ISCTE IUL

Instituto Universitário de Lisboa

IDENTIFICAÇÃO E CARACTERIZAÇÃO DE SITUAÇÕES DE "CHURN" EM SISTEMAS DE TELECOMUNICAÇÕES

João Pedro Azevedo Oliveira

Tese submetida como requisito parcial para obtenção do grau de

Mestre em Ciências e Tecnologias da Informação

Especialidade em Sistemas Integrados de Apoio à Decisão

Orientador(a):

Doutor Orlando Belo, Professor Associado,

Departamento de Informática, Escola de Engenharia da Universidade do Minho

Co-orientador(a):

Doutora Maria José Trigueiros, Professora Associada,

Departamento de Ciências e Tecnologias da Informação, Instituto Superior de Ciências do Trabalho e da Empresa

Outubro, 2009

“We are drowning in information,
but starving for knowledge”
John Naisbett

ISCTE IUL
Instituto Universitário de Lisboa

Formação em Sistemas Integrados de Apoio à Decisão

João Pedro Azevedo Oliveira

Outubro
2009

Agradecimentos

Uma dissertação, apesar do processo solitário a que qualquer investigador está sujeito, reúne contributos de várias pessoas. No presente estudo, estes contributos foram essenciais para a conclusão deste trabalho.

Ao meu orientador, Prof. Dr. Orlando Belo, pelo constante apoio, incentivo e ensinamentos que tornou este trabalho uma realidade.

A todos os elementos da empresa em questão, pela disponibilização da informação e ferramentas para a execução deste trabalho.

Aos meus familiares, em particular para a minha namorada Paula, pelo suporte emocional e ajuda incansável.

Resumo

Nunca a relação cliente - fornecedor teve uma importância tão relevante no mercado económico como nos dias de hoje, sendo a área de Telecomunicações um dos segmentos de mercado onde os fornecedores de serviços orientam, cada vez mais, as suas decisões estratégicas na gestão do cliente e na sua satisfação, com a finalidade de promover e consolidar a fidelização deste com a organização. Esta dissertação de mestrado está inserida num projecto profissional na área das telecomunicações, um dos mercados mais dinâmicos e estratégicos da actualidade, fértil na construção e definição de sistemas que permitem classificar e explicar o cliente *churn,* tornando a escolha deste mercado uma mais-valia. O objectivo desta dissertação de mestrado é aprofundar o relacionamento entre fornecedor e cliente, utilizando os sistemas de mineração de dados para detecção e definição do conceito de cliente *churn* na área das Telecomunicações. Recorrendo à utilização de ferramentas de descoberta de conhecimento em bases de dados (KDD - *Knowledge Discovery and Data Mining*), procurou-se extrair conhecimento através da exploração de um elevado volume de informação, habitualmente residente nas bases de dados das empresas de telecomunicações. A pesquisa foi realizada numa base real de dados de uma empresa de telecomunicações do Médio Oriente (que guarda o seu direito de não ser identificada), após ter sido minuciosamente limpa e devidamente auditada quanto à robustez dos dados e sua veracidade no mercado em que se insere. O resultado deste trabalho é a aquisição de conhecimento na construção, compreensão e definição do cliente *churn*, começando na definição dos clientes com probabilidade elevada de deixar a empresa até aos clientes leais à empresa, representativos de valores de lucro elevados.

Palavras-chave:

- Telecomunicações, *Churn*, Descoberta de Conhecimento em Bases de Dados, Mineração de Dados, Médio Oriente;

Classificação ACM:

- M.1.0. Business Administration – General, M.3.0. Marketing and Advertising – General, M.4.1. Accounting and Auditing – Accounting;

Abstract

Never a relationship between customer-supplier has had such an economic importance as in the modern times, being the telecommunications area one of the market segments where the service suppliers, focus even more their strategic priorities on the customer's management and satisfaction, with the aim to promote and consolidate the customer loyalty with the organization. This thesis dissertation has been introduced in a professional project in the telecommunications area, which is one of the most proactive and strategic markets of our times, regarding definition and built in systems that define and explain churn client and due to the fact of this industry being one of the most affected by the churn phenomenon, in part, as a consequence of the technology that generates its own obsolescence in a short amount of time. The thesis dissertation aims to study the relation between customer and supplier, using the data mining systems to detect and define the concept of the churn client in the Telecommunications area. By using tools of Knowledge Discovery and Data Mining (KDD), it was possible to explore big volumes of data, normally available at Telecommunication IT systems databases. The research was carried out over on a real database of a Telecommunications company in the Middle East (which reserves the right of confidentiality) after thorough data cleansing and auditing to certificate the validity of the information being analyzed within its market. The outcome of this research is an acquisition of acknowledgment regarding the definition and understanding of the churn concept, by being able to identify customers with higher probability of leaving the company to identifying loyal customers that represent a higher value for the operator.

Índice

Índice de Tabelas

Índice de Figuras

Lista de Siglas e Acrónimos

DM	Data Mining
CRISP-DM	CRoss Industry Standard Process for Data Mining
SCP	Sistemas de Comunicação Pessoal
GSM	Global System for Mobile
ROAMING	Termo utilizado no sector das telecomunicações para referir á extensão do serviço de conectividade num determinado local que é diferente do local onde o serviço foi adquirido.
BD	Base de Dados
SO	Sistemas Operacionais
KDD	Knowledge Discovery in Databases
MD	Mineração de Dados
ARPU	Average Revenue Per User
LBS	Location-Based Services
OSS	Operational Support Systems
BSS	Business Support Systems
HLR	Home Location Register
MSC	Mobile Switching Center
CRM	Customer Relationship Management
ERP	Enterprise Resource Planning
ROI	Return Of Investment
SQL	Structured Query Language
IT	Information Technology
CART	Classification and Regression Trees
CHAID	Chi-Square Automatic Interaction Detector
SMS	Short Message Service
MMS	Multimedia Messaging Service
GPRS	General Packet Radio Services
PL/SQL	Procedural Language/Structured Query Language
MSISDN	Mobile Station Integrated Services Digital Network: Termo utilizado para referir número de telefone.
SPSS	Statistical Package for the Social Sciences

1. Introdução

1.1. Churn em Sistemas de Telecomunicações

No final do século XX e início do século XXI, as telecomunicações transformaram-se num dos sectores mais dinâmicos e emergentes da economia mundial. Em menos de 100 anos, 60% da população mundial tem acesso a alguma forma de telecomunicação. Com uma taxa de crescimento surpreendente, as telecomunicações sem fios deixaram de ser apanágio de instituições secretas e nicho de milionários para passarem a um bem comum nos mais diversos cantos do mundo. Os serviços de telecomunicações adquiriram já estatuto de símbolo de juventude, sofisticação e modernidade junto da sociedade. A conjuntura mundial de globalização e abertura de mercados, a estimulação da concorrência como o meio privilegiado de defesa dos consumidores, proporcionando melhores serviços a preços mais competitivos, são os principais factores que contribuíram para o rápido crescimento, impulsionado por uma constante modernização tecnológica e consequente portfólio de novos e mais complexos serviços. Este crescimento criou nos operadores a necessidade e a preocupação em criar uma rede de suporte à distribuição e divulgação de produtos e serviços com uma abrangência geográfica de acordo com a estratégia de expansão, seja ela orientada a nichos de consumidores ou à massificação do produto. Qualquer que seja a estratégia definida, a qualidade da oferta é fundamental como forma de garantir quer a lealdade por parte do consumidor, quer a competitividade na aquisição e retenção de clientes de maior valor.

Os mesmos factores que conduziram ao crescimento meteórico das telecomunicações são também os que contribuem para o aparecimento do fenómeno de *churn – o abandono dos clientes*. Os avanços tecnológicos constantes, que possibilitaram a generalização do serviço móvel são, também, responsáveis por suscitar no cliente a vontade de mudar. O aparecimento ininterrupto, nos mercados, de alternativas novas, mais sofisticadas e financeiramente mais acessíveis, numa procura de melhoria e diversificação contínua das ofertas, geram as condições favoráveis que conduzem os clientes a equacionar a mudança. A tecnologia ligada às telecomunicações gera, actualmente, a sua própria obsolescência num período de tempo incrivelmente curto.

Uma das regiões que tem apresentado maior crescimento no sector das telecomunicações é o Médio Oriente, onde a liberalização do mercado a par com a extensão dos serviços pelos aglomerados multinacionais e a forte concorrência que se faz sentir, contribuíram para uma verdadeira revolução nas comunicações. Como demonstra a figura 1 e 2, a Arábia Saudita, em particular, apresentou nos últimos 5 anos um crescimento de cerca de 100% em Sistemas de Comunicação Pessoal (SCP), sendo esperado que este crescimento se acentue ainda mais durante os próximos anos. Em países emergentes e de acentuado crescimento económico - onde por norma se verifica um *Average Revenue Per User (ARPU)* mais elevado - os SCP têm um impacto ainda mais significativo na economia, podendo ser duas vezes superior quando comparado com países desenvolvidos [6].

	2002	2003	2004	2005	2006	2Q 2007	2007F	2008F	2009F	2010F	2011F
Bahrain	390	446	624	767	907	890	922	980	1,009	1,032	1,048
Iran	2,279	3,450	4,271	8,525	13,659	17,043	19,806	27,312	35,396	43,679	51,541
Iraq	20	50	1,383	4,572	7,790	9,204	10,626	13,951	16,867	19,111	20,830
Israel	6,334	6,618	7,222	7,767	8,419	8,677	8,899	9,246	9,440	9,572	9,649
Jordan	1,220	1,325	1,624	3,138	4,343	4,775	5,216	5,842	6,233	6,439	6,600
Kuwait	1,227	1,420	2,076	2,383	2,530	2,624	2,717	2,897	3,073	3,255	3,434
Lebanon	775	800	881	998	1,099	1,148	1,218	1,367	1,511	1,630	1,745
Oman	465	593	828	1,333	1,818	2,135	2,502	2,894	3,152	3,322	3,485
Palestine	227	278	450	567	822	909	1,004	1,165	1,328	1,486	1,631
Qatar	267	377	515	716	920	1,108	1,227	1,310	1,379	1,422	1,461
Saudi Arabia	5,008	7,238	9,150	13,500	20,000	23,890	27,000	31,131	34,680	36,934	38,411
Syria	400	1,185	1,963	2,975	4,569	5,392	6,310	8,329	10,736	13,356	15,773
Turkey	23,323	27,888	34,708	43,609	52,663	56,580	60,562	66,679	70,880	73,077	74,977
UAE	2,428	2,972	3,683	4,530	5,519	6,520	7,186	7,998	8,406	8,658	8,762
TOTAL	44,363	54,640	69,378	95,380	125,058	140,895	155,193	181,101	204,091	222,972	239,347

Source: ITU, Regulators, Operators
Data in red is estimated

Figura 1 – Mercado Móvel no Médio Oriente – Estatísticas e Previsões por País

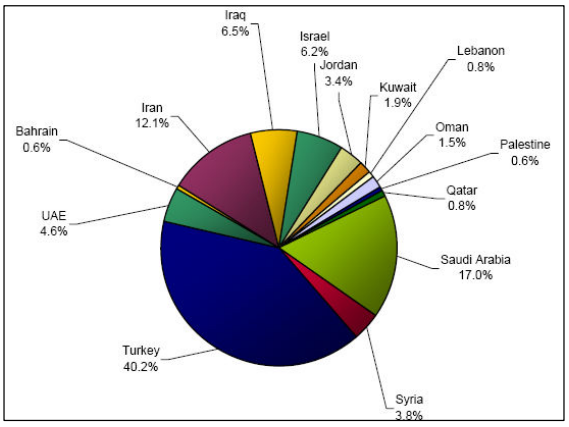

Figura 2 – Mercado Móvel no Médio Oriente – 2º Trimestre de 2009 por País

A comunicação no Médio Oriente caracterizada, num passado recente, por um foco quase exclusivo no serviço de transmissão de voz, apresenta hoje uma oferta mais diversificada de serviços, em particular, com a introdução da internet, *roaming* ou *Location-Based Services* (LBS) resultado, maioritariamente, da evolução do mercado e incremento da concorrência. Esta evolução serve de impulsionadora a um conjunto de movimentos que beneficiam os consumidores e contribuem para a efervescência do mercado.

Dada a natureza tecnológica do sector, todos os processos se encontram sistematizados, sendo, na sua grande maioria, suportados por sistemas de informação, desde a sua componente menos abstracta, como é o caso da mediação, até níveis superiores de abstracção, como a automação de campanhas de marketing e interacção com o cliente. Toda a informação utilizada para suportar a cadeia de gestão do cliente é, quantitativamente enriquecida, resultado da natureza transaccional de um operador de telecomunicações, mais importante tem um carácter qualitativo que advém da quantidade de elementos informativos fornecidos pelos sistemas *Operational Support Systems* (OSS) (engenharia, tipo *Home Location Register* (HLR), *Mobile Switching Center* (MSC)) e *Business Support Systems* (BSS) (sistemas de *Billing*, *Data Warehousing* ou *Customer Relationship Management* (CRM), por exemplo) que permitem ao operador um conhecimento pormenorizado de cada um dos seus clientes. Esta conjuntura é fundamental na análise do comportamento da base de clientes e do conhecimento das características chave que permitem identificar os tipos de cliente a angariar ou a reter. A existência de um sistema de análise central para suporte à decisão tem provado ser fundamental para atingir este objectivo, nomeadamente, pela verificação de duas directivas:

1. O objectivo que cada um dos elementos da cadeia de *OSS* (*HLR, MSC, SMSc*) e *BSS (Billing, CRM, ERP, etc.)* é orientado para uma ou várias áreas específicas de suporte à disponibilização de serviço e gestão do cliente; regendo-se por preocupações como a segurança, a integridade, a robustez e a fiabilidade do processamento da informação. Por isso, não se encontram orientados para a análise da informação ou para a produção de relatórios especializados, com informação transversal a várias áreas e departamentos de negócio.
2. Só será possível garantir a existência de uma visão única e cooperativa da organização, com a criação de um sistema corporativo que unifique a informação considerada relevante de cada um dos elementos da cadeia e seja a base, por excelência do sistema de suporte à decisão. Segundo Inmon [5], estes sistemas

corporativos podem ser definidos como uma colecção de dados orientada por assuntos, integrada, variante no tempo e não volátil, designados de *Data Warehouse*. Estes também podem ser caracterizados por um conjunto de cópias de dados transaccionais guardados numa estrutura específica (modelos dimensionais) e orientada para consultas e análises (baseadas em histórico) de grandes volumes de informação, sendo o *Data Mart* a estrutura analítica orientada à disponibilização da informação que satisfaz um conjunto de necessidades específicas. Estes sistemas disponibilizam uma vista conceptual privilegiada para os seus utilizadores, já que muitos deles pensam e actuam, no seu dia-a-dia, de forma multidimensional e, alguns deles (os gestores) tendem a colocar variadíssimas questões *ad-hoc* que envolvem a conjugação simultânea de diversos aspectos de análise, requerendo a utilização de dados de várias das suas perspectivas de negócio e modos de actuação que podem ser satisfeitas, através da análise e exploração de *hipercubos*. Os *Data Warehouses* são orientados à:

- Recolha de dados dos diversos sistemas de uma organização.
- Organização dos dados operacionais em sistemas de suporte à decisão.
- Armazenamento e colecção de um histórico das informações obtidas.
- Disponibilização de um modelo de dados comum a toda a organização.
- Independência entre os diversos sistemas operacionais e o *Data Warehousing*.

Após organização e consolidação da informação num *Data Warehouse*, esta deve ser manipulada e é, neste âmbito, que surgem os processos de mineração de dados, cujo objectivo é a extracção de conhecimento a partir de grandes volumes de informação. Existe no meio, quem utilize o conceito de KDD para referir a Mineração Dados, no entanto, o primeiro conceito é mais abrangente. "*KDD é o processo de identificação de válidos, novos, potencialmente úteis e compreensíveis padrões em dados*" [3]. Nos processos de KDD, dados são conjuntos de factos existentes (ex: Base de Dados) e o modelo é uma expressão que descreve um seu subconjunto ou uma expressão que lhe pode ser aplicada, a inferência de um padrão significa um modelo construído a partir dos dados. O significado de processo no âmbito do KDD implica diversas etapas, tal como se pode verificar na figura 3, envolvendo a:

- Limpeza de ruídos e de dados inconsistentes.

- Integração dos dados dispersos num único repositório.
- Selecção das variáveis importantes para análise.
- Transformação dos dados através da aplicação de algoritmos de mineração.
- Mineração de dados com o objectivo de detecção de padrões semelhantes que viabilizem a extracção de conhecimento.
- Avaliação e identificação de padrões a partir de regras de negócio.
- Apresentação de resultados recorrendo a técnicas de representação de conhecimento.

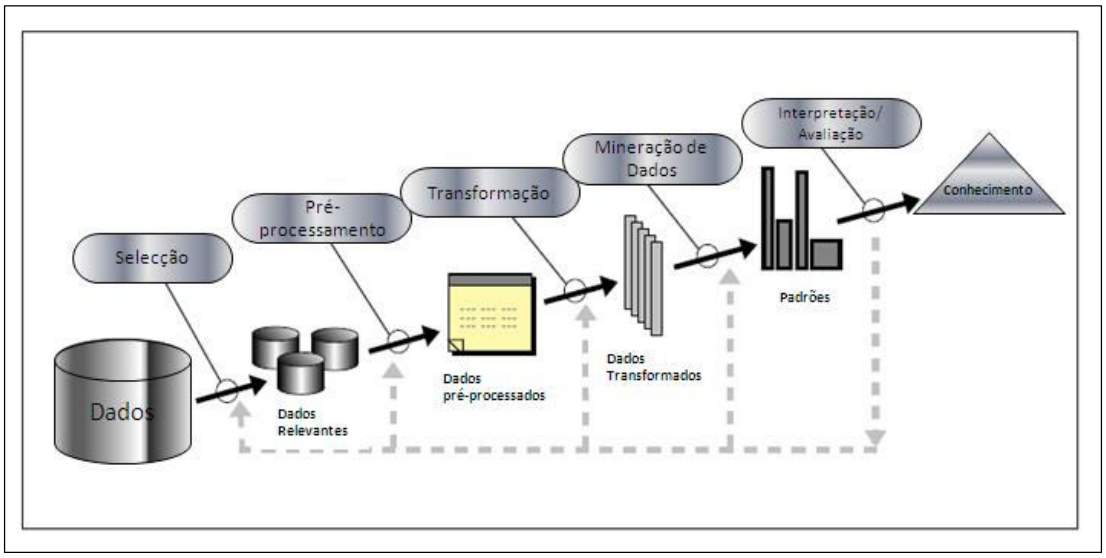

Figura 3 – Passos que compõe o processo KDD [3]

1.2. O estudo do churn em telecomunicações: motivação e objectivos

Churn é, actualmente, um dos tópicos mais importantes para qualquer empresa de telecomunicações no mundo, ganhando uma relevância crescente em mercados maduros e liberalizados. Encontram-se autores com diferentes opiniões sobre o tema, mas todos concordam que o *churn* é nefasto à organização e que afecta directamente o lucro das empresas, não só pelo custo associado à perda de clientes, como também, à sua eventual recuperação. É neste âmbito que a identificação e a compreensão do cliente *churn* se revestem de vital importância para a sobrevivência das empresas. Desta forma a principal

motivação desta dissertação é desenvolver técnicas, processos, modelos que permitam prever e analisar o potencial cliente *churn.*

Sendo o objectivo principal deste trabalho de dissertação, o estudo detalhado do conceito de cliente *churn* em telecomunicações, recorrendo à aplicação de técnicas de mineração de dados em processos com vista à sua detecção, importa concretizar este objectivo em detalhes mais específicos:

- O estudo detalhado sobre o conceito de *churn* (definição e conceitos fundamentais, casos de estudo; e soluções adoptadas por empresas de telecomunicações).
- O desenvolvimento de aplicações de técnicas de mineração de dados em processos de detecção de clientes *churn* (modelos baseados em classificação - previsão).
- A aplicação e avaliação das técnicas aplicadas sobre um conjunto seleccionado de casos de estudo.
- A criação de um sistema de informação que permita à empresa conhecer quais os seus clientes *churn* e em que condições.

Em suma, esta dissertação visa unir a construção de um modelo de mineração de dados, funcional, construído sobre a base de informação de uma empresa de telecomunicações, com o intuito de criar informação inteligente sobre o negócio, servindo de base ao processo de tomada de decisão pela equipa de Marketing de Retenção.

1.3. Metodologia de Trabalho

Relativamente à metodologia, de uma forma simples e objectiva, este documento seguiu a estrutura normal de um projecto de investigação, que, como se pode verificar na secção seguinte, foi reflectida na estrutura deste documento. Numa primeira fase, foi realizada uma introdução ao objectivo de estudo com a exposição do problema e alguns resultados informativos. De seguida, foi executado um estudo detalhado sobre o fenómeno do *churn* e suas implicações. Além disso, abordaram-se também várias das técnicas e algoritmos de mineração de dados utilizados usualmente em modelos de previsão, tendo-se realizado ainda o estudo de alguns casos reais de *churn* em ambientes empresariais – o estado da arte. Na fase seguinte, procedeu-se à análise e definição dos requisitos, ao desenvolvimento e à apresentação dos modelos e apresentou-se uma análise crítica dos

mesmos. Por fim, na última fase deste estudo, efectuou-se uma reflexão geral do trabalho desenvolvido e propôs-se algumas tarefas de trabalho futuro, que devem ser tidas em consideração. Importa salientar a forte influência que a metodologia CRISP-DM teve no desenvolvimento deste estudo de mineração de dados, reflectida na descrição dos capítulos e exposta na secção seguinte.

1.4. Estrutura do documento

Este documento, como vimos, começa por enquadrar os problemas da tese no seu meio envolvente, através de uma breve análise ao conceito em estudo nas empresas de telecomunicações e pela apresentação de dados referentes à região do Médio Oriente. Nos capítulos seguintes, é descrito o contexto científico da dissertação - Sistemas de Apoio à Decisão em ambientes de telecomunicações com o apoio dos Sistemas de Mineração de dados.

O segundo capítulo é essencialmente composto por uma exposição do conceito *churn*, descrevendo-se o problema do *churn*, as suas causas e consequências, formas e técnicas de detecção, com destaque para os métodos de classificação e previsão, através do estudo das técnicas de Redes Neuronais, Árvores de decisão e Regressão Logística e, por fim, apresentam-se alguns trabalhos relacionados com o estudo.

De seguida, no capítulo três, define-se a metodologia e as regras de negócio, procedendo-se à caracterização do cliente *churn* e à definição dos *targets;* para determinação dos objectivos do sistema de mineração de dados.

No capítulo seguinte, identifica-se e caracteriza-se o *churn* descrevendo-se as várias fases de construção dos modelos. Neste capítulo, é exposto o conjunto de dados utilizado, modelos e testes de detecção de *churn* efectuados, as suas características e limitações, fazendo-se ainda uma apresentação dos resultados da aplicação dos diferentes algoritmos de previsão. Ainda neste capítulo, é feita uma análise crítica sobre o desempenho de cada um dos algoritmos utilizados, destacando-se, sempre que possível e oportuno, os seus aspectos positivos e negativos, e fazendo-se uma comparação com outros casos de estudo sobre situações reais de ocorrência de *churn*.

Por fim, no sexto capítulo, apresentamos as conclusões finais e algumas propostas para desenvolver num futuro próximo.

2. O Mundo do *Churn*

2.1. *Churn*

Um vasto número de autores tem vindo a estudar o fenómeno do *Churn* de várias formas e perspectivas de negócio. Segundo Strouse, "*Churn* é o acto de infidelidade de um cliente, isto é, a perda de um cliente em favor de uma empresa concorrente" [7], finalizando-se, desta forma, a ligação com a empresa antiga. *Churn* pode ser definido pela "taxa anual de rotatividade da base de clientes" [7]. Strouse afirma também que, se numa análise anual, a taxa de *churn* é de 5%, isto representa que a empresa durante um ano perdeu 5% dos clientes da empresa por estarem insatisfeitos, em princípio, com algum dos seus serviços. Num mercado tão competitivo como o das telecomunicações, existe uma necessidade muito grande, por parte das empresas, de procurar manter uma relação de fidelidade, na tentativa de evitar o abandono e, consequentemente, evitar elevados custos de reaquisição desses clientes. Os sistemas de aviso prévio apresentam-se como uma possível solução que permitem, com antecedência, evitar o abandono do cliente, através da aplicação de técnicas e medidas de retenção, possibilitando à empresa não só manter os seus clientes mais lucrativos, mas também aumentar o seu nível de fidelização. Um outro objectivo, não menos importante, é o da aprendizagem com situações que envolvam clientes que abandonam a empresa, com a finalidade de aumentar a eficácia de previsão e evitar abandonos futuros. De acordo com Lejeune [9], outro factor a ter em conta é a restrição da base de clientes em contraposição à análise global dos clientes; por motivos de eficiência de resultados. Devem ser considerados os clientes que apresentam valor significativo para a empresa. Nestas situações, pretende-se aferir que tipo de cliente está em análise, uma vez que este pode ser um cliente com elevado valor para a empresa e com grande impacto – financeiro, por exemplo - em caso de abandono. Nesta fase, é necessário estabelecer algumas regras de causa - efeito para perceber a motivação para o abandono e em que medida faz sentido, de acordo com a política da empresa, reter ou não estes clientes.

A análise do cliente *churn* em empresas de telecomunicações permite não só identificar quais os clientes com índices de abandono elevados, mas também identificar quais os clientes de elevado valor. Desta forma, conseguimos aplicar proactivamente um conjunto seleccionado de acções com o intuito de garantir a retenção e, também, a fidelização do

cliente. É importante ter em conta que o comportamento entre o operador e o cliente são mutáveis ao longo do tempo e que estas mudanças advêm de factores causais, sociais, culturais, pessoais ou psicológicos [10]. Ainda segundo Strouse, a análise do cliente *churn* apresenta-se com uma análise multi-variada, que pode ser aplicada a um vasto número de aplicações, nomeadamente [7]:

- Identificar a sua permanência através de uma análise financeira e identificar quais os segmentos alvo.
- Analisar os segmentos mais propensos ao abandono, especificando e implementando algumas políticas preventivas ou de retenção que permitam minimizar as perdas e, obviamente, aumentar o lucro.
- Aferir a lealdade dos consumidores, analisando e aferindo quais os clientes que são "fiéis" e a razão porque o são.
- Recorrer a análises probabilísticas, analisar a tendência de aumento e estabilizar ou diminuir o *churn*.
- Identificar e melhorar as características responsáveis pelo abandono.
- Adoptar e desenvolver estratégias de comunicação, orientando-as para os segmentos mais vulneráveis à perda de clientes.
- Identificar e analisar quais as empresas e os produtos concorrentes, identificando-os e caracterizando-os claramente, bem como fazer uma análise de aquisições em massa dos clientes da empresa.
- Medir o impacto das campanhas de aquisições de novos clientes.
- Analisar e aferir as necessidades do cliente quanto à qualidade e necessidade do serviço.
- Avaliar a eficácia dos processos de retenção de clientes - se os processos estão a ser aplicados aos clientes correctos e se estão de acordo com o resultado esperado.

Torna-se claro que este conhecimento não pode ser obtido exclusivamente através de *Data Warehouses* nem através de ferramentas de gestão de clientes (CRM). Este tipo de conhecimento só pode ser conseguido com recurso ao crescimento de informação referente ao cliente que permita, através de ferramentas de mineração de dados, conhecer e classificar o cliente no que diz respeito a todos os itens descritos anteriormente.

2.1.1. Churn Voluntário

Quando se tenta compreender com exactidão as motivações que levam um cliente a abandonar o operador – gerando *churn* – apercebemo-nos do pouco esforço realizado na interpretação do seu perfil, enquanto cliente, e na sua motivação inicial para adesão ao serviço, ou às razões que o levam a querer mudar de operador. Sendo, o factor preço, a principal motivação na origem do abandono de um cliente, a decisão que o leva a mudar de operador é um processo complexo no qual estão presentes inúmeros factores (tecnológicos, económicos, qualidade, etc.), tal como comprovam as figuras 4 e 5. O *churn* voluntário acontece quando um cliente inicia o processo de terminação do contrato [2].

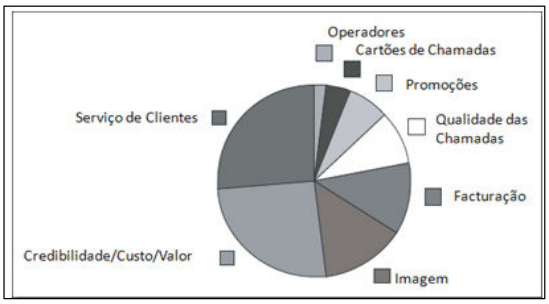

Figura 4 – Razões do *Churn* Voluntário no mercado móbil dos EUA [2]

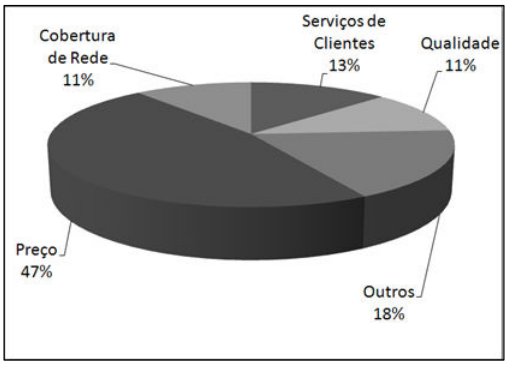

Figura 5 – Razões da satisfação dos clientes no mercado móbil – Longa Distância dos EUA [2]

O *churn* do tipo voluntário pode ser subdividido em 2 grandes conjuntos que se apresentam a seguir:

- ***Churn* Acidental.**

  Este tipo de *churn* acontece não por ter sido previamente planeado pelo cliente, mas porque, devido a determinado acontecimento na vida do cliente, este encontra-

se obrigado a cessar o contrato. Esta cessação de contrato é vista como um efeito colateral desse acontecimento [2].

- ***Churn* Deliberado.**

  Enquanto no *churn* acidental, não existe uma motivação objectiva do cliente que suscitasse uma acção preventiva por parte do operador - situação que acorre num número reduzido de casos, no caso do *churn* deliberado há um processo de tomada de decisão por parte do cliente sustentado num ou diversos factores [2], tais como:

  - Tecnologia - Cliente espera uma nova ou melhor tecnologia.
  - Económicos - Sensibilidade ao preço.
  - Qualidade - Melhor qualidade dos produtos/serviços.
  - Sociais - Questões do grupo social onde o cliente se insere.
  - Psicológicos - Questões meramente psicológicas (vontade).
  - Conveniência - Factores pessoais, familiares ou contratuais.

2.1.2. Churn Involuntário

Para grande parte das operadoras de telecomunicações, este é o tipo de *churn* mais frequente e com maior impacto financeiro. Neste tipo de *churn*, não se verifica uma intenção por parte do cliente em mudar de operador, mas sim um conjunto de circunstâncias que forçam a finalização da subscrição [2].

As motivações ou circunstâncias que geram este *churn* podem ser divididas em 3 subcategorias, nomeadamente:

1) **Fraude** – Quando são realizadas actividades proibidas por lei.
2) ***Churn* por não pagamento** – Clientes com problemas financeiros.
3) ***Churn* por subutilização** – Clientes que não fazem uso do serv*iço (ex.* **Clientes pré-pagos** que não pagam uma assinatura mensal e não utilizam o serviço).

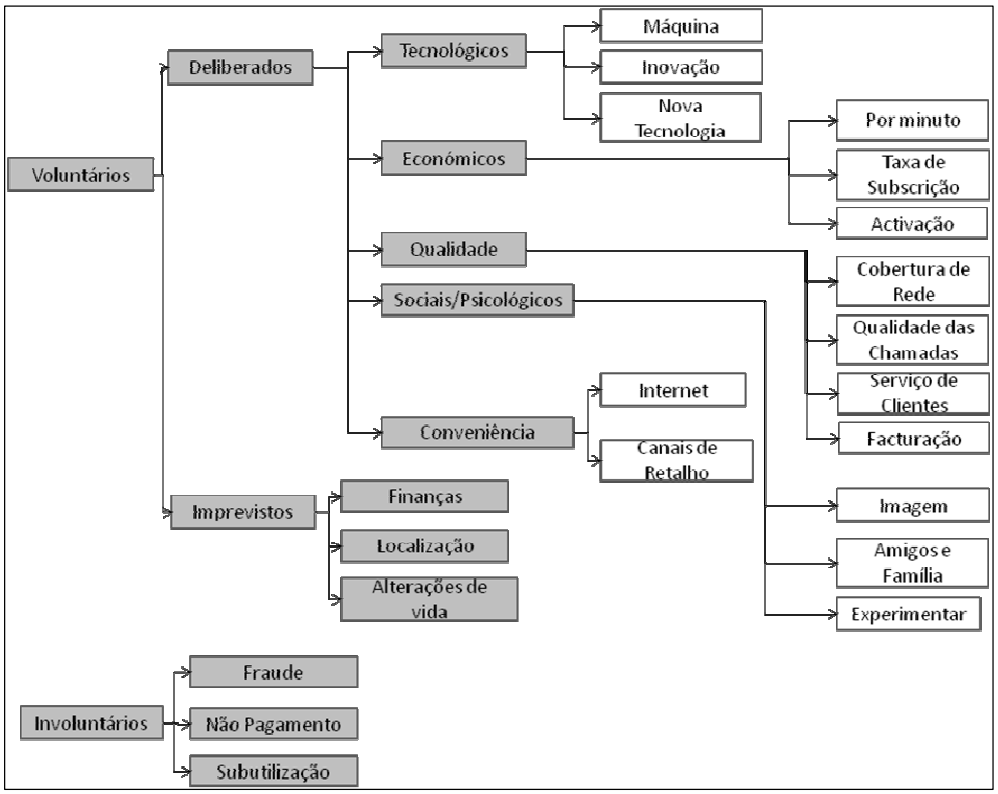

Figura 6 – Taxonomia do Cliente *Churn* [2]

De acordo com a informação apresentada na figura 6, as situações de *churn* no sector das telecomunicações podem acontecer de várias maneiras, em que cada operador, a partir do conhecimento *standard* sobre o cliente em *churn*, aplica à sua realidade, a definição mais apropriada. Neste processo, são tidas em conta diversas variáveis demográficas, geográficas, económicas ou de comportamento.

Considerando que o mercado da empresa de telecomunicações em estudo tem um conjunto de características particulares, o foco deste estudo desenvolveu-se em torno de questões como:

- Qual o comportamento típico de um cliente *churn*?
- Quais os clientes *churn* voluntários deliberados e suas motivações?
- Quais os clientes *churn* involuntários por não pagamento?

2.1.3. Churn Interno

O termo *churn* interno é usado de forma diferente por diferentes operadores de telecomunicações, estando associado, normalmente, a clientes que pretendem terminar determinado serviço para aderir a um outro produto ou método de pagamento. Exemplo disso é a migração do cliente de um plano de tarifário pós-pago para um plano de tarifário pré-pago. Este é um dos acontecimentos mais comuns nas telecomunicações derivado de constrangimentos financeiros dos consumidores. Dos três tipos de *churn* apresentados, este é aquele que se apresenta com menor importância, uma vez que o cliente não deixa efectivamente a empresa. No âmbito desta tese, este tipo de *churn* não será objecto de análise.

2.2. Sistemas de Mineração

Sabendo-se dos diversos tipos de *churn,* importa agora analisar e discutir as formas e os meios que hoje temos à nossa disposição para a sua detecção. Existem variadíssimos tipos de abordagem a este tipo de fenómeno. É pena, contudo, que nem todas sejam suficientemente eficazes ou úteis. No presente trabalho, não abordaremos todos os tipos de estratégias e técnicas para a identificação e tratamento de casos de *churn*. Isso seria, tal como já tivemos a oportunidade de referir no capítulo de abertura desta dissertação, fugir à sua linha de estudo e de realização de tarefas, uma vez que nos interessa, exclusivamente, abordar a forma como as técnicas e modelos da área de Mineração de dados podem ser aplicados e estudar a sua efectividade num domínio tão complexo e multifacetado como é o *churn*. Os métodos de agrupamento e de associação [14] ficam para um eventual estudo posterior, deste modo, técnicas como *K-Means* e *Apriori* [14] não serão consideradas. Visto isso, neste estudo será dado ênfase a técnicas relacionadas com sistemas de classificação, tais como árvores de decisão, redes neuronais e regressões logísticas [14], por serem as técnicas mais usadas, nos estudos analisados (secção 2.5), sobre previsão de *churn*. No entanto, antes de se exporem as técnicas em estudo, é necessário classificar os sistemas de mineração, tendo em contas a sua dimensão. Segundo Michael Berry [8], uma alternativa para a classificação dos sistemas de mineração de dados é a sua divisão em duas dimensões: a natureza da tarefa e o objectivo a atingir.

Classificando os sistemas em cada uma das duas dimensões podem agrupar-se sistemas semelhantes, aplicando algoritmos e procedimentos similares.

2.5.1 A Natureza da tarefa

Em relação à natureza da tarefa, esta pode ser dividida em três grupos [15], nomeadamente:

1. *Aprendizagem Supervisionada*: situação em que determinado indivíduo fornece a identificação (rótulos) de cada objecto de base de dados. Por isso, o conjunto de treino é acompanhado dos rótulos indicativos das classes das observações. O classificador é treinado de forma a replicar a decisão correcta para todos os padrões de treino.

    a. *Classificação*: uma tarefa de classificação consiste em associar um item a uma classe, de várias opções pré-definidas. A tarefa classifica os dados (constrói o modelo) com base no conjunto de treino e nos valores do atributo classificador, de forma a determinar a classe dos novos dados. A tarefa do utilizador será a de seleccionar qual a classe que melhor representa um registo. No caso de um conjunto de clientes de telecomunicações, cada registo contém os atributos de cada cliente: variáveis demográficas, financeiras e de tráfego, entre outras. O objectivo é classificar cada cliente, consonante o valor para a empresa, em alto, médio ou baixo.

    b. *Predição*: pode ser definida como a tarefa de prever um valor com base em outros atributos que explicam o registo. Modela funções sobre valores contínuos, com o objectivo de prever valores desconhecidos ou perdidos. Exemplo: a previsão do cliente *churn* numa empresa de telecomunicações; com base nos atributos do cliente, o objectivo é prever se este vai ou não deixar a empresa.

2. *Aprendizagem Não Supervisionada*: situação em que não se conhecem as categorias envolvidas, logo, os rótulos das classes no conjunto de treino é desconhecida. Sendo assim, o algoritmo de aprendizagem procura uma estrutura nos dados que permita dividi-los em classes.

    a. *Clustering ou Agrupamento*: é uma tarefa de classificação sem classes pré-definidas. O objectivo consiste em agrupar registos semelhantes e separar registos diferentes. Quando se tratam grandes volumes de dados, é

comum que padrões concorrentes se anulem ao serem observados como um todo. Separar os dados em vários subconjuntos permite ao utilizador detectar estas diferenças e identificar novas oportunidades.

b. Associação: devido à sua vasta aplicabilidade, as regras de associação são de grande importância no processo de mineração de dados. Segundo Han e Kamber [14], a tarefa das regras de associação é encontrar padrões num conjunto de dados que contêm itens relacionados ao acontecimento de outros itens. Exemplo de telecomunicações:

Campanha SMS's Grátis >> Campanha de MMS´s Grátis

Esta regra pode ser descrita da seguinte forma, os clientes que aderem à campanha SMS's Grátis tipicamente também aderem à campanha de MMS's Grátis.

2.5.2 Objectivos a atingir

No que se refere aos objectivos a atingir, existem dois tipos, a saber:

1. *Produção de um modelo*: o objectivo da previsão é construir um modelo que possa ser aplicado a novos dados, com intuito de obter determinada classificação para explicar determinada acção.
2. *Produção de informação*: neste caso, procura-se descrever os dados com o intuito de proporcionar ao utilizador uma nova perspectiva dos mesmos.

No presente estudo, serão considerados os dois tipos de objectivos enunciados anteriormente, quer para disponibilizar um modelo capaz de classificar os clientes *churn*, quer para proporcionar uma outra visão dos dados.

2.3. Técnicas de Modelação

2.3.1. Árvores de Decisão

Após a caracterização das dimensões dos sistemas de mineração de dados, é necessário analisar as técnicas/algoritmos mais usados na construção de modelos preditivos. Uma dessas técnicas é a árvore de decisão. Esta caracteriza-se por ser um modelo preditivo

apresentado em forma de árvore, em que cada ramo representa uma questão de classificação e as folhas representam um conjunto de dados para essa classificação [15]. Se usarmos, como exemplo, o conjunto de clientes que não vão renovar os seus contratos de serviços móveis, uma possível árvore de decisão está apresentada na figura 7.

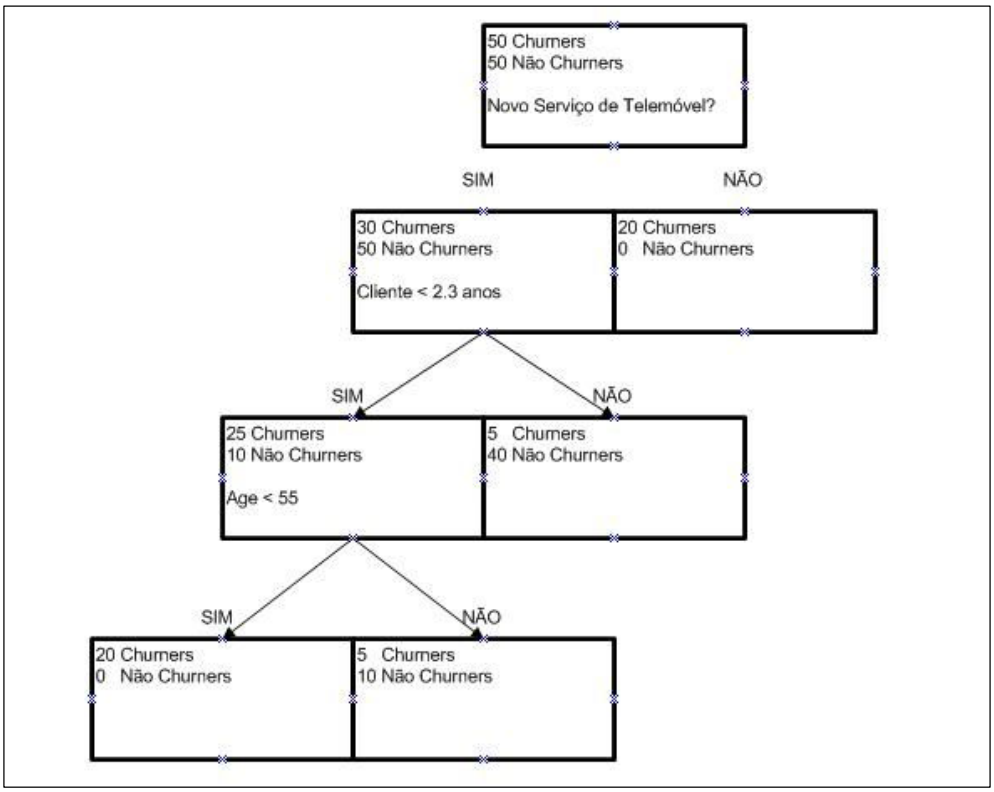

Figura 7 – Um exemplo de árvore de decisão

Numa árvore de decisão (figura 7) podem ser observadas as seguintes situações:

- A informação é dividida em cada ramo sem haver perda da informação.
- Os números de clientes *churn* mantêm-se inalterados em qualquer ponto da árvore.
- O modelo em árvore é intuitivo.
- O modelo é facilmente aplicável a campanhas de marketing de retenção.

Este tipo de modelo permite, de uma forma simples, aferir factos importantes, tais como os clientes que estão fidelizamos há algum tempo e têm os serviços móveis actualizados e são fiéis. Numa perspectiva de negócio, as árvores de decisão são segmentações da

informação em que cada folha da árvore é composta por informação relacionada entre si, representando um segmento. Devido à sua estrutura em árvore e facilidade de geração de regras, as árvores de decisão são uma das técnicas mais usadas na construção de modelos. Além do mais, a sua simplicidade permite o desenvolvimento de modelos complexos de *Return On Investment* (ROI) que podem facilmente ser usados, posteriormente, na construção dos modelos preditivos tão importantes para o negócio. Importa referir que, devido ao seu elevado nível de automatização, à facilidade de processamento e limpeza dos dados e à facilidade de tradução das árvores de decisão em SQL para implementação em base de dados relacionais, as árvores de decisão têm provado ser de simples integração com os restantes processos de IT.

2.3.2. Árvores de Decisão em Modelos Preditivos

No passado, grande parte das aplicações de árvores de decisão foram usadas como complemento a técnicas de estatística, como as regressões logísticas, para refinar e processar dados. Todavia, nos dias de hoje, esta técnica é cada vez mais usada nos modelos de previsão. A ideia de que as árvores de decisão servem apenas para a exploração de dados está claramente em mutação.

O primeiro passo no desenvolvimento de uma árvore de decisão é o processo de crescimento da árvore. De forma mais específica, o algoritmo de construção da árvore procura criar uma estrutura que represente, o mais fielmente possível, todos os dados disponíveis, mas a maioria das vezes não é atingível o grau de perfeição, por causa do ruído que está sempre presente na informação. Exemplo disso é a existência de variáveis não seleccionadas da base de dados. Contudo, estas variáveis são sobejamente importantes para a classificação alvo que estamos a tentar prever.

O crescimento das árvores assemelha-se a um jogo em que o objectivo é encontrar a melhor questão possível para fazer em cada nódulo da árvore. No final da árvore, os resultados serão nódulos que pretendemos que sejam de tipos distintos. Uma questão como: "Os clientes com mais de 40 anos?" não será, provavelmente, suficiente para distinguir os clientes que são *churn* daqueles que não o são - digamos que se obtém uma divisão na ordem dos 40% / 60%.

Na realidade, pode existir um conjunto de questões que faça uma divisão mais interessante. Nesta perspectiva, tomando como exemplo a questão "É cliente há menos de um ano?" que define o segmento onde o *churn* é 90%, a questão é tanto mais eficaz quanto maior for a sua capacidade de organização dos dados ou, neste caso específico, a probabilidade do cliente *churn* existir neste segmento. Num universo de 100 clientes, em que 50% dos clientes são *churn* e 50% não *churn*, a pergunta é boa se o modelo for capaz de identificar os clientes *churn*. De facto, o modelo ordena os dados que estavam, previamente, desordenados.

Um algoritmo de árvore de decisão "olha" para todas as possíveis perguntas que podem dividir o grupo de dados disponível em segmentos homogéneos, com respeito às diferentes classes a serem previstas. Alguns algoritmos de árvores de decisão podem usar também algumas heurísticas para escolher as questões. Dois desses algoritmos, talvez os mais usados nesta área, são os seguintes:

- *CART* – Os "*predictors*" são escolhidos à medida que diminui a desorganização dos dados. Na construção da árvore CART, cada "*predictor*" é escolhido com base na capacidade de divisão das diferentes previsões. Por exemplo, a medida "*entropy metric*" é usada para escolher o melhor "*predictor*". Uma das grandes vantagens do CART é a do algoritmo conter a validação do modelo e a descoberta do modelo óptimo. Isto consegue-se, construindo uma complexa árvore e, posteriormente, podando-a no sentido inverso até à obtenção da árvore óptima, e com base nos resultados da validação cruzada ou da validação de teste. Este algoritmo é relativamente robusto no que diz respeito a variáveis nulas - se existir um registo com variáveis nulas, este não é utilizado para a escolha do melhor "*predictor*".
- *CHAID* – Este algoritmo é similar ao CART na construção das árvores de decisão, diferindo, contudo, na forma de escolha dos "*predictors*". Para a escolha da divisão óptima é utilizado o teste do Qui-quadrado e tabelas de contingência, para determinar qual o melhor "*predictor*".

Os algoritmos de árvore de decisão devem ser interrompidos quando acontece um dos seguintes critérios:

- Quando o segmento contém apenas um registo.
- Quando os registos no mesmo segmento têm características idênticas.

- Quando a melhoria não é substancial para prosseguir com a divisão.

2.3.3. Redes Neuronais

As redes neuronais são outra das técnicas mais usadas na construção de modelos preditivos, caracterizando-se por serem uma estrutura de processamento de informação distribuída e paralela, formada por unidades de processamento, normalmente designadas por nós, neurónios ou células interconectadas por arcos unidireccionais, também chamados de ligações ou conexões. Os nós possuem memória local e podem realizar operações de processamento de informação localizada. Cada célula possui uma única saída, a qual se pode ramificar em muitas ligações colaterais (cada ramificação possui o mesmo sinal de saída do neurónio). Todo o processamento, que se realiza em cada unidade, deve ser completamente local, isto é, deve depender apenas dos valores correntes dos sinais de entrada que chegam dos neurónios, através das conexões. Estes valores actuam sobre os valores armazenados na memória local da célula [26].

Para explicar como as redes neuronais detectam padrões em grandes volumes de informação, diz-se usualmente que elas aprendem a detectar e a fazer melhores previsões da mesma forma que os humanos o fazem. As redes neuronais aprendem usando algoritmos e técnicas semelhantes aos encontrados em estatística e outros algoritmos de *Data Mining*. No entanto, importa dizer que as redes neuronais, embora aprendam e evoluam com o tempo, não superam os algoritmos estáticos. Na realidade, estes métodos estáticos aprendem, utilizando informação de histórico, da mesma forma que os evolutivos e podem até ser mais eficientes; uma vez que processam todos os registos como um todo ao invés dos evolutivos que processam registo a registo.

A vantagem da utilização das redes neuronais é poder serem usadas sem conhecimento profundo de como funcionam, ou de como o modelo preditivo é construído ou sem conhecimentos sobre base de dados. A maior parte das redes neuronais pode ser usada sem reorganização ou modificação dos dados, contudo, o contrário é muitas vezes verdade, uma vez que existem a necessidade de tomar decisões importantes no desenho do modelo para a utilização eficaz das redes neuronais, nomeadamente:

- Como deverão os nódulos da rede serem conectados.
- Quantos neurónios (unidades de processamento) devem ser usados.

– Quando é que o treino deve ser parado para evitar o "*overfitting*".

Existem muitos passos importantes necessários para o pré-processamento dos dados que alimentam as redes neuronais. Um dos requisitos mais comuns é a normalização dos dados numéricos entre 0.0 e 1.0. As categorias também devem ser transformadas em "*predictors*" virtuais entre 0 e 1. Para assegurar o sucesso na utilização deste método, é necessário compreender o significado da informação existente na base de dados e ter uma noção clara de qual é o problema de negócio a ser resolvido. O conceito fundamental é que as redes neuronais não disponibilizam atalhos. O modelo é representado por valores numéricos através de um método de cálculo complexo que requer que todos os valores do "*predictor*" sejam numéricos. O resultado das redes neuronais é numérico e, como tal, precisa de ser traduzido, caso o valor da categoria não o seja.

2.3.4. Redes Neuronais em Modelos Preditivos

Com o objectivo de criar uma previsão, as redes neuronais aceitam os valores para os "*predictors*" naquilo que são chamados os nódulos de "*input*".

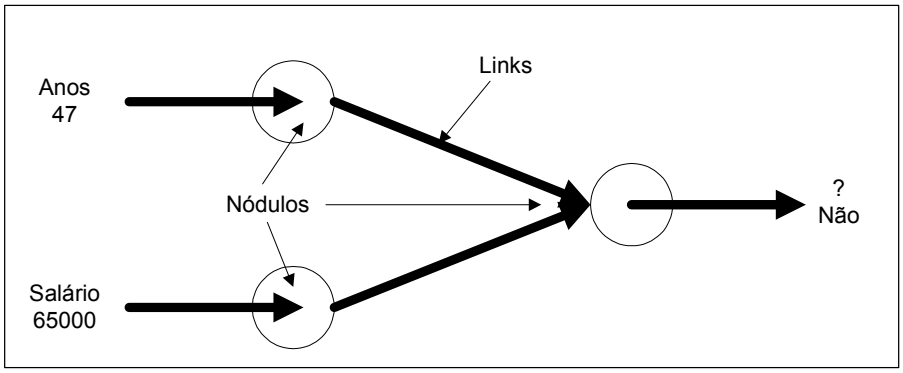

Figura 8 – Rede Neuronal

Estes passam a ser os valores dos nódulos que são multiplicados pelo valor do conector, em tudo semelhante aos pesos no método do vizinho mais próximo. Estes valores são, posteriormente, adicionados no nódulo de "*output*", onde uma função de níveis é aplicada para se obter o resultado numérico da predição.

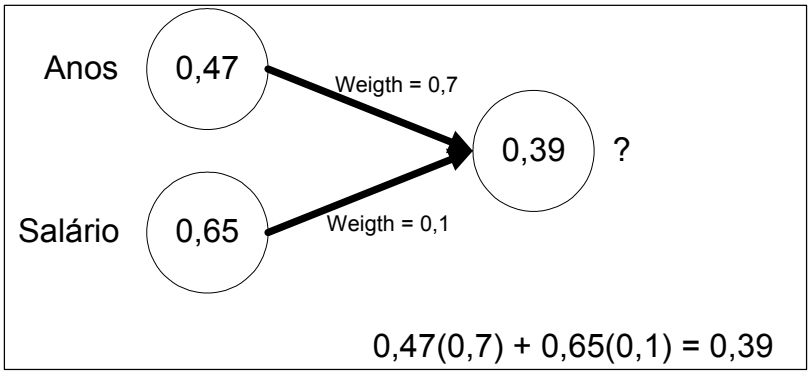

Figura 9 – Cálculo do valor da predição

No caso do exemplo, da Figura 9, que demonstra uma rede neuronal para prever o *churn* numa empresa de telecomunicações, se o resultado for próximo de 0, o registo é considerado como sendo de baixo risco de *churn*. Se o resultado for próximo de 1, o registo é considerado como sendo de alto risco de *churn*.

Os nódulos de *input* e de *output* são definidos pelo utilizador com base num determinado problema que se pretende solucionar. No entanto, à partida, os nódulos escondidos não têm definição e são as redes neuronais que durante o treino lhes atribuem significado, pressupondo no final dois problemas:

- É complicado acreditar no resultado de uma predição através de uma rede neuronal, se o significado implícito nestes nódulos escondidos não é perfeitamente entendido.
- Quando a predição é obtida no último nódulo e, consequentemente, é calculada a diferença ente a predição e o valor actual, como é possível corrigir o erro através dos nódulos escondidos para modificar o peso dos *links* que os conectam?

O significado dos nódulos escondidos não é sempre bem entendido, no entanto, a simples observação da actividade ou não dos nódulos permite retirar algum significado destes nódulos. O processo de aprendizagem para as redes neuronais foi definido para funcionar com os pesos nos *links* conectores na camada escondida. Uma boa metáfora para explicar este funcionamento é a analogia com uma operação militar, onde existem diversas linhas de comando [14]:

- **General**: responsável pela tomada de decisão de avançar ou retirar.

- **Majores**: responsáveis por informar o General.
- **Coronéis**: responsáveis por informar os Majores.
- **Soldados**: responsáveis por informar os Coronéis.

Esta hierarquia é muito semelhante à estrutura da rede neuronal constituída por várias camadas de nódulos escondidos e com um nódulo de *output*. Os conselhos provêm dos *inputs* dos nódulos escondidos. O peso do *link* conector corresponde ao valor de confiança que o General tem nos seus conselheiros; sendo a competência do conselheiro numa determinada matéria outro factor importante a ter em conta. Embora o conselheiro possa ser fidedigno, em determinada matéria este pode não ser especialista, pelo que não terá tanta influência na decisão do General.

Nesta analogia, o peso do *link* conector da rede neuronal é a confiança que o comandante tem nos seus conselheiros e o valor do nódulo representa o quão importante é a opinião deste conselheiro em determinada situação. Numa tomada de decisão, o General considera a validade e fiabilidade do conselho bem como o nível de conhecimento e confiança de cada conselheiro. A decisão do General é baseada no conjunto de informações dos conselheiros. Do mesmo modo, o nódulo de *output* irá construir a sua previsão tendo em conta os *inputs* de todos os nódulos com os quais está conectado. No caso das redes neuronais, esta decisão é alcançada multiplicando o peso do *link* pelo valor de *output* do nódulo e somando estes valores para todos os nódulos. Se a previsão estiver incorrecta, os nódulos com maior influência na tomada de decisão têm o seu peso alterado para que a previsão errada seja menos provável de acontecer numa próxima iteração. Este método de aprendizagem é muito similar ao que acontece quando a decisão tomada pelo General é incorrecta. A confiança que o General tem nos seus conselheiros decresce, quando estes lhe dão uma recomendação errada e ainda mais para aqueles que estão confiantes na sua recomendação, por outro lado, qualquer conselheiro que faça uma recomendação correcta vê a sua influência aumentar. Este processo de redefinição da influência propaga-se para os níveis inferiores. O procedimento de aprendizagem "*backpropagation*" é o mais usado, de fácil percepção, simplicidade e eficácia na grande parte dos problemas. O algoritmo "*backpropagation*" é usado no treino de redes multi-camada; baseia-se na aprendizagem supervisionada por correcção de erros, e é constituído por:

1. *Propagação*: depois de apresentado o padrão de entrada, a resposta de uma unidade é propagada como entrada para as unidades na camada seguinte, até à camada de *output*, onde é obtido o resultado da rede e consequentemente o erro é calculado.

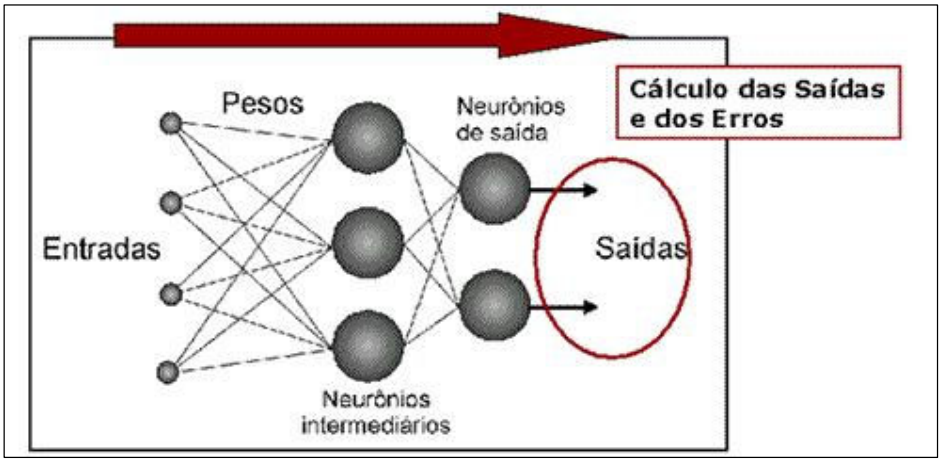

Figura 10 – Fase de Propagação

2. *Retro propagação* (*backpropagation*): desde a camada de saída até a camada de entrada, são realizadas as alterações nos pesos.

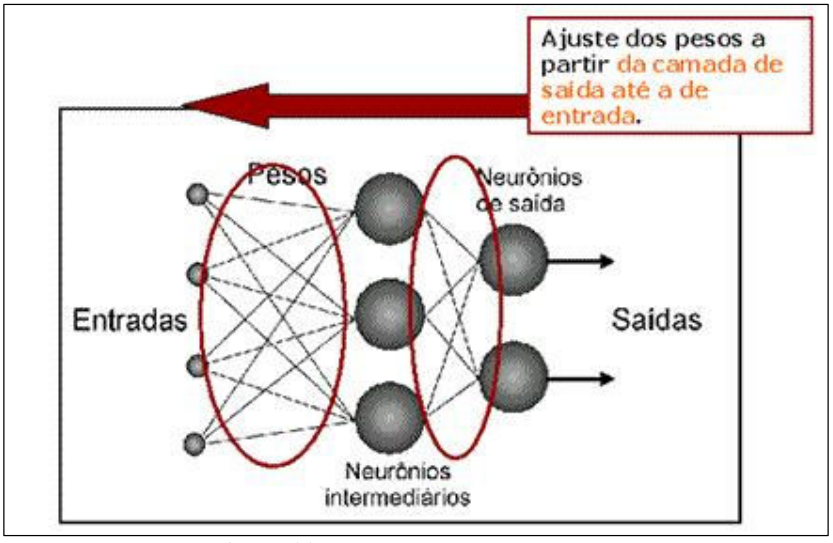

Figura 11 – Fase de Retro Propagação

Durante a fase de treino deve apresentar-se um conjunto formado pelo par entrada para a rede e valor desejado para a resposta da entrada. A saída será comparada ao valor desejado

e será calculado o erro global da rede, que influenciará na correcção dos pesos no passo de retro propagação. Apesar de não haver garantias que a rede forneça uma solução óptima para o problema, este processo é muito utilizado para apresentar uma boa solução para o treino de *Perceptons* multi-camada.

Existem duas outras arquitecturas de redes neuronais, também, utilizadas com frequência, designadamente:

- O modelo de KOHONEN é uma rede neural tipo "*feed forward*", não supervisionada, que usa simples neurónios adaptativos para receber sinais de um evento espacial. Segundo KOHONEN, o esquema básico do modelo faz com que neurónios da camada de saída disputem entre si a representação da informação apresentada aos neurónios de entrada [16].

– *Radial Basis Function* é uma rede neuronal supervisionada híbrida, baseando-se no método do vizinho mais próximo e na classificação das redes neuronais.

Como em todos os modelos preditivos, as redes neuronais requerem uma especial atenção para evitar o "*overfitting*". Neste aspecto, as redes neuronais são, relativamente, boas com dados de treino, mas não o são com novos dados; esta é, aliás, a grande problemática das redes neuronais, uma vez que é muito difícil entender como o modelo funciona. Enquanto as árvores de decisão ou os métodos do vizinho mais próximo podem atingir rapidamente 100% de eficácia de predição na bateria de dados de treino, as redes neuronais podem ser treinadas eternamente sem atingirem os 100%. Apesar de ser um facto interessante, os resultados obtidos na amostra de treino não são muito relevantes e não representam a eficácia da validade da base de dados.

Um dos métodos de controlo do "*overfitting*" nas redes neuronais é a limitação do número de conexões, uma vez que é representativo da complexidade do modelo, e modelos mais complexos aumentam a tendência de "*overfit*". Infelizmente, não existe nenhuma teoria/técnica para escolher o número ideal de conexões; uma bateria de teste pode ser usada para a validação do "*overfitting*", construindo redes neuronais numa parte da bateria de treino e usando a outra para detectar a eficácia da predição. O máximo valor da eficácia será obtido em determinado ponto do treino e à medida que o treino prossegue diminui, enquanto a eficácia do treino na base de dados irá continuar a aumentar. Os pesos dos *links* conectores para a rede devem ser guardados quando a eficácia atingir o seu pico.

2.3.5. Regressões Lineares

Por fim, das três técnicas em estudo, as regressões lineares são técnicas poderosas e muito comuns em estatística e mineração de dados, sendo utilizadas, normalmente, em processos de previsão. Existe uma variedade de diferentes tipos de regressões em estatística, no entanto, o principal objectivo comum a todos os tipos é estimar um valor condicional esperado. A forma mais simples da regressão é a regressão linear, um método estatístico para estimar a condicional (valor esperado) de uma variável Y, dados os valores de algumas variáveis X [17]. Em geral, a regressão trata da questão de estimar um valor condicional esperado. A regressão linear é chamada de linear porque se considera que a relação da resposta às variáveis é uma função linear de alguns parâmetros composta por um *predictor* e uma previsão. A relação entre os dois pode ser mapeada num espaço bidimensional em que os valores do *predictor* são mapeados no eixo do X e os valores da previsão no eixo do Y. O modelo de regressão linear pode ser visto como a linha que minimiza o rácio do erro entre o valor actual da previsão e o ponto da linha (previsão do modelo).

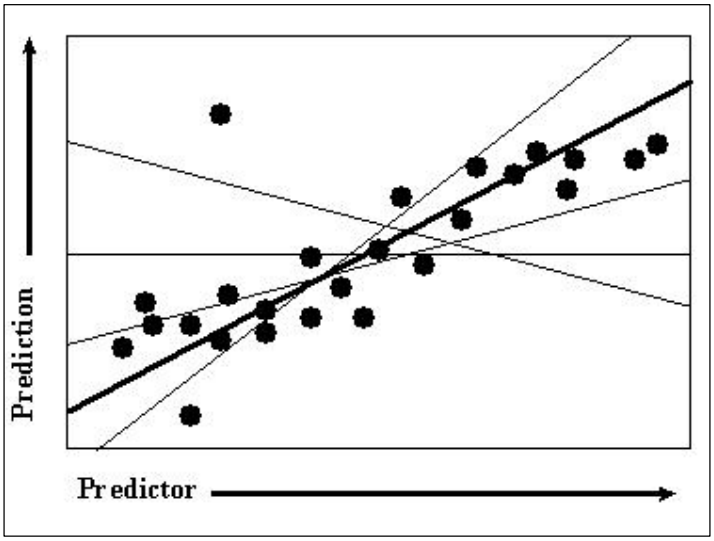

Figura 12 – A regressão linear.

O forma mais simples de regressão procura construir um modelo preditivo que é a linha que mapeia o valor do *predictor* com o valor da previsão. A equação será:

Previsão (Y) = a + b * *predictor* (X) **(4)**

Nos modelos preditivos, o truque consiste em encontrar o modelo que melhor minimize o erro. A forma mais comum para calcular o erro é o quadrado da diferença entre o valor previsto e o valor actual, calculado desta forma, os pontos mais afastados da linha terão um maior efeito na aproximação da linha para si, numa tentativa de reduzir o erro. Os valores de a e b, que minimizem o erro na equação de regressão, podem ser calculados directa e rapidamente a partir dos dados.

A regressão pode ser mais complexa do que a simples regressão linear, para encontrar um modelo melhor para determinados problemas de base de dados. Existem três modificações principais que podem ser introduzidas, nomeadamente:

1. A utilização de mais do que um *predictor*.
2. Aplicar transformações aos *predictors*.
3. Os *predictors* podem ser multiplicados e utilizados como termos na equação.

Podem ser aplicadas modificações para adaptar previsões de valores binários. A adição de *predictors* à equação linear pode produzir linhas complexas que têm em conta mais informação e, portanto, contribuem para uma melhor previsão; designada de regressão linear múltipla, podendo ser representada pela equação que se segue:

$$Y = a + b1(X1) + b2(X2) + b3(X3) + b4(X4) + b5(X5) \quad \textbf{(5)}$$

Esta equação (exemplo de utilização com cinco *predictors*) continua a descrever uma linha, mas num espaço hexadimensional, ao transformar os *predictors* utilizando o quadrado, o cubo ou a raiz quadrada é possível usar a mesma metodologia de regressão e criar modelos mais complexos que não podem ser representados por linhas. A isto se chama regressão não linear, onde um modelo de um *predictor* pode ser representado pela seguinte equação:

$$Y = a + b1\ (X1) \quad + b2\ (X12) \quad \textbf{(6)}$$

Em muitos casos reais, os analistas irão aplicar uma grande variedade de transformações nos dados para os treinar. Se os *predictors* não contribuírem para um modelo válido, os seus coeficientes na equação tenderão para zero podendo até ser removidos. Outra transformação aplicável aos valores dos *predictors* é a sua multiplicação. Por exemplo, um *predictor* criado pela divisão do salário/hora pelo salário mínimo poderá ser mais eficaz do que o *predictor* do salário/hora. Quando se tenta prever a resposta de um cliente com os

valores sim ou não, a forma *standard* de uma linha não funciona uma vez que existem apenas dois valores possíveis; o modelo será sempre igual, independentemente dos *predictors* utilizados ou dos dados. Tipicamente, nestas situações o valor do *predictor* é transformado para fornecer um modelo mais eficaz. Este tipo denomina-se de regressão logística sendo dos mais utilizados no meio, quando a variável *target* consiste num valor binário (*churn* ou não *churn*). Considerando k variáveis independentes no vector X = [X1,X2,…, Xk], a regressão logística descreve a probabilidade de ocorrência da variável Y, codificada como 0 ou 1. Se P(Y=1) = pi, então o modelo de regressão é definido pela equação apresentada na figura abaixo:

$$\log\left[\frac{p_i}{1-p_i}\right] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

$$\Pr(Y=1) = \frac{1}{1+\exp[-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k)]}$$

$$f(z) = \frac{1}{1+e^{-z}} \qquad (7)$$

$$\text{onde, } z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

Figura 13 – Derivação da fórmula matemática da Regressão Logística [17]

Nos modelos logísticos, o coeficiente *βk* descreve o relacionamento da variável independente com a variável dependente.

2.4. Limitações da utilização dos modelos Churn

Uma das grandes desvantagens do desenvolvimento de uma análise de *churn* é o custo associado e os investimentos necessários para armazenar, processar e inferir conhecimento a partir da informação, por este motivo deve ser realizada uma análise financeira prévia para avaliar se o investimento inerente a este tipo de projecto irá produzir lucro e benefício justificativos do investimento da empresa. Outra desvantagem, não menos importante, é o facto dos modelos de *churn* serem de aplicação arriscada quando não implementados e analisados por especialistas [8]. Além disso, os modelos *churn* necessitam de um

acompanhamento constante por parte de especialistas, isto porque, como a informação muda ao longo do tempo, implica que os modelos também mudem, existindo a necessidade de reajustes. Tendo em conta as desvantagens enunciadas, reveste-se de elevada importância o estudo de viabilidade e de suporte de um projecto *churn* antes da sua implementação, para avaliar os ganhos e perdas associados a um projecto deste tipo.

2.5. Casos de estudo

Ao longo desta extensa secção, é feita uma caracterização do conceito *churn*, dos sistemas de mineração usados na classificação de *churn* e das técnicas mais utilizadas nos sistemas de mineração de dados preditivos. De seguida, apresenta-se três casos de estudo que se relacionam com os modelos que serão desenvolvidos nesta dissertação e que serviram para confrontar quer opções, quer resultados.

Um desses casos de estudo foi um trabalho apresentado numa empresa de telecomunicações brasileiras intitulado "*Uma Análise de Cancelamento em Telefonia utilizando Mineração de Dados*" [19]. Neste estudo, analisou-se uma base real de dados de telecomunicações caracterizada por um segmento de clientes com elevado tráfego mensal e com tendência para abandonarem a empresa. Através de uma análise ao comportamento das variáveis, o autor inferiu que entre o oitavo e nono mês, existem variáveis que, por si só, conseguem explicar e classificar o cliente relativamente a *churn*. Esta constatação permitiu a redução da selecção inicial de 12 meses de dados para uma amostra mais reduzida. Os três métodos analíticos distintos utilizados, nomeadamente a evolução temporal da média standard, a evolução temporal do determinante da matriz de co-variância e a análise de auto valores, confirmaram a validade da redução da amostra, facilitando o custo do processamento no processo de análise. Para a classificação das variáveis, o autor recorreu a uma análise preditiva através de três variáveis sintetizadas: patamar, tendência e volatilidade, cada uma das características utilizadas na sintetização das novas variáveis demonstrando extrema eficiência na representação e separação dos grupos.

Na análise dos modelos não supervisionados utilizados, identificou-se uma movimentação dos *centroides*. Isto significa que, na ausência de uma modelagem controlada (classe *churn* e não *churn*) estes podem auxiliar na identificação dos clientes que mudaram de *cluster*, indicando mudança no perfil de consumo. O facto de ser possível a eliminação do último mês de amostra para a construção dos modelos demonstra a robustez e a eficiência

do modelo, assumindo particular importância para a realização de análises e consequente tomada de acções de retenção dos clientes em tempo útil.

O segundo caso de estudo analisado "*Applying data mining to Telecom churn management*" foi efectuado num operador de telecomunicações sem fios em Taiwan. O âmbito deste projecto foi o estudo do cliente *churn*, baseado numa amostra aleatória de 160000 clientes pós-pagos com subscrição mensal e activos há mais de três meses antes de 1 de Julho de 2001 [20]. Foram utilizados 11 meses de transacções, incluindo informação de facturação, detalhe de chamadas e informação de clientes (residente no CRM). As variáveis foram seleccionadas a partir de reuniões com os especialistas em telecomunicações, marketing, vendas, clientes e através de outros estudos anteriormente efectuados para identificar causas e sintomas anteriores ao facto de o cliente decidir deixar a empresa. A modelação utilizou técnicas de árvores de decisão com e sem segmentação e redes neuronais com *backpropagation*, para testar as diferenças entre os vários modelos e técnicas. Realizaram-se testes *T-test*, após avaliação do conjunto de dados disponíveis, e verificou-se que todos os modelos demonstram uma eficácia estável para os primeiros 6 meses de informação, existindo uma degradação acentuada no oitavo mês, que se assumiu como justificável pelo acontecimento do novo ano chinês. Os resultados comprovam que nos primeiros 6 meses, as redes neuronais com *backpropagation* são mais eficientes que as árvores de decisão e as árvores de decisão sem segmentação mais eficientes que as que têm segmentação. Esta segunda observação contrariou a teoria, segundo a qual a análise de segmentos de clientes é mais eficaz que a análise de toda a base de clientes. A validação da robustez dos modelos recorreu à utilização da técnica do LIFT, demonstrando que as redes neuronais e as árvores de decisão têm um elevado rácio de eficácia para os primeiros 10% de clientes *churn* previstos (valor 10).

O terceiro caso de estudo avaliado, com o título "*UMA APLICAÇÃO DE MINERAÇÃO DE DADOS NO GERENCIAMENTO DO CHURN EM SERVIÇOS DE BANDA LARGA DE TELECOMUNICAÇÕES*" [24], foi realizado numa grande empresa de telecomunicações a actuar no Brasil, com dados referentes a serviço de banda larga entre Janeiro e Dezembro de 2006, com o objectivo de encontrar a relação entre a qualidade de serviço e a taxa de abandono de clientes – *Churn*. O estudo focou-se nas causas voluntárias de abandono relacionadas com a qualidade do serviço prestado pelo operador, para obtenção de uma equação que represente a taxa de abandono em função da qualidade

de serviço, recorrendo à regressão múltipla. Este método de regressão foi escolhido por fornecer um meio de avaliar o grau e o carácter da relação entre as variáveis dependentes e independentes, através da variável estatística, a partir da variável independente. A ferramenta utilizada foi a SPSS pela sua facilidade de utilização, por estar amplamente documentada e por ser bastante utilizada na análise de regressão múltipla. Através da análise da matriz de correlação, verificou-se que as variáveis independentes não possuem valores superiores a 0.5 concluindo que não existe multicolinearidade (variáveis independentes que possuem relações lineares exactas ou aproximadamente exactas). Os valores do factor de inflamação de variância são bastante próximos de um, pelo que se conclui a não existência de multicolinearidade. Recorrendo ao ajustamento e à distribuição normal, verifica-se que existe uma distribuição consistente dos resíduos em relação à distribuição teórica, isto é, aproximam-se da curva normal. A análise dos dados obtidos demonstrou a existência de correlação entre os indicadores de qualidade apurados e a taxa de abandono. A partir de um conjunto de 12 variáveis, identificaram-se as quatro mais relevantes para a taxa de abandono: Tempo médio de reparação, Tempo Médio de Instalação, Tempo Médio de Prevenção e Taxa de Reparo; permitindo construir um conjunto de recomendações e acções de melhoria que contribuem directamente para o decréscimo do *Chur*n.

Nas empresas de telecomunicações, a previsão de clientes *churn* é extremamente difícil, devido não só ao elevado volume de dados, mas também à normal e frequente segmentação dos *Data Marts* o que dificulta a consolidação da informação necessária para classificação do cliente *churn*. Tomando os casos de estudo como exemplo, reveste-se de uma enorme importância, a exploração dos dados, a escolha das variáveis e a escolha das técnicas mais adequadas para obter os modelos mais eficientes e que apresentem os melhores resultados.

3. Metodologia e Preparação dos processos de Mineração de Dados

Os vários modelos preditivos que serão expostos nesta dissertação visam a extracção de padrões e a descoberta de conhecimento sobre informação de histórico de clientes de uma empresa de telecomunicações, com o objectivo de classificar e prever quais os clientes denominados *churners*, com elevada probabilidade de deixarem a empresa para uma outra concorrente.

A estrutura base dos modelos costuma estar baseada na metodologia *"KDD, o processo de identificação de válidos, novos, potencialmente úteis e compreensíveis padrões em dados" [3]*. No entanto, e devido à sua flexibilidade e iteração no desenvolvimento das várias etapas, optou-se pelo uso da CRISP-DM [4], dado esta metodologia apresentar-se como uma abordagem/modelo de apoio ao desenvolvimento de projectos de mineração de dados, focada nas questões e análises técnicas dos requisitos de negócio. Existem outras opções de metodologia apresentadas pelos fornecedores de software de mineração de dados, sendo uma das mais utilizadas a *SEMMA* (Sample, Explore, Modify, Model, Assess) da *SAS Enterprise Miner*. A *SEMMA* baseia-se no processo principal da realização da mineração de dados [18] a partir de uma representação estatística de uma amostra de dados. Esta facilita a aplicação da exploração estatística e das técnicas de visualização, a selecção e transformação das variáveis preditivas mais significantes e a modelação das variáveis para previsão e confirmação da eficácia do sistema. A metodologia CRISP-DM foi a seleccionada pelo cliente. Embora, estas duas metodologias sejam muito parecidas, a CRISP-DM foi preferida por se ter decidido dar mais ênfase às questões e análises técnicas de requisitos de negócio.

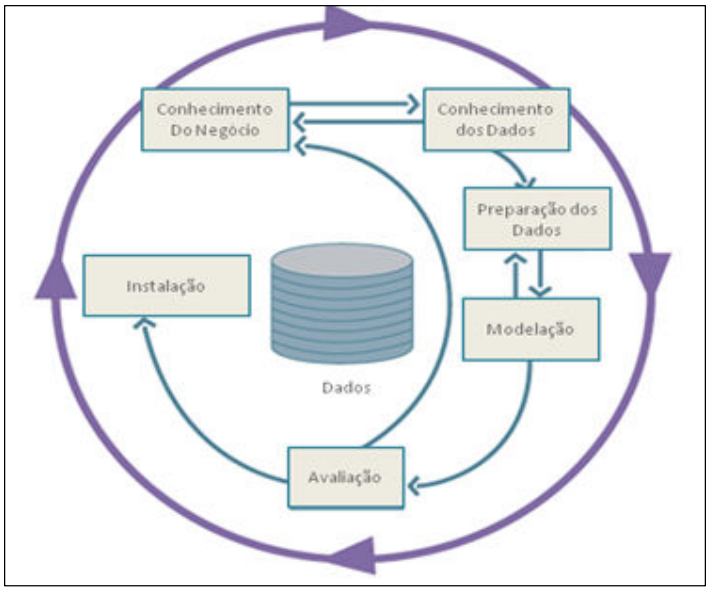

Figura 14 – Metodologia *CRISP-DM* [4]

A metodologia CRISP-DM é composta por 6 fases, a saber: Conhecimento do Negócio, Conhecimento dos Dados, Preparação dos Dados, Modelação, Avaliação e Instalação Todas estas etapas serão descritas ao longo desta e da secção seguinte, com a descrição do processo de caracterização e identificação de clientes *churn* em empresas de telecomunicações.

3.1. Conhecimento das Regras de Negócio

Esta fase centra-se sobretudo na compreensão dos objectivos e requisitos do projecto na perspectiva do negócio. Este conhecimento é, posteriormente, convertido na definição de um problema de mineração de dados.

No ano transacto a empresa de telecomunicações alvo deste trabalho obteve uma taxa anual de *churn* na ordem dos 4,8%; de acordo com os resultados do primeiro trimestre do presente ano a taxa de *churn* foi de 0,9%. Tendo em consideração estes dados/valores, o departamento de Marketing e Retenção decidiu implementar um projecto que permitisse identificar os clientes de valor com maior probabilidade de abandono e implementar acções de retenção que promovam a fidelização e lucro provenientes destes clientes.

Para a implementação deste projecto é necessário fazer um estudo de negócio que permita identificar pormenorizadamente o problema, para que, posteriormente, este possa ser transformado num problema de mineração de dados.

Por esse motivo, existem três conceitos de negócio a considerar:

1. *Produto*: item comercializado pela Organização;
2. *Cliente*: indivíduo que detém o produto no âmbito de uma conta;
3. *Conta*: abstracção comercial que representa um conjunto de produtos (ou apenas um).

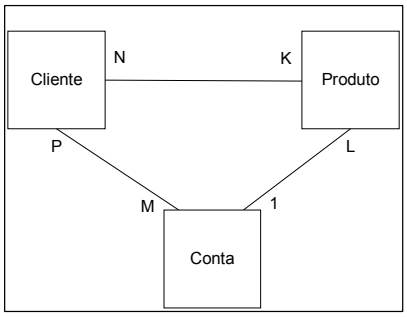

Figura 15 – Modelo dos Conceitos de Negócio

No mundo real um indivíduo poderá ter três estatutos distintos na sua relação com a empresa de telecomunicações, nomeadamente:

1. *Decisor*, individuo que toma a decisão de adquirir o produto.
2. *Detentor*, individuo que adquire o produto.
3. *Utilizador*, individuo que utiliza e explora o produto.

Habitualmente, a organização só consegue identificar o detentor, porque só recolheu informação acerca do indivíduo no momento da aquisição do produto. Contudo, em termos de marketing uma empresa de telecomunicações deve ter em consideração os diferentes tipos de clientes, dado tratar-se de um projecto de modelação para o departamento de Marketing de Retenção. Sendo assim, neste documento, a referência a 'cliente' corresponde ao indivíduo escolhido (decisor, detentor ou utilizador). Normalmente, as análises de mineração de dados são geralmente realizadas à volta do caso 'cliente'. Considerando que o cliente tem vários produtos no âmbito de uma ou mais contas, é necessário realizar um trabalho que permita a construção de uma representação

holística do cliente a partir desta realidade, ou seja, o perfil de um caso deve conter a informação pessoal do cliente, a informação respeitante ao(s) produto(s) a que este esteja associado e eventualmente, da(s) conta(s) a que este pertença. O caso poderá representar, no entanto, qualquer um dos três conceitos acima mencionados (Produto, Cliente e Conta). A escolha interfere, essencialmente, na definição do perfil de um caso e na definição da acção do estudo de modelação. Refira-se que na definição do perfil deverá ter-se em consideração que as características de um 'caso produto' não serão idênticas as características de um 'caso cliente' e que no estudo de modelação a acção é entendida como a representação de um acto realizado pelo caso. Os produtos podem ter papéis diferentes no estudo de modelação, sendo importante a definição das relações cliente - produto para se compreender quais as implicações que tem cada possível definição de caso, tanto na definição da acção como na definição do perfil. Um cliente (ou um grupo de clientes) que tenha vários produtos perfil pode ter vários produtos do mesmo tipo e/ou vários produtos de tipos diferentes.

Neste estudo, foram construídas várias hierarquias de produtos, com o intuito de facilitar a inclusão das características dos produtos no perfil do cliente. Assim, em vez de uma grande quantidade de tipos de produto perfil, optou-se por agregar esses tipos de produto em grupos definidos de acordo com as suas próprias características. Quando o caso é um cliente, este possui um ou vários produtos, logo a quantidade de características presentes no perfil do produto quando associadas ao cliente deverão ser do mesmo número. Note-se que, neste estudo, quando se refere que um cliente pode ter um ou vários produtos, na abstracção do negócio, significa que um cliente pode ter varias contas e que uma conta tem um produto, deste modo, uma conta corresponde a um contrato.

Considerando que o objectivo de estudo é criar um modelo preditivo que possibilita a retenção de clientes, foram identificadas as duas possibilidades: uma centrada na conta e outra no cliente. No caso da primeira, os aspectos menos positivos serão os que se seguem:

- O mesmo indivíduo irá aparecer em várias linhas da tabela de desenvolvimento, logo, a sua informação será repetida em várias linhas.
- O indivíduo pode ter contas com e sem evento de modelação, logo, este caso poderá ser considerado tanto positivo como negativo.

- O indivíduo pode possuir vários serviços com evento de modelação, o que tornará um caso positivo várias vezes com perfis diferentes, o que terá implicações nas ofertas.

Já no segundo caso, as desvantagens são as seguintes:

- Uma linha da tabela de desenvolvimento irá ter informação sobre o cliente, sobre cada uma das contas e, eventualmente, terá informação agregada de todas as contas (isto pode levar a um perfil muito extenso).
- O número de variáveis presentes no perfil não pode variar em função do número de contas que cada cliente tem.

Tendo em conta, os factores acima descritos referentes ao 'caso conta' e ao 'caso cliente', neste estudo, a opção adoptada para a modelação de identificação e classificação do cliente *churn* foi a segunda.

Relativamente ao estudo de regras de negócio, é necessário ter em consideração as seguintes questões:

- Duas *tabelas*: uma de desenvolvimento para se desenvolver o modelo e uma tabela de produção para utilização do modelo;
- *Caso*: é a entidade representada por uma linha da tabela de desenvolvimento e de produção, sendo que a definição do caso pode ser distinta para cada uma das tabelas;
- *Segmento*: é o conjunto de casos seleccionados para alvo do estudo de modelação;
- *Perfil*: é o conjunto de características do caso; é representado na tabela de desenvolvimento e de produção pelas variáveis independentes;
- *Acção*: é o acto realizado pelo caso e que será alvo do estudo de modelação. Suponha-se o seguinte conceito de um estudo de modelação: os clientes que, nos três meses seguintes ao momento da utilização do modelo, vão enviar uma carta à empresa a solicitarem o cancelamento de um serviço; neste caso, a acção é o envio da carta por parte de um cliente à empresa a solicitar o cancelamento de um serviço e o conceito do estudo de modelação é a acção mais o período de *output*, que no exemplo atrás referido, são os três meses durante os quais a carta de cancelamento pode ser enviada.

– *Conceito*: é representado na tabela de desenvolvimento pela variável dependente, pelo que não pode ser representado na tabela de produção. Em primeiro definir-se-á a acção, e só mais tarde será possível chegar ao conceito. Considere-se uma acção que representa apenas dois actos distintos, o cancelamento e o não cancelamento de um produto. Deste modo, a acção positiva é a realização da acção e a acção negativa é a não realização da acção. Um caso que esteja associado a uma acção positiva é um caso positivo e um caso que esteja associado a uma acção negativa é um caso negativo.

Os conceitos relacionados com os limites temporais utilizados no estudo de modelação são vários e devem ser considerados em dois momentos: a utilização e o desenvolvimento do modelo.

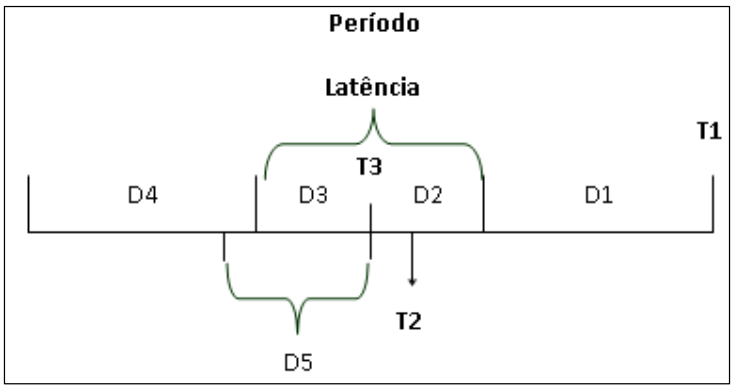

Figura 16 – Limites temporais na utilização do modelo de previsão

T1 – instante final do período de *output*.

D1 – duração do período de *output*.

D2 – duração do período de pro-actividade.

T2 – instante em que a tabela de produção já está construída, já se utilizou o modelo e já se realizou a campanha de marketing.

T3 – instante em que se inicia a construção da tabela de produção.

D3 – duração do período de indisponibilidade.

D4 – duração do período de *input*.

D5 – duração do período de desfasamento.

O período de *output* é a janela de tempo utilizada para construir a variável de *output* do modelo (variável dependente). O período de pro-actividade é a janela de tempo que a empresa tem para utilizar os resultados do modelo com capacidade de influência, de acordo com os objectivos que pretende atingir. Note-se que o período de pro-actividade poderá não estar completamente disponível, uma vez que é ainda neste período que se utiliza o modelo (*T2*-*T3*), ou seja, quanto mais tarde acontecer *T2*, menor será o período de pro-actividade útil. Na prática, o período de pro-actividade vai ser extensivo de acordo com o modo de armazenamento dos dados no DW. Define-se o período de indisponibilidade como o tempo que se deve recuar para se chegar à data dos dados mais recentes disponíveis no DW. Para a construção do perfil, normalmente (para não dizer nunca), não existem dados num DW relativos a '*hoje*'; no melhor cenário podem existir alguns dados relativos a '*ontem*'. O período de *input* é a janela de tempo utilizada para a construção das variáveis de *input* do modelo (variáveis independentes), e o período de latência corresponde ao tempo entre o fim do período de *input* e o início do período de *output*, resultando da soma dos períodos de pro-actividade e de indisponibilidade. Já o período de desfasamento não é mais do que o período de indisponibilidade da tabela usada na determinação dos casos. É necessário distinguir entre os períodos de indisponibilidade das tabelas utilizadas para a construção do perfil (*D3*) e da tabela utilizada para a determinação dos casos (*D5*), porque o valor de *D5* é calculado antes do valor de *D3* durante o estudo de modelação.

Na figura 16, o instante final dos períodos de desfasamento e de indisponibilidade são sempre coincidentes, mas o instante inicial destes períodos pode acontecer por qualquer ordem cronológica. Dependendo dos períodos de indisponibilidade individuais das tabelas envolvidas, existe, no entanto, a possibilidade do período de *input* e do período de desfasamento se intersectarem.

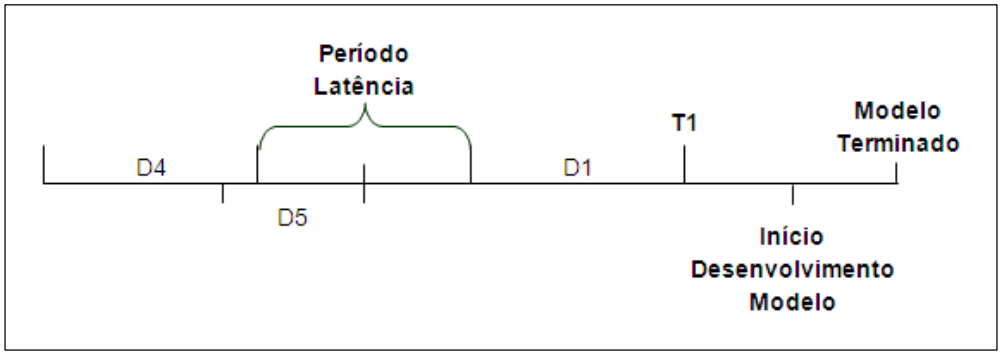

Figura 17 – Limites temporais no desenvolvimento do modelo de previsão

Para a implementação de um processo de desenvolvimento e utilização de modelos contínuos no tempo, apenas se deverá garantir que o instante em que o modelo está terminado (figura 17) nunca acontece depois da tabela de produção estar terminada, o que acontece entre *T3* e *T2* (figura 16). Por este motivo, é importante que não se utilizem os dados relativos ao período de latência.

Após a definição dos conceitos de negócio e limites temporais associados, é necessário proceder-se à caracterização e identificação do cliente *churn*.

3.2. Caracterização do Cliente Churn

O *Churn* é definido como a acção de cessar o serviço principal a pedido de um cliente ou por incumprimento da sua parte de alguma questão contratual. Refira-se que, neste caso, um serviço é equivalente a um contrato que, por sua vez, corresponde a uma conta. Os principais serviços considerados para a cessação de contrato são: o serviço móvel (Pré-pago e Pós-pago) e o serviço fixo (ADSL e Dial Up). O serviço móvel pré-pago refere-se a chamadas de voz, SMS e MMS associados a um carregamento periódico obrigatório e o pós-pago corresponde igualmente a chamadas de voz, SMS e MMS com uma assinatura mensal de preço com base fixa. Quanto ao serviço fixo, pode distinguir-se a *internet* ADSL e DIAL UP, sendo que a primeira é uma tecnologia de transmissão rápida de dados por cabo e a segunda é uma transmissão de dados por conexão à linha telefónica.

Como referido, a cessação de um serviço pode ser apresentada por dois motivos, a saber:

1. *Cessado por incumprimento contratual*: Suspensão do contrato devido ao não pagamento das facturas. O cliente entra num processo multi-etapa até ser declarado como não pagador e a sua subscrição será cessada. Os procedimentos

a adoptar na cessação do contrato diferem de pós para pré-pago. Deste modo, para os serviços Pós-Pagos, o cliente tem, inicialmente, as chamadas efectuadas barradas por um período de graça de 14 dias devido ao não pagamento das facturas; posteriormente se o não pagamento se prolongar, o cliente é submetido a um novo processo de barramento ficando impossibilitado de efectuar e receber chamadas durante o período de 30 dias, após o qual o serviço é cessado. No que concerne aos Pré-pagos, existe, normalmente, a obrigatoriedade de renovação anual do serviço. Caso esta renovação não seja efectuada, procede-se ao barramento de efectuar chamadas por um período de graça de três meses, após o qual o serviço é cessado.

A informação acima descrita pode ser identificada no *Data Mart* F_CVM_ACCOUNT_SNAP_M recorrendo ao atributo *Account Status*. Este atributo pode assumir vários valores, nomeadamente: 2 - Cessado; 448 - Confirmação de cessão por não pagamento; 450 - Confirmação de pagamento e consequente anulação de barramentos; 451 e 3993 - A dívida é considerada não colectável; quando a variável *Sub Request* possui o valor 22 - Identificação de cessação por não pagamento.

2. <u>*Cessado a pedido do cliente*</u>: É u*m*a das causas mais comuns para a desconexão de um serviço e é designada de desconexão voluntária, normalmente, classificada *a posteriori*, com o objectivo de analisar e controlar o *churn*. As classificações mais comuns são as seguintes: o cliente deixa o país, ou muda para um concorrente devido ao preço, serviço ou oferta, ou não necessita mais do serviço.

   A informação acima descrita pode ser identificada no *Data Mart* F_CVM_ACCOUNT_SNAP_M recorrendo ao atributo *Account Status*. Este atributo pode assumir vários valores, nomeadamente: 2 - Cessado; 448 - Confirmação de cessão por não pagamento; 450 - Confirmação de pagamento e consequente anulação de barramentos; 451 e 3993 - A dívida é considerada não colectável; quando a variável *Sub Request* possui o valor 21 - Identificação de cessação.

Relativamente à cessação, é de referir ainda o caso de *churn* interno que, neste caso de estudo, clientes que cessam o serviço, para fazer *upgrade* ou *downgrade* de serviço na mesma linha de produto não são tratados como clientes que cessam.

Após a caracterização de como identificar o cliente *churn* nos sistemas operacionais, é chegada a etapa final da análise com a definição do conceito ou *target*. *O Target* é o resultado de uma acção despoletada pelo cliente. Nesse sentido, a acção pode ser caracterizada pelo acto de mudança efectuada pelo cliente, onde o *target* é uma variável binária que representa a acção positiva ou a acção negativa por parte do cliente:

- 1 – O cliente vai deixar a empresa (*Churn*) – Acção Negativa.
- 0 – O cliente vai permanecer na empresa – Acção Positiva.

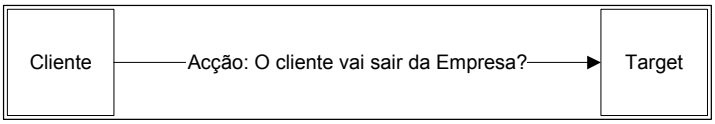

Figura 18 – Acção do cliente

3.3. Definição dos objectivos da mineração

A mineração de dados ou prospecção de dados pode ser definida como o "processo de extracção e apresentação de conhecimento novo, não detectável anteriormente, seleccionado a partir de base de dados para a tomada de decisões futuras" [11]. Após finalizado o processo de analise deste caso de estudo, o objectivo desta mineração é prever quais os clientes *churn* e o seu perfil. A previsão, deverá ser efectuada com uma antecedência de três meses, para os clientes com o serviço *Mobile* Pós-pago, contactáveis, com antiguidade superior a 6 meses e pertencentes ao segmento *star 5* (segmento de valor mais elevado). Para atingir o objectivo final, serão utilizados, o histórico de tráfego de rede, valores de facturação, tarifários, produtos e informação demográfica. A escolha deste segmento, enunciado em cima, obedeceu aos requisitos de negócio, justificando-se da seguinte forma:

- Clientes com o Serviço Mobile: Serviço com maior retorno financeiro.
- Clientes pós-pagos: Clientes que possuem um processo de fidelização com retroactivos financeiros em base mensal.

- Clientes pertencentes ao segmento *star 5*: Os clientes deste segmento são considerados pela empresa como os clientes de elevado valor e os mais importantes numa primeira fase de análise dado que a sua perda implica elevados custos e perdas financeiras.
- Clientes com antiguidade superior a 6 meses: Dado que a empresa está num país vocacionado para o turismo possuindo muitos clientes sazonais com um tempo médio de vida entre um e três meses, não sendo estes, clientes importantes para alvo de campanhas de retenção.
- Clientes contactáveis: do ponto de vista do negócio não faz sentido realizar-se um processe de mineração de dados para se prever clientes *churn* não contactáveis, uma vez que o resultado esperado do projecto é fornecer ao marketing de retenção a lista de clientes *churn* e respectivos contactos para uma abordagem de retenção.

Após a definição dos objectivos da mineração, a etapa de análise e requisitos de negócio fica concluída, sendo que as próximas etapas serão a descrição de todo o desenvolvimento associado aos modelos e resultados da mineração de dados para previsão do cliente *churn*.

4. Identificação do Cliente *Churn*

4.1. Conhecimento dos Dados

Uma das etapas mais relevantes para se efectuar a identificação e caracterização de uma situação de *churn* passa, obrigatoriamente, pela análise e conhecimento efectivo dos dados que suportarão esse processo. Esta etapa começa com uma selecção inicial de uma amostra significativa dos dados disponíveis, prosseguindo com outras actividades como a identificação de problemas relacionados com a qualidade e a compreensão dos dados e a detecção de subconjuntos interessantes para a formulação de hipóteses referentes a peças de informação eventualmente escondidas.

4.1.1. Colecção de um grupo de dados iniciais

Nesta fase de conhecimento dos dados, além da identificação das fontes de informação, é necessário começar a preparar o processo de extracção que, com a periodicidade definida (neste caso mensal), irá "refrescar" a informação. Neste processo foram identificadas várias fontes distintas, na sua generalidade *data marts*. De referir:

- F_CVM_INVOICE_M: uma tabela de factos do *data mart* de facturação, com informação sobre as facturas associadas às contas pós-pagas, com periodicidade mensal;
- F_CVM_NETWORK_EVENT_M: uma outra tabela de factos, do *data mart* de tráfego, com informação dos eventos de chamadas de todas as contas com periodicidade mensal;
- F_CVM_ACCOUNT_SNAP_M: uma tabela de factos do *data mart* de *Customer Value Management* (CVM), com os segmentos dos clientes com periodicidade mensal;

e, por fim, uma tabela de dimensão, associada a todos os *data marts* da empresa de telecomunicações, nomeadamente:

- D_PARTY_KEY: que contém informação dos vários clientes com periodicidade diária.

Esta informação foi reunida numa única tabela de desenvolvimento, designada de D_SOURCE_CHURN (em produção P_SOURCE_CHURN), contendo todos os atributos que caracterizam cada um dos *data marts* (a sua descrição pode ser encontrada na secção 4.1.2 e 4.2.1). O processo de extracção e carregamento da informação das diferentes fontes na tabela de desenvolvimento foi realizado por um procedimento em PL/SQL que será executado todos os meses, após a actualização dos *data marts*.

4.1.2. Descrição dos dados

Na descrição dos dados foi feita a identificação dos atributos e a volumetria dos dados que constituem cada tabela. O número total de contas da empresa em estudo é de aproximadamente 10 milhões, não correspondendo este valor ao número de contas activas. Assim, é esperado que cada uma das seguintes tabelas contenha cerca de 10 milhões de registos, mas no caso em que as contas estão inactivas os atributos da tabela estarão vazios. De seguida, são apresentados os atributos constituintes de cada uma das tabelas referidas anteriormente:

- F_CVM_INVOICE_M: *esta tabela integra* 23 medidas, agrupadas mensalmente por conta:
  - COUNT_VOICE_OUT_NAC
  - COUNT_VOICE_OUT_NT
  - COUNT_VOICE_OUT
  - COUNT_VOICE_IN_ROAM
  - COUNT_VOICE_OUT_ROAM
  - COUNT_OUT_SMS
  - COUNT_OUT_DATA
  - AIRTIME_VOICE_OUT_NAC
  - AIRTIME_VOICE_OUT_INT
  - AIRTIME_VOICE_OUT
  - AIRTIME_VOICE_IN_ROAM
  - AIRTIME_VOICE_OUT_ROAM
  - IN_USAGE_ROAM
  - OUT_USAGE_VOICE_NAC
  - OUT_USAGE_VOICE_INT

  - OUT_USAGE_VOICE
  - OUT_USAGE_ROAM
  - OUT_USAGE_SMS
  - OUT_USAGE_EVENTS
  - OUT_USAGE_DATA
  - MODULES
  - RENTS
  - OPENING_BALANCE
- F_CVM_NETWORK_EVENT_M: *esta tabela é* composta por 15 medidas agrupadas mensalmente por conta:
  - TAC_MOST_USED
  - TAC_LAST_USED
  - FIRST_IN_CALL_DATE
  - LAST_IN_CALL_DATE
  - FIRST_OUT_CALL_DATE
  - LAST_OUT_CALL_DATE
  - LAST_CALL_CHARGED_DATE
  - LAST_CALL_FREE_DATE
  - LAST_SMS_IN_DATE
  - LAST_SMS_OUT_DATE
  - LAST_SESSION_DATE
  - LAST_RECHARGE_DATE
  - END_OF_MONTH_NO_BALANCE
  - LAST_3M_USAGE_AVG
  - PREV_3M_USAGE_AVG

O número de clientes aproximado da empresa é de cerca de 6 milhões, o que faz com que as tabelas que se encontram ao nível do cliente apresentem também um número de registos igual. Neste caso, a tabela será constituída pelos seguintes atributos:

- D_PARTY_KEY: é uma tabela de dimensão de caracterização do cliente composta por 18 atributos:
  - DW_PARTY_ID

  - PARTY_ID
  - PARTY_TITLE
  - PARTY_CONTACT_FAX
  - PARTY_CONTACT_EMAIL
  - PARTY_GENDER
  - PARTY_BIRTHDAY
  - PARTY_MARITAL_STATUS
  - PARTY_PROFESSION
  - PARTY_ACTIV_DATE
  - PARTY_CESSATION_DATE
  - PARTY_HOME_ADDRESS_STREET
  - PARTY_HOME_ADDRESS_CITY
  - PARTY_HOME_ADDRESS_PO_BOX
  - PARTY_HOME_ADDRESS_EMIRATE
  - PARTY_HOME_ADDRESS_UNIT_NO
  - PARTY_SERVICE_LANGUAGE
  - PARTY_BILL_LANGUAGE

Existe uma outra tabela, desnormalizada por conjugar informação de contas e de clientes, na qual se podem encontrar também cerca de 10 milhões de registos, ao nível do cliente. Deste modo, para os casos em que um cliente possua duas ou mais contas, existirá um registo por cada uma dessas contas. Tal significa que esse cliente irá surgir na tabela tantas vezes quantas as contas que ele possui.

- F_CVM_ACCOUNT_SNAP_M: *Data Mart* composto por 41 medidas agrupadas mensalmente ao nível da conta.
  - CS_TOTAL_CALLS
  - CS_TOTAL_AIRTIME
  - CS_COMPLAIN_CALLS
  - CS_OPEN_COMPLAINTS
  - CS_INFORMATION_CALLS
  - CS_TT_AVERAGE_TIME
  - CS_SERVICE_CALLS
  - ACCOUNT_DW_REGION_ID

- ACCOUNT_NUMBER
- ACCOUNT_ACTIVATION_DATE
- ACCOUNT_CESSATION_DATE
- ACCOUNT_STATUS_ID
- ACCOUNT_STATUS_DATE
- ACCOUNT_CVM_STATUS_ID
- ACCOUNT_DW_PRODUCT_ID
- ACCOUNT_DW_PRODUCT_GROUP_ID
- ACCOUNT_CAP_LIMIT
- ACCOUNT_DW_RATE_PLAN_ID
- ACCOUNT_CESS_REASON_ID
- ACCOUNT_CESS_SUB_REQUEST_ID
- DW_PARTY_ID
- DW_PARTY_CATEGORY_ID
- PARTY_PAYM_METHOD
- ACCOUNT_ACTIV_STOR_MEDIUM
- PARTY_NATIONALITY
- DW_PARTY_TYPE_ID
- ACCOUNT_RP_TYPE
- ACCOUNT_STAR_SEGMENT_ID
- ACCOUNT_ROAMING_SEGMENT
- ACCOUNT_DATA_SEGMENT
- ACCOUNT_SMS_SEGMENT
- ACCOUNT_MMS_SEGMENT
- ACCOUNT_ACTIVITY_SEGMENT
- ACCOUNT_IDD_SEGMENT
- ACCOUNT_STD_SEGMENT
- ACCOUNT_RENEWAL_DATE
- ACTIVE
- CHURN
- GROSS_ADD
- GRACE_PERIOD
- DORMANT_ACTIVITY

4.1.3. Exploração dos dados

Nesta fase do estudo realizou-se uma análise da qualidade da informação para perceber quais os atributos a agregar, os campos a serem removidos, por não adicionarem valor, e o nível de redundância e duplicação de atributos. A redundância ocorre quando existem diferentes atributos com a mesma informação ou quando os atributos podem ser calculados a partir de outros já existentes, em "tempo real", sem prejudicar o desempenho do modelo. Exemplo disso é o atributo AIRTIME_OUT_TOT (tempo total de chamadas originadas) que representa a soma dos atributos seguintes:

- AIRTIME_OUT_FIX: Tempo total de chamadas para MSISDNs da rede fixa.
- AIRTIME_OUT_MC: Tempo total de chamadas para MSISDNs pertencentes aos *Major Countries*.
- AIRTIME_OUT_WRLD: Tempo total de chamadas para MSISDNs pertencentes a outros países que não os *Major Countries*.
- AIRTIME_OUT_OTMB: Tempo total de chamadas para MSISDNs de outras operadoras do mesmo país.

O processo de análise realizado integrou, também, a identificação de padrões ou de valores por omissão (*default*) - representativos de falta de informação ao invés de informação relevante. Neste estudo não se identificaram quaisquer atributos nestas condições.

4.1.4. Verificação dos dados

Após concluir a exploração dos dados descrita no ponto anterior, foi necessário proceder à sua validação no sentido de garantir que todos os dados foram correctamente escolhidos das diferentes fontes. A validação da sua relevância para o caso em estudo e identificação da existência de erros, bem como a frequência com que estes acontecem, é, como se sabe, crucial para qualquer processo de mineração de dados. A existência de valores desconhecidos (nulos) é assumida como um erro que pode ter origem num sistema informático ou na intervenção humana. Neste caso de estudo, optou-se por eliminar ocorrências para as quais as variáveis não têm valor, dado não existir um valor padrão que possa ser aplicado correctamente a estas situações. O objectivo subjacente a esta técnica é garantir que as variáveis significativas do modelo contenham algum valor. Tal justifica-se

porque não existe nenhum algoritmo que descubra padrões através de variáveis vazias e, como tal, não é possível construir modelos com significado.

4.2. Preparação dos Dados

A fase de preparação de dados incluiu todas as actividades necessárias para a construção do conjunto de dados que constituem a fonte do modelo. Fazem ainda parte desta fase, as tarefas de selecção dos atributos (perfil) e de transformação e de limpeza dos dados. As tarefas de preparação de dados são susceptíveis de ser executadas várias vezes, sem uma ordem pré-estabelecida.

4.2.1. Selecção dos dados

A colecção dos dados disponível obrigou, numa primeira fase, à modelação da informação e, posteriormente, à implementação de um processo adicional periódico de extracção e de transformação dos dados. A responsabilidade do carregamento dos dados para os modelos de mineração foi da ferramenta escolhida. A estratégia adoptada para a selecção do volume de informação consistiu no carregamento de treze meses completos de dados, ao qual se adicionaram uma "fotografia" do estado do cliente no décimo quarto mês. Isto acontece por se tratar de uma operadora de telecomunicações com grandes volumes de dados, em que o tempo necessário para a colheita, tratamento e disponibilização da informação ocupa grande parte do décimo terceiro mês.

A selecção dos atributos é uma questão pertinente na mineração de dados, e consiste em encontrar um subconjunto de dados no qual se aplica o algoritmo de aprendizagem [21]. Os algoritmos de mineração de dados normalmente utilizam amostras pequenas (normalmente inferiores a 20000 registos) com um número restrito de atributos (usualmente em número inferior a 30). Por isso, a selecção de uma amostra reduzida trouxe as seguintes vantagens:

- A selecção de atributos melhorou o desempenho do modelo, dado que grande parte dos algoritmos demonstra problemas de desempenho quando trabalham com um elevado número de atributos.

- O pequeno número de atributos facilitou a interpretação e o conhecimento induzido através dos algoritmos de mineração.
- Em casos como este, em que existiu alto custo na selecção dos dados, os métodos de selecção de atributos diminuíram o custo da aplicação.

No que diz respeito aos atributos, após várias reuniões com o cliente, as variáveis escolhidas para a construção do modelo foram as seguintes:

- F_CVM_ACCOUNT_SNAP.DW_PARTY_ID: Identificador único de cliente.
- F_CVM_ACCOUNT_SNAP.CS_COMPLAIN_CALLS: Número total de chamadas efectuadas para o *Contact Center* para reclamações.
- F_CVM_ACCOUNT_SNAP.CS_OPEN_COMPLAINTS: Número total de reclamações em aberto associadas ao cliente.
- F_CVM_ACCOUNT_SNAP.CS_INFORMATION_CALLS: Número total de chamadas de pedidos de informação efectuadas pelo cliente.
- F_CVM_ACCOUNT_SNAP.CS_TT_AVERAGE_TIME: Tempo médio de espera associado à resolução de cada reclamação.
- F_CVM_ACCOUNT_SNAP.CS_SERVICE_CALLS: Número total de chamadas efectuadas para os serviços do *Contact Center*.
- F_CVM_ACCOUNT_SNAP.ACCOUNT_ACTIVATION_DAYS: Data de activação da conta (serviço).
- F_CVM_ACCOUNT_SNAP.PARTY_PAYM_METHOD: Método de pagamento associado à conta do cliente.
- F_CVM_ACCOUNT_SNAP.PARTY_NATIONALITY: Nacionalidade do cliente.
- F_CVM_ACCOUNT_SNAP.CHURN: Identificador de *churn*.
- F_CVM_ACCOUNT_SNAP.GRACE_PERIOD: Indicador quando o cliente ultrapassou a data limite de pagamento da factura.
- F_CVM_INVOICE_M.COUNT_VOICE_OUT_NAC: Número total de chamadas classificadas como *Calls Outgoing National*.

- F_CVM_INVOICE_M.COUNT_VOICE_OUT_NT: Número total de chamadas classificadas como *Calls Outgoing International*.
- F_CVM_INVOICE_M.COUNT_VOICE_OUT: Número total de chamadas classificadas como *Outgoing Calls.*
- F_CVM_INVOICE_M.COUNT_VOICE_OUT_ROAM: Número total de chamadas classificadas como *Calls Outgoing Roaming*.
- F_CVM_INVOICE_M.AIRTIME_VOICE_OUT_NAC: Tempo total de chamadas classificadas como *Outgoing National*.
- F_CVM_INVOICE_M.AIRTIME_VOICE_OUT_INT: Tempo total de chamadas classificadas como *Outgoing International*.
- F_CVM_INVOICE_M.AIRTIME_VOICE_OUT: Tempo total de chamadas classificadas como *Outgoing Calls*.
- F_CVM_INVOICE_M.AIRTIME_VOICE_OUT_ROAM: Tempo total de chamadas classificadas como *Outgoing Roaming*.
- F_CVM_INVOICE_M.OUT_USAGE_VOICE_NAC: Custo total de chamadas classificadas como *Outgoing National*.
- F_CVM_INVOICE_M.OUT_USAGE_VOICE_INT: Custo total de chamadas classificadas como *Outgoing International*.
- F_CVM_INVOICE_M.OUT_USAGE_VOICE: Custo total de chamadas classificadas como *Outgoing Calls*.
- F_CVM_INVOICE_M.OUT_USAGE_ROAM: Custo total de chamadas classificadas como *Outgoing Roaming*.
- F_CVM_INVOICE_M.MODULES: Custo total de chamadas classificadas como *Hired Modules*.
- F_CVM_INVOICE_M.RENTS: Custo mensal do serviço.
- F_CVM_INVOICE_M.OPENING_BALANCE: Balanço transferido do último mês.
- F_CVM_NETWORK_EVENT_M.FIRST_OUT_CALL_DATE: Data da primeira chamada efectuada pelo cliente, classificada como *Outgoing Call*.

- F_CVM_NETWORK_EVENT_M.LAST_OUT_CALL_DATE: Data da última chamada efectuada pelo cliente, classificada como *Outgoing Call*.
- F_CVM_NETWORK_EVENT_M.LAST_CALL_CHARGED_DATE: Data da última chamada cobrada ao cliente classificada como *Outgoing Call*.
- F_CVM_NETWORK_EVENT_M.LAST_3M_USAGE_AVG: Média dos últimos três meses do tráfego do cliente classificado como *Outgoing Call*.
- D_PARTY.PARTY_GENDER: Género do cliente.
- D_PARTY.PARTY_BIRTHDAY: Data de nascimento do cliente.
- D_PARTY.PARTY_MARITAL_STATUS: Estado civil do cliente.
- D_PARTY.PARTY_PROFESSION: Profissão do cliente.

Depois de concluída a selecção dos dados, os períodos de informação, os volumes das amostras e a selecção dos atributos, ficámos preparados para passar à fase da limpeza dos dados.

4.2.2. Limpeza dos dados

A limpeza dos dados tem como principal objectivo garantir a qualidade mínima necessária para o desenvolvimento dos modelos. Tornou-se, necessário, nesta fase, tomar acções que permitissem a limpeza dos dados envolvidos no processo de forma a obter os resultados esperados. Um dos problemas mais comuns nesta fase é a presença de valores desconhecidos nos atributos, como é o caso dos registos com dados incompletos por falhas no processo de selecção ou de validação.

Várias técnicas têm vindo a ser usadas para tratar esses valores desconhecidos, tais como a sua substituição pela média ou a moda do atributo, ou a substituição por valores previstos, utilizando o algoritmo de aprendizagem. Em casos de previsão, os algoritmos disponibilizados pelas ferramentas não apresentam resultados efectivos sobre dados com valores desconhecidos, mesmo recorrendo a técnicas simples, como a substituição de todos os valores desconhecidos por uma variável global, pode trazer resultados indesejados, desvirtuando o processo de predição. Neste caso, o valor desconhecido foi interpretado como um factor positivo que não pode ser justificado [22]. Outro problema habitual são os atributos que possuem valores extremos, atípicos ou com características bastante distintas, denominados de discrepantes ou *outliers*. Sempre que se verificou que

os dados representavam um erro de observação ou medida, os registos que continham valores *outliers* foram removidos da amostra, depois de cuidadosamente analisados, pois embora atípicos poderiam ser reais.

4.2.3. Construção dos dados

A construção dos atributos, que só pode ser elaborada após a conclusão da limpeza dos dados, é o processo de composição de atributos ditos primitivos em novos atributos relevantes para a descrição de um conceito. Este processo de construção pode ser dividido em duas abordagens, a saber:

- *Indução Automática*: processos de construção de atributos guiada automaticamente pelo método de construção, ou seja, os atributos são avaliados em relação aos dados e podem ou não ser descartados no conjunto de dados [22].
- *Indução do utilizador*: processo de construção que utiliza o conhecimento do utilizador para guiar a construção dos atributos [22].

Os dados utilizados no processo de mineração deste estudo recorreram à indução do utilizador para a construção dos atributos na definição da amostra final. Descrevem-se a seguir os atributos construídos com base no conjunto de informação original:

- Utilizando dois atributos pertencentes ao *Data Mart* F_CVM_ACCOUNT_SNAP_M, construíram-se dois novos atributos:
  - Utilizando o atributo ACCOUNT_ACTIVATION_DATE que contêm informação da data da activação do serviço, construiu-se o atributo SNAP_ACTIVATION_DAYS que representa o número de dias que decorrem desde a data de activação até à data de medição do *churn*.
  - Usando o atributo PARTY_NATIONALITY que contêm a nacionalidade do cliente, converteram-se os valores qualitativos em valores inteiros, gerando a variável SNAP_PARTY_NATIONALITY.
- Recorrendo a três dos atributos do *Data Mart* F_CVM_NETWORK_EVENT_M, foram construídos outros novos atributos, a saber:
  - Usando o atributo FIRST_OUT_CALL_DATE que representa a data da primeira chamada *Mobile Originated Call*, foi convertida

(EVE_FIRST_OUT_CALL_DATE) em número de dias que decorrem desde a data da primeira chamada até à data de medição do *churn*.

- Utilizando o atributo LAST_OUT_CALL_DATE que representa a data da última chamada *Mobile Originated Call*, foi convertida (EVE_LAST_OUT_CALL_DATE) em número de dias que decorrem desde a data da última chamada até à data de medição do *churn*.
- Através do atributo LAST_CALL_CHARGED_DATE que representa a data da última chamada *Mobile Originated Call* facturada, foi convertida (EVE_LAST_CALL_CHARGED_DATE) em número de dias que decorrem desde a data da última chamada facturada até à data de medição do *churn*.

- Tendo por base a tabela D_PARTY, alguns dos seus atributos foram reconstruídos, nomeadamente:
  - Usando o atributo PARTY_GENDER da dimensão D_PARTY, que representa o género do cliente, converteram-se os valores qualitativos em valores inteiros, gerando a variável PAR_PARTY_GENDER.
  - Utilizando o atributo PARTY_BIRTHDAY pertencente à dimensão D_PARTY, que contém a informação da data de nascimento do cliente, construi-se o atributo DAYS_OF_AGE que representa o número de dias que decorrem desde a data de nascimento do cliente até à data de medição do *churn*.
  - Usando o atributo PARTY_MARITAL_STATUS da dimensão D_PARTY, que representa o estado civil do cliente, converteram-se os valores qualitativos em valores inteiros, gerando a variável PAR_PARTY_MARITAL_STATUS.
  - Usando o atributo PARTY_PROFESSION da dimensão D_PARTY, que representa a profissão do cliente, converteram-se os valores qualitativos em valores inteiros, gerando a variável PAR_PARTY_PROFESSION.
- Explorando o *Data Mart* F_CVM_INVOICE_M que contém os dados de utilização, consumo e facturação, foram construídas seis novas variáveis:

- INV_AVG_AIRTIME_CALL_OUT_NAC: representando a média do tempo de chamadas nacionais, calculada a partir do total do atributo AIRTIME_VOICE_OUT_NAC a dividir pelo total do atributo COUNT_VOICE_OUT_NAC.
- INV_AVG_USAGE_CALL_OUT_NAC: representando a média do custo das chamadas nacionais, calculada a partir do total do atributo OUT_USAGE_VOICE_NAC a dividir pelo total do atributo COUNT_VOICE_OUT_NAC.
- INV_AVG_AIRTIME_CALL_OUT_INT: representando a média do tempo de chamadas internacionais, calculada a partir do total do atributo AIRTIME_VOICE_OUT_INT a dividir pelo total do atributo COUNT_VOICE_OUT_INT.
- INV_AVG_USAGE_CALL_OUT_INT: representando a média do custo das chamadas internacionais, calculada a partir do total do atributo OUT_USAGE_VOICE_INT a dividir pelo total do atributo COUNT_VOICE_OUT_INT.
- INV_AVG_AIRTIME_CALL_OUT_ROAM: representando a média do tempo de chamadas em *roaming*, calculada a partir do total do atributo AIRTIME_VOICE_OUT_ROAM a dividir pelo total do atributo COUNT_VOICE_OUT_ROAM.
- INV_AVG_USAGE_CALL_OUT_ROAM: representando a média do custo das chamadas em *roaming*, calculada a partir do total do atributo OUT_USAGE_VOICE_ROAM a dividir pelo total do atributo COUNT_VOICE_OUT_ROAM.

As variáveis do tipo data foram transformadas em valores inteiros, por se demonstrarem mais representativas para os algoritmos de classificação; o mesmo se aplica às transformações efectuadas nos atributos qualitativos.

4.2.4. Integridade dos dados

A integridade dos dados pressupõe a análise das relações permitidas entre os atributos. Depois do estudo dos requisitos de negócio efectuado no ponto 3.1. e da criação do processo de extracção dos dados explicado na secção 4.2.1, a integração da informação foi assegurada, podendo afirmar-se que a tabela fonte para os modelos considerados continha um registo por cliente. Tal aconteceu porque a informação das várias contas que o cliente possui foi agregada, utilizando para o feito o campo DW_PARTY_ID (identificador único de cliente). Posteriormente, foi ainda realizada uma identificação de casos extremos para validar a integridade da informação.

4.2.5. Transformação dos dados

O objectivo da fase de transformação é a representação dos dados com o objectivo de superar as limitações dos algoritmos utilizados na mineração. A decisão das transformações a adoptar depende dos algoritmos usados no processo de mineração, envolvendo, normalmente, funções matemáticas na obtenção dos dados, para evitar as restrições dos algoritmos. As técnicas de transformação mais comuns são as seguintes: a normalização pelo desvio padrão, a normalização pela variação, a transformação dos atributos quantitativos em qualitativos e vice-versa, a transformação de dados complexos e a redução de dados.

Em relação à normalização, este é um processo que transforma os intervalos dos valores dos atributos, num intervalo específico (-1,1). Este tipo de transformação é importante para métodos que calculam distâncias entre vizinhos [23], apesar de não ser de grande utilidade para a maioria dos métodos que induzem representações (árvores de decisão), dado que a normalização diminui a compreensão do modelo gerado. Quando aplicada às redes neuronais permite realizar melhores processos de treino.

$$y = \frac{x - \mu}{\sigma} \quad (1)$$

Figura 19 – Normalização pelo desvio padrão

As normalizações pelo desvio padrão obtêm-se através da equação apresentada na figura 19, em que y representa o novo valor normalizado; x, o valor actual; μ a média; e σ o desvio padrão da variável. No anexo A deste documento é apresentada uma tabela com os dados da amostra original e os dados após normalização, recorrendo à função logaritmo e utilização da média e desvio padrão.

$$y = \frac{x - \min}{\max - \min} \quad (2)$$

Figura 20 – Normalização pela variação

Na figura 20 apresenta-se a equação utilizada para calcular a normalização pela variação. Nesta equação, y representa o novo valor normalizado; x, o valor actual e *máx.* e *min.* os valores de máximo e mínimo, respectivamente.

No que se refere à transformação dos atributos quantitativos em qualitativos e vice-versa, existem algoritmos que possuem limitações quando trabalham com este tipo de atributos. Consoante as limitações dos algoritmos, há a necessidade de conversão dos valores. Para este efeito, existem diversas abordagens. No caso da transformação de variáveis qualitativas em quantitativas usa-se, frequentemente, a substituição por valores numéricos. O contrário é recíproco para as transformações quantitativas em qualitativas [23]. Apesar de nenhum dos algoritmos utilizados exigir apenas atributos nominais na amostra de treino, optou-se por realizar um processo de transformação de todos os atributos qualitativos em quantitativos.

Relativamente à transformação de dados complexos, a grande maioria dos algoritmos utilizados na mineração de dados não consegue trabalhar com atributos do tipo data e hora. Geralmente, o método de transformação utilizado neste tipo de dados é a conversão em atributo do tipo inteiro. Um exemplo prático deste tipo de transformação é o caso do atributo *SUBSCRIPTION_DATE* que representa a data em que o cliente subscreveu o serviço com o formato do valor "20080201". A transformação inicial converte o valor de formato *date* em formato inteiro, procedendo-se à construção de uma nova variável (*SNAP_ACTIVATION_DAYS*). Esta irá conter o valor da diferença de datas entre o valor do atributo *SUBSCRIPTION_DATE* e o dia actual. Outras variáveis com o mesmo tipo de

transformação são as variáveis *EVE_FIRST_CALL_DATE*, *EVE_LAST_CALL_DATE*, *EVE_LAST_CALL_CHARGED_DATE* e *DAYS_OF_AGE*.

No que diz respeito à redução de dados, existem claras vantagens na utilização de grandes volumes de informação. Contudo, na prática, o conjunto pode tornar-se excessivamente grande superando a capacidade da própria ferramenta de previsão [22]. A técnica de redução de dados é de suma importância neste estudo, visto que a grande quantidade de registos com um custo de processamento elevado prejudicaria a construção dos modelos e a análise de resultados.

Concluída a validação do modelo, a aplicação de técnicas de redução de dados sobre o conjunto total dos mesmos ficou bastante simplificada. Na figura 21 está esquematizado o modelo de redução de dados proposto por Weiss [22].

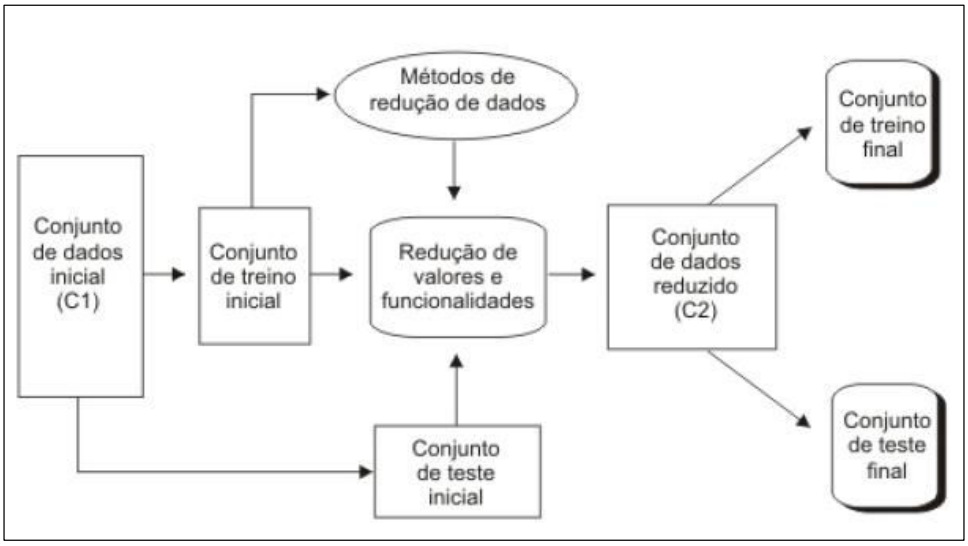

Figura 21 – Modelo de redução de dados proposto por Weiss [22]

Dado um conjunto (C1) de dados iniciais, os dados são reduzidos em quantidades de valores (ou atributos), e um novo conjunto é produzido (C2). Quando as dimensões do conjunto inicial se enquadram nos limites aceitáveis não são necessárias técnicas de redução de dados. Após a obtenção do grupo de dados reduzido, este pode ser dividido em conjunto de treino e conjunto de teste. É importante referir que a redução dos dados não pode afectar a qualidade dos resultados. Weiss detalha um conjunto de técnicas para a redução de dados em processos de mineração de dados. Neste estudo aplicou-se a técnica de amostragem em *cluster* em que os elementos de um conjunto de dados formam um

cluster (clientes pertencentes ao segmento *star* 5). No caso em estudo, a quantidade da amostra é de 5,000 registos, sendo 4,701 (94.20%) pertencentes ao segmento não *churn* e 299 (5.98%) pertencentes ao segmento *churn*, como demonstra a figura 22.

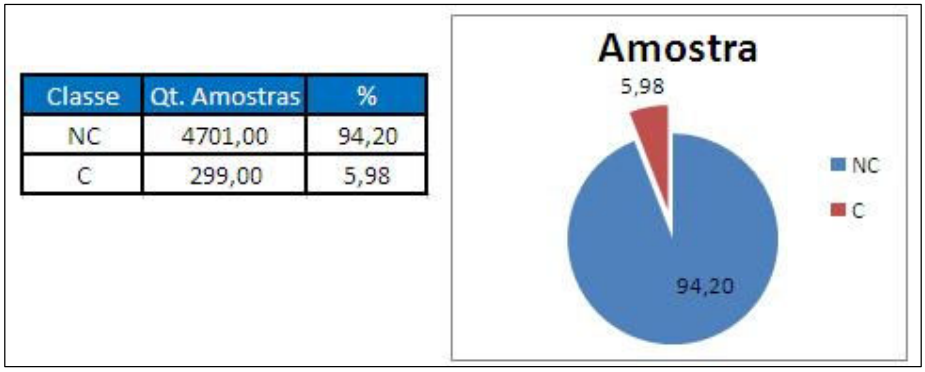

Figura 22 – Amostra em Estudo

4.3. Modelação

Na fase de modelação, são seleccionadas e aplicadas várias técnicas de modelação juntamente com a configuração e calibração dos seus parâmetros. Existem variadas técnicas para o mesmo tipo de problema de mineração, algumas delas têm requisitos específicos associados ao tipo de dados, existindo, por este motivo, a necessidade frequente de retroceder à fase de preparação dos dados.

4.3.1. Desenho e Planeamento de Testes

Antes de criar o modelo preditivo, é necessário criar um modelo de testes para testar a qualidade e a validade do modelo. Este modelo deve incluir um conjunto de dados para treino, outro para testes e outro para validação dos resultados.

A amostra foi dividida em 3 grupos, nomeadamente:

- 40% de dados para o treino (classe dos clientes *churn*: 120 casos e classe dos cliente *não churn*: 1880 casos).
- 30% de dados para o teste (classe dos clientes *churn*: 90 casos e classe dos clientes *não churn*: 1410 casos).

- 30% de dados para a validação (classe dos clientes *churn*: 90 casos e classe dos clientes *não churn*: 1410 casos).

Concluídas as fases de conhecimento e preparação dos dados e seleccionadas as técnicas de modelação descritas no ponto 2.6, temos garantidas as condições para o desenvolvimento do treino, do teste e da validação dos modelos.

4.3.2. Ferramenta Utilizada

A ferramenta utilizada neste estudo para o desenvolvimento dos modelos foi o *Statistical Package for the Social Sciences Clementine* (*SPSS Clementine*) [27]. Esta ferramenta foi a escolhida não só por ter sido um requisito do cliente, mas também por ser bastante usada na análise de regressões múltiplas e redes neuronais, pela sua rapidez de processamento em grandes conjuntos de dados e pela facilidade de manipulação de um grande número distinto de atributos.

4.3.3. Construção dos Modelos

Existem vários algoritmos para previsão em mineração de dados. Os que vão ser utilizados neste estudo foram já descritos anteriormente na secção 2.3. O primeiro método estudado foi a Rede Neuronal. Para o efeito, foram realizados vários estudos com recurso a diferentes métodos e utilizando amostras normalizadas e não normalizadas. O método com melhor comportamento foi o *Prune* que tem como principal característica o facto de começar com uma grande rede e, à medida que o treino vai avançando, é efectuada a poda nas unidades escondidas e de entrada mais fraca. Embora seja um dos métodos mais lentos de entre os algoritmos de redes, é, normalmente, o que apresenta melhores resultados. A figura 23 mostra como é feita a selecção do método *Prune* na ferramenta de mineração de dados escolhida.

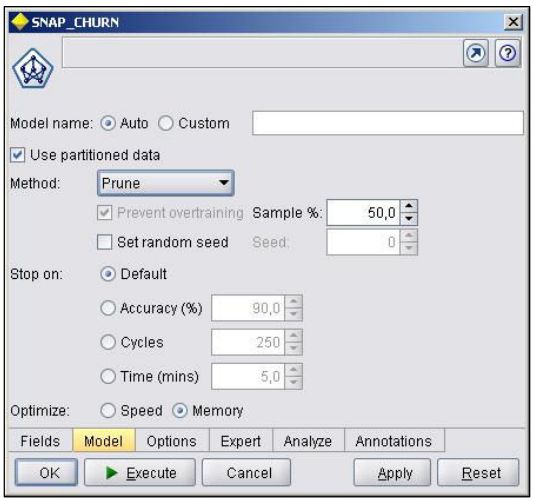

Figura 23 – Método *Prune* utilizado nas redes neuronais.

A configuração dos parâmetros das redes neuronais influencia o desempenho dos algoritmos. Dos vários testes realizados para identificar os melhores parâmetros de configuração das redes neuronais no SPSS Clementine, os que se apresentam a seguir foram aqueles que obtiveram os melhores resultados:

- *Use partitioned data*: No caso de existir um campo particionado, esta opção assegura que apenas os dados de treino serão utilizados para a construção do modelo.
- *Stop on*: Uso do valor por *default*, o treino só pára quando a rede atingir o estado óptimo.
- *Optimize*: Uso do valor *Memory* para obrigar o algoritmo a usar o disco na fase de processamento.
- Para as outras opções assumiram-se os valores por defeito da ferramenta.

A árvore de decisão foi o segundo método a ser estudado e o algoritmo *CHAID* foi aquele que apresentou melhor performance. Este algoritmo *CHAID – Chi-Squared Automatic Interaction Detection* é um algoritmo de classificação para árvores de decisão onde o *Chi-Square* é utilizado para identificar os pontos óptimos de corte. A figura 24 apresenta a selecção do método CHAID como o algoritmo das árvores de decisão.

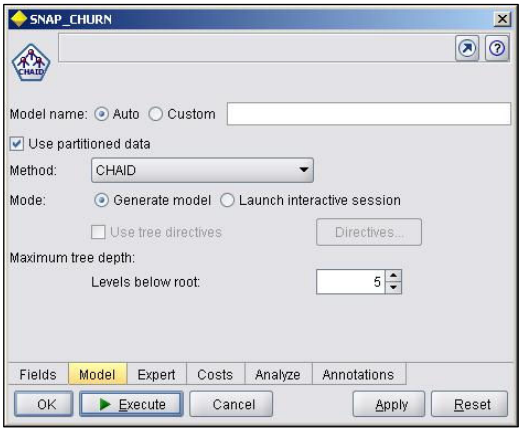

Figura 24 – Método *CHAID* utilizado nas árvores de decisão.

Neste algoritmo, os melhores parâmetros seleccionados para obtenção do melhor modelo foram:

- *Use partitioned data*: No caso de existir um campo particionado, esta opção assegura que apenas os dados de treino serão utilizados para a construção do modelo.
- *Mode*: Escolheu-se o *Generate model* para criação do modelo.
- *Levels below root*: No caso do algoritmo necessitar de aumentar o número de níveis para construção da árvore, o limite será 15 níveis.
- Para as outras opções assumiram-se os valores por defeito da ferramenta.

Finalmente foram estudadas as regressões logísticas, utilizando-se os algoritmos *Stepwise*, *Backwards* e *Forwards* [17]. O *Stepwise* foi o método que apresentou melhor comportamento, dado que começa escolhendo a melhor variável discriminatória. A variável inicial faz par com cada uma das outras variáveis independentes, uma a uma, e é escolhida a variável mais adequada para melhorar o poder discriminatório da função em combinação com a primeira variável. A terceira e as demais variáveis são seleccionadas pelo mesmo processo.

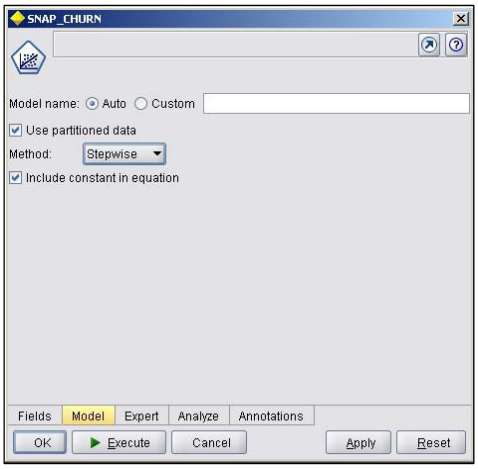

Figura 25 – Método *Stepwise* utilizado na regressão logística.

A figura 25 apresenta a forma como foi escolhido o método Stepwise utilizado na regressão logística. Neste caso utilizaram-se os parâmetros por omissão definidos pela ferramenta utilizada.

4.4. Avaliação dos Resultados

Construídos os modelos, torna-se necessário analisar e avaliar os resultados obtidos em cada um deles, de forma a poder escolher aquele que apresenta um melhor desempenho. Consequentemente, após a avaliação e antes de se proceder à instalação final, é necessário avaliar e rever cuidadosamente o modelo e todos os passos executados para a sua construção, assegurando a concretização de todos os requisitos de negócio. No final desta fase, é tomada a decisão do uso ou não dos resultados da mineração.

4.4.1. Apreciação Geral

Após a conclusão das etapas anteriores, é necessário proceder à avaliação da taxa de acerto de cada algoritmo. Para este efeito, foram utilizadas duas metodologias de validação e teste descritas abaixo:

- Validação utilizando o conjunto de treino: este método usou o mesmo conjunto de dados utilizado no treino para efectuar os testes; a utilização desta metodologia foi bastante optimista e provocou entropia na análise dos resultados.

- Validação utilizando uma percentagem da divisão do conjunto de treino: este método dividiu a informação em três amostras, uma amostra para treino do modelo, uma para teste e outra para validação; como este método apresentou amostras desconhecidas ao modelo (dados diferentes dos utilizados no treino) produziu um teste e validação mais eficientes; no presente estudo, 40% da amostra foi usada no treino, 30% no teste e 30% na validação.

O recurso a estas duas técnicas teve por objectivo validar o modelo e garantir a coerência dos resultados.

4.4.2. Resultados Redes Neuronais

O primeiro método testado utilizou *Redes Neuronais* e a técnica que demonstrou melhor desempenho foi o *Prune* com dados normalizados por desvio padrão. As variáveis que melhor explicam este modelo são: EVE_LAST_CALL_CHARGE_DATE, EVE_LAST_OUT_CALL_DATE e EVE_FIRST_OUT_CALL_DATE. Numa análise mais exaustiva do modelo, apura-se que a eficácia estimada é de 99,899% com 3 neurões de entrada, 2 neurões escondidos e 1 neurão de saída. A topologia desta rede neuronal é *Feedforward*, caracterizando-se por ser uma rede com um grafo sem ciclos (rede directa), normalmente, designada rede de camadas onde os neurónios que recebem sinais de excitação são chamados de camada de entrada e os neurónios que tem a sua saída como a saída da rede pertencem à camada de saída.

Rede Neuronal												
	Treino				Teste				Validação			
Métodos	Cor.	Incor.	% Cor.	% Incor.	Cor.	Incor.	% Cor.	% Incor.	Cor.	Incor.	% Cor.	% Incor.
Prune	1994	6	99,70	0,30	1495	5	99,67	0,33	1404	96	93,6	6,4

Tabela 1 – Resultados Redes Neuronais

A tabela 1 apresenta os resultados das redes neuronais, no que respeita ao treino, ao teste e à validação. Tomando como exemplo os resultados obtidos no treino, verifica-se que foram correctamente classificados 1994 casos, correspondendo a 99,7% e 6 casos foram indevidamente classificados, correspondendo a 0,3%. Deverá ser feita o mesmo tipo de leitura no que toca aos testes e à validação.

Rede Neuronal - Método Prune						
	Treino		Teste		Validação	
Classes	Não Churn	Churn	Não Churn	Churn	Não Churn	Churn
Não Churn	1878	2	1409	1	1330	60
Churn	4	116	4	86	36	74

Tabela 2 – Resultados Redes Neuronais pelo método Prune

A matriz de confusão (tabela 2) apresenta a quantidade de acerto e erros de cada classe, para a respectiva técnica utilizada na previsão. Por exemplo, no caso da validação, a classe não *churn* apresenta 1390 registos classificados correctamente e 20 registos classificados incorrectamente como *churn*. No caso da classe *churn* 16 registos são classificados incorrectamente como não *churn* e 74 registos são classificados correctamente como *churn*. Para os dados de treino e teste, a tabela acima exibe o mesmo grau de detalhe dos resultados obtidos. A elevada taxa de acerto nas redes neuronais pode resultar do factor *overfitting*.

4.4.3. Resultados Árvores de Decisão

As árvores de decisão apresentam os resultados num formato hierárquico de acordo com a prioridade das regras. Deste modo, a regra mais importante é apresentada na raiz da árvore e as regras menos relevantes são mostradas nos nós subsequentes. A vantagem das árvores de decisão são que elas tomam decisões tendo em consideração as regras mais importantes. A figura *26* mostra a árvore de decisão do caso em estudo.

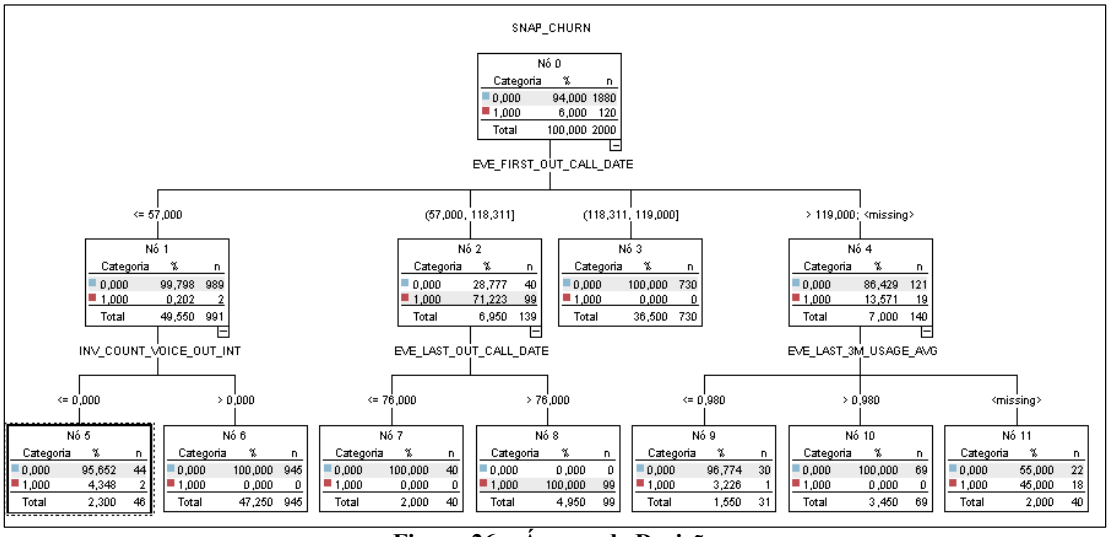

Figura 26 – Árvore de Decisão

A árvore tem dois níveis de profundidade e oito folhas. O nó raiz seleccionado pelo algoritmo CHAID foi o EVE_FIRST_OUT_CALL_DATE que representa a data da primeira chamada efectuada pelo cliente. Uma leitura possível do caso em estudo é quando EVE_FIRST_OUT_CALL_DATE tem valores inferiores a 57, a variável subsequente com maior importância para a classificação do *churn* é INV_COUNT_VOICE_INT, isto é, quando a data da primeira chamada é inferior a 57 dias, o número de chamadas internacionais que o cliente realiza é a variável mais importante para classificação do *churn*. Da mesma forma, pode-se aplicar a analogia às restantes variáveis. De seguida, é apresentado o significado das restantes variáveis que tem importância para o modelo, a saber:

- EVE_LAST_CALL_DATE: Última chamada efectuada pelo cliente. Um valor elevado nesta variável significa que o cliente não está a utilizar o serviço à bastante tempo.
- EVE_LAST_3M_USAGE_AVG: Média das chamadas dos últimos 3 meses, o valor baixo desta variável reflecte a baixa utilização do serviço pelo cliente.

De notar que a variável mais importante na árvore de decisão é resultado da transformação aplicada na fase de transformação dos dados, o que evidencia a necessidade de concentração de esforços nesta fase de construção do sistema de mineração de dados.

Árvore de Decisão												
	Treino				Teste				Validação			
Métodos	Cor.	Incor.	% Cor.	% Incor.	Cor.	Incor.	% Cor.	% Incor.	Cor.	Incor.	% Cor.	% Incor.
CHAID	1979	21	98,95	1,05	1477	23	98,47	1,53	1465	35	97,67	2,33

Tabela 3 – Resultados da Árvore de Decisão

A tabela 3 apresenta os resultados das árvores de decisão, no que respeita ao treino, ao teste e à validação. Tomando como exemplo os resultados obtidos no teste, verifica-se que foram correctamente classificados 1477 casos, correspondendo a 98,47% e 23 casos foram indevidamente classificados, correspondendo a 1,53%. Deverá ser feita o mesmo tipo de leitura no que toca aos testes e à validação.

Árvore Decisão - CHAID						
	Treino		Teste		Validação	
Classes	Não Churn	Churn	Não Churn	Churn	Não Churn	Churn
Não Churn	1880	0	1410	0	1409	1
Churn	21	99	23	67	34	56

Tabela 4 – Resultados da Árvore de Decisão pelo método Prune

Neste caso (tabela 4) a matriz de confusão apresenta resultados muito bons. Nota-se uma elevada taxa de sucesso na classificação das classes não só de treino, mas também do teste e validação, o que demonstra a elevada importância que têm as variáveis escolhidas pelo algoritmo. Por exemplo, no caso do teste, a classe não *churn* apresenta 1410 registos classificados correctamente e 0 registos classificados incorrectamente como *churn*. No caso da classe *churn,* 23 registos são classificados incorrectamente como não *churn* e 67 registos são classificados correctamente como *churn*. Os valores de Qui-quadrado elevados demonstram a robustez das variáveis e a forte relação entre as variáveis independentes e a variável dependente, garantindo consistência nos valores previstos – o Anexo B contém a descrição detalhada dos resultados obtidos na Árvore de Decisão.

4.4.4. Resultados Regressão Logística

O último método testado foi a regressão logística, sendo a técnica *stepwise* aquela que apresentou o melhor comportamento. Nesta técnica, utilizaram-se variáveis normalizadas

pelo desvio padrão, variação entre o mínimo e o máximo e função logaritmo, sendo a normalização pela variação entre o mínimo e o máximo a que apresentou melhores resultados, expostos na tabela abaixo.

Regressão Logística												
	Treino				Teste				Validação			
Métodos	Cor.	Incor.	% Cor.	% Incor.	Cor.	Incor.	% Cor.	% Incor.	Cor.	Incor.	% Cor.	% Incor.
Stepwise	1943	57	97,15	2,85	1445	55	96,33	3,67	1422	78	94,8	5,2

Tabela 5 – Resultados da Regressão Logística

A tabela 5 apresenta os resultados das regressões logísticas, no que respeita ao treino, ao teste e à validação. Tomando como exemplo os resultados obtidos na validação, verifica-se que foram correctamente classificados 1422 casos, correspondendo a 94,8% e 78 casos foram indevidamente classificados, correspondendo a 5,2%. Deverá ser feita o mesmo tipo de leitura no que toca aos testes e à validação.

Regresão Logistica - Stepwise						
	Treino		Teste		Validação	
Classes	Não Churn	Churn	Não Churn	Churn	Não Churn	Churn
Não Churn	1868	12	1407	3	1390	20
Churn	45	75	52	38	48	42

Tabela 6 – Resultados da Regressão Logística pelo método Stepwise

A matriz de confusão da regressão logística (tabela 6) apresenta bons resultados na fase de treino, mas ao longo das amostras de teste e validação nota-se uma perda de eficácia. Por exemplo, no caso do teste, a classe não *churn* apresenta 1407 registos classificados correctamente e 3 registos classificados incorrectamente como *churn*. No caso da classe *churn,* 52 registos são classificados incorrectamente como não *churn* e 38 registos são classificados correctamente como *churn*. Neste caso o valor do Qui-quadrado é muito elevado, 1037, o que demonstra variáveis muito robustas, contribuindo fortemente para a consistência dos valores previstos. Pelo contrário, o valor do R^2 é relativamente baixo, o que demonstra a baixa relação que existe entre as variáveis independentes. A taxa de eficácia do modelo é de 97,15% e a equação matemática obtida na regressão que representa o modelo é a que se segue:

Churn = 2,727 * SNAP_ACTIVATION_DAYS +

21,08 * SNAP_GRACE_PERIOD +

24,2 * INV_COUNT_VOICE_OUT +

13,75 * INV_OUT_USAGE_VOICE +

50,52 * EVE_FIRST_OUT_CALL_DATE +

-56,32 * EVE_LAST_OUT_CALL_DATE +

11,62 * INV_AVG_AIRTIME_CALL_OUT_INT +

25,56 * INV_AVG_USAGE_CALL_OUT_ROAM +

+ 0,3971

No Anexo C pode-se encontrar uma descrição detalhada dos resultados obtidos na regressão logística.

4.5. Abordagem Crítica

Ao longo das secções anteriores expuseram-se os processos de mineração de dados realizados com vista à identificação de situações de *churn* numa empresa de telecomunicações. A forma como estes processos foram realizados, com base na metodologia CRISP-DM, permitiu revelar e alcançar resultados bastante curiosos. Na realidade, a aplicação de mais do que um tipo de algoritmo de mineração de dados, opção de trabalho assumida desde o primeiro momento, contribuiu para que esses resultados e situações emergissem de uma forma natural. É verdade que, os dados escolhidos também tiveram o seu "quê" de responsabilidade nos resultados obtidos como demonstra a tabela 7.

Métodos	Treino				Teste				Validação			
	Cor.	Incor.	% Cor.	% Incor.	Cor.	Incor.	% Cor.	% Incor.	Cor.	Incor.	% Cor.	% Incor.
Rede Neuronal - Prune	1994	6	99,70	0,30	1495	5	99,67	0,33	1404	96	93,6	6,4
Árvore Decisão - CHAID	1979	21	98,95	1,05	1477	23	98,47	1,53	1465	35	97,67	2,33
Regressão Logística - Stepwise	1943	57	97,15	2,85	1445	55	96,33	3,67	1422	78	94,8	5,2

Tabela 7 – Resultados obtidos dos algoritmos de previsão

Comparando os vários métodos com a amostra da validação, nota-se um desempenho elevado e muito próximo entre os diferentes algoritmos utilizados, nomeadamente:

- 93,60% na rede neuronal pelo algoritmo *Prune*.
- 97,67% na árvore de decisão pelo algoritmo CHAID.
- 94,80% na regressão logística pelo algoritmo *Stepwise*.

Com base nestes indicadores, verifica-se que os resultados do estudo, embora coerentes, são demasiado elevados, quando comparados com os modelos de previsão de *churn,* nos

casos de estudo apresentados durante a secção 2.5, cujos resultados se encontram entre os valores de 50% e 70%. Os resultados obtidos dos algoritmos de previsão apresentados na tabela 7 são demasiado altos, consequência de uma mediação de *churn* efectuada muito em cima do acontecimento, quando grande parte dos clientes já deixou a empresa, sendo que uma maioria o fez de forma involuntária, por razões de não pagamento.

Método	Classe	Treino	Teste	Validação
Rede Neuronal - Prune	Não Churn	99,89	99,93	94,33
	Churn	96,67	95,56	82,22
Árvore Decisão - CHAID	Não Churn	100,00	100,00	99,93
	Churn	82,50	74,44	62,22
Regressão Logística - Stepwise	Não Churn	99,36	99,79	98,58
	Churn	62,50	42,22	46,67

Tabela 8 – Percentagem de acerto em cada classe

Tendo em consideração a amostra usada na validação apresentada na tabela 8, pode verificar-se que a Árvore de Decisão foi o método que melhor classificou a classe *Não Churn* (99,83%), sendo a Rede Neuronal a que apresenta melhores resultados na classificação do *Churn* (82,22%). No entanto, ponderando os resultados obtidos pelas classes *Churn* e *Não Churn* do método da Rede Neuronal e da Árvore de Decisão e tendo em conta as características específicas de cada um dos métodos, conclui-se que este último método é aquele que melhor satisfaz as necessidades do cliente. Refira-se que não existe um modelo perfeito para as diferentes situações, mas sim, modelos que têm melhor desempenho consoante os diferentes tipos de dados apresentados e os requisitos do cliente.

No caso das Redes Neuronais e tendo por base o algoritmo que melhores resultados apresentou (algoritmo *Prune*), verifica-se, através da análise da tabela 8, que existe um decréscimo acentuado no acerto da classificação de *churn* quando comparamos as amostras de teste (95,56%) e de validação (82,22%), o que não acontece quando a comparação é feita entre as amostras de treino (96,67%) e de teste (95,56%), o que indicia a existência de *overfitting*. Na utilização deste método, constatou-se que as variáveis que melhor representam a classificação do cliente *churn* são as seguintes: EVE_LAST_OUT_CALL_DATE (0,3954), EVE_LAST_CALL_CHARGED_DATE (0,3946) e EVE_FIRST_OUT_CALL_DATE (0,21). Pode-se interpretar os coeficientes apresentados da seguinte forma: o valor do coeficiente para a variável data da última chamada efectuada EVE_LAST_OUT_CALL_DATE é de 0,3954, o que significa que,

isolando-se as outras variáveis do modelo, para um aumento de 10% normalizado na variável EVE_LAST_OUT_CALL_DATE ocorre um aumento normalizado de 4% na taxa de *churn*, que em valores não normalizados se refere a um aumento em média de 60 clientes. De modo similar os outros indicadores podem ser interpretados. De facto, através da análise do significado de cada variável, constata-se que um cliente cuja data da última chamada é bastante antiga, que a data da última chamada paga é também antiga e que a data da primeira chamada efectuada foi há bastante tempo, conclui-se que se trata de clientes com fortes probabilidades de deixar a empresa. Contudo, esta informação não é por si só suficiente para identificar se a causa do *churn* se refere uma situação voluntária ou involuntária, uma vez que o modelo não seleccionou nenhuma variável de tráfego que é essencial para permitir identificar a causa deste facto. De seguida apresenta-se uma comparação entre este estudo e o caso "*Applying data mining to Telecom churn management*" analisado na secção 2.5. Neste ultimo, foram utilizados 11 meses de dados para a classificação do cliente *churn*, tendo-se verificado que o método das redes neuronais foram bastantes eficazes nos primeiros seis meses. O mesmo não acontece no estudo apresentado na tabela 8, em que, tendo por base 14 meses de dados, o mesmo método revelou ter um óptimo desempenho e constante até ao momento da validação.

Considerando o método de Árvore de Decisão, nomeadamente o algoritmo CHAID, cujos resultados já foram descritos na secção 4.4.3, constatou-se que as variáveis mais importantes foram: EVE_FIRST_OUT_CALL_DATE (0,722), INV_COUNT_VOICE_INT (0,221), EVE_LAST_CALL_DATE (0,03) e EVE_LAST_3M_USAGE_AVG (0,028). A interpretação dos coeficientes apresentados deve ser feita de modo semelhante ao já mencionado no modelo das Redes Neuronais. Analisando o significado de cada uma das variáveis acima mencionadas, verificou-se que clientes que efectuaram a primeira chamada há muito tempo e que não fizeram chamadas num passado recente, possuindo uma média trimestral de chamadas efectuadas muito baixa, são claramente clientes *churn* involuntários por não pagamento. Isto é explicado pelo barramento das chamadas efectuadas por parte do operador, após o não pagamento da factura por parte do cliente. Poder-se-á concluir que as variáveis relacionadas com o tráfego do cliente, neste caso, se revestem de enorme importância, dado que nos permitem explicar o motivo do *churn*. Quando analisamos novamente o caso de estudo referido no parágrafo anterior ("*Applying data mining to Telecom churn management*"), verifica-se que a pós a utilização das Redes Neuronais nos primeiros seis meses, as Árvores de

Decisão revelaram melhor desempenho nos 5 meses seguintes, que o modelo atrás referido. Comparando com o presente estudo, conclui-se que os resultados nas amostras de teste e validação têm um desempenho bastante similar.

Relativamente aos processos de regressão logística, o algoritmo que apresentou melhores resultados foi o *Stepwise* tal como já foi referido na secção 4.4.4. Na equação disponibilizada pelo referido algoritmo, as variáveis com maior relevância e que constituem a equação matemática são as seguintes:

Churn = 2,727 * SNAP_ACTIVATION_DAYS +

21,08 * SNAP_GRACE_PERIOD +

24,2 * INV_COUNT_VOICE_OUT +

13,75 * INV_OUT_USAGE_VOICE +

50,52 * EVE_FIRST_OUT_CALL_DATE +

-56,32 * EVE_LAST_OUT_CALL_DATE +

11,62 * INV_AVG_AIRTIME_CALL_OUT_INT +

25,56 * INV_AVG_USAGE_CALL_OUT_ROAM +

+ 0,3971

Os valores associados a cada variável correspondem à ponderação que cada variável tem, enquanto o ultimo valor da equação refere-se ao valor residual. Pode-se interpretar os coeficientes do modelo *Stepwise* da seguinte forma: o valor do coeficiente para a variável dias de activação do serviço até ao momento da medição de *churn* SNAP_ACTIVATION_DAYS é de 2,727, o que significa que, isolando-se as outras variáveis da equação, para um aumento de 10% normalizado na variável mencionada ocorre um aumento normalizado de 3% na taxa de *churn*, que em valores não normalizados se refere a um aumento em média de 45 clientes na taxa de *churn*. Os outros indicadores podem ser interpretados de modo semelhante. Tendo por base o estudo "*UMA APLICAÇÃO DE MINERAÇÃO DE DADOS NO GERENCIAMENTO DO CHURN EM SERVIÇOS DE BANDA LARGA DE TELECOMUNICAÇOES*", verificou-se a inexistência de multicolinearidade, uma vez que as variáveis independentes possuem valores inferiores a 0,5, garantindo a inexistência de relação lineares exactas ou aproximadamente exactas. O mesmo sucede neste estudo, dado que o valor de R^2 é de 0,2. Além disto, também se verificou em ambos os estudos a uma distribuição consistente dos resíduos em relação à distribuição teórica, aproximando-se, por isso, da curva normal. Refira-se ainda que das 37

variáveis fonte utilizadas neste estudo, foram identificadas oito como as mais relevantes para a taxa de *churn*, já referenciadas na equação do algoritmo.

Em suma, suportados pelos resultados alcançados, pelos modelos e estratégias de mineração de dados seguidos, pelos casos de estudo analisados e pelas políticas de gestão empresarial seguidas pela empresa alvo, verifica-se que as situações de *churn* mais críticas foram detectadas através do modelo de Regressão Logística. Na tabela 8 podemos ver um resumo desses mesmos casos, onde, apesar das Redes Neuronais apresentarem uma percentagem de desempenho mais elevada que os restantes modelos, a questão do *overfitting* inerente a estas, os bons resultados obtidos nas Árvores de Decisão, bem como a sua facilidade de interpretação e os valores obtidos no segundo caso de estudo que sustentam a validade do presente estudo, vieram suportar a escolha do modelo das Árvores de Decisão como aquele que apresenta um melhor desempenho, tendo em conta os requisitos de negócio.

Os resultados alcançados, embora credíveis, não são representativos da realidade, porque possuem uma percentagem elevada na classificação do cliente *churn*, consequência de uma medição efectuada muito em cima do acontecimento, tal como já foi mencionado. Por este motivo, dever-se-ia, na fase de análise, considerar a construção de dois modelos de classificação de *churn*, isto é, um modelo para classificar os clientes *churn* involuntários por falta de pagamento ou fraude e outro para classificar os *churn* voluntários, uma vez que do ponto de vista de negócio é importante e necessário analisar e classificar os diferentes tipos. Se esta abordagem fosse considerada na fase de análise, além dos algoritmos apresentarem resultados diferentes, estes seriam com certeza mais realistas e representativos da realidade.

5. Conclusões e Trabalho Futuro

É um facto que o impacto que as situações de *churn* têm nas empresas de telecomunicações é enorme, dada a forte concorrência que se verifica nos mercados actuais. Esta deve-se ao facto de as empresas apresentarem um leque muito variado de ofertas, bastante atractivas na maioria dos casos, que pela relação qualidade-preço motivam a constante mudança, por parte dos clientes, entre as diferentes operadoras de telecomunicações. É por isso que se torna importante estudar o conceito de *churn*, dado que o custo associado à reaquisição de novos clientes é superior ao custo da sua retenção. A definição de *churn* varia de empresa para empresa, não existindo uma regra *standard* de definição deste conceito. Todavia, qualquer que seja a interpretação de *churn* haverá sempre um dado comum: trata-se de algo nefasto e que prejudica de forma clara os lucros da empresa.

Neste estudo, optou-se por subdividir o *churn* em 3 grupos distintos: voluntário, involuntário e interno. Todavia, neste trabalho, o nosso objectivo principal teve por base a identificação dos dois primeiros tipos de *churn*, uma vez que, enquanto estes não forem devidamente estudados, o *churn* interno não qualquer tipo de relevância. Adicionalmente, pretendia-se, também, identificar os clientes com elevada probabilidade de trocar a empresa, neste caso particular uma empresa de telecomunicações (cliente *churn*) Deste modo, a identificação do cliente *churn* só é possível recorrendo a técnicas de análise particulares, tais como as que a mineração de dados nos disponibiliza.

De forma a garantir a efectividade dos processos desenvolvidos e um acompanhamento adequado da sua realização, utilizou-se a abordagem CRISP-DM, uma vez que esta é, segundo a generalidade dos seus seguidores, uma abordagem flexível e iterativa no desenvolvimento das várias etapas de um processo de mineração de dados. De acordo com esta abordagem, o primeiro passo realizado foi a análise de negócio da empresa em questão, na qual foram identificados os agentes do estudo (Decisor, Detentor e Utilizador), os conceitos de negócio a considerar (Produto, Cliente e Conta), limites temporais, os perfis (variáveis independentes), a acção (o cliente vai sair da empresa?) e o conceito (definição do *target*). No final desta etapa, estava transformado um problema de negócio num problema de mineração de dados. Neste estudo específico, o problema de negócio em questão é o elevado número de clientes que abandonaram a empresa e a consequente perda

de lucro. Quanto ao problema de mineração de dados, este foi definido como a previsão dos clientes com elevada probabilidade de deixar a empresa - prever o cliente *churn*. Após esta fase, procedeu-se à identificação das tabelas fontes de informação e dos seus atributos, explorou-se a qualidade dos dados, verificando-se a sua redundância e a sua eventual duplicação, e os campos a serem removidos, por não adicionarem valor. De seguida, efectuou-se a verificação dos dados, analisando os atributos nulos e implementando acções para a sua eliminação, sendo que, neste caso, a decisão foi não considerar ocorrências para as quais as variáveis não tinham valor. Depois de definida e verificada a informação seleccionada, desenvolveu-se um processo automático de extracção e tratamento da informação segundo as opções descritas anteriormente para alimentar os modelos.

No que diz respeito à limpeza dos dados, foram realizadas análises para identificar dados incompletos e *outliers* e procedeu-se à eliminação dos mesmos, quando caracterizados como valores irreais. Posteriormente, passou-se à construção dos atributos ditos primitivos em atributos relevantes, através da indução das variáveis por parte do utilizador. Desta forma, vários atributos foram definidos para melhorar o desempenho dos algoritmos de previsão. Também por este motivo, na fase de transformação dos dados, efectuou-se a normalização, não só através da utilização da função logaritmo, como também pela normalização através do desvio padrão e pela variação. Além disso, realizou-se ainda a transformação dos atributos quantitativos em qualitativos e vice-versa, à transformação de dados complexos e à redução de dados.

Concluídas todas estas fases, foram desenvolvidos os modelos preditivos com base nos métodos de Redes Neuronais, Árvores de Decisão e Regressão Logística. A escolha destes três métodos deveu-se ao facto destes serem aqueles que tradicionalmente são aplicados na previsão do cliente *churn* [20]. Recorrendo à ferramenta SPSS Clementine, para cada um destes métodos, foram testados todos os algoritmos disponibilizados pela ferramenta, embora apenas tenham sido apresentados os resultados referentes aos algoritmos que evidenciaram melhor desempenho (secções 4.4.2, 4.4.3 e 4.4.4). Dos três métodos estudados, a árvore de decisão foi o seleccionado quer, por motivos de requisitos de negócio, desempenho demonstrado nas três amostras testadas (método mais eficiente) e pela sua mais-valia na análise e suporte ao processo de tomada de decisão. Um outro factor que contribui para a escolha deste modelo foi a comparação dos resultados deste estudo com os resultados do estudo "*Applying data mining to Telecom churn*

management" [20] uma vez que os resultados obtidos são bastante idênticos, apesar de o presente estudo apenas ter tido em consideração um segmento bem definido, impossibilitando a análise do comportamento dos algoritmos na base global dos clientes. Este modelo classificou correctamente todas os casos para a classe *não churn* nas amostras de treino e teste e na amostra da validação classificou correctamente 99,93% dos casos. Já no que se refere à classe *churn*, verificou-se um decréscimo relativamente constante na classificação correcta dos casos, ao longo das três amostras, tal como se pode verificar na tabela 8. A justificação para tais resultados tão elevados advém do facto de ter sido efectuada uma medição do cliente *churn* muito em cima do acontecimento, sendo grande parte deles clientes *churn* involuntários por não pagamento. Contudo, caso na fase de análise esta situação tivesse sido considerada, dever-se-ia proceder à criação de dois tipos de modelos que permitissem distinguir os clientes *churn* voluntários dos involuntários. Se tivéssemos procedido desta forma, os valores apresentados pelos algoritmos seriam certamente muito diferentes. No que concerne às variáveis de classificação de *churn*, tal como já foi mencionado na secção 4.5, aquelas que melhor representam este modelo permitem inferir que um dos principais factores de classificação do cliente *churn* é o facto de este ter efectuado a primeira chamada há muito tempo, não ter efectuado chamadas num passado recente, e possuir uma média trimestral de chamadas efectuadas muito baixa. Neste caso trata-se de um cliente *churn* involuntário, dado que, por não pagamento da factura, o cliente viu o seu serviço barrado. Outros factores poderão ser explicativos de cliente *churn* caso sejam analisados todos os ramos da árvore. Sendo assim, a análise efectuada às árvores de decisão permite-nos concluir que as variáveis temporais relacionadas com o tráfego têm elevada importância na previsão do cliente *churn*. Convém ainda salientar que embora não tenha sido caso de estudo, para uma avaliação mais minuciosa das Árvores de Decisão, dever-se-ia aplicar este método a todos os segmentos, pois desta forma os resultados seriam mais credíveis.

Outra das questões não abordadas, pelo presente trabalho, foi a análise de métodos não supervisionados (*clustering*). Este tipo de análise é importante para a identificação de clientes que mudaram de *cluster* e identificação de mudanças no perfil de consumo, o que permitiria ter uma visão clara da mudança interna dos clientes (*churn* interno).

Como conclusão final e tendo em conta a não consideração na fase de análise da construção de dois modelos distintos para suportar os clientes voluntários e involuntários por não pagamento, os resultados embora limitativos, não inviabilizam a utilização do

referido modelo (Árvore de Decisão) noutros problemas de classificação e predição, que envolvam questões de abandono ou mudança de clientes para empresas concorrentes. Este estudo, além de ter servido para identificar e resolver questões relacionadas com o conceito *churn*, demonstrou também qual o comportamento típico do cliente *churn* involuntário por não pagamento e que pode, futuramente, ser uma arma importante no combate a questões relacionadas com fraude. Sobre os processos desenvolvidos, destaca-se a elevada importância das etapas de transformação, limpeza e normalização dos dados, nos resultados obtidos. No que se refere ao primeiro, as variáveis transformadas tiverem uma grande evidência nos modelos desenvolvidos o que explica a eleva importância desta etapa e a forma como ela deve ser abordada em trabalhos futuros. Em relação à limpeza dos dados, a identificação e substituição de valores omissos é sempre um problema nos processos de mineração de dados. Neste estudo, a opção adoptada foi a eliminação de registos com atributos omissos. Embora esta opção tenha possibilitado bons resultados, eliminou informação importante para a classificação do cliente *churn*. Soluções como a verificação e correcção destes atributos nos sistemas operacionais, ou a substituição pela média ou a moda, apresentam-se como alternativas viáveis para a substituição de valores omissos. No que respeita à normalização dos dados, os bons resultados obtidos pela regressão logística e pela rede neuronal só foram possíveis recorrendo a dados normalizados pela variação e pelo desvio padrão respectivamente. Em relação aos modelos estudados, a árvore de decisão foi o método escolhido para a classificação do cliente *churn*. No caso da rede neuronal, embora os bons resultado apresentados nas etapas de treino e teste, a questão do *overfitting* reduziu o seu desempenho na etapa de validação. No caso da regressão, o desempenho obtido foi muito reduzido, tornando obvia a escolha da arvore decisão, não só pelos resultados obtidos mas também pela sua fácil compreensão.

Para terminar, e como sugestões para eventuais acções futuras, precisaríamos de fazer algumas coisas para melhorar os processos implementados como os resultados alcançados, em especial, fazer:

- A diferenciação nas amostras do treino dos clientes *churn* voluntários dos involuntários.

- O aperfeiçoamento das amostras, com inclusão de mais variáveis derivadas e variáveis contínuas relacionadas com informação de chamadas de *incoming* - informação proveniente dos CDRs de bypass.
- A aplicação dos sistemas de mineração a todos os segmentos no sentido de se obter informação acerca de toda a base de clientes.
- A utilização de diferentes algoritmos para melhorar os resultados obtidos.
- A aplicação de outras técnicas de descoberta de padrões, como a análise associativa ou de agrupamento.
- A inserção no modelo de variáveis externas, nomeadamente, informação referente a empresas concorrentes e ao mercado onde estas empresas se inserem.

Bibliografia

[1] CHAPMAN, Pete (NCR), CLITON, Julian (SPSS), KHABAZA, Thomas (SPSS), REINARTZ, Thomas (DaimlerChrysler), WIRTH, Rüdiger (DaimlerChrysler), The CRISP-DM Process Model, CRISP-DM Discussion Paper, March 1999.

[2] MATTISON, Rob, The Telco Churn Management Handbook, USA-2005.

[3] FAYYAD, Usama; PIATETSKY-SHAPIR, Gregor; SMYTH, Padhraic, From Data Mining to Knowledge Discovery in Databases, January 1997.

[4] SHEARER C,"The CRISP-DM Model: The New Blueprint for DataMining". JOURNAL of Data Warehousing, Volume 5, Number 4, p. 13-22, 2000.

[5] INMON, W. H., Building the data warehouse, 3rd Edition, Wiley, 2002.

[6] HOTTELECOM, Middle East – Mobile subscribers Statistics and Forecast by country, December 2007.

[7] STROUSSE, K. G., Marketing Telecommunications Services: New Approaches for a Changing Environment. Norwood: Artech House, 1999, p.271.

[8] BERRY, M. J. A.; LINOFF, G. S., Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management. New York: John Wiley & Sons, 1997, p.311.

[9] LEJEUNE, M. A. P. M., Measuring the Impact of Data Mining on Churn. Internet Research, v11 n5, 2001, p375-87.

[10] KOTLER P.; ARMSTRONG G., Princípios do Marketing (9. Edição). Prentice Hall, 1999, p.77.

[11] SWIFT R. S., Accelerating Customer Relationships: Using CRM and Relationship Technologies. Prentice Hall, 2001, p97-110.

[12] FRANK, E.; HALL, M.; HOLMES G.; KIRKBY, R.; PFAHRINGER, B.; WITTEN I. H., WEKA A Machine Learning Workbench for Data Mining. New Zealand.

[13] DAVENPORT, T. H.; HARRIS, J. G., Competing on Analytics: The New Science of Winning. Harvard Business School Press, 2007.

[14] HAN, J.; KAMBER, M., Data Mining: Concepts and Techniques, 2ª ed. Morgan Kaufmann Publishers, March 2006.

[15] FRANK, E.; WITTEN, I. H., Data Mining Practical Machine Learning Tools and Techniques (2ª Edition). USA: Morgan Kaufmann Publishers, 2005, p61-82.

[16] LIPPMANN, R. P., An introduction to computing with neural nets. New York: ACM SIGARCH Computer Architecture News, 1988.

[17] MONTGOMERY, D. C.; PECK E. A.; VINING, G. G., Introduction to Linear Regression Analysis (3rd Edition). Wiley Series In Probability and Statistics, 2006.

[18] Sítio oficial da SAS, *website*: http://www.sas.com/offices/europe/uk/technologies/analytics/datamining/miner/semma.html Acesso em 12 de Setembro de 2009.

[19] ANDRADE, D. F., Uma Analise de Cancelamento em Telefonia Utilizando Mineração de Dados. Rio de Janeiro, 2007.

[20] HUNG, S.; YEN, D. C.; WANG, H., Applying data mining to telecom churn management. Taiwan: Department of Information Management, National Chung Cheng University, 2006.

[21] BARANAUSKAS, J.A; MONARD, M. C., Metodologias para a Selecção de Atributos Relevantes, website: http://dfm.ffclrp.usp.br/~augusto/publications/1998-sbia.pdf Acesso em 12 de Setembro de 2009.

[22] WEISS, S. M.; INDURKHYA, N., Predictive Data Mining – a practical guide. San Francisco: Morgan Kaufmann Publishers Inc., 1998, p230.

[23] BATISTA, G. E., Pré-processamento de Dados em Aprendizagem de Máquina Supervisionada. São Carlos: Instituto de Ciências Matemáticas e de Computação, 2003.

[24] NETO, A. P, F.; MATTOZO, T. C.; COSTA, J. A. F., Uma Aplicação de Mineração de Dados no Gerenciamento do Churn em Serviços de Banda Larga de Telecomunicações. Niteroi, Brasil, 2008.

[25] FRANK E; WITTEN, I. H., Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Newark: Morgan Kaufmann Publishers Inc., 2000, p376.

[26] HECHT-NIELSEN, R. Applications of Counterpropagation Networks. Neural. Networks Journal, 1988. p. 131-140.

[27] Sítio oficial do SPSS Clementine, *website*: http://www.spss.com/software/modeling/modeler/ Acesso em 11 de Outubro de 2009

Anexo A

As seguintes tabelas reflectem a normalização efectuada. Cada tabela está dividida em dados não normalizados e dados normalizados pela respectiva função. Os valores, das colunas da secção dos dados não normalizados, correspondem aos valores necessários para o cálculo da respectiva normalização. As outras colunas correspondem às características das variáveis após a normalização.

	Dados Não Normalizados					Dados Normalizados pela Função Logaritmo				
Variável	Mínimo	Máximo	Média	Desvio Padrão	Enviesamento	Mínimo	Máximo	Média	Desvio Padrão	Enviesamento
SNAP_CS_COMPLAIN_CALLS	0,000	1,000	0,001	0,022	44,721	0,000	0,000	0,000	0,000	--
SNAP_CS_OPEN_COMPLAINTS	0,000	0,000	0,000	0,000	--	0,000	0,000	0,000	0,000	--
SNAP_CS_INFORMATION_CALLS	0,000	7,000	0,210	0,733	4,908	0,000	0,845	0,022	0,102	5,012
SNAP_CS_TT_AVERAGE_TIME	0,000	51620,000	25810,000	36500,852	--	0,000	4,713	0,002	0,105	4,472
SNAP_CS_SERVICE_CALLS	0,000	15,000	0,390	1,153	4,970	0,000	1,176	0,043	0,149	3,825
SNAP_ACTIVATION_DAYS	193,000	5391,000	2210,580	1530,553	0,433	2,286	3,732	3,211	0,368	-0,382
SNAP_PARTY_PAYM_METHOD	-1,000	3,000	0,999	0,326	-0,107	0,000	0,477	0,006	0,054	8,545
SNAP_PARTY_NATIONALITY	0,000	153,000	85,011	57,695	-0,241	0,000	2,185	1,575	0,822	-1,253
SNAP_CHURN	0,000	1,000	--	--	--	0,000	1,000	--	--	--
SNAP_GRACE_PERIOD	0,000	1,000	0,060	0,238	3,708	0,000	0,000	0,000	0,000	--
INV_COUNT_VOICE_OUT_NAC	0,000	10193,000	455,846	506,560	6,782	0,000	4,008	2,300	0,833	-1,786
INV_COUNT_VOICE_OUT_INT	0,000	1045,000	75,844	88,552	3,243	0,000	3,019	1,504	0,737	-0,938
INV_COUNT_VOICE_OUT	0,000	10193,000	531,690	520,557	6,094	0,000	4,008	2,404	0,822	-2,066
INV_COUNT_VOICE_OUT_ROAM	0,000	582,000	14,367	41,961	5,760	0,000	2,765	0,398	0,697	1,450
INV_AIRTIME_VOICE_OUT_NAC	0,000	0,000	0,000	0,000	--	0,000	0,000	0,000	0,000	--
INV_AIRTIME_VOICE_OUT_INT	0,000	253831,000	19496,140	22492,235	3,262	0,000	5,405	3,610	1,435	-1,802
INV_AIRTIME_VOICE_OUT	0,000	484212,000	81146,947	65232,750	1,498	0,000	5,685	4,422	1,381	-2,675
INV_AIRTIME_VOICE_OUT_ROAM	0,000	20302,280	280,598	997,388	8,419	0,000	4,308	0,728	1,204	1,236
INV_OUT_USAGE_VOICE_NAC	0,000	0,000	0,000	0,000	--	0,000	0,000	0,000	0,000	--
INV_OUT_USAGE_VOICE_INT	0,000	9897,810	731,232	817,085	2,931	-0,155	3,996	2,362	1,015	-1,418
INV_OUT_USAGE_VOICE	0,000	10006,050	1089,721	868,603	2,453	-0,523	4,000	2,715	0,897	-2,347
INV_OUT_USAGE_ROAM	0,000	20806,930	298,265	1030,529	8,172	-0,174	4,318	0,752	1,223	1,192
INV_MODULES	0,000	778,000	1,170	24,917	30,408	0,000	2,891	0,021	0,185	9,787
INV_RENTS	-225,330	23150,000	90,773	529,219	41,450	0,000	4,365	1,631	0,447	0,919
INV_OPENING_BALANCE	-10921,420	1189254,460	3969,752	26977,729	42,518	0,866	6,075	3,363	0,416	-0,616
EVE_FIRST_OUT_CALL_DATE	0,000	363,220	69,843	59,933	0,552	0,000	2,560	1,495	0,702	-0,779
EVE_LAST_OUT_CALL_DATE	-1,000	304,000	47,738	45,147	0,374	0,000	2,483	1,179	0,848	-0,349
EVE_LAST_CALL_CHARGED_DATE	-1,000	304,000	47,738	45,147	0,374	0,000	2,483	1,179	0,848	-0,349
EVE_LAST_3M_USAGE_AVG	0,000	1766,540	52,132	109,677	7,012	-1,222	3,247	1,152	0,779	-0,399
PAR_PARTY_GENDER	0,000	2,000	0,198	0,415	1,810	0,000	0,301	0,002	0,025	1,184
DAYS_OF_AGE	7363,000	18950,000	12984,323	2159,995	-0,115	0,000	4,278	0,197	0,878	4,235
PAR_PARTY_MARITAL_STATUS	0,000	1,000	0,001	0,022	44,721	0,000	0,000	0,000	0,000	--
PAR_PARTY_PROFESSION	0,000	0,000	0,000	0,000	--	0,000	0,000	0,000	0,000	--
INV_AVG_AIRTIME_CALL_OUT_INT	1,000	4503,000	261,075	230,065	4,688	0,000	3,654	2,105	0,814	-1,930
INV_AVG_USAGE_CALL_OUT_INT	0,000	103,000	9,195	6,827	2,781	0,000	2,013	0,834	0,380	-0,940
INV_AVG_AIRTIME_CALL_OUT_ROAM	0,000	336,000	5,965	13,271	9,863	0,000	2,526	0,323	0,541	1,298
INV_AVG_USAGE_CALL_OUT_ROAM	0,000	337,000	6,764	14,948	8,801	0,000	2,528	0,348	0,568	1,220

Dados Não Normalizados						Dados Normalizados pela Variação				
Variável	Mínimo	Máximo	Média	Desvio Padrão	Enviesamento	Mínimo	Máximo	Média	Desvio Padrão	Enviesamento
SNAP_CS_COMPLAIN_CALLS	0,000	1,000	0,001	0,022	44,721	0,000	1,000	0,001	0,022	44,721
SNAP_CS_OPEN_COMPLAINTS	0,000	0,000	0,000	0,000	--	0,000	0,000	0,000	0,000	--
SNAP_CS_INFORMATION_CALLS	0,000	7,000	0,210	0,733	4,908	0,000	1,000	0,030	0,105	4,908
SNAP_CS_TT_AVERAGE_TIME	0,000	51620,000	25810,000	36500,852	--	--	--	--	--	--
SNAP_CS_SERVICE_CALLS	0,000	15,000	0,390	1,153	4,970	0,000	1,000	0,026	0,077	4,970
SNAP_ACTIVATION_DAYS	193,000	5391,000	2210,580	1530,553	0,433	0,000	1,000	0,388	0,294	0,433
SNAP_PARTY_PAYM_METHOD	-1,000	3,000	0,999	0,326	-0,107	--	--	--	--	--
SNAP_PARTY_NATIONALITY	0,000	153,000	85,011	57,695	-0,241	0,000	1,000	0,556	0,377	-0,241
SNAP_CHURN	0,000	1,000	--	--	--	0,000	1,000	--	--	--
SNAP_GRACE_PERIOD	0,000	1,000	0,060	0,238	3,708	0,000	1,000	0,060	0,238	3,708
INV_COUNT_VOICE_OUT_NAC	0,000	10193,000	455,846	506,560	6,782	0,000	1,000	0,045	0,050	6,782
INV_COUNT_VOICE_OUT_INT	0,000	1045,000	75,844	88,552	3,243	0,000	11,000	0,073	0,085	3,243
INV_COUNT_VOICE_OUT	0,000	10193,000	531,690	520,557	6,094	0,000	1,000	0,052	0,051	6,094
INV_COUNT_VOICE_OUT_ROAM	0,000	582,000	14,367	41,961	5,760	0,000	1,000	0,025	0,072	5,760
INV_AIRTIME_VOICE_OUT_NAC	0,000	0,000	0,000	0,000	--	0,000	0,000	0,000	0,000	--
INV_AIRTIME_VOICE_OUT_INT	0,000	253831,000	19496,140	22492,235	3,262	0,000	1,000	0,077	0,089	3,262
INV_AIRTIME_VOICE_OUT	0,000	484212,000	81146,947	65232,750	1,498	0,000	1,000	0,168	0,135	1,498
INV_AIRTIME_VOICE_OUT_ROAM	0,000	20302,280	280,598	997,388	8,419	0,000	1,000	0,014	0,049	8,419
INV_OUT_USAGE_VOICE_NAC	0,000	0,000	0,000	0,000	--	0,000	0,000	0,000	0,000	--
INV_OUT_USAGE_VOICE_INT	0,000	9897,810	731,232	817,085	2,931	0,000	1,000	0,074	0,083	2,931
INV_OUT_USAGE_VOICE	0,000	10006,050	1089,721	868,603	2,453	0,000	1,000	0,109	0,087	2,453
INV_OUT_USAGE_ROAM	0,000	20806,930	298,265	1030,529	8,172	0,000	1,000	0,014	0,050	8,172
INV_MODULES	0,000	778,000	1,170	24,917	30,408	0,000	1,000	0,002	0,032	3,041
INV_RENTS	-225,330	23150,000	90,773	529,219	41,450	0,000	1,000	0,014	0,023	41,450
INV_OPENING_BALANCE	-10921,420	1189254,460	3969,752	26977,729	42,518	0,000	1,000	0,012	0,022	42,518
EVE_FIRST_OUT_CALL_DATE	0,000	363,220	69,843	59,933	0,552	0,000	1,000	0,191	0,165	0,559
EVE_LAST_OUT_CALL_DATE	-1,000	304,000	47,738	45,147	0,374	0,000	1,000	0,160	0,148	0,374
EVE_LAST_CALL_CHARGED_DATE	-1,000	304,000	47,738	45,147	0,374	0,000	1,000	0,160	0,148	0,374
EVE_LAST_3M_USAGE_AVG	0,000	1766,540	52,132	109,677	7,012	0,000	1,000	0,028	0,061	7,111
PAR_PARTY_GENDER	0,000	2,000	0,198	0,415	1,810	0,000	1,000	0,099	0,208	1,810
DAYS_OF_AGE	7363,000	18950,000	12984,323	2159,995	-0,115	-0,635	1,000	-0,582	0,243	4,422
PAR_PARTY_MARITAL_STATUS	0,000	1,000	0,001	0,022	44,721	0,000	1,000	0,001	0,022	44,721
PAR_PARTY_PROFESSION	0,000	0,000	0,000	0,000	--	0,000	0,000	0,000	0,000	--
INV_AVG_AIRTIME_CALL_OUT_INT	1,000	4503,000	261,075	230,065	4,688	0,000	1,000	0,058	0,051	4,688
INV_AVG_USAGE_CALL_OUT_INT	0,000	103,000	9,195	6,827	2,781	0,000	1,000	0,089	0,066	2,781
INV_AVG_AIRTIME_CALL_OUT_ROAM	0,000	336,000	5,965	13,271	9,863	0,000	1,000	0,018	0,039	9,863
INV_AVG_USAGE_CALL_OUT_ROAM	0,000	337,000	6,764	14,948	8,801	0,000	1,000	0,020	0,044	8,801

Dados Não Normalizados						Dados Normalizados pelo Desvio Padrão				
Variável	Mínimo	Máximo	Média	Desvio Padrão	Enviesamento	Mínimo	Máximo	Média	Desvio Padrão	Enviesamento
SNAP_CS_COMPLAIN_CALLS	0,000	1,000	0,001	0,022	44,721	-0,045	45,409	-0,023	1,016	44,721
SNAP_CS_OPEN_COMPLAINTS	0,000	0,000	0,000	0,000	--	0,000	0,000	0,000	0,000	--
SNAP_CS_INFORMATION_CALLS	0,000	7,000	0,210	0,733	4,908	-0,286	9,263	0,000	1,000	4,908
SNAP_CS_TT_AVERAGE_TIME	0,000	51620,000	25810,000	36500,852	--	--	--	--	--	--
SNAP_CS_SERVICE_CALLS	0,000	15,000	0,390	1,153	4,970	-0,338	12,671	0,000	1,000	4,970
SNAP_ACTIVATION_DAYS	193,000	5391,000	2210,580	1530,553	0,433	-1,318	2,078	0,000	1,000	0,433
SNAP_PARTY_PAYM_METHOD	-1,000	3,000	0,999	0,326	-0,107	--	--	--	--	--
SNAP_PARTY_NATIONALITY	0,000	153,000	85,011	57,695	-0,241	-1,473	1,178	0,000	1,000	-0,241
SNAP_CHURN	0,000	1,000	--	--	--	0,000	1,000	--	--	--
SNAP_GRACE_PERIOD	0,000	1,000	0,060	0,238	3,708	-0,252	3,950	0,000	0,998	3,708
INV_COUNT_VOICE_OUT_NAC	0,000	10193,000	455,846	506,560	6,782	-0,900	19,222	0,000	1,000	6,782
INV_COUNT_VOICE_OUT_INT	0,000	1045,000	75,844	88,552	3,243	-0.856	10,944	0,000	1,000	3,243
INV_COUNT_VOICE_OUT	0,000	10193,000	531,690	520,557	6,094	-1,021	18,560	0,000	1,000	6,094
INV_COUNT_VOICE_OUT_ROAM	0,000	582,000	14,367	41,961	5,760	-0,342	13,528	0,000	1,000	5,760
INV_AIRTIME_VOICE_OUT_NAC	0,000	0,000	0,000	0,000	--	0,000	0,000	0,000	0,000	--
INV_AIRTIME_VOICE_OUT_INT	0,000	253831,000	19496,140	22492,235	3,262	-0,867	10,418	0,000	1,000	3,262
INV_AIRTIME_VOICE_OUT	0,000	484212,000	81146,947	65232,750	1,498	-1,244	6,179	0,000	1,000	1,498
INV_AIRTIME_VOICE_OUT_ROAM	0,000	20302,280	280,598	997,388	8,419	-0,281	20,074	0,000	1,000	8,419
INV_OUT_USAGE_VOICE_NAC	0,000	0,000	0,000	0,000	--	0,000	0,000	0,000	0,000	--
INV_OUT_USAGE_VOICE_INT	0,000	9897,810	731,232	817,085	2,931	-0,895	11,219	0,000	1,000	2,931
INV_OUT_USAGE_VOICE	0,000	10006,050	1089,721	868,603	2,453	-1,255	10,265	0,000	1,000	2,453
INV_OUT_USAGE_ROAM	0,000	20806,930	298,265	1030,529	8,172	-0,289	19,901	0,000	1,000	8,172
INV_MODULES	0,000	778,000	1,170	24,917	30,408	-0,047	31,177	0,000	1,000	30,408
INV_RENTS	-225,330	23150,000	90,773	529,219	41,450	-0,597	43,572	0,000	1,000	41,450
INV_OPENING_BALANCE	-10921,420	1189254,460	3969,752	26977,729	42,518	-0,552	43,936	0.000	1,000	42,518
EVE_FIRST_OUT_CALL_DATE	0,000	363,220	69,843	59,933	0,552	-1,165	4,895	-0,008	1,001	0,559
EVE_LAST_OUT_CALL_DATE	-1,000	304,000	47,738	45,147	0,374	-1,080	5,676	0,000	1,000	0,374
EVE_LAST_CALL_CHARGED_DATE	-1,000	304,000	47,738	45,147	0,374	-1,080	5,676	0,000	1,000	0,374
EVE_LAST_3M_USAGE_AVG	0,000	1766,540	52,132	109,677	7,012	-0,475	15,631	-0,019	0,985	7,111
PAR_PARTY_GENDER	0,000	2,000	0,198	0,415	1,810	-0,477	4,342	-0,001	1,001	1,810
DAYS_OF_AGE	7363,000	18950,000	12984,323	2159,995	-0,115	-6,011	2,762	-5,723	1,304	4,422
PAR_PARTY_MARITAL_STATUS	0,000	1,000	0,001	0,022	44,721	-0,045	45,409	-0,023	1,016	44,721
PAR_PARTY_PROFESSION	0,000	0,000	0,000	0,000	--	0,000	0,000	0,000	0,000	--
INV_AVG_AIRTIME_CALL_OUT_INT	1,000	4503,000	261,075	230,065	4,688	-1,130	18,438	0,000	1,000	4,688
INV_AVG_USAGE_CALL_OUT_INT	0,000	103,000	9,195	6,827	2,781	-1,347	13,740	0,000	1,000	2,781
INV_AVG_AIRTIME_CALL_OUT_ROAM	0,000	336,000	5,965	13,271	9,863	-0,449	24,869	0,000	1,000	9,863
INV_AVG_USAGE_CALL_OUT_ROAM	0,000	337,000	6,764	14,948	8,801	-0,453	22,092	0,000	1,000	8,801

Anexo B

A figura 27 representa o resultado da árvore de decisão pelo algoritmo CHAID. Do lado esquerdo da imagem, observa-se a estrutura da árvore de decisão de uma outra perspectiva (horizontal) e respectivos valores dos pontos de corte. Do lado direito pode-se analisar a importância das várias variáveis e respectivos valores.

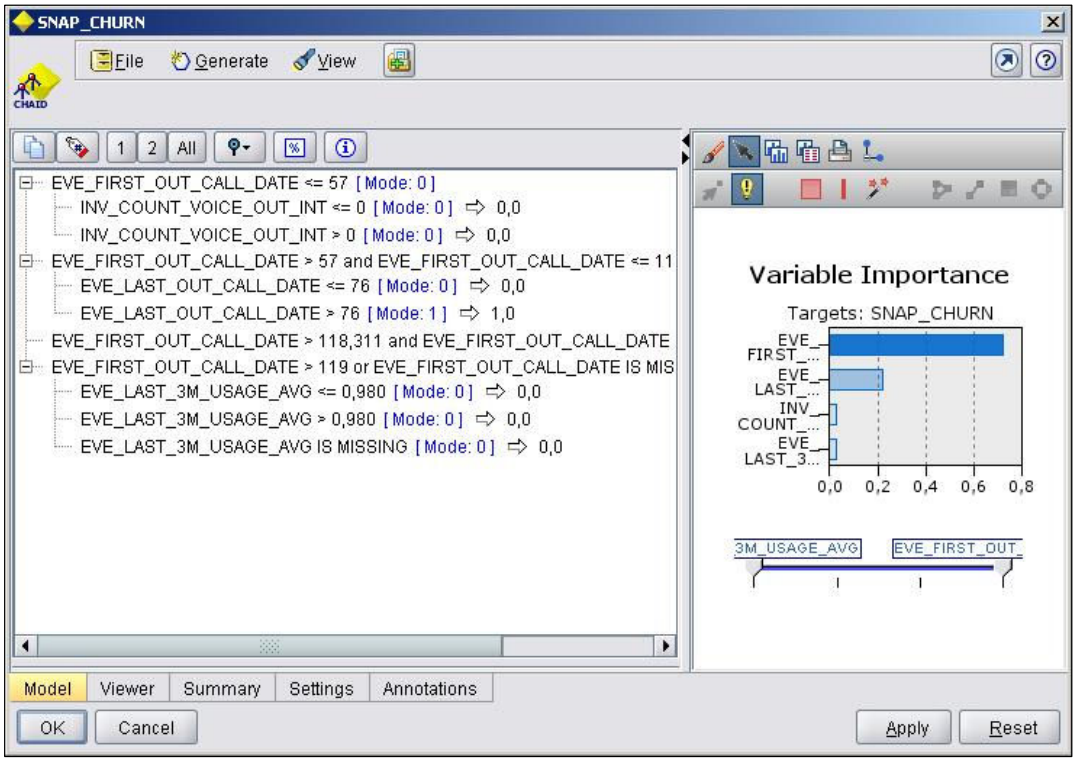

Figura 27 – Resultados da Árvore de Decisão

Anexo C

A figura 28, inicialmente, apresenta um sumário dos casos avaliados. No quadro *Step Summary*, apresenta-se a lista dos efeitos adicionados ou removidos em cada etapa da criação do modelo. No quadro *Model Fitting Information*, exibe-se o teste *likelihood-ratio* do modelo contra um teste em que todos os coeficientes dos parâmetros são zero e por fim a estatística *Chi-Square*. O quadro *Pseudo R-Square,* demonstra os resultados de três tipos diferentes de técnicas de estatística, aplicadas no ajuste do modelo. Estas técnicas são de certa forma análogas à técnica de estatística *R-Square*. No último quadro, mostra-se as estimativas dos coeficientes da equação, os testes dos coeficientes e intervalos de confiança para os rácios.

Case Processing Summary

		N	Marginal Percentage
SNAP_CHURN	,000	1880	94,0%
	1,000	120	6,0%
Valid		2000	100,0%
Missing		0	
Total		2000	
Subpopulation		2000(a)	

a. The dependent variable has only one value observed in 2000 (100,0%) subpopulations.

Step Summary

Model		Action	Effect(s)	Model Fitting Criteria	Effect Selection Tests		
				-2 Log Likelihood	Chi-Square(a,b)	df	Sig.
Step 0	0	Entered	Intercept	907,870	.		
Step 1	1	Entered	INV_OUT_USAGE_VOICE	599,559	141,712	1	,000
Step 2	2	Entered	EVE_LAST_OUT_CALL_DATE	564,896	31,720	1	,000
Step 3	3	Entered	EVE_FIRST_OUT_CALL_DATE	429,773	44,980	1	,000
Step 4	4	Entered	SNAP_ACTIVATION_DAYS	404,829	22,277	1	,000
Step 5	5	Entered	INV_AVG_USAGE_CALL_OUT_ROAM	390,237	7,002	1	,008
Step 6	6	Entered	SNAP_GRACE_PERIOD	378,109	7,074	1	,008
Step 7	7	Entered	INV_COUNT_VOICE_OUT	372,802	4,465	1	,035
Step 8	8	Entered	INV_AVG_AIRTIME_CALL_OUT_INT	365,190	6,107	1	,013

Stepwise Method: Forward Stepwise

a. The chi-square for entry is based on the score test.

b. The chi-square for removal is based on the likelihood ratio test.

Model Fitting Information

Model	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	907,870			
Final	365,190	542,680	8	,000

Pseudo R-Square

Cox and Snell	,238
Nagelkerke	,651
McFadden	,598

Parameter Estimates

SNAP_CHURN(a)		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp(B)	
								Lower Bound	Upper Bound
,000	Intercept	,397	,438	,823	1	,364			
	SNAP_ACTIVATION_DAYS	2,727	,610	19,952	1	,000	15,281	4,619	50,554
	SNAP_GRACE_PERIOD	21,076	,000	.	1	.	1423448731,800	1423448731,800	1423448731,800
	INV_COUNT_VOICE_OUT	24,201	9,395	6,636	1	,010	32383561201,705	326,338	3213527804522777000,000
	INV_OUT_USAGE_VOICE	13,752	5,417	6,443	1	,011	938017,986	22,955	38330753583,378
	EVE_FIRST_OUT_CALL_DATE	50,525	6,266	65,024	1	,000	876106558776403000000000,000	4066639020018803 0,000	1887462099666478000000000000,000
	EVE_LAST_OUT_CALL_DATE	-56,319	6,815	68,289	1	,000	3,48E-025	5,50E-031	2,20E-019
	INV_AVG_AIRTIME_CALL_OUT_INT	11,619	4,589	6,411	1	,011	111165,655	13,807	895066389,567
	INV_AVG_USAGE_CALL_OUT_ROAM	25,559	8,637	8,758	1	,003	125929672241,211	5606,701	2828451675796480000,000

a. The reference category is: 1,000.

Figura 28 – Resultados Obtidos na Regressão Logística