

# Estudo sobre a Predição de Churn

José Saias\*, Luis Rato e Teresa Gonçalves

Dep. de Informática, ECT, Universidade de Évora

2018-12-23

## Resumo

A monitorização detalhada que os meios digitais permitem tem vindo a acentuar o interesse no estudo do risco de abandono de um serviço por parte de um cliente, por forma a suportar o atempado acionamento de medidas de retenção. *Churn* é um termo muito usado nos negócios, para a perda de clientes.

Neste artigo apresentamos um levantamento do estado da arte sobre análise preditiva de *Churn*, com técnicas baseadas em *Machine Learning*, na área das telecomunicações, no sector financeiro e seguros, e em negócios em meio digital, assentes em SaaS e outros. Nem todos os setores são afetados da mesma forma pelo abandono de clientes. Na área das telecomunicações, por exemplo, o estudo e predição de *Churn* tem sido bastante trabalhado, como denotam as publicações relativas a dados desse ramo.

Pela análise bibliográfica, não se observa que exista um algoritmo ou método globalmente ótimo para a predição de *Churn*, com excelente desempenho em todos os tipos de dados. Denota-se alguma heterogeneidade nos dados, em cada cenário, e a não existência de uma só métrica de avaliação de desempenho, observando-se três métricas mais frequentes: AUC, precisão e exatidão.

## 1 Introdução

No início de 2018, um estudo<sup>1</sup> sobre o futuro da computação em nuvem, baseado em inquéritos a especialistas e *opinion-makers* realizado pelo grupo LogicMonitor apontava que 83% da computação empresarial em 2020 seria já efetuada em ambiente Cloud. A crescente utilização do ambiente Cloud, juntamente com o potencial dos dados recolhidos por ferramentas de monitorização detalhada, têm suscitado interesse no estudo sobre o risco de desistência ou abandono de um serviço.

Um dicionário online<sup>2</sup> na área de Gestão define *Churn* como "*attrition or turnover of customers of a business or users of a service*". Por outras palavras, *Churn* é a perda de clientes. A análise de *Churn* tem sido trabalhada no mais variados setores de negócio, incluindo banca, seguros, comércio eletrónico, telecomunicações, jogos online, media e TV digital *on-demand* ou venda a retalho [16]. A taxa de *Churn* é uma medida usualmente mencionada em estudos sobre fidelização e relatórios sobre a atividade de um negócio. Corresponde à percentagem de clientes perdidos num período de tempo específico, sem contar com os novos clientes angariados nesse período. Este período de tempo não é estático, podendo variar bastante em função

---

\*jsaias@uevora.pt

<sup>1</sup><https://www.forbes.com/sites/louiscolumbus/2018/01/07/83-of-enterprise-workloads-will-be-in-the-cloud-by-2020>

<sup>2</sup><http://www.businessdictionary.com/definition/churn.html>

do negócio. Veja-se por exemplo uma companhia aérea, onde o ciclo médio de repetição da compra será muito diferente do que sucede num jornal semanal. De acordo com [26], é comum existirem taxas de *Churn* entre 1% e 5% na área de SaaS. Mas esses valores podem flutuar e ultrapassar 15%, o que representa um significativo impacto na receita. Já relativamente à área de telecomunicações móveis e ISPs, o mesmo relatório aponta taxas de *Churn* inferiores, não ultrapassando 1% para empresas de topo.

*Up-Selling*, ou realização de *Up-Sell*, é uma estratégia que visa estimular o cliente a adquirir uma versão de produto ou serviço que é nova ou que é superior, em comparação com a versão que já tem ou que planeava originalmente comprar. *Cross-selling* é um conceito relacionado, que corresponde à oferta de produtos adicionais, como complemento a uma compra específica em que o cliente já tem interesse. Como exemplo, podemos pensar na sugestão de um capacete quando o cliente compra uma bicicleta.

Tanto *Up-Selling* como *Cross-selling* procuram aumentar vendas junto de clientes já existentes. A análise de dados necessária à redução de *Churn* poderá ser útil também para estas operações, nomeadamente se houver ganho de conhecimento sobre perfis de cliente. O presente documento sintetiza uma análise ao estado da arte em análise preditiva de *Churn*, com técnicas baseadas em *Machine Learning*.

## 2 Estado da Arte em Predição de *Churn*

Esta secção é dedicada a um levantamento de publicações sobre trabalho relacionado recente. Alguns dos aspetos chave serão recapitulados adiante, na secção 3. São referidas várias abordagens de *Machine Learning*, algoritmos ou métricas de avaliação, cuja descrição pode ser consultada em [18].

### 2.1 Sector de telecomunicações

Em 2017, Aditya Kapoor publicou um relatório sobre retenção de clientes e *Churn* no mercado americano de telecomunicações [2]. É referido um estudo, que indica uma taxa de *Churn* de 1,9% nos quatro operadores de topo (AT&T, Verizon, T-Mobile, Sprint). Estes operadores abrangem 100 milhões de clientes. Em média, o *Churn* ocorre aos 19 meses. É referido que o tempo de vida normal de um cliente é 52 meses, pelo que a receita perdida em cada caso de *Churn* é estimada em mais de 1100 dólares, de  $\$34 \times 33$  meses.

O autor descreve a modelação de um sistema preditivo para *Churn*, com base num repositório de cerca de 100.000 registos e 150 atributos. Os dados incluem informação sobre telefonemas (quantidade, duração, para os últimos 3 e para os últimos 6 meses); informação contratual; detalhes sobre o consumo de dados; e perfil socioeconómico do cliente. Com um processo do tipo *Filter*, foram selecionadas 20 características. Os dados foram divididos entre conjuntos de treino e teste, com tamanhos 70% e 30%, respetivamente. Através da plataforma Azure, são criados modelos de classificação com *Logistic Regression*, *Boosting*, *Random Forest*, redes neurais e SVM.

Não são indicados os resultados da avaliação. O autor menciona a otimização dos modelos através da procura de um limiar associado à maximização de lucro.

Num trabalho muito recente para encontrar clientes com maior probabilidade de *Churn*, descrito em [21], o autor descreve um processo que iniciou com aprendizagem não supervisionada, usando depois aprendizagem supervisionada, sobre dados de uma empresa de telecomunicações holandesa. Na primeira parte, o objetivo foi a identificação de grupos homogêneos de

utilizadores, mediante dois processos: segmentação com base em indicações de especialistas, por um lado, e também *clustering* com três técnicas: *t-Distributed Stochastic Neighbor embedding*, *Gaussian Mixture Model* e *Latent Class Analysis*. A segunda parte foi um processo de classificação, com os algoritmos *Logistic Regression*, *Random Forest*, *XGBoost* e um *ensemble* heterogéneo dos anteriores.

Foram usados dados demográficos, dados contratuais e relativos a comportamento online, de clientes com telefone móvel ou fixo, serviço de internet ou TV. O comportamento online inclui dados sobre itens em que o utilizador clicou e páginas visitadas. Os classificadores foram pensados para três classes: clientes que terminam o contrato (*Churn*), clientes que não tomam nenhuma iniciativa e deixam cessar o contrato atual (*Sleepers*), e clientes que renovam o contrato.

É mencionado que o custo de uma predição de *Churn* errada, com falso positivo, é maior do que custo de prever Não-*Churn* erradamente. No pré-processamento, são reportados alguns procedimentos, como obter a idade a partir da data de nascimento, ou o cálculo do número de dias até ao final da subscrição ativa. Em termos de classes, 10% das instâncias são da classe *Churn*. A métrica usada para avaliar estes modelos foi AUC, com a metodologia *5 fold cross-validation* e um conjunto de teste com 20% das instâncias, mantendo proporção entre classes. Para reduzir o risco de sobreajustamento, o conjunto de 80% para treino foi ainda subdividida entre conjunto de treino efetivo e conjunto de validação, e foi usada uma estratégia de *randomized search cross validation* com a medida macro F1 para escolher os hiperparâmetros. O trabalho inclui uma comparação destes modelos com um modelo referência treinado sem os inputs da fase de *clustering*, funcionando como *benchmark*.

Relativamente a usar a abordagem híbrida ou não, só a técnica de LR beneficiou com a fase de *clustering*. Os *clusters* revelaram-se pouco informativos para a predição de *Churn*. E ao mesmo tempo, é dito que os *clusters* tinham mais ruído que os grupos da segmentação feita por especialistas. Em termos de AUC relativamente à classe *Churn*, o valor mais alto foi para o classificador com XGBoost, com 0,95 no classificador baseline e valores entre 0,82 até 0,84 na abordagem híbrida. Estes últimos, apesar de inferiores ao observado para a versão base do mesmo algoritmo, foram os melhores na versão híbrida. A conclusão do autor indica que, naquele cenário, a melhor escolha para a predição de *Churn* seria XGBoost com a abordagem não híbrida, onde foi obtido um valor de AUC de 0,95, para a classe *Churn*.

Considerando as diferenças no custo de perda de cada tipo de cliente, em função dos seus gastos, o trabalho de Abbasimehr et al. [1] propõe uma abordagem em duas fases para a predição de *Churn* em clientes de elevado valor, aplicado sobre os dados de uma empresa de telecomunicações. A primeira fase visa encontrar os clientes de maior valor e os seus dados mais relevantes, para serem empregues na segunda fase, de criação dos modelos de *Churn*. A identificação do perfil de clientes de maior valor contempla dados sobre as suas interações e interligações numa rede social, e envolve uma operação de *clustering* com o algoritmo K-Means.

Os algoritmos de aprendizagem usados na segunda fase enquadram-se numa família de técnicas híbridas, *neuro-fuzzy*, que combinam redes neuronais com lógica difusa. Em particular, o trabalho usou *Adaptive Neuro-Fuzzy Inference System* (ANFIS) e *Locally Linear Neuro-Fuzzy* (LLNF), em MATLAB, aplicados juntamente com o algoritmo *Locally Linear Model Tree* (LoLiMoT).

Os autores estabeleceram ainda uma comparação entre os modelos com estas técnicas e um modelo treinado com redes neuronais do tipo *multi layer perceptron*, com vantagem para os primeiros, *neuro-fuzzy*. Os modelos preditivos foram avaliados, tendo sido obtidos valores de exatidão entre 44 e 52%, e resultados de sensibilidade entre 63 e 70%.

Em 2016, Georgina Esteves [15] comparou seis abordagens para identificar clientes que estão mais perto de abandonar o fornecedor de telecomunicações, num estudo com dados reais fornecidos por uma empresa consultora portuguesa, extraídos de um repositório com mais de cinco milhões de clientes.

Nos dados originais, apenas 18% dos casos correspondiam a *Churn*. Para minimizar o desequilíbrio entre classes, foi aplicada uma técnica de amostragem designada *Synthetic Minority Over-sampling Technique* (SMOTE) [5]. Este processo gera novos casos da classe minoritária com base em vizinhos próximos a casos existentes, e reduz casos da classe maioritária. O conjunto de treino corresponde a 70% das instâncias disponíveis. Destas, 38211 são casos *Churn* e 50948 correspondem a casos Não-*Churn*. O conjunto de teste corresponde a 30% do conjunto inicial.

Foram desenvolvidos modelos de classificação com os algoritmos *Naive Bayes*, árvores de decisão com C4.5, *Random Forest*, *AdaBoost* e redes neuronais e KNN.

A avaliação dos modelos envolveu três métricas: AUC, sensibilidade e especificidade, com um procedimento de *10-fold cross validation* com 3 repetições. A técnica com melhores resultados corresponde ao algoritmo *Random Forest*, com 0,9654 em AUC e 0,8417 para sensibilidade. O algoritmo C4.5 levou ao segundo melhor resultado de AUC, com 0,9073.

Tendo a aplicabilidade e a interpretabilidade como aspetos chave, Garcia Gomez desenvolveu a sua tese de doutoramento experimentando técnicas para *clustering*, análise e visualização de dados alusivos ao *Churn* no sector de telecomunicações no Brasil [14]. O primeiro grupo de experiências relatado envolve dados de outro sector de atividade. A partir de uma coleção de questionários aplicados junto de vários milhares de clientes de empresas petrolíferas, é feito um primeiro estudo sobre os fatores de satisfação dos clientes. Com uma abordagem de aprendizagem supervisionada, os dados são classificados com rede neuronal do tipo *feedforward* multicamada, treinada com um método bayesiano. Este método, na procura dos pesos ótimos, considera uma distribuição de probabilidades para os pesos, refletindo a incerteza. As duas classes nesta tarefa são *Satisfeito* e *Não-Satisfeito*.

Através de um procedimento *Orthogonal Search Rule Extraction* (OSRE), são identificadas as regras de classificação associadas ao modelo da rede neuronal. Isto introduz intelegibilidade ao processo, uma vez que as redes neuronais costumam ofuscar a lógica de classificação. Para a classe *Satisfeito*, as regras de classificação identificadas têm associada uma avaliação com especificidade entre 0,92 e 0,96, sensibilidade entre 0,10 e 0,21, e precisão entre 0,69 e 0,85.

Num segundo estudo procurou-se uma relação entre os padrões de consumo ou uso do serviço telefónico e a propensão para o abandono do serviço. Os dados usados eram relativos a chamadas telefónicas de 60596 pequenas e médias empresas brasileiras, clientes de um grande operador telefónico do estado de São Paulo, no Brasil. O procedimento executado compreendeu dois passos. Primeiro foram identificados micro-segmentos, com uma metodologia *Generative Topographic Mapping* (GTM), com uma grelha de  $8 \times 8$ . Os 64 micro-segmentos foram depois agrupados, para uma visão de nível superior, com uma operação de *clustering* através do algoritmo K-Means. O resultado foi um procedimento de *two-tier clustering*, com a primeira parte baseada em GMT, com interesse para o *clustering* e para a componente de visualização de dados.

Para a componente de visualização dos dados apurados pelo processo analítico e pelos modelos desenvolvidos, e a propósito da interpretabilidade, o autor apresenta técnicas inspiradas em métodos de representação geoespacial de elementos estáticos e dinâmicos. Na sequência dos estudos anteriores, foi desenvolvida uma ferramenta analítica para visualização e exploração dos dados multivariados. Através do suporte para representação de fluxos, a ferramenta foi

usada para análise da migração de clientes entre fornecedores de serviços, ao longo do tempo, relativos a duas coleções: dados de operadoras de telecomunicações no Brasil, e dados de um serviço espanhol de televisão *pay-per-view*.

Stephen Nabareseh desenvolveu um trabalho recente sobre modelos de predição de *Churn* [30]. Enquadrado na sua tese de doutoramento, o trabalho foi aplicado a dados reais de seis empresas de telecomunicações num país da África ocidental.

Os dados incluem respostas a questionários aplicados a clientes de serviço de voz e dados. Num tratamento prévio, foi efetuada uma análise exploratória de dados, para apurar atributos mais relevantes, *clustering*, e mineração de regras de associação.

As experiências realizadas envolveram o treino de modelos com três algoritmos: árvores de decisão com C4.5, *Logistic Regression* e *Linear Discriminant Analysis*. Os modelos foram comparados através de dois indicadores AUC e exatidão. Os valores de AUC foram 0,984, 0,663 e 0,663, respetivamente para C4.5, LR e LDA. Na medida exatidão, os resultados foram 0,747, 0,653 e 0,628, respetivamente para os mesmos modelos. É salientado o melhor desempenho com árvores de decisão, C4.5.

Entre as variáveis mais relevantes, são mencionadas a região do cliente, género, ocupação e gastos mensais com o serviço. É reportado o uso das ferramentas IBM SPSS Modeler e RapidMiner.

Noutro trabalho na área das telecomunicações [32], Wouter Verbeke usou técnicas de indução de regras para predição de *Churn*, com ênfase na intelegibilidade dos modelos. Essas técnicas, AntMiner+ e *Active Learning Based Approach* para SVM (ALBA), foram comparadas com modelos de classificação baseados em C4.5, Ripper<sup>3</sup>, SVM e *logistic regression*. AntMiner procura inferir regras a partir dos dados, com uma metodologia de *Ant Colony Optimization*. A técnica combinada de ALBA com C4.5 ou Ripper resulta no melhor valor de exatidão, aproximadamente 93,8%. O melhor valor no indicador sensibilidade foi 82,29%, obtido para C4.5 num modelo treinado com sobreamostragem.

Noutra experiência, foi usado o algoritmo RULEM para induzir modelos de classificação baseados em regras. A avaliação foi realizada com cinco conjuntos de teste em *hold-out*, com valores de exatidão para *Churn* entre os 87,26% e os 93,68%.

Noutra experiência sobre a associação entre clientes, ou redes sociais de conhecimentos, o autor conclui que existe uma influência grande dessas redes no *Churn* em telecomunicações, indicando mesmo que não apenas os amigos, mas também os amigos dos amigos têm influência na probabilidade de *Churn* de um cliente.

Ole-Christofer Granmo aplicou processos gaussianos na predição de *Churn* sobre os dados de uma empresa de telecomunicações francesa [25]. Os dados, pertencentes a um conjunto usado também na competição *Knowledge Discovery in Data*, incluem 230 atributos numéricos e categóricos, para 50000 instâncias com anotação binária sobre o *Churn*.

Com muitos valores em falta, as instâncias têm uma distribuição por classes desequilibrada, com 7,3% de casos *Churn*. A primeira operação de limpeza e eliminação de constantes reduziu o número de atributos para 176. O segundo passo de seleção foi baseado na correlação entre variáveis, com o coeficiente de correlação de Pearson com limiar 0,9, o que reduziu os atributos considerados para 147.

O modelo com o processo gaussiano foi implementado em MATLAB. Na avaliação, o modelo baseado nas cinco características numéricas mais informativas obteve um valor de AUC de 0,67, sendo que o modelo com melhor avaliação obteve 0,77, no mesmo indicador. A abor-

---

<sup>3</sup>de *Repeated Incremental Pruning to Produce Error Reduction*, uma técnica de indução de regras.

dagem assente em processos gaussianos foi comparada com métodos alternativos sobre os mesmos dados, com árvores de decisão, redes neuronais, KNN, LR, SVM e *ensembles*. O melhor valor de AUC entre estes métodos alternativos foi 0,73 para um *ensemble*, seguido de 0,7 para SVM. Estes modelos alternativos foram treinados com Knime e Weka.

Naren Tanneedi realizou um estudo sobre os motivos para *Churn*, com dois processos de análise: um estatístico, e outro baseado em aprendizagem automática, com árvores de decisão J48 [22]. Foram realizadas experiências com três conjuntos de dados: sobre seis meses de atividade de clientes de telecomunicações; sobre dados provenientes de inquiridos; e dados semanais sobre o uso do serviço de dados em dispositivos Android. Os dados incluem indicações sobre a qualidade, o “aborrecimento” do cliente (*annoyance*) e *Churn*. O modelo treinado para predição de *Churn* com árvore de decisão J48 foi avaliado, tendo-se apurado uma exatidão entre 71 e 95%. O autor treinou o modelo J48 com a ferramenta Weka.

Noutro trabalho na área das telecomunicações móveis, na Islândia, Emilia Nguyen defende que os elementos cruciais para a retenção de clientes são a predição de *Churn* e estratégias eficazes de prevenção do abandono. A autora fez um estudo para a construção de um modelo preditivo que indicasse as probabilidades de *Churn* para um cliente daquele sector [10]. Os dados iniciais dividem-se em seis categorias: demográficos; sobre faturação (em particular para tarifários pós-pagos); recarregamentos (sobre intervalos entre recarregamento de saldos e valores usados); padrão de comunicações (quantidade e duração dos telefonemas, redes de destino dos telefonemas); CDR ou registo de chamadas e consumos efetivamente faturados; e rede (com dados dos conhecidos ou interlocutores com quem um cliente mais interage, do *Churn* desses interlocutores da rede, e medida de centralidade do cliente nesse grupo). A predição assenta numa tarefa de classificação. Foram utilizados diversos métodos de seleção de características, incluindo análise do ganho de informação e análise de correlação de variáveis, e métodos *Wrapper* com árvores de decisão, Naive Bayes e *Logistic Regression*. Das categorias de dados mencionadas, a que teve maior proporção de atributos selecionados foi a categoria dos dados demográficos, com 47% dos atributos escolhidos. Os atributos do tipo CDR são mais explicativos para o *Churn* em pós-pago que em pré-pago, com 29% dos atributos usados apenas no modelo para pós-pago, e 14% dos atributos escolhidos no cenário dos pré-pagos.

Foram constituídos conjuntos de treino e de teste para pós-pago e pré-pago. Para diminuir o desequilíbrio entre classes, usou-se subamostragem, tendo resultado em conjuntos de treino com 4380 instâncias para pós-pago e 8464 instâncias para pré-pago. Os conjuntos de teste ficaram com a distribuição por classes original, para a avaliação refletir a situação real. Os modelos para classificação foram construídos com quatro algoritmos de aprendizagem: árvores de decisão C4.5 (J48 com Weka), *ensemble* com *boosting* de árvores de decisão, Naive Bayes e *Logistic Regression*. Em termos de exatidão, o modelo com o algoritmo árvore de decisão J48 é avaliado com o melhor resultado em pós-pagos, com 67,2%. Já em pré-pago, o valor mais alto em exatidão foi para o modelo com LR, com 73,8%. Com a medida AUC, os melhores modelos foram Naive Bayes em pós-pago e o modelo com árvores de decisão em *ensemble* com *boosting*, em pré-pago, com 74,4% e 83,9%, respetivamente.

Na sua recente tese de mestrado na Universidade de Toronto, em 2018, Yue Cui explorou técnicas de classificação com métodos *ensemble* para predição em *Customer Relationship Management* [33]. Inserido numa empresa que já usava técnicas de predição, o autor procurou soluções baseadas em ML que atingissem melhor desempenho na identificação dos destinatários mais adequados a campanhas de marketing.

O conjunto de dados usado está relacionado com a operadora de telecomunicações Orange, usado também na *Knowledge Discovery in Data Competition* de 2009. Nas 50000 instâncias, existem variáveis numéricas e categóricas, e ainda as anotações binárias para *Churn*, apetência e *Up-Selling*. No pré-processamento, é reportado o tratamento de valores em falta, a remoção de variáveis constantes e maioritariamente constantes (contendo o mesmo valor em mais de 99,9% das instâncias). Os valores de alguns atributos categóricos que possuíam mais de 100 categorias foram transformados para um conjunto inferior de categorias agrupadas. O autor menciona a não aplicação explícita de uma técnica de seleção de características, por usar a seguir algoritmos baseados na entropia, como árvores de decisão (em *ensemble*), que priorizam automaticamente os atributos.

Foram testados modelos preditivos para cada uma das variáveis alvo através de três métodos: *Random Forest*, *Gradient Boosting* de árvores de decisão e um *Ensemble Selection* (variante de método *ensemble* que testa várias combinações de diferentes algoritmos e dos pesos a atribuir a cada um). Adicionalmente, usou-se um modelo de *benchmark*, com o algoritmo *Naive Bayes*. Os hiperparâmetros foram afinados através de um varrimento exaustivo, em grelha, e *cross-validation*. A avaliação mostra que os três classificadores *ensemble* superaram o *benchmark*, aumentando a taxa de verdadeiros positivos. O resultado da avaliação com o indicador AUC, na predição de *Churn* foi: 0,6472 para *Naive Bayes*; 0,7097 para *Random Forest*; 0,7418 para *Gradient Boosting*; e 0,76 para *Ensemble Selection*. Nos modelos para predição de apetência, os valores de AUC foram: 0,6463 para *Naive Bayes*; 0,8374 para *Random Forest*; 0,8304 para *Gradient Boosting*; e 0,8725 para *Ensemble Selection*. E para os modelos de *Up-Selling*, os valores de AUC foram: 0,72 para *Naive Bayes*; 0,8426 para *Random Forest*; 0,8914 para *Gradient Boosting*; e 0,9064 para *Ensemble Selection*.

É mencionado o elevado tempo de construção do modelo para o método *Ensemble Selection*, devido à procura de combinações de algoritmos que o método acarreta, levando aproximadamente quatro vezes mais tempo. Por outro lado, foi essa a técnica que obteve os melhores resultados.

Noutro trabalho sobre *Churn* na área das telecomunicações [7], o autor usa *clustering* na análise exploratória de dados, para apreciar até que ponto são explícitas as diferenças entre casos de *Churn* e Não-*Churn*. Nos dados usados, apenas 3,39% dos casos correspondem a *Churn*. *Logistic Regression*, *Lasso*, *Adaptive Boosting*, *Naive Bayes* e *Random Forest* foram os algoritmos usados nos vários modelos de predição. O autor aplicou os algoritmos sobre duas modalidades dos dados: na proporção de classes original; e sobre uma coleção de dados balanceada, após aplicação de sobreamostragem. O indicador AUC foi uma das medidas de avaliação consideradas. Alegando motivos de confidencialidade, o autor não disponibilizou os valores com o resultado da avaliação.

Em 2014, uma publicação de Federico Castanedo et al. sobre predição de *Churn* na área das telecomunicações, refere o esforço colocado por diversos trabalhos na fase de *feature engineering* ou processo de seleção de características, e como é altamente dependente de cada contexto e respetivos dados [12].

Os autores apresentam várias soluções de predição de *Churn* com base em redes neuronais multicamada, com várias configurações, em particular do tipo *feedforward* com 4 camadas. Esta abordagem é apontada como uma das primeiras propostas de utilização de DL para *Churn* documentadas.

Os dados incluem milhares de milhões de registos telefónicos de um sistema empresarial. É reportado um resultado médio de 0,779 no indicador AUC, o que melhora substancialmente o desempenho de modelos anteriores, com *Random Forest*, cuja avaliação era 0,732 em AUC,

e com extenso trabalho na seleção de características.

Noutro trabalho com DL, Wangperawong et al. aplicaram redes neuronais convolucionais à análise de *Churn* em dados de telecomunicações [31]. Os autores procuraram transpor técnicas já testadas com sucesso em classificação de imagens para a predição de *Churn*. Para isso, descrevem a representação dos dados de clientes e respetivo comportamento ao longo do tempo na forma de imagens, que são depois processadas por uma rede CNN. Este processo de aprendizagem supervisionada foi aplicado para treinar um modelo com os dados de 6 milhões de utilizadores. Os autores reportam um resultado de 0,743 no indicador AUC.

Numa publicação [28] em *datascience.com*, de abril de 2018, Sowmya Vivek descreve o uso do algoritmo LDA para predição de *Churn* com dados de clientes de uma empresa de telecomunicações, com foco especial em fatores relacionados com o serviço. Usando o LDA como mecanismo de segmentação, procura-se dividir os clientes entre os grupos *Churn* e *Não-Churn*. Foi realizada uma análise multivariada da variância (MANOVA) nos dados, e foi estudada a importância relativa de cada variável independente. A exatidão reportada para o modelo de predição base é 85,67%, mas com uma sensibilidade de 12,7%, muito baixa. Após alguns testes no sentido de equilibrar os indicadores, o limiar a usar para a classificação com o algoritmo LDA ficou em 0,16, o que permitiu estabilizar ambas as medidas exatidão e sensibilidade em 74%.

## 2.2 Sector financeiro e seguros

Num trabalho na área financeira, já de 2018, Erdem Kaya et al. investigaram padrões espaciotemporais e entropia nas decisões de carácter financeiro, e a relação que podem ter com o *Churn* [11]. Com inspiração em trabalhos de base computacional na área das ciências sociais, os autores conceberam um modelo preditivo sensível não apenas aos aspetos temporais e espaciais, mas também aspetos comportamentais sobre o modo como os consumidores gastam dinheiro.

Os dados incluem informação demográfica, transações com cartões de crédito ou transferências bancárias, organizadas em dois conjuntos, A e B. O conjunto A tem cerca de 45 milhões de transações, relativas a 100.000 clientes. O conjunto B tem cerca de 22 milhões de transações, relativas a 60.000 clientes. O conjunto B foi extraído da mesma fonte, mas com critérios específicos: clientes com 10 ou mais transações com cartão de crédito, e em que 60% ou mais dessas transações estivessem associadas a equipamentos POS do banco que cedeu os dados. Por existir desequilíbrio entre classes, a avaliação dos modelos foi feita com a medida AUC, com um método de *8-fold cross-validation*. Na avaliação dos vários modelos treinados com *Random Forest*, os valores de AUC indicados estão entre os 0,513 e os 0,79.

Em [23], Nie et al. descrevem a aplicação de dois algoritmos na predição de *Churn* sobre dados relativos a uso de cartão de crédito de um banco chinês. Os autores estudam a influência que as variáveis sobre o utilizador, sobre o cartão de crédito, sobre o risco, e sobre as transações têm relativamente ao *Churn*. O conjunto inicial de 135 atributos é reduzido, mediante uma seleção de características orientada pela correlação, mas também por fatores de ordem económica ou financeira.

Os algoritmos usados foram *Logistic Regression* e árvores de decisão, com vantagem para o primeiro, na análise de desempenho realizada, que considerou a medida exatidão, mas também uma medida de custo de erro na classificação, onde custo é visto no sentido económico/financeiro

das potenciais perdas associadas aos dois tipos de erro de classificação. A avaliação do melhor modelo com o algoritmo LR indica uma taxa de erro de 15,3%, e 84,7% para exatidão. Com árvores de decisão, o melhor modelo foi avaliado com taxa de erro igual a 16,9%, e 83,1% para exatidão.

Na área financeira, o trabalho de Nina Olofsson incidiu na procura de indicadores finos, ou explicações locais (e não globais), para o *Churn*, mediante a comparação de indicadores relativos a diferentes grupos de utilizadores [24]. Os grupos de base para diferentes modelos foram constituídos a partir das diferenças de idade, de densidade populacional da área de residência, e de antiguidade. Os dados foram extraídos de uma *App* financeira desenvolvida por uma empresa do tipo *fintech*. Depois de alguns filtros, foram apurados os dados relativos a 173312 utilizadores, sendo aproximadamente 70% da classe *Churn*. Foi usada sobreamostragem com a técnica *Adaptive Synthetic Sampling Approach*, que considera *clusters* e distâncias dos casos da classe minoritária face à classe maioritária.

Os modelos para predição de *Churn* foram treinados com *Random Forest*. O melhor modelo na avaliação, com processo *5-fold cross-validation*, teve um valor médio de AUC igual a 0,93, com precisão em torno de 86%. As conclusões referem que não foram encontradas diferenças significativas nos preditores mais relevantes de cada grupo.

Em [6], é descrito o trabalho de Chantine Huigevoort na predição de *Churn* sobre dados de seguros. Um dos objetivos foi a identificação das variáveis ou atributos mais relevantes na análise de *Churn*. Foi apurado que a idade, o número de vezes que um cliente é segurado na empresa, e os gastos totais com saúde, são as características mais importantes.

Numa segunda fase de experiências, procurou-se a identificação de grupos similares, para reduzir eventual ruído e testar o ajuste de modelos específicos para cada grande grupo de clientes. O algoritmo K-Means, com a distância de Manhattan, foi usado para a identificação de perfis. Para valores de K até 8, os *clusters* resultantes não pareciam suficientemente diferentes. Para K=9 e K=10 é reportado um resultado melhor, onde são reconhecidos quatro perfis diferentes, em relação ao *Churn*: perfil médio; clientes mais antigos e com consumos acima da média; clientes jovens, que não pagam eles próprios o prémio do seguro e estão inseridos num grupo; clientes jovens com gastos em saúde inferiores à média e que pagam o próprio seguro.

Com o objetivo de prever um valor binário, foram testados classificadores com LR, com árvores de decisão, com SVM e com redes neuronais. Os dados foram extraídos de um sistema real, de uma companhia de seguros, tendo sido organizados em conjuntos com 10 000 instâncias cada. O primeiro tem instâncias de *Churn* e Não-*Churn*, na proporção original do sistema. O segundo tem apenas casos de *Churn* e o terceiro tem apenas casos de Não-*Churn*. A partir destes, foram depois constituídos os conjuntos de treino e o conjunto de teste usado na avaliação de todas as variantes do modelo. Os modelos contemplam dados sociodemográficos, a indicação de área geográfica de residência, incluindo um atributo sobre a classificação urbano/rural. Adicionalmente, havia dados sobre interação cliente - empresa, incluindo queixas, contactos de marketing, tempos entre momentos de contacto, ou duração dos contratos. Sobre o produto contratado, existem ainda dados sobre o valor a pagar, método de pagamento, as ativações do serviço e são considerados também possíveis *switching barriers*, os dados sobre barreiras à mudança.

Foram pensados cinco modelos, em função da política de treino usada. No Modelo 1 o conjunto de treino inclui instâncias na proporção original das classes. O Modelo 2 foi treinado com uma distribuição 80% Não-*Churn* e 20% *Churn*. E os Modelos 3 a 5 usaram diferentes distribuições de casos entre classes, respetivamente, 70/30, 66/33 e 50/50. Não foram usadas técnicas de sobreamostragem ou subamostragem.

O tipo de algoritmo com melhor resultado na métrica AUC é árvores de decisão, ainda que o mesmo tenha valores relativamente baixos de AUC. Com base nestes valores, os autores sublinham a importância de serem consideradas várias métricas de avaliação. Em termos globais, o uso de redes neurais resultou num melhor desempenho para o Modelo 3, obtendo 0,73 em AUC. É referido que, para todos os modelos treinados com a distribuição de casos entre classes original, têm pior desempenho. Considerando os indicadores de Lift, o melhor resultado é obtido com LR e redes neurais, onde 50% dos casos de *Churn* seriam alcançados numa franja de 20% da população em causa.

Num trabalho de 2018, Eero Sahlberg trabalhou em modelação de *Churn*, na área dos seguros [8], procurando relacionar a mudança de morada com o abandono de serviços, como o seguro de casa e outros. Tendo começado com um conjunto de dados relativos a 50 000 clientes de uma seguradora finlandesa, a fase de pré-processamento filtrou casos sem alteração de morada durante o último seguro válido, de onde resultaram 24230 casos.

O autor descreve três experiências de modelação: *Churn* relativamente ao seguro da casa; *Churn* de todos os seguros subscritos, tomando como universo os casos de *Churn* do seguro de casa; e a modelação de *Churn* total a partir dos dados iniciais, dos 24230 clientes considerados.

Na análise de atributos relevantes, foram apurados o poder económico, a idade dos clientes e a sua antiguidade como clientes em seguros, com os clientes de mais longa data e mais velhos a terem menos propensão para o *Churn*.

Este é um projeto de modelação econométrica, com *Logistic Regression* a considerar 21 variáveis. Depois de uma análise ANOVA, foram apurados 11 preditores. Conclui-se que os produtos de seguros de carácter pessoal contribuem positivamente para a retenção.

Dos resultados da segunda modelação, o autor conclui que clientes mais velhos preferem ter todos os seguros num só fornecedor, e assim quando mudam o seguro de casa, é muito provável que alterem todos os restantes seguros. Ao contrário, nas camadas mais jovens, o *Churn* do seguro de casa não está tão frequentemente associado ao *Churn* da restante linha de seguros. No terceiro modelo, que funcionou como uma espécie de validação das conclusões anteriores, é efetuada uma análise de variância sobre 24 variáveis. São apurados 9 regressores, ou variáveis com relevância.

Os valores de AUC sobre a correção da predição binária para cada modelo são respetivamente 0,601, 0,793 e 0,722.

Na tese de mestrado de Kim Hendrikse, encontramos um relatório do trabalho realizado na predição de *Churn* na área dos seguros de saúde, numa empresa holandesa [19]. Os dados existentes incluem: caracterização sociodemográfica dos clientes; acessos via Web ao serviço; os momentos de contacto entre cliente e seguradora a propósito de pedidos ou reclamações (referidos como *touchpoints*); e *Net Promoter Score* (indicador resultante de questões como “De 1 a 10, como avalia a probabilidade de recomendar este serviço a um amigo?”).

Foram desenvolvidos modelos preditivos com o algoritmo *Logistic Regression*, que diferem no conjunto de características usadas no ajuste. A seleção de características foi realizada com técnicas baseadas na significância estatística, e outras, incluindo métodos *Wrapper* com *10 fold cross validation* para uma escolha que maximizasse o indicador AUC. O resultado da avaliação dos modelos indicou valores entre 0,706 e 0,807 para exatidão; entre 0,524 e 0,670 para AUC; entre 0,136 e 0,265 na precisão; entre 0,275 e 0,397 no *recall*; e entre 0,025 e 0,092 para o indicador AUK. O método de seleção de características que resultou no menor número de variáveis foi o método *Filter*, e foi também o que obteve melhor valor para exatidão. O método *Wrapper* apurou um elevado número de características. Ao mesmo tempo, levou ao valor mais elevado de AUC e AUK.

## 2.3 Outras áreas de negócio

Num artigo de junho de 2018, Simion-Constantinescu et al. propõem um método para predição e mitigação de *Churn*, que visa ir além do estado-da-arte e que recorre ao processamento em paralelo, incluindo GPUs<sup>4</sup> [3]. Com um vasto repositório de dados de um grupo farmacêutico, os autores descrevem vários modelos, desde a predição da probabilidade de *Churn*, até um modelo *time-to-event* para estimar o número de dias até determinado evento para um cliente. Os clientes foram segmentados em quatro grandes grupos, com base nas características RFM, desde os considerados “mais fracos” até aos “mais fortes”. Estes quatro segmentos foram posteriormente subdivididos em micro-clusters, para maior capacidade de ajuste dos modelos. Existe um repositório com dados de clientes (27% assinalados com a classe *Churn*) e um repositório com transações relativas a vários anos, qualquer um com milhões de registos.

Os modelos são treinados com redes neuronais profundas, através da ferramenta Keras<sup>5</sup>, um framework de alto nível para *Deep Learning* sobre TensorFlow<sup>6</sup>. A predição de *Churn* foi testada com uma rede neuronal *fully-connected*, com uma rede do tipo RNN com *Long-Short Term Memory* (LSTM), e com um *ensemble* que combina ambos os tipos de rede. A rede *fully-connected* foi usada com 2 camadas escondidas ou intermédias, onde a primeira delas tem número de unidades igual ao tamanho do input, e a segunda tem metade das unidades. A última configuração usada foi uma solução com RNN, com uma pilha de três camadas LSTM, com 256 unidades cada, seguida de uma rede *fully-connected*.

As medidas usadas para avaliação foram precisão e *recall*. Os resultados obtidos para os vários modelos apontam precisão entre 0,411 e 0,516, e *recall* entre 0,83 e 0,93.

Em 2017, Robert Aman desenvolveu um estudo para identificar parâmetros relevantes para a solidez do vínculo com os clientes, numa empresa de *Software-as-a-Service* em Estocolmo, no negócio de marketing digital para B2B [27]. O autor descreve a aplicação de processos estatísticos e técnicas de *Data Mining* a dados provenientes de inquéritos e relativos à utilização de software. Através de ferramentas de monitorização, captaram-se dados detalhados da utilização dos produtos por parte dos utilizadores relativos a cada empresa cliente. Os dados foram tratados com técnicas de análise ANOVA<sup>7</sup>, regressão linear, *logistic regression*. O modelo para classificação do estado do cliente foi avaliado em termos de exatidão, com a melhor variante a obter 0,745 nesse indicador.

Os resultados mostram padrões diferentes de utilização de software, entre clientes que se mantém ativos e casos em que há abandono, com uma maior utilização de software geralmente associada a maior propensão para continuar como cliente. É referido que o valor percecionado pelo cliente, pela utilização do produto, explica, ou está associado, à solidez do vínculo com o cliente. Se esse vínculo for frágil, haverá maior probabilidade de abandono.

Na sua tese de mestrado [29], Stef Koeslag estudou fatores associados à satisfação de clientes, e indiretamente à intensidade da ligação do cliente ao fornecedor, através de dados de comportamento sobre compra e sobre reclamações efetuadas, num contexto de B2B. Adicionalmente, foram também previamente considerados os dados demográficos de cada cliente. Depois de uma análise, os dados sobre reclamações e demográficos não foram selecionados

---

<sup>4</sup> *Graphics Processing Unit* - processador gráfico, passível de utilização para cálculo numérico massivo

<sup>5</sup> <https://keras.io/>

<sup>6</sup> biblioteca de software de código aberto para computação numérica e ML - <https://www.tensorflow.org/>

<sup>7</sup> <https://statistics.laerd.com/statistical-guides/one-way-anova-statistical-guide.php>

como preditores, por motivos de qualidade desses dados para as instâncias disponibilizadas. A modelação de *Churn* foi feita com *Logistic Regression* e 14 preditores, entre os quais alguns atributos do tipo RFM (recência, frequência e valor monetário dos gastos) do cliente. Destas variáveis, as associadas aos valores gastos revelaram pouca capacidade discriminatória entre classes. No extremo oposto, as variáveis com maior potencial de predição em análise multivariada foram a frequência de compras nos últimos seis meses e o número de dias desde a última compra. O classificador resultante foi avaliado com uma exatidão de 87,5%.

O autor apontou dificuldades relacionadas com a janela temporal dos dados, que foi considerada muito curta, o que somado à qualidade de dados em algumas variáveis dificultou um estudo com dados mais abrangentes.

Em 2017, Jana Van Haver realizou um trabalho de predição de *Churn*, como variável dependente binária, em contexto de *Business-to-Business* (B2B). Um particularidade desta área é a não existência de subscrição, ou de uma obrigação periódica entre comprador e vendedor. Isto significa que uma parte não tem de anunciar à outra a sua desistência relativamente ao produto ou serviço que lhe compra. Assim, é mencionado o foco em abandono parcial, como possível antecedente de *Churn* definitivo [16].

A base de dados fonte continha dados sobre transações efetuadas por cerca de 10 000 clientes, entre 2011 e 2016. A fatia dos casos de *Churn* é 25%, descrita como relativamente alta face a outros estudos.

As oito técnicas de aprendizagem supervisionada usadas incluem os algoritmos *Logistic Regression*, árvores de decisão, *Naive Bayes*, *Random Forest*, *Support Vector Machines*, *Bagging*, *Boosting* e redes neuronais. As métricas de avaliação empregues foram exatidão, sensibilidade, especificidade, medida F, AUC e TDL. Usa-se o método *3-fold cross validation*, mantendo a proporção entre classes, com 5 repetições.

Os melhores resultados foram conseguidos com os modelos treinados com os métodos *Boosting* e *Logistic Regression*, sendo reportado que nas experiências realizadas, estes métodos alcançaram um bom compromisso entre capacidade preditiva, interpretabilidade e reduzido custo computacional. Concretamente, o melhor valor de TDL foi 3,78, para os modelos desenvolvidos com os métodos *Boosting*, LR e SVM. O melhor valor de AUC foi 0,9448, obtido com o algoritmo *Boosting*, e com uma vantagem de aproximadamente 1% face aos modelos de LR e SVM. No indicador exatidão, os valores situaram-se entre os 82 e os 89%.

Na análise de relevância de atributos, as características relacionadas com recência são apontadas como muito importantes neste estudo.

Num trabalho na área dos jogos [4], o autor aponta como objetivo o colmatar da lacuna de dados sobre um jogo recente, quando se pretende construir um modelo preditivo para esse jogo. O trabalho procura um modelo independente das especificidades de um jogo particular, que possa usar-se entre jogos. Assim, foram selecionados dados relativos a atributos transversais a três jogos populares em dispositivos móveis, e independentes do design particular de cada tipo de jogo. É referida a captação de dados de 100.000 dispositivos por cada um dos três jogos, e onde tenha ocorrido pelo menos uma sessão de jogo. São considerados dados relativos à instalação do software, relativos ao momento e duração das sessões de jogo, e relativamente à pontuação obtida (ou similar) em cada sessão de jogo.

Foram usados os algoritmos KNN, árvores de decisão, *Random Forest* e *Logistic Regression*. O parâmetro K, para o algoritmo KNN, foi otimizado através de uma procura em grelha, associado a um processo de *10-fold cross-validation* para maximizar a medida exatidão.

A exatidão observada na avaliação dos diferentes modelos situou-se entre 0,773 e 0,816. O método *Random Forest* obteve o melhor valor para exatidão, com 0,816, mas com pouca di-

ferença para os demais modelos. Para o indicador AUC, os diversos modelos foram avaliados com resultados entre 0,825 e 0,878, com o melhor valor novamente para o algoritmo *Random Forest*. Foi mencionada a utilização da linguagem de programação Python, com alguns pacotes de software: PANDAS, NumPy, Scikit-learn e Matplotlib.

O trabalho de Kawale et al. incide sobre o risco de *Churn* em jogo online [17], em particular no jogo *EveryQuest II*. Os dados, obtidos via SONY Corporation, são relativos apenas a jogadores que jogam em rede, de forma conjunta. Alguns atributos utilizados foram a influência que o jogador tem em relação aos outros, podendo esta ser positiva ou negativa, o tempo de jogo, o número de *logins*, o tempo para completar algumas *quests*, os pontos recebidos, entre outras. Foram considerados clientes que fizeram *Churn* nos meses de Agosto, Setembro e Outubro de 2006.

A abordagem dos autores assenta na premissa de que a predição de *Churn* neste tipo de jogos pode fazer-se de modo eficaz, juntando elementos sobre o envolvimento que o jogador tem relativamente ao jogo com pistas sobre as influências trocadas entre esse jogador e a sua rede de parceiros de jogo. Foram estudados três tipos de modelos: *Simple Diffusion Model* (sobre a influência social entre jogadores); *Network and Player Engagement* (baseado nas características do jogador e do seu envolvimento no jogo); *Modified Diffusion Model* (combina relações de influência social com dados do envolvimento do jogador). Para verificar se um jogador vai desistir do jogo, esta abordagem foca-se principalmente no decréscimo de tempo jogado, e no número de jogadores que desistiu do jogo, de entre aqueles com quem o primeiro jogador já jogou.

Nos segundo e terceiro tipos de modelo, os autores utilizaram os algoritmos *AdaBoostM1*, *AD-tree*, *JRip*, *J48* e *NaiveBayes*. Foi com o terceiro tipo de modelo, *Modified Diffusion Model*, e com o classificador *AdaBoostM1* que foi alcançado o melhor valor para a precisão (50,1%), e o segundo melhor valor para recall (29,8%). Os autores concluem que a combinação de dados sobre a influência social e dados sobre jogadores e respetiva dinâmica de jogo são importantes para a predição de *Churn*.

Num trabalho conjunto de Lovisa Gronros e Ida Janér, de 2018, é analisada a viabilidade de utilização dos dados relativos às primeiras 24 horas de atividade de um utilizador, para prever a respetiva probabilidade de *Churn*, na indústria do jogo online [20]. A empresa que cedeu os dados constatou que os primeiros dois dias após o registo de um utilizador são determinantes para o conquistar, sendo que a maior parte das desistências acontece nesse período. Assim, surgiu a necessidade de prever rapidamente o risco de *Churn* para os utilizadores recém registados, logo após as primeiras horas.

Os dados disponíveis incluem atributos sobre o utilizador e atributos sobre as transações que realizou. Como cada instância descreve um jogador, as respetivas transações foram sumariadas em novos atributos, incluindo o primeiro jogo que experimentou, jogo ou canal mais visitado, juntamente com número de apostas feitas em intervalos de 6 horas, quantidades e valores envolvidos nas transações monetárias ou o tempo decorrido deste a última transação. Os primeiros modelos foram treinados com todas as variáveis e foram baseados nos algoritmos KNN, com diferentes valores de K, *Quadratic Discriminant Analysis* (QDA), e Naive Bayes e SVM. Os resultados indicaram valores muito baixos de exatidão, pelo que não foram usados. Adicionalmente, é referido o elevado tempo de treino para SVM, que dificultava a aplicabilidade do algoritmo no curto tempo disponível para casos reais.

Os modelos posteriores usaram os algoritmos *Logistic Regression*, *Random Forest* e LDA. A seleção de características envolveu uma técnica de eliminação recursiva, com testes envolvendo conjuntos alternados de atributos. Os dados originais possuíam 101 atributos, sendo

80 numéricos e 21 categóricos. Depois de um processo de codificação *one-hot encoding*<sup>8</sup> das variáveis categóricas, o número total de atributos subiu para 234. A avaliação indicou valores de exatidão de 0,7336 para LR, 0,7540 para RF, 0,7269 para LDA e 0,7595 para um método *ensemble* que inclui os três modelos anteriores com decisão em voto pesado, respetivamente, pelos valores 0,005, 0,80 e 0,015, e que obteve o melhor desempenho. É referido o uso da linguagem de programação Python e bibliotecas `Pandas` e `ScikitLearn`.

### 3 Discussão

Na secção anterior, foram enumerados diversos trabalhos sobre análise de dados e modelos preditivos para *Churn*. Talvez por ser um sector com grande crescimento e transformação nas últimas décadas, a área de telecomunicações é muito referida em trabalhos de aprendizagem automática relacionados com a qualidade do negócio, e isso foi evidente também neste estudo, como mostra a distribuição de trabalhos por sectores indicada na Tabela 1.

Recapitulando alguns elementos chave mencionados antes, a Tabela 2 mostra o algoritmo usado no melhor modelo testado em cada trabalho referido na secção anterior. Na terceira coluna, temos indicação da métrica mais relevante indicada em cada trabalho e o respetivo valor para o modelo de melhor desempenho com aquele método, referido pelo autor da referência.

A dispersão de valores no resultado de avaliação dos classificadores de *Churn* é notória. A Tabela 3 apresenta valores de avaliação com AUC e exatidão, organizados por medida e por ordem decrescente de desempenho. Constatamos que os níveis de desempenho variam bastante, em função dos dados em cada contexto.

Pela análise bibliográfica, não se observa que exista um algoritmo ou método globalmente ótimo para a predição de *Churn*, com excelente desempenho em todos os tipos de dados. Por outro lado, a existência de uma biblioteca de cálculo ou software avançado de ML não garante automaticamente bons resultados. Veja-se por exemplo um artigo de 2017, publicado em `DZone.com`<sup>9</sup>, onde o autor descreveu uma abordagem baseada em redes neuronais para a predição de *Churn* [9]. As experiências relatadas envolvem redes LSTM, com o objetivo de separar casos onde haverá alguma ação (abrir um e-mail, aceder a determinada área) sobre o sistema no mês seguinte, dos casos em que não há nenhuma ação, e que estão em elevado risco de *Churn*. A experiência tem resultados inconclusivos, não sendo apresentados valores de avaliação. E o autor menciona possíveis explicações, como a possível qualidade e quantidade dos dados. Este documento retrata a dificuldade de aplicação de mecanismos de ML a dados reais que podem carecer de uma representação especial, adaptada aos algoritmos em que são processados, ou que podem, por si só, ter problemas de qualidade, onde nem a melhor técnica

<sup>8</sup>Nas variáveis categóricas não ordinais, cada valor é transformado num atributo binário, que assume o valor 1 nas instâncias com aquele valor, ou 0 caso contrário.

<sup>9</sup>uma comunidade online sobre desenvolvimento de software

Sector	Referências
telecomunicações	[21] [1] [15] [14] [30] [32] [25] [22] [10] [33] [7] [12] [31] [2] [28]
financeiro e seguros	[16] [4] [29] [27] [3] [20]
B2B, jogos, farmácia	[6] [8] [19] [24] [11] [23] [17]

Tabela 1: Distribuição de referências analisadas por sector de atividade

Método do melhor modelo	Medida	Avaliação	Referência
AdaBoost	precisão	0,501	[17]
AD	exatidão	0,95	[22]
AD (C4.5)	AUC	0,984	[30]
AD (C4.5)	exatidão	0,938	[32]
AD, LR	AUC	0,839	[10]
Boosting	AUC	0,9448	[16]
ensemble	AUC	0,76	[33]
ensemble	exatidão	0,759	[20]
LDA	exatidão	0,74	[28]
LR	exatidão	0,875	[29]
LR	exatidão	0,745	[27]
LR	AUC	0,793	[8]
LR	exatidão	0,847	[23]
LR	AUC	0,67	[19]
LR, Boosting, RF	AUC	?	[7]
LR, Boosting, RF, RN e SVM	?	?	[2]
processo gaussiano, ensembles	AUC	0,77	[25]
RF	AUC	0,9654	[15]
RF	AUC	0,878	[4]
RF	AUC	0,93	[24]
RF	AUC	0,79	[11]
RN	exatidão	0,52	[1]
RN	precisão	0,85	[14]
RN	AUC	0,779	[12]
RN	AUC	0,743	[31]
RN	precisão	0,516	[3]
RN, LR	AUC	0,73	[6]
XGBoost	AUC	0,95	[21]

Tabela 2: Método e avaliação com melhor desempenho em cada trabalho referido

de ML terá eficácia.

## 4 Conclusão

Em geral, os trabalhos analisados, e referidos na secção 2, lidam com a deteção de possível *Churn* como uma tarefa de classificação binária (*Churn*, *Não-Churn*) a partir dos dados do histórico associado aos clientes. Variam os dados em cada caso, mas também o leque de métodos experimentados e o pré-processamento e critérios eventualmente empregues na seleção de características.

O *Churn* num serviço baseado em subscrição não é analisado do mesmo modo que num serviço sem subscrição nem obrigação contratual do cliente informar a empresa sobre a sua intenção de cessar atividade [16]. Nos últimos, é mais difícil determinar se houve abandono ou é apenas uma fase de inatividade, o que é comum em áreas como a venda a retalho.

Como referido em [26], na área de SaaS o acompanhamento da satisfação e nível de fidelização

Medida	Avaliação	Método do melhor modelo	Referência
AUC	0,9840	AD	[30]
AUC	0,9654	RF	[15]
AUC	0,9500	XGBoost	[21]
AUC	0,9448	Boosting	[16]
AUC	0,9300	RF	[24]
AUC	0,8780	RF	[4]
AUC	0,8390	AD, LR	[10]
AUC	0,7930	LR	[8]
AUC	0,7900	RF	[11]
AUC	0,7790	RN	[12]
AUC	0,7700	processo gaussiano, ensembles	[25]
AUC	0,7600	ensemble	[33]
AUC	0,7430	RN	[31]
AUC	0,7300	RN, LR	[6]
AUC	0,6700	LR	[19]
exatidão	0,9500	AD	[22]
exatidão	0,9380	AD	[32]
exatidão	0,8750	LR	[29]
exatidão	0,8470	LR	[23]
exatidão	0,7590	ensemble	[20]
exatidão	0,7450	LR	[27]
exatidão	0,7400	LDA	[28]
exatidão	0,5200	RN	[1]

Tabela 3: Variabilidade de resultados da avaliação de desempenho em dois indicadores

de clientes e a medição da efetiva utilização que fazem dos serviços são importantes para validar modelos de precificação personalizados. Os clientes com propensão para *Churn* poderão então beneficiar de medidas especiais, como descontos, ofertas ou privilégios que reduzam o risco de abandono do serviço. Perante um elevado número de clientes nestas condições, a estratégia de retenção pode ser estratificada com base em prioridades, em função não apenas do risco apurado de *Churn* em cada caso, mas também do valor que o cliente tem para o negócio. Este valor do cliente é estimado a partir de receitas do presente, acrescidas de receitas potenciais ou futuras. Este valor pode ainda influenciar o esforço ou intensidade da ação de fidelização, como por exemplo determinar um valor limite para o desconto a oferecer [13]. A ponderação de vários indicadores, oriundos de modelos preditivos do risco de *Churn*, do peso do cliente, ou outras prioridades, pode ser realizada por Sistemas de Recomendação, que poderão auxiliar tarefas na área de CRM. Estes sistemas poderão ajudar à implementação de uma política de preços e níveis de serviço com carácter mais dinâmico, com interesse para cliente e fornecedor.

## Agradecimentos

Este artigo relata trabalho desenvolvido no âmbito do projeto “*Admin Portal & Reporting Analytics for Cloud Providers resellers - V2*” (APRA-CP.v2), com a referência ALT20-03-0247-FEDER-038500, apoiado pelo Programa Operacional Regional do Alentejo 2014/2020.

## Referências

- [1] Abbasimehr, H., Mostafa, S., Soroor, J. (2012). *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
- [2] Aditya Kapoor (2017). *Churn in the Telecom Industry – Identifying customers likely to churn and how to retain them*. Technical report. <https://wp.nyu.edu/adityakapoor/>
- [3] Andrei Simion-Constantinescu et al. (2018). *Deep Neural Pipeline for Churn Prediction and Mitigation*.
- [4] Bastiaan van der Palen (2017). *Predicting player churn using game-design-independent features across casual free-to-play games*. Master’s thesis. Tilburg University
- [5] Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). *SMOTE : Synthetic Minority Over-sampling Technique*. Journal of Artificial Intelligence Research, 16:321–357
- [6] Chantine Huigevoort (2015). *Customer churn prediction for an insurance company*. Master’s thesis. Eindhoven University of Technology
- [7] D.S. de Groot (2017). *Churn prediction in telecommunication: Classification problem*. Master’s Thesis. Delft University of Technology
- [8] Eero Sahlberg (2018). *Modeling customer churn among movers in the Finnish insurance market*. Faculty of Social Sciences, University of Helsinki - <https://helda.helsinki.fi/handle/10138/233797>
- [9] Eleni Markou (2017). *Recurrent Neural Networks for Email List Churn Prediction*. Technical report. <https://dzone.com/articles/recurrent-neural-networks-for-email-list-churn-pre-1>
- [10] Emília Nguyen (2011). *Customer Churn Prediction for the Icelandic Mobile Telephony Market*. Master’s Thesis. Faculty of Industrial Engineering, Mechanical Engineering and Computer Science, University of Iceland - <http://hdl.handle.net/1946/10159>
- [11] Erdem Kaya et al. (2018). *Behavioral Attributes and Financial Churn Prediction*. EPJ Data Science, vol. 7, no. 1
- [12] Federico Castanedo et al. (2014). *Using Deep Learning to Predict Customer Churn in a Mobile Telecommunication Network*
- [13] García, D.L., Nebot, À. & Vellido, A. (2017). *Intelligent Data Analysis Approaches to Churn as a Business Problem: a Survey*. Knowledge and Information Systems 51: 719 <https://link.springer.com/article/10.1007/s10115-016-0995-z>
- [14] Garcia Gomez, D. (2014). *Exploration of customer churn routes using machine learning probabilistic models*. Tesi doctoral, Universitat Politècnica de Catalunya, Departament de Llenguatges i Sistemes Informàtics. Available at: <http://hdl.handle.net/2117/95309>
- [15] Georgina Cunha Esteves (2016). *Churn prediction in the telecom business*. Tese de Mestrado. Faculdade de Engenharia da Universidade do Porto
- [16] Jana Van Haver (2017). *Benchmarking analytical techniques for churn modelling in a B2B context*. Master’s Dissertation. Faculteit Economie en Bedrijfskunde.
- [17] Jaya Kawale, A. Pal and J. Srivastava (2009). *Churn prediction in MMORPGs: A social influence based approach*. In Proceedings of the 12th IEEE International Conference on Computational Science and Engineering, 2009, Canada.

- [18] José Saias, M. Maia, L. Rato and T. Gonçalves (2018). *Machine Learning: um estudo sobre conceitos, tarefas e algoritmos relacionados com previsão e recomendação*. Relatório Técnico do Projeto APRA-CP
- [19] Kim Hendrikse (2017). *Predicting customer churn in the healthcare insurance market: a case study*. Master's Thesis. Eindhoven University of Technology
- [20] Lovisa Grönros and Ida Janér (2018). *Predicting Customer Churn Rate in the iGaming Industry using Supervised Machine Learning*. Master's thesis. KTH Royal Institute of Technology
- [21] Meeke Rijnen (2018). *Predicting Churn using Hybrid Supervised-Unsupervised Models*. Master's thesis. Leiden University
- [22] Naren Tanneedi (2016). *Customer Churn Prediction Using Big Data Analytics*. Master's Thesis. Faculty of Computing, Blekinge Institute of Technology
- [23] Nie, G., Rowe, W., Zhang, L., Tian, Y., & Shi, Y. (2011). *Credit card churn forecasting by logistic regression and decision tree*. *Expert Syst. Appl.*, 38, 15273-15285.
- [24] Nina Olofsson (2017). *A Machine Learning Ensemble Approach to Churn Prediction – Developing and Comparing Local Explanation Models on Top of a Black-Box Classifier*. Master's Thesis. KTH Royal Institute of Technology
- [25] Ole-Christofer Granmo (2013). *A Churn Prediction Model Based on Gaussian Processes*. Master Thesis, Faculty of Engineering and Science, University of Agder
- [26] PayPro Global Inc. (2014). *Tackling SaaS Churn*. <http://docs.payproglobal.com/documents/white-papers/PayPro-WP-Tackling-SaaS-Churn.pdf>
- [27] Robert Åman (2017). *Understanding when customers leave: Defining customer health and how it correlates with software usage*. Master's Thesis. Uppsala University
- [28] Sowmya Vivek (2018). *Using Linear Discriminant Analysis to Predict Customer Churn*. Disponível em: <https://www.datascience.com/blog/predicting-customer-churn-with-a-discriminant-analysis>
- [29] Stef Koeslag (2016). *Prediction of partial churners and behavioural loyal customers through behavioural historical customer data*. Master's Thesis. University of Twente
- [30] Stephen Nabareseh (2017). *Predictive analytics: a data mining technique in customer churn management for decision making*. Doctoral Thesis. Tomas Bata University
- [31] Wangperawong A., Brun C., Laudy O., Pavasuthipaisit R. (2016). *Churn analysis using deep convolutional neural networks and autoencoders*. arXiv:1604.05377
- [32] Wouter Verbeke (2012). *Profit driven data mining in massive customer networks: new insights and algorithms*. PhD thesis. Katholieke Universiteit Leuven
- [33] Yue Cui (2018). *Customer Relationship Management using Ensemble Methods*. Master's Thesis. University of Toronto