PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS
Programa de Pós-graduação em Engenharia Elétrica

Alexandre Reis Mundim

PROPOSTA DE ARQUITETURA PARA PREDIÇÃO DE PREÇOS UTILIZANDO MACHINE LEARNING EM CONJUNTOS E DADOS RUIDOSOS: UMA APLICAÇÃO EM PRODUTOS SEMI-ACABADOS

Belo Horizonte
2022

Alexandre Reis Mundim

PROPOSTA DE ARQUITETURA PARA PREDIÇÃO DE PREÇOS UTILIZANDO MACHINE LEARNING EM CONJUNTOS E DADOS RUIDOSOS: UMA APLICAÇÃO EM PRODUTOS SEMI-ACABADOS

Dissertação apresentada ao Programa de Pós-graduação em Engenharia Elétrica da Pontifícia Universidade Católica de Minas Gerais, como requisito para obtenção do título de Mestre em Engenharia Elétrica.

Orientador: Prof. Dr. Alexei Manso Correa Machado

Linha de pesquisa: Processamento e Análise de Sinais

Belo Horizonte
2022

FICHA CATALOGRÁFICA
Elaborada pela Biblioteca da Pontifícia Universidade Católica de Minas Gerais

Mundim, Alexandre Reis
M965p Predição de preços utilizando machine learning em conjuntos de dados com outliers: uma aplicação em produtos semi-acabados / Alexandre Reis Mundim. Belo Horizonte, 2022.
60 f. : il.

Orientador: Alexei Manso Correa Machado

Dissertação (Mestrado) – Pontifícia Universidade Católica de Minas Gerais. Programa de Pós-Graduação em Engenharia Elétrica

1. Inteligência computacional. 2. Aprendizado do computador. 3. Análise de regressão. 4. Algorítmos genéticos. 5. Redes neurais (Computação). 6. Teoria da previsão. 7. Negócios. I. Machado, Alexei Manso Correa. II. Pontifícia Universidade Católica de Minas Gerais. Programa de Pós-Graduação em Engenharia Elétrica. III. Título.

CDU: 681.3.091
Ficha catalográfica elaborada por Elizângela Ribeiro de Azevedo - CRB 6/3393

Alexandre Reis Mundim

PROPOSTA DE ARQUITETURA PARA PREDIÇÃO DE PREÇOS UTILIZANDO MACHINE LEARNING EM CONJUNTOS E DADOS RUIDOSOS: UMA APLICAÇÃO EM PRODUTOS SEMI-ACABADOS

Dissertação apresentada ao Programa de Pós-graduação em Engenharia Elétrica da Pontifícia Universidade Católica de Minas Gerais, como requisito para obtenção do título de Mestre em Engenharia Elétrica.

Linha de pesquisa: Processamento e Análise de Sinais

Prof. Dr. Alexei Manso Correa Machado (Orientador) – PUC Minas

---

Prof. Dr. Pedro O. S. Vaz de Melo – UFMG

Prof. Dr. Gustavo Luís Soares – PUC Minas

Belo Horizonte, 29 de Abril de 2022

Aos meus pais, Denise e Luís Ângelo, meu irmão, Luís Gustavo, e à minha esposa, Júlia.

AGRADECIMENTOS

Agradeço aos meus pais, Denise e Luís Ângelo, pelo incentivo e apoio durante todo meu desenvolvimento acadêmico. Vocês são para mim o retrato de como a educação pode provocar transformações.

Ao meu irmão, Luís Gustavo, pelas palavras de motivação nos momentos difíceis e que exigiram persistência durante o desenvolvimento dessa pesquisa.

À minha esposa, Júlia, pela compreensão durante o longo período de dedicação e desafios.

Ao meu orientador Prof. Dr. Alexei Machado, pelas orientações, provocações e dedicação. Agradeço sua compreensão pelos momentos difíceis dessa trajetória.

Aos professores do mestrado, em especial Profa. Dra. Flávia Magalhães e Prof. Dr. Gustavo Luís, por terem me apresentado conhecimentos que sozinho eu não teria descoberto.

Ao PPGEE PUC Minas e demais colegas pesquisadores, que persistem dedicando para o avanço da pesquisa, ciência, engenharia e tecnologia. Obrigado pelos conhecimentos compartilhados.

Aos meus colegas de trabalho, pelos incentivos e sugestões.

"When something is important enough, you do it even if the odds are not in your favor." Elon Musk

RESUMO

Bens intermediários são itens baseados em commodities, utilizados como insumos para a produção de outros produtos. Para sua negociação, o preço é o principal fator na decisão de compra pelos clientes, visto que a qualidade e a especificação técnica são equivalentes entre diversos fornecedores. Estratégias convencionais de precificação baseadas em regras de negócio ou no conhecimento de profissionais de venda podem considerar quantitativamente apenas uma parcela reduzida de variáveis, limitando a escalabilidade de operações comerciais e sua adaptabilidade, especialmente em mercados com mudanças bruscas de preço. Adicionalmente, a distribuição estatística resultante da coleta de dados em ambientes com baixo nível de digitalização pode ser assimétrica e abundante em *outliers*, que podem dificultar a análise e detecção de padrões. Este trabalho propõe a utilização de métodos baseados em aprendizado de máquina para a estimativa do preço de venda de bens intermediários, como aqueles compostos por metais não-ferrosos e plásticos de engenharia. As contribuições deste trabalho estão na proposição de uma abordagem baseada em *Automated Machine Learning* para a parametrização de modelos preditivos em cascata, incluindo etapa para pré-processamento de *outliers*. A proposta apresentou melhoria dos resultados alcançados se comparados quantitativamente com abordagens baseadas em *Stacking* e, se confrontada com redes neurais, obteve desempenho equivalente. Qualitativamente, a solução consegue identificar transações anômalas, auxiliando especialistas humanos a tomarem melhores decisões de precificação.

Palavras-chave: Precificação. Bens Intermediários. Distribuição Assimétrica. Aprendizado de Máquina.

ABSTRACT

Intermediate goods are commodity-based items used as input in the production of other products. Therefore, price is the main factor in the customer's purchase decision, since its quality is roughly the same from any vendor. Conventional pricing strategies based on business rules or the experience of sales personnel can only consider a limited amount of factors quantitatively, lacking scale and adaptation, especially in fast-changing markets. Furthermore, the resulting statistical distribution from collecting data in low-digitalization environments can be skewed and abundant in outliers, which can make pattern recognition and analysis difficult. In this paper, the application of learning-based methods for the selling price estimation of metal-based and engineering plastic goods is proposed. The contributions of this work lie in the proposal of an Automated Machine Learning based approach for the parameterization of cascading predictive models, including a preprocessing step for outliers. Improvements in the results were achieved when quantitatively comparing the proposal with Stacking based approaches and, if confronted with neural networks, it obtained a equivalent performance. Qualitatively, the solution is able to identify anomalous transactions, helping human experts to make better pricing decisions.

Keywords: Pricing. Intermediate Goods. Skewed Distribution. Machine Learning.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

A precificação é um dos principais elementos de qualquer relação comercial e cumpre papel fundamental na interação entre oferta e demanda. Ela impacta diretamente a receita e a lucratividade de empresas, visto que más decisões podem levar à redução de margens se preços forem ajustados abaixo; assim como preços excessivamente altos podem reduzir o ritmo de crescimento de negócios, conforme exposto por (KATSOV, 2017). Mudanças de preço podem fomentar o consumo, auxiliando no gerenciamento de estoque e, consequentemente, na saúde financeira das empresas. O problema de precificação já foi estudado historicamente e devido a fatores como a digitalização da economia, o aumento no volume de dados disponíveis e aumento de poder computacional, a adoção de algoritmos para a tarefa de precificação se popularizou, conforme (CAVALLO, 2018).

Apesar da evidente utilização deste tipo de solução, existem organizações que, devido às suas particularidades, precisam de soluções customizadas para resolver o problema. Negócios inteiramente digitais permitem a coleta de dados de potenciais clientes em maior escala e velocidade, por exemplo. Por outro lado, empresas voltadas ao mercado *offline* não possuem essas facilidades, portanto precisam ser mais eficientes ao lidarem com volumes menores de informação e distorções decorrentes da coleta ineficiente de dados.

1.1 Problema Motivador

Segundo (den Boer, 2015a), precificação dinâmica é o estudo da determinação de preços ótimos de venda de produtos ou serviços, em circunstâncias em que preços podem ser frequentemente ajustados. A exploração do conhecimento presente em dados históricos pode gerar vantagem competitiva a negócios, dado que estas contêm informações sobre o comportamento de consumidores e como estes respondem a diferentes preços. Há extensa literatura sobre o assunto, com contribuições oriundas de diversos campos do conhecimento, a exemplo da pesquisa operacional - que busca o preço que maximiza a receita e/ou a lucratividade de um vendedor; economia — normalmente preocupadas em explicar a formação e o comportamento de preços; e ciência da computação — que, em geral, provê o desenvolvimento de modelos de aprendizado de máquina, com o intuito de aprender de que maneira fatores como demanda, competidores e clientes estratégicos impactam na precificação, obtendo um retrato realista do funcionamento de um mercado.

Segundo (den Boer, 2015b), uma propriedade intrínseca deste problema de decisão é a ausência de informação para um agente vendedor: este não sabe como os consumidores responderão a diferentes preços de venda e, consequentemente, não é possível saber de antemão qual o preço ótimo. Dessa maneira, busca-se determinar a relação entre preço e sua respectiva absorção pelo mercado, sendo a utilização de modelos de *Machine Learning* uma opção viável para tal. Em modelagem preditiva, duas tarefas são candidatas para a resolução do problema: classificação e regressão. A primeira pode ser utilizada para

discernir as oportunidades que converteram daquelas que não para dado preço. Posteriormente, e com o auxílio de algoritmos de otimização, a busca do preço que maximiza a receita pode ser executada. Contudo, essa abordagem requer que dados de ambas classes sejam coletadas: vendas e não vendas. Caso apenas dados das vendas estejam disponíveis, é possível utilizar regressão. Nesse caso, é possível prever qual será o valor que determinada transação irá ocorrer, sendo essa informação valiosa para negócios.

Ainda que abordagens baseadas em aprendizado de máquina sejam viáveis, estas são sensíveis aos dados disponíveis. Em algumas ocasiões e não se limitando ao contexto de precificação, os dados podem representar apenas uma parcela da totalidade da realidade. Esse comportamento pode acontecer por fatores como coleta precária dos dados em frequência ou resolução; intervenções e ruído; e/ou erro humano. Observações impactadas ou geradas dessas maneiras são denominadas *outliers*, caracterizadas por (HAWKINS, 1980) pelo seu desvio substancialmente distinto do restante da amostra, a ponto de levantar suspeitas que foi gerada por outro mecanismo. Nesse contexto, o termo anomalia também é comumente empregado. Em outras circunstâncias, estes valores podem estar presentes nos dados de maneira fidedigna à realidade, naturais de sua própria distribuição. Consequentemente, não há porque questionar sobre a erroneidade desta observação (HAWKINS, 1980). Mesmo que este seja a conjuntura, caso as demais variáveis presentes sejam incapazes de estabelecer relações com a variável alvo, modelos preditivos tendem a ser prejudicados, conforme (FERNáNDEZ; BELLA; DORRONSORO, 2022). Isso impede com que os modelos generalizem adequadamente e, ao fornecer predições, tenham seu desempenho reduzido. Métodos podem ser empregados para identificar *outliers* antes de regressões ou classificações, com o intuito de reduzir sua influência. Ainda conforme (FERNáNDEZ; BELLA; DORRONSORO, 2022), a abordagem de detecção de *outliers* é utilizada para a identificação e posterior separação de observações com esse comportamento. Na sequência, uma base de dados livre de *outliers* pode ser utilizada para o treinamento de modelos. É importante ressaltar que a separação de *outliers* durante o aprendizado não impede que observações similares surjam em ambientes de fornecimento de predições, a exemplo de sistemas em produção. Portanto, é relevante que estratégias sejam aplicadas em ambos os contextos.

Dessa forma, o problema que norteia a presente dissertação é a definição de um método capaz de executar a tarefa de predição de preços através de regressão com o menor erro possível, sendo capaz de lidar com alta incidência de *outliers*. Além disso, o método deverá encontrar o subconjunto de variáveis disponíveis que mais impacta a qualidade da predição, assim como seus respectivos tratamentos e transformações.

1.2 Motivação

Dado que a ocorrência de *outliers* e anomalias pode ser decorrente de uma coleta ineficiente de dados, é razoável propor alterações nos processos que regem o fenômeno visando

a melhoria da captura de informações. Isso permitiria a criação de conjuntos de dados mais fidedignos à realidade, o que enriqueceria a qualidade dos modelos preditivos. Por outro lado, o esforço para a realização dessas modificações pode ser extremamente oneroso operacionalmente, visto que pode provocar alterações na execução de atividades, reduzindo velocidade e/ou aumentando custos no curto prazo. Ainda assim, o tempo para o reflexo de mudanças significativas nas bases de dados é de longo prazo. Em contrapartida, a construção de soluções baseadas em dados que exijam esforço limitado para aplicação e auxiliem no ganho de velocidade e/ou geração de informação são desejáveis, dado que a operação poderia dedicar menos tempo em precificação e investir mais em partes qualitativas do processo, como a construção de relacionamento com clientes e fornecedores, assim como a própria expansão do negócio. Se efetiva, a solução analítica pode estimular a modificação de processos e investimentos visando a coleta e o enriquecimento de dados.

As regras de precificação são difíceis de serem mapeadas de maneira lógica, tornando o problema excessivamente complexo para ser programado. Essas características, segundo Burkov (BURKOV, 2020), fazem com que esse seja o contexto ideal para a aplicação de algoritmos de aprendizado de máquina. Porém, abordagens tradicionais de aprendizado de máquina utilizando apenas um modelo não fornecem resultados satisfatórios nesse caso de uso, devido à distribuição mal comportada dos dados. Soluções de precificação baseadas em dados são amplamente exploradas na literatura, mas não na conjuntura exata estudada aqui: em um conjunto de dados com apenas informações das transações que converteram em compras (impedindo a construção de classificadores); com uma alta cardinalidade de categorias de produtos; e estes sendo divisíveis e personalizáveis. Além disso, grande parte dos processos comerciais não são representados no conjunto de dados, impedindo a detecção completa do motivo por trás da geração de *outliers*. Consequentemente, uma solução customizada se faz necessária.

1.3 Objetivo

Este trabalho tem como objetivo desenvolver uma solução capaz de fornecer predições que permitam prever o valor de fechamento de negociações em conjuntos de dados com alta incidência de *outliers* e com o menor erro de generalização possível. No contexto desta dissertação, *outliers* são observações caracterizadas pelo seu desvio substancialmente distinto do restante da amostra, conforme (HAWKINS, 1980). A proposta desenvolvida aqui possui conceitos derivados de Chen et al. (CHEN et al., 2018): *Automated Machine Learning* (AutoML) através de algoritmos evolucionários aplicados à hiperparametrização de modelos em cascata; Ramakrishnan et al. (RAMAKRISHNAN et al., 2019): conceitos de detecção de anomalias em precificação; e Fernández et al. (FERNáNDEZ; BELLA; DORRONSORO, 2022): detecção supervisionada de *outliers* para posterior treinamento de modelos de regressão. O desenvolvimento da pesquisa, bem como os testes de desempenho, foram realizados em um conjunto de dados proprietário, oriundo de operações

comerciais de uma empresa de metais não-ferrosos e plásticos industriais. Ademais, este trabalho tem os seguintes objetivos específicos:

- Analisar variáveis disponíveis assim como discutir seus respectivos tratamentos;
- Associar e aplicar múltiplas técnicas de aprendizado de máquina para circunstâncias em que conjuntos de dados que sofram com a contaminação de *outliers*;
- Desenvolver método baseado em algoritmos genéticos para ajuste de hiper-parâmetros de modelos em cascata;
- Possibilitar a geração de *score*, indicando quantitativamente anomalias e oportunidades em transações comerciais, guiando especialistas humanos rumo a uma precificação mais eficiente;
- Verificar o desempenho de abordagens baseadas em *Stacking* e redes neurais e compará-las com o método proposto em base de dados com distribuições assimétricas;

1.4 Estrutura do Trabalho

Além do presente capítulo, este trabalho apresenta outros sete capítulos. O Capítulo 2 apresenta os principais trabalhos relacionados a esta pesquisa. O Capítulo 3 apresenta uma revisão das principais técnicas utilizadas. O Capítulo 4 descreve a metodologia proposta. O Capítulo 5 apresenta o objeto do estudo de caso, assim como os materiais e tratamentos aplicados aos dados. O Capítulo 6 apresenta experimentos, resultados obtidos e discussões acerca da metodologia proposta. As conclusões, considerações finais e propostas de continuidade para a pequisa desenvolvida são apresentadas no Capítulo 7.

2 TRABALHOS RELACIONADOS

Neste capítulo, são analisados trabalhos relevantes e quais aspectos destes contribuíram para a presente pesquisa. Primeiramente, são abordados trabalhos sobre inteligência computacional aplicada à precificação, classificados conforme os métodos empregados, que por sua vez dependem da particularidade da operação modelada e da sua respectiva disponibilidade de dados. Fatores como mensuração da concorrência, interferência humana, atributos de produtos e clientes, qualidade e volume de dados trazem particularidades para cada ocasião. Finalmente, são abordados trabalhos com contribuições que não se limitam a aplicações voltadas à precificação, como detecção de *outliers* e ajuste de hiperparâmetros de modelos, utilizados na presente pesquisa para pré-processamento dos dados e otimização do desempenho de preditores, respectivamente.

2.1 Inteligência computacional aplicada à precificação

As abordagens podem ser divididas em (I) otimização via simulação, quando há ciclos definidos para ajuste de preços, conforme (DASGUPTA; DAS, 2000); (II) regressão, aplicadas em circunstâncias em que se fornecerá a predição de preços, a exemplo de (GUPTA; PATHAK, 2014); em ocasiões com a possibilidade de modelagem de demanda e/ou quando também há dados disponíveis sobre oportunidades que não converteram em vendas, é possível modelar por (III) classificação associada a otimização, conforme (YE et al., 2018), (SHUKLA et al., 2019) e (SCHLOSSER; BOISSIER, 2018) ou (IV) aprendizado por reforço, vide (NARAHARI et al., 2015) e (MAESTRE et al., 2018). Conforme (NARAHARI et al., 2015), as classificações não são mutuamente exclusivas.

O artigo de Dasgupta et al. (DASGUPTA; DAS, 2000) modela o processo de precificação dinâmica a partir da premissa que a disponibilidade de informações de competidores é limitada, o que faz jus à muitos cenários reais da economia. São propostas duas categorias de clientes: aqueles que conduzem uma pesquisa extensa dos preços disponíveis e sempre fecharão com o proponente mais barato e outros que comprarão aleatoriamente, desde que o valor de venda esteja dentro de seu orçamento. Assume-se que os produtos são indivisíveis e as compras são feitas ao longo de um período. No término deste, os preços podem ser reajustados. Por conta desses fatores, a aplicação dos métodos propostos perde relevância em circunstâncias em que há customização ou divisibilidade de produtos. Na análise, são comparados três algoritmos: o *Derivative Follower Pricing Strategy* (DF), *Adaptive Step-size Derivative Follower Strategy* (ADF) e *Model-Optimizer Strategy* (MO), todos baseados em otimização. O algoritmo DF é a mais estratégia mais simples: são aplicados incrementos no preço para cada intervalo de tempo e, na medida com que a rentabilidade observada cresce, o comportamento é reforçado e a direção da movimentação é mantida. Caso a rentabilidade reduza, a direção de movimentação do preço é invertida. O método ADF é a proposição de uma melhoria ao método DF, al-

terando dinamicamente o tamanho do incremento, evitando grandes flutuações no preço e, consequentemente, reduzindo a exposição a potenciais perdas de receita. O algoritmo MO destoa das abordagens anteriores, visto que este estende o intervalo de tempo e encontra relações históricas entre preço e rentabilidade através de polinômios, reajustado na medida com que novas observações são coletadas.

Gupta e Pathak (GUPTA; PATHAK, 2014) possui grande intersecção de características com a presente pesquisa. A principal diferença está no objetivo do método implementado, em que os autores buscam uma previsão de compra e aqui busca-se o valor de venda. O artigo apresenta com detalhes quais variáveis foram incorporadas ao modelo e esse descritivo foi utilizado aqui de maneira a agrupar as variáveis utilizadas. O algoritmo de agrupamento *K-Means* foi aplicado para a segmentação dos clientes, mas apenas para aqueles com perfil recorrente. Dessa forma, uma estratégia para novos clientes fica em aberto, sendo que estes poderão tornar recorrentes no futuro. No presente trabalho, buscou-se elaborar abordagens para clientes sem esse requisito.

O trabalho de Narahari et al. (NARAHARI et al., 2015) apresenta fundamentos da precificação dinâmica e categoriza as soluções a partir de sua modelagem: modelos baseados em inventário, análise estatística de dados, Teoria dos Jogos, *Machine Learning* e simulações. É destacado que as categorias não são mutuamente exclusivas e pode haver abordagens que englobem mais de uma delas. É apresentada uma comparação entre quatro estratégias de precificação dinâmica. Em um cenário heterogêneo, onde cada um dos proponentes possui uma estratégia de precificação diferente — que vai de encontro com o contexto explorado aqui — as abordagens *Myoptimal Pricing* (gulosa) e baseadas em teoria dos jogos superaram o algoritmo *Derivative-Following*, enquanto o algoritmo de Aprendizado por Reforço *Q-Learning* superou as demais estratégias.

Maestre et al. (MAESTRE et al., 2018) também apresenta uma solução baseada em Aprendizado por Reforço. O trabalho analisa, também, os efeitos do algoritmo considerando uma métrica de justiça. Dessa forma, o método propõe, por uma abordagem multiobjetivo, otimizar o preço e a percepção de justiça entre grupos de diferentes consumidores. O trabalho aplica o algoritmo de *Q-Learning* associado a uma rede neural artificial (*Deep Reinforcement Learning*). Ao modelar um algoritmo desse tipo, deve-se parametrizar o quão imediata será a recompensa ao agente. No artigo, este era recompensado se uma cotação se convertesse em uma venda. Ao utilizar-se métodos de Aprendizado por Reforço, é possível determinar o prazo das recompensas ao agente, permitindo o fomento a resultados de curto prazo em cenários de poucas vendas como também a recorrência de compras e fidelização de clientes se ajustado para um horizonte de tempo maior. Dado que na presente pesquisa estão disponíveis apenas dados de oportunidades que geraram conversão, modelar o problema de uma maneira análoga não é possível.

Outro artigo que apresenta uma métrica customizada é o trabalho de Ye et al. (YE et al., 2018), que analisa a estratégia de precificação para as listagens da plataforma de

aluguéis Airbnb. No modelo de negócios da empresa, os hospedeiros conseguem determinar o preço final que desejam para sua hospedagem e assim seu algoritmo fornece uma sugestão de preço. A presença da intervenção humana foi um dos fatores que levou os autores a definirem a métrica exclusiva. Dado que o hospedeiro possui atuação sobre o preço sugerido, entende-se que essa publicação possua elevada convergência com o presente trabalho em termos do processo modelado. Aqui, deseja-se estimar o preço que a venda irá concretizar, porém, a intervenção humana pode acontecer para garantia da transação, alterando o valor predito (análoga ao comportamento dos hospedeiros). Tecnicamente, são utilizados preditores para alimentar uma *Gradient Boosting Machine* (GBM), que entrega uma probabilidade de aluguel de determinada listagem. Em seguida, e utilizando também a predição da GBM como variável, um componente denominado modelo de estratégia sugere o preço ótimo para cada propriedade listada por uma regressão. Esta utiliza a métrica customizada para guiar o aprendizado. Finalmente, dados instantâneos do mercado, detectados no nível de um grupo de hospedagens, são incorporados com a intenção de ajustar o preço a objetivos e comportamentos do hospedeiro, assim como surtos de demanda. Segundo os autores, essa configuração adiciona robustez ao modelo, mas detalhes dos preditores ou do processo de agrupamento não são apresentados.

Shukla et al. (SHUKLA et al., 2019) apresenta uma solução para a precificação dinâmica de adicionais (espaço extra, bagagem, itens de conforto, etc.) para passagens aéreas. Diferentemente de Ye et al. (YE et al., 2018), os autores propõem uma modelagem de demanda que engloba tanto o preço pelo item como características do cliente, uma abordagem estendida à pesquisa corrente, por variáveis de comportamento comercial e financeiro de clientes. Foram desenvolvidos três modelos distintos para a estimativa da probabilidade de compra (classificador binário) e posterior precificação dos produtos: O primeiro modelo (*Ancillary purchase prediction with logistic mapping* - APP-LM) é composto por um algoritmo de *Gaussian Naive Bayes with Clustered Features* (GNBC) para estimativa de probabilidade seguido por função logística para mapeamento do preço de oferta. O segundo modelo (*Ancillary purchase prediction with exhaustive search* - APP-DES) funciona da mesma forma, porém o cálculo da probabilidade é realizado por uma rede neural. Em seguida, a probabilidade é utilizada como entrada em uma busca exaustiva pelo preço ótimo a ser ofertado. O último modelo desenvolvido (*End-to-End DNN with custom loss function* - DNN-CL) é inteiramente baseado em uma rede neural profunda, otimizada para a função customizada baseada em Ye et al. (YE et al., 2018). Finalmente, o trabalho apresenta um comparativo entre os modelos desenvolvidos e uma série de métricas para avaliação do desempenho técnico do modelo assim como para avaliação perante os requisitos do negócio. O modelo inteiramente baseado em redes neurais profundas possui o melhor desempenho nos experimentos. O trabalho apresenta alguns dos preditores utilizados para a modelagem dos algoritmos e seu teste foi realizado em um conjunto de dados de proporções equivalentes ao disponível aqui.

Schlosser e Boissier (SCHLOSSER; BOISSIER, 2018) apresentam uma solução que engloba diversas características em comum com o problema estudado aqui. A principal delas é que os vendedores não possuem informação completa a respeito do mercado e, sob essa incerteza, precisam definir estratégias de precificação. Para a previsão da demanda e estimativa da probabilidade de compra em determinados preço e condições de mercado, os algoritmos de Regressão Logística, *Least Squares*, *Gradient Boosting Trees* e *Multi-Layer Perceptron* (Rede Neural) tiveram uma boa performance, ao contrário dos métodos de *Random Forests* e *Support Vector Machines*. Em seguida, o cálculo do preço é feito a partir de programação dinâmica. Existem questões que impedem a replicação do trabalho sem modificações: são ofertados produtos unitários indivisíveis e sem customização, a modelagem é conduzida através da estimativa da curva de demanda e os dados a respeito da precificação adotada pelos concorrentes está disponível. Essa última característica faz parte da natureza de *marketplaces* como a Amazon, conforme apresentado no artigo.

Um artigo com a presença de dados esparsos e ruidosos é o de Bauer et al. (BAUER; JANNACH, 2018), característica que converge com a presente pesquisa. Ademais, são levadas em considerações situações em que não há dados de vendas anteriores, a exemplo da introdução de novos produtos ao portefólio. Nesses casos, propõe-se um preço ótimo de venda dada a categoria do produto, característica adotada aqui através da utilização de variáveis do grupo que cada item pertence. É salientado que informações adicionais a respeito dos consumidores não estavam disponíveis, diferindo do presente trabalho, visto que esses dados são acessíveis e sua consideração é beneficial. É utilizada uma abordagem que incorpora Inferência Bayesiana, *Kernel Regression* e Árvores de Decisão.

2.2 Detecção de *outliers*

O trabalho de Ramakrishnan et al. (RAMAKRISHNAN et al., 2019) não é voltado para a definição de preço, mas à detecção de anomalias no processo. Foi desenvolvida uma aplicação para a rede Walmart e é um trabalho pioneiro na intersecção entre detecção de anomalias nesse contexto. Ainda que seu foco não seja dedicado exclusivamente à predição de preços, ele provê informações relevantes em relação ao desenvolvimento e implantação de um modelo de Aprendizado de Máquina em um ambiente comercial de varejo: são apresentados detalhes quanto à seleção de preditores e, também, um comparativo entre algoritmos de aprendizado para a execução da tarefa. Diferentemente de outros trabalhos analisados, a utilização de atributos relativos ao preço de custo de seus produtos é evidenciada. Essa estratégia garante o lucro nas operações, mesmo que com margens baixas. Há negócios que adotam estratégias agressivas de precificação, obtendo prejuízo em itens, mas favorecendo o preço final do *mix* de produtos vendidos, o crescimento do negócio e/ou como combate à concorrência. O *Gaussian Naive Bayes* foi o escolhido como modelo de referência pelo reduzido tempo de execução e baixa complexidade. Buscando superar seu desempenho, foram testados modelos supervisionados (*Gradient Boosting Machine,*

Random Forest) e não supervisionados (*Isolation Forest* e *AutoEncoder*).

Também voltada para a detecção de *outliers* mas não se limitando à contextos de precificação, a pesquisa de Fernández et al. (FERNáNDEZ; BELLA; DORRONSORO, 2022) propõe a utilização de métodos para a detecção de *outliers*. Assim, é possível executar o pré-processamento de bases, antecedendo a aplicação de regressores ou classificadores. Para isso, são construídas marcações através da utilização de um limiar de contaminação, gerado por algoritmos de detecção de anomalias, que transformam os rótulos da tarefa final em duas classes: *inlier* ou *outlier*. Na medida com que hiper-parâmetros dos algoritmos de detecção de anomalias são ajustados, o limiar de contaminação é modificado e a definição quanto à anormalidade de cada observação é alterada. Ao obter essa informação para cada ponto de dado, o *dataset* é filtrado e fica livre de *outliers*. O modelo final de predição é treinado através do conjunto saneado. Para não ocorrerem fenômenos de vazamento de dados (*data leakage*), os conjuntos de dados são particionados para permitir o treinamento de detectores de *outliers* de maneira isolada à tarefa final de predição. O processamento favorece o desempenho de modelos preditivos em diversos conjuntos de dados *benchmarks*. Para a tarefa de detecção de anomalias, os algoritmos *Minimum Covariance Determinant* (MCD) e *Isolation Forest* (IF) forneceram desempenhos satisfatórios, enquanto *Principal Component Analysis* (PCA), *Histogram-based Outlier Detection* (HBOS) e *Local Outlier Factor* (LOF) tiveram boa performance computacional.

2.3 Ajuste de hiperparâmetros

Adicionalmente e com escopo de aplicação mais amplo que o de aprendizado de máquina aplicado ao contexto de precificação, o ajuste de hiperparâmetros de modelos é um passo importante para praticamente qualquer análise preditiva. Dessa forma, encontrar uma maneira efetiva de ajustá-los é explorada cientificamente. Comercialmente, soluções de aprendizado de máquina automatizado (*Automated Machine Learning - AutoML*) se baseiam em métodos de otimização para a tarefa. O trabalho de (HE; ZHAO; CHU, 2021) traz uma apresentação detalhada de técnicas estado-da-arte de *AutoML*.

Como proposto por (SAFARIK et al., 2018) e (WICAKSONO; SUPIANTO, 2018), é possível aplicar algoritmos genéticos (AGs) para a tarefa em questão. Os autores aplicam o método para o ajuste de um único modelo preditivo, mas dado que a solução para o problema estudado na presente dissertação não é facilmente modelado por um único algoritmo, a aplicação do AG e sua respectiva função objetivo devem ser adaptadas. Isso mitigará a ocorrência de sobreajustes e explorará, positivamente, a interação entre modelos, que não é necessariamente linear. Chen et al. (CHEN et al., 2018) propuseram o desenvolvimento de um sistema de *AutoML* aplicado ao cascateamento e combinação de modelos, método que se mostrou efetivo em uma grande gama de tarefas de aprendizado de máquina. Algoritmos genéticos não assumem convexidade da função objetivo e apresentam convergência, encorajando sua utilização. Além desses fatores, a combinação

de modelos preditivos aumenta o espaço de busca do problema de otimização, portanto reduzir a busca para um método com menos parâmetros como o AG se torna vantajoso.

2.4 Conclusão

Com base na revisão bibliográfica realizada, observa-se que há linhas de trabalho que adotam aprendizado por reforço, redes neurais e regressores baseados em aprendizado de máquina. Quando é possível modelar a demanda, classificadores são aplicados em conjunto a algoritmos de otimização. É possível visualizar não haver um consenso entre qual técnica utilizar para processos de precificação, se utilizadas ferramentas computacionais, vide a teoria que "não existe almoço grátis" em aprendizado de máquina e otimização (WOLPERT, 1996), (WOLPERT; MACREADY, 1997) e (WOLPERT, 2002). A não convergência dos artigos na escolha do método a ser aplicado reforça a tese que a definição do algoritmo é fruto direto dos dados disponíveis e da operação comercial modelada. Fatores como mensuração da concorrência, dados de oportunidades não convertidas, interferência humana no processo, dados dos produtos na compra e na venda, assim como perfis de clientes podem ser considerados. Infelizmente, poucas informações quanto ao tratamento dos dados foi disponibilizada nos artigos. Aqui, os processos de engenharia de variáveis foram detalhados.

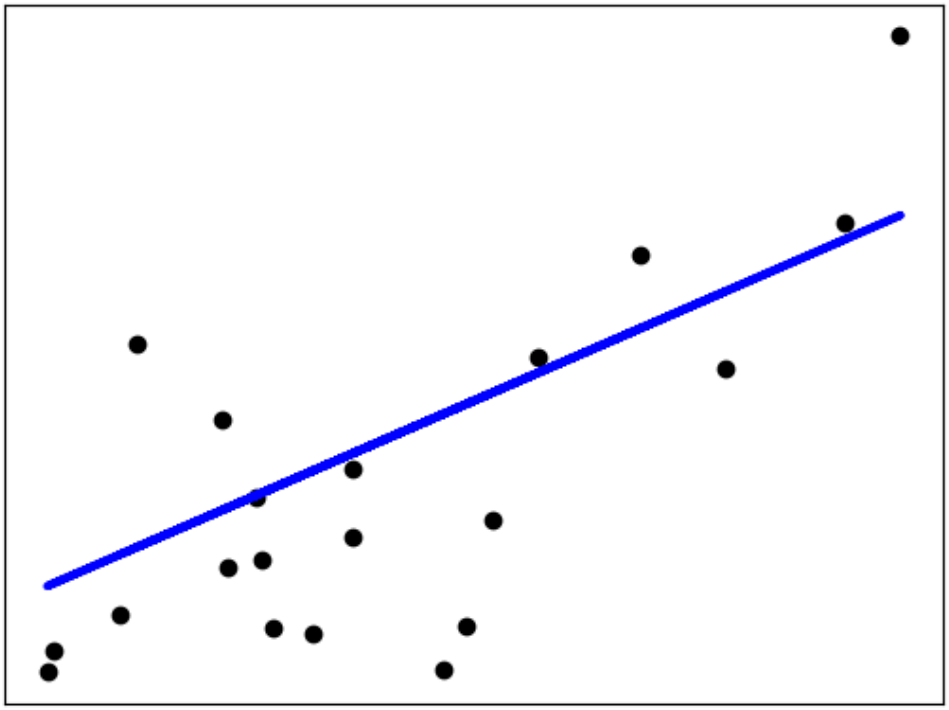
Na revisão, foram apresentadas oportunidades da aplicação de algoritmos de otimização para o ajuste de hiper-parâmetros (*AutoML*), em especial algoritmos genéticos. O trabalho de Chen et al. (CHEN et al., 2018) aplica a abordagem para cascatas de modelos preditivos. Segundo os autores, ainda que com a crescente popularização de *Big Data*, muitos conjuntos de dados são pequenos e esparsos. A aplicação de modelos em cascata colabora no combate aos desafios que surgem nessas circunstâncias, visto que a abordagem pode ser vista como a decomposição de um problema complexo nas partes que o constituem. Para a parte de detecção de anomalias em preços, o trabalho de Ramakrishnan et al. (RAMAKRISHNAN et al., 2019) exerceu grande influência na presente pesquisa, assim como Fernández et al. (FERNáNDEZ; BELLA; DORRONSORO, 2022). O trabalho de Ye et al. (YE et al., 2018) trouxe inspirações à modelagem de processos de precificação com intervenção humana e também como método, pela utilização de algoritmos de *Gradient Boosting*.

3 REFERENCIAL TEÓRICO

3.1 Regressão Linear

Segundo Deisenroth (DEISENROTH; FAISAL; ONG, 2020), regressão é o processo que encontra uma função que mapeia entradas a um valor numérico contínuo (Figura 1). Na maior parte dos casos, há a presença de ruído, mas o método busca encontrar a função que minimiza erros e explica a maioria do sinal nos dados. Ainda segundo Deisenroth (DEISENROTH; FAISAL; ONG, 2020), regressão é um dos problemas fundamentais de aprendizado de máquina, aparecendo em uma diversa quantidade de áreas de pesquisa e aplicações.

Figura 1 – Regressão Linear



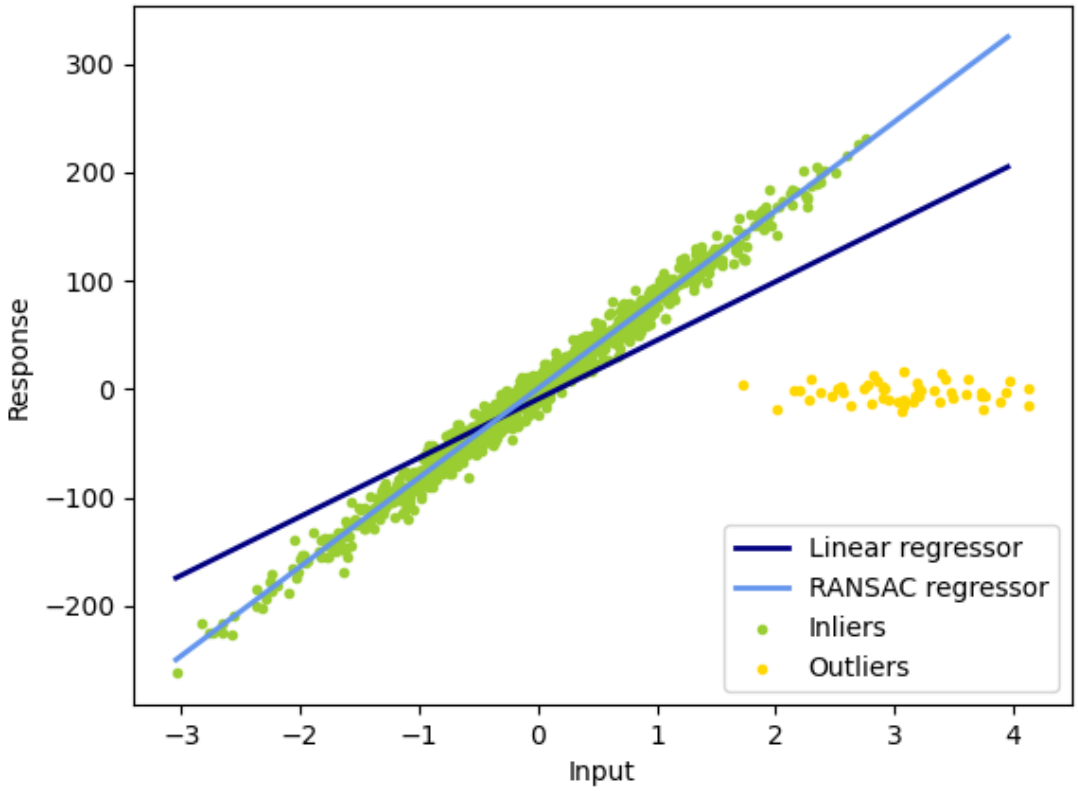
Fonte: Documentação da Biblioteca Sci-kit Learn.

Segundo Hastie (HASTIE; TIBSHIRANI; FRIEDMAN, 2001), modelos lineares de regressão foram desenvolvidos antes da aplicação da computação em estatística, mas ainda hoje existem boas razões para a sua utilização: são simples e fornecem certa interpretabilidade em como as entradas impactam a saída. Em casos com pequeno número de registros para treinamento e pouco ruído, métodos lineares podem ser mais eficientes que métodos não-lineares complexos. Finalmente, métodos lineares podem ser potencializados através

da aplicação de transformações nas entradas, aumentando seu escopo de aplicação. Esse último artifício será utilizado em nossa pesquisa.

No presente trabalho, onde há elevada presença de *outliers* (principalmente nos rótulos), é um contexto de atenção para modelos lineares tradicionais. Dessa forma, adotar essa categoria de modelo pode levar a erros na predição. Um dos métodos capazes de fornecer predições de qualidade nesse caso é a regressão RANSAC. É possível observar o método em comparação à regressão linear (Figura 2).

Figura 2 – Regressão Linear vs. Regressão RANSAC



Fonte: Documentação da Biblioteca Sci-kit Learn.

3.2 Regressão RANSAC

A regressão RANSAC (RANdom SAmple Consensus) é um método de regressão robusta, sendo um caso especial da regressão linear. Segundo a documentação da biblioteca de *software* Sci-kit Learn (PEDREGOSA et al., 2011), o algoritmo é eficiente em casos onde há forte incidência de *outliers* ou corrupções nos rótulos. No contexto de precificação, é possível que existam conjuntos de dados com distribuições assimétricas nos rótulos, o que pode atestar alta capacidade de especialistas comerciais na identificação de oportunidades de maximização de lucro. Por exemplo, clientes podem solicitar urgência em

um fornecimento, estando dispostos a pagar mais por isso. Fatores como esses não estão necessariamente representados em variáveis explicativas.

O método é um algoritmo iterativo e, de acordo com Choi et al. (CHOI; KIM; YU, 2009), contém duas fases realizadas até que o número de iterações seja completado: geração e avaliação de hipóteses. A primeira consiste em amostrar aleatoriamente e estimar parâmetros de uma regressão para esse subconjunto de dados. O outro processo calcula o erro da estimativa e conta a quantidade de candidatos dentro de uma margem de erro (ajustada pelo usuário). Se a quantidade de indivíduos na margem é o máximo até o momento, a melhor solução é atualizada.

Ainda segundo Sci-kit Learn (PEDREGOSA et al., 2011), é importante salientar que regressões robustas em ambientes de alta dimensionalidade é um problema extremamente difícil, visto que a esparsidade faz com que a maioria dos indivíduos seja *outliers*. Para atuar contra esse fator, é importante realizar a seleção de variáveis e, como foi feito aqui, utilizar métodos para redução de dimensionalidade, como a análise de componentes principais.

3.3 Análise de Componentes Principais

Segundo Deisenroth (DEISENROTH; FAISAL; ONG, 2020), trabalhar diretamente com dados de alta dimensionalidade é desafiador pelos seguintes agravantes: é difícil de analisar, sua interpretação e visualização são complexas e o armazenamento de grandes vetores pode ser dispendioso do ponto de vista prático. Por outro lado, a alta dimensionalidade possui características que compensam ser exploradas: frequentemente há variáveis redundantes que podem ser explicadas pela combinação de outros preditores. Além disso, a alta incidência de correlações pode indicar a presença de uma estrutura de baixa dimensão que norteia o fenômeno modelado. Técnicas de redução de dimensionalidade buscam encontrar estruturas menores dos dados, com perda de informação reduzida. A Análise de Componentes Principais (PCA), é uma técnica clássica e ainda é uma das técnicas mais utilizadas para compressão e visualização de dados.

A técnica visa elencar preditores conforme a sua variância: aqueles com menores variâncias são minimizados e ignorados de forma com que os dados comprimidos não sejam suficientemente diferentes da representação original. As componentes principais são ordenadas conforme a variância que explicam. A compressão se dá enquanto as componentes principais com menor contribuição são descartadas. A quantidade de descartes é definida pelo usuário. As componentes são ortogonais entre si.

3.4 Árvores de Decisão

Métodos baseados em árvores são constituídos por partições no espaço de variáveis ortogonais aos eixos e posterior definição de valor (como uma constante) ou classe para o grupo, para regressão e classificação, respectivamente. O algoritmo realiza partições

sucessivamente, definindo regras lógicas capazes de modelar o fenômeno estudado.

O algoritmo de árvore de decisão utilizado no presente trabalho é o *Classification and Regression Tree Algorithm* (CART), disponível na biblioteca Sci-kit Learn (PEDREGOSA et al., 2011). Segundo Géron (GERON; SAFARI, 2019), o algoritmo funciona primeiro dividindo o conjunto de dados em dois subconjuntos a partir de uma variável k e um limiar t_k. O algoritmo escolhe os valores do par k e t_k que produzirão os subconjuntos mais puros para tarefas de classificação ou aqueles que minimizem uma métrica para tarefas de regressão naquele ponto. Esse exercício é realizado recursivamente, criando nós de maneira subsequente. O fato do algoritmo apenas analisar o nó atual para a decisão de divisão, configura-o como guloso. Dessa forma, devemos nos satisfazer com uma solução sub-ótima ao utilizar o método, caso essa tenha sido a solução encontrada pelo modelo.

Segundo Hastie et al. (HASTIE; TIBSHIRANI; FRIEDMAN, 2001), árvores de decisão consistem no método mais próximo de uma solução *"off-the-shelf"*, podendo ser diretamente aplicadas a problemas de mineração de dados. Isso se dá por uma série de fatores, como a capacidade de lidar com diversas categorias de dados simultaneamente, inclusive valores faltosos; serem robustos a *outliers*; e apresentarem baixa sensibilidade a preditores irrelevantes. Outras propriedades relevantes das árvores de decisão são sua independência da distribuição de dados e a capacidade que o método possui de favorecer a interpretabilidade e explicabilidade de suas previsões.

Infelizmente, árvores de decisão possuem uma instabilidade: se o conjunto de dados de treinamento mudar, a árvore gerada pode ser completamente diferente, visto que os cortes no espaço de variáveis podem alterar. Para compensar esse efeito, existem os algoritmos de *Ensemble*.

3.5 Ensembles & Gradient Boosting Machines

Boosting se refere a qualquer método de *Ensemble* que combine métodos fracos de predição formando um método forte, de acordo com Géron (GERON; SAFARI, 2019). Na maior parte das vezes, esses métodos fracos são árvores de decisão. O principal racional é o aproveitamento das instabilidades das árvores de decisão para gerar múltiplas árvores descorrelacionadas (e até mesmo sobreajustadas aos dados) e, ao combiná-las, obtermos um algoritmo mais poderoso. Existem algumas variações dessa abordagem, como *Random Forests*, *Extra Trees* e algoritmos de *Gradient Boosting*.

O fundamento de algoritmos de *Gradient Boosting*, segundo (GERON; SAFARI, 2019), é o treinamento de árvores de decisão em sequência. *Gradient Boosting Machines (GBMs)* fazem isso através da adição de preditores com o intuito de corrigir a árvore predecessora, sendo cada modelo treinado nos resíduos do anterior.

Existem diversas implementações de GBMs, sendo o *Extreme Gradient Boosting (XGBoost)* uma das implementações mais populares. De acordo com Chen and Guestrin (CHEN; GUESTRIN, 2016), o algoritmo consegue resolver problemas de grande escala

utilizando uma quantidade mínima de recursos. Sua efetividade na resolução de problemas de precificação foi comprovada por Ye et al. (YE et al., 2018). Esse fator, juntamente à disponibilização do algoritmo implementado em código aberto, foram fatores que favoreceram sua utilização nesta pesquisa.

3.6 Stacking

Segundo (GERON; SAFARI, 2019), *Stacked generalization*, ou *Stacking* é uma abordagem que permite explorarmos o ponto levantado anteriormente. A visão por trás da técnica é treinar um modelo (chamado *Blender* ou *meta learner*) para agregar as predições de múltiplos modelos. Cada um dos modelos fornece diferentes predições conforme seu treinamento e seus parâmetros, para em seguida o *Blender* usar essas predições como entrada e concretizar uma predição final.

Como os modelos da cadeia, entende-se que a regressão RANSAC e o modelo de XGBoost são de naturezas distintas, trazendo descorrelação às suas predições e enriquecem a saída do *Blender*. Dessa forma, conseguiremos minimizar a interferência dos *outliers* quando necessário pela regressão robusta e, se compreendidos os fenômenos que geram oportunidades de aumento de preço, o GBM irá capturar.

3.7 Redes Neurais: Multi-Layer Perceptrons

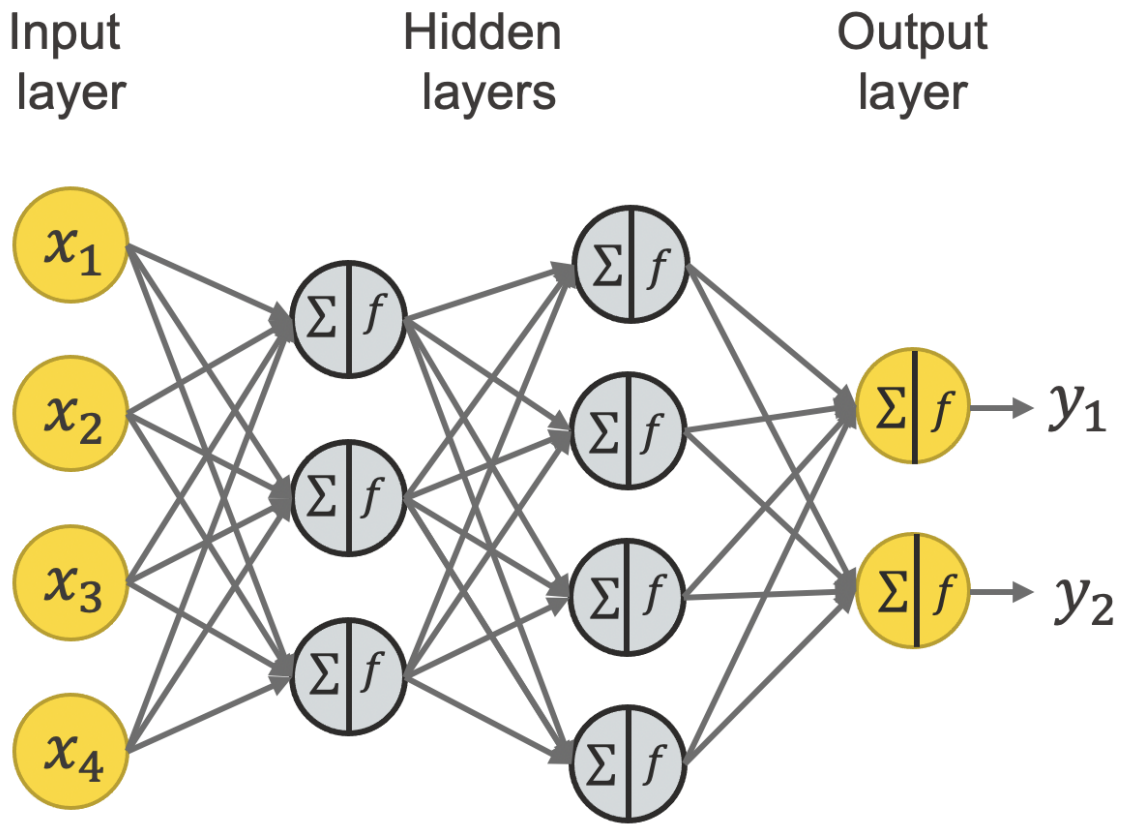
Segundo (GOODFELLOW; BENGIO; COURVILLE, 2016), o desenvolvimento de aprendizado profundo foi motivado parcialmente pela falha de algoritmos tradicionais em generalizar em tarefas com alta dimensionalidade. Ainda conforme os autores, *Multi-Layer Perceptrons* (MLPs), também conhecidos como *Feedforward Neural Networks*, são a essência de modelos de aprendizado profundo. Estes modelos possuem como objetivo serem aproximadores universais de funções, obtendo um mapeamento entre o vetor de entrada e o rótulo. Os fundamentos de modelos MLPs compõem a base de métodos mais avançados, como redes convolucionais ou recorrentes e estão presentes em muitas aplicações comerciais, atuando como os principais motores para serviços de reconhecimento de imagem, processamento de linguagem natural e muito mais.

Inicialmente, estes modelos possuíam semelhanças com componentes cerebrais, sendo vagamente inspirados pela neurociência. Na medida em que o desenvolvimento de modelos dessa natureza foram estudados e se tornaram um campo moderno de pesquisa em matemática e engenharia, estes métodos se distanciaram da sua inspiração biológica.

Seu funcionamento é dado por camadas de unidades de computação, chamadas neurônios que, por sua vez, formam uma rede. Sua versão inicial com uma camada se assemelhava a modelos lineares. Porém, a combinação de modelos lineares por si só é equivalente a um único modelo linear, o que não permite a aproximação de funções não-lineares. Para que a não-linearidade seja inserida no sistema, cada neurônio aplica uma função não-linear, chamada função de ativação, necessariamente derivável. Essa exigência faz com que seja

possível encontrar o ponto que minimiza os erros de predição através de otimização. Em um processo que promove uma redução de erros decremental, chamado retropropagação (*backpropagation*) (RUMELHART; HINTON; WILLIAMS, 1986), a contribuição de cada neurônio para a predição é atualizada, promovendo o aprendizado. O processo de aprendizado acontece por épocas, onde o conjunto de dados é exposto à rede integralmente. Um exemplo de rede neural pode ser visualizado (Figura 3).

Figura 3 – Exemplo de Rede Neural MLP



Fonte: https://www.knime.com/blog/a-friendly-introduction-to-deep-neural-networks.

3.8 Métodos para Parametrização de Modelos

Além da qualidade e quantidade dos dados a que são expostos, a performance dos algoritmos de aprendizado é dependente da especificação de seus hiper-parâmetros. No contexto de aprendizado de máquina, hiper-parâmetros não são aprendidos pelos algoritmos, mas definidos antes ou durante o treinamento. Os hiper-parâmetros atuam de forma com que os algoritmos encontrem o melhor ajuste aos dados, minimizando erros de predição e maximizando a generalização para dados não observados. Algoritmos capazes de modelar fenômenos complexos em geral possuem mais hiper-parâmetros, a exemplo de redes neurais (através do número de camadas e neurônios, taxa de aprendizado, funções

de ativação e outros), árvores de decisão e algoritmos derivados (através da profundidade da árvore, número mínimo de amostras por folha e outros).

3.8.1 Grid & Random Search

Ainda que existam algoritmos de aprendizado que ofereçam maior capacidade de modelagem, o espaço de busca de seus hiper-parâmetros é de dimensão elevada, de certa forma proporcional a seu poder preditivo. Dessa forma, a busca pela combinação de ajustes que minimize os erros do modelo é um problema de otimização, conforme Bergstra e Bengio (BERGSTRA; BENGIO, 2012). Ainda conforme os autores, mesmo com a publicação de diversos algoritmos para otimização de hiper-parâmetros, o método de *Grid Search* é um dos mais adotados. O método consiste na busca exaustiva de ajustes dado um conjunto de valores discretos de candidatos para cada um dos hiper-parâmetros. Os autores advogam a favor da técnica de *Random Search*, que consiste na busca aleatória de ajustes candidatos dado intervalo de possíveis valores. O racional por trás da abordagem é que ela apresenta as mesmas qualidades da *Grid Search* como a simplicidade conceitual, facilidade de paralelismo e implementação, com o bônus de ter uma performance melhor em espaços de busca maiores. Dado que múltiplos modelos serão treinados nessa aplicação e a técnica teve sua eficiência comprovada na publicação de Bergstra e Bengio (BERGSTRA; BENGIO, 2012), iremos adotá-la para a parametrização dos hiper-parâmetros de alguns modelos nessa pesquisa.

3.8.2 Algoritmos Genéticos

Segundo (BURKOV, 2020), a utilização de modelos de aprendizado em cascata é uma prática comum, mas deve ser aplicada com cautela, sendo uma providência para mitigação de riscos o treinamento e atualização dos modelos simultaneamente. Considerando esse fator, associado ao aumento no espaço de busca, uma maneira de executar esse procedimento é por algoritmos genéticos, método extremamente flexível, capaz de convergir em problemas de otimização sem garantia de convexidade, sejam estes restritos ou irrestritos, vide (YOUNG et al., 2015).

Algoritmos genéticos são uma heurística de busca inspirada na teoria da evolução natural de Charles Darwin. O algoritmo imita o processo evolutivo, em que indivíduos mais adaptados são eleitos para reprodução. Seus descendentes são submetidos a mutações, gerando indivíduos possivelmente mais adaptados na próxima geração. No processo artificial, os processos são denominados seleção, cruzamento e mutação. Diferentemente da seleção natural, a seleção artificial possui dispositivos que aceleram o processo evolutivo, sendo um destes o elitismo. O elitismo garante com que indivíduos com elevados níveis de adaptação sejam mantidos na população enquanto se mantiverem relevantes. O algoritmo genético aplicado no presente trabalho é derivado de (SOARES, 1997).

Conforme exposto por (YANG; SHAMI, 2020), os AGs podem ser ineficientes devido

à sua baixa velocidade de convergência. No entanto, a versatilidade dos AGs pode ser facilmente adaptada para o ajuste de funções objetivas complexas. Um exemplo de sistema com esse nível de complexidade é a proposta desta pesquisa, que contém ajuste de múltiplos modelos. Além disso, o espaço de busca do AG aqui não está limitado apenas ao escopo dos hiperparâmetros, mas também busca a melhor seleção de variáveis, conforme (JUNG; ZSCHEISCHLER, 2013) e (HONG; CHO, 2006). Esse fenômeno faz com que abordagens tradicionais de seleção de variáveis, como buscas recursivas, *Forward Selection* ou *Backward Elimination* sejam limitadas à seleção de features, sem trazer melhorias à busca dos hiper-parâmetros. Além disso, essas técnicas são extremamente exaustivas.

3.9 Detecção Supervisionada de *Outliers*

Segundo Fernández et al. (FERNáNDEZ; BELLA; DORRONSORO, 2022), a tarefa de detecção de *outliers* é usualmente abordada de maneira não-supervisionada, visto que não há conhecimento preliminar sobre o padrão que gerou estas observações. Em geral, estes algortimos fornecem uma função que quantifica o desvio de cada instância em relação à estrutura mais comportada da amostra, gerando um *score*. Esse, por sua vez, pode ser utilizado posteriormente, a exemplo de atuar como variável explicativa contínua para outros modelos preditivos; ou pode ser binarizado por um limiar com a mesma finalidade, marcando observações como *outliers* ou *inliers*.

Sem a presença de rótulos, a busca do conjunto de hiper-parâmetros que maximiza o desempenho da detecção é dificultado. Por outro lado, em aplicações em que há conhecimento sobre os padrões que contribuem para geração de observações anômalas, o problema pode ser trabalhado de maneira supervisionada, comumente como uma classificação. Assim como a abordagem não-supervisionada, a tarefa é utilizada como pré-processamento para outras atividades na sequência. Nessas circunstâncias, os hiper-parâmetros da detecção de *outliers* e do algoritmo aplicado na sequência podem ser ajustados simultaneamente, de maneira a obter vantagem da combinação dos modelos, conforme recomendado por Fernández et al. (FERNáNDEZ; BELLA; DORRONSORO, 2022). Dessa maneira, o subconjunto de hiper-parâmetros encontrado para ambos algoritmos minimizará o erro para a tarefa final, ao invés de minimizar o erro para a tarefa intermediária de detecção de anomalias.

3.10 Conclusão

Os conceitos apresentados no referencial teórico trataram acerca de técnicas de aprendizado de máquina que serão aplicadas para a tarefa de predição de preços, abordando suas vantagens e limitações, conforme a proposta deste trabalho. Foram apresentadas, também, técnicas para busca de hiper-parâmetros e seleção de features, assim como considerações a respeito de detecção de *outliers*.

4 METODOLOGIA E PROPOSTA DE DESENVOLVIMENTO

Neste capítulo, a solução proposta para o problema de predição de preços é apresentada, com enfoque nas etapas de treinamento e parametrização de modelos.

4.1 Seleção de Variáveis

Existem métodos (a exemplo de árvores de decisão e redes neurais) que conseguem realizar a seleção de variáveis automaticamente, porém executar a a tarefa anteriormente à execução do algoritmo promove ganho de velocidade e redução de esforços computacionais. Modelos baseados em regressão linear também conseguem executar a seleção de variáveis, mas através da redução dos coeficientes associados à cada variável. O algoritmo RANSAC, descrito em sessões anteriores, é um exemplo de método dessa categoria.

Ainda que esses algoritmos tenham capacidade de executar a seleção de variáveis, a existência destas em demasia pode desfavorecer o processo. Conhecido como "Maldição da Dimensionalidade", este efeito é amplamente estudado por diversas áreas da matemática, de acordo com (DONOHO, 2000), que explora seus aspectos negativos e positivos. Uma base de dados diagnosticadamente ruidosa, esparsa ou abundante em *outliers*, é um fator que agrava essa situação. Nesse contexto, a utilização de métodos como o PCA pode acelerar o processo de seleção de variáveis, eliminando preditores redundantes. No presente trabalho, todas as aplicações do método RANSAC foram precedidas por algoritmos de PCA. Em uma das abordagens, utilizamos algoritmos genéticos para a realização da seleção de variáveis, adicionalmente ao ajuste de hiperparâmetros.

4.2 Treinamento de Modelos

Três abordagens foram propostas para a solução do problema: metodologia baseada em *Stacking* de modelos; abordagem com modelos em cascata, otimizada por algoritmo genético; e abordagem baseada em redes neurais (MLP).

Os modelos escolhidos para as duas primeiras abordagens foram os mesmos, assim como o universo de busca de seus hiper-parâmetros, conforme exposto na Tabela 1. A distinção dos métodos se deu em como os modelos são aplicados. A abordagem baseada em redes neurais foi utilizada como comparativo aos demais métodos. Obviamente, seus hiper-parâmetros são distintos daqueles das demais abordagens.

4.2.1 Métrica de Avaliação dos Modelos

No contexto da aplicação de modelos preditivos em problemas de precificação, a métrica mais adequada para a medição de performance é o erro médio percentual absoluto (*Mean Absolut Percentage Error — MAPE*), por dois motivos principais. Primeiramente, preços são negociados e comparados dessa forma. É intuitivo para companhias que análises, medições, comparativos e descontos sejam abordados percentualmente. Em segundo

lugar, preços oscilam temporalmente e produtos diversos são negociados em valores com ordens de grandeza distintas. Uma mesma unidade de medida faz com que transações sejam analisadas atemporalmente, sem efeitos inflacionários e de maneira agnóstica em termos de produto. Dessa forma, ao invés de predizer o valor absoluto de venda do item, optou-se por modelar o valor de *markup*: determinado pela divisão direta do preço de venda pelo custo do produto (com frete e impostos). Essa modificação permite que efeitos de inflação sejam minimizados e todos os itens sejam modelados sob a mesma ordem de grandeza. Para obter o valor absoluto do item a ser vendido, basta multiplicar o *markup* pelo preço da última compra do item. Dado que não existem produtos com *markup* igual à zero, não há circunstâncias com ocorrência de erro na apuração da métrica. Sendo $\hat{y}_i$ a predição para a i-ésima observação e y_i o seu verdadeiro valor correspondente, o cálculo da métrica MAPE para n observações é dado por:

$$MAPE(y, \hat{y}) = \frac{1}{n} \sum_{i=0}^{n-1} \frac{|y_i - \hat{y}_i|}{y_i} \tag{1}$$

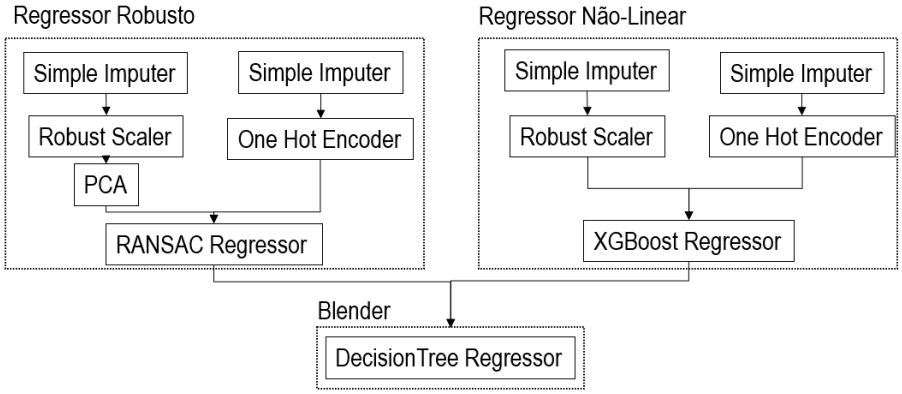
4.2.2 Abordagem baseada em Stacking de modelos

A abordagem baseada em *Stacking* explora positivamente diferenças entre algoritmos de naturezas diferentes, gerando um modelo final mais poderoso, conforme introduzido por (WOLPERT, 1992). Mais recentemente, os autores em (TRUONG et al., 2020) aplicaram a técnica para o contexto de predição de preços, na indústria imobiliária.

Em problemas onde a frequência de *outliers* é elevada, podem ser combinados um modelo robusto (Regressor RANSAC), precedido por PCA e um modelo não-linear (XGBoost). Antes do processamento pelos modelos, os dados devem ser tratados. Para este fim, são aplicadas ferramentas disponíveis na biblioteca Scikit-Learn: *Simple Imputer* para o preenchimento de valores faltosos; *Robust Scaler* para normalização com mediana igual à zero, adequado para conjuntos com alta ocorrência de outliers; e *One Hot Encoder* para criação de vetores binários para variáveis categóricas. O modelo definido como *Blender* foi uma árvore de decisão, atuando como mediador entre os demais algoritmos, aprendendo em qual ocasião cada um deve ser utilizado com maior veemência. A arquitetura pode ser observada (Figura 4).

Conforme comentado anteriormente, a técnica de *Random Search* foi utilizada para a parametrização dos modelos. A busca foi atualizada na medida que a função objetivo encontrou um valor menor ao anterior. A função objetivo é a mesma do algoritmo genético, a ser apresentada na próxima seção.

Figura 4 – Diagrama de blocos da abordagem baseada em *Stacking*



Fonte: Elaborado pelo Autor.

4.2.3 Abordagem com modelos em cascata, otimizada por algoritmo genético (AutoML)

Uma abordagem com modelos em cascata foi desenvolvida, inspirada em (RAMAKRISHNAN et al., 2019) e (CHEN et al., 2018), que apresentam detecções de anomalia em sistemas de precificação e treinamento automatizado de modelos preditivos através de algoritmos evolutivos, respectivamente. Conforme exposto por (BURKOV, 2020), modelos em cascata devem ser idealmente ajustados simultaneamente, portanto um algoritmo genético foi aplicado para a otimização de todo o *pipeline* de dados.

A proposta contempla diferentes tarefas de aprendizado de máquina, fazendo com que existam duas possibilidades para as observações: predição do preço por um regressor não-linear; ou entendimento pelo sistema que se trata de um *outlier*, não sendo fornecida uma predição e sim a condução para um especialista humano proceder com a precificação manualmente. A definição de qual procedimento será executado é feita pelo processo de busca do algoritmo genético.

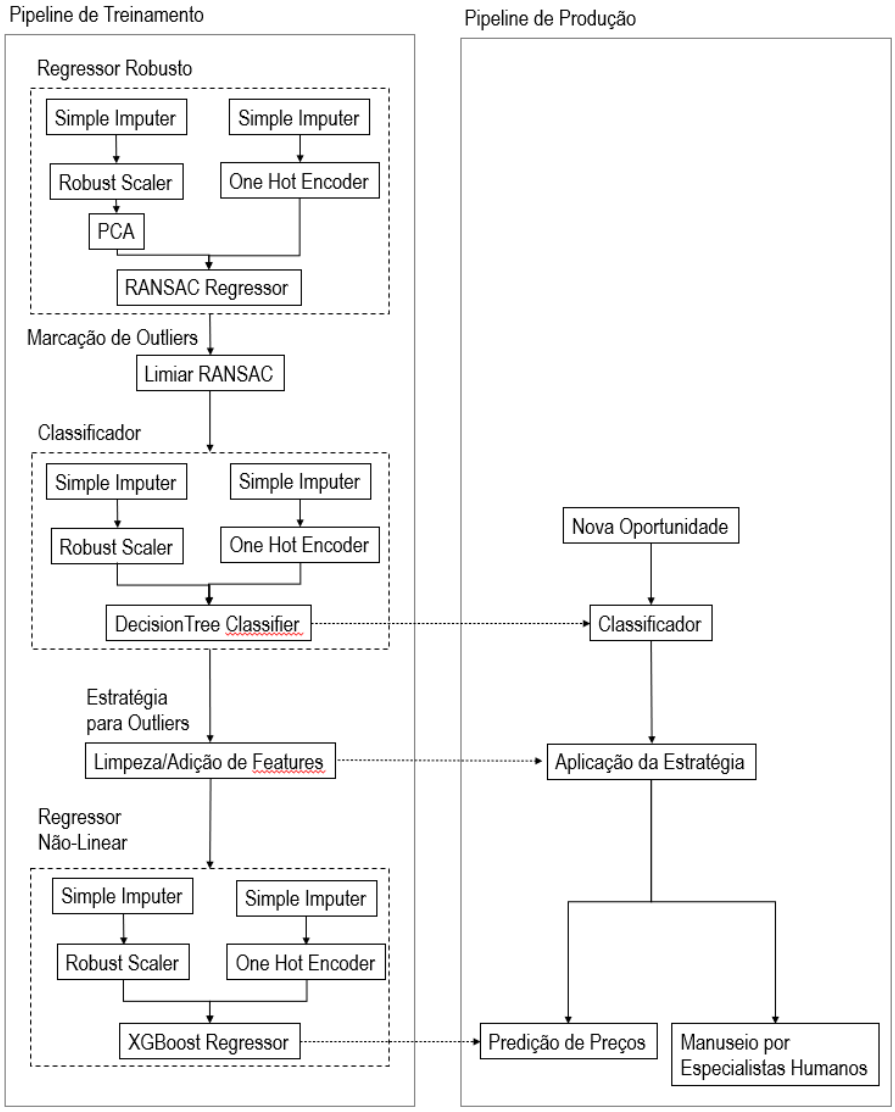
O fluxo na função objetivo do algoritmo genético segue conforme diagrama (Figura 5). É importante destacar que a proposta de solução é dividida em dois grandes blocos, sendo o *pipeline* de treinamento, onde os modelos são parametrizados; e o *pipeline* de produção, em que usuários especialistas humanos interagem com a solução construída anteriormente. Na sequência, as etapas serão detalhadas.

- **Individuo Genético:** Todos os hiper-parâmetros relevantes para os modelos preditivos e parâmetros adicionais ao fluxo de processamento dos dados estão contidos no indivíduo do algoritmo genético. O espaço de busca dos hiper-parâmetros é o mesmo utilizado para a abordagem baseada em *Stacking*. A configuração dos indivíduos genéticos é apresentada na Tabela 2.

- **Criação dos conjuntos de dados:** treino, validação e teste.

- **PCA + Regressão Robusta (RANSAC):** Treinamento do método de PCA,

Figura 5 – Diagrama de blocos da abordagem com modelos em cascata, otimizada por algoritmo genético



Fonte: Elaborado pelo Autor.

seguido da regressão robusta. Na sequência, o modelo de regressão fornece predições para o conjunto de treinamento. Essas predições são divididas pelos rótulos originais. Se a razão for maior que o limiar (limiar RANSAC) definido anteriormente, esse ponto de dados é marcado como *outlier* e vice-versa. Essa marcação indica que a regressão robusta não conseguiu predizer com precisão qual seria o preço para aquela transação e ela será endereçada de maneira específica, conforme será apresentado a seguir.

- **Treinamento de Classificador:** Após das marcações realizadas na etapa anterior, um algoritmo de árvore de decisão - com seus hiper-parâmetros definidos pelo indivíduo genético - é treinado para classificar se os indivíduos são ou não *outliers*. A depender da estratégia de tratamento de *outliers*, esse classificador poderá uti-

lizado em produção para direcionar oportunidades para precificação pelo regressor não-linear ou por um especialista humano. Uma função busca encontrar o ponto de corte ótimo entre as classes.

- **Aplicação de estratégia para *outliers*:**
  - **Estratégia *Score*:** Nessa estratégia, o classificador fornece *scores*, a serem utilizados como variável de treinamento do regressor não-linear. Em produção, todas as predições serão realizadas pelo modelo não-linear;
  - **Estratégia de Marcação Binária:** Nessa estratégia, o classificador fornece predições binárias, a serem utilizados como variável de treinamento do regressor não-linear. Da mesma forma que a abordagem anterior, todas as predições serão realizadas pelo modelo não-linear em ambiente de produção;
  - **Estratégia de Eliminação:** De maneira distinta às demais estratégias, nessa abordagem indivíduos preditos como *outliers* pelo classificador serão eliminados da base de treinamento do regressor não-linear. Seguindo a mesma lógica, os pontos de dados classificados como pertencentes à classe positiva serão direcionados a especialistas humanos para precificação em ambiente de produção;

- **Criação de conjunto de dados conforme estratégia de *outliers* para regressor não-linear:**

  Conforme a aplicação da estratégia para *outliers*, um novo conjunto de dados para o treinamento do regressor não-linear é criado, seja este com variáveis adicionais (predições do classificador, nas estratégias *Score* ou Marcação Binária) ou limpo, sem as amostras anômalas na base. Caso a estratégia seja a de Eliminação, a quantidade de indivíduos descartados será armazenada.

- **Treinamento de regressor não-linear & predição em dados de validação:** O treinamento do algoritmo de regressão não-linear (*XGBoost*) é executado na base de treinamento dedicada. Seus hiper-parâmetros foram definidos pelo indivíduo genético. Na sequência, este algoritmo é utilizado para a predição na base de validação. Sua performance em ambos conjuntos é registrada.

- **Apuração da função objetivo e apuração em dados de teste:** Finalmente, a função objetivo do algoritmo genético é computada e associada ao indivíduo, permitindo com que o processo evolutivo continue. Adicionalmente, as métricas de performance do regressor não-linear são computadas para o conjunto de teste.

A função objetivo *loss* foi customizada para considerar nuances do encadeamento dos modelos. Ela visa minimizar o erro de todo o sistema minimizando efeitos de sobreajuste. Para isso, são componentes da função objetivo:

- Performances t e v, representando o desempenho de regressão nos conjuntos de treino e validação, penalizadas pela quantidade q de observações descartadas em cada conjunto respectivamente, se a estratégia para *outliers* for a Estratégia de Eliminação. Um peso a atua para normalizar a ordem de grandeza entre os diferentes fatores. Este foi ajustado para 0.00001 na presente circunstância.

$$t = MAPE(y_{\text{train}}, \hat{y}_{\text{train}}) + aq_{\text{train}} \tag{2}$$

$$v = MAPE(y_{\text{val}}, \hat{y}_{\text{val}})) + aq_{\text{val}} \tag{3}$$

- Uma média m entre as performances t e v, fazendo com que o algoritmo favoreça a performance nos dados de validação sem que este seja sobreajustado. O valor de v recebe um multiplicador p para que a performance na generalização seja favorecida. Na presente ocasião, este fator foi ajustado para 2.

$$m = \frac{t + pv}{2} \tag{4}$$

- Uma penalização pela diferença d da performance entre os conjuntos, novamente favorecendo soluções que tenham o mínimo de sobreajuste e generalização. O cálculo de d contempla, também, a diferença nos valores de ROC AUC do classificador, também para combate à fenômenos de sobreajuste para este modelo.

$$d = t(1 - ROC\_AUC_{\text{train}}) - v(1 - ROC\_AUC_{\text{val}}) \tag{5}$$

- Uma regularização pela quantidade f de variáveis utilizadas, visando a geração de um modelo com menos preditores;

  Finalmente, a formulação da função objetivo é dada por:

$$loss = m + |d - 1| + \frac{fm + |d - 1|}{100} \tag{6}$$

4.2.4 Abordagem baseada em Redes Neurais (MLPs)

De maneira análoga aos outros métodos, redes neurais profundas (MLPs) foram aplicadas ao conjunto de dados. A lógica por trás dessa abordagem é trazer métodos de *Deep Learning* se comportam em conjuntos de dados com abundância de *outliers*. Para aplicações a partir de dados tabulares, a rede *Multi-Layer Perceptron* é suficiente; camadas e operações especializadas, como convoluções ou camadas recorrentes não se fazem necessárias. A utilização de métodos para seleção de variáveis também foi descartada, dado que redes neurais conseguem executar esse procedimento internamente através da

redução do peso de *features* menos relevantes. Além disso, parâmetros como o *Dropout* fomentam a regularização dos modelos, através da redução da dependência de neurônios específicos para a tarefa de predição. Os parâmetros do modelo foram ajustados por *Random Search*, sendo a busca atualizada enquanto a função objetivo encontra um valor menor ao anterior. A função objetivo da busca de parâmetros é a mesma da abordagem com algoritmo genético.

4.3 Conclusão

Neste capítulo, foram apresentados componentes para a tarefa de predição de preços, a exemplo de métodos para seleção de variáveis e da métrica de avaliação dos modelos. Destaca-se aqui, também, a definição do rótulo dos modelos de regressão, o *markup*. Através deste, os modelos podem fornecer predições para produtos de diferentes faixas de preço, sem a interferência de efeitos inflacionários. Foram descritos, também, três abordagens de treinamento: *Stacking*; modelos em cascata, otimizada por algoritmo genético; e redes neurais. A arquitetura de cada abordagem foi apresentada, assim como detalhes do tratamento de dados e o espaço de busca dos hiper-parâmetros de cada modelo. Para o caso da abordagem com otimização via algoritmo genético, a configuração dos indivíduos, assim as variáveis de decisão e função objetivo foram detalhadas.

Tabela 1 – Espaços de Busca dos Hiper-Parâmetros

Algoritmo	Parâmetro	Intervalo de Busca
PCA	n_components	(0.3, 0.99)
RANSAC	min_samples	(0.2, 1)
Árvore de Decisão	max_features	(0.1, 1)
Árvore de Decisão	max_depth	(3, 10)
XGBoost	n_estimators	(20, 200)
XGBoost	max_depth	(4, 32)
XGBoost	learning_rate	(0.0001, 0.5)
XGBoost	gamma	(0, 0.1)
Rede Neural (MLP)	Função de Ativação	(ReLU, Elu)
Rede Neural (MLP)	Número de Camadas Ocultas	(1, 6)
Rede Neural (MLP)	Número de Neurônios por Camada Oculta	(5, 100)
Rede Neural (MLP)	Taxa de Aprendizado	(0.00001, 0.03)
Rede Neural (MLP)	Otimizador	(RMSProp, Adam, SGD)
Rede Neural (MLP)	Taxa de Dropout	(0, 0.3)

Tabela 2 – *Encoding* do Indivíduo Genético

Aplicação	Parâmetro	Número de Bits
Regressor RANSAC	min_samples	3
Regressor RANSAC	Transformação do Target	1
Árvore de Decisão	max_features	3
Árvore de Decisão	max_depth	3
XGBoost	n_estimators	3
XGBoost	max_depth	3
XGBoost	learning_rate	3
XGBoost	gamma	3
XGBoost	Transformação do Target	1
PCA	n_components	3
Estratégia para *Outliers*		2
Limiar RANSAC		3
Seleção de features		1 para cada feature do conjunto de dados

5 ESTUDO DE CASO

Neste capítulo, um ambiente de estudo é apresentado. São detalhados aspectos dos produtos de uma companhia de matéria-prima, seus processos comerciais, como estes impactam a geração e coleta de dados e, consequentemente, a precificação.

5.1 Produtos Semi-Acabados: Metais Não-Ferrosos & Plásticos Industriais

Commodities são itens básicos com características uniformes, independente do seu produtor. Petróleo, grãos, gases e metais são alguns exemplos. Caso estes itens sejam utilizados como entrada ou sejam consumidos na produção de outros produtos ou serviços, são denominados bens intermediários ou semi-acabados. Metais não-ferrosos (cobre, latão, bronze, aço inoxidável, alumínio), e plásticos industriais de engenharia (poliamida, poliacetal, polietileno, etc.) fazem parte dessa categoria (Figura 6). Ambas categorias de produtos fazem parte da cadeia produtiva de diversas indústrias, como automobilística, civil, eletromecânica, moveleira, médica, alimentícia. Estes itens também são aplicados em processos de manutenção industrial, consumidos em atividades correlatas à produção, como em componentes e sobressalentes de equipamentos.

Figura 6 – Metais Não-Ferrosos e Plásticos Industriais

(a) Perfis, Barras Chatas e Tubos de Alumínio

(b) Barras Redondas de Latão

(c) Barras e Tubos Redondos de Nylon e Teflon

Fonte: Acervo do Autor.

Metais não-ferrosos são amplamente aplicados na construção mecânica, para a fabricação de eixos, calços, mancais, polias, parafusos, pinos, engrenagens, anéis e muito mais. Os itens mais comuns para essa aplicação são barras de bronze, aço inox, alumínio e latão. Bronze e aço inox são ofertados em diferentes ligas, que refletem em sua pureza, flexibilidade e resistência a esforços mecânicos. No caso do aço inox, por exemplo, existem ligas específicas para aplicações alimentares e hospitalares, por serem atóxicas. Outro metal não-ferroso bastante aplicado na indústria é o cobre. Por suas propriedades condutoras, o cobre é amplamente utilizado para a fabricação de equipamentos para energia elétrica.

Todos esses itens são fabricados em barras com secções redondas, quadradas, retangulares e sextavadas. A maioria destes itens é fabricada em comprimentos de 6000mm, com exceção do bronze e do latão, fabricados nos comprimentos de 500mm e 3000mm, respectivamente. Essa diferença se deve ao seu processo produtivo. O bronze, por exemplo, pode gerar fissuras internas (popularmente conhecidas como "brocas") se fabricado em comprimentos maiores que 700mm. Quanto menor a barra que será cortada, maior a possibilidade de geração de pontas, significando perda no processo. Normalmente, a equipe comercial calcula as perdas teoricamente, repassando-as aos clientes. Barras chatas muitas vezes são dobradas e/ou furadas por especificação dos compradores.

Outros formatos que metais não ferrosos são fabricados em tubos, chapas e arames. Tubos, principalmente de aço inox, são itens com uma extensa gama de especificações. Por exemplo, caso estes sejam fabricados através de chapas calandradas e subsequentemente soldadas, são mais baratos do que aqueles fabricados sem a solda. Além disso, quanto mais fina a sua parede, mais complexa é a sua fabricação e, como consequência, seu preço de mercado. Chapas são fabricadas desde espessuras finas (micrômetros), utilizadas como calços até chapas grossas (na ordem de centímetros) para usinagem de grandes peças. O corte de chapas geralmente é feito em formatos retangulares, mas também é possível ser cortado em círculos ou anéis. Essas últimas configurações geram muitas perdas de materiais, incluídas no custo. Peças alumínio também são fabricadas nos formatos de perfis, a partir do processo de extrusão. São largamente aplicados no mercado de construção civil e arquitetura. Esses itens são frequentemente pintados. Todas as modificações listadas acima geram novas informações à transação comercial. Como se tratam de uma gama extensa de atributos, os dados sofrem ganho de dimensionalidade em variáveis categóricas.

Plásticos industriais não possuem tanta cardinalidade em termos de modificações e ligas. São limitados aos cortes em chapas, tubos e barras. Suas aplicações são análogas aos itens de metal. Itens como polietileno e teflon são aplicados na indústria alimentícia devido à sua atoxidade e poliuretano é aplicado na indústria automobilística.

A produção desses itens em larga escala é padronizada em tecnicamente, garantindo uniformidade dimensional e resistência mecânica. Visto que a quantidade de aplicações desses itens é extensa, uma oportunidade de mercado se abre para empresas que compram esses itens em grandes volumes no atacado (diretamente de fábricas, importadores ou dis-

tribuidores) e os revendem no varejo, customizando os produtos conforme a especificação dos clientes (em sua maioria corte, mas também dobra, pintura e tratamentos térmicos). Dado que estes itens são customizados para cada aplicação, a venda de produtos idênticos é extremamente rara. Esse fator, juntamente à flutuação de índices macroeconômicos, consumo de insumos, mão-de-obra e serviços dedicados para aquele fornecimento, permite com que a precificação seja específica para transação. Visto que os produtos aqui estudados possuem forte perfil de commodity, o preço se torna um dos principais fatores na decisão de compra pelos consumidores. Dessa forma, explorá-lo com o apoio de métodos quantitativos pode gerar resultados positivos às organizações em margem e receita.

5.2 Processos de Comercialização de Itens Semi-Acabados

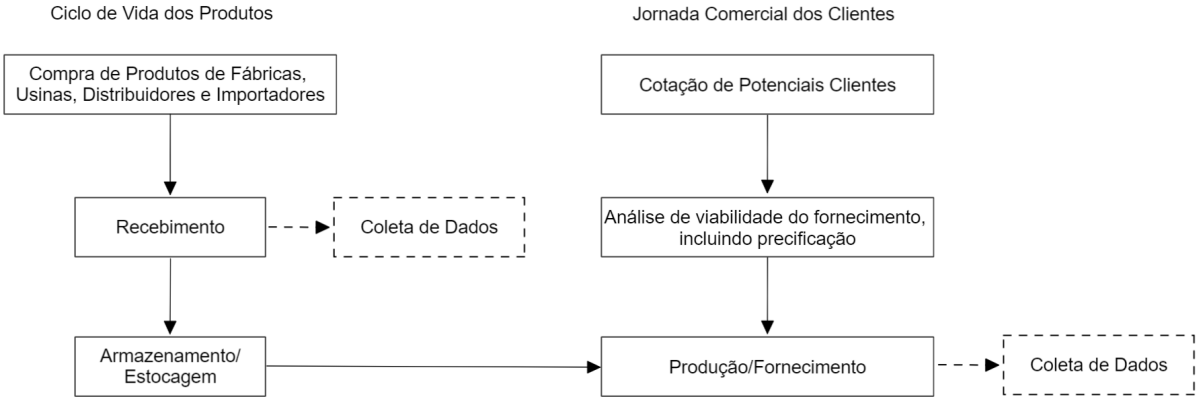
O processo comercial em uma companhia de semi-acabados se inicia quando um potencial cliente envia uma requisição de cotação, com o intuito de receber uma proposta comercial. A equipe comercial procede com uma análise de viabilidade do fornecimento, verificando quais são valores de mercado para a compra do item no atacado, cálculo de impostos, fretes e custos de produção. O relacionamento comercial com o potencial cliente também é considerado qualitativamente. Por exemplo, há clientes que enviam requisições de cotação para diversos fornecedores e possuem maior poder de negociação. Outros clientes atuam tanto como compradores quanto como orçamentistas para sua estrutura comercial. Dessa forma, esses clientes possuem taxas de conversão reduzidas e desperdiçam recursos profissionais em demandas com baixa probabilidade de compra. Há clientes que precisam dos produtos com urgência, se dispondo a pagar mais pelo bem para ter prioridade no atendimento e entrega por parte do fornecedor, muitas vezes buscando o reparo de alguma emergência em sua cadeia produtiva.

Os clientes também podem ser caracterizados pelo seu perfil de pagamento, seja por restrições de crédito por comportamentos de inadimplência no passado, como pelo perfil contrário, de pagamento à vista. Transações com clientes com perfil de inadimplência possuem riscos maiores associados, sendo este transferido ao preço. Levando-se em consideração todos os fatores elencados anteriormente, uma margem de lucro é adicionada aos custos visando a conversão. A presença de todos esses fatores faz com que estimar uma curva de demanda para os produtos seja uma tarefa complexa.

A coleta de dados acompanha dois processos: o ciclo de vida dos produtos, a partir da compra destes em grandes volumes para estoque e a jornada comercial dos clientes (Figura 7). Infelizmente, essa abordagem possui limitações que constantemente resulta em problemas.

Em primeiro lugar, o preço de compra dos itens é consolidado nos sistemas de gerenciamento apenas quando o item é adquirido. Dessa forma, se um item está estocado por um longo período, seu preço provavelmente estará defasado. Esse fator é parcialmente mitigado durante as operações comerciais, visto que a equipe de venda coleta essas infor-

Figura 7 – Ciclo de Vida dos Produtos & Jornada Comercial dos Clientes



Fonte: Elaborado pelo Autor.

mações com frequência através de e-mails ou telefonemas, mas essas informações não são sistematizadas caso a compra do item não se concretize por complexidade operacional. Como consequência, o volume de dados das tabelas relacionais com informações de compra é proporcional ao comportamento de crescimento de estoque. Períodos com restrição de custo geram menos compras e, consequentemente, uma menor coleta de dados. De maneira contraintuitiva, uma estratégia mais eficiente em suprimentos pode fazer com que organizações comprem com menos frequência, mas com mais precisão, evitando a geração de estoque indevido, mas comprometendo a frequência da coleta de dados.

Em segundo lugar, diferentes configurações de um mesmo item podem possuir a mesma quantidade por Unidade de Manutenção de Estoque (*Stock Keeping Unit* — SKU). Por exemplo, duas chapas de mesma espessura de um material podem possuir o mesmo peso, caso possuam largura e comprimento diferentes. Consequentemente, utilizar apenas a quantidade de venda como um preditor pode levar à erros. De maneira agravante, poucos detalhes são sistematizados após uma venda, de maneira com que a quantidade vendida é expressada unidimensionalmente, seja em quilograma, metro ou peça. Assim como outras etapas da coleta de dados apresentadas anteriormente, a criação de SKUs dedicados para cada item em diversas configurações é operacionalmente inviável, agravada pela explosão combinatória de SKUs que o processo geraria.

Pela complexidade operacional, é comum que os fluxos de coleta de dados sejam enxutos, sistematizando-se apenas informações das oportunidades que efetivamente converteram em vendas. Dessa forma, há uma grande parte de eventos que são censurados das bases de dados, o qual destaca-se aqui as oportunidades não convertidas. Esse fator impede a criação de curvas de elasticidade dos produtos, a medição de taxas de conversão e estimativas de preço com excesso de confiança ou cálculo sobredimensionado. Essas informações permitiriam a criação de modelos preditivos de classificação e o estabelecimento de curvas de demanda. Consequentemente, seria possível estimar — com o auxílio de técnicas de pesquisa operacional e otimização — qual preço que maximizaria a receita

esperada para dado item, respeitando a margem como restrição.

Finalmente, existem fatores que não são coletados através dos dados, como informações passadas via e-mail ou telefone, que possam indicar direções de negociação, como o preço de concorrentes. Além disso, as transações sistematizadas são realizadas após a venda, registrando sua configuração final, sem a informação de descontos. Como consequência desses comportamentos, obtém-se um conjunto de dados com elevada presença de *outliers*.

5.3 Materiais e Tratamento dos Dados

Foram coletados dados de onze tabelas do banco de dados relacional de uma empresa do ramo de metais não-ferrosos e plásticos industriais, agregadas em um único *dataset*. Cada uma das bases foi tratada individualmente, permitindo a criação de preditores para a variável alvo. Após a remoção de pontos de dados com preços de venda igual a zero ou margem negativa, o *dataset* possui cerca de 19,5 mil pontos de dados e apresenta o histórico de transações entre 2013 e o início de 2021. Os preditores podem ser classificados em uma das seguintes categorias:

- **Variáveis dos Produtos:**
  - **Categóricas:** material, configuração, liga, acabamento, unidade de estoque (quilograma, metro ou peça);
  - **Numéricas:** diâmetro, largura, espessura, tamanho da aba principal, malha, fio, peso por metro ou peso por metro quadrado e comprimento. Com o intuito de permitir a entrada de diferentes produtos em um mesmo algoritmo, os vetores de variáveis de atributos dimensionais foram concatenados, conforme recomendado por Burkov (BURKOV, 2020). Na abordagem, categorias de produtos que não possuem determinada dimensão têm o seu valor alocado como 0. Por exemplo, chapas não possuem a dimensão de diâmetro, portanto essa variável possui valor 0.
  - **Variáveis construídas como binárias:** parede fina, tubos padrão *Schedule* e material milimetrado;
- **Custo dos Produtos e Informações de Fornecimento:**
  - **Numéricas:** quantidade adquirida na última compra, coordenadas geográficas do último fornecedor (a localização impacta diretamente valores de frete), a diferença entre a data da venda e a data da compra (com o intuito de indicar o quão recente é o preço), custo médio de itens do mesmo material e categoria, assim como valores de custo médio em diferentes intervalos de tempo de itens similares (essa abordagem visa minimizar o impacto de preços antigos em itens com menor movimentação de estoque);

- **Informações referentes à transação:**
  - **Numéricas:** quantidade adquirida, comprimento total adquirido, coordenadas geográficas do destino, número de itens do mesmo material e configuração na compra. Esse último fator é considerado pelo seguinte: caso algum item se esgote durante o fornecimento, é uma prática da empresa ofertar uma medida acima para garantia da oportunidade. Posteriormente, durante o processo de usinagem do cliente, este reduz a peça para a medida desejada. Dessa forma, itens similares podem ter margens diferentes em um mesmo fornecimento;
  - **Variáveis construídas como binárias:** venda para matriz ou filial do cliente;
- **Comportamento histórico de pagamento do cliente:**
  - **Numéricas**: valores agregados em diferentes intervalos de tempo de valor devido, valor gasto, pontualidade de pagamentos;

5.4 Análise Exploratória de Dados

Ao se cruzarem informações de material, liga e configuração, observa-se que há cerca de 100 combinações. Seus valores de especificações variam significativamente, em distribuição com comportamento de cauda longa, conforme mostrado na Tabela 3. Uma estratégia para compensar esse fator seria a criação de conjuntos de dados específicos para cada grupo, mas o resultado seria de amostras pequenas. Além disso, utilizar um único *dataset* permite com que os modelos aprendam características compartilhadas entre os grupos ou sutilezas entre compras de um mesmo cliente com diferentes produtos em sua cesta. Com o intuito de ilustrar esses efeitos, barras de cobre são geometricamente iguais a barras de aço inox e os processos de corte são análogos. Adicionalmente, essa abordagem permite com que os modelos treinados forneçam predições a itens recém-cadastrados no estoque, evento relativamente comum na companhia estudada e em convergência ao proposto por Bauer et al. (BAUER; JANNACH, 2018).

Tabela 3 – Faixas de Valores para Atributos dos Produtos (em milímetros)

Atributo	Mínimo	Máximo	Média
Diâmetro Externo	0	381	34.882
Diâmetro Interno	0	273.05	6.742
Largura	0	1250	8.515
Altura	0	1000	4.212
Peso/Unidade de Comprimento	0	809.08	21.055

Outro aspecto relevante do processo refletido nos dados é a assincronia entre eventos de compra e venda dos produtos. Entre essas datas, os itens são estocados. O tempo em que cada produto fica armazenado depende puramente de sua demanda, podendo variar

entre dias e anos. Por se tratarem de metais, não há perda de valor ao longo do tempo, a menos que o produto sofra alguma avaria (não mapeadas nos dados). Dessa forma, sob a ótica comercial, não há problema em um item permanecer muito tempo em estoque. Interessante ressaltar que a partir de 2018 a companhia estudada alterou sua política de compras e se tornou mais eficiente, conseguindo manter níveis de estoque saudáveis realizando menos aquisições. Consequentemente, pela redução na frequência, a coleta de dados dos custos de compra dos itens diminuiu.

Uma estratégia para mitigar esse efeito seria a criação de processos de preços diariamente ou semanalmente, mas essa execução é onerosa do ponto de vista operacional. Além disso, fornecedores podem se sentir incomodados com a excessiva coleta de informações para poucas conversões de venda, gerando desconfortos no relacionamento comercial. Para endereçar esses fatores como variáveis, variáveis externas são trabalhadas para atuar como idade do preço. Para melhorar a estimativa quando a idade do preço é muito antiga, o preço atual de itens similares (mesma matéria-prima e configuração) são utilizados.

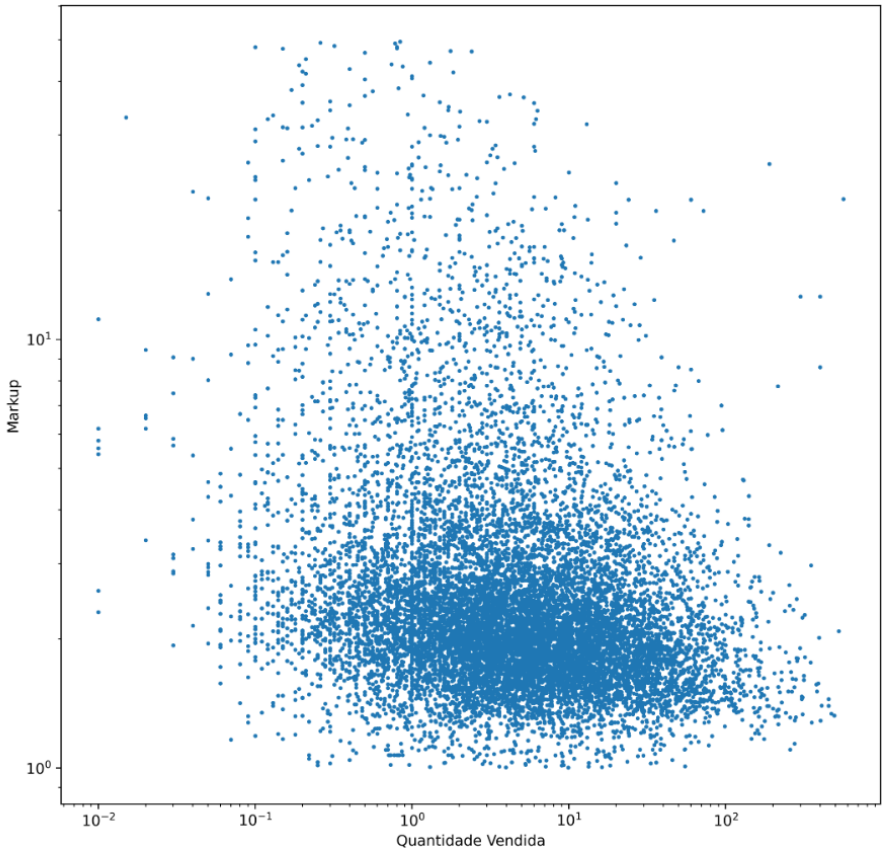
Obedecendo aos fatores de oferta e demanda, quanto menor a quantidade ofertada, maior o preço dos itens. Dessa forma, na medida em que os clientes compram peças cortadas em pequenas quantidades, maior a rentabilidade (*markup*) da operação. Isso acontece também pelo fato de que operações de corte possuem o mesmo esforço operacional, independente do comprimento das peças trabalhadas. Dessa forma, o preço dos itens cai em uma configuração similar a um decaimento exponencial, que pode ser visualizado (Figura 8). Dessa forma, a segunda transformação dos rótulos é para capturar esse efeito de não-linearidade.

Os preditores foram comparados ao valor da variável resposta através da correlação de *Spearman*. Optou-se pela utilização desta métrica uma vez que o conjunto de dados contém diversas categorias e configurações de produtos, fazendo com estes sejam negociados a patamares distintos de preço. Esse fator, associado a todas as outras variáveis categóricas, também traz não-linearidades ao conjunto de dados. A partir destas considerações, na Tabela 4, expõem-se as variáveis com maior correlação com a variável resposta.

5.5 Separação dos Conjuntos de Dados

Para a realização das operações de treinamento e validação dos modelos, o conjunto de dados foi separado em três subconjuntos: treinamento, validação e teste, com as proporções de 70%, 20% e 10%, respectivamente. O primeiro conjunto é utilizado para o treinamento dos modelos preditivos. O segundo é utilizado iterativamente junto ao primeiro, com o intuito de promover o ajuste de hiper-parâmetros. Esse processo deve ser executado com cautela, evitando o vazamento de informações de instantes de tempo futuros para instâncias do passado, melhorando artificialmente a performance do modelo. Na ocorrência dessa situação, ao disponibilizar os modelos para produção, é comum observar queda de performance na solução, visto que dados do futuro não estão disponíveis no ato

Figura 8 – Decaimento exponencial dos rótulos em relação à quantidade vendida



Fonte: Elaborado pelo Autor.

da predição. Adicionalmente, observar apenas a performance nos dados de validação pode gerar fenômenos de sobreajuste nesse conjunto, portanto uma comparação de desempenho dos conjuntos de treino e validação reduz essa possibilidade. O conjunto final — de teste — é utilizado apenas para avaliação da performance da solução em dados não antes observados pelos modelos, garantindo uma performance de qualidade em situações não antes vistas.

5.6 Tratamento & Transformação de Variáveis

Não foram apenas os rótulos que passaram por etapas de pré-processamento. Outras variáveis foram analisadas e tratadas para que a extração de valor dos dados seja mais efetiva pelas etapas de modelagem preditiva. A primeira dessas transformações se deu na variável de data, transformada do formato de calendário para o formato numérico. Aqui, destaca-se a importância da utilização do campo de data como uma variável: as transações ocorrem temporalmente e há um sentido lógico em dividir os conjuntos de treino, validação e teste respeitando a cronologia. Caso contrário, promover uma divisão com embaralhamento aleatório dos dados poderia transportar pontos de dados de pontos do futuro para o passado, causando vazamento de informações. Segundo (ZHENG; CA-

SARI, 2018), vazamento de dados significa que informações são erroneamente reveladas ao modelo, permitindo com que este adquira vantagens irrealistas no treinamento. Esse processo pode ocorrer de maneira sutil e ser dificilmente detectado.

Para garantia do funcionamento adequado dos modelos adotados junto ao pacote de *software* utilizado (biblioteca *Scikit-Learn* da linguagem *Python*) (PEDREGOSA et al., 2011), algumas ações foram tomadas. Por exemplo, amostras com dados numéricos faltosos tiveram seus valores substituídos pela média da variável no conjunto de dados de treinamento. Preditores que, durante a engenharia de variáveis passaram por alguma divisão por zero, a exemplo do montante pago em relação ao vendido para clientes em sua primeira compra, foram substituídos por valores fora de sua distribuição de dados original, com o intuito de capturar esse padrão, sendo -1 ou -99999, conforme prática de engenharia de variáveis sugerida por (BURKOV, 2020).

Amostras onde o valor de venda era menor que o valor de compra do produto foram excluídas do *dataset*, visto que estes são exemplos de prejuízo, sendo este um comportamento indesejável para o aprendizado do modelo. Todas as transações comerciais são orçadas com o intuito de geração de lucro. Caso o fornecimento seja inviável sob essa ótica, o produto não é ofertado. Entende-se, portanto, que essas amostras tratam-se de erro humano, seja na inserção dos dados ou na transação comercial em si.

Dados numéricos foram normalizados para mediana zero, conforme o objeto *Robust Scaler* da biblioteca *Scikit-Learn*, adequado para conjuntos com alta ocorrência de *outliers*. Variáveis categóricas foram processadas com a técnica de *One Hot Encoding*. A técnica cria vetores binários para cada categoria. O aumento da dimensionalidade é prejudicial para algoritmos de aprendizado, portanto uma seleção de variáveis eficiente pode se fazer necessária, como será visto adiante.

Tabela 4 – Principais Correlações

Variável	Correlação Spearman com a Variável Resposta
Valor Nominal Devido Acum	-0.996
Valor Nominal Pago Acum	-0.995
Valor Nominal Devido D365	-0.995
Valor Nominal Vencido Acum	-0.995
Valor Nominal Vencido D365	-0.994
Valor Nominal Pago D365	-0.994
Valor Nominal Devido D180	-0.992
Pessoa Jurídica	-0.991
Valor Nominal Vencido D180	-0.990
Valor Nominal Pago D180	-0.990
Soma Pagto Dentro Prazo Aberto Acum	-0.984
Quantidade Vendas Configuração Cliente	-0.984
Quantidade Vendas Item Cliente	-0.982
Valor Nominal Pago D60	-0.981
Valor Nominal Vencido D60	-0.980
Quantidade Vendas Material Cliente	-0.979
Contagem Pagto Dentro Prazo Aberto Acum	-0.978

6 EXPERIMENTOS & ANÁLISE DE RESULTADOS

Os valores de hiper-parâmetros que ofereceram melhores resultados são apresentados na Tabela 5, assim como os resultados da função objetivo (equação 6). Métricas de avaliação dos modelos nos conjuntos de dados são apresentados na Tabela 6. Os conjuntos de dados representam os mesmos indivíduos; apenas no caso da abordagem baseada em algoritmo genéticos que os dados foram enriquecidos com o auxílio do classificador, gerando uma *feature* auxiliar, o *Score* de propensão a *Outlier*.

Tabela 5 – Hiper-parâmetros encontrados & valores da função objetivo

Abordagem	Algoritmo	Parâmetro	Valor Encontrado
Stacking	PCA	n_components	0.25
	RANSAC	min_samples	0.37
	RANSAC	Transformação dos Rótulos	True
	Árvore de Decisão	max_features	0.22
	Árvore de Decisão	max_depth	9
	Árvore de Decisão	Transformação dos Rótulos	False
	XGBoost	n_estimators	35
	XGBoost	max_depth	17
	XGBoost	learning_rate	0.229
	XGBoost	gamma	0.1
	XGBoost	Transformação dos Rótulos	True
		Função Objetivo	3.88
AG/Modelos em Cascata	PCA	n_components	0.792
	RANSAC	min_samples	0.428
	RANSAC	Transformação dos Rótulos	True
	Árvore de Decisão	max_features	0.742
	Árvore de Decisão	max_depth	6
	Árvore de Decisão	Transformação dos Rótulos	True
	XGBoost	n_estimators	200
	XGBoost	max_depth	20
	XGBoost	learning_rate	0.005
	XGBoost	gamma	0.057
	XGBoost	Transformação dos Rótulos	True
		Função Objetivo	2.023
Redes Neurais	MLP	Função de Ativação	Elu
	MLP	Número de Camadas Ocultas	6
	MLP	Número de Neurônios por Camada Oculta	86
	MLP	Taxa de Aprendizado	0.001
	MLP	Otimizador	RMSProp
	MLP	Taxa de Dropout	0.25
		Função Objetivo	3.963

6.1 Abordagem baseada em *Stacking* de modelos

A busca aleatória para a abordagem de *Stacking* foi realizada em 200 iterações. A variância explicada pelas variáveis mantida pela redução de dimensionalidade do método PCA foi de 25%.

As performances de todos os modelos foram afetadas por fenômenos de sobreajuste, como é possível atestar na degradação de desempenho com a aplicação dos algoritmos em conjuntos de dados distintos. A distribuição das predições em relação aos rótulos para cada conjunto de dados é apresentada (Figura 9). É possível observar que a distribuição tende a valores à direita, o que indica que os modelos sobre-estimam valores dos rótulos. Esse fenômeno é compreensível pela presença de *outliers* e é insatisfatório, visto que predições mais altas que os rótulos não garantem a conversão. O fenômeno oposto seria melhor para as operações da companhia. O compartimento de estouro superior do histograma foi limitado para o percentil 99.

6.2 Abordagem com modelos em cascata, otimizada por algoritmo genético (***AutoML***)

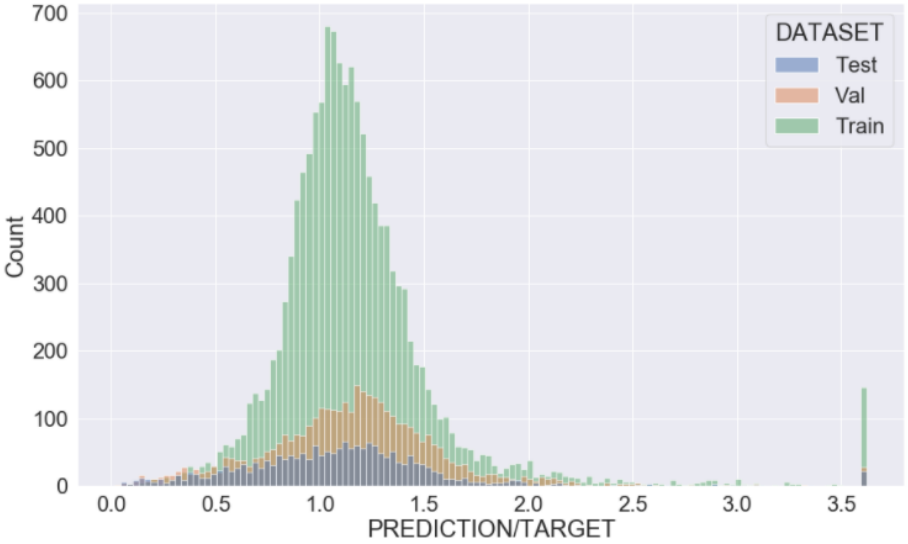
A busca através do algoritmo genético foi realizada com 100 indivíduos, por 50 gerações. O método de seleção utilizado foi o torneio, a probabilidade de cruzamento em um ponto foi de 60% e a probabilidade de mutação para cada *bit* foi ajustada para 10%. 15% dos melhores indivíduos foram mantidos a cada geração através do processo de elitismo. A evolução pode ser observada (Figura 10). Nota-se que o algoritmo genético trouxe melhorias durante quase todo o processo evolutivo, com um pequeno platô entre as gerações 23 e 31.

Diferentemente da abordagem baseada em *Stacking*, outros fatores e parâmetros foram definidos pelo algoritmo genético. Seus valores são apresentados na tabela 7. O algoritmo promoveu a seleção de variáveis, obtendo sua melhor performance com 46 delas. A variância explicada pelas variáveis mantida pela redução de dimensionalidade do método

Tabela 6 – Resultados

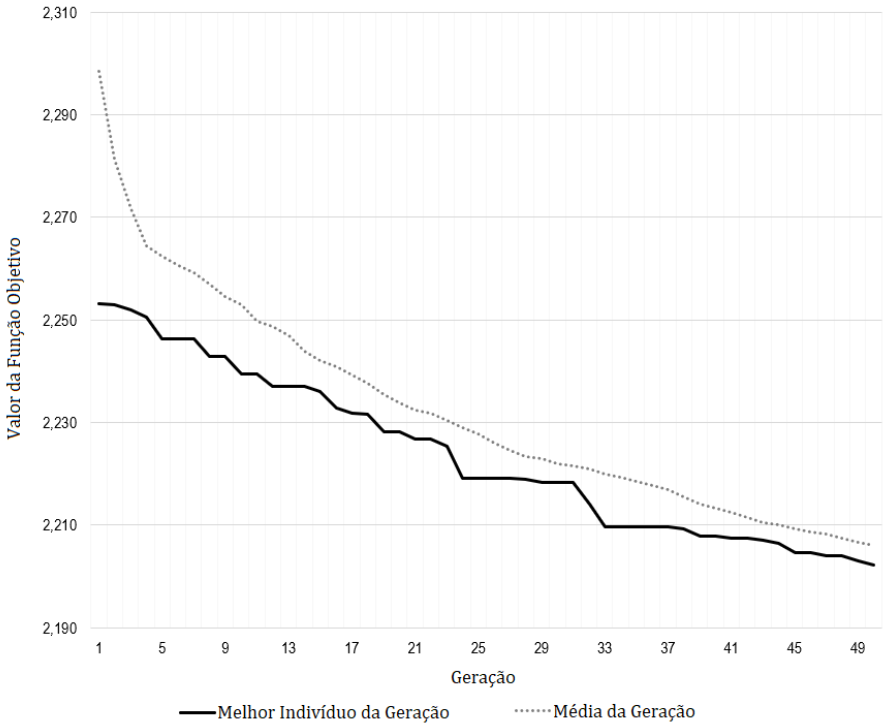
Conjunto de Dados	Abordagem/Modelo	MAPE
Treinamento	Stacking	77.610%
	Redes Neurais (MLP)	25.167%
	AG/Regressor Não-Linear	20.902%
Validação	Stacking	37.917%
	Redes Neurais (MLP)	26.324%
	AG/Regressor Não-Linear	26.060%
Teste	Stacking	200.541%
	Redes Neurais (MLP)	29.592%
	AG/Regressor Não-Linear	29.466%

Figura 9 – Histograma das razões entre predições e rótulos para a abordagem baseada em *Stacking*



Fonte: Elaborado pelo Autor.

Figura 10 – Evolução da população através das gerações e função objetivo



Fonte: Elaborado pelo Autor.

PCA foi de 79,2%.

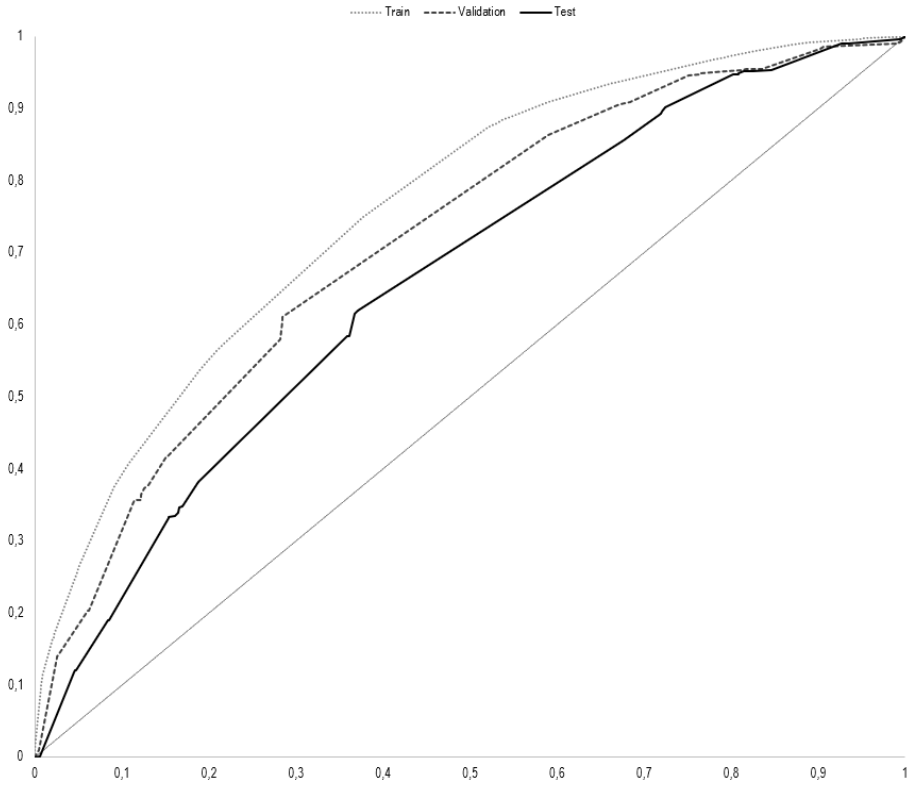
A performance do classificador é apresentada na curva *Receiver Operating Characteristic (ROC)* (Figura 11) e na tabela 8, onde é possível observar a presença de sobreajuste. Apesar desse fenômeno, é possível visualizar que o classificador conseguiu detectar padrões melhores que uma abordagem aleatória, capazes de indicar a anomalia de uma transação.

Essa anomalia, por sua vez, pode ser compreendida como uma oportunidade de aumento de preço. A estratégia encontrada pelo algoritmo genético para o tratamento de *outliers* foi a estratégia por *Scores*. Isso significa que manter observações com *outliers* na base de treinamento é benéfica para as tarefas de regressão, desde que estas sejam enriquecidas com a *feature* de *score*, oriunda do classificador.

Tabela 7 – Parâmetros adicionais definidos pelo Algoritmo Genético.

Parâmetro	Valor
Estratégia Para Processamento de *Outliers*	Score
Limiar RANSAC	1.1
Quantidade de Features	46

Figura 11 – ROC AUC Classificador

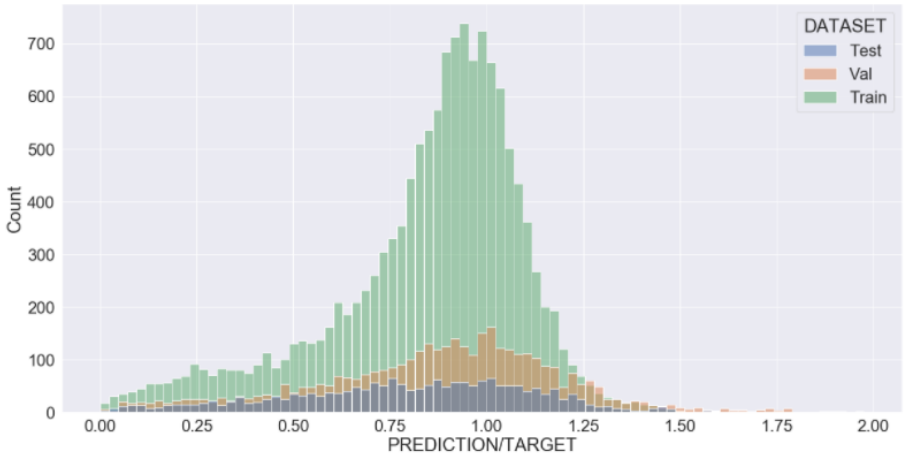


Fonte: Elaborado pelo Autor.

Assim como na abordagem utilizando *Stacking* mas de maneira substancialmente sutil, a performance de todos os modelos foram afetadas por fenômenos de sobreajuste, (Figura 12), que apresenta a distribuição das predições em relação aos rótulos para cada conjunto de dados. Diferentemente da abordagem baseada em *Stacking*, é possível observar que a distribuição tende a valores à esquerda, o que indica que os modelos subestimam o valor do *target*. Esse fenômeno é compreensível pela presença de *outliers* e melhor do que se fosse ao contrário (preços sobre-estimados), visto que preços mais baixos que aqueles que

efetivamente aconteceram também levariam à conversão de venda. Para preços acima, esse comportamento não é garantido, o que reduziria os ganhos da companhia.

Figura 12 – Histograma das razões entre predições e rótulos para a abordagem baseada em algoritmos genéticos e modelos em cascata



Fonte: Elaborado pelo Autor.

6.3 Abordagem baseada em Redes Neurais

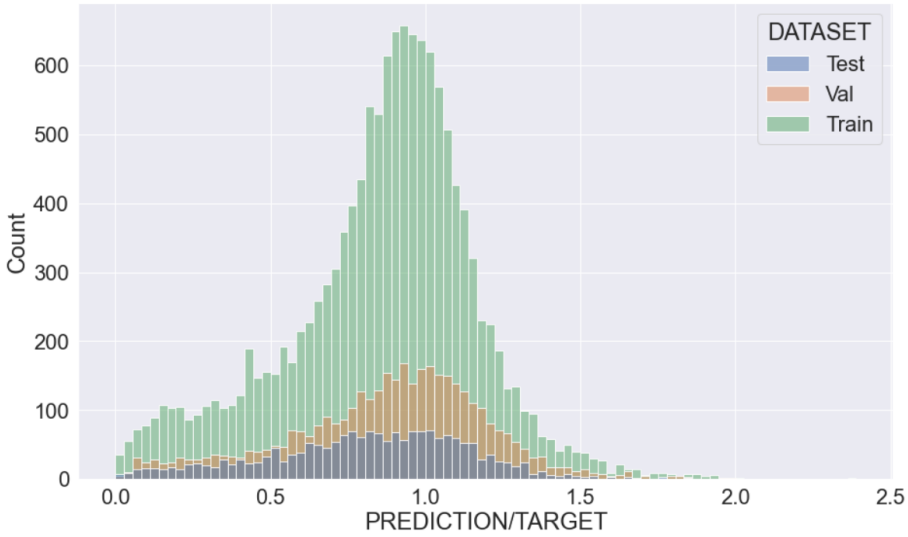
A busca aleatória para a abordagem baseada em redes neurais também foi realizada em 200 iterações. Se comparado aos outros métodos, o modelo obteve a performance estável na mudança de *datasets*, mostrando a capacidade do método de generalizar e sua robustez em relação aos *outliers*. O processo de busca também favoreceu a utilização do recurso de *dropout*, com taxa de 25%. Este favorece a regularização do modelo, reduzindo fenômenos de sobreajuste. O histograma da razão entre as predições da rede neural e os rótulos do conjunto de dados é apresentado (Figura 13). É possível observar que o modelo subestimou mais que sobre-estimou o *markup*, o que é uma categoria de erro mais segura que o contrário, dado que preços mais baixos garantem a conversão e o contrário não. Além disso, subestimar os preços é esperado pela própria natureza do *dataset*: margens, em ocasiões específicas, são levadas a elevados patamares por informações não necessariamente contidas nos dados. Ademais, dispositivos de regularização, a exemplo do *dropout*, garantem que o modelo favorece a estabilidade perante a performance em dados de treino e, como consequência, prioriza a generalização em dados de validação e teste. Ainda que não tão proeminente nos dados de teste, o pico da razão entre as predições e

Tabela 8 – Desempenho do Classificador

Conjunto de Dados	ROC AUC
Treinamento	67.815%
Validação	66.311%
Teste	62.441%

os rótulos em todos os conjuntos de dados, é por volta de 1.

Figura 13 – Histograma das razões entre predições e rótulos para a abordagem baseada em Redes Neurais



Fonte: Elaborado pelo Autor.

7 CONCLUSÕES

Com base na revisão bibliográfica realizada, foi possível atestar que a complexidade da tarefa de precificação é dependente dos pormenores de cada mercado, visto que a demanda, os produtos, os clientes e as informações disponíveis variam conforme cada situação. A partir dos trabalhos investigados, pôde-se observar os principais mecanismos utilizados para a precificação dinâmica, bem como os métodos para análise de demanda. Foram apresentadas características de diversos mercados associados a métodos de Inteligência Artificial, promovendo subsídio para o desenvolvimento de uma pesquisa com um conjunto de características não explorado desde então. Aspectos referentes ao impacto e tratamento de anomalias e *outliers* também foram abordados. As propostas elaboradas aqui são aplicáveis a outros contextos, desde que estes possuam distribuições assimétricas e dados com baixa qualidade, a exemplo de ambientes com baixa digitalização.

Na aplicação aqui estudada, a utilização de múltiplos modelos (*Stacking*) foi proposta, dado que é um conjunto de dados com distribuição mal comportada e de alta cardinalidade. Foram executados processos para transformação, tratamento, normalização e redução de cardinalidade dos dados antes destes serem processados pelos modelos preditivos, com o intuito de permitir que as informações presentes nos dados estejam mais explícitas para os algoritmos de regressão. A proposta sobre-estimou os valores dos rótulos, o que faz com que sua adoção em ambiente de produção seja questionável. Ao se aumentarem em demasia os valores, maior será a probabilidade de clientes optarem por outro fornecedor, reduzindo a performance comercial da empresa.

Da mesma forma, uma abordagem utilizando modelos em cascata parametrizados por um algoritmo genético foi proposta. Nesse esse caso, a própria arquitetura consegue determinar em qual estratégia as oportunidades serão processadas, permitindo com que o sistema decida por fornecer predições para todas as ocasiões ou por direcioná-las a especialistas humanos. Adicionalmente, o algoritmo genético encontra qual subconjunto de *features* é o mais adequado para os modelos, reduzindo o tempo para o fornecimento de predições em ambiente de produção. A proposta obteve melhores resultados que *Stacking* e subestimou os valores dos rótulos. Ainda que isso seja ruim, pois diminui margens, é uma alternativa melhor à da outra proposta, dado que preços mais baixos ainda garantem conversão. A metodologia foi eficiente em seleção de variáveis e ajuste de modelos em cascata, através da otimização pelo algoritmo genético.

Com esforço computacional menor, redes neurais (MLPs) apresentaram resultados similares à abordagem baseada em algoritmos genéticos. Modelos dessa natureza lidam muito bem com dados não-tabulares, como imagens ou áudio, mas aqui sua performance também foi garantida em conjuntos de dados oriundos de bancos relacionais. No ponto de vista de aplicação, a abordagem baseada em algoritmos genéticos se destaca não pela performance, da mesma magnitude que as redes neurais, mas sim pela possibilidade do

sistema indicar qual observação é um *outlier*, quantificando uma informação preciosa para especialistas humanos.

Os modelos foram treinados mantendo a presença de *outliers* na base de dados, dado que valores elevados podem representar oportunidades raras de lucros significativos. Porém, visto que existem sinais exógenos aos dados que dão ao time comercial indicativos de oportunidades de estresse de preços, entende-se que dificilmente será possível obter modelos com ajustes que gerem desempenhos maiores que aqueles apresentados aqui. Adicionalmente, o baixo nível de digitalização do negócio estudado dificulta a geração de variáveis adicionais sem a captura de novas fontes de dados. Apesar disso, acredita-se que a solução pode atuar como uma recomendação, fazendo com que os analistas gastem menos tempo na precificação, e enfatizando as etapas de negociação e construção de relacionamento com clientes.

Em resumo, dentre as contribuições deste trabalho para o processo de precificação, podem ser destacadas:

- Combinação de detecção de anomalias em processos de precificação, conforme proposto por Ramakrishnan et al. (RAMAKRISHNAN et al., 2019) com a precificação em si;
- Desenvolvimento e validação de arquitetura de AutoML, similar àquela proposta por (CHEN et al., 2018), aplicável a contextos de precificação com dados abundantes em *outliers*;
- Possibilidade de geração de *score*, indicando quantitativamente anomalias e oportunidades em transações comerciais, guiando especialistas humanos rumo a uma precificação mais eficiente;
- Verificação de performance insatisfatória de abordagens baseadas em *Stacking* em conjuntos de dados com distribuições assimétricas; e constatação da boa performance de redes neurais no contexto.

7.1 Propostas de Continuidade

Com base na proposta aqui desenvolvida e nos trabalhos relacionados, entende-se que ainda há oportunidades de tentativa de propor soluções mais efetivas. Uma alternativa é a construção de modelos que forneçam preços de custo, trazendo uma *proxi* do preço atual dos produtos nos fornecedores. Além dessa alternativa, a aplicação de técnicas não-supervisionadas (a exemplo de *AutoEncoders*) para detecção de anomalias podem ser úteis na detecção de oportunidades fora da normalidade. A criação de modelos de especialistas para cada material ou configuração também é uma abordagem a se cogitar, mas nesse caso os dados se tornariam ainda mais esparsos. Idealmente, essa oportunidade deve ser testada em uma situação com volumetria maior de dados. Um possível complemento para

combater essa ausência de dados seria o agrupamento hierárquico de produtos, buscando o atingimento de um tamanho amostral razoável através de diferentes granularidades.

Visto que todos os modelos sofreram com o fenômeno de sobreajuste e os conjuntos de dados foram particionados temporalmente, é possível que haja a presença de *Concept Drift* nos modelos. Isso significa que, na medida que o tempo passa, variáveis mudam suas características. A comprovação deste fenômeno pode ser avaliada em estudos adicionais.

Obviamente, a maneira mais efetiva de reduzir esforços de engenharia e processamento de dados é através da melhoria de processos e coleta de informações. Informações das transações que não converteram em vendas iriam gerar uma fotografia precisa de como os clientes reagem aos preços, permitindo com que a modelagem seja efetuada sob a ótica de classificação e otimização. Para o caso da companhia em questão, a recomendação é que esforços nessa direção sejam realizados com a visão de longo prazo, permitindo a consolidação de conjuntos de dados ricos e, consequentemente, uma precificação mais eficiente. A ferramenta desenvolvida neste projeto de pesquisa fomenta a cultura de dados e é um passo importante nessa direção.

REFERÊNCIAS

BAUER, J.; JANNACH, D. Optimal pricing in e-commerce based on sparse and noisy data. Decision Support Systems, v. 106, p. 53 – 63, 2018. ISSN 0167-9236. Disponível em: <http://www.sciencedirect.com/science/article/pii/S016792361730221X>. 18, 42

BERGSTRA, J.; BENGIO, Y. Random search for hyper-parameter optimization. J. Mach. Learn. Res., JMLR.org, v. 13, n. null, p. 281–305, fev. 2012. ISSN 1532-4435. 27

BURKOV, A. Machine Learning Engineering. [S.l.: s.n.], 2020. ISBN 978-1-7770054-6-7. 13, 27, 31, 41, 45

CAVALLO, A. More Amazon Effects: Online Competition and Pricing Behaviors. [S.l.], 2018. (Working Paper Series, 25138). Disponível em: <http://www.nber.org/papers/w25138>. 11

CHEN, B. et al. Autostacker: A compositional evolutionary learning system. In: Proceedings of the Genetic and Evolutionary Computation Conference. [S.l.: s.n.], 2018. p. 402–409. 13, 19, 20, 31, 54

CHEN, T.; GUESTRIN, C. Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: Association for Computing Machinery, 2016. (KDD '16), p. 785–794. ISBN 9781450342322. Disponível em: <https://doi.org/10.1145/2939672.2939785>. 24

CHOI, S.; KIM, T.; YU, W. Performance evaluation of ransac family. In: BMVC. [S.l.: s.n.], 2009. 23

DASGUPTA, P.; DAS, R. Dynamic pricing with limited competitor information in a multi-agent economy. Cooperative Information Systems. CoopIS 2000, Lecture Notes in Computer Science, Springer, v. 1901, 2000. 15

DEISENROTH, M. P.; FAISAL, A. A.; ONG, C. S. Mathematics for Machine Learning. [S.l.]: Cambridge University Press, 2020. 21, 23

den Boer, A. V. Dynamic pricing and learning: Historical origins, current research, and new directions. Surveys in Operations Research and Management Science, v. 20, n. 1, p. 1–18, 2015. ISSN 1876-7354. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1876735415000021>. 11

den Boer, A. V. Tracking the market: Dynamic pricing and learning in a changing environment. European Journal of Operational Research, v. 247, n. 3, p. 914–927, 2015. ISSN 0377-2217. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0377221715005925>. 11

DONOHO, D. High-dimensional data analysis: The curses and blessings of dimensionality. AMS Math Challenges Lecture, p. 1–32, 01 2000. 29

FERNáNDEZ Ángela; BELLA, J.; DORRONSORO, J. R. Supervised outlier detection for classification and regression. Neurocomputing, v. 486, p. 77–92, 2022. ISSN 0925-2312. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0925231222002090>. 12, 13, 19, 20, 28

GERON, A.; SAFARI, a. O. M. C. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd Edition. [S.l.]: O'Reilly Media, Incorporated, 2019. 24, 25

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. Deep Learning. [S.l.]: The MIT Press, 2016. ISBN 0262035618. 25

GUPTA, R.; PATHAK, C. A machine learning framework for predicting purchase by online customers based on dynamic pricing. Procedia Computer Science, v. 36, p. 599 – 605, 2014. ISSN 1877-0509. Complex Adaptive Systems Philadelphia, PA November 3-5, 2014. Disponível em: <http://www.sciencedirect.com/science/article/pii/S187705091401309X>. 15, 16

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. The Elements of Statistical Learning. New York, NY, USA: Springer New York Inc., 2001. (Springer Series in Statistics). 21, 24

HAWKINS, D. M. Identification of outliers. [S.l.]: Springer, 1980. v. 11. 12, 13

HE, X.; ZHAO, K.; CHU, X. Automl: A survey of the state-of-the-art. Knowledge-Based Systems, v. 212, p. 106622, 2021. ISSN 0950-7051. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0950705120307516>. 19

HONG, J.-H.; CHO, S.-B. Efficient huge-scale feature selection with speciated genetic algorithm. Pattern Recognition Letters, v. 27, n. 2, p. 143–150, 2006. ISSN 0167-8655. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0167865505002035>. 28

JUNG, M.; ZSCHEISCHLER, J. A guided hybrid genetic algorithm for feature selection with expensive cost functions. Procedia Computer Science, v. 18, p. 2337–2346, 2013. ISSN 1877-0509. 2013 International Conference on Computational Science. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1877050913005486>. 28

KATSOV, I. Introduction to Algorithmic Marketing: Artificial Intelligence for Marketing Operations. [S.l.: s.n.], 2017. ISBN 0692989048. 11

MAESTRE, R. et al. Reinforcement learning for fair dynamic pricing. In: IntelliSys. [S.l.: s.n.], 2018. 15, 16

NARAHARI, Y. et al. Dynamic pricing models for electronic business. Sadhana, Springer India, v. 30, p. 231 – 256, 2015. 15, 16

PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, v. 12, p. 2825–2830, 2011. 22, 23, 24, 45

RAMAKRISHNAN, J. et al. Anomaly detection for an e-commerce pricing system. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery Data Mining. New York, NY, USA: Association for Computing

Machinery, 2019. (KDD '19), p. 1917–1926. ISBN 9781450362016. Disponível em: <https://doi.org.ez93.periodicos.capes.gov.br/10.1145/3292500.3330748>. 13, 18, 20, 31, 54

RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R. J. Learning representations by back-propagating errors. Nature, v. 323, p. 533–536, 1986. 26

SAFARIK, J. et al. Genetic algorithm for automatic tuning of neural network hyperparameters. In: DUDZIK, M. C.; RICKLIN, J. C. (Ed.). Autonomous Systems: Sensors, Vehicles, Security, and the Internet of Everything. SPIE, 2018. v. 10643, p. 168 – 174. Disponível em: <https://doi.org/10.1117/12.2304955>. 19

SCHLOSSER, R.; BOISSIER, M. Dynamic pricing under competition on online marketplaces: A data-driven approach. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery Data Mining. New York, NY, USA: Association for Computing Machinery, 2018. (KDD '18), p. 705–714. ISBN 9781450355520. Disponível em: <https://doi.org.ez93.periodicos.capes.gov.br/10.1145/3219819.3219833>. 15, 18

SHUKLA, N. et al. Dynamic pricing for airline ancillaries with customer context. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery Data Mining. New York, NY, USA: Association for Computing Machinery, 2019. (KDD '19), p. 2174–2182. ISBN 9781450362016. Disponível em: <https://doi.org.ez93.periodicos.capes.gov.br/10.1145/3292500.3330746>. 15, 17

SOARES, G. L. Algoritmos genéticos: estudo, novas técnicas e aplicações. 1997. 27

TRUONG, Q. et al. Housing price prediction via improved machine learning techniques. Procedia Computer Science, v. 174, p. 433–442, 2020. ISSN 1877-0509. 2019 International Conference on Identification, Information and Knowledge in the Internet of Things. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1877050920316318>. 30

WICAKSONO, A. S.; SUPIANTO, A. A. Hyper parameter optimization using genetic algorithm on machine learning methods for online news popularity prediction. Int. J. Adv. Comput. Sci. Appl, v. 9, n. 12, p. 263–267, 2018. 19

WOLPERT, D.; MACREADY, W. No free lunch theorems for optimization. IEEE Transactions on Evolutionary Computation, v. 1, n. 1, p. 67–82, 1997. 20

WOLPERT, D. H. Stacked generalization. Neural Networks, v. 5, n. 2, p. 241–259, 1992. ISSN 0893-6080. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0893608005800231>. 30

WOLPERT, D. H. The Lack of A Priori Distinctions Between Learning Algorithms. Neural Computation, v. 8, n. 7, p. 1341–1390, 10 1996. ISSN 0899-7667. Disponível em: <https://doi.org/10.1162/neco.1996.8.7.1341>. 20

WOLPERT, D. H. The supervised learning no-free-lunch theorems. In: _____. Soft Computing and Industry: Recent Applications. London: Springer London, 2002. p. 25–42. ISBN 978-1-4471-0123-9. Disponível em: <https://doi.org/10.1007/978-1-4471-0123-9_3>. 20

YANG, L.; SHAMI, A. On hyperparameter optimization of machine learning algorithms: Theory and practice. Neurocomputing, v. 415, p. 295–316, 2020. ISSN 0925-2312. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0925231220311693>. 27

YE, P. et al. Customized regression model for airbnb dynamic pricing. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York, NY, USA: Association for Computing Machinery, 2018. (KDD '18), p. 932–940. ISBN 9781450355520. Disponível em: <https://doi.org.ez93.periodicos.capes.gov.br/10.1145/3219819.3219830>. 15, 16, 17, 20, 25

YOUNG, S. R. et al. Optimizing deep learning hyper-parameters through an evolutionary algorithm. In: Proceedings of the Workshop on Machine Learning in High-Performance Computing Environments. New York, NY, USA: Association for Computing Machinery, 2015. (MLHPC '15). ISBN 9781450340069. Disponível em: <https://doi.org/10.1145/2834892.2834896>. 27

ZHENG, A.; CASARI, A. Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists. 1st. ed. [S.l.]: O'Reilly Media, Inc., 2018. ISBN 1491953241. 45