AUTOMAÇÃO DE PROCESSOS: A DETECÇÃO DE ANOMALIAS COMO FORMA DE OTIMIZAR CAMPANHAS DE *MARKETING* DIGITAL

Jhonatan Figueiredo Cardoso[1]

Princya Nélida de Oliveira[2]

Jaqueline Brigladori Pugliesi[3]

Resumo

No contexto do *Marketing* Digital, a crescente expansão dos canais digitais e o consequente aumento no volume de dados gerados por esses meios tem trazido muitas oportunidades, mas também desafios. Um destes desafios é a otimização de campanhas de *Marketing* Digital, pelo volume e complexidade dos dados a serem analisados. Neste sentido, este estudo propõe a automação de processos que podem ser realizados sem a supervisão humana, por meio da aplicação de um algoritmo de detecção de anomalias em séries temporais como forma de mensurar a eficiência ou não de campanhas de *Marketing* Digital. O benefício esperado com a aplicação desta técnica de Aprendizado de Máquina é que um gestor de *Marketing* Digital não tenha que contar apenas com a sua experiência e intuição para analisar os dados coletados, podendo assim, focar em tarefas que exigem a intervenção humana e que sejam mais estratégicas para o negócio.

Palavras-chave: Aprendizado de Máquina. Ciência de Dados. Inteligência de *Marketing*.

Abstract

In the Digital Marketing context, the growing expansions of digital channels and the consequent increase in the volume of data generated by these means have brought a lot of opportunities, but also challenges. One of these challenges is the optimization of Digital Marketing campaigns, due to the volume and complexity of the data to be analyzed. For this reason, this study proposes the automation of processes that can be performed without the human being supervision, through the application of an anomaly detection algorithm in time series as a way of measuring the efficiency or not of Digital Marketing campaigns. The expected benefit with the application of this Machine Learning technique is that a Digital Marketing manager does not have to rely only on his experience and intuition to analyze the collected data, thus, being able to focus on tasks that require human intervention and that are more strategic for the business.

Keywords: *Machine Learning. Data Science. Marketing Intelligence.*

[1] Graduando em Análise e Desenvolvimento de Sistemas pela Fatec "Dr. Thomaz Novelino" – Franca/SP. Endereço eletrônico: jhonatancardoso17@hotmail.com.
[2] Graduanda em Análise e Desenvolvimento de Sistemas pela Fatec Dr. Thomaz Novelino" – Franca/SP. Endereço eletrônico: princyanoliveira@gmail.com.
[3] Doutora em Ciências da Computação pela USP – São Carlos/SP. Endereço eletrônico: jbpugliesi@gmail.com.

1 Introdução

Um dos objetivos do *Marketing* Digital é a otimização de campanhas, impulsionando o ROI (Retorno Sobre Investimento) e guiando os processos de criação de novas campanhas. Desenvolver um modelo para medir com precisão o impacto de ações de *Marketing* já era necessário em canais de mídia tradicionais, como televisão, jornais, panfletos etc. A crescente difusão de canais digitais (mídia digital), além de novos desafios, trouxe inúmeras possibilidades, permitindo tratar com maior precisão o problema da atribuição, uma vez que os profissionais podem acessar dados desagregados e a nível individual das reações dos consumidores às campanhas, o que não era possível com as mídias *offline* (ABHISHEK; DESPOTAKIS; RAVI, 2017).

Neste contexto, a análise de dados tem tido papel fundamental para a mensuração e otimização de resultados de estratégias de *Marketing* Digital, o que leva a problemática deste artigo.

Imagine o seguinte cenário: um gerente de campanha de *Marketing* Digital gerencia mais de uma dezena de campanhas. Normalmente, ele inicia seu dia olhando para um painel de dados, avaliando quais campanhas estão funcionando bem e quais não estão. Este gestor, contando apenas com os dados brutos oferecidos por relatórios quantitativos, precisa revisar e analisar uma grande quantidade de dados de forma manual, demandando um esforço intelectual que interfere na produtividade e no desempenho das ações, contando com a otimização de suas campanhas baseando-se principalmente em sua própria experiência e intuição.

A partir deste cenário é fácil intuir que ao automatizar tarefas que podem ser realizadas sem a supervisão humana, por meio da aplicação do Aprendizado de Máquina, os profissionais de *Marketing* liberam tempo para focar em problemas que exigem a intervenção humana, tendo como resultado um processo mais simplificado e eficaz, com menos erros humanos e com aumento na capacidade de profissionais de *Marketing* Digital de lidar com tarefas complexas e estratégicas.

A oportunidade de usar esses dados e aplicar técnicas de Aprendizado de Máquina para prever tendências de consumidores, oferece aos profissionais de Marketing Digital a possibilidade de uma tomada de decisão mais inteligente. Os algoritmos podem processar e analisar de maneira rápida e precisa dados de

ISSN: 2595-5942

campanhas, acionando notificações quando certas tendências ou picos incomuns ocorrem (MIKLOSIK et al., 2019).

Assim, a proposta deste artigo é discorrer sobre a possibilidade de utilização de técnicas de Aprendizado de Máquina, para que utilizando dados de quantidade de acessos ou vendas em um site, por exemplo, seja possível mapear um padrão de comportamento destes consumidores, verificando a eficiência ou não de uma campanha de *Marketing* para aquele produto ou serviço, no que se refere a captura e/ou conversão de *leads*. Busca-se alcançar este objetivo por meio da contextualização do tema proposto, em que, com um levantamento bibliográfico é possível compreender um pouco mais sobre o universo dos dados. Neste sentido, o Capítulo 2 traz um *overview* sobre a era dos dados e temas mais específicos que se relacionam, como a Ciência de Dados e o conceito de *Big Data*. Após isto, o Capítulo 3 aborda o conceito de Inteligência Artificial e suas subáreas, focando principalmente em Aprendizado de Máquina, relacionando-o ainda com o *Marketing* Digital e conceituando temas úteis a compreensão das discussões e resultados que serão apresentados no Capítulo 4, que buscará demonstrar o funcionamento do algoritmo criado para a detecção de anomalias em séries temporais, sugerindo de que forma ele poderia ser aplicado no contexto do *Marketing* Digital. Por fim, serão apresentadas as considerações finais sobre o estudo realizado e as referências utilizadas.

2 A era dos dados e a ciência de dados

Antes de conceituar o termo Ciência de Dados e de que forma ele se relaciona com este trabalho, faz-se necessária a compreensão do que é um dado. Neste sentido, o CCSDS (*Consultative Committee for Space Data Systems*, ou Comitê Consultivo para Sistemas de Dados Espaciais) definiu Dado como:

> Uma representação reinterpretável de informações de uma maneira formalizada adequada para comunicação, interpretação ou processamento. Exemplos de dados incluem uma sequência de *bits*, uma tabela de números, os caracteres em uma página, a gravação de sons feitos por uma pessoa falando ou um espécime de rocha lunar (CCSDS, 2012, p.10).

Davenport (1998, p.18), contribui com a definição de dados dizendo que estes são observações sobre o estado do mundo, facilmente estruturados e obtidos por

máquinas e com frequência são quantificados e de clara transcrição, concluindo sua fala dizendo que nada se perde quando representado em bits.

Além disso, o dado é considerado a base da pirâmide conhecida como DIKW (*Data, Information, Knowledge, Wisdom*, que em português significa, Dados, Informação, Conhecimento, Sabedoria) ou pirâmide do conhecimento. Nesta pirâmide do conhecimento, a informação é entendida como o dado processado ou interpretado, conhecimento como informação processada ou interpretada e sabedoria como conhecimento interpretado (HJØRLAND, 2018).

Desta forma, com o estudo e o processamento dos dados, visando seus processos de captura, preparação para análise e análise propriamente dita desses dados, originou-se a Ciência de Dados, que de acordo com Carvalho (2016):

> Estuda princípios, métodos e sistemas computacionais capazes de extrair de forma eficiente conhecimento novo, útil e relevante presente em conjuntos de dados. Para isso, ela faz uso de técnicas de mineração de dados, particularmente de construção automática de modelos, capazes de extrair esse conhecimento. A construção automática de modelos permite que funções, hipóteses e regras sejam extraídas a partir de experiências passadas, representadas no conjunto de dados (CARVALHO, 2016, p.63).

A Ciência de Dados abrange diversas áreas, como Matemática, Computação e Estatística, porém se difere da Ciência da Computação e da Estatística Analítica, pois utiliza princípios científicos aos dados coletados. Isso se justifica pelo que se chama, *Big Data*, que é responsável por exigir o uso de tecnologias diversas em relação à análise estatística. Assim, os profissionais do ramo da estatística que atuam no mercado há anos não conseguiriam realizar a análise profunda de dados de massa em tempo quase real, como acontece hoje nas grandes empresas, e será estudado de forma mais aprofundada na próxima seção.

2.1 *Big data*

Na tradução literal, se tem que *Big Data* significa “Grandes Dados”, o que de fato é. Este termo diz respeito aos dados com grande variedade, volume e velocidade cada vez maiores.

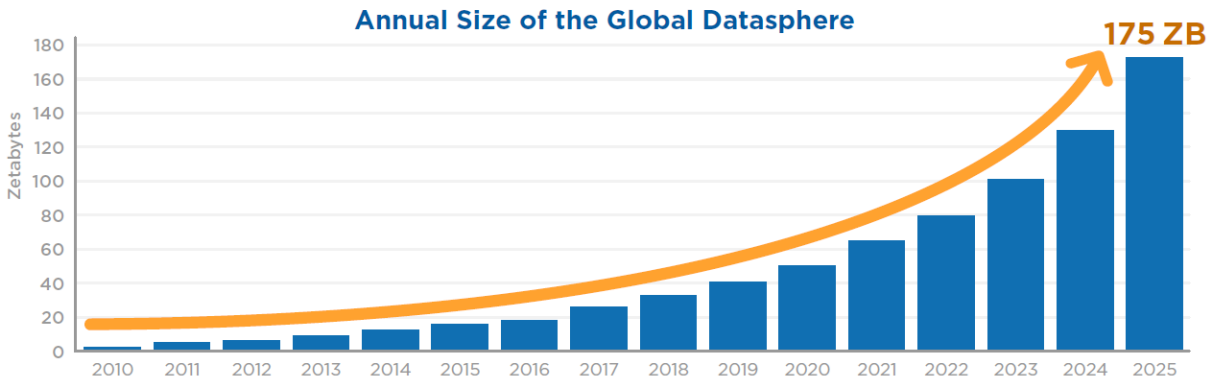
Em outras palavras, *Big Data* é um conglomerado de dados cada vez maior e cada vez mais complexo, sobretudo de novas fontes. Tal conglomerado, pela sua extensão, torna-se quase impossível para um *software* tradicional processá-lo.

ISSN: 2595-5942

Todavia, esse conglomerado de dados, tem importância significativa para as empresas de forma geral, pois pode ser utilizado para solucionar problemas que não se conseguia antes.

Segundo um estudo de 2018 de John Gantz, David Reinsel e John Rydning realizado e publicado pela empresa IDC - *International Data Corporation*, a esfera global de dados irá crescer de 33 *zettabytes* em 2018 para 175 *zettabytes* em 2025 (GANTZ; REINSEL; RYDNING, 2018, p.6), conforme Figura 1.

Figura 1: Tamanho anual da esfera de dados global



Fonte: GANTZ; REINSEL; RYDNING, 2018, p.6.

Para se ter uma noção mais concreta de quão grande é 175 *zettabytes*, seguem alguns exemplos:

- Um *Zettabyte* é equivalente a 1 trilhão de *Gigabytes*.
- Se fosse possível armazenar todos esses dados em DVDs, teríamos uma pilha de DVDs que alcançaria a lua 23 vezes ou daria 222 voltas no planeta Terra.
- Se uma pessoa fosse fazer o *download* de 175 ZB em uma taxa de *download* de 25 Mb/s, levaria 1,8 bilhões de anos para finalizar esse *download*, ou se cada pessoa do planeta pudesse realizar o *download* sem descansar, levaria ‘apenas’ 81 dias.

Destarte, para o tratamento de todos esses dados é imprescindível o uso de Inteligência Artificial, mais especificamente, o Aprendizado de Máquina, que oferece um gerenciamento otimizado dos dados, temas esses que serão conceituados no próximo capítulo.

3 Inteligência artificial

O conceito de Inteligência Artificial, mesmo em pleno ano de 2022, ainda não é universalmente definido e aceito, tendo sido explicado de forma diversa por vários autores.

Knight, Rich e Nair (2009, p.3), iniciam seu estudo no livro *Artificial Intelligence* dizendo que, embora a maioria das tentativas para definir com precisão termos complexos e de utilização ampla seja exercício de futilidade, é útil desenhar pelo menos uma fronteira aproximada em torno do conceito, para que se tenha ideia sobre a discussão. Dessa forma eles propõem a seguinte definição: “A Inteligência Artificial é o estudo de como fazer os computadores realizarem tarefas em que, no momento, as pessoas são melhores”.

Na tentativa de uma definição do que é a Inteligência Artificial, Luger (LUGER, 2013, p.1) diz que “a inteligência artificial (IA) pode ser definida como o ramo da ciência da computação que se ocupa da automação do comportamento inteligente”. Tendo isso em vista, ele complementa que tal definição é apropriada pois enfatiza a convicção de que a IA:

> [...] faz parte da Ciência da Computação e que, desse modo, deve ser baseada em princípios teóricos e aplicados sólidos nesse campo. Esses princípios incluem as estruturas de dados usadas na representação do conhecimento, os algoritmos necessários para aplicar esse conhecimento e as linguagens e técnicas de programação usadas em sua implementação (LUGER, 2013, p.1).

Assim, pode-se dizer que, a Inteligência Artificial abrange os sistemas que visam simular a inteligência humana para realizar tarefas, analisando dados em maior quantidade e velocidade do que um ser humano é capaz, com uma capacidade de pensamento e processamento superpoderoso. A Inteligência Artificial não tem como objetivo substituir o ser humano, mas sim melhorar as habilidades humanas.

A Inteligência Artificial está em grande expansão e evolução, sendo que novas subáreas são, cada vez mais, desenvolvidas e implementadas. Estas subáreas não estão isoladas e se relacionam de várias formas. Algumas das subáreas mais relevantes são as Redes Neurais, o Processamento de Linguagem Natural e o Aprendizado de Máquina (SOLVIMM, 2019, *online*).

As Redes Neurais são algoritmos especializados em reconhecimento de padrões baseadas no cérebro humano. Elas agrupam e classificam dados de acordo

com a percepção de similaridades, sejam esses dados imagens, vídeos, textos ou sons, que devem ser traduzidos em números e contidos em vetores (LUGER, 2013).

O Processamento de Linguagem Natural é a subárea da Inteligência Artificial que visa a criação de geradores e processadores de textos, com habilidade de interpretação e processamento da linguagem humana. Os textos interpretados oferecem uma interação mais natural e são utilizados em sistemas que convertem falas de uma pessoa em texto escrito, além de sistemas de tradução automática. Alguns exemplos são os *chatbots* e tradutores como o *Google Translate (RUSSEL; NORVIG, 2013).*

Também, tem-se o Aprendizado de Máquina (*Machine Learning*) que é a subárea que se dedica a criar sistemas cujos algoritmos e técnicas permitam a ele aprender com suas experiências, de forma automática e sem intervenção humana direta, tomando decisões e agindo com base em seu aprendizado. Para que isso aconteça, esses sistemas são alimentados com dados e instruções, buscando padrões para tomar decisões baseadas nos exemplos que lhes foram fornecidos, sendo que à medida que vai aprendendo, vai melhor desempenhando suas tarefas.

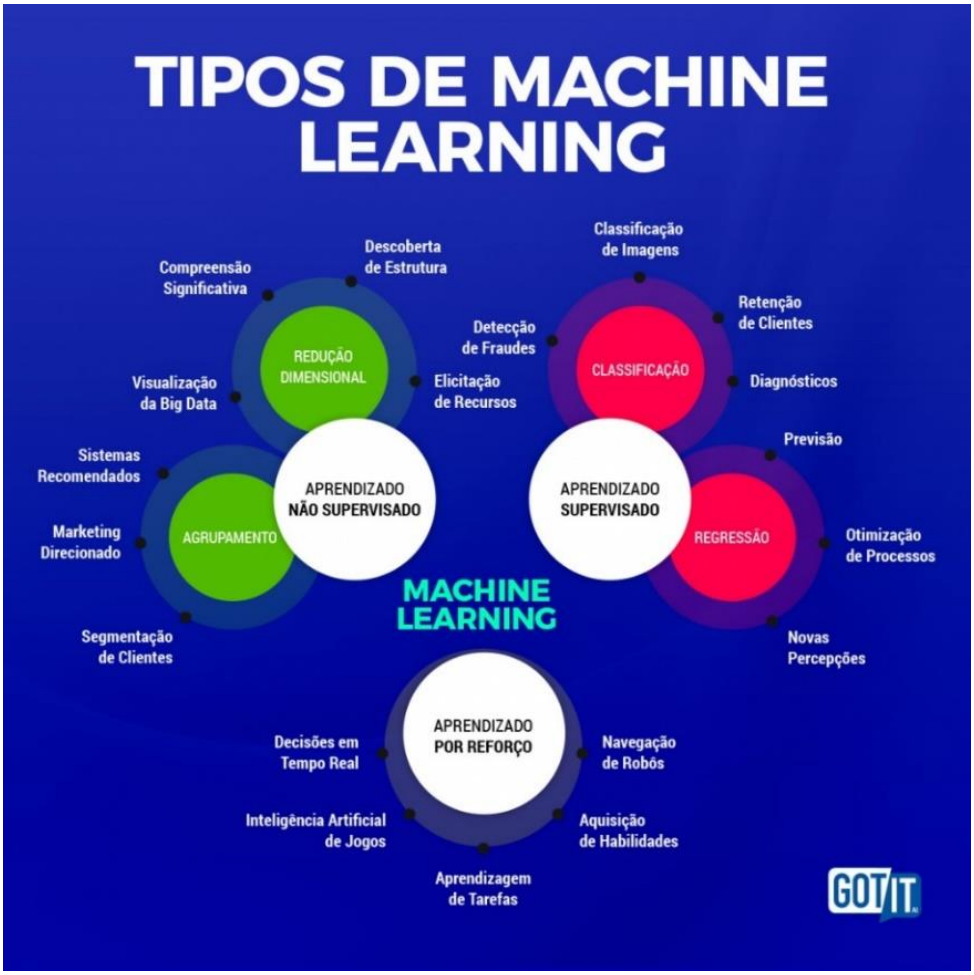
3.1 Aprendizado de máquina

Nesta seção será analisado, com mais profundidade, o funcionamento do Aprendizado de Máquina, buscando relacioná-lo com o objetivo deste estudo, que é a detecção de anomalias em séries temporais como forma de avaliar o desempenho de campanhas de *Marketing* Digital. Contudo, a título de informação, deve-se observar que existem categorias ou subcategorias dentro do Aprendizado de Máquina e uma forma muito utilizada para classificar essas subcategorias é a divisão pelo tipo de aprendizado, como na Figura 2.

Neste artigo, embora haja a proposição de técnicas de metrificação que se relacionam ao aprendizado supervisionado, a proposta do algoritmo de detecção de anomalias trazido é de aprendizado não supervisionado, em que os dados que serão aprendidos pelo algoritmo não são marcados ou rotulados, devendo ser aprendidos apenas por seu padrão.

ISSN: 2595-5942

Figura 2: Tipos de Aprendizado de Máquina (*Machine Learning*)



Fonte: GOT IT AL, *sd*, *online.*

Porém, antes de detalhar aspectos do algoritmo em si, julga-se importante a visualização de um panorama de utilização do Aprendizado de Máquina relacionado ao *Marketing* Digital.

3.1.1 O Aprendizado de máquina no *marketing* digital

Para os profissionais de *Marketing*, o aprendizado de máquina é uma oportunidade de tomar decisões cruciais rapidamente com base em *Big Data*. O Aprendizado de Máquina pode ser utilizado para encontrar padrões nas atividades do usuário em um site, por exemplo, ajudando a prever o comportamento dos usuários e otimizar rapidamente as ofertas de publicidade.

Quando centenas de parâmetros são coletados, os dados obtidos ganham valor porque contêm padrões de comportamento e dependências. Eles escondem o

enorme potencial dos dados comportamentais, permitindo complementar os dados do usuário com os parâmetros ausentes com base nos dados que já se tem para outros usuários.

Por exemplo, a maneira mais simples de definir um público-alvo é por sexo e idade, mas e se os usuários preencherem esses dados apenas em 20% dos casos? Como se pode entender quantos usuários do site se enquadram no público-alvo? Padrões de comportamento podem ajudar.

Pode-se usar dados de sexo e idade de 20% dos usuários para determinar padrões específicos para um determinado sexo e idade. Em seguida, usar esses padrões para prever o sexo e a idade dos 80% restantes dos usuários.

Com dados completos sobre sexo e idade, pode-se fazer ofertas personalizadas para todos os visitantes do *site*.

O Aprendizado de Máquina torna possível responder mais rapidamente às mudanças na qualidade do tráfego trazidas por campanhas publicitárias. Como resultado, pode-se dedicar mais tempo à criação de hipóteses do que à execução de ações rotineiras.

O valor dos resultados depende da relevância dos dados sobre os quais a análise foi realizada. À medida que os dados se tornam obsoletos, seu valor diminui. Uma pessoa simplesmente não consegue processar o volume de dados coletados a cada minuto pelos sistemas analíticos. Os sistemas de aprendizado de máquina podem processar centenas de solicitações, organizá-las e fornecer resultados na forma de uma resposta pronta para uma pergunta.

Pensando nisso, verifica-se a motivação para o estudo do tema e buscando aprofundar um pouco mais nos aspectos da Aprendizagem de Máquina, especificamente na detecção de anomalias em series temporais, as seções 3.1.2 e 3.1.3 trazem respectivamente, os conceitos de séries temporais e detecção de anomalias, para que na sequência possa ser desenvolvida a discussão do algoritmo desenvolvido e as conclusões deste estudo.

3.1.2 Séries temporais

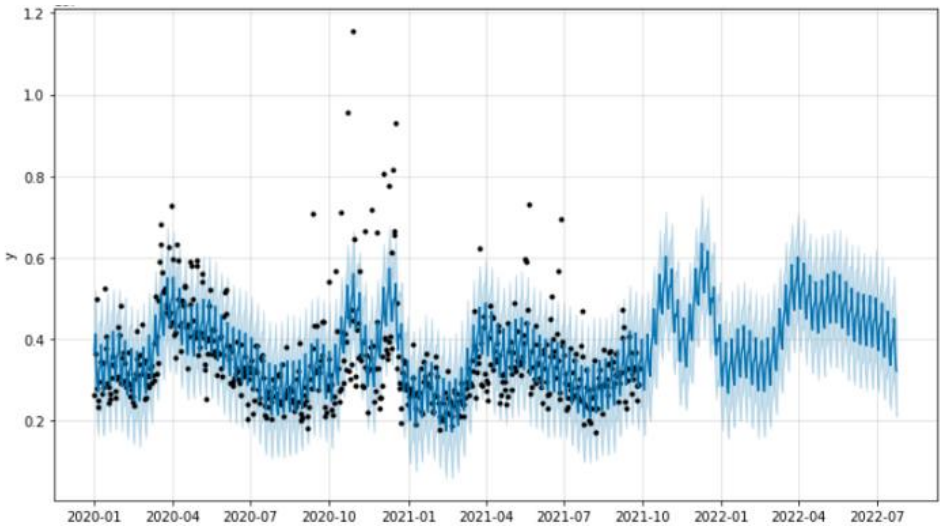
Para que se possa compreender onde poderia ser aplicado um algoritmo de detecção de anomalias é que se faz necessária a compreensão do que vem a ser uma série temporal.

As séries temporais são um conjunto de observações sobre uma variável ordenada no tempo e registradas em intervalos regulares. Como exemplo, podem ser citadas as seguintes séries temporais: vendas mensais de uma empresa, produto ou serviço, *leads* captados em um mês etc. O objetivo da análise de séries temporais é identificar padrões não aleatórios nas séries temporais das variáveis de interesse (STEVENSON, 2001).

Ao analisar uma variável, é possível verificar que ela tem um histórico que pode ajudar a identificar períodos de declínio/crescimento e sazonalidade, além de predizer observações futuras e orientar a tomada de decisões. Portanto, os modelos de séries temporais são ideais para avaliar o comportamento das variáveis ao longo do tempo.

Na Figura 3, estão representados os dados de acesso a um site com base no histórico dos usuários, em que, os pontos pretos, que são usados para treinamento de modelos, representam o número de usuários no site em dado período, em azul escuro tem-se a previsão alcançada e em azul claro estão os valores mínimos e máximos previstos. Nota-se ainda que os dados de acesso (pontos pretos) vão até o ano de 2021, verificando-se que o período posterior se trata da previsão alcançada a partir do padrão encontrado nos dados dos anos anteriores. Importante ainda mencionar, que a Figura 3 foi gerada a partir de um conjunto de dados com finalidade acadêmica e está sendo apresentado a título de visualização de séries temporais.

Figura 3: Número de usuários em um site com base em seu histórico



Fonte: Os autores

Com esse contexto em mente, cabe ainda conceituar a ideia de detecção de anomalias para que, com essas informações, possam ser feitas as discussões pertinentes no próximo capítulo deste artigo.

3.1.3 Detecção de anomalias

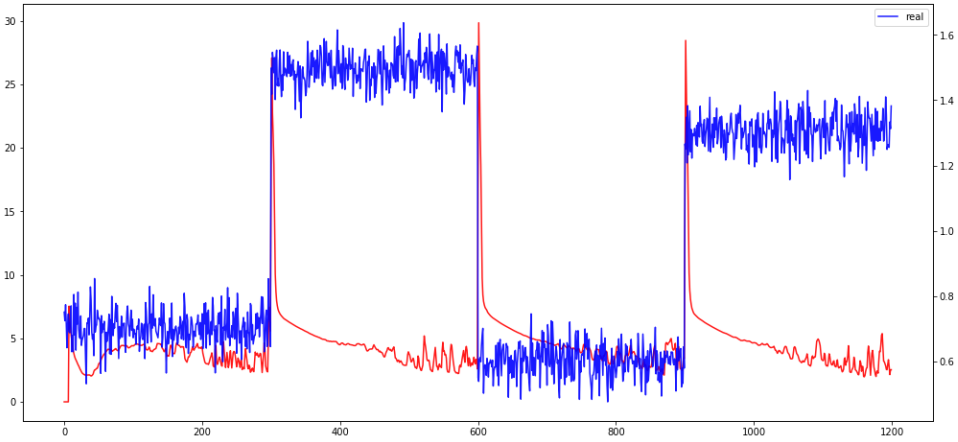
A detecção de anomalias ou detecção de *outliers* trata-se da identificação de variáveis discrepantes em uma base de dados. Para a aplicação desta técnica de Aprendizado de Máquina existem três abordagens, a saber:

- **Supervisionado**, em que os dados para a criação do conjunto de dados de treino devem ser estar previamente rotulados (normais e anormais).
- **Não supervisionado**, em que há um conjunto de dados que apresentam comportamento normal e, a partir destes, o algoritmo pode detectar àqueles dados anormais.
- **Semi supervisionado**, em que é construído um modelo representando o comportamento normal de um determinado conjunto de dados de treinamento normal, que na sequência é testado quanto a probabilidade de uma instância de teste ser gerada pelo modelo aprendido.

De forma geral, todas as técnicas para detecção de anomalias apresentam estrutura semelhante que consiste nas etapas de parametrização, treinamento e detecção. A detecção compara o modelo gerado no estágio de treinamento com a porção de dados selecionados para a parametrização. Critérios de limite serão selecionados para determinar instância de dados anômala (HAWKINS, 1980).

ISSN: 2595-5942

Figura 4: *Score* detecção de anomalias em série temporal



Fonte: Os autores

A Figura 4 apresenta o resultado gráfico da aplicação de um algoritmo de detecção de anomalias. Onde o que está em azul representa o padrão da série temporal real e em vermelho está o padrão anômalo.

Assim, tendo conceituado detecção de anomalias e séries temporais, passa-se então às discussões referentes ao algoritmo proposto para a detecção de anomalias em séries temporais aplicado a campanhas de *Marketing* Digital.

4 Resultados e discussão

Os dados utilizados neste artigo compõem um conjunto de dados de domínio público e com finalidade acadêmica disponíveis em Kaggle (2022) e serão utilizados para demonstrar o funcionamento do algoritmo de detecção de anomalias. Este conjunto de dados, de 150 exemplos, possui duas variáveis numéricas: X1 variável aleatória e X2 variável aleatória.

A título de abstração, imagine que este conjunto de dados representa duas medidas de teste para verificação do sucesso ou não de uma campanha de *Marketing* Digital. Além disso, considere a existência de uma base histórica com estas variáveis de outras campanhas, que já foram testadas e foram consideradas bem-sucedidas. A partir destas abstrações, a pergunta que se buscará responder é: “Essa nova campanha lançada poderia ser considerada uma anomalia?”. Para responder a essa pergunta, a seguinte metodologia será aplicada:

- 1º Modelar os dados de treino com uma distribuição específica.

- 2º A partir desta distribuição, calcular a probabilidade de encontrar essa nova campanha.
- 3º Se a probabilidade for baixa (menor do que um limiar fixado), então diz-se que a nova campanha é uma anomalia. Caso contrário, diz-se que não é uma anomalia.

Utilizando a base de dados de treino, X1 e X2 foram modelados com distribuições gaussianas, cada um com uma média e desvio padrão calculados a partir dos dados.

Agora, considerando um vetor bidimensional x = [x1, x2], pode-se calcular a densidade ou a probabilidade de se encontrar uma campanha com aquelas características (X1 e X2):

$p(x) = p\,(x1) \cdot p\,(x2)$

onde $p(x1)$ é a densidade de $x1$, de acordo com a gaussiana definida e analogamente para $x2$.

Desta forma, se $p(x)$ for menor do que um limiar ϵ, diz-se que o vetor x é uma anomalia. Caso contrário, não é uma anomalia. Assim, observa-se que este algoritmo pode ser generalizado e se for considerada a existência de uma base de dados com p variáveis X_1, ..., X_p, pode-se calcular ainda:

μ_1, σ_1: média e desvio padrão de X_1

μ_2, σ_2: média e desvio padrão de X_2

...

μ_p, σ_p: média e desvio padrão de X_p

E modelar:

$X_1 \sim Normal\,(\mu_1, \sigma_1)$, $X_2 \sim Normal\,(\mu_2, \sigma_2)$, ..., $Xp \sim Normal\,(\mu p, \sigma p)$

Então, definir:

$p\,(x) = p\,(x1) \cdot p\,(x2) \cdot \ldots \cdot p\,(x_p)$ para um vetor x = [x_1, x_2, ..., xp].

Assim, x será anomalia caso $p(x) < \epsilon$, sendo ϵ o limiar. Implementando em Python, que é uma linguagem de programação orientada a objetos, amplamente utilizada no contexto da Ciência de Dados, tem-se a classe apresentada na Figura 5.

ISSN: 2595-5942

Figura 5: Algoritmo de detecção de anomalias implementado

```
class DetectorAnomalias():

    def __init__(self, epsilon):
        self.epsilon = epsilon

    def fit(self, X):
        medias = X.mean(axis = 0)
        desvios = X.std(axis = 0)
        gaussianas = [st.norm(loc = m, scale = d) for m, d in zip(medias, desvios)]
        self.gaussianas = gaussianas
        self.X = X

    def prob(self, x):
        p = 1
        for i in range(self.X.shape[1]):
            gaussiana_i = self.gaussianas[i]
            x_i = x[i]
            p *= gaussiana_i.pdf(x_i)
        return p

    def isAnomaly(self, x):
        return int(np.where(self.prob(x) < self.epsilon, 1, 0))
```

Fonte: Os autores

A classe implementada é “DetectorAnomalias()”, que inicia-se com a declaração do método construtor parametrizado __init__, recebendo necessariamente o epsilon (limiar).

Na sequência, o método fit() receberá a base de dados (X) que conterá duas colunas, X1 e X2. Neste método serão calculadas as médias e os desvios padrão das duas colunas do *dataset.* Calcula-se também as distribuições gaussianas utilizando o scipy.stats(), instanciando uma distribuição normal que receberá “loc = m” como média e “scale = d” que será o desvio padrão, onde m e d estão variando dentro das médias e desvios padrões que acabaram de ser calculados. Dessa forma, tem-se uma lista de gaussianas, caso tenha-se “n” variáveis, serão construídas “n” distribuições gaussianas. Finalizando o método fit(), todas as gaussianas serão salvas e passa-se ao próximo método da classe.

O método prob(), calculará a probabilidade (p), recebendo uma nova instância qualquer (x), que no caso será um *array* (vetor com elementos pertencentes ao mesmo tipo de dados) de tamanho “n”. Inicialmente, atribui-se a variável p (probabilidade) o valor 1, pois por padrão a probabilidade deve sempre variar entre 0 e 1. Depois, para cada uma das gaussianas, será calculada a pdf correspondente, ou seja, a densidade daquele ponto, fazendo isso com cada coordenada em um *loop*. Pegando a i-ésima coordenada do vetor x (x_i) e por fim, calcular o p (x1), p

ISSN: 2595-5942

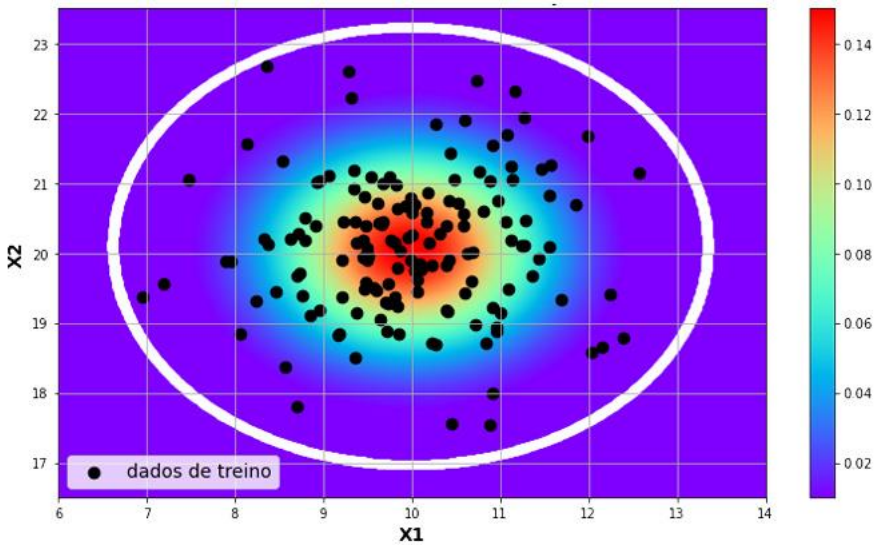
($x2$) etc., e ao final multiplicando todos estes valores para retornar o resultado (*return* p).

Por fim, foi construído o método isAnomaly(), para verificar se a variável é ou não uma anomalia, onde foi usada a probabilidade retornada no método anterior e caso essa probabilidade seja menor que o epsilon (limiar) pré-fixado, diz-se que sim, é uma anomalia; se for maior que epsilon, diz-se que não é uma anomalia.

O gráfico apresentado na Figura 6 representa um *grid* entre as variáveis X1 e X2 com vários pontos, e para cada ponto neste *grid* busca-se medir se aquela representação deve ser interpretada como uma anomalia ou não. Então, fixando o epsilon (limiar) em 0.001 tem-se, neste gráfico, no eixo x a variável X1 e no eixo y a variável X2, em preto o *scatter plot* dos dados de treino, ao lado do gráfico tem-se um *color bar* indicando qual a densidade de cada um dos pontos do *grid*, onde pontos mais afastados apresentam uma densidade muito baixa e pontos mais ao centro tem densidades mais altas, tendo em vista que a média da variável X1 ficou próxima de 10 e da variável X2 ficou próxima de 20.

Pode-se ainda verificar que foi colocada uma elipse em branco, indicando quais são os pontos em que a probabilidade, ou seja, a densidade estimada foi muito próxima do valor epsilon (limiar) fixado.

Figura 6: Detector de anomalias (limiar 0.001): probabilidades

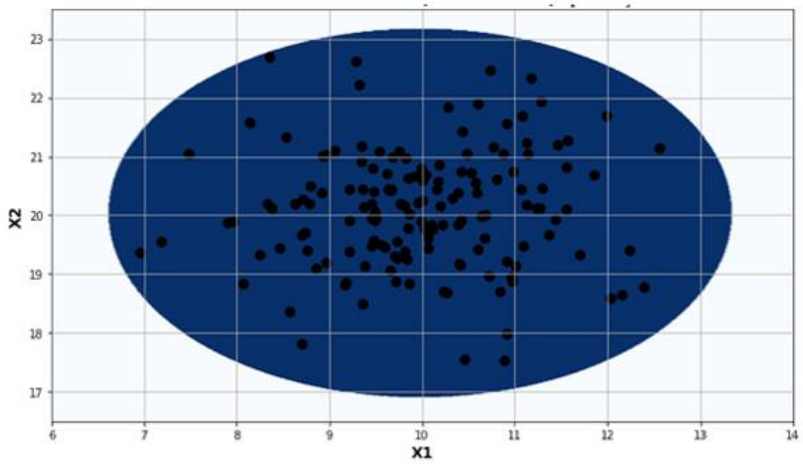


Fonte: Os autores

Em resumo, essa seria a margem em que se pode detectar se aquela coordenada é ou não é uma anomalia, assim, as coordenadas que estiverem além

da elipse branca, serão consideradas anomalias, as que estiverem dentro da elipse, serão consideradas normais e este é exatamente o *plot* da Figura 7, onde tudo o que está dentro da elipse azul não é uma anomalia e o que está fora, será caracterizado como uma anomalia.

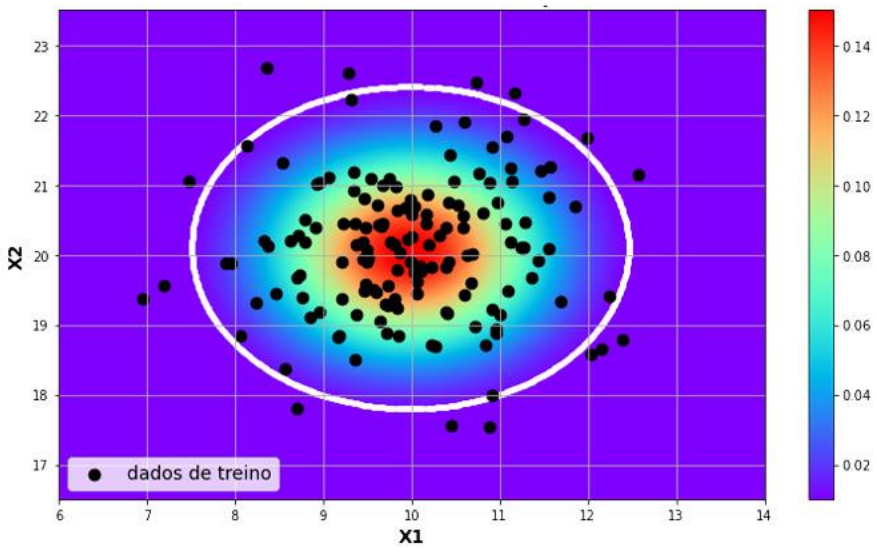
Figura 7: Detector de anomalias (limiar 0.001): predição



Fonte: Os autores

Alterando o limiar em 10x, conforme Figura 8, passando de um limiar de 0.001 para 0.01, verifica-se que a elipse branca se mostrou menor, não mais englobando todos os pontos, como na Figura 6, sendo os pontos pretos localizados além da elipse branca detectados como anomalias.

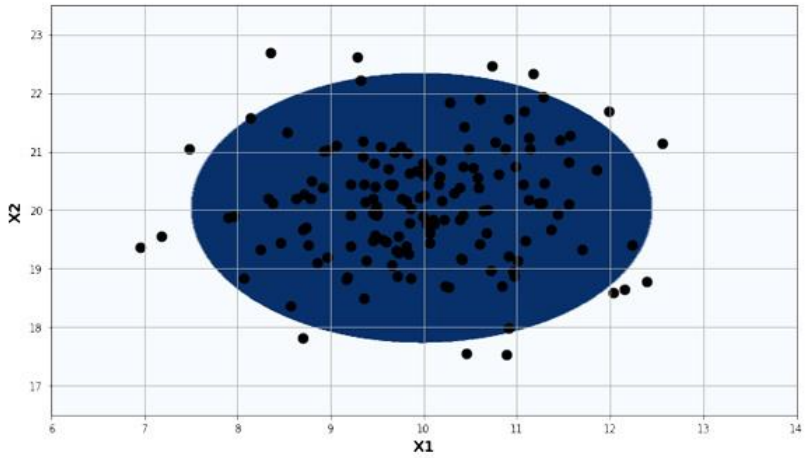
Figura 8: Detector de anomalias (limiar 0.01): probabilidades



Fonte: Os autores

Observa-se assim, que ao variar os valores do epsilon (limiar), encontram-se modelos completamente diferentes, sendo que no primeiro caso apresentado nenhuma anomalia foi detectada e neste modelo, demonstrado na Figura 9, uma vez variado o valor fixado para epsilon, tem-se a detecção de algumas anomalias. O que leva a pergunta, “qual é o critério para escolher o valor de epsilon (limiar) ótimo?”

Figura 9: Detector de anomalias (limiar 0.01): predição



Fonte: Os autores

Para escolher o valor de epsilon (limiar) ótimo, a seguinte metodologia pode ser considerada: suponha uma base de dados histórica com 10.000 campanhas normais, isto é, não são anomalias; e 50 campanhas que podem ser caracterizadas como anomalias. Poderia ser feita a seguinte divisão nos dados:

- ***Dataset* de treino**: 6.000 instâncias não anômalas.
- ***Dataset* de validação**: 2.000 instancias não anômalas e 25 instâncias anômalas.
- ***Dataset* de teste**: 2.000 instâncias não anômalas e 25 instâncias anômalas.

Neste contexto, os dados de treino seriam utilizados para calcular as médias e os desvios padrões de cada uma das variáveis explicativas, a fim de calcular p(x). Usando o conjunto de validação, poderia ser definido o melhor limiar ϵ e, finalmente, avaliar nos dados de teste se o modelo apresentado está bem generalizando.

Uma vez, sabendo-se que certas instâncias são anômalas e que certas outras não são, poderia se utilizar métricas clássicas de classificação binária, como: AUC

(Área Sob a Curva ROC), F1-score etc., verificando se o valor de épsilon fixado foi assertivo ou não.

O exemplo descrito pode ser interpretado como um problema de classificação binária, visto que se tem a variável *target* (marcação de anomalia) com exemplos positivos (anomalias) e negativos (não anomalias). Desta forma, teoricamente, é possível treinar modelos supervisionados e tentar reconhecer as anomalias. No entanto, alguns pontos de atenção precisam ser levantados:

- 1º Caso se tenham poucos exemplos da classe positiva, pode ser complicado que um algoritmo supervisionado aprenda o padrão dos dados. Neste caso, a abordagem não supervisionada pode ser interessante.
- 2º Podem existir vários tipos diferentes de anomalias. Novamente, com poucos dados rotulados, pode ser difícil detectar todos os padrões.

Em conclusão, o que se observa é que é necessário atenção ao problema em particular que se deseja verificar, não existindo um único método para a verificação, cabe encontrar a alternativa que melhor se encaixe na resolução do problema proposto. Porém, observando estas indicações, o algoritmo apresentado se mostra como uma possibilidade para a classificação do sucesso ou não de campanhas de *Marketing* Digital, podendo ser implementado com poucas alterações no que se refere à sua estrutura.

Considerações finais

Inicialmente buscou-se contextualizar o problema da mensuração de resultados no *Marketing* Digital e quais os desafios encontrados atualmente pelos profissionais da área no processo de otimização de campanhas tendo em vista a crescente no volume de dados gerados e coletados na divulgação dessas campanhas, apresentando assim a problemática desse artigo. Neste sentido, buscou-se ainda conceituar termos como *Big Data* e o consequente surgimento da Ciência de Dados, área que tem se tornado cada vez mais estratégica para as empresas com o objetivo de processar e analisar o grande volume de dados coletados, transformando-os em informação útil na tomada de decisão.

Nos capítulos que se seguiram, iniciou-se a abordagem de conceitos de Aprendizado de Máquina, fomentando a utilização de suas técnicas em processos

de automatização de processos de processamento e análise de dados, principalmente voltadas ao *Marketing* Digital.

Após a conceituação dos temas pertinentes, apresentou-se o algoritmo proposto para a automação do processo de análise de dados por meio da detecção de anomalias em séries temporais como forma de identificar o sucesso ou não de campanhas de *Marketing* Digital.

O algoritmo construído foi apresentado por meio da utilização de uma base de dados com finalidade acadêmica e os resultados apresentados demonstram a sua capacidade de detectar campanhas anômalas em uma dada série temporal. Para fins de aprofundamento, buscou-se ainda apresentar outra possibilidade de aplicação do algoritmo, agora com uma abordagem supervisionada bem como quais seriam os benefícios e pontos de atenção para as aplicações sugeridas.

Observa-se ainda que o algoritmo apresentado, sofrendo apenas algumas mudanças em sua estrutura, poderia ser utilizado em um problema de classificação não binária, com mais de duas variáveis a serem analisadas.

Neste sentido, conclui-se que este trabalho poderia ainda desdobrar-se na aplicação do algoritmo apresentado em uma base de dados real, promovendo um estudo de caso mais amplo sobre os resultados obtidos a partir de sua implementação. Para a proposta deste estudo, que era o desenvolvimento e proposição do algoritmo de detecção de anomalias e sua aplicação no *Marketing* Digital, acredita-se que os objetivos foram alcançados.

Referências

ABHISHEK, V.; DESPOTAKIS, S.; RAVI, R***. Multi-channel attribution: The blind spot of online advertising***, 2017. Disponível em: https://ssrn.com/abstract=2959778. Acesso em: 30 set. 2022.

CARVALHO, A.C.P.L.F. Interdisciplinaridade da Ciência de Dados. **Computação Brasil**, v.31, p.62-65, 2016. Disponível em: https://www.sbc.org.br/images/flippingbook/computacaobrasil/computa_31/Comp_Brasil_02_2016.pdf. Acesso em: 29 set. 2022.

CCSDS. CONSULTATIVE COMMITTEE FOR SPACE DATA SYSTEMS. ***Reference Model for an Open Archival Information System (OAIS)***. 2012. Disponível em: https://public.ccsds.org/pubs/650x0m2.pdf. Acesso em: 27 set. 2022.

DAVENPORT, Thomas H. **Ecologia da Informação**. São Paulo: Futura, 1998.

GANTZ, John; REINSEL, David; RYDNING, John. ***The Digitization of the World***. 2018. Disponível em: https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf. Acesso em: 27 set. 2022.

GOT IT AL, **Tipos de *machine learning***, sd. Disponível em: https://www.programaria.org/wp-content/uploads/2020/04/01_Conceitos_Tipos_Machine_Learning.jpg. Acesso em: 27 set. 2022.

HAWKINS, D. **Identification of Outliers**. Monographs on applied probability andstatistics. Chapman and Hall, 1980.

HJØRLAND, Birger. ***Data (with big data and database semantics),*** 11/10/2018. Disponível em: http://www.isko.org/cyclo/data. Acesso em: 30 set. 2022.

KAGGLE. ***Kaggle: Your machine learning and data science community***, *sd*. Disponível em: https://www.kaggle.com. Acesso em: 30 set. 2022.

KNIGHT, Kevin; RICH, Elaine; NAIR, Shivashankar B. ***Artificial Intelligence***. 3 ed. New Delhi: Tata McGraw Hill, 2009.

LUGER, George F. **Inteligência artificial**. 6 ed. São Paulo: Pearson Education do Brasil, 2013.

MIKLOSIK, A; KUCHTA, M.; EVANS, N; ZAK, S. ***Towards the adoption of machine learning-based analytical tools in digital marketing***, IEEE Access, vol. 7, pp. 85705–85718, 2019.
RUSSEL, Stuart; NORVIG, Peter. **Inteligência Artificial**. Tradução da 3ª ed. Rio de Janeiro: Elsevier, 2013.

SOLVIMM, **O que é Inteligência Artificial (IA)**, 03/01/2019. Disponível em: https://solvimm.com/blog/o-que-e-inteligencia-artificial/. Acesso em: 27 set. 2022.

STEVENSON, Willian J. **Estatística Aplicada à Administração**. – São Paulo: Harbra, 2001.