USO DE REDES NEURAIS NO CÁLCULO DA PROBABILIDADE DE COMPRA NO CENÁRIO DE COMÉRCIO ELETRÔNICO BRASILEIRO

Raphael Oliveira Lourenço

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Engenharia Elétrica, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Mestre em Engenharia Elétrica.

Orientador: Luiz Pereira Caloba

Rio de Janeiro
Fevereiro de 2022

USO DE REDES NEURAIS NO CÁLCULO DA PROBABILIDADE DE COMPRA NO CENÁRIO DE COMÉRCIO ELETRÔNICO BRASILEIRO

Raphael Oliveira Lourenço

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM ENGENHARIA ELÉTRICA.

Orientador: Luiz Pereira Caloba

Aprovada por: Prof. Luiz Pereira Caloba, Ph.D.
Prof. José Gabriel Rodriguez Carneiro Gomes, Ph.D.
Prof. Augusto Cerqueira Santiago, D.Sc.

RIO DE JANEIRO, RJ – BRASIL
FEVEREIRO DE 2022

Lourenço, Raphael Oliveira

USO DE REDES NEURAIS NO CÁLCULO DA PROBABILIDADE DE COMPRA NO CENÁRIO DE COMÉRCIO ELETRÔNICO BRASILEIRO/Raphael Oliveira Lourenço. – Rio de Janeiro: UFRJ/COPPE, 2022.

XIII, 65 p.: il.; 29,7cm.

Orientador: Luiz Pereira Caloba

Dissertação (mestrado) – UFRJ/COPPE/Programa de Engenharia Elétrica, 2022.

Referências Bibliográficas: p. 63 – 65.

1. Data Driven Marketing. 2. Softmax. 3. Probability Prediction. 4. Neural Network. I. Caloba, Luiz Pereira. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia Elétrica. III. Título.

A Deus, a minha amada esposa,
minha querida família e ao meu
querido pai

Agradecimentos

Agradeço a Deus, em quem tenho depositado a direção de meus sonhos, e que tem me ajudado a guiar cada um dos passos da minha vida.

Agradeço também a minha esposa, que me motivou em cada uma das etapas do curso, não deixando que nenhum dos obstáculos que surgiram pudessem me fazer desistir, e que compreendeu carinhosa e pacientemente toda a quantidade de tempo e esforço que depositei nessa empreitada.

Agradeço aos meus pais e familiares, que deram todo o suporte que eu necessitei durante toda a minha vida acadêmica. Com certeza, eu não teria chegado até aqui sem eles. Em particular, dedico esse trabalho ao meu paizão, que nos deixou durante a pandemia de Covid-19, mas que está comigo em meu coração o tempo todo.

Agradeço ao Professor Fábio Ramos, do departamento de Matemática Aplicada da UFRJ, por ser o primeiro a me incentivar fortemente a fazer um curso de mestrado.

Agradeço ao meu orientador, Professor Luiz Calôba, pelo suporte durante o curso e também durante a elobaração deste trabalho.

Agradeço ao Mauricio Machado, membro da secretaria acadêmica do PEE, que por incontáveis vezes prestou ajuda e orientação em situações administrativas complicadas. Não é nenhum exagero dizer que este trabalho não sairia sem seu auxílio.

Agradeço também ao meu amigo Michel, pelas caronas, conversas, listas de exercícios, e por todos os *Megazords* que fizemos para chegarmos ao fim do curso.

Por fim, agradeço aos meus amigos de graduação, Gabriel, Rodrigo, Jonathas, Pedro, Diogo, Lucas e Victor, que também me acompanharam em toda a minha

caminhada acadêmica e que, no fim do dia, se tornaram não só colegas de turma, mas amigos para a vida.

Resumo da Dissertação apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

USO DE REDES NEURAIS NO CÁLCULO DA PROBABILIDADE DE COMPRA NO CENÁRIO DE COMÉRCIO ELETRÔNICO BRASILEIRO

Raphael Oliveira Lourenço

Fevereiro/2022

Orientador: Luiz Pereira Caloba

Programa: Engenharia Elétrica

Nos dias atuais, a evolução nos processamentos de dados faz com que as máquinas se tornem muito mais ágeis e eficazes ao coletar, manipular e gerar dados de informação. Associado a isso, com a propagação da utilização dos meios digitais como internet, smartphones e dispositivos móveis, há uma grande facilidade na troca de informações entre pessoas ao redor do mundo. Além disso, essencialmente, qualquer navegação pela internet gera rastros de utilização, que são registrados pelos servidores de diversas empresas interessadas em entender mais sobre os comportamentos dos indivíduos.

Por outro lado, o cenário de comércio eletrônico é extremamente desafiador do ponto de vista de geração de vendas. Em média, uma compra é feita a cada cem visitas registradas nas principais plataformas de vendas online no Brasil. No entanto, com tantas informações disponíveis para serem utilizadas, dentro do contexto do Marketing 3.0, é possível associar o uso de Redes Neurais dentro do contexto de otimização de vendas.

Utilizando-se de dados reais de uma grande plataforma de vendas, esse trabalho se propôs a explorar o cenário do Cálculo de Probabilidades de compra dentro do e-commerce, pensando em se utilizar dessa informação para influenciar a jornada de compra dos indivíduos e, com a implementação de uma rede neural feed-forward com uma camada intermediária, obteve resultados interessantes tanto do ponto de vista quantitativo, com 0.904 de acurácia, 0.903 de recall, 0.906 de precisão e 0.905 de F1, quanto do ponto de vista qualitativo, onde as distribuições de probabilidades de compra e não-compra foram bastante condizentes com a realidade dos eventos.

Abstract of Dissertation presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

USE OF NEURAL NETWORKS IN THE CALCULATION OF PROBABILITY OF PURCHASE IN THE BRAZILIAN ELECTRONIC COMMERCE SCENARIO

Raphael Oliveira Lourenço

February/2022

Advisor: Luiz Pereira Caloba

Department: Electrical Engineering

Nowadays, the evolution in data processing makes machines become much more agile and efficient when collecting, manipulating and generating information data. Associated with this, with the spread of the use of digital media such as the internet, smartphones and mobile devices, there is a great ease in the exchange of information between people around the world. In addition, essentially, any internet browsing generates usage traces, which are recorded by the servers of several companies interested in understanding more about the behavior of individuals.

On the other hand, the e-commerce landscape is extremely challenging from a sales generation point of view. On average, one purchase is made for every one hundred visits registered on the main online sales platforms in Brazil. However, with so much information available to be used, within the context of Marketing 3.0, it is possible to associate the use of Neural Networks within the context of sales optimization.

Using real data from a large sales platform, this work proposed to explore the scenario of the Calculation of Probability of Purchase within e-commerce, thinking about using this information to influence the purchase journey of individuals and, using a Feed-forward neural network with one hidden layer, obtained results interesting both from a quantitative point of view, with 0.904 of accuracy, 0.903 of recall, 0.906 of precision and 0.905 of F1, and from a qualitative point of view, where the distributions of purchase and non-purchase probabilities were quite consistent with the reality of the events.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Com o decorrer do tempo, o desenvolvimento tecnológico permitiu que diversas invenções ocorressem na área da computação, como por exemplo a criação de sistemas de inteligência artificial, que trouxe grandes mudanças na maneira como pessoas e empresas se relacionam com a tecnologia e com o modo de funcionamento das coisas. Por meio dele, é possível fazer com que uma máquina possua a capacidade de realizar operações matemáticas que se baseam no que se entende como a forma de pensar dos seres humanos, tendo o poder de aplicar regras lógicas a um conjunto de dados, de aprender com os erros e acertos, de reconhecer padrões e de decidir de forma racional.

Estudos sobre métodos de inteligência artificial já existem há muito tempo. Em 1956, John McCarthy, um cientista da computação estadunidense, criou este termo para descrever um mundo em que as máquinas poderiam resolver os tipos de problemas que hoje são reservados para humanos. Porém a tecnologia daquela época não era desenvolvida o suficiente e não possuía um valor acessível. Os computadores necessitavam de bons modelos de dados para processar e analisá-los de forma eficiente, precisavam ter acesso a um banco de dados grande o suficiente para poder alimentar os modelos.

Nos dias atuais, a evolução na capacidade de processamento de dados, na potência dos microcomputadores e de suas memórias fez com que sistemas microprocessados se tornem muito mais ágeis e eficazes ao coletar, manipular e gerar dados de informação. Tal fato implicou no aumento da capacidade de armazenamento de informações, o que possibilitou o acesso a bancos de dados cada vez maiores e completos. Devido a esses progressos, tornou-se possível realizar e aplicar, em larga escala comercial e tecnológica, os estudos de (IA) já existentes desde décadas anteriores, pois com um processamento rápido e uma gama de informações, é possível alimentar modelos já desenvolvidos e implementar outros,

mais completos e robustos.

Para uma (RNA) ser bem estruturada, isto é, estar pronta para ser aplicada na resolução de problemas de escala comercial, atuando em sistemas de apoio a processos de tomada de decisão, são necessários computadores com um bom processamento, visto que para executar uma rede neural é necessário criar um modelo bastante complexo do ponto de vista de matemático, realizando milhares de cálculos por segundo, e que deverá ser treinado para realizar a tarefa de classificar ou prever os dados. Para isso, é preciso ter uma grande quantidade de dados, a fim de que uma parte seja utilizada no processo de aprendizado e outra para testes, onde é feita a averiguação do modelo, verificando se o mesmo foi bem construído para a resolução de um determinado problema.

Redes Neurais Artificiais são implementadas em diversas áreas, como por exemplo no setor financeiro, em que podem ser utilizadas para realizar previsões econômicas ou mesmo na área da saúde, para previsões sobre a ocorrência de certas doenças como hepatite, diabetes ou mesmo alguns tipos de câncer. Neste trabalho, temos como objetivo usar as redes neurais para calcular a probabilidade de um usuário de um site de comércio eletrônico realizar uma compra.

1.1 Contexto do Trabalho

A era da transformação digital, que segundo SIEBEL (2019) [20] é o contexto histórico de tecnologia em que vivemos atualmente e que ainda está em evolução mas que é caracterizado pela insterseção constante entre a computação em nuvem, o grande volume de informações disponíveis e a massiva utilização de equipamentos eletrônicos, impactando em cada aspecto do ciclo de vida e negócios das empresas e fazendo com que a utilização de dados esteja cada vez mais em voga nas mesmas.

Diante disso, a ciência de dados, que segundo CLEVELAND (2001) [21] é o conjunto de técnicas, teorias, análises, parâmetros de observação, algoritmos e princípios que dão suporte ao trabalho com dados, se tornou um importante campo de trabalho para toda companhia que faça uso de softwares, sistemas e plataformas de interação com usuários das mais diversas possíveis, ao aliar o armazenamento massivo de informações a uma capacidade de processá-los, interpretá-los e utiliza-los em seu processo decisório de negócios.

Captar, estruturar e analisar esses dados é parte de um trabalho fundamental,

permitindo que as empresas trabalhem de forma estratégica. Para que tudo isso seja possível, as práticas e o trabalho de Ciência de Dados devem estar devidamente difundidos em cada negócio, para o melhor aproveitamento das informações. Essa nova área de estudos propõe a utilização dos dados em prol de todas as áreas da empresa e, entre elas, o marketing consegue colher resultados estratégicos muito vantajosos. As análises e percepções através dos dados ajudam a preparar melhor as ações e a lidar com as preferências e os hábitos de consumo do público. Assim, a segmentação de clientes e o relacionamento passam a ser mais certeiros e com melhores resultados.

No entanto, para falarmos sobre Marketing Digital, que é a aplicação final deste trabalho, precisamos entender um pouco melhor o contexto do Marketing, e, principalmente do Comércio Eletrônico. Por isso, faremos uma revisão sobre esses temas.

1.2 Introdução ao Problema

Segundo ROSS (1992) [16], a história do comércio eletrônico começa quase um século antes do surgimento da própria internet. A Sears, uma empresa norte-americana que naquela época vendia relógios, começou a desenvolver o conceito e era muito diferente daquele que conhecemos hoje. O princípio básico seria vender produtos à distância. As encomendas chegavam através do telégrafo, após os consumidores escolherem os produtos que pretendiam no catálogo da empresa. A Sears procedia então ao envio do produto... longe de adivinhar que acabava de lançar as bases para uma nova forma de vender e comprar.

O conceito desenvolvido por Richard Sears é reconhecido por ser o primeiro serviço recorrente de catálogo de pedidos por correio, oferecendo relógios, diamantes e jóias, chegando a ter mais de 320 páginas em 1894. Apesar de não ter sido a primeira vez que catálogos por correio haviam sido utilizados para vender algum tipo de produto, a iniciativa de Sears foi a que mais obteve sucesso em médio prazo. Nesse ponto se inicia a história do *e-commerce*.

Esse tipo de serviço de compra à distância seguiu com bastante popularidade até a década de 1990, com diversos catálogos sendo fundados e obtendo sucesso em países como os Estados Unidos, Alemanha, França e Inglaterra. No entanto, com a criação da Internet, não tardou a surgir a ideia de replicar o conceito de comércio à distância neste novo meio em crescimento. Assim, logo em 1979, Michael Aldrich começou a desenhar o conceito de venda eletrônica que tanto se podia aplicar a negócios B2B (*business to business*) como a negócios B2C (*business to consumer*).

Todos os modelos assentavam na participação de consumidores que, a partir dos seus computadores, podiam adquirir bens e serviços de qualquer vendedor na Internet. O engenheiro informático dedicou-se a esta área nos anos que se seguiram.

O início da última década do século XX marcou o começo de um novo capítulo na história do e-commerce. Foi em 1990 que surgiu o primeiro navegador web, o WorldWideWeb – mais tarde renomeado como Nexus – que funcionava como interface para aceder à Internet através de um formato gráfico. Aqui, são lançadas também todas as condições necessárias para a popularização da Internet. Bastariam 10 anos para que a Internet se tornasse um meio importante em todo o mundo, quer a nível pessoal, quer profissional. Com essa popularização rápida da internet não demorou muito até que o processo de comprar através da rede começasse a decolar: Em 1994, uma pizza de cogumelos e pepperoni com queijo extra tratou-se do primeiro produto vendido na Internet, pela PizzaHut. O então PizzaNet – o sistema de encomenda de pizzas da maior cadeia mundial de pizzas – registou a sua primeira venda.

O ano de 1995 foi marcado pelo nascimento de dois dos maiores gigantes do comércio eletrónico. A Amazon e o eBay são lançados pela primeira vez, revolucionando em poucos anos a forma como se vende na Internet. Desde artigos de cozinha, a livros, filmes e até mesmo roupa, estes dois websites provaram que na Internet é possível vender tudo e para qualquer canto do mundo. Esse movimento não demorou a chegar ao Brasil e, já no início dos anos 2000, começam a surgir versões eletrônicas das lojas de grandes varejistas do país como as Lojas Americanas, Submarino e o Grupo Pão-de-Açucar.

Desde então o comércio Eletrônico tem crescido fortemente no país, e segundo GUASTI (2018) [19], mesmo com a forte recessão econômica que veio sobre o Brasil nos últimos 5 anos, o volume de vendas gerado por esse canal de negócios saltou de 28,8 Bilhões de Reais em 2013 para 47,7 Bilhões de Reais em 2017, um aumento de 65,6%. Mais de 55 Milhões de brasileiros realizaram uma compra on-line no último ano, e esse já é o método de compra preferido de quase metade da população brasileira.

Apesar disso, estima-se que a cada 100 visitas que os sites de comércio recebem apenas uma delas se torna uma compra. Por isso, encontrar metodologias que ajudem o processo de conversão das lojas a ser otimizado é de extrema importância, pois, pequenas melhorias percentuais podem levar a ganhos de receita da ordem de

Milhões de Reais.

Para a realização deste trabalho, foram coletados dados de usuários de um site de vendas eletrônicas ao longo do período de julho até agosto de 2018. Os dados coletados representam o momento em que os usuários estão visitando algum produto e estão tomando a decisão de compra. O objetivo final é utilizar as técnicas de redes neurais para calcular a probabilidade de um usuário realizar uma compra. Se esse processo puder ser feito de forma assertiva, é possível identificar usuários com probabilidade de compra mais alta, e assim, oferecer a eles algum tipo de incentivo para que a compra seja de fato realizada, aumentando o número de vendas que a empresa realiza.

1.3 Materiais e Métodos

Este trabalho irá se valer de duas ferramentas pra a construção dos resultados desejados:

i) Software utilizado: Para realizar o presente trabalho utiliza-se o software R, onde serão estruturados os códigos para manipulação de dados e treinamento dos modelos preditivos. R é um ambiente computacional e uma linguagem de programação que vem progressivamente se especializando em manipulação, análise e visualização gráfica de dados. Na atualidade é considerado o melhor ambiente computacional para essa finalidade. O ambiente está disponível para diferentes sistemas operacionais: Unix/Linux, Mac e Windows.

Foi criado originalmente por Ross Ihaka e por Robert Gentleman no departamento de Estatística da Universidade de Auckland, Nova Zelândia. Posteriormente, foi desenvolvido pelo esforço colaborativo de pessoas em vários locais do mundo.

R é altamente expansível com o uso dos pacotes. Os pacotes são bibliotecas com dados e funções para diferentes áreas do conhecimento relacionado a estatística e áreas afins. O R disponibiliza uma ampla variedade de técnicas estatísticas e gráficas, incluindo modelação linear e não linear, testes estatísticos clássicos, análise de séries temporais, classificação, agrupamento e outras.

O R é facilmente extensível através de funções e extensões, e a comunidade de R é reconhecida pelos seus contribuintes ativos em termos de criação de novos pacotes e novas formas de manipular e fazer inteligência com dados. Existem diferenças importantes, mas muito do código que era escrito para a linguagem S

segue inalterado. Muitas das funções padrão do R são escritas no próprio R, o que torna fácil para os usuários seguir as escolhas algorítmicas feitas. Para tarefas computacionais intensivas, códigos em C, C++, e Fortran podem ser utilizados e chamados durante a execução. Usuários experientes podem escrever código C ou Java para manipular diretamente objetos do R.

Outra força do R são os gráficos estáticos, que podem produzir imagens com qualidade para publicação, incluindo símbolos matemáticos. Gráficos dinâmicos e interativos estão disponíveis através de pacotes adicionais. Os gráficos que serão utilizados na ilustração dos procedimentos realizados nesse trabalho foram todos produzidos pelo R, através do pacote de gráficos chamado Plotly.

ii) Redes Neurais Artificiais (RNN): Uma Rede Neural Artificial é construída por meio de modelos computacionais que tentam realizar tarefas ou funções de interesse de forma semelhante ao sistema nervoso central (cérebro), sendo capazes de armazenar conhecimentos, realizando aprendizado de máquina bem como reconhecendo padrões.

RNAs em geral são apresentadas como um sistema de neurônios que captam valores de entrada e geram valores de saída, simulando assim, o comportamento de uma rede neural natural biológica.

Para o desenvolvimento deste trabalho, é esperada a utilização de Redes Neurais com ao menos uma camada intermediária, utilizando funções de ativação do tipo Logística para as camadas de entrada e aplicando a função softmax na saída pois isso irá permitir o calculo de probabilidades como saída da Rede para as entradas produzidas. Dedicaremos um capítulo a parte para tratarmos de uma revisão bibliográfica a respeito desse tema.

Capítulo 2

Marketing Orientado a Dados

O marketing orientado a dados (*data-driven marketing*) é uma estratégia na qual os profissionais de marketing obtêm informações e tendências ao analisar dados gerados pela empresa ou pelo mercado, transformando essas ideias em decisões acionáveis e estratégias de comunicação com os clientes. No geral, trata-se de tomar decisões mais objetivas sobre a estratégia de marketing ao se basear em dados. No entanto, para entendermos melhor como o Marketing avançou ao longo dos anos, é preciso olhar um pouco para a história.

2.1 O que é Marketing?

Marketing ou mercadologia ou, mais raramente, mercância, é a arte de explorar, criar e entregar valor para satisfazer as necessidades do mercado. São usados produtos ou serviços que possam interessar aos consumidores. Para isso é necessário criar uma estratégia definida que será utilizada nas vendas, comunicações e no desenvolvimento do negócio. A finalidade do marketing é criar valor e chamar a atenção do cliente, gerando relacionamentos lucrativos para ambas as partes.

Vários autores definiram o marketing de várias maneiras. A Associação Americana de Marketing define marketing como o processo de planejamento e execução da concepção, preço, promoção e distribuição de idéias, bens e serviços para criar trocas que satisfaçam os objetivos individuais e organizacionais.

Segundo CRONJE E DU TOIT (2004) [7], o marketing consiste em tarefas e decisões de gerenciamento direcionadas para o êxito em oportunidades e avaliação de ameaças em um ambiente dinâmico, desenvolvendo e transferindo efetivamente um oferta de mercado aos consumidores, de forma que os objetivos do negócio, do consumidor e da sociedade sejam alcançados.

2.2 História do Marketing ao Longo do tempo

Vários são os estudos que levam a visões diferenciadas sobre a História do Marketing, fato é que elas se complementam e apresentam uma visão mais ampliada quanto a sua origem.

Segundo DONATO (2012) [8], em meados do século XVIII, na Inglaterra, inicia-se a Revolução Industrial, onde a produção de bens artesanais é tomada pela produção mecanizada. Em decorrência a isso, com uma produção mais acelerada, a população começa a ter acesso a produtos com preços mais interessantes. Nessa época a concorrência ainda era inexistente, o que deixava os consumidores alheios a poucos fornecedores ou até mesmo um só, não tendo muitas opções de escolha. Nessa época inicia-se o estudo do mercado, focado em logística e produtividade.

No entanto, após a Segunda Guerra Mundial, a expansão econômica dos países se torna mais notável, com isso, muitos setores antes com características de monopolio, passam a enxergar concorrência. Logo, ideias de atração e relacionamento com o consumidor começam a ser apresentadas pelos mercadólogos. Assim, dá-se início ao "vender a qualquer preço". O Marketing ainda não era visto como a ação para satisfazer desejos e necessidades, mas sim como uma ferramenta cujo objetivo era a necessidade de vender.

Em 1940 iniciam-se discussões sobre a possibilidade em se desenvolver uma teoria mercadológica em termos científicos. Em 1954, é lançado o livro "A Prática da Administração", autoria de Peter Drucker, em que atenta os administradores sobre a importância do Marketing, revelando-o como uma ferramenta importante para se conhecer o mercado.

Em 1960, é enfatizada a importância da satisfação dos clientes, por Theodore Levitt (professor de Havard Business School). Em 1970, com o crescimento da concorrência, as empresas se atentam e começam a implantar departamentos de marketing e até mesmo os setores governamentais também começam a adotar as estratégias de marketing, como o que pode ser amplamente visto durante o período de governo militar no Brasil.

Em 1980, amplia-se ainda mais a atenção ao cliente, especialmente pelo lançamento do livro "Em busca da Excelência" de Tom Peters e Bob Waterman. Nessa mesma época o marketing passa a receber mais destaque e atenção pelas grandes corporações. Na década de 1990 com a internet revoluciona-se a logística,

a distribuição e as formas de pagamento, a gestão de relacionamento com o cliente é fortalecida, e também a responsabilidade social.

Na década de 2000, na visão de KOTLER, KARTAJAYA E SETIAWAN (2010) [10], há expansão dos meios de comunicação e da utilização de tecnologias que oferecem o poder da informação ao consumidor, trazendo uma nova visão do cliente sobre o processo de compra.

Segundo KOTLER, KARTAJAYA E SETIAWAN (2010) [10], o marketing evoluiu ao longo dos anos, passando por três fases que eles chamam de marketing 1.0, 2.0 e 3.0. GRACIOSO (1997) [11] coloca a ideia de que as mudanças no marketing são consequências de três revoluções vividas pela sociedade, a revolução industrial, a revolução industrial tecnológica e a revolução da informação, cada uma influenciando diretamente a forma das empresas fazerem marketing. Já BOONE E KURTZ (2011) [12] dividem a evolução do marketing em quatro eras distintas: da produção, das vendas, do marketing e do relacionamento, sendo que a era da produção representa o marketing 1.0, as eras das vendas e do marketing representam o marketing 2.0, e a era do relacionamento a do marketing 3.0.

Na fase do marketing 1.0, as estratégias eram centradas no produto, e seu objetivo era vender os produtos da fábrica a todos que quisessem comprá-los. A comunicação era feita da empresa para o cliente de forma vertical, sem a troca de informações e o foco era padronizar e ganhar em escala, reduzindo assim os custos e, consequentemente, os preços. O exemplo que melhor resume esta estratégia é a frase de Henry Ford: "O carro pode ser de qualquer cor, desde que seja preto". Nesse período, a atenção dos fabricantes era apenas com a qualidade de seus produtos.

Segundo GRACIOSO (1997) [11], nessa fase o marketing era guiado apenas pela teoria da oferta e procura, sem reconhecer que a demanda deve ser criada, ajustando diferenças entre oferta e procura apenas com o ajuste de preços. Este pensamento se devia muito ao padrão de vida que a população da época tinha, pois a grande maioria das pessoas nas cidades não tinha boas condições financeiras, consequentemente, não tinham dinheiro sobrando para escolher comprar o que mais lhe agradava. As compras eram basicamente por necessidade e as empresas se preocupavam, basicamente, em melhorar seus processos de produção, com o objetivo de conseguir reduzir custos e, consequentemente, reduzir seus preços.

O cenário citado começa a mudar ao redor do mundo a partir da década de 1920, impulsionado por uma elite em evolução, que começava a exigir produtos e serviços

de melhor qualidade. A segunda revolução industrial, chamada por GRACIOSO (1997) [11] de *revolução da alta tecnologia aplicada à produção*, melhorou o padrão de vida da população. Com mais dinheiro no bolso os clientes passaram a comprar mais e, consequentemente, exigir mais das empresas. Graças a uma renda discricionária mais alta, as pessoas passaram a comprar mais bens e serviços supérfluos. Mas, principalmente, passaram a exercer mais poder de escolha das marcas e dos fornecedores. Com o tempo, as empresas começam a perceber que existem outros fatores que influenciam na decisão de compra do cliente, dentre eles impulsos e motivações psicológicas que não estão atrelados à lei da oferta e da procura.

KOTLER, KARTAJAYA E SETIAWAN (2010) [10] chamam esta fase de marketing 2.0, em que o cliente é bem informado e a concorrência é maior, o que aumenta a quantidade de produtos que o consumidor tem a sua escolha. Isto faz com que a comunicação se inverta, então é o consumidor quem estipula como quer o produto e quanto quer pagar, a interação entre empresa e cliente fica maior. Sendo assim, o profissional de marketing precisa segmentar o mercado e criar produtos cada vez mais especializados.

As campanhas de marketing visam um lado mais emotivo, tentando atingir o coração e mente dos consumidores, criando neles, desejos de compra, mesmo sem existir a necessidade. Isto faz com que os profissionais de marketing tenham uma necessidade de conhecer melhor seu público alvo. As campanhas de marketing buscam personificar as marcas, tentando criar uma relação mais pessoal entre cliente e empresa. As redes sociais surgem como principais ferramentas utilizadas para atingir este objetivo. O cliente não aceita mais a relação vertical, ele quer fazer parte do processo. Ainda segundo esses autores, há uma mudança da atual era do marketing 2.0 entrando no marketing 3.0.

Nesta nova fase, não basta apenas a empresa incluir o cliente no processo de produção e venda, ela precisa retornar algo para a sociedade, a lucratividade tem como contrapeso a responsabilidade corporativa, segundo KOTLER, KARTAJAYA E SETIAWAN (2010) [10]. Essa terceira fase é chamada de revolução da informação, mas dá maior importância para a ampliação da possibilidade de captação e interpretação dos dados, aumentando a segmentação de mercado de uma forma jamais vista, tratando esta nova fase como *a era do indivíduo*. Nesta fase os clientes estão melhor informados e têm acesso aos dados das empresas, isso mostra que uma empresa que comete deslizes sociais, potencialmente será descoberta e terá tais deslizes divulgados, o que pode afetar negativamente sua imagem.

BOONE E KURTZ (2011) [12] chamam essa fase de era do relacionamento e enfatizam a necessidade das empresas criarem relacionamentos fortes e duradouros, não só com os consumidores, mas com todos os steakholders da empresa, essa nova era coloca os consumidores no centro das atenções da empresa, com o objetivo de que eles se tornem clientes fixos, pois se percebeu que conquistar novos clientes é mais caro para a empresa do que manter os mesmos.

A evolução do marketing exposta, deu-se em função das mudanças existentes na sociedade e teve como objetivo destacar a empresa na preferência dos consumidores. Como já foi explanada, a forma encontrada pelas empresas para se destacar na preferência dos consumidores tem sido investir num estreitamento do relacionamento entre eles, isso é melhor visto na teoria da era do relacionamento, defendida por BOONE E KURTZ (2011) [12], mas também aparece nas evoluções defendidas por KOTLER, KARTAJAYA E SETIAWAN (2010) [10]. Para melhorar esse relacionamento, as empresas começaram a criar novas formas de se aproximar do cliente e de promover seus produtos ou serviços de uma forma mais pessoal

2.3 Marketing Digital

Segundo COBRA (2010) [23], o Marketing digital é um conjunto de ações de comunicação que as empresas podem utilizar por meio da internet, da telefonia celular e outros meios digitais, para assim divulgar e comercializar seus produtos, conquistando novos clientes e melhorando a sua rede de relacionamentos. Ele engloba a prática de promover produtos ou serviços pela utilização de canais de distribuição eletrônicos, para então chegar aos consumidores rapidamente de forma relevante, personalizada e com mais eficiência.

Esse tipo de marketing traduz-se em ações adaptadas aos meios digitais, de forma a obter, nestes canais, a mesma eficiência e eficácia do marketing direto e, simultaneamente, potencializar os efeitos do marketing tradicional. Na sua operacionalização são, normalmente, utilizados canais, meios e ferramentas digitais.

Toda a construção de um plano de ação se inicia com uma visão, e, junto a ela, um objetivo a ser alcançado. A partir deste traço são definidos os meios a serem utilizados, para enfim atingir os resultados. O tratamento singular a cada cliente é o segredo para que a escolha de um plano de ação seja de fato apropriada, somente o conhecer de um cliente ou projeto em questão fará com que os esforços de mídia potencializem o trabalho de marketing, o tornando assim digital.

Um exemplo de sucesso, segundo CUEN (2014) [24], é o Facebook, que conquistou rapidamente o público, e, em pouco tempo, ultrapassou os seus concorrentes. A chave foi a simplicidade, a rapidez de execução e a liberdade que o usuário tem ao manusear as suas funções. O uso da rede fez tanto sucesso, que as empresas passaram a investir na criação de um ambiente de comunicação com o cliente, permitindo que criassem uma página de negócios, as *fan pages*.

Com o marketing aplicado ao meio digital um novo consumidor surgiu: o consumidor 2.0. Este gênero de consumidor é um ser exigente, informado, atento, e que procura sempre registros que lhe possibilitem saber mais antes de sair do conforto de sua casa. A troca de ideias com outros consumidores também é natural.

Segundo MACHADO E DAVIM (2016) [13], nos dias de hoje, além de um subtipo do marketing convencional, o marketing digital tem se tornado um novo fenômeno que reúne personalização e distribuição em massa para atingir as metas de negócio. A convergência tecnológica e a multiplicação de dispositivos levaram a uma abertura das maneiras pelas quais pensamos marketing na Internet e empurraram as fronteiras para um novo conceito de marketing digital - centrado no usuário, mais mensurável, onipresente e interativo. As estratégias de desenvolvimento de marketing digital oferecem muito potencial para marcas e organizações. Alguns deles são os seguintes:

- Fortalecimento de Marca: Plataformas e serviços online são uma grande oportunidade para construir uma forte imagem da marca na internet, devido ao seu escopo, presença e atualizações constantes.
- Completude: As possibilidades de disseminar informações através de links oferecem aos consumidores a chance de abordar a organização de maneira mais ampla e personalizada.
- Usabilidade - funcionalidade: A internet oferece plataformas simples e fáceis de usar em quase todo tipo de problema visando melhorar a experiência do usuário e permitir suas atividades.
- Interatividade: No contexto em que as organizações tentam forjar, a longo prazo, um relacionamento com seu público, a Internet oferece a possibilidade de ter uma conversação mais direta e, portanto, de gerar uma experiência positiva com a marca. Essa interatividade pode ser básica, como avaliação do produto, ou tornar-se um experiência abrangente.
- Comunicação visual: Alinhado ao pensamento visual, o marketing digital oferece aos profissionais diferentes ferramentas baseadas em imagem e vídeo. Esta

é uma maneira atraente de atingir públicos que podem levar a um maior envolvimento com os produtos e serviços.

- Publicidade relevante: Fácil segmentação e personalização de publicidade na Internet maximizam os resultados. Além disso, livre das limitações de outras mídias, como Televisão, rádio ou jornais, esse ambiente permitiu uma publicidade mais atraente ao público.

- Viralização: A essência da Internet como uma rede de nós interconectados torna a expansão exponencial de qualquer conteúdo possível. Tomando como exemplo o modelo de comunicação *boca a boca*, a comunicação viral se torna mais relevante devido à conectividade, instantaneidade e capacidade de compartilhamento de plataformas on-line que melhoram a divulgação do conteúdo.

- Medição de Resultados: As plataformas online estão em primeiro lugar na disponibilidade de acompanhamento de campanhas realizadas tornando extremamente viável a possibilidade de avaliar a produção de uma campanha, algo que é muito mais complicado de se realizar em outros meios de comunicação.

2.4 Big Data

Big data é um termo recente e por isso não existente na maior parte de dicionários de estatística. São dados multivariados e de elevada dimensão, geralmente criados em tempo real e apresentam um crescimento exponencial (na escala temporal), nomeados de megadados.

Quanto mais dados são gerados, maior é o esforço para extrair informações relevantes, e os centros de dados tiveram que aprender a lidar com o crescimento exponencial dos dados gerados e tiveram que desenvolver ferramentas que fossem para além de bancos de dados relacionais e sistemas paralelos de bancos de dados.

O uso dos primeiros equipamentos para processar dados datam de 1890, durante a realização do Censo dos Estados Unidos, conduzido pelo *U.S. Census Bureau*. Na ocasião, a Máquina de Tabulação construída por Herman Hollerith diminuiu o tempo de processamento dos dados para apenas 6 semanas segundo BOHME (1991) [25]. Entretanto, somente no século XX começaram a surgir os primeiros sistemas para armazenamento de informações. Em 1927, o engenheiro Fritz Pfleumer criou um método para guardar informações em fitas magnéticas.

Durante a Segunda Guerra Mundial, foi criada a primeira máquina digital de processamento de dados. Foi em 1943, quando os Britânicos desenvolveram um sistema para decifrar códigos nazistas durante a Segunda Guerra Mundial. O nome da máquina era Colossus, que podia interceptar mensagens a uma taxa de 5000 caracteres por segundo. O primeiro órgão público criado especificamente para o processamento de dados, a Agência Nacional de Segurança (NSA) dos EUA, foi fundado em 1952, com o objetivo de processar dados automaticamente para obter informações relativas a inteligência durante a Guerra Fria.

Um dos primeiros Centros de Dados foi criado em 1965, também pelo governo americano, com o objetivo de controlar o pagamento de impostos e as impressões digitais dos americanos. Este Centro de Dados possuía o mesmo padrão dos bancos de dados criados até a década de 1970. Eram bancos de dados centralizados, onde uma mesma máquina era responsável pelo uso, armazenamento e análise dos dados. Com o aumento da quantidade de dados, começaram a surgir novas arquiteturas de dados que permitissem processar e analisar esses dados. Nas década de 80 começaram a surgir os Sistemas de Bancos de Dados Paralelos. Nesse caso, ao invés de um banco de dados centralizado, cada processador se comunica com os outros apenas enviando mensagens através de uma rede interconectada. Os primeiros bancos de dados paralelos possibilitaram a criação do primeiro banco de dados com capacidade em terabytes, pela KMART, em 1986.

Em 1989, o cientista britânico Tim Berners-Lee criou a World Wide Web, para facilitar a troca de informações entre as pessoas dentro do contexto de gerenciamento de experimentos no CERN. Uma transcrição do artigo original de Tim propondo a solução que daria origem ao mundo *online* que temos hoje pode ser encontrada no seguinte link https://www.w3.org/History/1989/proposal.htm. O que Tim não sabia era que sua invenção iria revolucionar a forma como os dados eram gerados e a quantidade de dados criados. O termo big data foi usado pela primeira vez em 1997, entretanto o nome começou a ser usado oficialmente em 2005, quando Roger Mougalas, da O'Reilly Media publicou um artigo mencionando o tema.

Os dados que agregam o conjunto do big data são provenientes de várias fontes. Desta maneira, normalmente não apresenta uma estrutura bem definida, ou seja, não pode ser armazenada nos sistemas padrões de banco de dados, como o Sistema Gerenciador de Banco de Dados Relacional (SGBDR), onde os dados são representados por meio de tabelas, com diversas linhas e colunas.

Os engenheiros de dados começaram a verificar que bancos de dados relacionais não conseguiriam suportar essa grande quantidade de dados não estruturados. Desta maneira, novas tecnologias e processos tiveram que ser desenvolvidos para permitir que esses dados não estruturados fossem analisados, já que os mesmos podem representar até 80% do total de dados. Foi quando a Google criou o MapReduce, em 2004, que é um modelo de programação que permite processar grandes quantidades de dados em paralelo, dividindo o trabalho em um conjunto de tarefas independentes, geralmente executado em um cluster de computadores.

Posteriormente, foi desenvolvido o Hadoop, que é uma implementação em código aberto do MapReduce. O Hadoop foi criado pelo Yahoo em 2005 e pode ser considerado uma das maiores invenções de *data management* desde o modelo relacional. Entretanto, o Hadoop não é considerado uma base dados como o SGBDR. Ele é um sistema de distribuição de arquivos utilizado para processar e armazenas grande quantidade de dados (big data) por meio de clusters, onde os mesmos são processados paralelamente e podendo ser executados em servidores sem muito esforço. Atualmente, esse tipo de processamento é o mais utilizado por empresas que trabalham com big data e diversas empresas vêm contribuindo com código para seu desenvolvimento, como a Yahoo, Facebook, Cloudera, IBM e outras.

Segundo a IBM, em 2008 foram produzidos cerca de 2,5 quintilhões de bytes todos os dias e surpreendentemente 90% dos dados no mundo foram criados nos últimos dois anos, decorrente a adesão das grandes empresas à internet, como exemplo as redes sociais, dados dos GPS, dispositivos embutidos e móveis.

Atualmente, a Internet das Coisas, revolução tecnológica que tem como objetivo conectar os itens usados do dia a dia como eletrodomésticos, meios de transporte e até mesmo tênis, roupas e maçanetas à rede mundial de computadores, mudou a forma como os dados são gerados, aumentando de forma abrupta a quantidade de dados gerados.

Todos esses objetos físicos da Internet das Coisas são capazes de coletar e transmitir dados, gerando dados que contribuem para facilitar a vida das pessoas, fazendo com que as *coisas* conheçam bem seus donos e saibam adaptar suas funções de modo a oferecer experiências personalizadas de uso.

2.5 Marketing Orientado a dados

O que vimos até aqui é que o Século XXI trouxe algumas novidades para o mundo do Marketing:

i) Com o advento da internet e da democratização do acesso a informação, qualquer pessoa que queira comprar alguma coisa está muito mais exigente: Ela conhece qual a loja tem o menor preço, qual possui o melhor atendimento ou a melhor logística de entrega. Ou seja, os consumidores estão cada vez mais bem preparados.

ii) Novas tecnologias de processamento de dados fazem com que seja possível armazenar muito mais informações sobre qualquer que seja o assunto independemente da origem da informação.

iii)Quanto mais as pessoas usam a internet, entram em diferentes sites e fazem diferentes interações, mais informações sobre elas está disponível.

Aqui, o marketing orientado a dados se cria. Segundo KOTLER (2005) [9], O objetivo do marketing é tornar as vendas supérfluas, conhecendo e compreendendo o cliente de tal forma que o produto se venda por si próprio. E para que isso aconteça, nada melhor do que poder analisar o comportamento e as informações do cliente. Com isso, a união dos três fatores supracitados faz nascer o conceito que estamos buscando. Afinal, se até mesmo os produtos estão aprendendo sobre as pessoas, é fundamental que quem se dispõe a vende-los também o faça.

De forma prática, SAURA (2020) [26] sustenta que o Marketing orientado a dados pode ser visto como uma união entre os campos do Marketing e da Ciência de Dados. Com ambas as áreas trabalhando de forma unida, é possível gerar maximização de satisfação para clientes e para as empresas, conforme descreve JEFFERY (2010) [27]. Nesse sentido, todas as iniciativas que visam a utilização por parte das empresas das informações disponíveis a respeito de seus clientes, e dos dados atrelados a eles, visando a criação de um momento de compra mais personalizado, rico, otimizado e encantador, pode ser entendido como Marketing Orientado a Dados.

Capítulo 3

Redes Neurais Artificiais

O alvo principal deste trabalho é utilizar-se das técnicas de Redes Neurais para fazer previsões. Com isso, faz-se necessária a realização de uma revisão a respeito do tema. Esse capítulo, portanto, será dedicado ao estudo desse teoria.

3.1 Redes Neurais Biológicas

Segundo TAFNER, XEREZ e RODRIGUES FILHO (1996) [6], o cérebro humano possui, pelo que se conhece, cerca de 10 bilhões de neurônios. São eles as células mais diferenciadas do organismo, pois apresentam a maior complexidade estrutural e funcional. Um neurônio é capaz de criar até 10.000 sinapses, ou seja, até 10.000 conexões com neurônios adjacentes. São células que tem como principal característica ter a capacidade de gerar e conduzir impulsos nervosos.

Um neurônio tem quatro componentes principais: dendritos, corpo celular (soma), axônio e terminais sinápticos, representados na Figura 3.1. Os dendritos são mecanismos responsáveis por receber impulsos nervosos de neurônios vizinhos e conduzir esses impulsos até o corpo celular. O corpo celular é também um local de recepção de estímulos, através de contatos sinápticos. O axônio é muito longo e fino, é especializado em transmitir impulsos, como cargas elétricas, que emergem dos neurônios, gerando potenciais de ação. Por esse motivo, este componente possui uma alta resistência elétrica e uma capacitância grande. Por último, os terminais sinápticos são locais de contato com a célula vizinha, podendo ser outro neurônio ou uma célula muscular. É o ponto de contato entre a terminação axônica de um neurônio e o dendrito de outro, onde ocorrem as sinapses.

Sinapse é uma ligação em que o terminal do axônio faz contato com outro neurônio. Por meio das sinapses são feitas as transferências de informação de um neurônio para outro, que se dão por meio de moléculas de íons. Esse processo é

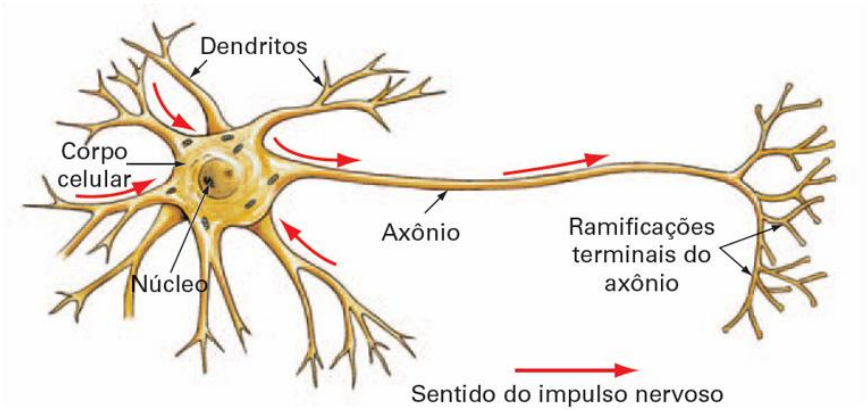

Figura 3.1: Estrutura base de um neurônio

chamado de transmissão sináptica. O lado pré-sináptico geralmente consiste de um axônio terminal, enquanto que o lado pós-sináptico pode ser o dendrito ou o soma de outro neurônio.

Os neurônios não realizam divisão celular, sendo assim, quando um neurônio é destruído, sua perda é permanente. Por outro lado, seus prolongamentos podem se regenerar e podem se modificar. O sistema nervoso possui plasticidade, isto é, possui capacidade de se remodelar em função de novas experiências ou lesões, reformulando suas conexões para se adaptar ao meio ambiente. Em um cérebro adulto, a plasticidade pode ser atribuída a dois mecanismos: a criação de novas conexões sinápticas entre neurônios e a modificação das sinapses existentes. Logo, a plasticidade pode estar relacionada a um processo de aprendizado.

De forma resumida, conforme HAYKIN (2009) [5] apresentou, o sistema nervoso pode ser visto como um sistema de três estágios, conforme se observa na Figura 3.2. O cérebro é o centro do sistema (rede neural): ele recebe as informações, as interpreta e toma decisões apropriadas. As informações são transmitidas pelos receptores, que transformam estímulos do corpo humano ou do ambiente externo em impulsos elétricos. Já os atuadores convertem impulsos gerados pela rede neural em respostas, como saídas do sistema.

Figura 3.2: Representação dos blocos que compõem o sistema nervoso

Dois conjuntos de setas são mostrados na figura acima: as que estão na parte superior indicam a transmissão da informação. As na parte inferior mostram a realimentação do sistema. Ele também é realimentado, isto é, com base nas informações resultantes, há o aperfeiçoamento da próxima tentativa, como um processo de aprendizagem. Sendo assim, os atuadores enviam informações que são captadas pela rede neural, que gera impulsos que são captados pelos receptores.

3.2 Redes Neurais Artificiais

Uma Rede Neural Artificial é construída por meio de modelos computacionais que tentam realizar tarefas ou funções de interesse de forma semelhante ao sistema nervoso central (cérebro), sendo capazes de armazenar conhecimentos, realizando aprendizado de máquina bem como reconhecendo padrões. RNAs em geral são apresentadas como um sistema de neurônios que captam valores de entrada e geram valores de saída, simulando assim, o comportamento de uma rede neural natural biológica.

Segundo HAYKIN (2009) [5], uma rede neural é um processador paralelamente distribuído e se assemelha ao cérebro em dois aspectos: o conhecimento é adquirido por ela a partir de seu ambiente por um processo de aprendizagem e as ligações entre os neurônios, chamados de pesos sinápticos, são utilizadas para armazenar o conhecimento adquirido. Para o processo de aprendizagem é realizado um método chamado de algoritmo de aprendizagem, que modifica os pesos sinápticos da rede, de forma ordenada, para que assim, um objetivo específico seja alcançado. O autor também assinala que as RNAs possuem características muito úteis e que são muito relevantes, são elas:

- Não-linearidade: Um neurônio artificial pode ser linear ou não-linear. Se a rede é constituída de neurônios que não são lineares, então a rede como um todo será não-linear. A não-linearidade possibilita considerar o comportamento nãolinear dos fenômenos físicos responsáveis pela geração dos dados de entrada.

- Mapeamento entre entradas e saídas: Os pesos sinápticos de uma rede neural podem ser modificados para reduzir a diferença entre a resposta desejada e a resposta real da rede. Esse procedimento é denominado *aprendizagem supervisionada* ou aprendizagem obtida por meio de um tutor. A modificação desses pesos é feita por meio da aplicação de amostras de treino, que possuem respostas de saída previamente desejadas. O treinamento é repetido para muitos conjuntos até que a rede neural se estabilize e não haja mais diferenças significativas nos pesos sinápticos. Para o conjunto de amostras utilizadas no

treinamento, não são feitas suposições prévias sobre o modelo estatístico desses dados, pois se houver uma relação implícita entre os dados, mesmo que não seja conhecida sua distribuição, as redes são capazes de apresentar um bom desempenho.

- Adaptabilidade: As redes neurais possuem a capacidade de adaptar seus pesos sinápticos em ambientes dinâmicos. Elas aprendem, de forma rápida, padrões complexos e possíveis tendências presentes nos dados. Uma RNA treinada para operar em um ambiente específico pode ser facilmente "retreinada" para lidar com alterações nesse meio. Quando operam em um ambiente onde as estatísticas mudam com o tempo (não-estacionário), uma rede neural pode ser planejada para alterar seus pesos em tempo real.

- Generalização: O processo de generalização está relacionado ao fato da rede neural produzir saídas (respostas) adequadas para valores de entradas que não estavam presentes durante o treinamento.

- Resposta Evidencial: Em contextos de classificação de comportamentos, uma RNA pode ser construída de forma a fornecer informações sobre qual o padrão escolhido e qual a relevância da escolha, o que possibilita uma melhor análise em casos de ambiguidade de escolha.

- Tolerância às falhas: Uma RNA é capaz de realizar computação robusta, apresentando resultados aceitáveis, no caso de falhas de alguns neurônios. O dano nesses neurônios deve ser extenso para que a resposta global da rede seja danificada seriamente. Isso ocorre pois as informações da ree acabam ficando distribuídas pelos neurônios, ou seja, todos eles absorvem uma parte do conhecumento adquirido pela rede.

- Informação contextual: Um neurônio é potencialmente afetado pela atividade de todos os outros neurônios da rede neural.

3.2.1 Modelagem de um Neurônio

Um neurônio artificial, segundo HAYKIN (2009) [5], é uma unidade de processamento de informação, sob forma matemática, que busca simular o comportamento de um neurônio biológico. Na figura 3.3 podemos ver um tipo de modelo de neurônio muito utilizado na criação de RNAs. Podemos identificar tmabém os principais componentes desse modelo de neurônio. São eles:

- Conjunto de sinapses: são caracterizados pelos seus pesos associados. Diferentemente de uma sinapse cerebral, o peso sináptico pode assumir valores negativos além de valores positivos.

- Função soma: Auxilia na sumarização dos sinais, somando-os ponderadamente com relação aos pesos sinápticos.
- Função de Ativação: Atua limitando a amplitude da saída do neurônio

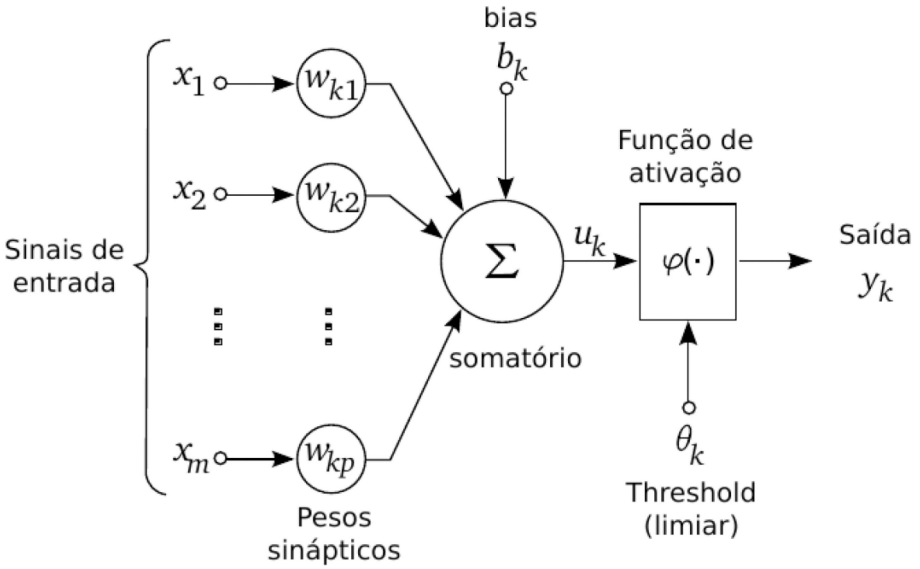

Figura 3.3: Modelo não-linear de um neurônio artificial

Sendo assim, em termos matemáticos, podemos descrever um neurônio k através do seguinte par de equações:

$$\mu_k = \sum_{j=1}^{p} w_{kp} \cdot x_j \tag{3.1}$$

$$y_k = \varphi(\mu_k + b_k) \tag{3.2}$$

Onde x_1, x_2,$\cdots$, x_p são os sinais de entrada, w_{k1}, w_{k2},$\cdots$, w_{kp} são os pesos sinápticos do neurônio k, μ_k é uma combinação linear entre os pesos e os sinais de entrada, b_k é o viés [1], φ é a função de ativação e y_k é o sinal de saída do neurônio.

3.2.2 Processos de Aprendizado

Nesta seção, começamos explicando os paradigmas de aprendizado das redes. O processo de aprendizado de uma rede neural é um fator de extrema importância. Nesse processo a rede possui a capacidade de aprender com as mudanças ocorridas no seu meio e com isso, ela melhora seu desempenho, visto que se adequa a essas

[1] O viés é um parâmetro adicionado para ajustar ainda mais a rede às entradas. Funciona de forma similar ao intercepto que é encontrado em equações lineares.

alterações, passando a funcionar da forma mais ajustada possível. Uma RNA aprende sobre seu meio através de um processo de ajustes aplicados em seus pesos sinápticos e essa técnica pode ser vista como uma forma de treinamento atribuído à rede. O aprendizado só ocorre quando a rede neural atinge uma solução generalizada para uma classe de problemas.

De acordo com HAYKIN (2009) [5], a aprendizagem é um processo pelo qual os parâmetros livres de uma rede neural são adaptados através de uma prática de estimulação pelo ambiente no qual a rede está inserida. O tipo de aprendizagem é determinado pela maneira pela qual a modificação dos parâmetros ocorre. Esse processo ocorre da seguinte maneira: a rede é estimulada pelo ambiente; depois, sofre alterações nos seus parâmetros livres, como resultado da estimulação ocasionada; por fim, a rede neural responde de uma outra forma ao ambiente, devido às alterações ocorridas na sua estrutura interna. O procedimento de alteração dos pesos, ou dos parâmetros livres, ocorre diversas vezes, pois no processo de aprendizagem a rede precisa passar por diversas iterações. Desse modo, após o estímulo inicial, são gerados pesos sinápticos iniciais, seguidos de novas estimulações, que geram novas mudanças nos pesos, até chegar em um ponto aceitável e o modelo conseguir se tornar generalizado.

Todo processo de aprendizagem está relacionado aos dados. Logo, a rede se fundamenta nos mesmos para extrair um modelo geral. A fase de aprendizado deve ser rigorosa e verdadeira, para que não surjam modelos falsos e a rede não consiga entender e aprender sobre o meio em que está inserida. Desse modo, de 50% a 90% dos dados devem ser separados para serem utilizados no treinamento da rede, para que assim a mesma aprenda, de fato, as regras. Os dados que sobram são utilizados para teste, tendo a finalidade de verificar se a rede se ajustou de forma correta para aquele meio.

Existem diversas formas de aplicar o aprendizado em uma RNA. Denomina-se algoritmo de aprendizado um conjunto de regras bem definidas para a solução de um problema de aprendizado. Existem muitos tipos de algoritmos específicos para determinados modelos de redes neurais, eles diferem entre si principalmente pelo modo como os pesos são modificados. Serão apresentados dois paradigmas de aprendizado: o aprendizado supervisionado ou aprendizado com um professor e o aprendizado não supervisionado. Abaixo, na Figura 3.4, segue um diagrama explicando para quais finalidades cada tipo de aprendizado é útil.

Na aprendizagem supervisionada, ocorrem duas fases. Na primeira, a rede

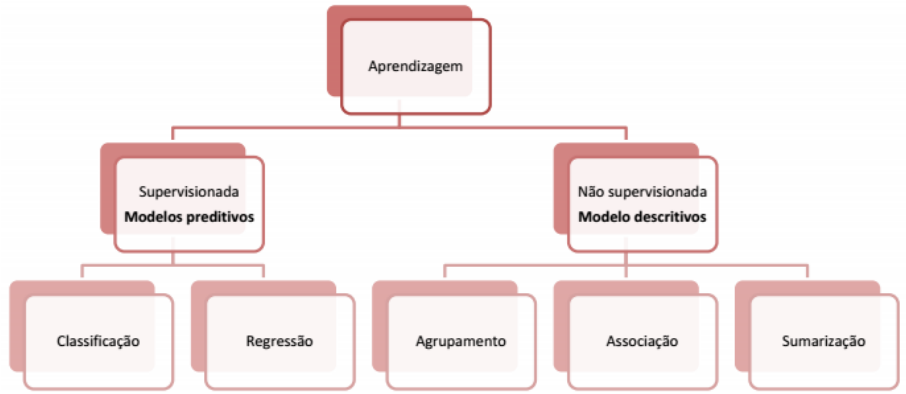

Figura 3.4: Tipos de aprendizagem e tipos de problemas normalmente associados

neural é treinada para retornar valores desejados, por um professor que já conhece o ambiente, ela recebe exemplos de valores de entrada e saída. Em virtude do seu conhecimento, o professor consegue fornecer respostas desejadas (saída) após receber um vetor de treinamento retirado do meio em que a rede está inserida (entrada). A resposta desejada representa a ótima ação realizada pela rede, que retornará respostas que o professor acha adequadas para o problema. Os parâmetros da rede são ajustados levando em consideração o vetor de treinamento e o sinal de erro, que é representado pela diferença entre a resposta desejada e a resposta real retornada pela rede. O sinal de erro é realimentado no sistema. Esse procedimento é realizado até que a rede neural consiga simular o professor, isto é, consiga gerar respostas que sejam adequadas, ajustando seus parâmetros de forma correta.

Após a rede concluir seu treinamento, o professor é dispensado e ela por si só lida com o ambiente, representando a segunda fase.

No aprendizado não supervisionado não existe um professor ou uma supervisão. Nesse tipo de aprendizagem, a rede descobre sozinha relações, padrões nos dados que são apresentados. O objetivo desse método é fazer com que a rede detecte padrões e ajuste seus parâmetros a partir de diversas iterações feitas com os dados de entrada, provenientes do ambiente. Sendo assim, após realizar o aprendizado, a rede será capaz de codificar as entradas e realizar as classificações conforme vai percebendo novos padrões.

3.2.3 Modelo Perceptron

A partir daqui, abordaremos o algoritmo conhecido como Perceptron, uma versão mais simples para a rede neural. Este método, composto por apenas um neurônio, procura separar do hiperplano duas classes linearmente separáveis (ω_1, ω_2). Posteri-

ormente, descrevemos os conceitos teóricos e os detalhes do algoritmo de aprendizado para treinar uma rede neural.

Em duas classes linearmente separáveis (ω_1, ω_2), existe pelo menos um hiperplano, $\mathbf{w}_o^T\mathbf{x} = 0$,que divide as duas regiões e consegue classificar corretamente o conjunto de amostras de treinamento, $\{(\mathbf{x}(1), y(1)),(\mathbf{x}(2), y(2)), \cdots, (\mathbf{x}(M), y(M))\}$, tal que:

$$y(m) = \text{sign}(\mathbf{w}_o^T\mathbf{x}(m)) = \begin{cases} -1, & \text{if } \mathbf{x}(m) \in \omega_1, \\ +1, & \text{if } \mathbf{x}(m) \in \omega_2. \end{cases} \tag{3.3}$$

O termo de viés do hiperplano foi incluído em $\mathbf{w}$ para simplificar a notação.Pela definição acima da função de sinal, temos que, $y(m)(\mathbf{x}^T(m)\mathbf{w}) > 0$ se e somente se $\{\mathbf{x}(m), y(m)\}$ está classificado corretamente. Um dos principais exemplos de aplicação desse algoritmo é o modelo de pontuação de crédito, no qual o objetivo é identificar se uma pessoa pode receber crédito de informações armazenadas em um banco de dados.

No perceptron, o objetivo é formular um algoritmo que ao final das iterações adquira um hiperplano que divide duas classes linearmente separáveis. No algoritmo de aprendizado proposto, o parâmetro do valor inicial $\mathbf{w}(0)$é um vetor nulo ou aleatório. A regra de atualização por iteração é dada por:

$$\mathbf{w} = \begin{cases} \mathbf{w} + \mu y(m)\mathbf{x}(m) & \text{se } \mathbf{x}(m) \text{ é mal classificado por } \mathbf{w}, \\ \mathbf{w} & \text{caso contrário,} \end{cases} \tag{3.4}$$

Onde $\mu > 0$ é o parâmetro de tamanho da etapa, que controla a convergência do algoritmo. Uma boa maneira de entender essa operação é usar sua interpretação geométrica, como mostrado na Figura 3.5. Como $\mathbf{x}(m)$ é mal classificado por$\mathbf{w}$, adicionando o termo $(\mu y(m)\mathbf{x}(m))$, temos uma atualização que move o hiperplano para que o coeficiente atualizado $\mathbf{w}$ classifique $\mathbf{x}(m)$ corretamente.

Após um número finito de iterações, uma solução ótima $\mathbf{w}_o$ é encontrada. Este resultado é válido independentemente de qual critério é escolhido para selecionar os dados em cada iteração e também como o vetor de peso no algoritmo é inicializado. O algoritmo perceptron está descrito na Tabela 3.1.

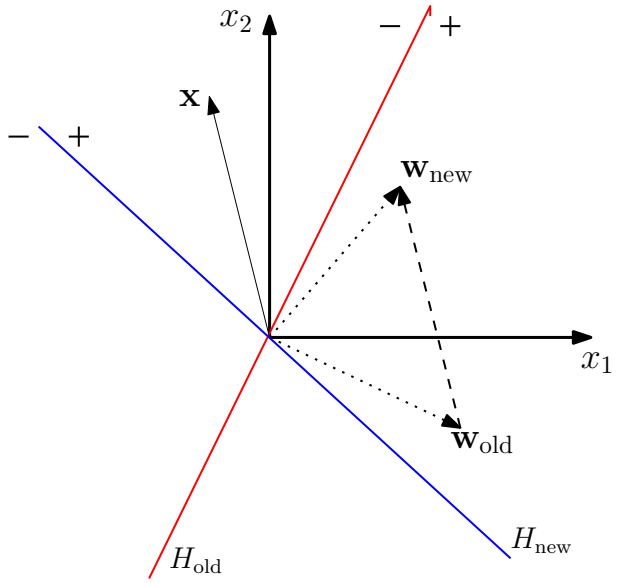

Figura 3.5: O vetor $\mathbf{x}$ é mal classificado pelo hiperplano H_{old} (em vermelho). A regra de atualização adiciona $\mathbf{x}$ ao coeficiente do vetor, $\mathbf{w}_{\text{new}} = \mathbf{w}_{\text{old}} + \mathbf{x}$, onde $\mu = 1$. O novo hiperplano H_{new} (em azul) classifica corretamente o vetor $\mathbf{x}$.

Tabela 3.1: Algoritmo do Perceptron

Algoritmo do Perceptron
Initialize
$\mathbf{w}$ = vetores aleatórios ou nulos,
select $\mu > 0$,
Do for $t > 0$ (época)
count = 0
Do for $m = 1 : M$ (para a iteração m, selecione os dados $\{\mathbf{x}(m), y(m)\}$)
if $y(m)(\mathbf{x}^T(m)\mathbf{w}) < 0$
$\mathbf{w} = \mathbf{w} + \mu y(m)\mathbf{x}(m)$
count = count + 1
end if
end
if count = 0
break; (If count = 0, então todos os dados estão classificados corretamente)
end if
end

3.2.4 Rede Neural Feed-Forward Multi-camadas

Há classificações um pouco mais complexas, como a que é ilustrada na figura 3.6, que não podem ser resolvidas pelo perceptron. Para esse tipo de problema, podemos aplicar o algoritmo conhecido como Multi-layer Perceptron (MLP), que consiste em uma versão do perceptron com mais camadas. As camadas adicionadas entre entrada e saída são chamadas de camadas ocultas. Cada camada é composta de neurônios, também conhecidos como nós nesse tipo de problema. Em cada um dos nós pertencentes às camadas ocultas e de saída, uma função de sinal é calculada, como no algoritmo Perceptron.

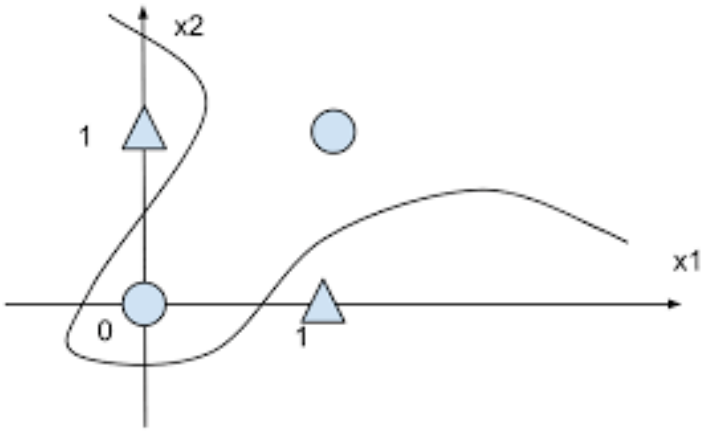

Figura 3.6: Duas classes são representadas, mas não há uma clara separação linear entre elas. A configuração nesta imagem está relacionada à função booleana XOR.

No MLP, um dos obstáculos é a dificuldade de aprender os pesos, já que estamos lidando com um problema combinatório bem mais complexo. Uma solução é a aproximação suave e diferencial para a função de sinal que nos permitirá usar métodos analíticos para encontrar os pesos. Essas são chamadas funções de ativação, que acabam introduzindo uma não linearidade na rede. Alguns exemplos de funções de ativação que podemos adotar aqui são a função ReLU, a função logística sigmoide,a função tangente hiperbólica, e a função soft-max. Durante este trabalho, serão objeto de estudo e utilização as funções de ativação logística e a função soft-max. Esses modelos de redes são conhecidos como feed-forward porque os dados avançam da camada inicial para a última camada. Além disso, quanto maior o número de camadas e o número de nós, mais complexidade é alcançada no modelo.

As camadas são denotadas por $l = 0, 1, 2, \cdots, L$, onde $l = 0$ é a camada de entrada e $l = L$ é a camada de saída. As camadas entre elas, $0 < l < L$, são conhecidas como camadas escondidas. Em todas as camadas, exceto na camada

de saída, o primeiro nó é considerado o nó de polarização e seu valor é definido como 1. Em cada camada, temos d^l sem contar o nó de polarização, rotulado por $1, \cdots, d^l$. A comunicação entre o i-ésimo nó da camada $l-1$ com o j-ésimo nó da próxima camada l é feita através do peso w_{ij}^l.É importante ressaltar que não há outro tipo de comunicação. Em cada nó nas camadas ocultas, exceto nos nós de viés, há uma função de ativação f aplicado no valor de entrada. Na camada de saída, a função calculada é g. A imagem 3.7 ilustra o esquema de uma rede neural

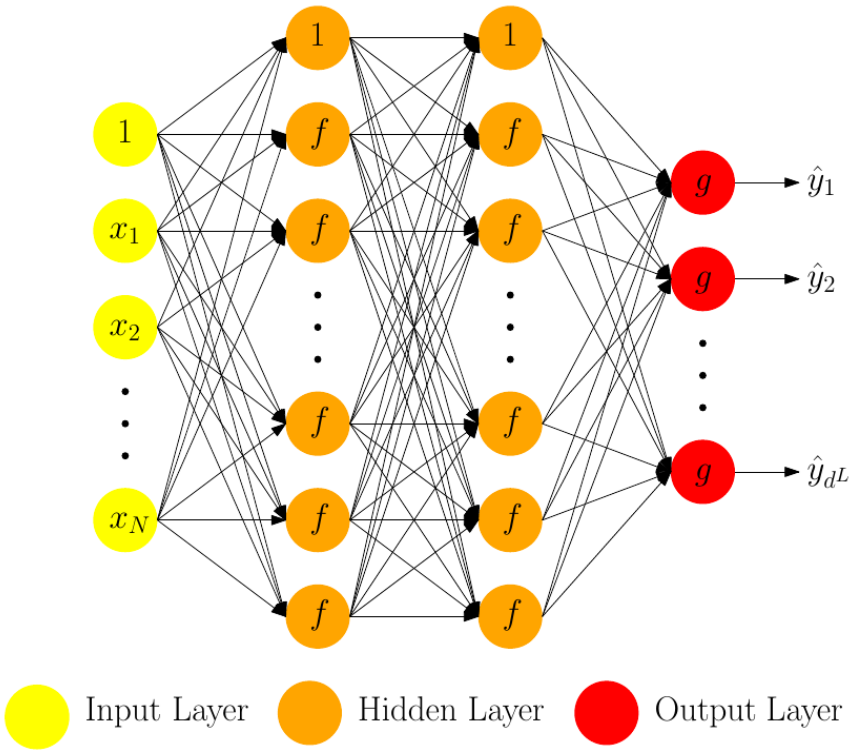

Figura 3.7: Nesta figura, a rede neural tem $L = 4$ camadas (Uma de entrada, 2 escondidas e a de saída.)

Em problemas de regressão, o número de nós na camada de saída é $d^L = 1$, retornando um valor numérico contínuo $\hat{y}$ na saída a ser comparada com o valor verdadeiro y. Enquanto nos problemas de classificação o valor d^L é igual ao número de classes. O sinal desejado $\mathbf{y}$ significa que o valor 1 aparece na entrada correspondente à categoria correta, caso contrário, o valor 0 aparece, e o sinal de saída $\hat{\mathbf{y}}$ retorna uma distribuição de probabilidade que soma 1.

Para simplificar os cálculos, introduzimos a notação para vetor e matriz. Em cada camada l, o vetor de entrada é $\mathbf{e}^l$ e o vetor de saída é $\mathbf{s}^l$. Na passagem da camada $l-1$ para a camada l, temos a matriz de pesos $\mathbf{W}^l$. Essa estrutura é mostrada na tabela 3.2.

Tabela 3.2: Configuração da rede neural

Vetor de entrada na camada l	$\mathbf{e}^l$	d^l-dimensional (vetor)
Vetor de saída na camadar l	$\mathbf{s}^l$	(d^l+1)-dimensional (vetor)
Matriz de pesos entre $l-1$ e l	$\mathbf{W}^l$	$(d^l+1)\times(d^{l+1})$-dimensional (matriz)

Como todas as definições e notações são apresentadas, podemos apresentar o processo de feed-forward que calcula uma função não linear paramétrica, $h_{\mathbf{W}}(\mathbf{x}) = \hat{\mathbf{y}}$, onde $\mathbf{W} = \{\mathbf{W}^1, \mathbf{W}^2, \cdots, \mathbf{W}^L\}$. Todo o processo é baseado na repetição de duas etapas em cada camada oculta, a soma das saídas ponderadas da camada anterior com os pesos e uma função aplicada na entrada de uma camada l para obter o vetor de saída:

$$\mathbf{e}^l = (\mathbf{W}^l)^T \mathbf{s}^{l-1}, \qquad \mathbf{s}^l = \begin{bmatrix} 1 \\ f(\mathbf{e}^l) \end{bmatrix}, \text{ para } 0 < l < L-1, \qquad \mathbf{s}^L = \Big[g(\mathbf{e}^L) \Big] \tag{3.5}$$

Onde a expressão $f(\mathbf{e}^l)$ é um vetor cujas componentes são $f(e_j^l)$ com $e_j^l = \sum_{i=0}^{d^{l-1}} w_{ij}^l s_i^{l-1}$, for $j = 1, \cdots, d^l$. Inicializando na camada de entrada, $\mathbf{s}(0) = \mathbf{x}$, o processo de feed-forward é esquematizado na seguinte cadeia:

$$\mathbf{x} = \mathbf{s}^0 \xrightarrow{\mathbf{W}^1} \mathbf{e}^1 \xrightarrow{f} \mathbf{s}^1 \xrightarrow{\mathbf{W}^2} \mathbf{e}^2 \xrightarrow{f} \mathbf{s}^2 \cdots \xrightarrow{\mathbf{W}^L} \mathbf{e}^L \xrightarrow{g} \mathbf{s}^L = \hat{\mathbf{y}}. \tag{3.6}$$

3.2.5 Back-propagation

Segundo HAYKIN (2009) [5], os principais fluxos de sinais que ocorrem na utilização de um MLP são o fluxo para a frente, que chamamos de feed-forward, que ocorre com os sinais de entrada e o fluxo para trás, chamado de *back-propagation*, pois leva os erros obtidos na saída para reajustar as camadas anteriores. Sendo assim, nosso próximo passo após obter o valor de saída na última camada é a propagação dos erros nos retornos, para atualizar os pesos $\mathbf{W}$. No processo, é necessário definir uma função objetivo ou função perda, $J(\mathbf{W})$. Alguns exemplos da função objetivo são os Mínimos quadrados ou erro médio quadrático, Entropia cruzada e Entropia relativa. Além disso, a partir das pesquisas realizadas na rede neural, foi demonstrado que uma boa combinação entre a função de ativação na

última camada e a função objetivo pode alcançar um melhor desempenho.

Na etapa do back-propagation, o objetivo é minimizar a função objetivo com relação ao parâmetro $\mathbf{W}$ assumindo que $\{(\mathbf{x}(1),\mathbf{y}(1)), (\mathbf{x}(2),\mathbf{y}(2)), \cdots, (\mathbf{x}(M),\mathbf{y}(M))\}$ sendo as amostras de treinamento [2], onde M é o número de dados nesse conjunto. O algoritmo escolhido neste capítulo para minimizar a função objetivo é a descida do gradiente, que atualiza os pesos $\mathbf{W}$ usando a direção negativa do gradiente:

$$\mathbf{W}^l(k+1) = \mathbf{W}^l(k) - \mu \frac{\partial \mathbf{J}(\mathbf{W})}{\partial \mathbf{W}^l}\Big|_{\mathbf{W}(k)} \tag{3.7}$$

Onde μ é o tamanho da etapa. A função objetivo total é a soma do ponto de vista sobre as amostras de treinamento.

$$J(\mathbf{W}) = \frac{1}{M}\sum_{m=1}^{M} J_m(\mathbf{W}) = \frac{1}{M}\sum_{m=1}^{M}\sum_{n=1}^{d^L} J_m^n(\mathbf{W}). \tag{3.8}$$

Onde J_m^n é a função objetivo aplicada ao nó de saída n para a amostra $\mathbf{x}(m)$. O processo de back-propagation lida com a dificuldade de calcular todas as derivadas em (3.7) a partir de um processo simples. O método começa calculando a derivada na última camada e obtém os outros valores recursivamente para as camadas anteriores, de modo que os valores obtidos na camada l influencia no resultado da camada $l-1$, de modo que o processo é chamado de retropropagação, em inglês, back-propagation. Na derivada parcial, a regra da cadeia é aplicada para que o valor obtido seja particionado em duas novas expressões:

$$\frac{\partial \mathbf{J}_m}{\partial \mathbf{W}^l} = \frac{\partial \mathbf{e}^l}{\partial \mathbf{W}^l}\frac{\partial \mathbf{J}_m}{\partial \mathbf{e}^l} = \mathbf{s}^{l-1}(\boldsymbol{\delta}^l)^T \tag{3.9}$$

Onde o primeiro termo é calculado usando a equação (3.5) e o segundo termo é alcançado a partir do processo de back-propagation. Essa expressão $\boldsymbol{\delta}^l$ é conhecida como vetor de sensibilidade para a camada l, que em teoria significa o gradiente (sensibilidade) do custo J_m em relação à entrada $\mathbf{e}^l$.

[2] Em alguns casos, essas amostras de treinamento são divididas em um conjuntos de validação com o objetivo de ajustar os hiperparâmetros de uma tarefa de aprendizagem.

O back-propagation é um processo semelhante ao feed-forward, mas com algumas diferenças ao calcular o valor $\boldsymbol{\delta}^l$. O vetor de sensibilidade $\boldsymbol{\delta}^{l+1}$ é multiplicado pelos pesos $\mathbf{W}^{l+1}$ e o componente de polarização é excluído. A seguir, a transformação aplicada é o elemento de multiplicação por $\boldsymbol{\epsilon}^l = [\mathbf{W}^{l+1}\boldsymbol{\delta}^{l+1}]_1^{d^l}$ e a derivada da função de ativação f em $\mathbf{e}^l$:

$$\boldsymbol{\delta}^l = f'(\mathbf{e}^l) \otimes \boldsymbol{\epsilon}^l \tag{3.10}$$

A cadeia abaixo mostra como o procedimento funciona:

$$\boldsymbol{\delta}^L \xrightarrow{\times \mathbf{W}^L\big|_1^{d^{L-1}}} \boldsymbol{\epsilon}^{L-1} \xrightarrow{\otimes f'(\mathbf{e}^{L-1})} \boldsymbol{\delta}^{L-1} \xrightarrow{\times \mathbf{W}^{L-1}\big|_1^{d^{L-2}}} \boldsymbol{\epsilon}^{L-2} \xrightarrow{\otimes f'(\mathbf{e}^{L-2})} \cdots \xrightarrow{\times \mathbf{W}^2\big|_1^{d^1}} \boldsymbol{\epsilon}^1 \xrightarrow{\otimes f'(\mathbf{e}^1)} \boldsymbol{\delta}^1. \tag{3.11}$$

Onde $\big|_1^{d^l}$ contém apenas os componentes $1, 2, \cdots, d^l$ do vetor $\mathbf{W}^{l+1}\boldsymbol{\delta}^{l+1}$. Essa fórmula para obter o vetor de sensibilidade pode ser explicada de maneira intuitiva:

- $\boldsymbol{\delta}^{l-1} \propto f'(\mathbf{e}^{l-1})$: inclinação maior significa uma pequena alteração em $\mathbf{e}^{l-1}$;
- $\boldsymbol{\delta}^{l-1} \propto \mathbf{W}^l$: e os pesos aumentarem, isso implica uma pequena alteração em $\mathbf{e}^{l-1}$;
- $\boldsymbol{\delta}^{l-1} \propto \boldsymbol{\delta}^l$: se o custo J_m for mais sensível à camada l, será também seja mais sensível à camada $l-1$.

Como a derivação do algoritmo está completa, ilustramos o algoritmo de retropropagação por gradiente descendente em 3.3.

Comentários relacionados ao algoritmo:

- Em cada época, é possível calcular a função de objetivo estabelecida no algoritmo para o conjunto de dados de treinamento J_{train} e o conjunto de dados de validação J_{val}, se esse conjunto estiver definido anteriormente.
- A escolha conjunta da função de ativação da saída e da função objetivo é significativamente importante para as tarefas de aprendizado da RN. Além de simplificar os cálculos na atualização de pesos, também podemos melhorar o desempenho do algoritmo. A combinação mais usada em problemas de regressão é a função linear com erro médio quadrático; para problemas de classificação, é função softmax com entropia cruzada.

- Até há algum tempo, a função de ativação mais utilizada era a função sigmóide, mas apresentava algumas desvantagens, como saturação nas caudas e centralização em um valor diferente de zero. Atualmente, as funções mais empregadas são a tangente hiperbólica (tanh) e a "unidade linear retificada " (ReLU).

- O procedimento padrão para inicializar os pesos $\mathbf{W}$ é selecionando aleatoriamente os valores. Se iniciada com números muito grandes ou muito pequenos, a função de ativação ficará saturada, resultando em gradientes baixos e, em seguida, uma convergência mais lenta. O procedimento atual para lidar com esse tipo de problema é inicializar os pesos com uma distribuição uniforme ou normal.

- Um critério de parada é treinar o modelo até uma certa quantia da época E, e, após esse período, verificar se o valor da função objetivo é menor que um limite pré-selecionado. Uma alternativa adicional é verificar se a função objetivo está diminuindo de iteração para iteração. Caso contrário, o processo pode ser interrompido.

- No algoritmo para adaptar uma RNA, para cada iteração *iter* é considerada apenas uma amostra de dados do tamanho b (mini-batch). Dependendo dessa escolha no conjunto de dados, aleatória ou determinística, isso pode resultar em uma convergência mais rápida.

- Ao selecionar o número de camadas ocultas em uma rede de treinamento, o critério usado para escolher esse valor é adicionar camadas até que o erro de validação não mostre melhora. Enquanto para o número de neurônios nas camadas ocultas, a regra é sempre colocar mais do que o necessário, ou seja, um número alto de neurônios por camada.

Tabela 3.3: Algoritmo Gradient descent back-propagation

Algoritmo Gradient descent back-propagation

Initialize

$\{(\mathbf{x}(1),\mathbf{y}(1)), (\mathbf{x}(2),\mathbf{y}(2)), \cdots, (\mathbf{x}(M),\mathbf{y}(M))\}$,

$\mathbf{W} = \{\mathbf{W}^1, \mathbf{W}^2, \cdots, \mathbf{W}^L\}$ (vetores aleatórios),

Select: step size $\mu > 0$, número da épica E, batch size b, número de camadas L,

número de nós $(d^1, \cdots, d^L)$, função de ativação f, função de saída g,

função objetivo J

$iter = M/b$

Do for $t = 1 : E$

Do for $i = 1 : iter$ (para cada iteração, selecione aleatoriamente b examplos

no conjunto de dados de treinamento $\rightarrow \mathbf{X}_i^t = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_b]$, $\mathbf{Y}_i^t = [\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_b]$)

[Forward Propagation]:

$\mathbf{S}^0 = [\mathbf{s}_1^0, \cdots, \mathbf{s}_b^0] = [ones(1, b); \mathbf{X}_i^t]$ ($ones(1, b)$ são o termo de viés)

Do for $l = 1 : L - 1$

$\mathbf{E}^l = (\mathbf{W}^l)^T \mathbf{S}^{l-1}$

$\mathbf{S}^l = [ones(1, b); f(\mathbf{E}^l)]$

end

$\mathbf{E}^L = (\mathbf{W}^L)^T \mathbf{S}^{L-1}$

$\hat{\mathbf{Y}} = [\hat{\mathbf{y}}_1, \cdots, \hat{\mathbf{y}}_b] = \mathbf{S}^L = g(\mathbf{E}^L)$

[Back-propagation]:

$\boldsymbol{\Delta}^L = [\boldsymbol{\delta}_1^L, \cdots, \boldsymbol{\delta}_b^L] = \frac{\partial J}{\partial \mathbf{E}^L}$

Do for $l = L - 1 : -1 : 1$

$\boldsymbol{\Delta}^l = f'(\mathbf{E}^l) \otimes [\mathbf{W}^{l+1} \boldsymbol{\Delta}^{l+1}]_1^{d^l}$

end

[Atualizando os pesos]:

Do for $l = 1 : L$

$\mathbf{W}^l = \mathbf{W}^l - \frac{\mu}{b} \mathbf{S}^{l-1} (\boldsymbol{\Delta}^l)^T$

end

end

$J_{\text{train}}(\mathbf{W}) = \frac{1}{M} \sum_{m=1}^{M} \sum_{n=1}^{d^L} J_m^n(\mathbf{W})$ (and J_{val} se este conjunto estiver definido anteriormente)

end

3.2.6 Função Softmax

Em matemática, a função softmax, também conhecida como softargmax ou função exponencial normalizada, é uma função que recebe como entrada um vetor de K números reais e a normaliza em uma distribuição de probabilidade composta por K proporcionalmente. para os exponenciais dos números de entrada. Ou seja, antes da aplicação do softmax, alguns componentes do vetor podem ser negativos ou maiores que um; e pode não somar 1; mas depois de aplicar o softmax, cada componente estará no intervalo (0,1) e os componentes adicionarão até 1, para que possam ser interpretados como probabilidades. Além disso, os componentes de entrada maiores corresponderão a probabilidades maiores. A função Softmax é freqüentemente usada em redes neurais, para mapear a saída não normalizada de uma rede para uma distribuição de probabilidade sobre as classes de saída previstas.

A formulação matemática padrão da função Sofmax pode ser dada da seguinte forma:

$$\sigma(\mathbf{z})_{\mathrm{i}} = \frac{\mathbf{e}^{z_{\mathrm{i}}}}{\sum_{j=1}^{K} \mathbf{e}^{z_{\mathrm{j}}}} \tag{3.12}$$

para $i = 1, 2, \cdots, K$ e $z = (z_1, \cdots, z_{\mathrm{n}}) \in \mathbb{R}_{\mathrm{K}}$

3.3 Métricas de Avaliação

Nosso problema será focado em avaliar a capacidade de uma rede neural prever a classe a que um evento está associado. Sendo assim, para esse tipo de problema, as principais métricas de avaliação estão listadas abaixo:

- **Verdadeiro Positivo (VP):** indica a quantidade de registros que foram classificados como positivos corretamente, ou seja, a resposta do classificador foi que o comentário era positivo e o comentário realmente era positivo.
- **Verdadeiro Negativo (VN):** indica a quantidade de registros que foram classificados como negativos de maneira correta, ou seja, a resposta do classificador foi que o comentário era negativo e o comentário realmente era negativo.
- **Falso Positivo (FP):** indica a quantidade de registros que foram classificados como comentários positivos de maneira incorreta, ou seja, a resposta do classificador foi que o comentário era positivo, mas o comentário era negativo.
- **Falso Negativo (FN):** indica a quantidade de registros que foram classificados como comentários negativos de maneira incorreta, ou seja, a resposta do classificador foi que o comentário era negativo, mas o comentário era positivo.
- **Acurácia:** A acurácia é o indicador mais simples de se calcular. Ela é simplesmente a divisão entre todos os acertos pelo total de registros. Esse indicador pode acabar enganando um avaliador. Para essa pontuação significar algo, de fato, é necessário que a base de dados que foi utilizada para avaliação tenha uma boa variedade de resultados. Usaremos um exemplo hipotético. Imagine que criamos um classificador para responder se um determinado exame contém ou não uma doença. Nesse cenário, nossa base de dados será composta por 90% de registros em que a doença não ocorre e apenas 10% de registros onde ocorre. Imagine agora que nosso modelo sempre responde que não há doenças. Qual seria a acurácia desse modelo? Como nossa base é composta por 90% de registros onde de fato não há doença, nosso modelo teria uma acurácia de 90%, mesmo ele sendo completamente descartável.
- **Precisão:** A Precisão ou *precision* em inglês é utilizada para indicar a relação entre as previsões positivas realizadas corretamente e todas as previsões positivas (incluindo as falsas). A métrica de precisão nos dá informação sobre falsos positivos, então trata-se de identificar um determinado resultado de maneira precisa.

  Imagine que nossa base de dados contenha 1000 registros, onde apenas 100 deles são positivos. Caso o modelo responda positivo apenas para um destes

casos, a precisão ainda estaria 100%. Isso porque os falsos negativos não são considerados nessa métrica. A principal utilização dessa métrica é para modelos onde é preciso minimizar os falsos positivos. Neste caso, quanto mais perto dos 100% chegarmos, melhor. É dado por:

$$Precisao = \frac{VP}{TP + FP} \tag{3.13}$$

- **Recall:** A métrica *Recall* é utilizada para indicar a relação entre as previsões positivas realizadas corretamente e todas as previsões que realmente são positivas (Verdadeiros Positivos e Falsos Negativos). A Recall nos dá informações sobre falsos negativos. Essa métrica por si só não é diretamente vinculada a classificar todos os casos corretamente. Ela indica o quanto nosso modelo está identificando os casos positivos corretamente. A Recall é bastante útil quando precisamos minimizar os falsos negativos. Isso é especialmente útil para casos de diagnósticos, onde pode haver um dano muito maior em não identificar uma doença, do que identificá-la em pacientes saudáveis. É dada por:

$$Recall = \frac{TP}{TP + FN} \tag{3.14}$$

- **Score F1 (F1):** Talvez a métrica F1 Score seja a menos intuitiva em seu entendimento. De forma bastante simples, ela é uma maneira de visualizarmos as métricas precisão e Recall juntas. Uma maneira de unir as duas métricas seria simplesmente calcular a média aritmética. O problema disso é que existem casos onde a precisão ou a Recall podem ser muito baixas enquanto a outra permanece alta. Isso indicaria problemas na geração de falsos positivos ou negativos, conforme já vimos nos tópicos anteriores. Para ajustar isso, o cálculo é um pouco diferente, mas ainda acaba sendo uma média entre as duas métricas anteriores. A média que iremos calcular é a média harmônica, quando os dois valores do cálculo são iguais. Essa média gera resultados muito próximos da média "comum". No entanto, sempre que os valores são diferentes, essa média se aproxima mais dos valores menores. É dado por:

$$F1 = \frac{2 * precisao * recall}{precisao + recall} \tag{3.15}$$

3.4 Notas históricas

Segundo HAYKIN (2009) [5], o estudo sobre redes neurais artificiais começou com o trabalho de MCCULLOCH E PITTS (1943) [14]. McCulloch era um psiquiatra e neuroanatomista, que há muito tempo estudava sobre a representação do sistema nervoso, e Pitts um matemático. Juntos publicaram um artigo denominado "A Logical Calculus of Ideas Immanent in Nervous Activity", no qual eles apresentavam cálculo lógico das redes neurais que reunia os estudos de neurofisiologia e da parte matemática. Eles assumiram que seu modelo de neurônio seguia uma lei "tudo ou nada", que com números suficientes de neurônios e com conexões sinápticas ajustadas de forma adequada, a rede neural seria capaz de realizar qualquer função computável.

Atualmente, os estudos realizados buscam por métodos de treinamento das redes artificiais, porém na época, o foco era em encontrar um modelo representativo de um neurônio e descrever suas capacidades computacionais. Em 1949, Donald Hebb publicou "The Organization of Behavior", onde foi apresentada pela primeira vez uma formulação explícita de uma regra de aprendizagem fisiológica para a modificação sináptica. Hebb propôs que a conectividade do cérebro é continuamente modificada conforme um organismo vai aprendendo tarefas funcionais diferentes e que agrupamentos neurais são criados por tais modificações. Ele assume que a aprendizagem do conhecimento é alcançada por meio de conexões entre neurônios adjacentes sempre que os mesmos estiverem excitados.

Outro avanço histórico foi o trabalho descrito em ROSENBLATT (1958) [15], que apresenta seu novo modelo *Perceptron*, um modelo inovador de aprendizagem supervisionada, que era composto de unidades sensoriais conectadas a uma única camada de neurônios de McCulloch e Pitts, que seria capaz de aprender tudo o que pudesse representar. Rosenblatt mostrou que, acrescentando sinapses ajustadas corretamente, as redes neurais de McCulloch e Pitts teriam a possibilidade de serem treinadas para classificar padrões, um tipo de tarefa que os seres humanos fazem sem nenhum esforço aparente e de forma quase instantânea. Porém, é um dos problemas mais difíceis de serem resolvidos por uma máquina.

O Perceptron simples descrito por Rosenblatt possui três camadas: a primeira recebe as entradas do exterior e possui conexões fixas (retina); a segunda recebe impulsos da primeira através de conexões cuja eficiência de transmissão (peso) é ajustável e, por sua vez, envia saída para a terceira camada (resposta). Nos anos 60, Windrow e Hoff desenvolveram o ADALINE (Adaptative linear Element) e o

MADALINE (Many Adaline) para o reconhecimento de padrões.

Em 1969, Minsky e Papert chamaram a atenção para algumas tarefas que o Perceptron não era capaz de executar, já que este só resolve problemas lineares separáveis, ou seja, problemas cuja solução pode ser obtida dividindo-se o espaço de entrada em duas regiões através de uma reta. O Perceptron, por exemplo, não consegue detectar paridade, conectividade e simetria, que são problemas não lineares separáveis, como o citado na figura 3.6. O principal argumento de Minsky e Papert era de que o problema do crescimento explosivo, tanto de espaço ocupado como do tempo requerido para solução de problemas complexos afetaria, cedo ou tarde, as RNAs, inclusive os Perceptrons, como descrevem BRAGA, CARVALHO e LUDERMIR (2000) [4]. O impacto desta publicação foi devastador, praticamente desaparecendo o interesse em redes neurais artificiais nos anos seguintes.

O interesse por redes neurais artificiais retornou com mais força apenas na década de 80, com o trabalhos como os de Hopfield, Kohonen e Grossberg, por exemplo. Por meio dos estudos dos pesquisadores supracitados foi possível esclarecer boa parte das dúvidas existentes em relação ao processo executado por certas redes neurais. As redes de Hopfield, por exemplo, atraíram muita atenção em estudos no campo da física, trazendo diversos pesquisadores novos para a área.

Em 1986, Rumelhart, Hinton e Williams aperfeiçoaram a utilização do Perceptron, criando o algoritmo de retropropagação, Back-propagation, levando a uma grande onda de interesse em redes neurais. O sucesso desse algoritmo estimulou o desenvolvimento de muitas pesquisas em redes neurais artificiais e de uma variedade de novos modelos baseados no cérebro humano.

Capítulo 4

Treinamento do Modelo

4.1 Definição do Problema

Como abordado em cápitulos anteriores, a era digital permite que as empresas saibam muito sobre os seus clientes, mas também faz com que as pessoas tenham acesso a uma quantidade de informação inimaginável sobre qualquer assunto.

Quando falamos do processo de compra seja de um produto ou de um serviço, um possível cliente pode pesquisar fatores como marcas, histórico e flutuação de preços, qualidade do fornecedor. Apesar dessa gama enorme de informações, segundo dados de pesquisa feita pelo Serviço de Proteção ao Crédito (SPC), 6 em cada 10 consumidores agem por emoção na hora de realizar uma compra. No entanto, como visto anteriormente, dentro do cenário do comércio eletrônico, para cada 100 visitas feitas aos sites de compras, apenas 1 venda é registrada.

Podemos então entender que, uma boa forma de influenciar a emoção dos compradores na hora da tomada de decisão, é conhece-lo de forma ampla. Dentro do espectro do que chamamos de Marketing orientado a dados isso pode ser feito de algumas formas: oferecendo o produto que mais combina com o cliente, descobrindo seus interesses, oferencendo uma experiência personalizada.

Nosso objetivo é identificar, no momento em que um cliente visita um produto em um site, com base nas intereações que ele teve com com essa loja no passado, qual a probabilidade do cliente realizar a compra. Esse cálculo é importante pois temos algumas interpretações a serem feitas: Se o cliente realizou a compra, temos sucesso em nosso objetivo. Mas, quando o cliente não faz a compra, temos o caso interessante de analisar. Se soubermos que um cliente com alta probabilidade de compra por algum motivo não finalizou o pedido do produto, podemos utilizar téc-

nicas de comunicação direcionada para esses clientes, que podem incluir descontos, promoções ou outros incentivos.

Logo, o objetivo final desse trabalho é utilizar as técnicas de redes neurais para maximizar as vendas dentro do comércio eletônico, através do cálculo da probabilidade de compra.

4.2 Coleta dos Dados

Os dados foram coletados diretamente das bases de um site de comércio eletrônico famoso no Brasil. O dado básico consiste em coletar informações sobre usuários que visitaram algum produto. Logo, assim que um cliente chega até a página de um determido item, coletamos infomações sobre o tipo de dispositivo que o cliente esá usando, qual produto ele está vendo, de que categoria é o produto, de que lugar do Brasil é o cliente, o valor do frete para entrega do produto, e também o prazo para a entrega. Por fim, registramos a informação que nos diz se ao final daquela visita, o cliente realizou ou não uma compra. É importante citar que todas essas informações estão disponíveis em tempo de navegação, isto é, no momento em que a visita está acontecendo é possível coletar as informações e utilizá-las.

A partir desses dados, é possível enriquecer as informações do cliente com os seguintes itens: há quanto tempo esse cliente é cadastrado, quantas visitas ele já fez no site, quantas compras esse cliente já fez, quantas vezes ele já visitou esse mesmo produto, qual o histórico de preço desse produto e quanto ele está maior ou menor em relação ao preço atual.

Para a coleta das informações sobre a visita atual do cliente, foram utilizados dados coletados entre Julho e Agosto de 2018. No entanto para enriquecer os dados com as informações do histórico de utlização do cliente, consideramos dados desde o início de 2015, que era o período disponível para consulta no momento em que os dados foram coletados. Ao final da coleta, obtivemos 49.954 registros em nossa base.

4.2.1 Dicionário de Dados

Para que possamos começar a explorar as variáveis, é importante introduzir um dicionário que contenha os nomes das variáveis e o que cada uma delas significa:

- PedidoFeito: Variável binária. Quando a visita se tornou uma compra assume valor 1, caso contrário, assume valor 0.

- Device: Representa a forma pela qual o cliente estava usando a loja. Pode assumir 3 valores: App (aplicativo da loja para celular), Desktop (site utilizado através de um computador) ou Mobile Site (site utilizado através de um celular).
- ProductViews: Indica quantas vezes o usuário já visitou o produto em questão. Caso seja a primeira visita, a variável assumirá o valor 1.
- Region: Indica a região do Brasil de onde o usuário é. Pode ser Norte, Nordeste, Centro-Oeste, Sudeste e Sul, uma vez que o site atende em todas as regiões do país.
- Loja: O site em questão possui o modelo de negócio de Marketplaces, isto é, lojistas maiores que vendem os produtos de outros lojistas menores. Essa variável indica de qual lojista é o produto que o cliente está visualizando. Essa variável foi agrupada em 6 grupos: Loja 1, Loja 2, Loja 3, Loja 4, Loja 5 e Outras Lojas. Os nomes estão mascarados por motivos de confidencialidade.
- PrecoProduto: Indica o preço em Reais atribuído ao produto no momento da visita.
- MenorPreco60dias: Indica o menor preço em Reais atribuído ao produto no período de 60 dias antes da visita.
- DistanciaMenorPreco: Diferença absoluta entre PrecoProduto e MenorPreco60dias
- Distancia%MenorPreco: Diferença percentual entre PrecoProduto e MenorPreco60dias
- Frete: Valor de frete de entrega do produto para o CEP informado pelo usuário
- PercentualFretePrecoTotal: É dado pelo seguinte cálculo: $Frete/(Frete + PrecoProduto)$
- daysToDeliver: Tempo em dias estimado para que a entrega do Produto seja feita no CEP do cliente.
- Categoria do Produto: Indica a qual categoria de produtos o Item visitado pertence. Pode assumir os seguintes valores: Celular, Eletrodomésticos, Eletrônicos, Computador e Acessórios,Acessórios de Video Game,Video Game, outros.
- Data de Cadastro: Indica a data em que o cliente utilizou o site pela primeira vez.

- DataVisita: Indica a data da atual visita que o cliente está fazendo.
- Visitas: Indica a quantidade de visitas que o cliente já fez ao site até aquele momento. Caso seja a primeira visita, assumirá o valor 1.
- Produtos Vistos: Aponta a quantidade de diferentes produtos que o cliente já visitou.
- Compras Feitas: Indica quantas compras o cliente já fez com essa empresa. Caso o cliente não tenha feito nenhuma compra, assume o valor 0.
- CategoriasCompradas: Conta em quantas categorias de produto diferentes o cliente já fez uma compra.
- PrecoMedioPedidos: Média de preço de todos os pedidos feitos pelo cliente. Caso ele não tenha nenhum pedido feito, assume valor 0.
- PrecoMaximoPedidos: Máximo preço de todos os pedidos feitos pelo cliente. Caso ele não tenha nenhum pedido feito, assume valor 0.
- UltimoPedido: Data do último pedido feito pelo cliente. Caso ele não tenha nenhum pedido feito, assume valor 0.

É importante notar que as variáveis device, Region, Loja e Categoria do Produto são consideradas variáveis categóricas. Elas serão alvo de transformações ao longo do processo de tratamento dos dados.

4.3 Tratamento dos Dados

A base original de visitas continha 49.953 registros de visitas, com 22 variáveis disponíveis para estudo. No entanto, como dito anteriormente, 4 dessas estão em formato categórico. Para utilizarmos essas informações em nossa modelagem precisaremos transformar essas variáveis para o formato de indicadora. Sendo assim, ao total, teremos a substituição dessas 4 por novas 21 colunas em nossa tabela. Cada uma delas assumindo o valor 1 ou 0 de acordo com a ocorrência de cada um dos valores categóricos supracitados. Um exemplo do procedimento aplicado pode ser visto na Figura 4.1.

Id de Registro	Categoria
1	A
2	A
3	B
4	B
5	C
6	C

Id de Registro	Categoria A	Categoria B	Categoria C
1	1	0	0
2	1	0	0
3	0	1	0
4	0	1	0
5	0	0	1
6	0	0	1

Figura 4.1: Exemplo do processo para transformar uma variável categórica em variáveis numéricas indicadoras

Outro ponto muito importante é a análise de possíveis *outliers* existentes em nossos registros. Para analisar a distruibuição das variáveis do conjunto de dados, utilizamos o conhecido método de estatística descritiva chamado *box-plot*[1]. Nesse trabalho, seguiremos a forma de utilização do box-plot para deteção de outliers proposta por TUKEY (1977) [28], onde um valor é considerado outlier caso seja maior que $1.5 \times (Q_3 - Q_1)$, sendo Q_1 o 1^{o} quartil da distribuição da variável e Q_3 o 3^{o} quartil da distribuição da variável.

Ao aplicar esse critério em todas as variáveis de nosso catálogo, três delas chamaram a atenção: PrecoProduto, Frete e daysToDeliver. A primeira apresenta valores de preços acima de $R\$50.000$. A segunda chega a apresentar valores de frete na casa de milhões de reais, o que provavelmente foi gerado por algum erro sistêmico. A terceira apresenta tempos de entrega superiores a 100 dias, o que se mostra inviável comercialmente e provavelmente está ligado a algum erro de sistema também. Essas constatações podem ser observadas nas figuras 4.2, 4.3, 4.4.

Após a identificação dos outliers e a posterior remoção desses registros, nossa base possui 44.978 registros. Além disso, seguindo o padrão de boas práticas para treinamento de redes neurais, realizaremos a padronização dos dados, isto é, normalizaremos coluna por coluna para obtermos dados com distribuição próxima a uma normal padrão, isto é, as variáveis foram transformadas para apresentar uma média zero e um desvio padrão unitário, antes de serem utilizadas no treinamento do modelo.

[1] O box-plot é um método estatístico para análise gráfica do comportamento, variabilidade, propagação e assimetria de uma amostra numérica de dados.

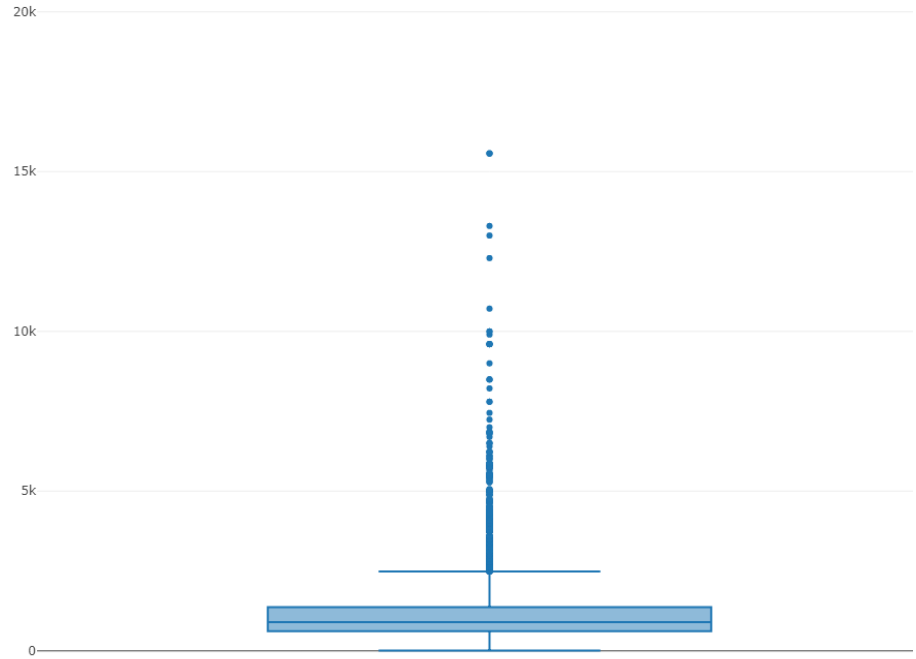

Figura 4.2: A análise do boxplot da variável precoProduto nos permitiu identificar que preços acima $R$$2.489,00 podem ser considerados outliers dentro de nossa amostra

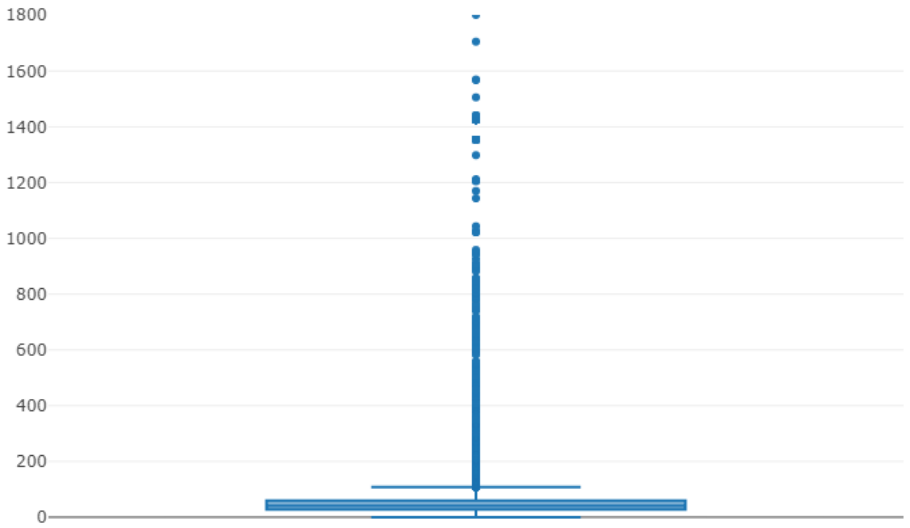

Figura 4.3: A análise do boxplot da variável Frete nos permitiu identificar que valores acima de $R$$107,97 podem ser considerados outliers dentro de nossa amostra

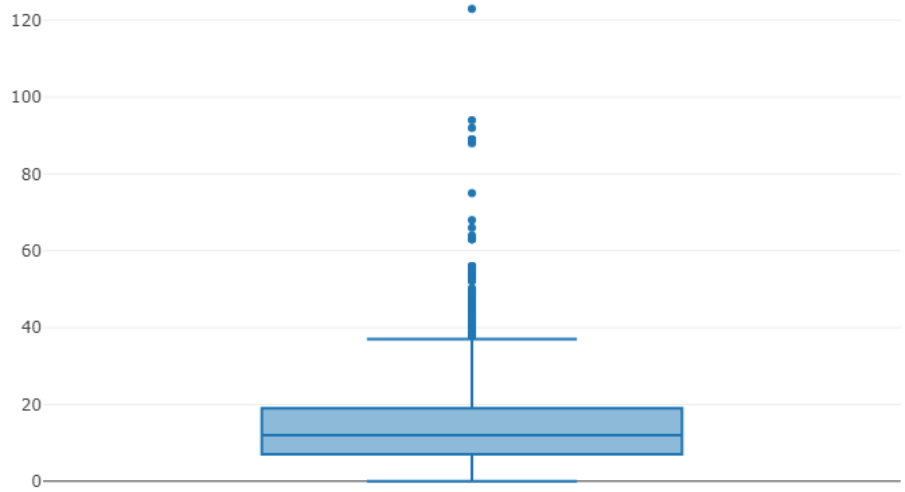

Figura 4.4: A análise do boxplot da variável daysToDeliver nos permitiu identificar que prazos de entrega acima de 37 dias podem ser considerados outliers dentro de nossa amostra

4.4 Balanceamento dos Dados

Como foi dito anteriormente, no comércio eletrônico brasileiro, temos em média 1 compra para cada 100 visitas em algum site de vendas. Logo, se nossa base de dados consiste exatamente nesse tipo de registros, isto é, visitas a sites de venda, e a nossa variável alvo é justamente o evento de compra, é de se esperar que nossa base esteja bastante desbalanceada no evento de compra ou não-compra. E de fato está!

Ao analisarmos a base, chegamos a conclusão de que dos 44.978 registros que apenas 1.894 registros são relativos a compras efetuadas. Ou seja, apenas 4.21% dos registros válidos representam eventos de compra, o que nos dá um problema com alto desbalanceamento entre as classes. Isso nos leva a questão de como fazer o balanceamento dessa base para que o modelo não fique excessivamente treinado na classe dominante.

4.4.1 Técnicas para Balancear Dados

As formas mais populares para balancear dados na utilização em técnicas de aprendizado de máquina são conhecidas, segundo SHELKE, DESHMUKH E SHANDILYA (2017) [17], por under-sampling, ou sub-amostragem e over-sampling, ou super-amostragem e serão explicadas abaixo:

- Sub-amostragem: Consiste em separar a classe com maior cardinalidade, e escolher uma amostra, preferencialmente aleatória, de tamanho similar a cardinalidade da classe de menor cardinalidade. Isso faz com que as classes se tornem balanceadas artificialmente.

- Super-amostragem: Consiste em tomar a classe de menor cardinalidade e acrescentar a ela cópias dela mesma até que a cardinalidade da classe se torne similar a cardinalidade da classe mais representativa.

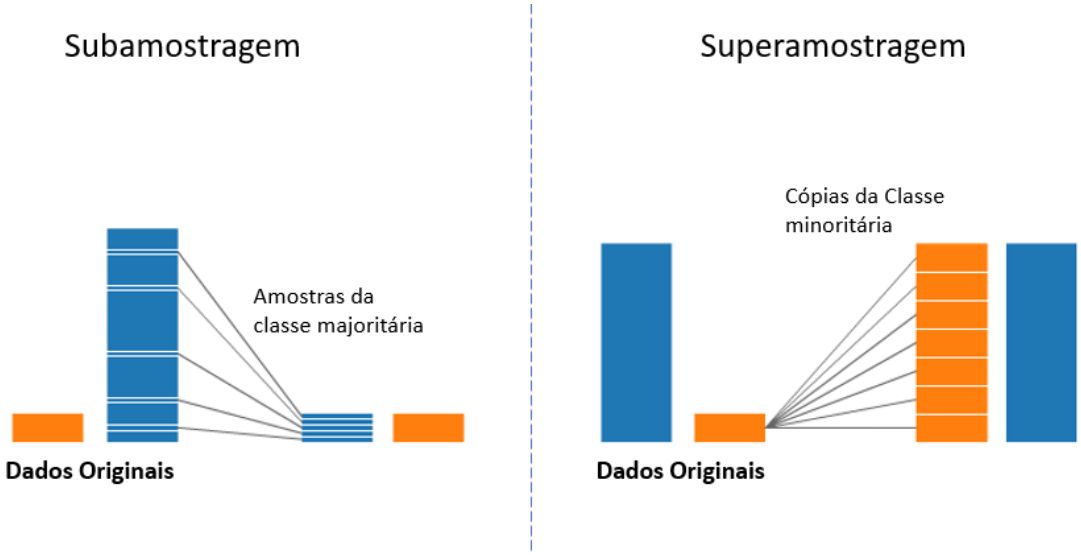

Figura 4.5: Técnicas de balanceamento de dados

Ainda segundo SHELKE, DESHMUKH E SHANDILYA (2017) [17], os métodos aleatórios de sub e sobre amostragem têm suas várias deficiências. O método aleatório de subamostragem pode remover potencialmente alguns exemplos importantes de nossa base de dados, impedindo que nosso modelo se torne tão generalista quanto poderia. Já amostragem aleatória excessiva podem levar ao excesso de ajustes do modelo, o que pode acarretar no temido overfitting.

Para nossa base, como a classe de compra é extremamente menor que a classe de não compra, decidimos utilizar a metodologia de Over-sampling pois, ao fazermos isso, impedimos que uma grande quantidade de informações sobre o processo de não compra seja ignorado pelo algoritmo de aprendizado do modelo.

Sendo assim, após a aplicação do método escolhido, adicionamos 23 cópias dos dados da classe compra, obtendo assim, um dataset com 43.562 registros de compra e 43.084 registros de não compra, fazendo com que a base de dados final possua 86.646 registros e classes de interesse bem balanceadas.

4.5 Seleção de variáveis

Para entendermos se todas as variáveis do nosso conjunto de dados são de fato relevantes para resolver o problema proposto, é preciso fazer algumas análises. A mais comum entre eles é a análise de correlações amostrais. Essa análise consiste em construir uma matriz cujas entradas representam a correlação, medida estatística amplamente conhecida, entre cada uma das colunas do nosso conjunto de dados.

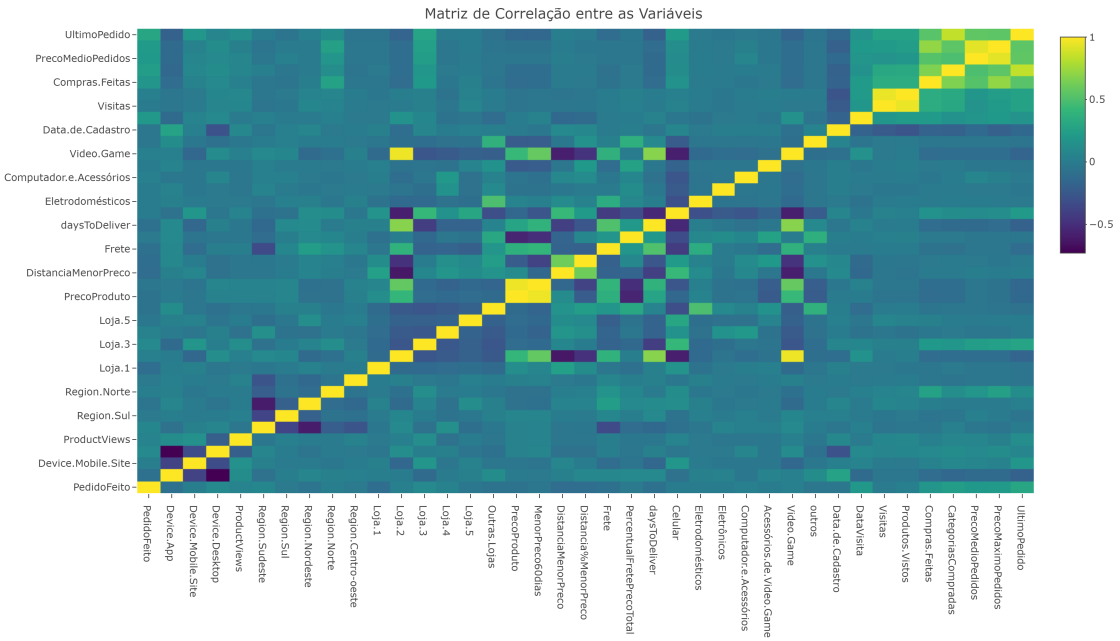

Figura 4.6: Matriz de correlação entre as variáveis. Valores de correlação estão variando entre -1 e 1. Correlações negativas e próximas a -1 estarão em azul com tom mais escuro, enquanto que correlações positivas e próximas a 1 estarão em tons próximos a verde e amarelo. Valores próximos a 0 estarão em tons de azul mais claro.

A análise da matriz apresentada na figura 4.6 nos permite tirar algumas conclusões:

- Os pares de variáveis PrecoProduto/MenorPreco60dias, Visitas/Produtos Vistos e PrecoMedioPedidos/PrecoMaximoPedidos tem correlação entre si próxima de 1. Com isso, segundo TRIANTAPHYLLOU E FELICI (2006) [29], é prudente que façamos a remoção de uma das variáveis de cada par. Com isso, as variáveis MenorPreco60dias, Produtos Vistos e PrecoMaximoPedidos não farão parte do processo de modelagem.

- As variáveis Celular, Eletrodomésticos, Eletrônicos e Acessórios de Video Game, todas relacionadas a variável categórica Categoria do Produto, possuem correlação 0 com a variável resposta. Com isso, segundo SUDHAMATHY E

VENKATESWARAN (2018) [30], também podem ser desconsideradas da modelagem.

- As demais variáveis do conjunto têm, em sua maioria, correlação baixa com a variável de saída. No entanto, serão consideradas na tentativa de construção do modelo. Sendo assim, recapitulando as manipulações de variáveis, começamos com 22 e após a transformações de categóricas em numéricas chegamos em 39. No entanto, com a remoção de 7 variáveis na etapa de tratamento dos dados, temos 32 variáveis, sendo uma associada a ocorrência do evento de compra e 31 variáveis que podem ser usadas como entradas.

4.6 Treinamento da Rede Neural

Para este trabalho, treinar uma rede neural envolve otimizar seus parâmetros para minimizar as perdas. A perda a ser minimizada é a perda média, isto é, o erro de previsão da saída da rede quando comparada com o que é visto no conjunto de treinamento. Os parâmetros são otimizados com métodos baseados em gradientes das funções de perda. Mais especificamente, os otimizadores iterativamente atualizam os parâmetros para avançar no espaço dos parâmetros em direção aos mínimos locais. Eles fazem isso a partir de uma configuração de peso inicial. A inicialização do peso é feita por atribuindo pequenos valores aleatórios a cada peso.

Em particular, durante a execução desta etapa do trabalho, estamos usando as seguintes premissas:

- A Rede utilizada foi do tipo Feed-Forward totalmente conectada com uma camada de entrada, uma camada escondida e uma camada de saída. O número de nós da camada intermediária será definido nos passos adiante.
- Foram utilizadas funções de ativação do tipo sigmoide (ou logística).
- Visando o cálculo de probabilidades associadas ao evento de compra, usaremos a função softmax na camada de saída.
- O método de aprendizado utilizado é o back-propagation de gradiente descendente com uma função perda quadrática.

Uma vez que nossos dados estão plenamente preparados para o modelo, após as etapas de tratamento, balanceamento e seleção de variáveis, podemos prosseguir para o treinamento da Rede, após a qual devemos encontrar o modelo que satisfaça nossa necessidade de falar sobre a probabilidade de compra dos clientes de um site.

4.6.1 Validação Cruzada

Segundo HAYKIN (2009) [5], A essência do aprendizado de uma rede neural é a codificação de um mapeamento entre entradas e saídas nos pesos e limiares sinápticos de um modelo. A esperança é que a rede se torne bem treinada para aprender o suficiente sobre o passado de tal forma que seja possível generalizar sua utilização em dados do futuro. Nessa perspectiva, o processo de aprendizagem equivale a parametrização de uma rede para um determinado conjunto de dados fixado. O que não é tão claro, é a forma sob a qual podemos garantir que essa generalização futura esperada possa acontecer.

Nesse contexto, uma ferramenta padrão em estatística, conhecida como validação cruzada, fornece um princípio orientador bastante razoável. Em sua essência, a metodologia propõe que o conjunto de dados disponível seja aleatoriamente particionado em uma amostra de treinamento e um conjunto de testes. A motivação aqui é validar o modelo em um conjunto de dados diferente daquele usado para estimativa dos parâmetros da rede.

Dentre os diversos métodos de se aplicar a validação cruzada, podemos destacar alguns deles:

- **Deixe-p-Fora** Esse método, como descreve ALPAYDIN (2004) [34], envolve o uso de p observações como o conjunto de validação e as observações restantes como o conjunto de treinamento. Isso é repetido em todas as maneiras possíveis de cortar a amostra original em um conjunto de validação de p observações e um conjunto de treinamento. Com $p > 1$ e com um número de amostras razoavelmente grande, esse processo fará com que a Rede precise ser treinada por um número alto de vezes, o que pode tornar o método extremamente custoso do ponto de vista computacional.

- **Método Holdout** Nesse caso, segundo OLSON E DELEN (2008) [33], atribuímos aleatoriamente pontos dos dados a dois conjuntos D_0 e D_1, geralmente chamados de conjunto de treinamento e de teste, respectivamente. O tamanho de cada um dos conjuntos é arbitrário, embora normalmente o conjunto de testes seja menor que o conjunto de treinamento. Em seguida, treinamos o modelo em D_0 e testamos (avaliamos seu desempenho) em D_1. O método de validação envolve uma única execução. Deve ser usado com cautela, pois sem a média de várias execuções, é possível obter resultados enganosos.

- **Validação em k-dobras** Segundo BISHOP (2006) [3], a amostra de dados original e completa é dividida aleatoriamente em k subamostras de tamanho

igual. Dessas, uma única subamostra é retida como conjunto de dados de validação para testar o modelo, e as $k-1$ subamostras restantes são usadas como dados de treinamento. O processo é repetido k vezes, com cada uma das k subamostras usadas exatamente uma vez como dados de validação. Os k resultados obtidos podem então ser calculados como média para produzir uma única estimativa. A vantagem deste método é que todas as observações são usadas para treinamento e validação, e cada observação é usada para validação exatamente uma vez. A validação cruzada de 10 vezes é comumente usada, mas, em geral, k segue sendo um parâmetro livre para ser utilizado da forma que se queira.

Neste trabalho, a validação cruzada utilizada para avaliar o modelo será a validação em k-dobras com $k = 10$. Ou seja, dividiremos nosso conjunto de dados tratado em 10 partes iguais, e realizaremos 10 rodadas diferentes de treinamento, onde cada uma das 10 partes será por 1 única vez o conjunto de teste e por 9 vezes fará parte do conjunto de treinamento.

4.6.2 Dimensionamento da Rede

Segundo BISHOP (2006) [3], O número de nós na entrada e na saída de uma rede neural é geralmente determinado pela dimensionalidade do conjunto de dados, enquanto o número M de nós nas camadas ocultas é um parâmetro que pode ser ajustado para oferecer o melhor desempenho preditivo. Observe que M é responsável por controlar o número de parâmetros existentes na rede e, portanto, podemos esperar que, em uma configuração de máxima probabilidade, haja um valor ideal de M que oferece o melhor desempenho de generalização, correspondendo ao melhor equilíbrio entre a sub-adaptação (under-fitting) e super-adaptação (over-fitting).

Sendo assim, temos como premissa que a camada de entrada de nossa rede terá 31 nós, e que a camada de saída terá 2 nós, pois isso reflete a quantidade de variáveis que restaram após a limpeza dos dados e o número de classes associadas a variável resposta. Um detalhe importante é que, apesar de termos 2 categorias como saída, ao utilizarmos a ativação softmax, a rede terá como respostas a probabilidade de pertencimento a cada uma das classes. Como são apenas 2, as respostas são complementares, fazendo com que somente um valor de saída seja necessário.

Nosso desafio, no entanto, é tentar descobrir qual o número de nós ideal para a camada oculta do modelo. Seguindo a linha citada acima, podemos fazer testes de forma a tentar encontrar o número ótimo. Uma forma de realizar esse teste é através do treinamento da rede variando apenas o número de Neurônios na camada oculta. Isso nos permite comparar o valor final da função de perda após a convergência do processo de treinamento para as diversas arquiteturas diferentes. O resultado dessa comparação, quando aplicadas sessões de treinamento da rede em todo o conjunto de dados, pode ser visto na figura 4.7:

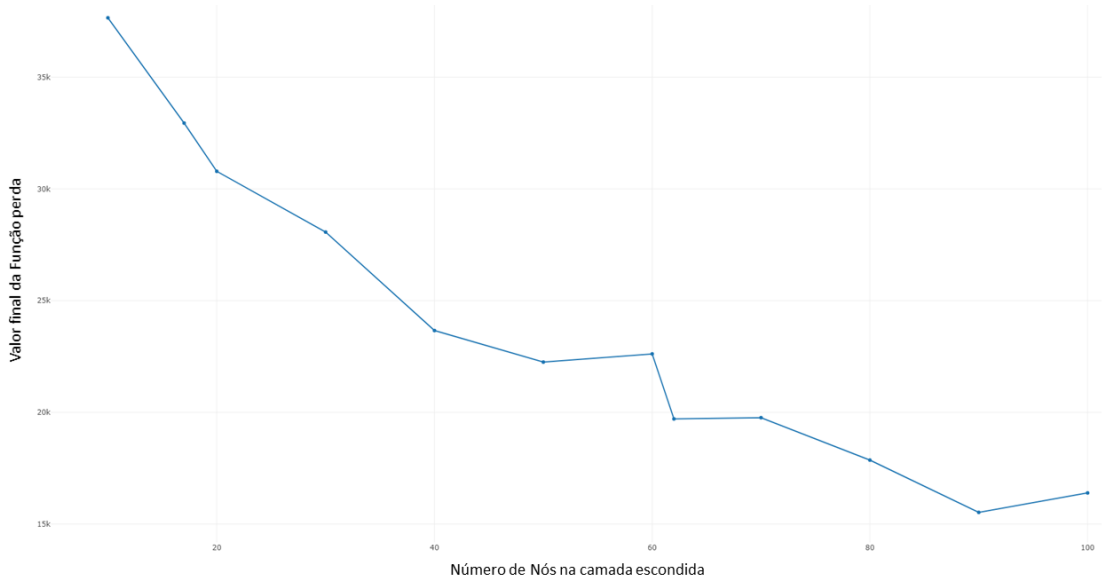

Figura 4.7: Desempenho da Rede em relação ao número de nós na camada oculta

Há uma queda relevante no valor dos critérios de treinamento da rede ao passar o número de nós da camada oculta para 62. Soma-se a isso a teoria proposta por HECHT-NIELSEN (1990) [32] e reforçada por REMESAN E MATHEW (2015) [31] sobre a utilização de duas vezes mais neurônios na camada intermediária do que na camada de entrada, e concluimos por utilizar o número de 62 neurônios na camada escondida. Logo, a arquitetura da rede se dará da forma descrita na figura 4.8.

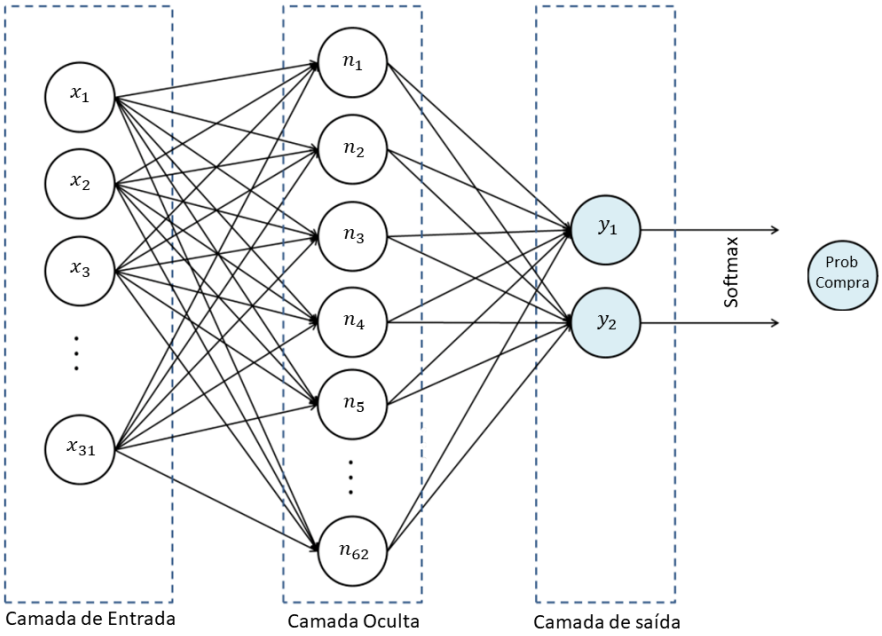

Figura 4.8: Aqui vemos a arquitetura final esperada da rede construída.

4.6.3 Definição do limite decisório

O limite decisório, mais frequentemente referido por sua nomenclatura em inglês *threshold*, é um dos parâmetros mais importantes quando se trata de utilizar probabilidades para fazer previsões a respeito de um problema classificatório binário. A saída da rede neural que treinaremos nos dará a probabilidade de uma visita se tornar ou não uma compra. Definir um valor para o limite decisório, significa que para um valor acima desse limite indica pertencimento a classe correspondente a saída; um valor abaixo indica pertencimento a classe restante. É tentador supor que o limite de classificação sempre deve ser 0.5, mas os limites dependem do problema e dos dados e, portanto, são valores que devemos ajustar.

O ponto principal é que para avaliarmos se o modelo está oferecendo boas respostas, precisamos transformar as probalidades que obtivimos na saída, através da função softmax, em classes. Só assim podemos comparar os valores obtidos na rede com os valores reais do conjunto de dados de treinamento.

Esse problema também foi tratado de forma heurística. A abordagem foi fazer o treinamento da rede, e após isso, calcular os valores de das métricas Acurácia, F1, Precisão e Recall, sendo as 3 últimas para as duas classes, coletando esses indicadores para todos os limites entre 0 e 1, utilizando duas casas decimais. É importante frisar que o limite será aplicado para a probabilidade de pertencer a classe de compra. Ou seja, se tomarmos um limite ξ, temos que um registro i será classificado como compra se $ProbCompra_i > \xi$. Caso contrário, será classificado

como um registro de Não-Compra.

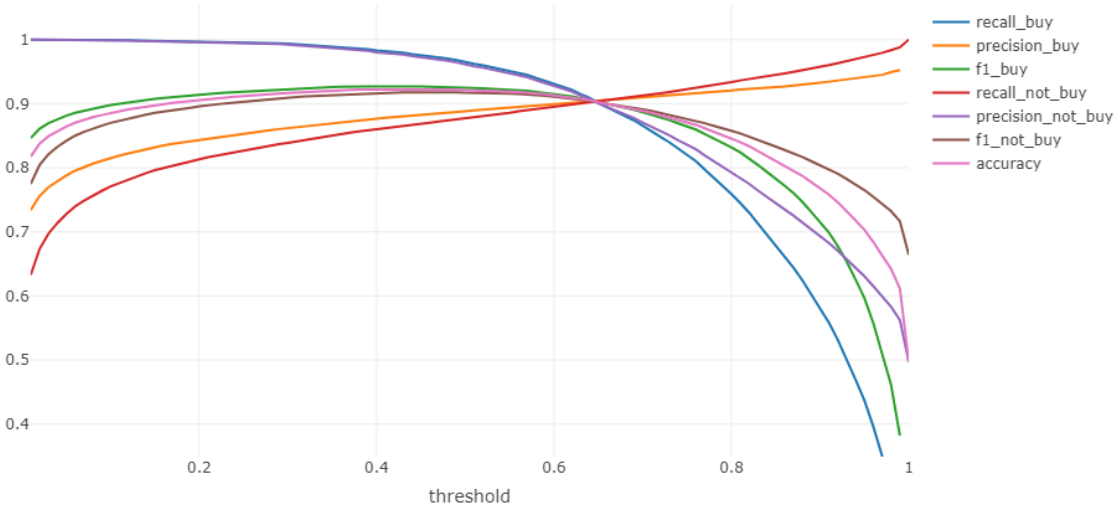

Figura 4.9: Avaliação da Rede com relação as métricas de desempenho variando o limite decisório

Como se pode ver no gráfico exibido na figura 4.9, há uma variação enorme de cenários quando variamos o limite decisório. No entanto, chama a atenção o fato de que para o valor de 0.65 absolutamente todas as métricas estão avalidas acima de 0.90, o que significa que dessa forma, temos um bom indicativo de que o modelo proposto irá apresentar bons resultados para todas as classes envolvidas. Sendo assim, o valor escolhido para ser o limite decisório do modelo, no que diz respeito a avaliação de desempenho será 0.65.

Capítulo 5

Análise de Resultados

Dedicaremos este capítulo para tratar da análise dos resultados obtidos após o treinamento do modelo, parte a qual cobrimos com o capítulo anterior. Nosso objetivo será avaliar nosso modelo e seu potencial enquanto um classificador, primeiramente. Uma vez que seja comprovado que o modelo se comporta como um bom classificador para o problema proposto, seguiremos com a discussão do uso das probabilidades de compra, saída primária do modelo, para realizar inferências sobre o processo de compra de um cliente.

5.1 Indicadores de Performance

Como dito anteriormente, para analisarmos a qualidade das probalidades calculadas pelo modelo, utilizaremos as métricas de Precisão, Recall, Acurácia e F1, todas introduzidas no capítulo 3. Além disso, consideramos o limite de 0,65 para definição das classes, e lançaremos mão do uso de validação cruzada em 10 dobras.

Para cada dobra de validação, um modelo foi treinado, probabilidades foram calculadas, classes foram definidas, e as métricas acima citadas puderam ser calculadas. Os resultados obtidos podem ser vistos na tabela 5.1.

Dobra	**Recall Compra**	**Precisão Compra**	**F1 Compra**	**Recall Não Compra**	**Precisão Não Compra**	**F1 Não Compra**	**Acurácia**
1	0,908	0,921	0,915	0,921	0,909	0,915	0,915
2	0,924	0,896	0,910	0,892	0,920	0,906	0,908
3	0,902	0,908	0,905	0,907	0,902	0,904	0,905
4	0,916	0,905	0,910	0,903	0,914	0,908	0,909
5	0,879	0,898	0,889	0,900	0,881	0,890	0,889
6	0,901	0,909	0,905	0,909	0,901	0,905	0,905
7	0,896	0,902	0,899	0,902	0,896	0,899	0,899
8	0,900	0,902	0,901	0,901	0,899	0,900	0,900
9	0,891	0,912	0,902	0,913	0,892	0,903	0,902
10	0,912	0,909	0,911	0,908	0,911	0,909	0,910
Média	0,903	0,906	0,905	0,905	0,902	0,904	0,904

Tabela 5.1: Resultados obtidos após a Validação Cruzada

Pela definição do limite decisório feita no capítulo anterior, era esperado que obtivéssemos as métricas de avaliação acima de 0,90, o que, de fato, se confirma após todo o processo de treinamento. Obter esse resultado nos dá algumas seguranças:

- Valores de Recall altos indicam que a probabilidade de termos falsos negativos na previsão do modelo é baixa, ou seja, se um cliente fez uma compra, a probabilidade de nosso modelo afirmar que a compra não vai ser feita é baixa.
- Valores de Precisão altos indicam probabilidade de termos falsos positivos na previsão do modelo é baixa, ou seja, quando classificamos uma visita como sendo uma compra, a probabilidade de ela não ser uma compra é baixa.

- A ocorrência das observações acima para ambas as classes em questão faz com que nosso modelo se mostre ainda mais confiável. Isso se reflete no valor $F1$, que também se mostrou alto para ambas as classes.

As observações acima se tornam ainda mais claras quando analisamos a Matriz de confusão do modelo, juntando todas as previsões feitas em cada uma das 10 etapas de validação cruzada:

		Valores Reais		
		Compra	Não-Compra	Total
Valores Previstos	Compra	39.336 (90,3%)	4.072 (9,4%)	43.408
	Não-Compra	4.226 (9,7%)	39.012 (90,6%)	43.238
	Total	43.562	43.084	86.646

Tabela 5.2: Matriz de Confusão do Modelo Obtido

5.2 Avaliando qualitativamente as probabilidades geradas

Na seção anterior, conseguimos consolidar a certeza de que o modelo construído é útil se quisermos um classificador que responda se uma nova visita no site será ou não uma compra. No entanto, o objetivo principal desse trabalho não é utilizar o modelo como um classificador e sim utilizar sua prerrogativa de devolver como saída as probabilidades de pertencimento a cada uma das classes de interesse.

Agora, passaremos a ter como objetivo analisar outros fatores que nos permitam ter a segurança de as probabilidades geradas fazem sentido com o problema que temos, como por exemplo, a dispersão que as probabilidades de compra possuem para registros de cada uma das classes.

É possível perceber, pela análise das figuras 5.1 e 5.2 que as probabilidades de compra estão concentradas próximas a 1 se o registro for um registro de compra e próximas a 0 se o registro for de não compra. Isso significa que o modelo proposto faz total sentido no que diz respeito ao cálculo das probabilidades de compra.

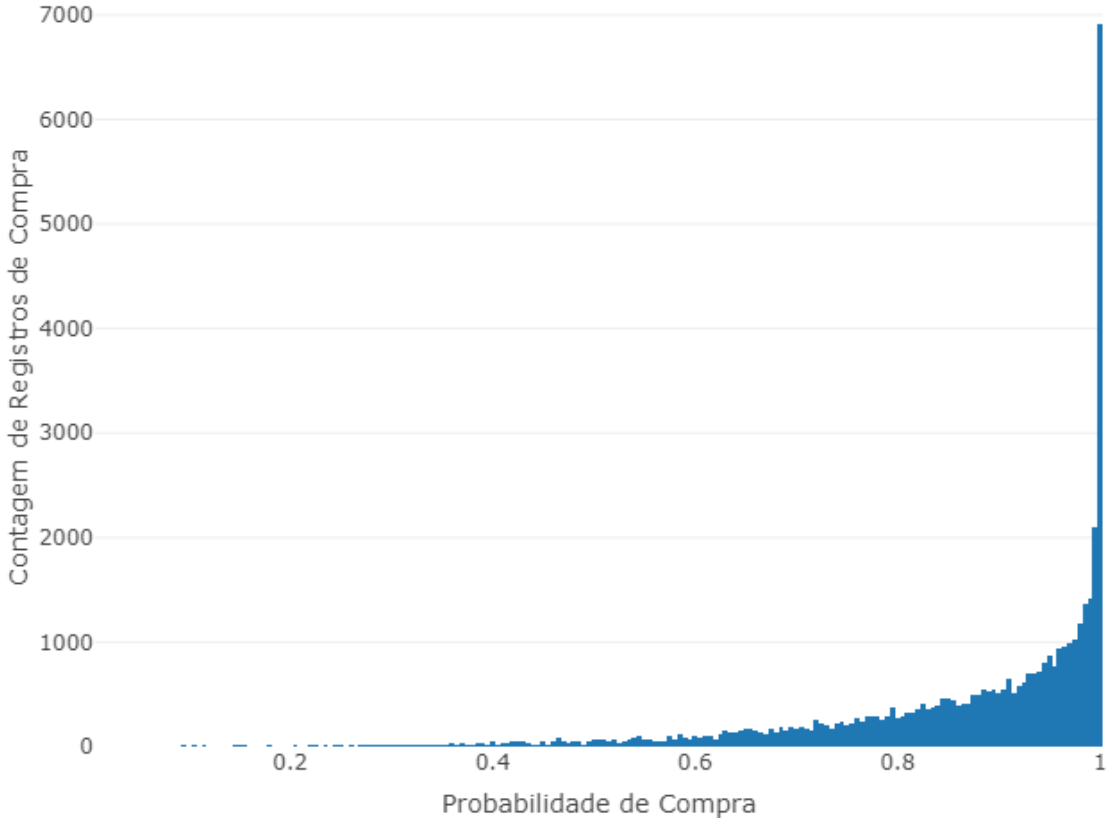

Figura 5.1: Histograma das Probabilidades de Compra calculadas para todos os registros onde houve compra

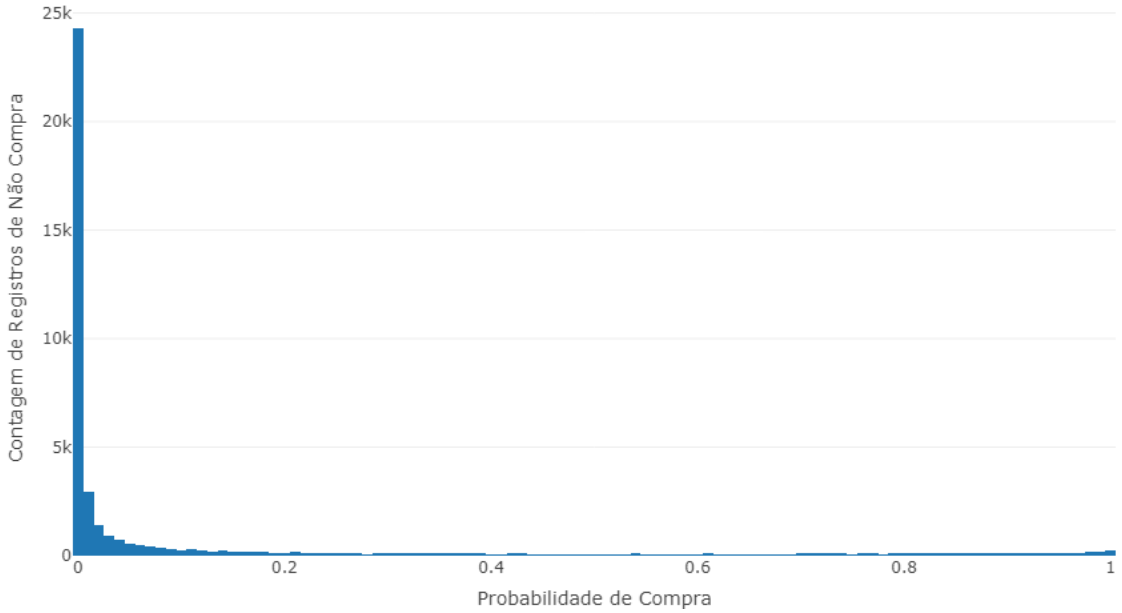

Figura 5.2: Histograma das Probabilidades de Compra calculadas para todos os registros onde não houve compra

Capítulo 6

Conclusão

6.1 Como é possível aplicar os resultados Obtidos?

Segundo CHURCHILL E PETER (2010) [1], A estratégia de preços influencia o comportamento de compra quando o consumidor está avaliando alternativas e chegando a uma decisão. Muitas vezes, os consumidores dão preferência a um produto mais barato: eles podem pensar em comprar um determinado bem porque ele está em liquidação ou porque eles têm um cupom de desconto para aquela marca. Para consumidores que tomam decisões rotineiras ou limitadas, o preço será especialmente importante caso seja um dos atributos do produto que entra em avaliação.

Nessa situação, os profissionais de marketing terão de cobrar menos, reduzir outros custos de compra ou convencer os consumidores a tomar decisões com base em outros atributos, ou munir os clientes com descontos que alterem a percepção deles sobre o preço do produto. E nesse conceito, muitas empresas acabam se perdendo, concedendo descontos sem que haja qualquer tipo de inteligência por trás.

Na figura 6.1, podemos ter um exemplo sobre uma das formas como o comércio eletrônico negocia preços com seus clientes hoje em dia. Ao visitar um produto no site da empresa, que teve suas logomarcas removidas por questões de confidencialidade, caso o cliente esteja próximo da área onde se pode clicar em fechar a página, a janela mostrada se abre, dando a oportunidade do cliente testar qualquer tipo de preço possível, até que algum deles seja aceito.

Essa ação, além de permitir que o cliente sempre encontre o menor preço aceito para um produto, pode gerar frustração em alguns clientes que não consigam os preços que gostariam. Além disso, todo e qualquer cliente que chegue a página pode estimular sua navegação para que o simulador de ofertas seja aberto. Em médio

Figura 6.1: Simulador de ofertas utilizado por um grande varejista do mercado brasileiro

prazo, a empresa não venderia mais seus produtos sem que o cliente utilizasse o artifiício de encontrar o menor preço possível.

Outro exemplo de política de descontos que vicia o cliente é feita por um grande aplicativo de entrega de comida:

Figura 6.2: Cupom de desconto para não pagar a taxa de entrega

Essa empresa tem uma política de cobrar taxas de entrega quando se pede alguma refeição através dela. No entanto, caso um cliente comece a fazer um pedido, mas não queira pagar a taxa de entrega, basta deixar seu pedido incompleto por alguns minutos, e o cupom exemplificado na figura 6.2 é enviado ao cliente.

Este trabalho tem como sua principal contribuição, a demonstração de que é possível identificar, utilizando os conceitos de redes neurais, a probabilidade de

compra de um dado cliente. Esse tipo de informação poderia auxiliar nos exemplos supracitados, identificando cliente que devem ou que não devem receber algum tipo de incentivo para realizar sua compra.

É preciso relembrar que nossa motivação final é fazer com que o número de vendas de um site aumente, utilizando o conceito de Marketing Orientado a dados. Sendo assim, o que fizemos até aqui foi utilizar uma rede neural com função de ativação softmax para nos devolver a probabilidade de compra de um usuário ao visitar determinado produto. Vimos também que essa rede se comporta bem como um classificador para as classes dadas no conjunto de dados e que as probabilidades geradas pelo modelo estão concentradas próximas as suas classes de pertencimento real.

Para o problema de vendas no e-commerce, o ponto central é: o que acontece com os Falsos Positivos, isto é, usuários que tiveram probabilidades de compra em nível razoável ou até mesmo alto, mas que por algum motivo, não fizeram uma compra? A resposta é que reconhecer esses usuários é o maior ganho que o modelo pode fornecer.

Quando uma empresa oferece descontos de forma indiscriminada, ela está se prejudicando em diversas formas. Além de estar sacrificando sua margem de lucro e sua geração de caixa de forma direta, ela também pode estar atingindo os alvos errados, isto é, clientes que já fariam a compra de qualquer forma, o que faz com o que a empresa acabe se tornando refém de fazer campanhas de descontos.

A metodologia que foi proposta nesse trabalho permite que o departamento de marketing de uma empresa de comércio eletrônico identifique quais clientes tinham uma boa probalidade de compra mas que, por algum motivo, deixaram de realizar o pedido. Esses clientes podem ser trabalhados de diversas formas, seja com campanhas de desconto, seja com ofertas de produtos associados, ou até mesmo com lembretes quando o preço do produto de interesse possa subir. Note que, devido a robustez do processamento da rede neural e a alta disponibilidade dos dados envolvidos no processo, é possível inclusive utilizar essa modelagem de forma *online*, influenciando a experiência de compra no momento em que ela está acontecendo, possibilitando a utilização dessa probabilidade como gatilho para variações de preço, ou condições especiais como os cupons ou, mais recentemente, as políticas de cashback[1]

[1]Política em que uma parte do valor pago em uma compra é devolvido ao cliente após certo tempo, normalmente cerca de 30 dias após a data da compra

É interessante dizer que, como a modelagem estará sendo feita de forma inperceptível ao usuário, diminui-se muito o risco de criar vícios de utilização nos clientes, o que protege a margem de lucro da empresa, e, no fim das contas, o mais importante, que é poder de falar com o cliente certo na hora certa, já estará sendo possibilitado pelo modelo desenvolvido.

6.2 Trabalhos Futuros

Dentro da modelagem proposta, como pode ser visto nas seções anteriores, pudemos criar um rede neural com bons resultados e poderia servir como um bom caminho para calcular probabilidades de compra e ser utilizada como ferramenta de conhecimento sobre o cliente, abrindo possibilidades para a sua utilização em campanhas de marketing orientadas a dados. No entanto, há diversas possibilidades a serem exploradas pensando em possíveis melhorias na modelagem proposta, dentre as quais podemos citar:

- Arquitetura de rede: Optamos por utilizar uma rede *feed-forward* com somente uma camada escondida. É possível que a inclusão de mais camadas intermediárias possa gerar resultados diferentes, melhorando as métricas de sucesso. Nesse caso, também devemos ficar atentos ao tempo de treinamento, que na arquitetura original não foi um problema mas que pode se tornar objeto de preocupação em redes mais profundas.

- Variáveis removidas: Durante o processo de limpeza dos dados, 7 variáveis foram deixadas de fora da modelagem. Podemos nos questionar se alguma delas traria ainda mais poder de generalização ao nosso modelo. Sendo assim, um possível passo é introduzir essas variáveis uma a uma no processo de treinamento e teste a fim de garantir que foi uma boa decisão deixa-las de fora ou se alguma delas deveria ser incluída novamente.

- Função de ativação: nesse trabalho, utilizamos a função sigmoide como função de ativação da rede nas camadas anteriores a saída. No entanto, existe uma grande variedade de funções que podem cumprir esse mesmo papel. Testar outros tipos de ativação, então, parece ser um caminho que vale a pena explorar.

6.3 Outras possibilidades de utilização da metodologia proposta

O chargeback ou estorno é o processo de cancelamento de uma venda realizada por meio de cartão (crédito ou débito), solicitada pelo seu titular quando ele não reconhece a compra ou ela está em dissonância com o que foi combinado/ prometido. Essa reclamação é feita ao banco ou à administradora do cartão, que impede a creditação ao vendedor ou o estorno do valor se ele já foi creditado. Geralmente esse é um processo rápido, sem aviso prévio e quase sempre decidido pró-consumidor. Por isso, é tão impactante para o varejo — principalmente quando falamos de lojas virtuais.

O chargeback sempre acontece a partir de uma reclamação do cliente ou de um conflito de informações no momento da consolidação do pagamento. Existem alguns casos bastante comuns que acontecem especialmente no comércio eletrônico:

- Erro no valor cobrado: Esse é o chargeback menos impactante para um negócio e de responsabilidade total do varejista. É um erro de digitação em negócios físicos ou de sistema em lojas virtuais e acontece, na maioria das vezes, quando um zero é adicionado por equívoco ao fim do valor — assim, R$100,00 se tornam R$1000,00. Se a empresa tem um bom controle financeiro com uma ferramenta de gestão, esse erro é facilmente identificado e a própria loja pode corrigi-lo.

- Não recebimento da mercadoria: Esse é outro chargeback comum em comércio eletrônico, que pode causar prejuízo, mas é facilmente controlado. Nesse caso, o cliente pede a devolução do dinheiro porque o produto não chegou ou não é compatível com o item comprado.

- Fraude deliberada: Agora entramos nos exemplos de chargebacks mais complicados. A fraude deliberada é realizada por meio de dados roubados. Um criminoso consegue acesso às informações da vítima e realiza a compra sem o seu consentimento.

- Auto-Fraude: Na auto-fraude, o próprio titular do cartão abusa do chargeback. O cliente faz o pedido, recebe a mercadoria e, propositalmente, aciona a administradora dizendo não reconhecer a compra. O valor então é estornado e a loja fica sem o dinheiro e sem o produto.

- Fraude amigável: A fraude amigável também é realizada pelo titular, mas sem má fé envolvida. Acontece principalmente quando ele não reconhece o nome

da empresa na fatura ou a compra é feita por alguma pessoa conhecida sem o seu consentimento.

De todos os itens listados, os que envolvem fraudes são os mais prejudiciais às empresas pois, se levados à cabo, fazem com que a empresa fique sem o dinheiro e sem o produto.

Um possível trabalho futuro envolve olhar o outro lado da moeda em que nos focamos durante o desenvolvimento desse trabalho. É possível estudar a relação entre os eventos de fraude e as compras realizadas cuja probabilidade de serem feitas segundo o modelo proposto eram baixas? Aqui teremos um paradigma totalmente invertido. Se antes queríamos identificar vendas que eram pra ocorrer mas não aconteceram, aqui, teremos que tentar entender se algumas das vendas quen fizemos sequer deveriam existir.

Referências Bibliográficas

[1] CHURCHILL JR, G.A., PETER, J.P. **Marketing: Criando Valor para os clientes**. 2 ed. São Paulo, Saraiva, 2010.

[2] BISHOP, C. **Neural Networks for Pattern Recognition**. 1 ed. Nova York, Oxford University Press, 1995.

[3] BISHOP, C. **Pattern Recognition and Machine Learning**. 1 ed. Singapura, Springer, 2006.

[4] BRAGA, A., CARVALHO, A., LUDERMIR, T. **Redes neurais artificiais: teoria e aplicações**. 1 ed. Rio de Janeiro, Livros Técnicos e Científicos, 2000.

[5] HAYKIN, S. **Neural Networks and Learning Machines**. 3 ed. Ontario, Prentice Hall, 2009.

[6] TAFNER, M., XEREZ, M., RODRIGUES FILHO, I. **Redes neurais artificiais: Introdução e princípios de neurocomputação**. 1 ed. Blumenau, Ekon, 1996.

[7] CRONJE, G., DU TOIT, G.S. **Introduction to Business Management**. 6 ed. Cape Town, Oxford University Press Southern Africa, 2004.

[8] DONATO, A.M. **Marketing de Serviços**. 1 ed. Curitiba, IESDE Brasil, 2012.

[9] KOTLER, P. **O Marketing sem segredos**. 1 ed. São Paulo, Bookman, 2005.

[10] KOTLER, P., KARTAJAYA, H., SETIAWAN, I. **Marketing 3.0: as forças que estão definindo o novo marketing centrado no ser humano**. 1 ed. Rio de Janeiro, Elsevier, 2010.

[11] GRACIOSO, F. **Marketing – O sucesso em 5 movimentos**. 1 ed. São Paulo, Editora Atlas, 1997.

[12] BOONE, G., KURTZ, G.S. **Marketing Contemporâneo**. 8 ed. São Paulo, Editora LTC, 2011.

[13] MACHADO, C., DAVIM, J.P. **Theory and Application of Business and Management Principles**. 1 ed. Basel, Springer, 2016.

[14] MCCULLOCH, W., PITTS, W. "A Logical Calculus of the Ideas Immanent in Nervous Activity", **The bulletin of mathematical biophysics**,v. 5, n.4, p.115-133, Grã-Bretanha, 1943.

[15] ROSENBLATT, F. "The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain", **Cornell Aeronautical Laboratoryl**, Nova York, 1958.

[16] ROSS, N. "A History of Direct Marketing", **Direct Marketing Association**, Nova York, 1992.

[17] SHELKE, M., DESHMUKH, P., SHANDILYA, V. "A Review on Imbalanced Data Handling Using Undersampling and Oversampling Technique", **International Journal of Recent Trends in Engineering and Research**,v. 3, n.4, p.444-449, India, 2017.

[18] NIA, M. *Big Data Analytics of Consumer Choices: A Two Sided Online Platform Perspective*, Tese de D.Sc., University of Texas at Dallas, Texas, Estados Unidos, 2016.

[19] GUASTI, P., MONTEIRO, K., OLIVEIRA, J., et al. *Webshoppers*, E-bit e Nielsen Brasil, São Paulo, Brasil, 2018.

[20] SIEBEL, T. **Digital Transformation: Survive and Thrive in an Era of Mass Extinction**, 1 ed. Nova York, RosettaBooks, 2019.

[21] CLEVELAND, W. "Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics", **International Statistical Review 69**, n.1, p.21-26, Estados Unidos, 2001.

[22] SIQUEIRA, A.C.B. **Marketing empresarial, industrial e de serviços**, 1 ed. São Paulo, Editora Saraiva, 2005.

[23] COBRA, M. **O Novo Marketing**, 1 ed. Rio de Janeiro, Editora Elsevier, 2010.

[24] CUEN, D. *As 10 chaves do sucesso do Facebook*, BBC News Brasil, 2014. Disponível em: https://www.bbc.com/portuguese/noticias/2014/02/140204_dez_chaves_sucesso_facebook_lgb Acesso em: 05 de Janeiro de 2020.

[25] BOHME, F.G. **100 Years of Data Processing: The Punchcard Century**, 3 ed. U.S. Department of Commerce, Bureau of the Census, Data User Services Division, Estados Unidos, 1991.

[26] SAURA, J.R. "Using Data Sciences in Digital Marketing: Framework, methods, and performance metrics", **Journal of Innovation and Knowledge**,v. 6, Issue 2, p.92-102, Espanha, 2020.

[27] JEFFERY, M. **Data-driven marketing : the 15 metrics everyone in marketing should know**, 1 ed. Estados Unidos, John Wiley & Sons, 2010.

[28] TUKEY, J.W. **Exploratory Data Analysis**, 1 ed. Estados Unidos, Addison-Wesley, 1977.

[29] TRIANTAPHYLLOU, E., FELICI, G. **Data Mining and Knowledge Discovery Approaches Based on Rule Induction Techniques**, 1 ed. Estados Unidos, Springer, 2006.

[30] SUDHAMATHY, G., VENKATESWARAN, C.J. **R Programming: An Approach to Data Analytics**, 1 ed. India, MJP Publishers, 2018.

[31] REMESAN, R., MATHEW, J. **Hydrological Data Driven Modelling**, 1 ed. Reino Unido, Springer, 2015.

[32] HECHT-NIELSEN, R. **Neurocomputing**, 1 ed. Estados Unidos, Pearson, 1990.

[33] OLSON, D.L., DELEN, D. **Advanced Data Mining Techniques**, 1 ed. Estados Unidos, Springer, 2008.

[34] ALPAYDIN, E. **Introduction to Machine Learning**, 1 ed. Estados Unidos, MIT Press, 2004.