DAS Departamento de Automação e Sistemas
CTC Centro Tecnológico
UFSC Universidade Federal de Santa Catarina

Inferência Bayesiana para apoio à tomada de decisões acerca da geração de *Whitelists* e *Blacklists* em Mídia Programática

Relatório submetido à Universidade Federal de Santa Catarina

como requisito para a aprovação da disciplina:

DAS 5511: Projeto de Fim de Curso

Andjara Pedrero Consentino

Florianópolis, Março de 2017

Inferência Bayesiana para apoio à tomada de decisões acerca da geração de *Whitelists* e *Blacklists* em Mídia Programática

Andjara Pedrero Consentino

Esta monografia foi julgada no contexto da disciplina

DAS 5511: Projeto de Fim de Curso

e aprovada na sua forma final pelo

Curso de Engenharia de Controle e Automação

Prof. Dr. Eduardo Camponogara

Andjara Pedrero Consentino

Título: Inferência Bayesiana para Apoio à Tomada de Decisão Acerca da geração de *Whitelists* e *Blacklists* em Mídia Programática

Esta monografia foi julgada no contexto da disciplina DAS5511: Projeto de Fim de Curso e aprovada na sua forma final pelo Curso de Engenharia de Controle e Automação.

Florianópolis, 30 de março de 2017

Banca Examinadora:

<Artur Pereira/Publya>
Orientador na Empresa
Nome da Empresa

Prof. <Eduardo Camponogara>
Orientador no Curso
Universidade Federal de Santa Catarina

Prof. <Marcelo de Lellis Costa de Oliveira>
Avaliador
Universidade Federal de Santa Catarina

<Maísa Beraldo Bandeira>
Debatedor
Universidade Federal de Santa Catarina

<Murilo Ramos Carraro>
Debatedor
Universidade Federal de Santa Catarina

Agradecimentos

Agradeço primeiramente a Publya pela oportunidade fornecida para a realização deste trabalho mas principalmente por todo o conhecimento compartilhado. Em especial ao Artur Pereira e Luiz Kozma por terem acredito e confiado no meu potencial sempre.

Não menos importante, gostaria de agradecer a minha família por todo suporte prestado durante os anos de faculdade e por ter feito de mim a pessoa que sou hoje. Minha mãe por todo o amor e conforto que só ela soube dar, meu irmão por todo o carinho, meu pai, que onde estiver está torcendo por mim mas principalmente meu padrasto, o qual com muito orgulho sempre pude chamar de pai, esse diploma é seu!

Ao meu namorado Rafael, por ter me incentivado nos momentos mais difíceis, não me deixando esquecer dos meus objetivos nunca.

Meus profundos agradecimentos ao Professor Eduardo Camponogara por toda a orientação e por ter tirado tempo para me auxiliar na realização deste trabalho.

Agradeço também a todos os meus amigos e colegas que fizeram das longas noites de estudos, momentos agradáveis e de intensa troca de conhecimento.

Resumo

O presente trabalho foi desenvolvido na empresa Publya, pioneira no mercado de mídia programática no Brasil e que já veiculou mais de 1000 campanhas ao longo de seus 3 anos de funcionamento. Dentre os vários tipos de mídia que a empresa oferece, focou-se nas mídias *display* e no seu processo de geração de *Whitelists* e *Blacklists* para a configuração de campanhas de anunciantes. Realizou-se então a criação de um sistema para aprendizado de máquina de fácil utilização para todos da empresa que, além de gerar automaticamente uma lista de sites, fornece também uma base sólida para a escolha de sites nos quais se realiza a compra de espaço publicitário. Utilizou-se inferência de Redes Bayesianas acerca das probabilidades condicionais de sucesso e insucesso de cada site, a partir de dados de campanhas passadas.

Palavras-chave: Mídia Programática, Redes Bayesianas, Inferência Bayesiana, Raciocínio Probabilístico, Aprendizagem de máquina, Hipótese de máximo posteriori.

Abstract

This project was developed at Publya, pioneer in the Brazilian programatic advertising industry, running over 1000 campaigns throughout the years. Among the several types of advertising managed by Publya, the project focused specially on banners display advertising and its process of generating whitelists and blacklists to an advertiser campaign setting. It was created an user friendly learning system that besides automatically generating the websites lists is also a rock solid tool for the decision making towards where to buy advertising space. For that purpose it was used Bayesian Inference to calculate the conditional probabilities of success and not success of each website, according to past campaigns data.

Keywords: Programatic Advertising, Bayesian Networks, Bayesian Inference, Probabilistic Reasoning, Machine Learning, Maximum a Posteriori Estimation.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

Lista de Siglas

RTB - Real Time Bidding

SSP - Suply-side Platform

DSP - Demand-side Platform

CPM - Custo por Mil Impressões

CTR - Click Trough Rate

CPA - Custo por Ação

CPC - Custo por Clique

MAP - Hipótese de máximo a posteriori

TPC - Tabela de Probabilidade Condicional

FP - Falso Positivo

FN - Falso Negativo

PV - Positivo verdadeiro

Sumário

1 Introdução

A Publya, empresa que atua no mercado de mídia programática há mais de 3 anos, já veiculou mais de 1000 campanhas ao longo desse período. Os dados gerados nessas campanhas porém, não são utilizados para melhoria em futuras campanhas, devido à impossibilidade de serem analisados constantemente por algum membro da empresa.

Esse grande número de campanhas já veiculadas deixam para trás uma quantidade de dados valiosíssimos que poderiam estar sendo analisados e empregados para tomada de decisões cada vez mais acertivas na configuração e gestão das campanhas.

Uma dessas importantes tomadas de decisões é a escolha de sites nos quais se compra espaço para mídia. Essa é uma escolha de extrema importância pois corresponde a cerca de 40% do investimento total de mídia de um anunciante. Pela impossibilidade de um único funcionário realizar essas análises, atualmente a criação de listas de sites é feita baseada no conhecimento geral e pessoal desse único funcionário, o que pode afetar na credibilidade do cliente para com estas listas e também em uma performance de campanha diferente da esperada.

Em situações como essa é perceptível que mesmo em um mercado tão singular quanto o marketing, a tecnologia vem abrindo seu espaço e se tornando cada vez mais importante. Reportagem recente da revista *online* Mundo do Marketing afirma:

> Dentre tantas reinvenções do Marketing, poucas delas terão mais impacto na vida do profissional da área do que a chegada da Tecnologia no setor. Isso porque as ferramentas de métricas ou para relacionamento com o consumidor vêm exigindo um olhar cada vez mais analítico e de processamento de dados. Mais do que se preocupar com a forma de comunicação de uma marca, o desafio é compreender as particularidades de cada inovação e saber aplicá-las no dia a dia.[...]O hibridismo no Marketing já é algo que vem ocorrendo em grandes empresas e startups, com pessoas de diferentes formações pensando em uma só estratégia. Se pelo lado do Marketing não há muito conhecimento em plataformas e leitura de dados, a equipe de TI chega para somar forças - o que fornece a empresa mais algum tempo de preparação para que se aprofunde em novas ferramentas. [1]

Nesse cenário, para criar uma base sólida de apoio a essa decisão será utilizado o aprendizado de máquina, subcampo da inteligência artificial, para que a aprendizagem seja constante e acompanhe as tendências do mercado. Esse aprendizado será apoiado sempre em uma estrutura de Redes Bayesianas que fará inferências acerca desses dados gerados pelas campanhas da empresa. Resumidamente, Redes Bayesianas são modelos gráficos para raciocínio baseados em incerteza, onde os nós representam as variáveis e os arcos representam conexões diretas entre eles e por isso vem se tornando a metodologia

padrão para a construção dos sistemas que confiam no conhecimento probabilístico e tem sido aplicadas em uma variedade de atividades do mundo real [2].

Com essas ferramentas então trabalhou-se para que os dados disponíveis possam ser facilmente analisados e usados para elevar a qualidade das próximas campanhas configuradas, fortalecendo assim o relacionamento entre o cliente e a empresa.

1.1 Justificativa

Não é novidade que o mundo está cada vez mais digital. Em um estudo realizado pela empresa de marketing digital *We Are Social* [3], no começo de 2015 chegava a cerca de 3 bilhões o número de usuários ativos na Internet, com uma média de uso de cerca de 7 horas por dia, tanto em laptops quanto smartphones. É irrefutável que esse cenário torna o marketing digital uma ferramenta poderosa para que as marcas consigam se fazer presentes onde os consumidores passam a maior parte do seu tempo, destacando sua imagem no mercado.

Com essa evolução do comportamento de navegação dos usuários na Internet, a publicidade teve que se reinventar e descobrir formas mais eficientes de alcançar o público online. E foi assim que surgiu a Mídia Programática, a combinação perfeita entre marketing e tecnologia, otimizando a compra de mídias por meio de softwares.

Segundo projeções, em 2016, dois terços do total de gastos com anúncios *display* nos Estados Unidos serão comprados via mídia programática, um valor que chega a US$ 22,1 bilhões [3]. Um mercado com capital tão alto assim requer uma análise constante para que o investimento seja cada vez mais preciso e com o maior retorno possível.

Porém é impossível prever exatamente o comportamento de uma campanha e os resultados que serão obtidos com a escolha de cada site. Nunca se terá acesso a todos os fatos sobre o universo no qual estão inseridas e, portanto, uma abordagem lógica do problema não trará resultados satisfatórios. Pode-se, entretanto, tratar o conhecimento incerto e tomar decisões racionais acerca deste.

Além disso, a qualidade de um site está atrelada à situação na qual este foi inserido, é impossível dizer em sua totalidade se um site é bom ou ruim. A escolha pode ser extremamente satisfatória para um certo tipo de campanha e causar prejuízos em uma outra completamente diferente.

Grandes nomes da mídia programática atualmente, como a MediaMath e sua ferramenta de *cross-devide* [4] já utilizam de ferramentas probabilísticas para classificar o comportamento dos usuários de Internet, visto que uma abordagem determinística em algumas situações torna-se infactível.

Por esse motivo, escolheu-se por atacar o problema com um raciocínio probabilístico,

que envolve a Teoria da Probabilidade e principalmente probabilidade condicional, a qual pode ser modelada por Redes Bayesianas.

1.2 Objetivo Geral

O objetivo geral desse trabalho é criar um sistema de fácil utilização e entendimento que sirva para o apoio à tomada de decisão na hora de gerar listas de sites com bom desempenho, também chamadas de *Whitelists* e de desempenho ruim, também chamadas de *Blacklists* para diferentes tipos de campanhas, criando assim uma base sólida que suporte essas escolhas.

Por fim, o resultado esperado é que esse sistema gere automaticamente listas de sites ótimos e sites ruins, baseando sua decisão no conhecimento adquirido ao longo do tempo pela inferência de uma Rede Bayesiana, mas também que forneça informações pertinentes sobre as chances de cada site para que os gestores de campanha possam tomar suas próprias decisões acerca das mesmas.

As probabilidades de sucesso - e também de insucesso - estarão condicionadas às características da campanha.

1.3 Metodologia do Trabalho

O presente trabalho está dividido de forma gradual, iniciando com uma contextualização no qual o problema está inserido, uma breve introdução da empresa na qual o trabalho foi realizado e a problemática dentro dela. A partir da descrição mais detalhada do problema, apresenta-se a solução proposta para resolvê-lo, o seu desenvolvimento, resultados e conclusões.

2 Revisão Bibliográfica

2.1 Marketing Digital

2.1.1 A Transição do Modelo de Marketing

O marketing de maneira geral, não é difícil de se definir. Pode-se recorrer ao dicionário - ou ao Google - e facilmente descrevê-lo como:

> Estratégia empresarial de otimização de lucros por meio da adequação da produção e oferta de mercadorias ou serviços às necessidades e preferências dos consumidores, recorrendo a pesquisas de mercado, design, campanhas publicitárias, atendimentos pós-venda etc. [5].

Há 5 anos, algumas das estratégias mais comuns de marketing eram comerciais de TV, anúncios em jornal e anúncios em outdoors, estando todas estas presentes no cotidiano da população. Porém, o marketing sofreu mudanças signicativas com o avanço da tecnologia, mobilidade e mídias sociais, causando uma transformação digital nos negócios e relacionamentos pessoais. E não só isso, também está abalando os modelos tradicionais e até mesmo o funcionamento da economia.

Porém, não se deve erroneamente pensar que o marketing como foi concebido no passado e essa nova maneira de fazer marketing são concorrentes, pelo contrário. Nota-se ao longo da história que as tecnologias vencedoras não competem com as anteriores e sim chegam para substituí-las. Foi assim com a iluminação elétrica em relação ao lampião dos postes a gás, com a tração mecânica em relação à tração animal, com a impressora em relação à máquina de escrever, com o CD em relação ao disco de vinil [6] e não seria diferente com o marketing digital em relação ao marketing offline.

2.1.2 Um panorama do uso Atual de Internet

Segundo uma pesquisa realizada pelo Grupo Banco Mundial [7], o mundo encontra-se em meio à maior revolução de informação e comunicação da história da humanidade. Mais de 40% da população do mundo têm acesso à Internet, e novos usuários entram *on-line* todos os dias. Entre os 20% dos domicílios mais pobres, quase 7 de cada 10 têm celular. É mais provável que os domicílios mais pobres tenham acesso a telefones celulares do que a sanitários ou água potável. Assim, considerando uma população mundial em 2015 de cerca de 7 bilhões de pessoas [8], esses 40% remetem a quase 3 bilhões de pessoas no mundo com acesso à Internet.

Nesse contexto, é fácil perceber o porquê do Marketing Digital ter uma previsão de crescimento só no Brasil de 12% em 2016 em relação ao ano anterior, chegando a movimentar cerca de R$9,3 bilhões no ano de 2015, de acordo com um estudo realizado pela IAB [9]. Para um melhor entendimento da proporção desse crescimento, pode-se compará-lo com a estimativa de crescimento do setor agropecuário no Brasil para os próximos 5 anos de 3,3% ao ano, sendo este o setor com uma das maiores previsões de crescimento devido, principalmente, ao aumento da produção de biocombustíveis [10].

São mais de 110 milhões de brasileiros conectados à Web, estando entre os países onde os internautas gastam mais tempo online. Embora a televisão aberta ainda seja um importante veículo de mídia no Brasil, existe uma clara e forte tendência de migração para a *Web*, que cresce de forma relevante ano a ano. O Brasil também é o terceiro país no mundo em números de usuários de Facebook, dos quais 80% acessam a rede social via dispositivos móveis. O país também se encontra na segunda colocação mundial de consumo de vídeos pelo Youtube e ocupa o terceiro lugar na rede de negócios Linkedin, logo depois da Índia e dos Estados Unidos [6].

Se o foco das empresas é, e sempre será, se fazer presente nos locais onde seus potenciais clientes passam a maior parte do tempo, tendo assim uma maior visibilidade, com tudo isto posto, a internet se torna então a vitrine mais cobiçada do momento, impulsionando cada vez mais o Marketing Digital.

2.1.3 O que é, de fato, o Marketing Digital

O marketing digital pode ser definido por estratégias de marketing aplicadas e adaptadas para a internet. Em outras palavras, o marketing digital é o nome que se dá para as ações de comunicação de uma empresa que utiliza a internet e dispositivos móveis para divulgar e vender seus produtos e serviços, além de ampliar o seu relacionamento com os clientes [11]. Empresas vêm utilizando o marketing digital nos negócios desde que a internet passou a ser considerada a maior fonte de informação em nível global. Trata-se de uma plataforma conveniente, acessível e que oferece oportunidades competitivas para negócios de todos os tamanhos [12]. Embora a digitalização seja um desafio para as empresas, ela oferece uma grande oportunidade para que as organizações revejam o seu modelo tradicional de negócios [6]. A importância da marca na internet vai além do *e-commerce*. Mesmo quando falamos de venda offline, a internet excerce muita influência na decisão de compras, já que, de acordo com uma pesquisa do Google, 64% dos brasileiros dizem ter feito uma pesquisa online antes de realizar sua última compra, sendo elas tanto *online* quanto *offline* [13].

Com o crescimento da mobilidade, o celular passou a alcançar diretamente o indivíduo, seja ele o pequeno consumidor ou o presidente de uma grande corporação, trazendo um alto potencial de alcance interativo para ser explorado. Diferentemente de

um comercial de TV que é exibido para todos os telespectadores de um certo canal, independente do público de interesse do anunciante, esse avanço digital permite a segmentação do público-alvo usando diferentes tecnologias. Alguns exemplos são o uso de *cookies* que caracteriza o comportamento na Web de um usuário, sem identificações pessoais, o tipo de conexão com a internet, etc.

No marketing digital, as ferramentas mais usadas pelas empresas para aumentar a visibilidade de páginas de internet e consequentemente atrair mais consumidores são os *banners* de *displays* em portais, links patrocinados e a otimização de páginas. Dentre estes, o presente trabalho está inserido no contexto dos *banners display*, mais precisamente em um formato de compra dessas mídias, mais conhecido como mídia programática. A Figura 1 ilustra um exemplo de *banner display* utilizado em uma campanha de divulgação da Publya.

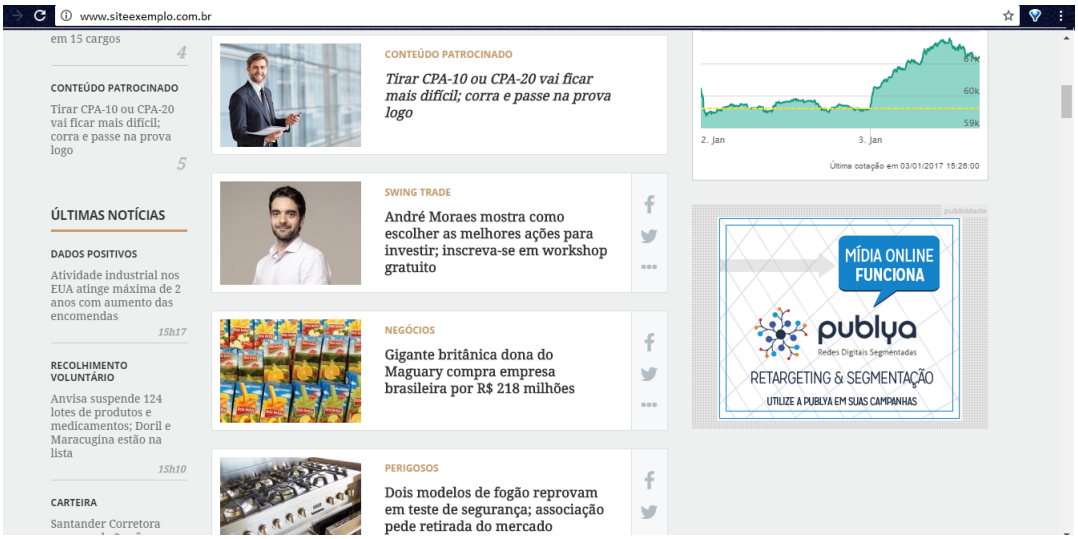


Figura 1 – Exemplo de Banner Display em site da internet.

2.2 Mídia Programática

2.2.1 O que é mídia programática

Até alguns anos atrás, o processo de compra de mídia online era totalmente manual e consumia muito tempo. Para anunciar na internet era preciso negociar separadamente com cada página na Web. Isso sacrificava bastante a produtividade e a eficiência dos resultados. E foi para facilitar esse processo de compra que surgiu a mídia programática. Sua chegada revolucionou o mercado publicitário e vem transformando o modo como as marcas se comunicam com seus consumidores.

A mídia programática pode ser definida como a compra de mídia através de *software* e orientada por dados de comportamento das pessoas na internet. Ou seja, é um processo automatizado de comprar e exibir anúncios online. E que, graças a esses dados, torna possível identificar características demográficas, interesses e intenções de compra dos usuários na internet, e atingir apenas com um público-alvo específico – gerando mais relevância e assertividade nas campanhas. Diferente do que parece, apesar do nome, a programática não é uma mídia e sim uma forma de comprar mídia. Com ela, é possível exibir campanhas de *display*, *mobile* e vídeo.

Essa nova forma automatizada facilitou o processo de compra de mídia online, já que com apenas uma operação o anúncio é exibido em vários sites. Outro grande diferencial são as segmentações, que permitem que a mensagem seja bastante personalizada e direcionada para cada *target*. As marcas também podem exibir banners display diferentes para cada um dos interesses ou intenções de compra dos consumidores, por exemplo. Com tudo isso, o retorno sobre o investimento é maior, já que a mídia é muito mais assertiva pois é direcionada, diminuindo dispersões de verba. E o melhor: as diferentes estratégias podem ser testadas e otimizadas durante a campanha, aumentando as chances de melhores resultados.

2.2.2 Como funciona

Em resumo, a compra dessas mídias se da pela integração entre diversas plataformas, tanto do lado dos anunciantes quanto do lado dos *publishers* - nome usado para se referir aos sites que disponibilizam espaço para anúncio em seus site. Compara-se muitas vezes o estilo de compra com o da bolsa de valores, já que funciona também como um leilão em tempo real - conhecido como *Real Time Biding* (RTB).

Para que tudo isso seja possível, os *publishers* disponibilizam seu inventário para *Ad Exchanges*, empresas que agregam uma enorme quantidade de espaços de mídia para venda, por meio de uma plataforma conhecida como *Suply-side Platform* (SSP). Do lado da compra, os anunciantes precisam de uma *Demand Side Platform* (DSP) para então ter acesso a esse inventário. Através desta plataforma, é possível configurar além do público-alvo, lances de CPM (Custo por Mil Impressões) para cada campanha.

O que acontece então é um leilão em tempo real dos espaços publicitários disponíveis, onde ganha o leilão aquele anunciante que programou o maior lance máximo através da DSP. A Ad Exchange, a cada vez que um usuário visita uma página na Internet, verifica quais anunciantes estão interessados na demografia desse usuário e quais destes deu o maior lance, definindo assim a marca vencedora – que exibirá sua mensagem naquele momento. Todo esse processo ocorre em menos de 70 milissegundos.

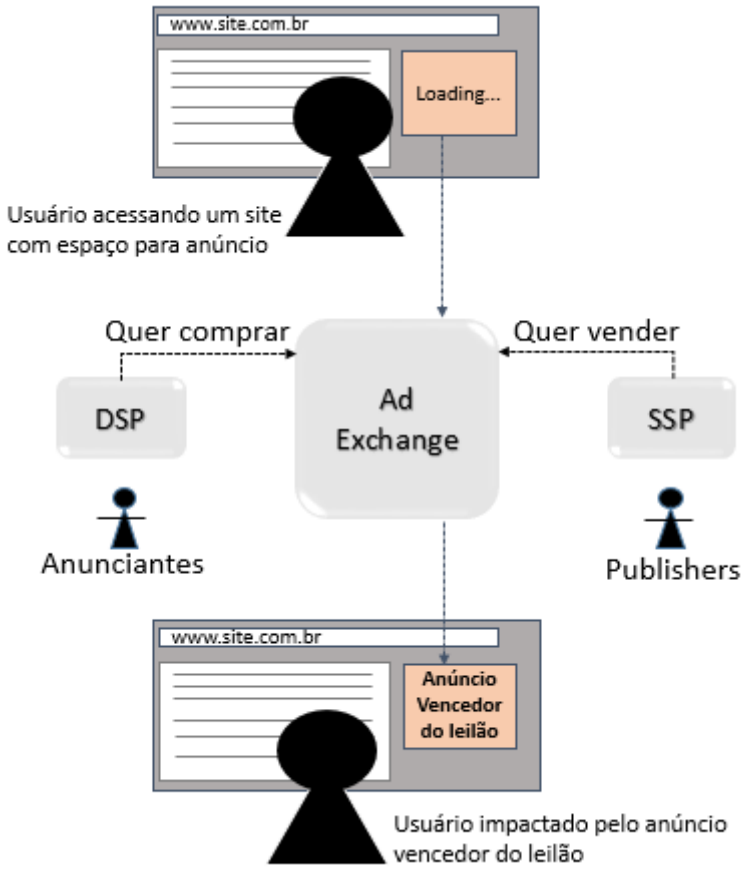


Figura 2 – lustração do funcionamento simplificado da compra de mídia via programática.

2.2.3 Trading Desks

Como pode ser observado pela Figura 2, os anunciantes precisam de uma DSP para anunciar e ter acesso a uma plataforma dessas é caro, já que elas exigem grandes volumes de investimento mensais. Por isso, existe um setor intermediário chamado *Trading Desks*. Estas são empresas especialistas em mídia programática que com acesso a uma DSP, operam campanhas de diversos clientes que contratam seus serviços. Estas empresas são verdadeiras facilitadoras, já que aliam o acesso às tecnologias necessárias para a execução de campanhas com a expertise e pessoas capacitadas para operá-las e buscar os melhores resultados, sem que as marcas precisem despender tempo e dinheiro para preparar um setor interno que opere estas campanhas [14].

A Figura 3 ilustra a mesma compra da Figura 2 porém com o intermédio de uma *Trading Desk*.

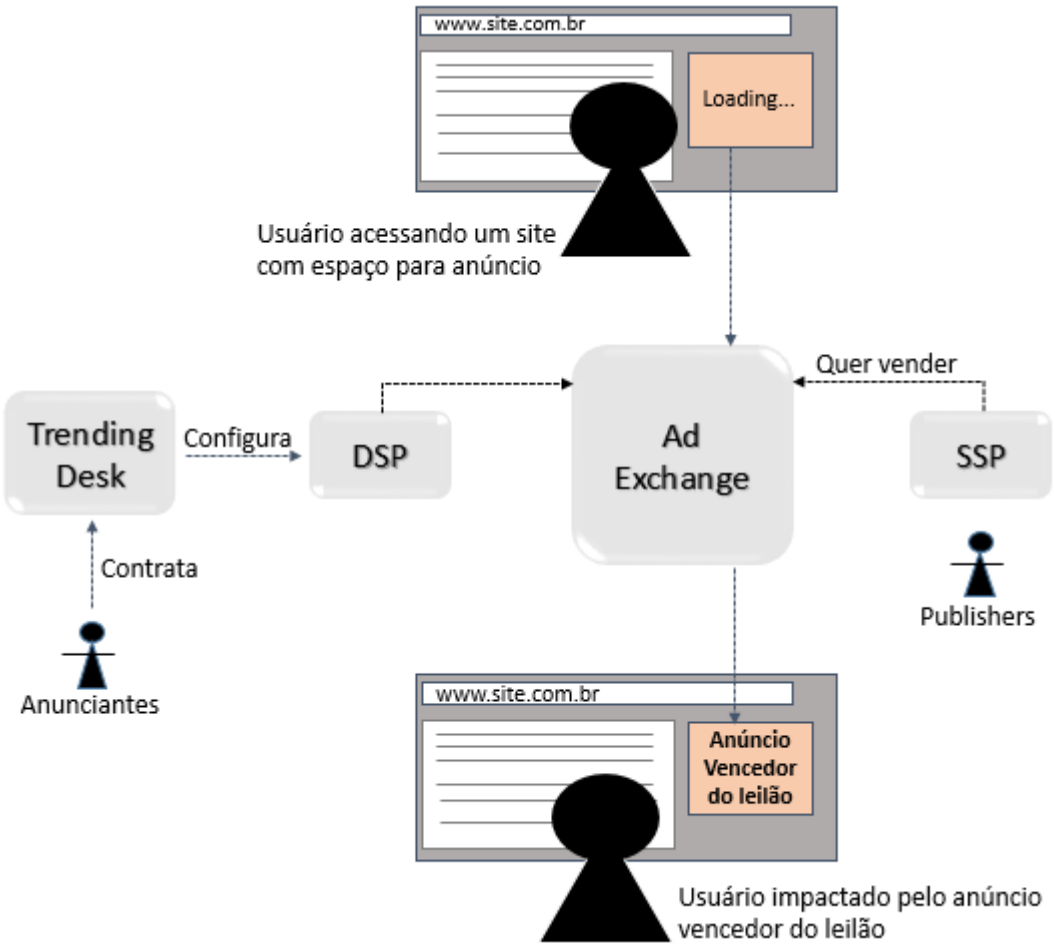


Figura 3 – Ilustração do funcionamento da compra de mídia via programática com intermédio de uma Trading Desk.

A Publya, empresa onde o presente trabalho foi desenvolvido, é uma *Trading Desk* especialista em Mídia Programática. A empresa, com sede em Florianópolis, começou suas operações em 2013 e até então já operou mais de 1000 campanhas de cerca de 400 marcas de todo o Brasil. Nos próximos capítulos seu funcionamento será melhor detalhado.

2.3 Probabilidade

Em muitos ambientes de trabalho, pessoas com certa experiência em determinado assunto, começam a assumir verdades sobre o mundo no qual estão inseridos sem embasamento teórico algum que o suporte, a certo ponto que uma observação sobre uma pequena parte desse mundo os levam a tomar essa observação como verdade absoluta e isso muitas vezes pode não ser o caso. O conhecimento do agente pode, na melhor das hipóteses, fornecer apenas um grau de crença nesses assuntos que se acredita ter conhecimento.

A principal ferramenta para lidar com graus de crença é a teoria da probabilidade,

que atribui a cada preposição um grau numérico de crença entre 0 e 1. A probabilidade proporciona um meio para resumir a incerteza que vem da ignorância de um agente. Atribuir a probabilidade 0 a uma dada proposição corresponde uma crença inequívoca de que a proposição é falsa, enquanto a atribuição da probabilidade 1 corresponde a uma crença inequívoca de que a proposição é verdadeira. As probabilidades entre 0 e 1 correspondem a graus intermediários de crença na veracidade das proposições. É importante observar que um grau de crença é diferente de um grau de verdade. Uma probabilidade de 0,8 não significa "80% verdadeira", mas sim um grau de crença igual a 80% - uma expectativa bastante forte [15].

Todas as declarações de probabilidade devem então indicar a evidência de acordo com a qual a probabilidade está sendo avaliada. À medida que o agente recebe novas percepções, suas avaliações de probabilidade são atualizadas para refletir a nova evidência. Antes de a evidência ser obtida, refere-se a ela como probabilidade a priori ou incondicional. Depois que a evidência é obtida refere-se a ela como probabilidade posteriori ou condicional.

Antes de especificar melhor a probabilidade condicional, será posto a seguir algumas notações básicas de probabilidade, fundamentais para o entendimento do trabalho realizado.

2.3.1 Notações básicas

Os graus de crença sempre são aplicados a proposições, que são afirmações de que tal situação está ocorrendo. O elemento básico da linguagem da teoria da probabilidade é a variável aleatória, que pode ser imaginada como algo que se refere a uma parte do mundo cujo status é inicialmente desconhecido. Cada variável aleatória tem um domínio de valores que pode assumir. Por exemplo em um consultório dentário, o domínio da variável Cárie poderia ser {verdadeira,falso}. A espécie de proposição mais simples afirma que uma variável aleatória tem um valor específico extraído de seu domínio. Por exemplo Cárie = verdadeira poderia representar a proposição de que tenha de fato uma cárie no dente de um paciente [15].

2.3.1.1 Variáveis Aleatórias

As variáveis aleatórias em geral se dividem em três espécies, dependendo do tipo de domínio:

- Variáveis aleatórias Booleanas: Como o exemplo anterior da Cárie, têm domínio {verdadeiro,falso}. Algumas vezes pode-se representar a variável no status "verdadeira"por seu nome com letra minúscula, como no caso de Cárie, *cárie* e seu status "falso"por $\neg cárie$.

- Variáveis aleatórias discretas: incluem variáveis aleatórias Booleanas como um caso especial, enumerável. Por exemplo uma variável aleatória Tempo poderia ter um domínio {ensolarado, chuvoso, nublado, nevoento}. Os valores no domínio devem ser mutuamente exclusivos e exaustivos.

- Variáveis aleatórias contínuas: o tipo mais comum de variáveis, onde assumem valores a partir dos números reais. O domínio pode ser a linha real inteira ou algum subconjunto como o intervalo [0,1].

2.3.1.2 Eventos atômicos

A noção de eventos atômicos é útil para que possa-se entender os princípios básicos da teoria da probabilidade. Um evento atômico é uma especificação completa do estado do mundo sobre o qual o agente está inseguro. Ele pode ser considerado uma atribuição de valores específicos a todas as variáveis das quais o mundo é formado. Por exemplo, se um certo mundo é composto por apenas duas variáveis aleatórias do tipo Booleanas Cárie e DorDeDente, então existem apenas quatro eventos atômicos distintos possíveis: $cárie \wedge dordedente$, $\neg cárie \wedge \neg dordedente$, $\neg cárie \wedge dordedente$ e $cárie \wedge \neg dordedente$.

Os eventos atômicos têm algumas propriedades importantes:

- São mutuamente exclusivos ou seja, no máximo um deles pode ocorres em cada instante. Por exemplo $cárie \wedge dordedente$ e $cárie \wedge \neg dordedente$ não podem ocorrer simultaneamente.

- O conjunto de todos os eventos possíveis é exaustivo, ou seja, pelo menos um dele tem que ocorrer.

- Qualquer evento atômico específico impõe a verdade ou falsidade de toda proposição, seja ela simples ou complexa [15].

2.3.2 Probabilidade a priori

A probabilidade incondicional, ou a priori, associada a uma proposição a é o grau de crença acordado para a proposição na ausência de quaisquer outras informações e é representada por $P(a)$. Por exemplo, se a probabilidade a priori que afirma que aquele mesmo paciente tem uma cárie é 0,1, então escreve-se:

$$P(Cárie = verdadeiro) = 0,1$$

ou

$$P(cárie) = 0,1$$

Muitas vezes, por praticidade, refere-se às probabilidades de todos os valores possíveis de uma variável aleatória. Nesse caso, tomando como exemplo a descrição do tempo, é comum usar a expressão $P(Tempo)$, que denota um vetor de valores para as probabilidades de cada estado meteorológico individual. Desse modo, ao invés de escrever as quatro equações:

P(Tempo = ensolarado) = 0,7

P(Tempo = chuvoso) = 0,2

P(Tempo = nublado) = 0,08

P(Tempo = nevoento) = 0,02

Escreve-se somente:

P(Tempo) = $\langle 0,7; 0,2; 0,08; 0,02 \rangle$

Usa-se comumente também expressões como $P(Tempo, Cárie)$ para denotar as probabilidades de todas as combinações de valores de um conjunto de variáveis aleatórias. Nesse caso, $P(Tempo, Cárie)$ pode ser representada por uma tabela de 4x2 de probabilidade. Isso se chama distribuição de probabilidade conjunta de Tempo e Cárie.

No caso de variáveis contínuas, não é possível representar a distribuição inteira como uma tabela, porque existem infinitos valores [15]. Muitas vezes, com uma variável aleatória contínua, divide-se o domínio em subdomínios para facilitar os cálculos de probabilidade.

É importante lembrar que $P(a)$ pode ser usado somente quando não existe nenhuma outra informação. Assim que algumas informações novas são conhecidas, deve-se raciocinar com a probabilidade condicional de a, dadas essas novas informações.

2.3.3 Probabilidade Condicional

Uma vez que o agente obtém alguma evidência relativa às variáveis aleatórias anteriormente desconhecidas que constituem o domínio, as probabilidades a priori não são mais aplicáveis. Ao invés disso usa-se então as probabilidade condicionais ou posteriores. A notação usada é $P(a|b)$ onde a e b são proposições quaisquer. Essa expressão pode ser lida como "a probabilidade de a, dado que tudo que sabemos é b". Usa-se como exemplo novamente o dente com cárie em um certo paciente onde dizer

$$P(cárie|dordedente) = 0,8$$

indica que, se for observado que o paciente tem uma dor de dente e ainda não houver nenhuma outra informação disponível, a probabilidade do paciente ter cárie é de 80% [15].

No tratamento de probabilidades condicionais, deve-se ter em mente que a probabilidade de um evento ocorrer está condicionada a algum outro fato conhecido.

Para exemplificar, toma-se um dado lançado: a probabilidade de obter um 6, no senso comum, é de $\frac{1}{6}$, porém só se pode afirmar que essa é a real probabilidade se assumirmos que o dado não está viciado, que é um dado "justo", caso contrário a probabilidade não seria mais a mesma. Nessa linha de raciocínio, pode-se dizer então não mais que a probabilidade de a acontecer é x mas sim que dado o fato conhecido b, a probabilidade de a ocorrer, é x. Voltando ao exemplo do dado, diz-se que, sabendo que o dado é um dado justo (b) a probabilidade de o dado cair com a face 6 para cima (a), é de $\frac{1}{6}$.

As probabilidades condicionais podem ser definidas em termos de probabilidades incondicionais. A equação de definição é:

$$P(a|b) = \frac{P(a \wedge b)}{P(b)} \tag{2.1}$$

que é valida sempre que $P(b) > 0$. Essa equação também pode ser escrita como:

$$P(a \wedge b) = P(a|b)P(b) \tag{2.2}$$

que denomina-se regra do produto. Também pode-se escrevê-la no sentindo contrário tendo então:

$$P(a \wedge b) = P(b|a)P(a) \tag{2.3}$$

Além disso, seja D a variável discreta que tem o domínio $\langle d_1, ..., d_n \rangle$. Então é fácil mostrar que:

$$\sum_{i=1}^{n} P(D = d_i) = 1$$

Ou seja, qualquer distribuição de probabilidade sobre uma única variável deve somar 1 assim como qualquer distribuição de probabilidade conjunta sobre qualquer conjunto de variáveis deve somar 1 e a probabilidade de uma proposição deve ser igual a soma das probabilidades dos eventos atômicos em que ela é válida, isto é:

$$P(a) = \sum_{e_i \in e(a)} P(e_i) \tag{2.4}$$

Essa regra facilitará as operações de probabilidade em diversas ocasiões, removendo das operações todos os eventos atômicos nas quais a probabilidade a ser calculada não é válida.

2.4 Redes Bayesianas

A componente amostral é comum aos modelos clássicos e Bayesianos, mas com interpretações diferentes. Embora os modelos Bayesianos passem por uma extensão dos

modelos clássicos, existe uma divergência fundamental entre os dois enfoques: no modelo clássico o parâmetro é um escalar ou um vetor desconhecido, porém fixo, ao passo que no modelo Bayesiano, o parâmetro é considerado como escalar ou vetor aleatório (não observável), pois para os Bayesianos tudo o que é desconhecido é incerto e, portanto, toda a incerteza deve ser quantificada em termos de probabilidade [16]. Em função disto, os modelos Bayesianos tratam formalmente a informação a priori, através da distribuição de probabilidade (subjetiva ou lógica) a priori. As informações a priori e amostrais permitem a atualização periódica da distribuição de probabilidade a posteriori e, portanto, permite modificar e atualizar as estimativas dos parâmetros. O teorema de Bayes é uma regra para atualização de probabilidades e pode ser utilizado para cálculo de probabilidades a posteriori em mais que um estágio [17].

2.4.1 Regra de Bayes

Como mostrado nas equações 2.2 e 2.3, a regra do produto pode ser escrita de duas maneiras devido à comutatividade da conjunção. Igualando os dois membros da direita e dividindo por $P(a)$, obtém-se:

$$P(b|a) = \frac{P(a|b)P(b)}{P(a)} \tag{2.5}$$

Essa equação é conhecida como regra de Bayes e é a base de todos os sistemas modernos de Inteligência Artificial para inferência probabilística. O caso mais geral de variáveis multi valoradas pode ser escrito na notação de P como:

$$P(X|Y) = \frac{P(X|Y)P(Y)}{P(X)} \tag{2.6}$$

Para ilustrar a regra de Bayes, pode-se usar um exemplo extraído de [17]:

Seja A uma variável aleatória com n estados, a_1, ..., a_n, e $P(A)$ a distribuição de probabilidades para estes estados, $P(A) = (x_1, ..., x_n)$; $x_i \geq 0$; $\sum_{i=1}^{n} x_i = 1$, onde x_i é a probabilidade de A estar no estado a_i, $P(A = a_i)$. Se a variável B possui os estados $b_1, ...b_m$, então $P(A|B)$ representa uma tabela $n \times m$ contendo os valores $P(a_i|b_j)$, a qual segue abaixo:

	b_1	b_2	b_3
a_1	0.4	0.3	0.6
a_2	0.6	0.7	0.4

Tabela 1 – Probabilidades Condicionais $P(A|B)$

A conjunção de probabilidades para as variáveis A e B, ou $P(A, B)$, é também uma tabela $n \times m$, representada pela probabilidade de cada configuração (a_i, b_j).

Se for imposto que $P(B) = \langle 0.4, 0.4, 0.2 \rangle$, aplicando-se a regra fundamental da equação 2.1 e a Tabela 1 para as variáveis A e B, obtem-se:

	b_1	b_2	b_3
a_1	0.16	0.12	0.12
a_2	0.24	0.28	0.08

Tabela 2 – Probabilidades Conjunta $P(A, B)$

A probabilidade $P(X)$ pode ser então calculada a partir da Tabela 2, calculando-se: $P(a_i) = \sum_{j=1}^{m} P(a_i, b_j)$. Este cálculo é chamado Marginalização de B em $P(A, B)$.Usando esse cálculo pode-se então calcular $P(A)$ como segue:

	$P(X)$
$A1$	0.4
$A2$	0.6

Tabela 3 – Probabilidades Marginal de $P(A)$

Com as tabelas anteriores então, pode-se usar a regra de Bayes para calcular $P(B|A)$, resultando na tabela abaixo:

	$A1$	$A2$
$b1$	0.4	0.4
$b2$	0.3	0.47
$b3$	0.3	0.13

Tabela 4 – Probabilidades Bayesiana $P(B|A)$

O exemplo acima traz a impressão de que seria muito fácil calcular todas as probabilidades condicionais através da junção de tabelas mas em problemas reais, muitas vezes o número de variáveis e estados é muito maior, dificultando assim a construção de todas as tabelas de probabilidade.

2.4.2 Redes Bayesianas

Redes Bayesianas são também chamadas Redes de Bayes, Redes Probabilísticas Causais (CPNs), Redes de Crença Bayesiana (BBNs) ou simplesmente redes de crença. A melhor forma de entender Redes Bayesianas é procurar modelar uma situação na qual a causalidade desempenha um papel importante, mas onde o entendimento sobre o que realmente está ocorrendo não é muito claro, de tal forma que é necessário descrever as coisas probabilisticamente.

Redes Bayesianas são Gráficos Acíclicos Diretos compostos por nós e setas. Os nós são variáveis aleatórias, e valem premissas de independência entre elas. Frequentemente,

as variáveis aleatórias podem ser imaginadas como estados de um determinado assunto, podendo assumir estados Booleanos, discretos ou contínuos. As setas especificam as premissas de interdependência entre as variáveis aleatórias, suas relações de causa e efeito dentro do domínio. Portanto, as premissas de independência entre os nós da rede determinam qual informação de probabilidade é requerida para especificar a distribuição de probabilidades entre as variáveis aleatórias na rede [17].

Fundamentalmente, as Redes Bayesianas são usadas para atualizar probabilidades quando a informação chega. A base matemática para isto é o Teorema de Bayes, já definido no capítulo anterior.

Para especificar a distribuição de probabilidades de uma Rede de Bayes, é necessário obter as probabilidades de todos os nós-raiz (nós sem predecessores) e as probabilidades condicionais de todos os nós-não-raiz, dadas todas as possíveis combinações de seus predecessores diretos.

Para exemplificar a criação de uma rede simples, pode-se usar um exemplo extraído de [18], chamado O Exemplo da Macieira.

O exemplo consiste em uma pequena plantação que pertence ao Jack. Um dia Jack percebe que sua melhor macieira está perdendo as folhas e ele quer saber o porquê disso estar acontecendo. Ele sabe que se as folhas estão secas devido à um período de seca, não há mistério - é bem comum as folhas caírem em um período assim. Por outro lado ele também sabe que perder as folhas pode ser uma indicação de doença na árvore.

Essa situação pode então ser modelada pela Rede Bayesiana abaixo:

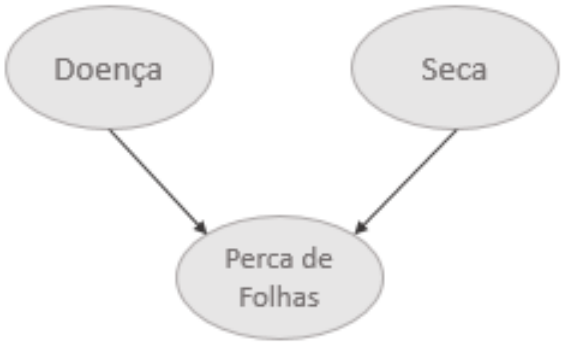


Figura 4 – Exemplo da representação qualitativa do domínio do problema por Redes Bayesianas

A setas indicam a dependência causal do nó "Doença"com "Perca de Folhas"e a dependência causal do nó "Seca"com "Perca de Folhas". Pela rede fica claro que é doença e seca que causam a queda das folhas e não a queda de folhas que causa doença ou seca.

O grafo representa qualitativamente o domínio mas para representar o domínio quantitativamente precisa-se das tabelas de probabilidade de cada nó.

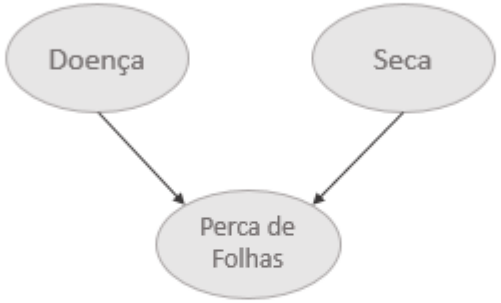
Ainda no exemplo da macieira a representação quantitativa da Rede pode ser representada como mostrado na Figura 5

P(Doença)

Doença = sim	Doença = não
0.1	0.9

P(Seca)

Seca = sim	Seca = não
0.1	0.9



	Seca = sim		Seca = não	
	Doença = sim	Doença = não	Doença = sim	Doença = não
Perca de folha = sim	0.95	0.85	0.90	0.02
Perca de folha = não	0.05	0.15	0.10	0.98

P(Perda de Folha|Doença, Seca)

Figura 5 – Exemplo da representação quantitativa do domínio do problema por Redes Bayesianas

Com a tabela de probabilidade condicional - também chamada de TPC - Jack terá acesso a qualquer tipo de probabilidade relacionada com sua macieira, seca e doença.

2.4.3 Inferência em Redes Bayesianas

Com a rede Bayesiana definida, pode-se extrair conhecimento nela representado através de um processo de inferência. Existem vários métodos para realização de inferência, dentre os métodos tradicionais destacam-se o de propagação em poliárvores e o de eliminação de variáveis [19]. Inferências podem ser realizadas sobre redes Bayesianas, em quatro maneiras distintas:

- Diagnósticos: partindo dos efeitos para as causas.
- Causa: partindo das causas para os efeitos.

- Intercausal: entre causas de um efeito comum.
- Mistas: combinação de dois ou mais tipos descritos acima.

2.5 Aprendizado de máquina

Segundo Norvig e Russel [15], a ideia por trás da aprendizagem é que as percepções devem ser usadas não apenas para agir, mas também para melhorar a habilidade do agente para agir no futuro. A Aprendizagem ocorre à medida que o agente observa suas interações com o mundo e com seus próprios processos de tomada de decisão. A aprendizagem pode variar desde a memorização trivial da experiência até a criação de teorias científicas inteiras.

Existem vários tipos de aprendizado e geralmente são classificados entre:

- Aprendizagem supervisionada, como por exemplo para treinamento de redes neurais e árvores de decisão, sendo essa a principal técnica utilizada para casos de classificação.
- Aprendizagem não-supervisionada, que envolve a aprendizagem de padrões na entrada, quando não são fornecidos valores de saída específicos e é usada principalmente no contexto de sistemas de raciocínio probabilístico.
- Aprendizagem por reforço, a qual baseia a aprendizagem em recompensas no tipo de ação tomada pelo agente, fornecendo assim indicações de que o comportamento atual é desejável ou não mais ou menos como quando uma criança aprende que quando recebe uma "bronca"é porque sua ação foi indesejável e quando recebe um presente é porque seu comportamento foi o desejável.

Dentro de cada uma dessas categorias há inúmeras subcategorias e inúmeros algoritmos para cada uma dessas, os quais podem - e devem - ter seu desempenho avaliado.

Para facilitar essa avaliação de desempenho pode-se seguir a metodologia a seguir, bem simples e genérica, podendo ser usada para todos os tipos de algoritmos de aprendizagem:

1. Coletar um grande número de exemplos.
2. Dividi-lo em dois conjuntos disjuntos: o conjunto de treinamento e o conjunto de teste.
3. Aplicar o algoritmo de aprendizagem ao conjunto de treinamento, gerando uma hipótese h.

4. Medir a porcentagem de exemplos no conjunto de teste que são corretamente classificados por h.

5. Repetir as etapas 1 a 4 para diferentes tamanhos de conjuntos de treinamento de cada tamanho selecionado aleatoriamente.

Ao seguir esse procedimento, pode-se então avaliar esse aprendizado em função do tamanho do conjunto de treinamento. Essa função pode ser representada em um gráfico e é chamada de curva de aprendizagem.

3 A Empresa: Publya

A Publya é uma *Independent Trading Desk* especializada em Mídia Programática. Através de plataformas DSP otimiza a compra de mídia *online* em *display* e vídeo, veiculadas em dispositivos desktop e mobile, entregando as campanhas para uma audiência qualificada e segmentada.

A empresa conta com o selo Google Partner e equipe certificada em Google Adwords, onde planeja, opera e otimiza campanhas de Links Patrocinados e Vídeos no Youtube. A Publya oferece também o serviço de planejamento e operação de campanhas em Facebook, Instagram, LinkedIn e Waze. Os anúncios são exibidos apenas em canais relevantes e para o segmento desejado, proporcionando melhores resultados para os anunciantes e agilidade para as agências.

A Empresa está em constante desenvolvimento, acompanhando assim o crescimento desse mercado de marketing digital, principalmente de mídia programática. Para ilustrar esse crescimento pode-se citar os investimentos feitos pelos anunciantes com a Publya em 2015 e 2016, sendo de R$ 3.772.584,94 para R$ 8.687.107,69 respectivamente, evidenciando um crescimento de 130% e a importância na empresa no mercado de mídia digital.

3.1 Campanhas

Quando o anunciante procura a Publya para uma campanha de *banners display*, seja diretamente ou por intermédio de uma agência, ele deve já ter muito bem definido o intuito dessa propaganda; qual tipo de público alvo quer atingir, o que espera dessa público alvo, em qual localização, por quanto tempo e com qual orçamento. Com essas informações, o time de gestão de campanhas pode começar então a configurar essa campanha para esse cliente, exclusivamente. Após configurada, essa campanha terá um acompanhamento diário para que se tenha certeza de que os *targets* estão corretos, que os lances máximos configurados para a *Ad Exchange* estão sendo vencedores e avaliando quais mudanças podem beneficiar a campanha.

Para analisar a campanha, algumas métricas muito comuns no marketing digital são utilizadas e serão descritas abaixo:

- Impressões: Número de vezes que o banner foi mostrado.
- Cliques: Número de vezes que houve um clique no banner mostrado.
- Conversões: Número de conversões da campanha. O tipo de conversão será definido pelo cliente e pode ir desde um preenchimento de formulário até uma compra

finalizada.

- CTR - Click Trough Rate: Taxa de cliques por impressões e é calculado por

$$\frac{Numero\quad de\quad Cliques}{Numero\quad de\quad Impressoes}\%$$

- CPC - Custo por Clique: Custo por cada clique e é calculada por

$$\frac{Verba\quad Gasta}{Numero\quad de\quad Cliques}$$

- CPA - Custo por Ação: Custo por cada conversão e é calculado por

$$\frac{Verba\quad Gasta}{Numero\quad de\quad Conversoes}$$

A métrica utilizada para analisar a performance de um site no qual está sendo comprado espaço para *banners* é o CTR. Quanto mais alto o CTR de um site melhor, sendo esses os escolhidos para se pagar CPMs (Custo por Mil Impressões) mais caros.

O anunciante precisa sempre saber exatamente o que espera desse público alvo ao fazer sua campanha. Por exemplo, este anunciante pode almejar apenas que seus clientes tenham ciência da existência da marca ou que fiquem sabendo de alguma novidade trazida pela mesma. Nesse caso, o objetivo da campanha será o CPC, ou seja, a plataforma DSP e todos os ajustes da campanha serão configurados para otimizar sempre essa métrica já que o cliente se interessa apenas por visualização e clique em seus *banners*. Um outro exemplo seria um outro anunciante que espera que esse público atingido, além de clicar nos *banners*, preencham um formulário de cadastro, no memento que são direcionados pelo clique. Nesse caso, o objetivo será o CPA e além dos cliques, essas conversões - preenchimento do formulário - serão contabilizadas e a plataforma DSP e todos os ajustes da campanha serão configurados para otimizar sempre essa métrica.

Diferentemente do CTR, as métricas de CPC e CPA, quanto menores melhores, já que tratam-se de custos.

3.2 A problemática dentro da empresa

Em uma campanha existem diferentes maneiras de se atingir o público alvo desejado e uma delas é chamada de *Whitelist*. Uma *Whitelist* nada mais é do que uma lista de sites selecionados pela empresa, separadas por assuntos ou segmentos. Por exemplo, pode-se tomar uma *Whitelist* negócios na qual estarão presentes websites como *cartacapital.com.br, administradores.com.br e endeavor.org*.

Por saber exatamente em quais sites será exibido os *banners display*, costuma-se pagar CPMs muito mais caros para que os lances na *Ad Exchange* sejam vitoriosos na

maioria das vezes. Porém, essas listas são feitas baseadas em achismos e suposições da pessoa que prepara a lista, sem qualquer análise sobre a eficácia desses sites em tipos específicos de campanhas. Um exemplo seria o site *cartacapital.com.br* ter sido adicionado na lista por parecer promissor para campanhas relacionadas com negócios, isso basicamente por abordar o tema negócios, mas não havendo embasamento algum sustentando essa informação.

Além das *Whitelists* a empresa também trabalha com *Blacklists*, que nada mais são do que listas de sites dos quais não se deseja comprar *banners* pois percebeu-se ao longo do tempo que tais sites não performam de maneira satisfatória ou porque é um site que o cliente não quer ter sua marca vinculada. Do mesmo jeito que as *Whitelist*, as *Blacklists* são baseadas em observações recentes de gestores de campanha ou em achismos.

Algumas coisas estão fora de nossas capacidades humanas, como tentar prever quais sites terão boa performance em determinadas circunstâncias. Profissionais da área de marketing podem ter uma boa intuição sobre quais desses sites podem ter melhor performance mas o problema é que essas regras orientadas pela intuição podem ser equivocadas e incompletas. A única maneira de chegar a regras corretas é pentear através de milhões de exemplos de sites e campanhas [20].

O problema torna-se então a não utilização de dados disponíveis de campanhas passadas para criar uma base sólida para a escolha de sites, enquanto um estudo detalhado sobre eles poderia estar sendo feito, gerando maior confiabilidade nas escolhas e fortalecendo o relacionamento com os clientes. Além disso, a geração dessas *Whitelists* e *Blacklists* de forma manual se torna muito onerosa à empresa, consumindo assim tempo que poderia estar sendo gasto na otimização das campanhas.

4 Formalização do Problema

4.1 O problema

Todas as campanhas que rodam na DSP geram dados que podem ser resumidos como abaixo.

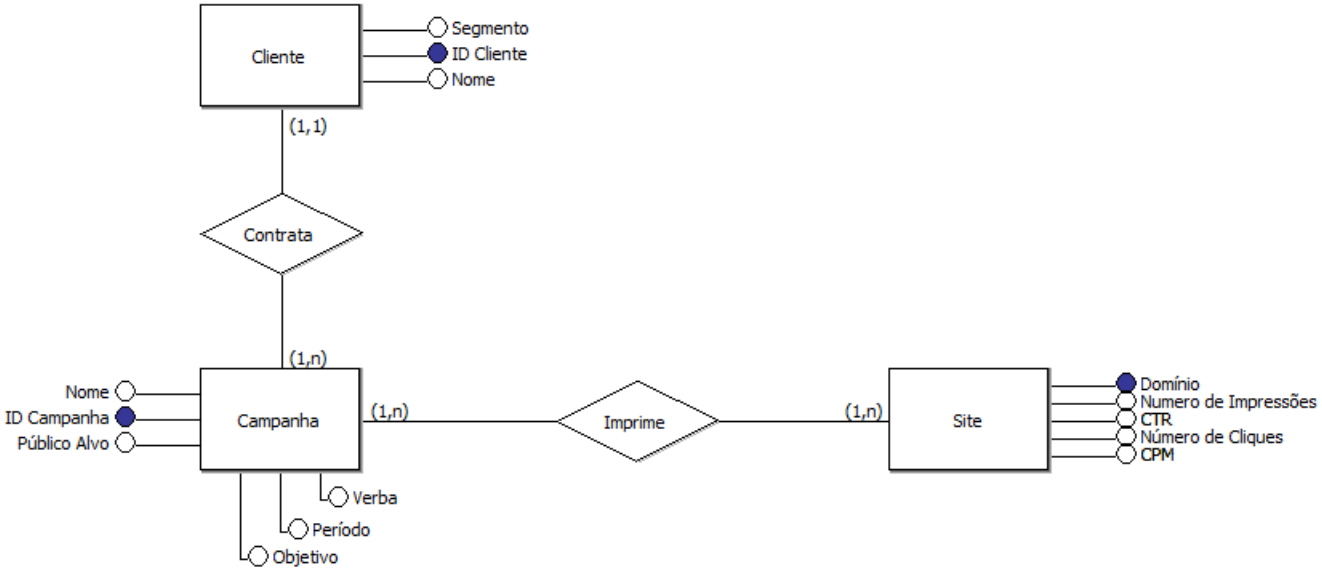


Figura 6 – Estrutura de dados resumida.

- CTR é uma variável contínua com domínio de valores que pode ir de 0 a 100%.
- O Objetivo é uma variável discreta que pode ser CPC ou CPA.
- O segmento é uma variável discreta que representa o segmento comercial de um cliente e é sempre composto por uma sigla de 3 letras. Os principais atualmente na empresa são AUT - automóveis, IMO - Imóveis, EDU - Educação e SHO - shoppings.
- Domínio é uma variável discreta que representa o domínio do site.

É preciso então, a partir de um aprendizado constante, criar um sistema que gere automaticamente *Whitelists Blacklists* baseados na inferência Bayesiana a partir de dados de campanhas anteriores, criando assim uma ferramenta capaz de julgar os domínios em suas probabilidades de obter uma performance satisfatória, de acordo com seus CTRs, para uma campanha de determinado segmento e objetivo.

4.2 Solução Proposta

O aprendizado de máquina vem sendo cada vez mais utilizado na área de marketing digital pois descreve a prática de um computador que se adapta sem ter que ser programado. Quanto mais dados forem alimentados ao programa de computador, mais inteligente ele será. Ele pode fazer previsões melhores e evoluir continuamente sem que o engenheiro ou programador tenha que fazer ajustes no código com base nas saídas [20].

No campo da análise de dados, o aprendizado de máquinas é um método usado para planejar modelos complexos e algoritmos que prestam-se para fazer predições - no uso comercial, isso é conhecido como análise preditiva. Esses modelos analíticos permitem que pesquisadores, cientistas de dados, engenheiros, e analistas possam "produzir decisões e resultados confiáveis e repetíveis"e descobrir os "*insights* escondidos"através do aprendizado das relações e tendências históricas nos dados [21].

Antes de escolher um método de aprendizado, é necessário se escolher o que de fato se quer aprender e como esse conhecimento será adquirido. Se tratando de informações que não são sempre exatas e que se tem total conhecimento sobre o mundo no qual está inserido, optou-se por um método de raciocínio probabilístico. Será utilizado inferência em Redes Bayesianas, vertente essa que tem como uma das suas principais características a adaptabilidade, podendo, a partir de novas informações, e com base em informações de cunho verdadeiro, gerar alterações nas dependências e nos seus conceitos. Permite, dessa forma, que as probabilidades não sejam meros acasos, podendo confirmar e criar novos conceitos [22].

A principal vantagem desse raciocínio probabilístico sobre algum outro raciocínio lógico é o fato de que os agentes podem tomar decisões racionais mesmo quando não existe informação suficiente para se provar que uma ação funcionará [17].

A escolha pode trazer a dúvida de porquê não usar o Data Mining - mineração de dados - para a resolução do problema. Em termos simples, enquanto o aprendizado de máquina usa muitos dos mesmos algoritmos e técnicas que a mineração de dados, uma diferença está no que as duas disciplinas preveem. A mineração de dados descobre padrões e conhecimento previamente desconhecidos e o aprendizado de máquina é usado para reproduzir padrões e conhecimento conhecidos, automaticamente aplicar isso a outros dados, e, em seguida, aplicar automaticamente esses resultados à tomada de decisões e ações [21]. Usar o data mining poderia prever os padrões de sites com melhor performance mas não daria a liberdade de fornecer ao agente quais são as chances de cada um desses sites.

5 Desenvolvimento

Tendo formalizado todas as principais bases de conhecimento para a resolução do problema levantado, pode-se começar a descrever o desenvolvimento da solução proposta.

5.1 Dados Disponíveis

Como dito anteriormente a Publya trabalha com uma DSP que realiza todas as transações de compra de mídia programática. Essa mesma DSP disponibiliza então para a empresa uma seção de *reports*, onde pode-se criar relatórios de campanhas, com diversas métricas e períodos. Esses dados são escolhidos e a plataforma da DSP gera um arquivo.csv com todas as informações. Para atingir a solução proposta, será preciso os seguintes dados de cada campanha:

- Segmento
- Objetivo
- Domínios dos Sites nos quais houve impressão dos banners da campanha
- CTR de cada site

Pelo fato de o segmento ser uma característica atribuída para um cliente, pela empresa, e que para a DSP não há necessidade dessa informação, o segmento das campanhas terá que ser obtido de forma indireta. Uma campanha possui sempre um cliente atrelado a ela, e quando esse cliente é cadastrado na plataforma seu nome é seguido por um código de 3 letras, que indicam o segmento no qual o cliente foi classificado. Para exemplificar, "Cliente Teste - EDU"traz a informação de que o cliente pertence ao ramo da educação e portanto também sua campanha. Para a informação de segmento da campanha então, será preciso trazer também os dados de nome do cliente.

O Objetivo de uma campanha é de extrema importância para a otimização da mesma dentro da plataforma, portanto esta informação é sempre passada para a DSP e portanto poderá ser obtida de forma direta.

Os domínios dos sites são sempre mostrados pela plataforma, já que tal transparência é fundamental, portanto também poderá ser obtida de modo direto.

O CTR de cada site é uma taxa já calculada pela plataforma utilizando-se das informações de impressões e cliques de cada site. Seu domínio é qualquer valor entre 0 e 100% e também pode ser obtida de forma direta. Porém, como o CTR será a métrica

utilizada na seleção de sites bons e ruins, seria impossível fazê-lo com um domínio contínuo assim, portanto decidiu-se por separar os dados em classes, transformando assim essa variável contínua em discreta.

O número de impressões pode ser obtido de forma direta e também deverá utilizado para evitar problemas de fraude, explicado melhor na próxima seção.

Um outro dado disponível de extrema importância é a data de veiculação das campanhas, já que para construir um modelo é necessário uma conjunto de dados para treino e um conjunto de dados para teste. Decidiu-se por não separar os dados aleatoriamente e sim cronologicamente e para isso será preciso separar as informações por data. Porém, a plataforma só permite a separação de dados por datas que estão inseridas em algum período dos últimos 4 meses ou um resumo de uma campanha durante todo seu período de veiculação, não importando qual seja esse período. Isso deixará a coleta desses dados um pouco mais trabalhosa, já que será preciso então fazer um levantamento inicial de quais campanhas rodaram em um primeiro período escolhido e quais em um segundo período escolhido, buscando depois, uma por uma, dentro da plataforma e trazendo então as informações desejadas.

5.2 Tratamento de dados

Antes de separar os dados entre conjunto de treinamento e conjunto de teste, criou-se uma base de dados com todas as informações mencionadas na seção anterior.

Por facilidade, decidiu-se por realizar todos os tratamentos de dados em Planilhas Google, uma variação de planilha de Excel com algumas funções extras, consideradas de extrema importância para futuras operações com os dados coletados.

Para tratar a variável Nome do Cliente, que traz a informação de segmento da campanha, acrescentou-se uma coluna a mais na base de dados e foi utilizada uma fórmula simples que extrai apenas as três primeiras letras da direita para esquerda, para cada linha.

Há um segundo tratamento que deve ser feito, correspondente ao número de impressões de um site. No ramo de compra de mídias online há diversos sites envolvidos em fraudes na compra e venda de seus espaços para propaganda, resultando em sites com baixíssimo número de impressões e que por ficarem em posições incômodas do site, acabam gerando cliques "acidentais", resultando assim em CTRs erroneamente altos. Tratando-se da compra programática de mídias, a venda de espaço se dá por lotes de mil impressões, por isso é de costume da empresa só começar a tomar qualquer tipo de decisão acerca de um site, após o mesmo ter tido mais de 5000 impressões. Por esse motivo será aplicado um filtro nos sites, eliminando os registros com menos de 5000 impressões, também realizado

de forma simples na própria filtragem da planilha.

Para transformar a variável CTR de contínua para discreta, optou-se por agrupá-la em quatro classes, as quais foram escolhidas de acordo com as necessidades da empresa e estimativas de mercado. As classes escolhidas foram

- CTR 1 = [0,0.03[
- CTR 2 = [0.03,0.06[
- CTR 3 = [0.06,0.1[
- CTR 4 = [0.1,1]

e qualitativamente representam CTR inaceitável, CTR ruim, CTR na média e CTR muito bom, respectivamente. Optou-se por não fazer uma separação igualitária, visto que no mercado de programática a maioria dos CTRs estão abaixo de 0,1 e fazer uma divisão igualitária comprometeria a qualidade dos resultados.

5.3 Construção da Rede

Como definido anteriormente, o problema será modelado por um Rede Bayesiana.

Sabe-se do domínio que o CTR é influenciado diretamente pelo site e pelo objetivo e segmento da campanha.

Para modelar o problema, deve-se ter em mente que para cada variável A que tem pais $B_1, ..., B_n$ existe uma tabela $P(A|B_1, ...B_n)$.

Para melhor definição da rede pode-se usar um algoritmo simples :

1. Escolher um conjunto de variáveis X_i que descrevam o domínio.
2. Escolher uma ordem para as variáveis.
3. Enquanto existir variáveis:
   a) Escolher uma variável X_i e adicionar um nó na rede.
   b) Determine os nós Pais(X_i) dentre os nós que já estejam na rede e que tenham influência direta em X_i.
   c) Defina a tabela de probabilidades condicionais para X_i.

Para montar a rede, começando pelo item 1 escolhe-se as variáveis já citadas anteriormente: Site, Segmento, Objetivo e CTR podendo já escolher essa ordem para o item de número 2. Para a primeira variável Site, a) adiciona-se um nó na rede b)

determina-se que não há pais para a variável Site. O item c será feito posteriormente para todas as variáveis. Para a segunda variável Segmento, a) adiciona-se mais um nó na rede b) determina-se que não há pais para a variável Segmento. Para a terceira variável Objetivo, a) adiciona-se um nó na rede b) determina-se que não há pais para a variável Objetivo. Nesse momento a rede está como mostra a Figura 7.

Figura 7 – Representação parcial da rede.

Agora para a quarta e última variável CTR a)adiciona-se mais um nó na rede b)verifica-se que CTR tem 3 nós pais, sendo eles todas as variáveis anteriores e portanto haverá uma seta de cada um desses nós, representando a dependência condicional de CTR com as ouras variáveis, como mostra a Figura 8.

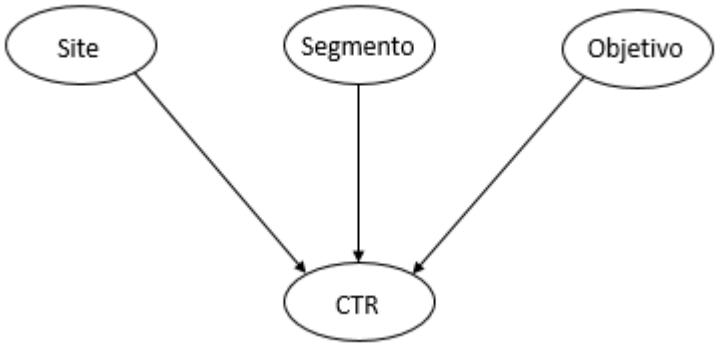


Figura 8 – Representação da rede com suas dependências.

A rede pode parecer simples mas escolhendo apenas essas 4 variáveis, considerando que Segmento tem 10 estados, Objetivo tem 2, CTR tem 4 e Site aproximadamente 1000, a tabela de probabilidade condicional terá $10 * 2 * 4 * 1000 = 40000$ entradas. Por esse motivo decidiu-se por preservar a simplicidade da rede para não prejudicar sua funcionalidade.

Por não possuírem nós pais, as tabelas de probabilidade condicionais das 3 primeiras variáveis são simplesmente suas probabilidade a priori e são distribuídas como segue nas Tabelas 5, 6 e 7.

A Tabela da variável Site é muito longa e por isso não será mostrada completa e sim apenas um pedaço para exemplo.

Objetivo	P(Objetivo)
CPC	0.54
CPA	0.46

Tabela 5 – Probabilidade a priori da variável Objetivo $P(Objetivo)$.

Segmento	P(Segmento)
Educação (EDU)	0.21
Shoppings (SHO)	0.03
Setor Imobiliário (IMO)	0.15
Automóveis (AUT)	0.15
Governo (GOV)	0.08
Vida e Saúde (VID)	0.09
Turismo (TUR)	0.07
Moda (MOD)	0.14
Bancos e Seguros (BAN)	0.08
Entretenimento (ENT)	0.01

Tabela 6 – Probabilidade a priori da variável Segmento $P(Segmento)$.

Site	P(Site)
abril.com.br	0.006
9gag.com	0.004
globo.com	0.007
youtube.com	0.009
google.com	0.010
dailymotion.com	0.012
apple.com	0.008
infojobs.com	0.006
...	...

Tabela 7 – Parte da tabela de Probabilidade a priori da variável Site $P(Site)$.

A tabela de probabilidade condicional da variável CTR terá cerca de 40000 entradas e não pode ser representada facilmente como as tabelas anteriores. A Figura 9 ilustra a estrutura da tabela caso as variáveis dos nós pais fossem todas Booleanas.

Objetivo	Objetivo 1				Objetivo 2			
Segmento	Segmento 1		Segmento 2		Segmento 1		Segmento 2	
Site	Site 1	Site 2	Site 1	Site 2	Site 1	Site 2	Site 1	Site 2
CTR 1	P(CTR=1|Obj=1,Seg=1,Site=1)							
CTR 2								
CTR 3								
CTR 4								

Figura 9 – Ilustração da tabela de probabilidade de CTR caso todas as variáveis pais fossem booleanas.

Nos próximos capítulos será mostrado o método usado para calcular facilmente todas as probabilidades condicionais da variável CTR utilizando o arquivo base de amostras citado anteriormente.

5.4 Conjunto de treinamento

O conjunto de treinamento o e conjunto de teste serão divididos cronologicamente, para assegurar que o modelo faz - ou não - uma boa estimativa do futuro.

O período escolhido para os conjuntos foram janeiro de 2014 até junho de 2016 para o conjunto de treinamento e julho de 2016 até dezembro de 2016 para o conjunto de teste.

Com os dados de treinamento separados, criou-se uma base de dados com cerca de 30000 eventos atômicos, ou exemplos, e aproximadamente 3000 sites diferentes.

5.5 Aplicação da Regra de Bayes

Para os requisitos do trabalho será preciso calcular todas as probabilidades condicionais envolvendo as quatro variáveis da Rede modelada. Para exemplificar, uma entrada da TPC seria o cálculo da probabilidade mostrada na Equação 5.1. Em palavras, esse cálculo traz a probabilidade do CTR estar na faixa 1, visto que foi observado que o eventos Segmento = EDU, Objetivo = cpc e Site = abril.com.br ocorreram simultaneamente. Para facilidade no espaçamento das equações, as variáveis Segmento e Objetivo serão abreviadas para Seg e Obj, respectivamente. O site abril.com.br será abreviado para a.

$$P(CTR = CTR1 | Site = a, Seg = EDU, Obj = cpc) \tag{5.1}$$

Utilizando a regra de bayes, o cálculo dessa probabilidade se dá pela Equação 5.2

$$\frac{P(Site = a, Seg = EDU, Obj = cpc | CTR = CTR1) * P(Site = a, Seg = EDU, Obj = cpc)}{P(CTR = CTR1)} \tag{5.2}$$

Analisando cada termo da regra de bayes:

- $P(Site = a, Seg = EDU, Obj = cpc | CTR = CTR1)$ corresponde a probabilidade desse evento atômico, dado o fato de que CTR = CTR1 ocorreu. Sendo N o número de amostras nas quais o CTR está na classe 1, e E o número de eventos atômicos nos quais (Site = a, Seg = EDU, Obj = CPC, CTR = 1) então esse termo pode ser calculado por $\frac{E}{N}$.

- $P(Site = a, Seg = EDU, Obj = cpc)$ corresponde a probabilidade dessas variáveis nesses estados ocorrerem simultaneamente. Seja T o número de amostras totais, e F o número de amostras nas quais Site = a, Seg = EDU e Obj = CPC ocorrem simultaneamente, então esse termo pode ser calculado por $\frac{F}{T}$.

- $P(CTR = CTR1)$ corresponde a probabilidade da variável CTR ser da classe 1. Seja T o número de amostras totais e G o número de amostras nas quais CTR =1, esse termo pode ser calculado por $\frac{G}{T}$.

Calculado cada termo, a Equação 5.2 é reescrita por:

$$\frac{(0.0004) * (0.4582)}{0.0018} = 0.1132$$

Ou seja, a probabilidade do CTR estar na faixa 1, dado que o site escolhido foi o *abril.com.br*, o objetivo escolhido foi CPC e o segmento escolhido foi o de Educação, é de 11.32%.

5.6 Inferência Bayesiana

Como pode-se perceber pelo número de probabilidade possíveis, há uma certa intratabilidade em realizar essas operações cerca de 30000 vezes para completar todas as entradas desta TPC. Por essa razão um método de inferência aproximada será utilizado.

Ao invés de completar a TPC com todas as probabilidades possíveis, serão produzidas apenas amostras que dizem respeito a probabilidade condicional que se está tentando calcular, eliminando assim momentaneamente todas aquelas outras amostras que não possuem informações relevantes para o cálculo. Esse método é chamado de amostragem de rejeição em redes Bayesianas.

> Note que a amostragem de rejeição é muito semelhante à avaliação de probabilidades condicionais diretamente do mundo real. Por exemplo para estimar $P(Chuva|CeuVermelhoNoite = verdadeiro)$, pode-se simplesmente contar com que frequência chove depois que se observa um céu vermelho na noite anterior - ignorando-se as noites em que o céu não está vermelho. (Aqui, o próprio mundo desempenha o papel do algoritmo de geração de amostras.) É óbvio que isso poderia tomar um longo tempo, se o céu só muito raramente ficasse vermelho, e essa é a deficiência da amostragem de rejeição [15].

Usando esse método, para calcular as probabilidades de cada evento atômico.

Seja

$$P(CTR = c|Obj = O, Seg = S, Site = St) \tag{5.3}$$

a probabilidade condicional do CTR estar na faixa c, dado que o evento Objetivo ocorrido é O, o Segmento ocorrido é S e o Site ocorrido é St.

Pela regra do produto pode-se escrever a Equação 5.3 como:

$$\frac{P(CTR = c, Obj = O, Seg = S, Site = St)}{P(Obj = O, Seg = S, Site = St)} \tag{5.4}$$

Seja c' todos os c possíveis de CTR, logo, a equação 5.4 pode ser calculada por:

$$\frac{P(CTR = c, Obj = O, Seg = S, Site = St)}{\sum_{c'}(CTR = c', Obj = O, Seg = S, Site = St)} \tag{5.5}$$

Esse será o método de cálculo utilizado para todas as probabilidades condicionais envolvendo o domínio do problema, o qual será incorporado a um sistema para o cálculo automático desses valores. A criação e implementação do sistema serão explicados nas seções seguintes do trabalho.

A utilização deste método traz uma questão muito importante: qual o número de exemplos suficiente para ser possível chegar a uma hipótese válida sobre cada evento atômico? Por exemplo, se um evento ocorreu apenas uma vez e se tentar inferir uma probabilidade acerca de seu estado, 100% das vezes ele estará naquele estado e isso raramente será verdade.

5.6.1 Espaço de amostras

Segundo Russel e Norvig [15], seja X o conjunto de todos os exemplos possíveis, D a distribuição sobre a qual os exemplos são extraídos, H o conjunto de hipóteses possíveis e seja N o número de exemplos no conjunto de treinamento.

Uma hipótese h é dita aproximadamente correta se erro(h)$\leq \epsilon$, onde ϵ é uma constante pequena. Para prosseguir, é de interesse mostrar que após ver N exemplos, com alta probabilidade, todas as hipóteses consistentes serão aproximadamente corretas.

Imaginando o conjunto de hipóteses possíveis como um retângulo e a função verdadeira como um ponto dentro deste espaço, uma hipótese aproximadamente correta poderia ser qualquer hipótese pertencente a uma circunferência de raio ϵ em torno da função verdadeira, como mostra a Figura 10. Todo o espaço fora da circunferência mas dentro do retângulo, será chamado então de H_{ruim} e, como o nome diz, representa o conjunto de hipóteses ruins.

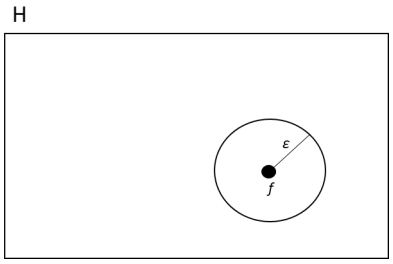


Figura 10 – Representação do espaço de hipóteses.

Pode-se calcular a probabilidade de uma hipótese seriamente errada $h_b \in H_{ruim}$ ser consistente com os primeiros N exemplos como segue. Sabe-se que erro$(h_b) > \epsilon$ e assim que a probabilidade de que ela concorde com um dado exemplo é no mínimo 1 - ϵ. O limite para N exemplos é:

P(h_b concorda com N exemplos)$\leq (1-\epsilon)^N$.

A probabilidade de H_{ruim} conter no mínimo uma hipótese consistente é limitada pelas probabilidades individuais:

P(H_{ruim} contém uma hipótese consistente $\leq |H_{ruim}| \leq (1-\epsilon)^N \leq H|1-\epsilon)^N$, onde usa-se o fato de que $|H_{ruim}| \leq |H|$. O que se busca então é reduzir a probabilidade desse evento até ficar abaixo de algum número pequeno δ:

$H|1-\epsilon)^N \leq \delta$.

Dado que $1-\epsilon \leq e^{-\epsilon}$, pode-se conseguir isso se for permitido que o algoritmo veja:

$$N \geq \frac{1}{\epsilon}(ln\frac{1}{\delta} + ln|H|) \tag{5.6}$$

Desse modo, se um algoritmo de aprendizagem retorna uma hipótese consistente com essa quantidade de exemplos, então com probabilidade no mínimo $1-\delta$, ele tem erro máximo ϵ.

Para determinar H faz-se 2^n, sendo n o número de combinações possíveis com as variáveis do domínio.

Agora, pode-se calcular o tamanho N da amostra mínima para a problemática do trabalho, escolhendo um valor ϵ pequeno, um valor δ pequeno fazendo o cálculo de H. Deseja-se encontrar um N para cada situação mostrada na Equação 5.5, logo H será calculado considerando-se que existem 4 (4 faixas de CTR) combinações possíveis para cada situação pré-definida.

Seja $\epsilon = 0.1$ e $\delta = 0.05$ valores aceitáveis e $H = 2^4$, utilizando a Equação 5.6:

$$N \geq \frac{1}{0.1}(ln\frac{1}{0.05} + ln|2^4|) \geq 57.6$$

Será preciso então cerca de 57 amostras para cada situação na qual Seg = S, Obj = O e Site = St para ter uma hipótese válida, com erro $\epsilon \leq 0.1$.

5.7 Sistema

Criar um sistema que mostre todos os eventos atômicos em um só lugar, ao mesmo tempo, além de inviável pode se tornar uma ferramenta de difícil utilização e então não servir o seu propósito de apoio à decisão. Por esse motivo, decidiu-se por criar um sistema que calcula as probabilidades *on demand*, ou seja, as probabilidades a serem calculadas serão escolhidas pelo agente utilizando o sistema.

Por facilidade de manuseio optou-se por utilizar também Planilhas Google para criação do sistema.

A principal funcionalidade que o sistema deve apresentar é o aprendizado independente ao passo que sua base de exemplos é esporadicamente atualizada, ou seja, não deve haver nenhum retrabalho ou interferência cada vez que a base é alimentada.

A primeira etapa então é criar uma interface simples para que o agente possa escolher para qual o tipo de campanha se deseja calcular as probabilidades. Não será limitado o site, podendo assim o agente ver dentro daquele tipo de campanha as probabilidades de cada site em uma lista. A interface deve também se adaptar aos novos tipos de segmentos que podem surgir.

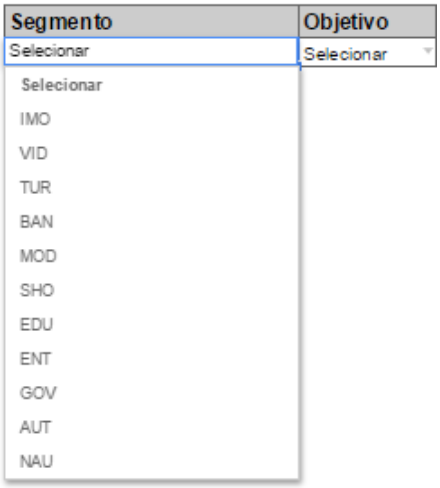


Figura 11 – Interface para a escolha das variáveis.

Com a interface pronta, pode-se então calcular o número de amostras N específicas para a combinação escolhida pelo agente. Já que os Sites se repetem para cada faixa de CTR decidiu-se por não posicioná-los na horizontal como de costume na TPC, visto que

isso criaria uma tabela com cerca de 12000 colunas e 4 linhas. Ao invés, posicionou-se os sites na vertical e as faixas de CTR na horizontal, criando assim então uma tabela com 4 colunas e cerca de 3000 linhas, reduzindo drasticamente o número de entradas. A listagem dos sites também deve se adaptar a novos sites que podem surgir na base de exemplos.

Após calculado o número de amostras para casa Site, para cada faixa de CTR, pode-se calcular a frequência de cada entrada e então calcular as probabilidades como especificado na Equação 5.5.

	0 <= CTR < 0,03	0,03 <= CTR < 0,06	0,06 <= CTR < 0,1	CTR >= 0,1		
site_domain					Total	Total geral
acasadocogumelo.com	0	0	0	0	0	1
accelerated-ideas.com	0	0	0	0	0	1
accuradio.com	0	0	0	0	0	2
accuweather.com	5	7	3	3	18	136
acessaber.com.br	0	0	0	0	0	1
acesseeveja.com	0	0	0	0	0	3
acheconcursos.com.br	0	0	0	1	1	4
acicri.com.br	2	2	0	1	5	13
acidezfeminina.com.br	0	0	0	0	0	7
aconteceuemjaragua.com.br	1	0	0	0	1	3
acordacidade.com.br	0	0	0	0	0	4
acrediteounao.com	0	0	0	0	0	6
acritica.uol.com.br	0	0	0	0	0	1
acsta.net	0	0	0	0	0	1
actugaming.net	0	0	0	0	0	3
ad.gamesow.com	0	0	0	0	0	1
addictinggames.com	0	0	0	2	2	6

Figura 12 – Tabela de cálculo do número de amostras de acordo com as variáveis escolhidas na interface.

Para evitar operações desnecessárias que podem deixar o sistema mais lento, acrescentou-se uma etapa de filtro nas amostras, continuando as operações somente com aqueles sites que cumprem a exigência mínima no número de amostras, calculado na seção anterior. Dessa maneira, calcula-se as frequências e probabilidades somente daqueles Sites sobre os quais pode-se obter hipóteses válidas.

Após a criação e validação do sistema, deverá ser criado um sistema especialista simples que, a partir das observações de probabilidades, gera *Whitelists* e *Blacklists* para cada tipo de campanha.

6 Resultados

6.1 Validação do Sistema

Usando diversas fórmulas e funcionalidades das planilhas Google, as quais estão evidenciadas nas Figuras 13 a 20, foi possível criar o sistema com todos os requisitos necessários, o qual traz as probabilidades de cada site, para cada faixa de CTR.

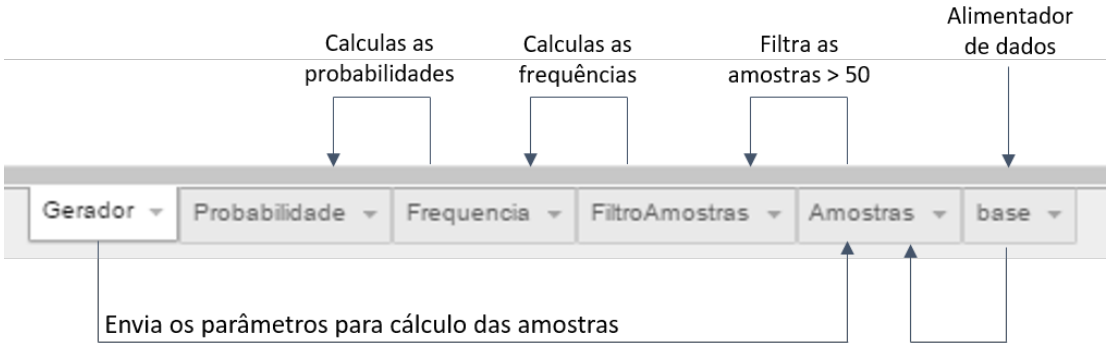
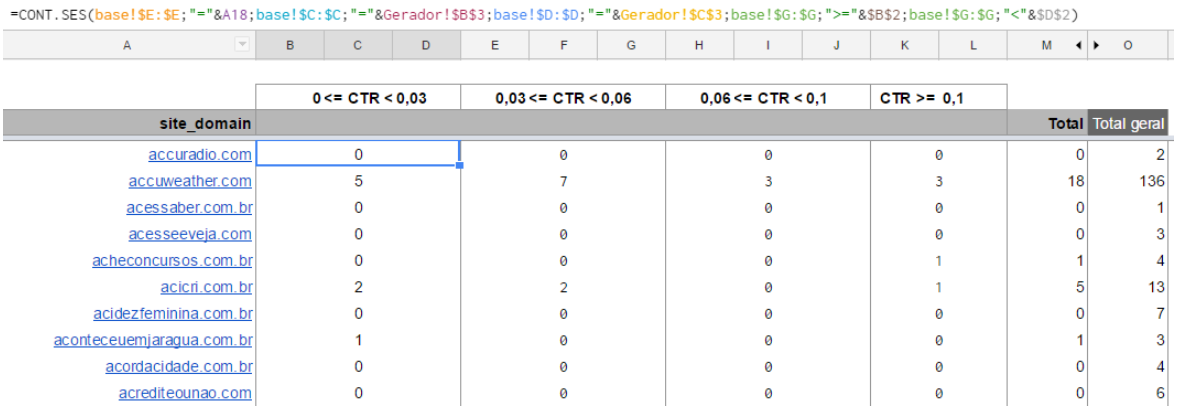


Figura 13 – Esqueleto principal do sistema criado.

intervalo	nome_cliente	Segmento	Objetivo	site_domain	CTR
CTD	[illegible] - SHO	SHO	cpc	abril.com	0,0758
CTD	[illegible] - SHO	SHO	cpc	clicrbs.com.br	0,0602
CTD	[illegible] - SHO	SHO	cpc	facebook.com	0,0272
CTD	[illegible] - SHO	SHO	cpc	freshome.com	0,0000
CTD	[illegible] - SHO	SHO	cpc	globo.com	0,0579
CTD	[illegible] - SHO	SHO	cpc	grooveshark.com	0,0221
CTD	[illegible] - SHO	SHO	cpc	ig.com.br	0,0656
CTD	[illegible] - SHO	SHO	cpc	mediashakers.co	0,0225
CTD	[illegible] - SHO	SHO	cpc	omelete.com.br	0,0000
CTD	[illegible] - SHO	SHO	cpc	playblasteroids.c	0,0000
CTD	[illegible] - SHO	SHO	cpc	terra.com.br	0,0692
CTD	[illegible] - SHO	SHO	cpc	uol.com.br	0,0628
CTD	[illegible] SHO	SHO	cpc	123rede.com	0,0382
CTD	[illegible] - SHO	SHO	cpc	90min.com	0,0447
CTD	[illegible] - SHO	SHO	cpc	9gag.com	0,0072
CTD	[illegible] - SHO	SHO	cpc	9gag.com	0,0338
CTD	[illegible] - SHO	SHO	cpc	9gag.com	0,0630
CTD	[illegible] SHO	SHO	cpc	abc.es	0,0000
CTD	[illegible] - SHO	SHO	cpc	abril.com	0,0758
CTD	[illegible] SHO	SHO	cpc	abril.com.br	0,0803
CTD	[illegible] - SHO	SHO	cpc	abril.com.br	0,0000
CTD	[illegible] - SHO	SHO	cpc	abril.com.br	0,0333
CTD	[illegible] - SHO	SHO	cpc	accuweather.com	0,0726
CTD	[illegible] - SHO	SHO	cpc	accuweather.com	0,1476
CTD	[illegible] - SHO	SHO	cpc	accuweather.com	0,0481

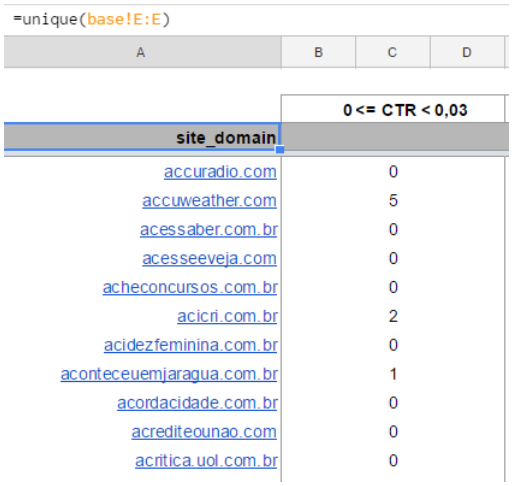
Figura 14 – Base de dados do sistema.



=CONT.SES(base!$E:$E;"="&A18;base!$C:$C;"="&Gerador!$B$3;base!$D:$D;"="&Gerador!$C$3;base!$G:$G;">="&$B$2;base!$G:$G;"<"&$D$2)

site_domain	0 <= CTR < 0,03	0,03 <= CTR < 0,06	0,06 <= CTR < 0,1	CTR >= 0,1	Total	Total geral
accuradio.com	0	0	0	0	0	2
accuweather.com	5	7	3	3	18	136
acessaber.com.br	0	0	0	0	0	1
acesseeveja.com	0	0	0	0	0	3
acheconcursos.com.br	0	0	0	1	1	4
acicri.com.br	2	2	0	1	5	13
acidezfeminina.com.br	0	0	0	0	0	7
aconteceuemjaragua.com.br	1	0	0	0	1	3
acordacidade.com.br	0	0	0	0	0	4
acrediteounao.com	0	0	0	0	0	6

Figura 15 – Aba de amostras do sistema.



=unique(base!E:E)

site_domain	0 <= CTR < 0,03
accuradio.com	0
accuweather.com	5
acessaber.com.br	0
acesseeveja.com	0
acheconcursos.com.br	0
acicri.com.br	2
acidezfeminina.com.br	0
aconteceuemjaragua.com.br	1
acordacidade.com.br	0
acrediteounao.com	0
acritica.uol.com.br	0

Figura 16 – Aba de amostras do sistema com detalhe para a lista de sites.

Foi observado que poucos Sites cumprem o requisito da amostragem, por isso decidiu-se por aumentar δ para 0.1, aumentando um pouco a probabilidade de uma hipótese ruim estar contida nas N amostras escolhidas mas aumentando assim a quantidade de Sites que poderão ser analisados de imediato. Com o novo δ a nova quantidade de amostras N passa a ser:

$$N \geq \frac{1}{0.1}(ln\frac{1}{0.01} + ln|2^4|) \geq 50.7$$

diminuindo para 50 o número de amostras necessárias.

Para checar os resultados, será comparado o valor da probabilidade calculada pelo sistema com a probabilidade calculada usando diretamente a regra de Bayes na Seção 5.5.

Relembrando, a probabilidade calculada foi de 11.32%.

O site *abril.com.br*, para o segmento de educação e objetivo CPC possui N amostras suficientes, como pode ser observado na Figura 19 e então foi possível calcular uma hipótese

=filter(Amostras!A3:O36104;Amostras!M3:M36104 >50)

site_domain	0 <= CTR < 0,03	0,03 <= CTR < 0,06	0,06 <= CTR < 0,1	CTR >= 0,1	Total	Total geral
abril.com.br	6	12	33	2	53	184
dailymotion.com	45	4	2	2	53	360
youtube.com	10	7	35	0	52	271

Figura 17 – Aba do filtro das amostras maiores que N.

=SEERRO(FiltroAmostras!B5/FiltroAmostras!O5;0)

site_domain	0 <= CTR < 0,03	0,03 <= CTR < 0,06	0,06 <= CTR < 0,1	CTR >= 0,1	Total
abril.com.br	0,0326	0,0652	0,1793	0,0109	0,29
dailymotion.com	0,1250	0,0111	0,0056	0,0056	0,15
youtube.com	0,0369	0,0258	0,1292	0,0000	0,19

Figura 18 – Aba de frequências.

válida para

$$P(CTR = CTR11|Seg = EDU, Obj = CPC, Site = abril.com.br)$$

a qual pode ser observada na Figura 20.

site_domain	0 <= CTR < 0,03	0,03 <= CTR < 0,06	0,06 <= CTR < 0,1	CTR >= 0,1	Total	Total geral
abril.com.br	6	12	33	2	53	184

Figura 19 – Linha na tabela de amostras referente ao site *abril.com.br*.

	P(CTR=1 |Site,Seg = EDU, Obj = CPC)	P(CTR=2 |Site, Seg = EDU, Obj = CPC)	P(CTR=3 |Site, Seg = EDU, Obj = CPC)	P(CTR=4 |Site,Seg = EDU, Obj = CPC)
	0 <= CTR < 0,03	0,03 <= CTR < 0,06	0,06 <= CTR < 0,1	CTR >= 0,1
abril.com.br	11,32%	22,64%	62,26%	3,77%

Figura 20 – Linha na tabela de probabilidades referente ao site *abril.com.br*.

Pode-se comprovar então que o sistema, com o método da amostragem, chegou a um resultado igual ao da aplicação direta da regra de Bayes, validando assim o método utilizado.

6.2 Aprendizagem MAP

Após calculadas as probabilidades será preciso validar a hipótese gerada a partir do conjunto de treinamento com o conjunto de teste que foi separado anteriormente. Porém,

o sistema calculou a probabilidade para cada classe de CTR. Como validar então os erros e acertos das hipóteses?

Uma aproximação muito comum, habitualmente utilizada na ciência, é fazer previsões com base em uma única hipótese mais provável, isto é, uma h_i que maximiza $P(h_i|d)$. Com muita frequência, essa aproximação é chamada de hipótese de máximo a posteriori ou MAP [15]. Denomina-se h_{map} como a hipótese com máxima probabilidade a posteriori e:

$$h_{map} = argmax_{h \in H} P(h|D)$$

onde D representa a base de dados utilizada para treinamento.

Essa aproximação é fácil de se entender por ser muito intuitiva. Quando, por exemplo, um agente se depara com uma escolha na qual há 90% de sucesso e 10% de derrota, uma aposta fácil seria na hipótese de sucesso. O mesmo será feito com as probabilidades calculadas para cada evento atômico, tomando como hipótese única a mais provável dentre as 4 classes de CTR.

No exemplo da Figura 20, h_{MAP} = CTR3, pois essa é a hipótese com maior probabilidade dentre as 4 classes. Espera-se então que a maioria dos dados no conjunto de teste para esse evento, estejam na faixa de CTR 3.

Dessa maneira, será possível quantificar e qualificar os acertos do modelo criado.

6.2.1 Avaliação do conhecimento

É de extrema importância que se valide o modelo criado na classificação dos CTR. Para classificações com variáveis discretas é costumeiro se utilizar matrizes de confusão e conceitos de Falso Positivo (FP), Falso Negativo (FN), Positivo Verdadeiro (PV) e Negativo Verdadeiro (NV). Como a classificação das 4 classes de CTR não é Booleana, ou seja, não será considerada positiva ou negativa somente, será usado apenas a medida de qualidade Positivo Verdadeiro, tendo então:

- Positivo Verdadeiro (PV): Número de exemplos corretamente classificados.
- Falso Positivo (FP): Número de exemplos que estavam em qualquer faixa de CTR acima daquela classificada pelo modelo.
- Falso Negativo (FN): Número de exemplos que estavam em qualquer faixa abaixo daquela classificada pelo modelo.

A partir dessas medidas cria-se uma matriz de confusão, que representa a distribuição dos resultados obtidos pela validação do modelo com o conjunto de teste. Cada

coluna da matriz representa as instâncias preditas de uma classe enquanto cada linha representa as instâncias reais de uma classe. Assim a diagonal principal representa os valores que foram preditos corretamente, as diagonais acima da diagonal principal representam os valores que foram preditos incorretamente e que são Falsos Positivos e as diagonais abaixo da diagonal principal representam os valores que foram preditos incorretamente e que são Falsos Negativos [23].

Para facilitar a visualização e análise dos resultados, serão registrados os resultados para os segmentos de maior presença dentro da empresa. As matrizes de confusão construídas levam em consideração resultados de CPC e CPA para cada segmento e apenas para os sites que atingiram o número mínimo de amostras necessárias.

	CTR1 Previsto	CTR2 Previsto	CTR3 Previsto	CTR4 Previsto
CTR1 real	**30**	1	0	2
CTR2 real	4	**15**	1	0
CTR3 real	2	0	**9**	0
CTR4 real	0	4	1	**11**

Tabela 8 – Matriz de confusão IMO

	CTR1 Previsto	CTR2 Previsto	CTR3 Previsto	CTR4 Previsto
CTR1 real	**16**	0	0	0
CTR2 real	0	**7**	0	0
CTR3 real	3	0	**0**	2
CTR4 real	0	3	0	**18**

Tabela 9 – Matriz de confusão SHO

	CTR1 Previsto	CTR2 Previsto	CTR3 Previsto	CTR4 Previsto
CTR1 real	**121**	3	3	2
CTR2 real	25	**20**	0	4
CTR3 real	5	5	**5**	16
CTR4 real	0	3	3	**92**

Tabela 10 – Matriz de confusão EDU

	CTR1 Previsto	CTR2 Previsto	CTR3 Previsto	CTR4 Previsto
CTR1 real	**40**	0	0	0
CTR2 real	12	**0**	0	0
CTR3 real	6	0	**2**	3
CTR4 real	0	0	0	**22**

Tabela 11 – Matriz de confusão AUT

Destas matrizes de confusão, se está interessado em calcular sua acurácia, ou seja, a porcentagem de amostras corretamente classificadas sobre a soma de todas as amostras e seu cálculo é descrito pela equação:

$$Acurácia = \frac{PV}{PV + FP + FN}$$

Para as matriz representadas pelas Tabelas 8, 9, 10 e 11 e suas acurácias podem então ser calculadas por:

$$Acurácia \quad Segmento \quad IMO = \frac{65}{65 + 4 + 11} = 0.8125$$

$$Acurácia \quad Segmento \quad SHO = \frac{41}{41 + 2 + 6} = 0.8723$$

$$Acurácia \quad Segmento \quad EDU = \frac{238}{238 + 12 + 41} = 0.8178$$

$$Acurácia \quad Segmento \quad AUT = \frac{64}{64 + 3 + 18} = 0.7529$$

Além disso, é importante calcular a porcentagem de classificações FP pois estas podem ser mais prejudiciais do que FN já que FP podem levar a uma "aposta"errada em um site, levando a um gasto de verba no site errado, enquanto um FN apenas priva esse gasto.

Para as acurácias calculadas acima, será calculada a porcentagem de FP dentre as classificações errôneas da forma:

$$\%FP = \frac{FP}{FP + FN} * 100$$

Para as mesmas matrizes, as porcentagens de FP são, respectivamente 26.66%, 25%, 22.65% e 14.28%. Esses resultados são positivos pois comprova que o modelo, quando errado, tende a classificar mais FN do que FP.

Para uma ideia mais geral, criou-se também uma matriz de confusão geral, dos 4 segmentos juntos, a qual segue abaixo:

	CTR1 Previsto	CTR2 Previsto	CTR3 Previsto	CTR4 Previsto
CTR1 real	**207**	4	3	4
CTR2 real	41	**42**	1	4
CTR3 real	16	5	**16**	21
CTR4 real	0	10	4	**143**

Tabela 12 – Matriz de confusão Geral

$$Acurácia \quad Geral = \frac{408}{408 + 37 + 76} = 0.7831$$

$$\%FP = \frac{37}{37 + 76} * 100 = 32.74\%$$

6.2.2 Curva de aprendizagem

Para avaliar o aprendizado do modelo, se utilizará a metodologia exposta na Seção 2.5, na qual separa-se o conjunto de treinamento em porções aleatoriamente e se contabiliza a porção de acertos no conjunto de teste, para cada tamanho de amostra.

Realizando o mesmo procedimento das matrizes de confusão para os 4 principais segmentos da empresa, chegou-se na curva de aprendizado abaixo:

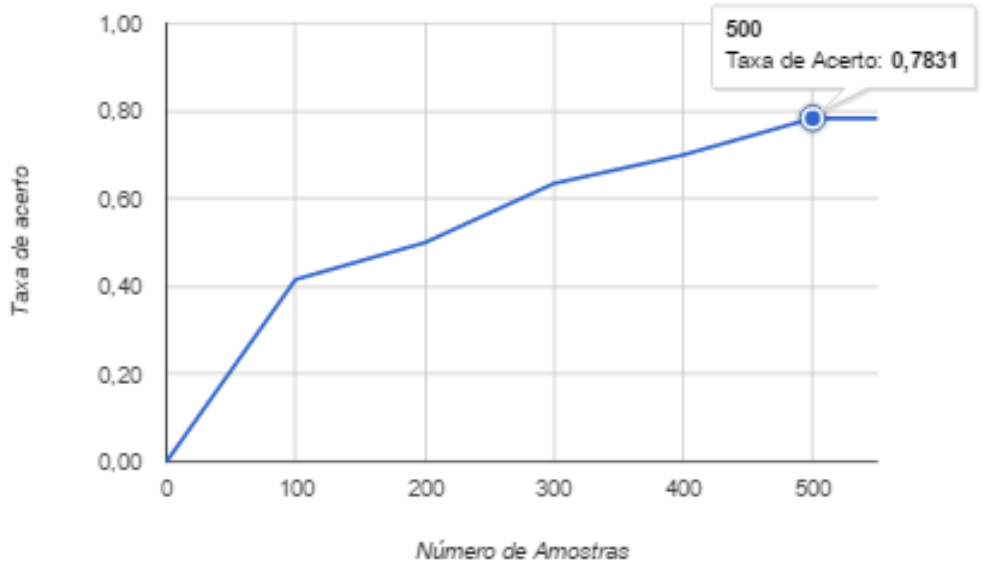


Figura 21 – Curva de aprendizado geral.

Observando a curva de aprendizagem é nítido que a grande quantidade de amostras é indispensável para a validação do modelo.

6.3 Geração de Whitelists e Blacklists

As probabilidades calculadas no sistema já podem ser usadas no apoio à decisão na hora de gerar *Whitelist* e *Blacklists*. Para facilitar esse processo, foi criado um sistema

especialista simples, baseado no tipo de ações costumeiras executadas pelos gestores das campanhas.

Na mesma interface na qual o agente escolhe o tipo de campanha e objetivo, será possível também escolher entre gerar uma *Whitelists*, uma *Whitelist Premium* ou uma *Blacklist*. Cada uma dessas opções seleciona sites da seguinte maneira:

- Whitelist: Sites com probabilidade somada maior que 50% de estar nos CTRs de faixa 3 e 4.
- Whitelist Premium: Sites com probabilidade maior que 50% de estar no CTR de faixa 4.
- Blacklist: Sites com probabilidade somada maior que 50% de estar nos CTRs de faixa 1 e 2.

Figura 22 – Interface para a geração de listas.

Com a lista gerada, o agente pode decidir, baseado nos cálculos de probabilidade se "aposta"ou não em algum outro site que não está na lista.

6.4 Análise dos resultados

Foi observado que muitos sites rodaram um número de vezes menor do que o estipulado para se poder tirar alguma conclusão acerca dos mesmos, impossibilitando momentaneamente a análise de todos os sites. Porém, como o sistema foi eficazmente projetado para aprender ao passo que mais dados são alimentados à base, tais sites eventualmente atingirão o número de amostras e poderão ser analisados.

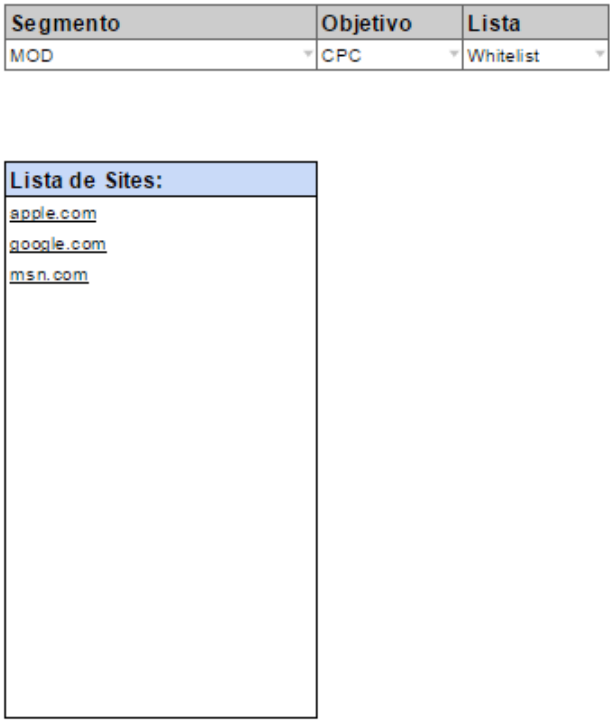
Segmento	Acurácia	%FP
EDU	0.8178	22.65
IMO	0.8125	26.66
AUT	0.7529	14.28
SHO	0.8723	25.00
Geral	0.7831	32.74

Tabela 13 – Comparação de Resultados por Segmento

Comparando os resultados por segmento, percebe-se que o segmento com a maior acurácia é o setor de Shoppings. Isso pode ser explicado pelo fato de que as campanhas deste segmento tendem a ser muito parecidas umas com as outras, o que leva a um padrão maior nos resultados.

No geral, os resultados são satisfatórios. As taxas de acerto em torno de 80% são relativamente altas, levando em consideração que existem muitas outras variáveis acerca de uma campanha que não podem ser medidas.

Comparando a Figura 23 com a Figura 24 pode-se perceber que dois segmentos distintos geram duas listas distintas. E até mesmo um mesmo segmento mas com objetivos diferentes também geram listas distintas, provando assim a dependência de CTR para com as variáveis de seus nós pais.



Segmento	Objetivo	Lista
MOD	CPC	Whitelist

Lista de Sites:
apple.com
google.com
msn.com

Figura 23 – Whitelist para Segmento EDU e Objetivo CPA.

Segmento	Objetivo	Lista
EDU	CPA	Whitelist

Lista de Sites:
apple.com
blogspot.com.br
globo.com
google.com
infoescola.com
outlook.com
youtube.com

Figura 24 – Whitelist para Segmento EDU e Objetivo CPA.

Pelas curvas de aprendizado também foi possível comprovar que há realmente um aprendizado cada vez maior à medida que mais amostras são apresentadas ao sistema.

Comprovado que pode-se confiar nos resultados obtidos pelo sistema, compara-se alguns sites contidos nas *Whitelist* existentes na empresa com os resultados destes sites gerados pelo sistema. Tomando como exemplo a *Whitelist* Imóveis, alguns sites que a compõe como *olx.com.br*, *infoimoveis.com.br* e *r7.com* foram classificados pelo sistema com alta probabilidade de terem seus CTRs nas faixas 1 e 2 em campanhas de imóveis, tanto CPC quanto CPA, como pode ser observado na Figura 25.

Figura 25 – Sites classificados para a *Blacklist* que atualmente estão na *Whitelist* da empresa.

Essa comparação já torna o sistema de grande valia para a empresa pois traz a informação de que se está investindo parte da verba do cliente em sites que se imaginou ter boa performance mas que foi comprovado que na maioria das vezes tende a ter uma performance ruim.

Colocando números nessa comparação: Se uma *Whitelist* de imóveis tem cerca de

15 sites, se cerca de 40% da verba do cliente é destinado à *Whitelist* e o investimendo médio de um cliente (no ano de 2016) foi de R$ 6.000,00. Eliminar esses 3 sites da lista de sites utilizados na campanha poupa cerca de R$ 480,00 que estavam sendo investidos em sites de baixa performance. Em um ano, essa verba mal alocada representa um montante de R$ 35.520,00, para apenas um segmento. A Tabela 14 estrutura estes dados para melhor entendimento.

Número de sites na *Whitelist* Imóveis	15
Número de sites equivocados na lista	3
Campanhas do setor de imóveis veiculadas em 2016	74
% Verba da campanha destinada à *Whitelist*	40%
Investimento médio por campanha (2016)	R$6.000,00
Economia por campanha	R$480,00
Economia em um ano	**R$35.520,00**

Tabela 14 – Valores para sites mal alocados em campanhas de imóveis.

7 Conclusões

Utilizando o sistema criado, a geração de listas de sites passará então a ser baseada em um conhecimento adquirido, processado e analisado, tendo uma base sólida para decisões e não mais achismos. Isso gera mais credibilidade da empresa com o cliente e se torna um diferencial dentre os concorrentes, além de poupar horas de trabalho manual para a mesma função. Mesmo que a lista de sites acabe por não ser muito extensa, se terá a certeza de que os sites escolhido trarão qualidade e não apenas quantidade, o que vai de encontro com as tendências dos anunciantes atuais, que buscam performance cada vez melhores e não entregas de quantidade mas sem qualidade. O sistema também traz uma base sólida para que os gestores de campanha possam tomar suas próprias decisões acerca da lista de sites.

Foi possível perceber que há uma tendência nas performances dos sites para cada tipo de campanha mas que nem todas as campanhas similares performam sempre da mesma maneira, o que já era esperado. Muitas outras variáveis estão envolvidas em todo o processo da compra programática de mídias e muitas dessas variáveis não estão disponíveis para análise. A escolha de um modelo probabilístico com certeza trouxe muitas vantagens para o trabalho, bem como a escolha de se usar Redes Bayesianas para modelar o domínio, já que essa permitirá a constante atualização das probabilidades nos nós, à medida que novas informações são alimentadas à base, e informação é o que nunca falta nesse mercado de mídias digitais.

Trabalhar com uma base grande de dados é sempre um desafio, tanto na parte de tratamento quanto no entendimento do que de fato aqueles dados significam no contexto da empresa, não sendo diferente na realização deste trabalho. Demorou-se muito tempo para definir o problema e qual seria a melhor forma de tratá-lo, trazendo um resultado que fizesse sentido tanto para a empresa quanto para as abordagens matemáticas e científicas escolhidas.

A escolha pelo aprendizado MAP facilitou a validação do modelo com o conjunto de teste, mesmo sendo tão radical, aproximando-se da realidade de um agente escolhendo dentre aquelas classes, em qual "apostar".

Demonstrou-se também ao fim do trabalho que muitas das decisões tomadas baseadas em intuições podem ser equivocadas e fazer uma grande diferença no andamento de campanhas. Por isso observa-se cada vez mais o marketing digital trazendo conhecimento de outras áreas como a de tecnologia para se fazer aproveitar dos valiosos dados que pairam sobre as operações deste mercado e que muitas vezes não são utilizados.

Nesse contexto então, percebe-se que a engenharia tem muito a acrescentar em

diversos tipos de mercado e diferentes áreas de trabalho, basta apenas um olhar crítico e uma busca constante pela inovação.

8 Trabalhos Futuros

O trabalho mostrou a necessidade da empresa em criar um banco dados para que as análises possam ser extraídos de forma mais rápida e organizada. Apesar de não ser o tema do trabalho, foi possível identificar essa necessidade ao longo do desenvolvimento do mesmo.

Para trabalhos futuros relacionados com o presente tema, seria de grande valia o estudo de mais variáveis acerca das campanhas e como cada uma delas influencia nos resultados das mesmas, trabalhando em acrescentar cada vez mais nós na Rede já criada.

Referências

1 OLIVEIRA, P. *Carência de Profissionais de Tecnologia no Marketing é Alta*. 2017. Https://www.mundodomarketing.com.br/reportagens/mercado/37124/carencia-de-profissionais-de-tecnologia-no-marketing-e-alta. Acessado em 26/01/2017. Citado na página 17.

2 BOBBIOA, A. et al. Improving the analysis of dependable systems by mapping fault trees into bayesian networks. *Elsevier*, v. 1, 2001. Citado na página 18.

3 SOCIAL, W. A. *Digital, Social e Mobile 2015 – Um Compilado de Dados e Tendências Digitais*. 2015. Http://www.b9.com.br/54482/social-media/digital-social-e-mobile-2015-um-compilado-de-dados-e-tendencias-digitais/. Acessado em 20/10/2016. Citado na página 18.

4 MEDIAMATH. *Fourth Source: Your Cross-Device Identification Questions Answered*. 2016. Https://www.mediamath.com/news/fourth-source-your-cross-device-identification-questions-answered/. Acessado em 26/01/2017. Citado na página 18.

5 GOOGLE. *Definição de Marketing*. 2016. Https://www.google.com/webhp?sourceid=chrome-instantion=1espv=2ie=UTF-8q=marketing+definiAcessado em 20/12/2016. Citado na página 21.

6 WSI. *Digital Minds - 12 Estratégias de Marketing Digital que toda Empresa Precisa Conhecer*. [S.l.]: ABOVE Publicações, 2016. Citado 2 vezes nas páginas 21 e 22.

7 MUNDIAL, G. B. *Relatório sobre o Desenvolvimento Mundial 2016: Dividendos Digitais*. 2016. Http://documents.worldbank.org/curated/pt/788831468179643665/pdf/102724-WDR-WDR2016Overview-PORTUGUESE-WebResBox-394840B-OUO-9.pdf. Citado na página 21.

8 WORLDBANK World Population. 2015. Http://data.worldbank.org/indicator/SP.POP.TOTL. Acessado em 13/12/2016. Citado na página 21.

9 NúMEROS de Investimento em Mídia Online 2015-2016 - AD SPEND BRASIL. 2015. Http://iabbrasil.net/guias-e-pesquisas/mercados. Acessado em 14/12/2016. Citado na página 22.

10 GOVERNO FEDERAL. *Caracterização do cenário macroeconômico para os próximos 10 anos (2016-2025)*. 2016. Http://www.epe.gov.br/mercado/Documents. Citado na página 22.

11 BORGES, L. *Marketing Digital*. 2013. Http://blog.luz.vc/o-que-e/marketing-digital/. Acessado em 17/12/2016. Citado na página 22.

12 INNOVATION, I. *Marketing Digital - Conceitos e Definições*. 2013. Https://www.internetinnovation.com.br/blog/marketing-digital-conceito-e-definicao/. Acessado em 17/12/2016. Citado na página 22.

13 GOOGLE. *Consumer Barometer 2015*. 2015. Citado na página 22.

14 KOZMA, L. *Descubra o Que É Mídia Programática*. 2016. Http://www.the-emagazine.com.br/categoria-92-marketing-online/2709-noticia-descubra-o-que-e-midia-programatica. Acessado em 23/12/2016. Citado na página 25.

15 RUSSEL, S.; NORVIG, P. *Inteligência Artificial*. [S.l.]: Editora Campus, 2004. Citado 7 vezes nas páginas 27, 28, 29, 35, 49, 50 e 58.

16 MURTEIRA, B. *Estatística: Inferência e Decisão*. [S.l.]: Lisboa:Imprensa Nacional, 1988. Citado na página 31.

17 CHAGAS, R. P. *Aplicação de Redes Bayesianas na Previsão de Crescimento de Fluxos de Caixa*. Tese (Doutorado) — Escola de Economia de São Paulo, 2007. Citado 3 vezes nas páginas 31, 33 e 42.

18 SOFTWARE, H. *Introduction to Bayesian Networks*. 2016. Http://hugin.com/wp-content/uploads/2016/05/Building-a-BN-Tutorial.pdf. Acessado em 24/01/2017. Citado na página 33.

19 GONçALVES, A. R. *Redes Bayesianas*. [S.l.], 2016. Acessado em 01/02/2017. Citado na página 34.

20 MEDIAMATH. *Marketing Wiki: Machine Learning*. 2016. Http://blog.mediamath.com/blog/technology/wtf-is-machine-learning/. Acessado em 06/01/2017. Citado 2 vezes nas páginas 39 e 42.

21 SAS. *Machine Learning: O Que É e Porque É Importante?* 2015. Http://www.sas.com/pt$_br$/*insights*/*analytics*/*machine* − *learning.html*.*Acessadoem*09/01/2017.*Citadonapágina*42.

22 WIKIPEDIA. *Redes Bayesianas*. 2015. Https://pt.wikipedia.org/wiki/Rede$_b$*ayesiana*.*Acessadoem*10/01/2017.*Citadonapágina*42.

23 SCHMITT, V. F. *Uma Análise Comparativa de Técnicas de Aprendizagem de Máquina para Prever a Popularidade de Postagens no Facebook*. Tese (Doutorado) — Universidade Federal do Rio Grande do Sul, Dezembro 2013. Acessado em 01/02/2017. Citado na página 59.