UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

1.5em

Predição de Deságios em Leilões de Transmissão da ANEEL com o Uso de Inteligência Artificial Interpretável

Luiz Felipe Casali Migliato

Dissertação de Mestrado do Programa de Mestrado Profissional em Matemática, Estatística e Computação Aplicadas à Indústria (MECAI)

ICMC USP SÃO CARLOS

Luiz Felipe Casali Migliato

Predição de Deságios em Leilões de Transmissão da ANEEL com o Uso de Inteligência Artificial Interpretável

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP, como parte dos requisitos para obtenção do título de Mestre – Mestrado Profissional em Matemática, Estatística e Computação Aplicadas à Indústria. *Versão revisada.*

Área de Concentração: Matemática, Estatística e Computação

Orientador: Prof. Dr. André Carlos Ponce de Leon Ferreira de Carvalho

USP – São Carlos
Fevereiro de 2023

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

C334p

Casali Migliato, Luiz Felipe
Predição de Deságios em Leilões de Transmissão da ANEEL com o Uso de Inteligência Artificial Interpretável / Luiz Felipe Casali Migliato; orientador André Carlos Ponce de Leon Ferreira de Carvalho. -- São Carlos, 2023.
72 p.

Dissertação (Mestrado - Programa de Pós-Graduação em Mestrado Profissional em Matemática, Estatística e Computação Aplicadas à Indústria) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2023.

1. Inteligência Artificial. 2. Árvores de Decisão. 3. Random Forest. 4. XGBoost. 5. CatBoost. I. Ponce de Leon Ferreira de Carvalho, André Carlos, orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Luiz Felipe Casali Migliato

Prediction of Discounts in ANEEL Transmission Auctions with the Use of Interpretable Artificial Intelligence

Dissertation submitted to the Instituto de Ciências Matemáticas e de Computação – in accordance with the requirements of the Professional Master's Program in Mathematics Statistics and Computing Applied to Industry, for the degree of Master in Science. *Final version.*

Concentration Area: Mathematics, Statistics and Computing

Advisor: Prof. Dr. André Carlos Ponce de Leon Ferreira de Carvalho

USP – São Carlos
February 2023

Dedico este trabalho à minha família, por sempre acreditarem e estarem ao meu lado.

AGRADECIMENTOS

Agradeço primeiramente à Deus por me possibilitar a ter todas as condições necessárias para desenvolver este trabalho.

Agradeço aos meus pais, Antonio Luiz e Iara Regina, e ao meu irmão, William, por todo o suporte e direcionamento que me foi oferecido durante essa jornada e por sempre estarem ao meu lado independentemente da situação, nada seria possível sem eles.

Agradeço pelo suporte e entendimento da minha namorada, Elisângela.

Agradeço aos professores e colaboradores da Universidade de São Paulo. Em especial, ao Prof. Dr. André Carvalho, pelo tempo e paciência durante a sua orientação neste trabalho.

Agradeço aos meus gestores de trabalho no Brasil, Geraldo Júlio e Adão, e ao atual no Canadá, Marco Barbosa, por todos os conselhos e flexibilidade concedida para a realização deste trabalho.

Agradeço aos meus inúmeros amigos, presentes da vida, e colegas que estiveram ao meu lado.

Agradeço a todos aqueles que, de alguma forma, colaboraram positivamente comigo.

“A menos que modifiquemos a nossa maneira de pensar, não seremos capazes de resolver os problemas causados pela forma como nos acostumamos a ver o mundo.”

(Albert Einstein)

RESUMO

MIGLIATO, L. F. C. **Predição de Deságios em Leilões de Transmissão da ANEEL com o Uso de Inteligência Artificial Interpretável**. 2023. 72 p. Dissertação (Mestrado – Mestrado Profissional em Matemática, Estatística e Computação Aplicadas à Indústria) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2023.

A predição de resultados futuros através do uso da inteligência artificial é uma aplicação relevante em diversas áreas, como a industrial, financeira, agronegócio, entre outras. A aplicação de inteligência artificial interpretável pode trazer um conhecimento adicional dos dados para os especialistas, além de poder ser traduzida em vantagem competitiva pelas empresas que a utiliza. Dessa forma, para os Leilões de Transmissão da ANEEL buscou-se investigar a capacidade preditiva de quatro algoritmos de Aprendizado de Máquina interpretáveis, mais especificamente Árvore de Decisão, *Random Forest*, XGBoost e CatBoost, em contextos gerados a partir de diferentes métodos de seleção de variáveis. A comparação e a avaliação do desempenho dos modelos gerados por esses algoritmos foram feitas a partir das métricas $RMSE$ e R^2, bem como o teste de hipótese de Friedman e o teste post-hoc de Nemenyi. Os resultados demonstraram que o contexto mais adequado foi o CatBoost com todas as variáveis. Assim, foi estudada a interpretabilidade do modelo através das árvores geradas e os atributos mais destacados, além de ser aplicado para predizer deságios em lotes de Leilões da ANEEL utilizando dados reais não visto.

Palavras-chave: Inteligência Artificial, Árvores de Decisão, Random Forest, XGBOOST, CatBoost.

ABSTRACT

MIGLIATO, L. F. C. **Prediction of Discounts in ANEEL Transmission Auctions with the Use of Interpretable Artificial Intelligence**. 2023. 72 p. Dissertação (Mestrado – Mestrado Profissional em Matemática, Estatística e Computação Aplicadas à Indústria) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2023.

The prediction of future outcomes through the use of artificial intelligence is a relevant application in several areas, such as industrial, financial, and agrobusiness, among others. The use of interpretable artificial intelligence may lead to additional knowledge of the data for the experts, besides bringing competitive advantage for the company which uses it. In this way, for the ANEEL Transmission Auctions it was investigated the prediction capacity of four machine learning algorithms, in particular, Decision Tree, *Random Forest*, XGBoost and CatBoost, in contexts devired from diffent variable selection methods. The comparison and performance evaluation of the models generated by this algorithms were analyzed through two metrics, *RMSE* e R^2, as well as through Friedman hypothesis test and Nemenyi post-hoc test. The results show that the context CatBoost with all variables was the most adequate one. Therefore, its interpretability was studied through the generated trees and the most important variables indicated by the model, besides being applicable to predict the discounts in lots of ANEEL Transmission Auctions using actual data unseen by the model.

Keywords: Artificial Intelligence, Decision Trees, Random Forest, XGBoost, CatBoost.

LISTA DE ILUSTRAÇÕES

LISTA DE QUADROS

LISTA DE TABELAS

SUMÁRIO

CAPÍTULO

1

INTRODUÇÃO

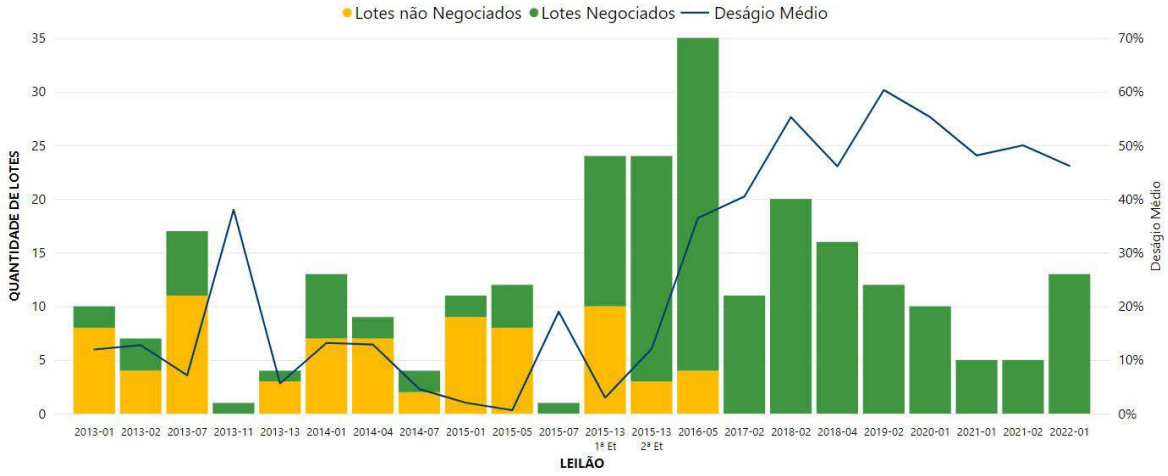
Anualmente, desde 1999, a Agência Nacional de Energia Elétrica (ANEEL) por delegação e em conformidade com as diretrizes do Ministério de Minas e Energia realiza leilões de Transmissão de Energia Elétrica. Tais leilões, vistos como obras de expansão da infraestrutura elétrica do Brasil, tem a finalidade de contratar concessões do serviço público, incluindo a construção, a operação e a manutenção de instalações de transmissão do Sistema Interligado Nacional (SIN). Como visto no último leilão aprovado para Junho de 2022, os investimentos totais estimados podem chegar na ordem de 15,3 bilhões de reais com estimativa de geração de mais de 30 mil empregos diretos para a construção de 5.289 quilômetros de linha de transmissão e 6.180 MVA em capacidade de transformação para subestações (EPE, 2022).

Cada leilão é composto por diversos lotes de empreendimentos independentes ao longo de todo o território nacional. Para cada lote de empreendimento é estabelecida, pela ANEEL, uma Receita Anual Permitida (RAP) para a prestação do serviço público durante um período de concessão de, geralmente, 30 anos. A RAP é a máxima receita anual que a empresa vencedora de um determinado lote terá direito a receber pela prestação de serviço público de transmissão aos usuários, a partir da entrada em operação comercial das instalações de transmissão. A empresa participante que apresentar a menor proposta de RAP será declarada vencedora do lote e terá o direto de celebrar o correspondente Contrato de Concessão. Em outras palavras, a empresa participante do leilão que apresentar, para um determinado lote, a proposta com o maior deságio em relação à RAP previamente estabelecida pela ANEEL, será consagrada vencedora do lote do leilão. Portanto, a variável decisória nesse processo é o valor do deságio que será vencedor, desconhecido até o encerramento da disputa do lote do leilão pelas empresas interessadas no mesmo (ANEEL, 2022).

As empresas participantes podem ser *utilities*, empresas diretamente ligadas ao setor de geração, transmissão e/ou distribuição de energia elétrica, construtoras ou fundos de investimentos, usualmente na modalidade individual ou em consórcio, podendo ainda ter ligações com

instituições internacionais. Nos últimos anos, os leilões têm contado com novos participantes, empresas entrantes no setor de transmissão de energia elétrica. Tal quadro com diversos tipos de empresas nas mais diferentes modalidades, induz a um cenário acirrado de maior competitividade como pode ser observado na Figura 1. É importante ressaltar que para participar em um lote do leilão, as empresas participantes iniciam os trabalhos de preparação de suas propostas com meses de antecedência, o que implica em um alto volume de investimento inicial pelas tarefas de visitas aos locais dos empreendimentos, elaboração dos projetos de engenharia, desenvolvimento e análise do modelo de negócio e até emissão de garantias financeiras exigidas pelo edital da ANEEL. Tal investimento é totalmente perdido caso a empresa não se consagre vencedora.

Figura 1 – Lotes não negociados, lotes negociados e deságio médio dos leilões desde 2013.



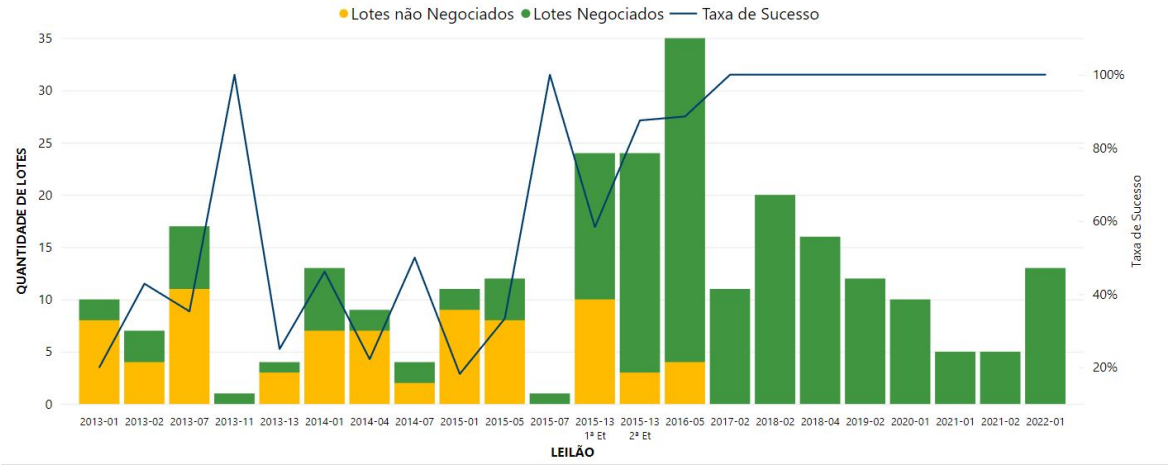
Fonte: https://www.gov.br/aneel/pt-br/centrais-de-conteudos/relatorios-e-indicadores/leiloes

A Figura 1 mostra esse acirramento através do aumento do deságio médio para próximo de 50% nas edições do leilão a partir de 2018, enquanto que anteriormente, esse valor não ultrapassava os 40% no período retratado no gráfico. O número de lotes negociados, que tiveram pelo menos uma empresa demonstrando interesse através de uma oferta de RAP, por edição do leilão também reforça essa condição. Como observado na Figura 2, a taxa de sucesso de lotes negociados sobre os lotes não negociados, desde 2017, foi de 100%.

Como consequência do crescimento econômico do país nas últimas duas décadas, houve o aumento da demanda por investimentos na infraestrutura do setor elétrico. Tal aumento de investimentos foi possível também pela capacidade das empresas de absorverem e executarem os empreendimentos requeridos, além da forte demonstração de interesse por parte das mesmas em concretizar esses investimentos (PRESTES *et al.*, 2019). A Figura 3 retrata o volume de investimento estimado crescente ao longo do período de realização dos leilões. Na mesma representação, é possível observar o investimento estimado atualizado pelo Índice Nacional de Preços ao Consumidor Amplo (IPCA) entre a data de acontecimento do leilão e a última edição em Junho de 2022.

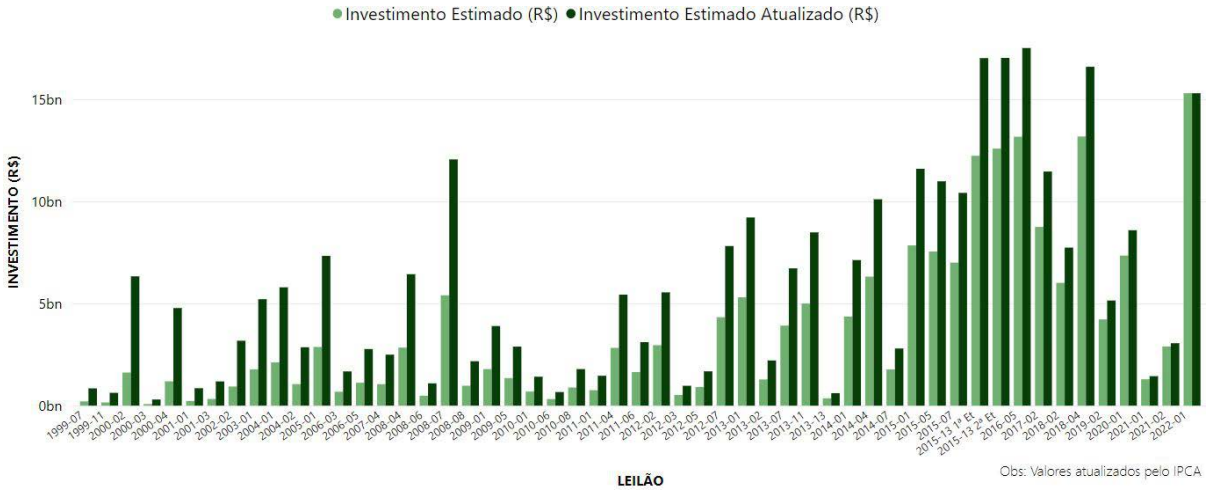
Caso uma empresa participante seja vencedora de um lote, todo o processo de recompen-

Figura 2 – Lotes não negociados, lotes negociados e taxa de sucesso desde 2013.



Fonte: https://www.gov.br/aneel/pt-br/centrais-de-conteudos/relatorios-e-indicadores/leiloes

Figura 3 – Investimento por edição do leilão.



Fonte: https://www.gov.br/aneel/pt-br/centrais-de-conteudos/relatorios-e-indicadores/leiloes

sação financeira pela ANEEL através da RAP ocorre na moeda corrente oficial da República Federativa do Brasil, o real. De acordo com o edital do leilão, a correção monetária do RAP é feita de forma anual pelo índice IPCA. Entretanto, uma parcela significativa dos custos para realização dos empreendimentos esta atrelada ao preço das matérias-primas, *commodities* e, de materiais elétricos, que tem como base de cotação internacional o dólar americano. Dessa

forma, as variáveis macroeconômicas como as taxas de juros, a taxa do câmbio e a inflação desempenham um papel fundamental no valor do deságio que será proposto pela empresa.

Portanto, se essa empresa conhecer antecipadamente o valor do deságio vencedor, poderá se preparar mais adequadamente e apresentar maior competitividade em sua proposta de deságio. Do ponto de vista operacional, o desenvolvimento de um modelo com um bom desempenho de prever o valor do deságio vencedor para os lotes do leilão ANEEL, com base nas informações técnicas e macroeconômicas, tem potencial para agregar um alto valor estratégico. Espera-se, inclusive, que o valor agregado seja refletido em ganhos financeiros para a empresa. Entretanto, estimar este ganho requereria acesso a dados e informações que se encontram fora do escopo deste trabalho.

1.1 Objetivos

Dessa forma, buscou-se: (i) investigar a capacidade preditiva de quatro algoritmos de Aprendizado de Máquina, mais especificamente Árvore de Decisão, *Random Forest*, XGBoost e CatBoost, em contextos gerados a partir de diferentes métodos de seleção de variáveis; (ii) comparar e avaliar o desempenho dos modelos gerados por esses algoritmos utilizando as métricas *RMSE* e R^2, bem como o teste de hipótese de Friedman e o teste post-hoc de Nemenyi; e (iii) aplicar o contexto mais adequado para predizer deságios em lotes de Leilões da ANEEL utilizando dados reais não visto.

1.2 Organização

Esta dissertação está organizada em cinco capítulos: Introdução, Fundamentação Teórica, Metodologia, Resultados e Discussão, e Conclusão. No segundo capítulo será apresentada a revisão teórica dos quatro algoritmos de Aprendizado de Máquina abordados nesse trabalho. No terceiro capítulo será apresentada a metodologia empregada na elaboração do experimento. No quarto capítulo serão apresentados análise e discussão dos resultados. E no quinto capítulo será apresentada a conclusão do trabalho.

CAPÍTULO

2

FUNDAMENTAÇÃO TEÓRICA

O objetivo deste trabalho foi comparar o desempenho de quatro algoritmos de Aprendizado de Máquina em uma situação real de predição de deságio em leilões da ANEEL. Os algoritmos comparados foram: Árvores de Decisão, Random Forest, XGBoost e CatBoost. Este capítulo descreve o funcionamento desses algoritmos.

2.1 Árvores de Decisões

Uma árvore de decisões é uma estrutura de dados hierárquica que implementa a estratégia "dividir e conquistar". É um método não paramétrico eficiente, que pode ser utilizado para classificação ou regressão. Com um método paramétrico, define-se um modelo sobre todo o espaço de entrada e seus parâmetros são aprendidos sobre o conjunto de treinamento. Em seguida, aplica-se o mesmo modelo e os mesmos parâmetros sobre dados de testes. Com um método não paramétrico, divide-se o espaço de entrada em regiões locais, definidas por medidas de distância, como, por exemplo, a norma euclidiana, e para cada entrada o modelo local calculado a partir dos dados de treinamento daquela região é utilizado. No entanto, esse processo pode ser custoso (ALPAYDIN, 2014).

Uma árvore de decisões é um modelo hierárquico utilizado em aprendizado supervisionado no qual a região local é identificada numa sequência de divisões recursivas (ALPAYDIN, 2014), ou seja, recursivamente, particiona-se o espaço de entrada e define-se um modelo local para cada região resultante dessa divisão do espaço de entrada (MURPHY, 2022). Em outras palavras, os modelos baseados em árvores dividem o espaço de características em um conjunto de retângulos, e em seguida ajustam um simples modelo para cada um desses retângulos (BISHOP; NASRABADI, 2006; HASTIE *et al.*, 2009). Mesmo sendo simples, esses modelos são bastante poderosos (HASTIE *et al.*, 2009).

Os modelos baseados em árvores podem ser vistos como um método de combinação de

modelos no qual somente um modelo é responsável por realizar predições em determinado ponto no espaço de entrada. O processo de selecionar um modelo específico, dada uma determinada entrada, pode ser descrito por um processo de tomada de decisões sequenciais, correspondente a uma árvore binária, que separa cada nó em dois ramos (BISHOP; NASRABADI, 2006).

Quando se constrói um modelo de árvore de decisões, primeiramente, procura-se pela característica com maior ganho de informação. Essa característica é colocada na raiz da árvore de decisões. Para cada ramo que resulta da raiz, procura-se novamente pela característica que possui o maior ganho de informação (HULL, 2021). Uma medida utilizada para o ganho de informação é a entropia, dada por:

$$Entropia = -\sum_{i=1}^{n} p_i \log(p_i) \tag{2.1}$$

Onde p_i é a proporção de dados que pertencem à classe i.

Para se quantificar o ganho de informação, uma alternativa à medida de entropia é a medida Gini, dada por:

$$Gini = 1 - \sum_{i=1}^{n} p_i^2 \tag{2.2}$$

Os modelos de árvore de decisões são populares por diversos motivos. O estabelecimento hierárquico das decisões permite uma localização rápida da região que cobre uma determinada entrada e possuem alta interpretabilidade, pois esse modelo pode ser convertido para um conjunto de regras "*if-then*", facilmente compreendido (ALPAYDIN, 2014; MURPHY, 2022). Além disso, as árvores de decisões conseguem lidar facilmente com entradas que possuem valores tanto discretos quanto contínuos, não há necessidade de conversão dos dados, a seleção de variáveis é automática e são relativamente robustas quanto a *outliers*. Somam-se à essas vantagens o fato de serem rápidas para treinamento e podem lidar com valores faltantes (MURPHY, 2022). Por essas diversas razões, as árvores de decisões são muito populares e, inclusive, algumas vezes preferidas mesmo em relação a modelos mais precisos, porém, menos interpretáveis e mais complexos (ALPAYDIN, 2014).

Entretanto, segundo (JAMES *et al.*, 2013) as árvores de decisão apresentam uma alta variância. Na prática isso significa que se o conjunto de treino for dividido em duas partes aleatórias e o modelo de árvores de decisão for aplicado em ambas, os resultados obtidos podem ser relativamente diferentes. Uma forma trivial para reduzir a variância consiste em gerar B subconjuntos a partir do conjunto de treino, procedimento esse chamado de *boostrap*, treinar o modelo em cada um desses subconjuntos de treino, com a finalidade de obter $\hat{f}^{*b}(x)$ para então se calcular a média dos valores preditos.

$$\hat{f}_{bag}(x) = \frac{1}{B} \sum_{b=1}^{B} \hat{f}^{*b}(x) \tag{2.3}$$

Esse processo é conhecido como *bagging*, que justamente significa *boostrap aggregating* e tem por objetivo reduzir a variância (FRANK; MARK, 2016). O número de árvores B a ser adotado é suficientemente grande para que o erro seja estabelecido e não ocorra *overfitting*, memorização do conjunto de teste e, consequentemente, de seu ruído, reduzindo sua capacidade de generalizar para dados não vistos (MARSLAND, 2011). Na prática, porém, a existência de uma forte variável preditora no conjunto de dados levará a maioria ou todas as árvores de decisões a utilizá-la em sua primeira divisão. Consequentemente, todas as árvores obtidas pelo *bagging* serão similares entre si e, portanto, as predições estarão altamente correlacionadas. O cálculo da média de várias árvores correlacionadas não direciona tanto para a redução da variância quanto a média de várias árvores não correlacionadas. Ou seja, isto significa que o *bagging* pode não levar a uma redução substancial da variância em comparação com uma única árvore desse conjunto (JAMES *et al.*, 2013).

O próximo algoritmo a ser discutido, *Random Forest*, busca superar essa questão por apresentar um método que descorrelaciona as árvores geradas.

2.2 Random Forest

Em muitos casos, o foco da área de Aprendizado de Máquina está em desenvolver um único modelo de predição que visa ser o mais preciso possível para uma dada tarefa. A ideia por trás do modelo de *Random Forest* desvia-se um pouco dessa abordagem. Ao invés de ajustar um único modelo, um conjunto com diversos modelos é treinado e as predições são geradas agregando-se os resultados desses vários modelos. Um modelo de predição que é composto a partir de um conjunto de modelos é chamado de *ensemble* (KELLEHER; NAMEE; D'ARCY, 2020). Um dos *ensembles* mais populares é o formado por árvores de decisões, chamado de *Random Forest*, ou Florestas Aleatórias. Nesse caso, treina-se não somente uma única árvore, mas diversas árvores de decisões, cada uma em um subconjunto aleatório de treinamento ou em um subconjunto aleatório de variáveis de entrada, e essas predições são, então, combinadas e a acurácia total pode ser consideravelmente ampliada (ALPAYDIN, 2014).

O modelo *Random Forest* em seu processo de elaboração das árvores de decisões, a cada momento de uma divisão, toma uma amostra aleatória de m variáveis preditoras escolhidas dentre o total das p existentes. Nesta divisão, então, pode-se utilizar apenas uma das m variáveis escolhidas e uma nova amostra m é feita a cada divisão. Tipicamente, é escolhida $m \approx \sqrt{p}$. Em outras palavras, o modelo a cada nova divisão em uma árvore de decisão não está permitido a considerar a maioria dos preditores disponíveis. Portanto, em média $(p\text{-}m)\,/\,p$ das divisões não considerarão um forte preditor, aumentando o uso dos outros preditores (JAMES *et al.*, 2013). Assim, o procedimento cria árvores descorrelacionadas, fazendo com que a média das árvores resultantes seja menos variável e, portanto, mais confiável (MARSLAND, 2011).

A principal diferença entre *bagging* e *Random Forest* está justamente na escolha do

tamanho do subconjunto m. De fato, se um modelo de *Random Forest* é construído utilizando-se $m = p$, os resultados das duas abordagens serão os mesmos (JAMES *et al.*, 2013).

2.3 XGBoost

Dentre os métodos de aprendizado de máquina utilizados em casos práticos, o *Gradient Tree Boosting* é um dos que mais se destaca em várias aplicações por sua escalabilidade e excelentes resultados. O *Extreme Gradient Boosting* (XGBoost) é ainda uma forma mais eficiente de implementação do *Gradient Tree Boosting* (CHEN; GUESTRIN, 2016).

Primeiramente, é necessário entender o conceito de *boosting*, também considerado um *ensemble*. Na abordagem *bagging* vista anteriormente, cópias múltiplas eram criadas do conjunto de treino original utilizando-se o *boostrap*, aplicadas separadamente em árvores de decisões e, então, combinava-se todas as árvores em ordem de criar um modelo único de predição. Já no procedimento *boosting*, as árvores de decisões são geradas sequencialmente. Isso significa que cada nova árvore utiliza informação de árvores já geradas. Em outras palavras, notavelmente as árvores eram geradas através de um conjunto de dados *boostrap* e, consequentemente, independentes entre si. *Boosting* não se utiliza desse conceito, ao invés, aplica cada árvore de decisão em uma versão modificada do conjunto de treino original, em específico nos resíduos do modelo, até que novas melhorias não sejam mais possíveis (JAMES *et al.*, 2013).

O modelo *Gradient Tree Boosting* avança um passo em relação ao *boosting*, pois iterativamente constrói novas, sempre melhores, árvores de decisões tentando minimizar o erro através do gradiente descendente. Considerando um modelo F que prediz $\hat{y}$ e busca otimizar uma função de desempenho, tipicamente o erro quadrático médio $(\hat{y} - y)^2$. A cada passo $1 \leq m \leq M$, o modelo F_m é aprimorado tentando obter-se $F_{m+1}(x) = F_m(x) + d(x) = y$. A fim de se encontrar $d(x)$, a equação pode ser reescrita como $d(x) = y - F_m(x)$, onde $y - F_m(x)$ é o gradiente negativo da função de perda do erro quadrático $\frac{1}{2}\,(y - F_m(x))^2$ (FRIEDMAN, 2001).

XGBoost implementa o *Gradient Tree Boosting* e introduz um termo de regularização para reduzir o *overfitting*. Levando em consideração a complexidade, uma melhor avaliação é obtida da qualidade das árvores e, portanto, uma avaliação da melhor divisão possível. Outra diferença é que, enquanto o *Gradient Tree Boosting* finaliza as suas divisões quando é encontrado uma perda na divisão, o XGBoost realiza as divisões até a profundidade máxima especificada e, então, poda os nós para além dos quais não existe um ganho positivo (GORMAN, 2017).

2.4 CatBoost

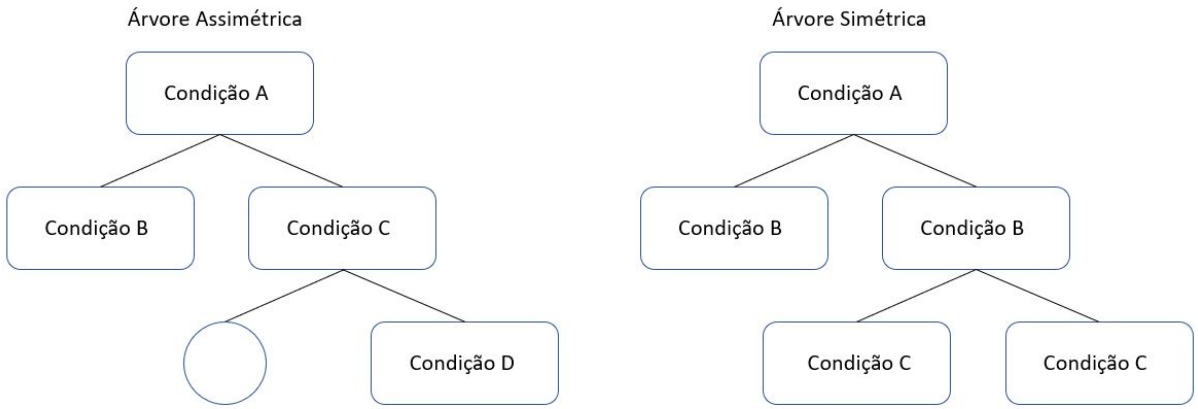
O modelo *Categorial Boosting* (CatBoost) é uma outra variação da implementação do algoritmo de *Gradient Tree Boosting*. O modelo recebe esse nome pois consegue de forma bem sucedida trabalhar com variáveis categóricas não-numéricas, exigindo o mínimo em transfor-

mação de variáveis, o que se opõe a maioria dos outros algoritmos de aprendizado de máquina como o próprio *Random Forest* e o XGBoost, onde a prática do *enconding*, transformação das variáveis em valor numérico, anteriormente a utilização do modelo se faz necessária (THIESEN, 2021).

A partir da utilização de uma nova forma para cálculo das folhas na formação das estruturas das árvores, o modelo CatBoost apresenta também um forte controle para *overfitting*. Inicialmente, o modelo gera permutações aleatórias independentes do conjunto de treino. As permutações, então, são usadas para a avaliação das divisões, definindo a estrutura das árvores, como por exemplo, os nós internos. A utilização de apenas uma permutação pode aumentar a variância das predições do modelo final, enquanto que o uso de diferentes permutações para treinar modelos distintos reduz esse efeito e não leva ao *overfitting* (PROKHORENKOVA *et al.*, 2018). Dessa forma, a aplicação de várias permutações melhora a robustez do algoritmo (DOROGUSH; ERSHOV; GULIN, 2018).

No processo de formação das árvores de decisão, o modelo CatBoost não segue os modelos similares de *Gradient Tree Boosting*. Ao invés, o modelo gera *oblivious trees*, também chamadas *decision tables* ou, em tradução literal, tabelas de decisão (LOU; OBUKHOV, 2017). O termo *oblivious* significa que o mesmo critério de divisão é aplicado ao longo de todos os nós do mesmo nível de profundidade da árvore como visto na Figura 4 (PROKHORENKOVA *et al.*, 2018). Tais árvores, conhecidas como árvores simétricas ou balanceadas, resultam na menor perda ao longo de todos os nós da mesma profundidade nas condições de divisão, uma vez que a estrutura da árvore opera como uma regularização para encontrar uma solução ótima (THIESEN, 2021). Os benefícios desse procedimento inclui uma computadorização e avaliação mais rápida, além de evitar o *overfitting* como mencionado anteriormente. Os modelos *Random Forest* e XGBoost são baseados em árvores assimétricas, onde condições de divisão para cada nó ao longo de um mesmo nível podem ser diferentes (WONG, 2022).

Figura 4 – Árvores Assimétricas e Simétricas.



Fonte: Elaborada pelo autor.

CAPÍTULO

3

METODOLOGIA

Conforme descrito no Capítulo 1, este trabalho visou alcançar os seguintes objetivos: (i) investigar a capacidade preditiva de quatro algoritmos de Aprendizado de Máquina, mais especificamente Árvore de Decisão, Random Forest, XGBoost e CatBoost, em contextos gerados a partir de diferentes métodos de seleção de variáveis; (ii) comparar e avaliar o desempenho dos modelos gerados por esses algoritmos utilizando as métricas $RMSE$ e R^2, bem como o teste de hipótese de Friedman e o teste post-hoc de Nemenyi; e (iii) aplicar o contexto mais adequado para predizer deságios em lotes de Leilões da ANEEL utilizando dados reais não visto.

Para alcançar esses objetivos, foi desenhado um Processo Experimental com sete etapas. Esse processo será descrito e detalhado no restante deste capítulo.

3.1 Processo Experimental

As sete etapas do Processo Experimental elaborado para viabilizar este trabalho são:

- Etapa 1: Coleta dos dados em fontes primárias;
- Etapa 2: Preparação dos dados coletados;
- Etapa 3: Análise exploratória dos dados;
- Etapa 4: Seleção de variáveis e geração dos conjuntos de treinamento;
- Etapa 5: Treinamento e avaliação dos modelos; e
- Etapa 6: Análise do modelo selecionado e da Importância das Variáveis; e
- Etapa 7: Predição da variável alvo "deságio".

3.1.1 Etapa 1: Coleta dos Dados em Fontes Primárias

Os dados utilizados neste trabalho dizem respeito aos Leilões de Transmissão de Energia Elétrica, envolvendo linhas de transmissão e subestações de alta tensão, realizados pela ANEEL. Esses dados foram coletados em fonte primária, diretamente no site da instituição (https://www.gov.br/aneel/pt-br/centrais-de-conteudos/relatorios-e-indicadores/leiloes). Nesse site, foram coletados dados desde o início da realização dos leilões, em 1999 até 2020, referentes a 13 variáveis, descritas no Quadro 1. Abaixo do nome da variável, entre parênteses, é mostrado as abreviações dos nomes conforme aparecem na base de dados e como serão utilizados em algumas tabelas e gráficos deste trabalho.

Quadro 1 – Nomes e descrições das variáveis coletadas neste estudo.

Variáveis	Descrição
Data (date)	Refere-se à data em que o leilão foi realizado. Os leilões abordados neste estudo ocorreram entre os anos de 1999 e 2020.
Linhas de Transmissão (lt)	Refere-se ao nível de tensão em kV das linhas de transmissão As linhas podem ter 6 níveis diferentes: 525, 500, 440, 345, 230 ou 130 kV.
Subestações (se)	Refere-se ao nível de tensão em kV das subestações. Os níveis de tensão são os mesmos das linhas de transmissão.
Número de Subestações (num_se)	Refere-se ao número de subestações ofertadas em cada lote.
Extensão (extension_km)	Refere-se à extensão em quilômetros (km) das linhas de transmissão presentes no lote.
Extensão total (extension_total_km)	Refere-se à extensão total das linhas de transmissão envolvidas no Leilão, também em quilômetros (km).
Regiões (north, south, center_west) (northeast, southeast)	Refere-se às regiões do Brasil onde os empreendimentos serão realizados.
Potencia Máxima do Transformador (max_pot_transformer_se)	Potência máxima de transformação das subestações (em MVA) presentes no Lote.
Prazo (term_months)	Prazo máximo para realização do empreendimento
RAP inicial (rap_intial_brl)	Receita Anual Permitida (RAP) estipulada inicialmente para o Lote pela ANEEL, em Reais (R$).
RAP total (rap_total_brl)	Receita Anual Permitida (RAP) total estipulada para o Leilão daquele Lote, em Reais (R$).
Deságio (target_perc)	Variável alvo a ser predita, representa o deságio vencedor em percentual de cada Lote.

Fonte: https://www.gov.br/aneel/pt-br/centrais-de-conteudos/relatorios-e-indicadores/leiloes

Além das variáveis listadas acima, para a originalidade do trabalho, foram também coletados dados macroeconômicos que poderiam afetar os lances recebidos nos leilões. Esses dados foram coletados levando-se em conta as datas do leilões. O Quadro 2 apresenta uma

descrição dessas variáveis.

Quadro 2 – Descrição das variáveis macroeconômicas.

Variáveis	Descrição
Dólar (v_usd)	Refere-se ao valor do dólar em reais na data anterior ao dia do leilão.
IPCA (ipca_acc_12months)	Refere-se à taxa de inflação acumulada nos últimos 12 meses anteriores ao mês de realização do leilão.
Taxa Selic (tx_selic)	Refere-se aos valores da taxa Selic nas datas em que os leilões foram realizados.

Fonte: www.valor.com

No total, foram coletadas dados acerca de 16 variáveis, que serão tratadas na etapa seguinte. A base de dados conta com 337 observações.

3.1.2 Etapa 2: Preparação dos Dados Coletados

Após a coleta, os dados foram tratados e reestruturados de forma a estarem apropriados para serem utilizados pelos algoritmos abordados neste estudo. A primeira transformação foi com relação às variáveis "Linha de Transmissão"e "Subestações", com nível de tensão medida em quilovolt (kV). As linhas de transmissão e as subestações ofertadas nos diversos leilões possuíam níveis diferentes, níveis estes padronizados. Por exemplo, enquanto alguns leilões possuíam linhas de transmissão com 525kV, outros ofertavam linhas com 440kV ou 230kV (ou uma das seis possibilidades descritas na Etapa 1).

Dessa forma, para cada nível de tensão foram criadas variáveis diferentes, referentes tanto às linhas de transmissão como também para as subestações. Por exemplo, para a linha de transmissão e subestação com nível de tensão de 525kV, foram criadas as variáveis lt525 e se525, respectivamente. Para os leilões que apresentaram linhas de transmissão e subestações com esse nível de tensão, essas variáveis receberam o valor de 525. Para os leilões que não apresentaram linhas de transmissão e subestações com essa tensão, essas variáveis receberam o valor zero. O mesmo foi feito para todos os outros cinco valores de tensão encontrados nos leilões. Optou-se por atribuir o valor da tensão a essas variáveis, ao invés de 1, para que as devidas proporções entre as tensões das diversas linhas de transmissão e subestações fossem resguardadas, uma vez que existe uma escala de tamanho físico proporcional dos equipamentos e materiais utilizados nesses valores.

Uma segunda transformação de variáveis realizada foi sobre a variável “Estados”. Um mesmo leilão poderia estar relacionado a vários estados em diferentes regiões do país. Dessa forma, essa variável foi convertida em cinco variáveis *dummies*, uma para cada região do país (ver Quadro 1). Assim, se um determinado lote envolvesse, por exemplo, 3 estados de uma dada região, para a variável referente à essa região seria atribuído o valor 3. Se para uma determinada

região, não houvesse nenhum estado envolvido, a variável referente à essa região receberia o valor zero. A variável original “Estados” foi removida da base de dados.

A última transformação realizada sobre os dados foi com relação à variável "Data". A partir dessa variável, foram extraídos os meses de realização dos leilões e atribuídos a uma nova variável, denominada na base de dados como “month”. Após essa transformação, a variável original "Data"foi retirada da base da dados.

Após as transformações realizadas nesta etapa, a base de dados passou a contar com 30 variáveis. Em seguida, a base de dados foi divididas em dois conjuntos. Um denominado conjunto de treinamento (“X_train”), composto por 80% dos dados, e outro denominado conjunto de teste (“X_test”), composto por 20% dos dados. Prática essa comumente adotada para a divisão da base de dados. Para cada conjunto, a variável alvo foi isolada, gerando as sequências “y_-train” e “y_test”, respectivamente. Ao final, o conjunto de treinamento foi composto por 269 observações e o conjunto de teste por 68 observações. O conjunto de teste foi isolado e retomado somente na última etapa da pesquisa, quando da predição da variável alvo (“deságio”). Para a análise exploratória (Etapa 3), seleção de variáveis (Etapa 4), bem como treinamento e avaliação dos modelos (Etapa 5), foi utilizado somente o conjunto de treinamento. Dessa forma, ficou assegurado que na etapa de predição, o modelo selecionado como mais adequado fosse aplicado sobre um conjunto de dados nunca visto (Etapa 7).

3.1.3 Etapa 3: Análise Exploratória dos Dados

Nesta seção, será descrita e apresentada a análise exploratória dos dados como uma forma de apresentar a base de dados de forma mais aprofundada. Primeiramente, para se ter uma compreensão melhor dos dados, buscou-se conhecer as médias das variáveis numéricas presentes na base de dados. A Tabela 1 apresenta esses valores.

Tabela 1 – Médias das variáveis quantitativas que compõem a base de dados deste estudo.

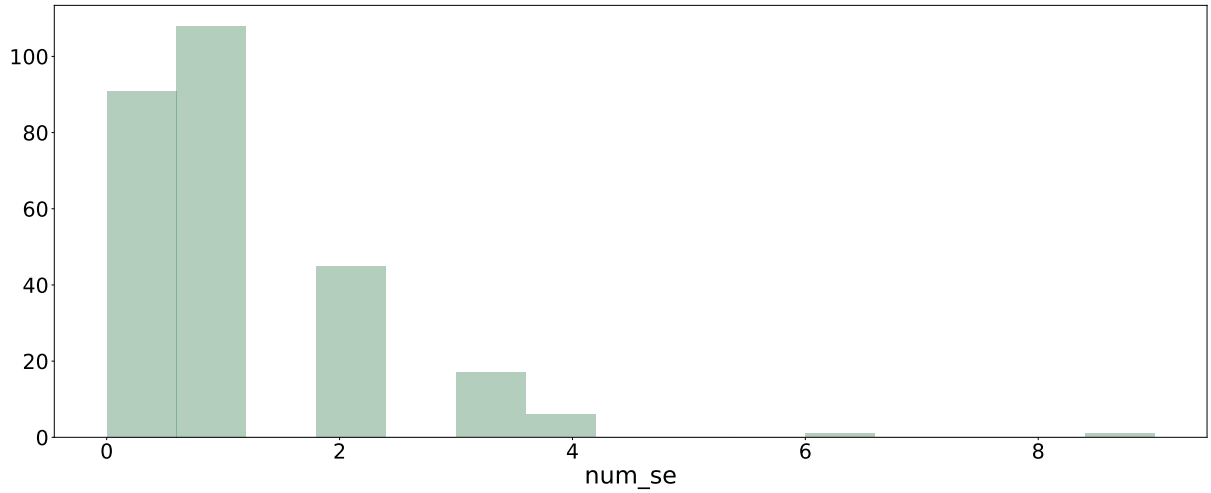
Variáveis	Média
Número de subestações	1.04
Potência máxima dos transformadores	539 kWh
Extensão total das linhas de transmissão	2.934 km
Prazo	38 meses
Dólar	U$2.70
Taxa Selic	11%
IPCA acumulado (12 meses)	5.82%
RAP inicial	R$59.520 milhões
RAP total	R$824.298 milhões
Deságio	4.27%

Fonte: Dados da pesquisa.

O número de subestações médio por leilão foi de 1.04. Pelo gráfico apresentado na

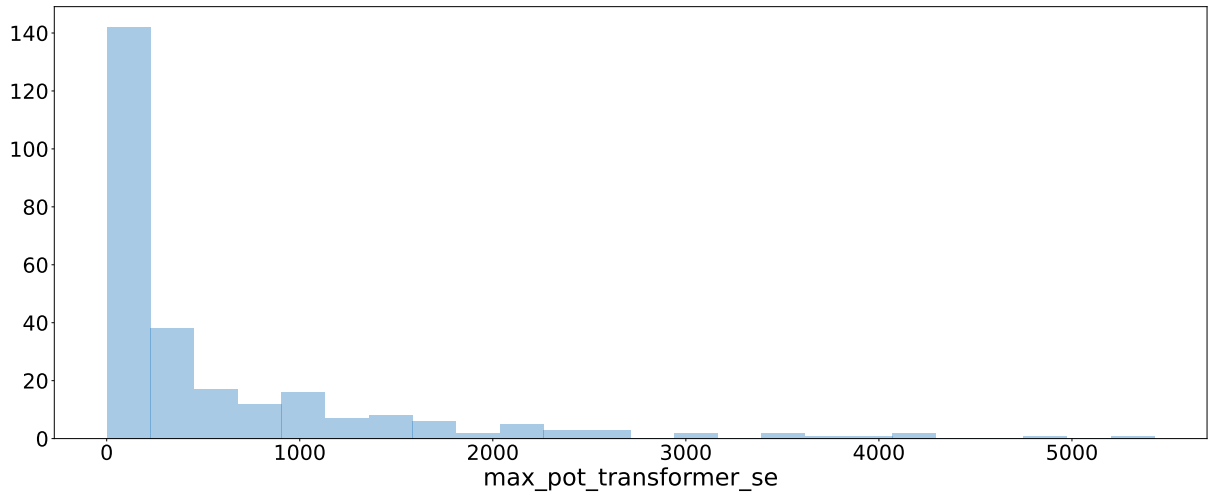
Figura 5, percebe-se que a grande maioria dos leilões contou com apenas uma subestação, sendo que a potência máxima mais frequente entre as subestações ofertadas foi de 300MVA, conforme mostra o gráfico da Figura 6.

Figura 5 – Gráficos de distribuição da variável “num_se”.



Fonte: Elaborada pelo autor.

Figura 6 – Gráficos de distribuição da variável “max_pot_transformer_se”.



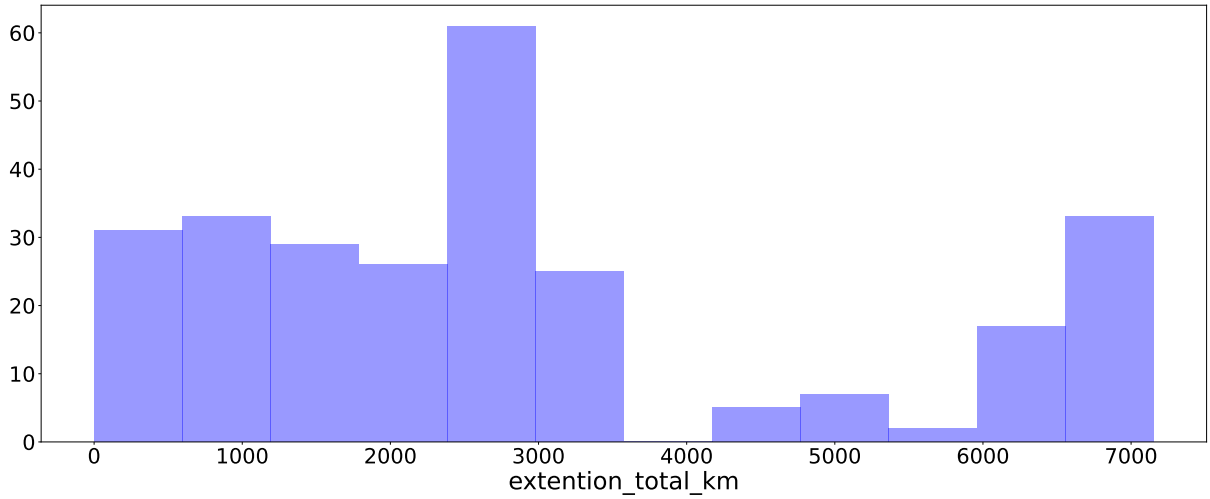
Fonte: Elaborada pelo autor.

A média da extensão total das linhas foi de 2.934km. O gráfico da Figura 7 mostra que essa extensão foi também a mais frequente entre os leilões. É interessante destacar que dentre as proponentes participantes do leilão existem algumas que apresentam uma natureza mais especializada para a construção de linhas de transmissão, enquanto que outras são voltadas mais para a construção de subestações. Dependendo do balanceamento entre a quantidade de subestações e a extensão das linhas de transmissão em um lote de leilão, pode-se favorecer

mais um determinado tipo de empresa do que outro e influenciar no deságio proposto por esses proponentes.

,

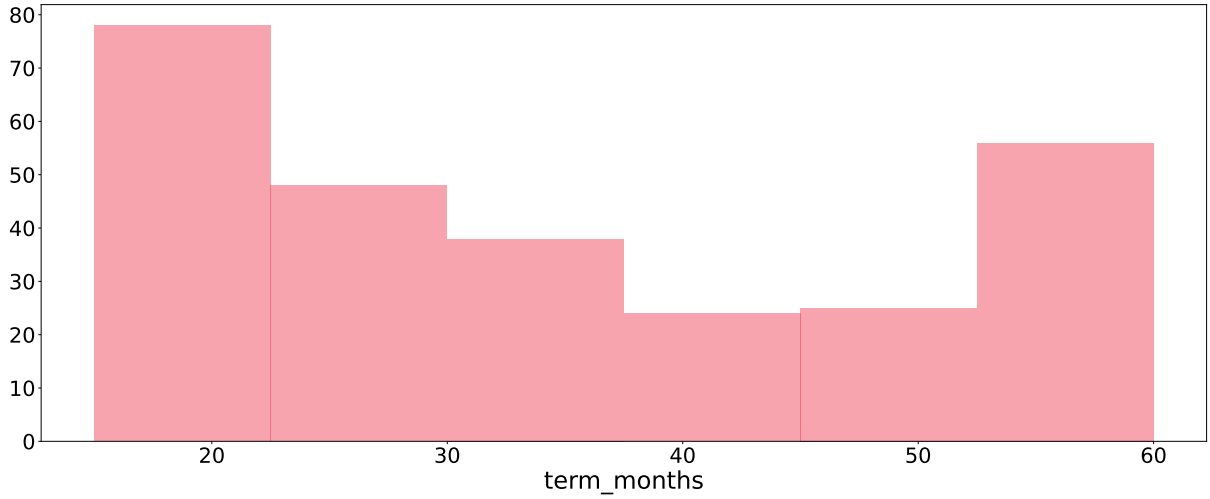
Figura 7 – Gráficos de distribuição da variável “extention_total_km”.



Fonte: Elaborada pelo autor.

Embora o prazo médio para entrega das obras seja 38 meses, o prazo mais frequente é de até 20 meses, conforme pode-se ver no gráfico da Figura 8.

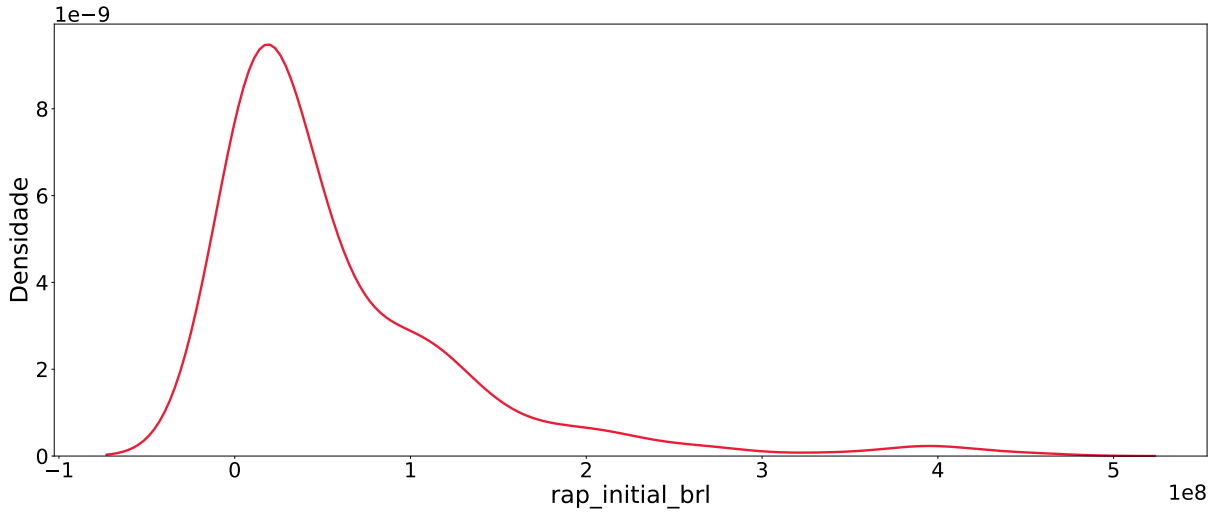
Figura 8 – Gráficos de distribuição da variável “term_months”.



Fonte: Elaborada pelo autor.

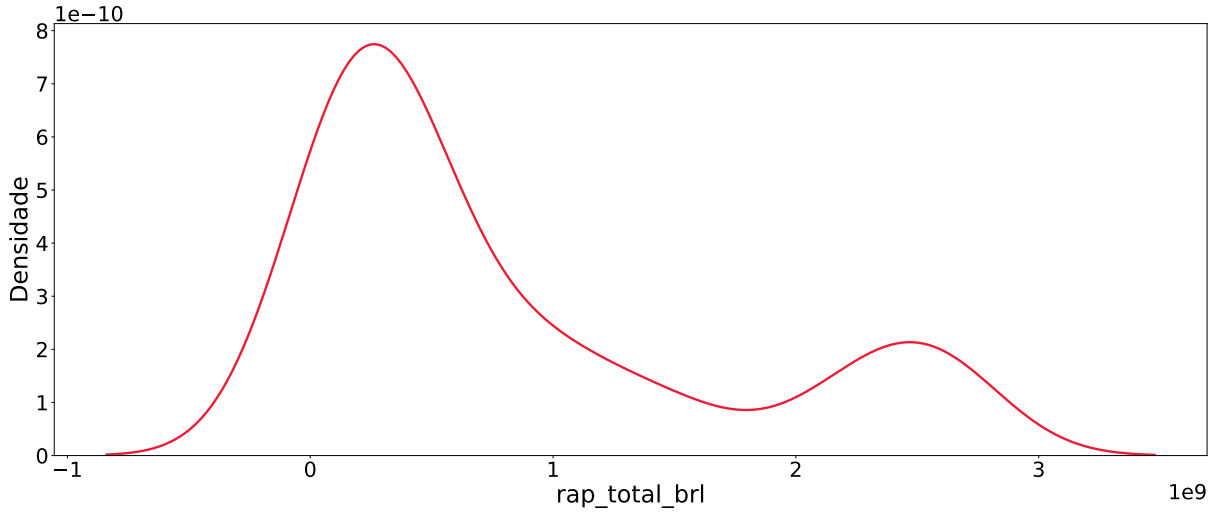
Com relação ao valor da receita anual inicialmente estabelecida pela ANEEL (rap_-initial_brl), o valor médio é de R$59.520 milhões, e sua distribuição é mostrada no gráfico da Figura 9. Já o valor médio da receita anual total (rap_total_brl) é de R$824.298 milhões, e sua distribuição aparece no gráfico da Figura 10.

Figura 9 – Gráficos de distribuição da variável “rap_initial_brl”.



Fonte: Elaborada pelo autor.

Figura 10 – Gráficos de distribuição da variável “rap_total_brl”.

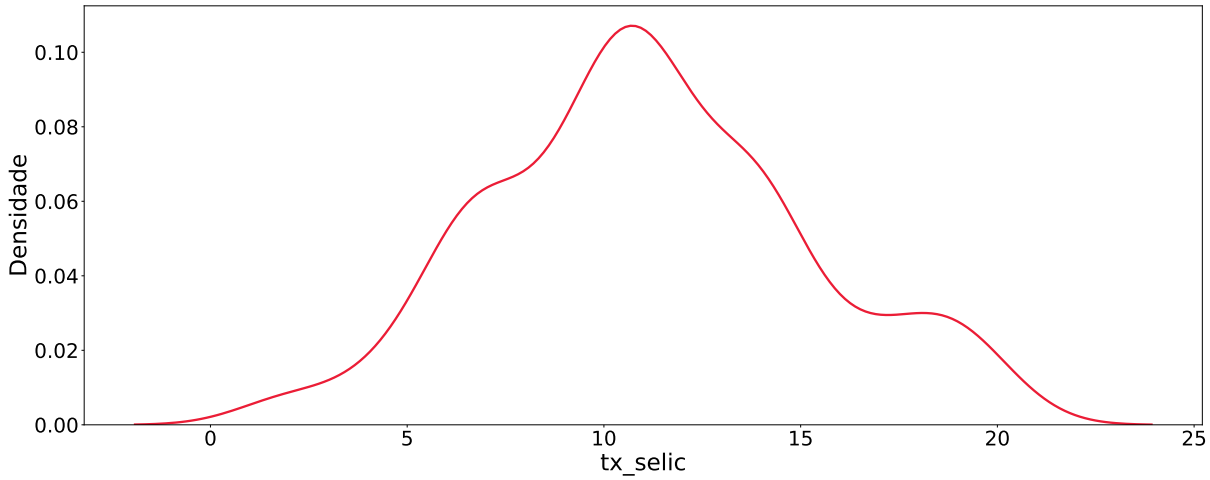


Fonte: Elaborada pelo autor.

As variáveis macroeconômicas Taxa Selic, Dólar e IPCA tiveram valores médios de 11%, R$2.70 e 5.82%, respectivamente. Os gráficos das Figuras 11, 12 e 13 mostram as distribuições dessas variáveis. Essas variáveis macroeconômicas foram incorporadas na base de dados de natureza mais técnica, pois acredita-se que por se tratar de investimentos, estas possam ser relevantes para ajudar no entendimento dos valores de deságio proposto para os lotes dos leilões. Interessante destacar que no momento da realização deste trabalho, tanto a Taxa Selic, quanto o IPCA apresentam os valores mais próximos das médias obtidas, 13.25% e 8.73% respectivamente. Enquanto que o dólar é o mais distante, cotado aproximadamente em R$5.15.

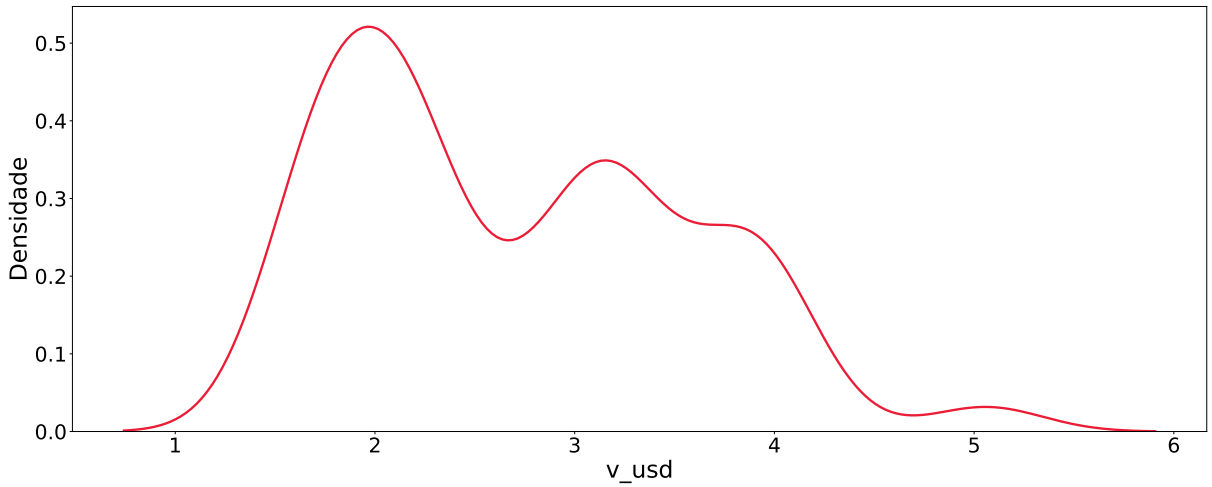
O gráfico da Figura 14 mostra em detalhes a distribuição da variável "linha de transmissão". É possível visualizar que a linha de transmissão com tensão de 230kV foi a que mais

Figura 11 – Gráficos de distribuição da variável “tx_selic”.



Fonte: Elaborada pelo autor.

Figura 12 – Gráficos de distribuição da variável “v_usd”.



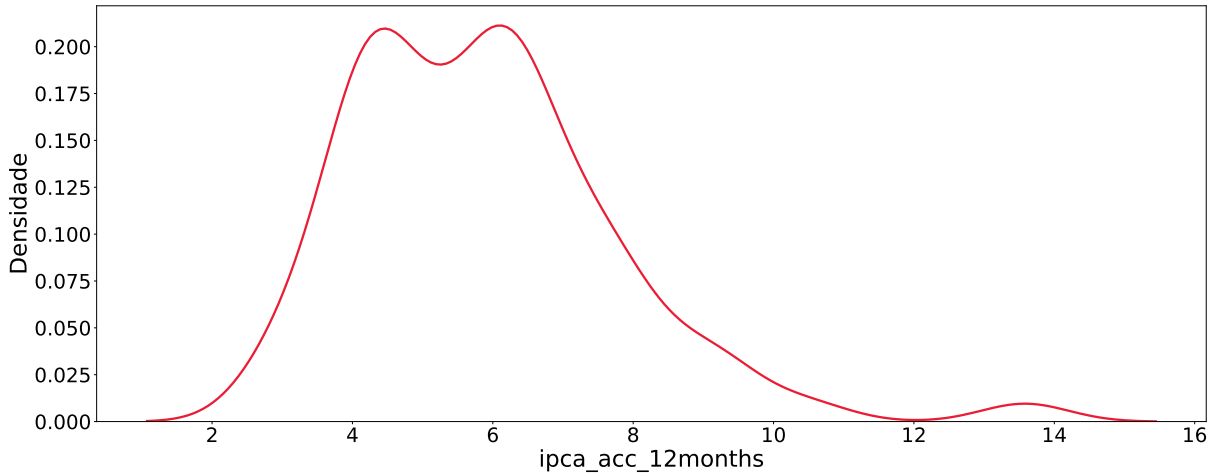
Fonte: Elaborada pelo autor.

apareceu entre os leilões, confirmando o entendimento de ser o nível mais aplicado no sistema de transmissão brasileiro.

A distribuição da variável “subestação” pode ser visualizada no gráfico da Figura 15. A subestação mais ofertada foi, assim como a linha de transmissão, a com nível de tensão de 230kV.

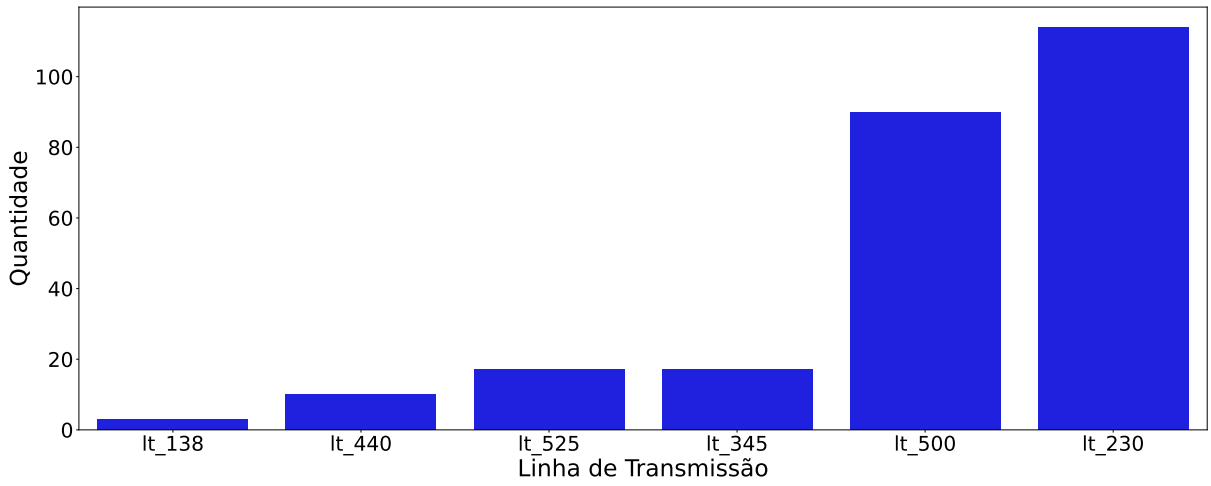
A Figura 16 mostra a matriz de correlação entre todas as variáveis presentes no estudo. As correlações positivas acima de 0.70 estão destacadas em vermelho. Não houve nenhuma correlação negativa abaixo de -0.70. Se enquadraram nesse quesito as variáveis lt525 e se525 (corr=0.73); lt440 e se440 (corr=0.84); extension_km e rap_initial_br (corr=0.77); extension_-total_km e rap_total_brl (corr=0.91); term_months e rap_total_brl (corr=0.82); e term_months e v_usd (corr=0.74).

Figura 13 – Gráficos de distribuição da variável “ipca_acc_12months”.



Fonte: Elaborada pelo autor.

Figura 14 – Gráfico de barras mostrando a distribuição da variável "linhas de transmissão".

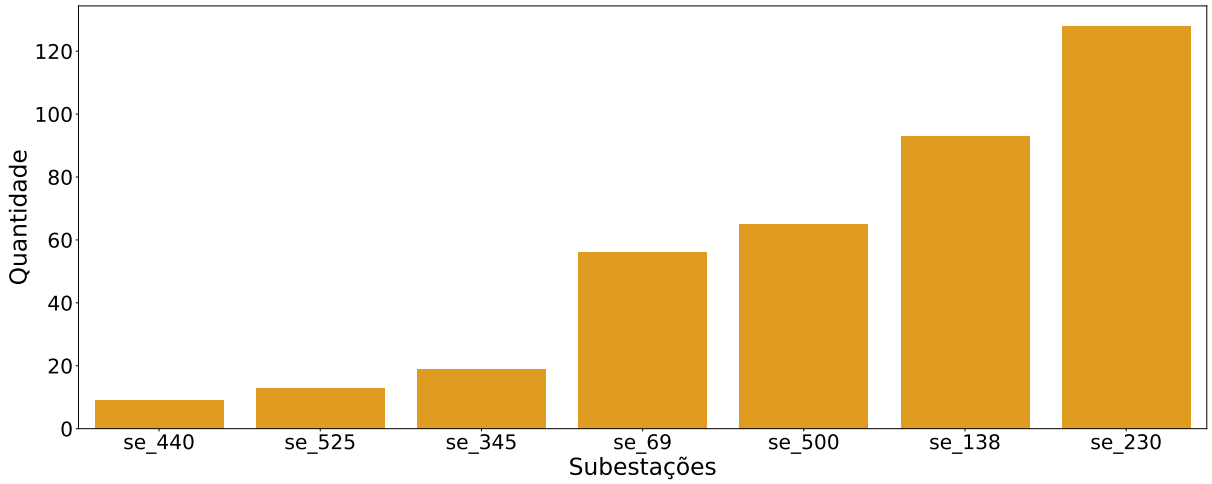


Fonte: Elaborada pelo autor.

A média do deságio nos leilões, a variável alvo “target_perc”, foi de 4.27%. O gráfico da Figura 17 mostra a distribuição dessa variável.

As Figuras 18, 19 e 20 mostram os gráficos de dispersão entre as variáveis macroeconômicas e a variável alvo. No três casos, parece não haver correlação alta entre essas variáveis. No caso da variável “ipca_acc_12months” pode haver uma correlação negativa, mas que seria considerada fraca.

Figura 15 – Gráfico de barras mostrando a distribuição da variável "subestações".



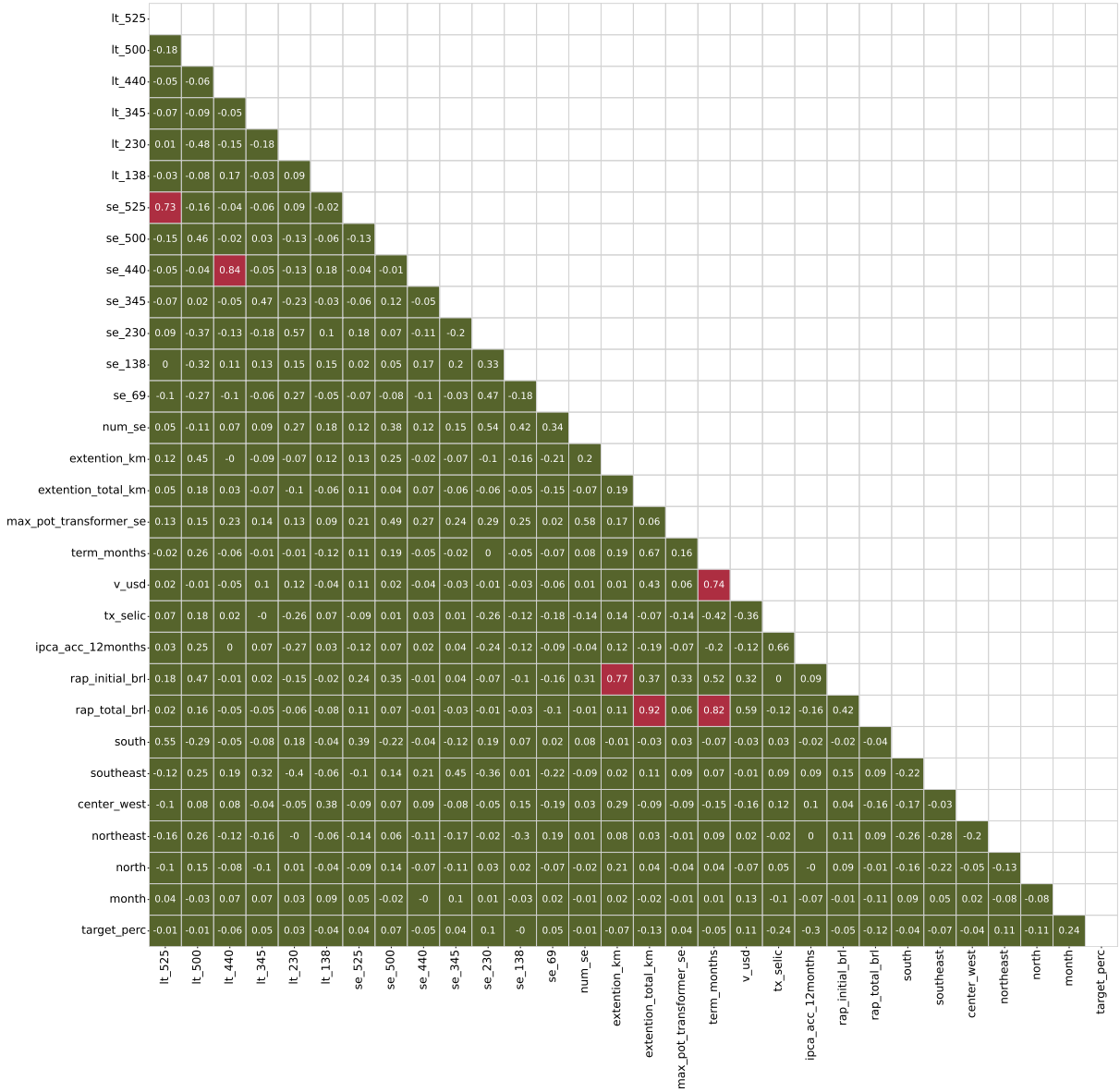
Fonte: Elaborada pelo autor.

3.1.4 Etapa 4: Seleção de Variáveis e Geração dos Conjuntos de Treinamento

A utilização de métodos de seleção de variáveis auxilia no entendimento de quais variáveis são mais importante para explicar a situação problemática para a qual se busca uma solução. Consequentemente, essas variáveis são as mais indicadas para serem utilizadas durante o treinamento e ajuste dos algoritmos de Aprendizado de Máquina. No entanto, os diversos métodos de seleção podem apresentar resultados diferenciados e, neste trabalho, buscou-se conhecer também o impacto desses métodos nos desempenhos dos algoritmos. Para isso, foram utilizados sete métodos para seleção de variáveis. Quatro desses métodos foram denominados, neste trabalho, métodos ou algoritmos pré-definidos, oferecidos por bibliotecas do Python. São eles: Boruta, Random Forest, Select from Model (SFM) e Recursive Feature Elimination (RFE). O método Boruta possui uma biblioteca própria, enquanto que para a implementação dos outros métodos foi utilizada a biblioteca scikit-learn. Todos esses métodos foram utilizados com os parâmetros padrão, conforme disponibilizados pelas bibliotecas. Os outros três métodos adotados neste estudo foram definidos pelo autor. São eles: a escolha aleatória de variáveis, a escolha arbitrária de variáveis pelo especialista (próprio autor) e a utilização de todas as variáveis presentes na base da dados.

Em geral, os métodos de seleção pré-definidos apresentam como resultado uma relação com todas as variáveis, porém listadas em ordem de importância. O método Boruta, entretanto, apresenta como resultado somente as variáveis que o algoritmo julga importante para explicação do problema. No caso do presente estudo, esse método considerou importante sete variáveis, dentre todas as variáveis presentes na base de dados. Dessa forma, foi convencionado que para todos os outros métodos, com exceção do método que utilizou todas as variáveis, seriam também consideradas as sete variáveis mais importantes.

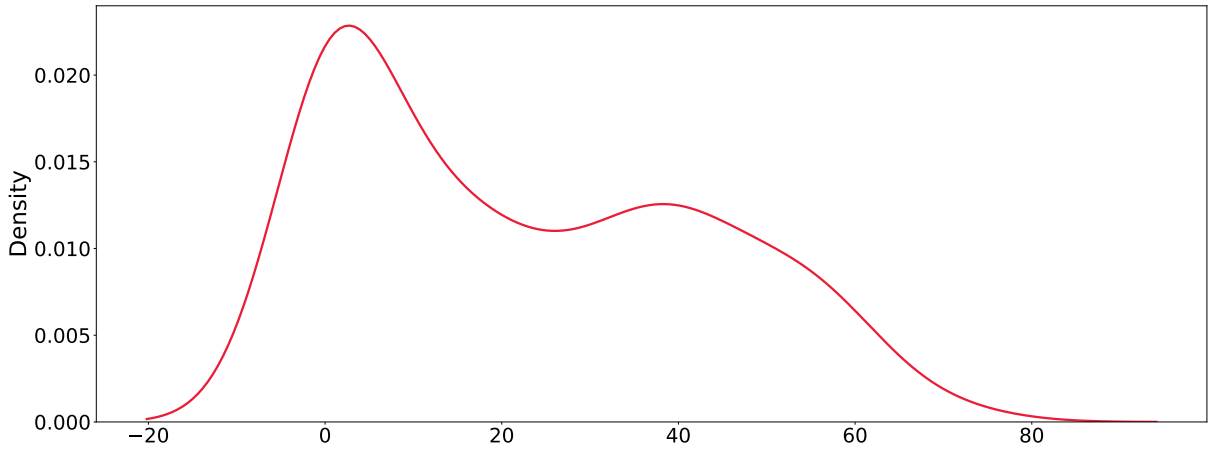
Figura 16 – Matriz de correlação entre as variáveis quantitativas abordadas neste estudo.



Fonte: Elaborada pelo autor.

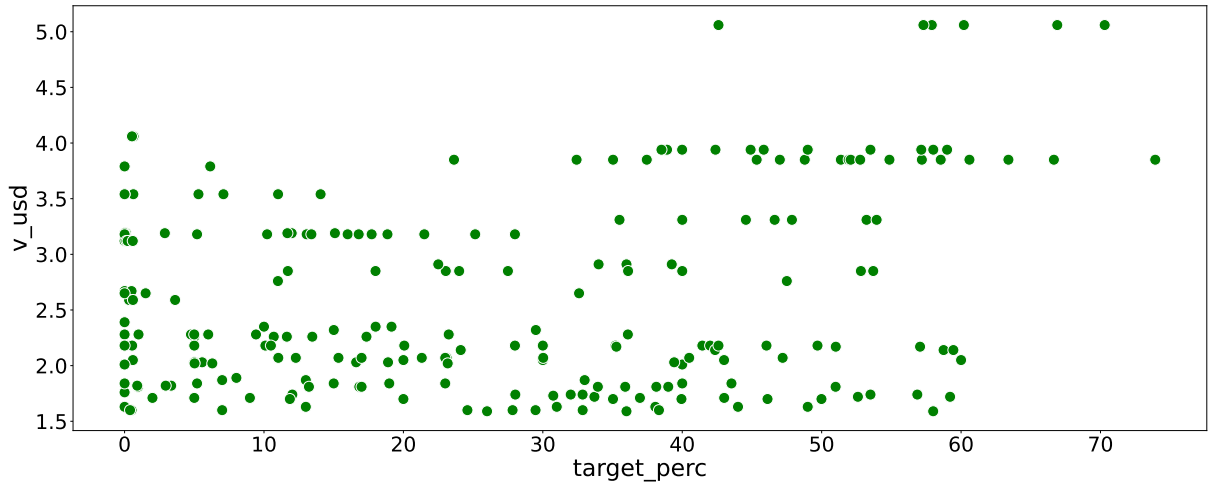
Outra atividade importante realizada nesta etapa foi a geração dos conjuntos de treinamentos, a partir das variáveis selecionadas pelos sete métodos de seleção. Na Etapa 2 ocorreu a divisão da base de dados, obtendo-se assim os conjuntos de treinamento e teste. A partir do conjunto de treinamento foram extraídos sete conjuntos também de treinamento, cada um composto somente pelas variáveis selecionadas por cada método de seleção. Assim, por exemplo, obteve-se um conjunto de treinamento com as variáveis selecionadas pelo método Boruta, outro conjunto com as variáveis selecionada pelo método RFE, outro pelas variáveis selecionadas pelo método SFM, e assim por diante. Esses conjuntos de treinamento foram utilizados na próxima etapa do trabalho.

Figura 17 – Gráfico de barras mostrando a distribuição da variável "subestações".



Fonte: Elaborada pelo autor.

Figura 18 – Gráfico de dispersão que mostra a relação entre a variável alvo “target_perc” e a variável “v_usd”.



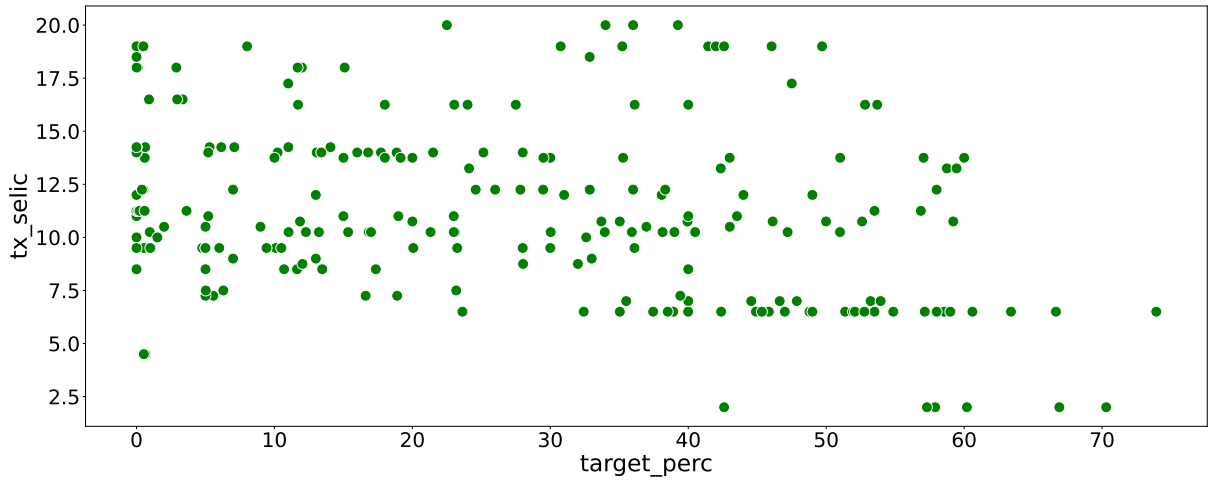
Fonte: Elaborada pelo autor.

3.1.5 Etapa 5: Treinamento e Avaliação dos Contextos

Nesta etapa, primeiramente, foram realizados os treinamentos dos quatro algoritmos abordados neste estudo: Árvore de Decisão, Random Forest, XGBoost e CatBoost. Esses algoritmos foram implementados com a linguagem Python por meio das bibliotecas Scikit-Learn (Árvores de decisão e Random Forest), XGBoost, e CatBoost.

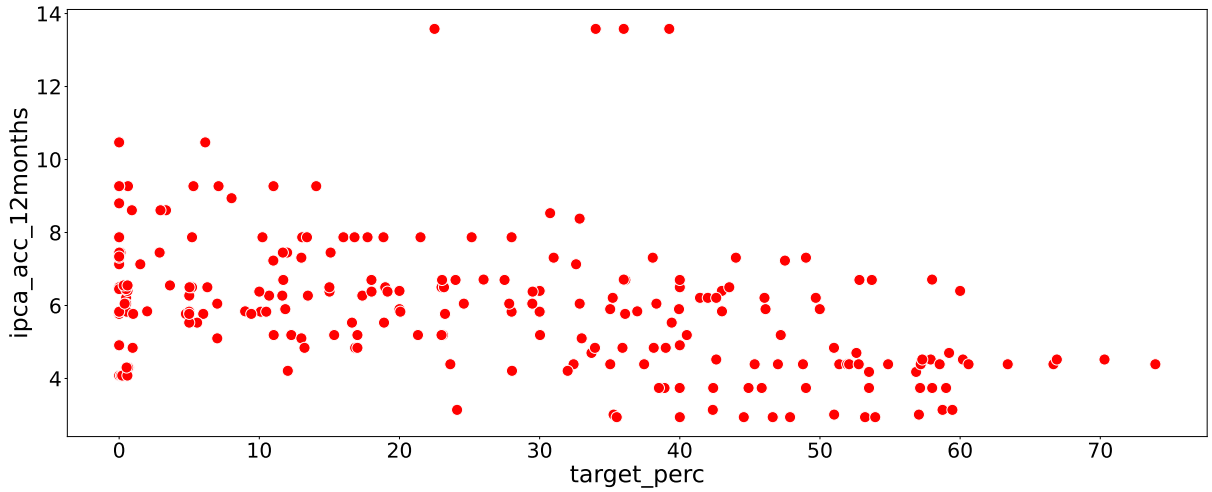
Os quatro algoritmos foram treinados com os sete conjuntos de treinamentos gerados na etapa anterior, referentes a cada método de seleção de variáveis. Cada par “algoritmo e conjunto de treinamento” foi denominado como "contexto", obtendo-se, portanto, 28 contextos no total (4 algoritmos X 7 conjuntos de treinamento). Por exemplo, o algoritmo Random Forest foi treinado com todos os sete conjuntos de treinamentos, gerados a partir dos métodos de seleção de

Figura 19 – Gráfico de dispersão que mostra a relação entre a variável alvo “target_perc” e a variável “tx_selic”.



Fonte: Elaborada pelo autor.

Figura 20 – Gráfico de dispersão que mostra a relação entre a variável alvo e a variável "IPCA acumulado nos últimos 12 meses".



Fonte: Elaborada pelo autor.

variáveis. O treinamento desse algoritmo com o conjunto de treinamento formado pela variáveis selecionadas pelo método SFM, por exemplo, forma um dos contextos; O algoritmo XGBoost também foi treinado com todos os sete conjuntos de treinamento. Quando treinado com o conjunto de treinamento gerado a partir do método Boruta, gerou um outro contexto, e assim por diante. Daqui em diante, para facilitar a escrita e compreensão, os contextos serão denominados pela relação “algoritmo/método de seleção”, como por exemplo: Random Forest/SFM (abreviado para RF/SFM), XGBoost/Boruta (abreviado para XG/BOR), CatBoost/RFE (abreviado para CB/RFE) etc.

Neste trabalho, a Árvore de Decisão teve o parâmetro “max_depth” ajustado para 3,

representando a profundidade da árvore. Os algoritmos Random Forest, XGBoost e CatBoost também tiveram o parâmetro "max_depth" ajustado para 3 e o parâmetro "n_estimators" ajustado para 400, representando o número de árvores geradas. Todos os outros demais parâmetros foram mantidos com os valores padrões.

O Quadro 3 mostra todos os 28 contextos "algoritmo/método de seleçã" gerados nesta etapa do trabalho. No decorrer do texto, bem como nas tabelas e gráficos, os contextos serão denominados da forma abreviada como aparecem na coluna "Contexto" deste Quadro.

Quadro 3 – Contextos gerados a partir dos algoritmos de Aprendizado de Máquina e dos métodos de seleção de variáveis.

Algoritmo	Método que originou os conjuntos de treinamento	Contexto
Árvore de Decisão	Boruta	AD/BOR
	Random Forest	AD/RF
	RFE	AD/RFE
	SFM	AD/SFM
	Aleatório	AD/Aleat
	Especialista	AD/Esp
	Todas as variáveis	AD/Todas
Random Forest	Boruta	RF/BOR
	Random Forest	RF/RF
	RFE	RF/RFE
	SFM	RF/SFM
	Aleatório	RF/Aleat
	Especialista	RF/Esp
	Todas as variáveis	RF/Todas
XGBoost	Boruta	XG/BOR
	Random Forest	XG/RF
	RFE	XG/RFE
	SFM	XG/SFM
	Aleatório	XG/Aleat
	Especialista	XG/Esp
	Todas as variáveis	XG/Todas
CatBoost	Boruta	CB/BOR
	Random Forest	CB/RF
	RFE	CB/RFE
	SFM	CB/SFM
	Aleatório	CB/Aleat
	Especialista	CB/Esp
	Todas as variáveis	CB/Todas

Fonte: Dados da pesquisa.

Para os treinamentos dos algoritmos, utilizou-se o procedimento conhecido como *Cross-Validation*, no qual o conjunto de treinamento é particionado em k partes, o modelo é treinado sobre $k-1$ partes e, então, testado na parte restante. Em seguida, as partes utilizadas para

treinamento e teste vão se alternando, obtendo-se novos resultado a cada alternância. Isso é feito até que todas as k partes tenham sido utilizadas como conjunto de teste. No caso deste trabalho, para o particionamento do conjunto de treinamento com *Cross-Validation*, adotou-se $k = 10$. Dessa forma, o conjunto de treinamento foi particionado em 10 partes, treinado com 9 partes e testado na parte restante. Como o conjunto de treinamento possuía 269 observações, o modelo foi treinado, no geral, com 242 observações e testado com 27 observações.

Para avaliação das predições, foram utilizadas duas métricas:

- a Raiz Quadrada do Erro Médio Quadrático, abreviada como *RMSE* (do inglês *Root Mean Square Error*); e
- o R-quadrado (R^2).

O *RMSE* é o desvio padrão dos resíduos, ou dos erros de predição, sendo uma medida de quão espalhado esses resíduos estão. Desta forma, pode ser uma interpretação de quão concentrado os dados estão em relação a curva de predição do modelo. A sua fórmula é dada por:

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}, \tag{3.1}$$

onde $\hat{y}_i$ são os valores preditos, y_i são os valores experimentais da variável resposta e n é o número de observações.

Dessa forma, o *RMSE* permite estimar o desvio padrão do erro para aquela observação em específico ao invés de um tipo de "erro total". Por dividir por n, a medição do erro permanece constante quando movimenta-se de uma observação de menor tamanho para uma maior. Assim, a acurácia aumenta com o número crescente de observações. Em outra palavras, *RMSE* é uma boa forma de indicar o quão distante o modelo é esperado de estar em sua próxima predição do valor observado, sendo uma boa métrica de acurácia para comparar erros de predição entre diferentes modelos. Se o valor do *RMSE* for baixo, isto geralmente significa que o modelo tem um bom desempenho em predizer os dados observados. Já um valor de *RMSE* alto, geralmente significa que o modelo não está sendo capaz de entender as importantes características implícitas nos dados (MOODY, 2019).

O R-quadrado, também conhecido como coeficiente de determinação, é uma medida estatística que indica o quão próximo os dados estão da linha de regressão ajustada. Pode ser interpretado como a porcentagem da variação da variável resposta que é explicada pelo modelo. Sua fórmula é dada por:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \tag{3.2}$$

onde $\hat{y}_i$ são os valores preditos, y_i são os valores experimentais da variável resposta, $\bar{y}$ é a média e n é o número de observações.

Dessa forma, o R-quadrado varia entre 0 e 1, onde 0 indica que o modelo não está sendo capaz de explicar em nada a variabilidade dos dados de resposta ao redor de sua média e 1 indica que o modelo explica toda a variabilidade dos dados de resposta. Ou seja, em geral, quanto maior o R-quadrado, melhor o modelo está se ajustando aos dados. Entretanto, o R-quadrado isoladamente não indica se um modelo é o mais adequado. Devendo ser avaliado em conjunto com os resíduos (FROST, 2020). Ambas as métricas discutidas anteriormente são amplamente utilizadas para avaliação de algoritmos de regressão.

Assim, cada contexto, adotando-se $k = 10$ para particionamento do Cross-Validation, foi treinado e testado 10 vezes, obtendo-se 10 medições de $RMSE$ e 10 medições de R^2. Esses grupos com 10 medições foram denominados amostras de medições ou somente amostras. No total, obteve-se 28 amostras de medições para o $RMSE$ e 28 amostras de R^2, cada amostra contendo 10 medições, lembrando que foram gerados 28 contextos.

Em seguida, foi realizado um teste de hipótese com a finalidade de verificar se alguma amostra de medições apresentava diferença significativa ($p - value \leq 0.05$) em relação a outras amostras, indicando assim se haveria um contexto mais adequado para tratamento do problema delineado neste trabalho. Por se tratar de comparações entre mais de duas amostras, foi utilizado o teste de hipótese de Friedman, implementado no trabalho por meio da biblioteca Scipy, do Python. O teste de Friedman avalia as seguintes hipóteses:

- Hipótese nula (H_0): Todas as amostras são provenientes de uma mesma distribuição;
- Hipótese alternativa (H_a): Pelo menos uma amostra é proveniente de uma distribuição diferente.

Portanto, o resultado desse método mostra se, entre todas as amostras, pelo menos uma delas apresenta diferença estatisticamente significativa em relação ao restante sem, entretanto, identificar qual amostra seria essa. Para se conhecer quais amostras apresentavam diferenças significativas, após o teste de hipótese de Friedman rejeitar a hipótese nula, foi utilizado o teste post-hoc de Nemenyi. Esse teste visa apontar quais amostras possuem diferenças significativas entre si através de desempenho em pares (NEMENYI, 1963) e foi implementado pela biblioteca Scikit-posthocs, também do Python. Assim, o teste de Nemenyi calcula os $p - values$ entre todas as amostras e evidencia aquelas que possuem diferença estatisticamente significativas.

Portanto, para avaliação e comparação dos diversos contextos, foram utilizadas as métricas $RMSE$ e R^2, bem como duas técnicas estatísticas, o teste de hipótese de Friedam e o teste post-hoc de Nemenyi. A partir dessas métricas e técnicas, foram elaborados três critérios que permitiram a escolha objetiva do contexto mais adequado para solucionar o problema levantado nesta pesquisa. Os três critérios são:

- Menor média da métrica $RMSE$;

- Média da métrica R^2 mais próxima de 1; e
- Maior número de contextos com diferenças estatisticamente significativas.

Para o modelo selecionado a partir dos critérios acima, foi também realizada um análise das variáveis que mais apareceram nas raízes das árvores e um levantamento das variáveis mais importantes apontadas por esse modelo nas 10 iterações do Cross-Validation.

3.1.6 Etapa 6: Análise do Modelo Selecionado e da Importância das Variáveis

O modelo selecionado na etapa anterior como sendo o mais adequado para realizar as predições sobre dados não vistos foi analisado de forma mais minuciosa nesta etapa. Primeiramente, foram identificadas e analisadas as raízes de todas as árvores geradas durante as 10 iterações do Cross-Validation. Essa análise mostrou quais variáveis mais apareceram nas raízes. Em seguida, foram identificadas e analisadas as variáveis classificadas como mais importantes pelo modelo selecionado. Isso também foi realizado para as 10 iterações do Cross-Validation. Finalizando essa etapa, foi realizada uma análise pelo especialista (autor) sobre as variáveis classificadas como mais importantes pelo modelo, justificando ou não essa classificação.

3.1.7 Etapa 7: Predição da variável alvo

A sétima e última etapa desta pesquisa compreendeu a predição da variável alvo "deságio" com dados ainda não vistos, ou seja, o conjunto de teste gerado na Etapa 2 e mantido isolado. O contexto utilizado para essa predição foi o que se mostrou mais adequado, conforme os critérios listados na etapa anterior. O algoritmo desse contexto será treinado novamente sobre todo o conjunto de treinamento. Nesse conjunto, foram mantidas as variáveis de acordo com o método de seleção do contexto selecionado.

Após o treinamento, o modelo foi utilizado para realizar predições sobre os dados não vistos, lembrando que esse conjunto de dados foi isolado para que não provocasse nenhum vazamento de dados, o que prejudicaria os resultados encontrados neste trabalho. Com as predições e os valores reais dos deságios nos leilões, foram calculadas as métricas *RMSE* e R^2. Essas métricas permitiram a avaliação final do contexto selecionado quando aplicado a novos dados.

Além disso, a partir da análise das variáveis mais importantes realizada na Etapa 7, foi escolhido treinar novamente esse modelo considerando apenas essas variáveis mais importantes, realizar novas predições, calcular as métricas *RMSE* e R^2 e comparar os resultados. Isso foi realizado em três situações diferentes: primeiramente com as três variáveis mais importantes, em seguida com as quatro mais importantes e por último com as seis mais importantes. Após as

predições, os resultados das métricas foram comparados entre si e com as predições anteriores, que utilizou todas as variáveis.

CAPÍTULO

4

RESULTADOS E DISCUSSÃO

Conforme descrito na Seção 3.1, o Processo Experimental deste trabalho envolveu sete etapas, listadas novamente abaixo:

- Etapa 1: Coleta dos dados em fontes primárias;
- Etapa 2: Preparação dos dados coletados;
- Etapa 3: Análise exploratória dos dados;
- Etapa 4: Seleção de variáveis e geração dos conjuntos de treinamento;
- Etapa 5: Treinamento, avaliação dos contextos e seleção do modelo;
- Etapa 6: Análise do modelo selecionado e da Importância das Variáveis; e
- Etapa 7: Predição da variável alvo "deságio".

Os resultados obtidos nas Etapas 1, 2 e 3 foram devidamente apresentados juntos com a descrição do Processo Experimental, no Capítulo 3. Neste capítulo serão apresentados e analisados os resultados obtidos nas Etapas 4, 5, 6 e 7.

4.1 Análise dos Resultados Obtidos na Etapa 4

Na Etapa 4 foram implementados os sete métodos de seleção de variáveis. Esses métodos foram: Boruta, Random Forest, Recursive Feature Elimination, Select From Model, Aleatório, Especialista e Todas as Variáveis. O Quadro 4 mostra os sete métodos de seleção de variáveis e as variáveis selecionadas.

Interessante notar que os quatro métodos de seleção pré-definidos selecionaram exatamente as mesmas variáveis, mudando somente a ordem de importância delas. Todas as variáveis

Quadro 4 – Métodos de seleção de variáveis utilizados neste estudo e as variáveis selecionadas. As variáveis aparecem em ordem de importância.

Método	Variáveis Selecionadas
Boruta	extention_km, term_months, v_usd, tx_selic, ipca_acc_12months, rap_initial_brl, rap_total_brl
Random Forest	rap_total_brl, tx_selic, term_months, ipca_acc_12months, rap_initial_brl, extention_km, v_usd
SFM	extention_km, term_months, v_usd, tx_selic, ipca_acc_12months, rap_initial_brl, rap_total_brl
RFE	extention_km, rap_total_brl, rap_initial_brl, ipca_acc_12months, tx_selic, v_usd, term_months
Aleatório	south, extention_total_km, se_69, se_230, se_525, lt_525, lt_230
Especialista	tx_selic, term_months, v_usd, ipca_acc_12months, rap_initial_brl, rap_total_brl, num_se
Todas	Todas as variáveis presentes no conjunto de treinamento

Fonte: Dados da pesquisa.

selecionadas pelos métodos pré-definidos também foram selecionadas pelo especialista, exceto a variável "extention_km". No lugar dessa, o especialista selecionou a variável "num_se" pois acreditou-se que poderia ser mais representativa para os modelos na predição da variável alvo por se tratar da quantidade de subestações envolvidas em um determinado lote e estar relacionada a uma complexidade de maior nível, visto o esforço exigido para a construção destas subestações.

As demais variáveis como "tx_selic", "v_usd" e "ipca_acc_12months" foram escolhidas pelo especialista justamente por refletirem as condições macroeconômicas que devem influenciar nas decisões de investimentos na infraestrutura energética do país. A variável "term_months" também foi considerada importante pelo especialista, pois representa o tempo para a realização do empreendimento e, consequentemente, o tempo para a realização do investimento sendo que um prazo mais curto ou mais longo pode influenciar no deságio apresentado por uma proponente. Já as variáveis "rap_initial_brl" e "rap_total_br", por representarem em termos financeiros o quanto o órgão governamental está disposto a pagar pela realização de um lote e do leilão como um todo, acredita-se que estão relacionadas ao tamanho e complexidade esperada para as obras e devem explicar. A variável "num_se" foi explicada no parágrafo anterior. Nenhuma das variáveis selecionadas pelo método aleatório foi selecionada pelos métodos pré-definidos ou pelo especialista.

Esses resultados mostram que houve bastante consistência entre as seleções realizadas pelos métodos pré-definidos e entre esses métodos e a seleção realizada pelo especialista. Somente o método Aleatório divergiu consideravelmente. Esse método servirá para mostrar se a seleção de variáveis é algo que traz algum ganho para o problema em questão ou, escolhendo-se aleatoriamente as variáveis, obtém-se resultados semelhantes. Isso ficará mais evidente com a análise dos resultados realizada na Etapa 5.

4.2 Análise dos Resultados Obtidos na Etapa 5

Na Etapa 5, todos os 28 contextos formados pelos 4 algoritmos e 7 conjuntos de treinamentos (originados a partir dos sete métodos de seleção de variáveis) foram treinados utilizando a técnica Cross-Validation com 10 partições ($k = 10$), conforme descrito no Capítulo 3. Conforme também já detalhado anteriormente, para cada contexto, obteve-se 10 medições de *RMSE* e 10 medições de R^2. A Tabela 2 mostra as médias das 10 medições dessas duas métricas para todos os 28 contextos.

Tabela 2 – Médias das métricas *RMSE* e R^2 para cada um dos 28 contextos.

Modelo/Método de seleção	Média *RMSE*	Média R^2
CB/Todas	12.388099	0.597434
XG/Todas	12.893714	0.561547
CB/RF	13.020520	0.550836
CB/RFE	13.053959	0.548466
CB/BOR	13.193672	0.537675
CB/SFM	13.193672	0.537675
CB/Esp	13.327393	0.517955
RF/Todas	13.599014	0.528807
RF/SFM	13.652632	0.523634
RF/BOR	13.652632	0.523634
RF/RFE	13.653555	0.523518
RF/RF	13.655216	0.523402
RF/Esp	13.688905	0.521728
XG/RFE	13.834243	0.491534
XG/RF	13.854223	0.488698
XG/BOR	13.894706	0.487109
XG/SFM	13.894706	0.487109
CB/Aleat	14.728629	0.447485
AD/RF	14.833847	0.436450
AD/BOR	14.833847	0.436450
AD/SFM	14.833847	0.436450
AD/RFE	14.833847	0.436450
AD/Esp	14.932312	0.430329
AD/Todas	15.129876	0.418060
XG/Esp	15.447098	0.355572
XG/Aleat	15.665852	0.347622
RF/Aleat	18.062237	0.186667
AD/Aleat	18.767291	0.121724

Fonte: Dados da pesquisa.

Na tabela, os contextos foram listados em ordem decrescente de desempenho de acordo com a média do *RMSE*. Em vários casos, a ordem de desempenho quando elencada pela média do *RMSE* é diferente da ordem de desempenho quando elencada pela média do R^2.

Primeiramente, será feita uma análise utilizando-se somente as médias das métricas

RMSE e R^2. De forma geral, percebe-se que existe uma certa consistência na ordem de desempenho. Essa consistência se mantém tanto pelas médias do *RMSE* quanto do R^2. Observando-se a Tabela 2, percebe-se que os modelos estão relativamente agrupados. O modelo CatBoost aparece predominante entre as primeiras posições, seguido pelo modelo Random Forest. Logo em seguida aparecem os contextos com o modelo XGBoost e, por último, os contextos com as Árvores de Decisão. Na parte de baixo da tabela, com os piores desempenho de acordo com as médias dessas duas métricas, estão contextos que utilizaram a seleção aleatória. O contexto com CatBoost que não aparece nas primeiras posições é justamente o que utilizou a seleção aleatória (aparece em 18^o lugar). Porém, entre os contextos que utilizaram a seleção aleatória, o que empregou o modelo CatBoost obteve melhor desempenho.

Ainda pela análise das médias, é possível constatar que o contexto $CB/Todas$ foi o que obteve melhor desempenho, tanto pela média do *RMSE* (12.38), como também pela média do R^2 (0.59). Em segundo lugar, também pelas médias das duas métricas, está o contexto $XG/Todas$, com *RMSE* médio de 12.89 e R^2 médio de 0.56. E em terceiro lugar, também pelas médias das duas métricas, está o contexto CB/RF, com *RMSE* médio de 13.02 e R^2 médio de 0.55. Desconsiderando a seleção aleatória, os piores desempenhos foram obtidos pelos contextos que utilizaram Árvores de Decisão. Pior do que as Árvores, somente os contextos que utilizaram a seleção aleatória. Inclusive, o pior desempenho entre todos os contextos foi obtido pelo contexto que utilizou Árvore de Decisão com a seleção aleatória.

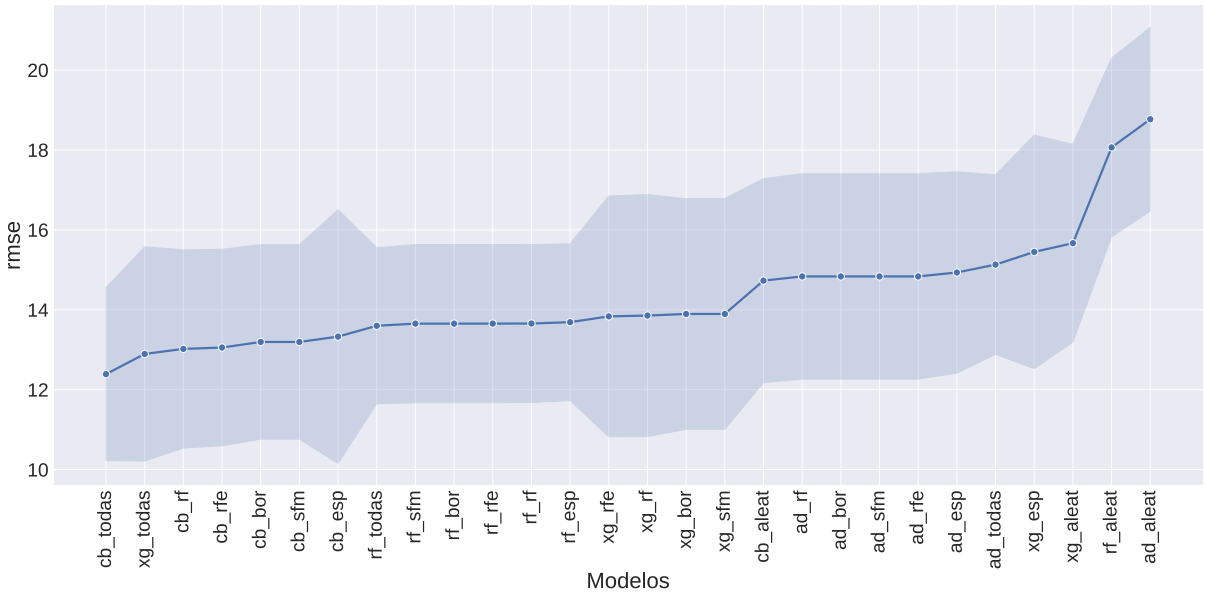
Os contextos que utilizaram os modelos CatBoost, Random Forest e XGBoost obtiveram melhor desempenho quando utilizaram a seleção com todas as variáveis. Isso só não ocorreu com os contextos que utilizaram Árvores de Decisão. O melhor desempenho desse modelo foi com a seleção de variáveis feita pelo método Random Forest. Com relação aos outros métodos de seleção, não foi possível observar nenhum padrão. Algo que parece não seguir uma determinada lógica é o fato do modelo XGBoost aparecer na segunda posição com a seleção com todas as variáveis, e os contextos que utilizaram esse modelo mas com as outras seleções, aparecer predominantemente com desempenhos inferiores aos contextos que utilizaram o modelo Random Forest, e não logo na sequência, após os contextos com CatBoost.

Outro ponto que convém salientar é que, se por um lado, tem-se os modelos se repetindo em determinadas posições de desempenho, como o caso do modelo CatBoost que aparece predominantemente nas primeiras posições, e o caso do modelo Árvore de Decisão, que aparece predominantemente nas últimas posições, por outro lado, percebe-se que os métodos de seleção de variáveis, com exceção da seleção aleatória, estão mais “dispersos” entre as posições de desempenho. Por exemplo, tem-se o método de seleção RFE aparecendo entre os contextos com melhores desempenhos, mas também esse método aparece entre os piores desempenhos. Isso sugere que o desempenho do contexto depende mais do modelo utilizado do que necessariamente do método de seleção.

A Figura 21 mostra o gráfico com as medições da métrica *RMSE* realizadas para os 28

contextos. A linha sólida no gráfico representa a média das 10 medições coletadas durante o treinamento dos contextos. Enquanto, que a área sombreada mostra o desvio padrão. Os contextos estão apresentados em ordem decrescente de desempenho. O contexto $CB/Todas$, que apresentou a menor média de $RMSE$, conforme discutido acima, aparece na extremidade esquerda do eixo x do gráfico, enquanto que o contexto $AD/Aleat$, que apresentou pior desempenho com essa métrica, aparece na extremidade direita desse eixo.

Figura 21 – Gráfico de linha dos resultados da métrica $RMSE$ para os 28 contextos.



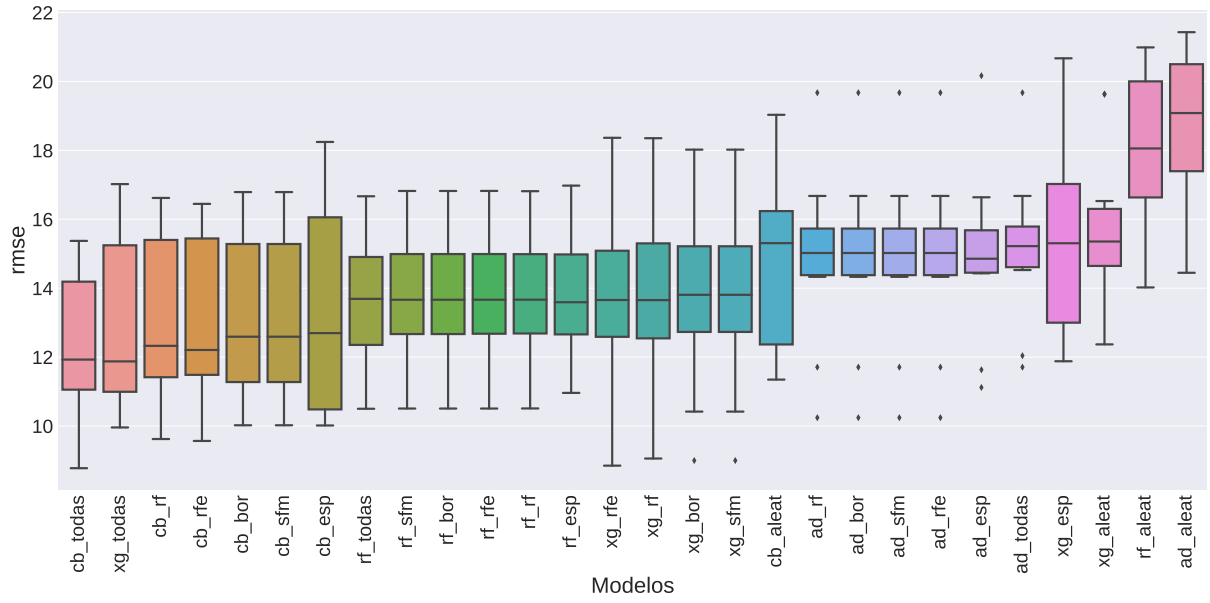
Fonte: Elaborada pelo autor.

Para se ter uma ideia da variabilidade dos valores obtidos em cada amostra de medições, foi também construído o gráfico de caixa para a métrica $RMSE$, apresentado na Figura 22. Pelo gráfico, é possível observar que a variabilidade entre os contextos é bastante semelhante. Por exemplo, os contextos que utilizaram Árvores de Decisão parecem ter variabilidade bem similar. O mesmo ocorre com os contextos que empregaram o modelo Random Forest.

A Figura 23 mostra o gráfico de linha com os resultados obtidos para a métrica R^2. Da mesma forma que no gráfico de linha da métrica $RMSE$, a linha sólida representa a média das 10 medições realizadas com essa métrica para cada contexto, e a área sombreada mostra o desvio-padrão. Os contextos estão listados em ordem decrescente de desempenho, de acordo com a média do R^2. A Figura 24 mostra o gráfico de caixas das medições realizadas com a métrica R^2. Assim como para a métrica $RMSE$, é possível observar que a variabilidade entre os contextos, de modo geral, é bastante similar. Entretanto, é notável que o primeiro contexto com maior valor de R^2, $CB/Todas$, e os dois contextos com menores valores, $RF/Aleat$ e $AD/Aleat$, respectivamente, apresentam a menor variabilidade, podendo indicar uma maior consistência dos resultados da métrica R^2 para estes contextos.

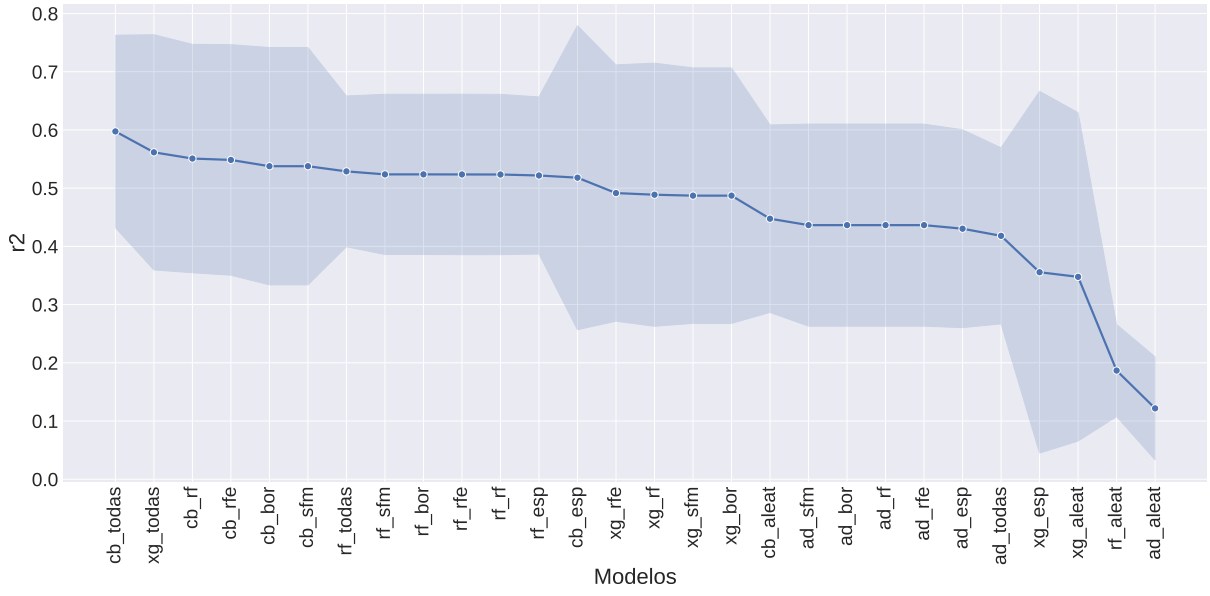
Com a análise anterior, baseada nos resultados das médias das métricas $RMSE$ e R^2,

Figura 22 – Gráficos de caixa dos resultados da métrica *RMSE* para os 28 contextos.



Fonte: Elaborada pelo autor.

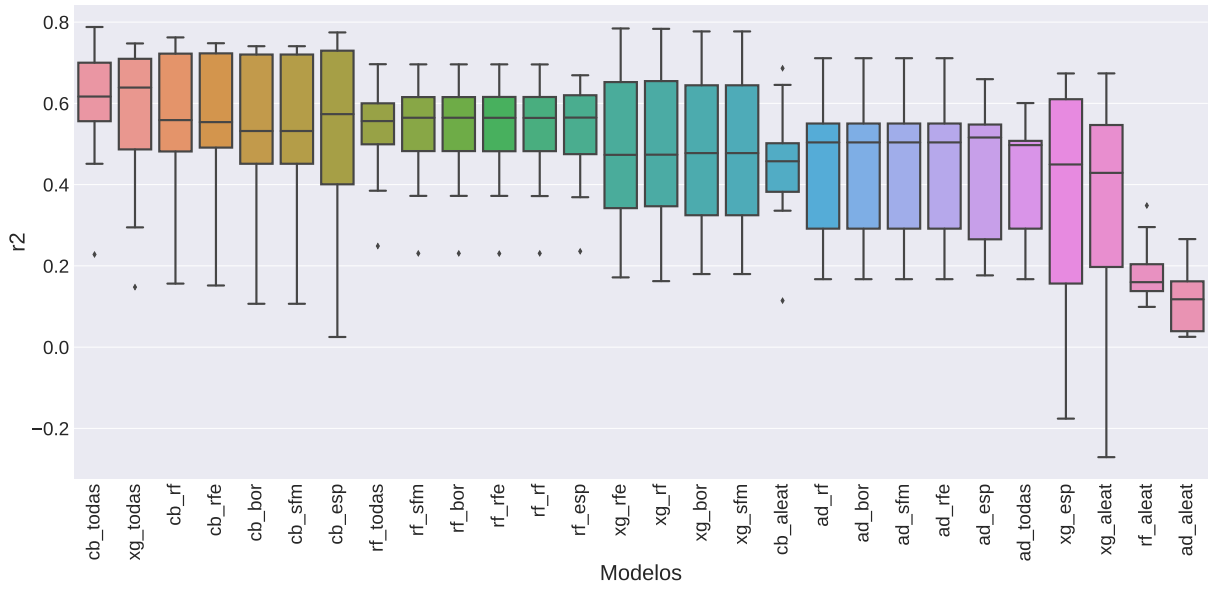
Figura 23 – Gráficos de linha dos resultados da métrica R^2 para os 28 contextos.



Fonte: Elaborada pelo autor.

não é possível avaliar se a diferença entre os diversos desempenhos apontados até aqui é estatisticamente significativa. Em outras palavras, não é possível afirmar que um determinado contexto apresenta desempenho superior a outro, ou se os resultados obtidos são simplesmente valores aleatórios. Isso significa que, caso sejam aleatórios, em outras rodadas de treinamentos, os contextos poderiam apresentar desempenhos diferentes daqueles observados até aqui, e a ordem de desempenho poderia ser bastante diferente.

Figura 24 – Gráficos de caixa dos resultados da métrica R^2 para os 28 contextos.



Fonte: Elaborada pelo autor.

Na busca por uma melhor compreensão dessa questão, foi realizado um teste de hipótese para verificar se há (ou não) diferença significativa entre as amostras de medições. Para tanto, foi utilizado o teste de hipótese de Friedman. Conforme descrito no Capítulo 3, esse método é utilizado quando se possui mais de duas amostras. Esse é o caso do presente estudo, onde foram obtidas 28 amostras de medições, conforme já descrito anteriormente. O teste de Friedman visa detectar se, entre essas diversas amostras, há pelo menos uma que apresente diferença estatisticamente significativa (ou seja, $p \leq 0.05$) em relação às amostras restantes. Esse teste foi realizado para ambas as métricas (*RMSE* e R^2) e os resultados obtidos são apresentados na Tabela 3.

Tabela 3 – $P-values$ para os testes de hipóteses de Friedman realizados para as 28 amostras de medições obtidas a partir das métricas *RMSE* e R^2.

Métrica	p-value
RMSE	$4.4821e^{-35}$
R^2	$3.1811e^{-37}$

Fonte: Dados da pesquisa.

Nota-se que os resultados dos $p-values$ para ambas as métricas é ≤ 0.05. Portanto, o teste mostra que pelo menos uma amostra de medição, dentre as 28, possui diferença significativa em relação a outros. Em outras palavras, de fato, pode-se afirmar que pelo menos um contexto apresenta desempenho significativamente diferente, podendo ser superior ou inferior em relação a outros contextos. Dessa forma, a hipótese nula desse teste, a qual afirma que todas as amostras são provenientes de uma mesma distribuição, foi rejeitada. Porém, embora o teste de Friedman indica que há pelo menos uma amostra com diferença significativa em relação a outras, o teste

não permite conhecer qual ou quais são essas amostras, e em relação a qual ou quais outras amostras essa diferença existe.

Uma vez que os $p-values$ encontrados nos testes de hipóteses de Friedman realizados sobre as amostras de medições foram estatisticamente significativas ($p-values \leq 0.05$), foi necessário realizar o teste post-hoc de Nemenyi para identificação das amostras estatisticamente diferentes. O teste mostrou que o contexto $CB/Todas$ foi o que apresentou diferença significativa com relação ao maior número de outros contextos. No caso, pode-se afirmar que esse contexto possui desempenho superior em relação a 11 outros contextos. Esses contextos estão relacionados na Tabela 4.

Tabela 4 – Diferenças estatisticamente significativas entre o contexto $CB/Todas$ e 11 outros contextos.

Contexto	Contexto com diferença significativa	$p-value$
cb_todas	ad_todas ad_bor ad_esp xg_esp ad_sfm ad_aleat rf_aleat xg_aleat cb_aleat ad_rf ad_rfe	≤ 0.05

Fonte: Dados da pesquisa.

Alguns contextos apresentaram diferença estatisticamente significativa em relação a somente dois outros contextos; e outros 3 contextos apresentaram diferença significativa em relação a somente um contexto. A Tabela 5 mostra quais são esses contextos.

Nesses casos, pode-se afirmar que os contextos mostrados na primeira linha da Tabela 5, apresentam desempenho superior ou inferior (depende do valor da média do $RMSE$) em relação aos contextos $AD/Aleat$ e $RF/Aleat$. Assim, também, pode-se afirmar que os três contextos da segunda linha possuem desempenho superior ou inferior em relação ao contexto $AD/Aleat$.

Dentre os 11 contextos que apresentam diferença significativa em relação ao contexto $CB/Todas$, listados na Tabela 4, 9 apresentaram diferença somente em relação ao contexto $CB/Todas$, e a mais nenhum outro. Esses contextos estão apresentados na Tabela 6

O teste de Nemenyi mostrou, ainda, que dois contextos com seleção aleatória apresentaram diferença significativa com relação a um grupo considerável de outros contextos. Esses contextos e relações estão apresentados na Tabela 7

Esses contextos são os que apresentaram piores desempenhos quando consideradas somente as médias, tanto do $RMSE$ quanto do R^2. Portanto, pode-se afirmar que os contextos

Tabela 5 – Contextos com diferenças estatisticamente significativas.

Contexto	Contexto com diferença significativa	$p-value$
xg_todas cb_rf cb_rfe cb_bor cb_sfm cb_esp rf_todas rf_sfm rf_bor rf_rfe rf_rf rf_esp xg_rfe	ad_aleat rf_aleat	≤ 0.05
xg_rf xg_bor xg_sfm	ad_aleat	≤ 0.05

Fonte: Dados da pesquisa.

Tabela 6 – Diferenças estatisticamente significativa entre o contexto $CB/Todas$ e 11 outros contextos.

Contexto	Contexto com diferença significativa	$p-value$
cb_aleat ad_rf ad_bor ad_sfm ad_rfe ad_esp ad_todas xg_esp xg_aleat	cb_todas	≤ 0.05

Fonte: Dados da pesquisa.

$RF/Aleat$ e $AD/Aleat$ possuem desempenho inferior aos diversos contextos mostrados nesta tabela; e é possível inferir que esses contextos seriam os menos adequados para solucionar o problema tratado neste estudo.

Conforme mencionado no Capítulo 3, neste trabalho, foram utilizados três critérios objetivos para escolha do contexto mais adequado para solução do problema tratado nesta pesquisa. Os três critérios foram:

- Menor média da métrica $RMSE$;
- Média da métrica R^2 mais próxima de 1; e
- Maior número de contextos com diferenças estatisticamente significativas.

Tabela 7 – Diferenças estatisticamente significantes entre os diversos contextos e seus respectivos $p-values$.

Contexto	Contexto com diferença significativa	$p-value$
rf_aleat	rf_todas xg_todas cb_todas rf_bor cb_bor rf_esp cb_esp rf_sfm cb_sfm rf_rf cb_rf rf_rfe xg_rfe cb_rfe	≤ 0.05
ad_aleat	rf_todas xg_todas cb_todas rf_bor xg_bor cb_bor rf_esp cb_esp rf_sfm xg_sfm cb_sfm rf_rf xg_rf cb_rf rf_rfe xg_rfe cb_rfe	≤ 0.05

Fonte: Dados da pesquisa.

Assim, o contexto que melhor atendeu a esses critérios foi o $CB/Todas$. Entre todos os contextos, esse foi o que apresentou a menor média do $RMSE$ (12.388) e a média do R^2 mais próxima de 1 (0.597). Além disso, esse foi o contexto que apresentou diferença estatisticamente significativa em relação ao maior número de outros contextos. Pode-se afirmar que o contexto $CB/Todas$ obteve desempenho superior em relação a 11 outros contextos. Portanto, esse contexto foi selecionado como sendo o mais adequado para realizar as predições sobre o conjunto de dados não vistos (conjunto de teste).

4.3 Análise dos Resultados Obtidos na Etapa 6

Na Etapa 6, foram identificadas e analisadas as frequências com que as variáveis apareceram nas raízes das árvores, considerando-se as 10 iterações do Cross-validation. Como o modelo selecionado foi um ensemble (CatBoost) com o parâmetro "n_estimators" ajustado em 400, para cada iteração do Cross-validation foram geradas 400 árvores. Portanto, ao final, foram geradas 4.000 árvores. A Tabela 8 mostra as 10 variáveis que mais apareceram nas raízes dessas 4.000 árvores e o número de vezes que isso ocorreu, considerando-se as 10 iterações do Cross-validation. Na Figura 25 é possível visualizar a frequência com que todas as variáveis apareceram nas raízes das árvores, por meio de um gráfico de barras.

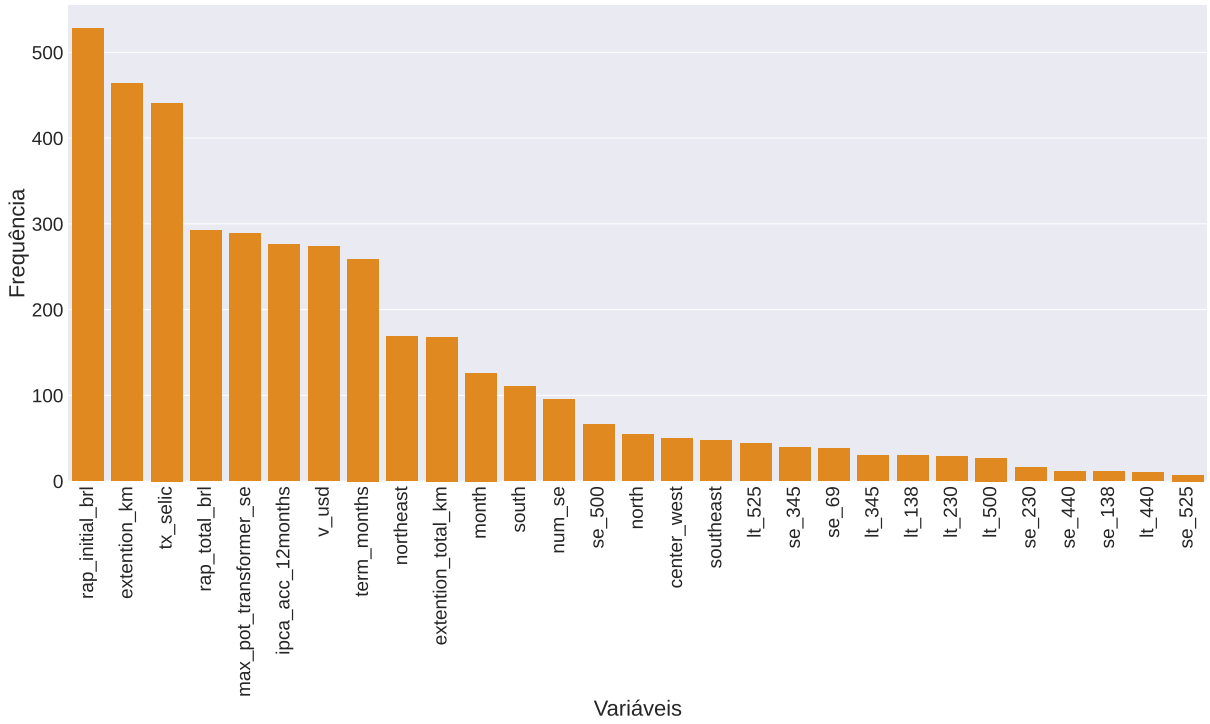
Tabela 8 – As 10 variáveis que mais apareceram nas raízes das árvores geradas durante as 10 iterações do Cross-validation.

Variável	Quantidade
rap_initial_brl	528
extention_km	464
tx_selic	441
rap_total_brl	292
max_pot_transformer_se	289
ipca_acc_12months	276
v_usd	274
term_months	259
northeast	169
extention_total_km	168

Fonte: Dados da pesquisa.

As variáveis listadas na Tabela 8 são as que mais consistentemente apareceram nas raízes das árvores, com destaque para a "rap_initial_brl", "extention_km", "tx_selic" e "rap_total_brl", que aparecem nas primeiras 4 posições. Das 4.000 árvores geradas nas 10 iterações do Cross-validation, essas variáveis apareceram 1.725 vezes nas raízes das árvores, o que corresponde por aproximadamente 43% das vezes. Isso significa que 43% das vezes essas variáveis foram consideradas mais puras do que as outras.

Figura 25 – Gráficos de barras mostrando a frequência com que cada variável aparece nas raízes das árvores.



Fonte: Elaborada pelo autor.

Uma análise das variáveis mais importantes para predição da variável alvo mostra que, dentre as quatro variáveis que mais apareceram nas raízes das árvores, somente a variável "extention_km" não se encontra entre as seis mais importantes, aparecendo na sétima posição. Cada iteração do Cross-validation gerou um "ranking" das variáveis consideradas mais importantes. Esses rankings são apresentadas nos 10 gráficos de barras horizontais (um para cada iteração) das figuras 26 e 27.

A Tabela 9 destaca que, dentre as 10 iterações do Cross-validation, somente duas variáveis foram consideradas como mais importantes, ocupando a primeira posição no ranking. Essa tabela mostra também a quantidade de vezes que isso aconteceu para cada uma dessas duas variáveis.

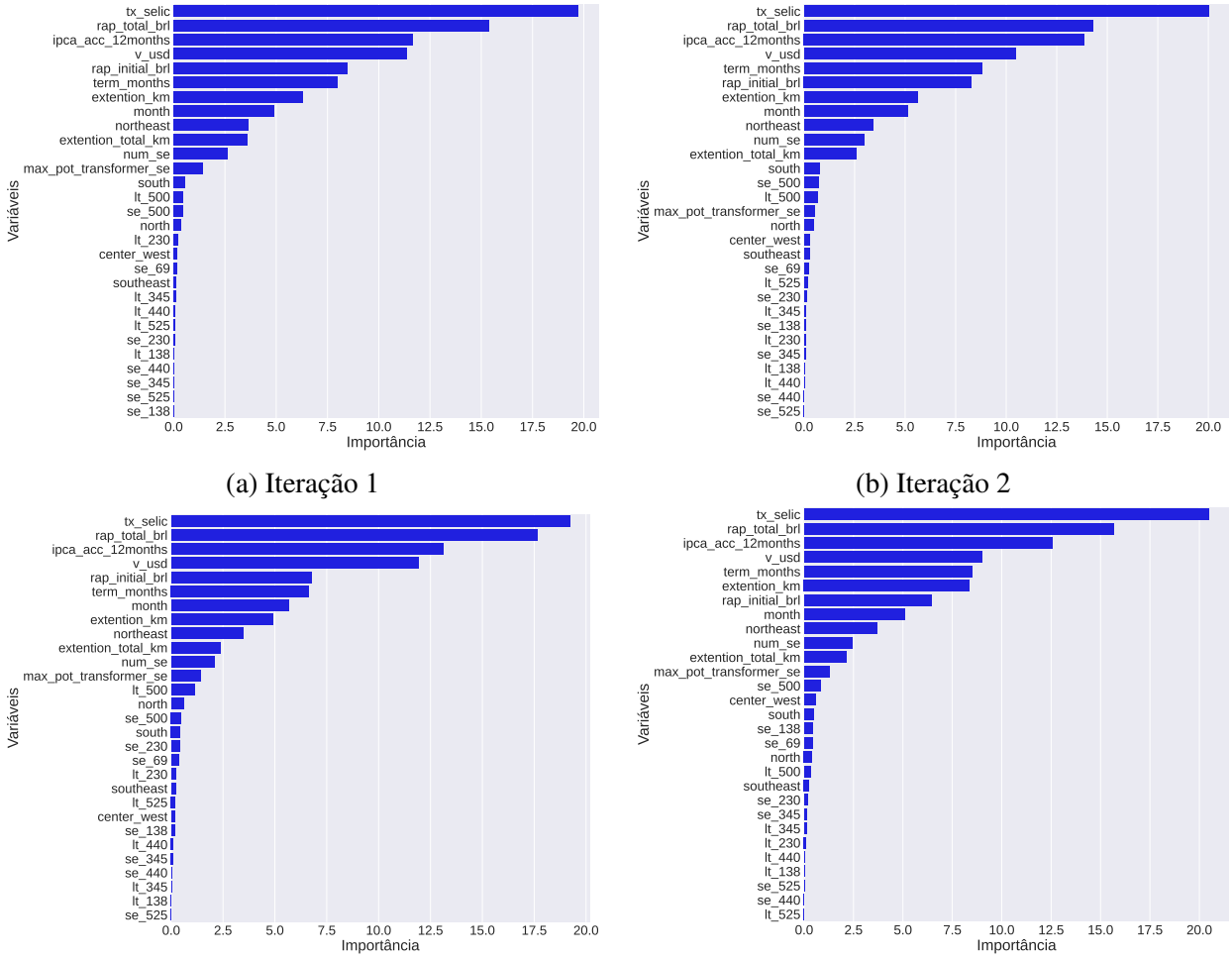
Tabela 9 – As duas variáveis que ocuparam a primeira posição como mais importantes para predição da variável alvo, dentre as 10 iterações do Cross-validation.

Variável	Quantidade
tx_selic	9
rap_total_br	1

Fonte: Dados da pesquisa.

Conforme já dito, cada iteração do Cross-validation gerou um ranking com as variáveis consideradas mais importantes para predição da variável alvo. Portanto, foi realizada uma contagem para se construir um "ranking geral" que considerou todas as iterações do Cross-

Figura 26 – Relação das variáveis mais importantes para cada iteração do Cross-validation.



(a) Iteração 1

(b) Iteração 2

(c) Iteração 3

(d) Iteração 4

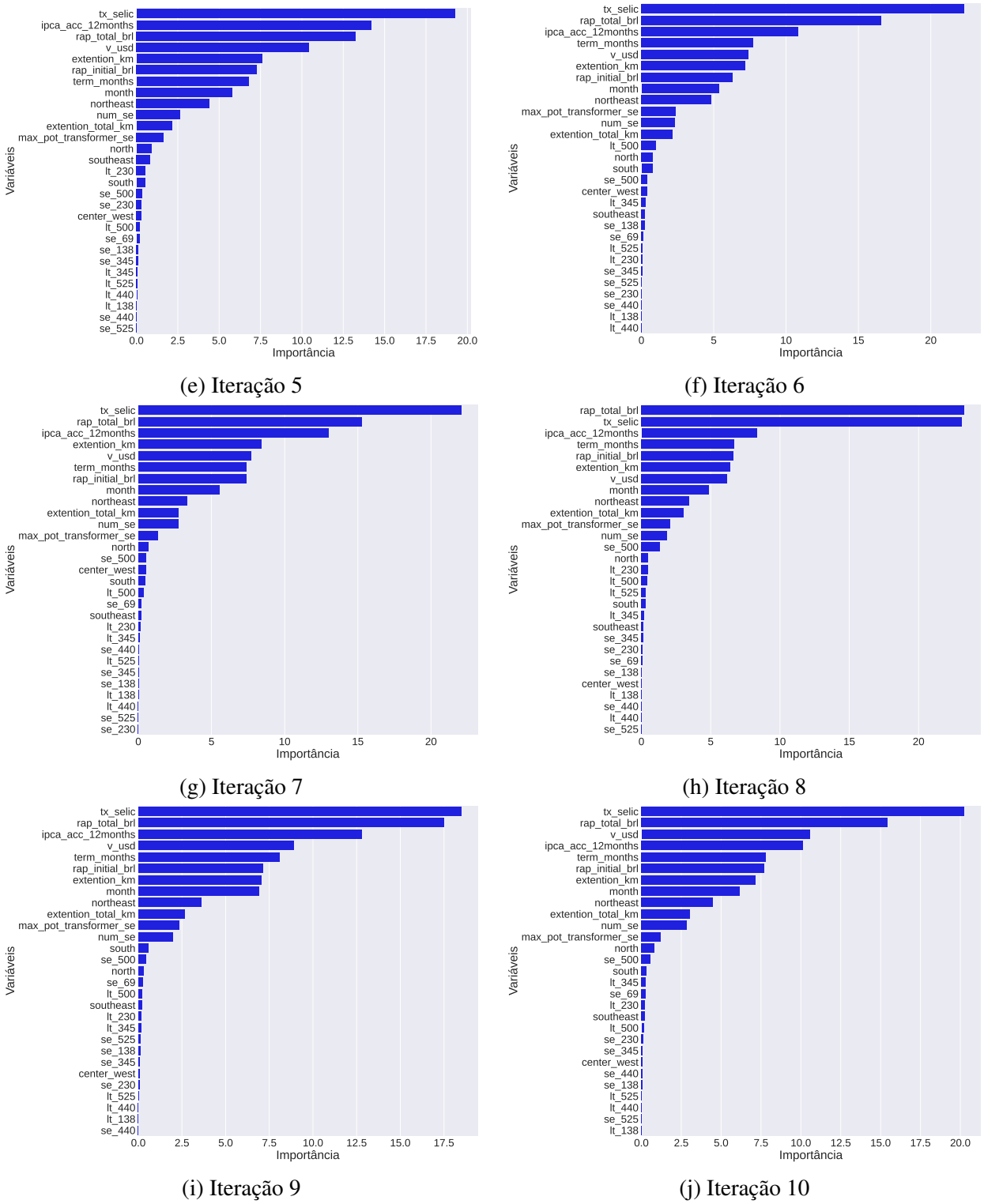
Fonte: Elaborada pelo autor.

validation. Essa contagem considerou a variável que mais apareceu em cada posição do ranking. Por exemplo, conforme mostrou a Tabela 9, considerando-se as 10 iterações, a variável "tx_selic" apareceu 9 vezes na primeira posição, e a variável "rap_total_brl" apareceu uma vez. Portanto, dessa forma, no ranking geral, a variável "tx_selic" ocupou a primeira posição, sendo considerada a variável mais importante para predizer a variável alvo. A Tabela 10 mostra o ranking geral para as 10 primeiras posições, ou seja, as 10 variáveis mais importantes para predizer a variável alvo quando considerada as 10 iterações do Cross-validation.

É possível notar que as sete variáveis mais consistentes, aquelas que apareceram mais vezes nas mesmas posições nas 10 iterações, foram as variáveis "tx_selic", "rap_total_brl", "ipca_acc_12months", "v_usd", "term_months", "rap_initial_brl" e "extension_km", respectivamente. Coincidentemente, essas são as mesmas sete variáveis escolhidas pelos métodos de seleção de variáveis pré-definidos na Etapa 4 e vistas no Quadro 4.

Acredita-se que a variável "tx_selic" tenha sido consistentemente o atributo indicado como mais importante, pois é a taxa básica de juros da economia brasileira, estando ligada

Figura 27 – Relação das variáveis mais importantes para cada iteração do Cross-validation.



(e) Iteração 5

(f) Iteração 6

(g) Iteração 7

(h) Iteração 8

(i) Iteração 9

(j) Iteração 10

Fonte: Elaborada pelo autor.

ao fluxo de emissão e de circulação de recursos financeiros no país. Ou seja, por nortear todos os demais juros, ela é extremamente importante para quem realiza algum empréstimo ou financiamento em instituições financeiras. Em outras palavras, podemos entender a taxa de juros como o preço do “aluguel” do capital financeiro. Uma vez que o modelo de negócio adotado

Tabela 10 – As 10 variáveis que mais apareceram nas raízes das árvores geradas durante as 10 iterações do Cross-validation.

Posição em importância	Variável	Quantidade
1^a	tx_selic	9
2^a	rap_total_brl	8
3^a	ipca_acc_12months	8
4^a	v_usd	6
5^a	term_months	4
6^a	rap_initial_brl	4
7^a	extention_km	4
8^a	month	9
9^a	northeast	10
10^a	extention_total_km	6

Fonte: Dados da pesquisa.

pelas empresas participantes dos leilões está totalmente ligado aos empréstimos e financiamentos dos empreendimentos, pode-se entender que o custo desse capital para a real concretização dos investimento é o fator mais importante para o deságio que será apresentado pelas mesmas.

Em segundo lugar, foi observado a variável "rap_total_br". Relembrando que essa variável condiz com a RAP total estipulada para o leilão de um lote em específico, pode-se interpretá-la como sendo a grandeza financeira de um leilão onde, então, essa ordem de grandeza influencia no interesse das proponentes participantes do leilão por determinado lote ou lotes.

Já as variáveis "ipca_acc_12months" e "v_usd" ocupam a terceira e quarta posição, respectivamente. Entende-se que essas variáveis também fazem parte da avaliação macroeconômica do país, juntamente com a Taxa Selic, e influenciam no deságio que será proposto. Pode-se concluir que o IPCA foi considerado nessa lista de variáveis, pois é o índice ao qual a correção do valor do RAP dentro do período de concessão está atrelado. Já o dólar, está atrelado ao custo de diversos equipamentos e materiais necessários para a realização dos empreendimentos, entrando no balanço dos custos do investimento total. Desta forma, ambas as variáveis também são importantes para justificar nossa variável alvo.

Em quinta, sexta e sétima posição foram apontadas, respectivamente, as variáveis "term_-months", "rap_initial_br" e "extension_km". A variável "term_months" indica qual o prazo máximo para a realização do empreendimento, ou seja, o tempo limite para entrada do lote em operação. Consequentemente, está relacionada à complexidade dos empreendimentos, bem como o tempo esperado para realização dos investimentos e início do recebimento da RAP pela proponente vencedora. Analogamente a variável "rap_total_br", a "rap_initial_br"mostra o montante de investimento inicialmente esperado para um lote específico dentro do leilão e, pela grandeza desse montante, pode influenciar no interesse e anseio das empresas participantes. Por último, foi apresentada a variável "extension_km" que refere-se a extensão em quilômetros das linhas de transmissão presente em um leilão e, como explicado na seção 4.1, pode retratar uma

complexidade técnica maior para um lote em específico, trazendo vantagens para empresas que possuam um domínio maior sobre linhas de transmissão ao contrário da escolha do especialista que buscava representar essa complexidade pela quantitade de subestações, através da variável "num_se".

Importante destacar que o modelo CatBoost foi capaz de compreender e destacar a importância dessas variáveis, como descrito nos parágrafos anteriores, apenas através da nossa base de dados.

4.4 Análise dos Resultados Obtidos na Etapa 7

Na Etapa 7, foi realizada a predição da variável alvo "deságio" com dados ainda não vistos. As análises da etapa anterior, as quais utilizaram as médias do $RMSE$ e do R^2, os testes estatísticos de Friedman e Nemenyi, e os três critérios objetivo, apontaram que o contexto mais adequado para realizar as predições foi o $CB/Todas$. Primeiramente, o algoritmo "CatBoost" foi novamente treinado sobre todo o conjunto de treinamento, composto por todas as variáveis e por todas as 269 observações.

Após o treinamento, o modelo resultante foi utilizado para realizar as predições sobre o conjunto de teste, ou seja, sobre dados ainda não vistos. Esse conjunto de teste havia sido gerado na Etapa 3 e seus dados foram isolados para que não houvessem vazamentos, prejudicando a construção dos modelos. Em seguida, com os valores preditos e os valores reais de deságios, foram calculadas as métricas $RMSE$ e R^2, apresentados na Tabela 11.

Tabela 11 – Resultados das métricas $RMSE$ e R^2 calculadas para as predições realizadas sobre o conjunto de teste.

Métrica	Valores
$RMSE$	13.18
R^2	0.60

Fonte: Dados da pesquisa.

Percebe-se que o $RMSE$ obtido com a predição sobre dados não vistos é um pouco mais elevado do que a média dessa métrica obtida durante a Etapa 5 (12.39). Já o R^2 apresentou praticamente o mesmo resultado tanto com os dados ainda não vistos, como na Etapa 5 (0.59). Essa diferença no $RMSE$ pode ser explicada exatamente pelo fato do modelo estar sendo aplicado, agora, sobre dados novos, simulando uma situação real.

Após a predição com o todo o conjunto de teste, a título de comparação, foram realizadas três novas predições sobre esse mesmo conjunto de dados, porém, utilizando-se seleções diferentes de variáveis. Primeiramente, foram utilizadas somente as três variáveis mais importantes, de acordo com o ranking geral da Tabela 10. Em seguida, foram utilizadas as quatro variáveis mais importantes. E, por último, foram utilizadas as seis variáveis mais importantes. Para cada

predição, o algoritmo CatBoost foi treinado novamente sobre todo o conjunto de treinamento, porém, obviamente, somente com as variáveis selecionadas. Para cada predição, também foram calculadas as métricas *RMSE* e R^2. Os resultados são apresentados na Tabela 12.

Tabela 12 – As 10 variáveis que mais apareceram nas raízes das árvores geradas durante as 10 iterações do Cross-validation.

Variáveis Utilizadas	*RMSE*	R^2
3 mais importantes	10.51	0.75
4 mais importantes	10.60	0.74
6 mais importantes	11.61	0.69

Fonte: Dados da pesquisa.

Essas novas predições mostraram que o melhor resultado foi alcançado quando se utilizou somente as três variáveis mais importantes, segundo o ranking geral. Inclusive, o desempenho do modelo quando se utilizou somente essas três variáveis foi consideravelmente melhor do que o alcançado quando se utilizou todas as variáveis, tanto para o *RMSE*, quanto para o R^2. Quando utilizadas todas as variáveis, essas métricas foram 13.18 e 0.60, respectivamente. Além disso, vale notar ainda que, também com as quatro e seis variáveis mais importantes, o desempenho do modelo foi superior do que quando foram utilizadas todas as variáveis.

CAPÍTULO

5

CONCLUSÃO

Neste trabalho, foi investigada a capacidade preditiva de quatro algoritmos de Aprendizado de Máquina para o deságio em lotes de leilões de Transmissão da ANEEL. Os quatro algoritmos utilizados foram: Árvore de Decisão, *Random Forest*, XGBoost e CatBoost. Em seguida, foram comparados e avaliados os desempenhos dos modelos gerados por esses algoritmos. Para isso, foram utilizadas as métricas *RMSE* e R^2, e os testes estatísticos de Friedman e post-hoc de Nemenyi. Por último, foi selecionado o contexto mais adequado para predizer deságios em lotes de Leilões da ANEEL utilizando dados reais não vistos.

Os resultados obtidos neste estudo mostraram que o modelo CatBoost com todas as variáveis ($CB/Todas$) obteve melhor média de *RMSE* e média de R^2 mais próxima de 1 quando comparado com todos os outros contextos. Além disso, o teste post-hoc de Nemenyi mostrou que é possível afirmar que esse contexto é o que possui desempenho superior em relação a um maior número de outros contextos. Dessa forma, ele foi selecionado como o mais adequado para ser utilizado para realizar as predições sobre dados não vistos. Antes das predições, foi realizada uma análise da importância das variáveis, de acordo como modelo CatBoost. Durante a etapa de treinamento, foi utilizado o Cross-validation com 10 iterações. Para cada iteração, o algoritmo gerou uma sequência ordenada com as variáveis mais importantes para explicar o problema em questão. Verificou-se que a variável "tx_selic" foi elencada como a mais importante para explicar o deságio nesses leilões, seguida pelas variáveis "rap_total_brl" e "ipca_acc_12months".

Sobre os dados não vistos, o CatBoost apresentou *RMSE* com valor superior à média obtida durante a etapa de treinamento e R^2 com valor praticamente igual à média obtida durante o treinamento. Portanto, pode-se dizer que, de forma geral, esse algoritmo apresentou desempenho um pouco inferior quando aplicado sobre os dados não vistos. No entanto, foram também realizadas predições sobre o conjunto de teste com as 3, 4 e 6 variáveis mais importantes. Isso revelou que o melhor desempenho para esse algoritmo seria obtido utilizando-se as 3 variáveis mais importantes.

No geral, os resultados obtidos com este trabalho mostram que predições de deságios de leilões de transmissão da ANEEL com o algoritmo CatBoost apresentam bons resultados, inclusive quando comparados aos outros modelos também abordados neste estudo, e portanto merece ser estudado com mais profundidade.

Em trabalhos futuros, podem ser considerados outros contextos com métodos de seleção diferentes, visando conhecer melhor o comportamento desse modelo em diversos cenários. Além disso, os métodos utilizados poderão ser analisados com profundidade, visando entender a seleção de variáveis realizada por eles. Outro ponto que poderá ser implementado no futuro é uma etapa de ajuste fino dos hiperparâmetros do algoritmo, testando e avaliando diversas outras possibilidades além daquela utilizada neste trabalho. Também será interessante executar a etapa de treinamentos, quando foram coletadas as métricas $RMSE$ e R^2 com o Cross-validation, mais de uma vez. Nesse caso, será possível verificar se todas as execuções confirmam o CatBoost com todas as variáveis como o contexto mais adequado para ser aplicado sobre dados reais. Com o intuito de explorar melhor o algoritmo Árvore de Decisão, o parâmetro "max_depth" pode ser ajustado para profundidades maiores, como 6, 9 ou 12, buscando aprimorar os resultados obtidos por esse algoritmo e compará-lo com os demais algoritmos.

REFERÊNCIAS

ALPAYDIN, E. **Introduction to Machine Learning**. [S.l.]: MIT Press, 2014. Citado nas páginas 27, 28 e 29.

ANEEL, A. N. de E. E. **Editais - Leilões de transmissão**. 2022. Https://www2.aneel.gov.br/aplicacoes$_l$$i$$f$$e$$r$$a$$y$/$e$$d$$i$$t$$a$$i$$s_t$$r$$a$$n$$s$$m$$i$$s$$s$$a$$o$/$e$$d$$i$$t$$a$$l_t$$r$$a$$n$$s$$m$$i$$s$$s$$a$$o$.$c$$f$$m$.$A$$c$$e$$s$$s$$a$$d$$o$$e$$m02J$u

BISHOP, C. M.; NASRABADI, N. M. **Pattern recognition and machine learning**. [S.l.]: Springer, 2006. v. 4. Citado nas páginas 27 e 28.

CHEN, T.; GUESTRIN, C. Xgboost: A scalable tree boosting system. In: **Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining**. [S.l.: s.n.], 2016. p. 785–794. Citado na página 30.

DOROGUSH, A. V.; ERSHOV, V.; GULIN, A. Catboost: gradient boosting with categorical features support. **arXiv preprint arXiv:1810.11363**, 2018. Citado na página 31.

EPE, E. de P. E. **Leilão de Transmissão nº 001/2022.** 2022. Https://www.epe.gov.br/pt/leiloes-de-energia/leiloes-de-transmissao/leilao-de-transmissao-n-001-2022 . Acessado em 28 Maio 2022. Citado na página 23.

FRANK, E.; MARK, A. **Hall, and Ian H. Witten (2016). The WEKA Workbench. Online Appendix for"Data Mining: Practical Machine Learning Tools and Techniques**. [S.l.]: morgan kaufmann,, 2016. Citado na página 29.

FRIEDMAN, J. H. Greedy function approximation: a gradient boosting machine. **Annals of statistics**, JSTOR, p. 1189–1232, 2001. Citado na página 30.

FROST, J. Regression analysis: An intuitive guide for using and interpreting linear models. Statistics By Jim Publishing, 2020. Citado na página 48.

GORMAN, B. **A Kaggle Master Explains Gradient Boosting**. 2017. Https://www.gormanalysis.com/blog/gradient-boosting-explained/ . Acessado em 16 Junho 2022. Citado na página 30.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. H.; FRIEDMAN, J. H. **The elements of statistical learning: data mining, inference, and prediction**. [S.l.]: Springer, 2009. v. 2. Citado na página 27.

HULL, J. **Machine learning in business: An introduction to the world of data science**. [S.l.]: Amazon Fulfillment Poland Sp. z oo, 2021. Citado na página 28.

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. **An introduction to statistical learning**. [S.l.]: Springer, 2013. v. 112. Citado nas páginas 28, 29 e 30.

KELLEHER, J. D.; NAMEE, B. M.; D'ARCY, A. **Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies**. [S.l.]: MIT press, 2020. Citado na página 29.

LOU, Y.; OBUKHOV, M. Bdt: Gradient boosted decision tables for high accuracy and scoring efficiency. In: **Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining**. [S.l.: s.n.], 2017. p. 1893–1901. Citado na página 31.

MARSLAND, S. **Machine learning: an algorithmic perspective**. [S.l.]: Chapman and Hall/-CRC, 2011. Citado na página 29.

MOODY, J. **What does RMSE really mean?** 2019. Https://towardsdatascience.com/what-does-rmse-really-mean-806b65f2e48e . Acessado em 28 Junho 2022. Citado na página 47.

MURPHY, K. P. **Probabilistic machine learning: an introduction**. [S.l.]: MIT press, 2022. Citado nas páginas 27 e 28.

NEMENYI, P. B. **Distribution-free multiple comparisons.** [S.l.]: Princeton University, 1963. Citado na página 48.

PRESTES, A. F.; BEZERRA, F. M.; MELLO, G. R. de; CASTRO, T. E. de. Investimento em infraestrutura energética e o crescimento econômico brasileiro no período de 2003 a 2018. **Revista Brasileira de Energia| Vol**, v. 25, n. 2, 2019. Citado na página 24.

PROKHORENKOVA, L.; GUSEV, G.; VOROBEV, A.; DOROGUSH, A. V.; GULIN, A. Catboost: unbiased boosting with categorical features. **Advances in neural information processing systems**, v. 31, 2018. Citado na página 31.

THIESEN, S. **CatBoost regression in 6 minutes**. 2021. Https://towardsdatascience.com/catboost-regression-in-6-minutes-3487f3e5b329. Acessado em 02 Julho 2022. Citado na página 31.

WONG, K. J. **CatBoost vs. LightGBM vs. XGBoost**. 2022. Https://towardsdatascience.com/catboost-vs-lightgbm-vs-xgboost-c80f40662924. Acessado em 03 Julho 2022. Citado na página 31.


ICMC
USP
SÃO CARLOS