

ANÁLISE PREDITIVA DO PERFIL DOS INVESTIDORES DO TESOIRO DIRETO PARA AÇÕES DE MARKETING

MATHEUS DA SILVA BIZZI

UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO (UFES)
matheusbizzi@gmail.com

CIRILO ANDRADE VIEIRA

UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO (UFES)
cirilo.andrade1@gmail.com

ALAMIR COSTA LOURO

UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO (UFES)
alamirlouro@gmail.com

ANÁLISE PREDITIVA DO PERFIL DOS INVESTIDORES DO TESOURO DIRETO PARA AÇÕES DE MARKETING

Resumo

Emerge um novo perfil de analista de *Marketing* com apoio na tomada de decisão baseada a dados. Observam-se, inclusive, muitas empresas que percebem sua maior riqueza naquilo que sabem a respeito de seus clientes. No entanto, a obtenção de conhecimento sobre clientes, exige a existência de dados e conhecimentos que extrapolam o perfil atual do Administrador de Empresas. Uma análise de um volumoso banco de dados torna necessária a utilização de ferramentas que extraiam conhecimento sobre os clientes. Nesse cenário, destacam-se as ferramentas de *Data Mining*. A partir de uma base de dados do cadastro dos investidores do Tesouro Direto no Brasil do site dados.gov.br, presente trabalho buscou utilizar a ferramenta R para entender o perfil dos investidores brasileiros, no intuito de otimizar ações de *marketing* de corretoras e bancos de investimento. Deseja-se estabelecer relações entre os atributos, de modo que essas instituições tenham suas decisões pautadas, não em sentimentos, mas sim em análises preditivas.

Palavras-chave: Investimentos, *marketing*, *data-mining*.

PREDICTIVE ANALYSIS OF TESOURO DIRETO INVESTORS PROFILE FOR MARKETING ACTIONS

Abstract

A new Marketing Analyst profile emerges with data-driven decision making support. There are many companies that realize their greatest wealth in what they know about their customers. However, obtaining knowledge about clients requires the existence of data and knowledge that goes beyond the current profile of a Business Administrator. An analysis of a big database makes necessary the use of tools that extract knowledge about customers. In this scenario, we highlight the tools of Data Mining. With a database of investors in Brazilian Tesouro Direto, from data.gov.br, this work sought to use R tool to understand the profile of the Brazilian investors, in order to optimize marketing actions of brokers and banks of investments. It is desired to establish relationships between attributes, so that these institutions have their decisions based not on feelings but on predictive analyzes.

Keywords: Investments, marketing, data-mining.

1 INTRODUÇÃO

O advento de tecnologias emergentes revolucionárias e mudanças sociais relativas a informações disponibilizadas por meio de grandes volumes de dados, do inglês “*Big Data*” e o uso inovador das informações já disponíveis dentro das próprias organizações ou mesmo a partir das mídias digitais ou de dados disponibilizados na Web para mineração de dados é um fenômeno recente, complexo, mas muito debatido (Wamba et al., 2017). Dessa percepção, uma premissa assumida no presente trabalho é que o uso da modelagem preditiva deve ser encarado como oportunidade para melhores tomadas de decisão na área do *Marketing*. Para tanto o novo perfil do analista de Marketing precisa abarcar conhecimentos sobre técnicas de modelagem do comportamento do cliente, de predição de respostas de campanhas, de otimização de marketing-mix e personalização (Wedel & Kannan, 2016).

O trabalho se baseia em um conjunto de dados que contém diversos atributos de uma lista de investidores cadastrados no programa Tesouro Direto do Tesouro Nacional. Tal programa é desenvolvido em parceria com a BMF&F Bovespa para venda de títulos públicos federais para pessoa física pela internet, os títulos públicos são ativos fixos, ou seja, o retorno pode ser estimado no momento na compra, fazendo dele um investimento conservador (de mais baixo risco) visto que o retorno é garantido pelo governo federal, além de ser de fácil acesso e flexível, o investimento no tesouro direto se tornou um investimento muito popular no Brasil. Nessa base de dados disponível na internet estão inclusos dados do perfil do investidor, do tipo qualitativos como Profissão, UF, data de adesão, estado civil, cidade e sobre as atividades da conta. A base de dados data de janeiro de 2002, e segue em curso até hoje, tendo sua atualização geralmente até o quinto dia útil do mês subsequente ao de referência estatística.

Em um contexto de alargamento das oportunidades para o marketing usando da ciência dos dados, apenas para exemplificar, a Netflix analisa milhões de dados de seus espectadores em tempo real, ajudando assim a determinar se um novo filme piloto se tornará uma opção bem sucedida (Xu, Frankwick & Ramirez, 2016). Ainda para esses últimos autores outros cenários do cotidiano atual são referidos na literatura como mudanças causadas pela análise de grande volume de dados. Mas todo esse volume de dados não é acessível e nem útil sem o conhecimento da ciência dos dados, disciplina que atua cada vez mais na interface entre a econometria, psicometria, estatística, ciência da computação e as diversas áreas do *business* (Wedel & Kannan, 2016).

Segundo Provost e Fawcett (2013), provavelmente, as aplicações mais claras de técnicas de mineração de dados estão no marketing, especialmente para facilitar o como marketing direcionado, a publicidade on-line e as recomendações para vendas cruzadas. A partir dessa visão o presente trabalho busca apontar por oportunidades na área do marketing usando de uma base de dados pública.

2 REFERENCIAL TEÓRICO

As análises preditivas com foco nos clientes podem transformar dados estruturados, ou não estruturados, em informações estratégicas, mas para tal elas exigem um conhecimento profundo de técnicas de modelagem de *Marketing* para predição de respostas de mercado, otimização de *marketing-mix* e personalização (Wedel & Kannan, 2016). O tradicional “4Ps” de *marketing* (preço, produto, promoção e praça) pode ser ajustado para melhor atender o cliente - preço único baseado em seu perfil ou padrões de consumo, ou um produto exclusivo que se adapte às necessidades do cliente, ou entrega em um momento e lugar que se adapte ao cliente (Cooke & Zubcsek, 2017).

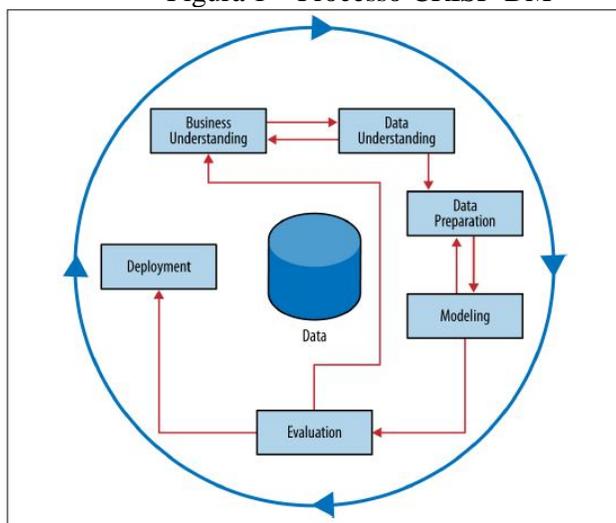
O *Marketing* direcionado pela ciência dos dados pode impulsionar interações excepcionais entre os funcionários e os clientes, alinhando o comportamento do cliente

com o funcionário certo e nas circunstâncias corretas. A compreensão das preferências dos clientes pode ajudar a alcançar esse objetivo, combinando fatores como disponibilidade e habilidades dos funcionários. O objetivo deste tipo de análise não é prescrever um certo tipo de comportamento do empregado, isto é ler um roteiro, mas alinhar funcionários e clientes de tal forma que o comportamento natural do empregado seja mais eficiente para o cliente, criando uma conexão e lealdade significativa entre organização e cliente (Cooke & Zubcsek, 2017).

3 METODOLOGIA

A metodologia utilizada na pesquisa segue os moldes utilizados em outras pesquisas que utilizam o processo CRISP-DM (Provost & Fawcett, 2013). Sendo dividida em etapas: pré-processamento, técnicas de *data mining*, interpretação de dados e avaliação da aplicação.

Figura 1 – Processo CRISP-DM



Fonte: Provost, F.; Fawcett, T., 2013, p. 27

3.1 Compreensão do Negócio e dos dados

Provost e Fawcett (2013) explicam que na compreensão do negócio é vital para definir o problema que será resolvido. Essa fase envolve o pré-processamento que consiste na aplicação de técnicas para organização, tratamento e a preparação dos dados. Nesta etapa, os procedimentos feitos buscam realizar a correção de dados inseridos de forma incorreta, exclusão de dados faltantes que comprometam futuras análises e a formatação dos dados. O objetivo é a maior qualidade dos dados, de modo que os mesmos se adequem aos algoritmos que serão utilizados e aumentem sua eficiência. Na presente pesquisa, as técnicas utilizadas se limitaram em limpeza e formatação de dados realizadas diretamente no *Microsoft Excel* ou por meio da ferramenta *Rstudio* que foi utilizada para todas as gerações de gráficos estatísticos e para execução do método supervisionado da árvore de indução, isto é, foi utilizada uma Classificação, que é quando a variável alvo é uma variável categórica (Provost & Fawcett, 2013).

3.2 Preparação dos dados

A base de dados original continha o campo “Data de Adesão” com registros extremos e com características bastante distintas dos demais registros, sendo então tratados como *outliers* e ou *missing values*. Dessa forma, foram excluídas 11.932 linhas registros num total de 1.048.576. Representando 1.14% dos dados, valor menor que 5% da população, o que segundo Hair, Black, Babin, Anderson e Tatham (2009) são dados podem ser classificados como dados perdidos e ignoráveis, entendidos como aleatórios e dessa forma foi escolhida a exclusão dos mesmos.

Após a etapa de limpeza, os registros do campo “Data de Adesão” passou pelo processo de transformação de dados. O procedimento realizado foi modificar o formato do campo de “dia/mês/ano” para “ano”, afim de que após formatação, as análises a serem realizadas fossem otimizadas e fizessem mais sentido na etapa de interpretação.

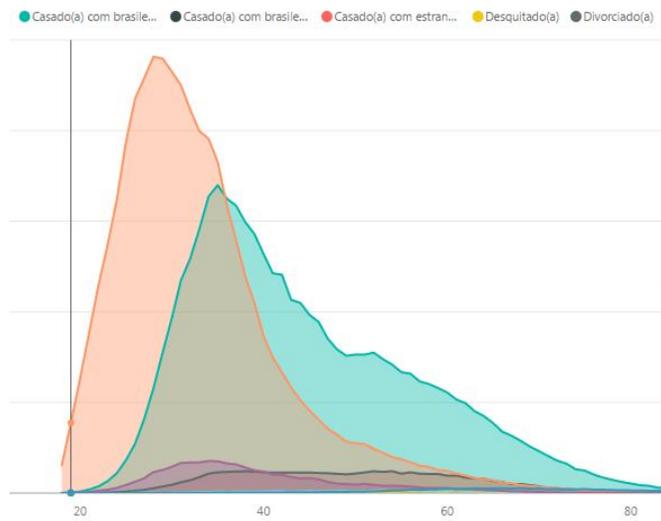
4 ANÁLISE DOS RESULTADOS

As técnicas realizadas no banco de dados foram aplicadas de modo que identificassem o perfil do investidor, sobretudo o de caráter ativo. Para tanto, as ferramentas utilizadas foram: *box plot* e algoritmo de árvore de indução, se preocuparam em responder as seguintes questões:

- A. Qual a relação entre a idade e a situação da conta?
- B. Qual a relação entre a idade e o tempo de investimento?
- C. Qual a média de idade das contas ativas?
- D. Quais os principais grupos formados dentro das contas ativas?
- E. Quais fatores influenciam a conta estar ativa ou inativa?

4.1 Análise gráfica

Figura 2 – Perfil do investidor, nível de investimento X estado cível X idade



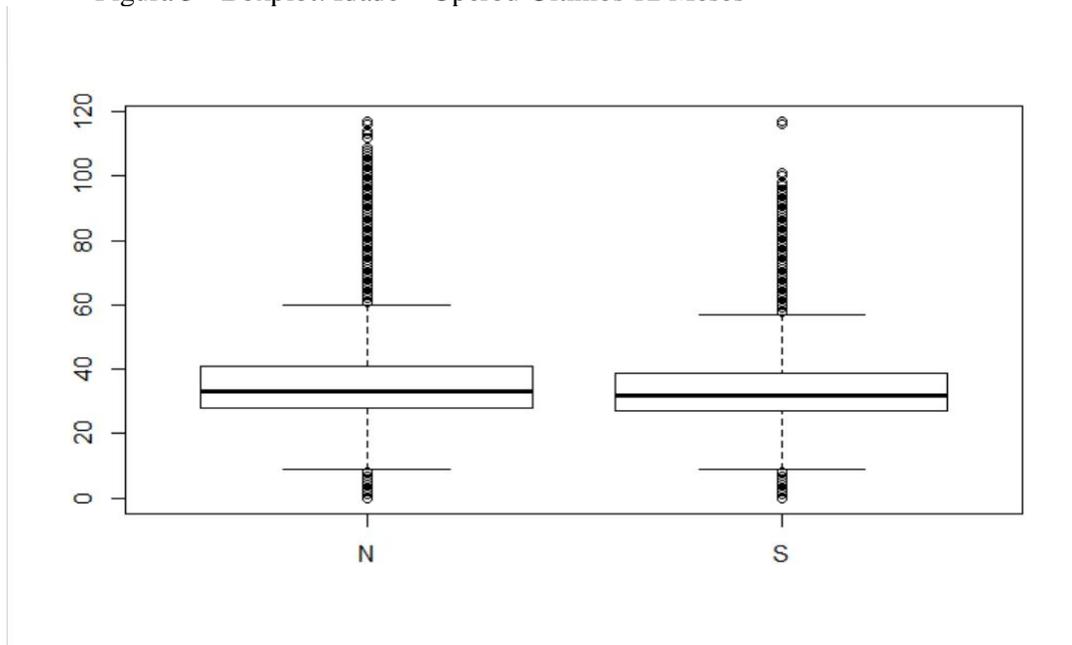
Fonte: Elaborado pelos autores (2017)

O gráfico gerado no software PowerBI da microsoft demonstra dois perfis em evidência como grupo de investidores ativos, sendo estes: “Solteiro, com média de 30 anos de idade” e “Casado com média de idade de 35 anos”.

4.1.1 Box plot

O boxplot é um gráfico que capta importantes aspectos de um conjunto de dados a partir dos seguintes valores: valor mínimo, primeiro quartil, segundo quartil, terceiro quartil, valor máximo e *outliers* (Hair, Black, Babin, Anderson & Tatham, 2009). É uma forma simples de visualizar como os dados se distribuem e foi utilizado na presente pesquisa para ilustrar a média de idade dos investidores em relação a situação da conta: ativa ou inativa e se houve operação nos últimos 12 meses. Dessa forma, os seguintes resultados foram observados:

Figura 3 - Boxplot: Idade ~ Operou Últimos 12 Meses



Fonte: Elaborado pelos autores (2017)

A análise visual em alguns casos não é suficiente para se chegar a conclusões, como exemplificado pela Figura 3. Para resolver essa questão foi utilizado o teste t.

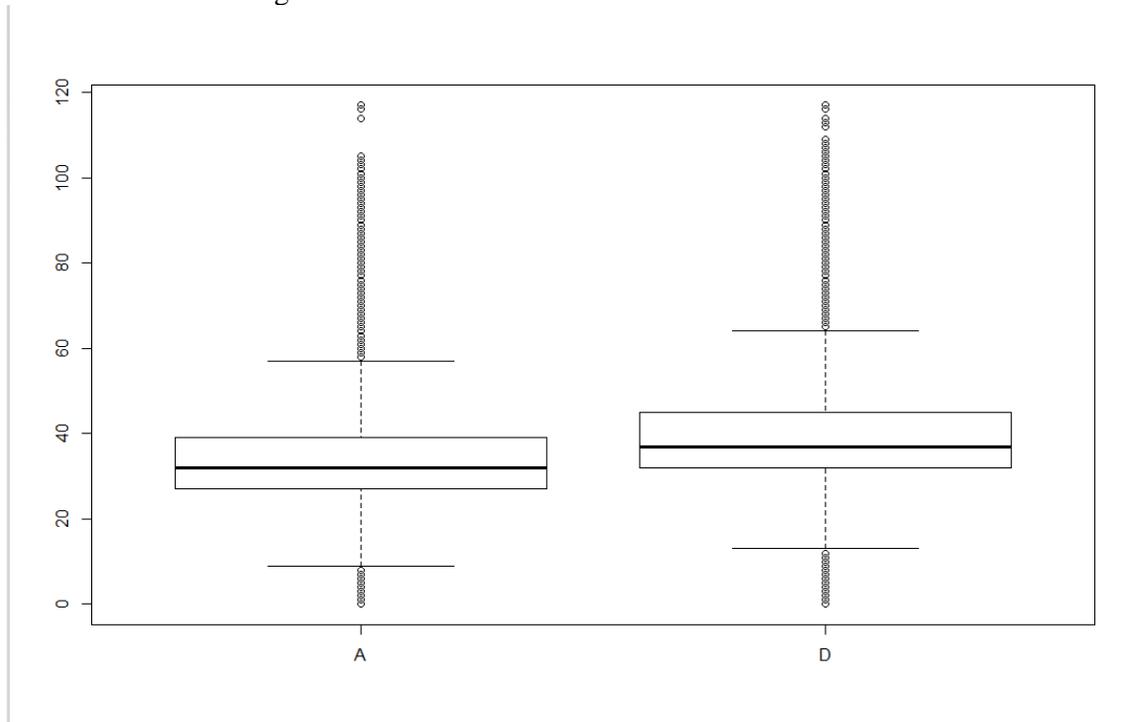
Welch Two Sample t-test

```
data: Idade by Operou.12.Meses
t = 47.068, df = 517930, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1.183476 1.286323
sample estimates:
mean in group N mean in group S
 36.02072      34.78582
```

O resultado demonstrou que os investidores que operaram nos últimos 12 meses, possuem idade em média 34.8 anos, já os que não operaram possuem média de idade significativamente diferente, p-value menor 0.001, com média de 36.0 anos. Não contribuindo muito para a análise por serem médias próximas.

Já para o caso de contas ativas versus inativas por idade o resultado pode ser melhor interpretado.

Figura 2 - Idade ~ Conta Ativa ou Conta Inativa



Fonte: Elaborado pelos autores (2017)

Welch Two Sample t-test

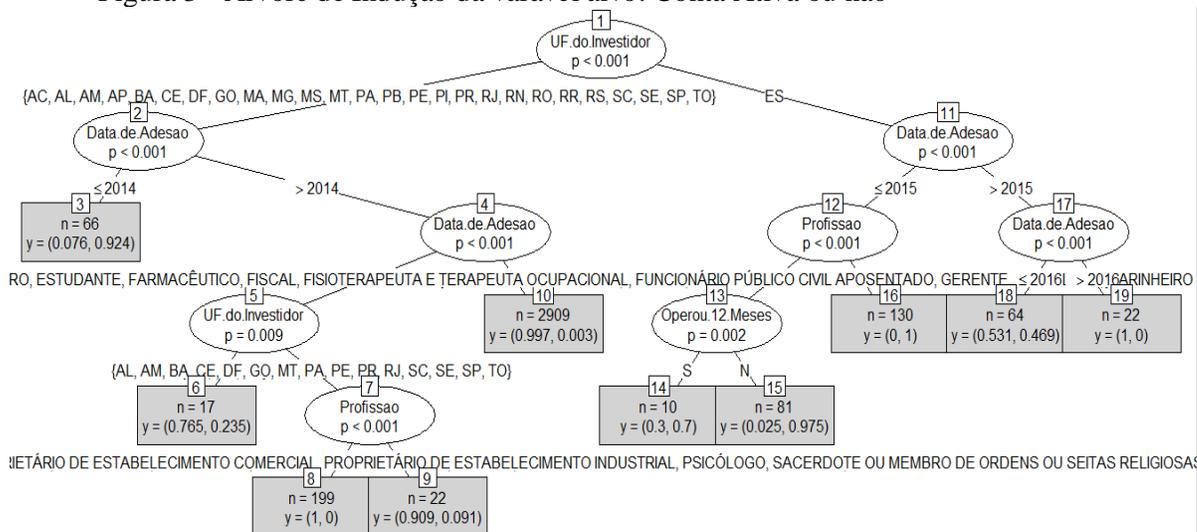
```
data: Idade by Situação da Conta
t = -162.04, df = 291260, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -4.998740 -4.879262
sample estimates:
mean in group A mean in group D
    34.75695      39.69595
```

O resultado demonstrou que os investidores que ativos, possuem idade em média 34.8 anos, já os inativos possuem média significativamente diferente, p-value menor 0.001, com média de 39.7 anos. Sendo possível extrair informação mais útil para uma possível campanha de *Marketing* entre os ativos ou entre os inativos.

4.2 Fase de Modelagem da Árvore de Indução

Árvore de indução é um modelo preditivo representado como próprio nome diz, a partir de uma estrutura semelhante a uma árvore. Dessa forma, cada galho da árvore é uma pergunta de classificação com as folhas ilustrando outros conjuntos de dados e suas classificações. Na presente pesquisa o algoritmo, *cree* do pacote R chamado “party”, em questão foi utilizado para determinar quais atributos podem determinar se a conta do investidor operou nos últimos 12 meses, obtendo os seguintes resultados, com uma acurácia de 0.7927546, consideravelmente maior que os 50% de uma avaliação randômica.

Figura 3 - Árvore de Indução da varável alvo: Conta Ativa ou não

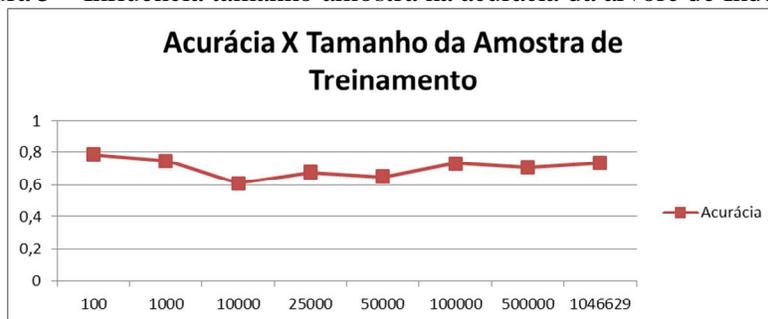


Fonte: Elaborado pelos autores (2017)

Uma primeira constatação é o comportamento diferenciado do estado do Espírito Santo em relação aos outros estados brasileiros. Outros destaques são o nó 3, 8, 9, 10, 16 e 19 que demonstram quase 100%, e as vezes 100%, de acerto. Demonstrando grande previsibilidade do perfil do investidor para alguns casos.

Já a figura 5 mostra uma instabilidade na acurácia quando o tamanho da amostra é pequeno, até 50 mil registros, e uma crescente da acurácia para amostrar maiores.

Figura 5 – Influencia tamanho amostra na acurácia da árvore de Indução



Fonte: Elaborado pelos autores (2017)

5 CONCLUSÕES

A tomada de decisão é um processo que exige embasamento e conhecimento. Dessa forma, a análise minuciosa de um mercado pode revelar a possibilidade de encontrar uma oportunidade até então oculta de ofertar produtos ou serviços, criar uma rede de distribuição melhor e mais eficiente. Em um cenário de mercado cada vez mais competitivo, todas as decisões devem ser tomadas com base em dados concretos, que traduzem a realidade e com análises que revelam o que estes escondem.

Estar atento às tendências do mercado e para onde ele se dirige não é mais a melhor estratégia. A competitividade aponta para a necessidade de que é possível e necessário encontrar uma oportunidade que as demais empresas não tenham visto. Neste contexto, as análises de *data-mining* trazem ao mercado a possibilidade de análises com aplicações em diferentes segmentos.

Na presente pesquisa, após as análises realizadas, foram encontrados dois perfis em evidência no grupo de investidores ativos, sendo estes: “Solteiro, com média de 30 anos de idade” e “Casado com média de idade de 35 anos”. Neste cenário, e retornando ao pressuposto de que o nível de competitividade do mercado cresce de forma cada vez mais acelerada, a clara identificação de perfis como os encontrados facilitam à tomada de decisão por parte dos gestores.

No que se refere especificamente ao marketing com foco na captação de clientes, saber quem são os consumidores de um determinado produto, sobretudo um em ascensão no mercado, otimiza eventuais campanhas publicitárias, uma vez que se sabe exatamente “para quem”, “o que” e “como” deve haver a melhor comunicação. Na presente pesquisa, por exemplo, sugerem-se campanhas específicas para os dois perfis encontrados, de modo que estes futuros clientes se identifiquem com a empresa. Além disso, sugere-se a criação de outros produtos criados com base nas características dos dois públicos, de modo a minimizar a evasão de clientes, inatividade nas contas ou baixos índices de operação.

4 REFERÊNCIAS

- Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2009). *Análise multivariada de dados*. (Bookman, Ed.), *Bookman* (6th ed.). Porto Alegre: Bookman.
- Provost, F., & Fawcett, T. (2013). *Data Science for Business* (1st ed.). Sebastopol, CA: O'Reilly Media, Inc. <https://doi.org/10.1007/s13398-014-0173-7.2>
- Wamba, S. F., & Mishra, D. (2017). Big data integration with business processes: a literature review. *Business Process Management Journal*, 23(3). <https://doi.org/10.1108/BPMJ-02-2017-0047>
- Wedel, M., & Kannan, P. K. (2016). Marketing Analytics for Data-Rich Environments. *Journal of Marketing*, 80(6), 97–121. <https://doi.org/10.1509/jm.15.0413>
- Xu, Z., Frankwick, G. L., & Ramirez, E. (2016). Effects of big data analytics and traditional marketing analytics on new product success: A knowledge fusion perspective. *Journal of Business Research*, 69(5), 1562–1566. <https://doi.org/10.1016/j.jbusres.2015.10.017>