Universidade Federal da Paraíba
Centro de Informática
Programa de Pós-Graduação em Modelagem Matemática e Computacional

TEORIA DOS LEILÕES NAS CONTRATAÇÕES PÚBLICAS PARAIBANAS - UMA ESTIMAÇÃO DOS CUSTOS DE TRANSAÇÕES ATRAVÉS DO APRENDIZADO DE MÁQUINA

Bradson Camelo

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Modelagem Matemática e Computacional, UFPB, da Universidade Federal da Paraíba, como parte dos requisitos necessários à obtenção do título de Mestre em Modelagem Matemática e Computacional.

Orientadores: Prof. Dr. Pedro Rafael Diniz Marinho
Prof. Dr. Rodrigo Bernardo da Silva

João Pessoa
Julho de 2024

TEORIA DOS LEILÕES NAS CONTRATAÇÕES PÚBLICAS PARAIBANAS - UMA ESTIMAÇÃO DOS CUSTOS DE TRANSAÇÕES ATRAVÉS DO APRENDIZADO DE MÁQUINA

Bradson Camelo

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DO PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM MATEMÁTICA E COMPUTACIONAL (PPGMMC) DO CENTRO DE INFORMÁTICA DA UNIVERSIDADE FEDERAL DA PARAÍBA COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM MODELAGEM MATEMÁTICA E COMPUTACIONAL.

Examinada por:

Presidente Prof. Dr., Pedro Rafael Diniz Marinho

Examinador interno: Prof. Dr., Rodrigo Bernardo da Silva

Examinador interno: Prof. Dr., Claudio Javier Tablada

Examinador externo: Prof. Dr., Joao Agnaldo Do Nascimento

JOÃO PESSOA, PB – BRASIL
JULHO DE 2024

Ata da Sessão Pública de Defesa de Dissertação de Mestrado de **BRADSON TIBÉRIO LUNA CAMELO**, candidato ao título de Mestre em Modelagem Matemática e Computacional, realizada no dia 19 de julho de 2024.

Aos dezenove dias do mês de julho do ano de dois mil e vinte e quatro, às 14h, via videoconferência, reuniram-se os membros da Banca Examinadora constituída para julgar o Trabalho Final do discente BRADSON TIBÉRIO LUNA CAMELO, vinculado a Universidade Federal da Paraíba sob a matrícula nº 20221018274, candidato ao grau de Mestre em "*Modelagem Matemática e Computacional*", na linha de pesquisa *"Modelagem Probabilística"* do Programa de Pós-Graduação em Modelagem Matemática e Computacional. A comissão examinadora foi composta pelos professores Pedro Rafael Diniz Marinho, Orientador e Presidente da Banca; Rodrigo Bernardo da Silva, Coorientador; Claudio Javier Tablada, Examinador Interno ao Programa; e João Agnaldo do Nascimento, Examinador Externo ao Programa. Dando início aos trabalhos, o Presidente da Banca cumprimentou os presentes, comunicou a finalidade da reunião e passou a palavra ao candidato para que fizesse, oralmente, a exposição do trabalho de dissertação intitulado *"TEORIA DOS LEILÕES NAS CONTRATAÇÕES PÚBLICAS PARAIBANAS - UMA ESTIMAÇÃO DOS CUSTOS DE TRANSAÇÕES ATRAVÉS DO APRENDIZADO DE MÁQUINA"*. Concluída a exposição, o candidato foi arguido pela Banca Examinadora, que emitiu o seguinte parecer: "***aprovado***". Do ocorrido, eu, Gean Paulo P. M. de Barros, secretário do Programa de Pós-Graduação em Modelagem Matemática e Computacional (PPGMMC), lavrei a presente ata, que vai assinada por mim e pelos membros da Banca Examinadora. João Pessoa, 19 de julho de 2024.

Gean Paulo Pereira Maurício de Barros
Secretário do PPGMMC
SIAPE 2326476

gov.br Documento assinado digitalmente
PEDRO RAFAEL DINIZ MARINHO
Data: 20/07/2024 16:12:55-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Pedro Rafael Diniz Marinho
Orientador (PPGMMC)

gov.br Documento assinado digitalmente
RODRIGO BERNARDO DA SILVA
Data: 22/07/2024 11:59:44-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Rodrigo Bernardo da Silva
Coorientador (PPGMMC)

gov.br Documento assinado digitalmente
CLAUDIO JAVIER TABLADA
Data: 20/07/2024 17:40:04-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Claudio Javier Tablada
Examinador Interno ao Programa (PPGMMC)

gov.br Documento assinado digitalmente
JOAO AGNALDO DO NASCIMENTO
Data: 20/07/2024 15:20:09-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. João Agnaldo do Nascimento
Examinador Externo ao Programa (UFPB)

Catalogação na publicação
Seção de Catalogação e Classificação

C181t Camelo, Bradson Tiberio Luna.
Teoria dos leilões nas contratações públicas paraibanas - uma estimação dos custos de transações através do aprendizado de máquina / Bradson Tiberio Luna Camelo. - João Pessoa, 2024.
111 f.

Orientação: Pedro Rafael Diniz Marinho, Rodrigo Bernardo da Silva.
Dissertação (Mestrado) - UFPB/CI.

1. Contratações públicas. 2. Custos de transação. 3. Aprendizado de máquina. 4. Teoria dos leilões. I. Marinho, Pedro Rafael Diniz. II. Silva, Rodrigo Bernardo da. III. Título.

UFPB/BC CDU 51:351.712(043)

Elaborado por CHRISTIANE CASTRO LIMA DA SILVA - CRB-15/865

Aos meus pais, minha esposa e meus filhos: meu passado, meu presente e meus futuros

Agradecimentos

Aos meus orientadores, os Profs. Drs. Pedro Rafael Diniz Marinho e Rodrigo Bernardo da Silva, que começaram como meus professores, tornaram-se amigos e, por fim, aceitaram me orientar. Obrigado pelas contribuições para o desenvolvimento deste trabalho, pela paciência e dedicação, mesmo quando eu passava algum tempo desaparecido.

À minha esposa Larissa, pela compreensão e apoio, mesmo quando eu passava intermináveis horas (todos os dias) no meu "bunker" com as minhas "matemáticas" e afins.

Aos meus três filhos, Felipe, Caio e Lucas, meus eternos três mosqueteiros.

Aos meus pais Anchieta e Eunice que sempre estarão presentes, mesmo na ausência.

Aos meus sogros Italo e Laura, pelo carinho e apoio nesta caminhada.

Aos professores da Pós-Graduação em Modelagem Matemática Computacional, em especial os Professores Claudio Javier Tablada e João Agnaldo Do Nascimento, pelas contribuições na qualificação, e aos Professores Luciano, Ana Paula, Marcelo e Hugo.

Aos colegas que ajudaram nas discussões acadêmicas e na evolução do conhecimento, em especial Valdemir e Jailson.

Por fim, mas não menos importantes, aos amigos que discutiram várias vezes essa pesquisa comigo, em especial Ronny Charles, Emiliano Zapata, Marcos Nóbrega, Fernando Baltar, Thayse Dias, Gabriela Galvão, Levi Pessoa, Ivo Cilento, Karlos Farias, Marcus Freire e Roberta do Bú.

Agradeço ao Tribunal de Contas do Estado da Paraíba pelo acesso aos dados e ambiente de trabalho que permite o aperfeiçoamento profissional, em especial aos membros e servidores deste órgão que me orgulho de fazer parte, no Ministério Público de Contas.

Minha gratidão e reconhecimento a todos vocês!

Resumo da Dissertação apresentada ao PPGMMC/CI/UFPB como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

TEORIA DOS LEILÕES NAS CONTRATAÇÕES PÚBLICAS PARAIBANAS - UMA ESTIMAÇÃO DOS CUSTOS DE TRANSAÇÕES ATRAVÉS DO APRENDIZADO DE MÁQUINA

Bradson Camelo

Julho/2024

Orientadores: Prof. Dr. Pedro Rafael Diniz Marinho
Prof. Dr. Rodrigo Bernardo da Silva

Programa: Modelagem Matemática e Computacional

Esta dissertação visa preencher uma lacuna específica na literatura ao investigar os custos de transação em contratações públicas, utilizando modelos matemáticos e técnicas de aprendizado de máquina. A motivação pessoal e prática para esta pesquisa surge da necessidade de melhorar a eficiência e a transparência nas licitações públicas, particularmente no estado da Paraíba, onde a otimização dos recursos públicos é de suma importância.

O objetivo principal deste estudo é desenvolver e aplicar um modelo matemático para analisar as contratações públicas e utilizar técnicas de aprendizado de máquina para prever os custos de transação que impactam os preços públicos.

A pesquisa é dividida em duas partes principais. A primeira parte dedica-se ao desenvolvimento de um modelo matemático (jogo-teórico), adaptando os leilões clássicos para as modalidades de contratação pública mais usuais, como Concorrência e Pregão. Esta seção explora matematicamente o impacto das estratégias e comportamentos dos participantes nos resultados dos leilões, com foco na presença de custos de transação, incluindo preços de entrada.

Na segunda parte, são aplicadas técnicas de aprendizado de máquina para prever os custos de transação em contratações públicas, utilizando dados como notas fiscais de entidades públicas do estado da Paraíba, bem como informações econômicas, geográficas, sociais e contábeis. Os métodos incluem o uso de *Random Forest* e *LASSO* para criar modelos preditivos, visando estimar os preços das contratações de forma mais precisa.

Os resultados da pesquisa indicam que o modelo *Random Forest* apresentou um coeficiente de determinação (R^2) de 0,97, explicando cerca de 97% da variabilidade nos custos de transação, com um erro quadrático médio (RMSE) de 0,14 desvios padrão dos preços normalizados. A análise revelou que fatores como o Tempo Médio de pagamento e prazo para adimplir precatórios são determinantes para os custos de transação. Esses resultados mostram que é possível prever os custos de transação nas contratações públicas com alta precisão, utilizando técnicas avançadas de aprendizado de máquina.

Na conclusão, destacam-se as consequências práticas da pesquisa, como a possibilidade de implementar modelos preditivos para melhorar a gestão das contratações públicas, promovendo maior eficiência e transparência no uso dos recursos públicos. A abordagem interdisciplinar adotada, que combina estatística, economia, matemática, ciência da computação e administração pública, reflete a complexidade e relevância do tema, oferecendo ferramentas práticas e teóricas para aprimorar os processos de licitação pública.

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

AUCTION THEORY IN PARAÍBA'S PUBLIC PROCUREMENT - A TRANSACTION COSTS ESTIMATION USING MACHINE LEARNING

Bradson Camelo

July/2024

Advisors: Prof. Dr. Pedro Rafael Diniz Marinho
Prof. Dr. Rodrigo Bernardo da Silva

Program: Computational Mathematical Modelling

This master's thesis aims to fill a specific gap in the literature by investigating transaction costs in public procurement, using mathematical models and machine learning techniques. The personal and practical motivation for this research arises from the need to improve efficiency and transparency in public tenders, particularly in the state of Paraíba, where the optimization of public resources is of utmost importance.

The main objective of this study is to develop and apply a mathematical model to analyze public procurements and use machine learning techniques to predict the transaction costs that impact public pricing.

The research is divided into two main parts. The first part is dedicated to developing a mathematical model (game-theoretic), adapting classic auctions to the most common public procurement modalities, such as Competitive Bidding and Auction. This section mathematically explores the impact of participants' strategies and behaviors on auction outcomes, focusing on the presence of transaction costs, including entry prices.

In the second part, machine learning techniques are applied to predict transaction costs in public procurements, using data such as invoices from public entities in the state of Paraíba, as well as economic, geographical, social, and accounting information. The methods include the use of Random Forest and LASSO to create predictive models, aiming to estimate procurement prices more accurately.

The research results indicate that the Random Forest model presented a coefficient of determination (R^2) of 0,97, explaining about 97% of the variability in transaction costs, with a root mean squared error (RMSE) of 0,14 standard deviations of

normalized prices. The analysis revealed that factors such as Average Payment Time and the timeframe for fulfilling judicial debts (precatórios) are crucial determinants of transaction costs. These results show that it is possible to predict transaction costs in public procurements with high precision using advanced machine learning techniques.

In conclusion, the practical implications of the research are highlighted, such as the possibility of implementing predictive models to improve the management of public procurements, promoting greater efficiency and transparency in the use of public resources. The interdisciplinary approach adopted, which combines statistics, economics, mathematics, computer science, and public administration, reflects the complexity and relevance of the topic, offering practical and theoretical tools to enhance public procurement processes.

Sumário

Lista de Figuras **xii**

Lista de Tabelas **xiii**

1 Introdução **1**

2 Fundamentação teórica **7**

2.1 Teoria dos Leilões 8

2.1.1 Pressupostos 9

2.1.2 Axiomas da Teoria da Decisão 11

2.1.3 Informação Privada x Comum 13

2.1.4 Tipos de Leilões Clássicos 13

2.2 Custos de Transação 15

2.3 Estatística Clássica e Aprendizado de Máquina 18

2.3.1 Algumas técnicas de aprendizado de máquina 27

2.3.2 Limpeza dos dados 32

2.4 Conclusão e Lacunas na Literatura 33

3 Modelo Matemático **36**

3.1 Pressupostos 36

3.1.1 Teoria dos jogos aplicada aos leilões 37

3.1.2 Descrição dos leilões 40

3.1.3 Valores Privados 40

3.2 Leilão aberto 41

3.2.1 Solução Analítica 45

3.3 Licitação de primeiro preço 46

3.3.1 Acrescentando preços de reserva e custos de transação - custos de entrada 49

3.4 Conclusão 52

4 Modelos de Aprendizado de máquina: *LASSO Random Forest* e *Gradient Boosting* 53

4.1 *LASSO* . . . 53

4.2 Árvores de decisão . . . 54

4.3 Técnicas de *Ensemble* . . . 58

4.3.1 *Random Forest* . . . 59

4.3.2 *boosting* . . . 60

4.4 Conclusão . . . 62

5 Treinamento dos modelos de regressão 64

5.1 Preparação e Análise Exploratória dos Dados . . . 65

5.2 Construção e avaliação do Modelo . . . 74

5.2.1 Modelos de Regressão . . . 76

5.2.2 Conclusão dos Modelos . . . 82

5.2.3 Importância das variáveis . . . 84

5.3 Limitações dos Métodos e Soluções Propostas . . . 85

5.4 Conclusão do aprendizado de máquina . . . 86

6 Conclusão 88

6.1 Resumo dos Resultados . . . 88

6.2 Implicações Teóricas e Práticas . . . 89

6.3 Limitações do Estudo . . . 89

6.4 Sugestões para Pesquisas Futuras . . . 90

7 Referências Bibliográficas 91

A Códigos em linguagem R - Link para github 98

Lista de Figuras

2.1 Separação do conjunto de dados. Fonte: Marinho, 2023, p. 88. 22
2.2 Validação Cruzada - *K-fold*. 24

3.1 Elaboração própria. 42

4.1 Exemplo de estrutura de uma árvore de regressão. 55
4.2 Fonte: Izbicki e Santos (2020), p. 77. 55

5.1 Frequência das modalidades de licitação nos municípios paraibanos. . 65
5.2 Atraso nos pagamentos em 2020. 67
5.3 Atraso nos pagamentos entre 2019-2023. 68
5.4 Densidade da distribuição do Custo de Transacao Normalizado. . . . 70
5.5 Matriz de correlação entre as variáveis independentes. 70
5.6 Dispersão e correlação das variáveis. 72
5.7 Distribuição das variáveis independentes. 73
5.8 Metricas para cada combinacao no *Grid Search* em cada Fold (*LASSO*). 77
5.9 Tunagem do modelo *Random Forest*. 79
5.10 Tunagem do *Gradient Boosting*. 81
5.11 Valores previstos x observados (*LASSO*). 82
5.12 Valores previstos x observados (*Gradient Boosting*). 83
5.13 Valores previstos x observados (*Random Forest*). 83
5.14 Importância das variáveis para o modelo *Random Forest*. 84

Lista de Tabelas

5.1 Definição das variáveis. 68
5.2 Resultados da Regressão Linear. 75
5.3 Melhores Hiperparâmetros do Modelo Lasso. 78
5.4 Valores dos hiperparâmetros do modelo Random Forest. 80
5.5 Melhores Hiperparâmetros Encontrados. 82
5.6 Resultados do modelo *Random Forest*, *Gradient Boosting* e *LASSO*. . 84

Capítulo 1

Introdução

A presente dissertação explora a intersecção de dois campos importantes no contexto das contratações públicas e formação de preços: a teoria dos leilões e o aprendizado de máquina. Este estudo visa aprimorar o entendimento e a eficiência dos processos de compras pública, especificamente no estado da Paraíba. Para isso, serão explorados modelos matemáticos de leilões e técnicas de aprendizado de máquina para prever os custos de transação, em especial reputacional, nas contratações públicas sem a analise de questões relacionadas com a legalidade do procedimento.

Para fornecer uma noção da importância mundial das contratações públicas, vale lembrar um relatório da Organização para a Cooperação e Desenvolvimento Econômico (OCDE) em 2011, que destacava que os processos de aquisição pública realizados nos países membros correspondiam, em média, a cerca de 13% do Produto Interno Bruto (PIB) e a 30% de todas as despesas governamentais (OCDE, 2011). Além dos problemas conhecidos sobre a licitude das contratações, existem sérios problemas de ineficiências nas contratações pública. Neste contexto, os países membros da OCDE adotaram várias ações para assegurar maior eficiência e transparência, incluindo licitações eletrônicas com informações disponíveis online, mecanismos de controle e a avaliação do impacto.

No contexto brasileiro, a efetividade dos valores das licitações é ainda mais comprometida não só por práticas ilícitas, mas também pela ineficácia do sistema atual de escolha da aquisição pública (Nobrega, 2020). Os procedimentos de contratações públicas vigentes atribuem muita importância ao aspecto formalista ao invés de uma abordagem com foco na eficiência, que utiliza uma metodologia científica do comportamento que visa estabelecer um critério normativo eficaz para o julgamento das políticas públicas (Cooter e Ullen, 2010).

A questão ganha maior relevância quando se considera o cenário de persistentes desigualdades regionais e sociais, em que os estados mais pobres costumam ser mais frequentemente afetados contratações ineficientes. O estado da Paraíba, por exemplo, encontra-se nessa situação. De acordo com dados do Instituto Brasileiro

de Geografia e Estatística (IBGE), em 2021, a Paraíba apresentava um Produto Interno Bruto (PIB) de R$ 77,5 bilhões, o que correspondia a 3% do PIB do estado de São Paulo (aproximadamente R$ 2,7 trilhões), o mais rico do país. Por outro lado, conforme dados do Tribunal de Contas do Estado da Paraíba (TCE-PB) para o mesmo período, ocorreram 61.461 licitações nos municípios paraibanos, totalizando licitações vencedoras no valor de 11,3 bilhões de reais, o que corresponde a 14,5% do PIB estadual naquele ano.

As contratações públicas são essenciais para garantir a alocação eficiente de recursos públicos, promovendo a concorrência e a transparência no processo de aquisição de bens e serviços pelo setor público. No entanto, este mecanismo, crucial para a governança e a integridade econômica, enfrenta desafios significativos em termos de fraudes e ineficiência. Neste trabalho de pesquisa, analisar-se-á apenas a eficiência (custos de transação).

Processos de aquisições públicas são conhecidos por terem custos de transação consideráveis (Strand et al., 2011; Libório et al, 2021). Entretanto, observa-se uma falta de trabalhos acadêmicos focados na avaliação da reputação dos compradores públicos, levantando algumas questões vitais. Por que os preços públicos são maiores que os preços privados? Será que a reputação de bom pagador de um ente público é capaz de impactar o preço das contratações? Há muitas variáveis impactando os custos de transação das contratações públicas.

Segundo North (1990), a eficiência das instituições pode ser ampliada pela otimização das normas que reduzam custos de transação (OECD, 2012; Guarnieri e Gomes, 2019). Análises da Lei de licitação e contratos administrativos por autores como Nóbrega (2012) e Libório et al. (2022) apontam para a associação dessa legislação com custos de transação elevados, que refletem a ineficiência do processo licitatório no Brasil. Como podem ser reduzidos esses custos? para responder a essa reflexão, usar-se-á a teoria dos leilões e será desenvolvido um modelo matemático para identificar, teoricamente, as variáveis que impactam os custos de transação.

Leilões têm sido, há muito tempo, de especial interesse para economistas devido a seus mecanismos explícitos para descrever como os preços são formados (Hendricks e Paarsch, 1995). Esses estudos enfatizam a relevância de entender as dinâmicas de leilão para otimizar os resultados em termos de custo e eficiência para o comprador. Contudo, poucas pesquisas foram realizadas para determinar os preços finais de leilões privados e nenhum estudo sobre preços públicos.

A previsão do preço final para leilões, que envolve a modelagem da incerteza relacionada ao processo de licitação, é uma tarefa desafiadora, principalmente devido à variedade de fatores que variam nas configurações dos leilões (Schapire et al., 2002). Mesmo que todos os fatores fossem contabilizados, ainda existe a incerteza no comportamento humano. A relação entre o preço final e os fatores relacionados

pode ser mais complexa do que uma simples linearidade. No entanto, essa questão está relacionada à utilidade dos licitantes, à receita dos vendedores e à eficiência da alocação do ponto de vista do bem-estar social como um todo.

Após a análise teórica, far-se-á um estudo empírico com o uso de aprendizado de máquina para prever os custos de transações das contratações públicas. Em finanças públicas, o tema tem ganhado atenção através da aplicação dessas técnicas para melhorar a detecção de fraudes em contratações públicas (ASH, GALLETTA e GIOMMONI, 2021). Essas tecnologias oferecem uma nova abordagem para lidar com a complexidade e a grande quantidade de dados envolvidos nas contratações públicas, permitindo análises mais precisas e eficientes.

O problema de pesquisa desta dissertação pode ser enunciado da seguinte forma: "Modelos matemáticos de leilões e técnicas de aprendizado de máquina podem prever custos de transação nas contratações públicas no estado da Paraíba?"

Esta dissertação tem como objetivo geral desenvolver e aplicar um modelo matemático para analisar as contratações públicas e utilizar técnicas de aprendizado de máquina para prever os custos de transação que impactam os preços públicos. E como objetivos específicos:

- Revisar a literatura sobre teoria dos leilões, custos de transação e aprendizado de máquina aplicados a contratações públicas.
- Inverter os modelos clássicos de leilão de venda para o modelo de leilão de compra, com preço de entrada e valor de reserva.
- Aplicar técnicas de aprendizado de máquina para prever os custos de transação com base em dados reais das contratações públicas na Paraíba.
- Propor recomendações práticas para a redução dos custos de transação nas contratações públicas.

A justificativa desta pesquisa encontra amparo prático e acadêmico. A aplicação de modelos matemáticos e técnicas de aprendizado de máquina para prever e reduzir os custos de transação nas contratações públicas pode trazer benefícios significativos para a administração pública. Primeiramente, ao sugerir caminhos para melhorar a eficiência dos processos licitatórios, esses modelos podem identificar e minimizar fatores que contribuem para os altos custos de transação, como distâncias, compras em pequeno volume e atrasos nos pagamentos entre os agentes envolvidos. A redução desses custos não apenas economiza recursos financeiros, mas também acelera a execução dos contratos, proporcionando serviços e produtos de forma mais rápida e eficiente à população.

Além disso, o uso de técnicas de aprendizado de máquina pode aumentar a transparência nos processos de contratação pública. Com a capacidade de analisar grandes volumes de dados e identificar padrões suspeitos, essas técnicas podem ajudar a detectar e prevenir contratações a preços muito acima da média de negociações similares. Isso é particularmente importante em estados com altos índices de desigualdade social e regional, como a Paraíba, onde a má gestão dos recursos públicos pode ter um impacto devastador na qualidade de vida da população.

Outro ponto importante é que, ao fornecer uma metodologia científica robusta para a previsão e análise dos custos de transação, a pesquisa pode servir como um modelo para outras regiões que enfrentarem desafios semelhantes. A implementação bem-sucedida dessas técnicas pode inspirar outras administrações públicas a adotarem abordagens semelhantes, amplificando o impacto positivo dessa pesquisa para a identificação de variáveis que impactam o preço público.

Assim, esta pesquisa oferece ferramentas práticas para os gestores públicos, possibilitando tomadas de decisão mais informadas e baseadas em dados concretos. Isso pode fortalecer a governança pública, melhorando o controle e alocação dos gastos públicos.

A relevância acadêmica desta dissertação reside na contribuição para a literatura sobre contratações públicas, teoria dos leilões e aprendizado de máquina. A pesquisa oferece uma análise dos custos de transação nas contratações públicas, explorando a abordagem através de aprendizado de máquina para a previsão desses custos, preenchendo uma lacuna existente na literatura acadêmica sobre leilões de compra, no modelo matemático, e sobre pesquisa empírica em contratações públicas.

A dissertação está estruturada da seguinte forma, iniciará com uma Revisão Bibliográfica que Abrange a teoria dos leilões, custos de transação e técnicas de aprendizado de máquina aplicados às contratações públicas. Neste capítulo, apresenta-se uma revisão abrangente da literatura, abordando a teoria dos leilões, custos de transação e aprendizado de máquina. A teoria dos leilões é discutida em termos de pressupostos, axiomas da teoria da decisão, e a distinção entre informação privada e comum, além de tipos clássicos de leilões. Em seguida, os custos de transação são analisados, destacando sua relevância econômica e a influência da reputação dos entes públicos como fator determinante nos preços das contratações. Por fim, são exploradas as técnicas de aprendizado de máquina e suas aplicações na previsão de custos de transação, incluindo métodos de regressão e técnicas de validação cruzada.

O capítulo seguinte será a exposição de um modelo matemático, desenvolvendo um modelo para identificar as variáveis que impactam os custos de transação. Aqui, é apresentado um modelo matemático invertido para leilões de compra, utilizando a teoria dos jogos bayesianos para analisar a eficiência e os custos de transação em contratações públicas. Inicialmente, são discutidos os pressupostos fundamentais

dos leilões, incluindo a aplicação da teoria dos jogos e o conceito de equilíbrio de Nash Bayesiano. Em seguida, são detalhados os diferentes tipos de leilões, como leilões abertos e leilões de primeiro preço, incorporando custos de transação e preços de reserva. A análise matemática demonstra como a implementação desses custos impacta a estratégia de lances dos participantes e a receita esperada do ente público. Através de análise funcional, conclui-se que a criação de custos de transação prejudica os preços finais das licitações, reforçando a necessidade de estratégias que minimizem esses custos para otimizar a eficiência das contratações públicas.

No quarto capítulo, far-se-á o treinamento dos modelos de regressão, com a aplicação de técnicas de aprendizado de máquina para prever os custos de transação com base em dados reais das contratações públicas na Paraíba. Aqui são apresentadas as aplicações de técnicas de machine learning para prever os custos de transação nas contratações públicas. A análise foi conduzida utilizando dois modelos principais de regressão: *Random Forest* e *LASSO*. Ainda detalha-se o processo de preparação dos dados, incluindo a organização de informações de notas fiscais, empenhos, licitações e dados socioeconômicos, além da criação de índices de reputação dos entes públicos. A avaliação dos modelos foi realizada com base em métricas de desempenho como RMSE, R^2 e MAE, com a *Random Forest* mostrando um desempenho superior na previsão dos custos de transação. A análise da importância das variáveis revelou que o tempo médio de pagamento e prazo para pagamento dos precatórios são fatores críticos. Os resultados demonstram que o uso de aprendizado de máquina pode significativamente melhorar a precisão das previsões de custos de transação, oferecendo insights valiosos para a otimização das contratações públicas.

Por fim, a conclusão apresenta uma síntese das principais conclusões do estudo e propostas de recomendações práticas para a redução dos custos de transação nas contratações públicas. Através de uma abordagem interdisciplinar que combina teoria dos leilões, custos de transação e aprendizado de máquina, foi possível fornecer uma análise detalhada das contratações públicas e desenvolver ferramentas práticas para aprimorar esses processos. A aplicação de um modelo matemático adaptado aos leilões de compra, complementado por modelos de regressão *Random Forest* e *LASSO*, revelou que algumas variáveis são cruciais na previsão dos custos de transação. Os resultados mostraram que o *Random Forest* é mais eficaz na previsão, contribuindo significativamente para a eficiência e transparência das contratações públicas. Esta pesquisa oferece insights valiosos para a formulação de políticas e estratégias que possam otimizar os processos licitatórios e reduzir os custos de transação, promovendo uma melhor alocação de recursos públicos.

O trabalho estatístico foi realizado com uma análise de dados reais do estado da Paraíba, usando as informações das notas fiscais de produtos comprados pelos entes públicos municipais entre 2019-2023. Além disso, trabalhou-se com dados

econômico-sociais dos municípios, dos processos licitatórios, empenhos, receita corrente e restos a pagar dos municípios entre 2018 e 2023.

Assim, a presente pesquisa possui diversas limitações que devem ser consideradas ao interpretar os resultados e ao planejar a implementação prática das recomendações.

Entre as questões internas, destacam-se a dependência da qualidade e consistência dos dados fornecidos pelos entes públicos ao TCE-PB. Inconsistências e erros nos dados de notas fiscais, empenhos e outras informações podem comprometer a precisão das análises e previsões. Além disso, os modelos matemáticos e algoritmos de aprendizado de máquina utilizados (*Random Forest* e *LASSO*) possuem limitações próprias. A seleção de variáveis independentes pode não capturar completamente todas as nuances dos custos de transação, e fatores relevantes podem ter sido omitidos ou inadequadamente representados, afetando a precisão das previsões.

Outro ponto que merece destaque é a generalização dos resultados, uma vez que os modelos desenvolvidos foram ajustados especificamente para o estado da Paraíba, o que pode limitar a aplicabilidade dos resultados para outras regiões ou contextos sem adaptações adicionais.

Essas limitações devem ser reconhecidas ao interpretar os resultados e ao considerar a implementação prática das recomendações desta pesquisa.

Capítulo 2

Fundamentação teórica

A revisão bibliográfica é um componente essencial desta dissertação, proporcionando uma base teórica sólida para a investigação sobre como modelos matemáticos de leilões e técnicas de aprendizado de máquina podem prever os custos de transação nas contratações públicas no estado da Paraíba. Este Capítulo examina detalhadamente as três áreas principais, destacando sua relevância e interconexão com o problema de pesquisa e os objetivos do estudo.

A teoria dos leilões é crucial para entender os mecanismos de formação de preços nas contratações públicas. Este segmento da revisão aborda os diferentes tipos de leilões, os pressupostos subjacentes e a aplicação da teoria dos jogos para modelar o comportamento dos participantes. A compreensão dessas dinâmicas é fundamental para desenvolver um modelo matemático que possa identificar as variáveis que impactam os custos de transação.

Os custos de transação, por sua vez, são um fator significativo nas contratações públicas. A revisão explora a literatura sobre a natureza e os determinantes desses custos, incluindo a reputação dos compradores públicos. Este entendimento é vital para analisar como esses custos influenciam os preços finais nas licitações e como podem ser reduzidos para aumentar a eficiência das contratações públicas.

O aprendizado de máquina representa uma abordagem para lidar com a complexidade e a grande quantidade de dados envolvidos nas contratações públicas. A revisão bibliográfica examina as principais técnicas de aprendizado de máquina, como modelos de regressão e métodos de validação, e suas aplicações na previsão de custos de transação. Estas técnicas são essenciais para desenvolver ferramentas práticas que possam melhorar a precisão das previsões e, consequentemente, a eficiência e transparência dos processos licitatórios.

O problema de pesquisa que guia esta dissertação é: "Modelos matemáticos de leilões e técnicas de aprendizado de máquina podem prever custos de transação nas contratações públicas no estado da Paraíba?" Para responder a essa questão, a dissertação estabelece os seguintes objetivos: desenvolver um modelo matemático

para identificar as variáveis que impactam os custos de transação e aplicar técnicas de aprendizado de máquina para prever esses custos.

A revisão bibliográfica destaca a lacuna na literatura existente, pois poucos estudos integraram essas abordagens para abordar a questão dos custos de transação nas contratações públicas. Ao fornecer uma análise detalhada e integrativa das teorias e técnicas relevantes, esta dissertação visa preencher essa lacuna e contribuir para a melhoria das políticas de contratação pública, tanto no contexto específico da Paraíba quanto potencialmente em outros contextos regionais e nacionais.

2.1 Teoria dos Leilões

Os leilões são um tipo de jogo, com estrutura composta por diversas regras e mecanismos que podem ser ajustados pelo comprador. Isso é feito para que os participantes (licitantes) atuem de uma maneira que favoreça os interesses de quem organiza o leilão (comprador). Conforme definido por Friedman (1984, p.48), uma instituição de mercado é "um conjunto de regras que especificam quais tipos de ofertas e outras mensagens são legítimas, e como e quando comerciantes específicos transacionam, dadas as mensagens escolhidas". Por sua vez, segundo McAfee e McMillan(1987, p. 701), um leilão é "uma instituição de mercado com um conjunto explícito de regras que determinam a alocação de recursos e os preços com base nos lances dos participantes do mercado". A escolha do método mais eficaz para essa alocação não é trivial, visto que diversos fatores influenciam o resultado. Desta forma, pode-se ver a contratação pública (via licitação) como um tipo de leilão, então, para simplificar, tratar-se-á como este sendo um gênero e aquela uma espécie em toda essa dissertação.

O processo de leilão é um aspecto crucial que determina a estrutura da negociação. Isso não abrange apenas a discussão sobre a divisão em lotes, mas também o critério de julgamento, que pode incluir fatores como preço, técnica, entre outros. A escolha do formato do leilão é igualmente significativa, pois pode influenciar a dinâmica da contratação. Cada formato tem suas peculiaridades e pode ser mais adequado, dependendo do contexto e do item em licitação. Toda a estruturação deve ser feita para criar incentivos para a revelação do melhor preço para os envolvidos (ente público e particulares vendedores). As regras de compartilhamento de informação são vitais para criar um ambiente que incentive os fornecedores a compartilhar informações relevantes.

Na estruturação do jogo da contratação pública, um dos pressupostos é de que os tomadores de decisão são racionais, ou seja, eles decidem de modo a buscar seus próprios objetivos. Assim, supõem-se que o propósito dos jogadores é maximizar o valor esperado de seus resultados, mensurados em uma escala de utilidade, seguindo

a ideia moderna de Von Neumann e Morgenstern (1944) do teorema de maximização da utilidade esperada. Para fins da Teoria, e lembrando que a ciência lida com modelos, considerar-se-á que, para finalidade do modelo, as pessoas são sempre racionais.

Quanto a maximizar o resultado da utilidade esperada, vale lembrar que não é necessariamente o mesmo que maximizar o resultado monetário esperado, pois a utilidade de cada unidade monetária varia de acordo com a aversão ao risco de cada indivíduo. Além disso, a utilidade não parece ser uma função linear (dependendo do estoque prévio do agente). Assim, os comportamentos dos tomadores de decisão - que satisfaçam certos axiomas intuitivos - podem ser descritos por um modelo quantitativo, já que sua conduta objetiva maximizar o valor matemático esperado da função utilidade, de acordo com uma distribuição de probabilidade subjetiva. Para maior aprofundamento, remete-se a von Neumann e Morgenstern (1944) ou a Luce e Raiffa (1957).

2.1.1 Pressupostos

Antes de tratar dos citados axiomas (Myerson, 1991) das escolhas racionais (fundamentos lógicos de todo o sistema), convém lembrar que as decisões costumam ser adotadas em ambientes de incertezas que são descritas por modelos probabilísticos e os agentes escolhem probabilidades (loteria) distribuídas sobre um conjunto de prêmios. Os modelos de probabilidade são adequados para descrever apostas com resultados que dependam de eventos objetivos e calculáveis, mas também existem as situações que não têm distribuição conhecida nem calculável que são denominadas de probabilidade subjetiva.

Em termos matemáticos, será adotada a terminologia de Myerson (1991). Assim, diz-se que para um conjunto finito Z, que é mapeado no conjunto $\mathbb{R}$ (co-domínio), é possível definir $\Delta(Z)$ como o conjunto de distribuições de probabilidade sobre Z. Em outras palavras:

$$\Delta(Z) = \{q : Z \to \mathbb{R} \mid \sum_{y \in Z} q(y) = 1 \text{ e } q(z) \geq 0, \forall z \in Z\}.$$

Designe-se por X o conjunto de todas as recompensas possíveis que o decisor pode receber, e por Ω o conjunto de todos os estados da natureza. Para simplificar, assim como Myerson (1991), assume-se que X e Ω são conjuntos finitos. Assim, é possível definir a probabilidade (loteria) como uma função f que designa um número real não negativo $f(x|t)$ para cada prêmio $x \in X$ e para cada estado $t \in \Omega$, tal que:

$$\sum_{b \in B} f(b|t) = 1, \ \forall t \in \Omega.$$

Finalmente, define-se L como o conjunto de todas essas probabilidades, isto é:

$$L = \{f : \Omega \to \Delta(X)\}.$$

Para qualquer estado $t \in \Omega$ e qualquer probabilidade $f \in L$, $f(\cdot|t)$ representa uma distribuição de probabilidade ao longo do conjunto X (de prêmios), atribuída por f no estado t. O que pode ser expresso pela equação:

$$f(\cdot|t) = (f(b|t))_{x \in X} \in \Delta(X).$$

Aqui, cada número real $f(x|t)$ é interpretado como a probabilidade condicional de se obter o prêmio x na probabilidade f e no estado da natureza t. Considerando 'prêmio' qualquer conjunto de bens que constitua uma alocação específica de interesse para o agente decisório.

O verdadeiro estado da natureza, conhecido pelos agentes decisórios, é denominado de 'evento' e é um subconjunto não vazio de Ω. Considere-se o conjunto de todos os eventos e chamado de Ξ, tal que:

$$\Xi = \{S | S \subseteq \Omega \text{ e } S \neq \emptyset\}.$$

Portanto, para qualquer função $f \in L$ e para qualquer evento $S \in \Xi$, pode-se escrever $f \succsim_s g$ nas situações em que a probabilidade f é, pelo menos, tão preferível quanto g para o agente decisor, se ele soubesse que o estado de natureza é um elemento do conjunto S. Ou seja, $f \succsim_s g$ *sse* o tomador de decisão estiver disposto a escolher f quando ele precisa escolher entre f e g e sabe apenas que o evento S ocorreu. Dadas estas relações ($\succsim_S$), definem-se as relações ($\succ_S$) e ($\sim_S$) de modo que:

$$f \sim_S g \text{ sse } f \succsim_S g \text{ e } g \succsim_S f;$$

$$f \succ_S g \text{ sse } f \succsim_S g \text{ e } g \not\sim_S f.$$

Isso significa que $f \sim_S g$ indica que o tomador de decisão seria indiferente entre f e g se ele tivesse que escolher entre eles após saber S; e $f \succ_S g$ indica que ele preferiria estritamente f sobre g nesta situação.

Deste modo, pode-se escrever $\geq$, $\succ$, e $\sim$ para $\succsim_\Omega$, $\succ_\Omega$, e $\sim_\Omega$, respectivamente. Ou seja, quando nenhum evento condicionante é mencionado, deve-se assumir que refere-se às preferências prévias antes que quaisquer estados em Ω sejam excluídos por observações.

2.1.2 Axiomas da Teoria da Decisão

Ainda será usada a descrição de Myerson (1991), deste modo, os axiomas básicos da teoria da decisão são aqueles que as preferências de um tomador de decisões racional devem satisfazer e eles podem ser apresentados como uma lista de axiomas. Segundo Roger Myerson, a menos que seja indicado de outra forma, espera-se que estes axiomas se apliquem a todas as loterias e, f, g, e h em L, para todos os eventos S e T em Ξ, e para todos os números α e β entre 0 e 1.

Os Axiomas 1.1A e 1.1B afirmam que as preferências devem sempre formar uma ordem transitiva completa sobre o conjunto de loterias.

AXIOMA 1.1A (COMPLETUDE). Para qualquer f e g, ou $f \succsim_s g$ ou $g \succsim_s f$.

AXIOMA 1.1B (TRANSITIVIDADE). Se $f \succsim_s g$ e $g \succsim_s h$, então $f \succsim_s h$.

É simples verificar que o Axioma 1.1B implica vários outros resultados de transitividade, como se $f \sim_s g$ e $g \sim_s h$, então $f \sim_s h$; e se $f \succ_s g$ e $g \succsim_s h$, então $f \succ_s h$.

O Axioma 1.2 afirma que apenas os estados possíveis são relevantes para o tomador de decisões, portanto, dado um evento S, ele seria indiferente entre duas loterias que diferem apenas em estados fora de S.

AXIOMA 1.2 (RELEVÂNCIA). Se $f(t) = g(t)$ para todo $t \in S$, então $f \sim_s g$.

O Axioma 1.3 assegura que uma maior probabilidade de obter uma loteria melhor é sempre mais preferível.

AXIOMA 1.3 (MONOTONICIDADE). Se $f \succ_s h$ e $0 \leq \beta \leq \alpha \leq 1$, então $\alpha f + (1-\alpha)h \succ_s \beta f + (1-\beta)h$.

Baseado no Axioma 1.3, o Axioma 1.4 afirma que à medida que γ aumenta, $\gamma f + (1-\gamma)h$ se torna continuamente melhor, de forma que qualquer loteria situada entre f e h é tão boa quanto alguma randomização entre f e h.

AXIOMA 1.4 (CONTINUIDADE). Se $f \succsim_s g$ e $g \succsim_s h$, então existe algum número γ tal que $0 \leq \gamma \leq 1$ e $g \sim_s \gamma f + (1-\gamma)h$.

Os axiomas de substituição, também conhecidos como axiomas de independência ou axiomas de decisão certa, são provavelmente os mais importantes neste sistema, no sentido de que impõem restrições fortes sobre a forma das preferências do tomador de decisões, mesmo sem a presença dos outros axiomas. Eles também devem ser

axiomas muito intuitivos. Expressam a ideia de que, se o tomador de decisões deve escolher entre duas alternativas e existem dois eventos mutuamente exclusivos, um dos quais deve ocorrer, tal que ele prefira a primeira alternativa em cada evento, então ele deve preferir a primeira alternativa antes de saber qual evento ocorrerá. (Caso contrário, ele estaria expressando uma preferência que certamente desejaria reverter após saber qual desses eventos era verdadeiro!) Nos Axiomas 1.5A e 1.5B, esses eventos são randomizações objetivas em um processo de seleção de loteria, conforme discutido anteriormente. Nos Axiomas 1.6A e 1.6B, esses eventos são desconhecidos subjetivos, componentes do Ω.

AXIOMA 1.5A (SUBSTITUIÇÃO OBJETIVA). Se $e \succsim_s f$ e $g \succsim_s h$ e $0 \leq \alpha \leq 1$, então $\alpha e + (1-\alpha)g \succsim_s \alpha f + (1-\alpha)h$.

AXIOMA 1.5B (SUBSTITUIÇÃO OBJETIVA ESTRITA). Se $e \succ_s f$ e $g \succsim_s h$ e $0 < \alpha \leq 1$, então $\alpha e + (1-\alpha)g \succ_s \alpha f + (1-\alpha)h$.

AXIOMA 1.6A (SUBSTITUIÇÃO SUBJETIVA). Se $f \succsim_s g$ e $f \succsim_s \tau g$ e $S \cap T = \emptyset$, então $f \succsim_{s \cup \tau} g$.

AXIOMA 1.6B (SUBSTITUIÇÃO SUBJETIVA ESTRITA). Se $f \succ_s g$ e $f \succ_\tau g$ e $S \cap T = \emptyset$, então $f \succ_{s \cup \tau} g$.

O Axioma 1.7 declara que o tomador de decisões nunca é indiferente a todos os prêmios. Este axioma é essencialmente uma condição de regularidade, assegurando que sempre há algo de relevante que possa ocorrer em cada estado.

AXIOMA 1.7 (INTERESSE). Para cada estado t em Ω, existem prêmios y e z em X tais que $[y] \succeq_t [z]$. Este axioma é uma condição de regularidade, garantindo que haja algo de interesse que possa ocorrer em cada estado.

AXIOMA 1.8 (NEUTRALIDADE DE ESTADO). Ele afirma que o tomador de decisões possui a mesma ordenação de preferências sobre jogos objetivos em todos os estados do mundo. Se este axioma não se verificar, é porque o mesmo prêmio pode ser valorizado de maneira diferente em diferentes estados.

AXIOMA 1.8 (NEUTRALIDADE DE ESTADO). Para quaisquer dois estados r e t em Ω, se $f(\cdot|r) = f(\cdot|t)$ e $g(\cdot|r) = g(\cdot|t)$ e $f \succeq_r g$, então $f \succeq_t g$.

Conforme afirmado inicialmente, segue-se a exposição de Myerson (1991), apesar de saber que outros autores possuem uma lista um pouco distinta dos axiomas, como Fishburn (1970).

2.1.3 Informação Privada x Comum

A teoria dos leilões, em grande parte da literatura existente (Cabal, Ferreira e Dias, 2016; Cassidy, 1967; Chwe, 1989; Fullerton e Mcafee, 1999; Klemperer, 1999; Lopomo, 1998; Mcafee e McMillan, 1987; Milgrom e Weber, 1982), analisa o modelo de valores privados independentes. Nesse modelo, cada licitante conhece o valor do objeto para si próprio, mas ignora o valor do objeto para os demais licitantes — isso é conhecido como a suposição de valores privados. Os valores são modelados como sendo independentemente retirados de alguma distribuição contínua. Pressupõe-se que os licitantes se comportem de maneira competitiva; assim, o leilão é considerado um jogo não cooperativo entre eles.

Ao escolher um leilão que se aplique o modelo de valores privados independentes, duas suposições são necessárias: que cada licitante conhece seu próprio valor e que esses valores sejam estatisticamente independentes. Elimina-se a possibilidade de que vários licitantes possam ter informações relevantes sobre a qualidade do bem.

Em contraste com os valores privados, o valor comum permite dependência estatística entre as estimativas de valor dos licitantes, mas não contempla diferenças nos gostos individuais.

Ao lidar com lances, esses dois paradigmas se destacam na avaliação do valor de um item. No mundo das licitações, é raro encontrar avaliações puras. Quase todas as negociações relacionadas a licitações misturam o valor privado (custo real de produção/entrega de bens ou serviços) e o valor comum, no qual incertezas sobre a avaliação ou competição do licitante influenciam sua percepção de valor. Ao desenhar leilões, é essencial considerar essa incerteza de avaliação e a divisão entre os componentes de valor privado e comum, pois eles podem determinar o sucesso do leilão.

2.1.4 Tipos de Leilões Clássicos

Existem quatro tipos básicos de leilões quando um único item está à venda: o leilão de lances ascendentes (também chamado de leilão aberto ou inglês), o leilão de lances descendentes (também chamado de leilão holandês), o leilão de primeiro preço com lances selados e o leilão de segundo preço com lances selados (também chamado de leilão Vickrey) (Milgrom e Weber, 1982; Cassidy, 1967; Klemperer, 1999; McAfee e McMillan, 1987). Dentre eles, o leilão inglês é o mais conhecido e mais amplamente utilizado. Uma vez que o leilão inglês é o formato de leilão mais comum na internet, este trabalho visa explorar a relação entre os principais atributos e o preço final para os leilões ingleses.

Leilão Inglês

O leilão inglês é um dos formatos de mais comuns e é caracterizado por sua natureza aberta e competitiva. Neste tipo de leilão, um item é oferecido para venda e os licitantes fazem ofertas públicas e progressivamente mais altas. O leilão começa com um preço mínimo estabelecido pelo vendedor, e os licitantes então aumentam suas ofertas em incrementos definidos. Todos os lances são feitos de maneira aberta, de forma que cada participante pode ouvir o lance dos outros, promovendo assim uma competição direta. O processo continua até que nenhum licitante esteja disposto a oferecer um preço mais alto pelo item. O licitante que fez a oferta mais alta quando o leilão é encerrado é o vencedor e compra o item pelo preço ofertado. No caso do leilão reverso, é para a compra e não para a venda, assim os lances são decrescentes, como ocorre nas licitações, mas é a mesma lógica.

Leilão Holandês

Originado do mercado de flores holandês do século XVII, este formato de leilão, também conhecido como leilão de relógio, possui características distintas quando comparado ao leilão inglês. Em um leilão holandês, o vendedor define o preço inicial. Após um intervalo de tempo pré-definido, o preço cai gradativamente até que um licitante aceite o preço, encerrando o leilão. Esse formato de leilão é decidido com apenas um lance – o primeiro fornecedor a fazer uma oferta "vence" o leilão.

Leilão de Proposta Selada

Os leilões de proposta selada são os modelos clássicos das concorrências, usado em concessões e grande contratos sujeitos à regulamentação governamental, como na indústria petrolífera. Os licitantes submetem duas propostas seladas: uma com sua habilitação e outra com sua oferta comercial. Apesar de serem abertos em momentos diferentes, no âmbito da teoria dos jogos, esta é considerada uma jogada simultânea, já que os jogadores não têm informações sobre as ações dos outros. É um leilão de lances fechados com regra de alocação para quem indica o maior (primeiro) preço.

Leilão de Vickrey

O Leilão Vickrey, cujo nome é uma referência ao economista William Vickrey, é um formato que destaca a estratégia de premiar o vencedor com o segundo menor lance, incentivando a licitação honesta, ou seja, os licitantes não precisam usar a margem de risco para depreciar o seu lance. Ele pode usar seu valor estimado da forma verdadeira.

Nesse leilão, todos os licitantes submetem uma proposta selada simultaneamente, sem conhecimento das ofertas dos outros. O vencedor é determinado pelo lance mais

baixo, mas o preço de venda é o segundo lance mais baixo. Portanto, o licitante está garantido de receber um preço melhor do que o seu próprio lance. Cada licitante é incentivado a licitar honestamente, pois são garantidos um lucro na diferença entre seu próprio lance e o segundo lance mais baixo.

O Leilão de Vickrey apresenta vantagens significativas que promovem uma licitação com um mecanismo de revelação de informação mais eficiente. Uma delas é o incentivo à licitação mais transparente com relação à valoração, visto que a estratégia dominante para os licitantes é oferecer seu verdadeiro valor de reserva.

Aqui, nessa dissertação, pretende-se analisar os caso concretos do modelo da legislação brasileira, em especial, a concorrência e o pregão. Lembrando que com a nova lei de licitações, a dispensa eletrônica é similar a um pregão mais ágil (prazos mais curtos).

2.2 Custos de Transação

No entanto, é importante observar que os próprios leilões envolvem custos, como taxas de inscrição ou comissões, que são uma consideração crucial tanto para vendedores quanto para compradores. Assim, embora os leilões possam oferecer uma forma eficiente de transação, eles também exemplificam como diferentes arranjos institucionais e de mercado influenciam os custos de transação. Ou seja, os leilões estão intrinsicamente ligados ao conceito de custos de transação na economia.

Os economistas vêem os preço como uma relação de troca entre dois bens escassos e a escolha do agente é realizada de modo a otimizar seu nível de satisfação. As premissas fundamentais utilizadas nessas análises são: 1) As pessoas respondem a incentivos; 2) O ordenamento jurídico é um meio de criar estes incentivos; 3) Os agentes objetivam maximizar suas utilidades, agindo racionalmente. As ferramentas apresentadas pela economia para esta abordagem são conhecidas como instrumentos baseados no mercado, pois nele se inspiram.

A análise de preços na microeconomia (Varian, 1992; Varian 2003; Eaton e Eaton, 1999; Mas-Colell et al., 1995; Pindick e Rubinfeld, 2005; Kreps, 1990; e Kreps, 2004) examina a interação entre compradores e vendedores em um ambiente denominado mercado. Dentro desse contexto, os preços emergem no ponto em que a oferta encontra a demanda, conhecido como equilíbrio de mercado.

Segundo essa análise, o preço ideal de mercado para um bem ou serviço é alcançado quando o valor que os consumidores estão dispostos a pagar iguala-se ao custo marginal de produção do fornecedor. Este conceito é particularmente relevante em mercados com numerosos compradores e vendedores. No entanto, a realidade do mercado é complexa devido a barreiras como localização, economia e regulamentações, que influenciam a participação no mercado, especialmente em contextos como

licitações.

Os custos transacionais desempenham um papel crucial na teoria econômica, destacando as despesas associadas à execução de uma transação — isto é, o processo de transferir bens ou serviços de uma parte para outra (Yeung e Camelo, 2023). Essas despesas podem abranger os custos relacionados à negociação, medição, monitoramento, segurança e administração. A relevância desses custos se manifesta no impacto que têm sobre a eficiência com que indivíduos e empresas conseguem alocar recursos.

Os custos transacionais são geralmente divididos em dois tipos: custos fixos e custos variáveis (Yeung e Camelo, 2023). Custos fixos são aqueles que se mantêm inalterados independentemente do número ou escala das transações, como as taxas para a elaboração de um contrato padrão por um advogado. Por outro lado, custos variáveis são aqueles que oscilam com o volume das transações, como os custos de envio de cartas para diversos compradores.

Ademais, os custos de transação estão sujeitos a fatores institucionais, incluindo leis e regulamentos que orientam os procedimentos transacionais (Yeung e Camelo, 2023). Por exemplo, regulamentações excessivamente rigorosas podem aumentar as despesas transacionais, enquanto uma regulamentação inadequada pode resultar em deficiências de confiança, elevando assim os custos transacionais.

A relevância da reputação no contexto das contratações tem sido extensivamente estudada na literatura corporativa (privada), demonstrando a influência significativa que a percepção de um ente como mal pagador pode ter sobre os preços cobrados em contratações. Walker (2010) destaca que a reputação corporativa é um ativo intangível, mas com implicações tangíveis sobre as medidas de desempenho e expectativas das partes interessadas. Esta visão é corroborada por Basdeo et al. (2006), que evidenciam o impacto das ações de mercado na reputação de uma firma e, consequentemente, em sua performance financeira.

Chu e Chu (1994) discutem como a qualidade dos produtos é sinalizada através da associação com varejistas de renome, um conceito que pode ser paralelamente aplicado à contratação pública; entidades com reputações mais fortes podem exigir preços mais altos devido à qualidade percebida. Stickel (1992) aborda a reputação no contexto de analistas de segurança, notando que a reputação pode ser um fator preponderante sobre o desempenho financeiro, um argumento que se estende ao âmbito das entidades públicas quando consideram-se seus hábitos de pagamento.

Góis et al. (2020) fornecem insights sobre a ligação entre reputação corporativa e risco de falência, sugerindo que uma reputação negativa pode ser um indicativo de riscos financeiros maiores, o que poderia levar fornecedores a aumentar os preços como uma forma de compensação pelo risco. Ahn e Choi (2009) ilustram o papel do monitoramento bancário na governança corporativa e como a gestão dos lucros

dos mutuários pode ser um sinal de reputação, o que tem implicações diretas na confiança e nas condições impostas pelos credores, incluindo o governo.

Trabalhos anteriores, como o de Beaver (1966), já apontavam para a capacidade de indicadores financeiros atuarem como preditores de falência, enquanto Butler e Fauver (2006) investigam como o ambiente institucional afeta as classificações de crédito soberanas, um reflexo da reputação de um país no cumprimento de suas obrigações financeiras. Ambos os estudos sugerem que uma reputação de mal pagador pode levar a condições mais onerosas em futuras contratações, pois os credores buscam compensar o risco elevado.

No âmbito dos mercados eletrônicos, Jolivet, Jullien e Postel-Vinay (2016) demonstram como a reputação pode influenciar os preços, com evidências de uma plataforma francesa. Baghai e Becker (2020) vinculam a reputação às classificações de crédito, mostrando como as reputações podem alterar a percepção de risco e as condições de crédito. Anginer et al. (2015) reforçam essa visão, argumentando que a reputação de uma firma afeta diretamente o custo do capital de dívida.

Assim, parece ser indiscutível que a reputação impacta os preços (de compra e de venda), mas surge a necessidade de definir o que seria a 'reputação'. Gotsi e Wilson (2001) procuram uma definição, entendendo-a como uma percepção coletiva que reflete a imagem histórica de uma empresa, colocando elementos objetivos, subjetivos e comparativos. Wartick (2002) avança na medição da reputação corporativa, enfatizando a necessidade de dados confiáveis para avaliar a reputação adequadamente. Esses dados são essenciais para entender como a reputação de uma entidade pública pode influenciar os custos de transação.

Esses estudos coletivamente ilustram uma realidade em que a reputação de um ente público como mal pagador parece afetar significativamente o custo das contratações públicas. O aumento dos custos de transação, uma forma de compensação pelo risco percebido, é uma consequência direta dessa reputação negativa.

No contexto de aquisições estatais, os custos de transação estão intrinsecamente ligados a normas e procedimentos que, embora visem a clareza e a justiça do processo, geram gastos prévios com a criação de contratos e a coleta de informações, além de despesas posteriores com a fiscalização e as consequências de não cumprimento das normas (Yakovlev et al., 2018). Apesar da ausência de dados precisos sobre o peso desses custos nas aquisições do governo brasileiro, a preocupação com essa questão é amplamente reconhecida no meio acadêmico e profissional (Libório et al., 2021).

Parte significativa dos custos de transação das contratações públicas nascem nas cláusulas contratuais exorbitantes. Conforme explicado por Cretella Jr (1989), o Governo possui uma diversidade de privilégios que o eleva a um patamar superior se comparado aos cidadãos comuns. Esse leque de benefícios é chamado de "cláu-

sulas exorbitantes", pois ultrapassam o padrão habitual dos contratos privados, atribuindo predominância à Administração. Um exemplo clássico desse privilégio é a possibilidade de atraso no pagamento contratual.

Assim, alguns autores (Niebuhr e Oliveira, 2018; Camelo, Nobrega e Torres, 2022) sugerem que cláusulas contratuais onerosas elevam os custos para os fornecedores, influenciando assim os preços finais. Essas cláusulas resultam em um desequilíbrio entre os deveres e direitos de compradores e fornecedores (Pereira e Nóbrega, 2012), e aumentam os custos de transação à medida que os fornecedores calculam seus preços considerando essas condições excessivas (Melo Barros e Barros, 2014).

Finalmente, valorizar a reputação dos fornecedores nas licitações públicas é visto como um mecanismo promotor de condutas éticas e confiança mútua entre as partes envolvidas (Spagnolo e Castellani, 2017). Os preços em economia refletem expectativas condicionadas a eventos específicos, significando que o valor esperado de um preço é baseado na ocorrência de um certo fato (Cooter e Ullen, 2010). Assim, o preço em qualquer momento espelha as informações públicas disponíveis, mas não obrigatoriamente as privadas, incorporando todos os custos de transação.

Assim, destaca-se a ausência de estudos empíricos sobre a reputação criando custos de transação nas contratações públicas. O único estudo encontrado é o de Libório et al. (2022), mas com limitações metodológicas importantes, pois o critério de reputação foi construído pela opinião de uma única empresa, no estado de Minas Gerais, de modo subjetivo.

2.3 Estatística Clássica e Aprendizado de Máquina

A fundamentação teórica do aprendizado de máquina, ou machine learning, perpassa a estatística, ciência da computação e inteligência artificial. Uma influente contribuição para este campo foi feita por Leo Breiman (2001) em seu trabalho "Statistical Modeling: The Two Cultures", que delineia duas abordagens principais na modelagem estatística: a tradicional, focada em modelos baseados em teorias e pressupostos, e a orientada por algoritmos, característica do aprendizado de máquina. Enquanto a modelagem estatística busca explicar fenômenos e estabelecer relações causais, o aprendizado de máquina prioriza a capacidade preditiva, mesmo que isso signifique usar modelos sem interpretação teórica direta.

Breiman argumentou pela importância de avaliar modelos mais pela sua eficácia preditiva do que pela precisão na descrição de relações entre variáveis. Esta perspectiva é central no aprendizado de máquina, que adota uma abordagem empírica, focada em resultados. Técnicas como redes neurais e máquinas de vetor de suporte, embora não tenham uma interpretação teórica tão clara quanto os modelos estatísticos tradicionais, são exemplos de como o aprendizado de máquina evoluiu para se

concentrar na precisão e utilidade prática.

A regressão para inferência é usada para entender a relação entre uma variável de resposta e uma ou mais variáveis explicativas. Nessa abordagem, o objetivo é determinar a natureza e a força dessa relação, bem como identificar quais variáveis explicativas são mais importantes na explicação da variação da variável de resposta. Essa análise é geralmente baseada em testes de hipóteses estatísticas e na obtenção de intervalos de confiança para os parâmetros do modelo (Izbicki e Santos, 2020).

Ao construir um referencial teórico robusto sobre regressão e aprendizado de máquina, precisa-se delinear claramente a distinção entre variáveis dependentes e independentes e sua aplicação em modelos preditivos. Conforme descrito por Seber e Lee (2012), a regressão lida com a estimação da relação funcional entre uma variável dependente contínua e um ou mais preditores.

Em termos matemáticos, a função de regressão $r(x) = \mathbf{E}[Y|X = x]$, é a expectativa condicional da variável resposta Y, dado o vetor de covariáveis X, e é uma ferramenta fundamental na previsão de valores não observados (Searle, 1997). O processo de treinamento de um modelo de aprendizado de máquina, como articulado por James et al. (2017), envolve a estimativa dessa função de regressão, denotada por $\hat{r}(x)$, a partir de dados existentes para fazer previsões acuradas sobre novas observações.

Diferencia-se a abordagem de regressão da classificação, sendo a última aplicável quando a variável resposta é qualitativa. Murphy (2022) esclarece que, enquanto a regressão busca prever um valor contínuo, a classificação objetiva identificar a qual categoria um novo ponto de dados pertence. Além disso, a escolha da função de perda adequada para a avaliação do desempenho de um modelo é essencial, sendo a função de perda quadrática L_2 e a função de perda absoluta L_1 exemplos notáveis discutidos por James et al. (2017). Estas são utilizadas para quantificar o erro entre as previsões do modelo e os valores verdadeiros, influenciando diretamente a construção de algoritmos de aprendizado de máquina eficientes.

James et al. (2017) também expõem a importância da função de risco preditivo, $R_{pred}(g) = \mathbf{E}[(Y - g(X))^2]$, no processo de avaliação e seleção de modelos de regressão. Esta função avalia a qualidade de um estimador g, baseando-se na esperança matemática do quadrado do desvio entre a predição e o valor real. O princípio subjacente é que um modelo ideal minimizará esta função de risco, levando a previsões que são tão próximas quanto possível dos valores verdadeiros.

Ao examinar esses conceitos no contexto da aprendizagem de máquina, a dissertação se beneficia enormemente das explanações do livro e dos vídeos de Izbicki e Santos (2020), que fornecem uma base sólida para a compreensão de modelos preditivos e suas aplicações.

Da mesma forma, Izbicki e Santos (2020) abordaram a questão da função de risco

das funções de predição, apontando que quanto menor o risco, melhor a função de predição "g()". Em geral, as funções de perda mais usadas são as quadráticas ou as absolutas. Após, analisaram o trade-off entre superajuste (*overfitting*) e o subajuste (*underfitting*), decorrentes da seleção dos modelos.

A seleção de uma função de perda adequada é essencial na construção de modelos de aprendizado de máquina. Conforme elucidado por Murphy (2022), uma função de perda $L(\cdot)$ é usada para quantificar o custo de erros na predição de um modelo. Propriedades importantes de $L(\cdot)$, tais como

$$L(0) \leq L(u) \leq L(v) \forall 0 < u < v,$$

garantem que a função seja uma medida confiável de discrepância (Murphy, 2023). Dentre as funções de perda comumente utilizadas, a função de perda quadrática $L(u) = u^2$ e a função de perda absoluta $L(u) = |u|$ são destacadas por sua eficácia em diferentes cenários de regressão.

No contexto de regressão, a minimização do risco preditivo $R_{pred}(g)$, na qual g é uma função de predição, equivale a encontrar a função de regressão $r(x)$ que melhor se ajusta aos dados, como descrito pela equação $R_{pred}(g) = \mathbb{E}[(Y - g(X))^2]$. Izbicki e Santos (2020) também abordam um teorema fundamental que estabelece uma relação entre o risco preditivo $R_{pred}(g)$ e o risco de um estimador de regressão $R_{reg}(g)$, mostrando que $R_{pred}(g)$ é composto por $R_{reg}(g)$ e a variância média do modelo.

Além disso, Murphy (2022) salienta a importância do Erro Quadrático Médio (EQM) em aprendizado de máquina, que é frequentemente usado para estimar o risco preditivo sobre uma nova amostra de observações.

A precisão de um método de aprendizado estatístico aplicado à regressão é frequentemente medida pelo erro quadrático médio (EQM ou *MSE - Mean Squared Error*), que avalia a proximidade das previsões do modelo aos valores observados. Um MSE pequeno indica que as previsões do modelo estão muito próximas das respostas verdadeiras, enquanto um valor grande sugere discrepâncias substanciais (James et al., 2017; Izbicki e Santos, 2020). O Erro Quadrático Médio é calculado como

$$EQM = \frac{1}{m} \sum_{i=1}^{m} (Y_{n+i} - g(X_{n+i}))^2,$$

sendo $(X_{n+1}, Y_{n+1}), \ldots, (X_{n+m}, Y_{n+m})$ as novas observações.

A busca por métodos que minimizem o EQM, e consequentemente o R, é uma das principais metas em aprendizado de máquina e é essencial para a criação de modelos com um bom desempenho preditivo (Murphy, 2022).

A escolha da função de perda em métodos de aprendizado de máquina é uma

decisão estratégica importante. A função de perda quadrática, frequentemente designada por L_2, é valorizada por penalizar de forma mais intensa os erros maiores, uma vez que o custo aumenta exponencialmente com o erro, característica menos acentuada na função de perda absoluta L_1 (James et al., 2017). Esta função L_2 também revela maior sensibilidade a *outliers*, oferecendo uma penalidade mais substancial em comparação com L_1 (James et al., 2017).

Murphy (2022) destaca que quando os erros seguem uma distribuição normal, a estimação por mínimos quadrados torna-se uma solução de máxima verossimilhança, resultando em estimativas lineares não viesadas com a menor variância possível. A função de perda quadrática também se beneficia de ser diferenciável, o que simplifica o processo de otimização (Murphy, 2022). Na cultura de machine learning, deve-se desconsiderar ϵ, não fazendo quais suposições sobre o erro.

Em modelos de regressão linear múltipla, a notação matricial é usada para simplificar a representação do modelo. O modelo é frequentemente descrito como

$$\mathbf{Y} = g(\mathbf{X}) = \boldsymbol{\beta}^\top \mathbf{X} + \boldsymbol{\epsilon},$$

com $\boldsymbol{\beta}$ sendo o vetor de coeficientes estimados pela minimização da soma dos quadrados dos resíduos, resultando em $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}$. A função de regressão estimada é dada por $g(\mathbf{x}) = \hat{\boldsymbol{\beta}}^\top \mathbf{X}$ (Wooldridge, 2000).

Muitos trabalhos estatísticos tradicionais focam em demonstrar que o método dos mínimos quadrados é um estimador de máxima verossimilhança eficiente. Esses estudos abordam não só a estimação de parâmetros, como também o desenvolvimento de testes de aderência, métodos para construção de intervalos de confiança e análises de resíduos. A suposição de que a verdadeira regressão $r(\mathbf{X})$ é igual ao valor esperado condicional de Y dado $\mathbf{X}$, isto é, $r(\mathbf{X}) = \mathbb{E}[\mathbf{X}|Y]$, é bastante forte e, em muitos casos, não é uma suposição realista. No entanto, Wooldridge (2000) destaca que ainda há justificativas para a aplicação do método dos mínimos quadrados na estimação dos coeficientes, mesmo quando a verdadeira relação de regressão $r(\mathbf{X})$ não satisfaz a linearidade.

Nas situações em que a linearidade é uma suposição questionável, existe a possibilidade de identificar um vetor $\boldsymbol{\beta}_*$ tal que a função $g_{\boldsymbol{\beta}_*}(\mathbf{X}) = \boldsymbol{\beta}_*^\top \mathbf{X}$ apresenta um bom poder preditivo. Nesses casos, os mínimos quadrados tendem a fornecer estimadores com risco baixo, convergindo para o melhor preditor linear, conhecido como o oráculo $\boldsymbol{\beta}_*$ (Izbicki e Santos, 2020). O oráculo $\boldsymbol{\beta}_*$ é definido como

$$\boldsymbol{\beta}_* = \arg\min_{\boldsymbol{\beta}} R(g_{\boldsymbol{\beta}}) = \arg\min_{\boldsymbol{\beta}} \mathbb{E}[(Y - \boldsymbol{\beta}^\top \mathbf{X})^2],$$

o que enfatiza que, mesmo sem a linearidade da regressão verdadeira $r(\mathbf{X})$, a técnica dos mínimos quadrados permanece relevante.

Para não ter o problema de selecionar um modelo super-ajustado (perfeito, pois usa todos os dados da amostra), separou-se os dados (*data splitting*[1]) em dois conjuntos *treinamento* e *teste*[2].

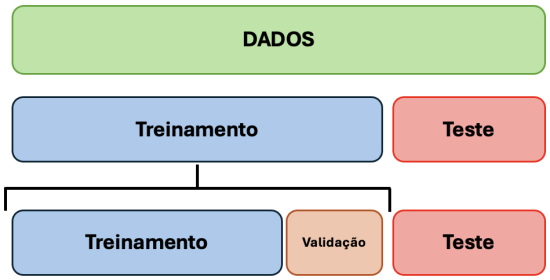

Figura 2.1: Separação do conjunto de dados. Fonte: Marinho, 2023, p. 88.

A seleção de observações para os conjuntos de treino e teste é conduzida de maneira aleatória, frequentemente utilizando estratificação baseada em certas variáveis para garantir representatividade. A randomização serve como uma estratégia para mitigar quaisquer viéses associados à ordenação dos dados. O objetivo é assegurar que ambos os conjuntos, de treino e de teste, exibam uma ampla gama de exemplos, refletindo a heterogeneidade do conjunto de dados completo.

Considera-se um processo de divisão de dados em que se dispõem de um conjunto completo com n instâncias. Uma fração s é selecionada de forma aleatória para formar o conjunto de treinamento, deixando o restante para o conjunto de teste. Esta prática busca preservar a representatividade e a diversidade em ambos os conjuntos, evitando quaisquer tendências ou ordenações prévias que possam influenciar o aprendizado do modelo.

A estratégia de validação cruzada é considerada uma melhoria significativa em relação ao simples particionamento de dados, especialmente por sua capacidade de utilizar eficientemente todas as observações disponíveis para treinamento e validação do modelo. Em particular, a técnica de validação cruzada leave-one-out (LOOCV) é uma forma extrema de validação cruzada, sendo cada amostra usada individualmente como um conjunto de teste, enquanto o modelo é treinado com o restante dos dados (Izbicki e Santos, 2020). Matematicamente, o procedimento para o LOOCV é descrito como:

$$\hat{R}(g) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - g_{-i}(X_i))^2,$$

[1]Uma variação desse procedimento é a validação cruzada, que usa toda a amostra, mas deixa espaços para testes não viesados.

[2]Em geral, usa-se cerca de 70% dos dados para treino e 30% para teste, mas isso varia de acordo com a quantidade de dados. O que é muito importante é que essa separação seja realizada de modo aleatório.

em que g_{-i} é o modelo treinado sem a i-ésima observação. Esta técnica garante que cada ponto de dados seja utilizado para testar a generalização do modelo, proporcionando uma estimativa abrangente do desempenho do modelo (Izbicki e Santos, 2020).

Quando o número de observações é grande, a abordagem *k-fold* cross-validation é frequentemente preferida. Esta abordagem divide o conjunto de dados em k subconjuntos distintos e realiza o treinamento e a validação k vezes, cada vez com um subconjunto diferente como conjunto de teste e os restantes como conjunto de treinamento (Murphy, 2022; Marinho, 2023). O risco do modelo é então calculado pela média dos resultados de cada um dos k *fold*s, fornecendo uma estimativa robusta do erro, dada pela fórmula:

$$\hat{R}(g) = \frac{1}{k}\sum_{j=1}^{k}\frac{1}{n_j}\sum_{i\in L_j}(Y_i - g_{-j}(X_i))^2,$$

de modo que L_j é o conjunto de índices do j-ésimo *fold* e g_{-j} é o modelo treinado sem as observações do j-ésimo *fold*. Ao aplicar a validação cruzada *k-fold*, pode-se alcançar um equilíbrio entre eficiência computacional e precisão na avaliação do modelo (Izbicki e Santos, 2020; Marinho, 2023).

Modelos avançados em aprendizado de máquina frequentemente utilizam hiperparâmetros, que são ajustes finos no algoritmo que não são diretamente aprendidos a partir dos dados. Estes são tipicamente selecionados através de um processo de validação cruzada, que é uma extensão do método de divisão dos dados, e visa proporcionar uma avaliação mais confiável do modelo através do cálculo do risco $R(g)$ (Izbicki e Santos, 2020). Este procedimento envolve a criação de várias partições dos dados em conjuntos de treino e validação para testar a generalização do modelo.

Na prática, realiza-se um grid search, explorando uma gama de possíveis valores para os hiperparâmetros. O modelo que apresenta o menor Erro Quadrático Médio (EQM) no conjunto de validação é selecionado como o hiperparâmetro ideal. O modelo é então treinado com o conjunto completo de treinamento utilizando este hiperparâmetro selecionado, com o objetivo de minimizar o EQM (Izbicki e Santos, 2020).

Para garantir que o modelo seja bem generalizado, utiliza-se o conjunto de teste, que permanece intocado durante o processo de validação cruzada, apenas sendo utilizado para a avaliação final do modelo. Esta abordagem confirma a eficácia do modelo em dados não vistos anteriormente, oferecendo uma estimativa do risco preditivo $R(g)$ (Izbicki e Santos, 2020).

A divisão inicial dos dados em conjuntos de treinamento e validação é um aspecto crítico da validação cruzada. Enquanto o conjunto de treinamento é utilizado para o ajuste do modelo, a validação é empregada para a seleção de hiperparâmetros.

Finalmente, o conjunto de teste é usado para avaliar a performance do modelo final, fornecendo uma medida do risco observado (Izbicki e Santos, 2020). Esse delineamento estruturado assegura que cada conjunto de dados desempenhe um papel específico, contribuindo para uma avaliação robusta do modelo. Como se pode ver na imagem abaixo:

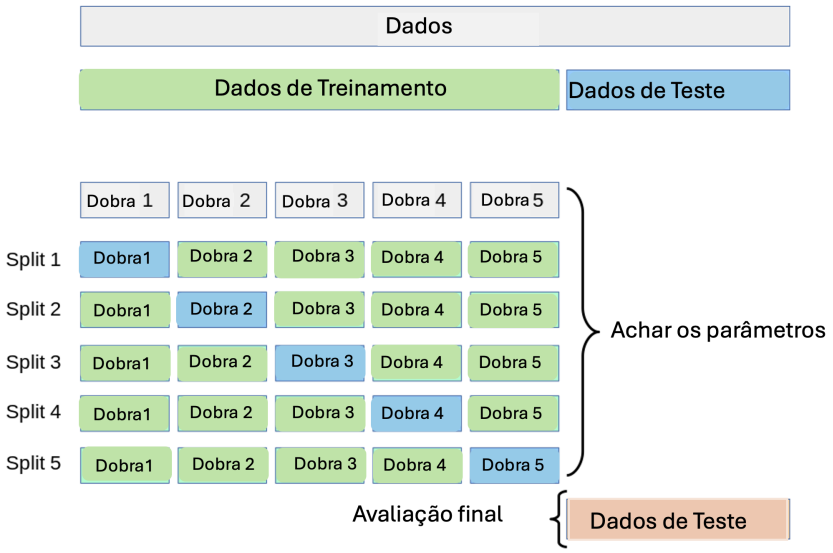

Figura 2.2: Validação Cruzada - *K-fold*.

Avaliar a performance do modelo preditivo g é crucial, e uma métrica comum para essa avaliação é o Erro Quadrático Médio (EQM). O EQM é calculado utilizando o conjunto de teste, que normalmente corresponde a uma parcela do conjunto total, como por exemplo, 30% das observações. Contudo, a proporção escolhida para o teste pode variar de acordo com o contexto específico da análise (Izbicki e Santos, 2020; Murphy, 2022). Utiliza-se o conjunto de dados de treinamento exclusivamente para a estimação de g (por exemplo, estimar os coeficientes em uma regressão linear) e o conjunto de teste é empregado com o objetivo de avaliar $R(\hat{g})$. A estimativa de $R(\hat{g})$, segundo Izbicki e Santos (2020, p.14) é dada por:

$$\hat{R}(\hat{g}) = \frac{1}{n-s} \sum_{i=s+1}^{n} (Y_i - \hat{g}(X_i))^2 \equiv R(\hat{g}),$$

sendo o erro quadrático médio calculado para o conjunto de teste.

Para determinar quais amostras compõem o conjunto de treinamento e quais formam o conjunto de teste, a seleção é feita de forma aleatória, utilizando um gerador de números aleatórios. Como o conjunto de teste não é utilizado na estimação dos parâmetros de $\hat{g}$, o estimador de $R(\hat{g})$ é consistente de acordo com a lei dos grandes números.

O erro quadrático médio mencionado acima é calculado utilizando os dados de

treino que foram empregados para ajustar o modelo, portanto, seria mais apropriado referir-se a ele como EQM de treinamento. Contudo, o interesse primordial não está na eficácia do método nos dados de treinamento. O que realmente valoriza-se é a precisão das previsões geradas pelo método quando aplicado a dados de teste inéditos. A razão dessa preocupação é que, frequentemente, o objetivo é aplicar o modelo aprendido em situações futuras e desconhecidas, o que justifica a ênfase na acurácia sobre novas observações.

Para expressar isso de maneira mais matemática, seguiu-se os passo de Hastie, Tibshirani e Friedman (2009), considerando que se ajustou o método de aprendizado estatístico às observações de treinamento $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ e obtive-se a estimativa $\hat{f}$. Em seguida, calculou-se $\hat{f}(x_1), \hat{f}(x_2), \ldots, \hat{f}(x_n)$. Se esses valores são aproximadamente iguais a $y_1, y_2, \ldots, y_n$, então o EQM de treinamento, conforme definido anteriormente, é pequeno. No entanto, o interesse real não é verificar se $\hat{f}(x_i) \approx y_i$, mas sim determinar se $\hat{f}(x_0)$ está próximo de y_0, sendo (x_0, y_0) uma observação de teste anteriormente não observada e não utilizada para treinar o método de aprendizado estatístico. O objetivo é selecionar o método que resulta no menor MSE de teste, ao invés do menor EQM de treinamento. Em outras palavras, se houvesse um grande número de observações de teste, seria possível calcular a média do erro quadrado de previsão para essas observações de teste (x_0, y_0).

Assim, é o EQM de teste, isto é, aquele calculado em um conjunto de dados previamente não visto, que fornece uma avaliação mais relevante da capacidade preditiva do modelo. Afinal, não se está interessados na precisão do modelo nos dados que já se conhece, mas sim na sua habilidade de prever corretamente novas observações. Porém, estimar o EQM de teste pode ser mais desafiador, já que frequentemente não se dispõe de dados de teste suficientes. Neste contexto, métodos como a validação cruzada são propostos para estimar o MSE de teste a partir dos dados de treino (James et al., 2017).

Tudo que foi tratado sobre o erro quadrático médio, vale para sua raiz, conhecido como Erro Quadrático Médio da Raiz (*Root Mean Squared Error*, RMSE), que é uma espécie de normalização, para sofrer menos impacto de outliers. Mas também encontra-se na literatura o uso de funções de erro usando o erro médio absoluto (*Mean Absolute Error*, MAE) para avaliação de eficiência dos modelos. Ambas são utilizadas para quantificar o erro entre valores previstos por um modelo e os valores observados.

O MAE mede a magnitude média dos erros em um conjunto de previsões, sem considerar sua direção, sendo a média das diferenças absolutas entre previsões e observações reais. Por outro lado, o RMSE é a raiz quadrada da média das diferenças ao quadrado entre as previsões e as observações. O MAE oferece uma medida direta e compreensível das discrepâncias médias em regressões, enquanto o RMSE, ao

aumentar o peso dos erros maiores, é mais sensível a valores atípicos e potencialmente mais informativo sobre grandes erros na previsão.

A escolha entre MAE e RMSE depende do contexto específico da regressão. O RMSE, ao penalizar mais fortemente os grandes erros, pode ser preferível em situações em que estes são particularmente indesejáveis (Hastie, Tibshirani e Friedman, 2009). No entanto, isso pode ser problemático ao comparar resultados de RMSE calculados em diferentes tamanhos de amostras de teste, pois o RMSE tende a aumentar com o aumento do tamanho da amostra. Por outro lado, o MAE oferece uma interpretação mais direta e transparente do erro médio, sendo muitas vezes mais fácil de compreender e comunicar.

Na busca por métricas, deve-se destacar que o uso das medidas AUC (*Area Under the ROC Curve*) e acurácia são tipicamente utilizadas em tarefas de classificação, não de regressão (James et al., 2017; Murphy, 2022). Em um contexto de regressão, como é o caso desta dissertação ao avaliar um modelos para previsão de valores contínuos de custos de transação, essas métricas não se aplicam diretamente, pois são destinadas a avaliar a performance de modelos em diferenciar entre classes categorizadas.

A AUC Mede a capacidade do modelo de distinguir entre classes binárias ou multiclasses. O cálculo da AUC envolve a plotagem da curva ROC, que é uma representação gráfica da sensibilidade versus especificidade para um sistema classificador binário enquanto seu limiar de discriminação é variado (Murphy, 2022). Por outro lado, a Acurácia mede a proporção de predições corretas (tanto verdadeiros positivos quanto verdadeiros negativos) em relação ao total de casos examinados. Para tarefas de regressão, não faz sentido falar em "predições corretas" sem antes definir um critério de correção, como um intervalo de tolerância ao redor do valor predito.

No capítulo 4, será usado prioritariamente o RMSE para avaliar a adequação do modelo de previsão, como recomendado pela literatura de Aprendizagem de máquina.

No desenvolvimento de modelos analíticos, é comum se enfrentar a questão de otimizar um conjunto de d parâmetros, que são bem definidos e sob controle. Nesse contexto, a aplicação de penalidades na função de custo é uma estratégia para impor uma medida de complexidade, auxiliando na obtenção de um modelo equilibrado, que mitigue o trade-off entre viés e variância (Izbicki e Santos, 2020). Esta abordagem é expressa na função de risco $R(g)$, que busca-se minimizar:

$$R(g) \approx EQM(g) + P(g).$$

O objetivo é minimizar $R(g)$ sem incorrer no risco de overfitting devido a um

número excessivo de parâmetros. Enquanto um EQM reduzido é desejável, é fundamental que a penalidade $P(g)$ seja suficientemente grande para evitar a complexidade desnecessária (Izbicki e Santos, 2020).

Entre as penalidades mais conhecidas, encontram-se o Critério de Informação de Akaike (AIC) e o Critério de Informação Bayesiano (BIC), que são calculados como:

$$\text{AIC: } EQM + \frac{2}{nd\hat{\sigma}^2},$$

$$\text{BIC: } EQM + \frac{\log(n)}{nd\hat{\sigma}^2}.$$

Aqui, $\hat{\sigma}^2$ é a estimativa da variância do erro e d é o número de parâmetros no modelo. Tais critérios ajustam o EQM para penalizar modelos com maior número de parâmetros, favorecendo assim modelos mais parcimoniosos (Murphy, 2022; James et al., 2017).

Izbicki e Santos (2020) também abordam a importância da penalização BIC e AIC como um critério para escolha do modelo. O critério de informação bayesiano (BIC) e o critério de informação de Akaike (AIC) são duas medidas muito usadas para selecionar modelos (Schwarz, 1978; Akaike, 1973). Quanto menor seu índice, mais preferidos eles são.

Quando se tem muitos parâmetros, a dificuldade computacional aumenta significativamente, então existem outras técnicas para lidar com essa dificuldade e selecionar o melhor subconjunto de covariáveis. Se for usar o AIC ou BIC será muito custoso "rodar" todos os modelos possíveis.

2.3.1 Algumas técnicas de aprendizado de máquina

Nesta parte da fundamentação teórica, serão apresentadas algumas técnicas de aprendizado de máquina, mas que não foram usadas na parte empírica, estas serão apresentadas em capítulo próprio.

A primeira forma para lidar com isso, é a aplicação técnica de um atalho (heurística) chamado de regressão *stepwise*, escolhendo quais variáveis devem ser incluídas em um modelo de regressão. Ela pode ser de frente para trás (ou de trás para frente). O *forward stepwise* vai acrescentando uma variável por vez, escolhendo o melhor parâmetro, e depois não testa mais as outras em modelos mais complexos. No *Backward* o processo começa do mais complexo e vai removendo uma variável de cada vez até chegar a um modelo final que tenha apenas as variáveis mais importantes (Izbicki e Santos, 2020). O processo do *stepwise* inicia com a avaliação de cada variável individualmente e seleciona aquela que minimiza o risco. Para $j = 1, ..., d$, ajuste a regressao na j-esima variavel Xj. Seja $\hat{R}(g_j)$ o risco estimado desta funcao (usando AIC ou validacao cruzada). Defina :

$$\hat{j} = \arg\min_{j} \hat{R}(g_j), \text{ sendo } S = \{\hat{j}\}.$$

Posteriormente, o modelo é ampliado de forma iterativa, incluindo uma nova variável a cada etapa, aquela que, junto ao conjunto já selecionado S, resulta no menor risco estimado. Assim, o modelo cresce até que a inclusão de novas variáveis não resulte em melhorias significativas. Para cada $j \in S^c$, ajuste a regressão $Y = \beta_j X_j + \sum_{s \in S} \beta_s X_s$, em que $\hat{R}(g_j)$ é o risco estimado desta função. Defina:

$$\hat{j} = \arg\min_{j \in S^c} \hat{R}(g_j), \text{ atualize } S \leftarrow S \cup \{\hat{j}\}.$$

Repete-se os passos anteriores até que todas as variáveis estejam em S ou até quando não seja mais possível ajustar o modelo de regressão. Por fim, seleciona-se o modelo com menor risco estimado.

Esse método tem a vantagem de ser computacionalmente mais viável do que avaliar todas as 2^d combinações possíveis de variáveis. Com forward stepwise, o número de modelos a considerar é drasticamente reduzido para aproximadamente $1 + d(d+1)/2$, simplificando o processo de escolha do modelo que oferece o melhor ajuste (Murphy, 2022).

A escolha do hiperparâmetro é feita selecionando o $\hat{j}$ que apresenta o menor Erro Quadrático Médio (EQM) durante a validação cruzada. Este método pragmático permite uma seleção eficiente de características, mantendo o equilíbrio entre precisão do modelo e complexidade computacional (Murphy, 2022).

Para decidir qual variável deve ser acrescentado (no *forward*) em cada etapa, a regressão stepwise utiliza testes estatísticos que avaliam a melhoria do modelo ao remover cada variável. Se uma variável melhora significativamente o modelo quando é acrescentada, ela fica no modelo. Esse processo continua até que todas as variáveis relevantes sejam colocadas e o modelo final seja alcançado. A regressão stepwise é útil porque permite que o pesquisador construa um modelo de regressão com apenas as variáveis mais importantes, sem testar todos os possíveis modelos.

Nos métodos não-paramétricos, Izbicki e Santos (2020) apresentam KNN (*K-Nearest Neighbors*), Nadaraya Watson. O KNN (K-vizinhos mais próximos, do inglês) é baseado na ideia de que observações semelhantes tendem a ter valores de saída semelhantes. Então se determina um numero K de vizinhos próximos que serão considerados para determinar o valor de saída de uma nova observação e a predição é a média desses K vizinhos.

O procedimento de k-vizinhos mais próximos, ou k-NN, destaca-se como um método essencial no domínio do aprendizado de máquina. Este algoritmo fundamenta-se na hipótese de que elementos semelhantes tendem a agrupar-se no domínio das características (Izbicki e Santos, 2020). Ao abordar a regressão com o k-NN, a meta

é projetar a função de regressão $r(x)$ que mais se aproxima das observações vizinhas de um dado ponto x.

Para uma dada instância x, a estimativa da função-alvo, denotada por $g(x)$, é a média aritmética dos valores de saída dos seus k vizinhos imediatos no espaço das características. A fórmula para $g(x)$, conforme descrito por Izbicki e Santos (2020), é expressa como:

$$g(x^*) = \frac{1}{k} \sum_{i \in N_{x^*}} y_i. \tag{2.1}$$

Neste contexto, N_x representa o conjunto das k amostras mais próximas de x, que é formalmente definido como:

$$N_{x^*} = \left\{ i \in \{1, \ldots, n\} : d(x_i, x^*) \leq d_{x^*}^k \right\}, \tag{2.2}$$

em que $d(x_i, x^*)$ mede a distância entre a amostra x_i e o ponto de interesse x, enquanto d_x^k é a distância até o k-ésimo vizinho mais próximo. Considerou-se a estimação do valor de uma função de regressão em um ponto específico baseada na localização dos seus vizinhos mais próximos no espaço de características. A expectativa condicional $\mathbb{E}[Y|X = \mathbf{x}^*]$ é aproximada pela média dos valores de resposta dos k vizinhos mais próximos de $\mathbf{x}^*$ (Izbicki e Santos, 2020). Assim, a estimação é dada por:

$$\bar{Y}_{N_{\mathbf{x}^*}} = \frac{1}{k} \sum_{i \in N_{\mathbf{x}^*}} y_i.$$

Aqui, $N_{\mathbf{x}^*}$ representa o conjunto de índices dos k vizinhos mais próximos de $\mathbf{x}^*$, e $\bar{Y}_{N_{\mathbf{x}^*}}$ é a média calculada dessas observações de resposta, fornecendo uma estimativa robusta para o valor de regressão no ponto de consulta (Izbicki e Santos, 2020).

A seleção cuidadosa do parâmetro (*tuning parameter*) k é decisiva para a eficiência do algoritmo e é comumente realizada por meio de validação cruzada, como sugerido por Izbicki e Santos (2020). Um valor elevado de$k$ tende a simplificar o modelo, potencialmente aumentando o viés mas reduzindo a variância. Inversamente, um$k$ menor pode diminuir o viés à custa de elevar a variância. A tarefa é encontrar um valor de$k$que harmonize essas duas métricas, proporcionando uma generalização robusta.

O Nadaraya-Watson é bem parecido, mas nele não se necessita que seja colhido um número k de vizinhos, mas se pensa em um intervalo (+ ou - h) em torno da observação e se calcula uma média dos resultados previstos, com um kernel de suavização para medir a similaridade entre as observações. Essa técnica é conhecida por suavizar os dados, reduzindo o ruído e removendo as flutuações aleatórias que podem estar presentes em dados brutos, já que vizinhos distantes não são computa-

dos. Ele é usado principalmente para problemas de regressão, nos quais é necessário prever um valor contínuo de saída.

A metodologia desenvolvida por Nadaraya (1964) e Watson (1964), conhecida como o estimador de Nadaraya-Watson, representa uma extensão do método k-NN e é bem avaliada na comunidade estatística para estimação de funções de regressão (Izbicki e Santos, 2020). Esta técnica aprimora o método convencional ao introduzir pesos na média das respostas, ponderando-as de acordo com a proximidade de cada observação ao ponto de interesse x.

A função de estimação $g(x)$ é definida como uma soma ponderada:

$$g(x) = \sum_{i=1}^{n} w_i(x) y_i, \tag{2.3}$$

de forma que $w_i(x)$ é um peso atribuído a cada observação baseado em sua similaridade com x, determinada por um kernel de suavização $K(x, x_i)$. A expressão de $w_i(x)$ é:

$$w_i(x) = \frac{K(x, x_i)}{\sum_{j=1}^{n} K(x, x_j)}. \tag{2.4}$$

A escolha de $K(x, x_i)$ pode variar entre várias formas funcionais, como o kernel uniforme, gaussiano, triangular e o de Epanechnikov, cada um com suas características e influências sobre a ponderação das observações (Izbicki e Santos, 2020).

O ajuste fino do parâmetro de ajuste h é essencial, já que impacta diretamente a variância e o viés da estimativa. Um h mais elevado normaliza os pesos e pode aumentar o viés, enquanto um h reduzido diferencia mais os pesos, possivelmente aumentando a variância. Adicionalmente, é destacado que, embora a escolha do kernel possa ter um efeito limitado, os parâmetros de ajuste associados são críticos para os resultados finais (Izbicki e Santos, 2020).

O estimador de Nadaraya-Watson para um ponto fixo x é o valor que minimiza a soma dos quadrados ponderados dos desvios das observações em relação a uma constante, que no caso é o valor estimado $\hat{\beta}_0$. Esta abordagem é uma forma de regressão linear ponderada sendo o valor estimado para $g(x)$ unicamente o intercepto $\hat{\beta}_0$, conforme a equação:

$$g(x) := \hat{\beta}_0 = \arg\min_{\beta_0} \sum_{i=1}^{n} w_i(x)(Y_i - \beta_0)^2. \tag{2.5}$$

Izbicki e Santos (2020) destacam que a técnica de *Support Vector Regression* (SVR) fundamenta-se nos *Reproducing Kernel Hilbert Spaces* (RKHS), que oferecem uma abordagem rica e generalizada para a modelagem de funções de regressão, $r(\mathbf{x})$. Tais espaços permitem a formulação de uma função objetivo otimizada para a precisão das previsões e, em seguida, identificar a função que melhor se adequa ao

subconjunto de funções, $\mathcal{H}$. Busca-se minimizar a expressão:

$$\arg\min_{g\in\mathcal{H}} \left(\sum_{k=1}^{n} L(g(\mathbf{x}_k, y_k)) + P(g) \right),$$

sendo L uma função de perda escolhida e P quantifica a complexidade da função g, dentro do subespaço $\mathcal{H}$. Este método é vantajoso em espaços funcionais de alta dimensão, em que uma solução direta pode ser inatingível. O uso de RKHS permite a simplificação do problema, tornando-o acessível para soluções práticas.

Conforme discutido por Izbicki e Santos (2020), a seleção cuidadosa do parâmetro λ é fundamental para o equilíbrio do modelo SVMR. A influência deste parâmetro na complexidade do modelo é refletida pelo termo de penalização $\lambda\|g\|^2_{\mathcal{H}_k}$, que controla a suavidade das funções permitidas no espaço RKHS. A estratégia de escolha de λ pode ser informada por uma variedade de técnicas de validação cruzada, buscando minimizar o risco de sobreajuste ao mesmo tempo em que se mantém a precisão preditiva.

O processo iterativo de otimização é direcionado por uma função de perda L, que mede a discrepância entre as previsões do modelo e os valores observados. Essa função pode ser adaptada para refletir diferentes sensibilidades a erros no modelo, como indicado pela estatística contemporânea.

$$\arg\min_{g\in\mathcal{H}_k} \left\{ \sum_{k=1}^{n} L(g(\mathbf{x}_k, y_k)) + \lambda\|g\|^2_{\mathcal{H}_k} \right\}, \tag{2.6}$$

de modo que a função de perda L e a medida de complexidade $\lambda\|g\|^2_{\mathcal{H}_k}$ são escolhidas para refletir a teoria subjacente e as necessidades práticas do modelo em desenvolvimento, utilizando funções kernel, em particular, o kernel de Mercer, como destacado por Izbicki e Santos (2020).

Nesse contexto, a suavidade das funções selecionadas do espaço RKHS é um fator crucial. Esta suavidade é ajustável através do parâmetro de penalização λ, que é cuidadosamente escolhido com base na validação cruzada, como sugerido por Izbicki e Santos (2020). A validação cruzada permite uma avaliação mais precisa do modelo, equilibrando a adequação ao conjunto de dados de treinamento com a generalização para dados não vistos. O procedimento envolve a experimentação com diferentes valores de λ para encontrar o que minimiza o risco de predição.

$$\arg\min_{g\in\mathcal{H}_k} \left\{ \frac{1}{n} \sum_{k=1}^{n} L(g(\mathbf{x}_k, y_k)) + \lambda\|g\|^2_{\mathcal{H}_k} \right\}. \tag{2.7}$$

A penalização $\|g\|^2_{\mathcal{H}_k}$ reflete a complexidade do modelo e age como um regulador e é definida em termos da função de kernel, conforme mencionado por Izbicki e

Santos (2020), prevenindo o sobreajuste.

2.3.2 Limpeza dos dados

Além disso, precisa-se ressaltar a importância da organização de dados, frequentemente descrita como 'tidy data'. Ela é uma etapa crucial na análise de dados e modelagem preditiva, conforme destacado em importantes referências como "R for Data Science" de Grolemund e Wickham (2017) e "Tidy Modeling with R" de Max Kuhn e Julia Silge (2022). No conceito de *tidy data*, apresentado por Grolemund e Wickham, os dados devem ser estruturados de tal forma que cada variável forma uma coluna, cada observação uma linha e cada tipo de unidade observacional uma tabela. Esta organização facilita imensamente as operações comuns de análise de dados, como agrupamentos, transformações e visualizações, tornando o processo de análise mais intuitivo e eficiente.

Essa visão de limpeza e organização dos dados foi essencial para a compreensão das melhores estruturações dos dados que foram trabalhados no capítulo 4, no uso das técnicas de aprendizado de máquina.

Kuhn e Silge (2022), em "Tidy Modeling with R", expandem essa noção para o contexto da modelagem preditiva. Eles destacam que uma organização cuidadosa dos dados é essencial não apenas para a análise estatística, mas também para a construção eficaz de modelos de aprendizado de máquina. Dados bem organizados são fundamentais para a aplicação e interpretação adequadas de modelos preditivos. Em muitos casos, a forma como os dados são preparados e apresentados pode influenciar significativamente a eficácia do modelo. Isso se deve à necessidade de algoritmos de aprendizado de máquina e modelos estatísticos de trabalhar com dados que estejam estruturados de maneira lógica e consistente. Assim, tanto em "R for Data Science" quanto em "Tidy Modeling with R", enfatiza-se a importância de dedicar tempo e esforço na organização dos dados, uma etapa fundamental para garantir análises e modelagens precisas e confiáveis.

O pacote **tidymodels** (KUHN et al., 2020) é uma coleção de pacotes do R destinada à modelagem preditiva e machine learning. Proporciona uma sintaxe consistente e ferramentas para diversas etapas do processo de modelagem, como preparação de dados, seleção de modelos, validação cruzada e ajuste de hiperparâmetros, integrando-se perfeitamente ao tidyverse. Esta é uma das principais características do **tidymodels**, permitindo uma manipulação de dados eficiente e intuitiva com pacotes como **dplyr** (WICKHAM et al., 2023) e **tidyr** (WICKHAM et al., 2023). Além disso, **tidymodels** utiliza uma sintaxe consistente e declarativa que facilita a especificação de modelos complexos e a experimentação com diferentes tipos de modelos sem alterar o código substancialmente.

Os componentes do tidymodels incluem: 1) **parsnip** (KUHN & VAUGHAN, 2024) que é um pacote para especificar modelos estatísticos e de machine learning de uma maneira independente de engine, permitindo a mudança de modelo ou plataforma de modelagem sem alterar a interface; 2) **recipes** (KUHN & WICKHAM & HVITFELDT, 2024) que é um pacote destinado ao pré-processamento de dados, permitindo a criação de especificações de como os dados devem ser processados antes da modelagem, como normalização, criação de variáveis dummy, e tratamento de dados faltantes; 3) **rsample** (FRICK et al., 2024) que fornece infraestrutura para resampling de dados, como validação cruzada e bootstrapping, que são essenciais para avaliar a eficácia de modelos estatísticos; 4) **tune** (KUHN, 2024), pacote que ajuda na otimização de parâmetros de modelos para melhorar a performance do modelo, usando grid search, random search entre outras técnicas; 5) **workflows** (VAUGHAN COUCH, 2024) que permite a criação de um objeto workflow que encapsula um pré-processador, como um recipe, e um modelo, facilitando o manejo conjunto desses dois componentes; e 6) **yardstick** (KUHN & VAUGHAN & HVITFELDT, 2024)que é um pacote para avaliação de modelos, oferecendo uma variedade de métricas de performance modelar para classificação, regressão e outros tipos de análises preditivas.

Por isso, usou-se nesta dissertação o **tidymodels**, pois é particularmente útil para uma abordagem sistemática e robusta para modelagem preditiva em R, proporcionando ferramentas que facilitam a modelagem, a análise e a interpretação de modelos complexos de forma mais acessível e reproduzível.

2.4 Conclusão e Lacunas na Literatura

Este capítulo revisou de forma abrangente a literatura relevante sobre teoria dos leilões, custos de transação e técnicas de aprendizado de máquina, estabelecendo uma base teórica sólida para a investigação desta dissertação. A teoria dos leilões foi explorada em termos de seus pressupostos, tipos e aplicação da teoria dos jogos, proporcionando uma compreensão fundamental dos mecanismos de formação de preços nas contratações públicas. A análise dos custos de transação destacou sua relevância econômica e a importância da reputação dos compradores públicos como fator determinante nos preços de contratação. Por fim, a revisão das técnicas de aprendizado de máquina demonstrou seu potencial para prever custos de transação e melhorar a eficiência dos processos licitatórios.

Os achados desta revisão são diretamente conectados ao problema de pesquisa central: "Modelos matemáticos de leilões e técnicas de aprendizado de máquina podem prever custos de transação nas contratações públicas no estado da Paraíba?" A literatura revisada indica que, embora existam modelos teóricos e aplicações práticas

isoladas, há uma lacuna significativa na integração dessas abordagens para abordar especificamente os custos de transação nas contratações públicas.

Após uma análise extensiva da literatura sobre contratações públicas, teoria dos leilões, custos de transação e aprendizado de máquina, torna-se evidente que existem lacunas significativas no conhecimento que esta pesquisa busca preencher. Embora haja uma riqueza de estudos focando individualmente em teoria dos leilões e aprendizado de máquina, a aplicação integrada destas áreas no contexto das contratações públicas representa um campo relativamente inexplorado. Esta interseção, crucial para entender as dinâmicas atuais e futuras de licitações públicas, oferece uma oportunidade única para contribuição acadêmica e prática.

Primeiramente, a maioria das pesquisas existentes sobre a teoria dos leilões em contratações públicas tende a focar em análises meramente teóricas, com suposições arbitrárias (principalmente no que se refere às distribuições de probabilidade), sendo dissociado, portanto, da realidade.

Além disso, a aplicação de técnicas de aprendizado de máquina para prever preços e analisar custos de transação em contratações públicas ainda é um campo emergente. A literatura existente oferece uma base sólida em termos de metodologias e técnicas, mas há uma escassez de estudos que aplicam estas técnicas no contexto específico das contratações públicas, considerando os custos de transação e, em especial, a reputação de pagamento.

A necessidade de uma análise mais aprofundada sobre os custos de transação nas contratações públicas também é evidente. A maioria dos estudos existentes se concentra nos aspectos gerais de custos e eficiência sem explorar detalhadamente como os custos de transação específicos impactam o processo de licitação e a formação de preços. Uma compreensão mais profunda desses custos pode levar a recomendações mais eficazes para a melhoria das práticas de contratação pública.

Portanto, esta pesquisa se justifica pela necessidade de explorar a interação entre teoria dos leilões e aprendizado de máquina em um contexto geográfico específico, e pelo potencial de oferecer insights práticos e teóricos para otimizar as aquisições públicas. Além de preencher as lacunas identificadas, a pesquisa tem o potencial de contribuir para a literatura acadêmica e para a prática administrativa, fornecendo direções claras para futuras investigações e aplicações práticas no campo das contratações públicas.

Essa lacuna justifica a necessidade de desenvolver um modelo matemático robusto que possa identificar as variáveis críticas que influenciam os custos de transação e aplicar técnicas avançadas de aprendizado de máquina para prever esses custos com precisão. Este esforço contribuirá para os objetivos gerais da dissertação, que incluem a melhoria da eficiência e transparência dos processos de contratação pública, potencialmente resultando em uma alocação mais eficaz dos recursos públicos.

Na próxima seção, será detalhado o desenvolvimento do modelo matemático para leilões de compra, baseado na teoria dos jogos bayesianos, e sua adaptação para considerar os custos de transação. Este modelo servirá como a base teórica para a aplicação subsequente das técnicas de aprendizado de máquina, abordadas nos capítulos posteriores, em que serão treinados e validados os modelos de regressão para prever os custos de transação com dados reais das contratações públicas na Paraíba. A integração dessas abordagens pretende fornecer uma solução prática e inovadora para os desafios identificados, alinhando-se com os objetivos da pesquisa e contribuindo para a literatura existente.

Capítulo 3

Modelo Matemático

O desenvolvimento de um modelo matemático de leilões de compra é crucial para esta dissertação, pois visa resolver o problema de como prever os custos de transação nas contratações públicas no estado da Paraíba. Este capítulo aborda os pressupostos fundamentais dos leilões, descreve a aplicação da teoria dos jogos bayesianos e o conceito de equilíbrio de Nash Bayesiano, e apresenta um modelo matemático invertido para leilões de compra. A relevância desses temas é destacada pela necessidade de entender os mecanismos de formação de preços e a influência dos custos de transação, fatores críticos que impactam a eficiência e a eficácia das contratações públicas.

Ao explorar os diferentes tipos de leilões, como leilões abertos e de primeiro preço, e incorporar custos de transação e preços de reserva ao modelo, este capítulo fornece uma base teórica robusta para a análise empírica subsequente. A lacuna enfrentada na literatura, com poucos estudos integram essas abordagens distintas para analisar especificamente os custos de transação nas contratações públicas, é enfretada através da criação de um modelo matemático adaptado para situações de compra. Este modelo não só contribui para a teoria dos leilões, mas também oferece ferramentas práticas para otimizar os processos licitatórios, alinhando-se aos objetivos da pesquisa de aumentar a eficiência e a transparência nas contratações públicas. Na próxima seção, será detalhada a aplicação dessas técnicas matemáticas em conjunto com métodos de aprendizado de máquina para prever os custos de transação, fornecendo uma solução prática e inovadora para os desafios identificados.

3.1 Pressupostos

Os leilões podem se caracterizar por uma de suas propriedades fundamentais: funcionam como mecanismos de equilibração de mercado, alinhando oferta e demanda. Outros instrumentos de mercado incluem vendas a preço fixo (como em supermercados) ou negociação (como na venda de uma casa ou carro usado) (Mas-Collel et

at., 1991). Entre as diversas formas de ferramentas de mercado que distribuem recursos escassos, um traço distintivo do leilão é o processo de formação de preços ser explícito. Ou seja, as regras que estabelecem o preço final são, em geral, claras e compreendidas por todas as partes envolvidas.

Leilões são frequentemente empregados na comercialização de bens que não possuem um mercado estabelecido. Oferecem maior flexibilidade que as vendas a preço fixo e podem ser menos demorados que o processo de negociação de preços.

No desenvolvimento dos modelos dos leilões, seguiram-se os modelos clássicos conforme os principais autores da teoria dos leilões (Bulow e Klemperer, 2009; Maskin e Riley, 2000; David et al., 2007; Fullerton e McAfee, 1999). O padrão apresentado pela literatura da teoria dos leilões é com modelos matemáticos que descrevem um leilão de venda, mas, aqui nesta seção da dissertação, almejou-se inverter matematicamente para um leilão de compra, sendo esta uma contribuição clara para a teoria dos leilões.

Para isso, partiu-se de modelos gerais, incluindo a descrição da teoria dos jogos bayesianos e seus equilíbrios, passando para um leilão do tipo aberto e concluindo com os leilões de primeiro preço e a incorporação dos custos de transação ao modelo.

3.1.1 Teoria dos jogos aplicada aos leilões

No decorrer deste capítulo da dissertação, seguir-se-á a descrição de Myerson (1992) sobre os jogos Bayesianos dos quais os leilões são espécies. Assim, será adotado o conceito de equilíbrio de Nash Bayesiano conforme estabelecido por Harsanyi (1967), que propõe uma maneira de lidar com jogos de informação incompleta. Nessa abordagem, um comprador sem informação completa é modelado como se ele estivesse incerto sobre as valorações dos outros compradores, como se "a natureza" introduzisse um jogador adicional responsável pela seleção dos tipos dos jogadores.

Os jogos com informação incompleta são analisados como jogos de dois estágios. Antes do início do jogo, "a natureza" seleciona um tipo para cada jogador, que é conhecido pelo próprio jogador mas não pelos outros. No segundo estágio, com o conhecimento de seu próprio tipo e da distribuição inicial de todos os tipos, cada jogador escolhe uma estratégia.

Para estabelecer formalmente a noção de equilíbrio, faz-se necessário introduzir algumas notações. O conjunto de jogadores é denotado por $J = \{1, 2, \ldots, n\}$, e o conjunto de tipos possíveis para cada jogador $i \in J$ é representado por X_i, que normalmente é um intervalo $[0, \bar{v}_i]$ ao longo deste trabalho. A distribuição de probabilidade sobre o conjunto de produtos $X = X_1 \times X_2 \times \ldots \times X_n$ é representada por $F(\cdot)$, refletindo as probabilidades associadas a cada combinação possível de tipos.

Para cada jogador $i \in J$, o conjunto de estratégias possíveis é denotado por S_i

e a função decisória de um jogador i é uma função $s_i : X_i \to S_i$, que mapeia o conjunto de tipos possíveis ao conjunto de estratégias possíveis. Em certos casos, pode-se ter $S_i = \mathbb{R}_+$ e a distribuição de probabilidade dos tipos dos outros jogadores, dado o tipo do jogador i, é denotada por $F_i(\cdot|x_i)$. Essa distribuição é atualizada pelo jogador i à medida que ele revisa suas informações prévias sobre a distribuição dos tipos dos outros jogadores, utilizando a regra de Bayes ao aprender seu próprio tipo x_i.

Define-se a função de "resultado" de um jogador i como $\pi_i(s_i, s_{-i}, x_i, x_{-i})$, que representa o ganho do jogador quando ele escolhe uma estratégia $s_i \in S_i$ e os demais jogadores adotam estratégias $s_{-i}(x_{-i})$, com $s_j(x_j)$ sendo a função decisória do jogador j, e os tipos são x_{-i}, escolhidos pela natureza. Para cada vetor de tipos $(x_1, x_2, \ldots, x_n)$ escolhido pela natureza, os jogadores atualizam suas crenças com base nas distribuições $\hat{F}_1(x_{-1}|x_1), \ldots, \hat{F}_n(x_{-n}|x_n)$.

Deste modo, um Jogo Bayesiano é caracterizado por uma quíntupla:
$G = \{J, \{S_i\}_{i\in I}, \{\pi_i(\cdot)\}_{i\in I}, X_1 \times \ldots \times X_n, F(\cdot)\}$.

consistindo em um conjunto de jogadores J, um conjunto de estratégias S_i para cada jogador i, uma função de lucro π_i para cada um, um conjunto de tipos possíveis e uma distribuição sobre o conjunto de tipos.

O equilíbrio de Nash [1] Bayesiano é então definido como uma lista de funções decisórias $(s_1^*(\cdot), \ldots, s_n^*(\cdot))$, tais que para todo jogador i em J, para todos os tipos x_i em X_i, e para todas as estratégias s_i em S_i, tem-se:

$$\int_{x_{-i}\in X_{-i}} \pi_i(s_i^*, s_{-i}^*, x_i, x_{-i}) d\hat{F}_i(x_{-i}|x_i) \geq \int_{x_{-i}\in X_{-i}} \pi_i(s_i, s_{-i}^*, x_i, x_{-i}) d\hat{F}_i(x_{-i}|x_i).$$

Ou seja, nenhum jogador pode melhorar seu lucro unilateralmente alterando sua estratégia, dado que os outros jogadores estão seguindo suas estratégias de equilíbrio.

No contexto de jogos de informação incompleta, cada participante estabelece sua estratégia com base em seu próprio tipo, isto é, aplica-se uma função de decisão fundamentada em princípios Bayesianos. Nesse cenário, a noção de equilíbrio de Nash é aplicada a essas funções de decisão: cada jogador desenvolve uma estratégia de melhor resposta, determinando as funções de decisão Bayesianas mais vantajosas, as quais são influenciadas pelas estratégias de melhor resposta dos demais jogadores, que também estão definindo suas funções de decisão Bayesianas.

Ao definir um leilão como um jogo Bayesiano G, será mantida a notação já estabelecida para o conjunto de licitantes potenciais $J = \{1, 2, \ldots, n\}$, em que $X_i =$

[1] O Equilíbrio de Nash é uma situação em um jogo em que nenhum jogador pode melhorar seu resultado mudando unilateralmente sua estratégia, assumindo que as estratégias dos outros jogadores permanecem constantes.

$[0, \bar{v}]$ representa o conjunto dos tipos possíveis para o licitante i, sendo $i = 1, \ldots, n$, e v_i o tipo recebido pelo jogador i. A distribuição conjunta de tipos é representada por $F(\cdot) : [0, \bar{v}]^n \to [0, 1]$ e a densidade associada por $f(\cdot) : [0, \bar{v}]^n \to \mathbb{R}_+$. O conjunto de lances ou estratégias possíveis para o licitante i, $i = 1, \ldots, n$, é denotado por $S_i = \mathbb{R}_+$.

Para simplicidade, supõe-se que a avaliação do vendedor seja zero (não aceitando, portanto, lances negativos). Assume-se também que não existe mercado secundário, nem possibilidade de revenda, uma vez que o propósito é oferecer uma explanação abrangente e didática da teoria de leilões para iniciantes, e não uma revisão exaustiva da pesquisa existente sobre o tema.

Finalmente, o retorno para o licitante i dependerá de sua disposição ao risco, baseando-se em uma função de utilidade ou avaliação $u_i(v_1, \ldots, v_n)$, e das regras estipuladas pelo leilão. A natureza precisa desta relação será explicitada ao longo do leilão.

Comumente, os modelos de leilão são categorizados em três tipos principais. No modelo de valores privados, cada licitante potencial conhece o seu próprio valor pelo objeto, o qual não é afetado pela valoração dos outros jogadores. Se os tipos dos indivíduos são independentes entre si — por exemplo, quando os tipos são determinados por sorteios independentes de uma distribuição fixa —, tem-se o modelo de valor privado independente (VPI). Caso as avaliações sejam dependentes, o modelo é o de valor privado correlacionado. De forma mais ampla, um modelo de valores privados pode ser mais adequado para bens não duráveis sem valor de revenda.

No modelo de valor comum, o objeto possui o mesmo valor para todos os licitantes, mas esse valor é desconhecido no momento da oferta. Geralmente, os indivíduos possuem alguma informação sobre o verdadeiro valor (desconhecido) do objeto. Se a informação é correlacionada entre os indivíduos, então se tem um modelo de valor comum dependente. Se a informação é independente entre os indivíduos, trata-se de um modelo de valor comum independente. O modelo de valor comum é frequentemente mais apropriado para analisar a venda de direitos minerários e concessões de perfuração de petróleo offshore.

Por fim, Milgrom e Weber (1982) introduzem o conceito de valores afiliados, que engloba tanto os valores privados quanto os comuns como casos particulares. De maneira simplificada, os valores afiliados capturam a ideia de que as avaliações individuais por um objeto possuem um componente privado, mas são influenciadas pela valoração que outras pessoas atribuem a ele. Na maioria das vendas que se pode imaginar, a avaliação de um licitante por um objeto possui um componente privado, mas essa avaliação também é influenciada pelas avaliações de outros indivíduos. Por exemplo, ao oferecer um lance por uma casa, considera-se tanto o valor pessoal do imóvel quanto a facilidade de revendê-lo no futuro. Contudo, afiliação é uma noção

de correlação positiva global e isso tem implicações particulares para a classificação dos formatos de leilão de acordo com a receita esperada que eles geram.

3.1.2 Descrição dos leilões

Considera-se um leilão em que n licitantes, avessos ao risco, competem por um único item ofertado por um leiloeiro igualmente avesso ao risco. Os licitantes têm valorações privadas independentes, v_i, extraídas de uma função densidade de probabilidade uniforme contínua[2] comum, $f(v)$, dentro do intervalo $[\underline{v}, \bar{v}]$. Esta função densidade possui uma função distribuição acumulada, $F(v)$, e pode-se estabelecer que $F(\underline{v}) = 0$ e $F(\bar{v}) = 1$. Propõem-se as seguintes premissas, seguindo os ensinamentos de Milgrom e Weber (1982):

Premissa 1: Há uma função utilidade, u, em $\mathbb{R}^{m+n}$ de tal modo que, para todo i, tem-se $u_i(\mathbf{S}, \mathbf{X}) = u(\mathbf{S}, X_i, \{X_j\}_{j \neq i})$. Com isso, a valorização de todos os proponentes depende de $\mathbf{S}$ de maneira idêntica, e a valorização de cada proponente é uma função simétrica dos indicativos dos demais proponentes.

Premissa 2: A função u é não-negativa, contínua e não-decrescente em relação às suas variáveis.

Premissa 3: Para cada i, espera-se que $E[V_i]$[3] $< \infty$.

Assim, se o licitante i adquire o item vendido e paga o montante b, seu ganho é simplesmente $V_i - b$.

Seja $f(\mathbf{s}, \mathbf{x})$ a densidade de probabilidade conjunta dos elementos aleatórios do modelo. Estabelece-se uma premissa acerca da distribuição conjunta de $\mathbf{S}$ e $\mathbf{X}$:

Premissa 4: A função f é simétrica[4] em seus últimos n argumentos.

3.1.3 Valores Privados

Um único objeto será comprado de um dentre n licitantes. Cada licitante i, em que $i = 1, \ldots, n$, recebe um tipo v_i e seu valor correspondente é igual a $u_i(v_i) = v_i$. A premissa subjacente é que os compradores são neutros ao risco, ou seja, são

[2] Esse é o uso padrão da literatura, pois facilita a parte matemática e é compatível com a teoria do valor privado.

[3] V_i é uma variável aleatória que representa a valoração v para o jogador i. Vale ressaltar que V_i pode variar de acordo com uma distribuição de probabilidade específica, refletindo o valor que o jogador i atribui a diferentes resultados possíveis do jogo.

[4] Na teoria dos jogos, uma função simétrica é uma função que mantém seu valor inalterado quando as posições dos jogadores são permutadas. Dessa forma, os resultados dependem unicamente das estratégias adotadas, e não da identidade dos jogadores. Formalmente, uma função f é simétrica se, para toda permutação σ dos jogadores, tem-se que $f(x_1, x_2, \ldots, x_n) = f(x_{\sigma(1)}, x_{\sigma(2)}, \ldots, x_{\sigma(n)})$, sendo x_i a representação da estratégia do jogador i. Este conceito é especialmente relevante em cenários em que presume-se que todos os jogadores são equivalentes, diferenciando-se apenas pelas estratégias escolhidas.

indiferentes entre uma loteria que gera um valor esperado de x e receber x com certeza (Menezes e Monteiro, 2008).

Cada licitante conhece sua própria avaliação v_i e sabe que as avaliações de seus oponentes são sorteadas independentemente da distribuição $F(\cdot)$ com densidade $f(\cdot) > 0$ no intervalo $[0, \bar{v}]$. Ou seja, $F(x)$ denota a probabilidade de que a variável aleatória v seja menor ou igual a um certo número x.

Este é o modelo VPI, em que o valor do objeto para um licitante depende apenas de seu próprio tipo. No entanto, o comportamento de lance depende das expectativas de cada um sobre as avaliações dos outros licitantes e sobre como eles darão seus lances. Embora o modelo de valor privado independente seja apenas apropriado para descrever casos em que o objeto não possui valor de revenda (ou é muito custoso para revender), ele permite derivar várias percepções importantes. Por simplicidade, assume-se que o vendedor estabelece o preço de reserva em zero e que não há taxas de entrada (Menezes e Monteiro, 2008).

3.2 Leilão aberto

Assim, os licitantes participam de um leilão com preços descendentes, de modo que os lances são restritos a níveis discretos determinados pelo leiloeiro. Presume-se que há $m + 1$ níveis de lances discretos, começando em l_0 e terminando em l_m[5]. Até o momento, não se estabelecem restrições quanto ao número real desses níveis de lance, nem quanto aos intervalos entre eles.

O leiloeiro propõe o primeiro nível de lance, l_0, e todos os licitantes dispostos a vender por esse preço, e assim continuar no leilão, indicam sua disposição ao leiloeiro. Neste ponto, o leiloeiro seleciona aleatoriamente um licitante dentre os dispostos. Este licitante é nomeado como o detentor do lance mais alto atual e este status é anunciado a todos os participantes. O leiloeiro então propõe o próximo nível de lance, l_1, e os licitantes novamente indicam sua vontade de permanecer no leilão. Mais uma vez, um licitante é selecionado aleatoriamente[6] como o detentor do lance mais alto atual. O processo continua, com o preço reduzindo através dos níveis de lance discretos, até que não haja mais licitantes dispostos a vender pelo preço de oferta mais baixo. O leilão é então encerrado e o item é vendido ao licitante com o lance mais alto no momento.

Os licitantes possuem uma estratégia dominante simples: devem continuar a participar no leilão e dar lances a cada nível de lance, até que o nível de preços corrente seja menor sua valoração privada para venda. Não há necessidade de estratégias

[5] Como se trata de um leilão de compra, os lances podems ser vistos aqui como desconto, então o maior lance leva o preço para um valor menor.

[6] A lei de licitações estabelecem regras de desempate, mas será tratada aqui como sendo algo aleatório.

complexas sobre as valorações dos outros licitantes, nem sobre o momento de seus lances. Assim, este protocolo de leilão é particularmente atraente em ambientes computacionais nos quais os licitantes provavelmente serão agentes de negociação automatizados com complexidade limitada, já que facilitará para fazer as simulações de monte carlo.

Apresentam-se três cenários nos quais o leilão se encerra no nível de lance l_v. Em cada cenário, os círculos representam a valoração privada de um licitante e a seta indica o nível de lance no qual o licitante foi selecionado como o detentor do lance mais alto atual:

- Caso 1: Dois ou mais licitantes têm valorações entre $[l_v, l_{v+1})$ e nenhum licitante possui valoração $v \geq l_{v+1}$;
- Caso 2: Um licitante possui uma valoração $v \geq l_{v+1}$, um ou mais licitantes têm valorações no intervalo $[l_v, l_{v+1})$ e o licitante com a valoração mais alta foi selecionado como o detentor do lance mais alto atual em l_i;
- Caso 3: Um licitante possui uma valoração $v \geq l_v$, um ou mais licitantes têm valorações no intervalo $[l_v, l_{v+1})$, e o licitante com a valoração mais alta não foi selecionado como o detentor do lance mais alto atual em l_{v-1}.

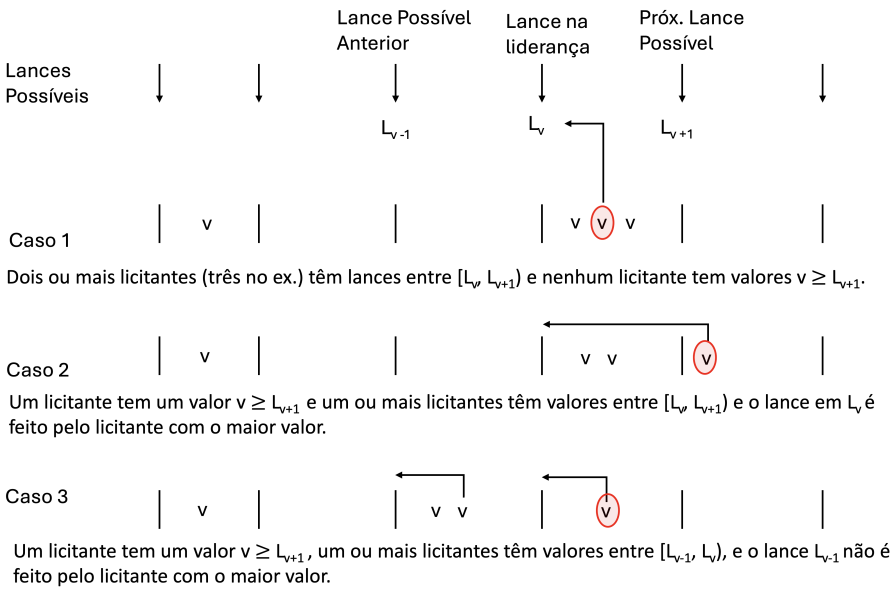

Figura 3.1: Elaboração própria.

Para determinar os níveis ótimos de lance, é imperativo inicialmente derivar uma expressão para a receita esperada do leiloeiro, considerando os níveis de lance discretos específicos empregados no leilão. Consoante com o trabalho de Rothkopf e

Harstad, pode-se caracterizar a probabilidade de conclusão do leilão em um determinado nível de lance por meio da análise de três casos exaustivos e mutuamente excludentes (Rothkopf e Harstad, 1994). Os casos podem ser descritos como segue:

Caso 1: Dois ou mais licitantes possuem uma valoração superior ao nível de lance l_v, porém nenhum destes tem valoração que exceda l_{v+1}. Portanto, uma vez que o preço do lance alcança l_v, nenhum licitante consegue propor um lance superior, e o item é atribuído ao licitante com o maior lance atual. Neste caso, a receita arrecadada pelo leiloeiro é inferior à que seria adquirida em um leilão contínuo (i.e., a segunda menor valoração) e o resultado pode ser alocativamente ineficiente, uma vez que o item não é necessariamente concedido ao licitante com a menor valoração;

Caso 2: Dois ou mais licitantes têm valorações entre l_v e l_{v+1}, e um único licitante tem uma valoração superior a l_{v+1}. Como esse licitante também era o atual detentor do lance mais alto quando o nível de lance atingiu l_v, nenhum dos outros licitantes possui valoração suficiente para aumentar o lance para l_{v+1}. Assim, o leilão se encerra no lance l_v e o item é comprado do licitante com a menor valoração. Novamente, a receita do leiloeiro é menor do que a que teria sido alcançada em um leilão contínuo, mas o desfecho é alocativamente eficiente;

Caso 3: Este caso é idêntico ao caso dois, exceto pelo fato de que o licitante com a menor valoração (lembrando que se trata de leilão de compra) não é o detentor do lance mais alto atual. Consequentemente, esse licitante é compelido a elevar o nível do lance e o leilão se encerra no nível de lance l_v, em vez de l_{v-1}. Este caso também é alocativamente eficiente; no entanto, a receita obtida pelo leiloeiro é de fato maior do que a adquirida em um leilão contínuo.

A receita esperada do leilão, portanto, depende da probabilidade de ocorrência de cada um desses três casos. Estas probabilidades podem ser expressas em termos da função de distribuição acumulada das valorações dos licitantes, $F(v)$. Assim, dado que $P(\text{caso1}, l_i)$ representa a probabilidade do primeiro caso ocorrer e o leilão encerrar no nível de lance l_i, pode-se descrever a probabilidade deste caso acontecer considerando k licitantes com valorações entre os níveis de lance l_i e l_{i+1}. A probabilidade de isso ocorrer é simplesmente $[F(l_{i+1}) - F(l_i)]^k$, enquanto a probabilidade de todos os outros $n - k$ licitantes terem valorações abaixo de l_i é $F(l_i)^{n-k}$. Portanto, $P(\text{caso1}, l_i)$ pode ser encontrada somando sobre todos os valores possíveis de k, resultando em:

$$P(\text{caso1}|l_i) = \sum_{k=2}^{n} \binom{n}{k} [F(l_{i+1}) - F(l_i)]^k F(l_i)^{n-k}. \tag{3.1}$$

Similarmente, para o caso dois, na qual existem k licitantes com valorações entre l_i e l_{i+1}, um licitante com valoração acima de l_{i+1} e $n - k - 1$ licitantes com valorações abaixo de l_i, a probabilidade também leva em conta que o licitante com

a maior valoração é o atual detentor do lance mais alto. Considerando que esta seleção é aleatória, a probabilidade é $\frac{1}{k+1}$. Portanto, a expressão completa é:

$$P(\text{caso2}|l_i) = \sum_{k=1}^{n-1} \binom{n-1}{k} \frac{n}{k+1} [F(l_{i+1}) - F(l_i)]^k F(l_i)^{n-k-1} [F(l_{i+1}) - F(l_i)][1 - F(l_{i+1})]. \tag{3.2}$$

Por fim, considera-se o caso três, que é idêntico na forma ao caso dois, com a exceção de que o licitante com a maior valoração não foi nomeado como o detentor do lance mais alto no nível de lance l_{i-1} e, portanto, deve elevar o preço para l_i. A probabilidade disto ocorrer é $\frac{k}{k+1}$, ao invés do fator $\frac{1}{k+1}$ que ocorreu no caso dois. Note que esta descrição implica que existe um nível de lance abaixo de l_i e, portanto, a expressão encontrada é válida apenas para os níveis de lance $l_1 \dots l_m$. Para incluir a instância na qual o leilão encerra no nível de lance l_0, faz-se isso separadamente e se observa que isso ocorre quando todos, exceto um licitante, têm valorações abaixo de l_0. Assim, a expressão final é descrita como:

$$P(\text{caso3}|l_i) = \begin{cases} nF(l_0)^{n-1}[1 - F(l_0)] & \text{para } i = 0 \\ \sum_{k=1}^{n-1} \binom{n-1}{k} \frac{k}{k+1} F(l_i)^{n-k-1} [F(l_i) - F(l_{i-1})]^k [1 - F(l_{i+1})] & \text{para } i > 0. \end{cases} \tag{3.3}$$

Agora, como estas três expressões descrevem completamente todas as maneiras possíveis pelas quais o leilão pode encerrar em um determinado nível de lance, pode-se encontrar a receita esperada do leiloeiro simplesmente somando sobre todos os possíveis níveis de lance e ponderando cada um pela receita que gera. Assim, a receita esperada do leilão é dada por:

$$E = \sum_{i=0}^{m} l_i [P(\text{caso1}|l_i) + P(\text{caso2}|l_i) + P(\text{caso3}|l_i)]. \tag{3.4}$$

A expressão resultante nesta fase é mais complexa devido às somas combinatórias nas equações 1, 2 e 3. No entanto, é possível simplificar significativamente esta expressão (notando que, sem perda de generalidade, pode-se definir $F(l_{m+1}) = 1$), para obter o resultado final:

$$E = \sum_{i=0}^{m} \left[\frac{F(l_{i+1})^n - F(l_i)^n}{F(l_{i+1}) - F(l_i)} \right] \left[l_i(1 - F(l_i)) - l_{i+1}(1 - F(l_{i+1})) \right]. \tag{3.5}$$

3.2.1 Solução Analítica

Para esse momento, apresenta-se a solução analítica para a equação do valor esperado do leilão. Assim, devem-se calcular as derivadas parciais da expressão de receita apresentada na equação 3.5, em relação a cada nível individual de lance l_i. Em seguida, pode-se resolver esta expressão para $\frac{\partial E}{\partial l_i} = 0$, e assim encontrar o valor de l_i que maximiza a receita.

Para realizar esta diferenciação, é preciso observar que cada l_i ocorre na soma da equação 3.5 duas vezes. Por exemplo, o nível de lance l_5 ocorre no termo de soma quando $i = 5$, como $F(l_i)$, e também no termo precedente quando $i = 4$, como $F(l_{i+1})$. Assim, para uma distribuição de valoração dos licitantes uniforme, substituem-se na expressão analítica $F(l_i) = \frac{l_i - \underline{v}}{\bar{v} - \underline{v}}$ nestes dois termos e diferencia-se para obter:

$$\frac{\partial E}{\partial l_i} = \frac{(l_{i+1} - \underline{v})^n - (l_{i-1} - \underline{v})^n}{(\bar{v} - \underline{v})^n} - \frac{n l_{i-1}(l_i - \underline{v})^{n-1} - n l_{i+1}(l_i - \underline{v})^{n-1}}{(\bar{v} - \underline{v})^n}. \tag{3.6}$$

Para encontrar o valor ótimo de l_i que maximiza a receita, deve-se tornar essa derivada parcial igual a zero (i.e., $\frac{\partial E}{\partial l_i} = 0$) e resolver a expressão resultante. Isso dá o resultado:

$$l_i = \underline{v} + \frac{1}{n-1} \sqrt[n]{\frac{(l_{i+1} - \underline{v})^n - (l_{i-1} - \underline{v})^n}{n(l_{i+1} - l_{i-1})}}. \tag{3.7}$$

O número de licitantes, n, desempenha um papel crucial. À medida que n aumenta, o termo $\frac{1}{n-1}$ diminui, sugerindo que com mais concorrentes, o incremento sobre a menor valoração para atingir o lance ótimo diminui. Isso implica que em leilões com muitos participantes, o aumento de cada lance individual além do valor mínimo tende a ser menor, mas, é sempre melhor que ter menos licitantes.

Um valor mais alto de $\underline{v}$ tende a elevar todos os lances, pois os licitantes devem começar a oferecer lances acima desta valoração. Além disso, a expressão sob a raiz n-ésima envolve os lances adjacentes l_{i+1} e l_{i-1}. A diferença elevada à potência de n entre esses lances, ajustada pelo valor mínimo $\underline{v}$, indica que variações significativas nos lances adjacentes podem levar a maiores variações no lance ótimo. Isto sugere que em um ambiente em que os lances variam amplamente, o potencial para um lance ótimo mais alto é maior.

A raiz n-ésima da expressão indica que a relação entre as variações dos lances não é linear, mas moderada pela quantidade de licitantes. Isto reflete uma interdependência complexa entre os lances dos participantes, que é central para a estratégia de lances.

3.3 Licitação de primeiro preço

A busca se inicia por um equilíbrio de Nash simétrico ao analisar o jogo pela perspectiva de um dos jogadores, por exemplo, o Jogador 1, invertendo o modelo de Menezes e Monteiro (2008) para situação de compra, ao invés de venda. Supõem-se que este jogador tenha uma valoração $v = v_1$ e acredita que os outros jogadores seguem uma estratégia de lance $b(\cdot)$. Conhecendo apenas seu valor e a distribuição das valorações dos jogadores $2, \ldots, n$, o Jogador 1 precisa deduzir qual é a sua melhor resposta. Suponha-se que o licitante $i = 2, \ldots, n$ tem valoração v_i. Assim, o licitante $i \geq 2$ faz lances $b_i = b(v_i)$. Se o Jogador 1 fizer um lance b_1 e este for menor que b_i para todo $i \geq 2$, ou seja, se $b_1 < \min\{b_2, \ldots, b_n\}$, ele ganha o contrato da licitação. Se $b_1 \geq \min\{b_2, \ldots, b_n\}$, o Jogador 1 não vence a licitação. Aqui, assume-se que, em caso de empate, isto é, se $b_1 = \min\{b_2, \ldots, b_n\}$, a licitação não se conclui. Portanto, o ganho do Jogador 1 é :

$$\begin{cases} b_1 - v & \text{se } b_1 < \min\{b(v_2), \ldots, b(v_n)\} \\ 0 & \text{se } b_1 \geq \min\{b(v_2), \ldots, b(v_n)\}. \end{cases}$$

Os lucros esperados ao fazer um lance de b_1 são dados por:

$$\pi(b_1) = \pi(v, b_1, b(\cdot)) = (b_1 - v)\Pr(b_1 < \min\{b(v_2), \ldots, b(v_n)\}).$$

Pode-se reescrever a expressão acima como:

$$\pi(b_1) = (b_1 - v)\Pr(b_1 < b(v_2), \ldots, b_1 < b(v_n)).$$

Suponha por um momento que a função $b(\cdot)$ seja estritamente decrescente e diferenciável. Assim, a faixa de $b(\cdot)$ é um intervalo: $[b(0), \bar{v}] = [\underline{b}, \bar{b}]$. O(A) licitante nunca oferecerá um lance menor que $\underline{b}$ pois ele(a) sairia no prejuízo com um pagamento menor que o valor do objeto contratado. Qualquer lance acima de $\bar{b}$ é um lance perdido. Portanto, pode-se supor, sem perda de generalidade, que $b_1 \in [\underline{b}, \bar{b}]$. Logo, existe um x em $[0, \bar{v}]$ tal que $b_1 = b(x)$.O problema do licitante 1 pode ser descrito como a escolha de x no intervalo $[0, \bar{v}]$ para maximizar a utilidade esperada:

$$\begin{aligned} \bar{\pi}(x) = \pi(b(x)) &= (b(x) - v)\Pr(b(x) < b(v_2), \ldots, b(x) < b(v_n)) \\ &= (b(x) - v)\Pr(x < v_2, \ldots, x < v_n). \end{aligned}$$

Utilizando o fato de que $b(\cdot)$ ser estritamente decrescente e todos os jogadores seguirem a mesma estratégia em equilíbrio, já que todos enfrentam o mesmo problema de maximização e considerando que as v_i são variáveis aleatórias independentes e idênticas, pode-se reescrever a equação acima como segue:

$$\bar{\pi}(x) = (b(x) - v)\Pr(x < v_2) \cdots \Pr(x < v_n) = (b(x) - v)F(x)^{n-1}.$$

Agora, a derivada de $\bar{\pi}$ é facilmente calculada:

$$\bar{\pi}'(x) = (b(x) - v)(n-1)f(x)F(x)^{n-2} + b'(x)F(x)^{n-1}. \tag{1}$$

No equilíbrio simétrico, o lucro esperado é maximizado em $x = v$, portanto a condição de primeira ordem é $\bar{\pi}'(v) = 0$. A partir da equação acima se obtém:

$$b'(v)F(v)^{n-1} = (b(v) - v)(n-1)f(v)F(v)^{n-2}. \tag{2}$$

O que se pode escrever:

$$\begin{aligned}(b(v)F(v)^{n-1})' &= b'(v)F(v)^{n-1} + b(v)(n-1)f(v)F(v)^{n-2} \\ &= v(n-1)f(v)F(v)^{n-2}.\end{aligned}$$

Pelo Teorema Fundamental do Cálculo se tem:

$$b(v)F(v)^{n-1} = \int_0^v x(n-1)f(x)F(x)^{n-2}dx + k,$$

em que k é a constante de integração. Se $v \to 0$, o lado esquerdo tende a zero já que $b(\cdot)$ é limitada. Portanto, conclui-se que $k = 0$. Isto é, a estratégia de lances de equilíbrio proposta é dada por:

$$b^*(v) = \begin{cases} \frac{(n-1)}{F(v)^{n-1}} \int_0^v xf(x)F(x)^{n-2}dx & \text{se } 0 < v \leq \bar{v}; \\ 0 & \text{se } v = 0. \end{cases} \tag{3}$$

Agora, é necessário verificar a continuidade de $b(v)$. Basta demonstrar isso para $v = 0$. Observe que, para $v > 0$,

$$b^*(v) = \frac{(n-1)}{F(v)^{n-1}} \int_0^v xf(x)F(x)^{n-2}dx.$$

é menor que

$$\frac{(n-1)}{F(v)^{n-1}} \int_0^v vf(x)F(x)^{n-2}dx = v.$$

Portanto, $b(v)$ é contínua em zero e, consequentemente, em todo lugar, além de que $b*(v) < v$. Agora, é necessário confirmar que b^* é de fato um equilíbrio. Das equações (1) e (2), observa-se que:

$$\begin{aligned}\bar{\pi}'(x) &= (b(x) - v)(n-1)f(x)F(x)^{n-2} + b'(x)F(x)^{n-1} \\ &= (x - v)(n-1)f(x)F(x)^{n-2}.\end{aligned}$$

Se $x < v$, então $\bar{\pi}'(x) > 0$. E se $x > v$, $\bar{\pi}'(x) < 0$. Fica claro, então, que $x = v$

maximiza a utilidade esperada. Se o valor v é o menor entre todos os jogadores, então em um equilíbrio simétrico em que as estratégias são crescentes, basta alguém dar um lance ligeiramente menor do que o segundo maior lance para vencer (Menezes e Monteiro, 2008).

O valor $b^*(v) - v$ indica o quanto um licitante esconde seu lance, no equilíbrio. Especificamente, ela demonstra o quanto o licitante diminui seu lance em relação à sua avaliação. Para calcular essa subestimação, integra-se a expressão (3) por partes.

Definindo $z = F(x)^{n-1}$, tem-se que $dz = (n-1)F(x)^{n-2}f(x)dx$. Da mesma forma, definindo $du = dx$, tem-se (pela integração) que $u = x$. Logo:

$$\begin{aligned}(n-1)\int_0^v xf(x)F(x)^{n-2}dx &= \int_0^v u dz \\ &= xF(x)^{n-1}\Big|_0^v - \int_0^v F(x)^{n-1}dx.\end{aligned}$$

Substituindo isso em (3), obtém-se:

$$b^*(v) = \frac{\int_0^v F(x)^{n-1}dx}{F(v)^{n-1}} - v.$$

Portanto, a subestimação é dada por:

$$\int_0^v (F(x)/F(v))^{n-1}dx.$$

Ela diminui com o aumento do número de licitantes. Quanto maior o número de meus oponentes, mais próximo do meu valor será o lance.

Agora que existe uma previsão de como os licitantes se comportarão em uma licitação que segue o modelo de primeiro preço. Assim, pode-se investigar qual será a ganho esperado do ente público a partir de um leilão de primeiro preço, denotado por G^1. O ganho esperado é simplesmente o valor esperado do menor lance, ou seja:

$$G^1 = E[\min\{b^*(v_1), \ldots, b^*(v_n)\}].$$

Sob a ótica do ente público comprador, os licitantes fornecedores são, ex-ante, idênticos. Assim, a probabilidade de que todas as avaliações estejam abaixo de um valor v é simplesmente $F(v)^n$ e sua densidade é $nF(v)^{n-1}f(v)$. Como resultado melhor descrito em Menezes e Monteiro (2008) de onde se partiu para se fazer a inversão para o modelo de compra, o ganho esperado pode ser escrita como :

$$G^1 = \int_0^{\bar{v}} nb^*(v)F(v)^{n-1}f(v)dv.$$

3.3.1 Acrescentando preços de reserva e custos de transação - custos de entrada

Nesta seção da dissertação, incorporam-se mais duas importantes variáveis ao modelo. Denomina-se o preço de reserva por r e a taxa de entrada por δ. Pressupõe-se que o preço de reserva seja conhecido por todos os licitantes, sendo um valor máximo que a administração está disposta a pagar pelo bem. A "taxa" de entrada é o custo de transação para participar da licitação.

É importante notar que estes dois mecanismos, preço de reserva e taxa de entrada, produzem efeitos antagônicos: eles diminuem o incentivo dos licitantes para participar da licitação; contudo, podem aumentar o ganho do ente público, visto que existe um impacto do preço de reserva no comportamento de lances (Menezes e Monteiro, 2008).

O primeiro passo é definir um valor de corte ρ para o jogador i tal que, se $v_i < \rho$, ele não participará do leilão. Se $v_i \geq \rho$ então i decide participar. Supõem-se que os jogadores $2, \ldots, n$ seguem essa regra de participação e dão lances de acordo com uma função $b(\cdot)$ estritamente decrescente e diferenciável. Calcula-se a melhor resposta do Jogador 1 encontrando um equilíbrio tal que $b(\rho) = r$. Considerando a valoração de Jogador 1 como $v_1 = v$, seu problema é escolher uma regra de participação e um lance b_1 para maximizar seus lucros esperados:

$$\pi_1 = E[(b_1 - v)I_{b_1} \leq \min\{b(Z), r\}] - \delta,$$

em que $Z = \min\{v_j; v_j \geq \rho, j = 2, \ldots, n\}$ se o conjunto for não-vazio e $Z = 0$ caso contrário. Portanto, para dar um lance, o lance de Jogador 1 deve ser menor ou igual a r. Assim, os lucros esperados de 1 podem ser reescritos como:

$$\begin{aligned}\pi_1 &= -\delta + (b_1 - v)Pr[b_1 \leq \min\{b(Z), r\}] \\ &= -\delta + (b_1 - v)Pr[b_1 \leq \min\{b(Z), b(\rho)\}].\end{aligned}$$

Se $b_1 = r$ então $\pi_1 = -\delta + (r - v)F^n - 1(\rho)$. Se $b_1 = b(s) < r$ então:

$$\begin{aligned}\pi_1 &= -\delta + (b(s) - v)Pr[s \leq Z] \\ &= -\delta + (b(s) - v)F^n - 1(s).\end{aligned}$$

Observe que $s < \rho$. Pode-se comparar essa última equação com o benefício esperado da seção anterior e concluir que a condição de primeira ordem é a mesma do caso em que tanto o preço de reserva quanto a taxa de entrada são iguais a zero. A única diferença é que a condição de contorno deve refletir o fato de que um licitante com valoração igual a ρ deve ser indiferente entre participar ou não, e, portanto, $b(\rho) = r$. Assim, a função de lance de equilíbrio é dada por:

$$b^*(v) = \begin{cases} \frac{\delta + \int_\rho^v F(x)^{n-1}dx}{F(v)^{n-1}} - v, & \text{se } v \geq \rho; \\ \text{"não dar lance"}, & \text{caso contrário.} \end{cases} \tag{3}$$

Agora é possível comparar essa estratégia de lances de equilíbrio com a estratégia quando tanto a taxa de inscrição quanto o preço de reserva são zero. É perceptível que as taxas de entrada influenciam somente a decisão do licitante de entrar ou não no leilão, enquanto um preço de reserva não nulo afeta tanto a decisão de entrar quanto a estratégia de lances — aqueles que entram tendem a dar lances mais agressivos no equilíbrio simétrico (Menezes e Monteiro, 2008).

Para caracterizar completamente o comportamento de equilíbrio, ainda precisa-se calcular o valor crítico de corte ρ. Lembre-se que um jogador com valor ρ é indiferente entre participar ou não:

$$-\delta + (r - \rho)F^{n-1}(\rho) = 0.$$

Isso significa que, quando o licitante indiferente decide participar, ele paga a taxa de entrada δ. Dado que ele só vence se for o único participante, o preço recebido (vencedor) na licitação é o preço de reserva r. $F^{n-1}(\rho)$ denota a probabilidade de que o licitante indiferente seja o único a dar um lance. Pode-se reescrever essa expressão como:

$$(r - \rho)F^{n-1}(\rho) = \delta.$$

É claro que $\rho < r$ e $\delta < r - \bar{v}$. A última desigualdade surge pelo fato de que o valor mínimo de $r - \rho$ é igual a $r - \bar{v}$ e que $F^{n-1}(\rho) < 1$. Assim, pode-se concluir que $r - \delta < \bar{v}$. Se essa desigualdade não fosse verdadeira, nenhum licitante jamais participaria desta licitação.

Destas duas últimas equações, nota-se que há dois efeitos no comportamento de equilíbrio ao impor uma taxa de entrada e um preço de reserva. Primeiramente, licitantes com avaliação mais altas não participarão da licitação. Em segundo lugar, aqueles que participam tenderão a dar lances de forma mais agressiva — isso se deve ao fato de que para vencer o certame agora o jogador deve dar um lance equivalente ao valor esperado do maior lance entre seus oponentes com valores entre ρ e $\bar{v}$, e não entre 0 e $\bar{v}$ como antes. Uma menor participação reduz a receita esperada do vendedor, mas lances mais agressivos tendem a aumentá-la. Portanto, pode-se questionar qual a combinação de preço de reserva e taxa de entrada maximiza a receita esperada do vendedor.

O ganho esperado G^1do ente público, retirou-se o segundo termo à direita do modelo de Menezes e Monteiro (2008) já que nas licitações não é possível o pagamento para entrada no certame, ou seja, os custos de transação não são receitas

para a administração pública, passa a ser expressa como:

$$\begin{aligned} G^1 &= \int_{\rho}^{\bar{v}} b^*(v) n F(v)^{n-1} f(v) dv \\ &= \int_{\rho}^{\bar{v}} v n F(v)^{n-1} f(v) dv - \int_{\rho}^{\bar{v}} n (\int_{\rho}^{v} F(x)^{n-1} dx) n f(v) dv. \end{aligned}$$

Alterando a ordem de integração na integral dupla, já que $\rho < x < v < \bar{v}$, quando se integra sobre v, lembrando que $\int_x^{\bar{v}} f(v)dv = 1 - F(x)$, tem-se:

$$G^1 = \int_{\rho}^{\bar{v}} v n F(v)^{n-1} f(v) dv - \int_{\rho}^{\bar{v}} n(1 - F(v)) F(v)^{n-1} dv.$$

Ao buscar o valor de ρ que maximiza G^1, diferencia-se a expressão acima e se obtém:

$$\frac{\partial G^1}{\partial \rho} = -\rho n F(\rho)^{n-1} f(\rho) + n(1 - F(\rho)) F(\rho)^{n-1}$$

$$= n F(\rho)^{n-1} \{-\rho f(\rho) + 1 - F(\rho)\}.$$

Em um máximo interior, tem-se:

$$n F(\rho)^{n-1} f(\rho) \left(-\rho + \frac{1 - F(\rho)}{f(\rho)} \right) = 0.$$

ou

$$\rho = \frac{1 - F(\rho)}{f(\rho)}.$$

Esta condição indica o nível do preço de reserva que maximiza o ganho esperado de um comprador (ente público) utilizando uma licitação com base no leilão de primeiro preço, quando não cobra taxas de entrada (custos de transação não são receita). Verifica-se que estabelecer um preço de reserva maximiza o benefício esperado do comprador. A lógica econômica por trás disso é muito simples, ela está relacionada à precificação em um monopsônio padrão: assim como um monopsonista padrão oferece um preço inferior ao custo marginal para extrair excedente dos vendedores de menor valoração, ao custo de excluir aqueles com maior valoração que não conseguem vender seus bens, um comprador estabelece um preço máximo de compra para extrair maior excedente esperado dos vendedores de menor valoração, mas exclui a participação daqueles com avaliações mais altas que não conseguem vender ao preço máximo estabelecido.

Obviamente, isso não é eficiente (ex-post) porque em algumas licitações o ente público não conseguirá adquirir o contrato, enquanto a eficiência dita que a contratação deveria ocorrer. Isso é análogo à perda de eficiência (peso morto) gerada quando um monopsonista padrão não adquire bens de vendedores de menor valoração.

3.4 Conclusão

Neste capítulo, desenvolveu-se um modelo matemático invertido para leilões de compra, aplicando a teoria dos jogos bayesianos e o conceito de equilíbrio de Nash Bayesiano. Explorando os diferentes tipos de leilões, incluindo leilões abertos e de primeiro preço, e incorporando custos de transação e preços de reserva ao modelo. Este desenvolvimento teórico é crucial para entender como os custos de transação impactam a formação de preços nas contratações públicas, um aspecto fundamental do problema de pesquisa: "Modelos matemáticos de leilões e técnicas de aprendizado de máquina podem prever custos de transação nas contratações públicas no estado da Paraíba?"

Os achados deste capítulo indicam que a criação de custos de transação pode prejudicar os preços finais das licitações, reforçando a necessidade de estratégias que minimizem esses custos para otimizar a eficiência das contratações públicas. A análise matemática demonstrou que variáveis como o preço de reserva e a taxa de entrada (que pode ser visto como custo de transação) têm um impacto significativo no comportamento dos licitantes e nos resultados dos leilões. Esses insights são diretamente conectados aos objetivos da pesquisa, que incluem desenvolver um modelo matemático para identificar as variáveis que impactam os custos de transação e aplicar técnicas de aprendizado de máquina para prever esses custos.

O próxim capítulo da pesquisa se concentrará na aplicação empírica dessas técnicas de aprendizado de máquina para prever os custos de transação nas contratações públicas. Utilizando os fundamentos teóricos estabelecidos nesta parte, serão treinados e validados modelos de regressão para analisar dados reais das contratações públicas na Paraíba. Essa abordagem integrativa permitirá não apenas a validação prática do modelo matemático desenvolvido, mas também a criação de ferramentas práticas que podem ser utilizadas para aumentar a eficiência e a transparência nos processos de contratação pública, alinhando-se com os objetivos gerais da dissertação.

Capítulo 4

Modelos de Aprendizado de máquina: *LASSO Random Forest* e *Gradient Boosting*

Neste capítulo, serão analisados os algoritmos de aprendizado de máquina mais avançados que foram utilizados na presente pesquisa, começando pela Heurística do *LASSO* e depois passando para as árvores de decisão que são pressupostos teóricos do *Random Forest* e do *Gradient Boosting*. As Árvores de decisão são algoritmos de aprendizado de máquina utilizados para tarefas de classificação e regressão. Como será visto, elas funcionam dividindo iterativamente os dados de entrada em subconjuntos baseados em atributos específicos, criando uma estrutura de árvore onde cada nó interno representa uma decisão baseada em um atributo e cada folha representa uma saída ou classe. A principal vantagem das árvores de decisão é sua simplicidade e interpretabilidade, pois os caminhos de decisão podem ser facilmente visualizados e entendidos.

Além das árvores de decisão, existem outros algoritmos de aprendizado de máquina baseados em conjuntos de árvores, como *Random Forest* e *Gradient Boosting*, que abordam algumas das limitações do algoritmo individuais. Ambos os métodos, ao agregarem múltiplas árvores, aumentam a robustez e a capacidade preditiva dos modelos, tornando-os populares em aplicações práticas de aprendizado de máquina.

4.1 *LASSO*

A técnica seguinte à regressão stepwise é o "*LASSO*" que costuma ser mais rápido e com mais garantias teóricas do que o *stepwise*, sendo considerada uma heurística). "o *LASSO* consiste em encontrar uma solucao que minimize a soma de seu erro quadratico medio com uma medida de complexidade de β" (Izbicki e Santos, 2020,

p.37). A ideia e reduzir a variancia do estimador de minimos quadrados.

Em termos topológicos, como está se medindo a distância (euclidiana) para os betas e de todos os parâmetros diferentes de zero, muitos parâmetros serão zerados e, portanto, não serão computados. Assim, o *LASSO* consegue diminuir bruscamente a quantidade de modelos a ser avaliado (Hastie, Tibshirani, Fiedman, 2009). Para o melhor uso do *LASSO*, é importante calcular o lambda (λ) mais eficiente, através do tuning parameter, podendo ser feito essa análise por validação cruzada.

O método *LASSO* propõe-se a aprimorar a estimação nos modelos de regressão linear, minimizando o risco de sobreajuste, e oferece benefícios notáveis em comparação com técnicas como a regressão stepwise (Izbicki e Santos, 2020). Pois, a eficiência computacional, que supera a regressão stepwise, especialmente em cenários com um grande número de variáveis, situação em que a busca exaustiva por 2^d modelos é impraticável. Além disso, a capacidade inerente do *LASSO* de realizar seleção de variáveis, identificando automaticamente os preditores mais significativos e, assim, reduzindo a dimensionalidade do problema.

A seleção de variáveis do *LASSO* é alcançada através da aplicação de uma penalidade que promove a esparsidade dos coeficientes, fazendo com que os menos significativos sejam reduzidos a zero (Izbicki e Santos, 2020). Formalmente, o *LASSO* soluciona o problema de otimização:

$$\hat{\beta}_{L1,\lambda} = \arg\min_{\beta_0,\beta\in\mathbb{R}^d} \left\{ \sum_{k=1}^{n} (y_k - \beta_0 - \sum_{i=1}^{d} \beta_i x_{k,i})^2 \right\} + \lambda \sum_{j=1}^{d} |\beta_j|,$$

de forma que λ é o parâmetro de ajuste que controla a intensidade da penalidade. À medida que λ aumenta, mais coeficientes são reduzidos a zero, conduzindo a modelos mais parcimoniosos (Izbicki e Santos, 2020).

Para λ suficientemente grande, observa-se que:

$$\sum_{k=1}^{n} (y_k - \beta_0 - \sum_{j=1}^{d} \beta_j x_{k,j})^2 + \lambda \sum_{j=1}^{d} |\beta_j| \approx \lambda \sum_{j=1}^{d} |\beta_j|,$$

resultando em que $\hat{\beta}_1 = 0, \ldots, \hat{\beta}_d = 0$. A seleção do parâmetro λ é geralmente realizada por meio de validação cruzada, procurando-se o valor que minimiza o erro de validação cruzada, garantindo assim a generalização do modelo (Izbicki e Santos, 2020).

4.2 Árvores de decisão

Na discussão dos métodos não paramétricos, Izbicki e Santos (2020) explicam as Árvores de regressão, *Bagging* e Florestas Aleatórias. Aqui, nessa dissertação serão usadas apenas as Florestas Aleatórias (*Random Forest*), mas a explicação inicial

Figura 4.1: Exemplo de estrutura de uma árvore de regressão.

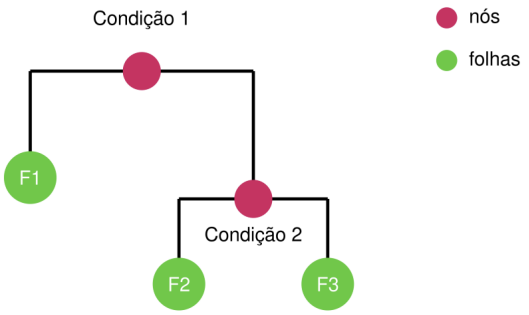

Figura 4.2: Fonte: Izbicki e Santos (2020), p. 77.

será de árvores de regressão.

As árvores de regressão, como descrito por Izbicki e Santos (2020), são um exemplo de metodologia de aprendizado de máquina não-paramétrica que permite a obtenção de modelos com alta interpretabilidade. O processo de construção de uma árvore inicia com a divisão recursiva do espaço das covariáveis em regiões distintas e disjuntas, cada uma delas correspondendo a um nó da árvore, e culmina em nós terminais denominados folhas.

A aplicação da árvore para a predição de novas observações começa no nó raiz, seguindo para a esquerda ou direita conforme a condição em cada nó seja satisfeita, até atingir uma folha que concede a predição, como ilustrado no livro de Izbicki e Santos (2020), na Figura:

Formalmente, uma árvore segmenta o espaço das covariáveis em regiões $R_1, \ldots, R_j$, e a previsão para a resposta Y de uma observação com covariáveis x que cai na região R_k é dada pela média das respostas das amostras de treinamento nessa região (Izbicki e Santos, 2020):

$$g(x) = \frac{1}{\|\{i : x_i \in R_k\}\|} \sum_{i:x_i \in R_k} y_i. \tag{4.1}$$

Para se chegar ao valor da variável resposta x, investiga-se em qual região está a observação x e, em seguida, com as amostras no conjunto de treinamento que pertencem àquela mesma região, calcula-se a média dos valores da variável resposta.

Como destacada por Izbicki e Santos (2020), a construção de uma árvore de regressão envolve duas etapas principais: (i) a criação de uma árvore completa e (ii) o processo de podagem dessa árvore, com o objetivo de evitar o overfitting (super ajuste).

Na primeira etapa, busca-se construir uma árvore que produza partições "puras", ou seja, partições nas quais os valores de Y sejam homogêneos, em cada uma das

folhas . Para isso, avalia-se a qualidade da árvore T através do seu erro quadrático médio (EQM), definido por:

$$\mathcal{P}(T) = \sum_{R} \sum_{i:x_i \in R} \frac{(y_i - \hat{y}_R)^2}{n}, \tag{4.2}$$

em que o valor predito ($\hat{y}_R$) para a resposta de uma observação está contida na região R. Encontrar a árvore T que minimize $\mathcal{P}(T)$ é computacionalmente inviável, mas O objetivo é encontrar uma árvore que minimize esse erro, um processo que é computacionalmente desafiador e frequentemente requer o uso de heurísticas, conforme explicado por Izbicki e Santos (2020).

Assim, utiliza-se uma heurística para encontrar uma árvore com um EQM baixo que seja adequada para as criação de divisões binárias recursivas. Em um primeiro momento, o algoritmo particiona em duas regiões o espaço das covariáveis. Para escolher essa uma dessas partições, o algoritmo avalia todas as possíveis covariáveis x_i e pontos de corte t_1, buscando a combinação que resulta na menor soma dos erros quadráticos das predições nas duas regiões R_1 e R_2:

$$\sum_{i:x_i \in R_1} (y_i - \hat{y}_{R_1})^2 + \sum_{i:x_i \in R_2} (y_i - \hat{y}_{R_2})^2, \tag{4.3}$$

de modo que $\hat{y}_{R_k}$ é a predição fornecida para a região R_k. Assim, a partição ideal minimiza essa soma dos erros quadráticos. As regiões R_1 e R_2 são definidas da seguinte maneira:

$$R_1 = \{x : x_i < t_1\} \quad \text{e} \quad R_2 = \{x : x_i \geq t_1\}, \tag{4.4}$$

sendo x_i a variável escolhida e t_1 é o ponto de corte que define a partição. Esta abordagem garante que cada divisão resulte nas regiões mais homogêneas possíveis em relação ao valor da variável resposta.

Essas escolhas (da covariável e do ponto de corte) são feitas de forma a minimizar o erro quadrático médio nas duas regiões, garantindo assim que a árvore de decisão obtida tenha um bom desempenho preditivo.

O procedimento continua de forma recursiva, até que se atinja uma árvore em que cada folha contenha um número reduzido de observações (por exemplo, o processo pode ser interrompido quando todas as folhas tiverem menos de dez observações). Embora a árvore criada por esse método produza bons resultados no conjunto de treinamento, é altamente suscetível ao superajuste. Isso significa que, embora a árvore se ajuste bem aos dados de treinamento, sua capacidade preditiva para novas observações será comprometida. Para mitigar esse problema, avança-se para a etapa (ii), conhecida como poda (Izbicki e Santos, 2020).

O objetivo da poda é simplificar a árvore de regressão, reduzindo seu tamanho e

complexidade, o que ajuda a diminuir a variância do estimador. Durante essa fase, cada nó da árvore é removido sequencialmente, e o impacto dessa remoção no erro de predição é avaliado usando um conjunto de validação. Com base nessa análise, decide-se quais nós devem ser mantidos na árvore para otimizar seu desempenho preditivo em dados novos. A poda, portanto, é crucial para obter um modelo que não só se ajuste bem aos dados de treinamento, mas também tenha uma boa capacidade de generalização.

Uma característica notável das árvores de regressão é sua alta interpretabilidade. No entanto, essas árvores geralmente apresentam baixo poder preditivo quando comparadas com outros estimadores. Métodos como *Bagging* e *Random Forest* são técnicas que superam essa limitação, combinando múltiplas árvores para fazer uma única predição para o mesmo problema.

Para ilustrar essa abordagem, usar-se-á o mesmo exemplo de Izbicki e Santos (2020). Assim, considere um contexto de regressão em que se tem duas funções de predição para Y, denotadas por $g_1(x)$ e $g_2(x)$. Os riscos dessas funções, condicionais em x, mas não nos dados de treinamento, são dados, respectivamente, por:

$$\mathbb{E}\left[(Y - g_1(x))^2 \mid x\right] \quad \text{e} \quad \mathbb{E}\left[(Y - g_2(x))^2 \mid x\right].$$

Agora, considere o estimador combinado $g(x) = \frac{g_1(x)+g_2(x)}{2}$. Têm-se a equação a seguir, apresentada por Izbicki e Santos (2020, p.83), que esclarece a expectativa do erro quadrático de uma previsão, evidenciando o benefício do uso combinado de vários estimadores sobre o uso individual.

$$E\left[(Y - g(x))^2|x\right] = Var[Y|x] + \frac{1}{2}Var[g_i(x)|x] \leq E\left[(Y - g_i(x))^2|x\right]. \tag{4.5}$$

Essa formulação ressalta a eficiência de combinar estimadores para reduzir o erro de previsão, reiterando a importância das técnicas de *ensemble*[1] na construção de modelos preditivos robustos. As estratégias de *ensemble*, incluindo *bagging*, *Random Forest* e *gradient boosting*, são empregadas para ampliar a acurácia das previsões de modelos isolados, uma abordagem amplamente discutida por Izbicki e Santos (2020). Esses métodos coletivos superam as limitações inerentes a estimadores únicos através da combinação inteligente.

[1] Os métodos de *ensemble* usam vários algoritmos de aprendizagem para obter melhor desempenho preditivo do que poderia ser obtido apenas com qualquer um dos algoritmos de aprendizagem constituintes.

4.3 Técnicas de *Ensemble*

O *ensemble* é uma técnica de aprendizado de máquina que combina diversos modelos que geram múltiplas árvores de decisão, cada uma aprendendo a partir de uma amostra bootstrap dos dados, gerando estimadores $g_b(x)$ cuja média forma a previsão final (Izbicki e Santos, 2020). O objetivo é diminuir a variância mantendo o viés controlado.

Inicialemente, analisar-se-á o *bagging* que é uma técnica que envolve a criação de múltiplas versões do modelo de aprendizado, cada uma treinada em diferentes subconjuntos dos dados de treinamento, utilizando a amostragem com reposição, o que resulta em diversos conjuntos *bootstrap*. Em seguida, diferentes modelos são treinados de forma independente em cada um desses conjuntos *bootstrap*. As previsões desses modelos são então agregadas para produzir uma previsão final, sendo que, para problemas de regressão, a média das previsões individuais é usada, enquanto para problemas de classificação, aplica-se a votação majoritária. A principal vantagem do bagging reside na capacidade de reduzir a variância do modelo sem aumentar significativamente o viés, resultando em uma melhoria substancial na precisão e na capacidade de generalização do modelo final.

A função de predição do método *bagging* é então definida, considerando $g_b(x)$ a função de predição obtida a partir da b-ésima árvore, como:

$$g(x) = \frac{1}{B}\sum_{b=1}^{B} g_b(x).$$

Embora o *bagging* produza preditores de difícil interpretabilidade, ele permite a criação de uma medida de importância para cada covariável. Essa medida de importância é baseada na redução da soma dos quadrados dos resíduos (*RSS - residual sum of squares*) em cada divisão da árvore.

Como se viu, as árvores de decisão são conhecidas por serem extremamente sensíveis ao ruído, o que faz com que se beneficiem significativamente da técnica de média. Além disso, como cada árvore gerada no processo de *bagging* é distribuída de forma idêntica (i.d.), a expectativa da média de B dessas árvores é a mesma que a expectativa de qualquer uma delas. Isso implica que o viés das árvores combinadas por *bagging* é o mesmo que o viés das árvores individuais, e a única possibilidade de melhoria reside na redução da variância.

Isso contrasta com o método de *boosting*, como será visto adante, em que as árvores são geradas de maneira adaptativa para reduzir o viés, e, portanto, não são distribuídas de forma idêntica. No *boosting*, cada árvore subsequente é ajustada para corrigir os erros das árvores anteriores, focando na melhoria da predição ao diminuir o viés iterativamente.

A média de B variáveis aleatórias i.i.d., cada uma com variância σ^2, possui variância $\frac{1}{B}\sigma^2$. Se as variáveis são apenas i.d. (identicamente distribuídas, mas não necessariamente independentes) com correlação par a par positiva ρ, a variância da média é dada por:

$$\rho\sigma^2 + \frac{1-\rho}{B}\sigma^2.$$

À medida que B aumenta, o segundo termo diminui, mas o primeiro permanece. Portanto, o tamanho da correlação entre pares de árvores em *bagging* limita os benefícios da média. A ideia das florestas aleatórias é melhorar a redução da variância do *bagging* ao diminuir a correlação entre as árvores, sem aumentar demasiadamente a variância. Isso é alcançado no processo de crescimento das árvores através da seleção aleatória das variáveis de entrada.

4.3.1 *Random Forest*

O Random Forest é um algoritmo de aprendizado de máquina baseado em um conjunto de árvores de decisão. Como será visto adiante, cada árvore no conjunto é treinada com um subconjunto aleatório dos dados de treinamento, usando uma técnica derivada do *bagging*.

Para construir uma floresta aleatória, segue-se o algoritmo abaixo (Hastie, Tibshirani e Friedman, 2009):

1. Para $b = 1$ até B:

(a) Para cada uma das B árvores, extrai-se uma amostra *bootstrap* $\mathbf{Z}$ de tamanho N a partir do conjunto de dados de treinamento.

(b) Constroe-se uma árvore T_b utilizando os dados *bootstrap*, repetindo recursivamente os seguintes passos para cada nó terminal da árvore, até que o tamanho mínimo do nó $n_{\min}$ seja atingido:

i. Seleciona-se m variáveis aleatórias dentre as d variáveis disponíveis.

ii. Escolhe-se a melhor variável/ponto de divisão entre as m selecionadas, para minimizar uma medida de impureza, como o erro quadrático médio (RMSE).

iii. Divide-se o nó em dois nós filhos.

2. O conjunto de árvores $\{T_b\}_{b=1}^{B}$ é então gerado.

Para fazer uma predição em um novo ponto x, no caso de regressão, a predição $\hat{f}_{\text{rf}}^{B}(x)$ é dada por:

$$\hat{f}_{\text{rf}}^{B}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x).$$

Esta abordagem garante que o modelo final seja robusto e tenha uma capacidade de generalização superior, devido à combinação de múltiplas árvores de decisão

treinadas em diferentes subconjuntos dos dados e variáveis.

Especificamente, ao crescer uma árvore em um conjunto de dados *bootstrap*, Antes de cada divisão, selecione $m \leq d$ das variáveis de entrada aleatoriamente como candidatas para a divisão[2] (Hastie, Tibshirani e Friedman, 2009).

Este método de seleção aleatória de variáveis antes de cada divisão reduz a correlação entre as árvores, permitindo que o *Random Forest* mantenha a variância reduzida do *bagging* enquanto melhora a robustez e a capacidade preditiva do modelo.

Assim, ao *Random Forest* é introduzida uma variação adicional entre as árvores ao selecionar aleatoriamente um subconjunto de m covariáveis em cada divisão dos nós, de forma que m é menor que o número total d de covariáveis (Izbicki e Santos, 2020). A determinação de m pode ser realizada por meio de validação cruzada, e uma escolha comum é $m \approx \sqrt{d}$, que tende a oferecer uma performance equilibrada.

De maneira geral para o *Random Forest*, realizar uma validação cruzada para determinar o valor ótimo de B não resulta em grandes benefícios, contrastando com o ajuste de parâmetros em outros métodos. A robustez do desempenho do *Random Forest* em relação à escolha de B é uma das principais vantagens desse método.

A razão para essa robustez é que o desempenho do *Random Forest* tende a estabilizar com um número relativamente grande de árvores. Isso ocorre porque, após certo ponto, adicionar mais árvores não contribui significativamente para a melhoria do modelo, uma vez que a média dos modelos tende a convergir para uma predição estável. Assim, escolher um B suficientemente grande geralmente é suficiente para garantir um bom desempenho, sem a necessidade de ajustes finos através de validação cruzada.

4.3.2 *boosting*

Assim como o *Random Forest* e o *bagging*, o *boosting* também se baseia na combinação de vários estimadores de regressão. No entanto, a maneira como essa combinação é feita é diferente. Existem várias versões e implementações de *boosting*, mas aqui será apresentada por Izbicki e Santos (2020).

No *boosting*, o estimador $g(x)$ é construído de forma incremental. Inicialmente, define-se $g(x) = 0$. Este estimador inicial tem um alto viés, mas variância nula. A cada iteração, o valor de g é ajustado para reduzir o viés e aumentar a variância. Isso é feito adicionando a g uma função que prevê os resíduos $r_i = Y_i - g(x_i)$. Uma maneira comum de fazer isso é utilizando uma árvore de regressão. É crucial que essa árvore tenha uma profundidade limitada para evitar o *overfitting*. Além disso, em vez de adicionar essa função integralmente, ela é multiplicada por um fator λ

[2] Quando m = d, o método do *Random Forest* se reduz ao *bagging*.

(conhecido como taxa de aprendizado), que varia entre 0 e 1 para controlar o ajuste excessivo. Formalmente, o algoritmo de *boosting* segue os seguintes passos:

1. Define-se $g(x) \equiv 0$ e $r_i = y_i$ para $i = 1, \ldots, n$.
2. Para $b = 1, \ldots, B$:
   (a) Treina-se uma árvore de decisão com p folhas utilizando os pares $(x_1, r_1), \ldots, (x_n, r_n)$. Denota-se a função de predição dessa árvore como $g^b(x)$.
   (b) Atualiza-se g e os resíduos: $g(x) \leftarrow g(x) + \lambda g^b(x)$ e $r_i \leftarrow Y_i - g(x)$.
3. Retorna-se o modelo final $g(x)$.

Esse procedimento de *boosting* permite a construção de um estimador final que é uma combinação de vários modelos simples, resultando em um modelo mais robusto e com melhor capacidade de previsão (Izbicki e Santos, 2020).

Tanto o *Boosting* quanto o *Random Forest* são métodos de *ensemble learning* que combinam múltiplos modelos para melhorar a capacidade preditiva em relação a modelos individuais. No entanto, eles diferem significativamente em seus princípios e abordagens. Ambos os métodos utilizam árvores de decisão como seus modelos base e aumentam a robustez e a capacidade de generalização dos modelos ao reduzir a variância e/ou o viés.

No *Random Forest*, várias árvores de decisão são construídas de forma independente. Cada árvore é treinada em um subconjunto aleatório dos dados (*bootstrap*) e, em cada nó de uma árvore, um subconjunto aleatório de características é considerado para a divisão. A predição final é obtida através da média das predições das árvores individuais. Por outro lado, no *Boosting*, os modelos são construídos sequencialmente, de modo que cada modelo tenta corrigir os erros do modelo anterior.

A principal melhoria no *Random Forest* vem da redução da variância do modelo através da média das predições de múltiplas árvores que são treinadas de forma independente. No *Boosting*, a principal melhoria vem da redução do viés do modelo através do ajuste sequencial aos erros das predições anteriores. Em termos de complexidade computacional, o *Random Forest* permite uma paralelização mais eficiente, já que as árvores são treinadas de forma independente, enquanto o treinamento no *Boosting* é sequencial, o que pode resultar em tempos de treinamento mais longos.

Gradient Boosting

No capítulo seguinte, do treinameto dos modelos, será utilizada uma das formas mais populares de Boosting que é o *Gradient Boosting*. Nesta técnica, cada novo modelo é

treinado para corrigir o resíduo (erro) do modelo combinado anterior. A abordagem do *Gradient Boosting* é baseada na otimização do erro de predição através da descida do gradiente.

Matematicamente, considerando um conjunto de dados com n observações $\{(x_i, y_i)\}_{i=1}^{n}$. No *Gradient Boosting*, começa-se com um modelo inicial $g_0(x)$, que pode ser uma predição constante, como a média dos valores alvo y. Em cada iteração m, ajustou-se um novo modelo $h_m(x)$ aos resíduos dos modelos anteriores. Especificamente, o m-ésimo modelo é treinado para minimizar a função do erro quadrático médio (já analisado no 2.3 da fundamentação teórica):

$$L_m = \sum_{i=1}^{n} \left(y_i - g_{m-1}(x_i)\right)^2,$$

sendo $g_{m-1}(x)$ a predição do modelo combinado até a iteração $m-1$. O novo modelo $g_m(x)$ é então ajustado aos gradientes dos resíduos, que são as derivadas da função de perda em relação às predições (Hastie, Tibshirani e Friedman, 2009):

$$g_{i,m} = \frac{\partial L(y_i, g_{m-1}(x_i))}{\partial g_{m-1}(x_i)} = y_i - g_{m-1}(x_i).$$

O modelo combinado é atualizado somando-se o novo modelo ajustado multiplicado por uma taxa de aprendizado η:

$$g_m(x) = g_{m-1}(x) + \eta h_m(x).$$

Aqui, η é um hiperparâmetro que controla a contribuição de cada modelo fraco para o modelo combinado final. O processo continua iterativamente até que um critério de parada seja atingido, como um número máximo de iterações ou uma melhoria mínima na função de perda.

4.4 Conclusão

Para concluir este capítulo, destacamos a importância dos métodos de ensemble, como o Random Forest e o Gradient Boosting, na construção de modelos de aprendizado de máquina robustos e eficazes. Estes métodos combinam múltiplos modelos base para melhorar a precisão preditiva e a capacidade de generalização, mitigando problemas como o overfitting e a variabilidade dos dados. Ao longo deste capítulo, foram apresentados os princípios teóricos e as vantagens dessas técnicas, evidenciando como cada uma contribui de forma distinta para a otimização do desempenho dos modelos.

No capítulo seguinte, serão detalhados os procedimentos de treinamento dos modelos, aplicando os conceitos discutidos aqui para desenvolver soluções práticas e

eficientes. O foco será em implementar e avaliar esses métodos, ajustando seus hiperparâmetros para maximizar a performance nos contextos específicos das contratações públicas.

Capítulo 5

Treinamento dos modelos de regressão

Este capítulo aborda a aplicação de técnicas de aprendizado de máquina para prever os custos de transação nas contratações públicas, um passo crucial para responder ao problema de pesquisa: "Modelos matemáticos de leilões e técnicas de aprendizado de máquina podem prever custos de transação nas contratações públicas no estado da Paraíba?" O objetivo principal é explorar e comparar modelos de regressão utilizando variáveis socioeconômicas e administrativas para criar previsões precisas dos custos de transação, contribuindo para os objetivos da dissertação de aumentar a eficiência e a transparência nos processos licitatórios.

A análise foi conduzida utilizando dois modelos principais de regressão: *Random Forest*, *Gradient Boosting* e *LASSO*. Cada modelo foi avaliado com base em métricas de desempenho, como o erro quadrático médio (RMSE), o coeficiente de determinação (R^2) e o erro absoluto médio (MAE). Além disso, a importância das variáveis no modelo *Random Forest* foi analisada para entender melhor os fatores que mais influenciam os custos de transação. A preparação dos dados envolveu a organização de informações detalhadas de notas fiscais, empenhos, licitações e dados socioeconômicos, resultando em uma base de dados robusta e extensa.

A relevância desses temas é destacada pela necessidade de identificar e prever os custos de transação com precisão, uma vez que esses custos impactam significativamente a formação de preços nas contratações públicas. A lacuna enfrentada na literatura, em que poucos estudos integram aprendizado de máquina com a teoria dos leilões para abordar especificamente os custos de transação, é suprida através desta abordagem empírica.

A próxima seção detalhará o processo de construção e ajuste dos modelos de regressão, os resultados obtidos e a análise da importância das variáveis, fornecendo uma compreensão abrangente do impacto das variáveis selecionadas nos custos de transação. Esta abordagem integrativa visa descobrir padrões ocultos nos dados

históricos de contratações públicas e produzir um modelo robusto que pode ser utilizado para otimizar os processos de contratação, alinhando-se com os objetivos da pesquisa de promover uma alocação mais eficaz dos recursos públicos.

5.1 Preparação e Análise Exploratória dos Dados

Apesar de ter tido todo o apoio do Tribunal de Contas do Estado da Paraíba na obtenção dos dados, sua organização não foi fácil nem trivial devido a algumas inconsistências e disposição dos dados.

Decidiu-se, por disponibilidade dos dados, fazer um recorte temporal de 2019 a 2023, conseguindo todos os dados de: Notas Fiscais (de compras dos entes públicos); Empenhos; Licitações; Contratos; Restos a Pagar; Receita Corrente líquida; e Precatórios. O tamanho da base de dados ultrapassa 10 GB, principalmente os arquivos com os empenhos.

Inicialmente, trabalhou-se na organização das informações das licitações para saber quais são as mais frequentes e se seria possível cruzar os dados com as informações de empenho e pagamento. Deste trabalho, identificaram-se as contratações mais frequentes:

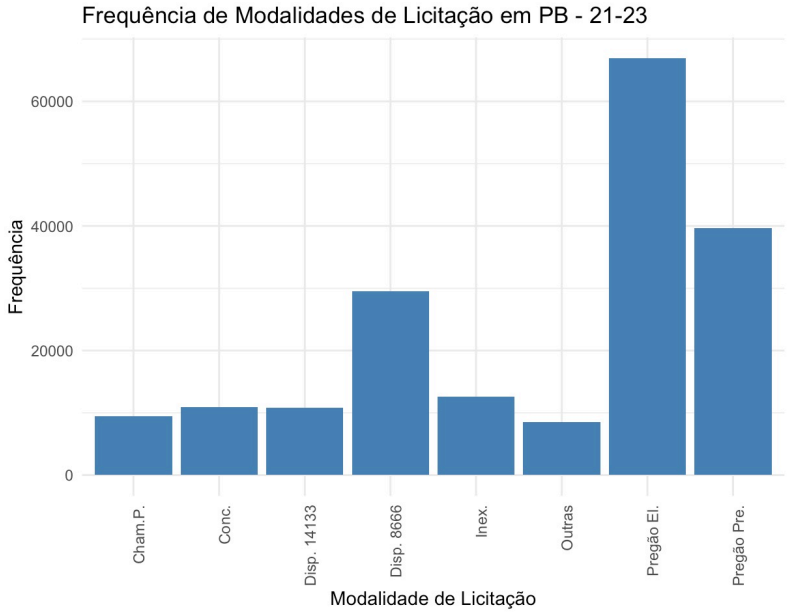

Figura 5.1: Frequência das modalidades de licitação nos municípios paraibanos.

Assim, vê-se que a maior parte das contratações é realizada na forma de pregão (seja presencial ou eletrônico), contratações diretas (dispensa e inexigibilidade) e concorrência. Vale salientar que a chamada pública não é um tipo de aquisição, mas de seleção pública, não sendo relevante para esse trabalho.

A análise dos dados de empenho foram trabalhosas, devido à demanda computacional, pois, por ano, há mais de 2 milhões de empenhos dos entes públicos. Era possível cruzar as informações com o arquivo das licitações, mas não se tem o dado sobre quais itens estavam sendo adquiridos, logo, não era possível fazer qualquer análise de dados, então foi preciso reformular a abordagem para avançar através das notas fiscais.

O TCE, através de convênio firmado com a SeFaz-PB, tem todas as notas fiscais emitidas após 2019, separada por ano de emissão. Ocorre que a base de dados está dividida entre os dados da nota fiscal/empresa e os dados com relação aos produtos. Esses dataframes foram unificados pela identificação da nota fiscal.

Como o objetivo é fazer uma previsão do custo de transação, a variável dependente (Y) passou a ser a variação, em desvios padrão, em torno dos preços deflacionados dos produtos. Assim, precisa-se isolar cada produto, para ver seu preço médio e o desvio padrão dele, considerando a unidade mais frequente daquele bem. Para isso, na base das notas/produto, têm-se três códigos: Código EAN, Código NCM e Código CEST.

O Código EAN de um produto, também conhecido como Código de Barras Europeu (European Article Number), é um sistema de codificação numérica padrão usado globalmente para identificar produtos de maneira única, sendo muito detalhada e ideal para identificar produtos específicos. Por sua vez, o Código NCM (Nomenclatura Comum do Mercosul) é mais usado para a classificação mercadorias no comércio internacional, sendo bem menos detalhado que o EAN, categorizando produtos em grupos mais amplos. Por fim, o Código CEST (Código Especificador da Substituição Tributária) é utilizado para especificar produtos sujeitos à substituição tributária do ICMS. Deste modo, foi feita a seleção dos produtos pelo EAN, excluindo os que não tinham esse cadastro.

Em seguida, foram acrescentadas ao dataframe informações geográficas e econômica do estado e dos municípios. Unificando os dados pelo código dos municípios (padronizado pelo IBGE), ficando todos em um dataframe. Assim, pode-se calcular as distâncias entre a cidade da emissão da nota fiscal e o destino do produto.

Outro trabalho de organização dos dados foi para excluir os erros, pois apesar de o número EAN ser único, foram identificados erros de alguns produtos, levando-os à exclusão. Além disso, foram observadas as variáveis independentes que continham valores ausentes, sendo retiradas da análise.

Na padronização dos produtos, foram considerados apenas os bens com a mesma unidade, usando a unidade mais frequente daquele produto. Para exemplificar, um bem X que apresentasse 10 notas fiscais, sendo 4 contratações cuja unidade registrada seja "*uni*", 3 com registro de "*comprimido*" e 3 de "*caixa*", neste exemplo hipotético, só foram selecionadas as contratações cuja unidade registrada foi "*uni*",

pois há uma maior quantidade de bens com esta unidade, implicando em uma maior probabilidade de que se trate da mesma grandeza.

Para acrescentar a informação sobre o histórico de pagamento (uma proxi para a reputação do ente público), foram usados dados de restos a pagar de 2018 a 2022, para criar um índice anual de dias médios de atraso de pagamento, para usar com um lag de 1 ano (ou seja, considerou-se que anualmente as expectativas dos vendedores se adaptam), retirando gastos com pessoal. Depois, unificou-se com o dataframe anterior.

Para demonstrar a diversidade de credibilidade dos municípios, pode-se ver o gráfico abaixo:

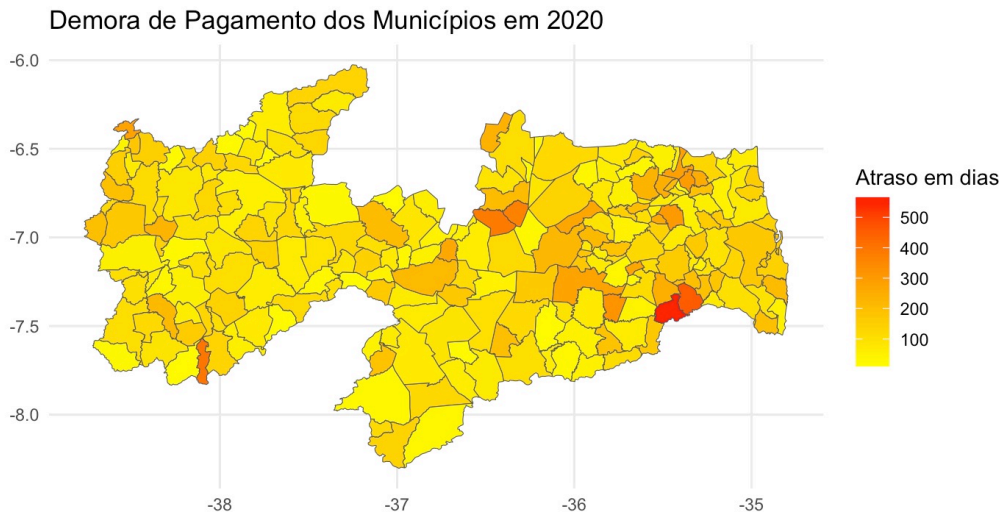

Figura 5.2: Atraso nos pagamentos em 2020.

Agora, todos os anos examinados conjuntamente:

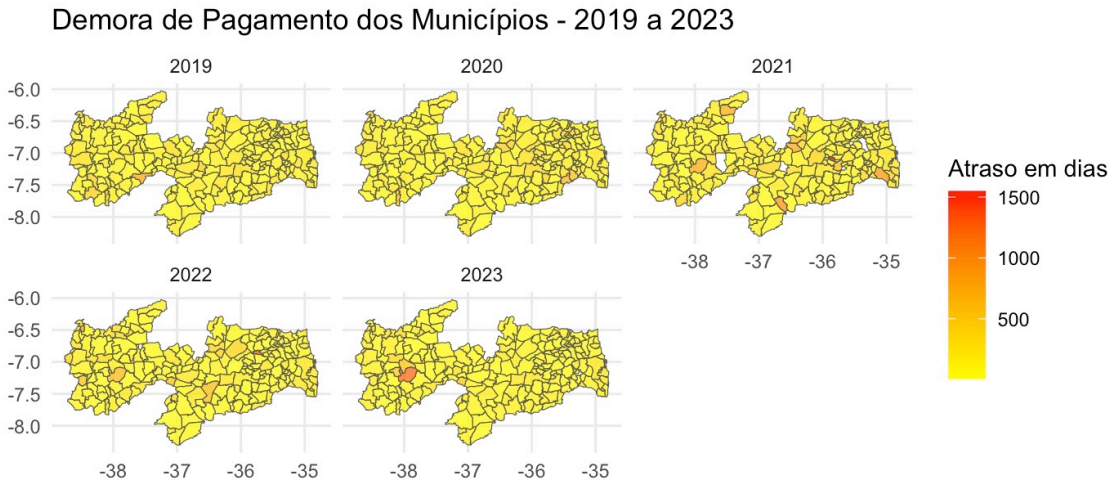

Figura 5.3: Atraso nos pagamentos entre 2019-2023.

Posteriormente, adicionou-se a mesma informação com o acréscimo dos precatórios, considerando a variável o tempo, em meses, que o ente passa para efetuar o pagamento desses títulos judiciais.

Ainda foram acrescentadas as informações sobre os prefeitos dos municípios, considerando as informações disponíveis no TRE-PB sobre os gestores eleitos nos anos em questão.

Outra variável acrescentada foi a receita corrente líquida mensal dos municípios. Como resultado, construiu-se um dataframe com as seguintes variáveis:

Tabela 5.1: Definição das variáveis.

Variável	**Descrição**
Código EAN Produto	Código de barras internacional do produto
Descrição Produto	Descrição detalhada do produto
Quantidade	Quantidade do produto
Unidade	Unidade de medida do produto
Valor Unitário	Preço por unidade do produto
Valor Produtos	Valor total dos produtos
Ano	Ano da transação
Mês	Mês da transação
Código Município	Código do município do emitente
Município Emitente	Nome do município do emitente
UF	Unidade Federativa do emitente

Código Município Destinatário	Código do município do destinatário
Município Destinatário	Nome do município do destinatário
UF Destinatário	Unidade Federativa do destinatário
Valor Total Produtos	Valor total de todos os produtos
Valor Total Nota	Valor total da nota fiscal
Chave	Chave única da nota fiscal
Cód. CNAE	Código CNAE da atividade econômica da empresa vendedora
CustoTran	Custo de transação bruto (variação para o menor valor contratado)
pct	Custo de transação Percentual
idhm	Índice de Desenvolvimento Humano Municipal
idhm edu	Índice de Desenvolvimento Humano Municipal - Educação
idhm long	Índice de Desenvolvimento Humano Municipal - Longevidade
idhm renda	Índice de Desenvolvimento Humano Municipal - Renda
gini	Coeficiente de Gini - concentração de renda
area km2	Área em quilômetros quadrados
populacao	População total
Tempo Medio	Tempo médio de pagamento
distancia	Distância entre o município emitente da NF e o destinatário
ReceitaCorrente	Receita corrente do município
Saúde	Índice de preços de saúde
Geral	Índice geral de preços (IPCA)
GeralAc	Índice geral de preços acumulado
Preço	Preço deflacionado do produto
Licit	Indicador da modalidade de licitação
n licit	Número de licitantes
Precatorio	Tempo para pagamento de precatórios
Sexo	Sexo do chefe do executivo municipal
Partido	Partido político do chefe do executivo municipal
Custo Tran norm	Custo de transação normalizado (quantidade de desvios-padrão em relação à media dos preços

A distribuição do custo de transação normatizado tem uma distribuição relativamente simétrica em torno do zero, como se seria de esperar, conforme se vê na

imagem a seguir:

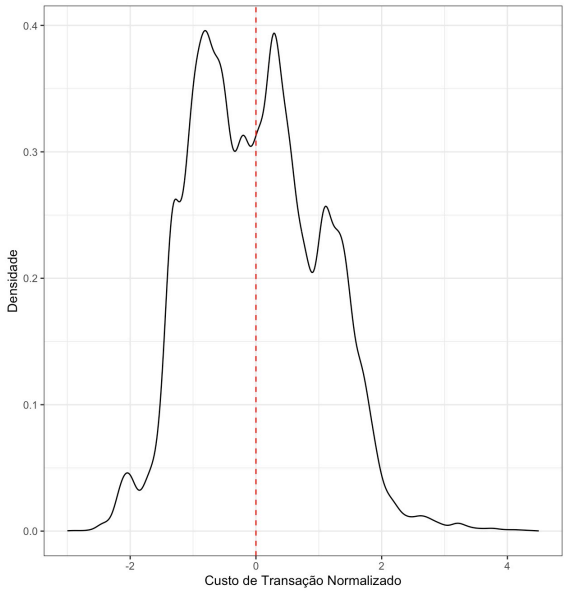

Figura 5.4: Densidade da distribuição do Custo de Transacao Normalizado.

Outra análise que foi realizada, permite ver que as variáveis independentes utilizadas para a explicar os custos de transação não parecem ter uma forte relação entre si, reduzindo o risco de colinearidade. O gráfico de dispersão e correlação das variáveis apresenta a relação entre diferentes indicadores socioeconômicos dos municípios da Paraíba. Os coeficientes de correlação são representados por círculos cuja cor e tamanho variam conforme a força e a direção da correlação entre as variáveis.

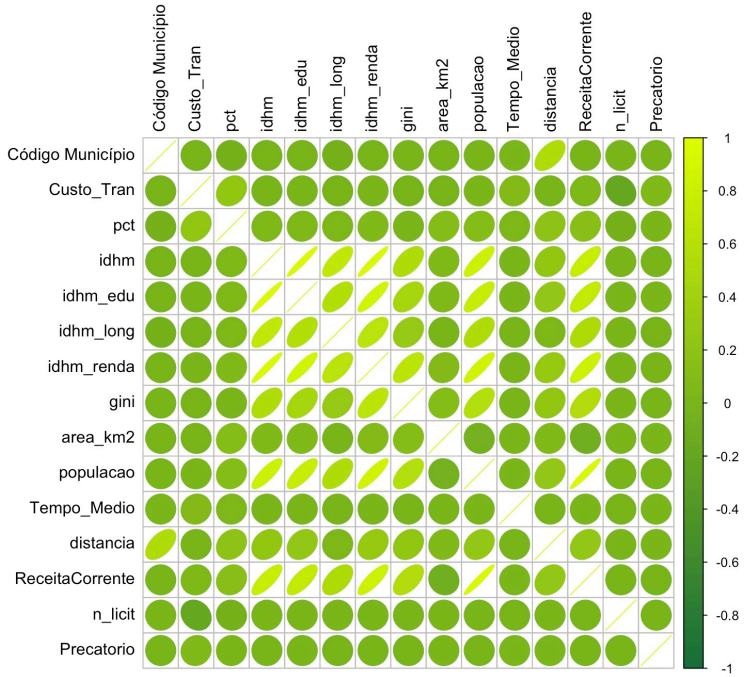

Figura 5.5: Matriz de correlação entre as variáveis independentes.

No gráfico acima, vê-se que a correlação entre a identificação do município e o índice de desenvolvimento humano municipal (IDHM) é desprezível, indicando não haver relaciomento. Existe uma correlação positiva significativa entre o índice de desenvolvimento humano na educação, o índice de desenvolvimento humano na longevidade e o índice de desenvolvimento humano na renda, sugerindo o que era intuitivo: que essas variáveis são correlacionadas. A correlação entre o IDHM e o coeficiente de Gini é baixa, indicando que a desigualdade de renda não tem uma forte relação com o índice de desenvolvimento humano. Há uma correlação positiva entre o IDHM e a população, sugerindo que municípios mais populosos tendem a ter um IDHM mais alto. A correlação entre a área do município e a população é baixa, indicando que a área do município não está fortemente relacionada com a sua população. A correlação entre a população e a receita corrente do município é positiva, sugerindo que municípios mais populosos tendem a ter uma receita corrente maior. Também existe uma correlação positiva entre a receita corrente do município e o número de licitantes nos certames, indicando que municípios com maior receita corrente tendem a realizar licitações com mais concorrentes. A correlação entre a receita corrente do município e o prazo de pagamento dos precatórios é baixa, indicando que a receita corrente do município não está fortemente relacionada com o prazo de pagamento dos precatórios.

Também foram analisadas as correlações entre as variáveis explicativas e a variável dependente (Custo de Transação), apresentando valores baixos, o que pode gerar resultados inesperados, sendo talvez necessário utilizar alguma outras variáveis importante para explicar os custos de transação, mas vale lembrar que esse gráfico só explicita as relações das variáveis numéricas.

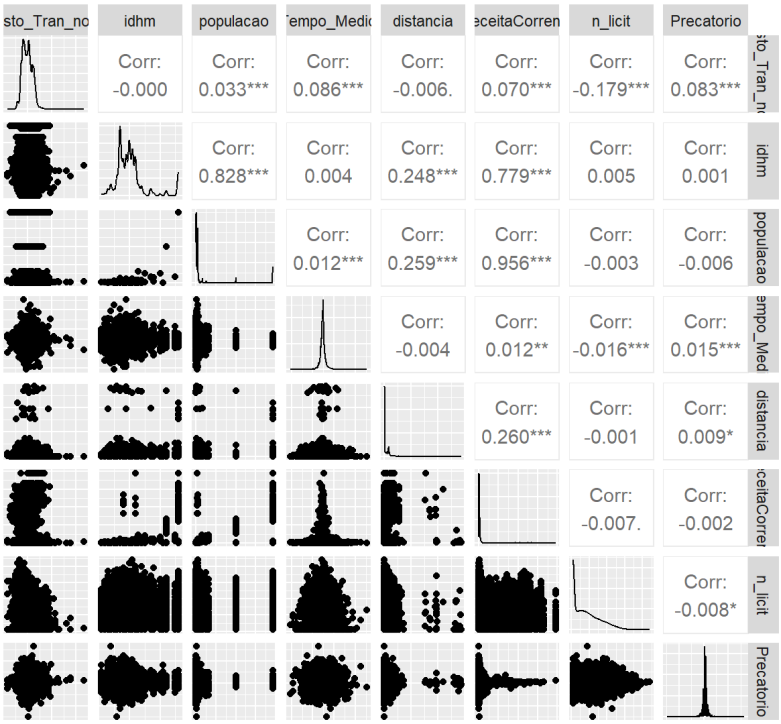

Figura 5.6: Dispersão e correlação das variáveis.

A figura acima demonstra a dispersão e correlação das variáveis, permitindo uma análise visual da força e direção das relações entre as variáveis.

Observa-se que a correlação entre os custos de transação normatizado e o índice de desenvolvimento humano municipal é nula (Corr: -0,000), indicando que não há uma relação direta. A correlação entre os custos de transação normatizado e a população é positiva (Corr: 0,033***), sugerindo que municípios mais populosos tendem a ter custos de transação normatizados ligeiramente mais altos.

A variável tempo medio de pagamento apresenta uma correlação positiva com os custos de transação normatizado (Corr: 0,086***), indicando que quanto mais demora para ter o pagamento, maiores são os custos de transação. Por outro lado, a correlação entre os custos de transação normatizado e as distancias é desprezível (Corr: -0,006), sugerindo que a distância não tem um impacto significativo nos custos de transação normatizados.

Há uma correlação positiva e significativa entre a Receita Corrente e os custos de transação normatizado (Corr: 0,070***), sugerindo que municípios com maior receita corrente tendem a ter custos de transação mais altos. A correlação entre os custos de transação normatizado e a quantidade de licitantes é negativa e significativa (Corr: -0,179***), ou seja, quanto mais licitantes, menores os custos de transação. A correlação entre o tempo de pagamento dos precatórios e os custos de transação normatizado é positiva (Corr: 0,083***), sugerindo que municípios demoram mais para pagar os precatórios tendem a ter custos de transação mais

altos.

Ainda foram avaliadas as distribuições das variáveis independentes o que fornecem uma visão detalhada dos diferentes fatores que podem influenciar os custos de transação das licitações. Pode-se ver nos gráfico a seguir:

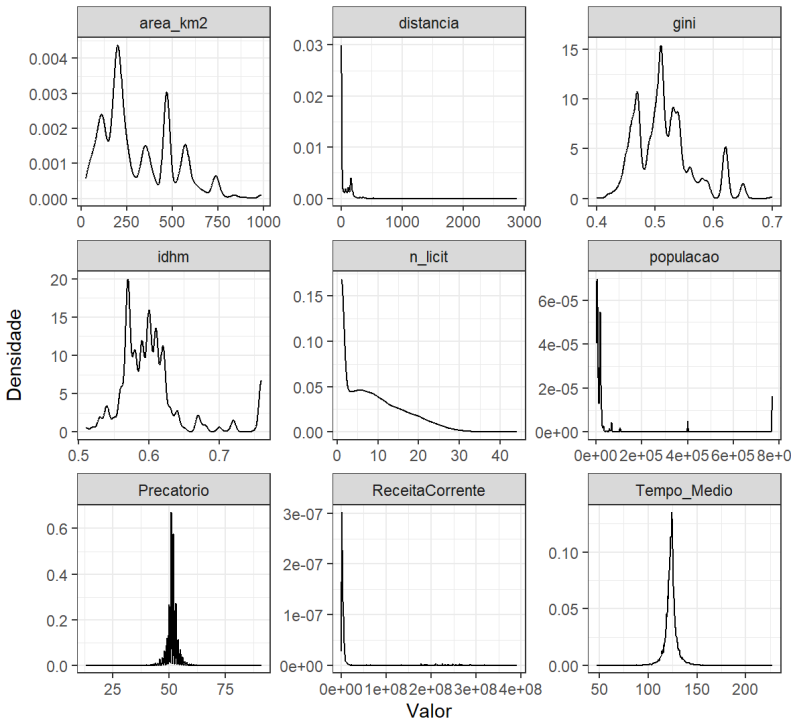

Figura 5.7: Distribuição das variáveis independentes.

Observa-se que a variável 'area km2', que representa o tamanho do município, apresenta uma distribuição multimodal, sugerindo a presença de municípios de diferentes tamanhos, com picos em áreas específicas. A variável 'distancia', que representa a distância entre o município e a sede da empresa emissora da nota fiscal, possui uma concentração elevada em valores mais baixos, indicando que a maioria das transações ocorre em distâncias relativamente curtas.

O índice de Gini ('gini'), que mede a desigualdade de renda no município, mostra uma distribuição com múltiplos picos, sugerindo variações significativas na desigualdade entre os municípios. O 'idhm', que reflete o índice de desenvolvimento humano de cada município, também apresenta uma distribuição multimodal, indicando a presença de diferentes níveis de desenvolvimento humano entre os municípios.

A variável 'n licit', que representa o número de licitantes que participaram da licitação, mostra uma distribuição decrescente, com a maioria dos municípios tendo poucas licitações muito concorridas. A distribuição da 'população' é concentrada em valores baixos, mas com alguns valores extremos elevados, indicando que a maioria dos municípios tem uma população pequena, mas há algumas exceções com populações significativamente maiores.

O tempo (em meses) para pagamento dos precatórios (‘Precatório‘) apresenta uma distribuição concentrada em um intervalo específico, sugerindo uniformidade nos prazos de pagamento dos precatórios em diversos municípios. A variável ‘ReceitaCorrente‘, que representa a receita corrente dos municípios, tem uma distribuição assimétrica, com a maioria dos municípios apresentando receitas relativamente baixas, mas alguns com receitas altas. Finalmente, a variável ‘Tempo Medio‘, que representa o tempo médio para o pagamento das contratações no município, apresenta uma distribuição com um pico acentuado, indicando um tempo médio comum entre os municípios.

Essas distribuições indicam uma heterogeneidade significativa entre os municípios, o que deve ser considerado na modelagem dos custos de transação das licitações públicas. Variáveis como ‘distancia‘, ‘população‘, ‘ReceitaCorrente‘ e ‘Tempo Medio‘ apresentam concentrações em valores específicos que podem ter um impacto substancial nos custos de transação, enquanto ‘area km2‘, ‘gini‘, ‘idhm‘, ‘n licit‘ e ‘Precatório‘ mostram variações que podem influenciar de diferentes maneiras dependendo do contexto específico de cada município.

5.2 Construção e avaliação do Modelo

Inicialmente, seguindo o realizado por Libório e outros (2022), vai-se analisar uma inferência linear simples, não para fazer uma comparação com os modelos de aprendizado de máquina, mas para mostrar que a linearidade não é suficiente:

Tabela 5.2: Resultados da Regressão Linear.

	Variável Dependente:
	Custo de Transação Normalizado
Quantidade	$-0{,}00001^{***}$
	(0,00000)
IDHM	-1.118^{***}
	(0,118)
População	$-0{,}00000^{***}$
	(0,00000)
Tempo Médio	$0{,}012^{***}$
	(0,001)
Distância	$-0{,}0002^{***}$
	(0,00004)
Receita Corrente	$0{,}000^{***}$
	(0,000)
Licit	$-0{,}157^{***}$
	(0,011)
Número de Licitantes	$0{,}091^{***}$
	(0,020)
Precatório	$-0{,}018$
	(0,015)
LicitPreg_ele	$0{,}714^{***}$
	(0,015)
LicitPreg_pres	$0{,}372^{***}$
	(0,013)
LicitTomada	$-0{,}090^{***}$
	(0,011)
n_licit	$-0{,}053^{***}$
	(0,001)
Precatorio	$0{,}029^{***}$
	(0,001)
Constant	-1.901^{***}
	(0,115)
Observations	78,939
R^2	0,102
Adjusted R^2	0,102
Residual Std. Error	0,953 (df = 78924)
F Statistic	642.239^{***} (df = 14; 78924)
Note:	*p<0,1; **p<0,05; ***p<0,01

O baixo valor da estatística R^2 indica que o modelo linear empregado não proporciona uma explicação adequada para o fenômeno estudado, revelando que tal abordagem pode não ser a mais apropriada para a modelagem em questão. Este resultado sugere a existência de limitações significativas devido à linearidade assumida na regressão. Apesar disso, observa-se que os custos de transação respondem de maneira negativa às deteriorações na reputação do ente público, corroborando as previsões teóricas.

Embora o modelo linear tenha sido adotado na pesquisa em referência, o desempenho insatisfatório do R^2 justifica a necessidade de explorar abordagens alternativas. Recomenda-se a implementação de técnicas avançadas de aprendizado de máquina que podem oferecer uma abordagem mais robusta e precisa para a análise dos custos de transação nas contratações públicas.

5.2.1 Modelos de Regressão

O primeiro passo foi dividir os dados em conjuntos de treino e teste, de forma a permitir a avaliação do desempenho dos modelos. Os dados foram divididos em 80% para o treino e 20% para o teste. Esta divisão foi feita de forma estratificada para assegurar que a variável alvo, `custo_tran_norm`, estivesse proporcionalmente representada em ambos os conjuntos.

Neste estudo, foram utilizados três modelos de regressão principais para prever os custos de transação nas contratações públicas: *Random Forest*, *Gradient Boosting* e *LASSO*. Nestes modelos foram ajustados e avaliados utilizando técnicas de validação cruzada e para otimização dos hiperparâmetros.

A modelagem foi conduzida utilizando o framework `tidymodels` do R, que integra diversas bibliotecas para o desenvolvimento de modelos preditivos. Inicialmente, o ambiente foi preparado com a limpeza de todas as variáveis e a carga das bibliotecas necessárias. Em seguida, definiu-se o diretório de trabalho apropriado, onde os dados e scripts relacionados à dissertação estão armazenados.

Após a carga, realizou-se a padronização dos nomes das colunas e a remoção de linhas e colunas vazias utilizando funções da biblioteca `janitor`. Esta etapa de limpeza foi fundamental para garantir a qualidade e a consistência dos dados antes da modelagem.

Tunagem do *LASSO*

Para o modelo *LASSO* utilizado neste trabalho, foi definida uma receita (*recipe*) para o pré-processamento dos dados de treinamento. Esta receita incluiu a transformação de variáveis nominais em variáveis dummy, a imputação de valores ausentes, a normalização de variáveis numéricas e a remoção de variáveis com zero variância

e variância próxima de zero. A imputação dos valores ausentes foi feita utilizando a mediana para variáveis numéricas e a moda para variáveis nominais. A normalização garantiu que todas as variáveis numéricas estivessem na mesma escala, enquanto a conversão de fatores para strings foi necessária para compatibilidade com o modelo.

A seguir, foi configurado um *workflow* que integrava o modelo de regressão *LASSO* com a receita de pré-processamento definida anteriormente. O modelo de regressão *LASSO* foi configurado para utilizar a biblioteca `glmnet`, com a penalização (*penalty*) definida como um hiperparâmetro a ser ajustado. A penalização é um parâmetro crucial no *LASSO*, pois controla a quantidade de regularização aplicada ao modelo, ajudando a evitar o sobreajuste e a selecionar automaticamente as variáveis mais relevantes.

Para ajustar os hiperparâmetros do modelo, foi definida uma grade de valores de penalização (*grid*) com 50 níveis diferentes. Esta abordagem permitiu uma exploração abrangente do espaço de hiperparâmetros, aumentando a chance de encontrar a configuração ótima. Aqui, também foi utilizado o *Grid Search* para otimização dos hiperparâmetros, da forma como foi visto na fundamentação teórica, especificamente o valor da penalização (λ). O *Grid Search* do *LASSO*, variando o penalty (que representa quantidade total de regularização), tem métricas bastante semelhantes para cada um dos folds, começando a piorar ao chegar próximo de 0,01. As métricas para cada combinação de hiperparâmetros em cada fold são mostradas na Figura:

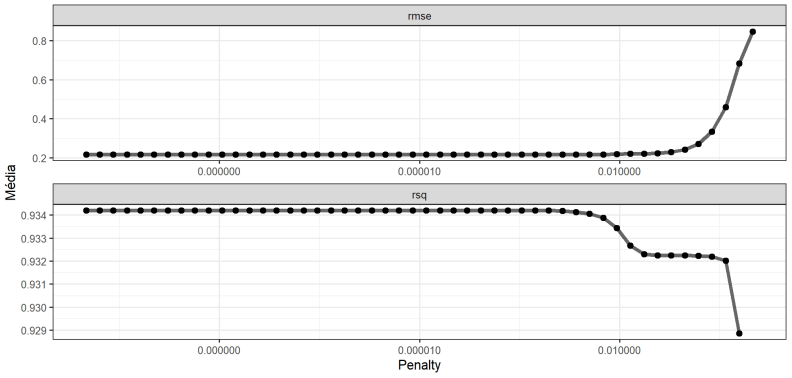

Figura 5.8: Metricas para cada combinacao no *Grid Search* em cada Fold (*LASSO*).

No gráfico superior, observa-se que o *RMSE* permanece praticamente constante e baixo para valores de penalidade até aproximadamente $0,01$. A partir desse ponto, o *RMSE* começa a aumentar rapidamente, indicando que penalidades mais altas degradam a precisão do modelo. Isso sugere que uma penalidade muito grande leva a um ajuste inadequado do modelo, aumentando o erro de previsão.

No gráfico inferior, o *RSQ* (R^2) também se mantém constante e relativamente alto até aproximadamente $0,01$, após o qual começa a diminuir drasticamente. Um *RSQ* mais alto indica que o modelo explica melhor a variabilidade dos dados. A

queda acentuada do *RSQ* com penalidades mais altas sugere que o modelo perde sua capacidade explicativa à medida que a penalidade aumenta, resultando em uma pior performance preditiva.

De modo geral, os resultados indicam que há um valor ótimo de penalidade (*Penalty*) em torno de $0,01$, sendo o *RMSE* minimizado e o *RSQ* maximizado. Este valor de penalidade oferece o melhor equilíbrio entre ajuste do modelo e penalização da complexidade, resultando em uma melhor performance preditiva para o modelo *LASSO*. Na tabela abaixo, pode-se ver o resultado:

Tabela 5.3: Melhores Hiperparâmetros do Modelo Lasso.

penalty	.config
0.0000000001	Preprocessor1_Model01

o resultado apresentado indica que o melhor valor encontrado para o hiperparâmetro penalty do modelo Lasso é 10^{-10}. Este valor é excepcionalmente baixo, sugerindo que a regularização imposta pelo termo de penalidade é quase nula, o que implica em um modelo que praticamente equivale a uma regressão linear ordinária.

A validação do modelo foi realizada utilizando validação cruzada com cinco dobras (*k-fold cross-validation*). Este método envolve a divisão do conjunto de dados de treinamento em cinco subconjuntos, permitindo que o modelo seja treinado em quatro deles e testado no subconjunto restante, repetindo este processo cinco vezes. A validação cruzada fornece uma estimativa robusta da performance do modelo, reduzindo o risco de sobreajuste e garantindo uma avaliação precisa da sua capacidade preditiva.

Tunagem do *Random Forest*

Iniciou-se com o pré-processamento dos dados, utilizando o pacote `recipes`. As etapas incluídas foram:

1. **Codificação de variáveis categóricas**: As variáveis nominais são transformadas em variáveis dummy utilizando `recipes::step_dummy(all_nominal_predictors())`. 2. **Remoção de variáveis com baixa variabilidade**: As variáveis preditoras com pouca variação são removidas com `recipes::step_nzv(all_predictors())`. 3. **Imputação de valores ausentes**: Valores ausentes nas variáveis numéricas são imputados usando a técnica de *k-nearest neighbors* com `recipes::step_impute_knn(all_numeric_predictors(), impute_with = imp_vars(quantidade, valor_total_nota, precatorio, preco))`. 4. **Remoção de variáveis altamente correlacionadas**: As variáveis numéricas com

alta correlação (acima de um limiar de 0.9) são removidas com `step_corr(all_numeric(), -all_outcomes(), threshold = 0.9)`.

Após o pré-processamento, os dados foram preparados e o modelo foi configurado. O modelo é definido com dois hiperparâmetros a serem ajustados: `min_n`, que é o número mínimo de observações em cada nó terminal, e `mtry`, que é o número de variáveis consideradas para cada divisão na árvore.

A configuração do modelo, nesta pesquisa, utilizou o pacote `ranger` com a importância das variáveis medida pela impureza (`importance = "impurity"`) e paralelização ativada (`num.threads = parallel::detectCores()`).

Em seguida, um *workflow* foi então criado, integrando o modelo configurado e a fórmula do modelo que especifica a variável alvo `custo_tran_norm` em função de todas as variáveis preditoras. Este *workflow* facilita a aplicação de todo o processo de modelagem de forma coesa e estruturada.

Para validar o desempenho do modelo, foi utilizada validação cruzada com cinco dobras, estratificada pela variável alvo. Este método garante que o modelo seja treinado e testado em diferentes subconjuntos dos dados de treinamento, proporcionando uma avaliação robusta da performance.

Finalmente, a tunagem dos hiperparâmetros foi realizada através de uma busca em grid, explorando diferentes combinações de valores de `min_n` e `mtry`. Foram utilizadas 20 diferentes combinações e as métricas de desempenho, como R^2 e RMSE, foram calculadas para cada combinação.. Os resultados da tunagem são apresentados na Figura:

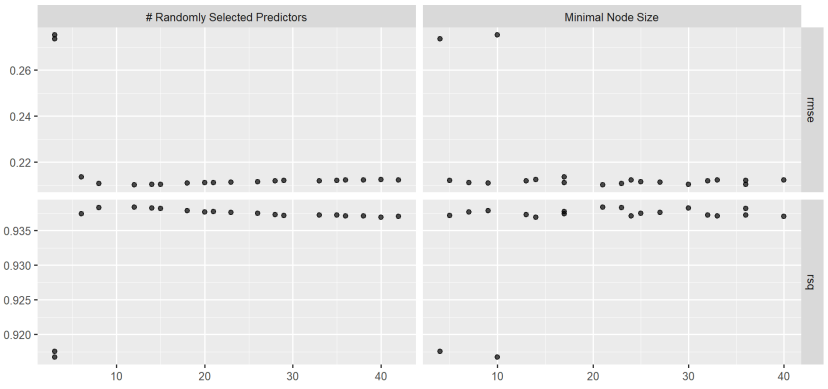

Figura 5.9: Tunagem do modelo *Random Forest*.

A figura acima apresenta os resultados da tunagem do modelo de *Random Forest*, mostrando o desempenho do modelo em termos das métricas RMSE (Root Mean Squared Error) e R^2 para diferentes combinações de hiperparâmetros. Especificamente, são exibidos os valores de RMSE e R^2 em função do número de preditores selecionados aleatoriamente (*Randomly Selected Predictors*) e do tamanho mínimo

do nó (*Minimal Node Size*).

Os gráficos permitem observar como a variação destes hiperparâmetros afeta a performance do modelo. A análise dos gráficos revela que, para diferentes números de preditores selecionados aleatoriamente, a variação do tamanho mínimo do nó tende a apresentar um comportamento consistente na métrica R^2, com ligeiras flutuações. No entanto, observa-se que há uma concentração de pontos com alta performance (alto R^2 e baixo RMSE) em determinadas combinações de hiperparâmetros, sugerindo que a escolha adequada destes parâmetros é fundamental para a otimização do modelo. Os hiperparâmetros obtidos após o treinamento e ajuste do modelo Random Forest foram:

Hiperparâmetro	**Valor**
`mtry`	12
`min_n`	21

Tabela 5.4: Valores dos hiperparâmetros do modelo Random Forest.

Relembrando que o parâmetro `mtry` determina o número de variáveis a serem consideradas em cada divisão de nó, sendo encontrado o ótimo como 12 variáveis. Isso significa que, o modelo avaliou 12 variáveis aleatórias para determinar a melhor divisão. Já o parâmetro `min_n` define o número mínimo de amostras em um nó terminal, tendo sido encontrado 21 como o ótimo. Esse valor ajuda a controlar a profundidade das árvores, prevenindo o sobreajuste e garantindo que cada nó terminal tenha um número adequado de observações para fazer previsões robustas.

É evidente que a tunagem dos hiperparâmetros possibilitou identificar configurações que maximizam a capacidade preditiva do modelo de *Random Forest*, resultando em previsões mais precisas dos custos de transação em contratações públicas.

Tunagem do *Gradient Boosting*

Na regressão que foi rodada usando essa técnica, utlizou-se a mesma receita usada no *Random Forest*, discutida acima. Os hiperparâmetros do modelo, tais como o número de árvores, profundidade das árvores, tamanho mínimo de amostras em cada nó, redução de perda, tamanho da amostra e taxa de aprendizado, foram configurados para serem ajustados (*tuning*) durante o processo de validação.

Para garantir a robustez e a generalização do modelo, implementou-se um esquema de validação cruzada com k-folds, estratificada. Esse procedimento permitiu avaliar o desempenho do modelo em diferentes subconjuntos dos dados, assegurando que o modelo final não estivesse superajustado aos dados de treinamento.

A etapa de tunagem dos hiperparâmetros foi conduzida utilizando uma grade de busca (*grid search*) com 20 combinações diferentes de hiperparâmetros. Os modelos

resultantes foram avaliados utilizando métricas de desempenho como o coeficiente de determinação (R^2) e o erro quadrático médio (RMSE).

Como foi visto acima, A tunagem de hiperparâmetros em *Gradient Boosting* é um processo essencial para otimizar o desempenho do modelo. A imagem abaixo ilustra como diferentes configurações de hiperparâmetros afetam duas métricas de erro: Erro quadrático médio (RMSE) e R^2. Este processo envolve ajustar parâmetros como o número de árvores, a taxa de aprendizado, o tamanho mínimo dos nós, a redução mínima da perda e a profundidade das árvores:

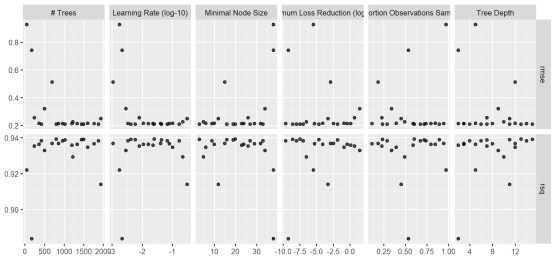

Figura 5.10: Tunagem do *Gradient Boosting*.

Primeiramente, o número de árvores (*# Trees*) influencia diretamente a capacidade do modelo de capturar padrões nos dados. No gráfico, observa-se que para um pequeno número de árvores, os valores de RMSE são altos, indicando um modelo subajustado. À medida que o número de árvores aumenta, o RMSE diminui até se estabilizar, sugerindo que adicionar mais árvores além desse ponto não melhora significativamente o desempenho e pode levar ao *overfitting*.

A taxa de aprendizado (*Learning Rate, log-10*) controla a contribuição de cada árvore adicionada ao modelo. No gráfico acima, uma taxa de aprendizado muito alta (valores próximos de zero no eixo log-10) está associada a maiores valores de RMSE, enquanto uma taxa de aprendizado muito baixa (valores muito negativos no eixo log-10) também pode não ser ideal devido à lenta convergência do modelo. O ponto ideal está em um balanço em que o RMSE é minimizado.

O tamanho mínimo dos nós (*Minimal Node Size*) determina o número mínimo de observações que um nó deve ter para ser dividido. Observa-se no gráfico que tamanhos mínimos muito pequenos ou muito grandes resultam em um aumento do RMSE. Portanto, existe um tamanho ótimo que minimiza o erro de predição.

A redução mínima da perda (*Minimum Loss Reduction, log-10*) especifica a redução mínima na função de perda necessária para realizar uma divisão. Valores muito baixos (no eixo log-10, valores mais negativos) resultam em um modelo mais complexo com menor RMSE, mas podem levar ao *overfitting*. Valores mais altos simplificam o modelo, aumentando o RMSE.

A proporção de observações amostrais (*Proportion Observations Sampled*) de-

fine a fração dos dados de treinamento utilizada em cada árvore. No gráfico, a variabilidade do RMSE é maior para proporções extremas (muito baixas ou muito altas), indicando que um valor intermediário é mais adequado para balancear viés e variância.

A profundidade das árvores (*Tree Depth*) controla o nível máximo de divisões em cada árvore. Conforme mostrado no gráfico, uma maior profundidade inicialmente reduz o RMSE, mas após um certo ponto, a profundidade adicional leva ao aumento do RMSE provavelmente devido a um *overfitting*.

Lembrando que a escolha ótima dos hiperparâmetros minimiza a função de perda, resultando em um modelo que generaliza bem para novos dados, foram encontrados os seguintes hiperparâmetros:

Tabela 5.5: Melhores Hiperparâmetros Encontrados.

trees	min_n	tree_depth	learn_rate	loss_reduction	sample_size	.config
798	20	13	0.025118	0	0.6042544	Preprocessor1_Model10

Assim como nas técnicas anteriores, foi utilizada uma validação cruzada com cinco dobras (*k-fold cross-validation*), para aumentar a robustez do modelo.

5.2.2 Conclusão dos Modelos

Os modelos *Random Forest*, *Gradient Boosting* e *LASSO* fornecem abordagens complementares para a previsão dos custos de transação nas contratações públicas. Enquanto o *Random Forest* e o *Gradient Boosting* são poderosos na captura de relações complexas entre as variáveis, o *LASSO* ajuda na seleção de variáveis importantes, simplificando o modelo e melhorando a interpretabilidade.

Os resultados do modelo *LASSO* foram visualizados através do gráfico de valores previstos versus valores observados, conforme mostrado na Figura:

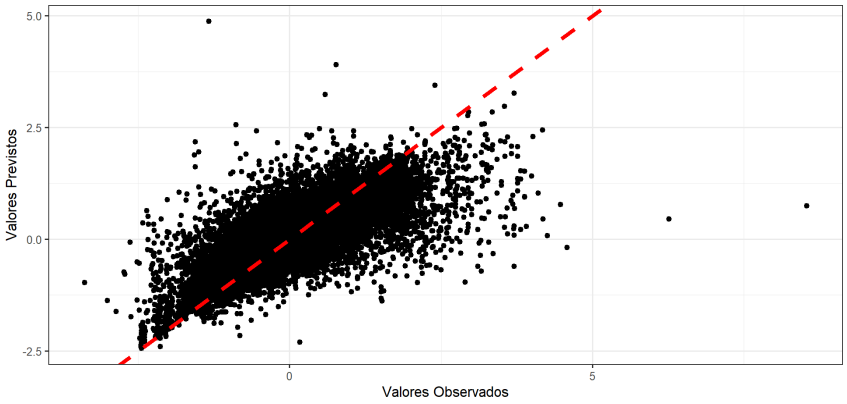

Figura 5.11: Valores previstos x observados (*LASSO*).

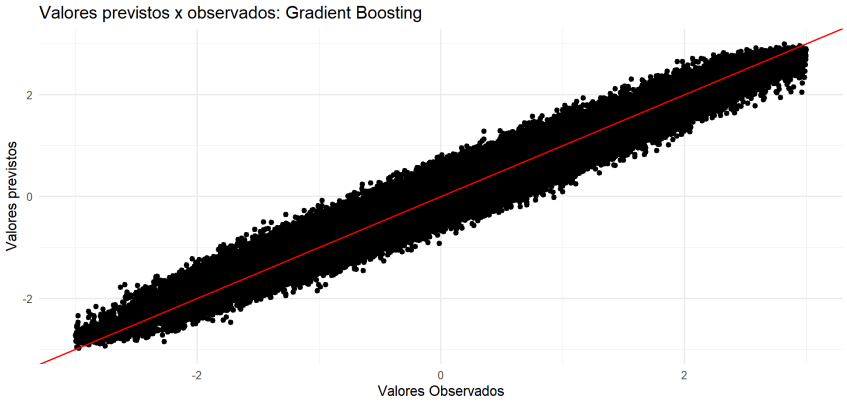

Figura 5.12: Valores previstos x observados (*Gradient Boosting*).

Os resultados do modelo *Gradient Boosting* foram apresentados no gráfico que compara os valores previstos com os valores observados, conforme ilustrado na Figura:

Por sua vez, os resultados do modelo *Random Forest* foram visualizados através do gráfico de valores previstos versus valores observados, conforme mostrado na Figura:

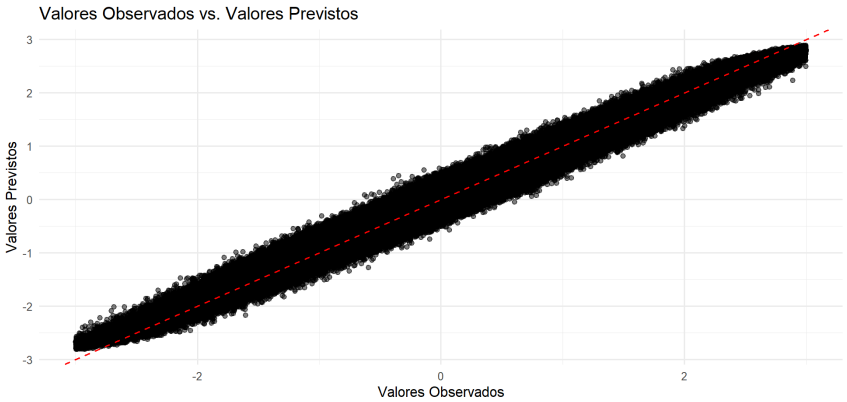

Figura 5.13: Valores previstos x observados (*Random Forest*).

Ao comparar os resultados dos três modelos de regressão – *LASSO*, *Gradient Boosting* e *Random Forest* – para prever os custos de transação de licitações públicas, foram observadas diferenças significativas em seus desempenhos. Vale ressaltar que a variável dependente (custos de transação) foi normalizada, representando a quantidade de desvios padrão em torno do preço médio dos produtos comprados.

No caso do *LASSO*, o gráfico de dispersão mostra os valores previstos pelo modelo em relação aos valores observados, sendo que a linha vermelha representa a reta ideal de 45 graus, local em que os valores previstos seriam iguais aos valores observados. No entanto, a dispersão dos pontos ao redor dessa linha é maior, indicando que o modelo *LASSO* possui erros de previsão mais elevados. Por outro lado,

o modelo *Gradient Boosting* e *Random Forest* apresentam um desempenho parecido e significativamente superior ao primeiro, pois a dispersão dos pontos ao redor da reta é menor em comparação com o modelo *LASSO*, o que sugere que estes últimos têm uma melhor capacidade de previsão e captura melhor a variabilidade dos dados.

Como graficamente os modelos *Gradient Boosting* e *Random Forest* apresentam resultados parecidos, é importante observar seus resultados de previsão na tabela abaixo:

Tabela 5.6: Resultados do modelo *Random Forest*, *Gradient Boosting* e *LASSO*.

Métrica	**Estimativa RF**	**Estimativa GB**	**Estimativa LASSO**
RMSE	0,13964	0,20876	0,973840
R^2	0,97304	0,93895	0,63468
MAE	0,10975	0,16424	0,790586

Deste modo, pode-se concluir que o modelo *Random Forest* é mais adequado para prever os custos de transação de licitações públicas, pois apresenta uma menor erro quadrático médio (RMSE), assim como um melhor desempenho nas outras métricas (Erro médio absoluto e coeficiente de determinação R^2, em comparação com os outros dois modelos *Gradient Boosting* e *LASSO*.

5.2.3 Importância das variáveis

A análise da importância das variáveis é essencial para entender quais fatores influenciam mais os custos de transação nas contratações públicas. Nesta seção, será discutida a importância das variáveis no modelo *Random Forest*, uma vez que este modelo fornece uma avaliação direta da relevância de cada variável no processo de previsão.

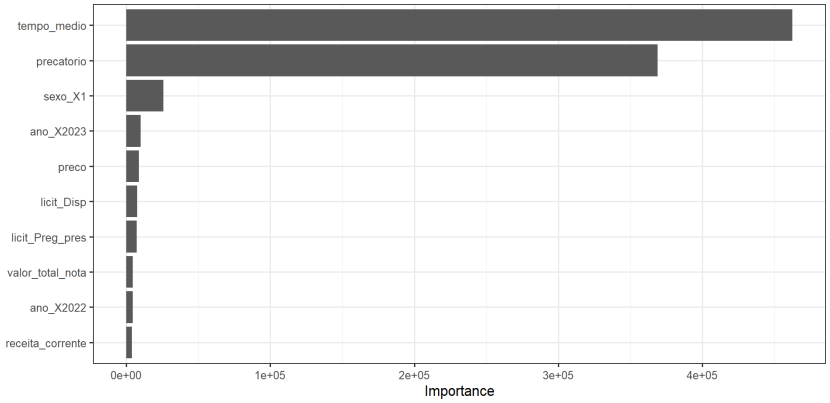

Figura 5.14: Importância das variáveis para o modelo *Random Forest*.

A análise da importância das variáveis para o modelo *Random Forest*, conforme

apresentada pela figura acima, com base na sua contribuição para a redução do erro ao longo das árvores de decisão, revela quais fatores têm maior influência na previsão dos custos de transação de licitações públicas. Cada uma dessas variáveis tem um papel distinto no desempenho do modelo, e a importância delas está refletida na capacidade de cada variável de reduzir o erro de previsão quando incluída no modelo.

Observa-se que a variável *tempo_medio* é a mais importante no modelo, indicando que o tempo médio para os pagamentos é um fator crucial na determinação dos custos de transação. Este resultado sugere que reduzir o tempo necessário para concluir as transações pode ser uma estratégia eficaz para minimizar os custos associados. A segunda variável mais importante é *precatorio*. A presença de precatórios, que são dívidas judiciais que o governo é obrigado a pagar, mostra uma forte influência nos custos de transação. Isso pode estar relacionado à complexidade e ao tempo adicional necessário para resolver estas situações.

Variáveis relacionadas ao tempo, como *ano_X2023* e *ano_X2022*, também são relevantes, sugerindo que as mudanças temporais influenciam os custos de transação. Isso pode incluir variações sazonais ou tendências de longo prazo que afetam a eficiência e os custos das contratações. O tipo de licitação, como *licit_Disp* e *licit_Preg_pres*, também apresentam importância considerável. Esses fatores diretamente associados às características das licitações e dos contratos impactam os custos de transação, refletindo as complexidades e especificidades de diferentes processos licitatórios. Ainda pode-se destacar variáveis como *valor_total_nota* e *receita_corrente* demonstram alguma importância, indicando que os valores financeiros envolvidos nas transações e a situação fiscal do órgão público também são fatores que afetam os custos de transação.

5.3 Limitações dos Métodos e Soluções Propostas

Embora a aplicação dos modelos de regressão *Random Forest*, *Gradient Boosting* e *LASSO* tenha mostrado resultados promissores na previsão dos custos de transação em contratações públicas, é essencial reconhecer algumas limitações inerentes aos métodos utilizados e aos dados empregados na pesquisa.

Uma das principais limitações está relacionada à qualidade e consistência dos dados. A organização dos dados provenientes de diversas fontes, como notas fiscais, empenhos e informações socioeconômicas, apresentou desafios significativos. Apesar dos esforços para padronizar e limpar os dados, ainda podem existir inconsistências que impactam a precisão dos modelos. A fusão de dataframes de notas fiscais com informações de produtos e a exclusão de registros com dados incompletos foram passos necessários, mas que também reduziram o tamanho da amostra disponível para análise.

Outro ponto crítico é a seleção e representatividade das variáveis independentes. Embora tenham sido incluídas variáveis significativas como o tempo médio de pagamento, precatórios, e indicadores socioeconômicos, outras variáveis potencialmente relevantes podem não ter sido consideradas devido à indisponibilidade de dados ou limitações na fase de coleta e pré-processamento. Adicionalmente, a baixa correlação entre algumas variáveis independentes e a variável dependente sugere que pode haver fatores importantes não capturados pelos modelos utilizados.

A própria natureza dos modelos de aprendizado de máquina, especialmente o *Random Forest* e o *Gradient Boosting*, apresentam desafios de interpretabilidade. Embora *Random Forest* e o *Gradient Boosting* sejam eficazes em capturar relações não-lineares complexas, a interpretação dos resultados e a compreensão do impacto individual de cada variável podem ser difíceis. Isso pode limitar a utilidade prática dos resultados para formuladores de políticas que necessitam de explicações claras sobre os fatores que influenciam os custos de transação.

Além disso, as técnicas de tunagem de hiperparâmetros, embora necessárias para otimizar a performance dos modelos, introduzem um nível adicional de complexidade e demandam poder computacional significativo. A validação cruzada com múltiplas dobras e a busca em grade aumentam a robustez dos modelos, mas também aumentam o tempo e os recursos necessários para a análise. A título de ilustração, para rodar o modelo *Random Forest* foram necessárias mais de 18 horas initerruptas em um computador com processador Intel Core i7-13700K com 64 GB DDR5 4400MT/s.

5.4 Conclusão do aprendizado de máquina

A análise dos resultados dos modelos de regressão *LASSO*, *Gradient Boosting* e *Random Forest* para prever os custos de transação de licitações públicas, considerando a variável dependente normalizada, revela diferenças substanciais em seus desempenhos. Especificamente, o modelo *Random Forest* demonstrou uma maior precisão.

O desempenho dos modelos foram avaliados utilizando as métricas de erro quadrático médio ($RMSE$), coeficiente de determinação (R^2) e erro absoluto médio (MAE), como demonstrado anteriormente, na conclusão dos modelos.

Esses resultados indicam que o modelo *Random Forest* possui um bom ajuste aos dados, com um R^2 de aproximadamente 0,97, sugerindo que cerca de 97% da variabilidade nos custos de transação é explicada pelo modelo. O valor de RMSE de 0,14 indica que o erro médio das previsões é de cerca de meio desvio padrão dos preços normalizados. Além disso, o MAE de 0,11 reflete que, em média, os erros absolutos das previsões são relativamente baixos, reforçando a precisão do modelo.

O Erro Quadrático Médio de 0,1396495 significa que, em média, a diferença entre os valores previstos pelo modelo e os valores observados é de aproximadamente 0,14 desvios padrão dos preços normalizados. Este valor fornece uma medida agregada da precisão das previsões do modelo, penalizando mais grandes erros.

O Erro Absoluto Médio de 0,1097531 indica que, em média, os erros absolutos das previsões são de cerca de 0,11 desvios padrão dos preços normalizados. Diferente do RMSE, o MAE não penaliza tanto os grandes erros, oferecendo uma medida mais direta e interpretável da precisão média das previsões. Um MAE mais baixo sugere que as previsões do modelo estão, em média, mais próximas dos valores reais.

Assim, o modelo *Random Forest* oferece uma maior precisão e um melhor ajuste aos dados. As métricas de desempenho obtidas reforçam que é um modelo robusto e confiável para esta aplicação, sendo capaz de fornecer previsões precisas e úteis para a análise de custos de transação no contexto das licitações públicas paraibanas.

Apesar das limitações identificadas, como a qualidade dos dados e a complexidade dos modelos, a pesquisa mostrou que é possível utilizar técnicas avançadas de aprendizado de máquina para prever custos de transação, alinhando-se com o objetivo geral de desenvolver e aplicar um modelo matemático para analisar as contratações públicas.

Capítulo 6

Conclusão

Esta dissertação teve como objetivo principal desenvolver e aplicar um modelo matemático para analisar as contratações públicas no estado da Paraíba e, em seguida, utilizar técnicas de aprendizado de máquina para prever os custos de transação que impactam os preços públicos. A partir de uma abordagem interdisciplinar que combinou teoria dos leilões, custos de transação e aprendizado de máquina, foi possível fornecer uma análise detalhada das contratações públicas, bem como desenvolver ferramentas práticas para aprimorar esses processos.

Além da parte empírica, merece destaque a inversão do modelo de leilão de venda para um modelo de leilão de compra, sendo uma contribuição significativa para a academia especializada, uma vez que os modelos existentes na literatura são focados em leilões de venda. Essa abordagem permite uma análise detalhada e prática dos custos de transação nas contratações públicas, proporcionando uma nova perspectiva sobre a formação de preços. Ao aplicar a teoria dos jogos bayesianos e o conceito de equilíbrio de Nash Bayesiano, o estudo oferece uma compreensão aprofundada dos impactos desses custos, especialmente em relação a variáveis como preços de reserva e taxas de entrada. Essa abordagem teórica não só enriquece a literatura acadêmica, mas também tem implicações práticas, ajudando a otimizar processos licitatórios.

6.1 Resumo dos Resultados

A análise dos dados revelou que tanto os modelos de regressão *Random Forest* quanto *Gradient Boosting* são mais eficazes que o *LASSO* na previsão dos custos de transação em contratações públicas. O modelo *Random Forest*, em particular, apresentou um coeficiente de determinação (R^2) de 0,97, explicando cerca de 97% da variabilidade nos custos de transação, e um erro quadrático médio (RMSE) de 0,14 desvios padrão dos preços normalizados, sugerindo uma boa precisão na previsão dos custos de transação. O erro absoluto médio (MAE) de 0,11 desvios padrão reflete que os erros absolutos das previsões são relativamente baixos, reforçando a precisão do

modelo.

A importância das variáveis no modelo *Random Forest* mostrou que o tempo para pagamentos e o tempo para adimplemento do precatório são os fatores mais influente na previsão dos custos de transação.

6.2 Implicações Teóricas e Práticas

Esta pesquisa contribui significativamente para a literatura existente ao fornecer uma análise das contratações públicas na Paraíba e ao desenvolver ferramentas que podem ser utilizadas para aprimorar esses processos, reduzindo os custos de transação associados às contratações. Os insights obtidos podem orientar políticas e estratégias para otimizar os custos de transação nas contratações públicas, focando nos fatores mais determinantes identificados pelo modelo .

A aplicação de técnicas de aprendizado de máquina, como *Random Forest*, *Gradient Boosting* e *LASSO*, mostrou-se eficaz na previsão de custos de transação, oferecendo uma abordagem robusta e confiável para a análise e melhoria dos processos de licitação pública. A combinação de um modelo teórico sólido com métodos avançados de análise de dados pode levar a uma gestão mais eficiente das contratações públicas, promovendo uma melhor alocação de recursos e a redução de ineficiências no setor público .

6.3 Limitações do Estudo

Apesar dos resultados promissores, a pesquisa apresentou algumas limitações. A qualidade e consistência dos dados foram desafios significativos. A organização dos dados provenientes de diversas fontes, como notas fiscais, empenhos e informações socioeconômicas, apresentou dificuldades, mesmo com esforços para padronizar e limpar os dados. A exclusão de registros com dados incompletos reduziu o tamanho da amostra disponível para análise .

A seleção e representatividade das variáveis independentes também foram limitadas pela disponibilidade de dados. Embora variáveis significativas tenham sido incluídas, outras potencialmente relevantes podem não ter sido consideradas devido à indisponibilidade de dados ou limitações na fase de coleta e pré-processamento .

Ademais, a natureza dos modelos de aprendizado de máquina, especialmente o Random Forest, apresenta desafios de interpretabilidade. A complexidade dos modelos pode dificultar a compreensão do impacto individual de cada variável, limitando a utilidade prática dos resultados para formuladores de políticas que necessitam de explicações claras sobre os fatores que influenciam os custos de transação .

6.4 Sugestões para Pesquisas Futuras

Para futuras pesquisas, sugere-se a ampliação do escopo de dados para melhorar significativamente os modelos preditivos. A inclusão de dados mais detalhados sobre itens específicos adquiridos, bem como a incorporação de outras variáveis econômicas e contextuais, poderia fornecer uma visão mais completa dos fatores que influenciam os custos de transação. Técnicas avançadas de imputação de dados podem ajudar a lidar com problemas de dados ausentes sem a necessidade de exclusão de registros importantes .

A pesquisa também pode se beneficiar da aplicação de métodos de aprendizado profundo (deep learning), especialmente se grandes volumes de dados se tornarem disponíveis. Redes neurais profundas têm o potencial de capturar padrões ainda mais complexos nos dados, embora também apresentem desafios de interpretabilidade e demanda por recursos computacionais .

Finalmente, investigações futuras poderiam focar em estudos longitudinais para entender como os custos de transação evoluem ao longo do tempo e em diferentes contextos econômicos. Analisar políticas específicas de gestão pública ou mudanças regulatórias e seus impactos nos custos de transação pode fornecer insights valiosos para a formulação de políticas públicas mais eficientes e transparentes .

Em conclusão, modelos matemáticos de leilões e técnicas de aprendizado de máquina podem prever os custos de transação nas contratações públicas no estado da Paraíba, oferecendo uma ferramenta poderosa para a promoção de uma alocação mais eficiente dos recursos públicos.

Capítulo 7

Referências Bibliográficas

AHN, S.; CHOI, W. The role of bank monitoring in corporate governance: Evidence from borrowers earnings management behavior. Journal of Banking & Finance, 33(2), 425-434. 2009

AKAIKE, H. A new look at the statistical model identification, IEEE Transactions on Automatic Control, 19 (6): 716–723, 1974

ANGINER, Deniz; MANSI, Sattar; WARBURTON, A. Joseph e YILDIZHAN, Çelim, Firm Reputation and the Cost of Debt Capital (June 2, 2015). Available at SSRN: https://ssrn.com/abstract=2024181

AROZAMENA, Leandro; CANTILLON, Estelle. Investment Incentives in Procurement Auctions. Review of Economic Studies 71, 2004.

ASH, Elliott; GALLETTA, Sergio; GIOMMONI, Tommaso. A Machine Learning Approach to Analyze and Support Anti-Corruption Policy, CESifo Working Paper, No. 9015, Center for Economic Studies and Ifo Institute (CESifo), Munich, 2021.

ATHEY, Susan; LEVIN, Jonathan; SEIRA, Enrique. Comparing open and Sealed Bid Auctions: Evidence from Timber Auctions, The Quarterly Journal of Economics, Oxford University Press, vol. 126(1), 2011.

BAGHAI, Ramin P. ; BECKER, Bo. Reputations and credit ratings: Evidence from commercial mortgage-backed securities, Journal of Financial Economics, Volume 135, Issue 2, 2020.

BASDEO, D.K., SMITH, K.G., GRIMM, C.M., RINDOVA, V.P. and DERFUS, P.J. 'The impact of market actions on firm reputation', Strategic Management Journal, 27, 1205–1219. 2006

BEAVER, W. H. Financial ratios as predictors of failure. Journal of Accounting Research, 4, 71-111. 1966

BREIMAN, Leo. "Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author)."Statist. Sci. 16 (3) 199 - 231, 2001.

BUTLER, A. W.; FAUVER, L. Institutional environment and sovereign credit

ratings. Financial Management, 35(3), 53-79. 2006.

BULOW, Jeremy; KLEMPERER, Paul. Why do Sellers (Usually) Prefer Auctions? American Economic Review, 99 (4): 2009.

CABAL, O.; FERREIRA, Luis.; DIAS, G. P. Adoption of reverse auctions In: public e-procurement: the case of Portugal, 11a. CRISTI, 2016.

CAMELO, Bradson; TORRES, Ronny Charles e NOBREGA, Marcos. Análise Econômica das Licitacoes e Contratos. Ed. Forum, Belo Horizonte. 2022.

CASSIDY, R. Auctions and Auctioneering. University of California Press, 1967.

CHWE, M. S.-Y.. The discrete bid first auction. Economics Letters, 31:303–306, 1989.

CHE, Yeon-Koo; GALE, Ian. Optimal Design of Research Contest. American Economic Review 93(3), 2003.

CHU, W. ; CHU, W. 'Signaling quality by selling through a reputable retailer: An example of renting the reputation of another agent', Marketing Science, 13(2), 177–189. 1994.

COOTER, Robert; ULEN, Thomas. Direito & Economia. Porto Alegre: Bookman, 2010.

DIAS, M. T. F. Análise Econômica do Direito nas contratações públicas: estudo de casos da legislação e experiência brasileira. In: II Congresso Internacional de Compras Públicas: Para um Crescimento da Economia Assente na Contratação Pública Sustentável, Inteligente e Inovadora. 2016

DAVID, Esther; ROGERS, Alex; JENNINGS, Nicholas R.; SCHIFF, Jeremy; KRAUS, Sarit; ROTHKOPF, Michael H. Optimal design of english auctions with discrete bid levels. ACM Trans. Internet Technol. 7, 2. May 2007.

EATON, B. C.; EATON, D. F. Microeconomia. São Paulo: Saraiva, 1999.

FIRKE S. janitor: Simple Tools for Examining and Cleaning Dirty Data. R package version 2.2.0, <https://CRAN.R-project.org/package=janitor>. 2023

FRICK H, CHOW F, KUHN M, MAHONEY M, SILGE J, WICKHAM H (2024). rsample: General Resampling Infrastructure. R package version 1.2.1, <https://CRAN.R-project.org/package=rsample>.

FRIEDMAN, Daniel. "On the efficiency of experimental double auction markets."The American Economic Review 74, no. 1, 1984.

FULLERTON, Richard L.; MCAFEE, R. Preston. Auctioning Entry into Tournaments. Journal of Political Economy 107(3), 1999.

GÓIS, A. D.; DE LUCA , M. M. M.; DE LIMA , G. A. S. F.; MEDEIROS, J. T. Corporate reputation and bankruptcy risk. Brazilian Administration Review, v. 17, n. 2, 2020.

GOTSI, M.; WILSON, A. M. Corporate reputation: Seeking a definition. Corporate Communications: An International Journal, 6(1), 24-30. 2001

GREENWELL, Brandon M.; BOEHMKE, Bradley C. Variable Importance Plots—An Introduction to the vip Package. The R Journal, 12(1), 343–366. URL https://doi.org/10.32614/RJ-2020-013. 2020

GUARNIERI, P.; GOMES, R. C. Can public procurement be strategic? A future agenda proposition. Journal of Public Procurement, 19(4), 2019.

HASTIE, T.; TIBSHIRANI,, R.; FRIEDMAN, J. The elements of statistical learning: data mining, inference, and prediction. Springer, 2009.

HENDRICKS, K.; PAARSCH, H. J., "A survey of recent empirical work concerning auctions," The Canadian Journal of Economics, 28, 2, 1995

IZBICKI, R.; SANTOS, T. M. dos. Aprendizado de máquina: uma abordagem estatística. 1^a edição. 2020. 272 páginas. ISBN: 978-65-00-02410-4

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. An introduction to statistical learning: with applications in R. Springer, 2017.

JOLIVET, Gregory; JULLIEN, Bruno; POSTEL-VINAY, Fabien. Reputation and Pricing on the e-Market: Evidence from a Major French Platform. International Journal of Industrial Organization, Volume 45, 2016.

KLEMPERER, P., "Auction Theory: A Guide to the Literature," Journal of Economic Surveys, 13, 3, 1999.

KLEMPERER, P. Auctions with almost common values: The 'Wallet Game' and its applications, European Economic Review, 42, issue 3-5, 1998.

KLEMPERER, P. "What Really Matters in Auction Design". Oxford University - England, agosto 2001.

KREPS, D. M. A Course in Microeconomic Theory. Cambridge: Princeton University Press, 1990.

KREPS, D. M. Microeconomics for managers. New York: W.W. Norton, 2004.

KUHN M (2024). tune: Tidy Tuning Tools . R package version 1.2.1, <https://CRAN.R-project.org/package=tune>.

KUHN M, VAUGHAN D, HVITFELDT E (2024). yardstick: Tidy Characterizations of Model Performance. R package version 1.3.1, <https://CRAN.R-project.org/package=yardstick>.

KUHN M, VAUGHAN D (2024). parsnip: A Common API to Modeling and Analysis Functions. R package version 1.2.1, <https://CRAN.R-project.org/package=parsnip>.

KUHN et al., (2020). Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles. https://www.tidymodels.org

KUHN M, WICKHAM H, HVITFELDT E (2024). recipes: Preprocessing and Feature Engineering Steps for Modeling. R package version 1.0.10, <https://CRAN.R-project.org/package=recipes>.

LIBÓRIO, M. P.; BERNARDES, P.; EKEL, P. I.; GICO JÚNIOR, I.T. A abor-

dagem da Análise Econômica do Direito em contratações públicas: uma revisão sistemática da literatura do Brasil. Economic Analysis of Law Review, 12(2), 2021.

LIBÓRIO, M. P., BERNARDES, P., REZENDE, S. F. L., EKEL, P. Y., ARAÚJO, E. C. A. d., GOMES, L. F. A. M., e GICO JÚNIOR, I. T. Efeito da reputação do comprador público nos preços de aquisição de produtos comuns em pregões eletrônicos Economic Analysis of Law Review, 13(2), 2022.

LIN, Chin-Shien; CHOU, Shihyu; WENG, Shih-Min; HSIEH, Yu-Chen. A final price prediction model for english auctions: a neuro-fuzzy approach. Qual Quant 47, 2013.

LOPOMO, Giuseppe. The English Auction Is Optimal Among Simple Sequential Auctions. journal of economic theory 82, 1998.

LUCE, R. D.; RAIFFA, H. Games and decisions: Introduction and critical survey. Wiley. 1957.

MCAFEE, R. P.; MCMILLAN, J., "Auctions and Bidding", Journal of Economic Literature, 25, 1987.

MARINHO, Pedro Rafael Diniz. Machine Learning / Aprendizagem de Máquina. 2023. Notas de aula. Não paginado.

MAS-COLELL, A.; WHINSTON, M. D.; GREEN, J. R. Microeconomic Theory. New York: Oxford University Press, 1995.

MASKIN, Eric; RILEY, John. Asymmetric Auctions, Review of Economic Studies, 67, 2000.

MICROSOFT Corporation, WESTON S (2022). doParallel: Foreach Parallel Adaptor for the 'parallel' Package. R package version 1.0.17, <https://CRAN.R-project.org/package=doParallel>.

MILGROM, P.; WEBER, R. A theory of auctions and competitive bidding, Econometrica 50, 1982.

Murphy, Kevin P. Probabilistic Machine Learning: An introduction", Cambridge, MA : MIT Press, 2022

MYERSON, Roger. Optimal auction design, Mathematical Operational Research. 6 1981.

MYERSON, Roger. Game Theory: Analysis of Conflict. Cambridge: Harvard University Press, 1991

NADARAYA, E. A. On estimating regression. Theory of Probability Its Applica- tions, 9(1), 1964

NETER, J., WASSERMAN, W. e KUTNER, M. H. Applied Linear Statistical Models. 4 ed. Chicago: IRWIN, 1996

NIEBUHR, P. M.; OLIVEIRA, C. L. O. "Custo da Administração" em Contratos Administrativos: uma análise da repercussão econômica das cláusulas exorbitantes e o abuso da posição da administração pública enquanto contratante. Justiça do

Direito, 32(3), 2018.

NÓBREGA, M. Análise Econômica do Direito Administrativo. In: Timm, L. B. (Org.). Direito e Economia no Brasil. 1ª. ed. São Paulo: Atlas, 2012.

NORTH, D. C. Institutions, institutional change and economic performance. Cambridge university press, 1990.

OECD, OECD Integrity Review of Brazil: Managing Risks for a Cleaner Public Service, OECD Public Governance Reviews, OECD Publishing. 2011

OLUSEGUN, S. Public Contract Execution: A Critical Review of Performance Hindering Factors Among Small and Medium Enterprises (SMEs) In Lagos State, Nigeria. Osogbo Journal of Management (OJM), 3(2), 2018

PEREIRA, M. M. F.; NÓBREGA, M. A. R. O diálogo entre público e privado na perspectiva do contrato incompleto. Revista da AMDE, 2012

PICCIONE, Michele; TAN, Guofu. Cost-Reducing Investment, Optimal Procurement and Implementation by Auctions. International Economic Review 37(3), 1996.

PINDICK, Robert S.; RUBINFELD, Daniel L. Microeconomia. 6. ed. São Paulo: Pearson Prentice Hall, 2005.

RILEY, J.; SAMUELSON, W. Optimal auctions, American Economic Review. 71, 1981.

ROTHKOPF, M. H.; HARSTAD, R. On the role of discrete bid levels in oral auctions. European Journal of Operations Research, 74:572–581, 1994.

SCHAPIRE, R. E.; STONE, P.; MCALLESTER, D.; LITTMAN, M. L.; CSIRIK, J. A., "Modeling auction price uncertainty using boosting-based conditional density estimation," In Nineteenth International Conference on Machine Learning, Sydney, 2002.

SEARLE, S. R. Linear Models. New York: Wiley, 1997.

SEBER, G. A. F; LEE, A. J. Linear Regression Analysis. New Jersey: John Wiley Sons.2012.

SPAGNOLO, G.; CASTELLANI, L. Introduction–Vendor rating, performance and entry in public procurement. In Law and Economics of Public Procurement Reforms (pp. 1-11). Routledge. 2017.

STICKEL, S.E. 'Reputation and performance among security analysts', Journal of Finance, 47(5), 1811–1837. 1992

STRAND, I.; RAMADA, P.; CANTON, E.; MULLER, P.; DEVNANI, S.; BAS, P. D.; DVERGSDAL, K. Public procurement in Europe. Cost and effectiveness. Bruss. PwC Lond. Econ. Ecorys. 2011

SCHWARZ, Gideon E. Estimating the dimension of a model, Annals of Statistics, 6 (2): 461–464, 1978

TAN, Guofu. “Entry and R & D in Procurement Contracting.” Journal of Economic Theory 58(1), 1992

TAYLOR, Curtis R. Digging for Golden Carrots: An Analysis of Research Tournaments. Amer- ican Economic Review 85(4), 1995.

VARIAN, H. R. Microeconomia: princípios básicos. 4. ed. Rio de Janeiro: Campus, 2003.

VARIAN, H. R., Microeconomic Analysis. 3. ed. New York: W.W. Norton, 1992.

VAUGHAN D, COUCH S (2024). workflows: Modeling Workflows. R package version 1.1.4, <https://CRAN.R-project.org/package=workflows>.

VON NEUMANN, J.; MORGENSTERN, O. Theory of games and economic behavior. Princeton University Press, 1944.

WALKER, Kent. A Systematic Review of the Corporate Reputation Literature: Definition, Measurement, and Theory. Corporate Reputation Review. Vol. 12. 2010

WARTICK, S. L. Measuring corporate reputation – Definition and data. Business & Society, 41(4), 371-392. 2002

WATSON, G. S. Smooth regression analysis. Sankhya: The Indian Journal of Statistics, Series A, 1964.

WICKHAM H (2023). stringr: Simple, Consistent Wrappers for Common String Operations. R package version 1.5.1, <https://CRAN.R-project.org/package=stringr>.

WICKHAM, H. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

WICKHAM H, HENRY L (2023). purrr: Functional Programming Tools. R package version 1.0.2, <https://CRAN.R-project.org/package=purrr>.

WICKHAM H, FRANÇOIS R, HENRY L, MÜLLER K, VAUGHAN D (2023). dplyr: A Grammar of Data Manipulation. R package version 1.1.4, <https://CRAN.R-project.org/package=dplyr>.

WOOLDRIDGE, Jeffrey M. Introductory Econometrics: A Modern Approach, South-Western College Publishing, 2000.

XIE Y (2024). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.47, <https://yihui.org/knitr/>.

YAKOVLEV, A.; BALAEVA, O.; TKACHENKO, A. Estimation of procurement costs incurred by public customers: a case study of a Russian region. Journal of Public Procurement, 18(1), 2018.

YEUNG, Luciana; CAMELO, Bradson. Introdução à Análise Econômica do Direito. São Paulo: Juspodivm, 2023.

YU, J. Discrete Approximation of Continous Allocation Machanisms. PhD thesis, California Institute of Technology, Division of Humanities and Social Science,

1999.

Apêndice A

Códigos em linguagem R - Link para github

O script e os dados estão disponíveis no link: https://github.com/BradCamelo/Custo_Trans_Licit_Pb.