UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

DEPARTAMENTO ACADÊMICO DE INFORMÁTICA

CIÊNCIAS DE DADOS E SUAS APLICAÇÕES

AUGUSTO CÉSAR DE SOUZA TANAKA

SATISFAÇÃO DO CONSUMIDOR E ÁRVORE DE DECISÃO, UM ESTUDO EMPÍRICO

CURITIBA

2021

AUGUSTO CÉSAR DE SOUZA TANAKA

SATISFAÇÃO DO CONSUMIDOR E ÁRVORE DE DECISÃO, UM ESTUDO EMPÍRICO

Monografia apresentada como requisito parcial à obtenção do título de especialista em Ciências de Dados e suas Aplicações, do Departamento de Informática, da Universidade Tecnológica Federal do Paraná.

Orientador: Prof. Dr. Matheus Garibalde

CURITIBA

2021

Ministério da Educação
UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
UTFPR - CAMPUS CURITIBA
DIRETORIA-GERAL - CAMPUS CURITIBA
DIRETORIA DE PESQUISA E PÓS-GRADUAÇÃO - CAMPUS CURITIBA
DEPARTAMENTO DE APOIO DAS ESPECIALIZAÇÕES LATO-SENSU DOS CURSOS DE INFORMÁTICA - CAMPUS CURITIBA
CURSO DE ESPECIALIZAÇÃO EM CIÊNCIA DE DADOS E SUAS APLICAÇÕES

TERMO DE APROVAÇÃO

SATISFAÇÃO DO CONSUMIDOR E ÁRVORE DE DECISÃO, UM ESTUDO EMPÍRICO

por

Augusto Cesar De Souza Tanaka

Este Trabalho de Conclusão de Curso foi apresentado às 19h00 do dia 02 de agosto de 2021 por videoconferência como requisito parcial à obtenção do grau de Especialista em Ciência de Dados e suas Aplicações na Universidade Tecnológica Federal do Paraná - UTFPR - Campus Curitiba. O aluno foi arguido pela Banca de Avaliação abaixo assinados. Após deliberação, a Banca de Avaliação considerou o trabalho aprovado.

_____

Prof. Msc. Matheus Garibalde Soares de Lima (Presidente/Orientador – Grupo BOTICÁRIO)

_____

Prof. Dr. Marcelo de Oliveira Rosa (Avaliador 1– DAELT-CT/UTFPR-CT)

_____

Profa. Dra. Rita Cristina Galarraga Berardi (Avaliadora 2 – DAINF-CT/ UTFPR-CT)

O Termo de Aprovação assinado encontra-se no sistema SEI- Processo nº 23064.031950/2021-70

Ao meu pai Teruo Tanaka
in memoriam

AGRADECIMENTOS

À minha família, amigos, minha esposa e ao meu filho. Razões de viver, motivos de esperança.

Sou cético, mas não me felicito por isso
desejo que os jovens tenham direito à
esperança.

(Jorge Luis Borges, 1960

RESUMO

TANAKA, Augusto César de Souza. **SATISFAÇÃO DO CONSUMIDOR E ÁRVORE DE DECISÃO, UM ESTUDO EMPÍRICO**. 2021. 30 f. Tese (Especialização em Ciências de Dados e suas Aplicações) - Universidade Tecnológica Federal do Paraná. Curitiba, 2021.

A obtenção de dados e o estudo da satisfação do consumidor são importantes para a otimização de processos, desenvolvimento e aprimoramento de novos produtos, bem como a otimização no uso de recursos para qualquer organização. Nesse contexto, empresas desenvolveram diversas formas de coletar dados, tais como aparelhos de coleta de dados de satisfação em lojas físicas. Com esses dados é possível utilizar formas de visualização, agrupamento e combinações de dados lançando mão de métodos estatísticos para se prover de insumos para mensuração da satisfação do consumidor e identificar Promotores e Detratores. Dessa forma, o objetivo deste trabalho é determinar as variáveis de satisfação do consumidor com maior probabilidade de apontar a ocorrência de um cliente do tipo Promotor. Após a análise e tratamento de dados, utilizando softwares estatísticos, se criou uma árvore de decisão, composta das categorias mais significativas que estão associadas à variável dependente dos consumidores insatisfeitos. Encontrou-se que as variáveis Atributos, Áreas e Disponibilidade de produtos foram, nesta ordem hierárquica, as mais relevantes para a probabilidade de obter clientes Promotores.

Palavras-chave: Satisfação do Consumidor. Árvore de Decisão. Detecção iterativa por método Chi-Quadrado.

ABSTRACT

TANAKA, Augusto César de Souza. **Customer Satisfaction and Decision Trees: An Empiric Study**. 2021. 30 p. Thesis (Specialization in Data Science and its Applications) - Federal Technology University - Parana. Curitiba, 2021.

Obtaining data and studying consumer satisfaction are important for optimizing processes, developing and improving new products, as well as optimizing the use of resources for any organization. In this context, companies have developed several ways to collect data, such as satisfaction data collection devices in physical stores. With this data, it is possible to use forms of visualization, grouping and combinations of data, making use of statistical methods to provide inputs to measure consumer satisfaction and identify Promoters and Detractors. Thus, the objective of this work is to determine the consumer satisfaction variables that are most likely to point out the occurrence of a Promoter-type customer. After analyzing and processing data, using statistical software, a decision tree was created, composed of the most significant categories that are associated with the dependent variable of dissatisfied consumers. It was found that the Attributes, Areas and Product Availability variables were, in this hierarchical order, the most relevant for the probability of obtaining Promoter type customers.

Keywords: Customer Satisfaction. Decision Tree. Chi-square automatic interaction detection.

LISTA DE ILUSTRAÇÕES

LISTA DE SIGLAS E ACRÔNIMOS

LISTA DE SIGLAS

NPS	*Net Promoter Score*

LISTA DE ACRÔNIMOS

CSAT	*Customer Satisfaction*
AtribCat	Atributo da Categoria
AreaCat	Área Física da Loja
Disponib	Resposta à pergunta sobre disponibilidade de produtos

SUMÁRIO

1 INTRODUÇÃO

Entender o consumidor, como ele pensa e se relaciona com um produto ou marca é chave para as decisões estratégicas (REICHHELD, 2011). Neste contexto surgem diversas metodologias de se fazer pesquisa e tentativas de entender de forma qualitativa e quantitativa o que os consumidores pensam (REICHHELD, 2011).

Clientes Promotores são aqueles que dão notas 9 ou 10 em uma escala de 0 a 10. São os clientes fiéis e capazes de reproduzir as boas impressões que se tem de uma marca. Os clientes Neutros são aqueles que dão nota 7 ou 8 e representam aqueles capazes de oscilar rapidamente para os mais fiéis ou capazes de repudiar a marca. Os clientes Detratores são aqueles que dão nota de 0 a 6 e constituem-se um foco de preocupação e necessidade de atenção, pois são capazes de denegrir e reproduzir más impressões do produto e da marca com vontade. Da proporção entre os clientes detratores e os clientes promotores tem-se o score NPS (*Net Promoter Score*) que representa esse conjunto de clientes e suas impressões da marca (REICHHELD, 2011).

Esse trabalho tem como objetivo determinar as variáveis de satisfação do consumidor com maior probabilidade de apontar a ocorrência de um cliente do tipo Promotor. Para tal, será utilizado o método de Árvores de Decisão.

Para o presente trabalho foi utilizado um conjunto de 718.260 amostras de satisfação do consumidor, com perguntas referentes a questões demográficas, experiência na loja e satisfação. Os totens são dispositivos imóveis que ficam dispostos em locais estratégicos em lojas com o intuito de fazer com que os consumidores avaliem o atendimento, produto, marca e/ou experiência através de perguntas simples, objetivas e que não demandam muito exercício mental ou tempo. São importantes, portanto, que as perguntas sejam inteligíveis e objetivas, já que o consumidor geralmente sempre tem limitação de tempo e atenção para responder às inúmeras perguntas da pesquisa.

Não raro, portanto, essas respostas feitas nesse contexto de falta de tempo podem gerar dados que se agrupam de maneira difícil de serem visualizados através de métodos de mais convencionais (gráficos de barra, pizza, linhas etc.), isto é, a variação não ocorre em montante adequado para se discernir padrões pertinentes para tomadas de decisão já que os clientes responderam sem discernimento, apenas se atendo às perguntas que geralmente criaram insatisfação.

É neste contexto que se propõe o uso de metodologias mais aptas a agruparem dados e separar aqueles estatisticamente significantes, tudo isso em um cenário de fácil inteligibilidade dos resultados aptos a serem postas em produção em um produto direcionado a administradores de empresa leigos em estatística.

Para o presente trabalho, foi utilizada a base com 30 variáveis de respostas advindas dos totens, as quais foram tratadas e selecionadas para chegar em um modelo de Árvores de Decisão. Os resultados apontam para a maior relevância da variável Atributo, seguindo de Área e por fim, Disponibilidade, para determinar a probabilidade dos clientes em se tornarem Promotores.

Este trabalho, por fim, permite *insights* para gestores da rede varejista que selecionam as variáveis de maior relevância dentro do negócio, que aumentam a probabilidade de obtenção de clientes Promotores e reduzem, naturalmente, a probabilidade de clientes Detratores.

Nas próximas seções iremos determinar o método de como os dados são coletados durante a pesquisa, a descrição da base de dados, a definição da árvore de decisão e do modelo adotado e, por fim, os resultados obtidos com o modelo, para finalmente expor as conclusões e comentários finais.

2 REFERENCIAL TEÓRICO

Nesta seção serão apresentados os principais fundamentos teóricos utilizados para a realização deste estudo. Será feito uma explanação (i) dos fundamentos do *Net Promoter Score* (NPS), que deu origem ao questionário utilizado pela rede varejista para que seja respondido por seus clientes nas lojas físicas, e (ii) o método de Árvore de Decisão, que foi utilizado como alternativa para reconhecer as variáveis que apresentam maior probabilidade dos clientes em se tornarem Promotores.

2.1 NPS

O estudo da satisfação do consumidor é uma métrica importante para toda empresa que deseje saber qual a proporção dos consumidores que estão satisfeitos ou insatisfeitos com a marca e o produto consumido. Neste contexto, surgiram algumas metodologias de análise de satisfação, notadamente NPS e CSAT (*Customer Satisfaction*, nomenclatura genérica para notas em poucas categorias, 5 ou menos, cujo valor maior representar uma melhor avaliação).

O primeiro foi inventado em 2003 por Fred Reichheld para avaliar como uma determinada marca seria vista por uma população tendo-se em vista aqueles que tiveram uma experiência satisfatória e os insatisfeitos. Para tanto, pergunta-se aos consumidores "Qual a probabilidade de você recomendar a marca X?". Desprende-se daí que os consumidores que respondem negativamente em uma escala de 0 a 10 tenham a propensão a não só não recomendar a marca, mas também ativamente denegri-la. Nas palavras de REICHFELD em seu livro *The Ultimate Question 2.0*, ele explica como as empresas podem serem afligidos com os comentários dos chamados detratores, isto é, aqueles que dão notas de 0 a 6 na escala NPS:

> Eles (os detratores) não são pessoas felizes. Estão insatisfeitos e até consternados com a forma como são tratados. Eles falam mal da empresa para seus amigos e colegas. Se eles não podem mudar facilmente de fornecedor - por exemplo, se eles têm contratos de longo prazo ou se não há concorrentes com ofertas semelhantes - eles se incomodam, registrando reclamação após reclamação e aumentando os custos. (REICHFELD, 2011, p. 45).

Em outras palavras, um consumidor em seu estado de satisfação desejado (Promotores) tem uma relação saudável com a marca, entende-se que possam até mesmo estar encantados e se fidelizar à marca. Já os detratores são capazes de denegrir a marca e expressar enfaticamente sua insatisfação.

Partindo-se do pressuposto que toda marca age em benefício próprio vis-à-vis o consumidor, em um cenário de consumidores 100% satisfeitos não haveria de se fazer nada, a não ser aprimorar o produto e eventualmente expandir a produção. Porém, em termos de informação e insights, os consumidores detratores tem um potencial maior de fazer a empresa revisar

2.2 ÁRVORE DE DECISÃO

Para se entender melhor essa massa de dados, tendo-se em vista as dificuldades já expostas, adotou-se como técnica o uso da Árvore de Aprendizado. Em sua acepção geral, uma árvore de decisão é um método de inferências estatística em que o algoritmo de aprendizado é supervisionado e uma variável alvo é definida. Conforme definição de SULLIVAN, 2011:

> Em uma linguagem simples, uma árvore de decisão pode ser definida como qualquer tipo de algoritmo de aprendizagem que seja supervisionado e tem uma variável alvo predeterminada que seja comumente usada em problemas de classificação. Problemas de classificação são o forte das árvores de decisão já que eles se adequam perfeitamente em problemas de classificação. As árvores de decisão também podem ser usadas tanto em variáveis contínuas como em variáveis com saídas em categorias. Nesse caso, a amostra é dividida em dois ou mais conjuntos homogêneos de dados divididos pelo divisor (*splitter*) ou diferenciador mais significante na variável de entrada. (SULLIVAN, 2011, p. 178).

Uma árvore típica é constituída por nós, folhas e raiz. Seu início se dá olhando-se de cima para baixo, sendo que a raiz, posicionada acima, contém a categoria mais significante para sua decomposição nos nós, sendo esses nós as categorias e os galhos as ramificações com cada um dos elementos possíveis, resultando em uma folha que conterá o percentual de cada um dos elementos de uma variável alvo

específica. Em outras palavras, trata-se de uma estrutura similar a um fluxograma na qual cada nó interno representa um "teste" em um atributo (por exemplo. no lançar de uma moeda, se o resultado é Cara ou Coroa), cada ramo representa o resultado do teste, e cada nó folha representa um rótulo de classe (decisão tomada após calcular todos os atributos). Os caminhos da raiz à folha representam as regras de classificação.

A incidência de nós, folhas e ramificações dependem da configuração adotada para o modelo, sendo esses o número mínimo de amostras para que um novo nó seja desdobrado, a significância estatística e a altura da árvore, isto é, quantas ramificações são admitidas desde a raiz até as folhas.

Considerando a complexidade potencial diante de elevado número de variáveis e categorias as árvores de decisão surgem como alternativas viáveis para entender como diferentes grupos de características se relacionam entre si e quais são as mais recorrentes e relevantes em relação a uma variável alvo. Isto é, no caso em questão. Uma árvore de decisão é uma ferramenta de suporte à decisão que usa um modelo de decisão e suas possíveis consequências, incluindo resultados de eventos fortuitos, custos de recursos e, evidentemente, utilidade. É uma maneira de exibir um algoritmo que contém apenas instruções de controle condicional, sem fazer considerações sobre a lógica de negócio.

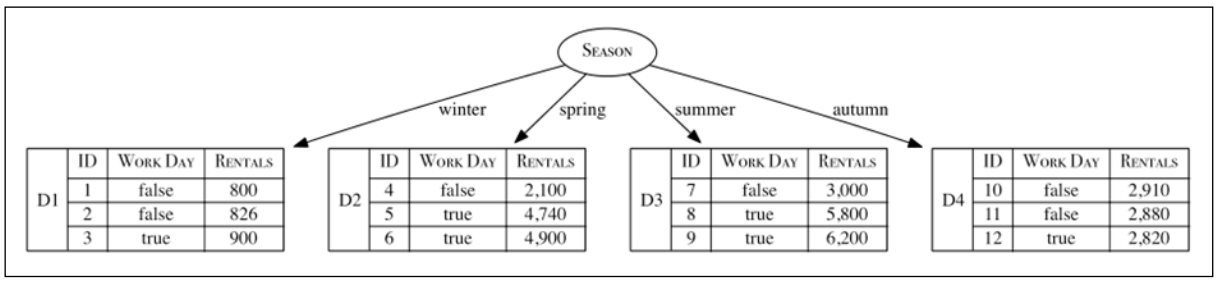

Figura 1 – Exemplo de árvore de decisão
Fonte: o autor

Quanto à validade e relevância das Árvores de Decisão para métodos de *machine learning*, SULLIVAN expõe com clareza sua capacidade de explorar e extrair informações de variáveis categóricas:

> "Os métodos de árvore de decisão fabricam um modelo de decisões que foram tomadas com base nos valores reais dos atributos nos dados disponíveis. Uma bifurcação é formada na árvore de decisão até que uma previsão seja feita para o registro fornecido. Frequentemente, eles são treinados para a classificação de dados e também para problemas de regressão. Eles são bastante rápidos e precisos, o que os torna altamente populares e os favoritos no mundo do aprendizado de máquina". (SULLIVAN, 2011, p. 178).

3 MÉTODO

Parte Para a análise dos dados em estudo é importante que algumas etapas sejam cumpridas para que o processo seja feito de maneira clara, replicável, auditável e com dados confiáveis, que não afetem a eficácia ou acurácia do modelo. Nesta seção iremos descrever o método utilizado para análise dos resultados com a finalidade de atingir o objetivo proposto. Serão apresentados os métodos de extração e limpeza dos dados seguido do processo de elaboração da árvore de decisão.

A coleta de dados foi realizada da seguinte forma: o cliente das lojas interage com uma plataforma imóvel, em uma tela interativa, em que diversas perguntas sobre a loja e experiência de consumo. Usualmente as pesquisas são respondidas durante ou após a compra. Esses totens de pesquisa estão distribuídos por inúmeras filiais da rede de varejo pelo Brasil e os dados são enviados por rede de celular para os servidores centrais, onde os dados são agrupados e organizados para servirem para tomadas de decisões dos gestores.

Os totens têm um dispositivo interativo similar a um *tablet* que coleta informações das respostas e as transmitem para uma base que serve como subsídios para estudos de satisfação, problemas, vícios e erros de planejamento e problemas relacionados ao que o produto oferece. Contudo, dada a questão de tempo, as perguntas tem que ser categorizadas de maneira a não demandar muito esforço mental do avaliador, isto é, ícones de sorrisos com diferentes nuances junto a perguntas objetivas compõe a maioria das respostas, em detrimento de respostas mais pormenorizadas e textuais, que possam servir para uma análise focalizada em um detalhe do serviço ou produto. Outro problema é que não é incomum um avaliador avançar rapidamente sobre diversas perguntas sem se ater ao que está sendo perguntado, apenas para se chegar a uma determinada resposta e ali sim, expressar seu contentamento ou descontentamento. Porém, de maneira a contornar este problema, pode-se agrupar os dados e se buscar tendências ou padrões que indiquem situações ou erros específicos em um determinado setor ou elemento do produto. Isto é, conforme as pesquisas vão sendo feitas, supondo-se que há algo efetivamente errado em um setor, ele tenderá a aparecer nas pesquisas conforme o número de dados coletados aumenta.

Para o presente estudo utilizamos dados coletados entre dezembro de 2020 e maio de 2021 referentes a rede varejista de abrangência nacional. Os dados

referem-se a pesquisa de natureza quantitativa e qualitativa, isto é, contém comentários diversos e também perguntas referentes a percepção de qualidade em setores da loja e atributos intangíveis tais como atendimento, tempo de fila e variedade de produtos.

As perguntas são feitas considerando a experiência de consumo a fim de coletar tanto informações mais abrangentes como o NPS, que se refere à experiência como um todo, quanto mais específicas, como a visita em áreas da loja, como hortifruti e padaria. Próximo ao final da pesquisa, perguntas sobre o perfil demográfico também são feitas, como faixa etária e gênero. A ordem das perguntas também tem relevância pois, por exemplo, o NPS sendo a pergunta mais genérica, deve estar à frente das outras, uma vez que muitas vezes a pessoa que interage com o totem não está com tempo disponível para responder a todas as perguntas, sendo as perguntas possivelmente extenuantes para alguém respondê-las após uma sessão de compras.

A fim de extrair informações em categorias que reúnam características em comum dos clientes detratores, optou-se por se utilizar o modelo estatístico de árvore de decisão, mais especificamente o modelo CHAID (KASS, 1980), sendo esperado entender o perfil de cliente que tiveram a pior experiência de loja a fim de ter *insights* sobre como resolver ou contornar esses problemas.

3.1 QUESTIONÁRIO

O questionário é composto por 9 etapas distribuídas entre perguntas referente à experiência como um todo, percepção da marca, questões referentes ao atendimento e visita a diferentes setores da loja bem como comentários abertos.

A pergunta mais abrangente é o NPS, em que é feita a seguinte pergunta: "Responda de 0 a 10, qual a chance de você recomendar esta loja a um amigo”. Esta resposta resultou nas categorias Promotor, Neutro e Detrator, sendo promotores aqueles que deram nota 9 ou 10, neutros os que deram nota 6 ou 7 e os demais, 6 e abaixo, detratores.

Após isso, pergunta-se se o cliente deseja continuar a avaliação. Caso ele responda não, o questionário é finalizado com uma última pergunta para coletar um comentário geral (150 caracteres). Caso ele responda sim, haverá perguntas

apresentadas em 3 possíveis respostas ilustradas por uma careta, um sorriso neutro e um sorriso feliz (Ruim, Regular e Bom, respectivamente). As perguntas pertencem a dois grupos: de Atributos (AtribCat) e de Área Física (AreaCat) da loja. Em relação aos Atributos, são 6 perguntas: Conforto da Loja, Tempo de Caixa, Atendimento, Preço, Variedade de Produtos e, por fim, Qualidade. Em relação às Áreas Físicas da loja, trata-se de 5 áreas: Perfumaria, Acessórios, Moda Masculina, Moda Feminina, Moda Infantil e Calçados.

Após as perguntas referentes a Atributos e Área Física, o cliente é levado a duas perguntas finais, às quais ele pode responder abertamente. A primeira é se algum produto específico não foi encontrado e a segunda para comentários gerais.

Conforme a Figura 2 é possível verificar a sequência de perguntas de forma esquemática:

Figura 2 – Diagrama do Questionário
Fonte: o autor

3.2 DESCRIÇÃO DA AMOSTRA

Os dados coletados referem-se ao período entre dezembro de 2020 e maio de 2021 de 19 estados, 9 regionais sendo que cada regional está sob uma gerência específica. As respostas foram coletadas em diversos períodos do dia (manhã, tarde e noite), totalizando 718.259 participantes, sendo 316.691 (44%) de mulheres, 292.748 (41%) homens e 108.820 (15%) que optaram por não responder.

Em relação à Faixa Etária, os respondentes totalizaram: abaixo de 17 anos 32.905 (5%), de 18 a 24 anos 79.977 (11%), de 25 a 34 anos 142.682 (20%), de 35 a 44 anos 150.933 (21%), de 45 a 54 anos 96.132 (13%) e acima de 55 anos 86.998 (12%). O número de não respondentes e valores vazios totalizaram 128.632 (17%). Quanto à distribuição das respostas pelos períodos do dia constaram os seguintes números: 197.966 (28%) pela manhã, 298.530 (42%) pela tarde e 221.763 e (31%) à noite.

Conforme Tabela 1, as categorias de NPS por sua vez ficaram distribuídas da seguinte forma: Promotor contou com 598.868 (83%) e Detratores 119.391 (17%). As notas atribuídas aos Atributos da loja ficaram distribuídos conforme a seguir: 281.928 (39%) Bom; 23.267 (3%) Ruim, sendo o total de dados vazios 413.064 (58%). Quanto às notas atribuídas às áreas físicas da loja estes ficaram distribuídos entre: 264.475 (37%) Bom, 27.780 (4%) Ruim e 426.004 (59%) dados vazios, sendo o total de dados vazios 388.146 (54%). Ao se analisar a distribuição de valores entre as variáveis, tem-se a seguinte configuração conforme observado na Tabela 1.

Tabela 1 - Tabela com valores e proporção entre as variáveis adotadas.

	Satisfação dos Atributos		Satisfação Áreas da Loja		Disponibilidade dos Produtos		NPS	
Bom	281.928	39%	264.475	37%	25.674	4%	119.391	17%
Ruim	23.267	3%	27.780	4%	299.490	42%	598.868	83%
Nulo	413.064	58%	426.004	59%	393.095	54%		

Fonte: o autor, 2021

A distribuição dos dados mostra uma proporção maior entre aqueles que avaliaram positivamente nos Atributos e na Área de Loja, sendo que apenas um pequeno percentual não encontrou todos os produtos que procurava. Isso reforça a necessidade de se estabelecer um limiar de corte pequeno para que a árvore seja capaz de abranger esses 25.674 consumidores que não encontraram todos os produtos e que estão distribuídos entre as diversas combinações de categorias. Para se configurar o modelo e estabelecer os limiares de tamanho, número mínimo de amostras e índices de probabilidade do modelo, adotou-se os seguintes parâmetros conforme imagem 1 a seguir, extraída do software R Studio.\

3.3 TRATAMENTO DO QUESTIONÁRIO

O tratamento de dados foi feito de maneira a agrupar as 6 respostas referentes aos Atributos e às Áreas Físicas em uma única variável para cada categoria. O cálculo para a criação desta variável única obedeceu a um método, já aplicado pela organização que executa a pesquisa, de criar um dado agregador médio, seguindo os seguintes critérios: cada resposta foi convertida em um número; a Bom foi atribuído o número 100, a Neutro, 0, e Ruim, -100. A partir desses números foi possível fazer uma média simples, resultando em valores que vão de -100 a 100, para cada um dos agrupamentos. A partir do resultado da média obtidos com as respostas de cada grupo criou-se uma nova variável respeitando-se os seguintes intervalos: -100 a -33,34 corresponde a ruim, de -33,33 a 33,33 corresponde a regular e de 33,34 a 100 corresponde a bom. Nas Equações 1 e 2 são apresentadas a lógica deste agrupamento.

Equação 1

$$Variável\ Atributos = \frac{\sum_i^n \quad Respostas\ referentes\ a\ Atributos}{n}$$

Equação 2

$$Variável\ Áreas\ Físicas = \frac{\sum_i^n \quad Respostas\ referentes\ a\ Áreas\ Físicas}{n}$$

4 RESULTADOS

Nesta seção apresentaremos a análise e discussão dos resultados. Na Figura 3 é possível observar a árvore de decisão das variáveis selecionadas conforme relatado na Seção 3. A árvore, além da raiz, gerou 8 nós, oriunda de 7 ramificações compreendendo as combinações de satisfação do consumidor referente a atributos, áreas da loja e percepção de produtos faltantes.

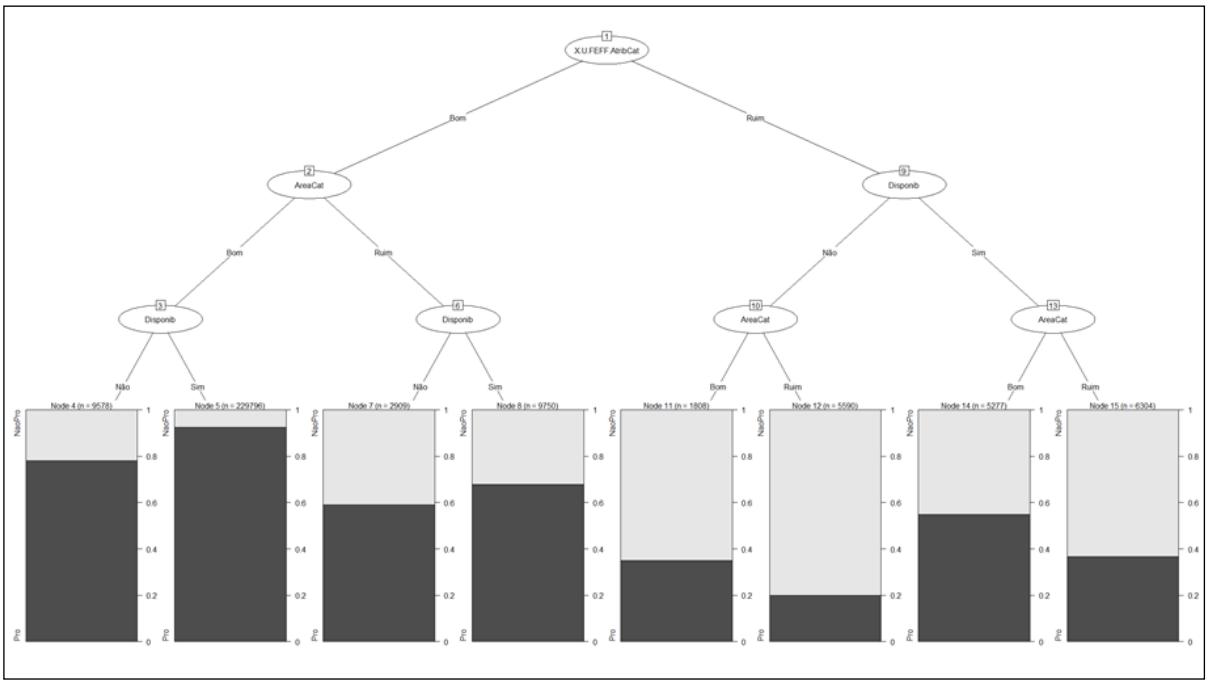

Figura 3 – Árvore de decisão oriunda da análise dos dados
Fonte: o autor

Na Figura 3 é possível avaliar os diferentes nós gerados pela árvore e os percentuais de probabilidade oriundo por determinado conjunto de atributos, necessariamente resultando em um consumidor Detrator ou Não-Detrator.

Em relação às variáveis constantes na Figura 3, as abreviações e definições estão definidas a seguir:

AtribCat: Satisfação geral de 6 Atributos do questionário (calculado conforme exposto no item 3.3).

AreaCat: Satisfação geral das 5 Áreas da Loja do questionário (calculado conforme exposto no item 3.3).

Disponib: Disponibilidade do Produto. Refere-se à resposta do cliente se ele encontrou todos os produtos que procurava na loja.

A seguir estão descritas as combinações de categorias de cada um dos nós finais da árvore:

● Nó 4: tem-se como resultado um consumidor com satisfação positiva em relação aos Atributos, satisfação Bom em Área e que não encontrou todos os itens de loja que procurava. Este consumidor tem uma probabilidade de 78,57% de ser um promotor.

● Nó 5: tem-se como resultado um consumidor com satisfação positiva em relação aos Atributos, satisfação positiva em Área e que encontrou todos os itens de loja que procurava. Este consumidor tem uma probabilidade de 93,47% de ser um promotor.

● Nó 7: tem-se como resultado um consumidor com satisfação positiva em relação aos Atributos, satisfação negativa em Área e que não encontrou todos os itens de loja que procurava. Este consumidor tem uma probabilidade de 58,41% de ser um promotor.

● Nó 8: tem-se como resultado um consumidor com satisfação positiva em relação aos Atributos, satisfação negativa em Área e que encontrou todos os itens de loja que procurava. Este consumidor tem uma probabilidade de 66,89% de ser um promotor.

● Nó 11: tem-se como resultado um consumidor com satisfação negativa em relação aos Atributos, que não encontrou todos os itens que procurava e que avaliou positivamente a Área. Este consumidor tem uma probabilidade de 38,66% de ser um promotor.

● Nó 12: tem-se como resultado um consumidor com satisfação negativa em relação aos Atributos, que não encontrou todos os itens que procurava e que avaliou negativamente a Área. Este consumidor tem uma probabilidade de 20,01% de ser um promotor.

● Nó 14: tem-se como resultado um consumidor com satisfação negativa em relação aos Atributos, que encontrou todos os itens que procurava e que avaliou positivamente a Área. Este consumidor tem uma probabilidade de 56,99% de ser um promotor.

● Nó 15: tem-se como resultado um consumidor com satisfação negativa em relação aos Atributos, que encontrou todos os itens que procurava e que avaliou negativamente a Área. Este consumidor tem uma probabilidade de 38,67% de ser um promotor.

Para facilitar a comparação entre as variáveis, tem-se a seguir nas Figuras 4 e 5 uma tabela resumindo da Figura 3 em relação às ramificações à direita e e à esquerda, contendo as diferentes combinações de avaliação positiva ou negativa nas diferentes variáveis, bem como suas respectivas influências na chance de determinado consumidor ser do tipo Promotor.

Atributo	Categoria	Disponibilidade	% Promotor	Influência Categoria
-	+	-	38,60%	18,59%
-	-	-	20,01%	
-	+	+	56,99%	17,32%
-	-	+	38,67%	
+	+	N/A	80,02%	17,37%
+	-	N/A	62,65%	

Figura 4 – Tabela comparativa entre as variáveis de satisfação (referente às ramificações à direita da Figura 3)
Fonte: o autor

Atributo	Categoria	Disponibilidade	% Promotor	Influência Categoria
+	+	+	93,47%	14,9%
+	+	-	78,57%	
+	-	+	66,89%	8,48%
+	-	-	58,41%	
-	N/A	+	47,83%	18,5%
+	N/A	-	29,34%	

Figura 5 – Tabela comparativa entre as variáveis de satisfação (referente às ramificações à esquerda da Figura 3)
Fonte: o autor

Dando continuidade à análise, o que melhor diferencia a probabilidade de detratores de não-detratores é a experiência geral do cliente, representada pela variável Atributos. Mesmo clientes que não encontraram todos os produtos e que avaliaram negativamente a Área, caso tenham uma boa experiência geral na loja (ex. decoração da loja, variedade de produtos), têm maior probabilidade de se tornarem promotores (Nó 7, 58,41%) do que, de forma inversa, um cliente que resultou negativamente na variável Atributos e tenha encontrados todos os produtos e avaliado positivamente a Área (Nó 14, 56,99%). Isto corrobora a hipótese de que a experiência de loja seja mais decisiva comparando-se com as demais variáveis adotadas para o modelo.

Ainda através da Figura 3, é possível observar o efeito da variável AreaCat para o aumento da probabilidade do cliente se tornar promotor. No caso que tenha uma má experiência na variável Atributos e não tenha encontrado todos os produtos disponíveis, é possível observar que a probabilidade do cliente se tornar um promotor se avaliar negativamente a variável AreaCat é de 20,01%; se ele tiver avaliado positivamente a variável AreaCat a chance sobe para 38,66%. Isso significa que a variável AreaCat influenciou positivamente 18,65% na probabilidade do cliente se tornar promotor (diferença entre os nós 11 e 12). No caso do cliente que tenha uma má experiência na variável atributos e tenha encontrado todos os produtos, é possível observar que a probabilidade se tornar promotor, se avaliar negativamente a variável AreaCat, é de 38,67%, Se ele tiver avaliado positivamente a variável AreaCat, a chance sobe para 56,99%. Isso significa que a variável AreaCat influenciou positivamente 18,32% na probabilidade de um cliente se tornar promotor (diferença entre os nós 14 e 15). Já para os clientes que tenham uma boa experiência na variável Atributos, o efeito da variável AreaCat para o aumento da possibilidade de um cliente se tornar promotor, desconsiderando as ramificações da disponibilidade de produtos, é de 23,37% (diferença entre a média 86,02% dos nós 4 e 5 e a média dos nós 7 e 8, 62,65%).

Dando continuidade à análise dos resultados que constam na Figura 3, observa-se o efeito da variável Disponib para o aumento da probabilidade do cliente se tornar promotor. Caso ele avalie positivamente a variável Atributos e tenha avaliado positivamente AreaCat, é possível observar que a probabilidade do cliente se tornar um promotor se não encontrar todos os produtos é de 78,57% e, caso tenha encontrado todos os produtos, a probabilidade é de 93,47%. Ou seja, a Disponibilidade do produto aumentou em 14,9% a probabilidade de um cliente ser promotor. No caso de o cliente ter considerado a variável AreaCat como ruim e não ter encontrado todos os produtos, a chance de ser promotor é de 58,41% e, caso tenha encontrado todos os produtos, a chance aumenta para 66,89%. Ou seja, nesse segmento o fato de um cliente ter encontrado todos os produtos aumenta em 8,48% a chance de ele ser promotor. Já para os clientes que tenham uma má experiência na variável Atributo, o efeito da variável Disponibilidade para o aumento da possibilidade de um cliente se tornar promotor, desconsiderando as ramificações de AreaCat, é de 18,50% (diferença entre a média 29,34% dos nós 11 e 12 e a média dos nós 14 e 15, 47,83%).

Comparando-se os resultados e análises acima é possível deduzir que a variável AtribCat é a que mais tem relevância para tornar-se um cliente promotor. Do ponto de vista visual, percebe-se um destaque maior na área correspondente aos promotores em cinza escuro nos quatro nós à esquerda da Figura 3 comparando-se com os quatros nós à direita. Comparando as outras duas variáveis, a saber, AreaCat e Disponib, a primeira é mais determinante para o aumento da probabilidade de tornar um consumidor promotor do que a segunda. Isso é explicitado quando comparamos a probabilidade do cliente ser promotor, em caso positivo ou negativo para ambas as variáveis, conforme apresentado acima.

5 CONCLUSÃO E CONSIDERAÇÕES FINAIS

Esse trabalho teve como objetivo determinar quais variáveis dentro de um modelo são as mais determinantes para a ocorrência de um cliente do tipo Promotor para um negócio varejista. Após a seleção das variáveis cruciais para tomada de decisão, executou-se o processamento do modelo de árvore de decisão CHAID para levantar-se quais são as variáveis mais determinantes e seu posicionamento hierárquico, bem como as probabilidades associadas ao tipo de consumidor.

Com o modelo pronto, partiu-se para uma análise gráfica e das probabilidades de cada nó e ramificação possíveis. Constatou-se que em termos de relevância estatística e predominância para a determinação de um cliente Promotor (ou, em outras palavras, evitando-se a existência de Não-Promotores), são as seguintes variáveis estão na ordem da mais relevante à menos relevante: Satisfação dos Atributos > Satisfação das Áreas da Loja > Disponibilidade dos produtos.

Dadas as contingências de tempo e interesse em fazer uma avaliação pormenorizada, encontrar formas mais precisas de se avaliar a qualidade de uma pesquisa de satisfação é de suma importância para se fazer jus aos investimentos alocados para obter os dados dos consumidores. Dentre os investimentos estão aqueles de ordem física, humana, tecnológica além do tempo dos administradores da empresa em questão.

Diante do exposto, os dados disponíveis podem servir de subsídios para que as decisões estratégicas e de melhoria de processos sejam aplicadas de maneira segmentada e ordenada, tendo-se como primeiro ponto de ação os mais importantes.

REFERÊNCIAS

REICHHELD, Frederick F. The Ultimate Question 2.0: How Net Promoter Companies Thrive in a Customer-driven World. 2. ed. New York: Harvard Business Press, 2011. 290 p. v. 1. ISBN 9781422173350, 1422173356.

HTTPS://RDRR.IO/ (Estados Unidos). Chaid. *In*: HOWSON, Ian. Chaid: CHi-squared Automated Interaction Detection. [*S. l.*], 2 maio 2019. Disponível em: https://rdrr.io/rforge/CHAID/man/chaid.html. Acesso em: 25 jul. 2021.

KELLEHER, John D. *et al*. Fundamentals of Machine Learning for Predictive Data Analytics. 2. ed. rev. New York: Mit Press, 2015. 624 p.

MACHINE Learning For Beginners: Algorithms, Decision Tree & Random Forest Introduction. 1. ed. Michigan: Healthy Pragmatic Solutions, 2017. 356 p. v. 1. ISBN B074YJYWDY.

BARROS, Rodrigo C. Automatic Design of Decision-Tree Induction Algorithms. 1. ed. New York: Springer, 2015. 459 p. v. 1. ISBN 978-3-319-14230-2.

RICHARDSON, Roberto Jarry. Pesquisa Social: Métodos e Técnicas. 3. ed. Sao Paulo: Atlas, 1999. 344 p. v. 1.

KASS, GORDON V. An Exploratory Technique for Investigating Large Quantities of Categorical Dat. **Applied Statistics**, Witwatersrand, África do Sul, ano 1980, v. 29, n. 2, p. 119-127, 1 dez. 1980.

Acesso em: 11 nov. 2011.

APÊNDICE A - Códigos utilizados no R Studio

```
install.packages("partykit")
install.packages("CHAID", repos="http://R-Forge.R-project.org")
require(rsample) # for dataset and splitting also loads broom and tidyr
require(dplyr)
require(ggplot2)
theme_set(theme_bw()) # set theme
require(CHAID)
require(purrr)
require(caret)

df <- read.csv("J:/07 Biblioteca/Trabalho Pós UTFPR 2021/datasets/dataset_reduzido.csv" , sep = ';' , header = TRUE, na.strings=c("", "NA"), encoding = "UTF-8" )

View(df)

df <- df %>% mutate(across( c(X.U.FEFF.AtribCat, AreaCat, Disponib,NPS),as.factor))

ctrl <- chaid_control(alpha2 = 0.01, alpha3 = -1, alpha4 = 0.01, minsplit = 250, minbucket = 7, minprob = 0.01, stump = FALSE, maxheight = -2)

chaid_df = chaid(NPS ~ ., data = df, control = ctrl)

plot(chaid_df)
```