentemente, boas previsões.

Como referido por Madigan e Nason citados por Klösgen e Zytkow, 2002, p. 205, a teoria da amostragem é umas das áreas com mais sucesso na estatística moderna, permitindo, com base em métodos de amostragem simples, a estimação fiável, não enviesada e eficiente das características de uma população.

O processo de amostragem não é mais do que criar um conjunto de dados retirados do universo em estudo mas com uma dimensão menor. A condição determinante para o sucesso deste processo é a representatividade da amostra.

"A precisão de um estimador depende em parte da variabilidade natural inerente à população. Contudo, o método de amostragem poderá aumentar a precisão de um estimador de duas formas. Em primeiro lugar, na maior parte das situações, a variabilidade da estimação é inversamente proporcional à raiz quadrada da dimensão da amostra. Portanto, a precisão de um estimador irá aumentar à medida que o tamanho da amostra aumenta. Em segundo lugar, uma escolha ponderada do método de amostragem poderá resultar em estimadores mais precisos." (Madigan e Nason citados por Klosgen e Zytkow, 2002, p. 205).

Uma técnica para a obtenção de uma amostra mais representativa, assenta em dividir o conjunto de dados em dois grupos distintos, tirando partido da variável de interesse ser binária: o grupo dos indivíduos com resposta positiva à campanha e o grupo com respostas negativas, e determinar proporções mais equilibradas entre os dois conjuntos.

Tendo em consideração que a dimensão do primeiro grupo é substancialmente menor do que a do segundo e insuficiente para a modelação, pode admitir-se que a totalidade dos indivíduos que o compõem se manterá na nova amostra. Por outro lado, torna-se necessário aumentar a representatividade destes indivíduos de modo a garantir o equilíbrio dos dados na amostra. Para desenvolver este passo, poderá optar-se por fixar uma percentagem de respostas positivas, passando este valor a representar a nova proporção de respostas positivas na nova amostra. Determinada a proporção respostas positivas, segue-se a selecção das respostas negativas pelo método de amostragem aleatória simples. Assim, garante-se a constituição de uma amostra mais equilibrada e definida com base em critérios úteis ao processo de modelação, e, em princípio, estimadores mais precisos.

Exemplificando, se nos dados disponíveis sobre a última campanha, existirem 1% de respostas positivas, em primeiro lugar considerar-se-á que todos esses indivíduos pertencerão à nova amostra. Posteriormente, estabelece-se que os elementos que responderam positivamente à última campanha devem corresponder, por exemplo, a 30% da nova amostra. Consequentemente, a amostra fica formada e dimensionada bastando seleccionar indivíduos da amostra inicial que não tenham respondido à última campanha garantindo que esse grupo constitui 70% da nova amostra.

Esta é a técnica aplicada no caso em estudo, todavia, não é a única forma de utilizar *oversampling*. Uma alternativa passa por replicar os eventos raros aumentando assim a sua dimensão e/ou dar mais "importância" aos eventos raros atribuindo-lhe um peso maior que os eventos mais comuns.

4.2.3 Conjunto de dados: Treino, validação, teste e *score*

A partição de dados é usualmente utilizada em modelos de previsão, e consiste na subdivisão do conjunto de dados, a servir de base para a modelação, em três tipos de conjuntos diferentes: de treino, de teste e validação.

"O conjunto de treino é utilizado para gerar uma explicação da variável dependente (*target*) em termos das variáveis independentes (*input*)" (Berry e Linoff, 1997). Do ajustamento feito sobre este conjunto surge o reconhecimento de padrões que se encontram ocultos nos dados. Contudo, deve ter-se presente que "todos os métodos de descoberta de conhecimento sofrem da tendência de ler em demasia os dados do conjunto de treino" (Berry e Linoff, 1997). A este fenómeno dá-se o nome de *sobreaprendizagem.*

Em termos estatísticos, *sobreaprendizagem* é o ajustamento a um modelo estatístico que tem demasiados parâmetros, ou seja, um número elevado de graus de liberdade. Se um modelo tiver complexidade suficiente para comparação de todos os dados disponíveis, apresentará um ajustamento perfeito. No entanto, esse mesmo modelo poderá não cumprir o objectivo de representar a realidade do fenómeno em estudo, uma vez que é concebido com base numa amostra, que, como qualquer amostra, dificilmente possuirá uma representatividade de 100% da população.

Este fenómeno poderá ser detectado confrontando o conjunto de treino com um outro conjunto: o de teste. Quando o primeiro conjunto apresenta um erro considerado pequeno, contrastando com um erro grande no conjunto de teste, está-se perante a *sobreaprendizagem.*

Bishop (1997) afirma que "o principal objectivo do reconhecimento de padrões é produzir um sistema que faça boas previsões para dados novos, noutras palavras, um sistema que exiba uma boa generalização. Para mensurar as capacidades de generalização da função, temos de gerar um segundo conjunto de dados, designado o conjunto de teste, o qual é produzido da mesma forma que o conjunto de treino, mas com novos valores para a componente de ruído". Em suma, o objectivo do conjunto de teste é medir a performance de generalização do modelo obtendo uma estimação não enviesada do erro.

Durante o processo de treino da rede, é utilizado um novo conjunto de dados para monitorização e melhoria da estimação do modelo – o conjunto de validação. O treino da rede é parado periodicamente para que esta seja testada sobre este novo conjunto medindo o seu erro. Este processo é repetido até à minimização do erro de validação, refinando, assim, o modelo e permitindo obter uma melhor generalização. Os resultados deste conjunto mostram como o modelo se comporta perante novos dados.

De acordo com Bishop (1997), a performance das redes neuronais é comparada através da avaliação do (menor) erro da função do conjunto de validação e, após selecção da considerada

"melhor" rede, a sua performance será confirmada pela medição do erro proveniente do conjunto de teste.

Importa referir que os três conjuntos descritos são independentes entre si, isto é, não contêm dados comuns. Quanto maior for a sua dimensão melhores serão os resultados obtidos, quer individualmente quer na obtenção do melhor modelo. Surge assim uma questão pertinente: que dimensão deverão ter estes conjuntos?

Apesar de não existir uma fórmula com percentagens fechadas, o conjunto de treino deverá ser sempre o maior, uma vez que a tarefa para a qual é utilizado requer a maior quantidade de informação possível.

Berry e Linoff (2000) afirmam que a divisão 60%-30%-10% (treino - teste -validação) revelou funcionar muito bem em casos práticos. Porém, quando perante amostras de dimensão reduzida, estes subconjuntos poderão não ser suficientemente representativos e, por conseguinte, torna-se infrutífera a divisão e prejudicará o ajustamento do modelo. Face a este cenário, poder-se-á prescindir do conjunto de teste admitindo assim que o modelo obtido tem uma boa generalização.

Após finalizada a modelação, é necessário introduzir um novo conjunto de dados aos quais será aplicado o modelo que foi construído com base nos conjuntos que referidos anteriormente.

Após o modelo construído, deverá ser seleccionado um conjunto de clientes que não foram alvo da campanha de marketing anterior e para os quais, consequentemente, o valor da variável *target* é desconhecido. A este conjunto será aplicado o melhor modelo resultante da estimação anterior com o objectivo de prever qual o valor da variável *target* para cada registo (cliente) – conjunto *score*.

Na presença de um estímulo como é um contacto no âmbito de uma campanha, haverá clientes que se mostrarão motivados à aquisição, chegando até à concretização do crédito.

4.3 Exploração dos Dados

A exploração dos dados consiste numa análise preliminar dos dados, que tem como objectivo a aquisição de algum conhecimento sobre o comportamento das variáveis e até sobre a

qualidade dos dados. Os resultados desta análise permitirão que o analista possa *a posteriori* proceder à correcta transformação dos dados.

A grande vantagem desta análise consiste no facto de a mesma assentar na rapidez e facilidade de interpretação pois, geralmente, o seu *output* é reflectido em termos de gráficos ou estatísticas descritivas.

Uma das tarefas mais importantes nesta etapa passa por visualizar a distribuição das variáveis, o que constitui uma grande vantagem quando se está perante um grande volume de dados, sendo o histograma a ferramenta mais utilizada para esta apreciação. Deste tipo de gráficos poderão retirar-se várias conclusões determinantes na preparação dos dados como a necessidade de agregação de valores ou até de variáveis, a previsão da importância das variáveis de *input* na previsão da variável *target*, a quantidade de *missing values*, em suma, um conjunto de informações potencialmente relevantes para a transformação e que serão temas a abordar na próxima secção.

Por outro lado, estatísticas descritivas como a média e o desvio-padrão também são relativamente simples de obter e interpretar, podendo descrever a distribuição associada a cada variável. Em particular, a dispersão dos dados pode ser visualizada graficamente através da evidência da mediana e dos quartis mais significativos ou até de possíveis *outliers*.

4.4 Transformação dos Dados

Os métodos preditivos têm potencial para encontrar padrões de comportamento nos dados e estimar medidas de performance dos modelos, determinando assim quais os que melhor servirão no cumprimento dos objectivos da análise.

Contudo, estes métodos partem do princípio que os dados se encontram num formato *standard* quando na maior parte dos casos reais isso não sucede. A preparação prévia dos dados pode influenciar a qualidade dos resultados e, aproveitando o exemplo dado pelos autores Weiss e Indurkhya (1998), uma simples especificação de um rácio poderá melhorar os resultados preditivos. Como referem estes autores, o *design* e a organização dos dados, incluindo o conjunto de objectivos e a composição das variáveis, são realizados pelos analistas, existindo dois objectivos principais na preparação dos dados: organizar os dados

numa forma *standard* e adequada para o processamento através de programas de previsão e preparar as variáveis que conduzem a uma melhor performance preditiva.

Na preparação dos dados, o analista depara-se com questões como a formatação, a conversão, a limpeza e extracção dos dados. Após obtenção dos dados, o objectivo é integrá-los numa única base de dados e representar toda essa informação num só formato, de forma coerente e consistente, reduzindo o *ruído*.

Os valores omissos, também designados *missing values*, poderão constituir um entrave importante ao desenvolvimento dos modelos. Se o conjunto de dados contiver muitos *missing values*, a tarefa de obter informação relevante e útil ou de previsão torna-se difícil. O tratamento deste tipo de dados poderá consistir em várias técnicas como a estimação do valor mais provável ou substituir pelo valor médio ou o mais frequente, ou criar uma classe específica para os *missing value*. É comum este tipo de constrangimento em colunas importantes como a do código postal quando são seleccionados potenciais clientes para o envio de *mailing*. Neste caso concreto, o melhor procedimento é a eliminação desses registos.

Segundo Berry and Linoff (2000), a insuficiência de dados históricos para muitos *inputs* é um outro problema de *missing data* que deve ser evitado. Segundo os autores, este problema surge, geralmente, porque os *inputs* dos modelos vão mais atrás no tempo para uns registos do que para outros, e a sua resolução passa pela redefinição do problema inicial. Um outro facto interessante, é que nalguns processos de modelação, é possível extrair muita informação se se perceber qual o padrão de variáveis que têm valores omissos.

A consistência e a coerência também se apresentam como requisitos que o conjunto de dados a modelar deve satisfazer. Tipicamente, a maior parte das inconsistências deve-se a erros manuais de entrada dos dados ou à forma de integração dos mesmos. A resolução para este tipo de problemas consiste na utilização de *querys* simples e rotinas como a verificação e o conhecimento dos domínios e formatos das variáveis.

Uma prática comum quando se analisam dados históricos é a criação de novas variáveis conhecidas por *variáveis derivadas*. Estas variáveis são combinações de outras variáveis que podem revelar-se mais interessantes no ajustamento do modelo aos dados o que as restantes.

As transformações referidas até aqui podem ser usadas para estabilizar as variâncias, remover

a não-linearidade das variáveis ou, ainda, normalizá-las. Saliente-se que o conceito de normalização[5] referido em todo o trabalho consiste na aproximação à distribuição Normal.

Como exemplo, se as variáveis de envolvimento de um cliente, respeitantes a vários períodos de tempo (meses) forem combinadas entre si, formando uma média desse espaço temporal, contribuirão para a redução de dados e, simultaneamente, para melhores resultados preditivos.

Um outro exemplo, mais complexo, consiste em calcular, para um determinado período de tempo, o montante total gasto em compras por cada cliente e classificar cada um dos clientes com o indicador do decil em que se encontra. Citando Berry e Linoff (2000): "estas variáveis são poderosas porque o comportamento passado é muitas vezes um preditor forte do comportamento futuro".

A existência de *outliers* é outro problema a resolver quando se realizam análises de dados, uma vez que estes registos, por terem valores tão distantes da maior parte dos dados, podem distorcer o resto dos dados reduzindo-os à insignificância (Pyle, 1999).

Uma das formas de suavizar o problema dos *outliers* consiste em transformar os dados, nomeadamente, através da utilização de funções que têm um efeito muito maior sobre valores elevados, como o logaritmo ou a raíz quadrada.

De acordo com Han e Kamber (2001), a transformação dos dados agrupa um conjunto de tarefas como a agregação, a generalização, a criação de atributos e a normalização, tendo esta uma importância relevante.

A normalização consiste na transformação de todas as variáveis, a incluir no modelo, com o objectivo de resolver problemas relacionados com as diferenças de amplitude entre as variáveis e com as diferenças de distribuição das mesmas.

No caso das redes neuronais, a normalização apresenta algumas vantagens entre as quais maior rapidez no processo de aprendizagem e o facto de evitar valores iniciais elevados para dadas variáveis (como montante disponível em determinado produto financeiro) *versus* valores iniciais muito baixos como para o caso das variáveis binárias. Possibilita também que os dados de treino de uma rede neuronal se distribuam ao longo de todo o domínio das funções de activação, potenciando o aumento da capacidade de assimilação da rede.

---

[5] Importa salientar que este conceito não se refere à estandardização de variáveis, isto é, supondo X uma variável aleatória com distribuição Normal, média μ e variância σ^2, a variável $Z = {(X-\mu)}/{\sigma}$ define-se como uma variável estandartizada possuindo uma distribuição $N(0,1)$.

Ao nível das distribuições das variáveis, a normalização facilita a remoção da distorção gerada pela existência de *outliers* e garante uma preditividade linear.

4.5 Técnicas de modelação

A modelação dos dados pressupõe a aplicação das técnicas mais adequadas à análise proposta no presente trabalho: a propensão dos clientes de uma instituição financeira ao Crédito Pessoal.

Com base em referências bibliográficas e em *case studies* com problemas semelhantes ao proposto anteriormente, encontra-se devidamente fundamentada na secção 1.2. a escolha das técnicas propostas – as redes neuronais – e outra técnica também muito mencionada nos problemas de classificação e previsão – a regressão logística.

4.6 Qualidade da modelação

Seguidamente, serão apresentadas algumas técnicas que permitem medir a qualidade dos modelos obtidos e, simultaneamente, comparar os vários métodos de modelação.

Na análise estatística é fundamental a avaliação da qualidade dos modelos construídos, determinando o sucesso do trabalho ou, na maioria dos casos, sendo decisivo aquando da escolha do melhor modelo. Groth (2000) confirma que "a qualidade dos dados sobre os quais são tomadas decisões em todo o mundo é, frequentemente, suspeita. Os resultados de *data mining* serão tão bons quanto os dados que representa."

Como referido no ponto 4.2.2., dada a natureza da amostra será utilizada a técnica de *oversampling* e, como tal, os resultados do modelo deverão ser alterados por forma a corresponderem aos valores reais da amostra inicial e a serem correctamente interpretados.

4.6.1 Lift

Através das técnicas de previsão mencionadas, atribui-se um *score* a cada cliente, sendo que os clientes com um valor *score* alto serão incluídos na campanha e os que possuírem um valor *score* baixo serão excluídos. Daqui sobressai uma questão essencial: qual o *threshold* óptimo de *score* que servirá de base para esta decisão?

Para campanhas com um orçamento pré-estabelecido, esta decisão poderá ser ultrapassada determinando o número limite de clientes a seleccionar, e utilizando como critério os *scores* mais elevados.

Piattetsky-Shapiro e Steingold (2000) defendem a necessidade de comparar a qualidade dos modelos em problemas no âmbito do marketing, justificando que a medida tradicional utilizada, a média da taxa de erro aplicada a todos os casos, é desadequada na decisão de contactos para uma campanha de marketing. Normalmente, os *marketeers* contactam os 10% ou 20% dos clientes com o *score* mais elevado, sendo que a precisão da previsão dos restantes 90% ou 80%, respectivamente, não tem muita relevância.

Por outro lado, os custos dos erros não serão iguais numa situação em que o cliente foi contactado mas não respondeu afirmativamente, como previsto, e noutra em que o cliente não foi contactado mas, se fosse, teria dado uma resposta afirmativa. No primeiro caso, o custo será o correspondente ao valor do envio do *mailing*, enquanto que no segundo poderá estar-se perante uma perda dos lucros inerentes à venda de um produto ou até, num caso mais extremo, à perda de um cliente.

A determinação do valor do *lift* auxilia neste tipo de decisões e é a medida mais utilizada no *marketing*.

A expressão do *lift* é dada por:

$$Lift\ (M,p) = \frac{\%T\arg ets(M,p)}{b},$$

onde o numerador expressa a % de clientes que se encontram no percentil p e para os quais o modelo M prevê uma resposta positiva, e b a *densidade* da amostra. A *densidade* é expressa pelo quociente entre o número de eventos raros e o número total de eventos do conjunto de dados em estudo.

Como referido anteriormente, as decisões das empresas são tomadas mediante custos e proveitos que poderão advir das acções de marketing, pelo que se torna necessário construir uma matriz onde são representados esses valores monetários (matriz de custos/benefícios).

Dada a matriz de custos/benefícios, os custos unitários são multiplicados por cada envio de mailing e o proveito por cada adesão ao produto (respostas positivas). O lucro será calculado pela diferença entre os dois valores obtidos, o que permitirá obter conclusões sobre a relação entre o conjunto de clientes a contactar (em percentis) e o lucro que poderá advir desses contactos.

Para as amostras construídas através da técnica do *oversampling,* será necessário estimar o verdadeiro *lift* (do conjunto de dados original) dividindo a densidade da amostra pela densidade do conjunto de dados original.

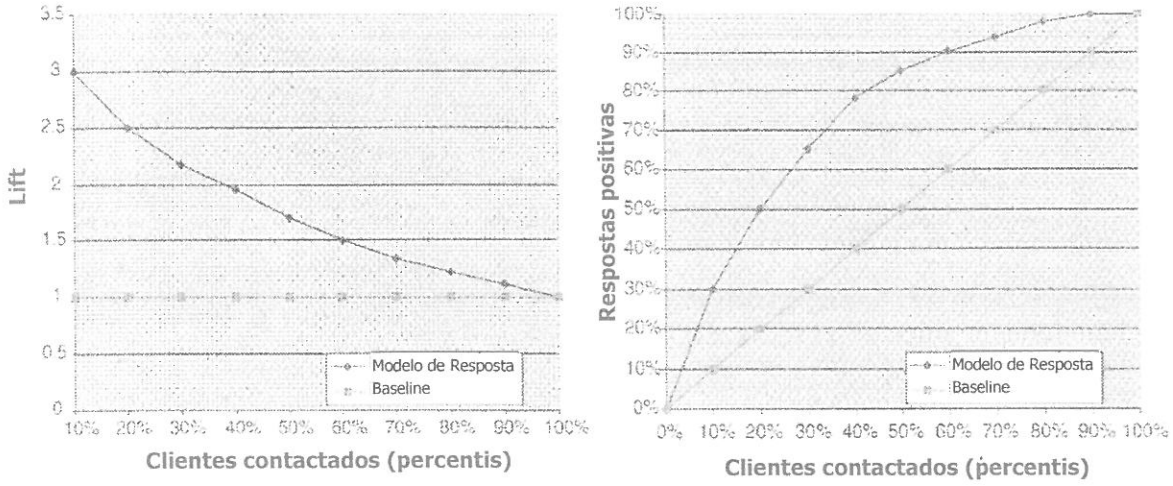

Figura 4 – Representação gráfica do *Lift* e dos Ganhos Cumulativos para o mesmo modelo.

A representação gráfica do *lift* tem como objectivo mostrar a performance do modelo de propensão. Este gráfico não prevê a taxa de resposta, no entanto, um valor do *lift* de 3 poderá indicar uma previsão de 3% de taxa de resposta como de 90%, dependendo de a taxa de resposta da população em estudo ser de 1% ou 30%, respectivamente (Berry e Linoff, 2000).

4.6.2 Matriz de Confusão

Outra forma de avaliar a qualidade de ajustamento de um modelo é medir a sua capacidade de previsão, calculando as percentagens de previsões correctas e incorrectas sobre os dados amostrais.

Isto significa que são comparadas as observações de um determinado acontecimento com as previsões desse mesmo acontecimento sendo o ajustamento tanto melhor conforme for maior a semelhança entre estas duas classificações.

Este método traduz-se na criação da matriz de confusão, na qual, de forma muito simples são expressos os valores de similitude entre a previsão e a realidade.

Para o problema da resposta a uma campanha de Crédito Pessoal dirigida a um universo de clientes, a matriz será preenchida com os seguintes valores:

True Positive (TP) – taxa de verdadeiros positivos, ou seja, os indivíduos para os quais a previsão indicava que responderiam positivamente à campanha, tendo se verificado que de facto responderam positivamente:

$$TP = \frac{d}{c+d}.$$

False Negative (FN) – taxa de falsos negativos, ou seja, os indivíduos para os quais a previsão indicava que não responderiam à campanha, tendo se verificado que afinal responderam positivamente à campanha:

$$FN = \frac{c}{c+d},$$

Para estas duas taxas, c é o número de previsões incorrectas de que a resposta seria negativa, quando de facto a resposta foi positiva, e d representa o número de previsões correctas de que a resposta seria positiva.

True Negative (TN) – taxa de verdadeiros negativos, ou seja, os indivíduos para os quais a previsão indicava que não responderiam à campanha, tendo se verificado que de facto não responderam:

$$TN = \frac{a}{a+b}.$$

False Positive (FP) – taxa de falsos positivos, ou seja, os indivíduos para os quais a previsão indicava que não responderiam à campanha, tendo se verificado que afinal não responderam à campanha:

$$FP = \frac{b}{a+b}.$$

Para estas duas taxas, a é o número de previsões correctas de que a resposta viria a ser negativa, e b é o número de previsões que davam a resposta como negativa, tendo-se verificado que as respostas afinal eram positivas.

		Valores previstos	
		Resposta Negativa	Resposta Positiva
Valores observados	Resposta Negativa	*TN* ou *a*	*FP* ou *c*
	Resposta Positiva	*FN* ou *b*	*TP* ou *d*

Tabela 2 – Matriz de Confusão

4.6.2.1 Matriz de Confusão com inserção de custos

Conforme foi explicado, a matriz apresentada na secção anterior, permite avaliar a qualidade do modelo de previsão, tendo em conta que possibilita comparar o valor previsto com o valor real e disponibiliza, assim, a informação sobre o que podemos designar "erros" de previsão.
Face à exigência do mundo empresarial, esta informação torna-se bastante útil na medida em que a sua utilização permite quantificar os custos incorridos por cada um dos "erros" cometidos na previsão face à observação, sejam estes custos de natureza financeira ou custos de oportunidade.
No problema de marketing em estudo neste trabalho, os dois tipos de erro materializam-se em enviar o *mailing* a um cliente não propenso e que, portanto, não responderá à campanha (falsos positivos), e não enviar o mailing a um cliente que estaria propenso a aderir à campanha e, portanto, a contratar um crédito pessoal com a instituição (falsos negativos).
Estas duas situações são totalmente distintas e têm impactos de grandezas muito diferentes, pois se, no primeiro caso, o custo em que a empresa incorre limita-se aos custos de envio da comunicação, podendo os mesmos ser deduzidos pela divulgação da marca ou do produto em causa e que poderá originar algum tipo de oportunidade de negócio futura, enquanto que no segundo caso a probabilidade de se estar a perder um negócio "quase certo" é muito elevada, incorrendo assim a empresa no que se designa por custo de oportunidade, ou seja, a não criação do lucro que estaria associado a este negócio. Naturalmente, este segundo tipo de erro, é muito mais preocupante para a empresa.
Assim, torna-se fundamental associar o custo de cada um dos tipos de erro mencionados, à matriz de confusão, para que se possa concluir sobre os custos decorrentes da má previsão.
Perante o exposto, claramente se define como objectivo a tomada de decisão que minimiza os custos esperados. Esta decisão será baseada na matriz de confusão com inserção de custos associados aos dois tipos de erro: os falsos negativos e os falsos positivos.
Designe-se, então, o custo associado à previsão do indivíduo x pertencer à classe i quando a classe correcta seria j como $C(i, j, x)$ e o valor óptimo da previsão poderá ser definido como

$$\sum_j P(j \mid x) C(i, j, x) .$$

Supondo que o custo por classificar um cliente incorrectamente como propenso ao Crédito Pessoal é o custo do envio do mailing (k_1 €) e que, por outro lado, o custo associado à previsão de um cliente como não propenso, não sendo seleccionado para a campanha e, por

conseguinte, representa uma perda equivalente ao lucro por adesão ao produto (k_2 €). Então a matriz de custos definir-se-ia como:

$$\begin{bmatrix} 0 & k_2 \\ k_1 & 0 \end{bmatrix}.$$

Os valores nulos correspondem aos custos associados às previsões correctas.

4.6.3 Curva ROC e a métrica F

A análise ROC (Receiver Operating Characteristic) foi uma técnica pioneira na década de 40, utilizada como uma forma de detecção e reconhecimento de sinais na presença de ruído (Egan, 1975; Green e Swets, 1974 citados por Wodon, 1997, p. 2083). Segundo Wodon (1997), esta técnica tem sido utilizada para a análise e evolução da performance da visão, memória, previsão meteorológica, detecção de mentiras através do polígrafo, imagens, radiografia na medicina dentária, etc.

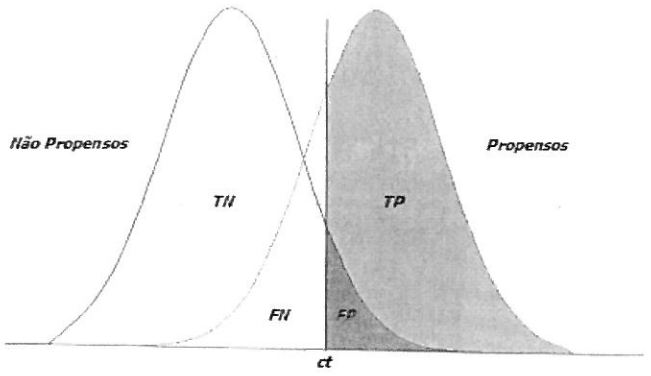

Figura 5 – Funções de densidade de probabilidade, para os indivíduos propensos e não propensos ao Crédito Pessoal

A curva ROC é uma representação gráfica da *sensibilidade* versus *especificidade*, sendo estas coordenadas determinadas pela variação do *valor de corte* ao longo de um eixo de decisão.

A *sensibilidade* define-se como a probabilidade do envio de *mailing* aos indivíduos que se encontram efectivamente propensos ao Crédito Pessoal e a *especificidade* como a probabilidade de decisão de não envio de *mailing* a um cliente quando este não iria aderir à campanha.

O *valor de corte* (*ct*) é a probabilidade a partir da qual o analista considera que o cliente se encontra propenso ou não ao produto e, com base nessa probabilidade, pode decidir enviar ou não um *mailing* ao cliente.

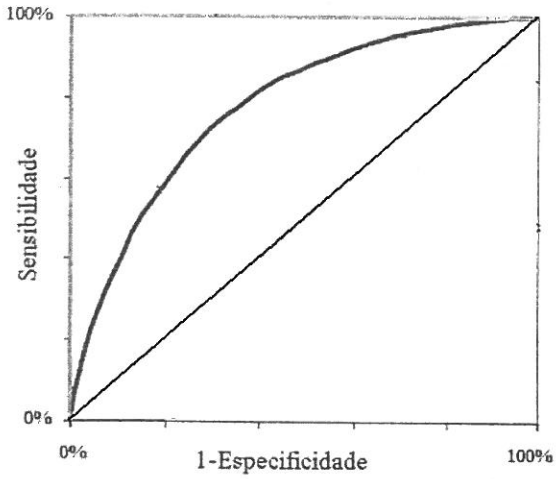

Figura 6 – Curva ROC: gráfico da especificidade *versus* sensibilidade para um domínio de valores de corte

Em termos de testes de hipóteses, admitindo

$$H_0 : O\,cliente\,adere\,à\,campanha \quad versus \quad H_1 : O\,cliente\,não\,adere\,à\,campanha,$$

tem-se,

$$\alpha = P(erro\,de\,tipo\,I) = P(rejeitar\,H_0 \mid H_0) = 1 - sensibilidade$$

$$\beta = P(erro\,de\,tipo\,II) = P(não\,rejeitar\,H_0 \mid H_1) = 1 - especificidade$$

onde o *valor de corte* é o *p-value* que define a região de rejeição.

Defina-se potência do teste:

$$k = \begin{cases} P(erro\,do\,tipo\,I) & sob\ H_0 \\ 1 - P(erro\,do\,tipo\,II) & sob\ H_1 \end{cases}.$$

A curva ROC mostra a relação entre a potência do teste e a probabilidade de ocorrer um erro do tipo I com a variação do valor crítico (*p-value*) do teste de hipóteses (Metz, 1986).

Na interpretação dos valores da figura 7, pode-se concluir que valores de corte baixos traduzem um teste muito sensível e pouco específico e valores de corte altos traduzem o contrário, sendo que o óptimo seria obter um teste muito sensível e muito específico.

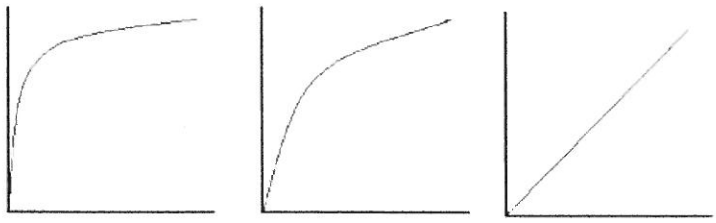

Figura 7 – Curvas ROC: representações com diferentes graus de previsibilidade (por ordem, gráficos de bom, moderado e pobre modelos de previsão)

De seguida, introduz-se dois conceitos importantes para determinar uma medida, designada métrica *F*, que será complementar à análise anterior:

$$precision = \frac{TP}{TP + FP} \quad \text{e} \quad recall = \frac{TP}{TP + FN} \quad .$$

O termo *precision* traduz a percentagem de classificações correctas de respostas positivas entre todas as previsões de respostas positivas enquanto que o *recall* exprime a percentagem de classificações correctas de respostas positivas sobre todas as respostas positivas observadas. Note-se que o conceito de *recall* converge para o da sensibilidade.

"Enquanto que as curvas ROC representam o *trade-off* entre os valores de TP e FP, a métrica *F* representa o *trade-off* entre os diferentes valores de TP, FP e FN" (Buckland e Gey, 1994, citados por Chawla, 2005, p. 857), exprimindo-se na seguinte forma:

$$Métrica\ F = \frac{(1 + \beta^2) * recall * precision}{\beta^2 * recall + precision},$$

onde β corresponde ao rácio entre *precision* e *recall* sendo, normalmente, substituído pelo valor 1. Pretende-se, portanto, quanto maior for o valor desta métrica melhor será a bondade de ajustamento do modelo perante a presença dos eventos raros.

4.6.4 Akaike's Information Criterion (AIC)

O AIC foi concebido por Akaike em 1973 para servir como medida de comparação da performance entre modelos diferentes, baseados em amostras de dados diferentes.

Para uma amostra de dimensão *n*, e com *k* parâmetros estimados, o AIC traduz-se na equação:

$$AIC = (n)\ln(SSE / n) + 2k,$$

onde SSE (*Sum of Squared Errors*) é a soma dos erros quadráticos.

Note-se que esta medida considera o facto de a soma dos erros quadráticos aumentar com o número de parâmetros estimados, k, e com o tamanho da amostra, n, tornando-se penalizador quando perante um modelo com uma dimensão grande e/ou com um elevado número de parâmetros.

Quanto mais baixo for este indicador, melhor será o ajustamento do modelo em análise, ou seja, numa situação de selecção do melhor modelo será escolhido o que apresentar o menor valor.

4.6.5 Schwarz's Bayesian Criterion (SBC)

Tal como o AIC, Shwarz's Bayesian Criterion surge como um indicador estatístico para avaliação do ajustamento dos modelos, criado em 1978 por Gideon E. Schwarz.

Esta medida baseia-se na hipótese que os dados em avaliação possuem uma distribuição pertencente à família da distribuição exponencial.

Sob a hipótese de que os resíduos do modelo são normalmente distribuídos, tem-se

$$\begin{aligned} SBC &= -2\ln L + k\ln(n) \\ &= (n)\ln(SSE/n) + k\ln(n) \end{aligned},$$

onde L é a função da máxima verosimilhança do modelo, n é a dimensão da amostra, k é o número de parâmetros estimados, *SSE* (*Sum of Squared Errors*) é a soma dos erros quadráticos.

Perante estimativas do SBC relativas a vários modelos, o modelo a ser seleccionado deverá ser o que possuir o menor valor.

A variação não explicada na variável dependente e o número de variáveis explicativas aumenta o valor de SBC o que significa que valores baixos desta medida implicam um melhor ajustamento ou um menor número de variáveis explicativas ou ambos.

5 Análise de Resultados

Como já foi referido anteriormente, o objectivo deste trabalho é a construção de um modelo de propensão, especificamente orientado para o produto financeiro que é o Crédito Pessoal, respondendo assim à grande questão: quem são os clientes propensos ao consumo deste produto? Para tal, foram utilizados dados reais fornecidos por uma instituição financeira e que são resultado de uma acção de marketing.

Através deste trabalho, pretende-se demonstrar, por um lado, em que aspectos as redes neuronais representam uma vantagem competitiva para o negócio comparativamente às técnicas mais tradicionais através da análise à performance de ambas e, por outro, mostrar em que pontos estas técnicas poderão não corresponder às expectativas esperadas.

Neste capítulo, serão mencionadas as decisões mais relevantes tomadas em vários momentos do processo de modelação, tanto na preparação dos dados como na aplicação das técnicas estatísticas e, sobretudo, será elaborada uma análise crítica sobre os resultados e o sucesso do modelo obtido.

É importante referir que, por uma questão de confidencialidade, as variáveis descritas adiante serão codificadas sempre que forem visíveis os seus resultados. Os resultados de teste e dimensões que constam das tabelas assim como as escalas nos gráficos também não corresponderão aos reais. Porém, garante-se que estas alterações não interferem na interpretação dos resultados e das conclusões daí provenientes.

5.1 A Amostra

A base de dados utilizada para a construção do modelo de propensão é composta por 82 269 registos de clientes que foram alvo de uma campanha de Crédito Pessoal que ocorreu no 2º trimestre do ano passado (2007).

As variáveis seleccionadas para a base inicial constituem um conjunto de 19 atributos, obtendo-se uma matriz com dimensão de 82 269 *x* 19.

Na fase inicial, antes da aplicação das técnicas de *data mining*, estes dados foram analisados e devidamente preparados envolvendo as actividades de selecção, exploração e transformação das variáveis.

5.1.1 Variáveis

As variáveis seleccionadas para estudar as diferenças entre os clientes propensos e os não propensos podem dividir-se entre as características pessoais e as que estão relacionadas com o envolvimento do cliente com o banco.

Natureza da Variável	Variável	Descrição	Tipo de Variável
Características Pessoais	Estado Civil	Solteiro Casado Viúvo Divorciado União de Facto Separado	Categórica
	Sexo	Masculino Feminino	Categórica
	Idade	Idade do cliente (em anos)	Contínua
	Profissão	Comerciante Liberais Operários Trabalhadores Administrativos Reformados Estudantes . . .	Categórica
	Situação Profissional	Por conta Própria Por conta outrém Outros	Categórica
	Nível Habilitações	S/ Escolaridade Ensino Básico Ensino Secundário Bacharelato Licenciatura . . .	Categórica
	Região	Região em que reside com base no código postal	Categórica
Envolvimento do Cliente com o Banco	Antiguidade	Antiguidade de cliente na instituição (em anos)	Contínua
	Segmento	Segmento do cliente (P/B/N/O)	Categórica
	Tipo Cliente	Indicação de que se trata de um cliente activo ou não	Categórica
	Nº Produtos	Nº produtos	Contínua
	Produto Tipo 1	Média do montante aplicado no produto 1 nos últimos 6 meses	Contínua

	Produto Tipo 2	Média do montante aplicado no produto 2 nos últimos 6 meses	Contínua
	Produto Tipo 3	Média do montante aplicado no produto 3 nos últimos 6 meses	Contínua
	Produto Tipo 4	Média do montante aplicado no produto 4 nos últimos 6 meses	Contínua
	Produto 7	Indicação se possui o produto 7 (S/N)	Categórica
	Produto 8	Indicação se possui o produto 8 (S/N)	Categórica
	Recursos	Média dos recursos do cliente nos últimos 6 meses	Contínua
	Aplicações	Média das aplicações do cliente nos últimos 6 meses	Contínua
Target	Crédito Pessoal	S/N (resposta à campanha)	Categórica

Tabela 3 – Identificação das variáveis do modelo e respectiva descrição e formato

Saliente-se o facto de terem sido excluídos desta selecção os clientes que não apresentam código postal por impossibilidade de futuro envio do *mailing*, e cuja profissão recai no grupo dos desempregados, estudantes e domésticas por questões inerentes ao risco do crédito, ou que, simplesmente, não apresentam valores válidos garantindo alguma coerência nos dados.

Quando analisada a distribuição das variáveis foram identificados vários aspectos chave como a dispersão de valores/categoria e os *missing values*. No seguimento desta análise, decidiu-se criar novas variáveis com base na frequência das respectivas classes descritas na tabela 3, agrupando-as em novas classes para o estado civil, profissão, nível de habilitações e o número de produtos que o cliente possui na instituição.

Relativamente aos *missing values*, apenas duas variáveis[6] apresentaram ocorrência de valores omissos com pesos de 3% e 22%. No entanto, foi necessário substitui-los por um valor estimado optando-se pelo valor com maior frequência uma vez que se trata de uma variável qualitativa. Para as variáveis cuja omissão de valor enviesaria a previsão (ex. estado civil ou sexo) ou impossibilitaria o envio de *mailing* como o código postal, a resolução consistiu na eliminação desses registos. No conjunto de treino e de validação não se verificou nenhum destes casos, contudo, no (eventual) conjunto de *score* o expurgo destes registos será considerado como uma das condições primordiais.

[6] A designação das variáveis não será referenciada por questões de compromisso de sigilo.

Importa referir que se expurgaram os clientes com antiguidade inferior a 6 meses uma vez que as variáveis de envolvimento não compreenderiam o mesmo espaço temporal que as dos restantes indivíduos.

Tratando-se de dados reais, é natural que estas variáveis não apresentem uma distribuição normal e, portanto, recorreu-se à sua modificação utilizando a função mais eficaz para cada uma delas especificamente. Na maior parte das variáveis, o logaritmo natural e a raíz quadrada foram as funções que se mostraram mais adequadas para a sua normalização. Estas transformações não só normalizam as variáveis como aumentaram também a correlação entre estas e a variável *target*. Adicionalmente, fica atenuado o efeito dos *outliers*, definidos como valores anómalos sem expurgar estas observações do modelo.

Para além da análise descritiva, verificou-se a correlação de cada uma das variáveis explicativas com a *target* através da regressão linear simples. De uma forma geral, as variáveis mostraram fraco poder explicativo quando relacionadas individualmente com a *target*. Porém, a variável de cariz pessoal x_7 e as variáveis de envolvimento x_2, x_3 e x_{12} destacaram-se por uma correlação com a *target* mais forte que as restantes.

Os dados que serviram de base para a construção do modelo estão actualizados à data de início da campanha e as variáveis de envolvimento referem-se a valores compreendidos entre Novembro de 2006 e Abril de 2007, existindo um mês entre o último mês considerado e o início da campanha.

5.1.2 Partição dos Dados

Como é comum nas campanhas de marketing, a taxa de resposta da campanha em estudo foi muito baixa, sendo inferior a 10%, o que torna a questão da representatividade crucial para o sucesso do modelo.

A repartição dos dados em três conjuntos (treino, validação e testes) era demasiado exaustiva quando a amostra da classe dos indivíduos que se pretende identificar é reduzida pelo que se optou por prescindir do conjunto de teste.

Os conjuntos de treino e de validação possuem uma dimensão de 60% e 40%, respectivamente e, através do método da amostragem estratificada[7], garantiu-se que, dentro de

[7] Ver capítulo 4, p. 84-108, em "Sampling Methods for Applied Research – Text and Cases" de Tryfos, P. (1996).

cada um deles, as proporções de clientes que aderiram à campanha é igual a 30%. No entanto, todos os resultados que serão apresentados nas próximas secções e que constam em tabelas e gráficos já se encontram corrigidos do *oversampling*, isto é, estão dimensionados às proporções da população em estudo.

5.2 Aplicação das técnicas de modelação

Antes da abordagem às técnicas utilizadas, é importante esclarecer que foram realizados três testes diferentes: no primeiro teste, foram utilizadas as variáveis sem qualquer transformação funcional e sem aplicação de qualquer método de selecção de variáveis; relativamente a este, no segundo teste realizou-se a transformação das variáveis conforme referido no ponto 4.4. e, por fim, o terceiro teste acrescenta um método de selecção das variáveis com base no teste do R^2[8] que, no essencial, expurga as variáveis que não mostrem correlação com a *target*.

5.2.1 Redes Neuronais

O número de nós, o número de camadas escondidas, o algoritmo e outro tipo de parametrizações necessárias na criação de uma rede neuronal artificial são algumas das variáveis a considerar nesse processo pelo que não existe uma receita para obter os melhores resultados seja para que tipo de problema for. Desta forma, resta experimentar várias combinações possíveis para começar a observar com quais o modelo de previsão mostra melhor performance.

Em cada teste, numa fase preliminar, realizaram-se várias combinações diferentes do número de neurónios por camada *versus* número de camadas intermédias. As redes neuronais mostraram ter melhor desempenho com um menor número de camadas intermédias tendo-se optado, na maioria dos casos, por apenas uma camada. Quanto ao número de neurónios, os resultados revelaram maior eficiência com dois neurónios.

Relativamente ao comportamento do algoritmo de Retropropagação e às suas variantes descritas na secção 2.2.6., os modelos resultantes do algoritmo Quasi-Newton e do gradiente conjugado apresentaram melhor performance com base no menor erro e na menor taxa de

[8] Método de selecção utilizado no Enterprise Miner do SAS, mais concretamente, através da ferramenta *Selection Variable*.

classificações incorrectas sendo este resultado comum aos três testes conforme se pode constatar na tabela 13 em anexo.

Comparando os dois algoritmos e como se pode observar pelos gráficos representados nas figuras 12, 13, 14 e 15, em anexo, onde se consegue visualizar a trajectória do erro médio *versus* número de iterações ocorridas, o algoritmo do gradiente conjugado converge mais rapidamente que o algoritmo Quasi-Newton. No entanto, quando comparado o erro do conjunto de validação, a melhor performance reside nos modelos provenientes do teste onde há uma prévia selecção de variáveis, principalmente, a exibida pelo algoritmo de Quasi-Newton. De realçar também que o modelo resultante deste algoritmo também mostra uma menor percentagem de classificações incorrectas. Note-se, ainda, que na figura 14 o gradiente conjugado mostra-se, praticamente, constante de iteração para iteração.

5.2.2 Regressão Logística

Como um dos pressupostos da regressão logística é a distribuição normal das variáveis explicativas, esta técnica apenas foi utilizada nos testes onde ocorreu a normalização das variáveis.

Tanto no teste em que ocorreu expurgo de variáveis como naquele em que todas as variáveis iniciais se mantiveram, o valor das estatísticas do teste da razão da verosimilhança da tabela 4 implica a rejeição da hipótese nula mostrando que ambos os modelos *logit* são modelos globalmente válidos e, portanto, as variáveis explicativas consideradas são conjuntamente relevantes. Em contrapartida, estes modelos possuem algum poder explicativo, situando-se os coeficientes de regressão (na tabela 4) em 15,5% e 18,2% em que o mais baixo se verifica no modelo em que foram expurgadas algumas das variáveis explicativas.

		Modelo 1		Modelo 4	
Teste	**Estatística Teste**	**Valor Estatística**	**Graus Liberdade**	**Valor Estatística**	**Graus Liberdade**
Teste Razão Verosimilhança	$-2\left(\ln \hat{L}_r - \ln \hat{L}\right)$	865,125	33	738,457	21
R^2 **McFadden**	$1 - L_\beta / L_0$	0,181833563		0,15521025	

Tabela 4 – Estatísticas de ajustamento e da capacidade explicativa dos modelos *logit*

Nas tabelas 5 e 6, podem-se observar os valores obtidos na construção do teste de Hosmer-Lemeshow, os quais mostram que os modelos obtidos revelam um ajustamento aos dados pobre. Perante o valor da estatística de teste, rejeita-se a hipótese nula a um nível de significância de 5%, concluindo-se que os valores observados são significativamente diferentes dos valores esperados. Este resultado não é completamente inesperado. Dadas as dimensões das classes da *target* na amostra inicial, será normal que as probabilidades dos clientes que respondem afirmativamente à campanha sejam baixas.

Teste - Variáveis Normalizadas - Modelo 1					
	Respostas afirmativas		Respostas negativas		G_{HL}
	Observados	Esperados	Observados	Esperados	
1	276	60,1350	113,6	329,4650	916,3198
2	211	25,2699	178,6	364,3301	1459,7730
3	177	16,5142	212,6	373,0858	1628,6463
4	150	11,5853	239,6	378,0147	1704,3921
5	118	8,3133	271,6	381,2867	1478,7738
6	83	6,1773	306,6	383,4227	970,7839
7	74	4,5480	315,6	385,0520	1073,1263
8	40	3,1798	349,6	386,4202	429,8573
9	24	2,0324	365,6	387,5676	238,6896
10	16	1,0277	373,6	388,5723	218,6941
Estatística de Hosmer - Lemeshow					**10 119,0564**

Tabela 5 – Tabela de teste de Hosmer-Lemeshow com diferenças entre valores observados e previstos em cada decil no modelo 1

Teste - Variáveis Normalizadas Relevantes - Modelo 4					
	Respostas afirmativas		Respostas negativas		G_{HL}
	Observados	Esperados	Observados	Esperados	
1	265	52,5908	124,6	337,0092	991,7771
2	198	23,2232	191,6	366,3768	1398,7351
3	174	15,5981	215,6	374,0019	1675,6898
4	138	11,2510	251,6	378,3490	1470,3568
5	126	8,5110	263,6	381,0890	1658,0779
6	98	6,5470	291,6	383,0530	1299,3062
7	71	4,9255	318,6	384,6745	897,7154
8	55	3,6187	334,6	385,9813	736,3987
9	25	2,4625	364,6	387,1375	207,5818
10	19	1,3398	370,6	388,2602	233,5798
Estatística de Hosmer - Lemeshow					**10 569,2187**

Tabela 6 - Tabela de teste de Hosmer-Lemeshow com diferenças entre valores observados e previstos em cada decil no modelo 4

No que respeita à relevância de cada variável individualmente, temos dois casos distintos cujos resultados do teste de Wald são visíveis nas tabelas 7 e 8. No modelo onde foram

excluídas algumas variáveis através do teste R^2 do SAS[9], ou seja, onde foi apurado previamente quais as variáveis estatisticamente significantes, os resultados da tabela 7 mostram (como era expectável) que todas elas eram relevantes individualmente a um nível de significância de 5%. Já no modelo constituído por todas as variáveis propostas inicialmente, ao mesmo nível de significância, a estatística de Wald da tabela 8 conduz à não rejeição da hipótese de que os coeficiente das variáveis x_1, x_6, x_8, x_{13} e x_{19} são nulos. Com excepção da última, todas as outras variáveis são referentes à actividade do indivíduo como cliente do banco e note-se que, em ambos os modelos, as variáveis que mostraram uma correlação mais forte com a *target* na análise exploratória apresentam-se como sendo estatisticamente significantes pelo teste da significância individual.

Em suma, os modelos estão em consonância sobre a relevância das variáveis explicativas, relacionadas com o envolvimento do cliente, que lhes são comuns. Os factores de cariz pessoal ganham mais importância no modelo cujos resultados constam da tabela 8.

Effect	DF	Wald Chi-Square	Pr > Chi-Square
X3	1	116.4647	<.0001
X16	3	31.1531	<.0001
X7	1	73.7559	<.0001
X12	12	178.5468	<.0001
X4	1	17.8276	<.0001
X10	1	5.7989	0.0160
X15	1	244.7861	<.0001
X2	1	46.0511	<.0001

Tabela 7 – Teste de Wald aplicado às variáveis do modelo *logit* com expurgo de variáveis

Effect	DF	Wald Chi-Square	Pr > Chi-Square
X1	1	3.5705	0.0588
X2	1	50.2637	<.0001
X3	1	132.3143	<.0001
X14	1	4.9622	0.0259
X16	3	34.2606	<.0001
X4	1	15.0882	0.0001
X5	1	4.8541	0.0276
X6	1	0.0150	0.9025
X7	1	74.3003	<.0001
X8	1	0.7215	0.3956
X10	1	10.8544	0.0010
X9	1	4.9611	0.0259
X12	12	187.8961	<.0001
X11	2	31.2636	<.0001
X17	1	13.0653	0.0003
X13	1	0.0610	0.8049
X15	1	95.7458	<.0001
X18	1	37.2381	<.0001
X19	1	3.5270	0.0604
X20	1	9.0350	0.0026

Tabela 8 – Teste de Wald aplicado às variáveis do modelo *logit* sem expurgo de variáveis

[9] Teste disponível na ferramenta do Enterprise Miner do SAS: "Variable Selection".

5.2.3 O Melhor Modelo

Perante os três testes elaborados no SAS, aplicaram-se as mesmas técnicas estatísticas com as mesmas parametrizações (número de camadas intermédias, nós e outras) e compararam-se os resultados dos vários modelos.

Dado o número elevado de modelos, aplicou-se como primeiros critérios de selecção os que registavam os menores *RMSE* e menor *Missclassification Rate* resultando num subconjunto de seis modelos cujas variáveis se encontram normalizadas como se pode verificar na tabela 9 em anexo. Definam-se esses modelos como na tabela abaixo:

Designação do Modelo	Técnica Modelação	Características Específicas	Teste
Modelo 1	*logit*		Com Normalização de Variáveis
Modelo 2	Rede neuronal com algoritmo do Gradiente Conjugado	- 1 camada escondida - 2 nós	Com Normalização de Variáveis
Modelo 3	Rede neuronal com algoritmo Quasi-Newton	- 1 camada escondida - 2 nós	Com Normalização de Variáveis
Modelo 4	*logit*		Com normalização e selecção de variáveis.
Modelo 5	Rede neuronal com algoritmo do Gradiente Conjugado	- 1 camada escondida - 2 nós	Com normalização e selecção de variáveis.
Modelo 6	Rede neuronal com algoritmo Quasi-Newton	- 1 camada escondida - 2 nós	Com normalização e selecção de variáveis.

Tabela 9 – Descrição dos modelos em análise

A partir deste momento, a análise recairá apenas sobre estes modelos comparando a sua bondade de ajustamento e capacidade preditiva.

De acordo com os valores da medida AIC (tabela 10), os modelos resultantes da utilização do algoritmo baseado no gradiente conjugado e o *logit* em que não ocorreu expurgo de variáveis mostram uma menor complexidade que os restantes. Em contrapartida, o critério Schwarz's Bayesian (tabela 8) indica os modelos *logit* como os que apresentam melhor ajustamento.

Teste	Diagrama	Modelo	Akaike's Information Criterion	Schwarz's Bayesian Criterion
Var. Normalizadas	Logit	Modelo 1	**3961**	**4168**
Var. Normalizadas	N_CG_1_2	Modelo 2	**3980**	4437
Var. Normalizadas	N_QN_1_2	Modelo 3	4017	4475
Var. Norm. Relevantes	Logit	Modelo 4	4066	**4204**
Var. Norm. Relevantes	N_CG_1_2	Modelo 5	3987	4444
Var. Norm. Relevantes	N_QN_1_2	Modelo 6	4066	4361

Tabela 10 – Valores das estatísticas AIC e SBC

A variável *target* Y é binária tomando o 1, se o cliente é propenso ao Crédito Pessoal ou 0, caso contrário. Este resultado é obtido após prever qual a probabilidade de o cliente aderir ao produto. Para atingir esse resultado a questão para a qual foi necessário obter uma resposta foi sobre qual o valor de *threshold* a considerar.

Terza (2006) mostrou que, para os modelos de regressão de resposta binária, a escolha do valor de *threshold* a partir do qual se define que a variável de interesse toma valor 1 ou 0 não deve ser arbitrária mas que, sob um critério de minimização do erro quadrático médio (de previsão) deverá ter o valor óptimo de ½. No entanto, o autor ressalva que a existência de informação que não está contida na amostra e, de certa forma, alteraria os resultados da previsão poderá influenciar os custos da previsão dos falsos zeros *versus* a dos falsos uns.

Neste caso em concreto, embora o SAS aplique o mesmo limite defendido por este autor tanto nas redes neuronais como na regressão logística é importante avaliar qual o *threshold* a aplicar quando seleccionado o alvo para a próxima campanha com base no modelo escolhido. Neste contexto e com base na advertência de Terza, foram analisados os possíveis valores para este ponto de *cutoff* sempre que analisados os vários indicadores de qualidade dos modelos.

Pela representação gráfica na figura 8, os modelos que prevêem maior taxa de resposta no 1º decil são o 3, 5 e o 6, enquanto que no 2º e 3º decis, os modelos 2 e 3 conseguem melhores previsões que os restantes. No 2º decil, verifica-se ainda que o modelo 6 permanece com um *lift* considerável. Sob a hipótese da decisão recair sobre o envio de *mailings* aos 20% indivíduos da amostra que possuem um *score* mais elevado, tanto o modelo 2 como 3 seriam candidatos para o melhor modelo, o que traduziria na previsão de 2,63 vezes mais respostas positivas do que num modelo aleatório.

Claramente, os modelos resultantes da aplicação das redes neuronais exibem maior capacidade preditiva nos primeiros decis do que a regressão logística. Ademais, a partir do 3º decil a queda do *lift* torna-se mais acentuada e as diferenças entre os modelos são pouco significativas.

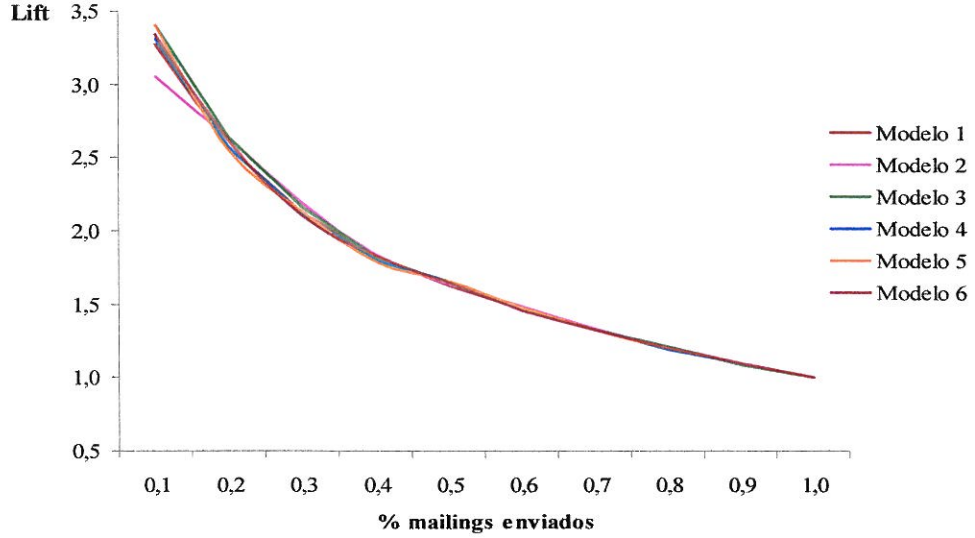

Figura 8 – Representação gráfica do Lift acumulado em cada decil

Todavia, para além da perspectiva de qual a proporção de *mailings* a enviar, o *lift* reflecte outra forma de determinar quais os clientes a seleccionar: definir o *threshold* a partir do qual os indivíduos devem ser integrados no alvo final através dos *scores* mais elevados registados nos modelos eleitos.

Perante as conclusões resultantes da figura 8, as escolhas no 1º decil direccionam para um *score* entre 0,067 e 0,081 enquanto que nos 2º e 3º decis a probabilidade limite estaria próxima de 0,05 e 0,04, respectivamente.

O objectivo deste tipo de modelos assenta sempre na previsão de respostas positivas mais elevadas do que na hipótese da não utilização de quaisquer modelos. Porém, a capacidade preditiva também é avaliada pela medição dos erros de previsão e estes são essenciais para a determinação do *threshold*. Ao analisar as matrizes de confusão, constata-se que as probabilidades de ocorrência do erro do tipo I são elevadas, sendo, no entanto, menos críticas se considerado um *threshold* igual a 10%. Ambos os modelo 1 e 2 (*logit* e o algoritmo gradiente conjugado) apresentam um menor erro nos 10%. Para um erro de 20%, o modelo 1 continua a possuir o melhor erro, competindo com o modelo 5.

Esta decisão poderia ser tão linear como aparenta, não se desse o caso de que, numa perspectiva de negócio, para além de se ponderarem os riscos de erro, é determinante a avaliação do lucro de acordo com as previsões obtidas por cada modelo e dos custos associados mediante a quantidade de *mailings* a enviar. Numa perspectiva financeira, a análise que se segue terá duas abordagens: os custos associados aos erros de previsão e o cenário custo/proveito tendo em conta a capacidade preditiva de cada modelo.

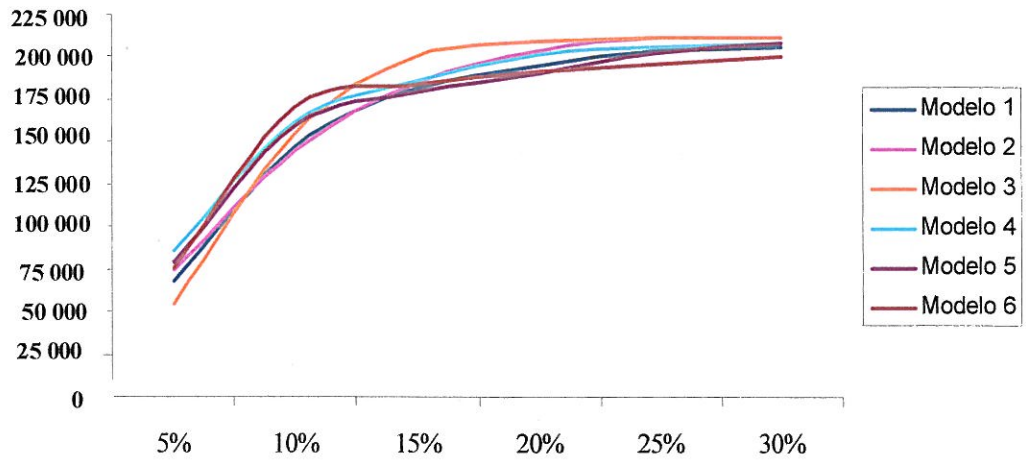

Figura 9 – Custos associados aos erros de previsão (tipo I + tipo II)

Através da representação gráfica da figura 9 construído a partir das matrizes de confusão com inserção de custos, constata-se que do *threshold* de 5% para 10% existe um aumento acentuado dos custos começando este valor a estabilizar a partir dos 15%. No limiar dos 5% os modelos 1 e 3 são os menos dispendiosos em caso de erro de previsão; nos 10% surge mais um candidato, o modelo 2.

Com base na população em estudo, a decisão pelo melhor modelo seria difícil se optasse por enviar *mailings* para metade dos indivíduos, uma vez que os proveitos são muito semelhantes neste decil como mostra a figura 9. Porém, se esta proporção fosse reduzida para 30% os proveitos seriam maiores para alguns modelos sendo mais vantajoso o modelo 2. Analisando o 2º decil, o lucro seria um pouco inferior sendo que os modelos 2 e 3 são os que registam valores mais elevados. No 1º decil, se a probabilidade de erro na classificação dos clientes é menor, os proveitos também o são. Adicionalmente, deve referir-se que da análise aos modelos não é evidente grande vantagem de uns em relação a outros. De qualquer forma, os modelos 3 e 5 são os que registam maior margem.

Através da figura 10, constata-se ainda que a opção por um outro modelo (com pior performance) já representaria valor acrescentado pois as diferenças entre os proveitos da previsão destes modelos e o modelo aleatório são positivas e bastante significativas.

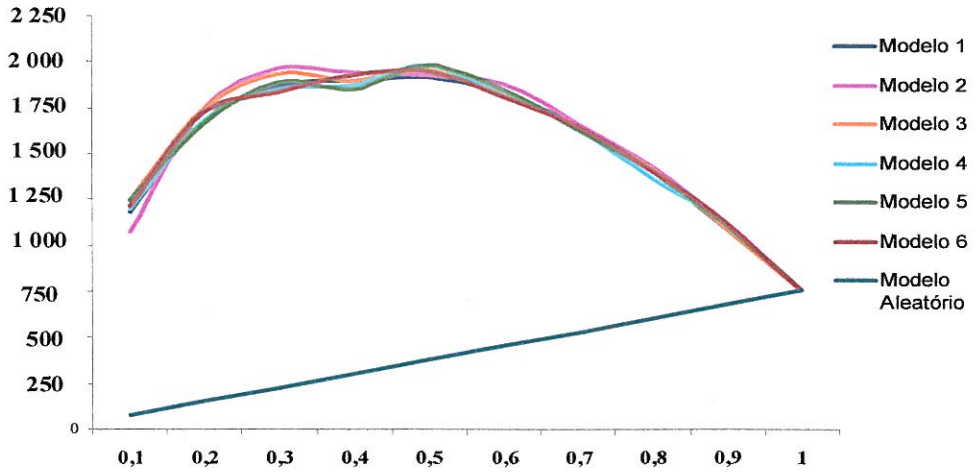

Figura 10 – Apresentação dos proveitos de acordo com a capacidade preditiva de cada modelo *versus* a % de mailigs enviados

A comparação através da curva ROC elege como melhor modelo o que apresentar maior área abaixo da respectiva curva. Observando o gráfico 4 onde se encontram delineadas as curvas ROC respeitantes a cada um dos modelos, conclui-se com considerável evidência que o modelo 3 é o que apresenta maior área distinguindo-se dos restantes. Seguindo a sua trajectória, evidencia-se um *tradeoff* bastante razoável entre a sensibilidade (% de 1's correctamente classificados) com um valor de 0,5 aproximadamente e um valor elevado da especificidade (% de 0's correctamente classificados). Os restantes modelos não mostram grandes diferenças entre si.

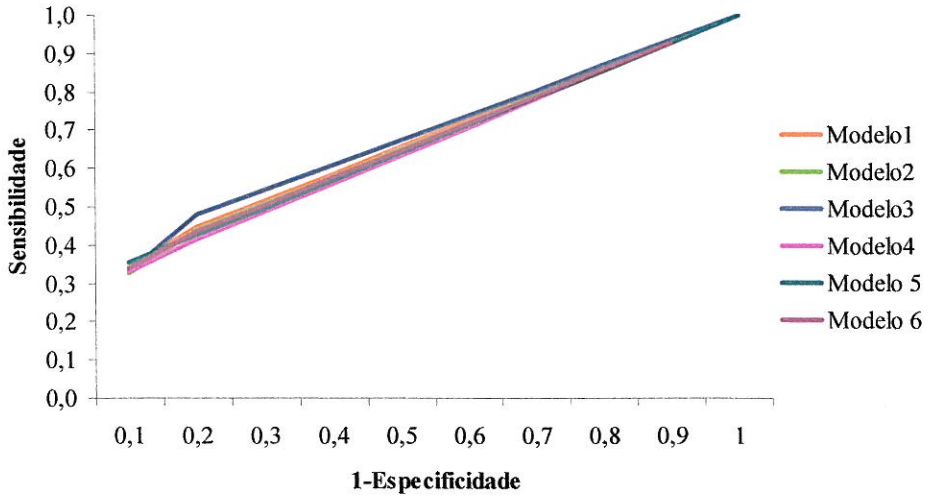

Figura 11 – Curva ROC dos modelos em análise

Todavia, como já mencionado anteriormente, o objectivo não consiste apenas na selecção do modelo com melhor performance mas na ponderação de um *threshold* considerado óptimo ou que represente mais valia em termos de proveitos e de taxa de erro.

Neste contexto, foram testados vários valores para este parâmetro, desde 5% a 35% determinando a medida F que mede a *recall versus* a *precision* apresentada nas tabelas 11 e 12. A *recall*, ou seja, a proporção de 1's correctamente classificados apresenta percentagens razoáveis para os *threshold* iguais a 5% ou 10%. Acima destes valores, as percentagens decaem significativamente. No que respeita à *precision* e à sensibilidade, as percentagens são baixas e aumentam consideravelmente a partir do limite igual a 20%. Estes dois indicadores ponderados na fórmula da medida F, indicam que a melhor combinação é obtida para um *threshold* de 0,1. Este facto é comum a todos os modelos em análise, sobressaindo mais nos modelos 1 e 5. Note-se que para o valor 0,05 os modelos concorrentes são os modelos resultantes do algoritmo do gradiente conjugado, ou seja, os modelos 2 e 5.

No que concerne à especificidade, este indicador apresenta proporções elevadas para qualquer que seja o valor de *threshold*, o que é expectável para uma amostra desproporcional relativamente às dimensões das possíveis classes da *target*.

	Modelo 1			Modelo 2			Modelo 3		
	5%	10%	15%	5%	10%	15%	5%	10%	15%
Sensibilidade/ Recall	38.126%	28.777%	7.574%	38.126%	30.117%	6.290%	43.774%	23.713%	3.338%
Especificidade	88.834%	96.645%	98.791%	87.899%	96.590%	99.010%	85.919%	97.250%	99.340%
Erro Tipo I	61.874%	71.223%	92.426%	61.874%	69.883%	93.710%	56.226%	76.287%	96.662%
Erro Tipo II	11.166%	3.355%	1.209%	12.101%	3.410%	0.990%	14.081%	2.750%	0.660%
Accuracy	87.634%	94.720%	96.706%	86.721%	94.690%	96.814%	84.922%	95.270%	97.068%
Precision	7.648%	10.019%	13.179%	7.099%	10.463%	13.351%	7.011%	10.714%	10.924%
F-value	0.1273984	**0.14863037**	0.096193684	0.1196882	**0.1553004**	0.0855137	0.1208573	**0.1475917**	0.0511305

Tabela 11 – Valores da medida F para *thresholds* de 5%, 10% e 15% testados nos modelos onde não há expurgo de variáveis

	Modelo 4			Modelo 5			Modelo 6		
	5%	10%	15%	5%	10%	15%	5%	10%	15%
Sensibilidade/ Recall	34.531%	27.101%	5.905%	35.173%	27.128%	7.574%	37.099%	21.254%	6.162%
Especificidade	89.439%	97.140%	99.230%	90.319%	97.360%	99.010%	88.834%	98.130%	99.505%
Erro Tipo I	65.469%	72.899%	94.095%	64.827%	72.872%	92.426%	62.901%	78.746%	93.838%
Erro Tipo II	10.561%	2.860%	0.770%	9.681%	2.640%	0.990%	11.166%	1.870%	0.495%
Accuracy	88.138%	95.144%	97.021%	89.012%	95.366%	96.845%	87.609%	96.044%	97.296%
Precision	7.347%	9.816%	15.681%	8.098%	10.736%	15.649%	7.457%	11.492%	23.187%
F-value	0.121162	**0.14412029**	0.085793874	0.1316505	**0.1538398**	0.1020746	0.1241774	**0.1491816**	0.0973611

Tabela 12 – Valores da medida F para *thresholds* de 5%, 10% e 15% testados nos modelos em que se consideram apenas as variáveis relevantes

Em todas as medidas de qualidade dos modelos interpretadas neste capítulo consegue-se extrair uma conclusão comum, do teste onde não houve expurgo de variáveis sendorelevantes, ou não, para a explicação da propensão dos clientes ao Crédito Pessoal proveio modelos com melhor performance.

No que concerne à bondade de ajustamento, os modelos *logit* e o algoritmo do gradiente conjugado evidenciaram uma melhor classificação dos dados, facto confirmado pelas matrizes de confusão e pela métrica F.

A capacidade preditiva ponderada pelos lucros inerentes aos diferentes modelos mostrou-se mais apurada nos modelos resultantes das redes neuronais através dos algoritmos do gradiente conjugado e o Quasi-Newton, em particular, os modelos 2 e 3.

A nomeação de um modelo como modelo "vencedor" não é fácil e depende das prioridades das instituições, se se centram na probabilidade do menor erro de classificação possível ou no maior lucro independentemente das classificações incorrectas. Neste contexto, apresenta-se o modelo 2 com menor probabilidade de ocorrência do erro do tipo I e o modelo 3 com maior lucro quer se decida pelo envio de *mailings* a 10% ou 20% dos clientes que constituem a base. Como as diferenças entre estes dois concorrentes não são significativas, qualquer um poderá ser considerado como um bom modelo de previsão.

6 Conclusões

Neste capítulo, em primeiro lugar serão enumeradas as conclusões mais importantes determinando em que medida os objectivos deste trabalho foram atingidos, quais as dificuldades encontradas e como se enquadraram os métodos de previsão escolhidos.
Numa perspectiva futura, serão ainda abordados os diferentes pontos de orientação que poderão tornar este trabalho mais enriquecedor em termos de diversidade de técnicas de previsão e de complementaridade do conhecimento extraído sobre o perfil de clientes, permitindo melhorar a precisão e a formulação do modelo assim como as condições específicas do produto a propor aos clientes que se mostram mais propensos.

6.1 Trabalho realizado

A primeira conclusão que se retira e que confirma as mais variadas teorias estatísticas consiste na obtenção de melhores resultados com a normalização das variáveis, até na aplicação de técnicas como as redes neuronais que não necessitam de informação prévia sobre a distribuição das variáveis.
Em segundo lugar, os modelos que registam menores erros e, simultaneamente fornecem melhores previsões, são claramente os que englobam todas as variáveis inicialmente propostas, comparativamente aos modelos em que houve o interesse de expurgar as variáveis que possuíam (individualmente) menor poder explicativo, independentemente da técnica de modelação a aplicar. Torna-se importante salientar que, neste caso concreto, não existia um número muito elevado de variáveis a considerar pois perante centenas de variáveis esta conclusão provavelmente não seria verosímil.
À medida que este projecto era realizado, dois desafios completamente diferentes iam-se evidenciando: comparar duas técnicas de naturezas distintas (as redes neuronais e a regressão logística) e trabalhar uma base de dados reais sobre a qual se sabe *a priori* que os resultados nem sempre são os esperados ou até conclusivos.
No primeiro desafio, as várias estatísticas e gráficos construídos mostraram quase sempre que os resultados provenientes dos modelos obtidos pelas redes neuronais eram melhores que os apresentados pelo modelo *logit*. A decisão recai nos modelos 2 ou 3, derivados dos algoritmos

gradiente conjugado e o Quasi-Newton e por um *threshold* entre 5% e 10%, confirmando que o envio de *mailings* para um universo compreendido no intervalo entre 10% e 20% da amostra inicial de clientes terá maior lucro do que se se optasse por enviar para 100% dos clientes. A fixação do *threshold* será sempre uma decisão ponderada entre a dimensão dos erros de previsão e os proveitos que poderão advir da capacidade preditiva do modelo.

Em termos práticos, para além do aspecto monetário, esta seria uma forma de maior e melhor aproveitamento da base de clientes alvo para campanhas, permitindo à instituição aproveitar os clientes não seleccionados para outro tipo de campanhas pois estes clientes poderão encontrar-se nesse momento mais propensos para outro tipo de produto. Por outro lado, se um departamento *responsável por* um determinado produto recorrer frequentemente a este tipo de campanhas para aumentar a sua carteira não verá a dimensão do seu alvo reduzir tão rapidamente e, no limite, evitará uma situação de escasseio de alvo.

Em termos estatísticos e computacionais, conclui-se que a grande vantagem das redes neuronais reside no facto de não necessitarem de pressupostos sobre a distribuição das variáveis e dos resíduos, possuindo uma grande rapidez de processamento. Esta última propriedade não se mostrou muito evidente neste trabalho devido à recorrência das técnicas de *oversampling,* dado que a base sobre a qual se actuou na maior parte do tempo era bastante mais reduzida do que a inicial. Outro aspecto importante e que deve ser realçado, é o facto da regressão logística possuir propriedades estatísticas particulares existindo, portanto, a necessidade de recorrer a testes adicionais e específicos para uma melhor análise da sua performance. Por um lado, este facto constitui um constrangimento quando se pretende comparar resultados entre vários modelos que sejam resultado de outra técnica de modelação como as redes neuronais, por outro, alguns testes poderão mostrar-se desajustados ao tipo de amostra em análise como é exemplo o teste de Hosmer-Lemeshow. Este teste não se adequa a amostras que apresentam dimensões nas classes de domínio da *target* desequilibradas uma vez que as probabilidades de evento (raro) previstas são sempre muito baixas e as distâncias entre estas e os valores esperados revelam-se grandes tal como se observou no caso em estudo.

Os dados utilizados como dados reais apresentavam uma consistência e integridade razoáveis tendo exigido, no entanto, uma exploração atenta e algumas transformações, especialmente, de natureza funcional. Conforme a análise realizada, os resultados foram conclusivos e indiciaram um possível aproveitamento para optimização dos alvos das campanhas.

Ao nível do conhecimento dos clientes propensos ao Crédito Pessoal, este trabalho mostrou com evidência a relevância do comportamento do indivíduo enquanto cliente. Tal como na bibliografia apresentada, fica o entendimento de que as necessidades e o perfil do indivíduo são o trilho que deve ser seguido para determinar o produto que melhor se adequa.

Torna-se importante lembrar que cada caso é um caso e que, apesar de neste alvo específico de clientes, as redes neuronais apresentarem melhor performance, tal não deve ser entendido como uma realidade extrapolável a outros casos, pois é perfeitamente admissível que para um alvo diferente ou numa campanha de outra natureza, técnicas diferentes se mostrem mais adequadas.

Importa ainda referir que, embora se eleja no presente trabalho os modelos resultantes das redes neuronais e que não excluem nenhuma das variáveis inicialmente propostas, a capacidade preditiva dos restantes não apresenta diferenças suficientemente relevantes para que se possa optar pela sua exclusão, de forma absoluta.

Na realização da tese, foram encontradas algumas dificuldades como a escassa bibliografia existente sobre modelos de propensão e a constante tentativa de convergência do *software* aos objectivos propostos.

6.2 Trabalho futuro

Atingida a fase final deste trabalho, torna-se necessária uma reflexão ao trabalho que foi realizado na perspectiva da criação de uma base de conhecimento, literatura e resultados, que poderão suportar trabalhos futuros em torno do tema da propensão.

Em primeira instância, a diversidade de técnicas de previsão a aplicar aos problemas da propensão poderão desenvolver um maior conhecimento sobre as suas diferentes performances assim como à condução a modelações mais eficientes e precisas. A maioria dos artigos consultados, atribuíam grande proeminência às redes neuronais e à regressão logística. No entanto, alguns complementavam com o sucesso de outra técnica de *data mining* bastante utilizada em tarefas de classificação – as árvores de decisão – como são exemplo os trabalhos desenvolvidos por Andreeva e Crook (2005) e Tam e Kiang, 1992, citados por Hsieh, 2005.

"Uma das principais vantagens das árvores de decisão é o facto de o modelo ser consideravelmente explicável desde que este toma a forma de regras explícitas. Isto permite que se analise os resultados identificando os atributos chave no processo. Esta técnica também

é útil quando os dados de *input* são de qualidade incerta – resultados espúrios tornam-se óbvios em regras explícitas" (Berry e Linoff, 1997, p. 122).

Com o objectivo de aumentar a capacidade preditiva dos modelos suportados quer em redes neuronais quer em árvores de decisão ou ainda na regressão logística, estes modelos poderão ser combinados por forma a originar um só modelo aproveitando o conhecimento captado por cada um individualmente, aumentando a precisão. Este tipo de modelos é conhecido por *ensemble models* ou previsões por *ensemble*.

Por outro lado, e saindo do raio de acção dos problemas de propensão, nos últimos meses a nossa sociedade foi assolada por uma crise financeira e económica que tem levado à alteração de uma série de postulados com que nos havíamos acostumado a viver. Entre outras consequências, estes fenómenos estão a suscitar a alteração do enfoque das áreas de produto e risco das instituições financeiras, com a consequente criação de novos problemas diariamente. Nesse sentido, fará todo o sentido que, num trabalho a iniciar nesta fase, se complemente o estudo efectuado, com a introdução de mecanismos que permitam atribuir a cada um dos clientes considerados como propensos, valores para a taxa de juro e montante de financiamento, que possam convergir para as suas expectativas e que, portanto, aumentem a sua propensão para o consumo deste produto, garantindo, naturalmente, os pressupostos de rentabilidade e risco impostos pela entidade financiadora.

Bibliografia

Andreeva, G., Ansell, J. & Crook, JN (2005). Modelling the purchase propensity: analysis of a revolving store card. *Journal of the Operational Research Society*, 56, 1041-1050.

Amemiya, T. (1985). *Advanced Econometrics*. Harvard University Press.

Atiya, A. (2001). Bankruptcy Prediction for Credit Risk Using Neural Networks: A Survey and New Results. *IEEE Transactions on Neural Networks*, 4, pp 929-935.

Baesens, B., Van Gestel, T., Stepanova, M., Van den Poel, D. & Vanthienen (2005). Neural network survival analysis for personal loan data. *Journal of the Operational Research Society*, 56, 1089-1098.

Bem-Akiva, M. & Lerman, S.R. (1985). Discrete Choice Analysis: theory in application to travel demand. *The MIT Press*.

Berry, M.J.A. & Linoff, G. (1997). *Data Mining Techniques: for Marketing, Sales, and Customer Support*. New York and Toronto: John Wiley & Sons, Inc.

Berry, M.J.A. & Linoff, G. (2000). *Mastering Data Mining: The Art and Science of Customer Relationship Management.* New York and Toronto: John Wiley & Sons, Inc.

Bishop, C. M. (1997). *Neural Networks for Pattern Recognition.* New York: Oxford University Press

Bodkin, R., Hsiao, C., Intriligator, M. (1996). *Econometric Models, Techniques, and Applications* . Prentice Hall International Editions.

Braga, A. C. S. (2000). *Curvas ROC: Aspectos Funcionais e Aplicações.* Tese de Doutoramento em Engenharia de Produção e Sistemas. Escola de Engenharia - Universidade do Minho, Braga. 267 pp..

Chawla, N. V. (2005). Data Mining for Imbalanced Datasets: An Overview. *Data Mining and Knowledge Discovery Handbook* . Springer US 40, 853-867.

Cheng, B. & D.M. Titterington (1994). Neural networks: a review from a statistical perspective. *Statistical Science* 9, 2-54.

Chorão, L. A. (2005,). *Credit Scoring: Logit vs Redes Neuronais Artificiais – Um exemplo aplicado a cartões de crédito*, Dissertação de Mestrado, Instituto Superior de Estatística e Gestão de Informação/Universidade Nova de Lisboa.

Cortez, P. (1997). *Algoritmos Genéticos e Redes Neuronais na Previsão de Séries Temporais.* Tese de Mestrado em Informática. Universidade do Minho, Braga. 91 pp..

Cortez, P. & Neves J. (2000). *Redes Neuronais Artificiais.* Departamento de Informática, Escola de Engenharia da Universidade do Minho, p.p. 1-50.

Cramer, J. S. (1986). *Econometric Aplications of Maximum Likelihood Methods*. Cambridge: Cambridge University Press.

Cramer, J. S. (2003). *Logit Models from Economics and other Fields*. Cambridge: Cambridge University Press.

Cramer, J. S. (2004). Scoring bank loans that may go wrong: a case study. *Statistica Neerlandica*, Vol. 58, nº3, p.p.365-380.

D'Agostino Jr, R. B. & Rubin, D. B. (2000). Estimating and Using Propensity Scores with Partially Missing Data. *Journal of the American Statistical Association,* 95, 749-759.

Davidson, R. & MaCKinnon, J.G. (1984). Convenient Specification Tests for Logit and Probit Models. *Journal of Econometrics*, 25, pp. 241-262.

Davidson, R. & MaCKinnon, J.G. (1993). *Estimation and Inference in Econometric.* Oxford: Oxford University Press Inc..

Eliashberg, J. & Lilien, G.L. (1993). *Handbooks in operations research and management science*, Vol. 5, Marketing, Elsevier Science Publishers.

Fausett, L. (1994). *Fundamentals of Neural Networks Architectures, Algoritms and Applications.* Prentice-Hall, Inc..

Freeman, J: A. & Skapura, D.M. (1992). *Neural Networks: Algoritms and Programming Techniques.* Addison-Wesley Publishing Company.

Goss, E. P. & Ramchandani, H. (1998). Survival Prediction in the Intensive Care Unit: a Comparison of Neural Networks and Binary Logit Regression. *Socio-Economic Planning Science*, 32, pp 189-198.

Groth, R. (2000). *Data Mining – Building Competitive Advantage.* Prentice – Hall, Inc.

Gurney, K (2001). *An Introduction to Neural Networks*. London: UCL Press.

Han, J. & Kamber, M. (2001). *"Data Mining: Concepts and Techniques"*. EUA: Morgan Kaufmann.

Hanssens, M., Parsons, L. & Scultz, R. (1992). *Market response models:econometric and time series analysis.* Norwell: Kluwer Academic Publishers.

Haykin, S. (1994). *Neural Networks: a comprehensive foundation.* EUA: IEE Computer Society Press.

Horowitz, J. L. & Savin, N. E. (2001). Binary Response Models: Logits, Probits and Semiparametrics. *The Journal of Economic Perspectives*, Vol. 15, 4, 43-56.

Hsieh, N. -C. (2005). *Hybrid mining approach in the design of credit scoring models*. Expert Systems with Applications 28, pp 655-665.

Hsieh, N. -C. (2004). An integrated *data mining* and behavioral *scoring* model for analyzing bank customers. *Expert Systems with Applications*, 27, 623-633.

Hoetker, G. (2007). The use of Logit and Probit models in Strategic Management Research: Critical Issues. *Strategic Management Journal*, 28, 331-343.

Huang, Y.-M., Hung, C.-M. & Jiau, H. C. (2006). Evaluation of neural networks and *data mining* methods on a credit assessment task for class imbalance problem. *Nonlinear Analysis: Real World Applications*, 7, pp 720-747.

Klösgen, W. & Zytkow, J. (2002). *Handbook of Data Mining and Knowledge Discovery.* New York: Oxford University Press.

Jacobs, R.A. (1988). Increased Rates of Convergence Through Learning Rate Adaptation. Neural Networks, 1, 295-307.

Lilien, G.L., Kotler, P. & Moorthy, K. S. (1992). *Marketing Models*, Prentice-Hall International Editions.

Liu, Y. & Schumann, M. (2005). Data mining feature selection for selection for credit *scoring* models. *Journal of the Operational Research Society*, 56, 1099-1108.

Long J.S. (1997). *"Regression Models for Categorical and Limited Dependent Variables"*, Sage Publications, Inc.

Metz, C. E. (1986). Statistical Analysis of ROC Data in Evaluating Diagnostic Performance. *Multiple Regression Analysis: Applications in the Health Sciences*, 13. Donald E. Herbert & Raymond H. Myers. 365-384. American Institute of Physics.

Mitchell, T. (1997). *Machine Learning.* McGraw-Hill International Edit

Negnevitsky, M. (2002). *Artificial Intelligence: a guide to intelligent systems*. Local: Addison Wesley.

Piatetsky-Shapiro, G. & Steingold, S. (2000). Measuring Lift Quality in Database Marketing. *ACM SIGKDD (Special Interest Group on Knowledge Discovery and Data Mining) Explorations 2*, 2, 76-80.

Pires, P. (2007). Comparação de Variantes de Redes Neuronais Artificiais e dos Modelos Mixed Logit e Logit Multinomial na Aquisição de Produtos em Supermercados. Em: *Actas do Congresso Hispano-Lusas de Gestão Científica – Conocimiento, Innovación y emprendedores: Camino al futuro.*Cogrõno, Fevereiro de 2007, Universidade de la Rioja, 2099 – 2113.

Pyle, D. (1999). *Data Preparation for Data Mining.* San Francisco, CA: Morgan Kaufmann.

Roszbach (2003). Bank lending policy, credit *scoring* and the survival of loans. Sveriges Riskbank Working Paper Series, 154.

Santos, I. e Ramos, M. (2003, Outubro). Data Mining no suporte à construção de Conhecimento Organizacional. *Actas da 4ª Conferência da Associação Portuguesa de Sistemas de Informação*. Porto.

Smith, K. & Jatinder, G. (2002). Neural Networks in Business: Techniques and Application. New York: Idea Group Publishing.

Steingold, S., Wherry, R. & Piatetsky-Shapiro, G. (2001). Measuring Real-Time Predictive Models. *ICDM (IEEE International Conference on Data Mining),* 649-650.

Swingler, K. (2001). *Applying Neural Networks: a pratical guide*. Morgem Kaufman Publishers, Inc.

Terza, J. V. (2006). Optimal discrete prediction in parametric binary response models. *Economic Letters 91*, 72-75.

Trippi, R. R. & Turban, E. (1993). *Neural Networks in Finance and Investing: Using Artificial Intelligence to Improve Real-World Performance.* Chicago e Cambridge: Probus Publishing Company.

Tryfos, P. (1996). Sampling Methods for Applied Research – Text and Cases. New York: John Wiley & Sons, Inc.

Weiss, S. M. & Indurkhya, N. (1998). *Predictive Data Mining: A Pratical Guide.* San Francisco, CA: Morgan Kaufmann.

Westphal, C. & Blaxton, T. (1998). *Data Mining* Solutions: Methods and Tools for Solving Real-World Problems. USA: John Wiley & Sons, Inc.

Wielenga, D. (2007). Identifying and Overcoming Common Data Mining Mistakes. *SAS Global Forum 2007: Data Mining and Predictive Modeling*. SAS Institute Inc.

Wodon, Q. T. (1997). Targeting the Poor Using ROC Curves. *World Development,* Vol.25, Nº 12, 2083-2092. Great Britain: Elsevier Science.

Wooldridge, J. M. (2003). *Introductory Econometrics: a modern Approach , 2e.* Mason: Thomson, South-Western.

Zou, K. & Hall, W. (2000). Two Transformation models for estimating na ROC Curve derived from continuous data. *Journal of Applied Statistics*, Vol. 27, Nº 5, p.p. 621-631.

Zufryden, F. (1982). A General Model for Assessing New Product Marketing Decisions and Market Performance. Em: Machol (eds.), *Studies in the Management Sciences*, vol.18, p.p. 63-83, North-Holland Publishing Company. New York.

Anexos

Os valores que constam das tabelas e gráficos desta secção são meramente indicativos.

Fit Statistics	Modelo 1 - Logit - Var_Norm		Modelo 2 - N_CG_1_2 - Var_Norm		Modelo 3 - N_QN_1_2 - Var_Norm		Modelo 4 - Logit - Var_Norm_Selec		Modelo 5 - N_CG_1_2 - Var_Norm_Selec		Modelo 6 - N_QN_1_2 - Var_Norm_Selec	
	Training	Validation	Training	Validation	Training	Validation	Training	Validation	Training	Validation	Training	Validation
Akaike's Information Criterion	3.961	.	3.980	.	4.017	.	4.066	.	3.987	.	4.066	.
Average Error Function	0,499889987	0,534217642	0,49202693	0,532687531	0,496835308	0,535301877	0,516136644	0,533690647	0,492908306	0,532980233	0,509770288	0,530169154
Average Profit for CI	0,919079604	0,906120876	1,004561815	1,037113417	1,004678969	.	0,996342549	1,030969769	1,008612871	1,037113417	1,06732389	1,128236709
Average Squared Error	0,166016122	0,178577036	0,163044107	0,178780061	0,165095425	0,178822271	0,172304291	0,178942911	0,163333786	0,178761882	0,170130732	0,177920206
Degrees of Freedom for Error	3863	.	3823	.	3823	.	3874	.	3823	.	3849	.
Divisor for ASE	7792	5194	7792	5194	7792	5194	7792	5194	7792	5194	7792	5194
Error Function	3.895	2.775	3.834	2.767	3.871	2.780	4.022	2.772	3.841	2.768	3.972	2.754
Final Prediction Error	0,168852535	.	0,169270746	.	0,171400403	.	0,174261284	.	0,169571487	.	0,174285653	.
Maximum Absolute Error	0,969456508	0,993374058	0,974013047	0,970601136	0,972170387	0,972350445	0,977853654	0,9561778	0,974085287	0,972032271	0,990298163	0,970828337
Mean Square Error	0,167434328	0,178577036	0,166157427	0,178780061	0,168247914	0,178822271	0,173282788	0,178942911	0,166452637	0,178761882	0,172208192	0,177920206
Misclassification Rate	0,300051335	**0,299576434**	0,300051335	**0,299961494**	0,300051335	**0,299961494**	0,300051335	**0,299961494**	0,300051335	**0,299961494**	0,295431211	**0,296495957**
Model Degrees of	33	.	73	.	73	.	22	.	73	.	47	.
Root Average Sum of Squares	0,40745076	0,422583762	0,403787205	0,422823913	0,406319363	0,422873824	0,415095521	0,423016443	0,404145748	0,422802415	0,412469068	0,421805887
Root Final Prediction Error	0,410916701	.	0,411425262	.	0,414005318	.	0,417446145	.	0,411790587	.	0,417475332	.
Root Mean Squared Error	0	**0,422583762**	0,407624124	**0,422823913**	0,410180343	**0,422873824**	0,416272492	**0,423016443**	0,407986074	**0,422802415**	0,414979749	**0,421805887**
Schwarz's Bayesian Criterion	4.168	.	4.437	.	4.475	.	4.204	.	4.444	.	4.361	.
Sum of Case Weights Times Freq	7.792	5.194	7792	5194	7792	5194	7792	5194	7792	5194	7792	5194
Sum of Frequencies	3896	2597	3896	2597	3896	2597	3896	2597	3896	2597	3896	2597
Sum of Squared Errors	1.294	928	1.270	9.286	1.286	929	1.343	929	1.273	928	1.326	924
Total Degrees of Freedom	3896	.	3896	.	3896	.	3896	.	3896	.	3896	.

Tabela 13 – Comparação das estatísticas AIC e SBC dos modelos concorrentes

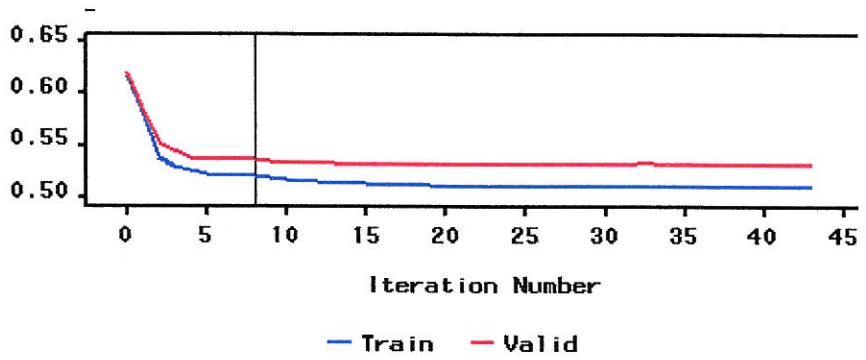

Figura 12 – Erro médio *versus* número de iterações na convergência do Gradiente Conjugado com expurgo de variáveis

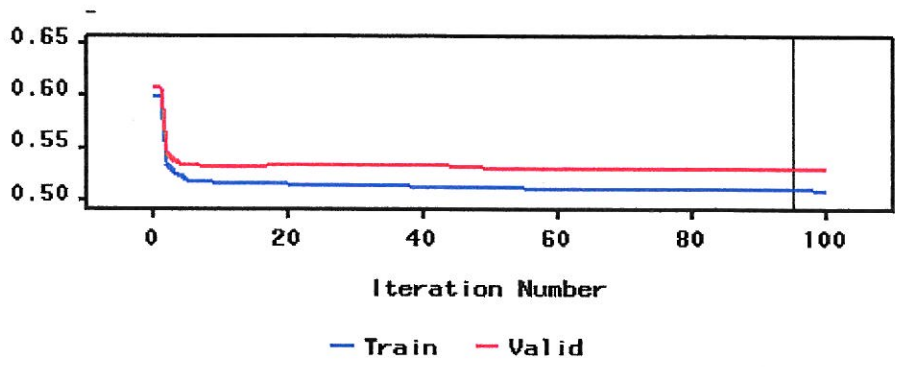

Figura 13 - Erro médio *versus* número de iterações na convergência do Quasi-Newton com expurgo de variáveis

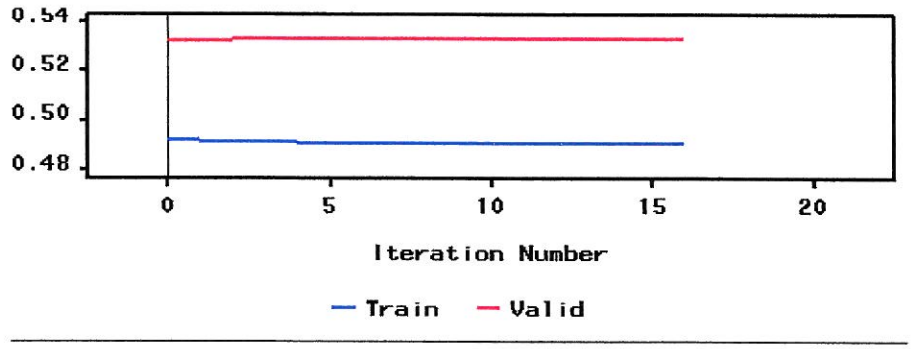

Figura 14 - Erro médio *versus* número de iterações na convergência do Gradiente Conjugado sem expurgo de variáveis

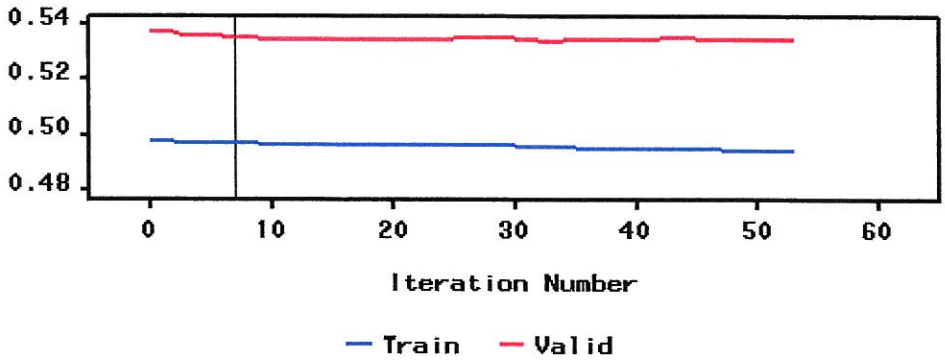

Figura 15 – Erro médio *versus* número de iterações na convergência do Quasi-Newton sem expurgo de variáveis

	Modelo 1		Modelo 2		Modelo 3		Modelo 4		Modelo 5		Modelo 6	
Decil	Lift Acumulado	Taxa Resposta Prevista p/ População	Lift Acumulado	Taxa Resposta Prevista p/ População	Lift Acumulado	Taxa Resposta Prevista p/ População	Lift Acumulado	Taxa Resposta Prevista p/ População	Lift Acumulado	Taxa Resposta Prevista p/ População	Lift Acumulado	Taxa Resposta Prevista p/ População
0,1	3,27	7,75%	3,06	7,23%	3,40	8,06%	3,31	7,84%	3,40	8,06%	3,34	7,90%
0,2	2,57	6,08%	2,63	6,23%	2,63	6,23%	2,57	6,08%	2,55	6,03%	2,61	6,19%
0,3	2,13	5,04%	2,19	5,18%	2,17	5,13%	2,11	5,00%	2,14	5,06%	2,10	4,97%
0,4	1,82	4,30%	1,84	4,36%	1,82	4,30%	1,80	4,27%	1,79	4,25%	1,83	4,34%
0,5	1,63	3,87%	1,64	3,87%	1,65	3,90%	1,66	3,92%	1,66	3,93%	1,65	3,90%
0,6	1,47	3,48%	1,49	3,52%	1,47	3,49%	1,47	3,47%	1,48	3,50%	1,46	3,46%
0,7	1,33	3,15%	1,33	3,16%	1,32	3,13%	1,33	3,14%	1,32	3,14%	1,33	3,14%
0,8	1,21	2,86%	1,21	2,87%	1,21	2,86%	1,20	2,83%	1,21	2,85%	1,21	2,86%
0,9	1,09	2,59%	1,10	2,61%	1,09	2,59%	1,10	2,61%	1,10	2,60%	1,10	2,61%
1	1,00	2,37%	1,00	2,37%	1,00	2,37%	1,00	2,37%	1,00	2,37%	1,00	2,37%

Tabela 14 – Comparação dos valores do lift e da taxa prevista para a população em cada decil

Decil	Total de clientes Contactados	Respostas Positivas com Modelo	Respostas Positivas sem Modelo	Custo Envio	Valor venda com modelo	Valor venda sem modelo	Lucro com modelo	Lucro sem modelo
0.1	41 135	3 375	974	411 345	1 687 455	487 000	1 276 110	75 655
0.2	82 269	4 989	1 948	822 690	2 494 414	974 000	1 671 724	151 310
0.3	123 404	6 351	2 922	1 234 035	3 175 727	1 461 000	1 941 692	226 965
0.4	164 538	7 202	3 896	1 645 380	3 600 878	1 948 000	1 955 498	302 620
0.5	205 673	7 990	4 870	2 056 725	3 994 861	2 435 000	1 938 136	378 275
0.6	246 807	8 515	5 844	2 468 070	4 257 354	2 922 000	1 789 284	453 930
0.7	287 942	9 015	6 818	2 879 415	4 507 672	3 409 000	1 628 257	529 585
0.8	329 076	9 365	7 792	3 290 760	4 682 505	3 896 000	1 391 745	605 240
0.9	370 211	9 577	8 766	3 702 105	4 788 671	4 383 000	1 086 566	680 895
1	411 345	9 740	9 740	4 113 450	4 870 000	4 870 000	756 550	756 550

Tabela 15 – Ganhos Cumulativos com base na matriz de custos e no lift acumulado do modelo 1 como exemplo.