Universidade Federal de Campina Grande
Centro de Engenharia Elétrica e Informática
Coordenação de Pós-Graduação em Ciência da Computação

Paul Monthaler

PROCESSO INTERDISCIPLINAR DE ANÁLISE E CLASSIFICAÇÃO DE DADOS: UM ESTUDO DE CASO EM E-MAIL MARKETING.

Campina Grande, Paraíba, Brasil
2018

Paul Monthaler

PROCESSO INTERDISCIPLINAR DE ANÁLISE E CLASSIFICAÇÃO DE DADOS: UM ESTUDO DE CASO EM E-MAIL MARKETING.

Dissertação submetida à Coordenação do Curso de Pós-Graduação em Ciência da Computação da Universidade Federal de Campina Grande – Campus I como parte dos requisitos necessários para obtenção do grau de **Mestre em Ciência da Computação.**

Orientador: Prof. Dr. Kyller Costa Gorgônio

Orientador: Prof. Dr. Frederico Moreira Bublitz

Campina Grande, Paraíba, Brasil

2018

M789p Monthaler, Paul.
Processo interdisciplinar de análise e classificação de dados : um estudo de caso em e-mail marketing / Paul Monthaler. - Campina Grande-PB, 2018.
91 f. : il. color.

Dissertação (Mestrado em Ciência da Computação) - Universidade Federal de Campina Grande, Centro de Engenharia Elétrica e Informática, 2018.
"Orientação: Prof. Dr. Kyller Costa Gorgônio, Prof. Dr. Fred Moreira Bublitz".
Referências.

1. Aprendizado de Máquina. 2. Ciência da Computação. I. Gorgônio, Kyller Costa. II. Bublitz, Fred Moreira. III. Título.

CDU 004.738.5:339(043)

FICHA CATALOGRÁFICA ELABORADA PELO BIBLIOTECÁRIO GUSTAVO DINIZ DO NASCIMENTO CRB - 15/515

"PROCESSÓ INTERDISCIPLINAR DE ANÁLISE E CLASSIFICAÇÃO DE DADOS: UM ESTUDO DE CASO EM E-MAIL MARKETING"

PAUL MONTHALER

DISSERTAÇÃO APROVADA EM 31/08/2018

KYLLER COSTA GORGÔNIO, Dr., UFCG
Orientador(a)

FREDERICO MOREIRA BUBLITZ, Dr., UEPB
Orientador(a)

EANES TORRES PEREIRA, Dr., UFCG
Examinador(a)

DANILO FREIRE DE SOUZA SANTOS, Dr., UFCG
Examinador(a)

CAMPINA GRANDE - PB

Agradecimentos

A Deus, pela existência.

Aos meus pais, Regina e Toni, pelas raízes, educação e valores.

A minha amada esposa, Larissa, pelas asas, incentivos e a paciência.

A Lorenzo, pelos sorrisos que me encorajaram.

Aos meus orientadores, pela oportunidade.

Aos colegas do projeto de análise de dados, pelo compartilhamento desta trajetória.

A todos que contribuíram direta e indiretamente para a conclusão desse trabalho.

A todas as dificuldades que enfrentei; não fosse por elas, eu não teria saído do lugar.

Muito obrigado!

Resumo

Conseguir respostas com Aprendizado de Máquina não é reduzido à pura matemática e estatística. As informações necessárias proveem dos especialistas do universo dos dados coletados, aqueles com competências no ambiente de aplicação dos resultados e de cientistas de computação. Nesse sentido, acredita-se que combinação de conhecimento das ferramentas, compreensão da área dos dados coletados e uma base de estudos na aplicação deles, consegue entregar melhores resultados. O problema é que os processos atuais falham na combinação dessas áreas. Assim, propõe-se nesse trabalho um processo, juntando os resultados de entendimento do negócio, análise do problema, coleta, análise e classificação de dados. Para validação, foi desenvolvido um estudo de caso em campanhas de e-mail marketing. O resultado da aplicação do processo é uma combinação entre aumento da visibilidade, retenção de assinantes e maximização do retorno. Estudos de marketing revelam o aumento do resultado com o aumento da frequência e personalização das mensagens, enquanto a ética e sociologia medem a percepção de invasão de privacidade com o aumento da personalização indevida.

Abstract

Solutions with Machine Learning are too often reduced to pure math and stats. But the information required for the solution comes from professionals from the context of the collected data, those with expertise in the environment where the solution is applied, and from computer scientists. This work suggests the combination of the different core competences and an interdisciplinary approach to achieve the best possible result. The proposed process will then be applied in a filed study of e-mail marketing. The result of the process application is a combination of increased visibility, retention of subscribers, and maximization of return. Marketing studies reveal the increase in the outcome with increased frequency and personalization of messages, while ethics and sociology measure the perception of invasion of privacy with increasing inappropriate personalization.

Lista de Figuras

Lista de Tabelas

Lista de Abreviações

AED	Análise Exploratória dos Dados
ARF	Abuse Reporting Format
BR	Bounce Rate
CPF	Cadastro de Pessoas Físicas
CAPEM	Código de Auto-regulamentação para Prática de E-mail Marketing
CRISP-DM	CRoss-Industry Standard Process for Data Mining
TFD	Teoria Fundamenta nos Dados
DAA	Digital Advertising Alliance
DMARC	Domain-based Message Authentication, Reporting, and Conformance
DKIM	DomainKeys Identified Mail
EM	E-mail Marketing
EDAA	European interactive Digital Advertising Alliance
IBGE	Instituto Brasileiro de Geografia e Estatística
ISP	Internet Service Provider
RG	Registro Geral
RGPD	Regulamento Geral de Proteção de Dados
ROI	Return Of Investment
SPF	Sender Policy Framework
SEMMA	Sample, Explore, Modify, Model and Assess
SSE	Sum of Squared Error

Sumário

Capítulo 1 – Introdução

Aprendizado de Máquina é um hiperônimo para soluções que usam algoritmos para analisar dados, encontrar padrões e determinar ou prever eventos. Já na primeira definição do termo, o autor escondeu a armadilha da área: obter automaticamente uma solução superior ao nível de conhecimento do programador – e tudo quase sem programar (SAMUEL, 1959). Infelizmente, essa visão da magia, dos resultados sem conhecimento da área, em que o especialista recebe os dados, o algoritmo faz tudo sozinho, e o cliente recebe os resultados desejados, é muito resistente. Mas, Aprendizado de Máquina não é mágica; ela consegue criar maior quantidade de pouca informação, mas não consegue tirar algo do nada (DOMINGOS, 2012).

O artigo mais famoso sobre mineração de dados, "*From Data Mining to Knowledge Discovery: An Overview*" usa um processo que inicia com a compreensão do domínio da aplicação e do conhecimento prévio relevante, necessários para identificar o objetivo do ponto de vista do cliente (FAYYAD et. Al., 1996). Mas, parece que essa primeira etapa caiu no esquecimento. Hoje, a informação é buscada diretamente nos dados. Livros sobre Aprendizado de Máquina e Análise de Dados costumam iniciar diretamente observando os dados que estão à disposição. Distribuição dos dados, histogramas, frequências e séries temporais podem ser muito reveladoras. Mas, somente os dados ordenados e agrupados, interpretados em relação a um determinado contexto, podem virar informação. Entender o contexto é cada vez mais importante na área de análise de dados (PENG, 2018). Sem contexto, os dados não têm – ou mudam – de valor. Um ótimo exemplo para isso é a observação do céu estrelado (YAU, 2013). Olhando para cima, é possível ver as estrelas como pontos numa superfície plana. Foi assim, que nasceram as figuras das constelações: conectando os pontos das estrelas representadas num plano. Mas, na verdade, cada um desses pontos é muito longe do outro, as distâncias podem chegar a anos luz. Na dimensionalidade, as constelações desaparecem, não são mais identificáveis.

O objetivo desse trabalho é mostrar um processo de análise de dados, um caminho que inicia com o entendimento do contexto e termina com a escolha do algoritmo de Aprendizado de Máquina. O processo será aplicado num estudo de caso. A ausência de um cliente verdadeiro, e um estudo de caso baseado em pesquisa bibliográfica, não permite avaliar a preparação e entrega dos resultados. Portanto, o processo apresentado não cobre essas duas

partes.

As fases que conduzem os dados puros até a sabedoria, na análise de dados, podem ser resumidas em poucos passos: coleta de dados, análise exploratória dos dados, análise estatística, limpeza, transformação e normalização, seleção das variáveis, seleção do tipo de algoritmo (preditivo, de classificação), ajuste, regularização do modelo, e, no final, a aplicação do modelo e interpretação dos resultados (MOHANTY, et. Al., 2013).

A contextualização fornece as informações sobre a área das perguntas, o domínio de ação e seu âmbito. Com ela, é possível reduzir o problema em partes, coletar os dados necessários, criar as hipóteses para cada uma delas, e no final, entregar uma resposta. Entretanto, a imersão no contexto pode ser bastante penosa, como será mostrado no exemplo da aplicação do processo no estudo de E-mail Marketing, pois precisa passar pela definição da área, entender seus indicadores usados, para compreender o problema e entender um caminho possível para a solução. Além disso, é necessário definir qual a participação dos stakeholders, como eles são envolvidos no processo de negócios, na definição do problema e o tipo de resultado esperado.

O restante do trabalho está organizado da seguinte maneira: o Capítulo 2 consiste na fundamentação teórica. Serão apresentadas, de maneira concisa, as diferentes etapas de um processo de análise de dados. Em seguida, serão identificados alguns dos processos mais usados na área e suas diferenças. Finalmente, será criado um processo que junta partes mais interessantes dos processos mencionados, criando a pipeline de análise de dados. No Capítulo 3, o processo será aplicado em um estudo de caso. Para isso, será feita uma extensa pesquisa bibliográfica da área de e-mail marketing, seguida pela redefinição do problema, a coleta e a análise de dados. Os problemas citados são expostos na aplicação em um problema real, da área de e-mail marketing. Capítulo 4 usará as informações coletadas no capítulo anterior, para criar a solução. Finalmente, no Capítulo 5, serão feitas as considerações finais, além de serem apontadas as dificuldades encontradas na aplicação.

Em termos de sua classificação, o trabalho será baseado em uma pesquisa bibliográfica, exploratória e aplicada.

Capítulo 2 – Pipeline de Análise de Dados e Aprendizado de Máquina

Aprendizado de máquina permite a automatização de construção de modelos analíticos. Praticamente todos os programas de inteligência artificial são construídos para solucionar um problema baseado em dados (GILHOOLY, 1989). Existem diferentes processos para conseguir a solução. Para entender melhor as soluções encontradas, é necessário descrever as possíveis etapas básicas. Embora já existam processos que tratem de cada uma dessas etapas, isso não é feito em um processo único. A seguir descrevemos as principais etapas do processo, necessárias para maximizar os resultados obtidos pelos modelos de Análise de Dados e Aprendizado de Máquina.

2.1. O contexto

Aprendizado de Máquina oferece os algoritmos, as ferramentas para solucionar os problemas, mas só o contexto oferece o conhecimento necessário para a interpretação de dados e resultados (TUKEY, 1977). As características usadas para os algoritmos podem ser divididas em três categorias: primários, contextuais e irrelevantes (TUMEY, 2002). Enquanto características irrelevantes não contribuem para o resultado, no máximo conseguem atrapalhar o caminho para o melhor resultado; as primárias são úteis quando são consideradas isoladamente. As características contextuais são importantes para a combinação de dados diferentes. Algoritmos de Aprendizado de Máquina conseguem extrair correlações entre os dados fornecidos, mas não tem o potencial de entender os fatos ligados aos dados elaborados. Infelizmente, a informação contextual é frequentemente omitida nos dados, e nem sempre é possível recuperar ou inferir informação contextual (TUMEY, 2002). O uso do contexto em soluções de Aprendizado de Máquina mostra efeitos positivos nos resultados, pode ajudar a selecionar variáveis certas quando a disponibilidade de dados é esmagadora (BRÉZILLON, 1999). O contexto não é limitado à informação vertical, focada no problema atual, mas envolve também informação horizontal, resultados de análises anteriores, similares, teorias e referências da área.

2.2. Definição do problema

Depois de conhecer o contexto, as circunstâncias, é possível enquadrar o entendimento do problema. Existem as mais variadas formas de problemas, e igualmente numerosas definições para cada tipo. Uma formulação conhecida é do psicólogo alemão Karl DUNCKER, 1945 : “Um problema existe quando um organismo vivo tem um objetivo, mas não sabe como chegar lá”. E, normalmente, Aprendizado de máquina quer resolver exatamente isso. Coletar e usar os dados para encontrar soluções, que permitam chegar a um objetivo definido. Para ROBERTSON, 2001, é possível distinguir entre os diferentes tipos de problemas resumidos na Tabela 1.

Tabela 1: Classificação de problemas

Categoria	Descrição
Bem definido	Todas as informações são fornecidas diretamente ou indiretamente (é possível inferir o que falta)
Mal definido	Alguns aspectos não estão definidos, informações vagas, faltantes.
Problema de dedução	Solução depende de um resultado anterior.
Conhecimento enxuto	Pouco conhecimento prévio necessário para solucionar o problema.
Conhecimento rico	Solução necessita de grande conhecimento prévio, como conhecimento de domínio.
Semântica enxuta	Quem tenta resolver, não tem experiência nesse tipo de problema.
Semântica rica	Quem está resolvendo o problema pode contribuir muito com a experiência nesse tipo de problema

FONTE: (ROBERTSON, 2001)

A principal distinção entre problemas está na definição. Problemas bem definidos proveem todas as informações necessárias para a solução, e os problemas mal definidos não tem nada além da definição e da solução desejada. Um exemplo de problema bem definido é a soma de dois números. No caso da análise de dados, ou Aprendizado de Máquina, mesmo um problema bem definido não implica no caminho certo, nas ações aplicadas para as diferentes variáveis, na escolha das ferramentas ou do algoritmo.

No início, o problema é o problema. Sua solução inicia com sua definição e metas das ações (operações) para resolver o problema, ou da série de problemas intermediários, das restrições previstas ou encontradas ao longo desse caminho e termina quando alcançar a meta (ROBERTSON, 2001). Examinando a classificação de problemas, é possível entender que sua representação inicial depende da definição do problema, da hipótese, das informações existentes, do caminho delineado para chegar até a meta, e do nível de experiência e conhecimento necessário das partes envolvidas. Cada uma das atividades necessita dos seus dados, das suas informações, ferramentas e perspectivas. Para chegar nisso, é necessário dividir o problema em partes. Cada parte terá a sua pergunta, os entregáveis, *stakeholders* e tipo(s) de resultado(s) esperado(s). Assim, solucionar um problema não é uma atividade isolada, mas uma série de ações correlacionadas, em que cada ação tem suas próprias características, um estado inicial, e um estado desejado (GILHOOLY, 1989).

Davenport dedica uma atenção especial para os stakeholders (DAVENPORT, 2013). Ao final, são eles que definem o problema, vão receber e analisar os resultados. Para o autor, é necessário identificar todos os interessados, com as próprias necessidades, interesses, expectativas e a influência deles.

2.3. Os dados

Definido o problema, é necessário entender os dados à disposição. A qualidade desses dados pode ser medida através dos quatro C definidos por GLEASON E MCCALLUM, 2012. Traduzido para o português, os dados devem ser completos, coerentes, corretos e censuráveis. Cada categoria responde a uma pergunta específica (Tabela 2).

Tabela 2: Categorias Qualidade de dados

Categoria	Pergunta associada
Completos	Tem todas as informações previstas?
Coerentes	As informações encontradas batem?
Corretos	Os valores encontrados estão certos?
Censuráveis	É possível rastrear os dados?

FONTE: (GLEASON E MCCALLUM, 2012)

Quando a coleta dos dados não faz parte do trabalho, isso implica uma série de

problemáticas inerentes: onde os dados foram coletados, qual a população, qual a frequência da coleta, como funciona o coletor usado, qual o valor de cada informação e qual a relação entre eles.

Durante cursos introdutórios de estatística, os estudantes são aproximados a métodos analíticos, teste de hipóteses, regressão, a toda a parte matemática e conceitual; na aplicação em dados reais, o foco passa do conteúdo deles para os objetos que eles representam, as correlações entre eles, e se a informação contida faz sentido (TUKEY, 1977).

2.4. Análise Exploratória dos Dados

Antes de iniciar com a análise estatística, é necessário fazer uma a Análise Exploratória dos Dados (AED) (STELMAN, 2018). A AED é um trabalho de detetive, de investigação, para encontrar e revelar as pistas contidas nos dados (TUKEY, 1977). A AED permite verificar se as informações recebidas sobre a natureza dos dados coincidem. Raramente, os dados são disponibilizados de forma adequada, muitas vezes a origem deles não é clara, falta documentação e informações sobre suas características (FINK, 2012).

Assim, é possível controlar a consistência dos dados baseado na natureza definida deles. Por exemplo, o Código de Endereçamento Postal brasileiro deve ser um número de cinco dígitos, sem digitais, o nome uma sequência de caracteres alfabéticos sem números, a idade de uma pessoa um valor positivo entre 0 e 130 anos. Todas as inconsistências encontradas devem ser informadas e solucionadas. Assim, é possível excluir ruído, informação errada, ou mesmo encontrar problemas com a coleta dos dados.

A AED pode ser dividida em duas grandes fases (STELMAN, 2018): a representação numérica e gráfica, em que cada uma pode ser para uma variável ou para mostrar relações de duas ou mais variáveis. Cada fase diferencia entre variáveis quantitativas (discretas e contínuas) e categóricas (nominais e ordinais). Variáveis quantitativas contêm valores numéricos que representam uma magnitude, enquanto as categóricas permitem uma classificação. As variáveis categóricas podem ser representadas por números. Nesse caso, é necessário ter uma tabela de referência, que permite traduzir o número usado para a categoria da variável.

2.4.1. Representação numérica

Dados categóricos contém um conjunto finito de valores. A primeira verificação mostra a frequência, a distribuição (contagem, proporção, porcentagem). Com isso, é possível encontrar estranhezas, valores faltantes ou valores fora do conjunto definido.

Para variáveis quantitativas, é interessante encontrar a amplitude, valores máximos e mínimos, o centro da distribuição, e valores atípicos. As métricas mais usadas são média, mediana, desvio padrão e coeficiente de variação (DEAN, 2014). Essas primeiras métricas estatísticas permitem entender se os dados encontrados fazem sentido, se o valor máximo ou mínimo de uma variável é plausível para aquele contexto (FINK, 2012).

2.4.2. Representação gráfica

Os números não mentem. Mas, quando é necessário ver o comportamento de uma variável e a relação entre as observações, a representação gráfica da distribuição pode ser uma grande ajuda (STELMAN, 2018).

Para dados categóricos, a representação gráfica mais usada é um histograma, mostrando a frequência ou proporção para cada valor encontrado. Para dados quantitativos, um *boxplot* ou *scatterplot* consegue mostrar a tendência central, os valores anômalos, e a dispersão ou simetria dos dados (YAU, 2013).

2.4.3. Série temporal

Se os dados representam uma série temporal de valores, é conveniente observar o comportamento das variáveis ao longo do tempo. Nesse contexto, é indispensável ter as informações sobre a coleta dos dados. Existem situações, em que dados são coletados somente em situações definidas, ou quando os valores coletados mudam.

A AED serve como base para todo o trabalho sucessivo. Ela é necessária para verificar todas as informações recebidas sobre a natureza dos dados. Qualquer situação que não corresponde às específicas necessita de esclarecimento.

2.5. Tratamento dos dados

O tratamento de dados consiste em quatro passos (TUKEY, 1977):

- Remover dados indesejados
- Consertar erros estruturais
- Filtrar valores atípicos
- Substituir dados perdidos (faltantes)

Antes de proceder a qualquer tratamento de dados, é importante definir a hipótese para a ação, o procedimento escolhido e a fundamentação que está por trás das escolhas. Cada passo precisa de uma descrição da situação antes do tratamento e o resultado de sua aplicação. Com essa documentação, é necessário consultar o dono dos dados, o especialista da área e passar com ele cada etapa. Isso permite ter maior confiança no tratamento, impede, com grande probabilidade, a perda de dados importantes e sua falsificação.

A remoção de dados indesejados pode ser feita por dois motivos. Na combinação de dados de diferentes fontes, é possível obter duas ou mais colunas com a mesma informação. O segundo motivo é a remoção de dados irrelevantes. Considera-se irrelevante um dado coletado, que não tem relação com o problema atual.

Erros estruturais normalmente são problemas relacionados a variáveis categóricas. As causas para esses erros são variados. A coleta de dados editável, a internacionalização de valores, a junção de fontes diferentes. Isso pode resultar em nomes diferentes para as mesmas classes, problemas de capitalização ou ortográficos. Outro problema estrutural pode ser o uso de idiomas diferentes para nomear a mesma categoria.

A filtragem de valores atípicos apresenta uma situação delicada. Raramente, existem razões para esse procedimento. Uma justificativa pode ser um problema com uma versão do coletor, com um tipo de respondente, seja este uma máquina, um sensor, uma pessoa ou um programa. Valores fora de um intervalo definido não são somente um problema de dados, mas em primeira instância uma complicação na coleta e devem ser consertados na fonte. Tipo ou frequência de valores atípicos podem até questionar a qualidade dos dados e impor uma nova coleta.

A substituição de dados é outra situação suscetível. Como no caso de valores atípicos, a substituição precisa de argumentos fortes e válidos para a aplicação. Cada intenção de

imputação deve ser referida aos especialistas dos dados e depende da origem dos dados, o uso que será feito e do resultado esperado. Como no caso de valores atípicos, é importante entender a causa da deficiência da informação. A solução mais simples pode ser a filtragem das observações incompletas. Um caso especial é constituído por dados coletados ao longo do tempo. Para alguns algoritmos, é necessário ter séries de coletas, uma para cada minuto, dia ou mês. Nesse caso, a falta da informação pode ser baseada na coleta. Como já mencionado na descrição de séries temporais, a arrecadação pode ser limitada para situações em que os valores coletados mudam. Nesse caso, é possível repetir o último valor coletado. Outras soluções aplicáveis seria a substituição da observação incompleta por outra completa, o uso da média dos valores observados ou o uso de um valor (randômico ou predefinido) de uma observação com valores parecidos nas outras variáveis, baseado em uma regressão, uma interpolação ou extrapolação.

Em geral, qualquer mudança de valores nos dados deve ser anotada diretamente nos dados. A maneira mais simples é uma coluna extra, com o uso de um valor booleano, que mostra se a observação foi imputada ou modificada. Isso permite filtrar essas observações e descobrir mudanças nos resultados baseados na inclusão ou exclusão delas.

Outra observação importante. Se for possível, é interessante entender os processos atrás da coleta de dados. Como, quando e com qual frequência os dados são coletados, qual o comportamento do coletor, caso falte uma informação na coleta, como as observações são registradas. Existem situações, em que coletores substituem dados faltantes com valores que estão dentro da amplitude prevista para os dados coletados. Por exemplo, coletar um 0 em lugar de um dado nulo, não existente, onde 0 é um dos valores que essa mesma variável pode assumir.

Seguidamente à elaboração de toda a documentação gerada, a aceitação ou rejeição das hipóteses geradas, e das ações previstas, pelos especialistas da área é necessário repetir a AED. Esse ciclo é repetido até encontrar somente os dados definidos, na amplitude certa ou com os valores previstos pela documentação.

2.6. Definição dos rótulos

Aprendizado de Máquina é dividido em duas grandes áreas: aprendizado não

supervisionado e supervisionado. Na versão não supervisionada, o algoritmo descreve dados existentes, encontrando valores estatísticos que conseguem descrever e agrupá-los, ou encontrando sobreposições de valores que permitem a redução dos dados usado (SPRING, 2015). Identificando as variáveis independentes, seria possível reduzir a quantidade de informações presentes. Assim, a análise dos componentes principais permite, em teoria, criar um novo reduzido arranjo de dados, que representa os dados originais pelos componentes principais. A análise dos componentes principais pode ser combinada com a verificação de correlações entre as variáveis. Os métodos mais usados são o coeficiente de Pearson, para dados com distribuição normal, ou para todos os outros o *Rho* de Spearman ou o *Tau* de Kendall.

Do outro lado, um sistema supervisionado toma decisões baseadas na experiência contida em exemplos solucionados com sucesso (SPRING, 2015). Para isso, é necessário definir rótulos, que permitem entender a natureza das linhas passadas para o algoritmo de Aprendizado.

A definição dos rótulos parece ser uma questão muito técnica, relacionada somente com a parte de Aprendizado de Máquina. Mas essa fase é a terceira grande armadilha no caminho de análise de dados, depois da descrição e compreensão do problema, e a análise e verificação dos dados. O rótulo apresenta a informação definida como *ground truth*. É a verdade que é usada para classificar as observações, para definir o que é verdadeiro ou falso. Um exemplo facilita o entendimento dessa verdade. O problema abordado é encontrar sensores antes de falharem. O rótulo define aqueles que falharam. Selecionando dados de sensores que falharam e outros que nunca falharam, o algoritmo pode aprender padrões encontrados nos dados. Com esses padrões, será possível encontrar sensores perto da falha para permitir uma troca proativa. Nesse caso, o *ground truth* marca os sensores que falharam. O que acontece, quando esse valor não é confiável? Quando os sensores marcados, pouco depois de uma primeira falha continuam a trabalhar, como se não estivesse com nenhum problema? O rótulo estaria comprometido, o algoritmo não conseguiria criar um modelo adequado.

A definição dos rótulos deve ser elaborada junto com os especialistas da área. São eles que podem descrever o comportamento, o valor das observações. Com as informações dos peritos, é possível identificar situações comprometedoras, estados não aceitáveis nos dados. O conselho para a criação dos rótulos é simples: não confiar no *ground truth*. Ou seja, usar todas

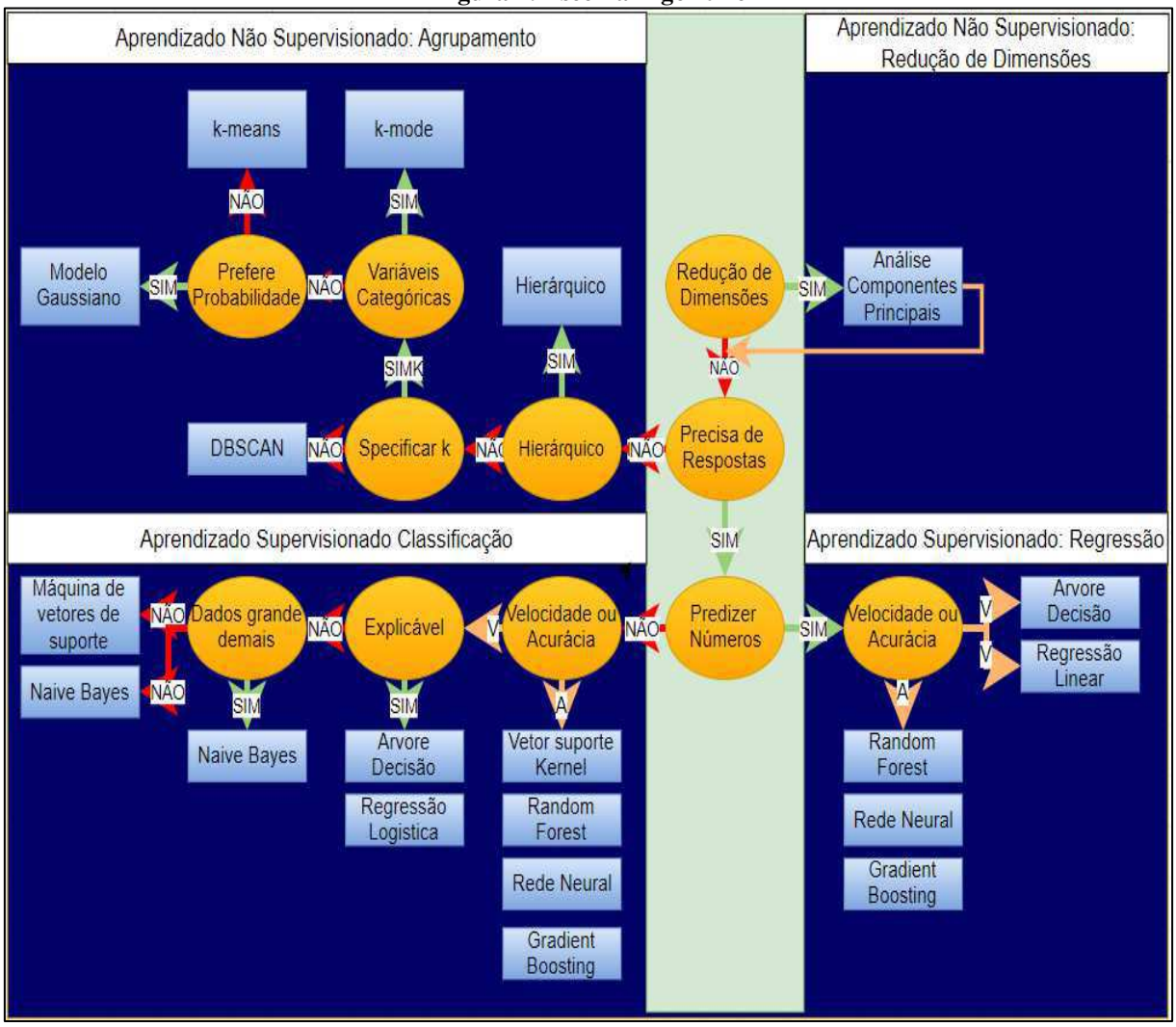
as possibilidades de representação de informação, para encontrar inconsistências nos rótulos. No caso dos sensores, poderia ser uma série temporal dos rótulos por sensor. Se a verdade, na qual foi baseada a seleção dele é fundamentada, não podem existir observações com o rótulo de “funcionante” depois de um acontecimento marcado como “falha”.

Suplementar à qualidade dos rótulos, é a sua distribuição. Quando a frequência de rótulos nos dados é mínima, será difícil para o algoritmo definir padrões para eles. Pode ser vantajoso sacrificar alguns rótulos, para melhorar a identificação de outros. Nesse caso, é necessário obter a aceitação para a escolha. Assim, não será o especialista, mas o stakeholder da pergunta, do problema específico, porque o abandono de um rótulo se equipara a exclusão desse problema.

2.7. Escolha do algoritmo

Raramente, o autor dos problemas se intromete na escolha do algoritmo. Mas, a escolha é definida indiretamente pelos requisitos não funcionais. Esses requisitos são os atributos de qualidade de um software e descrevem como a solução deve funcionar. No caso de Aprendizado de Máquina, existem três fatores importantes para a escolha do algoritmo (SPRING, 2015): a escolha entre acurácia e velocidade, a explicabilidade do resultado e a quantidade de dados. Por isso, é importante falar com o cliente sobre as possibilidades e limitações de cada modelo. Ao final, será o cliente a aplicar o modelo, por isso deve entender os problemas relacionados à quantidade (pouca, grande, excessiva) de dados, os requisitos necessários para a aplicação de modelos pesados e os custos relacionados ao poder de cálculo. E dependerá da aplicação dele, para entender qual tipo de resultado será imprescindível. Basta ter um número, uma probabilidade de um acontecimento, ou é necessário ter um resultado com uma boa explicabilidade, por exemplo, para possibilitar a busca da causa raiz de um problema? A Figura 1 mostra um panorama dos algoritmos de Aprendizado de Máquina – sem ser exaustiva – e o caminho para a escolha.

Figura 1: Escolha Algoritmo



FONTE: (Adaptado da LI, 2018).

Observando a figura, é possível perceber que antes de iniciar com a escolha do algoritmo, pode-se optar pela redução de dimensões. Isso pode ser uma análise dos componentes principais, para encontrar as variáveis com o mesmo valor informativo, sobre posicionadas a outras. Depois, se não precisar de respostas, pode ser usado um algoritmo de agrupamento, com um número definido ou ainda não conhecido de grupos, escolhendo a melhor maneira para o tipo de variável (categórica ou não). Se quisermos uma predição, vai depender se ela é feita para obter uma classificação ou um valor em um determinado instante. Como já mencionado no início, será os três fatores importantes, acurácia, explicabilidade e dimensão da base de dados, a definir a escolha do algoritmo.

2.8. Processo de análise de dados

A literatura mostra diferentes caminhos para uma análise de dados bem-sucedida. A primeira grande distinção pode ser feita entre abordagens com processo linear e outra com processo circular. Nas duas abordagens, as etapas são as mesmas (EAD, 2009):

- Propósito.
- Questões.
- Coleção dos dados.
- Procedimentos e métodos de análise de dados.
- Identificação e Interpretação dos resultados.
- Escrever, relatar e disseminar o resultado.
- Avaliação.

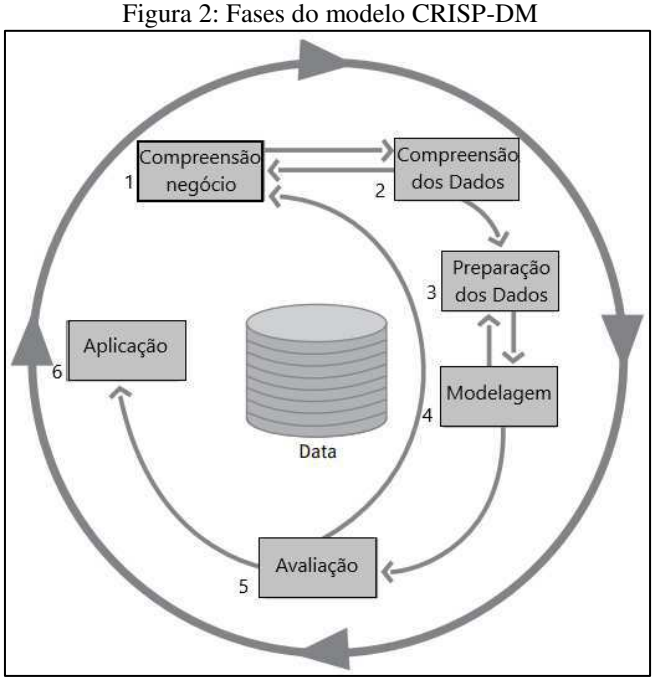
A primeira etapa, o propósito, é usada para definir o que fazer e as razões pelas quais o problema é importante. Para ser efetiva, uma análise de dados é funcional, ela adiciona valor. Para conseguir nesse intuito, é necessário juntar as pessoas certas para seguir todo o processo. Depois disso, é possível definir a pergunta. Com ela, será possível definir quais dados usar ou coletar, e qual tipo de análise aplicar. Na coleta dos dados tem uma parte crucial nesse processo. Dependendo do tipo de dados (observações, questionários, entrevistas, documentos, testes, etc.), é possível encontrar respostas para problemas diferentes. Na análise dos dados, será possível identificar as informações que eles têm. Dependendo do tipo de dado, diferentes níveis de análise serão aplicados. Obtido os resultados, será feita a interpretação e avaliação deles. Pode ser necessário juntar informação adicional, para facilitar a compreensão dos resultados. Juntado toda essa informação, será necessário escrever a história que os dados contam, da coleta, análise, interpretação dos resultados. O resultado deve ser mostrado para os stakeholders. No final, será feita uma avaliação de todo o projeto, para evitar repetir os mesmos erros na próxima vez, ou confirmar definições feitas anteriormente.

Davenport e Kim, 2013 usam uma abordagem linear, mas preveem a revisão dos resultados anteriores na etapa seguinte. Eles enumeram três etapas macro, o enquadramento do problema, a solução dele e a apresentação dos resultados. Pelo escopo do estudo, a análise será limitada às primeiras duas etapas. Os autores dividem o enquadramento do problema em reconhecimento do problema (entender completamente qual é o problema e a sua

importância) e a revisão das informações obtidas na etapa anterior. Traduzido para as etapas básicas descrita na contextualização, eles iniciam com o entendimento do contexto, passam para a definição do problema, com identificação dos stakeholders, do escopo do problema e a especificação do que deve ser descoberto. Depois disso, todas as informações coletadas são questionadas e procuradas evidências sobre a sua validade. A coleta de dados, a AED, o tratamento dos dados, a definição dos rótulos e a escolha do algoritmo, são resumidos na solução do problema.

Um dos modelos mais usados para mineração de dados é conhecido pela abreviação CRISP-DM (*Cross-industry standard process for data mining*) (CHAPMAN et. Al., 2000). Esse modelo hoje é tão conhecido, que raramente são citados os autores. As seis fases e as relações entre elas são representadas e enumeradas na Figura 2.

Figura 2: Fases do modelo CRISP-DM



FONTE: (EAD, 2009) modificado.

A primeira fase consiste na compreensão do negócio. Como nos outros modelos citados, é a etapa reservada para o contexto, a área do negócio, os stakeholders, o objetivo do projeto e os critérios do sucesso. Resumindo, a formulação do documento que define a primeira versão do CRISP-DM (CHAPMAN et. Al., 2000), é necessário ver o problema da perspectiva do negócio, para depois transformar as informações obtidas em um problema de mineração de dados. Fase dois, a compreensão dos dados, inclui toda a parte de coleta dos dados e da AED. Eventuais incongruências nos dados devem ser explicadas, por isso, o

gráfico prevê um apontador para a parte de compreensão dos negócios. A terceira e quarta fase, preparação e modelagem dos dados resume todos os passos necessários partindo dos dados coletados e entregando a base de dados final, que será usada para os próximos passos. No caso de aprendizagem de máquina, a modelagem dos dados seria contida na parte de preparação dos dados, e substituída pela parte de seleção e aplicação do modelo. Fase cinco faz a avaliação dos resultados e do processo usado e determina os próximos passos. Finalmente, na última fase, é definido o plano de aplicação dos resultados, de monitoramento e manutenção, a criação dos entregáveis finais e a revisão do projeto, para ser usada como retroalimentação nos próximos projetos.

Para completar as abordagens mais usadas para mineração de dados, é importante mencionar SEMMA (*Sample, Explore, Modify, Model and Assess*), introduzido pela empresa SAS (SAS, 2018). Os cincos passos de SEMMA são amostrar, explorar, modificar, modelar e avaliar os dados. Comparado com os modelos até agora citados, falta toda a parte de contexto, problema e coleta dos dados. Essa lacuna mostra, que o SEMMA, mesmo sempre citado, quando o assunto é mineração de dados, nada mais é que uma abordagem usada no software da empresa SAS, mas que não pode ser enxergada fora desse contexto (ROHANIZADEH, S. S., MOGHADAM, M. B., 2009).

Os processos encontrados têm passos parecidos, com granularidade e nomes diferentes. Para padronizar o resumo (Tabela 3), foram aplicadas as nomenclaturas usadas na fundamentação teórica.

Tabela 3: Resumo dos processos de análise de dados

Fonte	Contexto	Problema	Dados	AED	Tratamento	Rótulos	Algoritmo
AED	Propósito	Questões	Coleta	Procedimentos e métodos de análise de dados			
Davenport & Kim	Reconhecimento, Revisão, Reformulação Problema		Solucionar Problema				
CRISP-dm	Compreensão do negócio		Compreensão dos dados		Preparação dos dados, Modelagem		
SEMMA			Amostra	Explorar	Modificar e modelar		

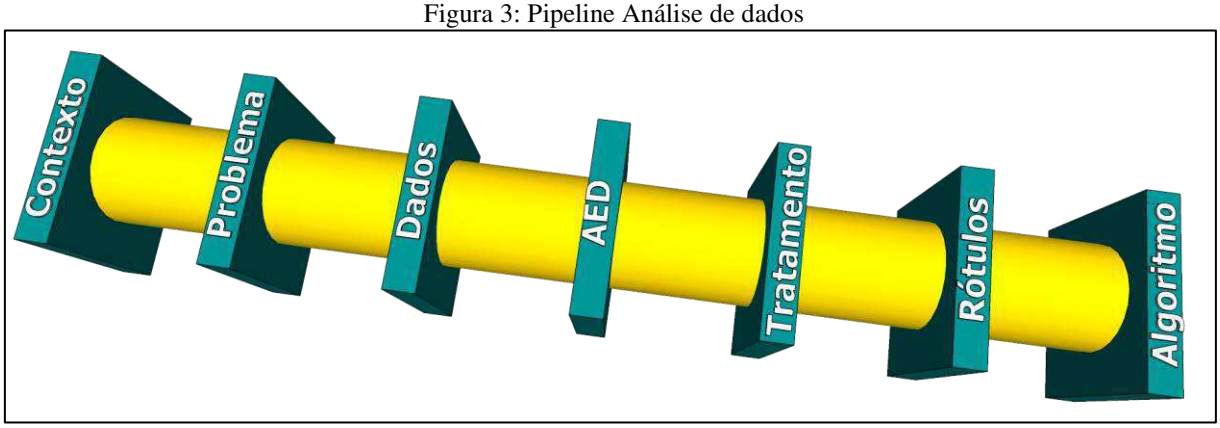
FONTE(PRÓPRIA)

Todos os processos encontrados são da área de mineração de dados, e por isso não incluem a parte de Aprendizado de Máquina, como a definição dos rótulos e a escolha dos algoritmos.

2.9. Pipeline de análise de dados

O termo processo deriva da palavra composta latim *procedere* (*pro* + *cedere*), e indica a ação de ir pra frente. O pipeline de análise de dados representa esse conjunto sequencial de ações com objetivo comum. Para esse trabalho, foi decidido usar passos definidos para cada problemática, sem misturar áreas como contexto e problema, ou coleta, análise e tratamento dos dados. Cada uma dessas etapas deve obter o espaço necessário para garantir o melhor resultado. Os passos necessários para a proposta do trabalho, da pergunta até o resultado da análise de dados podem ser resumidos como pipeline, acrônimo usado para descrever uma segmentação de instruções, de passos. O transporte dos dados através do tubo é garantido pelas bombas que transportam o conteúdo até a próxima estação. Em cada umas das estações é necessário passar por uma lista de verificação, para garantir uma passagem fluida até a próxima estação. O pipeline representado na Figura 3 mostra o fluxo de ações descritas até agora. Os passos são a definição do contexto, a descrição do problema, o entendimento, a AED e o tratamento dos dados, a definição dos rótulos e, ao final, a escolha do algoritmo.

Figura 3: Pipeline Análise de dados



FONTE: (PRÓPRIA)

Capítulo 3 – Estudo De Caso

O pipeline de análise de dados é constituído por poucas etapas bem definidas. A importância de cada etapa se manifesta em sua aplicação. A interação das partes envolvidas facilita cada passo e é a base para uma solução interessante. A complexidade do assunto é demostrada quando aplicado em um ambiente novo. A identificação do contexto, da área em geral, da área em concreto e dos fatores envolvidos são um conjunto muito grande em relação aos trabalhos necessários na análise dos dados, na geração do modelo e na entrega do resultado. Para mostrar isso, foi escolhido um estudo de caso na área de marketing, mais específico de e-mail nurturing. O problema colocado é simples: "Como aumentar o retorno financeiro de uma campanha de e-mail marketing?".

3.1. Contextualização na área de marketing

A contextualização é o primeiro passo. As informações necessárias para responder a pergunta podem ser derivadas diretamente da pergunta. No caso atual, a pergunta já contém o contexto: marketing, mais especificamente e-mail marketing. A pergunta é sobre uma medida, o "como aumentar". Para isso é necessário conhecer os indicadores usados e os valores padrão para entender se é possível o aumento. E, ao final, será necessário entender como é possível aumentar um resultado nessa área. Finalmente, será imprescindível entender quem são as partes interessadas e sua relação com as informações, o negócio e o resultado. Assim, obtemos um conjunto de perguntas que devem ser respondidas:

- O que é e-mail marketing?
- Quais são os indicadores usados para medir o resultado de uma campanha de e-mail marketing?
- Quais são os valores padrão para esses indicadores?
- Como é possível chegar a um resultado melhor?
- Quais são os passos a partir do e-mail até a geração de resultado?
- Quais são os passos para aumentar a qualidade de cada um desses passos?
- O que implica a aplicação das medidas encontradas?
- Como medir essa reação?
- Quais são as partes envolvidas?
- Qual é a relação delas com o negócio?

- Qual resultado cada uma dessas partes espera?
- Quais dados são necessários para construir um modelo?
- Como coletar os dados?
- Qual o modelo de Aprendizado de Máquina escolher?

3.2. Escolha das ferramentas

A biblioteca de software Apache Hadoop é uma estrutura que permite o processamento distribuído de grandes conjuntos de dados, projetado para escalar de servidores únicos para milhares de máquinas, cada uma oferecendo computação e armazenamento locais (APACHE HADOOP, 2018). Apache Spark funciona de maneira semelhante ao Hadoop, mas realiza as computações em memória, conseguindo reduzir o tempo necessário para uma regressão logística de até 100 vezes.

Databricks é uma plataforma de computação distribuída que unifica Análise de Dados, Ciência de dados e Aprendizado de Máquina (DATABRICKS, 2018). Assim, é possível examinar dados para encontrar padrões e derivar correlações significativas nos dados brutos ou naqueles preparados, limpos, para, no final, usá-los para fazer previsões no seu comportamento. A plataforma foi desenvolvida por criadores de Apache Spark e oferece as bibliotecas de Spark para SQL, captura de fluxos de dados, Aprendizado de Máquina, Aprendizado profundo e análise de gráficos (DATABRICKS, 2018). A plataforma se destaca por oferecer colaboração em tempo real, versionamento com integração com GitHub, e notebooks interativos que suportam o uso de Hive SQL, Python, R e Scala no mesmo notebook. Databricks foi especialmente criado para trabalhar com grande número de dados, mas é muito interessante pela facilidade de uso, já que permite acessar todas as ferramentas sem instalação local, diretamente através de qualquer navegador.

Embora a Databricks ofereça gráficos rápidos de qualquer tipo de dado, a falta de personalização das visualizações é, sem dúvida, o déficit da plataforma. É certo que há a possibilidade de usar várias bibliotecas de R, para uma apresentação gráfica mais elaborada e sofisticada dos dados. Mas, sobretudo pela fase de análise de dados, da análise exploratória dos dados, foi escolhido Tableau Desktop para a apresentação gráfica. Tableau permite a conexão com Databricks e consegue acessar diretamente as tabelas geradas naquele ambiente. A explicação gráfica dos dados acessados por Tableau é simples, veloz e eficaz. A criação de

gráficos participativos e a publicação dos dados em um servidor Tableau permite a entrega de resultados para um público mais exigente. Assim, o cliente recebe acesso a gráficos interativos, com a possibilidade de navegar através dos dados e resultados apresentados, interagindo com a ferramenta.

Resumindo: para a transformação e análise dos dados é usado Databricks, as visualizações, por sua vez, são feitas com Tableau Desktop.

3.3. Marketing

Antigamente, resumido em "contar e vender", hoje o marketing é entendido como o gerenciamento das relações lucrativas com o cliente, com o objetivo de criar valor - a fim de capturar o valor em troca (KOTLER; ARMSTRONG, 2014). Assim, a venda é uma consequência resultante do marketing, ou como Peter Drucker, o pai da administração moderna, definiu "o objetivo do marketing é tornar a venda supérflua" (DRUKER, 1973). Hoje, o consumidor é capacitado, e procura informações sobre um produto ou serviço para garantir a melhor escolha. Com isso, "o foco se afasta do poder de vendas e se torna uma batalha baseada na informação" (KOTLER, 2012). Um modo de falar diretamente com o consumidor, sem precisar de intermediário, é o marketing direto, que inclui catálogos, televisão, quiosques, internet, e permite espalhar informação direcionada para consumidores individuais, receber resposta imediata e cultivar um relacionamento duradouro com o cliente (COMPUTING, 2018). Uma das ferramentas do marketing direto, com finalidade de espalhar informação direcionada, é o marketing por correio eletrônico, o *e-mail marketing* (EM).

3.3.1. Marketing por correio eletrônico

Quando Ray Tomlinson mandou o primeiro correio eletrônico em 1971 (COMPUTING, 2018), ainda não era previsível a inundação de mensagens eletrônicas que iria gerar. Somente sete anos depois, Gary Thuerk, enviou 397 e-mails para promover computadores da DEC (RECORDS, 2018), resultando em uma venda de 13 milhões de dólares. Atualmente, as pessoas que trabalham em um escritório, em média, recebem 121 e enviam 40 e-mails por dia, com a previsão de alcançar 333 bilhões de e-mails enviados anualmente e, para o ano de 2022, há uma estimativa de alcance de 4,2 bilhões de endereços

de correio eletrônico (GROUP, 2018). Os dados da edição atual desse estudo estatístico sobre o uso de e-mail, que o Radicati Group efetua desde 1993, mostram o porquê desse meio ser tão interessante para fins de marketing: aproximadamente 14% dos usuários norte-americanos seguem os links contidos nas mensagens que remetem para sítios internet, e mais de 6% efetua compras como consequência de um e-mail recebido no dispositivo móvel.

EM é definido como mensagem de correio eletrônica enviada e recebida pela internet que tenha por objeto divulgar ou ofertar produtos ou serviços, manter relacionamento com base de destinatários ou, ainda, propiciar atendimento ao cliente (CAPEM, 2010). EM resume o envio direcionado de mensagens comerciais ou não para uma lista definida de destinatários. Ele é o instrumento de marketing para criar uma relação entre empresa e o (futuro) cliente. Mesmo com o uso intensivo de redes sociais, e-mail continua sendo o melhor meio para alcançar clientes (HOSTPAPA, 2012), porque, enquanto 94% das pessoas usam e-mail, só 61% usam as redes sociais. E mais de 70% do público prefere e-mail como meio de mensagens publicitárias (MARKETINGSHERPA, 2015). Olhando o rendimento, definido pelo retorno do investimento (*return of invest*ment, ROI), para cada dólar gasto em EM podem ser calculados até 44 dólares de retorno (VAN, 2016).

3.3.2. Lead Nurturing

Existem diferentes possibilidades e cenários possíveis para o uso de e-mail na área de marketing. Nesse trabalho, o ambiente escolhido é da nutrição de leads, melhor conhecido como "*lead nurturing*". Esse termo em inglês é usado para denominar uma estratégia de marketing. Nesse contexto, um lead, é um potencial contato de vendas que expressa interesse em seus produtos ou serviços. "*Nurturing*" significa alimentar e proteger (a prole), apoiar e encorajar, treinar e educar. Juntos, os dois termos são usados para descrever um conjunto de práticas, um processo usado para providenciar conteúdo pertinente de conteúdo educacional para apoiar e aproximar futuros clientes para a própria empresa, o próprio produto.

A grande diferença do Lead Nurturing com outras áreas de EM, é a atitude do destinatário. Em geral, como definido pela CAPEM, o assinante deve ter realizado a escolha em receber o e-mail. Mas, essa escolha pode ser induzida: a pessoa escolhe assinar uma lista de e-mails, porque não tem outra maneira para ter acesso a um desconto, poder participar de

um sorteio de prêmios ou receber informações. Esse trabalho é um ótimo exemplo para essas práticas forçadas. Raramente, era possível acessar estatísticas, estudos com indicações para o mercado ou informações de empresas que atuam no ramo de marketing, sem deixar, pelo menos, o endereço de e-mail. Automaticamente, todas essas empresas mandavam comunicações de marketing. A maioria com uma insistência e frequência, que as mensagens não eram definidas como SPAM.

Lead Nurturing é baseado no recíproco respeito. A empresa sabe da importância do (futuro) cliente, o seu tempo limitado e que sua assinatura permite nutrir a fome de informações específicas, sem ser uma carta branca para inundar a sua caixa de correio eletrônico. O assinante selecionou uma empresa e/ou um produto ou serviço, que está usando ou quer usar no futuro. Com a assinatura, afirma que deseja informações através de mensagens eletrônicas relativas à sua escolha. Essa opção está longe de ser considerada uma situação irritante.

3.4. Indicadores do E-mail Marketing

Cada EM tem um propósito, um alvo definido, escolhendo destinatários com especificações certas para a campanha da lista de assinaturas, um início e um fim. Sucesso e fracasso são definidos por diferentes métricas, e os valores usados como norteamento para cada métrica dependem do ramo da indústria, do país (da empresa e do destinatário), da dimensão da empresa, e do tipo de produto, entre outros. Antes de usar os indicadores de marketing, e do EM, em especial, vamos explicar quais são, como são medidos e qual a confiabilidade de cada um. Os indicadores mais conhecidos para esse ramo de distribuição de informação são:

- Taxa de e-mails não entregues
- Taxa de entrega
- Taxa de e-mails entregues ao destinatário
- Taxa de abertura
- Taxa de cliques nos links embutidos
- Taxa de conversão

- Retorno sobre Investimento
- Taxa de crescimento da lista de assinatura
- Taxa de cancelamento de assinatura
- Taxa de contestação
- Taxa de assinantes não comprometidos

3.4.1. Taxa de mensagens não entregues

A taxa de e-mails não entregues é formada pelos e-mails rejeitados, bloqueados e deferidos.

A taxa de e-mails rejeitados na literatura é chamada de *bounce rate* (BR). Para evitar equívocos, estamos na área de EM. No domínio de análise de Web, esse nome é usado pra mostrar o percentual de visitantes que acessa uma página Web, e, sem acessar qualquer outro link do site, sai dele (TAXA DE REJEIÇÃO, 2018). No EM essa taxa é usada pra medir a porcentagem de e-mails rejeitados pelos servidores, sem chegar ao destino. O capítulo sobre filtros e-mail e outros desafios na entrega (GROVES, 2009) é bastante exaustivo na descrição da BR: existem dois tipos de rejeição, definidos pelos tipos de erro encontrado, temporários ou definitivos. Erros são considerados temporários quando o endereço de correio eletrônico é válido, mas por alguma causa o e-mail não pode ser entregue. Isso pode ser causado por um servidor de destino temporariamente não disponível, uma política de e-mail restritiva, uma caixa de correios do destinatário sem espaço, um e-mail maior do que permitido para essa caixa de correio, ou uma mensagem de ausência do destinatário. Erros temporários são chamados de *Soft Bounce*.

Os erros permanentes são causados por endereços errados: domínio não existente, servidor de destino bloqueia permanentemente o envio para aquela caixa de correios ou por usuário inexistente. Maior o número de erros permanentes, pior a qualidade da lista de endereços usada para a campanha. Pra piorar, os fornecedores de serviços de correio eletrônico usam o número de endereços inválidos usados por um remetente, para definir sua reputação. Maior o número de erros permanentes, pior a reputação. Então, esses endereços deveriam ser retirados imediatamente da lista.

É bom lembrar que o uso da palavra “temporário” é bastante enganoso nesse contexto. Antigamente, devido à estabilidade das conexões, os servidores de correios ficavam fora do ar, e as caixas de correio eram bastante reduzidas pelo alto custo de armazenamento. Mas, essas duas situações hoje são bastante incomuns. Atualmente, o *soft bounce* não mostra mais um erro temporário na parte do destinatário, mas sim a rejeição da mensagem (BOYD, 2018), que pode ter sido devido à má reputação do remetente, do endereço IP ou do domínio ter sido inserido em uma lista de e-mails bloqueados; do domínio ou endereço IP, ser definidos como fonte de spam, ou ao uso de elementos que o servidor ou provedor, que recebe a mensagem, define como spam. Somente uma análise mais aprofundada pode ajudar a entender essa taxa.

Os e-mails deferidos podem ser classificados como um verdadeiro problema temporário. São mensagens de correio eletrônico que, no momento, não podem ser entregues, por exemplo, devido à indisponibilidade de um dos servidores de correio eletrônico até o destinatário. Devido à natureza do problema, os servidores não informam o remetente tempestivamente, mas somente quando a entrega parece ser mesmo impossível (que pode ser um limite de tempo ou de tentativas falhas de entrega).

3.4.2. Taxa de entrega

A taxa de entrega serve como base para muitas das métricas usadas no EM. Ela é definida pelo número de e-mails enviados e aqueles rejeitados de um *hard* ou *soft bounce*.

$$Taxa\ de\ entrega = \frac{(e-mail\ enviados) - (e-mail\ rejeitados)}{(e-mail\ enviados)}\ x\ 100\%$$

É surpreendente que a taxa de entrega, mesmo por empresas importantes do mercado, seja definida como a “porcentagem de e-mails que foram realmente entregues nas caixas de entrada dos destinatários” (HUBSPOT, 2017). Sabendo que a informação sobre esses dois tipos de rejeição vem do servidor de gateway, e não do destinatário final, é fácil deduzir que a taxa de entrega mostra somente a quantidade de e-mails aceitos pelo servidor, e não de e-mails que chegaram até a caixa postal. O servidor ainda pode decidir o que fazer com os e-

mails, se bloqueá-los, ou enviá-los para uma pasta spam, como descrito acima. Por essa razão, é bom distinguir entre entrega e entregabilidade, *delivery* e *deliverability*. A entrega é definida pela aceitação física do servidor de gateway. Entregabilidade define o destino final da mensagem, caixa de entrada, pasta spam ou outra pasta diferente (LEWKOWICZ, 2018). No artigo citado, a entregabilidade é definida através de identificação, reputação e conteúdo.

Identificação é o meio de um set de protocolos que consegue confirmar a identidade do remitente, como *Sender Policy Fram*ework (SPF), *DomainKeys Identified Mail* (DKIM), e *Domain-Based Message Authentication, Reporting, and Conformance* (DMARC).

A reputação já foi mencionada na rejeição. Cada campanha de e-mail e a reação dos destinatários contribuem para a reputação do remetente. O envio de e-mails relevantes para os endereços certos, e destinatários que marcam o remetente como confiável são essenciais para uma boa reputação.

O conteúdo define como os servidores de e-mail, os filtros spam, e, finalmente, os destinatários reagem ao recebimento da mensagem.

A extensão da problemática da entregabilidade é mostrada pelo *Deliverability Benchmark Report.* A edição de 2017 mostra que, por exemplo, no Brasil, os profissionais de marketing conseguiram uma taxa de colocação na caixa de entrada média de 79%, apenas ligeiramente abaixo da média global que é 80% (RETURN PATH, 2017). Esses números revelam que, globalmente, 20% dos e-mails não chegam ao destinatário! E que com essa informação na mão, o valor de taxa de entrega, a Delivery Rate, deveria ser corrigida para um percentual abaixo de 20%.

3.4.3. Taxa de abertura

A taxa de abertura, TA, exprime a porcentagem de destinatários que abrem a mensagem. Ela pode mostrar a relação saudável do destinatário com a empresa, o interesse no produto/serviço oferecido, se o remetente é considerado confiável, se a qualidade do e-mail é boa e se a lista de assinantes usada é saudável – já que escolhemos as pessoas certas para a informação enviada (PHRASEE, 2018). Mas, tudo isso pode ser resumido com confiança no remetente e ótima escolha para formular o assunto. Além disso, as métricas que mais

influenciam a taxa de abertura são dia e hora de recebimento da mensagem, a quantidade de mensagens e o sincronismo certo das ofertas. O funcionamento atrás da TA é simples (SHOULIDS, 2018) : no e-mail, é escondido um pixel de rastreamento, uma imagem invisível, normalmente de um pixel de altura e de largura. Quando o destinatário abre o e-mail, a imagem é baixada do servidor, e esse acesso conta como abertura do e-mail. Isso significa que o destinatário deve baixar as imagens contidas no texto. Desde quando grandes provedores, políticas de empresas e aplicativos de gerenciamento de e-mail passaram a não baixar as imagens por padrão, o destinatário consegue ler a mensagem, sem baixar o pixel de rastreamento e, assim, sem ser percebido pelo servidor e em consequência também passa despercebido pelo remetente. A situação piorou com a limitada franquia de tráfego de dispositivos móveis, forçando aos usuários a baixar imagens somente quando for absolutamente necessário. Com isso, a TA se revela uma métrica imprecisa, bastante enganadora, e não confiável para os profissionais de marketing, por subestimar os números reais de abertura.

$$Taxa\ de\ abertura = \frac{(e-mails\ abertos)}{(e-mail\ enviados)-(e-mail\ rejeitados)}\ x\ 100\%$$

3.4.4. Taxa de cliques

Uma métrica melhor pra determinar a expressividade da comunicação é a *Click Through Rate*, CTR. Ela mostra a taxa de pessoas que clicaram na oferta incluída no e-mail. Para chegar nisso, o destinatário precisa abrir o e-mail, ler o conteúdo, continuar clicando pela oferta/informação contida. Então, poderemos presumir que conseguimos colocar algo capaz de estimulá-lo.

$$Taxa\ de\ cliques = \frac{(links\ abertos)}{(e-mail\ enviados)-(e-mail\ rejeitados)}\ x\ 100\%$$

Quanto mais alta a taxa de cliques, melhor a campanha e-mail? A definição de uma

boa taxa de cliques depende, como praticamente para todas as taxas de Marketing, do ramo da indústria, do produto, do país, até da dimensão da empresa que assina a mensagem. Mesmo assim, maior o número de cliques, maior o número de potenciais clientes.

As métricas de entrega, abertura e cliques podem ser resumidas como indicadores de visibilidade. O brilho da campanha, a eloquência das palavras, e eventualmente a gráfica apelativa, conseguiram obter a visibilidade necessária, para o destinatário abrir o link embutido no e-mail.

3.4.5. Taxa de conversão

Toda campanha de e-mail tem um propósito – e a taxa de conversão mostra finalmente se foi possível alcançá-lo. O destinatário recebeu uma informação, abriu o e-mail, clicou no link embutido, chegou à página de destino e agora converte todo esse processo, fazendo exatamente o que era definido como objetivo de todo esse esforço. Isso pode ser, entre outros, a finalização de uma compra, a aceitação de um contrato, o preenchimento de um formulário, o fornecimento de informações ou o ato de baixar uma informação. O fato de ter conseguido o objetivo da campanha, além de constatar sua boa qualidade, pode ser usado como indicador da página de destino.

$$Taxa\ de\ conversão = \frac{(conversões)}{(e-mail\ enviados) - (e-mail\ rejeitados)}\ x\ 100\%$$

3.4.6. Retorno sobre Investimento

Chegou o momento de monetizar o custo da campanha e entender se valeu a pena investir nela. Uma das métricas usadas pra isso é o Retorno sobre Investimento (ROI), que mostra a relação entre o investimento e o custo de uma campanha de e-mail. Apesar de ser definida como uma das quatro métricas mais importantes do EM pelos especialistas da área publicitária dos Estados Unidos da América (VAN, 2016), é negligenciada por mais de 60% das empresas brasileiras, mesmo que quase 90% monitorem os resultados (FONSECA, 2018).

A maneira mais simplificada divide a receita pela diferença entre receita e custo.

$$Retorno\ sobre\ investimento = \frac{(receita)}{(receita) - (custo)}\ x\ 100\%$$

Mas, essa métrica também pode ser enganosa: um ROI alto pode resultar de uma campanha barata e/ou de baixo esforço. Não observar o impacto da campanha e a reação dos assinantes da lista de e-mail, e ver somente o lucro direto de uma campanha, podem ser uma visão estreita e irresponsável.

3.4.7. Taxa de crescimento da lista de assinatura

Uma maneira de enxergar a qualidade das mensagens, do produto oferecido e do índice de confiabilidade da empresa, é através do crescimento de assinantes pedindo informações. Na verdade, o cálculo apresentado aqui é bastante simplificado, porque observa somente o número total da lista. Desse jeito, um número alto de novas assinaturas consegue camuflar os cancelamentos efetuados durante a campanha.

$$Taxa\ crescimento\ assinaturas = \frac{(assinantes\ no\ fim\ da\ campanha)}{(assinantes\ no\ início\ da\ campanha)}\ x\ 100\%$$

3.4.8. Taxa de cancelamento de assinatura

A parte negativa na lista de assinaturas é devido aos cancelamentos efetuados durante a campanha. Para cada causa que leva à perda do assinante, tem algumas possibilidades de atuação para prevenir. O mais importante é escolher bem os destinatários para cada campanha. Quanto menos o assunto é interessante para o destinatário, maior a possibilidade dele se cansar da informação. Esse cansaço pode ser também provocado pela quantidade excessiva de mensagens ou um tamanho exagerado, que leva a um tempo de abertura

inadmissível (WAINWRIGHT, 2018). Até o simples fato de enviar e-mails não otimizados para dispositivos móveis, pode ser fatal: 80% dessas mensagens são deletadas e como consequência, três destinatários em dez, decidem cancelar a assinatura (LEGGATT, 2018). Mesmo querendo limitar os cancelamentos de assinaturas, é importante oferecer uma oportunidade bem visível para sair da lista em cada comunicação. Segundo um estudo sobre SPAM, 50% dos americanos marcam mensagens como SPAM, porque não conseguem encontrar a opção de cancelamento da assinatura (WHITE, 2018). Existem diversas leis que definem as regras de optar ativamente para participar (*opt-in*) ou cancelar (*opt-out*) uma assinatura de e-mail. Geralmente, elas se resumem em quatro pontos (VERTICALRESPONSE, 2018):

1. O assinante deve ter escolhido receber o e-mail.
2. O e-mail deve revelar a origem
3. O assunto deve refletir o conteúdo do e-mail
4. O e-mail deve incluir a possibilidade de cancelamento da assinatura

No Brasil, essas regras são incluídas no "Código de Auto-regulamentação para Prática de E-mail Marketing" (CAPEM, 2010), conduta amplamente aceita, mas sem nenhuma base legal.

$$Taxa\ cancelamento = \frac{(cancelamentos)}{(e-mail\ enviados) - (e-mail\ rejeitados)}\ x\ 100\%$$

3.4.9. Taxa de contestação

A métrica é conhecida como Complaint Rate. O acrônimo "Complaint" é usado para indicar "isso é SPAM". Consequentemente, a Complaint Rate, a taxa de contestação, mede o número de mensagens marcadas como spam. Um circuito de retorno de informação notifica o remetente quando o destinatário marca a mensagem como SPAM. A informação é enviada em formato ARF (*Abuse Reporting Format*), definido pela RFC 5965 e, consequentemente, a RFC 6650 (FALK, 2012). Infelizmente, o formato usado pode ser diferente, e a lista de Fornecedores de Acesso à Internet (*Internet Service Provider*, ISP) que suporta o circuito de retorno, é limitada. Quando for disponível, a taxa pode ser usada para encontrar padrões e causas das contestações.

$$Taxa\ de\ contestação = \frac{(número\ contestações)}{(e-mail\ enviados)}\ x\ 100\%$$

Com a finalidade de conseguir interpretar os resultados, é importante lembrar que é necessário calcular a taxa de contestação separadamente para cada um dos ISP.

3.4.10. Taxa de assinantes não comprometidos

Outra métrica que olha para a qualidade da lista dos assinantes, é a taxa de assinantes não comprometidos. São aqueles destinatários que aparentemente não reagem às mensagens recebidas. Infelizmente, é outra métrica ambígua. O destinatário pode ter marcado todas as mensagens como SPAM, sem informar o emissor, colocando às mensagens em uma pasta à parte, usada como um SPAM pessoal. Mas, ele pode ter lido todas as mensagens, sem baixar o pixel de rastreamento, e por isso, não ter informado sobre isso. Outra possibilidade: o destinatário lê a mensagem no trabalho, mas conclui as compras em casa por restrições do uso da Internet no trabalho.

A solução mais fácil pode ser utilizada quando as assinaturas são coletadas de maneira duvidosa. Nesse caso, os endereços podem não ter valor, e basta tirá-los da lista dos assinantes. Em todos os outros casos, o endereço tem um valor. Poder ser o desconto recebido no momento da assinatura, o trabalho investido em cada um dos assinantes, os negócios fechados até um determinado momento, ou a possibilidade de negócios no futuro.

3.4.11. Resumo das taxas do e-mail marketing

Resultados, números e estatísticas podem ser importantes para o negócio do EM. Mas, é vital, não perder o foco do objetivo do negócio e de cada campanha de EM. Às vezes, os números são usados para satisfazer o ego de alguém: o resultado não permite interpretação, é só um número, e por isso, inútil para aprimorar o serviço. Usando as informações de um estudo brasileiro (FONSECA, 2018), é possível resumir a aplicação do EM pelas empresas brasileiras e os hábitos de consumo nesse país (Tabela 4). 99% dos usuários acessam e-mail, quase todos deles assinaram algum newsletter, e mais da metade considera a frequência das

mensagens excessiva.

Tabela 4: Resumo hábitos EM no Brasil

Empresa	Usuário
77% adotam a estratégia do EM	99% possuem e acessam e-mail
97% delas acreditam na eficácia do EM	95% deles recebem algum newsletter
76% deles usam profissionais dedicados	84% deles querem receber conteúdo
Só 50% investem na nutrição de leads	55% consideram a frequência excessiva

FONTE: (FONSENCA, 2018)

Das empresas, dois terços adotam uma estratégia de EM, um terço tenta criar as campanhas sem profissionais dedicados, e somente a metade investe na nutrição de *leads*. Como já mencionado no início, para cada métrica tem uns valores que podem servir como norteamento. A única fonte encontrada com valores para o Brasil é a empresa Getresponse (LESZCZYNSKI, 2018). Para comparar os valores indicados da Getresponse, são usados os valores de Mailchimp, a maior plataforma de automação de marketing do mundo (E.M.B., 2018). Os números resumidos na Tabela 5, mostram a ambiguidade da informação disponível. Enquanto a taxa de abertura das duas empresas é bem parecida, a diferença da taxa de cliques é de quase 70%. É interessante ver que a taxa de cancelamento da assinatura no Brasil é mais baixa que no resto do mundo.

Tabela 5: Comparações taxas no mundo e no Brasil

Taxa	Mailchimp	Getresponse	
	Mundo	Mundo	Brasil
Taxa de abertura	22,79%	24,88%	26,97%
Taxa de cliques	2,73%	4,06%	4,20%
Taxa de cancelamento assinatura	0,29%	0,24%	0,19%
Taxa de contestação	0,02%	0,02%	0,02%

FONTE: (EMB, 2018)

Como aumentar os resultados das campanhas de EM? Partindo do pressuposto que a qualidade das campanhas já está num nível profissional de alta qualidade, com exceção do aumento de número de assinantes, as possibilidades são limitadas à maior frequência de mensagens e sua personalização.

3.5. Personalização

No mercado atual, determinado por uma queda no crescimento econômico, o reposicionamento das empresas da visão global para a regional e da regional para a local deve ser a tendência primordial (KLOTER, P., KLOTER, G., 2013). A análise precisa do mercado, a definição dos segmentos, a priorização e a dedicação exclusiva para cada um deles. KLOTER, 2004, tornam-se sempre mais importantes. Com a segmentação, a definição de grupos de perfis de clientes com características, necessidades e possibilidades semelhantes (DICKSON, P. R., GINTER, J. L., 2004), e a disponibilidade de informações sobre os (futuros) clientes, torna a personalização da comunicação cada vez mais íntima. Recorrendo mais uma vez aos números de diferentes pesquisas, e-mails com assunto personalizado aumentam a taxa de abertura em 26%; conteúdo personalizado melhora a taxa de cliques nos links embutidos em 14%, aumentando os resultados finais consideravelmente (STIGLITZ, 2018).

Publicidade personalizada não é uma novidade. A introdução de mensagens personalizadas já está documentada desde 1870 (ROSS, 1992), mas a importância da individualização aumenta com a emergente disciplina de marketing de conteúdo (F. C., 2016). Isso é devido ao aumento da percepção do conteúdo com a pertinência da informação (CHAMBERLIN, 1969). Assim, a eficácia da publicidade é diretamente afetada pela disponibilidade de dados utilizados (SOLOVE, 2008). Entretanto, com o advento de Big Data, a atual agregação de informações sobre os consumidores, aumenta a possível invasão na sua privacidade (JAI, T. M. C., KING, N.J., 2016). Nunca foi tão fácil conseguir dados dos consumidores. Basta o nome e o endereço de e-mail para ter acesso à história de compras, preferências exibidas nas mídias sociais ou dados demográficos (JAI, T. M. C., KING, N.J., 2016). Além disso, é possível comprar dados de terceiros para facilitar a personalização das comunicações com o cliente. O consumidor quer privacidade, mas a tecnologia coleta,

analisa, combina, usa e compartilha informações sobre ele (FTC, 2009). Isso leva a um paradoxo: maior a personalização das mensagens, maior a intrusão na privacidade (SUSTANO, J., et. Al., 2013). Essa sensação de violação da privacidade pode criar uma aversão do consumidor ao produto anunciado em uma mensagem publicitária. Estudos mostram que a percepção de invasão de privacidade pelo consumidor, pode implicar resultados opostos, obtendo-se o efeito inverso do desejado: em vez de incentivar a compra de um produto, o consumidor cria uma aversão a ele, e, na pior hipótese, para toda a marca que ele representa (BROWN, M.; MUCHIRA, R, 2004). A economia comportamental estuda como os vieses individuais, sociais, cognitivos e emocionais influenciam nas decisões econômicas (ACQUITI, A., GROSSKLAGE, J., 2006). Essa mudança na intenção de compra pode acontecer a qualquer momento, mesmo no último minuto (BROWN, M.; MUCHIRA, R, 2004). A importância dessa informação ganha ainda mais peso, quando é combinada com o número de pessoas, sensíveis à própria privacidade. Como exemplo podemos citar TUROW et al., 2009, mostrando que de 1.000 americanos, 68% dos entrevistados, definitivamente, e 19%, provavelmente, não permitiria que os anunciantes fossem rastreados on-line, se fosse dada uma escolha. Com o aumento da suscetibilidade na área da privacidade, evitar essas invasões fica cada vez mais importante.

3.6. Privacidade

Privacidade é uma das características fundamentais para a vida social (NISSENBAUM, 2011), a “reivindicação de indivíduos, grupos ou instituições para determinar por si mesmos quando, como e em que medida a informação sobre eles é comunicada a outros” (WESTIN, 1967). Privacidade define o "direito de ser deixado sozinho" (WARREN, S. D., BRANDEIS, L. D., 200), não ajuda apenas a esconder coisas, mas é necessária para ser o dono de si mesmo, a autonomia e integridade (GARFINKEL, 200). É um “espaço para respirar, para poder participar nos processos de gestão de fronteira que permitem e constituem o autodesenvolvimento” (COHEN, 2013). Visões mais modernas chegam numa visão pluralística da privacidade, baseada nos diferentes cenários específicos para diferentes contextos (SOLOVE, 2008).

A conceituação da privacidade tem sido sempre difícil (WEINTRAUB, 1997). Ela tem

um problema de imagem (COHEN, 2013). Pode ser vista por muitas perspectivas diferentes, incluindo as políticas, as pertinentes aos direitos do cidadão e do lado da proteção do consumidor (BARNES, 2006). Tem que vê o bem-estar da comunidade acima da privacidade (BAKER, 2010). Quem defende a ideia de que não existe privacidade em espaço público (ANDERSON, 2012), que o bem-estar da comunidade está acima da privacidade, porque senão esta pode ser a razão pela qual não conseguimos prender terroristas (BAKER, 2010). Quando colocamos na balança essa visão de privacidade com os valores da segurança nacional, da eficiência e do empreendedorismo, a privacidade sempre sai perdendo (COHEN, 2013). Tanto que na era da mídia digital, parece que privacidade nem existe mais (GANDY, 1993). O próprio ato de fornecer informações sobre si mesmo on-line, torna essa informação pública (READ, 2006). Pessoas e empresas, que vivem das informações dos usuários, percebem a discussão sobre a privacidade como um exagero (MAHAJAN, 2017), até considerar que ela nem é mais uma norma social (BLANKER, 2014). Em uma ecologia online florescente, em que indivíduos, comunidades, instituições e corporações geram conteúdo, experiências, interações e serviços, a moeda suprema é a informação, incluindo informações sobre pessoas (NISSENBAUM, 2011). Mas, mesmo que nós vivamos em um mundo em que compartilhamos informações pessoais, ainda tem a visão que devemos rejeitar que privacidade seja um valor fora de moda (TWH, 2013). Já no século XIX, WARREN e BRANDEIS, 1890, constataram que, mesmo estando diante de um princípio tão antigo quanto o direito comum, de tempo em tempo, é necessário definir de novo a exata natureza e a extensão da proteção da pessoa e da propriedade.

3.6.1. Percepção da privacidade

Privacidade não é uma constante, mas sim uma noção dinâmica, baseado no perfil da sociedade, nas necessidades das pessoas e no avanço tecnológico (ROBOTS, 2015). Os métodos usados para avaliar o nível de privacidade mostram a dificuldade de definir a sua percepção. Tem quem defenda diferenças na percepção da privacidade pelo critério idade do consumidor: que os mais jovens são mais preocupados com a privacidade que os mais velhos (MOSCARDELLI, 2004; BOYD, D., HARGITTAI, E., 2010; KOKOLASIS, 2017); outros que mostram os mais jovens se preocupando menos com a privacidade (KOKOLASIS, 2017; BAKER, 2010). Mesma situação, quando a divisão é feita por gênero dos usuários: de um

lado àqueles que mostram que as mulheres são mais conscientes quando o assunto é privacidade (MOSCARDELLI, 2004; BOYD, D., HARGITTAI, E., 2010; KOKOLASIS, 2017), do outro, quem não encontra diferenças significativas (BOYD, D., HARGITTAI, E., 2010). E continua quando a diferença é definida pela renda, o estado civil, a área geográfica, o país, as horas online, o número de sites acessados, o tipo de sites acessados, a rigidez da legislação existente (KOKOLASIS, 2017). DINEV et al. mostram a percepção da privacidade como resultado entre a percepção do controle (sigilo, confidencialidade e anonimato) e percepção do risco (grau de sensibilidade da informação, benefício alcançado com o fornecimento dela, importância de transparência e expectativas reguladoras) (DINEV, et. al.,2013). Ao final, parece que privacidade é um sentimento, uma percepção, que não é limitada pelas diferentes definições existentes.

Como exemplo disso, podemos mencionar dois casos diferentes: o anúncio pelo governo da Áustria em disponibilizar os dados dos registros eletrônicos de saúde, e o uso ilegal de dados pelo Facebook.

O registro eletrônico de dados de Saúde da Áustria, conhecido pela abreviação ELGA (*Elektronische Gesundheitsakte*) é um sistema criado para a padronização e comunicação eletrônica entre os prestadores de serviços de saúde (ELGA, 2018). É interessante notar, que no site da ELGA, esse objetivo principal é mencionado somente depois de elogiar a possibilidade dos pacientes de conseguir acessar os próprios dados sobre sua saúde em todo momento e qualquer lugar. Logo depois de publicar que pesquisas científicas terão acesso facilitado para esses dados de Saúde a partir de 2019, 5.000 pacientes deixaram o sistema do prontuário eletrônico – juntando-se aos outros 273.000, que desde sempre optaram por não enviar os dados para o sistema centralizado (ELGA, 2018). Em valor absoluto, esse número pode parecer negligenciável, mas pode ser só o início de uma série preocupante, já que, ao final, esse sistema tenta garantir o acesso rápido a dados essenciais em casos de urgência – e só isso salva vidas.

O segundo exemplo trata das reações ao uso ilegal de dados de 50 milhões de perfis do Facebook em março 2018 (SFGATE, 2018). Na Alemanha, poucos dias depois da publicação dos fatos pela Observer e o New York Times, uma pesquisa mostrou que 49% dos entrevistados, 53% dos homens e 44% das mulheres, já consideraram um cancelamento das contas de redes sociais. Nos Estados Unidos, os lesados pensam na reparação pecuniária do prejuízo sofrido. Já na primeira semana foram apresentados quatro processos à Corte Federal

em relação a esse vazamento de dados (SFGATE, 2018). Além disso, empresas reagem, excluindo as contas e limitando os anúncios na rede social (RED, 2018). E no Brasil? Com exceção da preocupação de interferências nas próximas eleições (GANHÃO, 2018), o terceiro país em números de usuários dessa rede social (STATISTA, 2018), não mostra significantes reações a respeito dessa situação.

Os exemplos citados acima mostram a amplitude de tipos de dados: de um simples post no Facebook, que fala sobre a situação climática da própria cidade, até os dados sobre o estado de saúde de uma pessoa.

3.6.2. Coleta de dados pessoais

Na conclusão de uma compra online, é necessário informar nome, endereço, telefone, e outros dados que identificam a pessoa e permitem a entrega do produto. Mas, mesmo nas compras em uma loja física, no supermercado, na loja de artigos esportivos, na farmácia, ou no abastecimento do carro, as empresas querem coletar a maior quantidade de dados possível sobre o cliente. Nome, CPF e e-mail são passados continuamente para diferentes tipos de transações. Essas informações podem ser necessárias para conseguir usufruir um produto "de graça". Na verdade, o usuário paga para a instalação de um aplicativo, para o uso do motor de busca, para o espaço de armazenamento na nuvem ou o para o uso de uma rede social, com as próprias informações. Cada passo na internet, hoje, é monitorado. As empresas querem saber qual produto os internautas olham, quanto tempo fica numa página, o que colocam na cesta, quais produtos compram juntos, e qual compra não é concluída. Além disso, é possível rastrear o caminho na rede, saber qual loja é acessada no mesmo momento, qual foi a proveniência e qual será a próxima loja que será consultada para encontrar o melhor preço. Todas essas informações permitem criar diferentes perfis dos usuários: social, econômico, racial, político, de saúde. Essas informações são valiosas e não ficam paradas. São passadas entre empresas que trabalham juntas, que tem o mesmo proprietário, ou são mesmo vendidas. A falta de uma legislação global e interpretações diferentes permite contornar normas e escapar de leis válidas somente para uma região. A legislação com bastante espaço para interpretação, a falta de fiscalização e as possibilidades técnicas permitem coletar dados sobre os internautas.

Como exemplo, podem ser citados dois exemplos da aplicação do novo Regulamento Geral de Proteção de Dados da União Europeia (RGPD) pelo grupo Facebook (EC,2018). O primeiro, é do aplicativo WhatsApp. Sobre a lista de contatos, nos termos de serviço do Whatsapp (ILW, 2018), encontramos a seguinte informação que o usuário fornece ao aplicativo: "os números de usuários do WhatsApp e de outros contatos em sua lista de contatos regularmente. Você confirma ter autorização para nos fornecer tais números de forma que possamos prestar os nossos Serviços." A coleta de contatos que nem são usadas no WhatsApp mostra a ganância de dados da empresa. Mas, passar a responsabilidade por essa coleta ao usuário, significa dizer que a empresa trabalha intencionalmente em uma área obscura. Se, por exemplo, um empresário salva o número de um cliente para poder contatá-lo, ele não precisa da permissão do cliente para fazer isso. Agora, se o WhatsApp coleta esse número, mesmo que ele não seja usado pelo aplicativo – e esse número não está registrado com WhatsApp (o que significa que o usuário não aceitou os termos de uso do aplicativo) – então, o proprietário do smartphone está infringindo a lei. Segundo a RGPD, o WhatsApp precisaria da permissão para poder passar o número do cliente para terceiros.

Fica ainda mais complicado, quando o WhatsApp é usado para enviar fotos, com a finalidade de mostrar o andamento de uma obra em construção, por exemplo. Sabendo que WhatsApp recebe acesso à foto, e os donos do aplicativo trocam às informações coletadas com outros aplicativos, a RGPD define que é necessário obter primeiro a permissão do dono da obra, para usar o aplicativo. Com esses dois exemplos é possível ver, como uma nova lei consegue mostrar os problemas no uso de aplicativos. Quando antes se deu pouca atenção aos termos de uso e serviço (OBAR, J. A., OELDORF-HIRSCH, A., 2016) agora, por obrigação da nova lei, os empresários na Europa são obrigados a aumentar a atenção. Como consequência pode ter sempre mais empresários, que desistem do uso deles, depois de entender que os aplicativos compartilham os dados com terceiros.

O outro exemplo vem do Facebook. Os dois pilares do RGPD são transparência e consentimento informado. Os dois são fortemente ignorados pela rede social. Acessando a rede na União Europeia, o usuário é informado que Facebook usa dados coletados dentro e fora do Grupo Facebook, que eles não vendem dados para ninguém, mas os usa para direcionar publicidade, sem dizer para as empresas, quem você é. O usuário pode escolher permitir o uso de dados provenientes de empresas parceiras, mas é logo informado que essa restrição diminui a relevância da publicidade mostrada – sem diminuir a quantidade dos

anúncios. E quando o assunto é o reconhecimento facial, Facebook informa ao usuário que não pode prevenir que outra pessoa usará a sua imagem e se passará por você, se ele não aceitar o uso dessa ferramenta pelo Facebook. Parece que Facebook, com essas formulações indutivas queria desafiar a nova legislação europeia, a fim de atenuar as novas regras tão limitantes para gigantes da comunicação.

3.6.3. Monitoramento dos usuários

O usuário deixa um rastro bem visível com as suas ações na Internet, seja pelos comentários que ele deixa em diferentes sites, pela escolha das pessoas, marcas ou grupos de interesse que ele está seguindo nas redes sociais, ou pelas curtidas espalhadas durante o dia. Todas essas Informações são usadas pelas empresas para publicidade comportamental. É facilmente compreensível que os sites coletem dados sobre o usuário enquanto ele navega neles. Mas fica menos transparente, quando sites conseguem acessar informações de comportamento do usuário em outro lugar. Para isso, existem diferentes métodos de acompanhamento do internauta que permitem inferir muito sobre sua personalidade.

3.6.4. Cookies

Um método comumente usado para armazenar as pegadas digitais dos internautas é a utilização de ficheiros guardados em seu dispositivo chamados de cookies. Primeiramente, *cookies* são usados para identificar usuários e memorizar as suas preferências em um site. Em vez de preencher sempre o mesmo formulário para poder acessar um site, os dados são salvos no *cookie* pela primeira vez, e recuperados todas as vezes que for necessário – sem que o usuário precise repetir para sempre as mesmas informações. Esses *cookies* são persistentes, eles ficam no computador, quando o usuário sai do navegador. Os *cookies* de sessão normalmente são usados somente durante a visita a um site. Eles podem, por exemplo, salvar produtos na cesta, enquanto o usuário continua navegando pelo site. *Cookies* de sessão são apagados quando o navegador é fechado.

Os *cookies* persistentes podem ser usados por qualquer site que o usuário visite.

Assim, é possível conhecer o usuário, os sites que ele visitou e suas preferências - sem perguntar a ele antes e sem informá-lo. Como definido por diferentes legislações, por exemplo, no Brasil pela Lei nº 12.965, de 23 de abril de 2014, conhecida como Marco Civil, e posteriormente, completado pelo Decreto Presidencial nº 8.771, de 11 de maio de 2016, ou na Europa (EC, 2018), a legislação obriga os administradores de sites, a informar ao usuário que eles usam cookies, quando a informação é armazenada. Essa regra geral tem muitas exceções. Mas, na maioria das vezes, é uma situação "pegar ou largar". Ou seja, o usuário não pode escolher se ele quer continuar a acessar a informação, ele tem que aceitar o uso deles, senão tem que deixar de visitar o site. O Regulamento Geral de Proteção de Dados da União Europeia tentou corrigir essa situação. Assim, para sites que atuam na Europa, agora é necessário:

- consenso claro e afirmativo do usuário para cada tipo de dado e cada uso,
- continuar a dar a mesma experiência de uso do serviço oferecido para quem não aceita o uso de cookies - se eles não são indispensáveis para o uso,
- e a possibilidade de desfazer a escolha, anulando o consenso anteriormente afirmado *(opt-out*).

Mas, a lei permite interpretações diferentes, como o consenso para cada tipo de uso de uma só vez, misturando as informações de cookies necessários para o funcionamento de um serviço com as de informações pessoais, até a interpretação de pesquisa de marketing como pesquisa científica, que tem mais liberdade no uso dos dados. Na situação atual, uma vez que o usuário permite o armazenamento das informações no dispositivo, ninguém precisa perguntar para acessar todas as informações contidas neles, mesmo aquelas que permitem identificar o usuário pessoalmente como nome, domicílio ou endereço de correio eletrônico. A ridicularizarão do assunto fica clara quando a European Interactive Digital Advertising Alliance (EDAA) informa aos próprios usuários, que a maioria dos prestadores não utiliza essas informações (EDAA, 2018) . Resumindo: sem leis para a preservação da privacidade do usuário, as empresas podem usar tudo que for tecnicamente possível.

3.6.5. Opt Out

Mesmo em países que ainda não tem a opção de *opt-out*, recentemente introduzido na

União Europeia, com a emancipação dos usuários e o aumento da assertividade, sempre mais empresas oferecem a possibilidade de optar ativamente em não receber publicidade baseado no comportamento e no histórico de navegação na Internet. A Digital Advertising Alliance (DAA, 2018) nos Estados Unidos, a AdChoice , 2018 no Canadá e a EDAA, 2018 na Europa, assim como a Google, 2018, fornecem aos consumidores mais transparência e controle na aplicação de publicidade comportamental. As empresas que participam dos programas se comprometem a não usar as informações salvas nos *cookies*.

3.7. Frequência das mensagens

A lógica à base do aumento do número de e-mail é simples: quanto mais mensagens, maior a possibilidade de pessoas abrirem a mensagem e clicar na oferta contida. Por outro lado, com aumento da frequência dos e-mails, pode aumentar o número de cancelamento de assinaturas. E, se a frequência é muito baixa, tem perigo de a empresa ser esquecida. Um exemplo atual mostrou essa situação. Em maio 2018, entrou em vigor o Regulamento Geral de Proteção de Dados da União Europeia (EC, 2018). A partir dele, todas as empresas que trabalham com dados na União Europeia, precisam atualizar o termo de compromisso de privacidade. Como primeira consequência, as empresas informaram aos clientes sobre o tratamento dos dados pessoais, pedindo uma autorização dos assinantes para conseguir continuar usando-os. O autor recebeu uma informação de todas as empresas com as quais ele entrou em contato até hoje. O resultado foi hilário: de agências de viagens, que não usou há decênios, hotéis onde estava alojado uma única vez, linhas aéreas, centros de mergulho que nunca mais irá visitar, vitícolas, até os escoteiros, que contatou uma única vez, bandas que não escutou há anos, etc. Todas essas empresas têm contato do usuário, as suas informações que nunca mais usou. Quem sabe, se não dava para o autor voltar em um dos centros de mergulho, se eles insistiam um pouco para lembrar as submersões, os cenários e os peixes encontrados.

Com o objetivo de concluir uma venda - partindo de um lead - são necessários de seis a oito interações com o cliente (GLYM, 2018). Para reforçar a ação, aumentar quantidade de lembretes, e com isso, aumentar a quantidade de e-mails parece ser a melhor solução (ROSS, 2018). Mesmo sabendo que mais de 40% dos destinatários queriam uma frequência menor

(WATSON, 2018), a quantidade de e-mails aumenta continuamente. Com o aumento da informação, aumenta a possibilidade de escolha e com ela, o paradoxo da escolha (SCHWARTZ, 2015): quanto maior o número de possibilidades, maior a inseguridade e a desmotivação intrínseca no momento da escolha (IVENGAR, S. S.; LEPPER, M. R., 200) e menor a sua satisfação (SCHWARTZ, 2015).

Existem diferentes possibilidades de verificar a cadência certa de e-mails. A mais dolorosa é dada pela interpretação das razões de cancelamento de assinatura. Quando o destinatário escolhe não receber mais informações, ele pode ser redirecionado para um site, onde é possível informar o motivo do cancelamento. Existem diferentes estatísticas sobre o assunto, muitas vezes indicando a exagerada frequência como principal razão. Mas, quando uma pessoa quer cancelar a assinatura, ainda está disposta a ler todos os motivos enumerados e escolher o adequado? Pode ser que pretenda só parar finalmente a avalanche de mensagens, escolhendo qualquer uma das possibilidades listadas.

Outra possibilidade é dada através da divisão da lista de assinatura em partes iguais, preferivelmente, respeitando a classificação dos componentes da lista, em seguida, utiliza uma frequência de mensagens diferentes para cada uma e mede os resultados depois de um intervalo definido. Mas, nos dois casos, é necessário aceitar um inconveniente grave: a perda de clientes.

Por que não perguntar diretamente ao cliente? No momento da assinatura, ele poderia ter a opção de escolher a frequência das mensagens certas para ele, ou mesmo o dia, quando ele deseja receber as informações. Mas, como fazer quando não tem informações relevantes para o cliente? E se tiver informações urgentes, mas o número definido de mensagens para o período já foi alcançado? E se for uma campanha de descontos?

3.8. As reações negativas

O uso de informações que o usuário nunca passou deliberadamente para as empresas pode terminar em uma personalização abusiva e uma reação negativa da parte do destinatário das mensagens. O mesmo pode acontecer, quando a quantidade de mensagens não obedece à frequência selecionada no momento da assinatura ou fica insuportável. Quando os

destinatários chegam nesse ponto, eles cancelam a assinatura. O cancelamento pode ser total, passando pela retirada do endereço de correio eletrônico da lista de assinantes. Nesse caso, o remetente consegue quantificar as desistências. Mas, quando não é possível localizar a ligação que aponta para o cancelamento, ou o procedimento do mesmo se revela complicado, com a necessidade de fornecer outras informações, o destinatário pode optar por marcar a mensagem como SPAM. Com isso, aumenta a taxa de contestação, e pode diminuir a confiança do emissor. Outra possibilidade seria o redirecionamento de todas as mensagens desse remetente para uma pasta particular para mensagens indesejadas. Em todos esses casos, a empresa perdeu um cliente, sem possibilidade de reverter sua decisão.

3.9. Como medir a percepção

Os três principais recursos usados na arte de persuasão já foram definidas por Aristóteles (384–322 a.C.) na sua obra Arte da Retórica (ROSS, 2018): *ethos*, a ética e credibilidade do orador, *pathos*, as emoções do público e *logos,* os argumentos usados na explicação. Mas, como aumentar as emoções do consumidor, sem perder em credibilidade? Como medir a percepção da privacidade, e perceber o ponto antes de entrar naquele cerco pessoal de não nos sentirmos invadidos? Como conhecer a quantidade certa de informações, antes de perder um assinante ou mesmo clientes? Para psicólogos cognitivos e psicólogos evolucionistas, grande parte do nosso comportamento pode ser explicada através de mecanismos psicológicos (S.U., 1997). Pessoas são essencialmente uma coleção de algoritmos biológicos, formados através de milhões de anos de evolução, que respondem ao ambiente (NOAH, 2017). Nesse trabalho, vamos tentar associar as informações, medi-las, quantificá-las e agrupá-las, para, finalmente, ter um modelo de previsão.

3.10. O custo de um cliente

A extensão dessa pesquisa, a perda de clientes por causa de personalização ou frequência exagerada de mensagens de correio eletrônico, tem uma única finalidade: não perder clientes por causa disso.

O valor de um cliente pode ser quantificado através de diferentes maneiras. Primeiramente, tem o custo de aquisição do cliente. Uma perspectiva simplista enxerga somente os custos diretamente visíveis: passei o meu endereço de e-mail para ter a possibilidade de participar de concurso promocional, receber um brinde, ou um desconto de 10% na aquisição de um produto de R$ 1.000,00. Mas, é mais do que os R$ 100,00 de desconto: é o custo de uma campanha em relação ao número de novos clientes adquiridos durante essa campanha. O investimento é a soma de todas as despesas necessárias: salários e treinamento de pessoas, comissões de venda ou descontos garantidos, custo de ferramentas, software e plataformas usadas, matérias para aquela campanha e custos de serviços profissionais usados.

$$\text{Custo aquisição } = \text{R\$ } \frac{(\text{investimento campanha})}{(\text{número de clientes adquiridos})}$$

Mas, perder um cliente não é somente perder um investimento inicial. Tem diferentes estudos sobre clientes satisfeitos, quanto eles compram a mais, e recomendam a empresa para os outros. Nesse trabalho, o foco é na importância de não perder clientes. Estudos mostram que o custo de retenção de clientes existentes é de cinco até vinte vezes menores que a de aquisição de clientes novos (STENE, 2002). Além disso, quatro em cinco clientes nunca mais voltarão para uma empresa, depois de virar as costas para ele (SDL, 2018).

3.11. Resumo das informações do contexto

E-mail nurturing é uma forma de nutrição de clientes (futuros e atuais), que assinam uma lista de e-mail. A assinatura não é fruto de uma enganação ou de iscas, o cliente escolheu receber as informações.

Retomando o problema inicial, como aumentar o retorno financeiro de uma campanha de e-mail marketing, pode-se constatar que:

- Maior o número das pessoas que seguem o link embutido na mensagem, maior o retorno esperado da campanha.

- Aumentar a quantidade de mensagens enviadas aumenta a probabilidade de seguir o link embutido.
- Com um número constante de assinantes, a quantidade de mensagens pode ser aumentada somente aplicando uma frequência maior.
- Quando a frequência de mensagens é alta demais, os destinatários podem desistir da assinatura.
- Personalização é importante para aumentar a visibilidade das mensagens.
- No momento da assinatura, é difícil obter informações dos assinantes.
- É bastante fácil obter informações pessoais de terceiros.
- O uso de personalização indevida pode ser sentido como invasão de privacidade.
- Percepção de invasão da privacidade e frequência insuportável de mensagens provoca a perda de clientes.
- O custo de manutenção do cliente é muito mais baixo comparado ao investimento necessário para a angariação de clientes novos.
- Um número alto de mensagens marcadas como spam diminui a fiabilidade do remetente.
- As métricas encontradas não tem validade universal, mas são fortemente dependentes do contexto.
- Pelas métricas encontradas, o cliente brasileiro não desiste facilmente de uma assinatura.

Capítulo 4 – Aplicação do processo no Estudo De Caso

A forma mais direta de coletar os dados para a classificação seria a medição das reações no campo. Basta dividir a lista de assinantes em diferentes partes, com características equilibradas entre elas, e variar frequência e personalização. Mas, pelo custo dos clientes, ninguém quer sacrificá-los em um experimento. Além disso, as reações dependem de fatores individuais, do ambiente e do momento do recebimento da mensagem. Já mencionamos alguns desses fatores falando da taxa de abertura. Resumindo: somente quando a pessoa certa, recebe a mensagem certa, na hora certa e na situação certa, o conteúdo dela será percebido da maneira esperada pelo remetente. O destinatário deve estar na condição de ler, de poder clicar nos link embutidos, ter a necessidade, além da responsabilidade e o dinheiro, para uma eventual aquisição. Com tantas variáveis, apresenta-se inviável o monitoramento de uma lista de e-mail real. É necessário criar uma situação que não seja influenciada pelo conteúdo e a apresentação dele na mensagem, assim como a situação do destinatário. Precisamos focar somente nos parâmetros que queremos identificar na pesquisa: personalização e frequência. Por essa razão, preferiu-se renunciar a uma pesquisa de campo e usar um questionário. Através dele, é possível usar uma introdução e definir vários cenários; colocar os respondentes em situações diferentes; e recolher as respostas para esses contextos descritos.

O questionário será aplicado de maneira anônima. Isso implica a renúncia a uma eventual entrevista, mas, como já mencionado antes, o anonimato aumenta a percepção de controle (DINEV, SMITH, HART, 2017] e pode resultar em respostas mais sinceras. Em nenhum momento, os participantes informam qualquer dado de identificação, como nome, RG, CPF ou endereço e-mail.

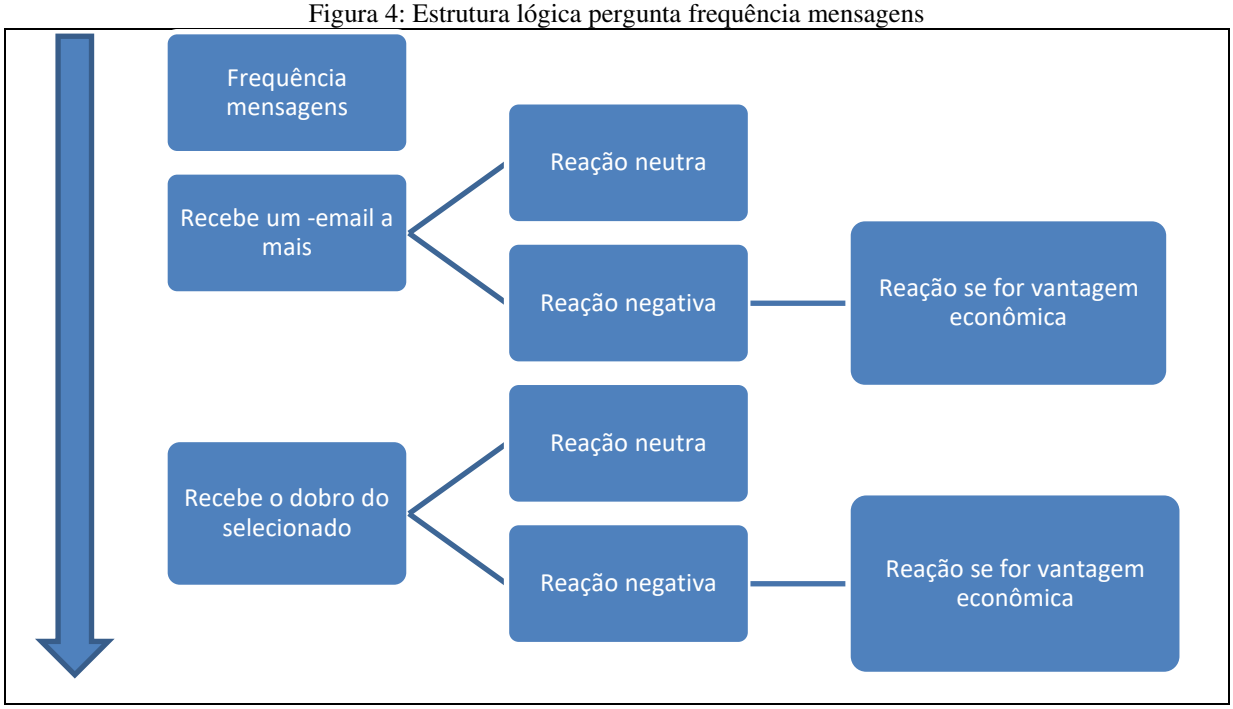
Para essa pesquisa, o anonimato tem um lado negativo. Queremos medir as respostas de brasileiros. Para minimizar as respostas de não brasileiros, todo o questionário será disponível somente em português. Isso não dá certeza na questão, mas ajuda a eliminar respostas de estrangeiros que não dominam esse idioma.

4.1. Definição das perguntas

A próxima etapa foi a definição das perguntas (apêndice1). Aqui, a primeira parte, é referente à melhor frequência de mensagens. As possibilidades oferecidas se estendem de um e-mail por mês até um por dia. Após o respondente informar a quantidade de mensagens

desejada no momento da assinatura, criamos situações para entender a reação para o aumento das mensagens. O fluxo das perguntas é representado na Figura 4. A primeira pergunta é sobre a reação se chegasse uma só mensagem a mais; a segunda se a quantidade de e-mail dobrasse. Em caso de uma atitude negativa por parte do respondente, queremos entender se a percepção muda, se as mensagens contivessem uma vantagem econômica.

Figura 4: Estrutura lógica pergunta frequência mensagens



FONTE(PRÓPRIA)

A definição das perguntas relativas à privacidade e ao manuseio das informações pessoais é mais complexa. O intuito é entender qual informação é mais valiosa, mais pessoal para o respondente e, consequentemente, medir sua reação, no momento que uma dessas informações fosse usada para a personalização de uma mensagem a ele dirigida. Para isso, primeiro é necessário entender o grau de resistência no momento do fornecimento da informação: quanto mais o respondente resiste ao passar a informação, mais alto será o valor:

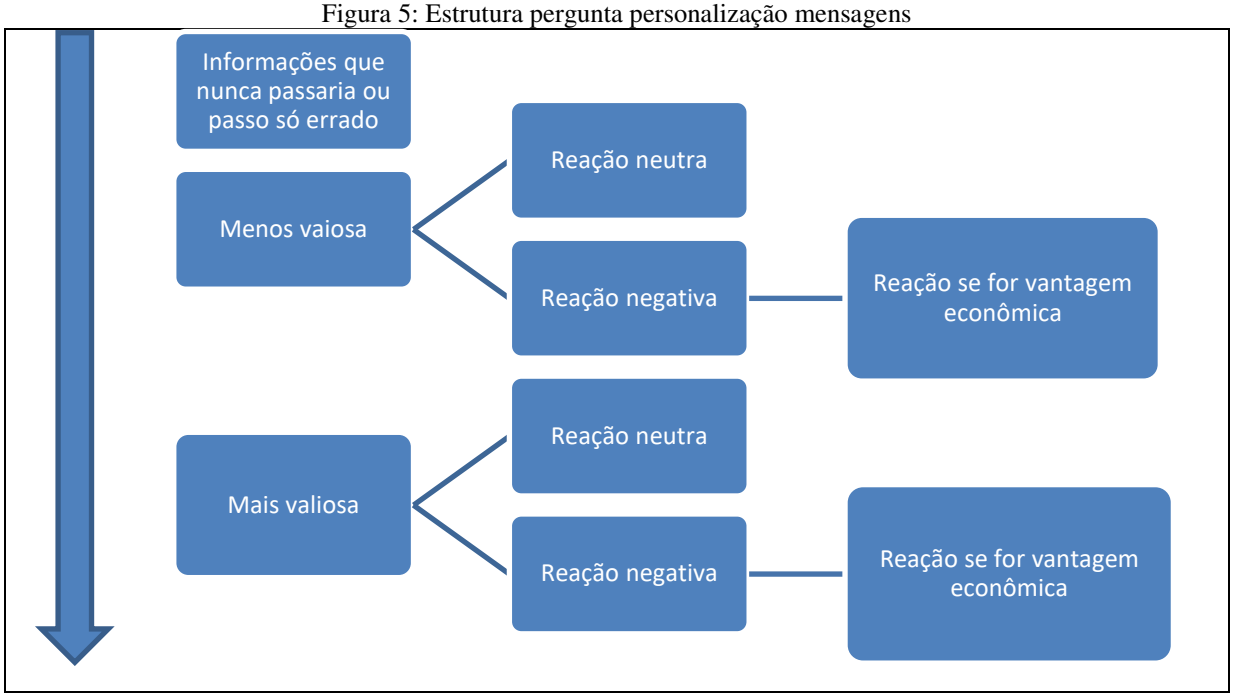
- Passo sem problema
- Passo somente se for necessário
- Nunca passaria
- Se for necessário, passaria informação errada.

Como representado na Figura 5, dentro do universo das informações que o respondente passaria somente de forma errada, ou nunca passaria, são escolhidas a

informação menos e mais valiosa, para medir a sua reação.

Como nas perguntas referentes à frequência, em caso de uma atitude negativa por parte do respondente, é medida a reação, se as mensagens contivessem uma vantagem econômica.

Figura 5: Estrutura pergunta personalização mensagens



FONTE: (PRÓPRIA)

Além da coleta de dados para as respostas de nossas perguntas principais, há informações demográficas como gênero, faixa etária, ocupação e estudo. Na definição das áreas de estudo, distinguimos entre computação e exatas, pressupondo que pessoas da área de computação tem mais afinidade com quesitos de privacidade na internet. A análise de dados vai ajudar a entender se isso é verdadeiro.

A definição das perguntas sobre o comportamento do usuário no dia a dia foi feita através de uma pesquisa qualitativa. O questionário usado para isso tinha uma única pergunta: Qual atitude testemunha o uso leviano da Internet? As respostas recebidas foram agrupadas e normalizadas, assim que respostas parecidas foram representadas de uma das respostas recebidas. Elas são de áreas diferentes, e permitem em combinação com o comportamento selecionado para cada um, a criação de um perfil do usuário relativo a essa sua atitude. As perguntas e os comportamentos usados na pesquisa são:

Atitudes:

- Limito a visibilidade das minhas postagens.
- Uso aplicativos do Facebook que acessam os meus contatos.
- Crio outras contas com Login do Facebook.
- Aceito convites de estrangeiros nas redes sociais.
- Permissões de aplicativo:
    - Android: controlo as permissões de um aplicativo antes da instalação
    - Iphone: limito as permissões dos aplicativos.
- Uso sites que precisam de cookies.
- Uso a possibilidade de opt-out para cookies.
- Uso motores de busca alternativos.
- Uso proxies para navegar.
- Uso redes Wi-Fi abertas
- Cubro a câmera do meu dispositivo.

Os possíveis comportamentos dão a possibilidade de mostrar falta de familiaridade ou uso ("Não sei", "Não se aplica"), e os cinco valores para a possível aplicação:

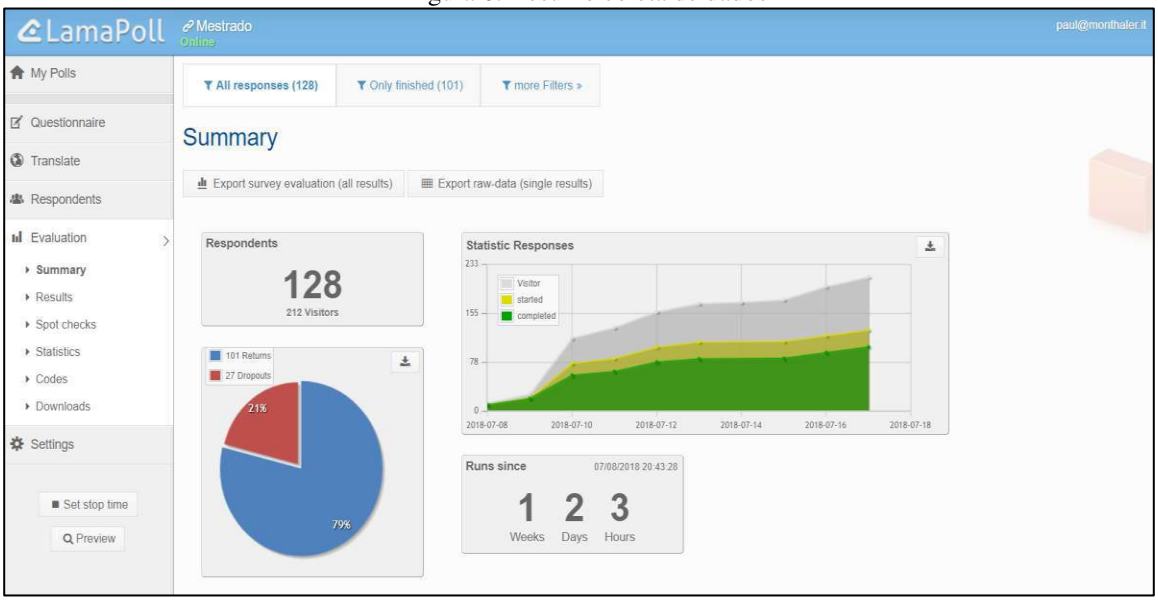
- Sempre, Na maioria das vezes, Meio a meio, De vez em quando, Nunca

A coleta de dados foi feita através do site LAMAPOLL, 2018. A ferramenta para a criação de enquetes é intuitiva, eficaz e potente. O design responsivo permitiu responder ao questionário através de um computador pessoal, notebook, tablet ou mesmo smartphone. Além disso, permite a criação de um projeto personalizado, e de questionários dinâmicos. Isso foi um pré-requisito para conseguir criar perguntas e respostas com visibilidade limitada e saltos lógicos, possibilitando o dobro da quantidade de e-mails selecionados, ou a definição da informação mais ou menos valiosa. A partir disso, foi possível limitar o tamanho do questionário, sem renunciar a informações complexas.

4.2. Análise da amostra

O sumário da coleta de dados é mostrado na Figura 6**: Resumo coleta de dados**.

Figura 6: Resumo coleta de dados

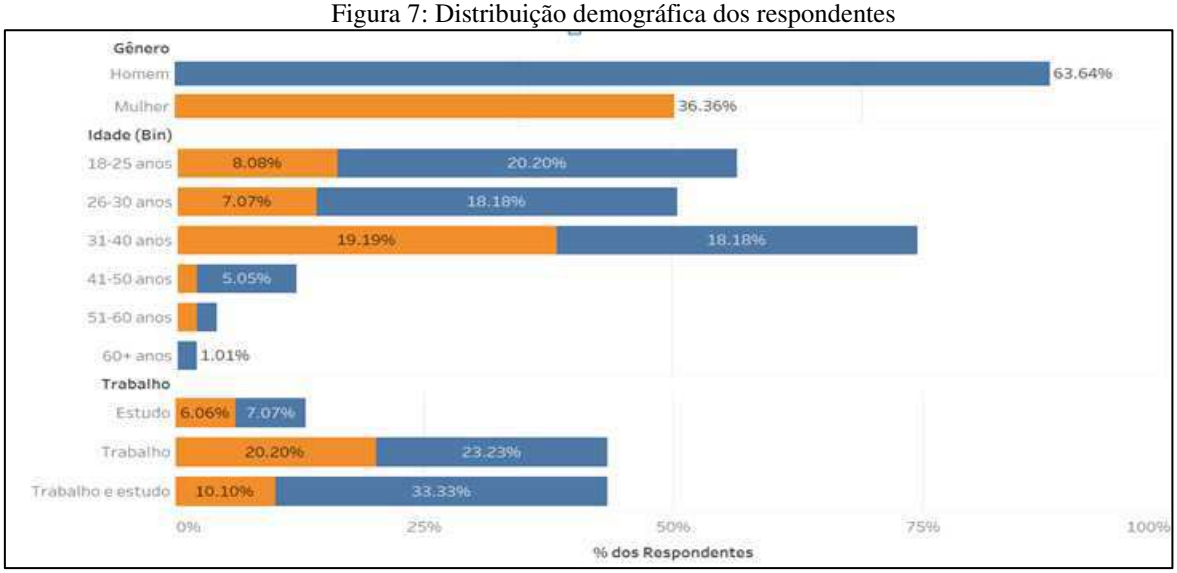


FONTE: (LAMAPOLL, 2018)

Os dados foram coletados entre dia 8 e 17 de julho de 2018. De 212 pessoas que acessaram o questionário, 128 responderam, dos quais 101 chegaram até o fim.

Das 101 pessoas, pouco mais de um terço (36%) são mulheres. Em geral, a faixa etária mais representada é entre 31 e 40 anos, mas especificamente para os homens tem mais respondentes entre 18 e 25 anos, e em seguida as faixas de 26 até 30 e 31 até 40 anos, sempre com a mesma proporção. A grande maioria das pessoas trabalha, sendo que somente 13% dos respondentes declararam se dedicar apenas aos estudos (Figura 7).

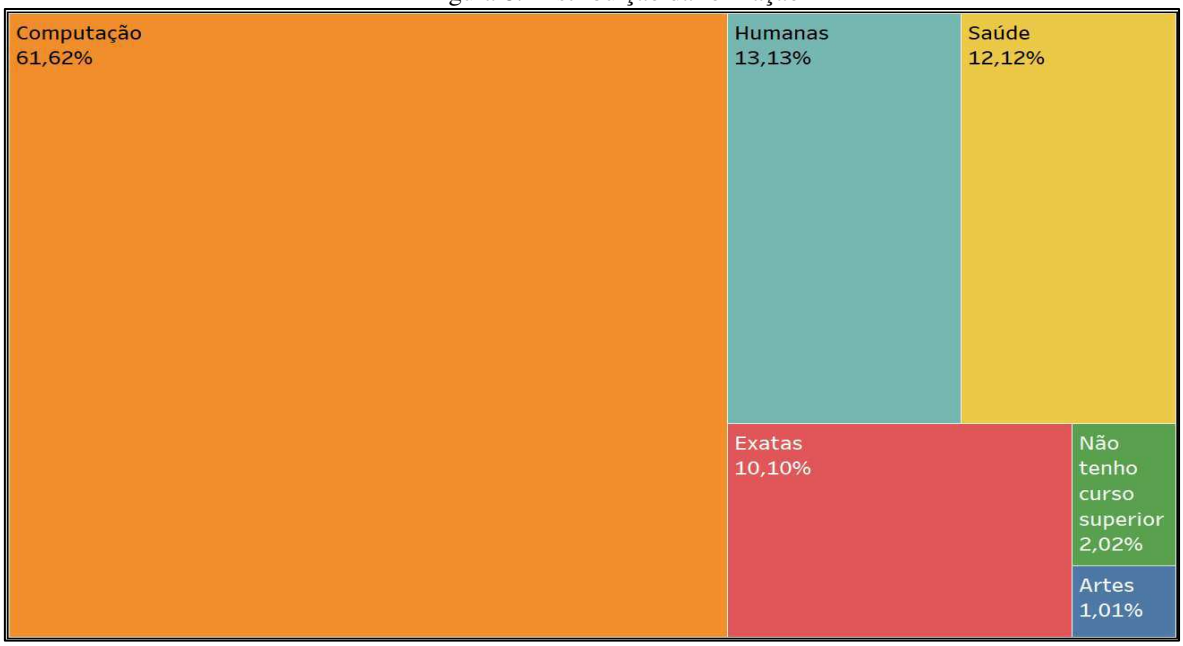
Figura 7: Distribuição demográfica dos respondentes



FONTE: (PRÓPRIA)

Entre as pessoas que trabalham, a metade ainda está cursando uma graduação. A distribuição do tipo de formação é visível na Figura 8. Foi decidido de dividir os ramos entre exatas e humanas, e as pessoas com e sem formação. A distinção de computação dentro de exatas foi baseada na pressuposição do comportamento diferente de pessoas que estudavam informática, sobretudo em quesitos de privacidade e segurança. No mesmo momento, a distribuição dos cursos, mostra um víeis dos dados. Mais de 60% de todos os respondentes que chegaram até o final do questionário, são da área de computação.

Figura 8: Distribuição da formação



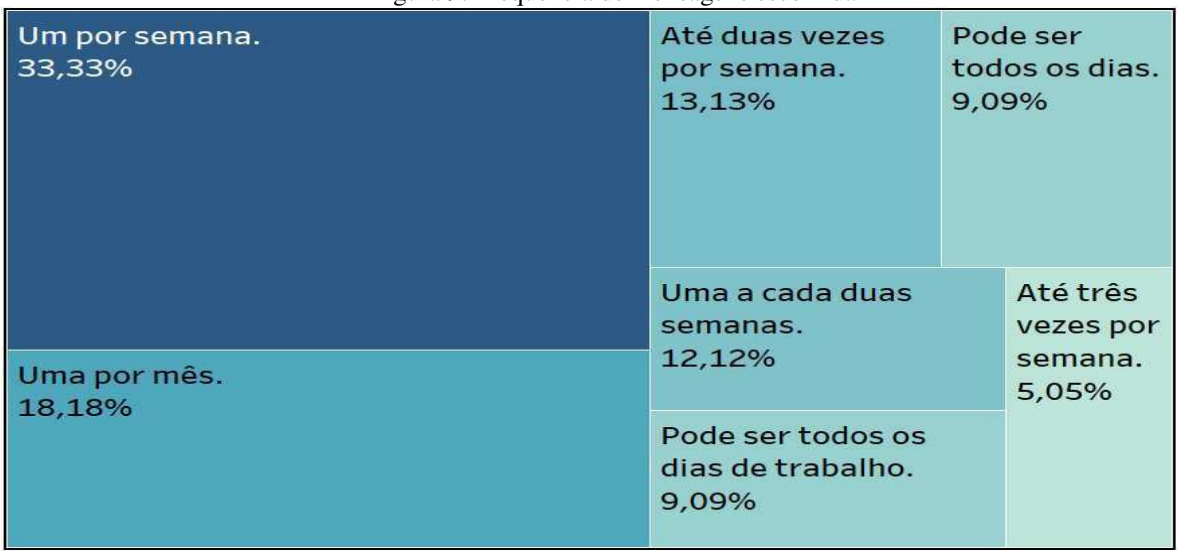
FONTE:((PRÓPRIA)

4.3. Distribuição dos valores encontrados

O aumento do número de mensagens é diretamente proporcional ao resultado positivo das campanhas – mesmo quando uma parte muito grande pretendia uma frequência menor. Isso foi mostrado na parte sobre a melhor frequência de mensagens – baseado em estudos estrangeiros. Isso vale também para o consumidor brasileiro? E qual seria a taxa de perda de assinantes se exageramos com a quantidade de mensagens? Segundo os dados mencionados na parte relativa aos indicadores do EM, em geral, a soma de taxa de cancelamento e da taxa de contestação é de 0,21% para o Brasil. Como apontado, esse valor é mais baixo que a média mundial (0,26%), indicando que o brasileiro não desiste facilmente das mensagens recebidas.

Iniciamos por partes. Antes de tudo, é necessário ver qual a frequência que os respondentes escolheram no momento da assinatura de uma lista de e-mail. Um terço das pessoas queria apenas até duas mensagens por mês (18% uma, 12% duas mensagens), e 32% até uma mensagem por semana (Figura 9).

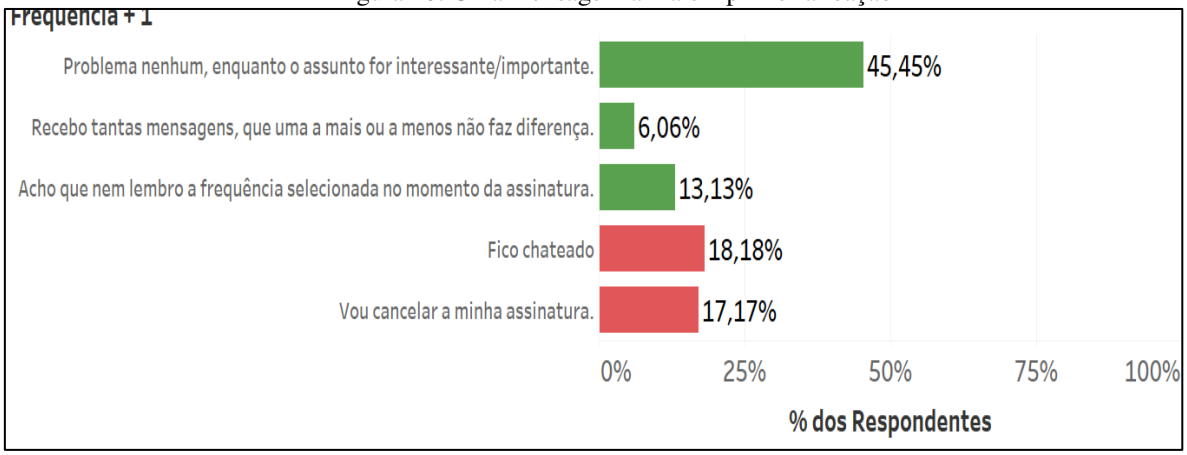
Figura 9: Frequência de mensagens escolhida



FONTE: (PRÓPRIA)

E como as pessoas reagem, quando a quantidade de e-mails aumenta, quando chega uma mensagem a mais do que foi escolhido no momento da assinatura da lista de e-mail? Olhando a distribuição das respostas na Figura 10, parece que os estudos encontrados e citados não estão certos: só 64% dos respondentes, marcados em verde, não tem problema, quando chega uma mensagem a mais.

Figura 10: Uma mensagem a mais – primeira reação

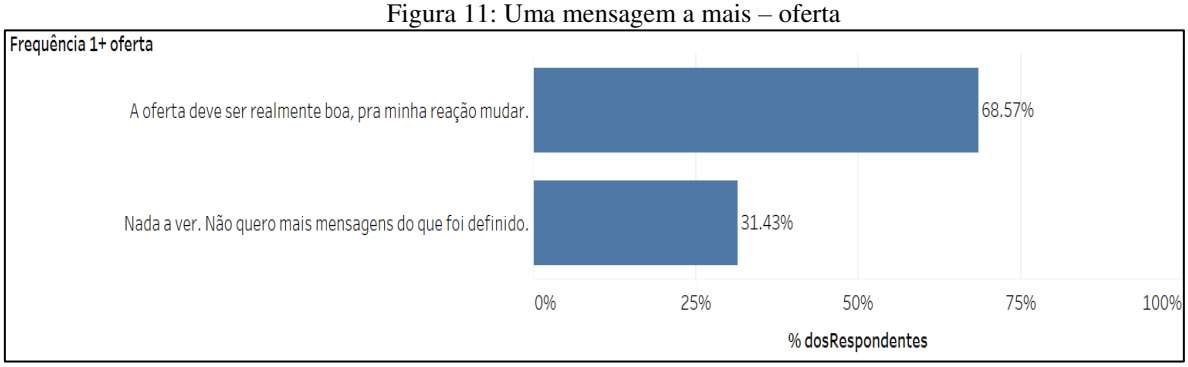


FONTE: (PRÓPRIA)

Com isso, significa que 36% dos consumidores brasileiros não gostam que

desrespeitem sua opção. Só por causa de uma única mensagem a mais, 17% optam por cancelar a assinatura. E como essas pessoas reagem, se a mensagem a mais contiver uma oferta (Figura 11)? Pelos estudos citados anteriormente, o consumidor tem um preço, e quanto melhor a oferta, mais ele perdoa.

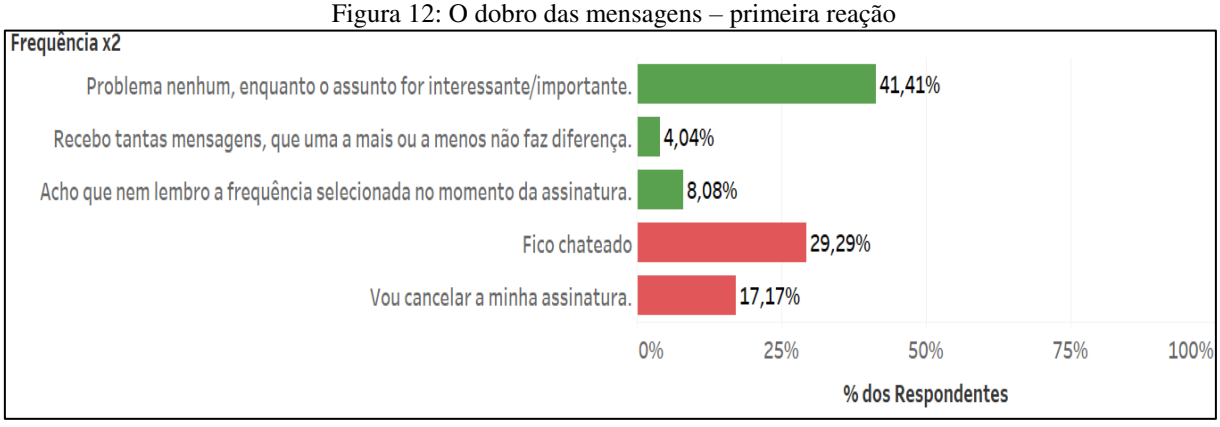
Figura 11: Uma mensagem a mais – oferta



FONTE: (PRÓPRIA)

Mais de 68% aceitam a mensagem – se a oferta for boa. Do outro lado, 31% recusam a mensagem a mais. Focando nas pessoas que já ameaçaram cancelar a assinatura na primeira resposta, quase 8% de todos os respondentes continuam firme na decisão. E se em vez de uma mensagem a mais, chegasse o dobro do que foi selecionado no momento da assinatura? A diferença com a primeira situação é considerável (Figura 12).

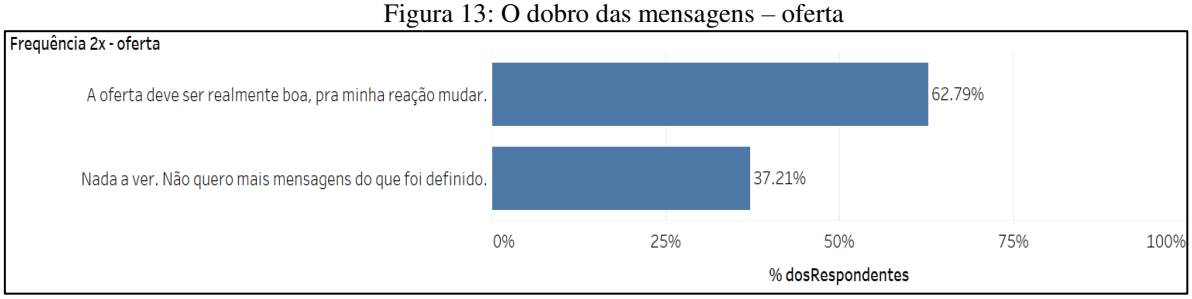
Figura 12: O dobro das mensagens – primeira reação



FONTE: (PRÓPRIA)

Das 35% de pessoas que tinham uma reação negativa, agora, 46% dos respondentes não gostam da inundação de mensagens. Mesmo assim, o número de pessoas que optaria por cancelar a assinatura fica em 17%. Como as pessoas reagem se as mensagens têm um cupom ou uma oferta? A resposta está explícita de maneira visual na Figura 13.

Figura 13: O dobro das mensagens – oferta



FONTE: (PRÓPRIA)

Nela, podemos ver que agora tem um aumento notável das pessoas que não aceitam a mensagem, mesmo com a possibilidade de ter uma vantagem econômica. Das 31% na primeira situação, agora 37% das pessoas recusam as mensagens. Olhando os respondentes que optaram por cancelar a assinatura na resposta anterior, mais de 70% deles continuam com a mesma intenção. Isso significa que quase 12% dos assinantes da lista seriam perdidos.

A primeira avaliação das respostas mostra que o brasileiro definitivamente não gosta de ser inundado com mensagens – mesmo quando escolheu assinar uma lista de e-mail. O número de cancelamento, quando o número de mensagens extrapola o desejado, é alarmante. Nessa situação, a desistência geral de campanhas de EM é quase 60 vezes a perda considerada normal.

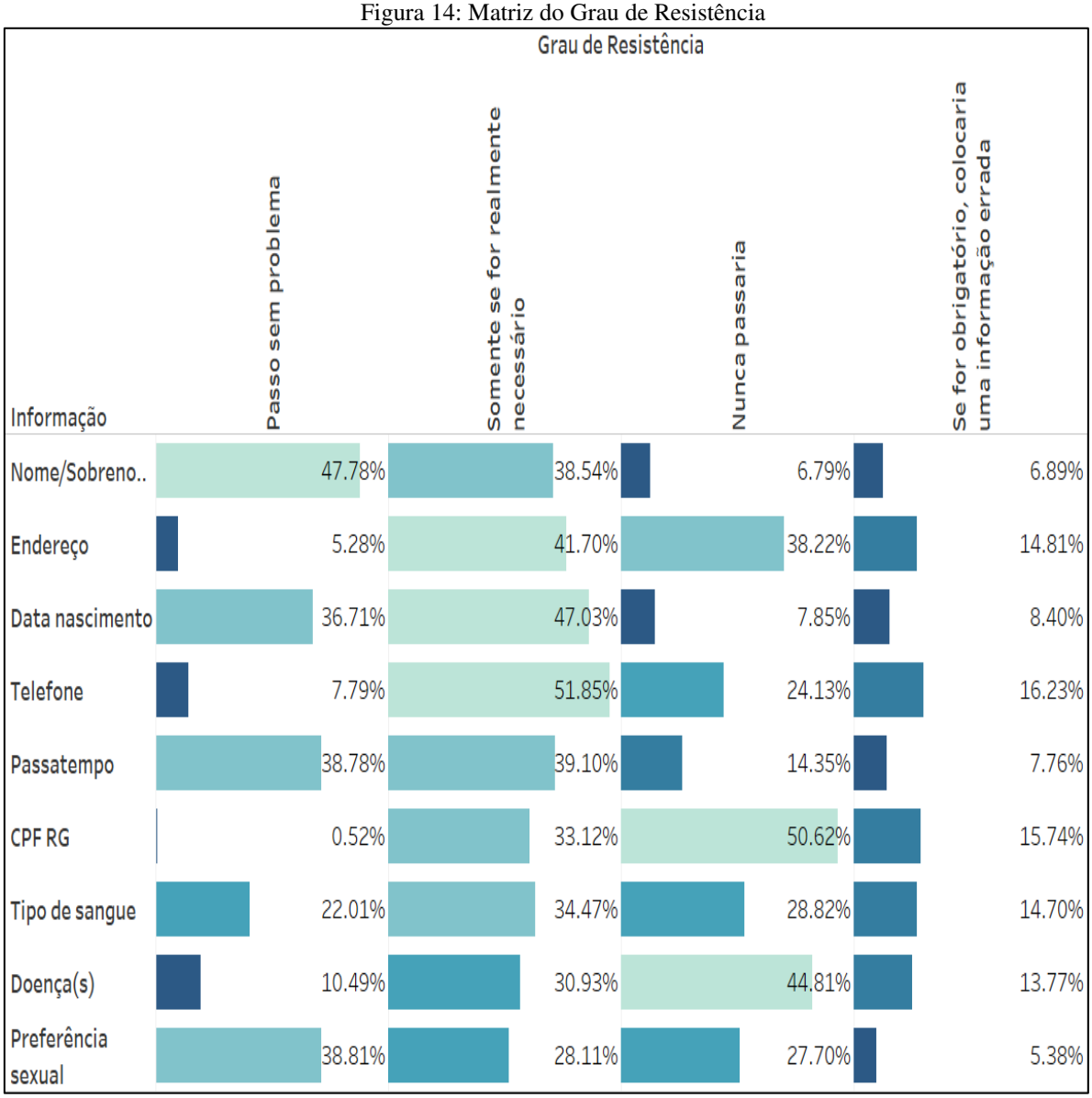
4.4. Personalização das mensagens

A classificação dos assinantes de uma lista ajuda na definição do público alvo e aumenta a probabilidade de ter os destinatários certos para a campanha. As informações usadas para essa classificação nem sempre são fornecidas pelos assinantes da lista. Além de usar os metadados, é possível a troca ou venda de informações de usuários. Enquanto essa informação é usada para criar grupos de assinantes com características parecidas para diferentes campanhas, ninguém consegue enxergar a existência dos dados de proveniência duvidosa. Mas, quando essas informações são usadas também para a personalização das mensagens, o destinatário pode ter uma sensação de privacidade invadida. Qual a percepção do consumidor brasileiro?

Perguntamos aos respondentes para diferentes tipos de informação, qual o grau de

resistência no fornecimento dela. As informações listadas são de caráter pessoal, e podem ser classificadas como demográficas (nome e sobrenome, endereço, data de nascimento e telefone), atividade de entretimento livre (passatempo preferido), documentos pessoais, (Cadastro de Pessoas Físicas, CPF e o Registro Geral, RG), dados de saúde (grupo sanguíneo e doenças) e da vida íntima (preferência sexual). A primeira surpresa na distribuição encontrada (Figura 14) está no comportamento em relação a informações sensitivas.

Figura 14: Matriz do Grau de Resistência



FONTE: (PRÓPRIA)

Existem diferentes definições de informação sensitiva (COUPER, M. P. et. al., 2008), mas, em geral, essa categoria contém informações sobre uso de drogas, preferência sexual, voto nas eleições ou sobre a renda. Para evitar constrangimentos, as pessoas preferem não

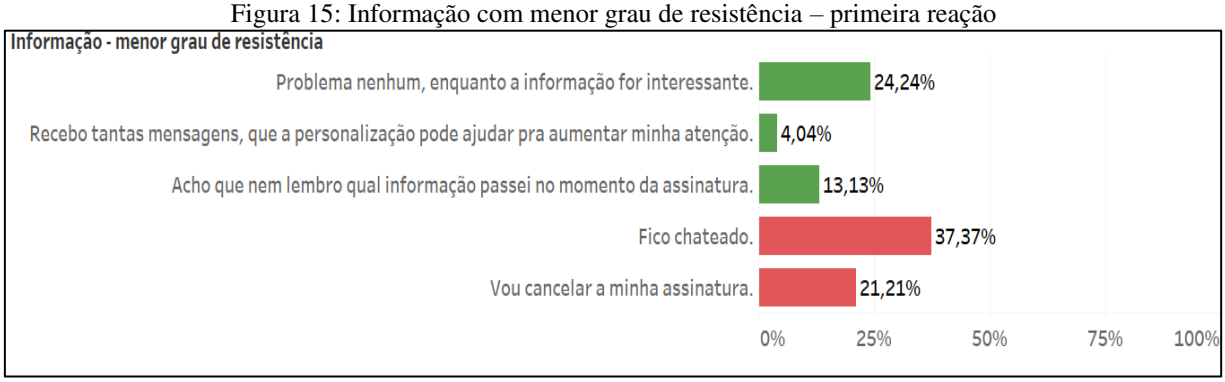
responder ou passar informação falsa em pesquisas (TOURANGEAU, R.; YAN, T., 2007).

Nos dados, 39% dos respondentes passariam informações sobre preferências sexuais e do passatempo preferido, ficando somente atrás da informação do nome (48%). A informação mais valiosa, dentre o universo das respostas oferecidas, são o CPF e o RG, porque 51% dos respondentes nunca passariam essa informação. Em seguida, as informações sobre doenças (45%) e o endereço (38%). O número de telefone também está em uma situação surpreendente: só 8% dos respondentes passariam o número de telefone sem problema, 52% só se for realmente necessário (valor mais alto dessa categoria) e 16% (de novo valor mais alto da categoria) forneceriam uma informação errada, quando o campo for obrigatório.

O questionário permitiu selecionar, dentro das informações que o respondente não passaria (ou passaria errada), a informação mais e menos valiosa. Dessa forma, foi possível qual fosse a reação dele, se esse tipo de informação fosse usado para personalizar um e-mail de EM, sem ter passado a informação.

De início, a reação com menos valor. Somente 42% dos respondentes aceitariam essa personalização indevida, enquanto 58% reagiram de forma negativa (Figura 15).

Figura 15: Informação com menor grau de resistência – primeira reação

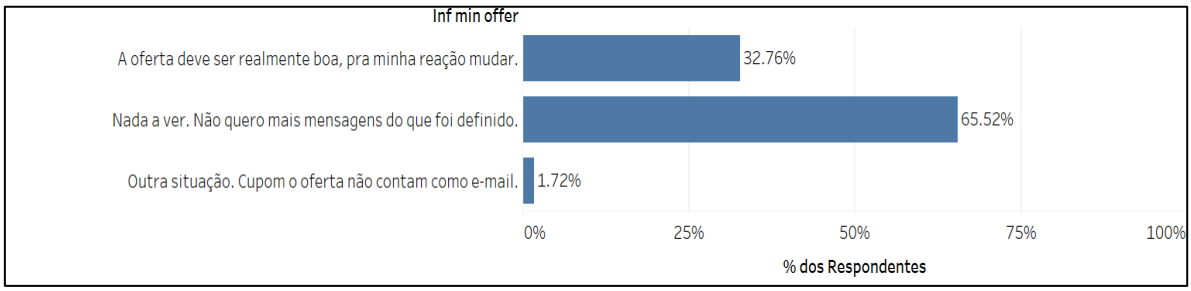


FONTE: (PRÓPRIA)

Como no caso da frequência de mensagens, os respondentes que discordam da prática, são confrontados com a possibilidade de ter recebido uma mensagem personalizada indevidamente, mas com uma vantagem financeira. O resultado é claro: 67% não estão agradados com a oferta. A verdadeira surpresa é a relação das pessoas que já queriam cancelar a assinatura depois da primeira pergunta e as que não aceitam a vantagem financeira incluída na mensagem, porque 94% delas continua com a mesma ideia e ficam com a primeira escolha. Com isso, quase 20% dos assinantes da lista de e-mail são perdidos já com a primeira tentativa de personalização indevida. E se a informação usada for àquela definida como a

mais sensível pelos respondentes? O gráfico da Figura 16 mostra a distribuição.

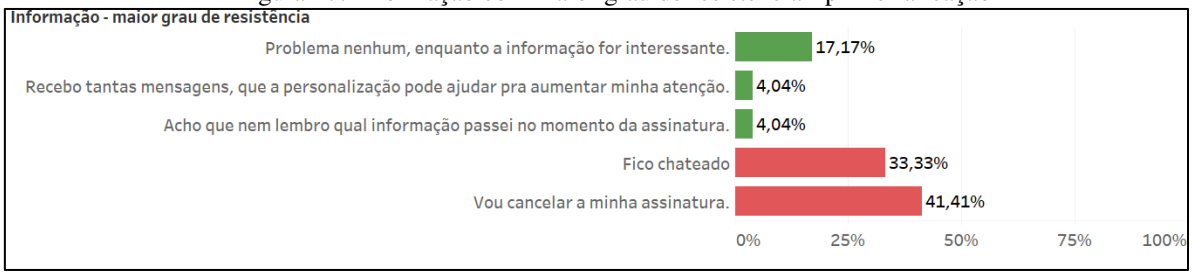
Figura 16: Informação com menor grau de resistência – oferta



FONTE: (PRÓPRIA)

A quantidade de pessoas que não aceitaria a indevida individualização cresceu de 10% e chega a 77%. E como essas pessoas reagem, quando as ofertas recebidas contêm uma oferta irresistível? Nesse caso, o aumento não é tão alto. O gráfico de barras (Figura 17) mostra um aumento de sete pontos, comparado com a mesma situação da informação menos sensível. O aumento nesse grupo é ainda maior, atingindo uma diferença de 11 pontos.

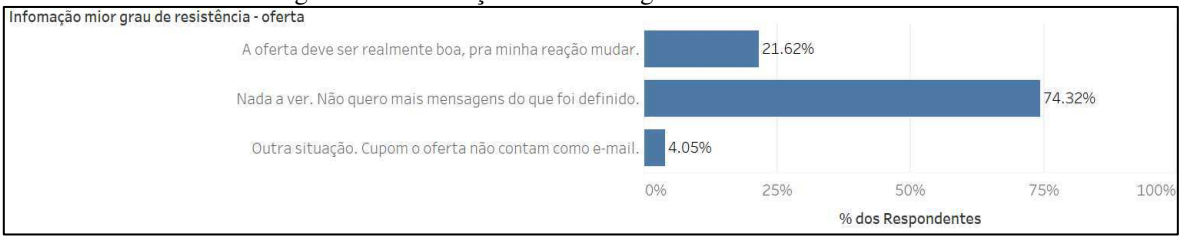
Figura 17: Informação com maior grau de resistência - primeira reação



FONTE: (PRÓPRIA)

O número que mais interessa é sempre o do cancelamento das assinaturas (Figura 18). 86% das pessoas que já ameaçavam cancelar a assinatura na primeira resposta, não se deixam levar pela oferta. Com isso, 40% dos assinantes seriam perdidos depois de usar a informação classificada como a mais sensível das respondentes.

Figura 18: Informação com maior grau de resistência - oferta



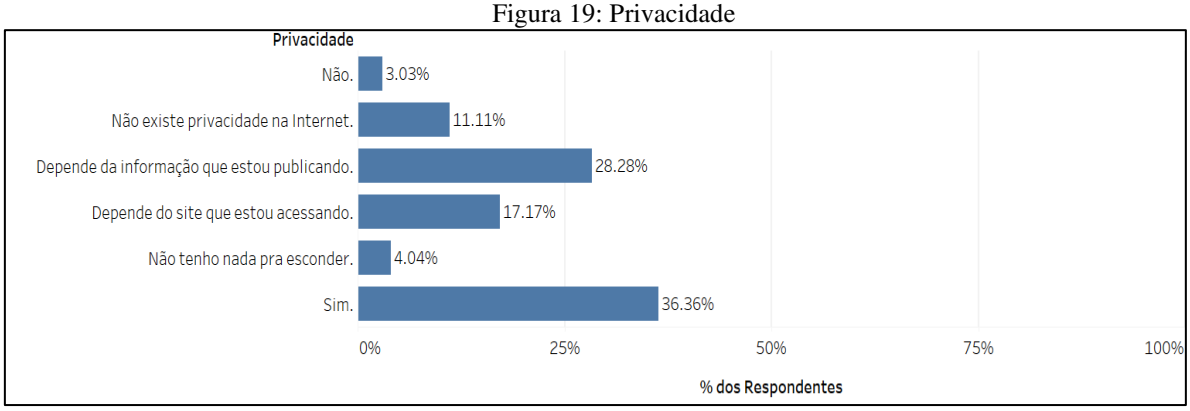
FONTE: (PRÓPRIA)

4.5. Privacidade

Além dos dados sobre frequência e grau de resistência a fornecer informações, foram coletados dados sobre a atitude dos respondentes sobre a privacidade e o comportamento deles em situações de redes sociais, uso de navegadores e preocupação com a própria privacidade na Internet.

Para entender como as pessoas classificam a si mesmos, foi perguntado se elas cuidam da privacidade na internet. As respostas são resumidas na Figura 19. Só uma minoria (3%) informa que não cuida da privacidade. Para quase 30%, é uma questão de natureza dos dados, enquanto a maioria, 36%, afirma cuidar da privacidade.

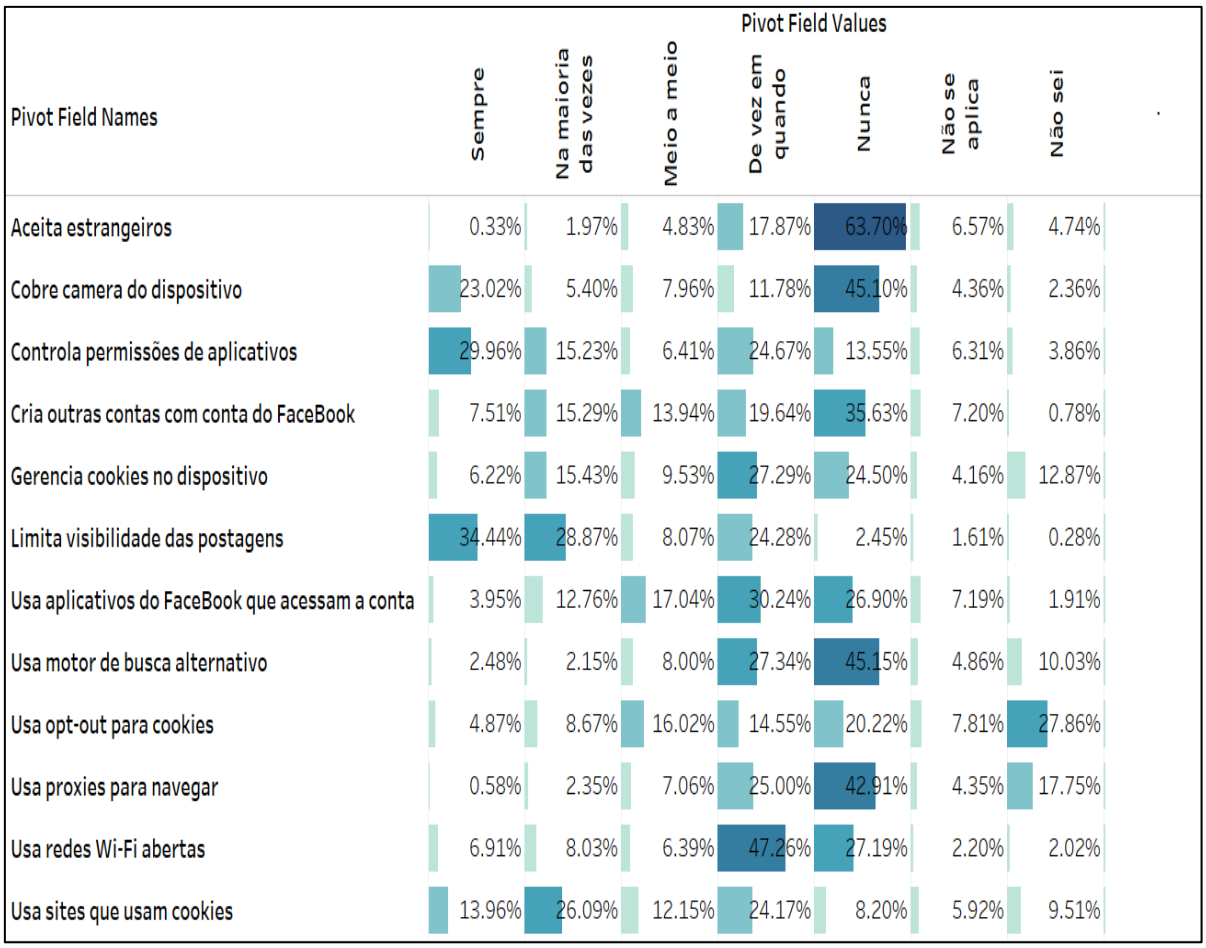
Figura 19: Privacidade



FONTE: (PRÓPRIA)

Depois da pergunta geral, os respondentes receberam a possibilidade de informar sobre algumas atitudes em relação à privacidade no dia a dia. Foram perguntas mais simples: sobre aceitar convites de estrangeiro nas redes sociais, limitar as visibilidades das postagens; até umas atitudes muito limitativas, como o uso de redes Wi-Fi abertas e o comportamento com os cookies. A formulação das perguntas foi até técnica, para ver se os respondentes tinham bastante conhecimento. Figura 20, representando a matriz das atitudes relativas à privacidade, resume as respostas recebidas.

Figura 20: Matriz atitude privacidade



FONTE: (PRÓPRIA)

4.6. A classificação de assinantes

As respostas analisadas permitem diferentes deduções. À primeira vista, chama atenção o grau de resistência medido no ato de fornecer informações pessoais. Mesmo cientes das limitações dos dados coletados, os resultados estão longe dos valores publicados na literatura. Quando o número de mensagens é maior que o selecionado, ou a personalização abusiva é percebida como intrusão na privacidade, uma notável parte dos destinatários decide cancelar a assinatura. Do outro lado, as pesquisas citadas acima mostram a relação direta entre o aumento da comunicação e da personalização e o aumento do resultado. Lembrando que mais da metade dos respondentes (53%) aceitam até o dobro de mensagens do que foi escolhido no momento da assinatura. Mas, a resistência é importante, a perda de assinantes sai

muito cara, quando deve ser compensada com a aquisição de novos clientes. Qual é o perfil dos assinantes que aceita mais mensagens? Como excluir as pessoas que reagem com o cancelamento da assinatura?

Optou-se pela criação de variáveis deduzidas das informações coletadas. As faixas etárias são um exemplo disso. A classificação usada para a idade prevê os seguintes grupos: 18–25, 26–30, 31–40, 41–50, 51–60 e acima de 60 anos.

As outras classes foram criadas baseadas nas distribuições de valores coletados. São as classes de velocidade de resposta, grau de resistência no fornecimento de informações, preocupação com a privacidade no uso de ferramentas, aplicativos e dispositivos no dia a dia. Para todos os valores foram usados os percentis que determinam os 25% e 75% dos dados.

A distribuição do tempo de resposta mostrou algumas observações com um grande afastamento dos demais (*outlier*). Já que as pessoas em questão completaram o questionário, esse valor foi reduzido para 60 minutos. Com uma média e um desvio padrão de 11 minutos, esse valor ainda consegue indicar um tempo muito superior ao normal, sem interferir demais na avaliação das outras observações.

Para classificar o grau de resistência no fornecimento de informações foram somadas as respostas de cada tipo por respondente. Quando uma pessoa passa menos que o valor dos 25 percentis, atualmente 4 respostas, então ela é enquadrada como fundamentalista, quando passa pelo menos, quanto os 75 percentis (9 respostas), então é definida como despreocupada. As classificações usadas para o grau de resistência são adotadas por WESTIN (1967), usando as expressões portuguesas (despreocupado, realista e fundamentalista) para as classificações instituídas por ele (*privacy unconcerned, pragmatic, fundamentalist*). As mesmas classes foram usadas para classificar os respondentes relativos ao comportamento com a privacidade nas redes sociais, na navegação do dia a dia e o uso dos dispositivos.

A última variável criada usando os dados existentes foi relativa ao comportamento nas redes sociais e o uso da internet no dia a dia. Com poucas ações restritivas, a pessoa é classificada como “despreocupada”, com muitas, “fundamentalista”. Os valores usados são resumidos na Tabela 6. As informações sobre as atitudes dos respondentes surpreenderam em alguns pontos. Primeiro, os dados não mostram o que o preconceito estava mandando. Das pessoas que responderam o questionário, a maioria é bastante restritiva nas atitudes do dia a dia.

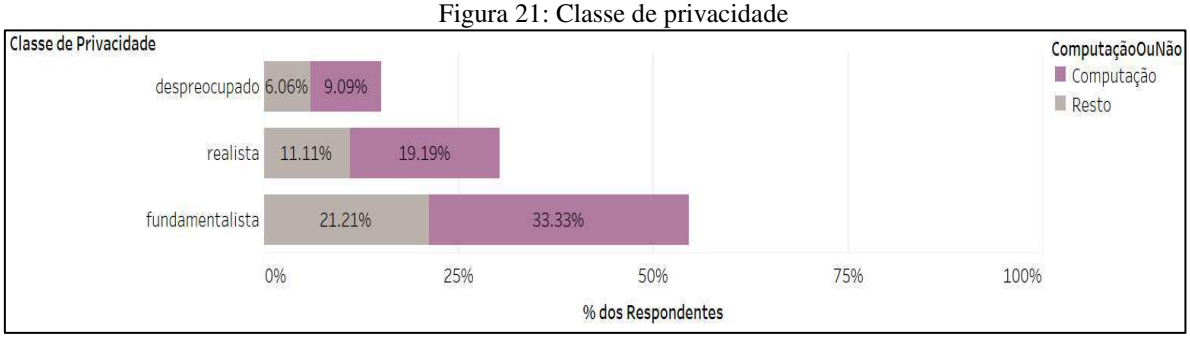
Tabela 6: Categorias inferidas de comportamento

Classificação	Valor usado	< 25%	< = 75%	> 75%
Velocidade de resposta	Tempo de resposta	Veloz	Normal	Lento
Preocupação com informações	Quantidade "Passo sem problema"	Fundamentalista	Realista	Despreocupado
Classe de privacidade	Quantidade das atitudes restritivas	Despreocupado	Realista	Fundamentalista

FONTE: (PRÓPRIA)

Observando as distribuições em relação ao gênero e a faixa etária dos respondentes (Figura 21), não foi possível encontrar grandes diferenças entre eles. Diferentemente para o tipo de formação das pessoas. Aqui, foi possível confirmar o que se havia dito anteriormente: respondentes que estudam – ou estudaram computação – são mais restritivos que o resto.

Figura 21: Classe de privacidade



FONTE: (PRÓPRIA)

Preferimos usar os dados brutos, com as respostas em formato de texto, em vez de recorrer à codificação automática dos resultados.

4.7. Aprendizado de Máquina –não supervisionado

Os dados presentes estão limpos, no entanto, a quantidade é reduzida. Mesmo assim, será criado um algoritmo, capaz de receber muito mais dados. Ao final, a arquitetura do framework selecionado é concebida para o uso com grandes volumes de dados.

O tempo depende do algoritmo, da quantidade dos dados, e dos recursos de cálculo a

disposição. De novo, os dados existentes e natureza dos clusters usados em Databricks ajudam a diminuir o tempo necessário para receber a resposta.

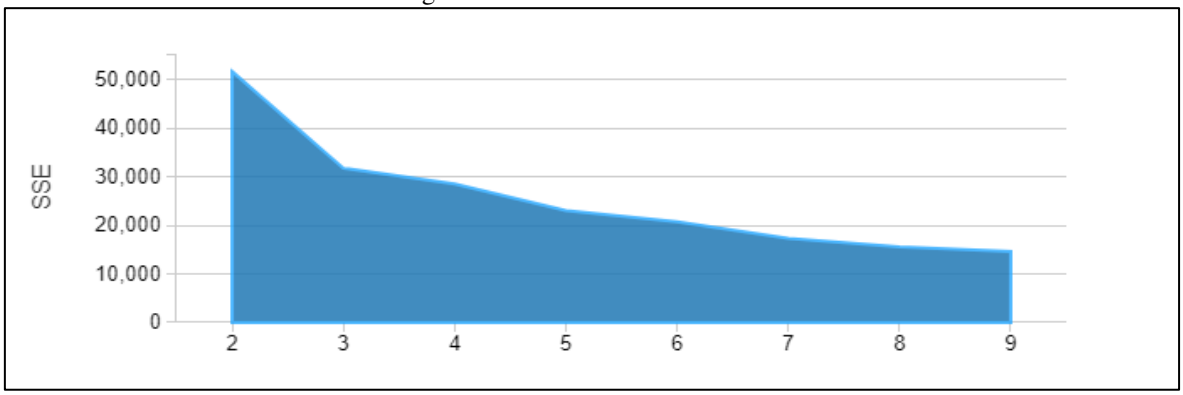
Falta a explicabilidade. O problema atual necessita de um resultado que tem alta interpretabilidade. Isso facilita explicar o resultado e usá-lo para a aplicação na escolha do público alvo. Além disso, um resultado que não se resume em um número e uma probabilidade, mas pode ser entendido pelas escolhas feitas pelo algoritmo, permite encontrar a causa de um problema, e como resolver uma situação.

Antes de definir ativamente rótulos para os dados, foi escolhido usar algoritmos não supervisionados. Databricks oferece diferentes soluções para isso.

4.7.1. Classificação com K-means

A primeira tentativa foi a classificação baseada em *k-means*. Pela simplicidade e velocidade na aplicação essa técnica é muito usada. Essa técnica é um método de agrupamento, cujo objetivo é dividir as observações em k grupos. Para cada um desses grupos, os k centroides representam a média de todos os valores do mesmo grupo. Foi usado um *k-means* evolutivo com valores para *k* de 2 até 9, para encontrar o melhor número de centroides. Para cada um deles é medido o decréscimo da distância até os pontos do agrupamento. Depois de calcular a soma das distâncias da média ao quadrado de cada ponto desses centroides (*sum of squared error*, SSE), foi criado um gráfico com os SSE (Figura 22). Nele, é possível ver como o SSE no início cai drasticamente, para depois, em 3,quase não ter alteração. Esse ponto, visível na figura como joelho, define o melhor valor para *k*.

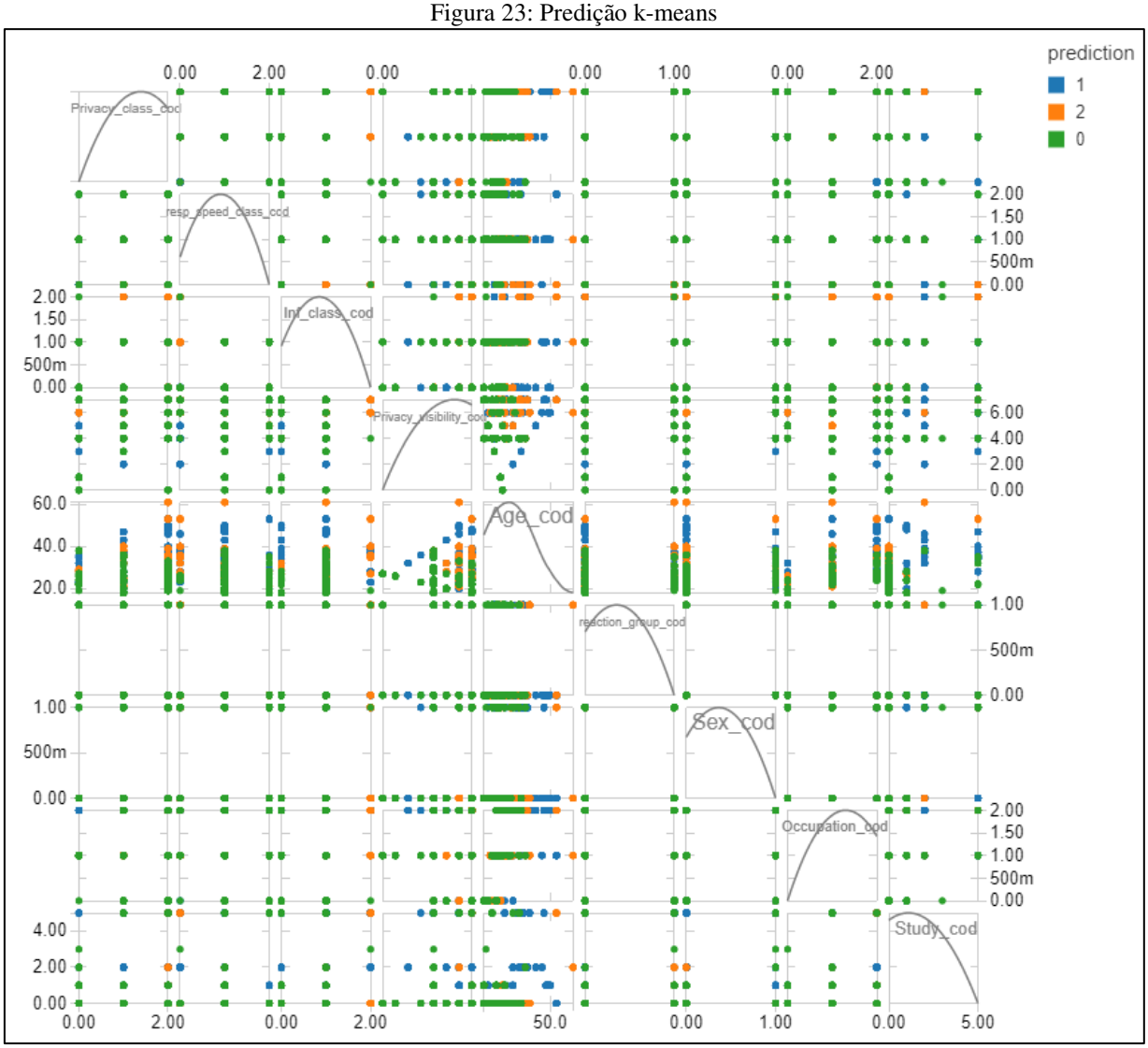
Figura 22: Resultado k-means evolutivo



FONTE: (PRÓPRIA)

O modelo resultante foi aplicado nos dados da pesquisa. O resultado foi plotado em forma de matriz, para ter a possibilidade de comparar visivelmente a distribuição dos valores de todas as colunas (uma parte do resultado é representada na Figura 23).

Figura 23: Predição k-means



FONTE: (PRÓPRIA)

O resultado não foi convincente, porque os agrupamentos representados não conseguiram mostrar correlações interessantes entre as variáveis.

4.8. Aprendizado de Máquina –supervisionado

A primeira tentativa foi com o uso de Databricks. Para simplificar, serão usados dois

algoritmos, um para a frequência e outro para a personalização das mensagens. Para cada problema, serão usados rótulos diferentes.

4.8.1. Árvore de decisão

O algoritmo selecionado será uma árvore de decisão. O algoritmo é aparentemente simples, mas tem alto poder de previsibilidade com um número pequeno de suposições (DEAN, 2014). Além disso, o resultado de uma árvore de decisão tem alta explicabilidade, porque permite entender as causas que levam a um determinado resultado (DAUME, 2012).

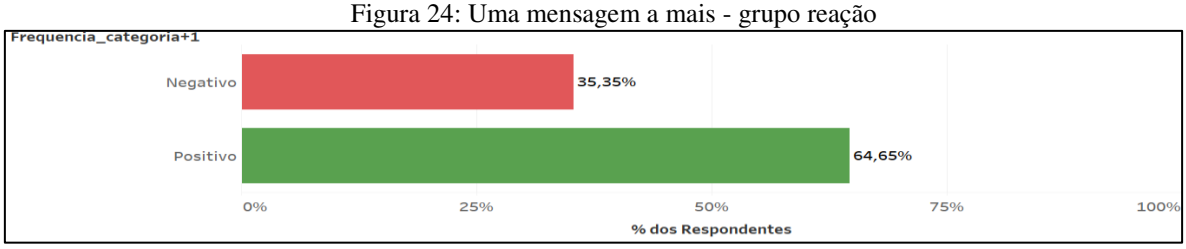
4.8.2. Frequência das mensagens

Como encontrar as pessoas que permitem o envio de mais mensagens do que elas mesmas escolheram no momento de assinar uma lista de e-mail? Lembrando, foram cinco as reações medidas com o questionário:

- Problema nenhum, enquanto o assunto for interessante/importante.
- Acho que nem lembro a frequência escolhida no momento da assinatura
- Recebo tantas mensagens, que uma a mais ou a menos não faz diferença
- Fico chateado
- Vou cancelar a minha assinatura

As primeiras três reações são consistentemente positivas, as últimas duas, negativas. A distribuição nos dados é representada na Figura 24. O objetivo é identificar aquele percentual de 65% das pessoas que suportam mais mensagens.

Figura 24: Uma mensagem a mais - grupo reação



FONTE: (PRÓPRIA)

O rótulo usado para a criação do modelo será exatamente esse. Quem reage negativamente será marcado com 1, quem de maneira positiva, com 0. A seleção das colunas e a representação dos valores usados para o algoritmo é fruto de um trabalho repetitivo. As variáveis usadas, com os valores da codificação das variáveis categóricas, são listadas na Tabela 7.

Tabela 7: Variáveis usadas pela arvore Frequência

Variável	Valores	
Faixa etária	1: 18-25 anos 2: 26-30 anos 3: 31-40 anos	4: 41-50 anos 5: 51-60 anos 6: 60+ anos
Idade	I: Idade em anos	
Gênero	0 masculino, 1 feminino 2 não informado	
Trabalho	0: estudo 1: trabalho e estudo 2: Trabalho	
Frequência escolhida	F: número de mensagens por mês	
Informação: passo	1: escolhido 0:não escolhido	
Informação: somente	1: escolhido 0:não escolhido	
Informação: nunca	1: escolhido 0:não escolhido	
Informação: falsa	1: escolhido 0:não escolhido	
Classificação privacidade	0: fundamentalista 1: realista 2 : despreocupado	
Estudo/Estudei Computação	1: sim 0:não	
Estudo/Estudei Exatas	1: sim 0:não	

FONTE: (PRÓPRIA)

Nas primeiras tentativas, foi usado um tipo de transformação para as variáveis categóricas, conhecido como "*one-hot encoding*". Essa transformação consiste basicamente em criar uma coluna para cada valor que uma variável categórica pode assumir, e atribuir 1, quando o elemento tem aquele valor, ou 0, em caso contrário. Mas, já que os resultados

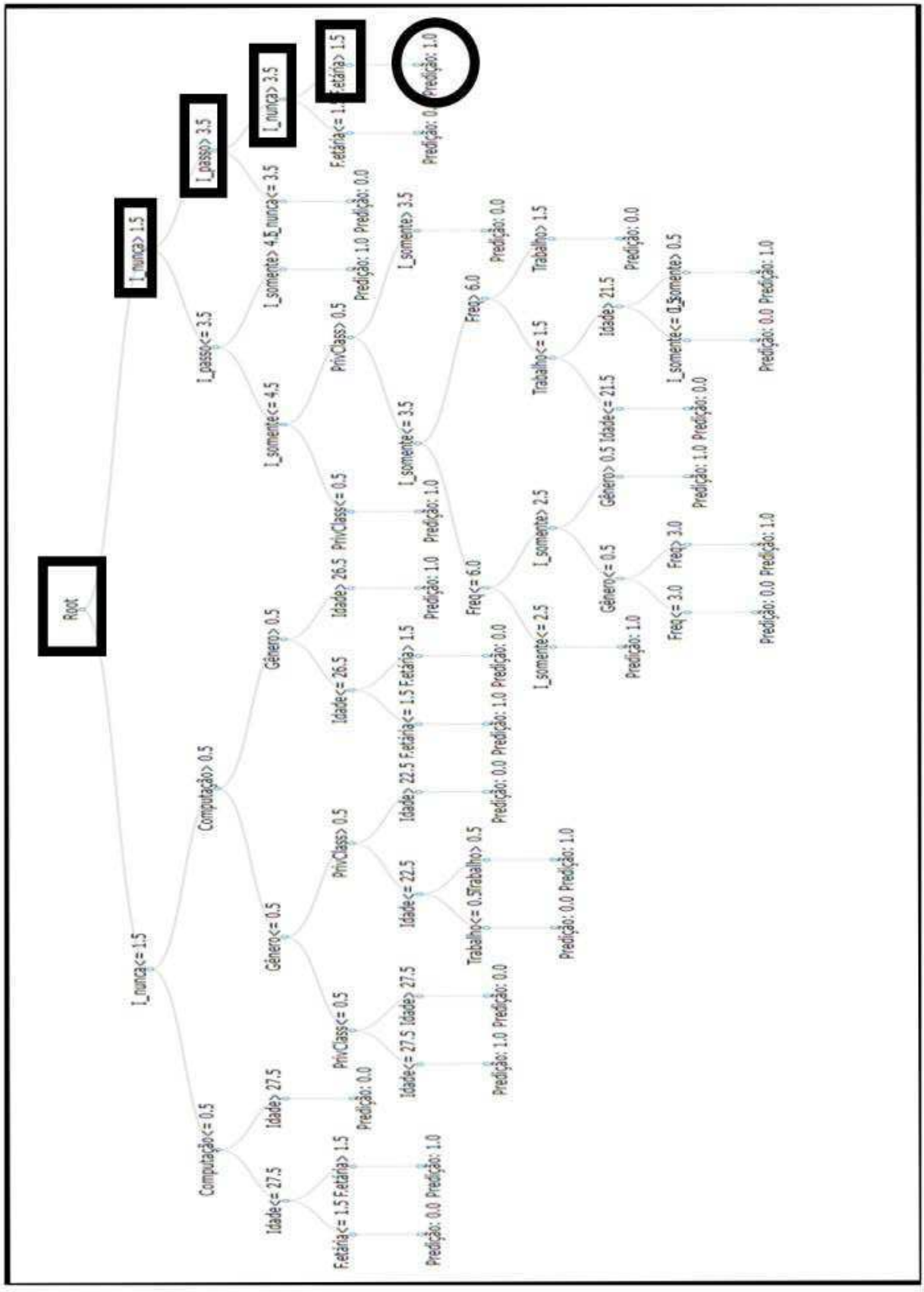
praticamente não mudaram, foi decidido manter a decodificação inicial, que permitiu a melhor legibilidade do modelo resultante. Após iniciar com todas as variáveis disponíveis, e avaliar os resultados da árvore de decisão, a seleção foi sempre mais restrita. Com Tableau, foi possível visualizar o resultado do modelo, agrupando as predições e o rótulo usado para encontrar distribuições gráficas nos dados. Ao final, as colunas usadas para a classificação foram as informações demográficas (idade, gênero, trabalho e estudo), a classificação baseada na preocupação com as informações e a classificação baseada na privacidade. Para avaliar o modelo resultante, é necessário dividir os dados em uma parte usada para treino e a outra para aplicar e testar o modelo. Para a divisão dos dados entre dados de treino e dados de teste foram feitos diferentes testes. O melhor resultado é baseado na divisão de 80% dos dados para treinar o modelo e os restantes 20% para medir a qualidade da predição. A divisão dos dados, usando a função "randomSplit", não foi tão exata como esperada. No final, somente 16 respostas foram usadas para a aplicação do modelo. O resultado foi uma árvore com 49 nós de profundidade 9.

Databricks tem a possibilidade de imprimir uma versão gráfica da árvore. Infelizmente, não tem implementação para imprimir o nome das variáveis usadas no lugar do número dela. Por isso, usamos uma representação alternativa, reproduzida na Figura 25. Nela, é possível enxergar o passo a passo para chegar à predição. Partindo de uma variável raiz, a ramificação depende do valor das variáveis, e termina na predição. Como exemplo, a descrição do caminho mais a direita da árvore, partindo da raiz:

- Nó 1 Escolheu mais de uma vez "Nunca passaria"
- Nó 2 Escolheu mais de 3 vezes "Passo sem problema"
- Nó 3 Escolheu mais de 3 vezes "nunca passaria"
- Nó 4 Faixa etária >= 26-30 anos
- Predição Não aguenta mais mensagens do que ele mesmo escolheu

Resumindo: se o respondente escolheu mais de três vezes "Passo sem problema", e o mesmo número de "Nunca passaria", e tem mais de 26 anos, o modelo indica que seria melhor não aumentar o número de mensagens. Mas, se outra pessoa, tem as mesmas características, mas menos de 26 anos, então ela não deveria reagir de maneira negativa a uma mensagem a mais.

Figura 25: Resultado árvore – Frequência



FONTE: (PRÓPRIA)

As métricas do modelo são resumidas na Tabela 8. Para o rótulo 0, a precisão e o Recall são aceitáveis, mas a predição para o rótulo 1 é praticamente inexistente.

Tabela 8: Métricas modelo frequência

Métricas do modelo		
	Rótulo 0	Rótulo 1
Precisão	83 %	25%
Recall	80%	33%

FONTE: (PRÓPRIA)

A precisão quantifica a porcentagem do acerto, medindo quantas das predições feitas são verdadeiras. Do outro lado, o recall é uma medida que mostra quantos dos acontecimentos foram identificados. Para melhor entender, basta ver a matriz de confusão para as 16 respostas usadas para a avaliação do modelo (Tabela 9). Ela repete as informações resumidas nas métricas de maneira mais intuitiva. Nas linhas, o rótulo, a situação verdadeira, e nas colunas, a predição do modelo. Para o rótulo 0, o modelo acerta 10 observações de 12; para o rótulo 1, das 4 observações, ele erra 3 e acerta somente 1 única predição.

Tabela 9: Matriz Confusão modelo Frequência

Matriz Confusão			
		Predição	
		0	1
Rótulo	0	10	2
	1	3	1

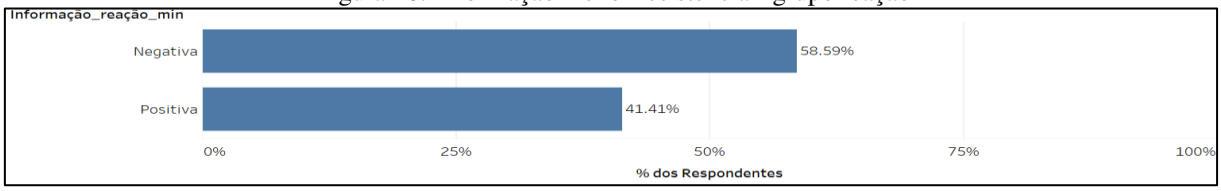
FONTE: (PRÓPRIA)

É importante lembrar, que o resultado é devido, em grande parte, à falta de dados. Com os dados presentes, só 20 observações são usadas para verificar a qualidade do modelo.

4.8.1. Personalização das mensagens

O procedimento para encontrar as pessoas que suportam uma personalização indevida é a mesma para a frequência dos resultados. Depois de definir os rótulos, 59% reagem de maneira negativa, já com o uso da informação com o menor grau de resistência (Figura 26).

Figura 26: Informação menor resistência - grupo reação



FONTE: (PRÓPRIA)

Como no caso da frequência, quem reage negativamente será marcado com 1, já quem reage de maneira positiva com 0. Depois do processo seletivo das colunas, foram usadas as mesmas colunas como no caso da frequência:

- As informações demográficas (idade, gênero, trabalho e estudo)
- A classificação baseada na preocupação com as informações
- E a classificação baseada na privacidade

As métricas do modelo (Tabela 10) mostram que esse modelo é pior do que o da frequência. Com uma precisão de 56%, é somente um pouco melhor que um "cara ou coroa" com uma moeda.

Tabela 10: Métricas modelo personalização

Métricas do modelo		
	Rótulo 0	Rótulo 1
Precisão	87 %	25%
Recall	54%	66%

FONTE: (PRÓPRIA)

A interpretação da matriz de confusão (Tabela 11) mostra a má-qualidade do modelo: nos rótulos 0, o modelo ainda consegue predizer 7 de 8. Mas, para o rótulo 1, o resultado é péssimo, pois somente 2 de 8 predições para esse rótulo estão certas. Por quanto foi avaliado e experimentado, não foi possível melhorar o desempenho dos modelos.

Tabela 11: Matriz Confusão modelo Personalização

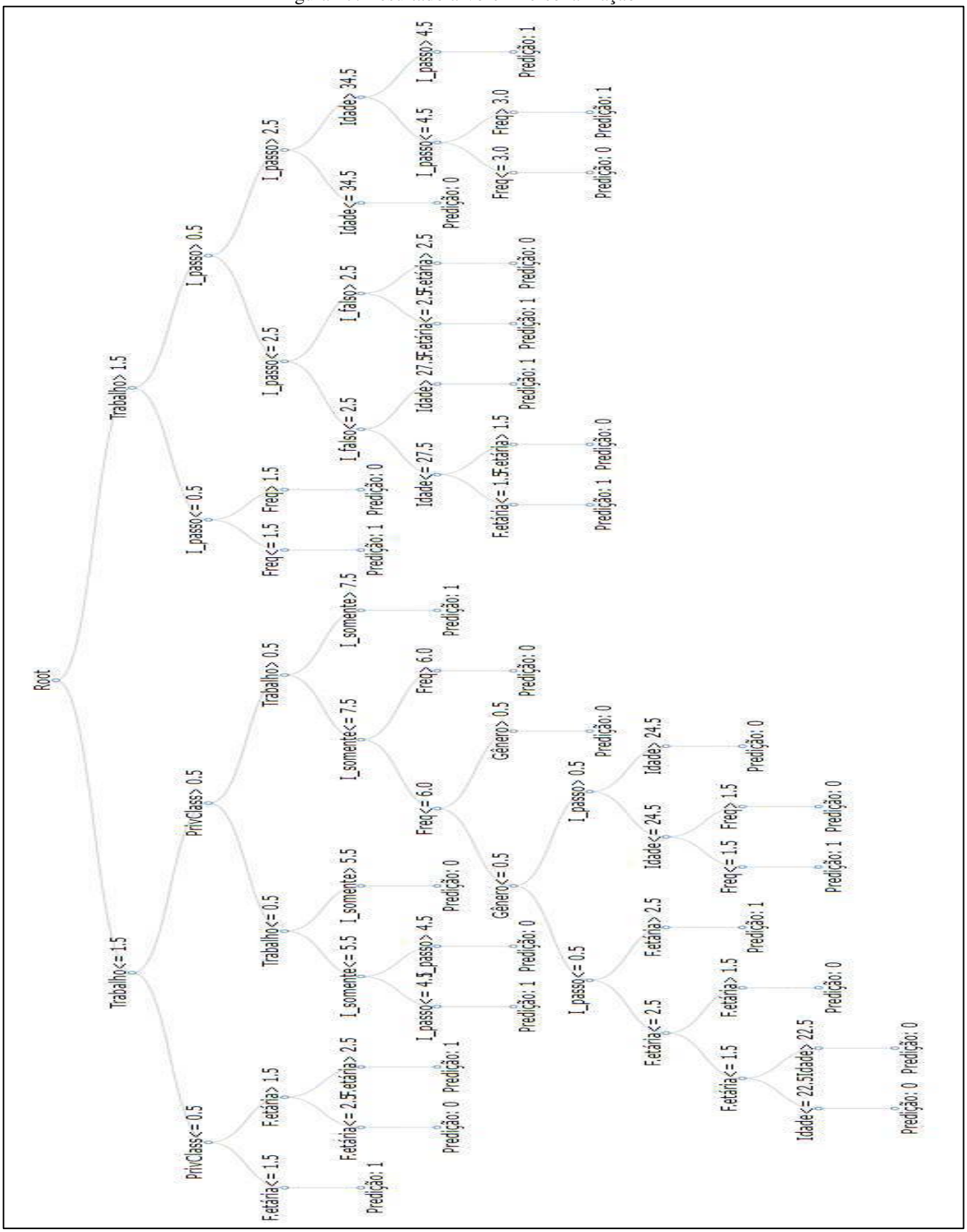
Matriz Confusão			
		Predição	
		0	1
Rótulo	0	7	1
	1	6	2

FONTE: (PRÓPRIA)

O resultado do modelo foi uma árvore com 53 nós e profundidade 10, representada na Figura 27: Resultado árvore - Personalização. Partindo da raiz, o modelo faz a primeira grande divisão pelo trabalho. Quem estuda, cai na ramificação à esquerda, quem trabalha à direita. Quem estuda e trabalha é dividido entre os dois caminhos. O lado dos estudantes em seguida é dividido pela classe de privacidade. Fundamentalistas pra esquerda e são logo marcados com 1, inferindo a reação negativa em caso de uso de personalização indevida. O mesmo vale para os "não fundamentalistas" acima de 30 anos, enquanto os mais jovens aguentam a personalização.

É interessante observar que, para o modelo da frequência, o segundo nível de um dos dois ramos é definido se a pessoa estuda ou estudou computação, enquanto essa informação não é usada no modelo de personalização.

Figura 27: Resultado árvore - Personalização



FONTE: (PRÓPRIA)

4.9. Resultado da aplicação

A aplicação do pipeline de análise de dados com o estudo feito na área de EM foi trabalhoso. As primeiras duas etapas, a contextualização com a área e a definição do problema representa o maior esforço nessa aplicação. A definição do contexto, mesmo pertinente à área de computação, mostra quanto seja importante entender o negócio, os parâmetros usados para medir sucesso e fracasso, e quanto é difícil encontrar valores padrão, valores de referência para cada um deles. É verdade, que esses valores dependem de muitos fatores, e com isso pode ser improvável encontrar o valor perfeito. Mas, parece que informações relativas à área são retidas e tratadas como segredos de negócio. Os valores publicados pelas empresas de EM parecem mostrar somente a qualidade da empresa, e em nenhum caso, foi possível encontrar valores muito dissonantes. A divisão do problema principal nas suas partes, a criação de tarefas diferentes, permitiu entender quais dados são necessários para responder ao problema. Além das informações correlacionadas diretamente com o aumento do retorno de uma campanha de EM, encontramos valores sobre o custo do cliente e o custo de cancelamento de assinaturas, assim como a perda de reputação com o aumento de mensagens marcadas como spam. Essas informações foram usadas para a criação do questionário e a definição dos rótulos. Após a coleta dos dados, já na primeira exploração deles, foi possível verificar que a informação contida neles, com todas as suas limitações, é muito diferente do esperado pela literatura. As diferenças maiores foram encontradas na quantidade de cancelamentos (por exemplo, quando a frequência fica maior do que foi definido no momento da assinatura) e na resistência de passar informações pessoais (as informações definidas como mais íntimas são passadas sem problema, enquanto para informar outras, definidas como menos importantes, foi encontrada uma resistência maior). Conseguimos ver também, que "o brasileiro" cuida mais da privacidade na internet, do que foi esperado. Obviamente, como já mencionado várias vezes, com as limitações dos dados coletados.

A análise exploratória dos dados permitiu a criação de classes baseadas nos dados coletados, usados nas abordagens não supervisionada e supervisionada de Aprendizado de Máquina. Os passos de classificação, redução de variáveis e Aprendizado de Máquina, não supervisionado e supervisionado, resultaram em dois modelos de árvores de decisão, que permitem classificar os usuários que suportam uma frequência maior de mensagens, e/ou uma personalização, mesmo com dados de fontes ambíguas. Foi possível mostrar um caminho de

classificação com um resultado com altíssima interpretabilidade e foi possível criar modelos suficientemente robustos.

Os modelos permitem uma classificação dos usuários, que pode ser usado no momento da criação das campanhas. Isso produz aumento de frequência e personalização para uma parte dos assinantes, com maior possibilidade de aumento do retorno, limitando as perdas de assinantes, e assim os custos necessários para compensar os clientes perdidos.

Alguns resultados secundários surpreenderam, como a resistência de passar informações definidas como “menos pessoais”, ou a facilidade de passar outras definidas como “íntimas”, ou a atitude, em geral, relativa à privacidade na Internet.

Capítulo 5 – Considerações finais

O esforço investido para aprender o contexto foi enorme. Mas, somente a dedicação para aprender esse contexto, permitiu entender o problema, dividi-lo em partes, encontrar as respostas para cada uma delas, e criar e coletar os dados. A definição dos stakeholders, com a destinação dos resultados esperados, ajudou na escolha do algoritmo de Aprendizado de Máquina.

O uso de uma pesquisa completamente bibliográfica, sem cliente real, impossibilitou mostrar a troca de informação necessária entre as duas partes, e os diálogos necessários foram reduzidos a um monólogo de pensamentos do autor. Isso afetou, sobretudo, a parte da investigação, deslocando a atenção da análise de dados recebidos pelo cliente, para a preparação da coleta de dados. Mesmo assim, foi possível notar que a informação coletada nem sempre ajuda no caminho da coleta e da interpretação dos dados. Toda informação recebida deve ter confirmação nos dados.

É nesse ponto que se encaixa o trabalho futuro. A definição do fluxo de informações entre cliente e cientista de dados. Cada inconsistência nos dados, cada passo de tratamento, limpeza, criação ou substituição dever ter a sua fundamentação, aceita pelo cliente e cientista, para assim, criar uma base para a criação de um modelo de Aprendizado de Máquina.

BIBLIOGRAFIA

ACQUISTI, A. and J. Grossklags, “What Can Behavioral Economics Teach Us About Privacy?,” in *ETRICS - International Conference on Emerging Trends in Information and Communication Security*, 2006.

AED, “Introduction to Data Analysis Handbook,” *Migr. Seas. Head Start Tech. Assist. Cent.*, 2009.

Return Path, “2017 Deliverability Benchmark Report - Analysis of Worldwide Inbox Placement Rates.” [Online]. Available: https://digital.returnpath.com/wp-content/uploads/main/2018/02/2017-Deliverability-Benchmark.pdf. [Accessed: 12-February 2018].

ANDERSON, H. R. “The Mythical Right to Obscurity: A Pragmatic Defense of No Privacy in Public*,” *A J. Law Policy Inf. Soc.*, no. Spring, 2012.

“Apache Hadoop.” [Online]. Available: https://hadoop.apache.org/. [Accessed: 23-Jun 2018].

BAKER, S. A. *Skating on Stilts: Why We Aren’t Stopping Tomorrow’s Terrorism*. Stanford, CA: Hoover Institution Press, 2010.

BARNES, S. B. “A Privacy Paradox: Social Networking in the United States,” *Firsmonday.org*, vol. 11, Sep. 2006.

BLANK, G., G. Bolsover, and E. Dubois, “A New Privacy Paradox: Young People and Privacy on Social Network Sites,” San Francisco, 2014.

BOYD, D. and E. Hargittai, “Facebook privacy settings: Who cares?,” *First Monday*, vol. 15, no. 8, 2010.

BOYD, W. “Delivered, Bounced, Blocked, and Deferred Emails Explained | SendGrid,” 2017. [Online]. Available: https://sendgrid.com/blog/delivered-bounced-blocked-and-

deferred-emails-what-does-it-all-mean/. [Accessed: 26-May-2018].

BILENKO, M., M. Richardson, and J. Tsai, "Targeted, Not Tracked: Client-Side Solutions for Privacy-Friendly Behavioral Advertising," in *TPRC*, 2011.

BRÉZILLON, P. "Context in problem solving: A survey," *Knowl. Eng. Rev.*, 1999.

BROWN, M., and R. Muchira, "Investigating the relationship between internet privacy concerns and online purchase behavior," *J. Electron. Commer. Res.*, vol. 5, no. 1, pp. 62–70, 2004.

CAMPAIGN MONITOR, "The 2016 Annual Email Marketing Report," 2017. [Online]. Available:https://www.campaignmonitor.com/company/annual-report/2016/. Accessed: 19-May-2018].

CHAMBERLIN, E. H. *The theory of monopolistic competition; a re-orientation of the theory of value.* Harvard University Press, 1969.

CHAPMAN, P. *et al.*, "Crisp-Dm 1.0," *Cris. Consort.*, 2000.

COHEN, J. E. "What Privacy Is for," *Harv. Law Rev.*, vol. 14, pp. 1–24, 2013.

"COMPARING DATABRICKS TO APACHE SPARK - DATABRICKS." [Online]. Available: https://databricks.com/product/comparing-databricks-to-apache-spark. [Accessed: 20-Jul-2018].

"COMPUTING HISTORY." [Online]. Available: http://www.computinghistory.org.uk/det/6116/First-e-mail-sent-by-Ray-Tomlinson/. [Accessed: 24-Mar-2018].

COUPER, M. P., E. Singer, F. G. Conrad, and R. M. Groves, "Risk of Disclosure, Perceptions of Risk, and Concerns about Privacy and Confidentiality as Factors in Survey Participation.," *J. Off. Stat.*, 2008

"DATABRICKS™." [Online]. Available: https://databricks.com/. [Accessed: 14-Jul-2018].

DAUME, H. *A course in machine learning*. 2012.

DAVENPORT, T. H. "Keep up with your quants," *Harvard Business Review*. 2013

DEAN, J. *Big data, data mining, and machine learning: value creation for business leaders and practitioners*. 2014.

Decreto nº 8771. 2016.

DICKSON, P. R. and J. L. Ginter, "Market Segmentation, Product Differentiation, and Marketing Strategy," *J. Mark.*, vol. 51, no. 2, p. 1, 1987.

"DIGITAL ADVERTISING ALLIANCE." [Online]. Available: https://digitaladvertisingalliance.org/. [Accessed: 02-Jun-2018].

DINEV, T. H. Xu, J. H. Smith, and P. Hart, "Information privacy and correlates: An empirical attempt to bridge and distinguish privacyrelated concepts," *Eur. J. Inf. Syst.*, 2013.

DOMINGOS, P. "A few useful things to know about machine learning," *Commun. ACM*, 2012.

DUNCKER, K. "On problem solving," *Psychol. Monogr.*, 1945.

DRUCKER, P. F. "Management: Tasks, Responsibilities, Practices," *Manag. tasks Responsib. Pract.*, p. 576, 1973.

CAPEM, "Código de Autorregulamentação para a prática de E-Mail Marketing," 2010.

"ELGA: Wissenswertes zu ELGA." [Online]. Available: https://elga.gv.at/faq/wissenswertes-zu-elga/index.html. [Accessed: 06-May-2018].

"ELGA- Forschungsdaten-Debatte," 2018. [Online]. Available: http://www.vienna.at/elga-daten-5-000-abmeldungen-nach-forschungsdaten-debatte/5774582. [Accessed: 06-May-2018].

HOSTPAPA, "EMAIL MARKETING KNOCKS OUT SOCIAL MEDIA IN 5 ROUNDS.", 2012. [Online]. Available: https://www.hostpapa.ca/email-vs-social/. [Accessed: 08-Apr-2018].

"EMAIL MARKETING BENCHMARKS." [Online]. Available: https://mailchimp.com/resources/research/email-marketing-benchmarks/. [Accessed: 20-May-2018].

EUROPEAN COMMISION, "2018 reform of EU data protection rules," 2018. [Online]. Available: https://ec.europa.eu/commission/priorities/justice-and-fundamental-rights/data-protection/2018-reform-eu-data-protection-rules_en. [Accessed: 01-Jun-2018].

EUROPEAN INTERACTIVE DIGITAL ADVERTISING ALLIANCE, "Your Online Choices," 2018. [Online]. Available: http://www.youronlinechoices.com/pt/. [Accessed: 02-Jun-2018].

EUROPEAN INTERACTIVE DIGITAL ADVERTISING ALLIANCE, "Your Online Choices - Perguntas frequentes," 2018. [Online]. Available: http://www.youronlinechoices.com/pt/faqs. [Accessed: 03-Jun-2018].

FALK, J. "Creation and Use of Email Feedback Reports: An Applicability Statement for the Abuse Reporting Format (ARF)," Jun. 2012.

FAYYAD, U. G. Piatetsky-Shapiro, and P. Smyth, "FROM DATA MINING TO KNOWLEDGE DISCOVERY IN DATABASES," *AI Mag.*, 1996.

FEDERAL TRADE COMMISSION, "Self-Regulatory Principles For Online Behavioral Advertising," 2009.

FINK, K. "Is It Just Me, or Does This Data Smell Funny?," in *Bad Data Handbook*, M. Loukides and M. Blanchette, Eds. Sebastopol, CA: O'Reilly Media, 2012, p. 264.

FORRESTER CONSULTING, "Inspire Customers With Emotionally Engaging Content. Contextual Relevance Sparks An Emotional Connection," 2016.

GANDY, O. H. J. "The Panoptic Sort: A Political Economy of Personal Information (Critical Studies in Communication and in the Cultural Industries)," Westview Press, Boulder, CO, 1993.

GANHÃO, M. "Índia e Brasil preocupados," *Expresso*, 2018. [Online]. Available: http://expresso.sapo.pt/internacional/2018-03-22-India-e-Brasil-preocupados-com-interferencias-da-Cambridge-Analytica-nas-suas-proximas-eleicoes. [Accessed: 25-Mar-2018].

GLEASON, K. and Q. E. McCallum, "Data Quality Analysis Demystified: Knowing When Your Data Is Good Enough," in *Bad Data Handbook*, M. Loukides and M. Blanchette, Eds. Sebastopol, CA: O'Reilly Media, 2012, p. 264.

GILHOOLY, K. J. "Human and machine problem solving," *New York, NY, US Plenum Press. (1989). xvi, 382 pp.*, 1989.

GLYNN, F. "It Takes 6 to 8 Touches to Generate a Viable Sales Lead. Here's Why - Salesforce Blog," 2015. [Online]. Available: https://www.salesforce.com/blog/2015/04/takes-6-8-touches-generate-viable-sales-lead-heres-why-gp.html. [Accessed: 30-Apr-2018].

Google opt-out." [Online]. Available: https://support.google.com/ads/answer/2662922?hl=pt. [Accessed: 03-Jun-2018].

GROVES, E. *The Constant Contact Guide to mail Marketing*. New Jersey: John Wiley & Sons, Inc., 2009.

HUBSPOT, “Inbound marketing analytics,” 2017.

“INFORMAÇÃO LEGAL DO WHATSAPP.” [Online]. Available: https://www.whatsapp.com/legal/?eea=0#terms-of-service. [Accessed: 02-Jun-2018].

IYENGAR, S. S. and M. R. Lepper, “When choice is demotivating: Can one desire too much of a good thing?,” *J. Pers. Soc. Psychol.*, vol. 79, no. 6, pp. 995–1006, 2000.

JAI, T.-M. (Catherine) and N. J. King, “Privacy versus reward: Do loyalty programs increase consumers’ willingness to share personal information with third-party advertisers and data brokers?,” *J. Retail. Consum. Serv.*, vol. 28, pp. 296–303, Jan. 2016.

KOKOLAKIS, S. “Privacy attitudes and privacy behaviour: A review of current research on the privacy paradox phenomenon,” *Computers and Security*, vol. 64. pp. 122–134, 2017.

KOTLER, P. *Marketing Management , Millenium Edition*, 10th ed. Boston: Pearson, 2002.

KOTLER, P. “Ten Deadly Marketing Sins: Signs And Solutions,” *Wiley*, pp. 1–12, 2004

KOTLER, P. and G. Armstrong, “Principles of marketing, 15th edition,” *J. Chem. Inf. Model.*, vol. 53, pp. 1689–1699, 2012

KOTLER, P. and G. Kotler, Milton, *Market - Your Way To Growth*. Hoboken, New Jeresy: John Wiley & Sons, Inc., 2013.

KOTLER, P. and G. Armstrong, *Principles of Marketing*, 15th ed. Cloth: Pearson, 2014.

“LAMAPOLL™,” 2018. [Online]. Available: https://www.lamapoll.de/. [Accessed: 14-Jul-2018].

LEGGATT, H. “BlueHornet: 80% delete emails not optimized for mobile - Email Marketing - BizReport,” 2013. [Online]. Available: http://www.bizreport.com/2013/03/bluehornet-302-

unsubscribe-on-receipt-of-a-badly-formatted-m.html. [Accessed: 27-May-2018].

"LEI DO MARCO CIVIL DA INTERNET NO BRASIL." [Online]. Available: http://www.cgi.br/lei-do-marco-civil-da-internet-no-brasil/. [Accessed: 03-Jun-2018].
L. Fonseca, "Email Marketing Trends 2017: Panorama do Email Marketing no Brasil," 2017. [Online]. Available: https://inteligencia.rockcontent.com/relatorios/email-marketing-trends-2017/. [Accessed: 27-May-2018].

LEI Nº 12.965, DE 23 DE ABRIL DE 2014. .

LESZCZYNSKI, M. "Email Marketing Benchmarks 2017," 2017. [Online]. Available: https://www.getresponse.com/resources/reports/email-marketing-benchmarks.html. [Accessed: 28-May-2018].

LEWKOWICZ, K. "Email Delivery vs. Deliverability: What's the Difference?" [Online]. Available: https://blog.hubspot.com/marketing/email-delivery-deliverability. [Accessed: 26-May-2018].

LI, H. "Which machine learning algorithm should I use? - Subconscious Musings," 2017. [Online]. Available: https://blogs.sas.com/content/subconsciousmusings/2017/04/12/machine-learning-algorithm-use/. [Accessed: 30-Jul-2018].

MADDEN, M., A. Lenhart, and S. Cortesi, "Teens, social media, and privacy," *Pew Internet* ..., p. 107, 2013.

MAHAJAN, N. "The Thinker Interview with Philip Kotler, the Father of Marketing," *October, 8*, 2013. [Online]. Available: http://knowledge.ckgsb.edu.cn/2013/10/08/marketing/philip-kotler-interview-four-ps-marketing/. [Accessed: 27-Jan-2017].

MARKETINGSHERPA. MarketingSherpa, Inc, 2015.

MOHANTY, S., M. Jagadeesh, and H. Srivatsa, *Big data imperatives: Enterprise big data*

warehouse, BI implementations and analytics. 2013.

MOSCARDELLI, D. M. "Teens Surfing The Net: How Do They Learn To Protect Their Privacy?," *J. Bus. {&} Econ. Res.*, vol. 2, no. 9, pp. 43–56, 2004.

NISSENBAUM, H. "A Contextual Approach to Privacy Online," *Deadalus 140*, vol. 4, no. Fall 2011, pp. 32–48, 2011.

NOAH, Harari Yuval *Homo deus: a brief history of tomorrow. Vintage Digital*. New York: HarperCollins Publishers, 2017.

OBAR, J. A. and A. Oeldorf-Hirsch, "The Biggest Lie on the Internet: Ignoring the Privacy Policies and Terms of Service Policies of Social Networking Services," *SSRN Electron. J.*, Aug. 2016.

PENG, R.. "Context Compatibility in Data Analysis · Simply Statistics," 2018. [Online]. Available: https://simplystatistics.org/2018/05/24/context-compatibility-in-data-analysis/. [Accessed: 25-Jul-2018].

PHRASEE, "Open rate: The most important email marketing metric ever," 2017. [Online]. Available: https://phrasee.co/open-rate-important-email-marketing-metric-ever/. [Accessed: 26-May-2018].

"REAÇÕES ESCÂNDALO DE DADOS." [Online]. Available: https://tecnologia.uol.com.br/noticias/redacao/2018/03/23/facebook-cambridge-escandalo-reacoes-empresas-google-uniao-europeia.htm. [Accessed: 25-Mar-2018].

READ, B. "Think before You Share.," *Chron. High. Educ.*, vol. 52, no. 20, 2006.

RECORDS, G. W. "Oldest electronic spam," 2018. [Online]. Available: http://www.guinnessworldrecords.com/world-records/oldest-electronic-spam. [Accessed: 06-May-2018].

RIJN, J. van, “National Client Email Report 2015,” *Direct Mark. Assoc.*, 2016.
SAS Institute Inc, *Data Mining Using SAS® Enterprise MinerTM: A Case Study Approach*, 4th ed. Cary, NC: SAS Institute Inc, 2018.

ROBERTSON, S. I. *Problem solving*. 2001.

ROBOTS IN ASSISTED LIVING ENVIRONMENTS, “Actual and perceived privacy considerations and ethical requirements I,” 2015.

ROHANIZADEH, S. S. and M. B. Moghadam, “A Proposed Data Mining Methodology and its Application to Industrial Procedures,” *J. Ind. Eng.*, 2009.

ROSENBERG, M. N. Confessore, and C. Cadwalladr, “How Trump Consultants Exploited the Facebook Data of Millions - The New York Times,” 2018. [Online]. Available: https://www.nytimes.com/2018/03/17/us/politics/cambridge-analytica-trump-campaign.html. [Accessed: 25-Mar-2018].

ROSS, N. “A History of Direct Marketing,” Direct Marketing Asociation, New York, 1992.

ROSS,W.D. “Aristotle, Rhetoric,” 1959. [Online]. Available: http://www.perseus.tufts.edu/hopper/text?doc=Perseus%3Atext%3A1999.01.0060%3Abook%3D1%3Achapter%3D1%3Asection%3D1. [Accessed: 30-Apr-2018].

ROSS, C. “Quantity or Quality Which is Better with Email? - SendPulse,” 2017. [Online]. Available: https://sendpulse.com/blog/quantity-or-quality-which-is-better-with-email. [Accessed: 30-Apr-2018].

SAMUEL, A. L. “Some Studies in Machine Learning Using the Game of Checkers,” *IBMJ. Res. Dev.*, 1959.

SCHWARTZ, B. “The Paradox of Choice,” in *Positive Psychology in Practice: Promoting Human Flourishing in Work, Health, Education, and Everyday Life: Second Edition*, 2015, pp. 121–138.

SDL, “The Global CX Wakeup Call,” 2017. [Online]. Available: https://www.sdl.com/download/research-the-global-cx-wakeup-call/82425/. [Accessed: 16-Jun-2018].

SELTMAN, H. “Experimental Design and Analysis,” *Online B.*, 2018.

SFGATE, “Facebook Lawsuits Over Data Harvesting,” 2018. [Online]. Available: https://www.sfgate.com/news/bayarea/article/Facebook-Users-And-Shareholders-File-Four-12775748.php. [Accessed: 25-Mar-2018].

GARFINKEL, S. „*Database nation: the death of privacy in the 21st century*“, vol. 2000, no. February. Sebastopol, CA: O’Reilly Media, 2000.

SHOULDIS, G. “What Does The Open Rate Mean In Email Marketing?,” 2013. [Online]. Available: http://3bugmedia.com/does-open-rate-mean-email-marketing/. [Accessed: 19-May-2018].

SPRING, M. L. *Machine Learning in Action*. 2015.

SOLOVE, D. J. “The End of Privacy?,” *Sci. Am.*, vol. 299, no. 3, pp. 100–106, Sep. 2008.

STANFORD UNIVERSITY., *Stanford encyclopedia of philosophy.* Stanford University, 1997.

STATISTA, “Facebook users by country,” 2018. [Online]. Available: https://www.statista.com/statistics/268136/top-15-countries-based-on-number-of-facebook-users/. [Accessed: 25-Mar-2018].

STERNE, J. *Web Metrics-Proven Methods for Measuring Web Site Success*. New York, New York, USA: Robert Ipsen, 2002

STIGLITZ, K. “Email Marketing Stats You Need to Know,” 2017. [Online]. Available:

https://www.campaignmonitor.com/blog/email-marketing/2016/01/70-email-marketing-stats-you-need-to-know/. [Accessed: 01-Jun-2018].

SUTANTO, J., E. Palme, C.-H. Tan, and C. W. Phang, "Addressing the Personalization-Privacy Paradox: An Empirical Assessment From a Field Experiment on Smartphone Users," in *MIS Quarterly*, 2013, vol. 37, no. 4, pp. 1141–1164.

"TAXA DE REJEIÇÃO - Ajuda do Google Analytics." [Online]. Available: https://support.google.com/analytics/answer/1009409?hl=pt-BR. [Accessed: 26-May-2018].

"THE CANADIAN ADCHOICES PROGRAM." [Online]. Available: https://youradchoices.ca/. [Accessed: 02-Jun-2018].

THE RADICATI GROUP, "Email Statistics Report, 2018-2022 - Executive Summary," *Email Stat. Rep.*, no. 0, p. 3, 2018.

THE WHITE HOUSE, *Consumer Data Privacy in a Networked World: A Framework for Protecting Privacy and Promoting Innovation in the Global Digital Economy*. 2013.

TOURANGEAU, R. and T. Yan, "Sensitive Questions in Surveys," *Psychol. Bull.*, 2007.

TUKEY, J. W. *Exploratory Data Analysis*. 1977.

TURNEY, P. D. "The Management of Context-Sensitive Features: A Review of Strategies," Dec. 2002.

TUROW, J. , J. King, C. J. Hoofnagle, A. Bleakley, and M. Hennessy, "Americans Reject Tailored Advertising and Three Activities that Enable It," *SSRN Electron. J.*, 2009.

VERTICALRESPONSE, "Global email anti-spam laws," 2017. [Online]. Available: https://www.verticalresponse.com/blog/email-anti-spam-laws-around-the-world-infographic/. [Accessed: 27-May-2018].

WAINWRIGHT, C. "9 Ways to Dramatically Reduce Email Unsubscribe Rates," 2017. [Online]. Available: https://blog.hubspot.com/blog/tabid/6307/bid/31034/9-ways-to-dramatically-reduce-email-unsubscribe-rates.aspx. [Accessed: 27-May-2018].

WARREN, S. D. and L. D. Brandeis, "The Right to Privacy," *Harv. Law Rev.*, vol. 4, no. 5, pp. 193–220, 1890.

WATSON, Z. "What Consumers Want From Marketing Emails," 2015. [Online]. Available: https://technologyadvice.com/blog/marketing/marketing-email-preferences-2015/. [Accessed: 15-Apr-2018].

WEINTRAUB, J. "The theory and politics of the public/private distinction," in *Public and private in thought and practice*, J. Weintraub and K. Kumar, Eds. Chicago, IL and London, England: The University of Chicago Press, 1997, pp. 1–42.

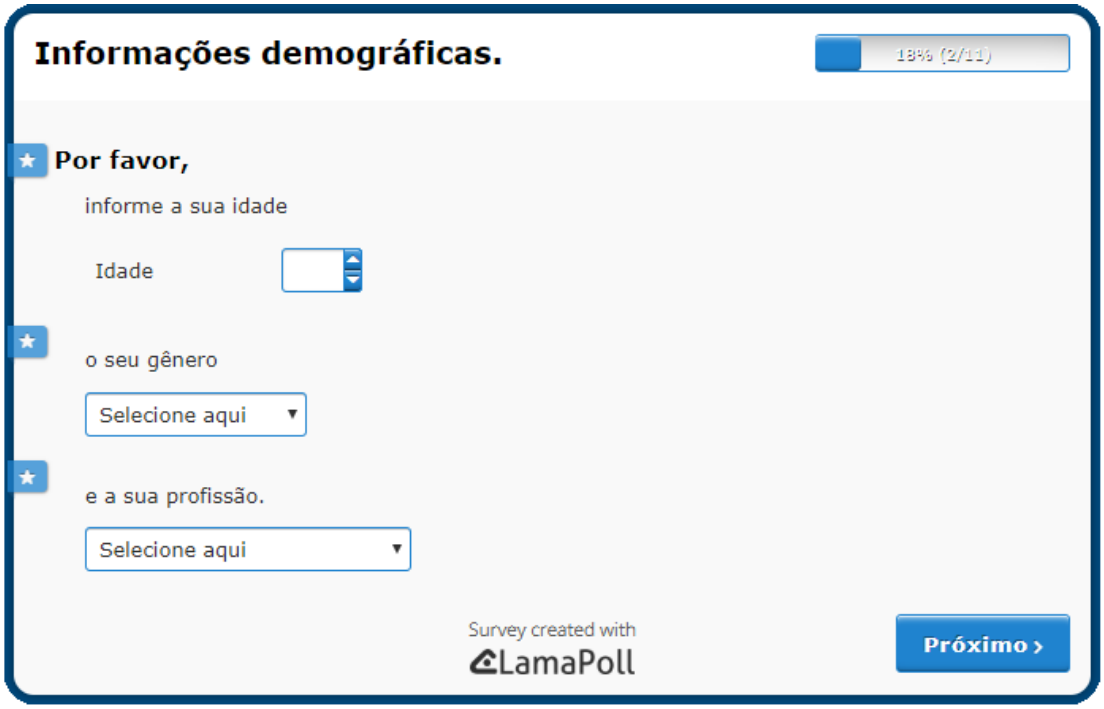
WESTIN, A. F. Privacy and Freedom. New York: Atheneum Press, 1967.

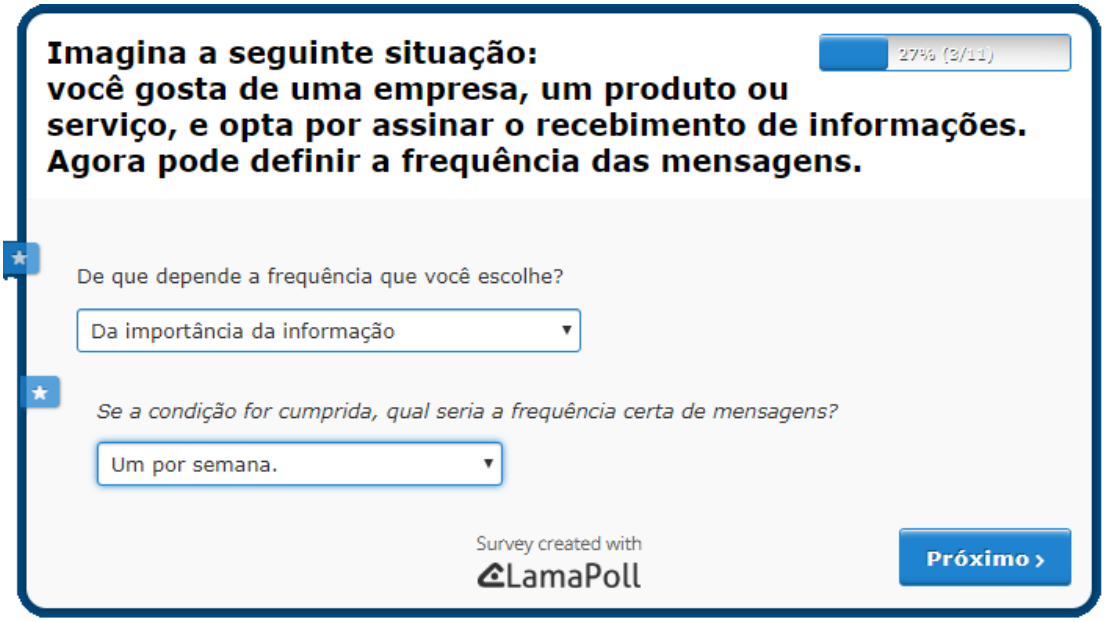
WHITE, C. S. "Millennials' Email Marketing Dislikes—Litmus Software, Inc.," 2016. [Online]. Available: https://litmus.com/blog/millennials-email-marketing-dislikes-are-mostly-the-same-as-everyone-elses. [Accessed: 27-May-2018].

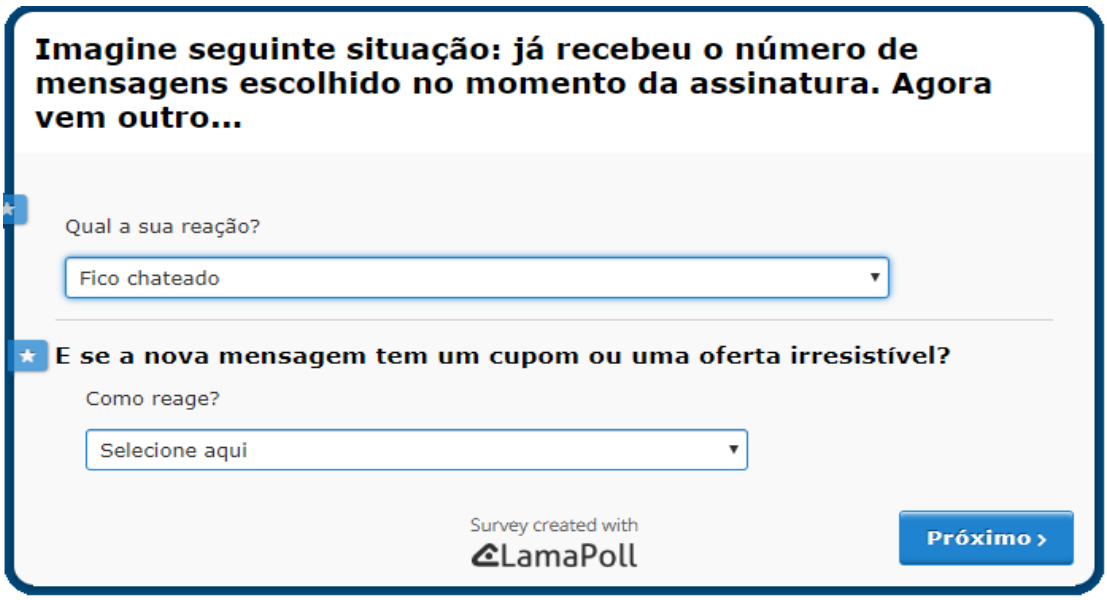
YAU, N. *Data Points. Visualization That Means Something*. Indianapolis: John Wiley & Sons, Inc., 2013.

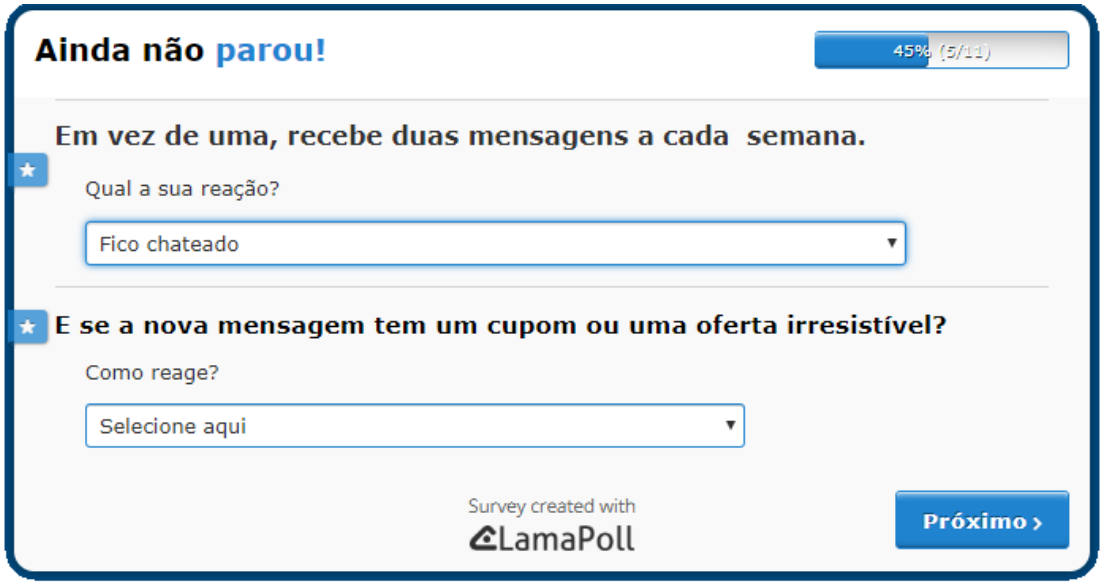
APÊNDICE

Telas do questionário online

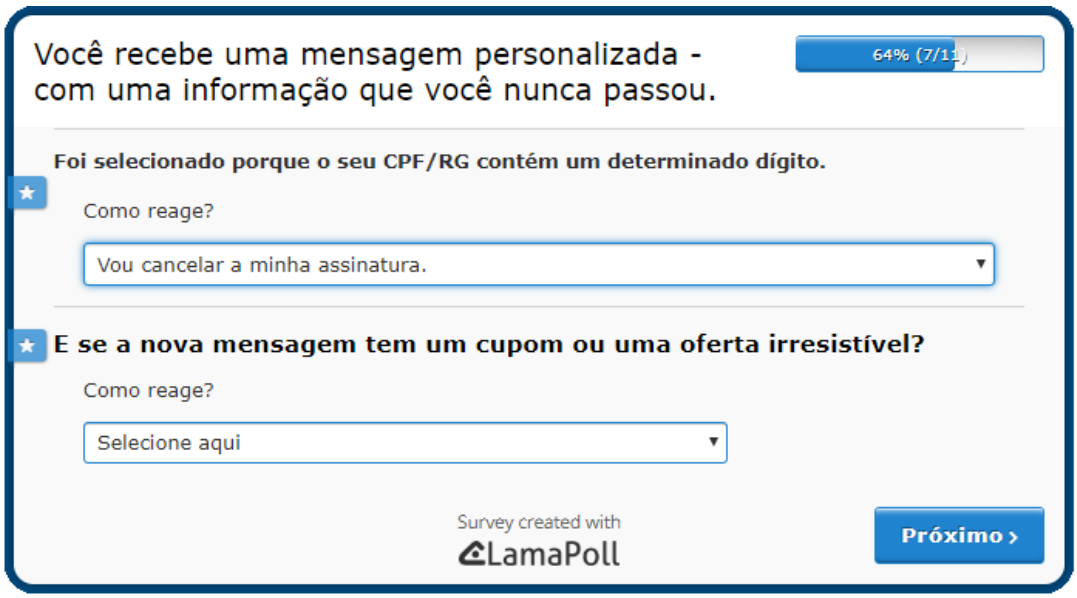

Informações demográficas.
18% (2/11)
Por favor,
informe a sua idade
Idade
o seu gênero
Selecione aqui
e a sua profissão.
Selecione aqui
Survey created with
LamaPoll
Próximo ›

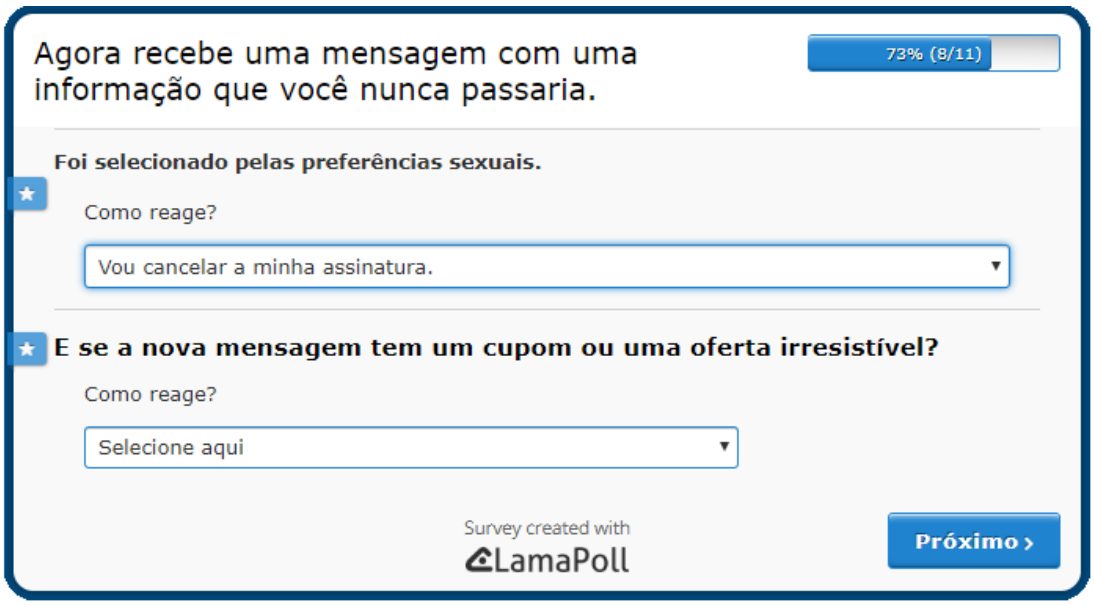

Imagina a seguinte situação:
você gosta de uma empresa, um produto ou serviço, e opta por assinar o recebimento de informações. Agora pode definir a frequência das mensagens.
27% (3/11)
De que depende a frequência que você escolhe?
Da importância da informação
Se a condição for cumprida, qual seria a frequência certa de mensagens?
Um por semana.
Survey created with
LamaPoll
Próximo ›

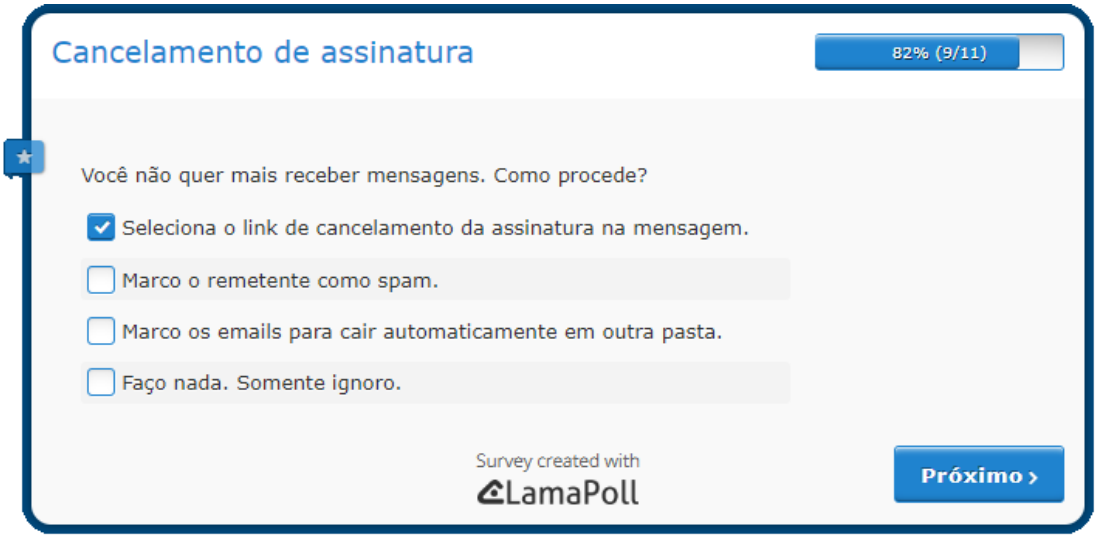

Imagine seguinte situação: já recebeu o número de mensagens escolhido no momento da assinatura. Agora vem outro...
Qual a sua reação?
Fico chateado
E se a nova mensagem tem um cupom ou uma oferta irresistível?
Como reage?
Selecione aqui
Survey created with
LamaPoll
Próximo ›


Ainda não parou!
45% (5/11)
Em vez de uma, recebe duas mensagens a cada semana.
Qual a sua reação?
Fico chateado
E se a nova mensagem tem um cupom ou uma oferta irresistível?
Como reage?
Selecione aqui
Survey created with
LamaPoll
Próximo ›


No momento da assinatura, é necessário informar somente o e-mail. Tem a possibilidade de aumentar a personalização.
55% (6/11)
Quais informações você passaria na assinatura de uma lista de e-mail?
Por favor, selecione a reação para todas as linhas.
Passo sem problema
Somente se for realmente necessário
Nunca passaria
Se for obrigatório, colocaria uma informação errada
Nome/Sobrenome
Data nascimento
Endereço
Telefone
Passatempo preferido
CPF/RG
Grupo sanguíneo
Doença(s)
Preferência sexual
Survey created with LamaPoll
Próximo ›


Você recebe uma mensagem personalizada - com uma informação que você nunca passou.
64% (7/11)
Foi selecionado porque o seu CPF/RG contém um determinado dígito.
Como reage?
Vou cancelar a minha assinatura.
E se a nova mensagem tem um cupom ou uma oferta irresistível?
Como reage?
Selecione aqui
Survey created with LamaPoll
Próximo ›


Agora recebe uma mensagem com uma informação que você nunca passaria.
73% (8/11)
Foi selecionado pelas preferências sexuais.
Como reage?
Vou cancelar a minha assinatura.
E se a nova mensagem tem um cupom ou uma oferta irresistível?
Como reage?
Selecione aqui
Survey created with
LamaPoll
Próximo ›


Cancelamento de assinatura
82% (9/11)
Você não quer mais receber mensagens. Como procede?
Seleciona o link de cancelamento da assinatura na mensagem.
Marco o remetente como spam.
Marco os emails para cair automaticamente em outra pasta.
Faço nada. Somente ignoro.
Survey created with
LamaPoll
Próximo ›


Privacidade
Últimas duas perguntas!
91% (10/11)
Você cuida da privacidade na Internet?
Não.
Não existe privacidade na Internet.
Não tenho nada pra esconder.
Depende da informação que estou publicando.
Depende do site que estou acessando.
Sim.
Qual o seu comportamento?
Por favor, selecione a reação para todas as linhas.
Sempre
Na maioria das vezes
Mais a mais
Ou vou em quando
Nunca
Não sei
Não se aplica
Limito a visibilidade das minhas postagens
Uso aplicativos do facebook que acessam os meus contatos
Crio outras contas com Login do Facebook
Aceito convites de estrangeiros nas redes sociais
Android: controlo as permissões de um aplicativo antes da instalação
iPhone: limito as permissões de um aplicativo
Uso sites que precisam de cookies
Uso a possibilidade de opt-out para cookies
Gerencio os cookies no meu dispositivo
Uso motores de busca alternativos
Uso proxies pra navegar
Uso redes Wi-Fi abertas
Cubro a câmera do meu dispositivo
Survey created with
LamaPoll
Terminar ›


100% (11/11)
ORDEM E PROGRESSO
Muito obrigado pela participação!
Survey created with
LamaPoll