UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO TECNOLÓGICO
DEPARTAMENTO DE ENGENHARIA DE PRODUÇÃO E SISTEMAS
CURSO DE GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO CIVIL

Filipe Ferreira Viella

Modelagem de Atribuição de Marketing Digital a partir de Cadeias de Markov

Florianópolis
2023

Filipe Ferreira Viella

Modelagem de Atribuição de Marketing Digital a partir de Cadeias de Markov

Trabalho de Conclusão de Curso submetido ao curso de Engenharia de Produção Civil do Centro Tecnológico da Universidade Federal de Santa Catarina como requisito para a obtenção do título de Engenheiro de Produção Civil.

Orientador: Prof. Ricardo Villarroel Dávalos, Dr.

Florianópolis

2023

Viella, Filipe Ferreira
Modelagem de atribuição de marketing digital a partir de cadeias de Markov / Filipe Ferreira Viella ; orientador, Ricardo Villaroel Dávalos, 2023.
73 p.

Trabalho de Conclusão de Curso (graduação) - Universidade Federal de Santa Catarina, Centro Tecnológico, Graduação em Engenharia de Produção Civil, Florianópolis, 2023.

Inclui referências.

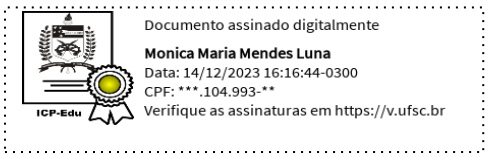
1. Engenharia de Produção Civil. 2. Cadeia de Markov. 3. Modelo de atribuição. 4. Marketing Digital. I. Dávalos, Ricardo Villaroel. II. Universidade Federal de Santa Catarina. Graduação em Engenharia de Produção Civil. III. Título.

Filipe Ferreira Viella

Modelagem de Atribuição de Marketing Digital a partir de Cadeias de Markov

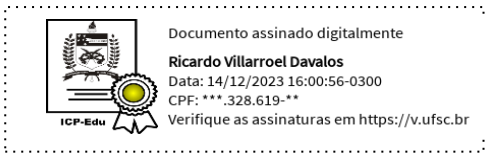
Este Trabalho de Conclusão de Curso foi julgado adequado para obtenção do título de Bacharel em Engenharia Civil com Habilitação em Engenharia de Produção e aprovado em sua forma final pelo Curso de Engenharia de Produção Civil.

Florianópolis, 5 de Dezembro de 2023.



Prof. Mônica Maria Mendes Luna, Dra.

Coordenadora do Curso



Prof. Ricardo Villarroel Dávalos, Dr.

Orientador

Florianópolis

2023

AGRADECIMENTOS

Primeiramente, agradeço a minha família e a meus pais, Maria dos Anjos e Sebastião, pelo amor, apoio incondicional e incentivo que foram fundamentais para concluir essa jornada. A confiança me deu forças para continuar.

Aos meus amigos, que compartilharam risos, desafios e inspirações, obrigado por estarem sempre ao meu lado. A amizade tornou os dias difíceis mais leves e as vitórias mais significativas.

Aos colegas de trabalho, que compartilharam suas experiências e conhecimentos, agradeço por enriquecerem meu aprendizado e me darem a oportunidade de evoluir cada vez mais. Suas perspectivas e colaborações foram inestimáveis para o desenvolvimento deste trabalho.

Por fim, agradeço a Universidade Federal de Santa Catarina, instituição que me proporcionou todas as ferramentas e oportunidades necessárias para meu crescimento acadêmico e pessoal. A qualidade do ensino e o apoio dos professores foram fundamentais para a minha formação. Carrego com muito orgulho tudo o que vivi e todas as lições que aprendi aqui.

RESUMO

Na era digital em constante expansão, o marketing desempenha um papel importante na estratégia de negócios de empresas de diferentes portes. A capacidade de atribuir com assertividade o valor de cada interação do cliente é crucial para manutenção da competitividade. Este trabalho tem como objetivo avaliar um modelo de atribuição de marketing baseado em cadeias de Markov, visando maior precisão e adaptabilidade a diferentes contextos. A revisão bibliográfica comparou modelos de atribuição existentes na literatura, justificando a escolha da abordagem baseada em cadeias de Markov devido à sua eficácia em lidar com o problema de atribuição, especialmente em relação à flexibilidade, escalabilidade e reutilização. A pesquisa se caracteriza como aplicada e descritiva, e a metodologia utilizada é de modelagem e simulação. A implementação prática do modelo foi feita em linguagem de programação R, e testes com dados simulados demonstraram sua capacidade de fornecer uma visão mais realista do problema de atribuição de marketing, aumentando a precisão. A partir dos resultados, análises comparativas foram propostas com o auxílio de gráficos e mapas de calor, a fim de aprimorar investimentos em campanhas de marketing e entender o comportamento dos clientes. Além disso, o trabalho destacou o reaproveitamento do código desenvolvido, permitindo sua aplicação em diversas indústrias e contextos.

Palavras-chave: Cadeia de Markov; Modelo de atribuição; Marketing digital.

ABSTRACT

In the constantly expanding digital era, marketing plays a significant role in the business strategy of companies of all sizes. The ability to accurately attribute the value of each customer interaction is crucial for maintaining competitiveness. This work aims to evaluate a marketing attribution model based on Markov chains, with the goal of achieving greater precision and adaptability in different contexts. The literature review compared existing attribution models, justifying the choice of the Markov chain-based approach due to its effectiveness in addressing the attribution problem, especially in terms of flexibility, scalability and reusability. The research is characterized as applied and descriptive, and the methodology used is modeling and simulation. The practical implementation of the model was carried out in the R programming language, and tests with simulated data demonstrated its ability to provide a more realistic view of the marketing attribution problem, increasing precision. Based on the results, comparative analyses were proposed with the assistance of graphs and heat maps to improve marketing campaign investments and understand customer behavior. Furthermore, the study highlighted the reusability of the developed code, allowing its application in various industries and contexts.

Keywords: Markov Chain; Attribution Model; Digital Marketing.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Nos últimos anos, houve um aumento significativo no investimento em campanhas e anúncios de marketing digital. Isso se deve, em grande parte, ao crescente número de pessoas que utilizam a internet para buscar informações e realizar compras. As receitas de publicidade online nos Estados Unidos cresceram 10,8% em 2022, totalizando US$ 209,7 bilhões, um aumento de US$ 20,4 bilhões se comparado ao ano anterior, apesar das altas taxas de inflação e incerteza econômica, confirmando a resiliência da indústria de publicidade na Internet (IAB, 2023). A pandemia do COVID-19 acelerou essa tendência, à medida que muitas empresas foram forçadas a adotar o comércio eletrônico como principal canal de vendas.

Com o aumento da competição no ambiente digital, as empresas têm buscado investir em estratégias de marketing mais sofisticadas e eficazes, incluindo campanhas de anúncios em mídias sociais, otimização de mecanismos de busca a partir de inteligência artificial e criação de modelos de atribuição de marketing, a fim de otimizar o retorno sobre o investimento dessas campanhas (Hubspot, 2023).

O marketing digital tem se tornado uma opção cada vez mais atraente para empresas de todos os tamanhos e setores. Em 2021, 70% das empresas aumentaram seu investimento em publicidade online e pretendem continuar aumentando, segundo pesquisa da MIT Technology Review (2022). Então é provável que o foco de investimentos em campanhas e anúncios online continue a crescer nos próximos anos, à medida que a internet se torna cada vez mais central para as atividades comerciais.

Segundo pesquisa do grupo Dentsu, com previsões fornecidas para 58 mercados cobrindo as Américas, Ásia-Pacífico, Europa, Oriente Médio e África, os gastos com anúncios digitais aumentaram em 32,4% no ano de 2021, após a pandemia de 2020, com cerca de 347 bilhões de dólares investidos, o que representa 52,5% do valor total de gastos com publicidade. Essa parcela tende a aumentar ainda mais nos próximos anos, à medida que os outros tipos de mídia perdem espaço, de acordo com as previsões indicadas na Tabela 1, que ilustra o gasto total com publicidade em bilhões de dólares e a participação em porcentagem deste total para cada categoria observada (Calladine, 2022).

Tabela 1 - Gastos globais com publicidade entre 2021 e 2025

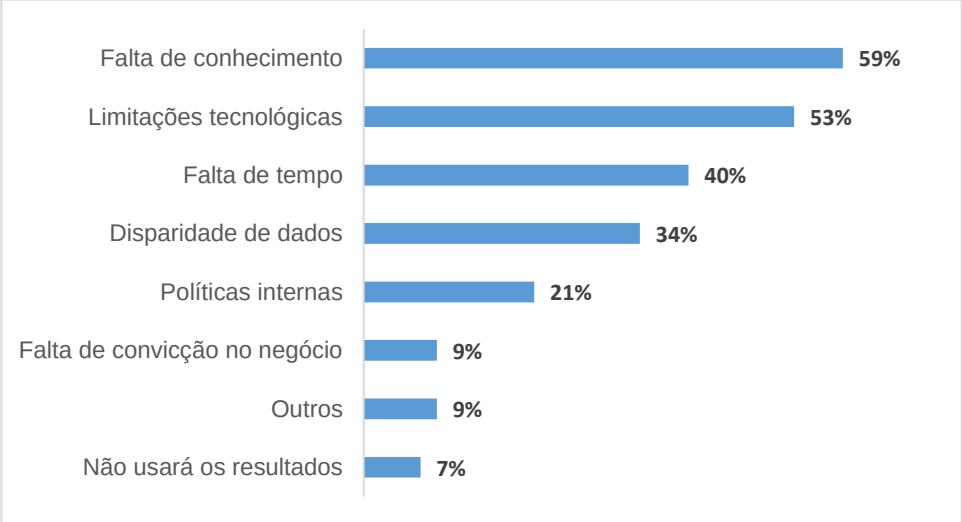
Ano	Gasto total (US$b)	Parcela do gasto total por categoria					
		Mídia Digital	Televisão	Mídia impressa	Mídia externa	Rádio	Cinema
2021	660.5	**52.5%**	27.1%	8.2%	5.3%	5.2%	0.3%
2022	713.6	**55.3%**	25.6%	7.2%	5.4%	5.0%	0.3%
2023	740.9	**57.1%**	24.7%	6.7%	5.3%	4.9%	0.4%
2024	776.9	**58.2%**	24.3%	6.2%	5.2%	4.8%	0.3%
2025	811.6	**59.5%**	23.7%	5.8%	5.1%	4.7%	0.3%

Fonte: Adaptado de Calladine (2022)

Aliado a esse crescimento, os anunciantes utilizam diversos canais de marketing digital a seu favor, como mídias sociais, e-mails, mensagens personalizadas, produção de vídeos, ferramentas de busca pagas, parceiros afiliados e muitos outros, porém falta transparência a respeito de quanto cada canal contribui para o sucesso de sua empresa. Os dados de uma pesquisa feita pela Econsultancy em parceria com a AdRoll, baseado em respostas de 987 profissionais da área de marketing empresarial na Europa, América do Norte e Ásia-Pacífico, mostram que 70% dos negócios, na época, apresentavam dificuldades em calcular e compreender a eficácia de suas campanhas (Econsultancy, 2017).

Os motivos mais comuns para que técnicas mais sofisticadas de atribuição de marketing não fossem aplicadas eram a falta de conhecimento e as limitações tecnológicas, como pode ser visto na Figura 1.

Figura 1 - Impedimentos de aplicação da atribuição de marketing



Fonte: Adaptado de Econsultancy (2017)

Nesse contexto global de diversificação das plataformas e canais de publicidade, é fundamental que as empresas tenham uma compreensão clara e precisa do impacto de suas estratégias de marketing. Surge então uma oportunidade de pesquisa para que essas estratégias sejam otimizadas, com o auxílio de análises probabilísticas e modelos algorítmicos, a fim de manter a competitividade de uma empresa no mercado e obter vantagens em um ambiente cada vez mais complexo e orientado por dados.

1.1 PROBLEMA DE PESQUISA

À medida que o mundo se torna cada vez mais digital, a quantidade de dados disponíveis aumenta, paralelamente com a complexidade de interpretá-los e extrair informações relevantes. A capacidade de analisar e tirar valor de um grande volume de dados solidifica a posição de empresas no mercado, permitindo uma tomada de decisão mais eficaz, identificando tendências, compreendendo o comportamento do cliente e trazendo inúmeras vantagens estratégicas.

A competição acirrada no cenário de publicidade digital impõe grandes desafios para as empresas que buscam maximizar seus resultados, uma vez que a alocação inadequada de recursos pode resultar em desperdícios de orçamento e, consequentemente, resultados insatisfatórios. Os modelos de atribuição desempenham um papel crucial na determinação e entendimento de como os canais de marketing contribuem para a conversão de um cliente, provando ser uma alternativa promissora para resolver problemas de alta complexidade.

Nesse cenário, os seguintes questionamentos serviram como motivadores para o trabalho em questão:

a) as empresas têm controle e entendimento sobre seus investimentos em marketing digital e conseguem definir como alocar de forma eficaz seu orçamento?
b) as ferramentas atuais mais utilizadas no mercado conseguem atribuir de forma satisfatória o crédito real de cada canal de publicidade?
c) ferramentas matemáticas, aliadas ao uso da tecnologia, podem auxiliar a diminuir essa lacuna e tornar menos complexo esse processo?

O uso de cadeias de Markov, uma ferramenta proposta pelo matemático russo Andrei Andreyevich Markov no início do século XX, consegue ilustrar bem

esse problema, em conjunto com os avanços tecnológicos e computacionais da atualidade (Rentola, 2014). A implementação de modelos de atribuição mais precisos torna-se uma realidade alcançável, provando ser um forte aliado na aplicação em grande escala, se utilizado e adaptado a linguagens de programação, sendo capaz de capturar a dinâmica complexa das interações dos clientes com os canais de marketing.

Assim, nessa direção, ficou estabelecido o seguinte problema de pesquisa: a aplicação de cadeias de Markov pode contribuir para uma atribuição de marketing mais precisa e eficiente, se aliada a tecnologias computacionais modernas?

1.2 OBJETIVOS

O objetivo geral deste trabalho é avaliar um modelo de atribuição de marketing mais preciso, construído a partir da aplicação de cadeias de Markov, que possa ser adaptado para diversas indústrias e contextos.

Os objetivos específicos consistem em:

a) comparar os modelos de atribuição mais utilizados na atualidade;
b) explicar as propriedades e características das cadeias de Markov aplicadas na modelagem de atribuição de marketing;
c) propor um modelo de atribuição baseado em cadeias de Markov e estruturado a partir de linguagem de programação;
d) validar o modelo proposto por meio de simulações com um conjunto de dados específico.

1.3 JUSTIFICATIVA

A atribuição de marketing é complexa e dinâmica, e sua abordagem tradicional tem algumas limitações. Muitas organizações ainda dependem de modelos de atribuição baseados em regras heurísticas ou na atribuição igualitária, que não capturam a dinâmica real do comportamento do consumidor (Zaremba, 2020). Além disso, a análise retrospectiva de dados históricos frequentemente deixa de considerar as transições de um canal para outro ao longo do tempo, ignorando informações valiosas (Jayawardane et al., 2015).

As cadeias de Markov são uma classe de modelos probabilísticos que permitem capturar as dependências sequenciais entre estados ao longo do tempo, o que é essencial para compreender como os consumidores percorrem a jornada do

cliente e interagem com os diversos pontos de contato de uma campanha (Anderl et al., 2014). Ao utilizar essa ferramenta na criação de um modelo de atribuição, é possível levar em consideração a natureza dinâmica e não linear das jornadas dos consumidores, bem como a influência de diferentes canais em momentos distintos (Kakalejcik, 2018).

Esta pesquisa se propõe a explorar a aplicação de cadeias de Markov na criação de um modelo de atribuição de marketing, visando contribuir para uma compreensão mais precisa e eficaz da alocação de recursos de marketing e do comportamento do cliente ao longo de suas jornadas.

Ao adotar uma abordagem baseada em cadeias de Markov, a pesquisa busca oferecer uma metodologia mais robusta, permitindo a tomada de decisão baseada em dados por parte das organizações, alocando recursos de maneira mais eficiente e otimizando o retorno sobre o investimento em marketing.

1.4 ESTRUTURA DO TRABALHO

O trabalho contém 5 capítulos, onde no primeiro é apresentada a introdução, juntamente com o problema de pesquisa e os motivadores da escolha do tema, assim como os objetivos, justificativa e estrutura da monografia.

O capítulo 2 é constituído pela fundamentação teórica, que busca entender os principais conceitos relacionados ao tema e explorar a literatura a respeito dos modelos de atribuição mais utilizados na atualidade, a fim de fazer um comparativo com modelos mais complexos e orientados a dados. Ainda neste capítulo, é explicada a escolha do modelo mais apropriado, alinhado com o objetivo do trabalho.

O capítulo 3 trata do método de pesquisa, ilustrando o roteiro metodológico proposto e as delimitações acerca do desenvolvimento do trabalho.

No capítulo 4 é definida a linguagem de programação a ser utilizada para a criação do modelo, uma revisão do conteúdo dos pacotes utilizados com um exemplo de funcionalidade comprovada, assim como a caracterização dos dados de entrada utilizados. A seguir, o modelo é avaliado a partir de um comparativo com outros modelos previamente apresentados e os resultados são discutidos.

Por fim, no capítulo 5 as conclusões são apresentadas, seguidas pelas recomendações para trabalhos futuros, identificadas ao longo do desenvolvimento.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 CONCEITO DE ATRIBUIÇÃO

No âmbito do marketing digital, a atribuição corresponde a um conjunto de processos que visam a identificar as ações de um determinado usuário que contribuem para a obtenção de conversões ou outros resultados desejados. A medição dessa atribuição é distinta daquela empregada na mídia tradicional, sobretudo em razão da existência de meios consistentes para a identificação de usuários no ambiente digital, o que viabiliza uma atribuição mais precisa. A importância dessa técnica reside na sua capacidade de avaliar o quanto cada anúncio influenciou o comportamento de um consumidor, permitindo a determinação do ROI realizado em publicidade digital (IAB Brasil, 2018).

As interações dos usuários com diversas campanhas e telas distintas no meio digital é o que será mensurado em um modelo de atribuição, onde cada interação é considerada como um evento e, a partir daí, é calculado e atribuído um valor para cada evento que contribuiu de alguma forma no resultado de conversão.

O objetivo de um modelo de atribuição é entender o comportamento de compra dos consumidores e determinar quais canais de marketing são mais efetivos para investir fundos. Sharma (2016) lista em sua obra algumas perguntas de negócios que podem ser respondidas com uma boa modelagem de atribuição:

a) como melhorar o retorno sobre o investimento de todos os canais de marketing?
b) como determinar o canal de marketing mais efetivo para investimento?
c) investindo em diversos canais de marketing, quanta incrementalidade cada canal pode trazer para o resultado final de um negócio?
d) como avaliar a performance de uma estratégia de marketing específica?
e) o que acontece com o cliente antes de finalizar sua compra?
f) como os canais trabalham juntos para gerar conversões?

Shao e Li (2011) definem o processo de atribuição em marketing digital como o problema de atribuir crédito a um ou mais anúncios para direcionar o usuário às ações desejáveis, como fazer uma compra, ressaltando a importância de construir modelos que se encaixam melhor na realidade de cada negócio. Por outra perspectiva, um modelo de atribuição também pode ser usado no intuito de entender

melhor o comportamento dos clientes, especialmente daqueles que não completaram sua jornada de conversão, no intuito de identificar que tipos de comportamento geram abandono e, consequentemente, a não conversão.

2.2 TIPOS DE MODELOS DE ATRIBUIÇÃO

Escolher um modelo de atribuição que atenda às necessidades de uma empresa pode ser uma tarefa desafiadora, é particularmente difícil definir os anúncios que performarão da melhor forma com uma quantidade apropriada de crédito. Escolher o modelo certo para um negócio pode ajudar a garantir soluções que mensurem as informações corretas da maneira mais precisa possível. É importante garantir que o modelo escolhido capture cada parte da jornada do cliente (Amazon, c2023).

Segundo revisão sistemática da literatura feita por Zaremba (2020), observou-se um número significativo de abordagens diferentes para a atribuição de conversão, desde regressão logística, passando por modelos bayesianos hierárquicos e probabilísticos até redes neurais e abordagem da teoria dos jogos. Zaremba (2020) alega que a partir de 2019, publicações com modelos baseados em cadeias de Markov e valor de Shapley começaram a despertar mais interesse entre os pesquisadores, o que levou a criação da classificação presente no Quadro 1.

Quadro 1 - Modelos populares de atribuição de conversão

(continua)

Categoria	**Tipo**	**Modelo**	**Regras gerais**
Heurística (crédito atribuído arbitrariamente)	*Single-touch*	Último toque	O impacto geral na conversão é atribuído à última interação.
		Primeiro toque	O impacto geral na conversão é atribuído à primeira interação.
	Multi-touch	Linear	O impacto na conversão é atribuído proporcionalmente a cada atividade no caminho.
		Baseado na posição	O impacto na conversão é atribuído dependendo da posição da atividade no caminho. Por exemplo, o Google Analytics atribui por padrão 40% do impacto para a primeira e última fonte, e os 20% restantes são divididos proporcionalmente entre as outras atividades.
		Peso ponderado	O impacto na conversão é atribuído arbitrariamente e subjetivamente a cada fonte (frequentemente com base em uma análise prévia do profissional/pesquisador).

Quadro 1 - Modelos populares de atribuição de conversão

(conclusão)

Categoria	Tipo	Modelo	Regras gerais
Algorítmica (crédito atribuído econometricamente)	*Multi-touch*	Linear	O impacto na conversão é estudado com base em regressão logística, que, por sua vez, se baseia na decomposição de todos os caminhos de conversão e na atribuição binária da presença ou ausência do canal no caminho.
		Cadeia de Markov	O impacto das fontes na conversão é determinado com base em uma análise do impacto incremental de toda a fonte na população. Com base em todos os caminhos de conversão, são criadas cadeias com a probabilidade de migração do usuário entre as fontes individuais atribuída. Durante a análise, fontes individuais são removidas da área de cálculo e os fluxos de probabilidade são examinados em cadeias sem a fonte excluída. A diferença resultante é um impacto incremental que ilustra o impacto real de uma determinada fonte na conversão final.
		Valor de Shapley	A abordagem da teoria dos jogos e o método do valor de Shapley são uma medida da contribuição marginal média de um canal para cada conjunto de canais (coalizão, que é um caminho único para o esquema de compra). A contribuição marginal de um canal específico é a diferença média entre os resultados de conversão dos conjuntos de canais (coalizão) com e sem um canal específico.

Fonte: Adaptado de Zaremba (2020)

Poutanen (2020) afirma que o desempenho calculado de cada canal pode variar drasticamente dependendo do modelo de atribuição escolhido. Isso pode levar a um dimensionamento ineficiente dos investimentos em marketing, direcionando esforços para canais que não representam o comportamento real dos clientes, desperdiçando recursos que não geram resultados otimizados. Diante disso, é importante compreender mais a fundo as peculiaridades de cada modelo.

2.2.1 Modelos Heurísticos

Também chamados de modelos de atribuição baseado em regras, os modelos heurísticos dependem de um conjunto pré-definido de regras a respeito de como dividir o crédito entre os canais na jornada do cliente. Como visto no Quadro 1, esses modelos podem ser divididos em dois tipos, o *single-touch*, em tradução literal significa toque único ou clique único, onde é atribuído crédito a apenas um dos canais da jornada, ou *multi-touch*, traduzido como múltiplos toques ou múltiplos cliques, onde vários canais podem receber uma parcela do crédito.

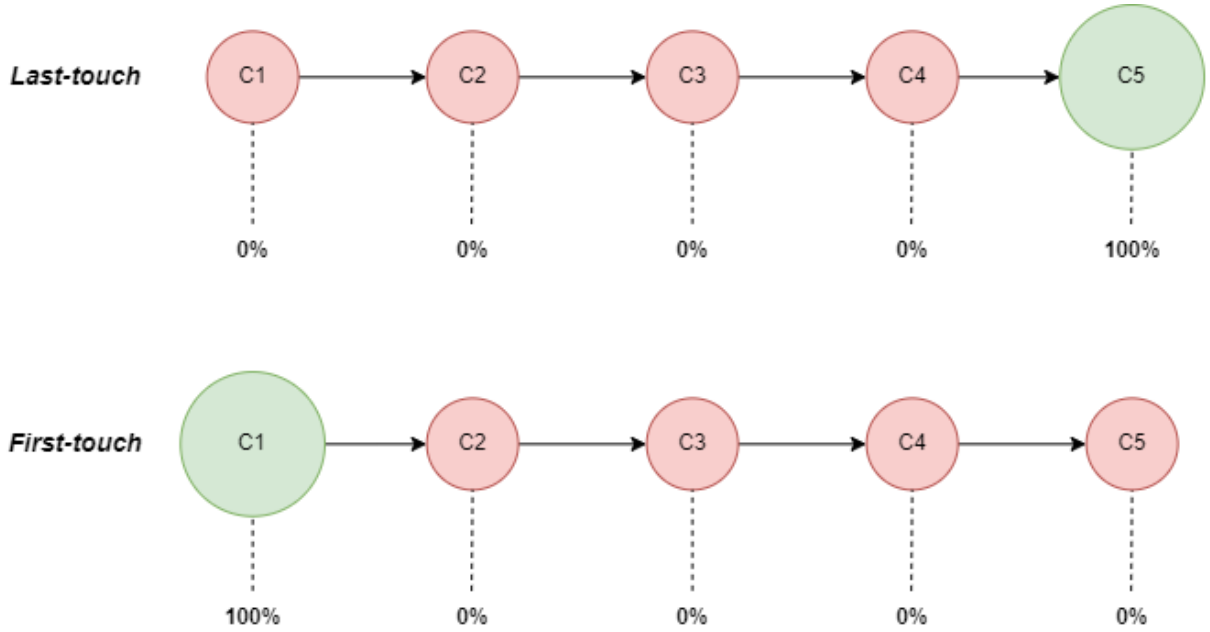
Como esses modelos alocam créditos de forma arbitrária, acabam ganhando maior popularidade e, muitas vezes, são os primeiros modelos adotados em empresas. Devido a sua simplicidade, são fáceis de entender e de implementar, não exigindo muito conhecimento ou técnicas mais robustas (Zaremba, 2020).

2.2.1.1 Single-touch

A metodologia mais utilizada nesta categoria é o modelo de *last-touch,* que atribui todo o crédito da conversão apenas ao último ponto de contato na jornada no cliente. Devido a sua ampla utilização, geralmente é considerado como base comparativa para outros modelos (Jayawardane et al., 2015). Em alguns casos, onde a jornada do cliente é curta, o uso do modelo *last-touch* pode ser justificado. Entretanto, ele certamente ignora um grande volume de informações de outros canais em jornadas mais complexas, simplificando excessivamente o comportamento do cliente (Poutanen, 2020). Outro modelo muito similar é o *first-touch*, sendo o oposto do *last-touch*, onde o crédito é atribuído apenas a primeira interação na jornada do cliente.

A Figura 2 ilustra esses dois casos de modelos *single-touch*, onde cada círculo representa uma campanha ou canal distinto, e a porcentagem abaixo é a parcela de crédito atribuída a esse canal na conversão do cliente.

Figura 2 - Modelos heurísticos *single-touch*



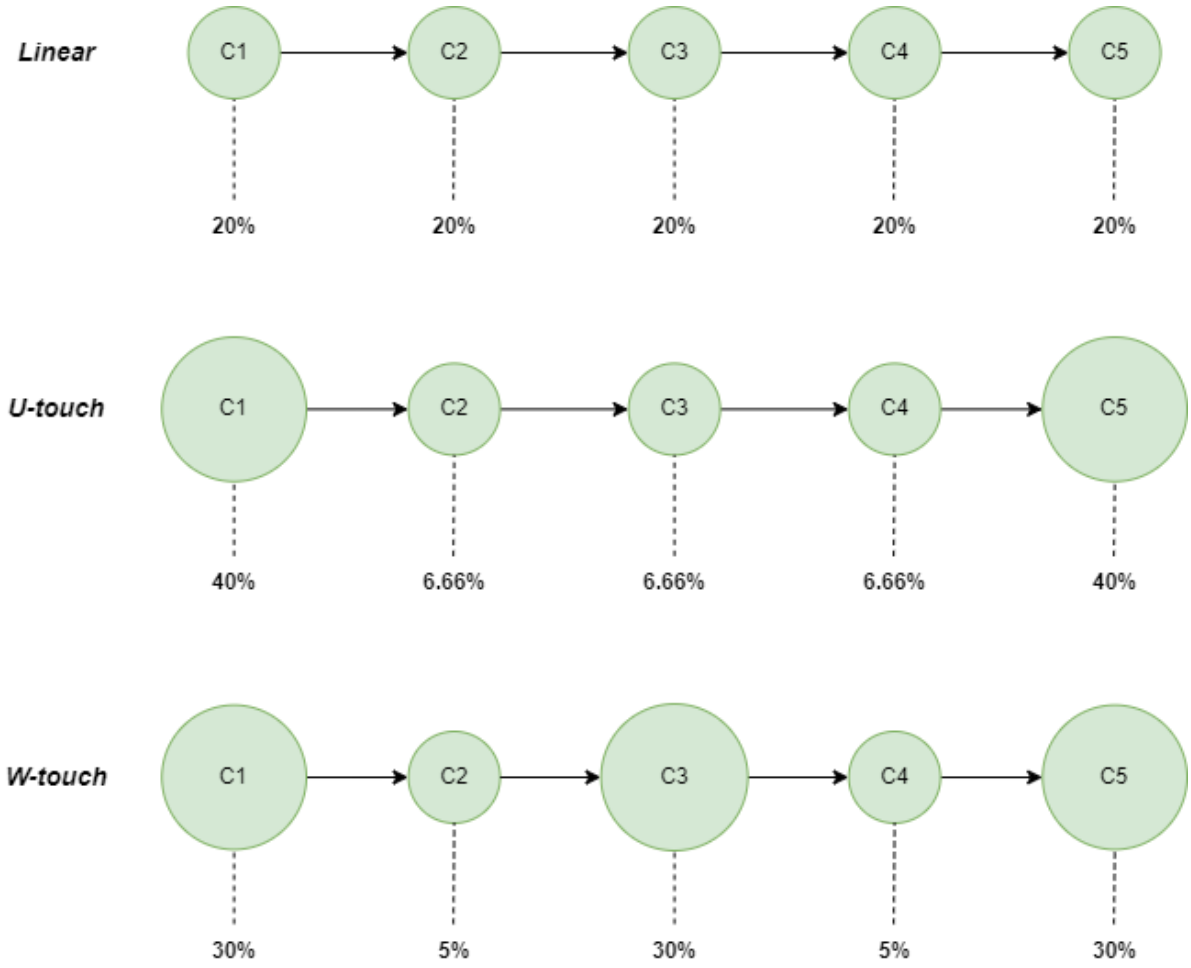
Fonte: Elaborado pelo autor

2.2.1.2 Multi-touch

Nos modelos heurísticos *multi-touch*, existe uma variedade ainda maior de exemplos, que foram propostos para sanar a principal limitação dos modelos simplistas, definindo regras fixas de modo que a distribuição de crédito passe por todos os canais que levam a uma conversão. Entretanto, como essas regras ainda não se apoiam em dados específicos das campanhas, mas sim em intuição, acabam retornando resultados genéricos (Jayawardane et al., 2015).

O modelo linear é o *multi-touch* mais simples, atribuindo crédito igualmente em todos os eventos ou canais ao longo da jornada do cliente. Se houver cinco campanhas mensuráveis em uma jornada, cada uma recebe 20% de crédito (IAB Brasil, 2018). O modelo baseado em posição mais conhecido é o padrão do Google Analytics, que também é chamado de U-*touch*, onde 40% do crédito é atribuído a primeira campanha da jornada, que despertou o interesse do cliente, outros 40% são atribuídos a última campanha que resultou em conversão, a porcentagem remanescente é distribuída igualmente entre as outras campanhas (Zaremba, 2020). Variações do modelo baseado em posição também são encontradas na literatura, como é o caso do W-*touch*, que ainda aloca mais crédito a uma campanha intermediária na jornada e está ilustrado na Figura 3.

Figura 3 - Modelos heurísticos *multi-touch*



Fonte: Elaborado pelo autor

Outras variações, que levam em conta janelas de tempo específicas de quando os eventos ocorrem, também são encontradas na literatura, podendo atribuir uma quantidade diferente de crédito a um canal que gerou uma conversão em um período de dias específico pré-definido pelo profissional/pesquisador (IAB Brasil, 2018).

2.2.2 Modelos Algorítmicos

Os modelos algorítmicos utilizam técnicas estatísticas e ferramentas mais sofisticadas, também chamados de modelos baseados em dados, têm uma abordagem probabilística e fornecem uma visão menos tendenciosa do que os modelos heurísticos. Em um cenário ideal, os dados usados para construir o modelo devem ser coletados da empresa sob avaliação para que representem o

comportamento real do cliente com a maior precisão possível (Jayawardane et al., 2015).

Um grande volume de dados é necessário para que os modelos algorítmicos possam estimar de forma confiável o comportamento dos usuários na jornada do cliente, portanto, não são uma boa opção para empresas menores que ainda não têm uma cultura e gerência de dados bem estruturada (Poutanen, 2020).

Rentola (2014) apresenta um resumo das abordagens algorítmicas encontradas em sua revisão bibliográfica, com uma breve descrição da ideia por trás de cada modelo. O Quadro 2 retrata esse resumo de forma simplificada, porém nos itens subsequentes esses métodos estão mais bem explicados.

Quadro 2 - Resumo de abordagens algorítmicas

Método	**Explicação**
Regressão logística	Classifica clientes em grupos de conversão e não conversão com base em suas interações nos canais. Estima a distribuição de crédito com base nos coeficientes do modelo
Modelo probabilístico	Estima o aumento da probabilidade de conversão de um cliente com base em suas interações com diferentes canais.
Abordagem Markoviana	Modela os caminhos de conversão do cliente com um diagrama de estados Markoviano. Estima as probabilidades de transição entre diferentes estados ou canais. O modelo de atribuição é obtido a partir das probabilidades de transição.

Fonte: Adaptado de Rentola (2014)

2.2.2.1 Regressão logística

Muitos problemas de classificação de duas ou mais variáveis podem ser resolvidos utilizando técnicas distintas de máquinas vetoriais, algoritmos *random forest* e redes neurais, mas estes acabam gerando um modelo complexo com solução de caixa preta, ou seja, não é facilmente compreendido para os tomadores de decisão de marketing e acaba não se tornando aplicável em casos de otimização, existe pouco espaço para adaptação (Poutanen, 2020).

Shao e Li (2011) então propõem um dos primeiros modelos de regressão logística para atribuição de marketing, utilizando um esquema que codifica a presença e ausência de cada canal na jornada do cliente como uma variável binária, a fim de aplicar regressão logística simples. Com os resultados positivos (conversão) ou negativos (não conversão), eles aproximam a jornada do cliente como uma classificação binária (Jayawardane et al., 2015). Essa abordagem permite que os

tomadores de decisão foquem mais na interpretação, porém perdendo precisão, se comparado a técnicas mais complexas (Shao; Li, 2011).

2.2.2.2 Modelo probabilístico

Shao e Li (2011) também propõem um segundo modelo probabilístico, alegando ser ainda mais simples que o de regressão, tal simplicidade se traduz em uma baixa variabilidade de estimativa, o que é uma característica esperada, e facilidade de interpretação, ao mesmo tempo que diminui a precisão. Este modelo consegue estimar valores de atribuição baseado em probabilidades condicionais de primeira e segunda ordem, ou seja, probabilidades de um evento acontecer considerando de um a dois eventos anteriores.

Ao considerar probabilidades de ordem superior, a precisão do modelo pode ser melhorada significativamente (Shao; Li, 2011). O problema é que isso aumenta a complexidade do algoritmo de estimativa de parâmetros do modelo. Assim, a precisão adicional pode não valer a complexidade extra que se deve pagar em termos de tempo de cálculo (Rentola, 2014).

2.2.2.3 Valor de Shapley

O método de valor de Shapley foi originalmente desenvolvido para modelar a contribuição de cada jogador em um jogo cooperativo na abordagem de teoria dos jogos, mas foi posteriormente aplicado em vários outros campos, como a publicidade. O valor de Shapley trata cada um dos canais de publicidade como um jogador em um jogo e assume que todos jogam juntos para influenciar o cliente a converter (Zhao et al., 2018).

Poutanen (2020) pontua que o modelo baseado em valor de Shapley não está muito alinhado com a jornada real do cliente, visto que ele parte do pressuposto que a sequência dos canais não afeta os resultados, o que certamente não condiz com a realidade, pois outros modelos obtêm resultados que variam muito quando a ordem dos canais é modificada (Dalessandro, 2012; Anderl et al., 2014).

Zhao et al. (2018) ainda afirmam que o cálculo pelo valor de Shapley é uma tarefa que exige muito poder computacional, visto que ele se repete 2^n vezes, onde n é o número de canais em cada jornada, tornando um cálculo com 15 ou mais canais, praticamente inviável.

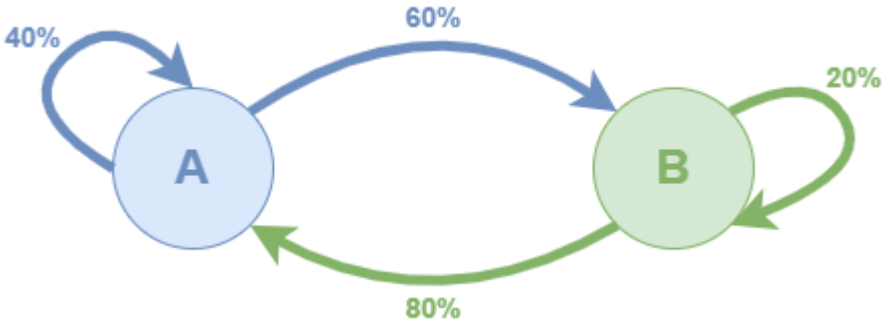
2.2.2.4 Cadeia de Markov

Modelo que ganhou popularidade nos últimos anos, especialmente dentro da comunidade de ciência de dados, abordando uma das maiores deficiências do modelo do valor de Shapley discutido anteriormente, que é justamente o fato de levar em consideração a sequência em que os canais aparecem na jornada no cliente (Poutanen, 2020).

Por ser um método que utiliza grafos markovianos, onde os canais são considerados como os estados e as linhas de transição representam a probabilidade de o cliente ir de um estado a outro, este método se torna mais amigável e de fácil entendimento por parte dos tomadores de decisão. Os grafos de cadeias de Markov se tornam maiores apenas com o aumento do número de estados, porém independem do número de jornadas de clientes consideradas (Anderl et al., 2014), o que permite trabalhar com um número mais elevado de dados sem trazer tanta complexidade computacional e visual ao modelo.

Em resumo, de forma simples, um grafo markoviano mostra a probabilidade de um estado “A” transitar para outro estado “B” ou inclusive se manter no mesmo estado, o que acaba sendo bastante pertinente para ilustrar o problema de atribuição que é discutido neste trabalho. A Figura 4 ilustra um exemplo de uma cadeia de Markov genérica para melhor compreensão.

Figura 4 - Cadeia de Markov genérica



Fonte: Elaborado pelo autor

As probabilidades de transição calculadas com um modelo de Markov podem englobar todos os pontos de interação na jornada do cliente. Para mensurar o crédito atribuído nestes pontos, usa-se uma técnica chamada de *Removal Effect*, ou efeito de remoção em tradução literal, onde a contribuição de cada canal é

calculada removendo-o da jornada do cliente e observando quantas conversões acontecem sem esse canal. (Gaur; Bharti, 2020).

2.3 ESCOLHA DO MODELO MAIS APROPRIADO

Barajas e Akella (2016) afirmam que o uso de métodos heurísticos na criação de modelos de atribuição não é propício, visto que eles ignoram reações hipotéticas sem que o usuário tenha contato com o anúncio. Além da baixa acurácia, existem opções mais sofisticadas nos modelos algorítmicos que são orientadas a dados e, consequentemente, entregam mais credibilidade nas decisões gerenciais.

Kakalejcik et al. (2018) levantam um ponto interessante quanto a disponibilidade de dados no meio empresarial: a exportação de jornadas do cliente, que consistem nos canais de marketing usados para acessar o site antes da compra, está entre os recursos padrão do *Google Analytics*, que é uma das ferramentas gratuitas de análise *web* mais populares da atualidade, independente do nicho (Caldwell, 2023). Isso garante que uma análise baseada em dados possa ser executada amplamente em empresas de diversos tamanhos, setores e orçamentos, inclinando ao uso de modelos algorítmicos.

Anderl et al. (2014) propõem seis critérios gerais para modelagem de atribuição que refletem rigor científico, bem como aspectos relevantes para a implementação na prática, embasado em conceitos vistos nas obras de Shao e Li (2011) e Dalessandro et al. (2012). O Quadro 3 resume de forma objetiva todos esses critérios, sua definição e importância no processo de modelagem de atribuição de marketing e no uso de seus resultados.

Quadro 3 - Critérios para uma boa modelagem de atribuição

(continua)

Critério	Definição	Importância
Objetividade	Os modelos devem ser capazes de atribuir crédito a canais ou campanhas individuais de acordo com sua capacidade factual de gerar valor, como contribuir para conversões ou aumentar as receitas.	- Permite cálculo de impacto relativo das variáveis de decisão e objetividade na avaliação das opções de decisão; - Recompensa um canal individual de acordo com sua capacidade de afetar a probabilidade de conversão (justiça).
Precisão preditiva	Os modelos devem ser capazes de prever eventos de conversão corretamente.	- Persuasão de gestores com a credibilidade do modelo; - Alta precisão na previsão de usuários ativos ou inativos.
Robustez	Os modelos devem fornecer resultados estáveis e reprodutíveis se forem executados várias vezes.	- Evita resultados ruins e instáveis (variabilidade).

Quadro 3 - Critérios para uma boa modelagem de atribuição

(conclusão)

Critério	Definição	Importância
Interpretabilidade	A estrutura do modelo deve ser transparente para todas as partes interessadas e os resultados devem ser interpretáveis com relativa facilidade.	- Resultados aplicáveis diretamente em decisões gerenciais; - Simples e fáceis de comunicar; - Gerentes geralmente não aceitam abordagens caixa preta.
Versatilidade	Versatilidade combina adaptabilidade e facilidade de controle. Adaptabilidade é a capacidade de incorporar novas informações que se tornam disponíveis ao longo do tempo. A facilidade de controle permite que os usuários ajustem as entradas para atender aos requisitos específicos da empresa e obtenham as saídas apropriadas.	- Capacidade de atualização do modelo assim que novas informações estiverem disponíveis; - Permite ajustes de entradas para modificar saídas.
Eficiência algorítmica	As saídas do modelo de computação são rápidas e estão disponíveis quando solicitadas.	- Capacidade de lidar com grande volume de dados de forma rápida e eficiente, sem se sobrecarregar; - Fornece resultados assim que necessário;

Fonte: Adaptado de Anderl et al. (2014)

Com base em estudos de revisão bibliográfica extensiva e na criação dos critérios do quadro acima, Anderl et al. (2014) avaliam que modelos com base em cadeia de Markov contemplam os requisitos propostos e performam melhor em relação a outros.

Outro ponto relevante levantado por Meyn e Tweedie (2009) em sua obra é o fato de que as cadeias de Markov são consideradas convergentes à medida que o tempo avança, ou seja, a cadeia atinge um estado estacionário onde sua distribuição de probabilidade permanece inalterada pelas transições ao longo do tempo. Em outras palavras, uma vez que a cadeia de Markov atinge a convergência, seus resultados não mudam conforme ela continua a evoluir, o que garante uma visão mais realista após múltiplas iterações.

A partir do uso de um pacote de algoritmos desenvolvido por Altomare (2023) chamado *ChannelAttribution,* em linguagem de programação R*,* um modelo de atribuição a partir de cadeias de Markov pode ser facilmente implementado, os dados exportados do *Google Analytics* se adaptam sem muitos ajustes à estrutura suportada por este pacote (Kakalejcik, 2018). Os resultados podem, inclusive, ser comparados com modelos heurísticos mais básicos na própria estrutura.

Agregando todos os fatores mencionados anteriormente, é definido que o uso de cadeias de Markov para a criação de um modelo de atribuição de marketing, que possa ser adaptado a outros contextos, é adequado e recomendado, devido aos avanços de conhecimento tecnológico na atualidade que incentivam e facilitam essa modelagem. Vale ressaltar que embora a modelagem de atribuição seja uma chave para o marketing otimizado, e tenha evoluído bastante nos últimos anos, não há uma metodologia exata e padronizada para modelar os efeitos de cruzamento de canais (Shao; Li, 2011).

2.4 CADEIA DE MARKOV APLICADA NA MODELAGEM DE ATRIBUIÇÃO

Poutanen (2020) descreve em sua obra a metodologia de aplicação de cadeias de Markov para o problema de atribuição, proposta por Anderl et al. (2014), que faz o uso de cadeias de primeira ordem ou de ordem superior.

Primeiramente, é importante definir que os canais de marketing da jornada do cliente são tratados como os estados da cadeia de Markov $(s_1, s_2, \dots, s_n)$, representados na fórmula (1):

$$S = \{s_1, s_2, \dots, s_n\} \tag{1}$$

e as probabilidades de transição entre os estados são tratadas como uma matriz de transição, no caso de cadeias de primeira ordem, representadas pela fórmula (2):

$$W = P\left(X_t = s_j \middle| X_{t-1} = s_i\right), 0 \leq w_{ij} \leq 1, \sum_{j=1}^{N} w_{ij} = 1 \forall i \tag{2}$$

onde W é a probabilidade de transição para o próximo estado s_j dado o estado atual s_i. A probabilidade de transição w_{ij} corresponde à probabilidade de uma interação no canal i ser seguida por uma interação no canal j e é sempre entre 0 e 1 e a soma de todas as probabilidades de transição é 1. Um grafo de Markov é a representação de todos os estados e probabilidades de transição $M = \{S, W\}$ (Anderl et al., 2014).

Com um simples exemplo fictício, considera-se uma empresa que possui três canais de publicidade, que serão os estados S1, S2 e S3 da cadeia. Para fins de modelagem, é necessário adicionar mais três outros estados para que a jornada do cliente fique completa (START, CONVERSION e NULL). START representa o ponto de partida de cada jornada, CONVERSION representa um estado de sucesso quando a jornada é contabilizada como uma conversão e NULL representa um estado onde não houve conversão no período observado (Poutanen, 2020).

Como exemplo, o Quadro 4 mostra três jornadas fictícias de clientes em uma determinada empresa, onde apenas a jornada número 1 leva a conversão.

Quadro 4 - Exemplo de jornadas de clientes

Jornada	Estados em ordem cronológica
1	START>S1>S2>S3>CONVERSION
2	START>S2>NULL
3	START>S1>S2>NULL

Fonte: Adaptado de Poutanen (2020)

As probabilidades de transição podem ser definidas agrupando todas as transições que têm o primeiro estado em comum e relacionando com o estado seguinte, de todas as jornadas, para a partir disso calcular a probabilidade. Nos casos em que o estado de partida e de chegada são iguais, deve-se somar as probabilidades, visto que é uma transição idêntica. A Tabela 2 ilustra melhor esse procedimento.

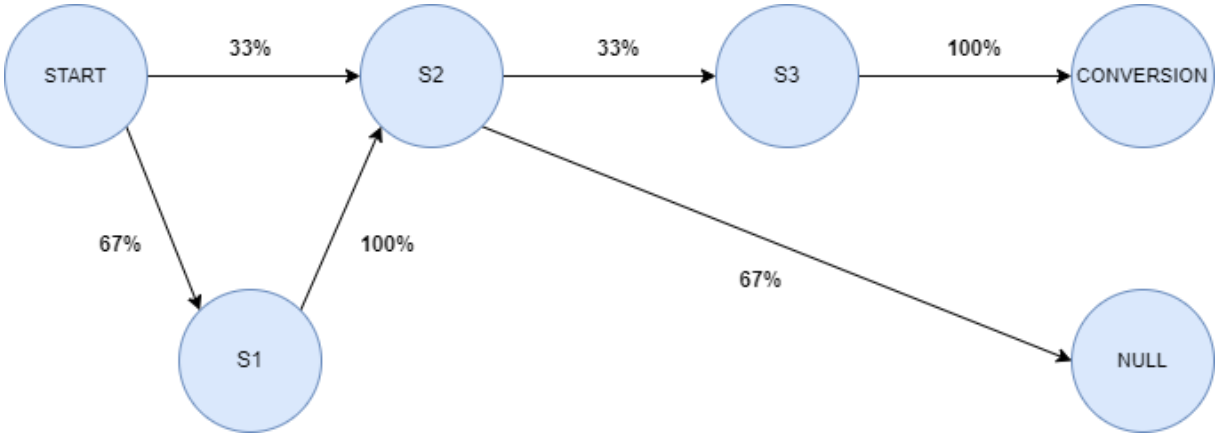
Tabela 2 - Cálculo das probabilidades de transição

De	Para	Probabilidade	Probabilidade Total
START	S1	2/3	**67%**
START	S1		
START	S2	1/3	**33%**
S1	S2	2/2	**100%**
S1	S2		
S2	S3	1/3	**33%**
S2	NULL	2/3	**67%**
S2	NULL		
S3	CONVERSION	1/1	**100%**

Fonte: Elaborado pelo autor

Com as probabilidades devidamente calculadas, pode-se plotar o modelo no formato clássico de um grafo markoviano, onde os estados são representados por círculos e as probabilidades por setas, juntamente com o seu valor em percentual, como pode-se observar na Figura 5.

Figura 5 - Jornadas traduzidas em um grafo markoviano



Fonte: Adaptado de Poutanen (2020)

A ordem da cadeia de Markov define quantos estados antes do estado atual são considerados ao calcular a probabilidade (Anderl et al., 2014), porém para fins de simplificar o entendimento da lógica por trás do modelo, o exemplo deste tópico foi construído apenas na visão de primeira ordem. A cadeia de Markov de primeira ordem leva em conta apenas o estado atual e a probabilidade de sair desse estado. A de segunda ordem olha para trás um estado, então leva em conta o estado atual e um estado antes do estado atual. A de terceira ordem olha para trás dois estados e assim por diante (Poutanen, 2020). A probabilidade de transição de uma cadeia de Markov de ordem k é calculada conforme a equação definida na fórmula (3):

$$w_{ij} = P\left(X_t = s_j \middle| X_{t-1} = s_{t-1}, X_{t-2} = s_{t-2}, \dots, X_{t-k} = s_{t-k}\right) \qquad (3)$$

A complexidade do modelo e o número de parâmetros independentes crescem exponencialmente à medida que a ordem da cadeia de Markov aumenta, exigindo mais poder computacional, porém melhorando a precisão (Poutanen, 2020). Em análise feita por Anderl et al. (2104), a ordem da cadeia é limitada a quatro, no intuito de permitir atualizações em tempo real, mantendo um bom balanço entre versatilidade e precisão.

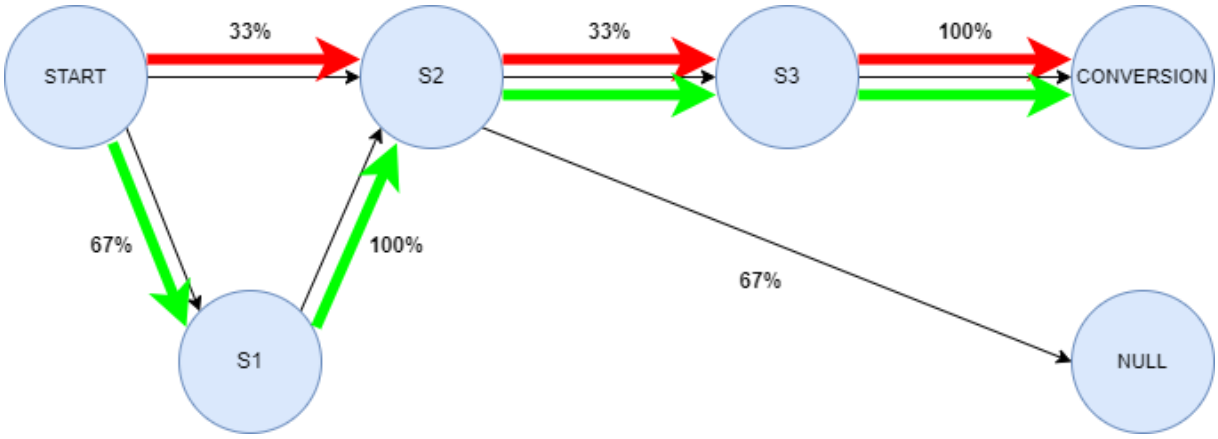
2.4.1 Removal Effect

O grafo da cadeia de Markov, por si só, não representa o potencial da contribuição de cada canal para a conversão, apenas as probabilidades de transição de um estado a outro. Para poder de fato atribuir um potencial de contribuição individual de cada canal, é necessário realizar um cálculo a parte. Anderl et al.

(2014) propõem uma análise de efeito de remoção, chamado de *Removal Effect*, que determina a mudança na probabilidade de o modelo gerar uma conversão, a partir do estado inicial, ao remover um dos canais. Esse processo é feito para cada canal individualmente, nos quais se deseja descobrir qual sua contribuição para a conversão.

Primeiramente, deve-se calcular a probabilidade de conversão do modelo como um todo, sem a remoção de nenhum canal. Para isso, faz-se uma soma e produto das probabilidades de todos os caminhos que levam a conversão. No caso do exemplo deste tópico existem dois caminhos, que estão representados pelas setas verdes e vermelhas na Figura 6.

Figura 6 - Caminhos que levam a conversão



Fonte: Elaborado pelo autor

Então a probabilidade de conversão do modelo como um todo é calculada a como mostra a lógica representada pela expressão da fórmula (4):

$$P(\text{conversão modelo}) = P(START > S2 > S3 > CONVERSION) + P(START > S1 > S2 > S3 > CONVERSION) \quad (4)$$
$$P(\text{conversão modelo}) = 0{,}33 * 0{,}33 * 1 + 0{,}67 * 1 * 0{,}33 * 1 = 0{,}33$$

Com isso temos que 33% das jornadas resultam em conversão. Porém, a probabilidade de conversão desconsiderando a existência do canal S1 no modelo é representada pela sequência START>S2>S3>CONVERSION, como mostra no cálculo da fórmula (5):

$$P(\text{conversão após remoção de } S1) = P(START > S2 > S3 > CONVERSION) = 0{,}33 * 0{,}33 * 1 = 0{,}11 \quad (5)$$

O cálculo do *Removal Effect* de S1 é feito calculando o complemento da divisão da probabilidade de conversão do modelo sem o canal S1 pela probabilidade de conversão total do modelo, conforme fórmula (6):

$$Removal\ Effect(S1) = 1 - \frac{0{,}11}{0{,}33} = 0{,}67 \qquad (6)$$

Em outras palavras, isso significa que 67% das conversões seriam perdidas caso o canal S1 fosse removido (Poutanen, 2020).

Pelo grafo da Figura 6, é fácil perceber que todas as jornadas que levam a conversão utilizam os canais S2 e S3, então seu efeito de remoção é 1, significando que todas as conversões seriam perdidas se qualquer um desses canais fossem removidos. Conhecendo os efeitos de remoção de todos os canais, é possível calcular o potencial de contribuição de cada canal com base no efeito de remoção relativo dos canais em comparação com a soma dos efeitos de remoção de todos os outros canais. Por exemplo, o coeficiente de atribuição do canal S1 é calculado por 0,67 / (0,67+1+1) = 0,25. Isso implica que se deve atribuir 25% de todas as conversões ao canal S1 (Bryl, 2016). Vale ressaltar que este é um exemplo bastante simplificado de jornadas de cliente, pois geralmente existem centenas ou até milhares de jornadas diferentes, especialmente se a empresa utiliza um número maior de canais de marketing.

De forma resumida, a Tabela 3 sintetiza as informações da probabilidade de conversão sem o canal, o efeito de remoção e o valor de conversão atribuído a cada um dos três canais do exemplo deste capítulo, conforme o que foi explicado até aqui.

Tabela 3 - Removal effect e conversão atribuida

Canal	**Probabilidade de conversão sem o canal**	***Removal Effect***	**Conversão atribuída**
S1	0,11	$1 - \frac{0{,}11}{0{,}33} = 0{,}67$	$\frac{0{,}67}{0{,}67 + 1 + 1} = 0{,}25$
S2	0	$1 - \frac{0}{0{,}33} = 1$	$\frac{1}{0{,}67 + 1 + 1} = 0{,}375$
S3	0	$1 - \frac{0}{0{,}33} = 1$	$\frac{1}{0{,}67 + 1 + 1} = 0{,}375$

Fonte: Elaborado pelo autor

Em cadeias de Markov de maior ordem, é possível calcular o efeito de remoção levando em conta a cronologia anterior dos canais de marketing, aumentando exponencialmente a complexidade. Nesses casos, o efeito de remoção de um único canal é calculado como uma média dos efeitos de remoção de todos os estados que têm aquele canal de comercialização específico como o último canal na sequência (Anderl et al., 2014). Quanto maior o efeito de remoção de um canal, maior é a importância dele para o modelo em questão.

2.5 CONSIDERAÇÕES FINAIS DO CAPÍTULO

Com base nos conceitos expostos ao longo dos tópicos anteriores, fica clara a importância de se atribuir crédito aos canais de publicidade no cenário de marketing digital, tanto para descobrir onde investir melhor o orçamento de uma empresa, quanto para entender mais profundamente o comportamento de seus clientes. Além disso, os modelos algorítmicos provaram trazer resultados mais precisos e próximos da realidade que os modelos heurísticos, de acordo com os critérios propostos por Anderl et al. (2014).

A cadeia de Markov em específico, mostrou ser uma ferramenta que, apesar de centenária, consegue lidar com o problema de atribuição de forma elegante e não muito complexa, se comparada a outros métodos algorítmicos mais abstratos. Aliando sua aplicabilidade com a evolução tecnológica da atualidade e os pacotes já existentes em linguagens de programação computacional, um modelo eficaz pode ser implementado e avaliado de forma lógica.

Os próximos capítulos explicam de forma mais aprofundada como funciona a lógica de programação utilizada para implementar tais modelos e como devem estar estruturados os dados de entrada para que os resultados possam ser atingidos. A partir disso o modelo em si é desenvolvido no intuito de avaliar sua eficácia perante outros modelos, com um conjunto de dados criados e padronizados para realizar as devidas simulações.

3 METODOLOGIA

Esse capítulo apresenta os procedimentos metodológicos adotados no desenvolvimento do modelo de atribuição proposto, caracterizando o tipo de pesquisa e as ferramentas utilizadas para alcançar o objetivo. As etapas do trabalho serão descritas de forma cronológica a fim de situar o leitor para melhor compreensão.

3.1 CARACTERIZAÇÃO DA PESQUISA

Pesquisas no âmbito científico podem ser classificadas segundo a sua finalidade ou natureza, que são divididas em pesquisa básica e pesquisa aplicada, segundo seus propósitos, que podem ser exploratórios, descritivos ou explicativos e ainda quanto aos seus procedimentos técnicos (Gil, 2017). A natureza desta pesquisa se caracteriza como aplicada, pois é voltada à aquisição de conhecimentos visando a aplicação em uma situação específica. Quanto ao seu propósito, se caracteriza como uma pesquisa descritiva, cuja preocupação é voltada a observar os fatos, registrá-los, analisá-los, classificá-los e interpretá-los, e o pesquisador não interfere neles (Andrade, 2002 apud Beuren; Raupp, 2006). Gil (2017) ainda afirma que a pesquisa descritiva tem como objetivo descrever as características de uma determinada população ou fenômeno, com a finalidade de identificar possíveis relações entre variáveis.

O procedimento técnico que mais se adequa ao conteúdo do trabalho é a metodologia de modelagem e simulação, com ênfase em pesquisa operacional e modelos quantitativos. Arenales et al. (2007) definem pesquisa operacional como uma abordagem científica que auxilia no processo de tomada de decisões, procurando melhorar o planejamento e projeto de sistemas sob condições que requerem alocações eficientes de recursos escassos. Modelos quantitativos também são definidos por Miguel (2012, p. 171) como:

> modelos abstratos descritos em linguagem matemática e computacional, que utilizam técnicas analíticas (matemáticas, estatísticas) e experimentais (simulação) para calcular valores numéricos das propriedades do sistema em questão, podendo ser usados para analisar os resultados de diferentes ações possíveis no sistema.

Isso permite compreender melhor o ambiente em análise, identificar problemas, formular estratégias e oportunidades, apoiar e sistematizar o processo de tomada de decisão.

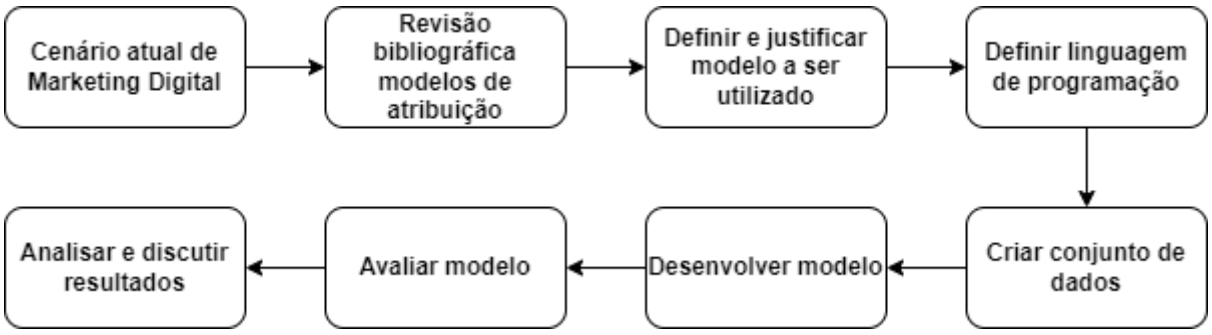
3.2 ROTEIRO METODOLÓGICO

Com o propósito de alcançar os objetivos estabelecidos e organizar o andamento deste trabalho, foi elaborado um diagrama sequencial das atividades realizadas, divididas nas seguintes etapas:

a) compreender o cenário atual de marketing digital e sua importância no âmbito profissional e no mercado de trabalho, observando onde existem oportunidades de melhoria para sustentar o problema de pesquisa;
b) realizar revisão da literatura sobre modelos de atribuição existentes a fim de entender quais são mais utilizados na atualidade e quais melhor se encaixam para resolver o problema de pesquisa;
c) definir e justificar o modelo a ser utilizado com base na revisão bibliográfica, levando em consideração os critérios definidos;
d) definir a linguagem de programação utilizada para construir o modelo com base nos pacotes já existentes no mercado;
e) criar e descrever o processo de criação de um conjunto de dados que possa simular jornadas realistas de clientes e que se enquadre nos requisitos e padrões da linguagem escolhida;
f) desenvolver o modelo na linguagem escolhida;
g) avaliar o modelo construído por meio de simulações utilizando como entrada o conjunto de dados preparado;
h) analisar e discutir os resultados obtidos, fazendo comparativos com os modelos de atribuição mais utilizados e acessíveis, descritos na revisão bibliográfica.

O fluxograma ilustrado pela Figura 7 representa, de forma resumida, as etapas do roteiro proposto.

Figura 7 - Etapas do trabalho



Fonte: Elaborado pelo autor

A escolha do procedimento metodológico descrito acima foi fundamentada na necessidade de desenvolver um modelo de atribuição de marketing de forma sistemática e coerente. Ao seguir esse roteiro metodológico, é possível estabelecer diretrizes claras para a realização da pesquisa, desde a revisão bibliográfica até a análise e discussão dos resultados.

3.3 DELIMITAÇÕES

O escopo deste trabalho abrange exclusivamente canais de marketing digital, pois desenvolver um modelo de atribuição que também considere canais offline em conjunto é uma tarefa quase impossível, visto que as campanhas fora do ambiente online dificilmente podem ser medidas ao nível do usuário (Poutanen, 2020). Devido às dificuldades na mensuração dos canais offline, a literatura tende a focar apenas em canais online.

É importante ressaltar que, em muitos casos, os dados disponíveis não estão limpos e organizados, o que implica em realizar um tratamento minucioso para padronização. Porém, a fim de simplificar e otimizar o tempo de pesquisa disponível, o desenvolvimento do modelo parte do pressuposto que os dados de entrada já estão bem estruturados e validados, independente da forma em que esse dado seja gerado, por auxílio de softwares ou mecanismos de terceiros, não havendo estudos ou análises de tráfego mais detalhadas em campanhas digitais neste trabalho.

4 DESENVOLVIMENTO

Modelos de atribuição utilizando cadeias de Markov, como aquele apresentado no tópico 2.4, podem ser implementados em linguagens de programação como forma de contornar as principais limitações do cálculo manual e atingir, com rigor científico, os critérios propostos por Anderl et al. (2014): objetividade, precisão, robustez, interpretabilidade, versatilidade e eficiência.

Com o aumento da complexidade e do volume de dados na modelagem matemática, fica inviável realizar cálculos manuais para o objetivo deste trabalho. As cadeias de Markov podem se tornar muito complexas em cenários com muitos estados e transições. A linguagem de programação permite lidar de forma mais eficiente com sistemas grandes e intricados que seriam praticamente impossíveis de gerenciar manualmente.

O presente capítulo define qual linguagem de programação melhor se encaixa nas condições deste trabalho e explica como o modelo é implementado e simulado, levando em consideração questões como escalabilidade, flexibilidade, reutilização e visualização.

4.1 LINGUAGEM DE PROGRAMAÇÃO ESCOLHIDA

Atualmente, diversas linguagens podem ser usadas para modelar problemas matemáticos complexos que envolvem as áreas de pesquisa operacional e ciência de dados. As mais populares são Python, R, C/C++, Julia e Matlab (Reis, 2021). A familiaridade do autor com a linguagem e a disponibilidade de bibliotecas e ferramentas complementares relacionadas a cadeias de Markov foram os fatores mais importantes utilizados como filtro para limitar esse conjunto, sendo Python e R as duas opções ponderadas.

Python e R são linguagens de código aberto, ou seja, seu código-fonte está disponível para o público em geral e qualquer pessoa pode visualizar, modificar e distribuir o código da linguagem de forma gratuita, garantindo liberdade de uso, modificação e distribuição (Kovacs, 2021). Uma das características mais relevantes de uma linguagem de código aberto é sua abordagem colaborativa, visto que a própria comunidade desenvolve e compartilha bibliotecas que auxiliam na resolução dos mais variados tipos de problemas, o que aumenta consideravelmente o número de pacotes existentes para esses softwares.

Python é uma linguagem de programação orientada a objetos, lançada em 1989, de fácil aprendizado e legibilidade, é amplamente usada por programadores e desenvolvedores, incluindo bibliotecas especializadas em *machine learning* e *deep learning* (IBM, 2021). Já o R é uma linguagem voltada para análise estatística e visualização de dados, desenvolvida em 1992. É bastante utilizada por estudiosos e pesquisadores no âmbito acadêmico e se inclina fortemente para modelos estatísticos e análises especializadas (IBM, 2021), oferece uma gama variada de bibliotecas para limpeza e manipulação de dados, criação de gráficos e visualizações (R Project, c2023).

Uma breve síntese de algumas das principais funcionalidades relacionadas a dados (coleta, exploração, modelagem e visualização) é apresentada no Quadro 5, levantando alguns pontos fortes e fracos das linguagens Python e R nesses aspectos. Vale ressaltar que, na maioria dos casos, existem bibliotecas externas desenvolvidas pela comunidade que acabam por preencher as lacunas e defasagens de cada linguagem, porém o foco é entender para onde cada linguagem é direcionada.

Quadro 5 - Recursos relacionados a dados de cada linguagem

Recursos	**Python**	**R**
Coleta de dados	Suporta todos os tipos e formatos de dados, incluindo tabelas SQL.	Projetado para que os analistas de dados importem dados de arquivos Excel, CSV e de texto.
Exploração de dados	Possui a biblioteca *Pandas*, bastante utilizada para explorar e tratar dados.	Otimizado para análise estatística de grandes conjuntos de dados.
Modelagem de dados	Possui bibliotecas como *Numpy* para análise numérica, *SciPy* para computação científica e *scikit-learn* para algoritmos de aprendizado de máquina.	O conjunto específico de pacotes conhecido como *Tidyverse* facilita a importação, manipulação, visualização e relatório de dados.
Visualização de dados	Não é um ponto forte, porém existem as bibliotecas *Matplotlib* para gráficos básicos e *Seaborn* para gráficos estatísticos.	Tem um foco maior em visualização, pode-se usar o *ggplot2* para criar gráficos de dispersão complexos com linhas de regressão.

Fonte: Adaptado de IBM (2021)

Em resumo, ambas as linguagens têm funcionalidades similares e estão aptas a lidar com modelos de atribuição baseados em cadeias de Markov, então a escolha norteia-se na existência de pacotes específicos direcionados ao problema em questão. A linguagem R possui o pacote *ChannelAttribution* desenvolvido por Davide Altomare, com última atualização em maio de 2023, que trata justamente do

problema de atribuição de multicanal e a abordagem se baseia na pesquisa de Anderl et al. (2014), já mencionada anteriormente neste trabalho (Altomare, 2023).

Outro pacote que é bastante relevante em termos de visualização para este problema é o *markovchain*, desenvolvido por Giorgio Alfredo Spedicato, com última atualização em setembro de 2023, permite criar e gerenciar cadeias markovianas em tempo discreto com mais facilidade, além de fornecer funções para análises estatísticas, probabilísticas e visualizações destas cadeias (Spedicato, 2023).

Atualmente, ambos os pacotes mencionados já possuem versões em Python, porém devido ao maior uso da linguagem R em âmbito acadêmico e seu maior foco em visualizações gráficas, opta-se por utilizá-la para o desenvolvimento e simulação do modelo.

4.2 PACOTE CHANNEL ATTRIBUTION

Intitulado “Modelo de Markov para atribuição multicanal online”, em tradução livre, o pacote *ChannelAttribution*, desenvolvido por Davide Altomare, atende o problema de atribuição para canais de marketing digital de forma direta e simplificada. O algoritmo foi implementado em C++, com sua primeira versão rodando em linguagem de programação R, e hoje encontra-se na versão 2.0.7 publicado em 17 de maio de 2023 no repositório CRAN (*Comprehensive R Archive Network*), uma rede de servidores web em todo o mundo que armazena versões atualizadas de códigos e documentação para R (CRAN, c2023).

O pacote contém funções para atribuição de canais em marketing digital e é descrito por Altomare (2023, p. 2) da seguinte maneira:

> [...] anunciantes utilizam uma variedade de canais de marketing online para alcançar os consumidores e querem saber até que ponto cada canal contribui para o seu sucesso de marketing. Isso é chamado de problema de atribuição multicanal online. Em muitos casos, os anunciantes abordam este problema através de métodos heurísticos simples que não levam em conta quaisquer interações com o cliente e muitas vezes tendem a subestimar a importância dos pequenos canais na contribuição de marketing. Este pacote fornece uma função que aborda o problema de atribuição de forma probabilística. Usando uma representação Markov de ordem k para identificar correlações estruturais nos dados da jornada do cliente. Isto permite fornecer uma avaliação mais confiável da contribuição de marketing de cada canal. A abordagem segue basicamente aquela apresentada em Anderl et al. (2014). Entretanto, resolvemos o processo de estimação utilizando simulações estocásticas. Desta forma também é possível ter em conta os valores de conversão e a sua variabilidade no cálculo da importância do canal. O pacote também contém uma função que estima três modelos heurísticos (abordagem de primeiro toque, último toque e toque linear) para o mesmo problema.

O Quadro 6 mostra as principais funções existentes dentro do pacote e uma breve descrição de sua funcionalidade.

Quadro 6 - Funções do pacote *ChannelAttribution*

Noma da função	Descrição
transition_matrix	Estima uma matriz de transição a partir dos dados da jornada do cliente.
choose_order	Encontra a ordem mínima do Modelo de Markov que forneça uma boa representação do comportamento dos clientes para os dados considerados, a partir de conceitos de ROC (*Receiver Operating Characterist*), métrica bastante utilizada em aprendizado de máquina. Requer caminhos que não levam à conversão como entrada.
markov_model	Estima um modelo Markoviano de ordem k a partir dos dados da jornada do cliente. Esta função itera a estimativa até que a convergência seja alcançada e permite o multiprocessamento.
heuristic_models	Estima três modelos heurísticos (*first-touch*, *last-touch* e linear) a partir dos dados da jornada do cliente.

Fonte: Adaptado de Altomare (2023)

O pré-requisito mínimo para que o pacote funcione corretamente é que os dados de entrada contenham três colunas que representem: as jornadas do cliente com a ordem correta de todos os canais por onde ele transitou, o número de conversões de cada uma dessas jornadas e o número de conversões que aquela jornada em específico deixou de trazer (não conversões). De forma simplificada, um exemplo do *input* de dados está no Quadro 7, onde cada canal fictício está representado pela letra C seguido de um número, o separador de canais escolhido é o caractere “>”, porém ele pode ser alterado se necessário.

Quadro 7 - Exemplo fictício de dados de entrada

Jornada	Número de conversões	Número de não conversões
C1 > C2 > C3	2	3
C1 > C3	4	2
C3 > C5 > C2	3	0
C1 > C2 > C4	5	3
C5 > C1	2	3

Fonte: Elaborado pelo autor

Cada função possui parâmetros a serem definidos, os quais devem estar relacionados com a formatação dos dados de entrada. A sequência de quadros a

seguir mostram os principais parâmetros de cada função que são utilizados nos tópicos seguintes, o tipo de dado que o parâmetro aceita, o dado padrão caso não tenha nenhum *input* para aquele parâmetro e sua descrição.

Quadro 8 - Parâmetros da função transition_matrix

Parâmetro	Tipo	Padrão	Descrição
Data	*data.frame*	*Input* obrigatório	*Data frame* ou o local do arquivo que contém todos os dados de jornadas e conversões
var_path	*string*	*Input* obrigatório	Nome da coluna que contém as jornadas do cliente
var_conv	*string*	*Input* obrigatório	Nome da coluna que contém a quantidade de conversões
var_null	*string*	Nulo	Nome da coluna que contém a quantidade de não conversões
flg_equal	booleano	TRUE	Se for TRUE, as transições de um canal para ele mesmo serão consideradas

Fonte: Adaptado de Altomare (2023)

A função *transition_matrix* (Quadro 8) é importante pois gera a matriz de transição dos estados da cadeia de Markov, ou seja, as probabilidades de o cliente transitar de um estado a outro. Com o resultado dessa função, o auxílio do pacote “markovchain” e algumas operações para pivotar os dados, é possível criar a visualização do grafo markoviano, que permite enxergar como a cadeia é estruturada.

Quadro 9 - Parâmetros da função choose_order

Parâmetro	Tipo	Padrão	Descrição
Data	*data.frame*	*Input* obrigatório	*Data frame* ou o local do arquivo que contém todos os dados de jornadas e conversões
var_path	*string*	*Input* obrigatório	Nome da coluna que contém as jornadas do cliente
var_conv	*string*	*Input* obrigatório	Nome da coluna que contém a quantidade de conversões
var_null	*string*	Nulo	Nome da coluna que contém a quantidade de não conversões
max_order	inteiro	10	Ordem máxima da cadeia de Markov a ser considerada
plot	booleano	TRUE	Se for TRUE, um gráfico com a curva AUC será exibido

Fonte: Adaptado de Altomare (2023)

De acordo com Altomare (2023), a função *choose_order* (Quadro 9) utiliza conceitos de aprendizado de máquina, criando conjuntos de treinamento e de teste a partir do input de dados, a fim de estimar um modelo de Markov para cada ordem considerada. Esses modelos são usados para prever o estado final (conversão ou não conversão) para cada jornada do cliente no conjunto de teste. Para cada modelo é definida uma curva ROC (*Receiver Operating Characterist*) e calculada a área sob a curva, que é chamada de AUC (*Area Under the Curve*). Este procedimento permite encontrar a ordem mínima que forneça uma boa representação do comportamento dos clientes para os dados considerados, porém não é o foco deste trabalho e não será aprofundado.

Quadro 10 - Parâmetros da função markov_model

Parâmetro	**Tipo**	**Padrão**	**Descrição**
Data	*data.frame*	*Input* obrigatório	*Data frame* ou o local do arquivo que contém todos os dados de jornadas e conversões
var_path	*string*	*Input* obrigatório	Nome da coluna que contém as jornadas do cliente
var_conv	*string*	*Input* obrigatório	Nome da coluna que contém a quantidade de conversões
var_value	*string*	Nulo	Nome da coluna que contém o valor total das conversões
var_null	*string*	Nulo	Nome da coluna que contém a quantidade de não conversões
order	inteiro	1	Ordem da cadeia de Markov
out_more	booleano	FALSE	Se for TRUE, além do resultado, serão retornados os efeitos de remoção

Fonte: Adaptado de Altomare (2023)

A função *markov_model* (Quadro 10) realiza 100.000 iterações como padrão, o que na maioria dos casos é suficiente para atingir a convergência (Altomare, 2023), porém esse número pode ser ajustado para atender outros casos. O resultado principal é a quantidade de conversões atribuída a cada um dos canais de marketing dos dados de entrada, com isso pode-se enxergar de fato quais canais foram mais relevantes e trouxeram o maior número de conversões. Porém, ao utilizar o parâmetro “out_more=TRUE”, é possível obter os efeitos de remoção considerados para cada canal, o que se torna um resultado auxiliar relevante para análises posteriores.

Vale ressaltar também que os parâmetros podem ser ajustados para testar cenários diferentes. Como exemplo, é possível utilizar o parâmetro “var_value” para analisar cenários que envolvam valor monetário e não somente quantidades de conversões, apesar de não ser obrigatório para o funcionamento do algoritmo. A ordem da cadeia de Markov obtida na função *choose_order* também pode ser utilizada como o parâmetro “*order*” para obter resultados mais precisos.

Quadro 11 - Parâmetros da função heuristic_models

Parâmetro	Tipo	Padrão	Descrição
Data	*data.frame*	*Input* obrigatório	*Data frame* ou o local do arquivo que contém todos os dados de jornadas e conversões
var_path	*string*	*Input* obrigatório	Nome da coluna que contém as jornadas do cliente
var_conv	*string*	*Input* obrigatório	Nome da coluna que contém a quantidade de conversões
var_value	*string*	Nulo	Nome da coluna que contém o valor total das conversões

Fonte: Adaptado de Altomare (2023)

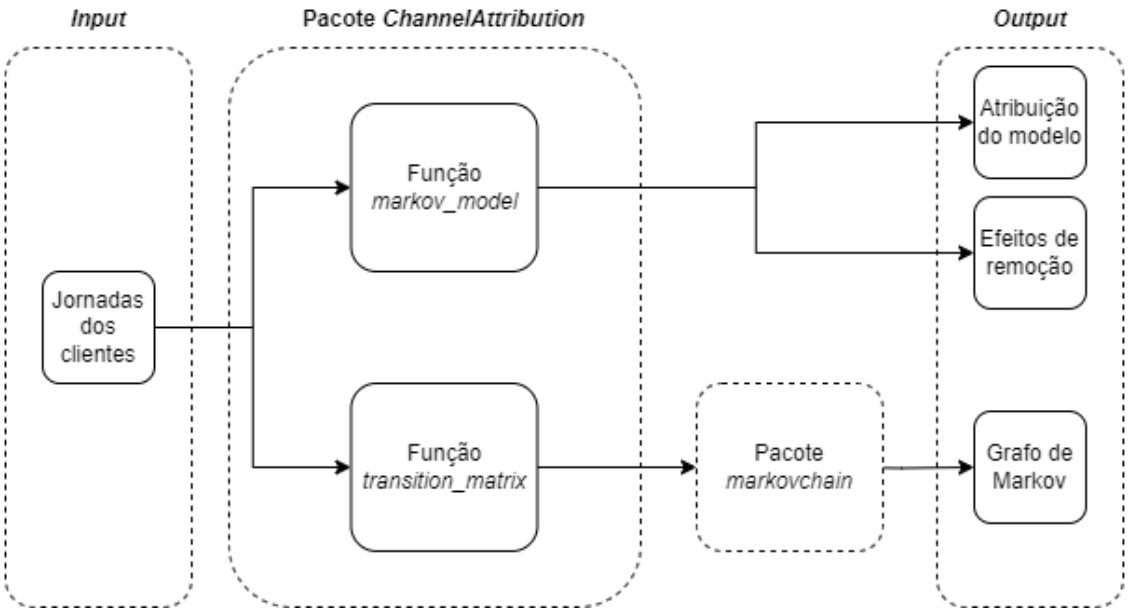
Por fim, a função *heuristic_models* (Quadro 11), de forma análoga a *markov_model*, também traz a quantidade de conversões atribuída a cada um dos canais, porém na visão de três modelos heurísticos mais simples: *first-touch*, *last-touch* e linear. Com esse resultado, é possível fazer comparativos gráficos com o modelo markoviano e analisar as diferenças de cada um dos modelos.

4.2.1 Teste de funcionalidade

Os processos internos usados na construção de todas as funções do pacote *ChannelAttribution* não são especificados minuciosamente na documentação disponibilizada por Altomare (2023), devido a isso, não se pode afirmar com exatidão quais métodos e cálculos foram utilizados na construção interna do algoritmo. Como forma de confirmar se o pacote atende as necessidades deste trabalho, testes foram feitos com o mesmo exemplo calculado de forma manual no tópico 2.4, em conformidade com a obra de Poutanen (2020), porém agora utilizando a linguagem de programação R em conjunto com os pacotes *ChannelAttribution* e *markovchain*.

O teste usa como entrada os dados de três jornadas ao longo de três canais distintos, S1, S2 e S3, conforme mostra o Quadro 4. Depois é aplicada a função *transition_matrix* para gerar a matriz de transição e transformá-la em um grafo markoviano com o auxílio do pacote *markovchain*. Por se tratar de um exemplo simples, é considerada uma cadeia de Markov de ordem 1. A seguir, aplica-se a função *markov_model* para gerar os resultados de atribuição do modelo e os efeitos de remoção. A Figura 8 ilustra todo o processo de forma esquemática, mostrando os dados de entrada, cálculos realizados pelos pacotes e dados de saída.

Figura 8 - Processo de utilização dos pacotes



Fonte: Elaborado pelo autor

O código em R utilizado para obtenção dos resultados encontra-se no Apêndice A, juntamente com comentários explicativos de cada linha de código. Nas figuras a seguir encontram-se os dados de saída obtidos referentes a atribuição do modelo (Figura 9), efeitos de remoção (Figura 10) e grafo de Markov (Figura 11).

Figura 9 - Conversões atribuídas do modelo teste

	channel_name	total_conversions
1	S1	0.2501376
2	S2	0.3749312
3	S3	0.3749312

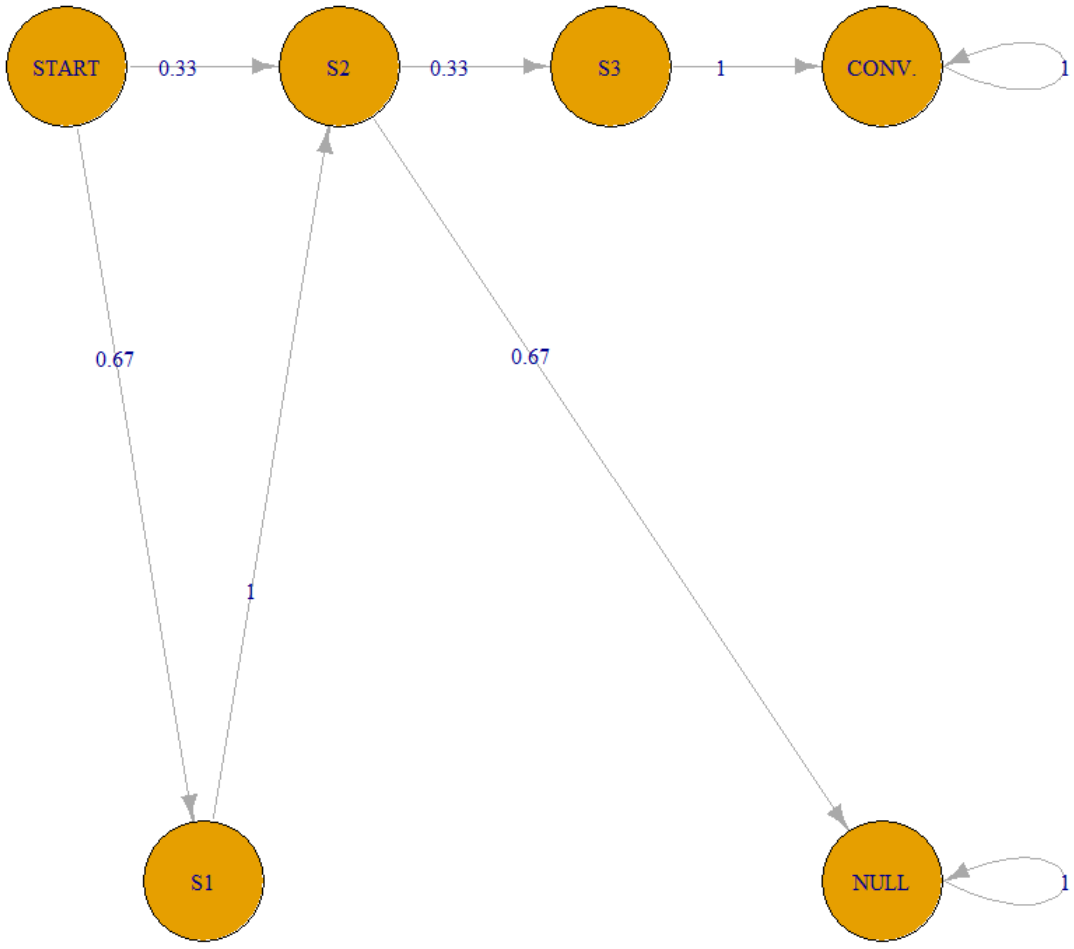
Fonte: *Output* software R Studio

Figura 10 - Efeitos de remoção do modelo teste

	channel_name	removal_effects
1	S1	0.6671559
2	S2	1.0000000
3	S3	1.0000000

Fonte: *Output* software R Studio

Figura 11 - Grafo do modelo teste



Fonte: *Output* software R Studio

Observa-se que os resultados mostrados na Figura 9 e Figura 10, se levado em consideração arredondamentos, são os mesmos calculados manualmente na Tabela 3 que se encontra no tópico 2.4.1. Foi obtido um valor de conversão atribuído de 0.25, 0.375 e 0.375 e um efeito de remoção de 0.67, 1 e 1 para os canais S1, S2 e S3, respectivamente. Já o grafo markoviano gerado a partir do pacote se

assemelha bastante ao grafo construído manualmente apresentado na Figura 5 do tópico 2.4, com as mesmas probabilidades de transição.

Com isso, constata-se que os resultados obtidos com os pacotes *ChannelAttribution* e *markovchain* em linguagem R são de fato os mesmos calculados manualmente seguindo a abordagem proposta por Anderl et al. (2014), tanto para o valor de conversão atribuído quanto para os efeitos de remoção. A partir disso, é possível seguir com segurança para análises envolvendo conjuntos de dados maiores e mais complexos.

4.3 CARACTERIZAÇÃO DOS DADOS UTILIZADOS

O conjunto de dados utilizados para realizar as simulações se baseiam nas campanhas de marketing de uma empresa que opera na indústria de hospedagem de sites online. A empresa comercializa produtos exclusivamente por canais online, portanto não existem lojas físicas e os esforços são voltados para atrair clientes a partir de canais digitais.

A padronização dos dados das jornadas dos clientes, para serem utilizadas como entrada no modelo, exige um tratamento minucioso e um trabalho extensivo de limpeza, transformação e integração desses dados, o que acaba se tornando um processo demorado. Por este motivo, um conjunto de dados de jornadas de clientes é gerado aleatoriamente, simulando jornadas reais, onde o cliente transita pelas diversas campanhas da empresa, gerando ou não uma conversão ao final de cada jornada.

Vale ressaltar que as campanhas escolhidas são baseadas nas já existentes dentro da empresa, o que posteriormente facilita com que os dados sejam adaptados para jornadas reais. O Quadro 12 mostra quais os canais de marketing utilizados, uma abreviação de quatro letras do nome do canal, apenas para fins de simplificação das jornadas e uma breve descrição do que se trata cada canal.

Quadro 12 - Canais de marketing considerados

Canal	Abreviação	Descrição
Paid Search	paid	Publicidade de busca paga, canal onde a empresa paga aos mecanismos de busca como Google e Bing para que seus anúncios sejam colocados em uma posição mais elevada nas páginas de busca. O Pagamento Por Clique (PPC) é a forma mais comum.
Organic Search	orgc	Canal de busca orgânica, onde o cliente acessa normalmente a página da empresa a partir de mecanismos de busca listados de forma gratuita, sem que a empresa pague pelos cliques.
Direct	dirc	Canal onde o cliente acessa a página da empresa diretamente pelo navegador, inserindo a URL da página, não utilizando mecanismos de busca de terceiros.
Social Network	socl	Tráfego pago provindo de mídias sociais como Facebook, Instagram, YouTube e outros.
Referral	refe	Tráfego gerado a partir de sites de terceiros, que não são de mecanismos de busca ou mídias sociais.
Display	disp	Tráfego provindo de anúncios gráficos em estilo *banner*, podem ser exibidos em diferentes sites.
Email	mail	Tráfego gerado a partir de publicidade ou comunicação via email.
Other	othr	Outros tipos de tráfegos que não se encaixam nos canais já mencionados.

Fonte: Elaborado pelo autor

Para criação das jornadas fictícias, foi levado em consideração a quantidade usual de tráfego em cada tipo de canal. Naturalmente alguns canais têm mais acessos que outros, então quantidades aleatórias de visitas foram estimadas levando em conta a existência dessa diferença, com um limite máximo de 8000 acessos. A Tabela 4 mostra o número de visitas estimado para cada canal, com base em observações feitas nos canais de marketing da empresa para um período de dois meses.

Tabela 4 - Tráfego em cada canal

Canal	Número de visitas
refe	7862
socl	7688
orgc	5115
paid	4944
disp	2189
othr	1008
dirc	825
mail	201

Fonte: Elaborado pelo autor

Com esses dados de tráfego, pode-se construir cada jornada individual de forma aleatória, respeitando o número total de visitas, a fim de preservar a coerência dos dados. Os seguintes pontos foram considerados para a criação da base de dados que é analisada nos próximos tópicos:

a) cada jornada pode acontecer mais de uma vez, gerando ou não uma conversão;
b) quando uma conversão é gerada, um valor monetário é atribuído a ela, com um limite máximo de 2000 unidades monetárias por conversão;
c) caso a conversão não aconteça para aquela jornada, nenhum valor monetário é considerado;
d) um mesmo canal pode ser revisitado mais de uma vez em sequência, não é obrigatório que o cliente transite apenas para canais distintos.

Com isso, 5000 jornadas diferentes foram randomizadas com o auxílio do software Microsoft Excel. O conjunto de dados possui quatro colunas: *path, total_conversions, total_conversion_value, total_null*. Essas quatro colunas representam as jornadas do cliente, o número total de conversões geradas por essa jornada, o valor monetário gerado a partir dessas conversões e o número total de não conversões, respectivamente. A Figura 12 mostra uma prévia de como o conjunto de dados foi organizado e o link para baixa-lo na íntegra encontra-se no Apêndice D, ao final deste trabalho.

Figura 12 - Conjunto de dados randomizado

	A	B	C	D
1	path	total_conversions	total_conversion_value	total_null
2	orgc > refe > paid > orgc	1	2.44	3
3	refe > refe > refe > refe	2	31.95	6
4	paid > refe > paid > paid > paid > refe > paid > paid > paid > paid > paid	2	67.54	6
5	socl > orgc	1	24.02	3
6	refe > orgc > socl > disp > disp > socl > disp	0	0	2
7	paid > orgc > paid > paid > orgc > refe > refe > refe > paid > paid > paid > paid > paid	1	50.44	3
8	refe > socl > orgc > socl > orgc	0	0	2
9	socl > refe > othr	1	68.2	2
10	orgc	1	33	3
11	refe > refe > socl	1	26	3
12	refe > refe > refe	10	307.35	38
13	socl > orgc	1	34.92	3

Fonte: Elaborado pelo autor

Em posse do conjunto estruturado e devidamente compreendido, é possível aplicar o modelo de Markov e analisar os resultados obtidos. O tópico seguinte

explica como são feitas as simulações e quais percepções podem ser extraídas do valor de conversão atribuído para cada canal, dos efeitos de remoção e das probabilidades de transição.

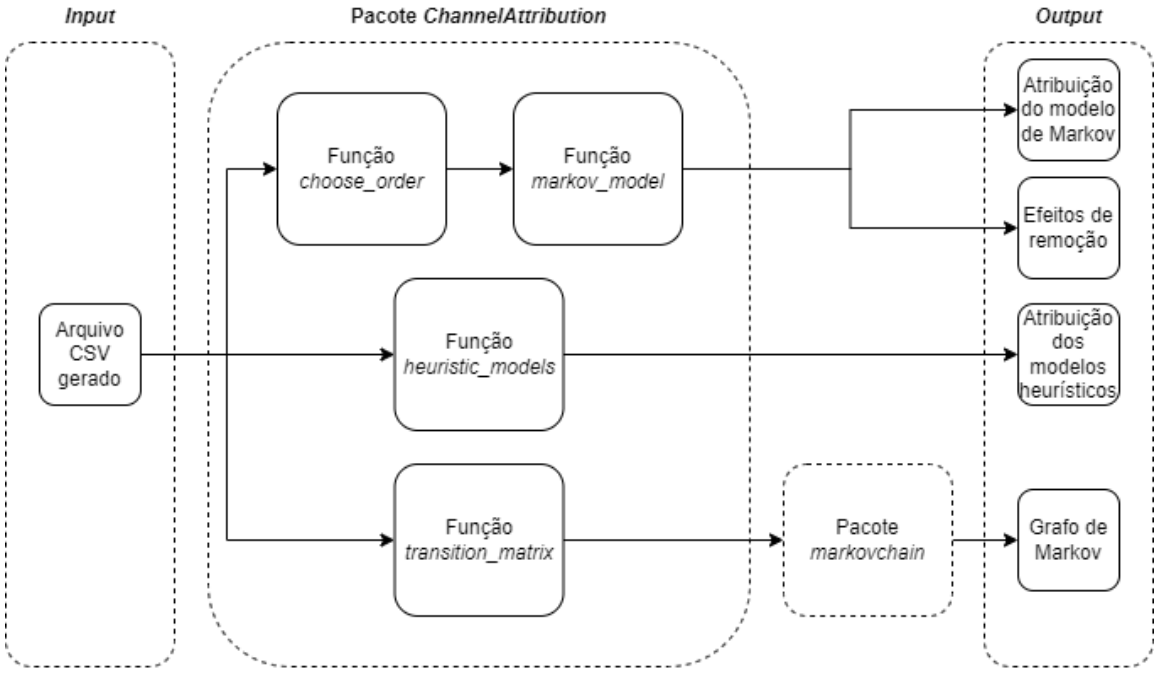
4.4 SIMULAÇÕES

Neste tópico, o modelo de Markov é avaliado através de simulações. A definição de simulação considerada neste trabalho é aquela proposta por White e Ingalls (2015) em sua obra, que conceitua simulação como uma abordagem particular para estudar modelos, fundamentalmente experimental, fornecendo uma representação imitativa de um processo ou sistema que poderia existir no mundo real. A partir disso, busca-se inferir comportamentos das jornadas de clientes, com base nos resultados das simulações, trabalhando em condições semelhantes às reais e controlando variáveis que podem ser ajustadas.

Para ser utilizado como entrada do modelo, o conjunto de dados apresentado no tópico anterior foi salvo como um arquivo CSV (*Comma Separated Values)*, formato amplamente adotado, que simplifica todas as colunas em uma única, separando os registros por vírgula. Essa conversão é feita de forma simples e automatizada no Microsoft Excel.

De forma análoga ao processo ilustrado na Figura 8, busca-se obter os mesmos *outputs* apresentados no teste de funcionalidade do tópico 4.2.1, porém dessa vez utilizando um conjunto de dados maior, escolhendo uma ordem da cadeia de Markov mais eficiente a partir da função *choose_order* e fazendo comparativos com modelos heurísticos a partir da função *heuristic_models*. A Figura 13 ilustra o processo seguido.

Figura 13 - Processo utilizado para realizar as simulações



Fonte: Elaborado pelo autor

4.4.1 Resultados

A função *choose_order*, a partir dos métodos de curva ROC e AUC, trouxe uma ordem ótima de 5 para os dados em questão. Com isso, um parâmetro de ordem 5 foi definido para a função *markov_model* e o resultado de atribuição do modelo encontra-se na Tabela 5, que representa a quantidade de conversões geradas por cada canal, em conjunto com o faturamento atribuído pelo modelo em unidades monetárias, de acordo com o valor de conversão de cada jornada. Observa-se que as jornadas trouxeram um total de 8.193 conversões que representam um faturamento total de 319.849 unidades monetárias. Os efeitos de remoção retornados encontram-se na Tabela 6.

Tabela 5 - Resultado de atribuição do modelo de Markov

Canal	Conversões atribuídas	Faturamento atribuído
Organic Search	1.678	64.266
Referral	1.543	63.160
Paid Search	2.002	73.254
Social Network	1.695	68.187
Display	598	24.704
Other	303	11.769
Direct	289	12.656
Email	85	1.852
TOTAL	8.193	319.849

Fonte: Elaborado pelo autor

Tabela 6 - Efeitos de remoção do modelo de Markov

Canal	Efeitos de remoção (%)
Organic Search	39,76%
Referral	36,56%
Paid Search	47,44%
Social Network	40,17%
Display	14,17%
Other	7,18%
Direct	6,86%
Email	2,01%

Fonte: Elaborado pelo autor

O resultado do número de conversões do modelo de Markov foi agregado ao *output* da função *heuristic_model* para fins comparativos. Cada canal recebeu uma quantidade atribuída de conversões de acordo com as regras dos modelos em específico, essas atribuições encontram-se na Tabela 7.

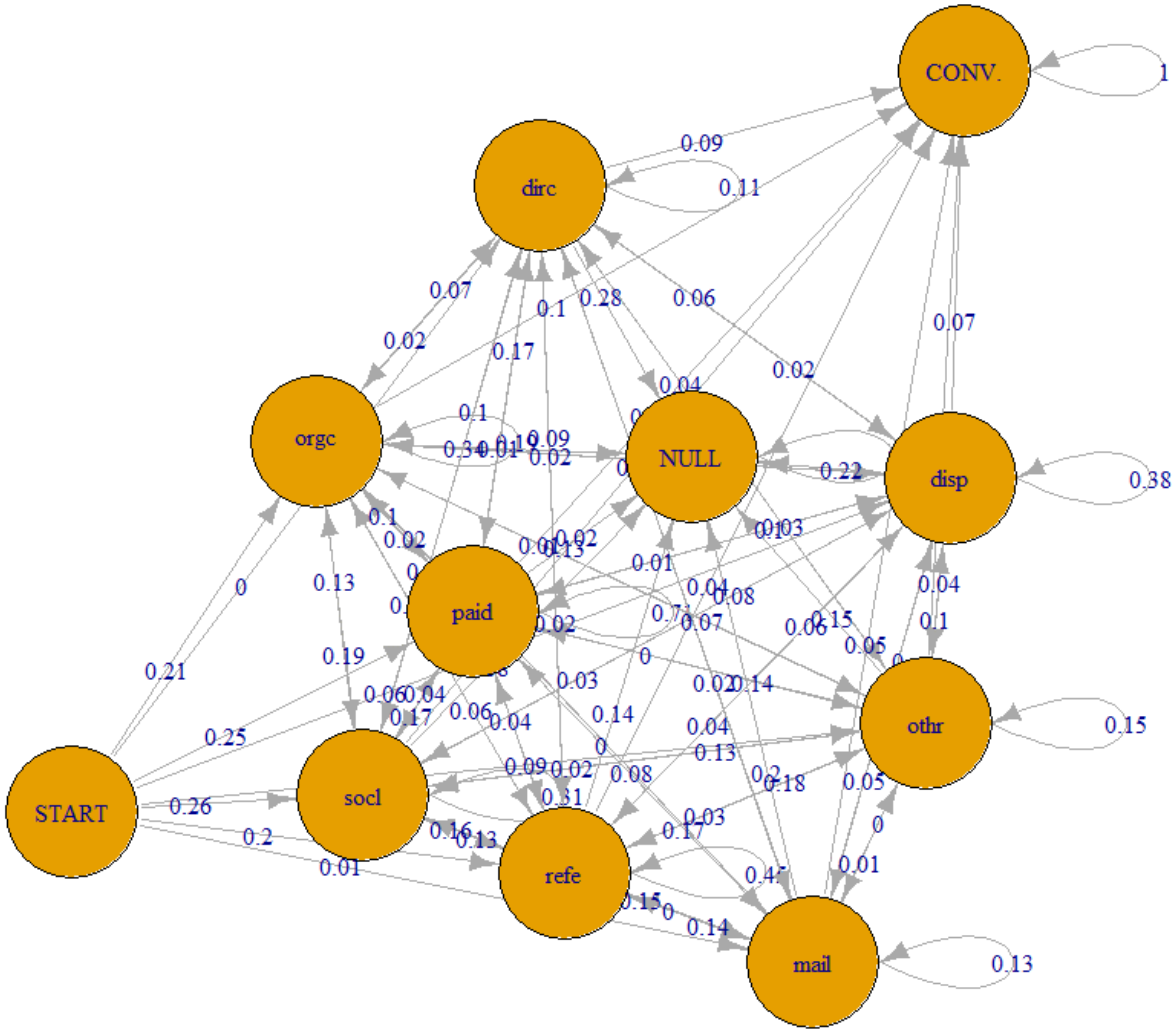
Tabela 7 - Conversões atribuídas por cada modelo

Canal	First touch	Last touch	Linear touch	Markov model
Organic Search	1.712	2.194	1.890	1.678
Referral	1.692	1.363	1.484	1.543
Paid Search	1.876	2.604	2.324	2.002
Social Network	2.161	766	1.497	1.695
Display	571	745	640	598
Other	61	178	144	303
Direct	37	296	148	289
Email	83	47	65	85
TOTAL	8.193	8.193	8.193	8.193

Fonte: Elaborado pelo autor

Na Figura 14 é apresentado o grafo markoviano do modelo, construído levando em consideração uma cadeia de Markov de primeira ordem.

Figura 14 - Grafo do modelo de Markov



Fonte: *Output* software R Studio

Apesar de ser possível representar graficamente uma cadeia de Markov de ordem mais elevada, nota-se que, devido a quantidade de estados e transições, o grafo acaba ficando bastante poluído mesmo em uma cadeia de primeira ordem, o que prejudica a interpretação dos resultados. Por esse motivo, uma alternativa de representação das probabilidades de transição é proposta no tópico seguinte, em conjunto com as discussões dos resultados obtidos.

Os códigos em linguagem R para obtenção dos resultados do tópico atual encontram-se no Apêndice B, com comentários explicativos das linhas de execução.

4.5 ANÁLISES E DISCUSSÕES

Neste tópico, são discutidas análises e sugestões com base nos dados de saída do modelo proposto. Por se tratar de informações geradas de forma aleatória em uma perspectiva simulada, ressalta-se que não há dados concretos de valores monetários investidos nas campanhas apresentadas, porém esse tipo de informação se torna bastante relevante ao ser relacionada com os resultados obtidos e será levada em consideração.

4.5.1 Conversão atribuída e efeitos de remoção

A partir dos resultados individuais do modelo de Markov da Tabela 5, cruzando com dados de investimentos em campanhas de marketing da empresa, é possível calcular o custo por conversão ao dividir o valor total investido no canal pelo número de conversões atribuídas, conforme fórmula (7):

$$Custo\ por\ Conversão = \frac{Investimento\ no\ canal}{Nº\ de\ conversões\ atribuídas} \quad (7)$$

Subtraindo o valor total de investimento no canal do faturamento atribuído a ele, também pode-se comparar a margem bruta de lucro gerada por cada canal, como mostra a fórmula (8):

$$Margem\ bruta = Faturamento\ atribuído - Investimento\ no\ canal \quad (8)$$

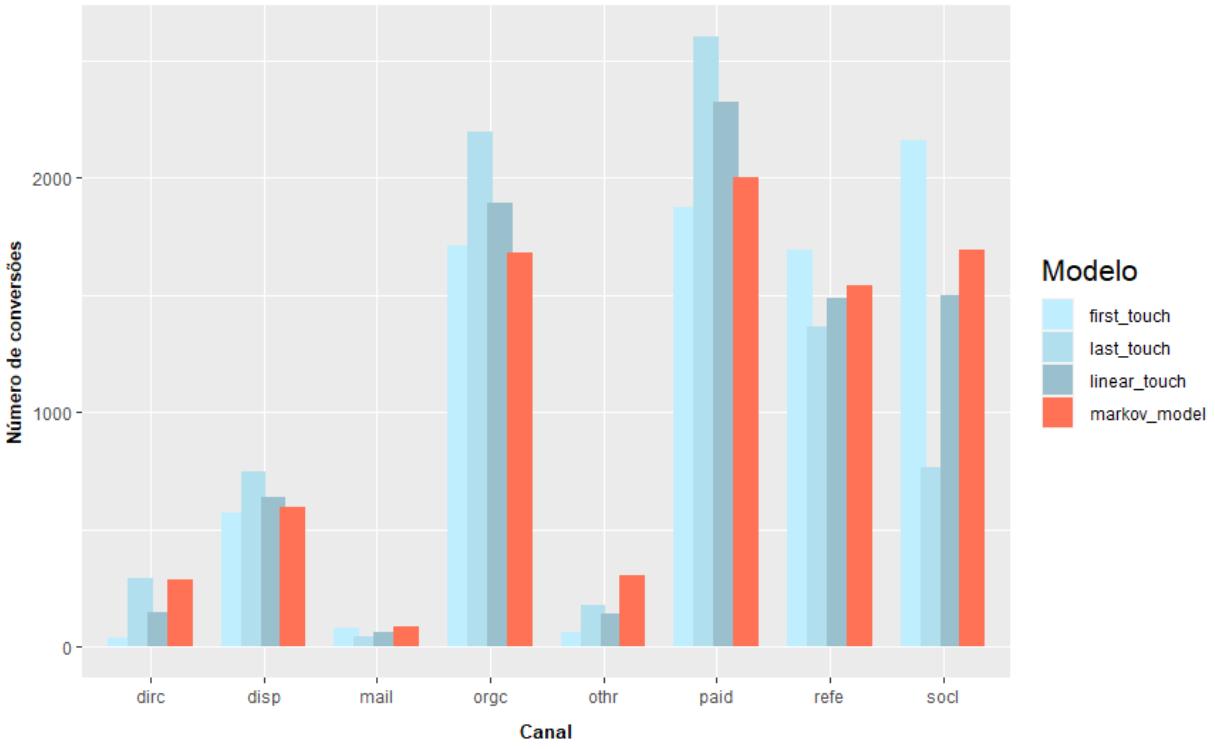
Essas duas métricas podem ser utilizadas em conjunto por tomadores de decisão para definir quais canais geram maior retorno sobre o investimento e planejar futuros investimentos.

O modelo também mostra o quanto a remoção de um único canal pode afetar a quantidade total de conversões, a partir dos dados de efeitos de remoção da Tabela 6. Um efeito de remoção alto, significa que aquele canal em específico fez parte de muitas jornadas que levaram a conversão e, consequentemente, é um canal importante para o negócio da empresa. Percebe-se que o canal *Paid Search* tem o maior efeito de remoção de 47,44%, ou seja, esse canal fez parte de praticamente metade de todas as jornadas que levaram a uma conversão, portanto é um canal de extrema importância para as vendas neste contexto, enquanto o canal *Email* praticamente não teve participação em jornadas que converteram, com um efeito de remoção de apenas 2%.

4.5.2 Comparação com modelos heurísticos

Outra análise comparativa importante é a diferença das conversões atribuídas entre os modelos heurísticos e o modelo markoviano. A fim de visualizar melhor as diferenças de atribuições, foi gerado um gráfico de barras com o auxílio do pacote *ggplot2* no próprio R Studio, a partir dos dados da Tabela 7. A Figura 15 mostra esse comparativo, onde as barras em tom de azul claro representam os resultados dos modelos heurísticos e as barras laranjas os resultados do modelo de Markov.

Figura 15 - Gráfico comparativo dos modelos



Fonte: Elaborado pelo autor

Observa-se que os canais com mais conversões atribuídas pelo modelo de Markov são justamente os mesmos que possuem os maiores efeitos de remoção, o que já era esperado visto que o efeito de remoção é um passo antecedente obrigatório no cálculo da atribuição do modelo, conforme proposto no trabalho de Anderl et al. (2014).

Entretanto, se observados os modelos heurísticos, percebe-se que existem maiores variações. No modelo *first touch*, por exemplo, o canal de Social Network possui o maior número de conversões entre todos os canais, já no modelo *last touch*

esse canal é o quarto maior. Esse padrão acontece comumente nos modelos heurísticos, visto que a simplicidade na regra acaba por ignorar interações com outros canais, não representando um comportamento próximo da realidade, como já constatado por Barajas e Akella (2016). O modelo de Markov é proposto com o intuito de calibrar de forma mais precisa as atribuições, não negligenciando outros canais que fazem parte da jornada como um todo.

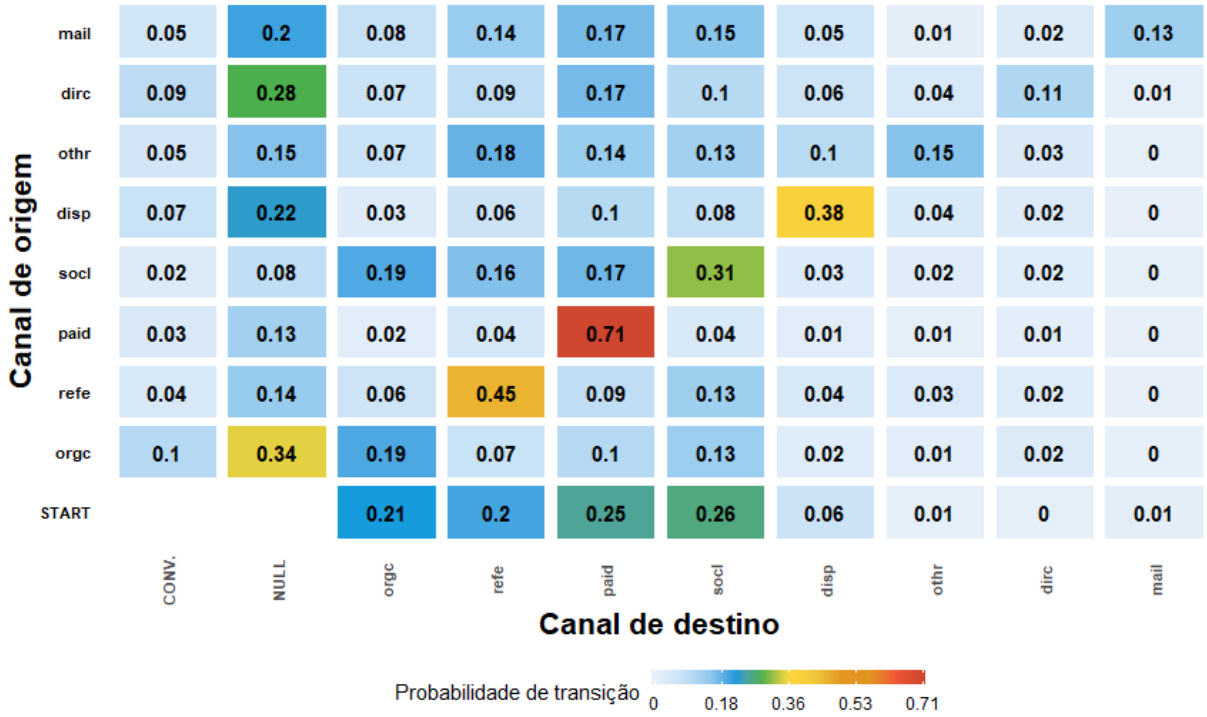
Neste exemplo em específico, ao observar os resultados das conversões no canal *Social Network* por uma perspectiva do modelo *first touch*, a empresa poderia estar investindo mais recursos nele de forma menos eficiente, pois o modelo *first touch* superestima a capacidade de conversão deste canal, enquanto o modelo de Markov considera o canal *Paid Search* como o maior gerador de conversões.

Ao enxergar um cenário de conversões mais realista com o modelo markoviano, a empresa pode direcionar investimentos aos canais que melhor performam, no intuito de trazer mais conversões. O modelo pode ser reaplicado de forma recorrente, reavaliando as novas jornadas em períodos pré-definidos, para trazer mais conversões com menos recursos. A longo prazo, isso gera uma economia monetária substancial para a empresa.

4.5.3 Probabilidades de transição

O grafo de Markov da Figura 14, devido a quantidade elevada de estados, dificulta a visualização e análises das probabilidades de transição entre os canais. Por este motivo, um mapa de calor é construído visando uma melhor interpretação, utilizando novamente o pacote *ggplot2* do R Studio a partir da matriz de transição do modelo, com base em um mapa de calor já criado por Bryl (2016) em suas análises. Nele é possível observar a probabilidade de um cliente transitar de um canal a outro, de forma similar ao grafo de Markov, porém em uma configuração menos poluída e com auxílio visual de intensidade de cores. O mapa encontra-se na Figura 16.

Figura 16 - Mapa de calor da matriz de transição

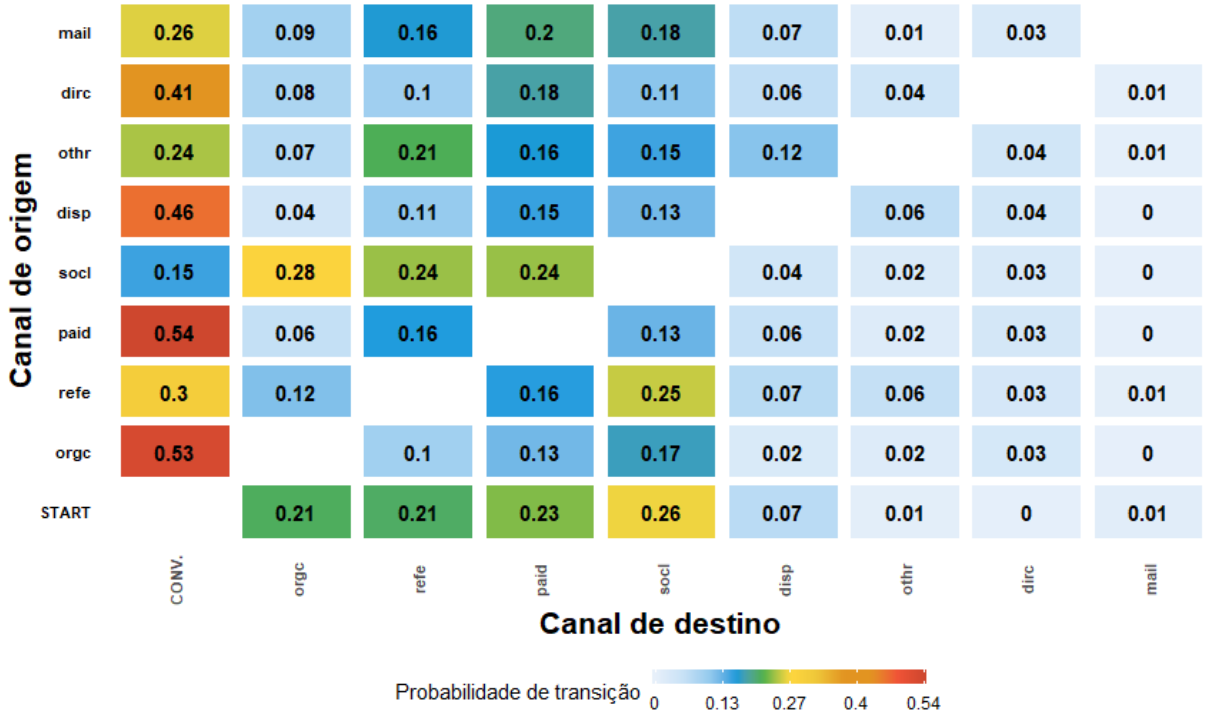


Fonte: Elaborado pelo autor

Primeiramente, ao analisar a diagonal da matriz, nota-se que a probabilidade de um canal transitar para ele mesmo é bastante alta, isso se deve a quantidade elevada de jornadas onde o cliente revisita o mesmo canal diversas vezes consecutivas (por exemplo paid>paid>paid>paid). Também se observa que as probabilidades de transição da coluna NULL são maiores que as da coluna CONV, isso indica que houve mais jornadas que não converteram do que jornadas que geraram uma conversão, o que se confirma observando o total de não conversões do conjunto de dados criado, onde existem 8193 conversões e 28985 não conversões.

Para uma análise direcionada a conversões, é possível encurtar as jornadas com sequencias repetidas (de dirc>dirc>dirc>paid>paid para dirc>paid, por exemplo) e desconsiderar todas as jornadas que não converteram. Dessa forma, um novo mapa de calor é criado para analisar apenas as jornadas de conversão simplificadas, apresentado na Figura 17.

Figura 17 - Mapa de calor simplificado



Fonte: Elaborado pelo autor

Dessa visualização, extrai-se algumas informações relevantes para o negócio, no âmbito dos dados estudados. Observando o canal de origem START e o canal de destino socl, é possível afirmar que 26% das jornadas que convertem, começam a partir do canal de *Social Network*, com essa informação, pode-se assumir que fazer investimentos em mídias sociais é uma boa forma de iniciar uma jornada com potencial de conversão. Porém, olhando por outra perspectiva, o canal *Social Network* por si só tem uma baixa probabilidade de transição para gerar uma conversão (15%), então investir somente neste canal não é uma boa opção, seria necessário combinar investimentos em outros canais com maiores probabilidades de transição para gerar uma conversão, como o canal paid, orgc ou disp.

É importante ressaltar que nem sempre um canal com uma alta probabilidade de transição para conversão, será o canal com a melhor conversão atribuída pelo modelo de Markov, pois são análises distintas que devem ser ponderadas em conjunto. Uma equipe de marketing da empresa, por exemplo, pode utilizar esses valores no planejamento de campanhas para identificar canais que normalmente trabalham juntos para gerar conversões e tornar a jornada do cliente o mais unificada possível, permitindo um controle e entendimento mais preciso do comportamento de sua base de clientes.

4.6 CONSIDERAÇÕES FINAIS DO CAPÍTULO

Para que o estudo pudesse ser aplicado com informações reais de uma empresa, seria necessário que o conjunto de dados utilizado como entrada do modelo estivesse padronizado no formato explicado nos tópicos 4.2 e 4.3, como ilustram o Quadro 7 e a Figura 12. Geralmente, para alcançar esse resultado de padronização, é necessária cooperação interfuncional entre diversas áreas da empresa, especialmente de marketing, engenharia e dados, o que acaba demandando um tempo maior de trabalho em conjunto e permeando questões burocráticas e de prioridades departamentais.

Por esse motivo foi definido trabalhar com dados fictícios, porém baseados em dados reais. Essa decisão não impacta no desenvolvimento de um modelo de atribuição flexível, mas restringe as análises e discussões que podem ser feitas em cima dos resultados, para trazer visões de negócio mais realistas e que sejam de fato aplicadas como melhorias.

Por outra perspectiva, os tempos de execução dos códigos, poder de processamento necessário para rodá-los e performance no geral não podem ser mensurados de forma realista com dados fictícios, visto que geralmente o volume de dados de tráfego de uma empresa é muito maior do que o utilizado nas simulações, isso acaba se tornando uma limitação que não foi possível ser aprofundada neste trabalho. Como parâmetro, o conjunto de dados utilizado neste estudo, contendo cinco mil jornadas fictícias distintas, tem um tempo de execução menor do que cinco segundos.

As análises e visualizações presentes no tópico 4.5, se observadas em conjunto, podem trazer percepções diversas sobre o negócio da empresa, englobando desde a otimização de recursos em canais de marketing até a compreensão do comportamento do cliente e de suas jornadas. Os códigos na linguagem de programação R foram construídos de forma que todos os gráficos e visualizações apresentados possam ser gerados de forma simples e ágil, permitindo alteração nos dados de entrada, sem muitos ajustes, para testar cenários variados, e se encontram no Apêndice C devidamente comentados.

5 CONCLUSÕES E RECOMENDAÇÕES

5.1 CONCLUSÃO

O presente trabalho buscou avaliar um modelo de atribuição de marketing mais preciso que possa ser facilmente adaptado a contextos diversos, utilizando conceitos e aplicações da ferramenta matemática de cadeias de Markov para o problema de atribuição eficiente de créditos a campanhas de marketing digital. Conclui-se que os objetivos, tanto geral quanto específicos, foram alcançados.

Uma revisão bibliográfica foi feita para entender o estado do marketing digital na atualidade e quais modelos de atribuição existentes na literatura são usualmente aplicados. Suas características, pontos positivos e negativos, diferenças entre modelos heurísticos mais simplistas e modelos algorítmicos com maior embasamento em dados, como é o caso do modelo de Markov, foram comparados a fim de justificar a escolha da ferramenta.

A partir disso, explicou-se como a ferramenta de cadeias de Markov lida com o problema de atribuição de marketing e como é possível aplicá-la de forma prática para trazer resultados mais precisos, considerando aspectos como flexibilidade, escalabilidade e reutilização. Para atingir o objetivo, o uso de uma linguagem de programação foi determinado como indispensável, a fim de suportar e gerenciar o constante aumento do volume de dados na atualidade.

Com o auxílio da linguagem de programação R escolhida, um modelo de atribuição baseado em cadeias de Markov foi proposto, implementado a partir dos pacotes *ChannelAttribution* e *Markovchain*, já disponibilizados pela comunidade na ferramenta de código aberto. O modelo foi então devidamente testado e simulações foram conduzidas, utilizando como entrada um conjunto de dados gerados aleatoriamente pelo autor, construído com base em observações de campanhas reais dentro de uma empresa do ramo de hospedagem de sites online.

Os resultados das simulações foram apresentados, destacando que de fato o modelo de Markov consegue trazer uma visão mais realista do problema de atribuição de marketing que os demais modelos heurísticos, ou seja, possui maior precisão. Sugestões de análises foram propostas, com o auxílio visual de gráficos comparativos e mapas de calor, a fim de direcionar possíveis percepções referentes a otimização de recursos e investimentos em campanhas de marketing e ao comportamento do cliente relacionado a suas interações nos diferentes canais.

Os códigos utilizados para implementar o modelo e gerar as visualizações apresentadas foram devidamente explicados e construídos de modo que adaptações possam ser aplicadas com poucos ajustes, reforçando a característica de reutilização e flexibilidade definida como parte do objetivo geral deste trabalho, permitindo que testes possam ser conduzidos para investigar cenários diversos e que o mesmo código possa ser reproduzido em outras indústrias e contextos, relacionados ao processo de atribuição de marketing.

5.2 RECOMENDAÇÕES PARA TRABALHOS FUTUROS

Ao longo do desenvolvimento do trabalho, identificou-se alguns cenários diferentes que podem ser explorados, como a inclusão de variáveis extras no conjunto de dados, que podem trazer um nível a mais de detalhamento e complexidade ao modelo e novas percepções de negócio.

É possível segmentar o conjunto de dados para fazer análises mais direcionadas, rodando o modelo individualmente para cada segmentação e observando se existem diferenças de atribuição nesses segmentos. Por exemplo, um tipo de produto em específico pode ser comercializado de forma mais eficiente em certos canais, rodando o modelo individualmente para conversões apenas de um produto específico, pode-se entender quais canais funcionam melhor nesses segmentos. Isso permite que a empresa possa focar em aumentar as conversões de um produto específico ao investir mais nos canais com maior conversão atribuída a ele.

Assim como é possível incluir informações de valor monetário para cada conversão gerada no modelo, estima-se que é possível incluir uma variável relacionada ao tempo da jornada. Uma forma de contabilizar essa variável seria a partir do número de interações que o cliente teve com cada canal ao longo de uma jornada específica, ou ainda o período em dias a partir do início da jornada até a conversão. Isso traria uma perspectiva diferente ainda não analisada neste estudo e não contemplada nas funções básicas dos pacotes utilizados.

Por fim, fugindo um pouco do tema de atribuição, mas mantendo o foco em marketing digital, é possível estruturar um modelo que utilize os mesmos princípios de cadeias de Markov para analisar a eficácia de *landing pages*, ou páginas de aterrissagem em tradução livre, termo bastante conhecido em marketing para se referir a páginas criadas com a função de direcionar os visitantes para uma ação

desejada, seja para conversões, formulários ou downloads de algum recurso. Estima-se que um modelo de Markov semelhante pode ser utilizado para testar qual *landing page* performa melhor em relação a outra, de forma similar a como testes A/B são conduzidos para este fim atualmente.

REFERÊNCIAS

ALTOMARE, Davide. **Markov Model for Online Multi-Channel Attribution**. 2023. Disponível em: https://cran.r-project.org/web/packages/ChannelAttribution/ChannelAttribution.pdf. Acesso em: 20 jun. 2023.

ANDERL, Eva et al. **Mapping the customer journey: A graph-based framework for online attribution modeling**. 2014. Disponível em: http://ssrn.com/abstract=2343077. Acesso em: 10 mai. 2023.

AMAZON. **What is marketing attribution? A beginner's guide**. [S.l.], c2023. Disponível em: https://advertising.amazon.com/library/guides/marketing-attribution. Acesso em: 2 abr. 2023.

ARENALES, M. et al. **Pesquisa Operacional**. Rio de Janeiro: Campus-Elsevier, 2011.

BARAJAS, J.; AKELLA, R. **Experimental Designs and Estimation for Online Display Advertising Attribution in Marketplaces**. Marketing Science, v. 35 n. 3, p. 465-483, 2016.

BEUREN, Ilse Maria; RAUPP, Fabiano Maury. **Metodologia da Pesquisa Aplicável às Ciências Sociais**. São Paulo: Atlas, 2006.

BRYL, Serhii. **Marketing Multi-Channel Attribution model with R (part 1: Markov chains concept)**. Analyzecore, 2016. Disponível em: https://www.analyzecore.com/2016/08/03/attribution-model-r-part-1/. Acesso em: 4 set. 2023.

CALDWELL, Alex. **Top 5 Web Analytics Tools for Improving Site Performance**. DevPro Journal, 2023. Disponível em: https://www.devprojournal.com/software-development-trends/marketing/top-5-web-analytics-tools-for-improving-site-performance/. Acesso em: 20 jun. 2023.

CALLADINE, Dan. **2023 Global Ad Spend Forecasts**. Dentsu, [S.l.], 2022.

CRAN. **The Comprehensive R Archive Network**. c2023. Disponível em: https://cran.r-project.org/. Acesso em: 28 set. 2023.

DALESSANDRO, B. et al. **Causally motivated attribution for online advertising**. Pequim, ACM, 2012.

ECONSULTANCY. **The State of Marketing Attribution 2017**. AdRoll, 2017. Disponível em: https://www.adroll.com/assets/pdf/guides-and-reports/AdRoll-State-of-Marketing-Attribution-2017.pdf. Acesso em: 10 abr. 2023.

GAUR, Jitendra; BHARTI, Kumkum. **Attribution Modelling in Marketing: Literature review and research agenda**. Academy of Marketing Studies Journal, n 24, p. 1-21, 2020.

GIL, Antonio Carlos. **Como elaborar projetos de pesquisa**. 6. ed. São Paulo: Atlas, 2017.

HUBSPOT. **Compreenda os relatórios de atribuição**. 2023. Disponível em: https://knowledge.hubspot.com/pt/reports/understand-attribution-reporting. Acesso em: 20 set. 2023.

IAB. **Internet Advertising Revenue Report 2022**. 2023. Disponível em: https://www.iab.com/wp-content/uploads/2023/04/IAB_PwC_Internet_Advertising_Revenue_Report_2022.pdf. Acesso em: 15 jun. 2023.

IAB BRASIL. **Modelos de atribuição em publicidade digital**. [S.l.]. 2018. E-book.

IBM. **Python vs. R: What's the Difference?**. IBM Cloud Team, 2021. Disponível em: https://www.ibm.com/blog/python-vs-r/. Acesso em: 20 set. 2023.

JAYAWARDANE, C. H. W.; HALGAMUGE, S. K.; KAYANDE, U. **Attributing Conversion Credit in an Online Environment: An Analysis and Classification**. 3rd International Symposium on Computational and Business Intelligence (ISCBI), IEEE, p. 68-73, 2015.

KAKALEJCIK, Lucas et al. **Multichannel Marketing Attribution Using Markov Chains**. Journal of Applied Management and Investments, v. 7, n. 1, p. 49-60, 2018.

KOVACS, Leandro. **O que é open source?**. Tecnoblog, 2021. Disponível em: https://tecnoblog.net/responde/o-que-e-open-source-software-de-codigo-aberto/. Acesso em: 20 set. 2023.

MEYN, S. P.; TWEEDIE, R. L. **Markov Chains and Stochastic Stability**. Cambridge University Press, 2 ed. 2009. Disponível em: https://ericmoulines.files.wordpress.com/2014/03/meyntweedie2009.pdf. Acesso em: 11 dez. 2023.

MIGUEL, Paulo Augusto Cauchick (org.). **Metodologia de Pesquisa em Engenharia de Produção e Gestão de Operações**. 2. ed. Rio de Janeiro: Campus-Elsevier, 2012.

MIT Technology Review. **Marketing digital na América Latina: o impulso no ecossistema Martech**. 2022. Disponível em: https://mittechreview.com.br/marketing-digital-na-america-latina-o-impulso-no-ecossistema-martech/. Acesso em: 20 jul. 2023.

POUTANEN, Riku. **Analysis of online advertisement performance using Markov chains: Case Fiksu Rouka Oy**. 2020. 59 p. Tese (Mestrado em Engenharia Industrial) – Tampere University, Tampere, 2020.

R Project. **What is R?**. c2023. Disponível em: https://www.r-project.org/about.html. Acesso em: 22 set. 2023.

REIS, Fábio dos. **7 Linguagens de programação para Ciência de Dados**. 2021. Disponível em: http://www.bosontreinamentos.com.br/ciencia-de-dados/7-linguagens-de-programacao-para-ciencia-de-dados-2021. Acesso em: 25 set. 2023.

RENTOLA, Olli. **Analyses of Online Advertising Performance Using Attribution Modeling**. 2014. 78 p. Tese (Mestrado em Sistemas e Pesquisa Operacional) – Aalto University, Espoo, 2014.

SHAO, Xuhui; LI, Lexin. **Data-driven Multi-touch Attribution Models.** San Diego, ACM SIGKDD, 2011.

SHARMA, Himanshu. **Attribution Modelling in Google Analytics and Beyond**. [S.l.: s.n.], 2016.

SPEDICATO, Giorgio Alfredo. **Easy Handling Discrete Time Markov Chains**. 2023. Disponível em: https://cran.r-project.org/web/packages/markovchain/markovchain.pdf. Acesso em: 2 out. 2023.

WHITE, Preston; INGALLS, Ricki G. **Introduction to simulation**. California, Winter Simulation Conference (WSC), 2015. Disponível em: https://www.researchgate.net/publication/302479579_Introduction_to_simulation. Acesso em: 20 out. 2023.

ZAREMBA, Arkadiusz. **Conversion Attribution: What Is Missed by the Advertising Industry? The OPEC Model and Its Consequences for Media Mix Modeling**. Journal of Marketing and Consumer Behaviour in Emerging Markets, v. 1 n. 10, p. 4-23, 2020.

ZHAO, K.; SEYED, H. M.; SAEED, R. B. **Shapley Value Methods for Attribution Modeling in Online Advertising**. 2018. Disponível em: https://arxiv.org/abs/1804.05327. Acesso em: 19 jun. 2023.

APÊNDICE A – Código R utilizado no tópico 4.2.1

```
library(ChannelAttribution)
library(markovchain)
library(igraph)

# Jornadas do tópico 2.4
my_data = data.frame(path = c('S1 > S2 > S3', 'S2', 'S1 > S2'),
                     total_conversions = c(1, 0, 0),
                     total_null = c(0, 1, 1))

# Aplica a função transition_matrix
df_trans = transition_matrix(Data = my_data,
                             var_path = 'path',
                             var_conv = 'total_conversions',
                             var_null = 'total_null')

df_trans = df_trans$transition_matrix

# Aplicam-se transformações necessárias para pivotar a matriz de transição
# Cria um data frame com novas colunas de probabilidade de transição de 100%
para null>null e conversion>conversion
# OBS: isso é necessário para que o data frame possa ser transformado em um
grafo markoviano
new_rows = data.frame(channel_from = c("(null)", "(conversion)"),
                      channel_to = c("(null)", "(conversion)"),
                      transition_probability = c(1, 1))

# Unifica-se os dois data frames
df_trans = rbind(df_trans, new_rows)

# Cria um grafo com a library(igraph) a partir do data frame
graph = graph.data.frame(df_trans, directed = TRUE)

# Converte o grafo para uma matriz de transição
transition_matrix = as_adjacency_matrix(graph, attr =
"transition_probability", sparse = FALSE)

# Renomeia os estados para ficar mais condizente com o exemplo
colnames(transition_matrix) = c("START", "S1", "S2", "S3", "NULL", "CONV.")
rownames(transition_matrix) = c("START", "S1", "S2", "S3", "NULL", "CONV.")

# Cria um objeto "markovchain"
mc = new("markovchain", transitionMatrix = transition_matrix)
```

```
# Cria um layout (posicionamento) específico para representar os estados do
grafo
layout = matrix(c(0,1,0.5,0,1,1,2,1,3,0,3,1), ncol = 2, byrow = TRUE)

# Plota o objeto "markovchain"
plot(mc, edge.arrow.size = 0.50, vertex.size = 30, layout = layout)

# Aplica o modelo markoviano no data frame
m_mod = markov_model(my_data,
                     var_path = 'path',
                     var_conv = 'total_conversions',
                     var_null = 'total_null',
                     order = 1,
                     out_more = TRUE)

# Extrai os resultados do modelo separadamente
attribution_result = m_mod$result
removal_effects = m_mod$removal_effects
```

APÊNDICE B – Código R utilizado no tópico 4.4.1

```
library(ChannelAttribution)
library(markovchain)
library(igraph)

# Define o diretório onde o arquivo CSV está armazenado
setwd('C:/xxxx')

my_data = read.csv('Jornadas.csv')

# Escolhe a ordem da cadeia (Resultado = 5 neste caso)
res = choose_order(my_data,var_path = 'path',var_conv =
'total_conversions',var_null = 'total_null', plot=FALSE)

# Aplica o modelo markoviano no dataframe
m_mod = markov_model(my_data,
                     var_path = 'path',
                     var_conv = 'total_conversions',
                     var_null = 'total_null',
                     # var_value = 'total_conversion_value',
                     order = 5,
                     out_more = TRUE)

# Extrai os resultados do modelo separadamente
attribution_result = m_mod$result
removal_effects = m_mod$removal_effects


# Aplica os modelos heurísticos
h_mod = heuristic_models(my_data,
                         var_path = 'path',
                         var_conv = 'total_conversions')

# Unifica o resultado dos modelos heuristicos e markoviano no mesmo local
merged_results = merge(h_mod,attribution_result, by='channel_name')
```

```
# ----------------- GRAFO MARKOVIANO -----------------

# Aplica a função transition_matrix
df_trans = transition_matrix(Data = my_data,
                             var_path = 'path',
                             var_conv = 'total_conversions',
                             var_null = 'total_null')

df_trans = df_trans$transition_matrix

# Aplicam-se transformações necessárias para pivotar a matriz de transição
# Cria um data frame com novas colunas de probabilidade de transição de 100%
para null>null e conversion>conversion
# OBS: isso é necessário para que o data frame possa ser transformado em um
grafo markoviano
new_rows = data.frame(channel_from = c("(null)", "(conversion)"),
                      channel_to = c("(null)", "(conversion)"),
                      transition_probability = c(1, 1))

# Unifica-se os dois data frames
df_trans = rbind(df_trans, new_rows)

# Cria um graph com a library(igraph) a partir do data frame
graph = graph.data.frame(df_trans, directed = TRUE)

# Converte o grafo para uma matriz de transição
transition_matrix = as_adjacency_matrix(graph, attr =
"transition_probability", sparse = FALSE)

# Renomeia os estados
colnames(transition_matrix) = c("START", "orgc", "refe", "paid", "socl",
"disp", "othr", "dirc", "mail", "NULL", "CONV.")
rownames(transition_matrix) = c("START", "orgc", "refe", "paid", "socl",
"disp", "othr", "dirc", "mail", "NULL", "CONV.")

# Cria um objeto "markovchain"
mc = new("markovchain", transitionMatrix = transition_matrix)

# Plota o objeto "markovchain"
plot(mc, edge.arrow.size = 0.50, vertex.size = 30)
```

APÊNDICE C – Código R utilizado no tópico 4.5

```
library(ChannelAttribution)
library(markovchain)
library(igraph)
library(reshape2)
library(ggplot2)

# Define o diretório onde o arquivo CSV está armazenado
setwd('C:/xxxx')

my_data = read.csv('Jornadas.csv')

# -------------- GRAFICO COMPARATIVO HEURISTIC VS MARKOV --------------

# Aplica o modelo markoviano no dataframe
m_mod = markov_model(my_data,
                     var_path = 'path',
                     var_conv = 'total_conversions',
                     var_null = 'total_null',
                     # var_value = 'total_conversion_value',
                     order = 5,
                     out_more = TRUE)

# Extrai os resultados do modelo separadamente
attribution_result = m_mod$result
removal_effects = m_mod$removal_effects

# Aplica os modelos heurísticos
h_mod = heuristic_models(my_data,
                         var_path = 'path',
                         var_conv = 'total_conversions')

# Unifica o resultado dos modelos heuristicos e markoviano no mesmo local
merged_results = merge(h_mod,attribution_result, by='channel_name')

# Ajusta o nome das colunas
colnames(merged_results) = c('channel_name', 'first_touch', 'last_touch',
'linear_touch', 'markov_model')

# Pivota os resultados para poder construir um gráfico
merged_results = melt(merged_results, id='channel_name')
```

```
# Constrói um gráfico com ggplot2
custom_colors = c("lightblue1", "lightblue2", "lightblue3", "coral1")

ggplot(merged_results, aes(channel_name, value, fill=variable)) +
  geom_bar(stat='identity', position = position_dodge(width = 0.7)) +
  ggtitle('Modelos Heurísticos vs Modelo Markoviano') +
  theme(axis.title.x = element_text(size = 10, face = "bold", vjust = -2)) +
  theme(axis.title.y = element_text(size = 10, face = "bold", vjust = +2)) +
  theme(title = element_text(size = 16)) +
  theme(plot.title = element_text(size = 20)) +
  xlab("Canal") + ylab("Número de conversões") +
  scale_fill_manual(values = custom_colors) +
  labs(fill = "Modelo")
```

```
# -------------------- TRANSITION MATRIX HEATMAP --------------------------
df_trans_hm = transition_matrix(my_data,
                        var_path = 'path',
                        var_conv = 'total_conversions',
                        var_null = 'total_null',
                        flg_equal = TRUE)

df_plot_trans = df_trans_hm$transition_matrix

cols = c("#e7f0fa", "#c9e2f6", "#95cbee", "#0099dc", "#4ab04a", "#ffd73e",
"#eec73a","#e29421", "#e29421", "#f05336", "#ce472e")

t = max(df_plot_trans$transition_probability)

ggplot(df_plot_trans, aes(y = channel_from, x = channel_to, fill =
transition_probability)) +
  theme_minimal() +
  geom_tile(colour = "white", width = .9, height = .9) +
  scale_fill_gradientn(colours = cols, limits = c(0, t),
                       breaks = seq(0, t, by = t/4),
                       labels = c("0", round(t/4*1, 2), round(t/4*2, 2),
round(t/4*3, 2), round(t/4*4, 2)),
                       guide = guide_colourbar(ticks = T, nbin = 50, barheight
= .5, label = T, barwidth = 10)) +
  geom_text(aes(label = round(transition_probability, 2)), fontface = "bold",
size = 4) +
  labs(x = "Canal de destino", y = "Canal de origem", fill = 'Probabilidade de
transição') +
  scale_x_discrete(labels = c("CONV.", "NULL", "orgc", "refe", "paid", "socl",
"disp", "othr", "dirc", "mail")) +
  scale_y_discrete(labels = c("START", "orgc", "refe", "paid", "socl", "disp",
"othr", "dirc", "mail")) +
  theme(legend.position = 'bottom',
        legend.direction = "horizontal",
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        plot.title = element_text(size = 20, face = "bold", vjust = 2, color =
'black', lineheight = 0.8),
        axis.title.x = element_text(size = 16, face = "bold"),
        axis.title.y = element_text(size = 16, face = "bold"),
        axis.text.y = element_text(size = 8, face = "bold", color = 'black'),
        axis.text.x = element_text(size = 8, angle = 90, hjust = 0.5, vjust =
0.5, face = "bold")) +
  ggtitle("Mapa de calor da matriz de transição")
```

```r
# -------------------- TRANSITION MATRIX HEATMAP SIMPLIFICADO ----------------

df_trans_hm2 = transition_matrix(my_data,
                                 var_path = 'path',
                                 var_conv = 'total_conversions',
                                 var_null = NULL,
                                 flg_equal = FALSE)


df_plot_trans = df_trans_hm2$transition_matrix

cols = c("#e7f0fa", "#c9e2f6", "#95cbee", "#0099dc", "#4ab04a", "#ffd73e",
"#eec73a","#e29421", "#e29421", "#f05336", "#ce472e")

t = max(df_plot_trans$transition_probability)

ggplot(df_plot_trans, aes(y = channel_from, x = channel_to, fill =
transition_probability)) +
  theme_minimal() +
  geom_tile(colour = "white", width = .9, height = .9) +
  scale_fill_gradientn(colours = cols, limits = c(0, t),
                       breaks = seq(0, t, by = t/4),
                       labels = c("0", round(t/4*1, 2), round(t/4*2, 2),
round(t/4*3, 2), round(t/4*4, 2)),
                       guide = guide_colourbar(ticks = T, nbin = 50, barheight
= .5, label = T, barwidth = 10)) +
  geom_text(aes(label = round(transition_probability, 2)), fontface = "bold",
size = 4) +
  labs(x = "Canal de destino", y = "Canal de origem", fill = 'Probabilidade de
transição') +
  scale_x_discrete(labels = c("CONV.", "orgc", "refe", "paid", "socl", "disp",
"othr", "dirc", "mail")) +
  scale_y_discrete(labels = c("START", "orgc", "refe", "paid", "socl", "disp",
"othr", "dirc", "mail")) +
  theme(legend.position = 'bottom',
        legend.direction = "horizontal",
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        plot.title = element_text(size = 20, face = "bold", vjust = 2, color =
'black', lineheight = 0.8),
        axis.title.x = element_text(size = 16, face = "bold"),
        axis.title.y = element_text(size = 16, face = "bold"),
        axis.text.y = element_text(size = 8, face = "bold", color = 'black'),
        axis.text.x = element_text(size = 8, angle = 90, hjust = 0.5, vjust =
0.5, face = "bold")) +
  ggtitle("Mapa de calor da matriz de transição")
```

APÊNDICE D – Link de acesso aos dados utilizados

Link para baixar o arquivo CSV com o conjunto de dados randomizados que representa as jornadas utilizadas neste trabalho:

https://www.4shared.com/s/feFJF6qAejq