2018

Performance in the Prediction of Dropout using the Machine Learning in Sport Services

Pedro Sobreiro
ESDRM-IPSantarem, sobreiro@esdrm.ipsantarem.pt

Paulo Pinheiro
Universidade Aberta, ppinheiro@cedis.pt

Abel Santos
ESDRM-IPSantarem, abelsantos@esdrm.ipsantarem.pt

Desempenho na Previsão do Abandono Recorrendo ao Machine Learning em Serviços de Desporto

Performance in the Prediction of Dropout using the Machine Learning in Sport Services

Pedro Sobreiro, ESDRM-IPSantarem, Portugal, sobreiro@esdrm.ipsantarem

Paulo Pinheiro, Universidade Aberta, Portugal, ppinheiro@cedis.pt

Abel Santos, ESDRM-IPSantarem, Portugal, abelsantos@esdrm.ipsantarem.pt

Resumo

O abandono de um cliente é um fenómeno que ocorre com frequência em clientes de serviços de desporto. Este estudo, pretende avaliar o desempenho das técnicas baseadas em modelos de aprendizagem com dados, a partir dos dados disponíveis dos clientes e o histórico de utilização dos serviços de desporto, para a realização da previsão do abandono. Foram aplicadas várias técnicas para realizar a previsão, de forma a identificar qual apresentava maior exatidão, bem como uma comparação da exatidão da previsão recorrendo a uma aproximação baseada em train/test e k-fold. O algoritmo com melhor desempenho é o *Gradient Boosting Classifier*, em ambas as aproximações, apesar do tempo de execução ser elevado. Os resultados obtidos indiciam que não existem diferenças significativas entre a exatidão da previsão e do tempo de execução entre as técnicas utilizadas em train/test e k-fold com $p>0.05$ recorrendo ao Mann-Whitney.

Palavras-chave: Previsão abandono; *Machine Learning*; Serviços de desporto; Gestão do desporto

Abstract

The customer dropout is a phenomenon that often occurs in customers of sports services. This study intends to evaluate the performance of machine learning algorithms to predict dropout, using the available data of the customers and the history of the use of sports services to predict dropout. Several techniques were applied to perform the prediction to identify which was more accurate, as well the comparison of the accuracy of the prediction using a train / test and k-fold approach. The best performance was achieved with the algorithm Gradient Boosting Classifier in both approaches, although the runtime is high. The results show that no significative differences in the accuracy of the prediction and execution time of the techniques used in train / test and k-fold with p> 0.05 using Mann-Whitney.

Keywords: *Dropout prediction; Machine Learning; Sports services; Sport management*

1. INTRODUÇÃO

A retenção de clientes é um problema transversal a várias áreas da atividade económica e que exige a exploração de abordagens inovadoras para as organizações ganharem vantagens competitivas e reduzirem o abandono. O setor de prestação de serviços desportivos apresenta taxas

ISSN 2183-489X

de desistência elevadas (Avourdiadou & Theodorakis, 2014), situação que não é diferente em Portugal onde a taxa de retenção se situa em 31% nos Ginásios e Academias (Associação de Empresas de Ginásios e Academias de Portugal, 2016). Emeterio, Iglesias-Soler, Gallardo, Rodriguez-Cañamero e García-Unanue (2016) referem que no primeiro ano de utilização dos ginásios a retenção é apenas de 50%.

A rentabilidade nos ginásios é um aspeto fundamental (Ferrand, Robinson, & Valette-Florence, 2010), onde a retenção tem muita importância (Hurley, 2004; MacIntosh & Law, 2015). Os menores custos associados à manutenção de um cliente, em relação à angariação de novos clientes (Ahmad & Buttle, 2002), demonstra a importância que a retenção têm. A retenção, de acordo com Bodet (2012), representa a intenção de um cliente voltar a comprar e manter-se membro do ginásio. Apesar da retenção no consumo de serviços de desporto ser associada à satisfação do cliente e qualidade das instalações (Avourdiadou & Theodorakis, 2014; Howat & Assaker, 2016), a previsão poderá permitir aos responsáveis das instalações desportivas desenvolverem ações para contrariar o abandono, uma vez que evitar a saída é mais rentável do que angariar novos clientes (Edward & Sahadev, 2011).

A eficiência na previsão do abandono está relacionada com a seleção das variáveis preditoras (Hall, 1998), que podem ser suportadas em estudos existentes. Cooil, Keiningham, Aksoy e Hsu (2007) consideram o género relevante para previsão do abandono, aspeto reforçado com a identificação da diferença entre géneros na maior realização do exercício físico (Pridgeon & Grogan, 2012). Pawlowski, Breuer, Wicker e Poupaux (2009) consideram também, para além da idade, o rendimento mensal que surge como um elemento relevante para a compreensão do consumo desportivo. A identificação das rotinas, na realização da prática desportiva, é outro aspeto fundamental para a manutenção do cliente (Ferrand et al., 2010; Pridgeon & Grogan, 2012), aumentando até a disponibilidade dos clientes para realizarem deslocações maiores, de forma a poderem utilizarem determinadas instalações desportivas (MacIntosh & Law, 2015). Apesar da importância das variáveis consideradas relevantes, a base de dados existentes na organização limita a sua seleção.

Machine learning é um processo automático que extrai padrões dos dados (Kelleher, Namee, & D'Arcy, 2015), que nos permite antecipar acontecimentos para desenvolvermos ações para os contrariar. O *machine learning* pode ser utilizado para suportar o desenvolvimento de estratégias de retenção de acordo com os dados existentes (Verbeke, Martens, Mues, & Baesens, 2011). Os dados disponíveis são utilizados treinar um modelo, com o intuito de se realizar generalizações (Domingos, 2012).

O objetivo deste estudo é prever o abandono dos clientes, recorrendo a várias técnicas de *machine learning* utilizando parte dos dados para treinar o modelo e os restantes dados para testar o

modelo, comparando o seu desempenho com a utilização do *k-fold*. No final comparamos o desempenho na precisão das várias técnicas utilizadas na aproximação treinar/testar e *k-fold* considerando o tempo a precisão.

2. TÉCNICAS PARA PREVISÕES

A previsão do abandono é um problema de classificação, abandona ou não abandona. Utiliza um conjunto de técnicas para aprenderem a modelar a relação entre um conjunto de características descritivas e uma característica alvo (Kelleher et al., 2015). Para realizar a avaliação do desempenho da classificação utiliza-se normalmente a exatidão da previsão através do cálculo do número de previsões corretas em relação ao número total de previsões, que depende da identificação do algoritmo adequado e na divisão dos dados em treino e teste do modelo (Kotsiantis, Zaharakis, & Pintelas, 2007).

2.1. Algoritmos para previsão

Os algoritmos foram analisados de acordo com a disponibilidade da biblioteca *scikit-learn* (Pedregosa et al., 2011), que foi selecionada considerando o desempenho, disponibilidade de documentação e variedade de algoritmos disponibilizados, de seguida descrevemos os algoritmos utilizados para realização de previsões.

O algoritmo *k-nearest neighbors (KNeighborsClassifier)* é um método não paramétrico utilizado para classificação e regressão (Altman, 1992). *KNeighborsClassifier* implementa um modelo de aprendizagem baseado nos k vizinhos próximos de cada ponto, onde k representa um valor especificado pelo utilizador.

Support vector networks (SVC) é um algoritmo que pode ser utilizado para classificação e regressão, utilizando dados associados a cada uma das categorias que pretendemos prever, a técnica que tenta encontrar um hiperplano linear ótimo de forma que margem de separação entre os elementos positivos e negativos é maximizada (Coussement & Van den Poel, 2008). Os dados são representados por pontos num espaço de forma que os dados que pertencem a cada uma das categorias sejam divididos por um intervalo que seja maior quanto possível. Durante a previsão são enquadrados no espaço criado e colocados numa categoria de acordo com o lado do hiperplano onde sejam colocados. O algoritmo original foi proposto por Vapnik em 1963 construindo um classificador linear, mais tarde tornou-se possível criar classificadores não linear através do *kernel trick* (Boser, Guyon, & Vapnik, 1992).

A *logistic regression* (LogisticRegression) é uma técnica de classificação muito utilizada para prever uma variável dicotómica dependente (Coussement & Van den Poel, 2008), considera as outras variáveis que estamos a utilizar para prever como independentes. As variáveis dependentes

são normalmente rotuladas como 0 ou 1, representando abandona/não_abandona. Apresenta como vantagem a sua simplicidade de interpretação face a outras técnicas como *neural networks* (Coussement & Van den Poel, 2008). Esta técnica apresenta a probabilidade de obter um resultado e escolhe um valor de corte para classificar as variáveis independentes quando superiores ou inferiores a este valor. O algoritmo foi desenvolvido inicialmente por Cox (1958).

Decision Tree Classifier (DecisionTreeClassifier) algoritmo que permite criar um modelo que prediz o valor de uma variável baseado em várias variáveis, aplicado a variáveis dependentes que têm um conjunto de valores finitos (Loh, 2011). A *decision tree* consiste na utilização de um nó inicial, nós interiores e nós folha (nós que terminam), os nós estão ligados por ramos, onde cada nó raiz ou interior especifica o teste a ser realizado numa variável dependente especificando o valor previsto na variável que se pretende prever (Kelleher et al., 2015).

Gaussian Naive Bayes (GaussianNB) implementa o classificador Naïve Bayes, que calcula a probabilidade de um determinado input pertencer a uma classe. O classificador é uma probabilidade simples de classificação que utiliza o teorema de Bayes e pressupostos de independência para prever o classe de uma instância (Milošević, Živić, & Andjelković, 2017).

Random Forest Classifier (RandomForestClassifier) são algoritmos de aprendizagem de dados para classificação e regressão que constroem uma multiplicidade de árvores de decisão durante a aprendizagem dos dados gerando uma classe de classificação (Ho, 1995). A simplicidade na interpretação deste modelo contribui para a sua popularidade (Coussement & Van den Poel, 2008). Segundo Gama, Oliveira, Lorena, Faceli e Carvalho (2017) utiliza um conjunto de preditores aleatoriamente selecionados para crescer cada árvore, o que origina várias árvores, já que os preditores selecionados para cada árvore são menores do que o número de variáveis. As árvores obtidas são utilizadas para calcular uma classificação considerando todas as árvores, esta aproximação reduz a instabilidade, já que os modelos em árvore são mais instáveis (Gama, Oliveira, Lorena, Faceli, & Carvalho, 2017).

Gradient Boosting Classifier (GradientBoostingClassifier) modelo iterativo que utiliza métodos de previsão recorrendo a vários *weak learners* de uma forma iterativa, normalmente árvores de decisão (Milošević et al., 2017). Cada iteração associa um peso diferente a cada exemplo do conjunto de treino, gerando vários classificadores em cada iteração, o que reduz o peso dos bem classificados e aumenta os *weak learners* (Gama et al., 2017). Na última etapa, o classificador final agrega os vários classificadores em que cada um obtém um peso em função da sua precisão, onde a principal vantagem deste algoritmo está na redução da variabilidade e variância das hipóteses individuais (Gama et al., 2017).

2.2. Treinar e testar o modelo

A previsão e teste do modelo baseia-se em treinar o modelo e testar a precisão do modelo treinado. Normalmente realiza-se a divisão de dados em dados treino e teste ou utilização de validação cruzada. A proporção recomendada para dividir os dados é de aproximadamente 2/3 para treinar e os dados restantes para testar (Kotsiantis et al., 2007), verificando-se uma utilização generalizada na literatura de uma proporção de 70/30% (Buntine & Niblett, 1992). Outra alternativa à proporção de dados para treinar/testar, é a validação cruzada é uma técnica estatística que divide os dados treino e teste, recorrendo ao *k-fold* (Refaeilzadeh, Tang, & Liu, 2016), onde dos dados são divididos em k grupos de dimensões iguais, onde se realiza k iterações para treinar e testar, o que obriga a que se treine o modelo k vezes, o que se torna dispendioso em termos de desempenho em modelos complexos e *datasets* maiores (Swersky, Snoek, & Adams, 2013).

O k-fold é uma técnica amplamente utilizada para estimar a generalização do erro de modelos de *machine learning*, que requer treinar os modelos k-vezes, o que se torna exigente em modelos complexos e *datasets* de maior dimensão (Swersky et al., 2013). Esta abordagem utiliza todos os dados disponíveis como treino e teste, testando com a fração de dados correspondente a 1/k-vezes.

3. METODOLOGIA

Neste estudo utilizamos os dados de 8376 clientes de uma instalação desportiva de Lisboa, obtidos através da aplicação informática e@sport durante o período de tempo de 01 de Junho de 2014 até 31 de Outubro de 2017, correspondendo à informação armazenada através da aplicação informática na adesão do clientes, histórico de pagamentos e acessos aos serviços disponibilizados. O tratamento de dados foi realizado com o Anaconda e IPython (Continuum Analytics, 2016), recorrendo ao Pandas (McKinney & others, 2010) e NumPy (Walt, Colbert, & Varoquaux, 2011).

Atendendo aos objetivos e ao problema que se pretende resolver, optou-se por utilizar a metodologia CRISP-DM (CRoss Industry Standard Process for Data Mining) (Wirth & Hipp, 2000), que consiste no standard mais utilizado devido à flexibilidade da sua implementação em qualquer domínio e à compatibilidade com qualquer ferramenta de *data mining*. A metodologia CRISP-DM contempla seis fases flexíveis, tendo sido desenvolvidos os seguintes passos (Wirth & Hipp, 2000): compreensão do negócio; compreensão dos dados; preparação dos dados, modelação, avaliação e implementação.

A compreensão do negócio permitiu identificar a necessidade de prever se um cliente está ou não em risco de abandono, permitindo o desenvolvimento de um programa de retenção aos clientes que tenham um risco maior.

A compreensão e preparação dos dados têm como objetivo a obtenção dos dados, a familiarização e a sua preparação, permitindo identificar problemas com a qualidade dos dados. O primeiro passo

envolveu a seleção das variáveis consideradas relevantes no contexto do problema, recomendações da literatura e disponibilidade dos dados. Depois procedemos ao tratamento dos dados, transformação e preparação, o que envolveu a deteção de *outliers* ou *missing values* por variável. As variáveis utilizadas no estudo foram selecionadas de acordo com os dados disponíveis na base de dados e considerando simultaneamente a sua relevância para a previsão do abandono. Os atributos extraídos encontram-se representados na Tabela 1. Por último, também foi associada uma classificação a cada cliente, representando se tinha abandonado ou não a instalação desportiva até momento que os dados foram extraídos (classe_desistencia).

As previsões foram realizadas recorrendo à biblioteca *scikit-learn*, onde testamos oito algoritmos em duas fases realizando a classificação do abandono. Os algoritmos utilizados foram: *k nearest neighbors classifier* (KNeighborsClassifier); *C-Support Vector Classification* (SVC); *Logistic Regression* (LogisticRegression); *Decision Tree Classifier* (DecisionTreeClassifier); *Gaussian Naive Bayes* (GaussianNB); *Random Forest Classifier* (RandomForestClassifier) e *Gradient Boosting Classifier* (GradientBoostingClassifier). Os atributos extraídos, exceto classe_desistencia (atributo alvo) foram utilizados para treinar o modelo com o objetivo de prever classe de desistência. A previsão foi realizada através da utilização dos dados que não foram utilizados para treinar para prever a classe_desistencia utilizando o modelo treinado. A exatidão da previsão baseou-se na confrontação da classe_desistencia prevista em relação à classe_desistencia real. Os algoritmos foram executados sequencialmente utilizando os parâmetros por defeito do *scikit-learn*.

A avaliação teve como objetivo analisar se os resultados obtidos cumprem os objetivos propostos. Foram analisados os algoritmos utilizados para identificar o algoritmo que tem o melhor desempenho no contexto do problema. A análise do desempenho baseou-se na utilização da matriz de confusão e determinação da exatidão da previsão, realizando a aprendizagem com 70% dos dados para treinar o modelo e 30% para testar a previsão, correspondendo a 5863 cliente para treino e 2513 para testar. A análise do desempenho também foi realizada utilizando uma validação cruzada através do *k-fold*, com um k=10.

A implementação será baseada na utilização do algoritmo que apresenta o melhor desempenho, permitindo a identificação do risco de abandono por cliente a partir dos dados obtidos no sistema de informação. O risco permite a realização de contramedidas para contrariar o risco de abandono e tentar reduzir a taxa de abandono.

Tabela 1 – Variáveis utilizadas.

Variável	Definição
Idade	Representa a idade do cliente e foi calculada a partir da data de nascimento existente nos dados dos clientes
Genero	Representa o género do cliente

Diassemfrequencia	O número de dias sem frequentar as instalações à data de 31 de outubro de 2017
Mesesinscricao	Tempo de inscrição do cliente em meses
Volnegocios	Valor faturado ao cliente durante os meses que esteve inscrito
Freqmedia	Frequência média semanal do cliente
utilizacao_livre	Variável que indica se o cliente esteve inscrito ou não uma utilização livre da instalação
atividade_aquaticas	Variável que indica se o cliente esteve inscrito ou não atividade aquática na instalação
atividade_fitness	Variável que indica se o cliente esteve inscrito ou não atividade de fitness na instalação
Natividades	Número de atividades que o cliente realizava
Nfrequencias	Número total de vezes que o cliente frequentou a instalação durante a sua inscrição
Freqcontratadasemanal	Número de vezes que o cliente que o cliente contratou para realizar
Nrenovações	Número de vezes que o cliente renovou o contrato
Nreferencias	Número de clientes referências que o cliente indicou
classe_desistencia	Indica se o cliente abandonou ou não

4. RESULTADOS

A exatidão da previsão de cada algoritmo foi calculada através da confrontação da previsão realizada face à situação real do cliente, se abandonou ou não abandonou a instalação desportiva. Estes valores foram obtidos através da matriz de confusão que representa os *True Positive* (TP - Não abandonou com resultado previsto de não abandono), *True Negative* (TN - Abandonou com resultado previsto de abandonar), *False Positive* (FP - Não abandonou com resultado previsto de abandonar), *False Negative* (FN - Abandonou com resultado previsto de não abandonar). A precisão da previsão calculada com TP/(TP+FP) referente aos clientes que não abandonaram. A precisão de cada algoritmo e tempo de execução está representada no Tabela 2. O melhor desempenho na exatidão da previsão na divisão de treino e teste foi obtido com o algoritmo *Gradient Boosting Classifier* (GBC) com uma precisão de 0.93, apesar de em termos de tempo de execução ter tido o segundo pior desempenho, o algoritmo mais pesado em termos da duração da sua execução nesta fase foi *Support vector networks*. Na utilização do k-fold o *Gradient Boosting Classifier* manteve o melhor desempenho com uma precisão na previsão de 0.926 e o segundo pior tempo de execução. O algoritmo SVC piorou muito em termos de tempo de execução.

Tabela 2 – Comparação da taxa de sucesso das previsões realizadas.

Algoritmo	Exatidão TT	Tempo TT	Exatidão k-fold	Tempo k-fold
KNN	0.879179	0.219	0.880611	0.323
SVC	0.788443	6.215	0.780919	67.686
LR	0.851480	0.147	0.853630	1.140

DT	0.888730	0.060	0.894583	0.519
GNB	0.781757	0.010	0.795128	0.076
RFC	0.924069	0.242	0.916551	0.943
GBC	0.933142	0.797	0.926700	6.065

Relativamente à precisão obtida nos que abandonaram (PrecisãoA) e não abandonaram (PrecisãoNA) na abordagem treinar/testar o KNeighborsClassifier apresentou uma precisão de 0.88 nos clientes que abandonaram e 0.95 nos que não abandonaram, o SVC de 0.75 nos clientes que abandonaram e 0.91 nos que não abandonaram. LogisticRegression de 0.93 nos clientes que abandonaram e 0.79 nos que não abandonaram; DecisionTreeClassifier de 0.72 nos clientes que abandonaram e 0.88 nos que não abandonaram; GaussianNB de 0.75 nos clientes que abandonaram e 0.93 nos que não abandonaram; RandomForestClassifier de 0.51 nos clientes que abandonaram e 0.86 nos que não abandonaram e GradientBoostingClassifier de 0.82 nos clientes que abandonaram e 0.94 nos que não abandonaram. Na Tabela 3 está representado os valores obtidos.

Tabela 3 – Comparação da taxa de sucesso das previsões realizadas.

Algoritmo	PrecisãoA	PrecisãoNA	Exatidão TT
KNN	0.88	0.95	0.879179
SVC	0.75	0.91	0.788443
LR	0.93	0.79	0.851480
DT	0.72	0.88	0.888730
GNB	0.75	0.93	0.781757
RFC	0.51	0.86	0.924069
GBC	0.82	0.94	0.933142

Após a comparação dos resultados da precisão e tempo de execução de *k nearest neighbors classifier*; *C-Support Vector Classification*; *Logistic Regression*; *Decision Tree Classifier*; *Gaussian Naive Bayes*; *Random Forest Classifier* e *Gradient Boosting Classifier* comparamos os testes realizados com *Mann-Whitney*. O teste de Mann-Whitney em relação à exatidão train/test (mediana = 0.865) e k_fold (mediana = 0.868) permite-nos concluir que não existiram diferenças significativas, U=21, p=0.35. Os resultados obtidos no teste de Mann-Whitney em relação ao tempo de execução no train/test (mediana = 0.219) e k_fold (mediana = 0.943) também não apresenta diferenças significativas, U=12, p=0.06, apesar tempo de execução ser superior no k_fold.

5. DISCUSSÃO E CONCLUSÃO

Os resultados obtidos apresentam um desempenho fraco do algoritmo *Support Vector Networks*, quer em termos de precisão ou em tempo de execução. O que nos permite excluir o algoritmo se pretendermos prever o abandono, contudo este fraco desempenho poderá estar associado a um

fraco balanceamento dos dados (Coussement & Van den Poel, 2008), i.e. número de pessoas que abandonaram versus clientes que não abandonaram. O bom desempenho demonstrado pelo *Gradient Boosting Classifier* pode ser encontrado noutros estudos (e.g. Milošević, Živić, & Andjelković (2017)) onde apresenta uma boa performance face a outros algoritmos. Outro aspeto a considerar é a precisão do *Logistic Regression* de 0.93 nos clientes que abandonaram e 0.79 nos que não abandonaram, isto poderá indiciar que o modelo poderá ter um comportamento eficiente na previsão do abandono, contudo um melhor balanceamento da amostra dos clientes que abandonaram versus não-abandonou poderá ser importante para ser desenvolvido uma melhor análise.

Na comparação do desempenho não encontramos diferenças significativas nos resultados obtidos nos algoritmos utilizando 70/30% para treinar e testar a previsão em relação ao k-fold, o que reflete a pouca diferença em termos resultados, considerando a precisão e tempo de execução. O algoritmo baseado no *Gradient Boosting Classifier* apresentou um bom desempenho em termos precisão, contudo se tivermos necessidade de um algoritmo com melhor desempenho em termos de tempo de execução, devemos considerar o *Random Forest Classifier* para a previsão do abandono.

O trabalho futuro a desenvolver poderá ser a aplicação de outros modelos como *Neural Networks* e *Deep Neural Networks* para avaliarmos a sua performance em termos de precisão na previsão e melhorar o balanceamento dos dados na proporção dos que abandonaram versus não abandonaram.

REFERÊNCIAS

Ahmad, R., & Buttle, F. (2002). Customer retention management: a reflection of theory and practice. *Marketing Intelligence & Planning*, *20*(3), 149–161. doi:10.1108/02634500210428003

Altman, N. S. (1992). An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. *The American Statistician*, *46*(3), 175–185. doi:10.1080/00031305.1992.10475879

Associação de Empresas de Ginásios e Academias de Portugal. (2016). Barómetro 2016.

Avourdiadou, S., & Theodorakis, N. D. (2014). The development of loyalty among novice and experienced customers of sport and fitness centres. *Sport Management Review*, *17*(4), 419–431. doi:10.1016/j.smr.2014.02.001

Bodet, G. (2012). Loyalty in Sport Participation Services: An Examination of the Mediating Role of Psychological Commitment. *Journal of Sport Management*, *26*(1), 30–42.

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A Training Algorithm for Optimal Margin Classifiers. Em *Proceedings of the Fifth Annual Workshop on Computational Learning Theory* (pp. 144–152). New York, NY, USA: ACM. doi:10.1145/130385.130401

Buntine, W., & Niblett, T. (1992). A further comparison of splitting rules for decision-tree induction. *Machine Learning*, *8*(1), 75–85. doi:10.1007/BF00994006

Continuum Analytics. (2016). Anaconda Software Distribution. Obtido 20 de Julho de 2017, de https://www.anaconda.com/download/

Cooil, B., Keiningham, T. L., Aksoy, L., & Hsu, M. (2007). A Longitudinal Analysis of Customer Satisfaction and Share of Wallet: Investigating the Moderating Effect of Customer Characteristics. *Journal of Marketing*, *71*(1), 67–83. doi:10.1509/jmkg.71.1.67

Coussement, K., & Van den Poel, D. (2008). Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques. *Expert Systems with Applications*, *34*(1), 313–327. doi:10.1016/j.eswa.2006.09.038

Cox, D. R. (1958). The Regression Analysis of Binary Sequences. *Journal of the Royal Statistical Society. Series B (Methodological)*, *20*(2), 215–242.

Domingos, P. (2012). A Few Useful Things to Know About Machine Learning. *Commun. ACM*, *55*(10), 78–87. doi:10.1145/2347736.2347755

Edward, M., & Sahadev, S. (2011). Role of switching costs in the service quality, perceived value, customer satisfaction and customer retention linkage. *Asia Pacific Journal of Marketing and Logistics*, *23*(3), 327–345. doi:10.1108/13555851111143240

Emeterio, I. C. S., Iglesias-Soler, E., Gallardo, L., Rodriguez-Cañamero, S., & García-Unanue, J. (2016). A prediction model of retention in a Spanish fitness centre. *Managing Sport and Leisure*, *21*(5), 300–318. doi:10.1080/23750472.2016.1274675

Ferrand, A., Robinson, L., & Valette-Florence, P. (2010). The intention-to-repurchase paradox: a case of the health and fitness industry. *Journal of Sport Management*, *24*(1), 83–105.

Gama, J., Oliveira, M., Lorena, A. C., Faceli, K., & Carvalho, A. P. de L. (2017). *Extração do Conhecimento de Dados Data Mining*. Edições Sílabo.

Hall, M. A. (1998). *Correlation-based Feature Selection for Machine Learning*.

Ho, T. K. (1995). Random decision forests. Em *Proceedings of 3rd International Conference on Document Analysis and Recognition* (Vol. 1, pp. 278–282 vol.1). doi:10.1109/ICDAR.1995.598994

Howat, G., & Assaker, G. (2016). Outcome quality in participant sport and recreation service quality models: Empirical results from public aquatic centres in Australia. *Sport Management Review*, *19*(5), 520–535. doi:10.1016/j.smr.2016.04.002

Hurley, T. (2004). Managing Customer Retention in the Health and Fitness Industry: A Case of Neglect. *Irish Marketing Review*, *17*(1/2), 23–29.

Kelleher, J. D., Namee, B. M., & D'Arcy, A. (2015). *Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies* (1 edition). Cambridge, Massachusetts: The MIT Press.

Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, *160*, 3–24.

Loh, W.-Y. (2011). Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *1*(1), 14–23. doi:10.1002/widm.8

MacIntosh, E., & Law, B. (2015). Should I stay or should I go? Exploring the decision to join, maintain, or cancel a fitness membership. *Managing Sport and Leisure*, *20*(3), 191–210. doi:10.1080/23750472.2015.1025093

McKinney, W., & others. (2010). Data structures for statistical computing in python. Em *Proceedings of the 9th Python in Science Conference* (Vol. 445, pp. 51–56). SciPy Austin, TX. Obtido de https://pdfs.semanticscholar.org/f6da/c1c52d3b07c993fe52513b8964f86e8fe381.pdf

Milošević, M., Živić, N., & Andjelković, I. (2017). Early churn prediction with personalized targeting in mobile social games. *Expert Systems with Applications*, *83*, 326–332. doi:10.1016/j.eswa.2017.04.056

Pawlowski, T., Breuer, C., Wicker, P., & Poupaux, S. (2009). Travel Time Spending Behaviour in Recreational Sports: An Econometric Approach with Management Implications. *European Sport Management Quarterly*, *9*(3), 215–242. doi:10.1080/16184740903023971

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*(Oct), 2825–2830.

Pridgeon, L., & Grogan, S. (2012). Understanding exercise adherence and dropout: an interpretative phenomenological analysis of men and women's accounts of gym attendance and non-attendance. *Qualitative Research in Sport, Exercise and Health*, *4*(3), 382–399. doi:10.1080/2159676X.2012.712984

Refaeilzadeh, P., Tang, L., & Liu, H. (2016). Cross-Validation. Em *Encyclopedia of Database Systems* (pp. 1–7). Springer, New York, NY. doi:10.1007/978-1-4899-7993-3_565-2

Swersky, K., Snoek, J., & Adams, R. P. (2013). Multi-Task Bayesian Optimization. Em C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems 26* (pp. 2004–2012). Curran Associates, Inc. Obtido de http://papers.nips.cc/paper/5086-multi-task-bayesian-optimization.pdf

Verbeke, W., Martens, D., Mues, C., & Baesens, B. (2011). Building comprehensible customer churn prediction models with advanced rule induction techniques. *Expert Systems with Applications*, *38*(3), 2354–2364. doi:10.1016/j.eswa.2010.08.023

Walt, S. van der, Colbert, S. C., & Varoquaux, G. (2011). The NumPy Array: A Structure for Efficient Numerical Computation. *Computing in Science & Engineering*, *13*(2), 22–30. doi:10.1109/MCSE.2011.37

Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. Em *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining* (pp. 29–39).