U.PORTO
FEP FACULDADE DE ECONOMIA
UNIVERSIDADE DO PORTO

Previsão de *churn* em telecomunicações

Por

Daniel José Araújo Meireles

Dissertação de Mestrado em Modelação, Análise de Dados e Sistemas de Apoio à Decisão

Orientado por:

Prof. Doutor João Manuel Portela da Gama

2014

Nota Biográfica

Daniel José Araújo Meireles nasceu em Porto – Portugal – a 15 de Janeiro de 1987. Em 2009, concluiu o Mestrado Integrado em Engenharia Industrial e Gestão na Faculdade de Engenharia da Universidade do Porto.

Nesse mesmo ano, iniciou a sua atividade profissional na empresa *Bosch Car Multimedia*, primeiro na forma de estágio curricular e seguidamente como funcionário contratado. Desde então, tem exercido várias funções ligadas ao departamento de produção, sendo que atualmente trabalha fundamentalmente em análise de dados e ferramentas para cálculo de indicadores.

Agradecimentos

À Inês, pela companhia e apoio constantes, que foram motivo para fazer mais e melhor.

Ao Professor João Gama, pelas sugestões, ideias e disponibilidade permanente.

Ao N., pelo acompanhamento, disponibilidade e abertura necessárias para esclarecer todas as dúvidas. Assim como, pelas questões levantadas que levaram a enriquecer este trabalho.

À operadora, pela possibilidade de trabalhar com dados reais, o que confere a este trabalho uma grande credibilidade.

Este trabalho foi apoiado pelo projecto de desenvolvimento Sibila (NORTE-07-0124-FEDER-000059), financiado pelo Programa Operacional Regional do Norte (ON.2 O Novo Norte), sob o Quadro de Referência Estratégico Nacional (QREN), pelo Fundo de Desenvolvimento Regional (FEDER), e por fundos nacionais, através da Agência de Financiamento Portuguesa, Fundação para a Ciência e a Tecnologia (FCT) e pela Comissão Europeia através do projecto MAESTRA (Nº ICT-2013-612944).

Resumo

Esta dissertação aborda o tema da previsão de *churn* em clientes pré-pagos de uma operadora de telecomunicações móveis. *Churn* significa a mudança de um cliente de uma empresa para outra. A capacidade de uma empresa para conseguir prever com precisão quando é que um cliente irá deixar de usar o seu serviço, para que possa tomar medidas, é fundamental. Este facto é principalmente relevante no mercado das telecomunicações móveis em que o custo de aquisição de um novo cliente ultrapassa largamente o custo de manter um cliente atual. O mercado de clientes pré-pagos abrange aqueles clientes que não estão ligados contratualmente à empresa, logo existe menos informação sobre esses clientes. Portanto, é também mais difícil prever o *churn* pois podem sair da operadora sem qualquer notificação, apenas deixando de utilizar o serviço. Nesta dissertação são utilizados dados reais anonimizados de uma operadora de telecomunicações móveis com milhões de clientes, sendo utilizadas duas abordagens distintas. A primeira baseada num método de dispersão de influência com origem nos clientes identificados como *churners*, a segunda utilizando os milhões de dados gerados pelos clientes nas suas interações construindo atributos de utilização, entre outros, de forma a produzir um modelo de classificação com o objetivo de detetar antecipadamente os potenciais *churners*. A eficácia da previsão é avaliada de acordo com a métrica de *lift*, que é a percentagem do total de *churners* encontrados, quando se escolhe apenas uma determinada percentagem do total de clientes.

Palavras-Chave: *Churn*, Telecomunicações, Pré-pago, Classificação, Influência

Abstract

This dissertation addresses the issue of churn prediction in prepaid clients of a mobile telecommunications company. Churn means the action of a client leaving a company for another one. The ability of a company to be able to accurately predict when a customer will stop using its services, so it can act on it, is paramount. This fact is particularly relevant in the mobile telecommunications market where the cost of acquiring new customers largely outweighs the cost of keeping a current one. The prepaid market includes those customers which are not contractually bound to the company, thus there is less information about them. This makes it harder to predict when they will churn, because they can leave the company without notification, they simply stop using the service. In this work, real anonymized data is used from a mobile telecommunications company with millions of clients, and two distinct approaches are tested. The first one based on an influence spread model originating from the churners, while the second one makes use of the millions of data generated by the clients in their interactions by computing utilization attributes, among others, aiming to build a classification model which can detect with anticipation the potential churners. The prediction effectiveness is evaluated according to the lift metric, which is the fraction of the total number of churners found, when just a certain fraction of all customers are chosen.

Keywords: Churn, Telecommunications, Prepaid, Classification, Influence

Conteúdos

Índice de figuras

Índice de tabelas

1. Introdução, Problema e Motivação

O problema a estudar é a previsão de *churn* em telecomunicações utilizando variáveis derivadas da rede de contactos dos clientes, isto é, a análise da rede social construída através das interações entre os utilizadores.

O *churn* é definido como o abandono de um cliente de um serviço para outro, e pode ter diferentes motivos. São vários os trabalhos, como Radosavljevik, Putten, & Larsen (2010), que abordam este tema de forma profunda e com diversas variações nas variáveis chave, como por exemplo, o período de inatividade a partir do qual é despoletado o evento de churn. Outras análises incluem a seleção de classificadores híbridos, no caso redes neuronais, utilizando dados de perfil dos clientes para atributos, ver Tsai & Lu (2009).

As motivações mais comuns para o abandono do serviço (*churn)* são: tarifários mais vantajosos na concorrência, má qualidade do serviço, número de amigos pertencentes a outros operadores, má assistência ao cliente, entre outros.

A abordagem a estes problemas que já são alvo de investigação pela comunidade científica há alguns anos, está agora a ser afetada, de certa forma, pelo crescimento exponencial da importância das redes sociais na sociedade de hoje em dia. Nesse sentido, vários artigos têm sido escritos recentemente, tais como Robins (2013) em que são abordadas as principais características das redes sociais e as suas mais populares formas de modelação, ou o trabalho, menos recente, de Bohn et al. (2009) sobre a melhor forma de representação do peso da ligação entre dois nós (utilizadores) numa rede social, aplicada à área das telecomunicações.

Na mesma direção vemos problemas como a exploração dos aspetos espaciais das redes sociais na descoberta de associações entre pessoas pela utilização de *smartphones*, ver Slingsby, Beecham, & Wood (2013), ou por exemplo, na resolução de problemas de recomendação de produtos, muito popular nos grandes retalhistas *online*, ver Sohn, Bae, & Chung (2013).

Com o crescimento das redes sociais tornou-se imperativa a análise aos grupos ou comunidades distinguíveis na rede, normalmente compostas por indivíduos que partilham um determinado número de características comuns. Portanto, a análise de redes sociais tornou-se essencial para a deteção dessas mesmas comunidades, ver Ma et al. (2013), onde os autores conseguem a deteção de comunidades através de diferentes formas de visualização da rede.

Um outro trabalho, neste caso específico na área do *churn* aplicado a um serviço de rede social, os autores propõem-se a prever não o *churn* individual mas sim o *churn* da própria comunidade, ver Oentaryo, Lo, Zhu, & Prasetyo (2012).

Os dados para construir a rede serão fornecidos por uma operadora de telecomunicações e incluirão as chamadas e interações dos seus clientes durante um período de 12 meses. A solução que será proposta na elaboração da tese poderá também ser testada e avaliada na rede de utilizadores da operadora durante ou após a realização da tese.

Neste estudo, o objetivo é encontrar futuros *churners* na base de clientes com tarifários pré-pagos. Uma das dificuldades em termos de conhecimento dos clientes é o facto destes utilizadores não serem obrigados, pois não existe nenhum vinculo contratual, a revelarem a sua idade, género, local de nascimento, morada, entre outros. O principal conhecimento que temos dos clientes será encontrado através da análise da sua rede de contactos, isto é, interações ao nível de chamadas, *SMS* e *MMS*, dados de recarregamentos e outros dados, como telemóvel utilizado ou modo como o cliente entrou para a operadora.

A grande motivação para este problema é a possibilidade de com a deteção antecipada do evento de *churn*, possibilitar à empresa tomar ações de retenção direcionada aos clientes mais propensos a fazerem *churn* de acordo com o modelo construído. A eficácia dessas campanhas de retenção estará também dependente da antecipação com que se consiga identificar os clientes porque quando forem abordados pela empresa, a decisão de abandonarem o serviço pode já ser irrevogável.

Outra das motivações para este trabalho é verificar o valor acrescentado da utilização de variáveis derivadas da rede social dos utilizadores, nomeadamente rácio entre chamadas para utilizadores de outras redes e chamadas totais, número de contactos que abandonaram o serviço num período seguinte ou número de contactos totais.

A abordagem proposta ao problema terá como objetivo cumprir os requisitos desejáveis num sistema de previsão de *churn*. Segundo Balle, Casas, & Catarineu (2011) os requisitos são os seguintes:

- Precisão: na avaliação do classificador este deve ter uma medida de *recall* elevado (pelo menos todos os *churners* são identificados) e precisão relativamente elevada (não haver muitos falsos positivos).
- Desempenho: a rapidez com que o modelo pode ser executado com novos dados é essencial para poderem ser tomadas as decisões certas no tempo certo.
- Flexibilidade: o modelo tem que conseguir manter-se com bons indices de previsão com a previsivel alteração nos padrões dos clientes que serão introduzidos cada vez que for necessario fazer uma previsao de *churn*.
- Escala: o modelo tem que reagir de forma aceitável ao aumento de dados com que poderá ser alimentado.
- Segmentação: esta característica prende-se com a capacidade de serem retirados dados concretos sobre o perfil de utilizadores mais propensos a deixarem o serviço e possivelmente incluir variáveis baseadas na experiência dos analistas que conhecem o negócio.

O modelo que será proposto deverá incluir atributos não-relacionais (dados de perfil do cliente), juntamente com variáveis retiradas da rede de contactos (dados de conetividade) e atributos relacionais, tais como, a influência acumulada por um determinado utilizador após eventos de *churn* num determinado período.

Este será então um modelo híbrido, em linha com as propostas mais recentes dos investigadores que se dedicaram a este problema, por exemplo, Verbeke, Martens, & Baesens (2014).

Por outro lado, poderá também ser testado um modelo de propagação de influência que terá em conta o resultado de um classificador normal (árvores de decisão, regressão logística, entre outros) para inicialização dos nós não-*churners*, ver Kusuma & Radosavljevik (2013).

A partir deste capítulo, esta dissertação está organizada da seguinte forma: no capítulo 2 é feita a revisão da literatura sobre o tema, no capítulo 3 pode-se encontrar a descrição dos atributos disponíveis e uma análise ao período de dormência a utilizar para caracterizar um cliente *churner*. No capítulo 4 são analisados em pormenor os atributos e selecionados aqueles que garantem um maior poder preditivo assim como a sua relação com a classe que queremos prever. No capítulo 5 são explicados os dois modelos utilizados para a previsão de *churn*. Finalmente, no capítulo 6, são retiradas as conclusões finais e sugeridos possíveis trabalhos futuros.

2. Revisão da literatura

De seguida será apresentado um resumo dos artigos com as abordagens entendidas como determinantes que têm sido feitas ao tema da previsão de *churn* em telecomunicações.

O artigo de Dasgupta & Singh (2008) é dos primeiros a abordar o problema do *churn* em telecomunicações utilizando análise de redes sociais. O objetivo era encontrar uma relação entre a probabilidade de um utilizador deixar de utilizar o serviço (*churn*) e o número de amigos, ou seja, pessoas com quem esse utilizador tinha contacto telefónico, que já tinham deixado o serviço. Eles propuseram uma técnica de dispersão pela rede (SPA) ativada pelos *churners* num determinado periodo de tempo de forma a simularem o efeito que existe na vida real de transmissão de influência para que os utilizadores na rede desses *churners* também abandonem o serviço.

Ao contrário de outros estudos, tais como Phadke & Uzunalioglu (2013) que abordam o *churn* no segmento pós-pago, este trabalho está direcionado para o segmento pré-pago, ou seja, não existem dados contratuais como nome, género, morada, idade, duração do contrato, entre outros. Os principais dados a serem utilizados são aqueles que derivam das CDR's (*Call Detail Record*).

Cada CDR contém, no mínimo, a identificação do emissor, do recetor, a duração da chamada, o tipo de contacto (voz, *SMS*, etc.).

Com estes dados das CDR's é possível construir a rede social da empresa de telecomunicações, em que os nós são os utilizadores, e as ligações entre eles são tão mais fortes quanto for o volume de contacto telefónico entre eles. Neste estudo, foi utilizado um mês de CDR's para construir o modelo e validá-lo.

Neste artigo, os autores pretendem criar um sistema de aviso prévio de forma a identificar antecipadamente os clientes mais propensos a sairem do serviço, ao contrário dos sistemas anteriores em que esses mesmos utilizadores eram encontrados através de alterações significativas na utilização do serviço, tais como, redução no volume de chamadas ou menor frequência de recarregamentos.

Características dos dados utilizados:

- Mais de 2 milhões de utilizadores e mais de 9 milhões de ligações.
- Chamadas com menos de 5 segundos foram retiradas.
- As ligações entre dois nós só são feitas quando existe reciprocidade de chamadas.
- O peso da ligação entre dois nós é calculado pela soma das chamadas entre nós.
- Foram retirados números de serviços, voice mail, entre outros.

Relativamente, ao peso da ligação entre dois nós, este é um assunto que tem uma resolução diferente conforme os autores, há até artigos, ver Motahari & Mengshoel (2012), em que o peso das ligações é categorizado e as diferentes categorias são avaliadas de acordo com o impacto que têm no padrão de chamadas dos clientes.

Voltando ao artigo em análise, baseado no número de contactos (amigos) que tinham abandonado no mês anterior, foi possível verificar que havia uma influência significativa do número de amigos que tinham abandonado o serviço e o abandono dos utilizadores a serem analisados. Para além disso, também verificaram que os utilizadores, cujos amigos que tinham deixado o serviço estavam ligados entre eles, tinham ainda maior probabilidade de também eles abandonarem o serviço.

Estes factos levaram os investigadores a testarem o método proposto de modelação de *churn* como um fenómeno de dispersão pela rede.

O método está dividido em quatro fases:

1. Ativação dos nós – durante cada iteração existe sempre um determinado número de nós, ao qual está associada uma determinada energia/influência. A quantidade de energia que é transferida de um nó X para um nó Y dependerá de um fator de dispersão (d) e de uma função de transferência de energia (F).
2. Fator de dispersão – este fator designado por (d) multiplicado pela energia do nó na iteração i determina quanta energia é transferida para os nós adjacentes. O restante fica no próprio nó. Um valor elevado de (d) significa que a energia transmitida fica em grande parte próxima da fonte, enquanto um menor valor permite uma maior dispersão da energia pela rede.

3. Distribuição de energia – a energia determinada pela multiplicação do fator (d) pela energia total do nó é distribuída para os nós vizinhos através de uma função linear, em que a energia transmitida para um determinado nó depende do peso da ligação entre esse nó e o nó transmissor, relativamente a todas as ligações entre o nó transmissor e os seus vizinhos.
4. Condição de término – quando após uma iteração não tiverem sido ativados novos nós e quando as alterações no valor da influência não ultrapassam a fronteira de precisão definida (E_T).

O resultado final será um nível de energia E por cada nó que caso esteja acima de um limite (T_c) o utilizador é classificado como *churner*, caso contrário será não-*churner*.

Principais conclusões:

- Através da análise da curva de *lift* e do gráfico do *hit rate* foi escolhido o parâmetro d=0,72 como fator de difusão que melhores resultados apresentava.
- Quando comparado o método apresentado com os métodos tradicionais de utilização de características dos utilizadores em forma de tabela, mesmo utilizando vários atributos por utilizador, não só de utilização do serviço, como de conectividade (1º grau) e inter-conectividade (2º grau), o SPA apresenta os melhores resultados. Apesar de haver uma melhoria significativa quandos se incluem os atributos de inter-conectividade.
- Este facto permite aos autores afirmarem que a decisão de *churn* não está apenas dependente das relações dos utilizadores com *churners*, mas sim na estrutura da rede circundante. Provam também que é essencial utilizar os dados da rede em detrimento de dados de utilização e conectividade pois estes adequam-se à forma como o fenómeno do *churn* ocorre.

O artigo de Breu, Guggenbichler, & Wollmann (2008) procura uma abordagem diferente utilizando uma construção do modelo em dois passos, uma fase de *clustering* e uma fase de classificação.

Numa primeira fase, foi necessário definir o que é o evento de *churn*. Eles consideram que não se deve estabelecer um mesmo período de tempo de inatividade para todos os tipos de clientes. Por exemplo, para pessoas que utilizam o telemóvel todos os dias, se esse cliente estiver uma semana sem usar o telemóvel, é provável que nao o volte a usar (*churn*). No entanto, caso seja um cliente que utilize menos frequentemente o serviço, então fará sentido ter um prazo para definição do evento de *churn* mais alargado.

Neste sentido os clientes foram divididos em quatro *clusters*, utilizando a metodologia da distância máxima, construidos baseados nas seguintes características:

1. Racio entre as chamadas com mais de um dia de distancia e o número total de chamadas.
2. Média da duração do período entre duas chamadas consecutivas.
3. A última data no período de observação que o cliente fez uma chamada.
4. A primeira data no período de observação que o cliente fez uma chamada.
5. O período de tempo em que cada cliente esteve ativo no período de observação.
6. A distância máxima entre duas chamadas de um cliente no período de observação.
7. Número de dias que um cliente fez ou recebeu uma chamada
8. Número total de chamadas recebidas de cada cliente
9. Número total de chamadas feitas de cada cliente
10. Valor total que foi cobrado ao cliente durante o periodo de observação.
11. Duração total das chamadas recebidas
12. Duração total das chamadas feitas

A partir daí são definidos períodos de observação, retenção e previsão para cada *cluster* individual.

Na segunda fase da classificação, o período de observação para cada *cluster* é dividido em dois sub-períodos e são consideradas para classificação quatro variáveis referentes ao primeiro sub-periodo (minutos de utilização de chamadas recebidas e feitas, frequencia de chamadas recebidas e feitas) mais quatro variáveis que representam a variação dessas quatro variáveis entre os dois períodos de tempo e finalmente uma variável binária que determina o estado do cliente (*churn*/não-*churn*) no período de previsão, se tiver ou não feito chamadas nesse período.

Depois dos atributos definidos são utilizados diferentes algoritmos de classificação para cada um dos *clusters* definidos anteriormente. Os algoritmos testados foram três variações de árvores de decisão (CART, CHAID e C5.0) e redes neuronais.

Em conclusão, utilizando uma medida de ganho para avaliar a performance dos algoritmos foi determinado que as árvores de decisão eram mais adequadas relativamente às redes neuronais.

O artigo escrito por Richter, Yom-Tov, & Slonim (2010) aborda uma metodologia diferente para deteção do *churn* baseada na dinâmica dos grupos encontrados na rede das chamadas dos clientes.

As metodologias tradicionais involvem de alguma forma a identificação dos *churners* num determinado periodo de tempo passado ou futuro, de forma a ser feita a previsão para um período mais à frente. O método proposto neste artigo não utiliza essa informação. Neste artigo, os autores procuram identificar os grupos de clientes que estão mais sujeitos a tornarem-se *churners*, antes de qualquer dos elementos do grupo ter abandonado o serviço.

Uma das premissas iniciais deste modelo é que a influência social na decisão de abandonar o serviço é muito mais dominante em grupos sociais mais unidos. Um dos motivos é a rapidez com que a informação circula, seja essa informação positiva ou negativa. Em segundo lugar, porque os membros de grupos sociais densos tendem a fazer mais chamadas entre si, e de forma a poupar dinheiro nas chamadas e outros serviços entre eles, têm tendência a manter-se no mesmo operador.

Finalmente, porque estes grupos pequenos e bastante unidos têm líderes dominantes que afetam as preferencias do grupo em geral e na decisão de deixar um operador em particular.

Portanto, estes investigadores procuram encontram pequenos grupos com um risco elevado de *churn*, antes de qualquer um dos membros ter abandonado o serviço.

Esta metodologia denominada por *group-first* segue os seguintes passos:

1. Treino do modelo – identificação dos grupos sociais e utilizá-los para construir um modelo de previsão.
    a. Quantificação das ligações sociais – a quantificação da ligação entre dois nós é muito importante já que pretende representar com precisão a verdadeira relação social entre duas pessoas. Para esta quantificação, neste artigo, é utilizada uma formula baseada no conceito da informação mútua. Mais especificamente, só são consideradas as 100 chamadas mais recentes no período de observação. Depois, é feita uma matriz de semelhança entre dois nós, de forma a verificar se essas duas pessoas ligam para as mesmas pessoas ou não. A quantificação é feita entre 0 e 1.
    b. Manter as ligações mais fortes – como o objetivo é encontrar grupos sociais unidos, é introduzido um parâmetro (p) que determina a percentagem das ligações mais fortes que são mantidas, sendo as restantes ligações ignoradas.
    c. Partição da rede – neste fase dois parâmetros novos (m) e (M) determinam que *clusters* com número de membros abaixo de (m) são eliminados e *clusters* com número acima de (M) são divididos em grupos mais pequenos e mais fortes.
    d. Ligar nós sem *cluster* – depois de no ponto b. terem sido eliminadas as ligações mais fracas e consequentemente terem sido desconsiderados vários nós. Nesta fase, depois dos *clusters* já estarem bem definidos, volta-se a considerar esses nós eliminados e se tiverem um grau de conetividade elevado em termos de número e duração de chamadas com os membros de algum dos *clusters*, esse nó é acrescentado ao respetivo, desde que não ultrapasse (M).

e. Análisar influência social – nesta fase é necessário analisar cada *cluster* individualmente de forma a determinar o peso social de cada membro no *cluster*. Para isso é usado um procedimento baseado em cadeias de markov.
f. Atributos-chave dos *clusters* – é feita a classificação dos *clusters* baseada em atributos, tais como, número de membros, rácio entre número de membros do operador em causa e de outros operadores e medidas de utilização por parte do líder do *cluster* e dos membros. O *cluster* é então classificado como *churn* ou não-*churn* caso mais de um terço dos seus membros tenham abandonado o serviço.
g. Treino com árvore de decisão – utilizando os atributos-chave e a classificação dos *clusters* é feito um treino (supervisionado) utilizando árvores de decisão.
h. Função de *churn* individual – é atribuída uma probabilidade de *churn* individual a cada membro do grupo inversamente proporcional ao ranking (ver e.) dentro do grupo.

2. Teste do modelo – na fase de teste do modelo o objetivo principal é prever *churn* de um grupo. O resultado é obtido através da árvore de decisão treinada que terá como output um parâmetro de puridade média dos nós multiplicado pela classe atribuida (ex: 1/-1). Para a previsão de *churn* individual é aplicada a função de *churn* individual multiplicada pelo resultado do *cluster* em que o cliente está inserido.

O objetivo era então prever os utilizadores que iriam abandonar o serviço nos 14 dias seguintes aos dados utilizados para treino. Nas experiências feitas foram apenas utilizados 3 dias de dados para treino e feito testes em 7 períodos de tempo semelhantes, sendo a performance avaliada através das curvas de *lift*.

As principais conclusões deste artigo são as seguintes:

- A probabilidade de *churn* para utilizadores que pertencem a grupos pequenos (<20) é 2,7 vezes maior do que se pertencessem a grupos grandes.
- A probabilidade de *churn* de um grupo em que o líder (caso exista) pertença a um outro operador é de 19,4 vezes mais do que se o líder for da operadora em análise. E os membros desses grupos têm 1,6 vezes mais probabilidade de deixarem o serviço.
- A probabilidade de *churn* está muito dependente da força da influência social de um membro no seu *cluster*. Sendo então que a pessoa mais provável a fazer *churn* é o líder (3 vezes mais). Em grupos em que duas ou mais pessoas tenham feito *churn* a probabilidade de ter sido o líder é cerca de 12 vezes mais do que aquela que seria de esperar se fosse escolhido um membro aleatoriamente.
- A análise às árvores de decisão revelou que grupos com mais de 4 membros são menos prováveis de se tornarem *churners*. Para além disso, os grupos em maior risco de *churn* são aqueles em que o lider fez e recebeu poucas chamadas, existem poucas interações sem ser chamadas (*SMS*, por exemplo) e em que o líder, como já tinha escrito, não pertence ao operador em análise.

Este artigo apresenta também como possibilidade de trabalhos futuros a integração desta metodologia *group-first* com os métodos baseados unicamente em *churn* dos individuos. Isto porque, diferentes tipos de *churn* são identificados por abordagens diferentes, o que indicia que as abordagens híbridas terão um maior sucesso na tarefa de previsão de *churn*.

O artigo apresentado por Huang & Kechadi (2013) apresenta uma técnica híbrida para abordagem à previsão de *churn*. Em primeiro lugar, é aplicado um algoritmo de *clustering* aos dados de treino, de seguida é utilizado um método de indução de regras para generalizar a classificação em cada *cluster*.

O algoritmo de *clustering* utilizado é uma variante do k-means, que é ponderado de forma a tornar mais evidente a relação entre os atributos e a classe em que está inserido o utilizador (*churn*/não-*churn*). Para além disso, a previsão feita sobre os dados de teste é conseguida através da utilização do classificador mais adequado às características de um cliente, de acordo com o *cluster* em que está inserido.

Este modelo tem quatro fases principais:

1. Discretização dos dados: todos os atributos contínuos são discretizados utilizando um algoritmo que tem a classe em consideração.
2. *Clustering* ponderado: os dados de treino são agrupados em *clusters* e é utilizado o método "path analysis" para calcular os pesos dos atributos.
3. Extração de regras: é aplicado um método de aprendizagem de regras para cada *cluster*, de forma a cada cada *cluster* tenha um classificador.
4. Previsão: as instâncias de teste são classificadas de acordo com o classificador encontrado para o *cluster* mais próximo dessa instância, calculado através de uma medida de distancia ou similaridade.

As experiências feitas neste estudo compararam este modelo híbrido com os classificadores mais comuns (árvores de decisão, SVM, regressão logística, entre outros) e os resultados foram bastante apreciáveis em termos da comparação das curvas ROC e da respetiva AUC.

Os autores propõem também algumas sugestões para trabalhos futuros como a remoção prévia de outliers e dados redundantes, a utilização e comparação de diferentes algoritmos de *clustering* na primeira fase e a substituição do algoritmo de aprendizagem de regras na segunda fase por outros modelos preditivos como árvores de decisão, redes neuronais, entre outros.

O artigo dos investigadores da universidade de Leiden, ver Kusuma & Radosavljevik (2013), tem como objetivo fazer a comparação entre vários modelos de previsão de *churn*. Para além disso, apresentam uma extensão do modelo apresentado em Dasgupta & Singh (2008), no qual utilizam a probabilidade indicada por um classificador sobre os atributos de um cliente como valor inicial para os nós no processo de difusão de influência pela rede. Por outro lado, na parte da classificação dos clientes, utilizam vários atributos derivados da rede em que estão inseridos.

Relativamente a estes últimos dados foi feita uma divisão em duas categorias. Os dados consideram uma rede não-direcionada.

Dados de conetividade:

1. Contagem de ligações de entrada/saída do nó.
2. Soma e média do peso da ligação para entradas/saídas do nó.
3. Contagem e média de chamadas, *SMS* e ambos de e para os vizinhos.
4. Total e média do peso das ligações.
5. Frequência e média da interação com vizinhos para chamadas e *SMS* em separado.
6. Contagem de ligações no primeiro, segundo e terceiro grau.

Dados de conetividade com *churners*:

1. Os mesmos que nos dados de conetividade mas considerando apenas nós *churners*.
2. Rácios entre os mesmos indicadores dos dados de conetividade dividindo os valores encontrados considerando apenas ligações para nós *churners* e aqueles relativamente ao total de nós que um determinado utilizador está ligado.

A outra novidade introduzida neste artigo está relacionada com um modelo de propagação avançado. No modelo de propagação simples, o valor inicial de energia/influência para os nós de não-*churners* é igual a 0.

No modelo avançado, a energia inicial é definida pelo resultado da classificação aplicada ao conjunto de atributos do utilizador.

Para além disso, a propagação de influência foi feita através de um gráfico de rede direcionado, em que a energia é apenas transmitida através das ligações do nó para os seus vizinhos e de um não-direcionado em que ambos os tipos de ligações do nó foram utilizados. Para uma análise mais detalhada dos efeitos da direcionalidade, ver Haenlein (2013).

Os dados utilizados correspondem a 1 mês de chamadas (fevereiro) e o objetivo é fazer a previsão do *churn* durante o mês de junho baseado nos dados da rede em fevereiro, março e abril. Para o modelo de propagação, são identificados como nós *churners* (2 meses consecutivos de inatividade) os utilizadores que abandonaram o serviço durante os meses de fevereiro, março e abril.

O cálculo do peso da ligação entre dois nós é calculado em segundos. Para isso, as *SMS* são convertidas em metade de uma chamada. Para além disso, são ponderadas com o dobro do peso todas as chamadas e *SMS* que sejam feitas à semana fora do periodo de trabalho e aos fins-de-semana. Uma outra ponderação que é utilizada de forma a conferir maior correspondencia com a realidade ao peso das ligações, é a ponderação da antiguidade das interações, isto é, existe um fator de redução de valor para as interações mais antigas de forma a favorecer as relações mais recentes.

Nas experiências realizadas foram testados 7 modelos.

Modelo 1: modelo simples com dados de perfil do cliente classificado com regressão logistica e CHAID (árvore de decisão)

Modelo 2: modelo que utiliza apenas os dados de conetividade derivados da rede.

Modelo 3: Modelo 1 + Modelo 2

Modelo 4: modelo de propagação simples numa rede direcionada

Modelo 5: modelo de propagação avançado numa rede direcionada

Modelo 6: modelo de propagação simples numa rede não-direcionada

Modelo 7: modelo de propagação avançado numa rede não-direcionada

As conclusões deste estudo são bastante surpreendentes, já que no geral, o modelo 1 é aquele que apresenta melhores resultados de ganho e de *lift*.

Para além disso, a diferença entre os modelos 1 e 3 é residual, ou seja, o facto de serem acrescentados aos dados de perfil, os dados de conetividade, não melhora a capacidade preditiva do modelo.

Relativamente aos modelos de propagação de influência, apresentam resultados bastante piores que o modelo 1 e o 3, no entanto, fica demonstrado que a utilização do resultado do modelo 1 (com árvores de decisão) como energia inicial dos nós não-*churners*, melhora significativamente a performance relativamente aos modelos simples. Uma outra conclusão um pouco surpreendente é que os melhores resultados nos modelos de propagação são obtidos considerando a rede não-direcionada.

3. Estudo de caso

Neste capítulo serão expostos os dados que serviram de matéria-prima para este trabalho, assim como, a análise detalhada sobre a definição do evento de *churn*.

3.1. Dados em estudo

Os dados recebidos para efetuar esta dissertação fornecidos pela operadora compreendem uma duração de 12 meses. Estes dados anonimizados estão divididos em tabelas que passarão a ser explicadas de seguida de forma a melhorar o entendimento futuro devido a referências às mesmas.

3.1.1. CDR *Raw*

Este é o principal conjunto de tabelas. Os dados foram fornecidos à semana. Cada entrada nesta tabela representa uma interação que involve pelo menos um cliente da operadora. Esta interação pode ser uma chamada, uma mensagem ou uma consulta de dados na *internet*. Para cada interação temos a data em que ocorreu, o tipo de chamada (*call type*) e o código do tarifário (*rate plan*) na altura em que se deu a interação.

Através desta tabela podemos obter dados importantes, tais como, distribuição da duração das chamadas.

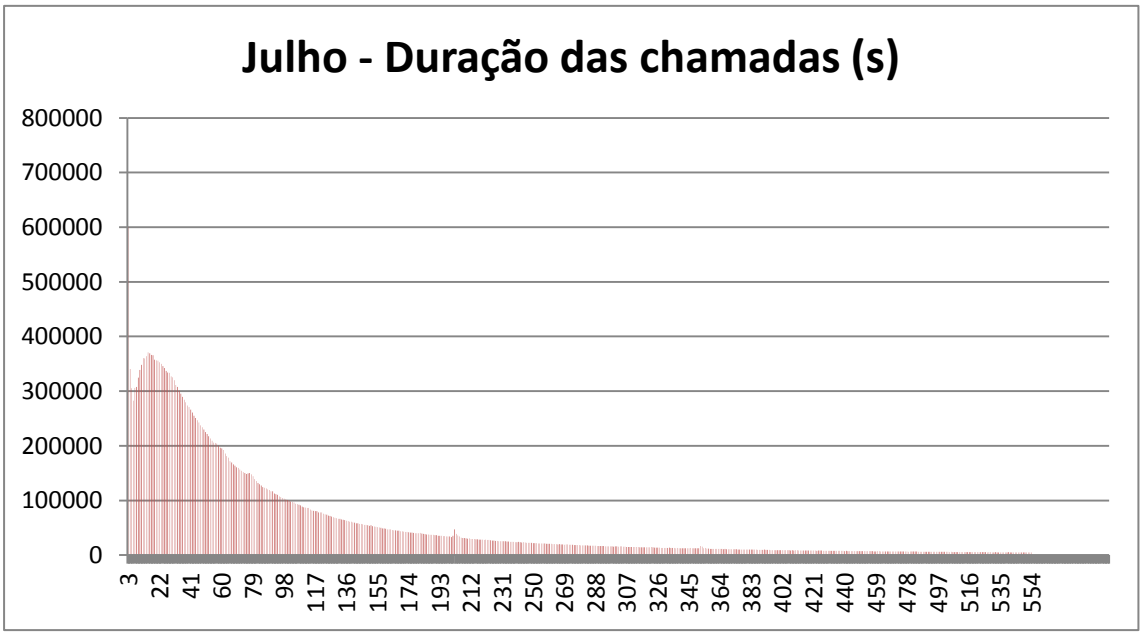

Fig. 1 - Julho - Duração das chamadas (s)

Por esta distribuição podemos inferir que as chamadas de duração menor ou igual a 3 segundos podem ser consideradas outliers já que a sua frequência difere bastante da distribuição apresentada. Pela análise do gráfico podemos também concluir que existe um grande volume de chamadas inferior a 1 minuto. Mais especificamente, aproximadamente 52% do número total de chamadas.

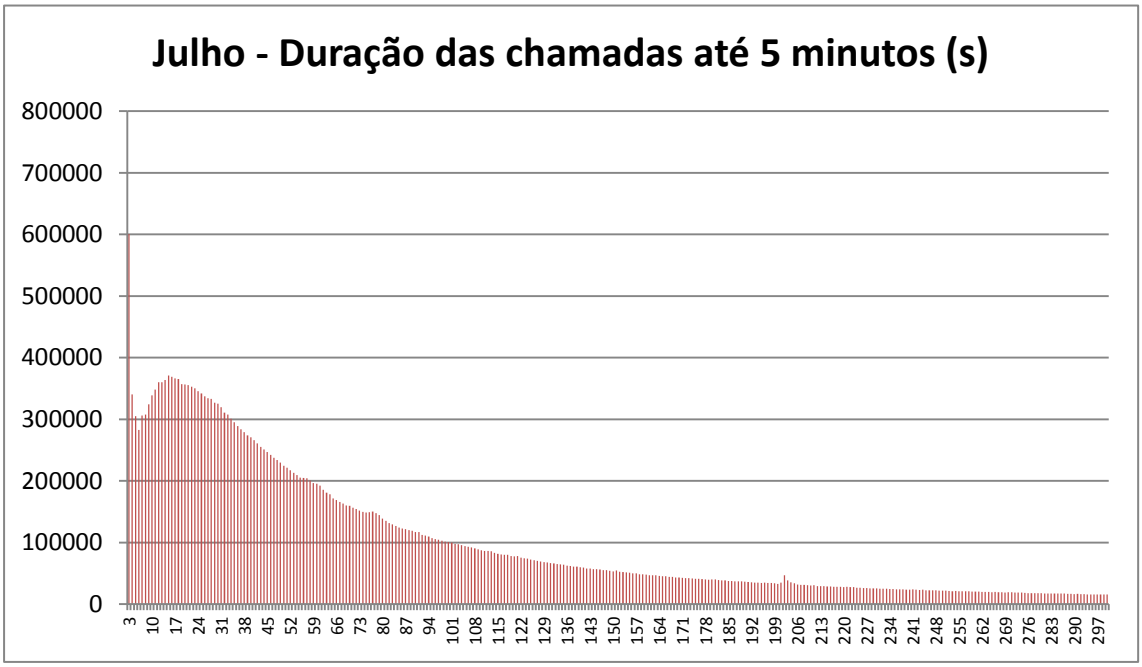

Fig. 2 - Julho - Duração das chamadas até 5 minutos (s)

Pela tabela cdr *raw* podem também ser feitas estatísticas relativamente ao período do dia e ao dia da semana nos quais existem mais eventos (chamadas e mensagens).

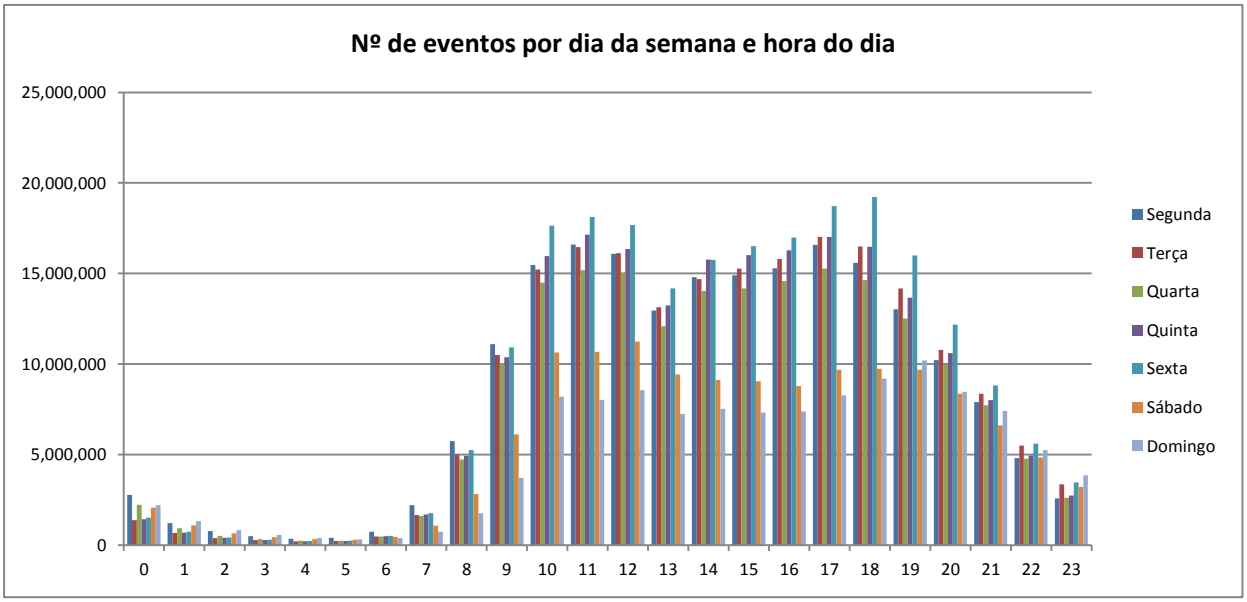

Fig. 3 - Número de eventos por dia da semana e hora do dia

Nesta análise feita para um período de 6 meses, pela análise do gráfico, podemos verificar um período entre as 10h e as 18/19h em que se verifica um período de elevada atividade relativamente contínuo e o restante período com menor atividade. Em termos de dias da semana, verifica-se um claro decréscimo na atividade ao fim-de-semana relativamente à semana. Para finalizar, pode-se concluir que a sexta-feira é o dia em que mais atividade existe na operadora para quase todas as horas do dia.

3.1.2. *Subscribers*

Neste conjunto de tabelas temos os dados referentes aos clientes (*subscribers*). Os dados foram fornecidos ao mês. Cada entrada nesta tabela representa um cliente. Para cada cliente temos disponíveis vários dados, entre eles, data desde que entraram para a operadora, origem da aquisição, tarifário, data da última recarga (*top up*) e data da última comunicação.

Neste estudo vai ser focada a base de clientes pré-pago para os quais existe pouca informação na base de clientes relativamente a atributos como género, idade, localidade e telemóvel utilizado. Como tal, estes atributos apesar de terem sido fornecidos não serão utilizados.

3.1.3. *Top up*

Neste conjunto de tabelas temos os dados referentes às recargas (*top up*) dos clientes. Os dados foram fornecidos à semana. Cada entrada nesta tabela representa uma recarga do cliente. Nesta tabela temos a informação do cliente, valor da recarga, data, meio de recarga e saldo antes da recarga.

3.1.4. *Call type*

Neste conjunto de tabelas temos os dados referentes aos tipos de chamadas. Os dados foram fornecidos mensalmente. Estes tipos de chamadas descrevem de forma genérica de que fonte foi emitada a chamada e de que fonte foi recebida.

Desta tabela foram retirados os tipos de chamadas que correspondem a interações que não indicam utilização do telemóvel, por exemplo, chamadas que vão para voice mail ou mensagens de serviços.

3.2. Análise períodos de dormência (*churn*)

Esta análise tem por objetivo definir o critério a considerar para período de dormência ou *churn*, isto é, a partir de quantos dias sem atividade é que será considerado que um cliente abandonou a rede ou passou a ter uma atividade muito reduzida. Em primeiro lugar, foi feita uma análise genérica a uma amostra da base de clientes segundo os critérios enunciados. Em segundo lugar, foi feita uma análise direcionada ao segmento de clientes, da qual será alvo esta dissertação, mais uma vez recorrendo a uma amostragem da base total.

3.2.1 Análise de dormência geral

Como critério para definição dos clientes a utilizar nesta análise, como o mínimo de período de dormência considerado foi de 3 semanas, foram retirados todos os clientes que tiveram atividade em pelo menos 26 semanas não consecutivas.

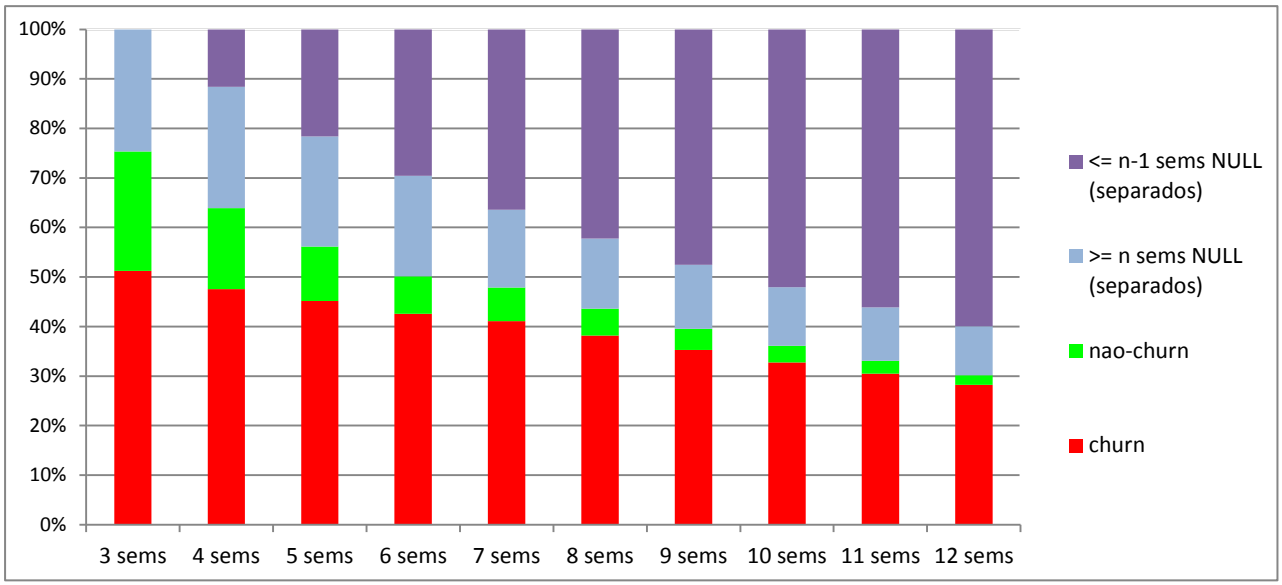

Fig. 4 - Análise da dormência geral

A legenda <=n-1 sems NULL (separados) significa, por exemplo no caso das 5 semanas, que existem 21% de clientes que não chegam a ter 5 semanas não-consecutivas sem atividade.

A legenda >= n sems NULL (separados) significa, por exemplo no caso das 7 semanas, que existem cerca de 13% de clientes que têm mais do que 7 semanas de dormência mas que essa dormência não é consecutiva.

A legenda *churn* e não-*churn*, corresponde à percentagem de clientes que cumprem o requisito de *x* semanas consecutivas de dormência. Sendo que, no caso de *churn* significa que são clientes que não voltaram a ter atividade após a dormência. No caso de não-*churn* significa que após o período de dormência voltaram a ter atividade em pelo menos uma semana.

Em termos destas duas variáveis é apresentado o gráfico de seguida considerando como base de clientes apenas aqueles que apresentam *x* semanas consecutivas de dormência.

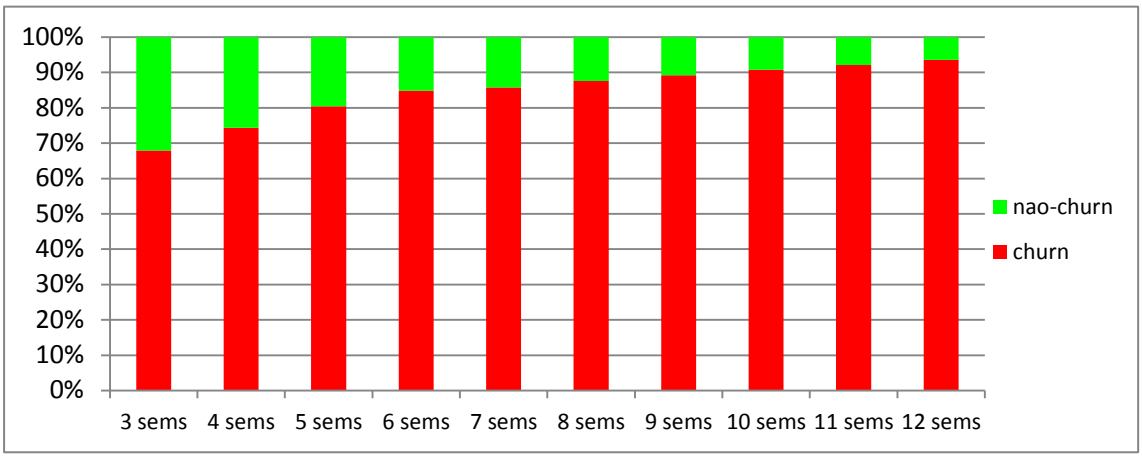

Fig. 5 - Análise da dormência geral (detalhe)

O facto a destacar é que podemos verificar que cerca de 7% dos clientes que fica 12 semanas (3 meses) sem atividade ainda regressa, pelo que, será à partida difícil definir um período que reduza essa percentagem. O âmbito desta tese será prever com antecipação a entrada em *churn* de um determinado cliente, para que a operadora possa agir a tempo de "segurar" esse cliente. Portanto, o período de dormência a utilizar nunca poderá ser tão grande como 12 semanas sem atividade já que isso implicaria uma redução drástica na probabilidade de conseguir manter o cliente tendo em conta o facto de ter estado tanto tempo sem comunicar.

Por outro lado, um período tão curto como 3 semanas não é significativo já que cerca de um terço volta após esse período, não podendo ser considerado como abandono.

O período ideal a definir será entre as 6 e as 7 semanas já que a diferença no retorno para as semanas seguintes não é muito significativa e já representa um período em que o cliente poderá estar perto de abandonar o serviço.

3.2.2 Análise de dormência – segmento particular

Foi definido como prioridade pela operadora a análise a um segmento específico de clientes.

Nesse sentido foi feita uma análise mais sensível aos dados destes clientes no sentido de encontrar o período de dormência ideal.

Nesta análise, foi considerado como atividade quando o cliente efetua uma chamada ou envia uma mensagem. Isto é, não foi considerado receber chamadas ou mensagens como atividade. Este tipo de análise encontra fundamento no facto do comportamento tipo de um cliente que abandona um serviço. Em primeiro lugar esse cliente deixará de usar o serviço ativamente como emissor, mas manterá o serviço ativo como recetor até efetuar a mudança para outra operadora. Como o objetivo é identificar atempadamente os casos de possível abandono, foi considerado que a partir do momento em que deixa de ativamente usar o serviço, mesmo estando a receber chamadas ou mensagens, já é considerado como dormência.

Começamos, então por verificar durante um período de 3 meses qual tinha sido a evolução das entradas em dormência para diferentes durações neste segmento.

Este gráfico demonstra a entrada nas dormência acumulada ao longo dos dias sobre o total da base no início do período (dia 0). No final podemos concluir, por exemplo, que cerca de 12% dos clientes no dia 0, tiveram um período de dormência de 20 dias durante os 90 dias seguintes. Podemos também notar que a diferença entre periodos de dormência diminui progressivamente sendo que se nota uma maior diferença entre os 20 e os 30 dias e quase não se nota diferença entre os 60 e os 70 dias.

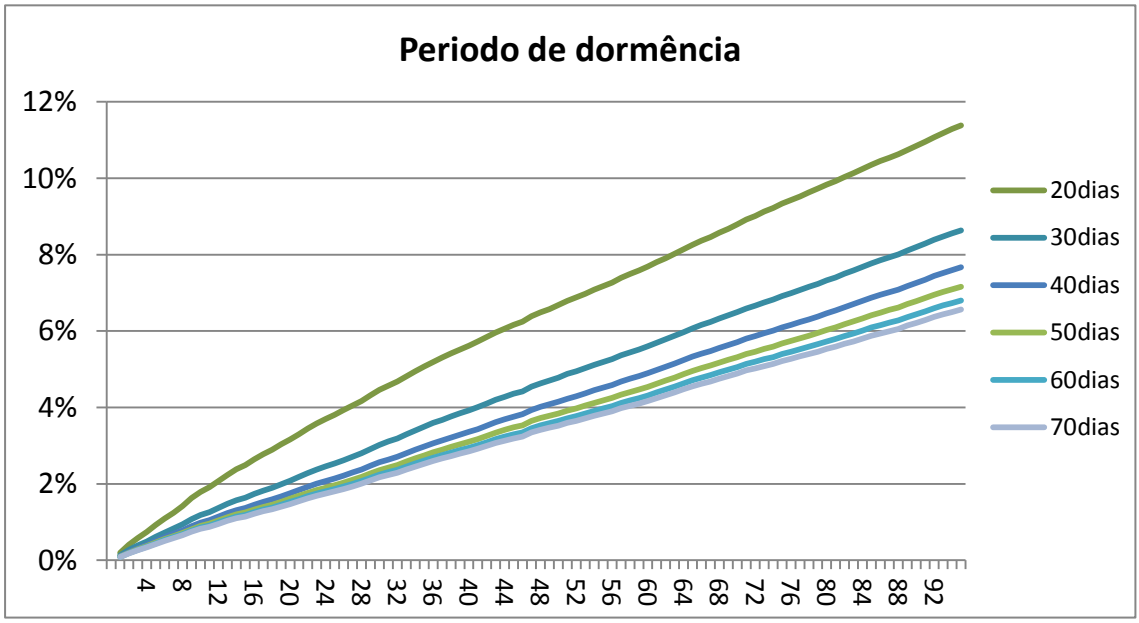

Fig. 6 - Análise dormência particular (Período de dormência)

Para complementar esta informação foi verificada também a percentagem de regresso à atividade destes mesmos clientes para os mesmos períodos de dormência.

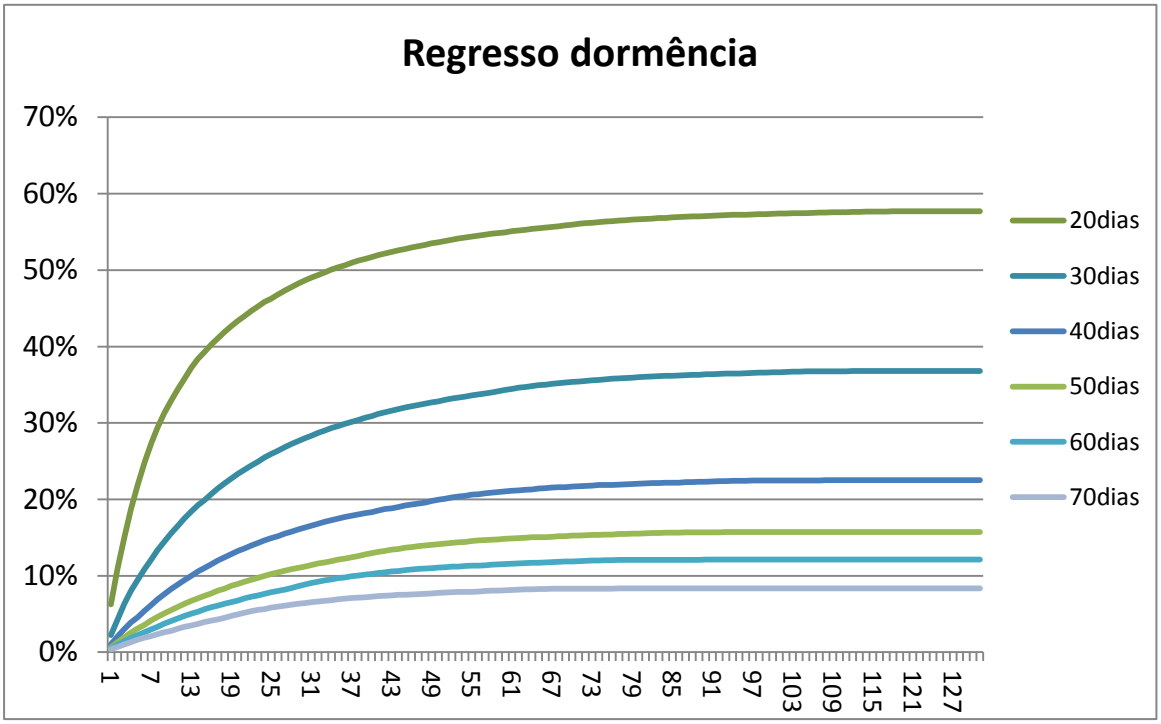

Fig. 7 - Análise dormência particular (Regresso de dormência)

Neste gráfico os dias no eixo das abcissas correspondem aos dias após ultrapassado o período de dormência em que o cliente voltou a manifestar atividade. Por exemplo, para os 20 dias, podemos verificar que cerca de 6% dos clientes voltaram no dia a seguir aos aos 20 dias de dormência.

Neste segmento podemos então verificar que não faz mesmo sentido considerar um período de dormência de 20 ou 30 dias já que as probabilidades de retorno são bastante elevadas. Mais uma vez, tal como na análise geral, em que o período ideal apontava para as 6 ou 7 semanas, neste caso para este segmento, podemos chegar a uma conclusão similar, já que a partir principalmente dos 50 dias, já não há grande alteração na taxa de retorno apesar de naturalmente continuar a diminuir.

No entanto, para validar esta hipótese podemos ainda analisar o perfil de recargas antes, durante e depois do período de dormência para verificar como voltam os clientes em termos de recargas após os períodos de dormência em estudo.

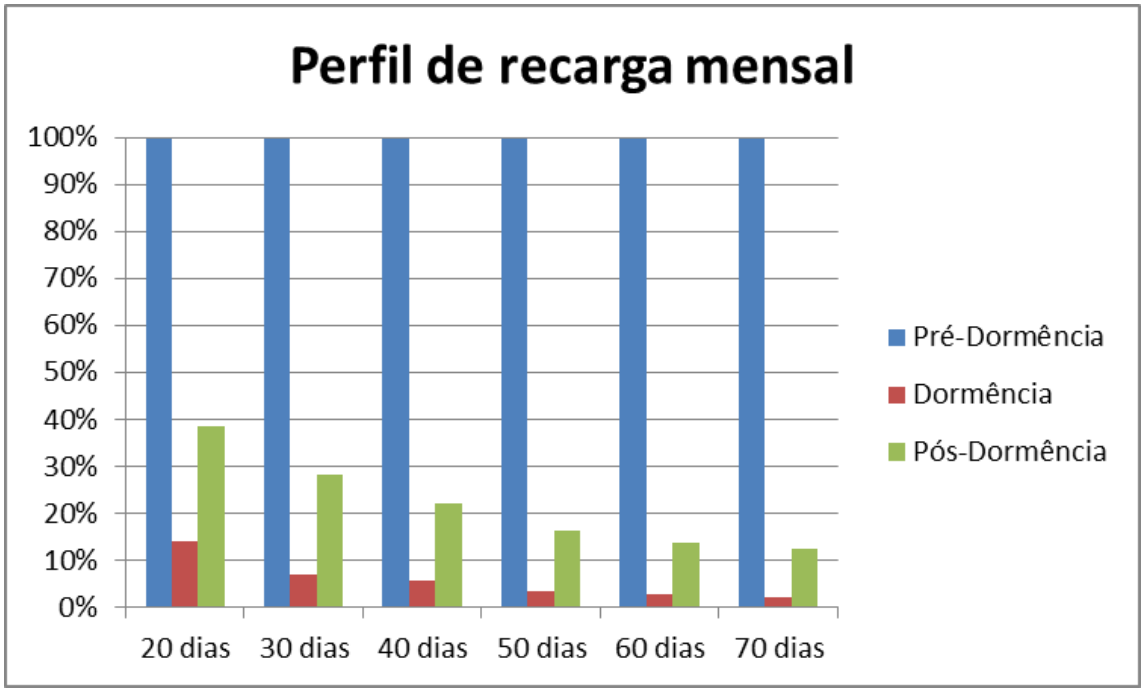

Fig. 8 - Análise dormência particular (Perfil de recarga mensal)

Por este gráfico podemos ver a percentagem relativamente à média de recargas mensais, no período pré-dormência, de cada cliente nos períodos indicados na legenda. Ou seja, como é possível constatar, no período de pré-dormência é de 100% para todas as classes.

Mais uma vez podemos confirmar que à medida que consideramos um período de dormência mais exigente os clientes que voltam têm um peso cada vez menor. Esta medida de perfil é fortemente influenciada pelos clientes que não voltam em cada uma das dormências e que consequentemente não fazem recargas. No entanto, esse é um espelho da realidade em que existe a perda real de volume de recargas. Mais uma vez, o periodo que deverá ser escolhido mantém entre os 40 e os 50 dias.

3.2.3. Conclusão - Análise de dormência

Pelas análises realizadas pode ser concluído que o período ideal para considerar como dormência está entre os 40 e os 50 dias de dormência em termos de atividade executada pelo cliente, no sentido de fazer chamadas ou enviar mensagens. Estes dois períodos serão tidos em conta na modelação que será feita durante a continuação desta dissertação.

No entanto, já que, em média, 83% dos clientes que estão dormentes 30 dias também estão dormentes 40 dias e 74% dos clientes que estão dormentes 30 dias também estão dormentes 50 dias, foi utilizado de forma a prevenir o *churn* de forma mais antecipada, por defeito, o critério de 30 dias de dormência.

4. Análise de Atributos

Neste capítulo terá lugar a descrição dos passos que levaram à construção do modelo para previsão de *churn*.

4.1. Atributos

Para este problema específico de *churn* na base de clientes pré-pagos, ao contrário do que sucede tradicionalmente na literatura referenciada, não serão utilizados atributos nominais. Existem dois motivos para isso. Primeiro, porque a maioria desses atributos prendem-se com características que não são facilmente obtidas devido à natureza destes clientes e portanto existem muitos valores em falta nas bases de dados das operadoras. Tal como mencionado anteriormente na secção 3.1.2, entre estes atributos estão o género, localidade e idade.

Em segundo lugar, porque para a utilização de algoritmos de classificação como a regressão logística e o perceptron multi-camada, o ideal é utilizar atributos contínuos já que ambos os métodos involvem uma ponderação associada a cada atributo, tendo como resultado final uma função que classificará baseado numa probabilidade da classe em causa. Devido a este facto, variáveis como o canal de venda, o tarifário específico, o fabricante do telemóvel ou o modelo do telemóvel do cliente, não foram utilizadas no modelo. Apesar de ser possível uma binarização destes atributos nominais, devido ao número de classes existentes em cada variável, o modelo ficaria demasiado pesado, o que poderia prejudica a sua performance.

Portanto, neste trabalho serão utilizados apenas atributos contínuos. Para a escolha do horizonte temporal a que os atributos se referem foi indicado pela operadora que a classificação de cliente é apenas utilizada para pessoas que utilizem o seu serviço há mais de 90 dias. Por isso, para os atributos abaixo referidos, foi sempre considerado a média mensal dos últimos 3 meses.

Estes atributos podem ser divididos em 5 categorias.

4.1.1. Atributos de perfil

Média de lucro – a média de lucro que a operadora tem com o cliente nos últimos meses.

Antiguidade – dias que passaram desde que o cliente usou o serviço da operadora pela primeira vez.

4.1.2. Atributos de atividade

Os atributos de atividade foram divididos em duas categorias: IN para chamadas/mensagens recebidas e OUT para chamadas/mensagens efetuadas. Para além disso, foi também dividida a origem/destino das chamadas e mensagens entre:
DENTRO: de/para dentro da operadora, CONC: de/para operadoras concorrentes e OUTROS: de/para outras operadoras (principalmente internacionais e rede fixa).

Segue agora em baixo uma tabela com os atributos considerados e a explicação.

		cont_voz	soma_voz	cont_sms	media_voz	per_cont_voz	per_soma_voz	per_cont_sms	per_cham_sms
IN	Dentro	X	X	X	X	X	X	X	
	Conc	X	X	X	X	X	X	X	
	Outros	X	X	X	X	X	X	X	
	Semana	X	X	X	X	X	X	X	X
	Fim-de-semana	X	X	X	X	X	X	X	
	Dia	X	X	X	X	X	X	X	
	Noite	X	X	X	X	X	X	X	
OUT	Dentro	X	X	X	X	X	X	X	
	Conc	X	X	X	X	X	X	X	
	Outros	X	X	X	X	X	X	X	
	Semana	X	X	X	X	X	X	X	X
	Fim-de-semana	X	X	X	X	X	X	X	
	Dia	X	X	X	X	X	X	X	
	Noite	X	X	X	X	X	X	X	

Tabela 1 – Atributos de atividade

Atributo	Definição
cont_voz	# chamadas
soma_voz	tempo da chamada
cont_sms	# mensagens
media_voz	duração média do tempo de chamada
per_cont_voz	% de # chamadas Dentro/Conc/Outros, Sem/Fds, Dia/Noite
per_soma_voz	% de tempo chamada entre Dentro/Conc/Outros, Sem/Fds, Dia/Noite
per_cont_sms	% de # mensagens Dentro/Conc/Outros, Sem/Fds, Dia/Noite
per_cham_sms	% de # chamadas relativas a # mensagens

Tabela 2 - Atributos de atividade (definição)

4.1.3. Atributos de rede

Os atributos de rede são atributos que representam o contexto em termos da rede de contactos do cliente em questão.

		cont_amigos	cont_amigos_prox	cont_amigos_fixos	per_amigos	per_amigos_prox	per_amigos_fixos	per_churn_15dias	per_churn_30dias
IN	Dentro	X	X		X	X			
	Conc	X	X		X	X			
	Outros	X	X		X	X			
OUT	Dentro	X	X	X	X	X	X	X	X
	Conc	X	X	X	X	X	X		
	Outros	X	X	X	X	X	X		

Tabela 3 - Atributos de rede

Atributo	Definição
cont_amigos	# contactos (critério >= 3 mensagens/chamadas mês)
cont_amigos_prox	# contactos (critério >= 10 mensagens/chamadas mês)
cont_amigos_fixos	# contactos (critério pelo menos 1 mensagem/chamada mês em 3 meses seguidos)
per_amigos	% contactos entre Dentro/Conc/Outros segundo critério em cima
per_amigos_prox	
per_amigos_fixos	
per_churn_15dias	% contactos que entraram em dormência entre 15 a 30 dias do dia em análise
per_churn_30dias	% contactos que entraram em dormência entre 30 a 60 dias do dia em análise

Tabela 4 - Atributos de rede (definição)

O critério escolhido para a escolha de um amigo normal foi o de 3 contactos num mês já que, em média, mais de 60% das relações num mês são de 1 e 2 interações mensais. Nessas interações o mais provável será tratarem-se de relações pontuais que não têm influência na opção ou não do cliente fazer *churn*.

O critério para a escolha do amigo próximo de 10 contactos teve a ver com a média de interações de cada cliente com os seus amigos. Em média cada relação de um cliente era de 14 interações, no entanto, este valor estava muito enviezado devido à presença de vários outliers e valores extremos à direita da média. Após removidos esses casos, a média ficou nas 10 interações.

4.1.4. Atributos de recargas

Os atributos de recargas têm a ver com o perfil de recarregamentos de um cliente, como pode ser consultado na tabela em baixo.

Atributo	Definição
cont_topup	# recargas
dias_ult_topup	# dias desde o último topup
media_dias_entre_topup	média de dias entre topup's
dias_atraso_topup	diferença de dias relativamente à media_dias_entre_topup
media_saldo_topup	média de saldo quando fez topup
media_sem_cont_topup	média semanal de # topup
media_topup	média de valor de topup
media_sem_topup	média de valor semanal de topup

Tabela 5 - Atributos de recargas

4.1.5. Atributos de variação

Os atributos de variação têm o objetivo de perceber a variação em algumas das variáveis definidas anteriormente. A variação é medida entre a média dos três últimos meses e o último mês. As variáveis que foram alvo de atributos de variação podem ser consultados na tabela em baixo.

		Atributos de Atividade						
		cont_voz	soma_voz	cont_sms	media_voz	per_cont_voz	per_soma_voz	per_cont_sms
IN	Dentro	X	X	X	X	X	X	X
	Conc	X	X	X	X	X	X	X
	Outros	X	X	X	X	X	X	X
OUT	Dentro	X	X	X	X	X	X	X
	Conc	X	X	X	X	X	X	X
	Outros	X	X	X	X	X	X	X

Tabela 6 - Atributos de variação (atividade)

		Atributos de Rede					
		cont_amigos	cont_amigos_prox	per_amigos	per_amigos_prox	cont_amigos_fixos	per_amigos_fixos
IN	Dentro	X	X	X	X	X	X
	Conc	X	X	X	X	X	X
	Outros	X	X	X	X	X	X
OUT	Dentro	X	X	X	X	X	X
	Conc	X	X	X	X	X	X
	Outros	X	X	X	X	X	X

Tabela 7 - Atributos de variação (rede)

4.2. Seleção de Atributos

No total foram calculados 226 atributos divididos em diferentes áreas conforme visto anteriormente. Em baixo pode ser visto o resumo:

Atributos gerais: 2
Atributos de atividade: 100
Atributos de rede: 32
Atributos de recargas: 8
Atributos de variação: 84

Devido ao elevado número de atributos foram utilizadas técnicas de seleção de atributos de forma a tornar mais eficiente o treino e teste do modelo, assim como, melhorar a compreensão sobre as variáveis que mais influência têm no *churn*.

Utilizando o programa WEKA foram testados três métodos de seleção de atributos. Para a escolha dos atributos foram utilizados os dados referentes aos clientes de Novembro. Os resultados são os mesmos quer consideremos os *churners* como dormência a 30, 40 ou 50 dias.

4.2.1. CfsSubsetEval

O método CfsSubsetEval avalia um conjunto de atributos e verifica quão relevante é para a previsão considerando a capacidade individual de cada um dos atributos assim como o grau de redundância entre eles. (Hall, 1999)

Através deste método foi possível chegar a um conjunto de 30 atributos considerados os mais relevantes, pois a partir da introdução do 31º atributo já não houve melhoria no resultado do algoritmo.

Destaque para as variáveis **antiguidade**, **variação_cont_voz_dentro_out, cont_voz_semana_out e per_amigos_fixos_dentro_out**. Estas são as que representam uma maior subida percentual no resultado do algoritmo.

4.2.2. Gain ratio

Este método avalia a relevância de um atributo através da medida de rácio de ganho de informação.

Os cinco atributos mais valiosos de acordo com este método são: **antiguidade**, **variação_cont_voz_dentro_out**, **cont_voz_semana_out**, **cont_voz_dia_out** e **per_amigos_fixos_dentro_out**.

4.2.3. Correlation Attribute Eval

Este método avalia os atributos pela correlação (método de Pearson) entre um atributo e a classe.

Os atributos mais correlacionados com a classe são então: **antiguidade**, **media_saldo_topup**, **cont_amigos_prox_dentro_out**, **per_churn_30dias** e **media_topup**.

4.2.4. Resultados

De seguida é possível verificar os diferentes resultados, em termos de *lift* do *top* 10%, quando se faz variar o número de atributos indicados por cada um dos algoritmos referidos em cima. Para estes testes foi utilizada a classificação da base de clientes de novembro, treinada com a base de outubro. Os algoritmos de classificação utilizados foram a Regressão Logística e o Perceptron Multi-Camada.

LIFT	Nº Atributos	Regressão Logistica	Perceptron Multi-Camada
Todos	226	3.16	3.38
Gain Ratio	top 50	2.98	2.90
	top 40	2.84	2.83
	top 30	2.78	2.77
	top 20	2.72	2.71
	top 10	2.60	2.54
CFS Subset	30	2.90	2.90
Correlation	top 40	2.98	3.11
	top 30	3.04	3.22
	top 20	3.12	3.35
	top 10	2.88	3.00

Tabela 8 - Comparação seleção de atributos

Da análise da tabela, pode-se concluir, em primeiro lugar, que o Perceptron Multi-Camada obtém consistentemente melhores resultados que a Regressão Logística. Em segundo lugar, apesar dos melhores resultados serem atingidos quando são utilizados todos os atributos, em termos de eficiência do treino do modelo, não é a melhor alternativa. Logicamente, quando são utilizados mais atributos o treino do modelo é bastantes vezes mais lento. Esse trade-off entre eficiência do algoritmo e o *lift* resultante, leva à escolha dos 20 atributos mais correlacionados com a classe, para utilizar no modelo de previsão.

4.3. Análise atributos para modelação

Nesta secção procura-se entender melhor a relação entre os atributos que serão utilizados para a modelação e a classe correspondente (*churn*/não-*churn*). Os dados utilizados para esta análise corresponderam à base de clientes de novembro. Nesta base, existem cerca de 3,4% de pessoas que entram em dormência durante os 30 dias seguintes.

4.3.1. Número de anos que o cliente está na rede

Este atributo é claramente um dos mais importante já que está presente no *top* de todos os algoritmos de seleção de atributos.

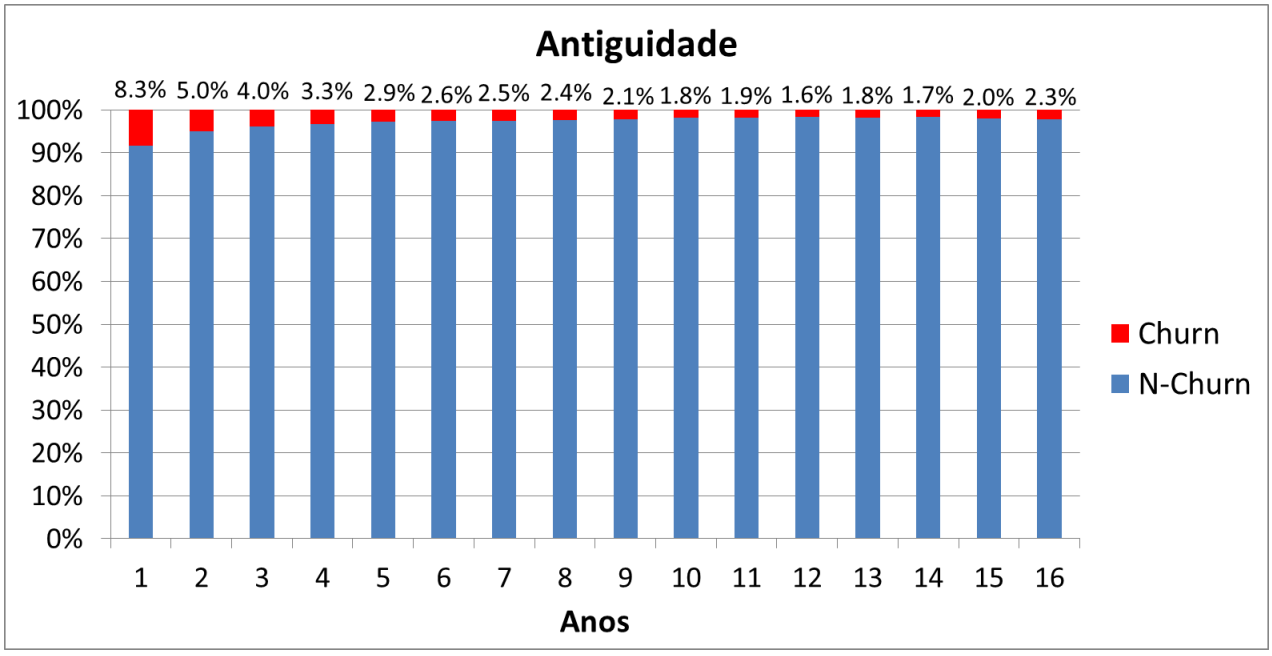

Fig. 9 - Antiguidade

Como podemos verificar pela análise do gráfico, clientes que se encontram na faixa dos menores ou igual a 1 ano, têm mais de duas vezes mais probabilidade de entrarem em dormência do que um cliente qualquer. Mesmo nas faixas de até 2 anos e até 3 anos, a taxa de *churn* é maior que na base geral. De notar também que, a partir dos 10 anos de permanência na operadora, a taxa desce para metade.

4.3.2. Variação do número de chamadas efetuadas para contactos da rede

Este atributos de variação é também um daqueles que é indicado pelos algoritmos como um dos mais relevantes.

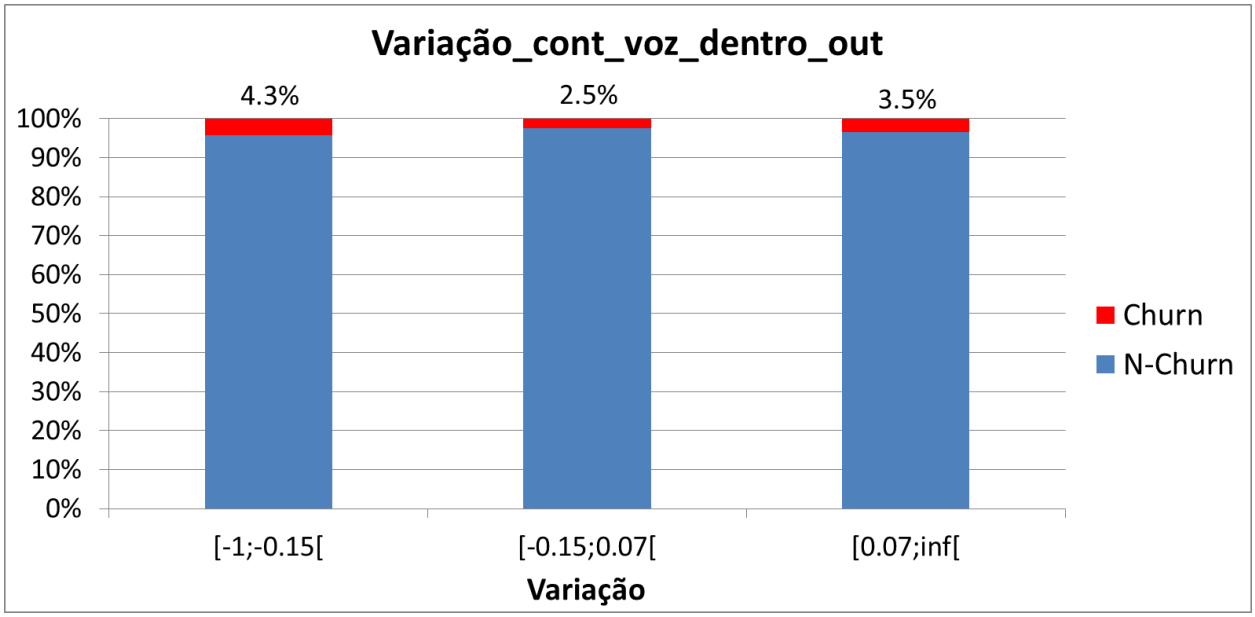

Fig. 10 - Variação_cont_voz_dentro_out

Pela análise do gráfico que foi dividido nos intervalos indicados de forma a que cada intervalo tivesse o mesmo número de individuos (discretização equal-binning), é possível constatar que existe uma probabilidade maior de *churn* quando existe uma variação negativa neste indicador, a partir dos 15%. Por outro lado, uma variação mais neutra, não é indicadora de *churn*, aliás quando este comportamento se verifica até existe uma menor probabilidade disso acontecer.

4.3.3. Número total de chamadas efetuadas durante a semana

Este atributos de atividade é também um daqueles que é indicado pelos algoritmos como um dos mais relevantes.

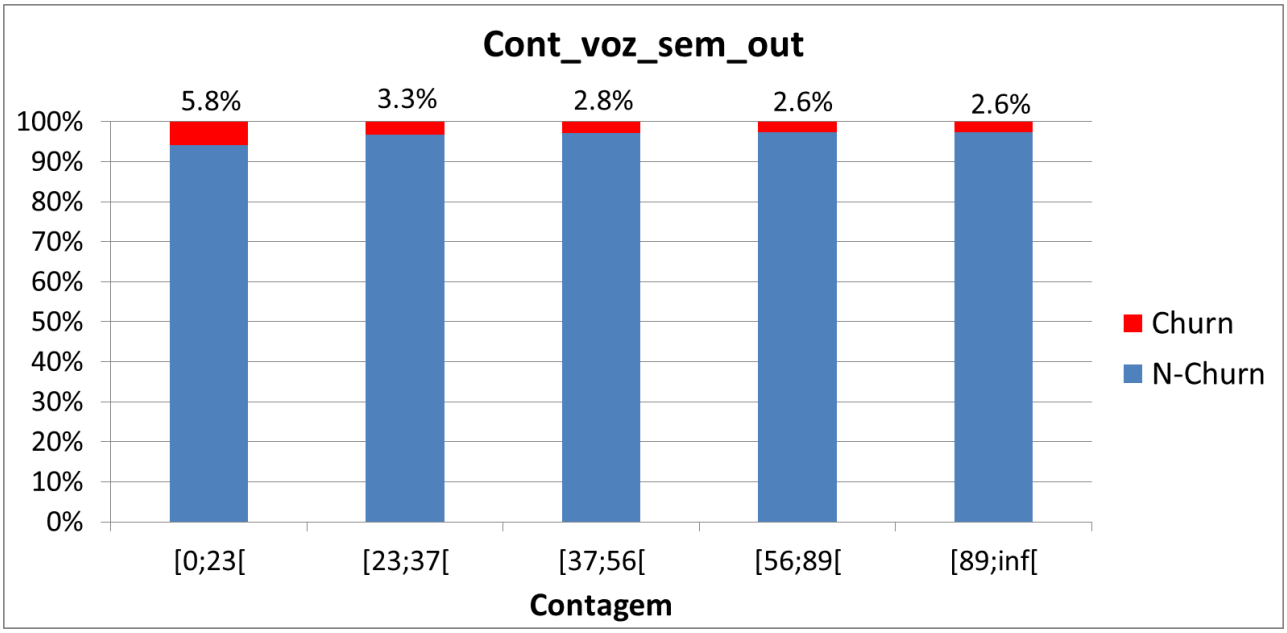

Fig. 11 - Cont_voz_sem_out

Utilizando novamente a discretização de forma a que cada intervalo tenha o mesmo número de individuos, podemos constatar que pessoas que utilizam menos o telemóvel, estão mais propensas a fazer *churn* que a média. Isso fica demontrado pela percentagem de 5,8%, acima dos 3,4% da base geral.

4.3.4. Percentagem de contactos da mesma rede que se mantêm

Este atributos relacionado com a rede de contactos do cliente é também um daqueles que é indicado pelos algoritmos como um dos mais relevantes. Esta variável indica a percentagem de contactos que se mantêm durante os 3 meses em análise, que pertencem à operadora, dentro do total de contactos do cliente.

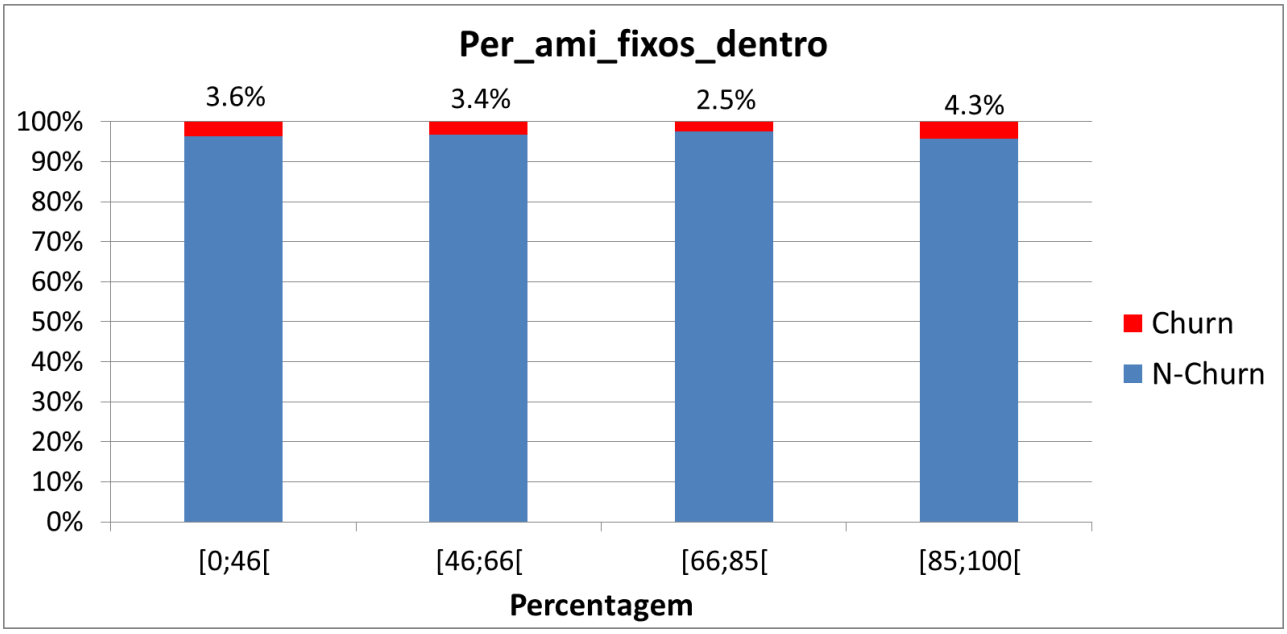

Fig. 12 - Per_amigos_fixos_dentro

A interpretação desta variável não é tão simples já que o lógico seria quanto mais ligado um cliente estivesse à operadora em termos de contactos, mais ele lhe seria fiel.

No entanto, uma possível explicação para o facto de, no intervalo entre 85 e 100% de amigos da operadora, a percentagem de *churn* ser acima da média e acima dos outros intervalos poderá ser o facto de na eventualidade de um ou mais contactos abandonarem a rede, estes clientes serem mais propensos a deixarem também, ao invés de clientes que têm a sua base de contactos mais distribuída por outras operadoras.

4.3.5. Média de saldo quando o cliente faz uma recarga

Este atributos relacionado com as recargas é também um daqueles que é indicado pelos algoritmos como um dos mais relevantes.

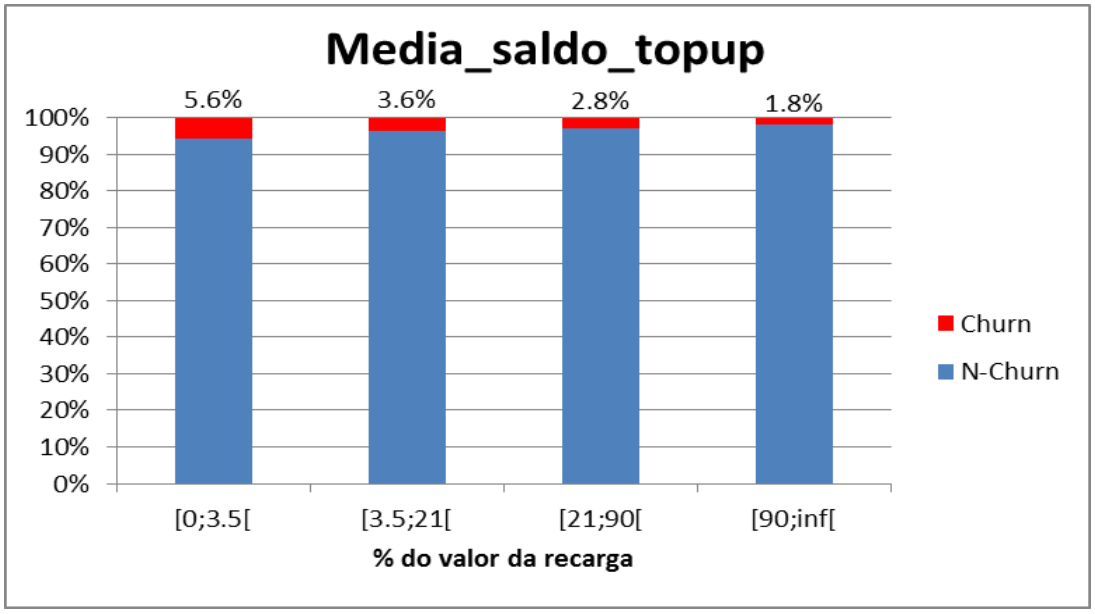

Fig. 13 - Média saldo *topup*

Pela análise desta variável é possível constatar que aqueles clientes que fazem uma recarga quando já praticamente não têm saldo, percentagem muito baixa relativamente à recarga, têm muito mais probabilidade de serem *churners*. No sentido inverso, quando os clientes recarregam ainda com muito saldo, percentagem elevada relativamente à recarga, indica uma propensão a não entrarem em dormência.

4.3.6. Percentagem de contactos que entraram em dormência entre 30 a 60 dias antes do dia em análise

Este atributo é também um daqueles que é indicado pelos algoritmos como um dos mais relevantes. Tal como enunciado anteriormente, indica a percentagem de contactos do cliente que entraram em dormência entre 30 a 60 dias do dia da análise.

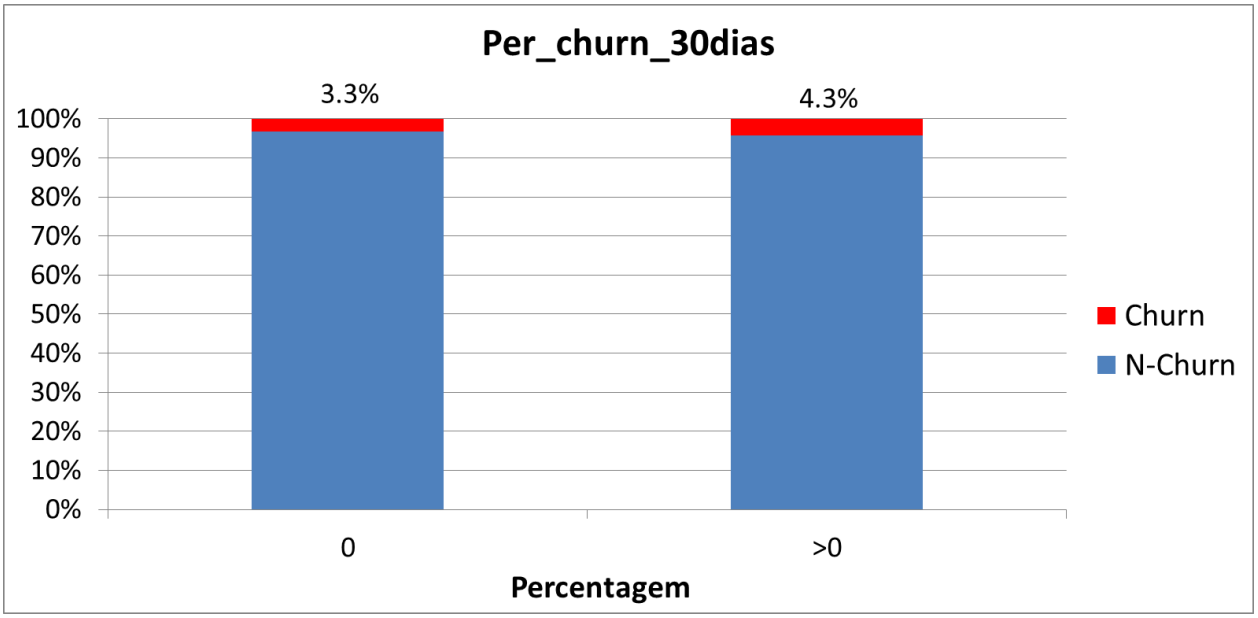

Fig. 14 - Per_churn_30dias

Para a análise desta variável não foi feita a distribuição de acordo com o número de indivíduos por classe já que apenas 13% dos clientes da base tinham contactos que deixaram a rede nestas circunstâncias. Portanto, a coluna com 0%, representa 87% da base de clientes, sendo que a outra coluna representa o resto. No entanto, apesar do desbalanceamento dos dados nesta variável, podemos constatar aquilo que seria expectável. A percentagem de *churn* é maior no segmento de clientes que têm contactos que entraram em dormência durante o segundo mês anterior ao mês no qual é feita a análise da dormência.

Na sequência da análise desta variável podemos verificar que no caso dos clientes que entram em dormência entre 15 a 30 dias do dia em análise, existem também evidências do mesmo fenómeno.

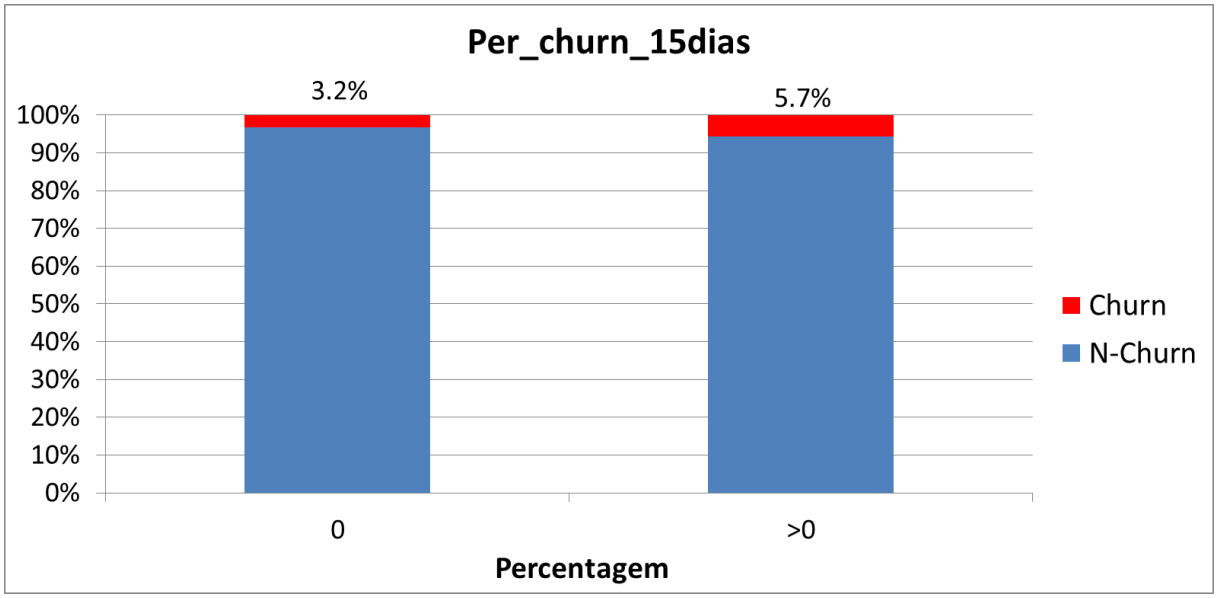

Fig. 15 - Per_churn_15dias

Neste caso, o desbalanceamento dos dados é ligeiramente superior (90-10), no entanto, as conclusões não deixam de ser válidas, indicando um efeito em cadeia verificado após um cliente entrar em dormência.

4.4. Efeito dos tipos de atributos

Nesta secção procura-se entender de forma seccionada o efeito de cada grupo de atributos na performance do algoritmo. Para esta análise foram usados o *top* 20 atributos conforme a conclusão da secção 6.2.4.

Os 20 atributos *top* são os seguintes:

Geral	**Rede**
antiguidade	cont_amigos_prox_dentro_OUT
	cont_amigos_dentro_OUT
Recarga	cont_amigos_rivais_OUT
media_saldo_topup	per_amigos_prox_rivais_OUT
media_topup	per_amigos_prox_dentro_OUT
media_sem_topup	cont_ami_fixos_dentro
	cont_ami_fixos_rivais
Variação	per_ami_fixos_dentro
delta_cont_voz_dentro_OUT	per_ami_fixos_rivais
delta_soma_voz_dentro_OUT	per_churn_30 dias
	per_churn_15 dias
Atividade	
cont_voz_SEM_OUT	
cont_voz_DIA_OUT	
cont_voz_dentro_OUT	

Fig. 16 - Atributos Finais

Nesta análise pretende-se verificar o impacto na previsão com e sem as variáveis de rede.

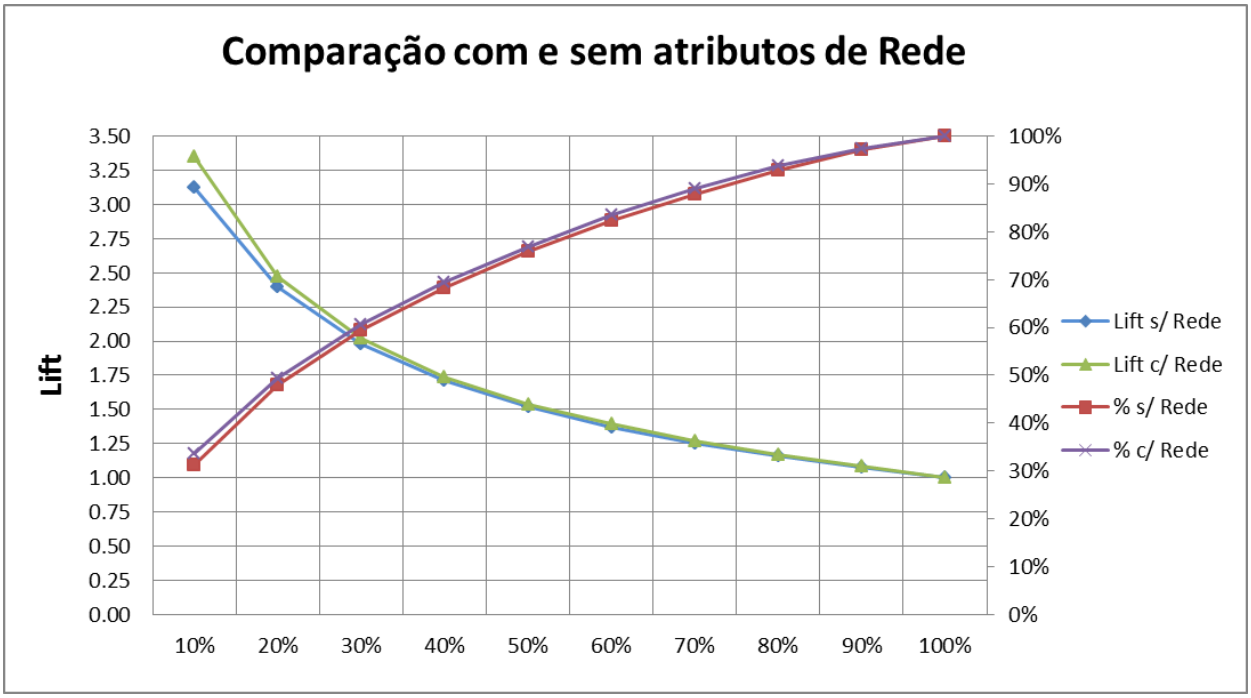

Fig. 17 - Comparação com e sem atributos de Rede

Como é possível verificar a grande diferença acontece ao nível do *top* 10 e 20 % de *lift*. No *top* 10% incluindo os atributos de rede é alcançado um *lift* de 3.35, enquanto que sem esses atributos o *lift* fica-se pelos 3.13.

Concluindo, neste capítulo foram analisados todos os atributos calculados e a opção para o capítulo seguinte na secção do modelo tradicional será utilizar os 20 atributos mais importantes de acordo com a análise feita neste capítulo.

5. Modelação

Neste capítulo serão abordados os testes referentes aos diferentes modelos utilizados, assim como aos algoritmos de classificação e parametros.

5.1. Modelo de propagação (SPA)

Este modelo introduzido por (Dasgupta & Singh, 2008) pretende simular a propagação de energia/influência que existe quando um cliente entra em dormência e transmite essa (má) influência aos seus contactos. Este modelo, tal como descrito na revisão da literatura, é composto por três componentes principais:

- **Fator de dispersão (d)**

Quanto maior o fator, mais influência é transmitida aos nós imediatamente a seguir aos nós ativos.

- **Função de dispersão**

A percentagem de influência transmitida aos nós seguintes é definida por uma função que alocará maior influência aos nós que em termos relativos tiverem uma relação mais forte com o nó ativo.

- **Força da ligação**

A força da ligação irá ser a variável utilizada pela função de dispersão para determinar a percentagem de influência a distribuir pelos nós circundantes.

No início do modelo é utilizada a rede de contactos dos clientes, sendo atribuido a cada nó (cliente) uma energia inicial determinada pelo seu estado no futuro (*churn* ou não-*churn*).

O processo de propagação começa tornando ativos os nós *churners*. Na primeira iteração estes nós propagam a energia para os nós vizinhos. A partir daí, os nós para os quais foi transmitida influência propagam para os seus vizinhos e por aí em diante.

Regras da propagação

1. Um nó não pode ser ativado mais do que uma vez, ou seja, cada nó só propaga a energia uma vez.
2. A propagação é feita da seguinte forma: o nó ativo fica com (1-d)*Ei, em que Ei é a energia do nó na iteração i, e propaga o restante d*Ei para a sua rede de vizinhos.
3. A propagação é feita pelos vizinhos de forma linear, isto é, a relação entre o nó ativo e os seus vizinhos é definida como uma percentagem do total da variável escolhida para força da ligação. A energia a propagar é então distribuída de acordo com essas percentagens.
4. A energia inicial é igual à energia final. A propagação termina quando não há mais nós para ativar.

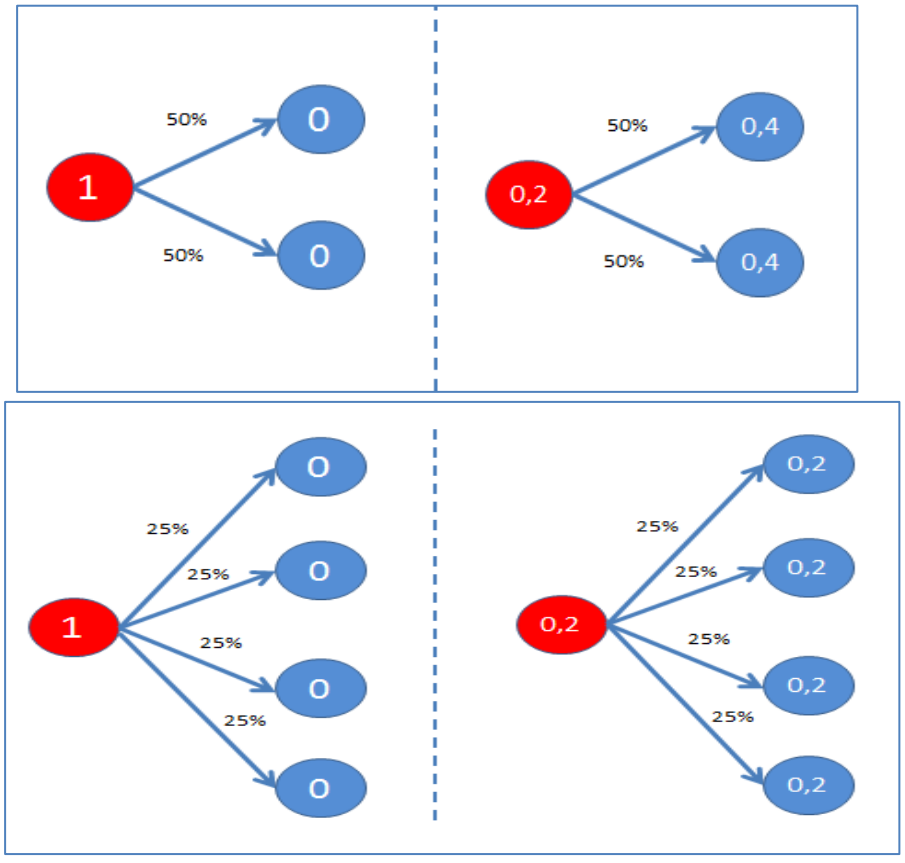

Fig. 18 - Modelo propagação SPA (d=0,8)

Experiências realizadas

Premissa: foi construída a rede baseada nos contactos no mês n-1 ou n-1 e n-2 (conforme ponto 3.), nessa rede foram identificados como nós ativos os *churners* no mês n servindo de início à propagação. Este modelo foi avaliado de acordo com os clientes que entraram em dormência no mês n+1. O modelo foi testado considerando diferentes cenários:

1. Variação do fator de dispersão (d): três valores diferentes 0.2, 0.5 e 0.8.
2. Variação da força da ligação: soma_voz (tempo falado entre clientes) e cont_interações (contagem de interações entre clientes).
3. Variação da rede utilizada: um mês, dois meses, contactos de primeiro grau, contactos de segundo grau.

   Em qualquer uma das variantes o ponto de início são os nós *churners*. A partir desses nós é construída a rede com os seus contactos e os contactos dos seus contactos. A primeira versão refere-se aos contactos de primeiro grau, a segunda aos contactos de segundo grau.

   Na variante de um mês significa que a rede de contactos é construída baseada nos contactos do último mês, na de dois meses significa os contactos dos últimos dois meses.

Resultados

Lift top 10%		d = 0,2	d = 0,5	d = 0,8
1 mês - 1º grau	soma_voz	1.45	1.44	1.35
	cont_interações	1.34	1.34	1.37
2 meses - 1º grau	soma_voz	1.57	1.56	1.50
	cont_interações	1.49	1.48	1.41
1 mês - 2º grau	soma_voz	1.28	1.24	1.25
	cont_interações	1.26	1.23	1.23
2 meses - 2º grau	soma_voz	1.33	1.30	1.29
	cont_interações	1.30	1.30	1.29

Tabela 9 - Resultados SPA

Numa primeira análise aos resultados, há dois pontos importantes que se podem salientar. Primeiro, os valores de *lift* obtidos estão muito abaixo do que seria de esperar, já que são pouco acima de 1. Em segundo lugar, verifica-se um decréscimo generalizado na performance do modelo quando se faz a propagação até ao 2º grau.

Em relação ao primeiro caso, pode-se afirmar que o facto de apenas considerarmos o efeito de um cliente entrar em dormência, não é por si só um fator determinante para a decisão de outros clientes entrarem em dormência. Esta inferência não implica que esta variável não tenha importância, até porque é uma das mais importantes segundo a classificação obtida pelos algoritmos de seleção de atributos (per_churn_30dias), no entanto, é de frisar que por si só pelos resultado obtidos não tem um grande efeito preditivo.

Em relação à segunda conclusão, uma das possíveis explicações surge no facto de ao considerarmos mais nós, de uma rede com cerca de 200 mil nós com os contactos de 1º grau, passa-se para uma de cerca de 4 milhões de nós. Esta alteração, conjugada com o facto de existirem cerca de 5% potenciais *churners* na primeira rede, para 0,25% na segunda, leva a que a energia se disperse mais por nós que não vão entrar em dormência, o que eleva ao *top* 10% mais nós que não estão em risco de dormência piorando o *lift*.

5.2. Modelo tradicional com variáveis de rede

Este modelo foi construído baseado nos atributos e respetiva seleção explicados nos capítulos anteriores. As diferentes bases para o modelo foram os diversos meses, sendo que os *churners* foram avaliados sempre durante os dias imediatamente a seguir ao último dia utilizado para o cálculo dos atributos.

Os critérios para a escolha dos clientes que entrariam para a base de treino e de teste foram ajustados de forma a cumprir dois requisitos chave. Primeiro, tratarem-se de clientes que utilizam regularmente o serviço e para isso foi considerado o critério de pelo menos 7 interações em dias diferentes por mês nos últimos 3 meses.

Para além disso, é importante garantir que estes clientes que estão na base e entram em dormência no mês seguinte, ainda são contactáveis. Para isso, é colocado mais um critério: tem que ter pelo menos uma comunicação nos últimos 3 dias.

Nas experiências realizadas foram utilizados diferentes bases de clientes conforme o mês escolhido, diferentes períodos de dormência (30, 40 e 50 dias) e diferentes algoritmos de classificação.

RL: regressão logística

PLC (x): perceptron multi-camada (número de camadas)

Os resultados apresentados são referentes ao *lift* no *top* 10%.

Treino	Teste	30 DIAS			40 DIAS			50 DIAS		
		RL	PLC (1)	PLC (2)	RL	PLC (1)	PLC (2)	RL	PLC (1)	PLC (2)
Setembro	Outubro	2.914	3.179	3.118	2.692	2.813	2.856	2.561	2.718	2.712
Setembro	Novembro	3.113	3.352	3.313	2.906	3.106	3.108	2.857	3.03	3.02
Setembro	Dezembro	2.948	3.153	3.211	2.697	2.931	2.956	2.607	2.767	2.858
Setembro	Janeiro	2.831	2.925	3.031	2.678	2.807	2.802	2.601	2.704	2.688
Outubro	Novembro	3.119	3.354	3.341	2.825	3.044	3.025	2.72	2.92	2.923
Outubro	Dezembro	2.848	3.164	3.178	2.627	2.901	2.766	2.507	2.708	2.731
Outubro	Janeiro	2.799	2.973	2.977	2.567	2.634	2.666	2.444	2.571	2.608
Novembro	Dezembro	2.855	3.215	3.168	2.69	2.948	2.983	2.577	2.781	2.847
Novembro	Janeiro	2.898	3.011	3.001	2.658	2.853	2.802	2.576	2.772	2.753
Dezembro	Janeiro	2.828	2.987	3.058	2.698	2.832	2.839	2.601	2.649	2.73

Tabela 10 - Resultados modelo tradicional

Em primeiro lugar, podemos constatar a grande diferença de performance entre a regressão logística e o perceptron multi-camada, sendo este último bastante superior relativamente ao anterior. Para além disso, podemos verificar que, em geral, há uma diminuição na performance em todos os algoritmos à medida que para base de treino se utiliza um conjunto de meses mais distante dos meses a prever. Por último, podemos verificar que a previsão tem melhores resultados para a dormência a 30 dias.

É possível também constatar isso no gráfico seguinte.

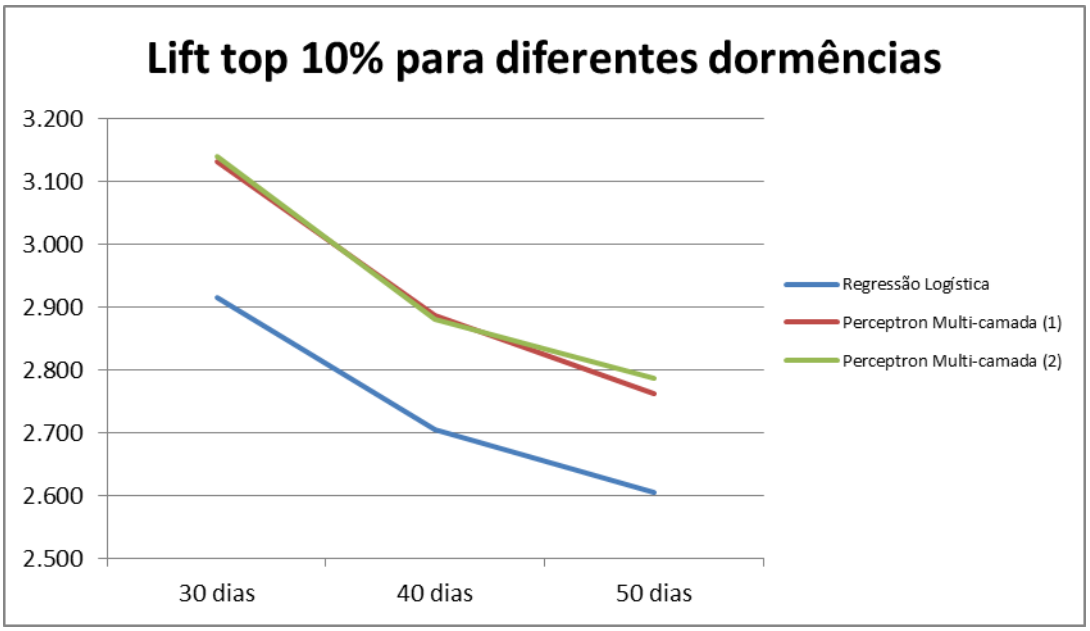

Fig. 19 - *Lift top* 10% para diferentes dormências

5.2.1. Modelo tradicional com variáveis de rede – análise *churners*

Com esta análise, pretende-se avaliar o impacto que tem no *lift* o facto de nem todos os clientes que se encontram no *top* 10% estarem ainda ativos, no momento em que é utilizado o modelo.

Para os exemplos em baixo, consideremos que estavamos no dia 1-Out. Nesta análise, devemos ter em conta duas situações:

a) Clientes que abandonam nos últimos dois dias do mês de treino (dias 29-Set e 30-Set). Apesar destes clientes serem excluídos da base de treino, não podem naturalmente ser excluídos da base de teste já que precisaríamos da informação do próprio dia em que estávamos para chegar a uma conclusão. Analisando as bases de teste disponíveis (em que temos essa informação) podemos chegar a uma estimativa do que acontecerá na realidade. Em média, nas bases de 30 dias de dormência, cerca de 1,9% dos clientes estão nessa situação, desses, cerca de 1,16% encontram-se no *top* 10%, isto dá um total de

cerca de 0,02% da base de clientes que tudo indica que já não serão recuperáveis.

b) Clientes que abandonam no dia em que é utilizado o modelo (dia 1-Out). Quando é feita a previsão para o mês seguinte, é necessário ter os dados até ao final do mês anterior (dia 30-Set). Logo, isso só será possível no dia 1-Out. Por motivos de simplicidade, nos testes efetuados, esse dia foi incluído no cálculo do *lift*. No entanto, considerando que só a partir do dia 2-Out, os clientes começariam a ser abordados, teremos forçosamente que excluir do *lift* os clientes que abandonam exatamente no dia 1-Out. Em média, nas bases de 30 dias de dormência, são cerca de 0,05% da base de clientes.

Conclusão: efetuando as correções para os três períodos de dormência, em média, houve uma redução de cerca de 6%, que pode ser visualizada no seguinte gráfico.

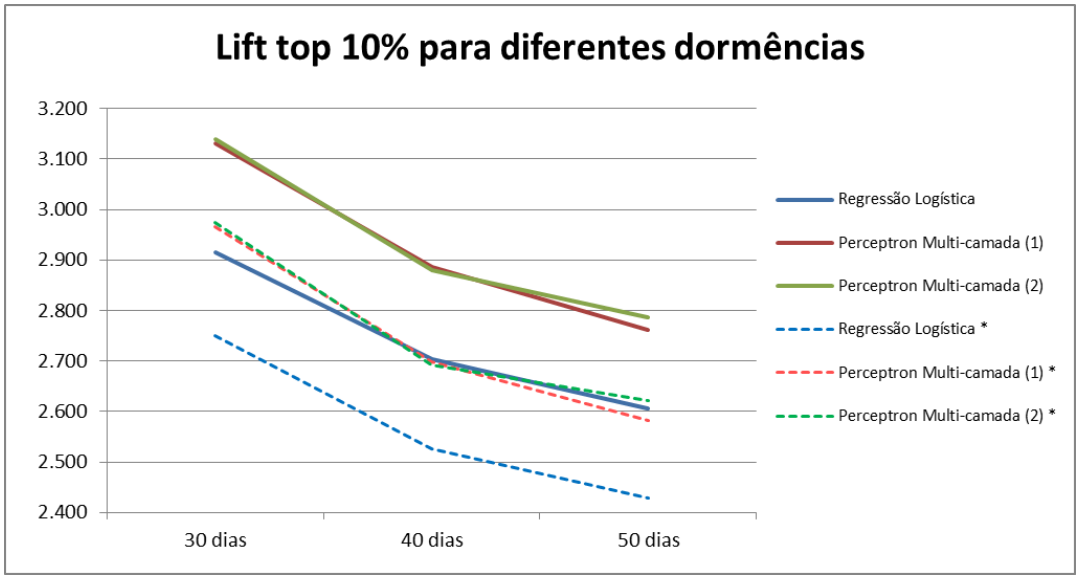

Fig. 20 - *Lift top* 10% para diferentes dormências (Corrigido)

De notar, que os valores corrigidos estão assim mais próximos da realidade que será encontrada quando o modelo for aplicado.

6. Conclusão e Trabalhos Futuros

Este trabalho tinha como objetivo construir um modelo de previsão de entrada de clientes em dormência com a antecipação necessária de forma a ser possível agir a tempo de impedir que a dormência ocorra e a perda de lucro associada a esse cliente.

Em termos de objetivos pode-se afirmar que este trabalho produziu um modelo que está em linha com aquilo que tem sido publicado nos *papers* mais recentes em termos de *lift*. No entanto, este trabalho vai mais além ao analisar os clientes que já não serão contactáveis aquando das campanhas de retenção, algo que influencia positivamente o *lift* mas que na verdade deveria ser excluído dos resultados.

Um dos factos que poderá ter sido mais surpreendente foi o pouco poder preditivo do modelo de propagação SPA. No entanto, não há dúvida que o fenómeno de *churn* pode ser explicado em termos da transmissão de informação entre contactos. Este facto é demonstrado pela importância das variáveis de rede no modelo de previsão tradicional.

Como complemento deste trabalho, seria importante acrescentar outros tipos de variáveis, nomeadamente relativas ao serviço a clientes, por exemplo, número de reclamações ou número de chamadas para linha de avarias. Estas variáveis poderiam ajudar a encontrar outros tipos de *churners*.

Para além disso, poderia ser interessante uma análise mais focado nos grupos ou comunidades de utilizadores e não nos utilizadores individualmente, isto é, seria necessário uma extensão das variáveis de rede para incluir, por exemplo, tipo de comunidade em que está inserido, tamanho da comunidade, evolução prevista da comunidade (expansão, retração), entre outros.

7. Referências

Abbasimehr, H., Setak, M., & Soroor, J. (2013). A framework for identification of high-value customers by including social network based variables for churn prediction using neuro-fuzzy techniques. *International Journal of Production Research*, *51*(4), 1279–1294.

Balle, B., Casas, B., & Catarineu, A. (2011). The Architecture of a Churn Prediction System Based on Stream Mining. *Lsi.upc.edu*.

Bohn, A., Walchhofer, N., Mair, P., & Hornik, K. (2009). ePub WU Institutional Repository Social Network Analysis of Weighted Telecommunications Graphs, (March).

Breu, F., Guggenbichler, S., & Wollmann, J. (2008). A DUAL-STEP MULTI-ALGORITHM APPROACH FOR CHURN PREDICTION IN PRE-PAID TELECOMMUNICATIONS SERVICE PROVIDERS. *Vasa*.

Dasgupta, K., & Singh, R. (2008). Social ties and their relevance to churn in mobile telecom networks. *Proceedings of the 11th international conference on Extending database technology: Advances in database technology*, 668.

Haenlein, M. (2013). Social interactions in customer churn decisions: The impact of relationship directionality. *International Journal of Research in Marketing*, *30*(3), 236–248.

Hall, M. A. (1999). Correlation-based Feature Selection for Machine Learning, (April).

Huang, Y., & Kechadi, T. (2013). An effective hybrid learning system for telecommunication churn prediction. *Expert Systems with Applications*, *40*(14), 5635–5647.

Kusuma, P., & Radosavljevik, D. (2013). Combining Customer Attribute and Social Network Mining for Prepaid Mobile Churn Prediction. *benelearn2013.org*.

Ma, Y.-X., Xu, J.-Y., Peng, D.-C., Zhang, T., Jin, C.-Z., Qu, H.-M., … Peng, Q.-S. (2013). A Visual Analysis Approach for Community Detection of Multi-Context Mobile Social Networks. *Journal of Computer Science and Technology*, *28*(5), 797–809.

Motahari, S., & Mengshoel, O. (2012). The Impact of Social Affinity on Phone Calling Patterns: Categorizing Social Ties from Call Data Records. *… Workshop on Social …*, *12*.

Oentaryo, R. J., Lo, D., Zhu, F., & Prasetyo, P. K. (2012). Collective Churn Prediction in Social Network. *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 210–214.

Phadke, C., & Uzunalioglu, H. (2013). Prediction of Subscriber Churn Using Social Network Analysis. *Bell Labs Technical …*, *17*(4), 63–75.

Radosavljevik, D., Putten, P. Van Der, & Larsen, K. (2010). The Impact of Experimental Setup in Prepaid Churn Prediction for Mobile Telecommunications: What to Predict, for Whom and Does the Customer Experience. *Trans. MLDM*, *3*(2), 80–99.

Richter, Y., Yom-Tov, E., & Slonim, N. (2010). Predicting Customer Churn in Mobile Networks through Analysis of Social Groups. *SDM*, 732–741.

Robins, G. (2013). A tutorial on methods for the modeling and analysis of social network data. *Journal of Mathematical Psychology*, *57*(6), 261–274.

Slingsby, A., Beecham, R., & Wood, J. (2013). Visual analysis of social networks in space and time using smartphone logs. *Pervasive and Mobile Computing*, *9*(6), 848–864.

Sohn, J.-S., Bae, U.-B., & Chung, I.-J. (2013). Contents Recommendation Method Using Social Network Analysis. *Wireless Personal Communications*, *73*(4), 1529–1546.

Tsai, C.-F., & Lu, Y.-H. (2009). Customer churn prediction by hybrid neural networks. *Expert Systems with Applications*, *36*(10), 12547–12553.

Verbeke, W., Martens, D., & Baesens, B. (2014). Social network analysis for customer churn prediction. *Applied Soft Computing*, *14*, 431–446.