FLÁVIO DE FALCÃO E HELENA
Engenheiro Civil, Universidade de São Paulo

OPORTUNIDADES NO MERCADO IMOBILIÁRIO COM APLICAÇÃO DE MODELOS DE APRENDIZAGEM DE MÁQUINA: UM ESTUDO DE CASO EM SÃO PAULO, BRASIL

Versão Original

Dissertação apresentada à Escola Politécnica da Universidade de São Paulo para obtenção do título de Mestre em Ciências

Área de concentração:
Sistemas Eletrônicos

Orientador: Prof. Dr. Flávio A. M. Cipparrone

SÃO PAULO
2023

Catalogação-na-publicação

Helena, Flávio
Oportunidades no mercado imobiliário com aplicação de modelos de aprendizagem de máquina: um estudo de caso em São Paulo, Brasil. / F. Helena -- São Paulo, 2023.
100 p.

Dissertação (Mestrado) - Escola Politécnica da Universidade de São Paulo. Departamento de Engenharia de Sistemas Eletrônicos.

1.Precificação de imóveis 2.Modelagem de inteligência de máquina 3.Base de dados I.Universidade de São Paulo. Escola Politécnica. Departamento de Engenharia de Sistemas Eletrônicos II.t.

AGRADECIMENTOS

"Don't underestimate the power of vision and direction. These are irresistible forces, able to transform what might appear to be unconquerable obstacles into traversable pathways and expanding opportunities."

Jordan Peterson

Primeiramente agradeço a Deus pelo dom da vida.

Esse trabalho não seria possível sem algumas pessoas. Agradeço inicialmente ao Professor e orientador Flávio Cipparrone, que esteve sempre presente para discussões e enriquecimento do tema deste trabalho.

Agradeço também ao estimado amigo Celso Azevedo, que contribuiu com a pesquisa e com ideias relevantes para este trabalho.

Agradeço também à minha mãe, Maria Luiza, por acreditar sempre em tudo que me propus fazer e me apoiar incondicionalmente.

Também agradeço a meu pai, Flávio, que sempre foi minha inspiração e o melhor engenheiro civil que já conheci.

E, de forma geral, agradeço a todos, direta ou indiretamente envolvidos neste trabalho.

Lista de Figuras

Lista de Tabelas

Sumário

Resumo

O objetivo deste trabalho é estudar a precificação de propriedades imobiliárias na cidade de São Paulo com o intuito de encontrar oportunidades sub precificadas. A modelagem utilizada considera variáveis intrínsecas (número de quartos, área construida, ano de construção, etc.), assim como variáveis extrínsecas (qualidade do asfalto, transporte público, florestamento, etc.) para estimar o preço de mercado de cada propriedade e, para isso, utiliza dados provenientes de listagens online de apartamentos e bases de dados públicas. Uma modelagem estatística inovadora é proposta para encontrar oportunidades, buscando explorar a robustez de distintos modelos de aprendizagem de máquina (Hedônico, KNN e XGBoost), ao ponderar os resultados segundo os seus respectivos erros percentuais médios. Os resultados indicam ser possível encontrar oportunidades, o que motiva futura pesquisa de aprofundamento na metodologia.

Palavras chave: Mercado imobiliário; XGBoost; Hedonic Price Modeling; Real Estate; oportunidade

Abstract

The objective of this work is to study the pricing of real estate properties in the city of São Paulo in order to find underpriced opportunities. The modeling used takes into account intrinsic variables (number of bedrooms, built area, year of construction, etc.), as well as extrinsic variables (quality of the asphalt, public transportation, afforestation, etc.) to estimate the market price of each property, using data from online apartment listings and public databases. An innovative statistical modeling is proposed to find opportunities, aiming to explore the robustness of different machine learning models (Hedonic, KNN, and XGBoost) by weighing the results according to their respective average percentage errors. The results indicate that it is possible to find opportunities, which motivates further research to deepen the methodology.

Keywords: Real estate market; XGBoost; Hedonic Price Modeling; Real Estate; opportunity

1. Introdução

Os imóveis representam um dos bens mais importantes que uma pessoa pode possuir ao longo da sua vida. De fato, eles não são apenas um lugar para as pessoas morarem, mas geralmente representam o componente mais relevante da riqueza privada. Portanto, modelar os preços dos imóveis é importante para diferentes atores do mercado: para os agentes de políticas governamentais, sendo fundamental para a previsão econômica de curto prazo e bem-estar social; para as empresas, uma vez que as construtoras se deparam com a relevante questão de construir ou não e precisam de informações para tomar uma decisão assertiva (Rafiei e Adeli, 2016); e para o público em geral e investidores, composto por proprietários e locatários, avaliando suas decisões com base nos preços atuais e futuros dos imóveis.

De forma mais específica, os preços das propriedades imóveis são relevantes para a pessoa vendedora na medida que o valor de anúncio de um imóvel afeta diretamente a sua liquidez, isto é, o tempo estimado de venda. De acordo com Cervero e Kang (1989) esse relacionamento é complexo, mas em geral, propriedades super precificadas levam mais tempo para vender, independentemente das condições gerais do mercado ou do subgrupo de preços. Especificamente, o tempo necessário para vender depende de quanto o preço de listagem excede o valor de mercado.

De acordo com Haurin *et al.* (2010), o número de ofertas que chega em um apartamento será, via de regra, inversamente proporcional a seu preço listado. No entanto, há uma oportunidade financeira em posicionar a propriedade em um valor superior, uma vez que as ofertas que efetivamente chegarão serão superiores. Esse relacionamento depende de uma série de fatores, como custo de manutenção do vendedor, urgência na venda, idade da propriedade, entre outros.

Nesse contexto, para tornar o apartamento mais líquido, muitos proprietários aceitam uma redução no valor listado de suas propriedades, com diminuição de preços abaixo daqueles que poderiam vender o apartamento em condições mais ideais de mercado ou se pudessem esperar mais. Esse cenário pode apresentar uma arbitragem, principalmente no contexto do mercado imobiliário brasileiro, que conta com participantes tradicionais com informações ainda esparsas e pouco confiáveis.

Assim, encontrar uma metodologia consistente de precificação para imóveis permite não apenas a possibilidade de vendedores alinharem o preço de listagem de uma propriedade com seu valor de mercado, mas também para determinar propriedades que apresentem preço de listagem abaixo do preço previsto, o que pode representar uma oportunidade de compra para investidores e participantes do mercado imobiliário em geral.

O objetivo deste trabalho, portanto, é estudar a precificação de propriedades, especificamente apartamentos, na cidade de São paulo, e definir uma metodologia consistente de modelagem de preços que permita encontrar propriedades sub precificadas. Na primeira parte do trabalho, é apresentada uma revisão de literatura sobre precificação de imóveis e trabalhos relevantes para referência. A próxima seção discute e detalha a metodologia e a construção das bases de dados utilizadas para modelar os preços para este estudo de caso de São Paulo, Brasil, a maior cidade da América Latina. A terceira seção traz uma análise exploratória da base de dados obtida no item anterior. A quarta seção traz a sugestão de implementação do processo de procura de oportunidades e a modelagem em um banco de dados considerando a base de dados gerada. Por fim, são tecidas conclusões e sugestões de próximos passos para o trabalho na última seção.

Cabe ressaltar que neste trabalho, a parte da metodologia, com o desenvolvimento e construção da base de dados, especificamente o Capítulo 3, é originário de artigo científico que foi publicado nos Anais do XXXVIII Simpósio Brasileiro de Bancos de Dados (SBBD, 2023), no qual o autor deste trabalho também figura entre a lista de autores do artigo. Assim sendo, cabe destacar que este referido capítulo foi uma construção conjunta entre este autor e Celso Azevedo Ribeiro.

2. Revisão Bibliográfica

Ventolo e Williams (2001) definem uma avaliação no âmbito do mercado imobiliário como, fundamentalmente, uma estimativa de valor, seguindo a definição estabelecida no USPAP (Padrão Uniforme de Práticas de Avaliação de Propriedades). Esse processo abrange não apenas a descrição da propriedade em questão, mas também a opinião do avaliador sobre a condição da propriedade, sua adequação para um propósito específico e o provável valor monetário no mercado aberto. Uma avaliação bem conduzida, que seja objetiva, baseada em pesquisa sólida e devidamente documentada, desempenha um papel fundamental no auxílio à tomada de decisões para todas as partes envolvidas, seja em uma negociação de compra, venda, locação ou qualquer outra transação imobiliária.

Ventolo e Williams (2001) também destacam a importância de obter uma estimativa confiável de valor, uma vez que diferentes partes envolvidas têm motivações distintas. O vendedor busca compreender o valor da propriedade a fim de estabelecer um preço de venda justo, enquanto o comprador deseja pagar o mínimo possível. Ao mesmo tempo, o corretor busca otimizar sua comissão. Além disso, as instituições financeiras, independentemente do valor do empréstimo, requerem avaliações para embasar suas decisões de subscrição, determinando a quantia a ser emprestada aos candidatos a crédito. As avaliações também desempenham um papel crucial na estimativa de valores para fins de impostos e seguros, bem como em processos de desapropriação.

Em uma abordagem mais ampla, os preços da habitação são influenciados por duas amplas categorias de fatores, conforme discutido por D'acci (2019). Essas categorias abrangem as características intrínsecas e extrínsecas das propriedades. As características intrínsecas referem-se a todos os fatores estritamente relacionados à própria propriedade, como o número de quartos, banheiros, janelas, varandas e o ano de construção, entre outros. Por outro lado, as características extrínsecas englobam os fatores geográficos, econômicos e políticos que não estão diretamente ligados à propriedade em si. Isso inclui a proximidade a parques e praças urbanas, a disponibilidade e qualidade do transporte público, a condição das vias públicas e calçadas, os níveis de poluição, a distância até importantes centros de trabalho e outros aspectos semelhantes.

Na literatura, é possível encontrar uma grande quantidade de trabalhos mostrando os efeitos de uma série de características extrínsecas na avaliação de imóveis. Benson et al.

(1998) realizaram um estudo que estimou o valor da vista de uma propriedade em mercados imobiliários residenciais unifamiliares. Focando em Bellingham, Washington, eles diferenciam a vista por tipo e qualidade. Seus resultados revelaram que as vistas de alta qualidade, como as panorâmicas para o oceano, podem aumentar os preços das casas em até 60%, enquanto mesmo as vistas de menor qualidade ainda acrescentam cerca de 8%.

Jim e Chen (2009) avaliaram a amenidade de paisagens naturais em Hong Kong, considerando vistas do porto e da montanha. Seu estudo indicou que uma vista ampla do porto poderia aumentar o valor de um apartamento em até 2,97%, enquanto uma vista limitada ainda elevava o preço em 2,18%. Curiosamente, uma vista ampla da montanha tinha um efeito oposto, reduzindo o preço em 6,7%. A distância entre o apartamento e essas vistas preferenciais também teve impacto nos preços, mostrando como as características extrínsecas podem influenciar as avaliações imobiliárias.

Jin et al. (2022) realizaram uma análise da relação entre a acessibilidade ao transporte público e os preços das habitações em Xangai, China, empregando o método GBRT. Os resultados destacam a variação no impacto da acessibilidade em casas de diferentes faixas de preço, revelando que o tempo de viagem até o centro de negócios central influencia mais as habitações de baixo e médio preço, enquanto a acessibilidade ao metrô em toda a cidade é de maior relevância para as residências de alto preço. Além disso, observaram-se efeitos não lineares e limítrofes da acessibilidade ao transporte público nos preços das habitações, incluindo variações negativas em certos casos. Essas descobertas apontam para a possibilidade de desigualdades na acessibilidade entre diferentes grupos de renda, destacando a necessidade de considerações políticas neste contexto.

Cordera et al. (2018) estudaram a relação entre acessibilidade e preços de propriedades urbanas nas cidades de Roma e Santander. Os resultados indicaram que a acessibilidade foi um fator com efeito positivo nos preços dos imóveis em ambas as cidades, embora em Santander isso tenha ocorrido para apenas um dos indicadores definidos no trabalho. Esses resultados apoiam, segundo os autores, a introdução de políticas de captura de valor.

Variáveis extrínsecas relacionadas à economia também foram estudadas por seus efeitos no mercado imobiliário. Alhodiry et al. (2021) conduziram uma pesquisa que revelou a influência positiva dos preços do petróleo no mercado imobiliário turco. Essa relação é mediada pelas taxas de juros domésticas. Além disso, os resultados indicam que as taxas de juros dos Estados Unidos exercem um impacto considerável no mercado turco, atuando por meio das

flutuações nos preços do petróleo e das taxas locais. Essa análise ressalta a sensibilidade do mercado imobiliário turco a eventos externos, devido à interconexão econômica e à presença significativa da moeda dos EUA na dívida turca.

Mais recentemente, Ma et al. (2023) conduziram uma pesquisa que revelou até mesmo a existência de influência do sentimento da mídia em relação às políticas governamentais nos preços do mercado imobiliário chinês. Os resultados indicam que o impacto desse sentimento na precificação de imóveis é heterogêneo, variando de acordo com a natureza do sentimento, positivo ou negativo, e o veículo de mídia responsável por divulgá-lo. Notavelmente, o estudo demonstrou que quando as políticas governamentais são promovidas pela mídia oficial, o impacto no preço dos imóveis é mais pronunciado.

Assim, uma das etapas mais importantes da avaliação imobiliária é a escolha de um modelo adequado que possa incluir, no processo de estimativa de preço, variáveis intrínsecas e extrínsecas. Uma das abordagens mais consolidadas é o uso de métodos hedônicos, que foi definido por Rosen (1974) como um modelo que decompõe os valores das propriedades "em pacotes de atributos de utilidade que contribuem para a heterogeneidade observada nos preços, considerado como a soma composta de elementos que representam preços estruturais e locacionais implícitos". Assim, os preços dos imóveis serão compostos por acréscimos incrementais de múltiplas características, incluindo variáveis intrínsecas e extrínsecas.

Adicionalmente, Herath e Maier (2010) afirmam que o Método de Precificação Hedônica (MPH) é um modelo que avalia uma mercadoria considerando suas propriedades constituintes, calculando e somando seus valores separadamente, de forma independente. De acordo com esta definição, dois requisitos precisam ser atendidos: o preço estimado da propriedade deve ser expresso por suas partes constituintes e cada parte constituinte deve ter um valor implícito no preço final da propriedade.

De fato, a classificação inicial baseada nas citações ISI Web of Knowledge, considerando 471 artigos, mostra que existe uma vasta linha de estudos utilizando o MPH para analisar diversos aspectos do mercado imobiliário e o impacto de diversos fatores nos preços dos imóveis na habitação preços (Herath e Maier, 2010).

Chau e Chin (2003) destacam que o método de precificação hedônica apresenta diversas limitações dignas de consideração. Em primeiro lugar, é crucial destacar a importância da seleção da forma funcional no modelo, uma vez que uma escolha inadequada pode resultar em estimativas inconsistentes. Além disso, a segmentação de mercado é um desafio a ser

enfrentado, pois os mercados imobiliários não possuem uniformidade intrínseca, embora a literatura careça de diretrizes claras quanto à definição e à mensuração dos submercados correspondentes. A especificação inadequada de variáveis também é uma preocupação relevante, pois pode contribuir para a ineficiência ou para o viés nas estimativas obtidas. Por fim, é importante destacar que o uso de variáveis proxy, quando não se dispõe de dados reais, pode introduzir erros de medição, comprometendo assim a integridade dos resultados, que podem se tornar tendenciosos e inconsistentes.

A precisão na estimativa dos preços de venda de imóveis é fundamental, uma vez que vendedores e compradores dependem dessas informações no momento da negociação. Como uma alternativa às limitações do modelo hedônico, Selim (2009) comparou os desempenhos de previsão dos modelos de regressão hedônica e de redes neurais artificiais e demonstrou que as Redes Neurais Artificiais (RNA) podem ser uma alternativa mais precisa e robusta para a previsão dos preços das casas na Turquia. McCluskey et al. (2013) mostraram que a RNA tem um desempenho muito bom em termos de poder preditivo e, portanto, precisão de avaliação, superando a tradicional análise de regressão múltipla (MRA) na avaliação de imóveis.

Na mesma linha, o estudo de Zhao et al. (2014) utilizou algoritmos relacionados ao KNN para analisar os preços de imóveis usados, que foram entendidos como um indicador confiável da demanda real do mercado imobiliário, devido ao foco na habitação, em vez de investimento. Isso inclui previsão de preços e a comparação entre os algoritmos KNN e KNN ponderado. Através da análise da importância dos atributos, pode-se demonstrar as influências dos diferentes atributos dos imóveis nos preços e nas principais preocupações dos compradores.

Outro estudo de Zeng (2021) comparou o desempenho do modelo de regressão XGBoost com o modelo tradicional de regressão múltipla, e os resultados destacaram a superioridade do XGBoost na avaliação de preços imobiliários, oferecendo previsões mais robustas em relação à abordagem de random forest. Além disso, Zaki et al. (2022) utilizaram o XGBoost para prever preços de casas e alcançaram uma notável precisão de 84,1%, em contraste com a regressão hedônica, que obteve apenas 42% de precisão.

Um estudo adicional, realizado por Guliker et al. (2022), concentrou-se no mercado imobiliário da Holanda, onde os preços das casas têm se mantido elevados. Neste estudo, três modelos foram comparados: regressão linear, regressão geograficamente ponderada e XGBoost. Os resultados revelaram que o XGBoost apresentou o desempenho mais preciso,

explicando 83% da variância com um RMSE de €65,312 e um MAPE de 6,35%. Vale ressaltar que esse estudo também destacou a importância do uso de dados geoespaciais na construção de um modelo de avaliação imobiliária abrangente em nível nacional.

Os trabalhos publicados recentemente, citados anteriormente, apenas aumentam a possibilidade de explorar os campos da modelagem de aprendizado de máquina para avaliação de imóveis.

Mais especificamente ao tema deste trabalho, com o intuito de encontrar propriedades abaixo do valor de mercado, Baldominos et al. (2018) desenvolveram uma aplicação de aprendizado de máquina que identifica oportunidades com um preço substancialmente abaixo do preço de mercado em Madri (Espanha). A aplicação é formalmente implementada como um problema de regressão que tenta estimar o preço de mercado de uma propriedade a partir de características recuperadas de listagens públicas online utilizando distintos modelos, incluindo árvores de regressão, KNN, Support Vector Machines e RNAs, identificando vantagens e desvantagens de cada um deles. O melhor modelo gerou um erro médio e mediano, respectivamente, de 16.8% e 5.71%.

Neste trabalho, é desenvolvida uma metodologia sistemática para encontrar oportunidades no mercado imobiliário da cidade São Paulo, Brasil, usando dados provenientes de sites de listagem de apartamentos e fontes fornecidas de forma aberta ao público. A modelagem será composta por uma Regressão hedônica, que será o modelo base de comparação, e também pelo uso do modelo XGBoost e pelo modelo KNN (K-nearest neighbors), com objetivo de comparar os preços aferidos de mercado e os valores de listagem, de acordo com o erro médio. Não foi possível encontrar nenhum trabalho com metodologia semelhante para a cidade de São Paulo do conhecimento do autor deste trabalho.

3. Dados e Metodologia de Tratamento

Vale ressaltar uma importante peculiaridade do Brasil quando se trata de transações imobiliárias. O Brasil utiliza um sistema de registro dessas transações por meio do Cartório de Notas, onde as informações são de propriedade do governo e não são disponibilizadas publicamente. (Tierno *et al.* 2007). Portanto, diferentemente de vários países desenvolvidos, onde não existe essa limitação de acesso às informações, uma vez que os cadastros imobiliários foram totalmente privatizados, no Brasil não existe um banco de dados oficial contendo os preços de transação ou mercado dos imóveis.

Diante desse detalhe, foi necessário utilizar uma forma alternativa de obtenção dessas informações, por meio de anúncios imobiliários acessados online. Além disso, outros dados sobre a infraestrutura e característica das cidades (qualidade do pavimento, calçadas, arborização) são escassos na maior parte do país, sendo a cidade de São Paulo uma das metrópoles que possui um vasto banco de dados com esse tipo de informação. Pelas razões acima expostas, a cidade de São Paulo foi escolhida como objeto deste trabalho.

Por se tratar de uma metrópole com cerca de onze milhões de habitantes, a escolha de áreas delimitadas foi igualmente necessária. Nessa escolha, o principal critério foi garantir tanto a variabilidade da realidade socioeconômica dos bairros escolhidos quanto a representatividade estatística da amostra. Segundo PMSP, 2007 – Índice de Desenvolvimento Humano por Distrito (IDHD) entre os 50 primeiros distritos colocados no ranking do município, foram escolhidos seis, totalizando 5% da população do município nesses distritos (PMSP, 2010). Os distritos escolhidos e seus dados demográficos são apresentados na Tabela 1.

Tabela 1

Distritos de São Paulo escolhidos, escopo do estudo

Distrito	IDH (2007)	IDH Ranking na cidade	População (2010)	% da população total
Moema	0.961	1º	83,368	0.74%
Pinheiros	0.960	2º	65,364	0.58%
Vila Mariana	0.950	7º	130,484	1.16%
Carrão	0.886	30º	83,281	0.74%
Vila Andrade	0.853	48º	127,015	1.13%
Jaguaré	0.849	50º	49,863	0.44%

Tão importante quanto a escolha dos bairros, foi necessária a seleção das características dos imóveis e respectivas fontes de dados. Aqui, partindo das referências já analisadas no Capítulo 2, foram selecionadas as variáveis comumente associadas ao preço e à qualidade de um imóvel, tanto intrínsecas quanto extrínsecas. A Tabela 2 apresenta um breve resumo das principais fontes de dados onde essas variáveis foram obtidas.

Tabela 2
Fontes de dados disponíveis utilizadas neste estudo.

Bases de dados intrínsecas	Bases de dados extrínsecas
◊ Viva Real (Web) ◊ IPTU ◊ Registro urbano	◊ SEADE ◊ GeoSampa

Os próximos itens irão discutir como cada dado foi obtido dessas fontes. Eles também abordarão as suposições e técnicas usadas para tratar os dados, remover itens problemáticos e elaborar um banco de dados final que será utilizado para modelagem neste trabalho.

3.1 Base de dados: Características Intrínsecas

As características intrínsecas são aquelas que estão contidas ou diretamente associadas a uma propriedade. Bons exemplos são o número de quartos, número de banheiros, área construída, vagas de garagem, entre outros. Estas características estão comumente ligadas ao preço do imóvel, pois afetam diretamente a sua usabilidade. (Park e Bae., 2015).

Diferentes fontes de dados foram usadas para obter essas características. A primeira fonte consistiu em dados obtidos por meio de anúncios de imóveis no site VivaReal. Foi desenvolvido um código webscrape em linguagem Python para obter todos os anúncios das 100 primeiras páginas do site, repetindo o processo para cada um dos seis distritos escolhidos, extraídos em 17 de janeiro de 2021. Após o processo, foram extraídos 5.479 anúncios. A Tabela 3 mostra as variáveis intrínsecas que foram extraídas de cada um dos anúncios.

Tabela 3
Variáveis utilizadas na base de dados do Webscrape

Variável	Descrição
Área	Área privada do imóvel em m2
Garagens	Número de espaços para estacionamento

Quartos	Número de quartos
Banheiros.	Número de banheiros
Preço	Valor de listagem
Distrito	Distrito onde a propriedade se localiza
Endereço Completo	Endereço da propriedade com número

A segunda fonte é baseada nos dados de arrecadação do IPTU, disponibilizados pelo município na plataforma GeoSampa. São dados disponíveis ao público, atualizados anualmente, que descrevem diversas informações sobre os imóveis da cidade, bem como seus proprietários. A Tabela 4 apresenta essas informações, referentes ao ano de 2020.

Tabela 4

Variáveis utilizadas na base de dados do IPTU

Variável	Descrição
Ano de construção	Ano de construção do imóvel.
Pavimentos	Número de andares do imóvel.
Esquinas	Número de curvas do terreno.
Fração ideal	Porcentagem da área que cada unidade tem de um edifício.
Área do terreno	Área do Terreno (milhares de metros).
Área construída	Equivalente à Área Útil.
Área usada	Área do Terreno ocupada por construção (milhares de metros).
Fator de obsolescência	Número que varia de 0 a 1, atribuído pela prefeitura, representando o percentual de depreciação do imóvel em função de sua idade e conservação.
Fachada	Tamanho da fachada do prédio (em metros).
Endereço completo	Completo com rua e número.
CodLog	Código relativo ao nome da rua.
CEP	Código postal da propriedade CEP
Número IPTU Número	Único de registro de cada imóvel na prefeitura.

Preço do terreno	Valor avaliado do terreno da propriedade

A terceira fonte também foi obtida da plataforma GeoSampa e possui a geolocalização do terreno da propriedade por meio de seu respectivo polígono geolocalizado. Esses dados estão disponíveis para cada uma das propriedades na cidade. Nesta base de dados, cada polígono está associado a um único número identificador, neste caso o CPF (Tabela 5). Esta será uma importante fonte para relacionar todas as características intrínsecas com sua respectiva posição geográfica na cidade.

Tabela 5
Variáveis utilizadas na base de dados Geo/Polygon

Variável	Descrição
Nº do Contribuinte (Terreno)	Único de registro de cada terreno na prefeitura.
Polígono	Geolocalizado do terreno da propriedade.

3.2 Base de dados: Características Extrínsecas

Em relação às características extrínsecas, é preciso avaliar as peculiaridades do bairro a que pertence cada imóvel. Dados como ciclovias, pontos de ônibus, metrôs, arborização da cidade, entre outros, podem ser caracterizados como extrínsecos. Vários trabalhos na academia têm relacionado esse tipo de característica como tendo impacto na determinação do preço dos imóveis por elas afetados. (D'acci, 2019).

Na cidade de São Paulo, esses recursos estão disponíveis gratuitamente em duas grandes fontes públicas. O primeiro é o SEADE, que reúne dados de diferentes esferas da administração pública. A partir dessa fonte primária, o objetivo foi obter informações relacionados a equipamentos de uso comum disponibilizados à população, mais especificamente sua localização na cidade, com latitude e longitude. Esses dados foram obtidos para as seguintes classes listadas na Tabela 6.

Tabela 6

Variáveis utilizadas na base de dados SEADE

Variável	Descrição
UBS	Latitude e longitude das Unidades Básicas de Saúde Públicas do Município.
CREAS & CRAS	Latitude e longitude dos Centros de Assistência Social, administrados pelo poder público.
Escolas Públicas (Federal, Estadual, Municipal)	Latitude e longitude das escolas públicas de ensino médio da cidade, divididas por administração.
Escolas Particulares e Outras Escolas.	Latitude e longitude dos colégios particulares da cidade.
Universidades Públicas	Latitude e longitude das universidades públicas.
Universidades Privadas	Latitude e longitude das universidades privadas.
Hospitais públicos	Latitude e longitude dos hospitais públicos.
Hospitais Privados	Latitude e longitude dos hospitais privados.
Consultórios Clínicos	Latitude e longitude dos consultórios clínicos que oferecem apenas consulta médica.
Clínicas Médicas	Latitude e longitude das clínicas médicas que oferecem consultas médicas, exames e pequenas cirurgias.
FATECs	Latitude e longitude das faculdades técnicas administradas pelo governo.
Poupatempo	Latitude e longitude das repartições públicas que emitem documentos e prestam outros serviços à população

Centros Populares	Latitude e longitude dos centros populares da cidade.
Museus	Latitude e longitude dos museus, privados e públicos.
Estações de Metrô	Latitude e longitude das estações de metrô.
Ônibus	Latitude e longitude de todos os pontos de ônibus do sistema de transporte público da cidade.

A segunda fonte de informação consiste na plataforma GeoSampa, mencionada anteriormente, que tem como foco a infraestrutura pública existente na cidade de São Paulo. Aqui, também, o foco foi em equipamentos de uso comum que não estavam disponíveis nas bases de dados do SEADE. Os dados de localização foram obtidos para as categorias apresentadas na Tabela 7.

Tabela 7
Variáveis utilizadas na base de dados do GeoSampa.

Variável	Descrição
Bicicletários	Latitude e longitude dos bicicletários públicos localizados na Cidade.
Ciclovias	Figura geométrica geolocalizada (linha) contendo a posição das ciclovias da cidade. As Ciclovias caracterizam-se como a melhor infraestrutura para bicicletas, com faixas exclusivas segregadas do trânsito, sinalização vertical e horizontal e pavimento específico.
Ciclofaixas	Figura geométrica geolocalizada (linha) contendo a posição das ciclovias

Variável	Descrição
	da cidade. As Ciclofaixas caracterizam-se por possuírem alguma infraestrutura para bicicletas, com faixas pintadas segregadas do trânsito e sinalização horizontal.
Ciclorrotas	Figura geométrica geolocalizada (linha) contendo a posição das ciclorrotas da cidade. As ciclorrotas são caracterizadas como ruas com baixo tráfego de automóveis, sendo bike friendly. Podem conter sinais horizontais.
Bus Rapid Transit (BRT)	Figura geométrica geolocalizada (linha) contendo a posição da infraestrutura de BRT da cidade.
Corredores Arteriais	Figura geométrica (linha) geolocalizada contendo a posição dos corredores Arteriais da cidade, aqueles que cortam a cidade e possuem grande fluxo de passageiros.
Ônibus Coletores	Figura geométrica geolocalizada (linha) contendo a posição dos corredores de Ônibus Coletores da cidade, aqueles que ligam regiões residenciais e comerciais.
Ônibus Locais	Figura geométrica (linha) geolocalizada contendo a posição dos corredores de Ônibus Locais da cidade, aqueles que geralmente ligam regiões distintas de um mesmo bairro.

Variável	Descrição
Árvores	Latitude e longitude das árvores em área urbana.
Imóveis e Locais Acessíveis	Latitude e longitude dos locais que atendem a todos os critérios de acessibilidade para pessoas com deficiência, principalmente em suas calçadas, e possuem selo de acessibilidade arquitetônica, reconhecido pela prefeitura.
Reparações no pavimento da rua	Latitude e longitude das reparações do pavimento da rua realizadas pela autarquia.

Conforme mencionado anteriormente, todos esses dados de características extrínsecas são geolocalizados, e serão avaliados segundo a sua proximidade com cada uma das propriedades estudadas. Os detalhes de como todas as informações foram coletadas e quais premissas foram adotadas serão abordados nos tópicos a seguir.

3.3 Unindo as Bases de Dados

Nota-se que todas as bases de dados citadas até aqui não são conectadas e não possuem nenhuma chave de ligação inequívoca. Analisando as Tabelas 3 a 7, apresentadas nos itens 3.1 e 3.2, pode-se observar a existência de pontos de convergência que permitem a junção das bases de dados. No entanto, esses pontos devem ser avaliados cautelosamente.

A ideia principal na lógica aqui adotada para a junção das bases de dados consiste em dividir o processo em etapas de um fluxo, em que cada uma das junções é processada sequencialmente, base a base, com a adoção de um critério conjunto. A Figura 1 ilustra a ordem de cada uma dessas etapas.

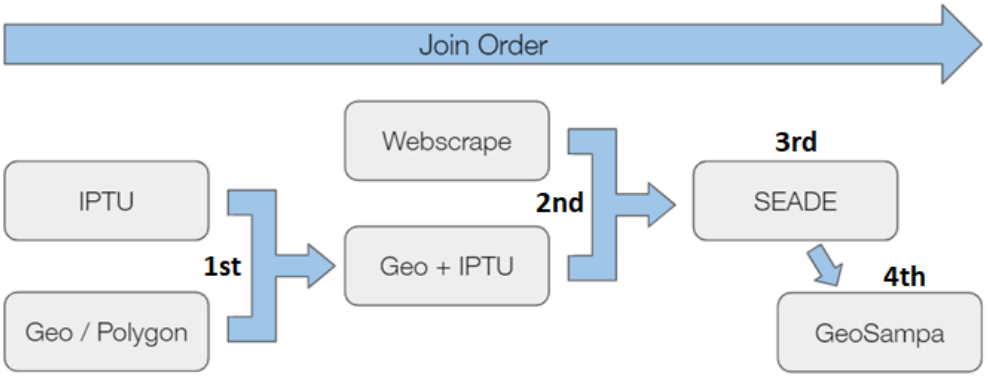


Fig. 1 – Diagrama de construção da base de dados

A Primeira Etapa da junção inicia-se com o banco de dados do IPTU, no qual foram adicionadas as geolocalizações dos polígonos representativos dos imóveis, utilizando para isso o CPF. Há uma particularidade a destacar: existe um número de contribuinte relativo ao terreno e outro relativo ao imóvel. Quando um terreno contém apenas uma propriedade, os números são iguais, mas se o terreno possuir mais de uma propriedade, um terreno possuirá mais de um número de contribuinte associado a ele. Por exemplo, prédios com vários apartamentos terão um CPF para cada unidade individual (presente na base de dados do IPTU). Já na respectiva base contendo os polígonos (banco de dados Geo/Polygon), há apenas um registro para cada edificação, pois o polígono representa o terreno e não cada apartamento.

Esse fato pode ser observado na formação dos números de contribuinte em ambos os casos, compostos pela combinação dos códigos das divisões administrativas do município. Quando se fala de terrenos ou residências, juntar as duas bases de dados (IPTU e Geo/Polígono) é simples, pois possuem o mesmo CPF. Porém, no caso de apartamentos, o número presente na base de dados do IPTU será formado pelo Código do Setor + Código do Bloco + Código do Lote Código da Unidade + Número do Condomínio dentro do bloco. Para uma mesma propriedade, o número do terreno no banco de dados Geo/Polygon, será formado apenas pela agregação dos Código do Setor + Código do Bloco + Código do Lote + Número do Condomínio no bloco. Ou seja, no segundo caso não há Código de Unidade formando o número.

Consequentemente, para ingressar nas bases de dados, o número do contribuinte é modificado na base de dados do IPTU, removendo o código da unidade toda vez que for identificado que o registro é um apartamento. Por fim, por meio de uma união que mantenha a integridade da base original nessa nova chave, as informações dos polígonos são trazidas para

o banco de dados do IPTU, garantindo que cada registro do IPTU tenha um polígono associado a ele.

Sumarizando:

- $if\ registro$ é $apartamento\ then$:

$$key = CodSector + CodBlock + CodLot + CodCond$$

- $else$: #casa, terreno

$$key = CodSector + CodBlock + CodLot + CodUnity$$

Após este primeiro passo, o próximo desafio é agregar as informações do banco de dados Webscrape. Inicialmente, é importante ter em mente que essas informações, por serem coletadas diretamente dos anúncios online apresentam diversas limitações, pois dependem das informações inseridas pelo usuário final no momento da criação do anúncio, ou seja, podem possuir erros de digitação ou mesmo informações incorretas. Portanto, é fundamental que essas informações sejam processadas e uma das primeiras ações nesse sentido é retirar do banco de dados todos os registros que possuem endereço incompleto, ou seja, sem o nome da rua ou o número do imóvel.

A Segunda Etapa da junção começa após este processo de limpeza. Com base no endereço completo disponível em cada anúncio e utilizando a biblioteca googlemaps em Python, é possível obter o CEP do imóvel, bem como sua Latitude e Longitude. Assim, com o CEP e o número do endereço, pode-se fazer a composição dos dois bancos de dados obtidos até o momento.

Porém, há mais um detalhe a ser considerado: no Brasil a segmentação dos CEPs apresenta problemas e limitações que podem levar a inconsistências. Por exemplo, existem ruas segmentadas em várias seções, cada uma com um CEP diferente. (Aranha, 1997). Este problema estava presente na biblioteca usada para obter códigos postais. Por essas ruas (com vários códigos postais) foi atribuído o "código postal padrão", desconsiderando o fato de que um pequeno trecho da rua tinha outro código postal.

Este fato pode levar a uma perda considerável de informações quando cruzadas com a base de dados "Geo IPTU", já que esta, por ser uma informação oficial do município, possui todos os CEPs corretos.

Para superar esse ponto, a partir do “Geo IPTU”, foi criado um dicionário contendo todos os CEPs e seus respectivos “CodLog” (conforme Tabela 4) do município. Embora uma mesma rua possa ter dois ou mais CEPs, ela nunca terá mais de um CodLog. O próximo passo foi associar, no banco de dados do Webscrape, as informações desse dicionário. Ao inserir, para cada CEP o seu respectivo CodLog, foi possível criar uma chave comum para ambas as bases de dados, composta pelo CodLog + Endereço + Número. A partir desta chave, foi implementada a junção interna dos dois bancos de dados.

Uma questão adicional deve ser destacada: Imagine dois apartamentos no mesmo prédio, um grande Rooftop e outro um pequeno Studio. Ambas as propriedades terão a mesma chave descrita acima, pois estão localizadas no mesmo endereço (CodLog + Endereço + Número). Se o anúncio da base de dados Webscrape fosse para o Studio, como seria possível atribuir o cadastro correto da base de dados Geo IPTU a este anúncio já que a chave não é única? A resposta é adotando uma condição de contorno usando a área do imóvel, assim sendo possível escolher, dentre todos os cadastros da base de dados Geo IPTU com esta mesma chave, aquele que tiver a área que mais se aproxima da área do apartamento Studio.

Dito isso, a união das duas bases de dados anteriores resultará na combinação de todas as possibilidades com a mesma chave, tanto na primeira base de dados (Geo IPTU), quanto na segunda base de dados (Webscrape). A condição de contorno descrita acima foi implementada calculando a diferença entre as duas variáveis, “Área Útil” (conforme Tabela 4 – do Geo IPTU) e “Área Construída” (conforme Tabela 3 – do Webscrape). Então é mantido apenas um registro por chave (CodLog + Endereço + Número), que é escolhido como aquele com a menor diferença (absoluta) entre as duas variáveis. Esse processo possui suas limitações, mas foi a forma encontrada para se obter a melhor aproximação dadas as restrições existentes nas bases de dados.

Para finalizar o processo de junção da Segunda Etapa, é feita uma validação final na base de dados resultante. A partir das informações dos polígonos, disponíveis no banco de dados resultante, todos os centróides dos respectivos polígonos são calculados com sua latitude e longitude. Eles podem ser vistos como os pontos amarelos na Figura 2. Em seguida, as distâncias entre esses pontos e os correspondentes à latitude e longitude obtidos do endereço completo (dos anúncios no banco de dados Webscrape) usando a biblioteca googlemaps – os pontos vermelhos em Figura 2 – são calculados. Foi adotado um limite de 300 metros devido ao tamanho médio das quadras da cidade, que é de aproximadamente 500 metros. Portanto, se

essa distância for maior que o limite, o registro é excluído. Com este procedimento, adiciona-se mais uma camada para garantir que a junção foi feita corretamente, sempre no mesmo imóvel.

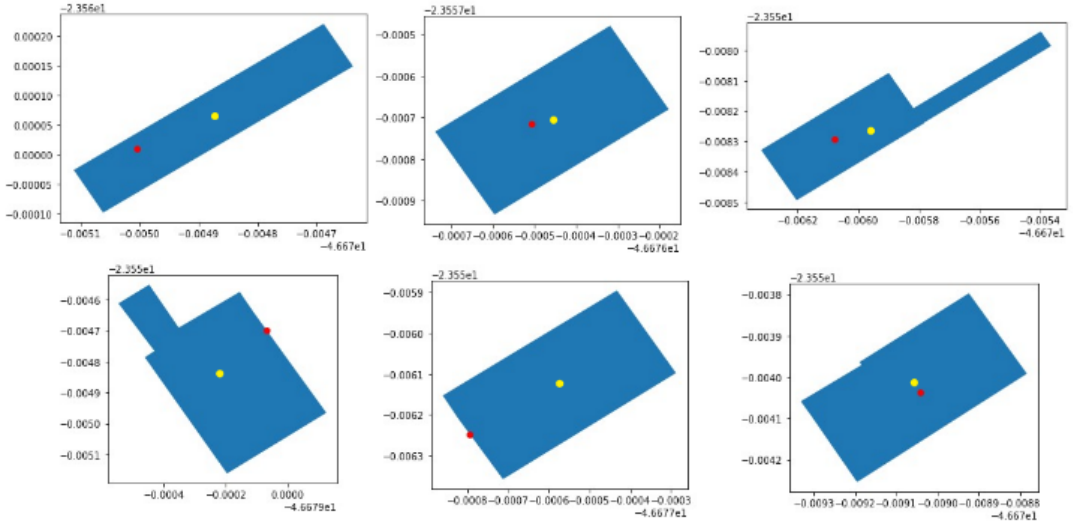


Fig. 2 – Exemplos de polígonos representando os terrenos da propriedade. Os pontos em amarelo são os centróides das figuras geométricas e os em vermelho são a geolocalização (lat/long) obtida com a biblioteca googlemaps a partir do endereço completo.

Então, o banco de dados contendo todas as características intrínsecas já mencionadas está pronto. Essa base se chamará “Geo IPTU Web”, passando para a Terceira Etapa do fluxo: agregar à base de dados Geo IPTU Web as informações sobre características extrínsecas disponíveis na base de dados do SEADE. Esta agregação é feita para cada uma das classes do banco de dados SEADE (conforme Tabela 6). A base de dados contém, para cada classe, as respectivas informações de latitudes e longitudes. Por exemplo, olhando para a classe “Universidades Públicas”, o banco de dados terá a latitude e longitude de todas as universidades públicas da cidade de São Paulo.

O processo de agregação dessas informações à base de dados “Geo IPTU Web” consiste em contar quantos pontos de cada classe existem nas proximidades de cada imóvel da base de dados “Geo IPTU Web”, considerando uma distância de até 500 metros do centroide do polígono (obtido anteriormente). Assim, tomando o exemplo anterior, o processo consistirá em contar quantas Universidades Públicas estão a 500 metros de cada imóvel. Neste processo foi adotada a distância de harversine. Contar a quantidade de pontos de cada classe que estão nas proximidades de um determinado imóvel é uma forma de quantificar o impacto da quantidade de itens de uma determinada classe no preço do imóvel.

Essa mesma lógica também foi adotada para incluir informações sobre características extrínsecas da base de dados do GeoSampa (Quarta Etapa), listando quantos itens de suas classes estão nas proximidades das propriedades. Existem apenas duas particularidades nesta base de dados. A primeira diz respeito ao fato de que determinadas classes, como Infraestrutura Rodoviária ou Infraestrutura Rodoviária, são linhas geolocalizadas na base de dados, contendo a latitude e longitude de seu contorno. Isso leva ao cálculo de quantas linhas existem nas proximidades de uma determinada propriedade, ao invés do cálculo de pontos, feito anteriormente. Além disso, uma vez que a geometria da linha está disponível no banco de dados do GeoSampa, seu comprimento total foi calculado, e essas informações foram agregadas ao cruzar os bancos de dados.

A segunda particularidade se deve ao fato de algumas classes estarem espalhadas pela cidade, por exemplo, há um pequeno número de paradas de BRT na cidade de São Paulo. Assim, a adoção de uma distância de contagem de até 500 metros, neste caso, resultaria em vários valores nulos, ou seja, muitos imóveis sem BRT em seu entorno. Portanto, diferentes distâncias para contar esses itens nas proximidades das propriedades foram adotadas, com base em cada classe, para obter menos valores nulos. A Tabela 8 contém os valores dessas distâncias.

Tabela 8

Distância de Contagem Adotada de cada variável no Banco de Dados GeoSampa.

Classe	Distância considerada	Justificativa
Árvores	Até 500 m	Alta densidade na cidade.
Bicicletários Ciclovias Ciclofaixas Ciclorrotas Problemas de Pavimentação	Até 1000 m	Densidade média na cidade.

BRT Faixas Arteriais Faixas Coletoras Faixas Locais Propriedades Acessíveis	Até 2.000 m	Baixa densidade na cidade.

Com esta última integração, todas as informações foram adicionadas ao banco de dados, garantindo a presença das características intrínsecas e extrínsecas que serão utilizadas no processo de modelagem.

3.4 Variáveis criadas

Todas as variáveis descritas nas Tabelas 3 a 7 dos itens anteriores foram obtidas diretamente das características contidas nos bancos de dados externos. Para melhorar o retrato de certas características das propriedades por meio de variáveis, algumas operações matemáticas, como escala e conversão, foram realizadas nos dados obtidos até o momento.

O primeiro exemplo é a variável objetiva a ser estudada: o preço por metro quadrado. Wolverton (1997) demonstrou que existe uma dependência direta entre o preço absoluto de um imóvel e sua área, ou seja, em um processo de modelagem do preço absoluto, a área do imóvel seria o principal fator determinante, ofuscando os demais fatores. Para eliminar esse problema, a estratégia aqui adotada foi dividir o preço pela área, obtendo o preço por metro quadrado de cada imóvel, que será a única variável de saída modelada através das características intrínsecas e extrínsecas do imóvel.

Além disso, certas variáveis presentes no banco de dados podem trazer uma representação mais precisa de sua influência quando são ponderadas. Exemplo disso são os dados relacionados à infraestrutura para bicicletas e ônibus. Encontrar a influência do número de ciclovias no entorno de um imóvel sobre seu preço, por exemplo, pode ser melhor representado se esse número for ponderado pelo comprimento total de cada ciclovia presente na área. Por exemplo, imagine que uma propriedade tenha três ciclovias muito curtas em sua vizinhança e outra propriedade tenha apenas uma ciclovia longa. Para ser consistente com a

representação do impacto dessas ciclovias no preço das duas propriedades, é necessário ponderar o número de ciclovias pelo comprimento das ciclovias. Assim, dessa forma, é possível quantificar se as ciclovias próximas são curtas, conectando assim menos regiões da cidade, ou se são longas, servindo como uma alternativa de transporte mais viável.

Por outro lado, modelos de Machine Learning podem ter seu desempenho consideravelmente melhorado quando as variáveis (características) são escalonadas para valores médios, no caso de variáveis que possuem ordem de grandeza maior que outras. (Juszczak *et al.*, 2015). No caso desse modelo, todas as feições que possuem ordem de grandeza cem vezes o valor médio global foram dimensionadas para evitar viés no modelo.

Por fim, para melhorar a representação das variáveis para quartos e banheiros, foi criada uma razão entre esses dois números. Calculando esta razão para estas duas variáveis existe uma forma de quantificar o impacto dos quartos com casa de banho, penalizando os imóveis que têm muitos quartos mas não têm banheiros suficientes. (Heidari *et al.*, 2021). A Tabela 9 mostra um resumo das variáveis criadas no banco de dados.

Tabela 9

Novas variáveis criadas baseadas em outras variáveis da base de dados

Variável	Cálculo	Descrição
Preço/m²	$Preço\ /\ Área\ Útil$	Variável objetivo do modelo.
Número ponderado de Ciclovias, Ciclovias, Ciclovias, BRT, Arterial, Coletor e Locais de Ônibus.	$N°_i * Total\_comprimentoh_i\ /\ Global\_avg(length)$ Por exemplo: $N°_{Ciclovias\ nas\ proximidades\ da\ prop}\ /\ Glob\_avg\ (comprimento\ total\ das\ ciclovias)$	Número ponderado de cada infraestrutura por seu respectivo comprimento, relativo à média global no banco de dados.
Quantidade de Árvores *Escaladas*, Consultórios *Escaladas*de	$Recurso\ /\ Global\_avg(Recur$	Variáveis dimensionadas por sua respectiva

Clínicas Médicas , *Escaladas*de Reparos em Pavimentos .		média global no banco de dados.
Número de quartos por Banheiros	$N°_{de\ quartos} / N°_{de\ banheiros}$	Número de quartos em relação aos banheiros.

Em todos os casos acima, as variáveis criadas substituem as respectivas variáveis utilizadas em seu cálculo. Por exemplo, o "Número ponderado de ciclovias" substitui tanto o "Número de ciclovias" quanto o "Comprimento de cada ciclovia" no modelo final, portanto, não há contagem dupla. Nos itens a seguir, será mostrada a base de dados resultante e as variáveis consideradas no processo de modelagem.

3.5 Base de dados resultante

Após realizar todos os Passos descritos na Figura 1, bem como incluir as variáveis, criadas ou modificadas, conforme descrito no item 3.4, o banco de dados resultante estava completo. Esta base de dados servirá de input para a implementação dos modelos de inteligência de máquina. Contém 44 variáveis que representam características do modelo (entradas) e 1 variável objetiva (saída). Um resumo dessas variáveis é apresentado na Tabela 10.

Tabela 10

Variáveis na base de dados resultante, para serem usadas no modelo.

Variável	Tipo
Preço por Metro Quadrado	Variável resposta
Preço Condomínio por Metro Quadrado	Variável de entrada
Área Construída	Variável de entrada
Quartos para Banheiros (rácio)	Variável de entrada
Nº de Garagens	Variável de entrada
Nº de Esquinas	Variável de entrada

Fração Ideal	Variável de entrada
Terreno	Variável de entrada
Área Usada	Variável de entrada
Ano de Construção	Variável de entrada
Fator de Obsolescência	Variável de entrada
Nº de Andares	Variável de entrada
Comprimento Frontal	Variável de entrada
Nº de UBS	Variável de entrada
CREAS	Variável de entrada
CRAS	Variável de entrada
Nº de Escolas Particulares	Variável de entrada
N.º de Escolas Públicas Estaduais	Variável de entrada
N.º de Escolas Públicas Municipais	Variável de entrada
N.º de Escolas Públicas Federais	Variável de entrada
N.º de Escolas (outras)	Variável de entrada
N.º de FATECs N.º de FATECs	Variável de entrada
N.º de Universidades Privadas	Variável de entrada
N.º de Públicas Universidades	Variável de entrada
Nº de Museus	Variável de entrada
Nº de *Poupa tempo* Escritórios	Variável de entrada
N.º de Centros Populares	Variável de entrada
N.º de Hospitais Públicos	Variável de entrada
N.º de Hospitais Privados	Variável de entrada
N.º Escalado de Consultórios Clínicos	Variável de entrada
N.º Escalado de Clínicas	Variável de entrada
N.º Escalado de Reparação de Pavimentos	Variável de entrada
Número Escalado de Árvores	Variável de entrada
Número de Imóveis Acessíveis	Variável de entrada
Número de Bicicletários	Variável de entrada
Número de Estações de Metrô	Variável de entrada
Número de Paradas de Ônibus	Variável de entrada

Número Ponderado de Ciclovias	Variável de entrada
Número Ponderado de Ciclofaixas	Variável de entrada
Ponderado Número de Ciclorrotas	Variável de entrada
Número Ponderado de Faixas de Ônibus Locais	Variável de entrada
Número Ponderado de Faixas de Ônibus Coletor	Variável de entrada
Número Ponderado de Faixas de Ônibus Arteriais	Variável de entrada
Número Ponderado de BRT	Variável de entrada

O banco de dados resultante contém 4.019 registros, cada um correspondendo a um único e único Imóvel e reunindo todas as informações descritas na Tabela 10. A Figura 3 mostra a distribuição desses imóveis pelos bairros da cidade, discutidos anteriormente na Tabela 1.

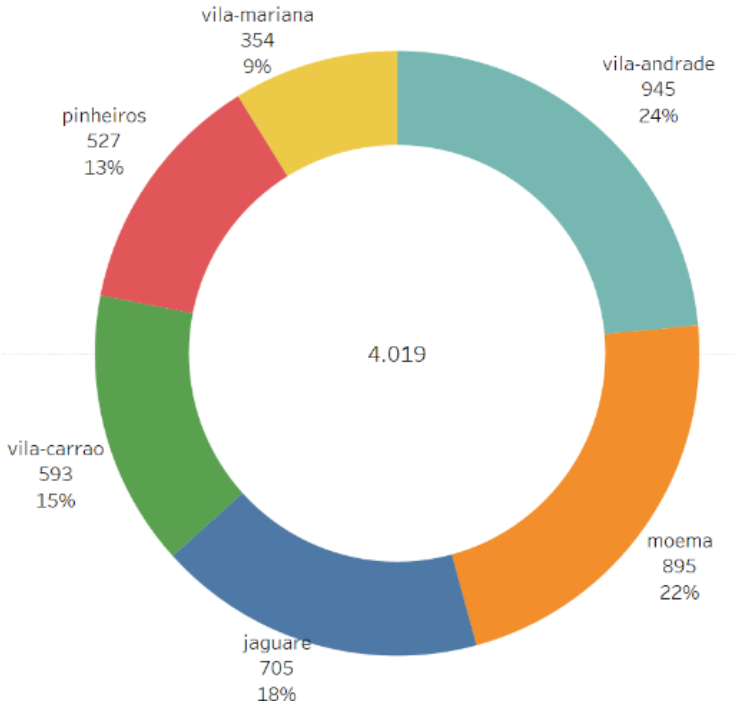


Fig. 3 – Propriedades Imobiliárias no Banco de Dados Resultante de acordo com os Distritos a que pertencem.

Vale ressaltar que o processo de webscraping do site VivaReal reuniu 5.479 registros na base de dados inicial (Webscrape Database) e, após todas as regras e validações para junção

dessas bases totalmente distintas, restaram 4.019 registros, resultando em um índice de assertividade para o processo de união de cerca de 73%.

4. Análise Exploratória dos Dados

Inicialmente, é relevante analisar e explorar o dataset obtido no último item de forma a entender o relacionamento das variáveis com o preço por m2, obtendo direcionamento inicial de quais são os relacionamentos mais relevantes entre variáveis independentes e a variável resposta.

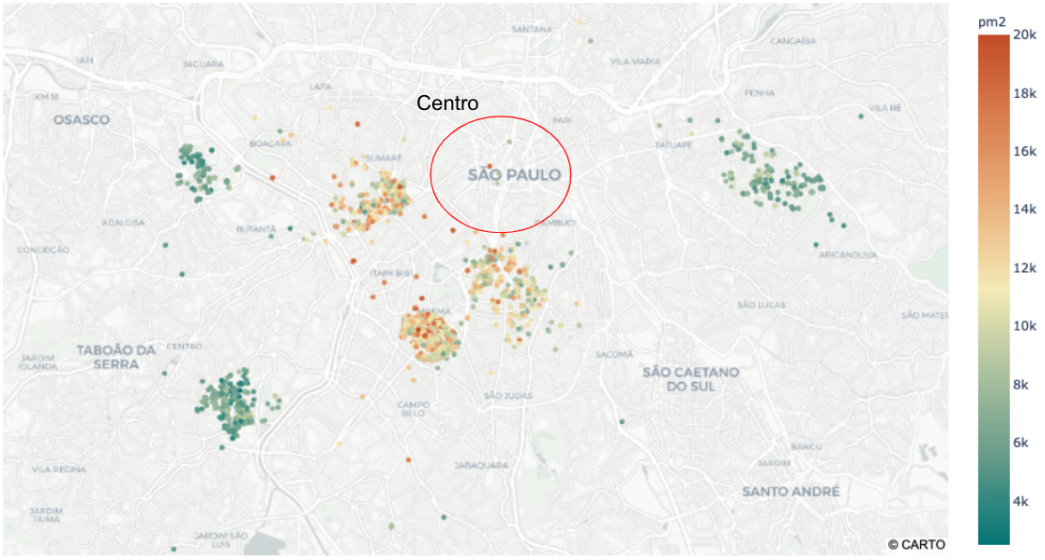


Fig. 4 – 4019 propriedades imobiliárias geolocalizadas em São Paulo, Brasil. A cor representa o preço por m2

Ao explorar a distribuição espacial, temos que os bairros mais centralizados da cidade tendem a possuir preços por m2 mais elevados, enquanto os bairros mais distantes têm a propensão a preços por m2 inferiores. Essa tendência de variação de preços em função da distância do centro da cidade já foi abordada e indicada por D'acci (2019) na cidade de Torino, citado anteriormente neste trabalho.

De forma mais sistemática, na figura 5 é possível analisar a correlação entre cada uma das variáveis extrínsecas e intrínsecas e o preço por m2, de forma a entender relacionamentos que sejam mais ou menos relevantes.

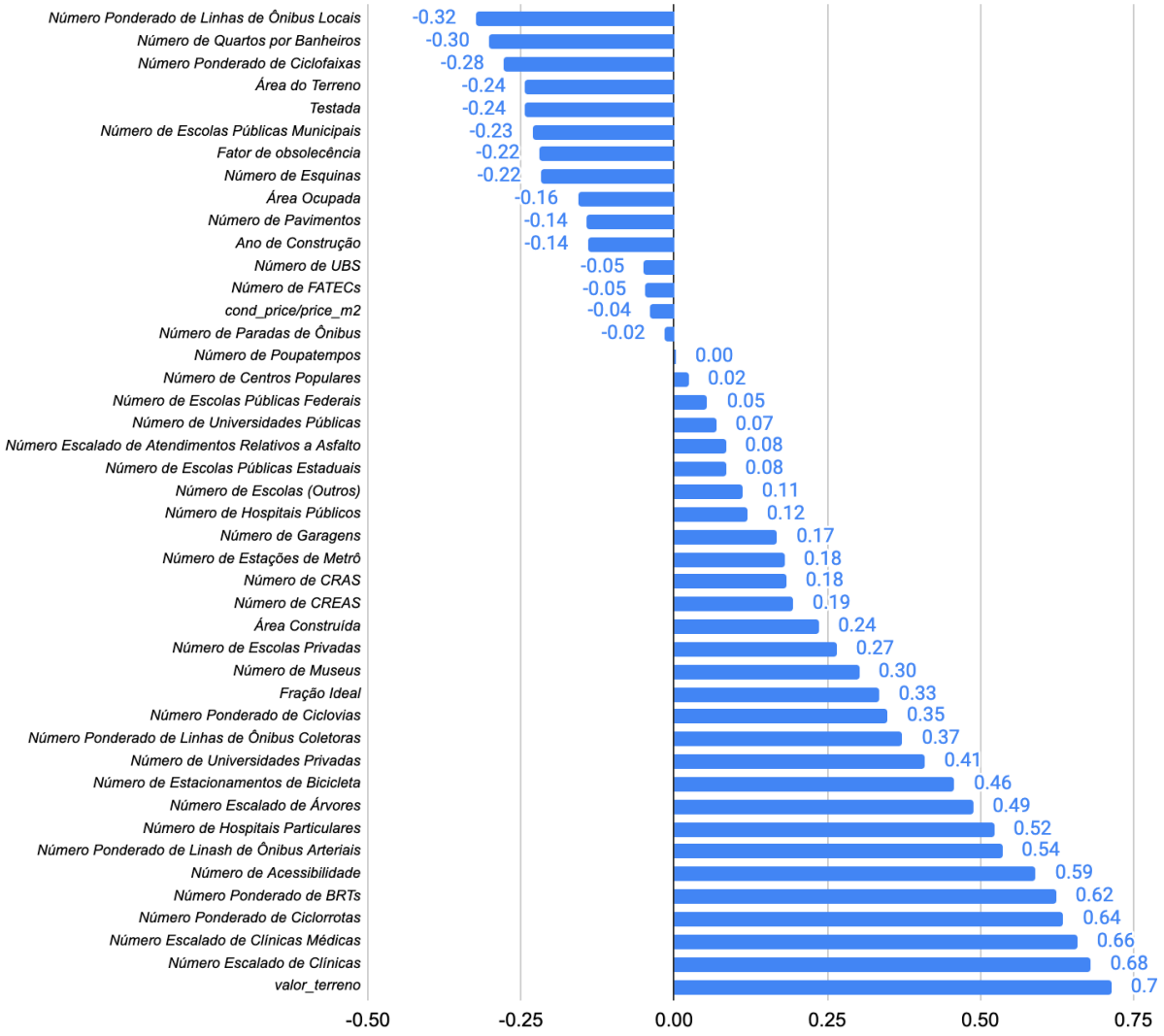


Fig. 5 – Correlação entre cada variável e preço por m2

Alguns pontos são importantes de serem observados no gráfico anterior. Em primeiro lugar, há maior número de correlações positivas do que correlações negativas. Em módulo, as correlações positivas também são maiores, tendo 5 variáveis que superam 0.6. É importante ressaltar, porém, que Schober *et al.* (2018) indicam que, por exemplo, um coeficiente de correlação de 0,65 pode ser interpretado como uma correlação "boa" ou "moderada", dependendo da regra prática aplicada. Também deve ser afirmado com cautela que um coeficiente de correlação de 0,39 representa uma associação "fraca", enquanto 0,40 é uma associação "moderada". Os coeficientes, portanto, devem ser usados apenas como referência, mas não como base para uma análise mais aprofundada

Em segundo lugar, temos importantes correlações tanto entre variáveis intrínsecas, como por exemplo a área construída e a relação entre quartos e banheiros, quanto extrínsecas, como exemplo o número de clínicas e de BRTs, em relação ao preço por m2, reforçando uma tendência de influência conjunta da localização e das características próprias de cada apartamento, como observado no trabalho de D'acci (2019).

Em terceiro lugar, pode-se observar a variável de maior destaque intrínseca como o valor do terreno, com correlação de 0.71. É importante destacar que ela possui certo grau de viés em relação ao preço do apartamento, uma vez que o preço do terreno é componente do preço da propriedade construída nele. No entanto, uma vez que o objetivo do trabalho é exatamente encontrar propriedades com valores de listagem inferiores aos de mercado, utilizar uma variável com alta correlação que não depende diretamente da variável resposta não torna a análise menos válida.

A fim de aprofundar a análise exploratória, é fundamental começar construindo segmentações dos dados. Uma das divisões mais naturais e amplamente utilizadas em estudos de precificação de propriedades é a divisão por distritos. Essa abordagem se baseia na premissa de que os distritos representam uma unidade territorial relevante. De acordo com Cazzolato (2005), distritos podem ser definidos como porções espaciais claramente delimitadas, reconhecidas e apropriadas sob três perspectivas principais: o substrato físico-urbano, as estruturas territoriais e locais, e a perspectiva dos cidadãos. Essa divisão por distritos é especialmente relevante como uma unidade de divisão territorial.

Assim, a divisão por distrito foi realizada, gerando a distribuição boxplot de preço para cada um, como pode ser visto na figura 6 a seguir.

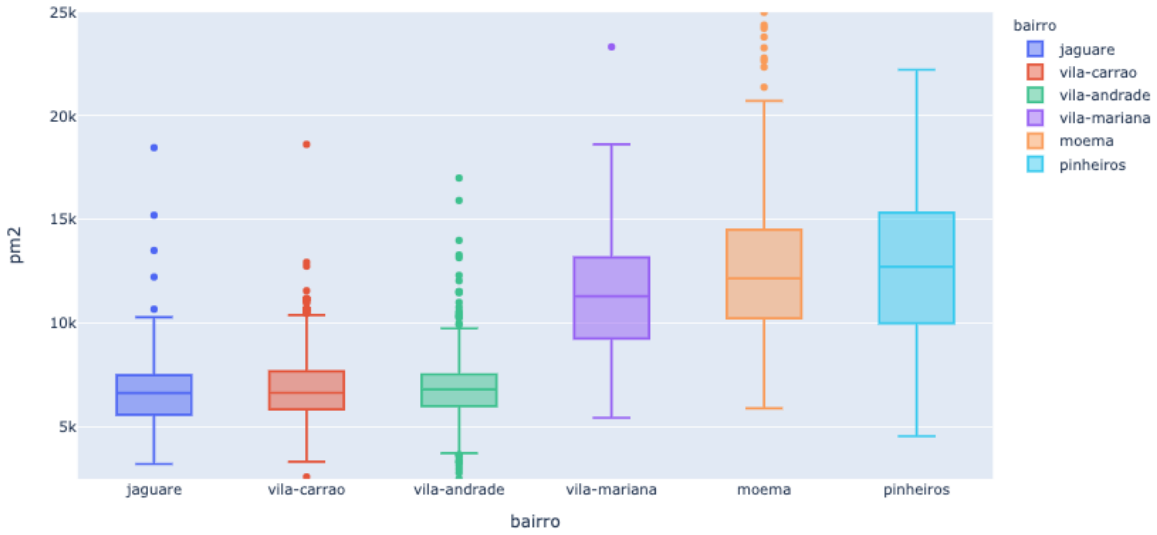


Fig. 6 – Distribuição Boxplot do preço por m2 para os 6 distritos de São Paulo considerados

De acordo com o que foi trazido no capítulo 3, vemos que existe uma divisão bastante nítida de preços entre esses distritos, com um grupo de preços natural formado entre Jaguaré, Vila Carrão e Vila Andrade, enquanto Vila Mariana, Moema e Pinheiros formam outro grupo. Isso pode ser observado também por índices oficiais. Segundo o ranking da prefeitura de São Paulo (PMSP, 2019), Vila Mariana, Moema e Pinheiros se encontram entre os 10 distritos menos prioritários em uma lista com 96 no que tange a vulnerabilidade de famílias. Por outro lado, Vila Andrade se encontra como o 33 em maior prioridade, Jaguaré em 51 e Vila Carrão em 81.

Para além da análise distrital, pode-se adentrar no relacionamento de variáveis intrínsecas específicas de forma exploratória, iniciando pela variável de maior correlação negativa, Número de Quartos por Banheiros. É interessante notar que essa variável possui uma tendência de relacionamento com o preço por m2 inversamente proporcional. De forma intuitiva, é possível entender que um apartamento que possua apenas um banheiro para dois ou mais quartos oferecerá conforto menor a seus moradores do que aquele que possui número menor de quartos para cada banheiro. Além disso, quando essa variável atinge o valor de um, podemos entender que o apartamento possui suítes ou, pelo menos, um banheiro para cada quarto.

É interessante notar que, quando comparadas as medianas de cada grupo definido para análise, pode-se observar que o preço por m2 varia de R$6.600 para mais de 2 até R$11.200,

quando essa variável vale entre 0 e 0,5, com a ressalva que a categoria com menos de 0,5 quarto por banheiro tende a pertencer a apartamentos de luxo.

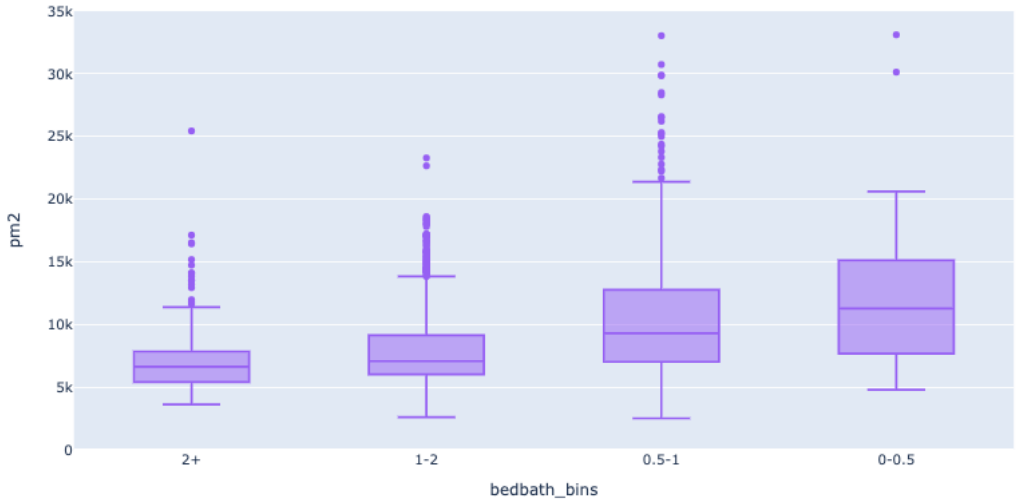


Fig. 7 – Distribuição Boxplot do preço por m2 para 4 categorias da variável Número de quartos por banheiros

Outra variável relevante ao preço do apartamento é seu número de garagens. Tendo em vista que a garagem no Brasil consta no registro de imóveis como área do apartamento, é lógico imaginar que quanto maior o número de garagens, maior o preço total do apartamento. No entanto, pelo gráfico de distribuição de preços por faixa de número de garagens apresentado na figura 8, é observado que o preço por m2 também é influenciado de forma diretamente proporcional, em que quanto mais garagens, maior a mediana de preços por m2 do grupo. De forma semelhante à distribuição de Quartos por Banheiros, apartamentos com seis ou mais garagens possuem uma mediana de preços muito superior aos outros, correlacionando com apartamentos de luxo.

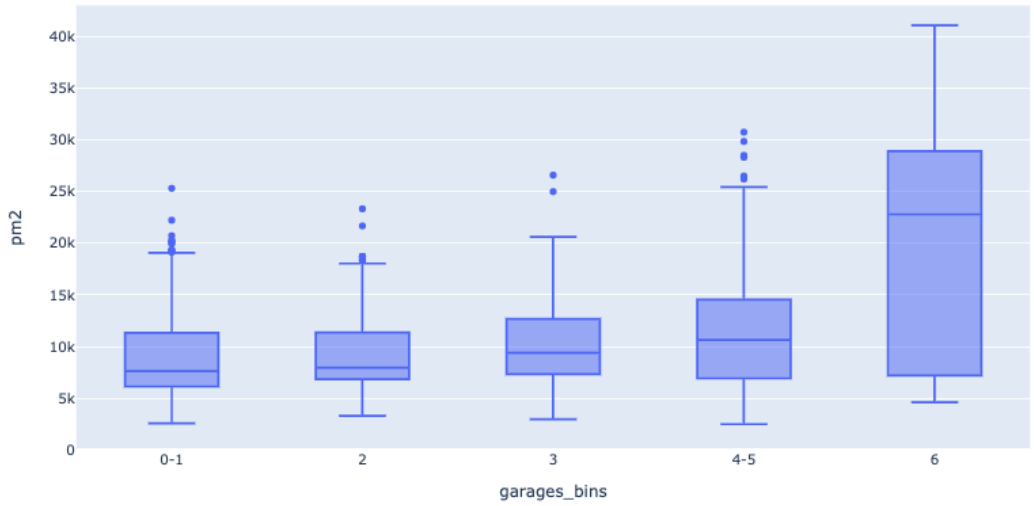


Fig. 8 – Distribuição Boxplot do preço por m2 para 5 categorias da variável Garagens

É analisado também o relacionamento entre a área total do apartamento e seu preço por m2. Observa-se na figura 9 existir uma tendência crescente, em que grupos de menores áreas possuem preços por m2 menores e uma menor dispersão de forma geral. No entanto, uma exceção se destaca. Quando são observados apartamentos muito pequenos, até 25m2, é encontrada uma distribuição de preço por m2 superior. Isso se deve principalmente aos studios, que são apartamentos pequenos que são bastante valorizados em São Paulo. É importante notar que parte do efeito é causado, também, por erros na digitação da área. Isso poderá ser explorado de forma mais nítida na modelagem de dados do próximo item.

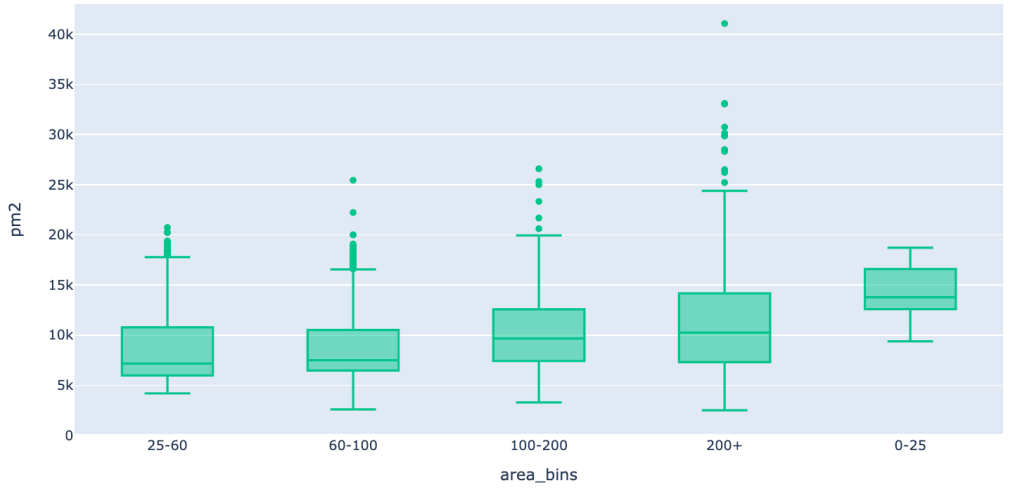


Fig. 9 – Distribuição Boxplot do preço por m2 para 5 categorias da variável Área Total

Na análise de variáveis extrínsecas relacionadas a instalações públicas e privadas na cidade de São Paulo, destacam-se insights valiosos ao segmentar esses elementos por bairro. A proximidade de instituições de ensino, serviços de saúde e opções culturais emerge como desejável, como confirmado por Wen et al. (2018), cujos resultados indicam que, sob a perspectiva da qualidade educacional ou acessibilidade, escolas, bem como universidades, exercem impacto significativo nos preços imobiliários.

Ao analisar os bairros e a presença de pelo menos uma dessas instalações públicas em um raio de 500 metros, observa-se uma correlação positiva entre a proximidade de universidades privadas, hospitais e museus e valores imobiliários mais elevados, conforme evidenciado na tabela a seguir.

Tabela 11

Informação sobre instalações públicas e privadas por propriedade em cada distrito

Bairro	Preço/m2 médio (R$)	Ao menos uma Universidade Particular	Ao menos um Hospital	Ao menos um Museu
Jaguaré	6619	39%	5%	20%
Moema	12627	99%	100%	3%
Pinheiros	12661	92%	94%	96%
Vila Andrade	6813	82%	69%	1%
Vila Carrão	6966	10%	67%	1%
Vila Mariana	11374	97%	97%	79%

Pode-se observar que nos bairros com o preço por metro quadrado mais elevado, a maioria das propriedades está localizada em proximidade de pelo menos uma universidade particular, hospital ou museu. Notavelmente, o distrito de Pinheiros se destaca, apresentando a maior proporção de propriedades com pelo menos uma dessas instituições, enquanto o Jaguaré se destaca de forma negativa. Curiosamente, esses dois distritos representam os extremos no que diz respeito ao preço por metro quadrado, com Pinheiros representando o maior preço e Jaguaré o menor. Embora esse dado não atinja significância estatística, oferece uma visão interessante da situação, destacando com dados um padrão intuitivo.

Direciona-se a análise à variável ano de construção. Ao examinar a correlação, representada na Figura 5, observa-se um valor de -0,14, indicando uma correlação negativa. Esse resultado contraria a suposição comumente aceita de que propriedades mais novas tendem a ter valores mais elevados, justificando, assim, a necessidade de uma análise mais aprofundada.

Ao explorar essa variável de acordo com os bairros, identifica-se um fenômeno intrigante. A divisão da variável ano de construção por bairro revela que bairros com médias de preço mais elevadas apresentam medianas de ano de construção mais baixas. Os resultados são apresentados na tabela a seguir, lançando luz sobre essa correlação inesperada.

Tabela 12

Análise da variável ano de construção por distrito

Bairro	Média (com valores nulos)	Média (sem valores nulos)	Mediana	% de valores nulos	Preço por m2 (R$)
Jaguaré	1967	2004	2009	1.84%	6616
Moema	1935	1991	1989	2.79%	12695
Pinheiros	1815	1981	1973	8.35%	12833
Vila Andrade	1961	2007	2012	2.33%	6802
Vila Carrão	2000	2000	2003	0.00%	6966
Vila Mariana	1904	1989	1989	4.24%	11476

É importante destacar, em primeiro lugar, a presença de valores nulos em alguns bairros, como Pinheiros e Vila Mariana, sendo respectivamente 8.35% e 4.24%. A proporção total de valores nulos na base de dados é de 2,96%, uma porcentagem relativamente baixa, que permite a eliminação dessas propriedades sem valor, a fim de aprimorar a precisão da análise. Esse raciocínio se tornará mais evidente na próxima explicação.

Em uma análise inicial, desprovida de contexto, poderia-se erroneamente inferir que uma correlação negativa entre o ano de construção e o preço por metro quadrado indica que propriedades mais antigas são mais caras, levando a criar hipóteses de causalidade. No entanto, essa análise revela que, ao comparar diferentes bairros, a diferença no ano de construção não afeta significativamente os preços.

No entanto, ao examinar internamente cada bairro, observa-se uma correlação positiva entre o ano de construção e o preço por metro quadrado, especialmente após a exclusão dos valores nulos da base de dados. Uma observação notável diz respeito à alteração na média dos valores devido a uma pequena quantidade de valores nulos. Por exemplo, no bairro do Jaguaré, apenas 1,84% de valores nulos quando eliminados alteram a média do ano de construção de 2004 para 1967. Isso muda completamente a interpretação dos dados, destacando a importância de analisar métricas complementares, como a mediana e % de valores nulos. Os resultados estão detalhados na tabela a seguir.

Tabela 13

Correlação ano de construção e preço/m2 com e sem valores nulos

Bairro	Correlação ano de construção e preço/m2 com valores nulos	Correlação ano de construção e preço/m2 sem valores nulos
Jaguaré	0.04	0.54
Moema	-0.09	0.43
Pinheiros	-0.17	0.24
Vila Andrade	0.05	0.35
Vila Carrão	0.41	0.41
Vila Mariana	-0.15	0.25

Um achado notável que merece destaque é a decisão de eliminar 119 amostras que continham valores nulos para o ano de construção. Essa intervenção teve um impacto drástico na análise, transformando o coeficiente de correlação geral entre ano de construção e preço por m2 de -0,14 para 0,18. Essa mudança substancial alterou significativamente a interpretação dos dados e, ao mesmo tempo, resultou em melhorias no desempenho dos modelos de previsão, como será detalhado no próximo item deste trabalho. Esse exemplo enfatiza a importância do conhecimento profundo da base de dados por parte do pesquisador, indo além da simples aplicação de modelos sem uma análise criteriosa das variáveis.

É crucial reconhecer que a análise exploratória de dados deve ser interpretada dentro de seu contexto. Embora alguns relacionamentos possam ser identificados, o propósito da exploração dos dados é, ao mesmo tempo, destacar variáveis de interesse e identificar possíveis inconsistências na base de dados. Por exemplo, ao examinar a variável "Suítes", observou-se que todos os valores estavam ausentes. Portanto, dentro do contexto atual, essa variável foi considerada irrelevante e, consequentemente, foi excluída da análise.

Com essa etapa concluída, podemos avançar para a modelagem proposta, com o objetivo de identificar relações estatísticas robustas entre as variáveis independentes e a variável de resposta.

5. Modelagem Proposta

O objetivo deste estudo é identificar e explorar oportunidades no mercado imobiliário. Para isso, é essencial começar com uma definição clara. Uma oportunidade no mercado imobiliário é definida como uma propriedade cujo preço está abaixo do seu valor de mercado atual. Teoricamente, uma oportunidade é considerada mais promissora quanto maior for a diferença entre o valor listado e o valor de mercado avaliado por modelos específicos.

Consequentemente, a metodologia adotada baseia-se na avaliação do valor de mercado de cada propriedade, por meio de diversas técnicas de Aprendizado de Máquina. Posteriormente, esses valores são comparados com os preços de listagem reais. As oportunidades são identificadas quando a discrepância entre o valor listado e o valor estimado ultrapassa um determinado limite. Esse limite é estabelecido considerando as limitações e possíveis erros associados a cada modelo de avaliação.

É relevante observar a relação direta entre o conceito de oportunidade, conforme definido anteriormente, e o conceito de outlier. De acordo com Aggarwal (2017), um outlier é uma amostra que se diferencia significativamente dos demais dados. Na área de mineração de dados e estatística, os outliers também são denominados anomalias, desvios ou discrepâncias.

No entanto, é crucial distinguir de forma clara esses dois conceitos. Outliers costumam ser indesejados e podem ocorrer tanto em valores excessivamente altos quanto baixos. Como afirmado por Osborne & Overbay (2004), outliers são pontos de dados extremos que podem levar a taxas de erro inflacionadas e distorções substanciais nas estimativas de parâmetros e estatísticas, e devem ser sempre verificados, pois sua presença pode afetar significativamente a precisão e os erros de inferência.

Em contrapartida, é fundamental destacar que as oportunidades no contexto em análise são altamente desejáveis e representam uma situação na qual os valores de listagem de propriedades estão estritamente abaixo dos preços de mercado estimados por meio de modelos específicos. Isso significa que essas oportunidades indicam propriedades que são possíveis alvos atrativos para investidores e compradores em busca de negócios vantajosos.

Embora a abordagem para identificar oportunidades seja semelhante à utilizada para detectar outliers, é estabelecido um método consistente para diferenciar essas duas possibilidades. Inicialmente, a metodologia seleciona as propriedades com preços inferiores aos valores de mercado. A partir desse ponto, é realizada uma análise minuciosa de cada

conjunto de dados da propriedade. Posteriormente, é executada uma avaliação manual. Isso resulta em uma lista de oportunidades com alto grau de confiabilidade.

Uma das ferramentas amplamente utilizadas para avaliar se uma propriedade está precificada de forma atípica ou não é a análise das transações ocorridas no mesmo edifício. Essa informação, disponível em alguns websites da internet, é obtida diretamente da base de dados do Imposto de Transmissão de Bens Imóveis (ITBI) e desempenha um papel relevante na análise de mercado. Isso ocorre porque essas transações antigas no mesmo edifício proporcionam uma visão valiosa sobre propriedades com características semelhantes, variando apenas em aspectos como o andar, área, estado de conservação e detalhes legais, mas mantendo as características extrínsecas iguais. Essas informações refinam a análise e fornecem uma base sólida para tomar decisões mais informadas e seguras no mercado imobiliário.

Ressalta-se que, uma vez que a lista de oportunidades é uma ferramenta para identificar possíveis bons negócios, surge uma etapa subsequente relevante para investidores ou compradores. Essa etapa envolve a verificação in loco da propriedade para confirmar sua existência e avaliar a ausência de fatores que possam impactar a compra. Embora o presente trabalho coloque sob escrutínio um amplo espectro de variáveis das propriedades em questão, ele não considera aspectos como o estado de conservação da propriedade, regras de condomínio e variáveis externas à propriedade, como ruído, vista e segurança, entre outros. Compete ao tomador de decisões ponderar esses fatores adicionais e incorporar sua perspectiva subjetiva à análise.

5.1 Passo a passo proposto

A metodologia que é proposta neste trabalho pode ser observada no esquema a seguir.

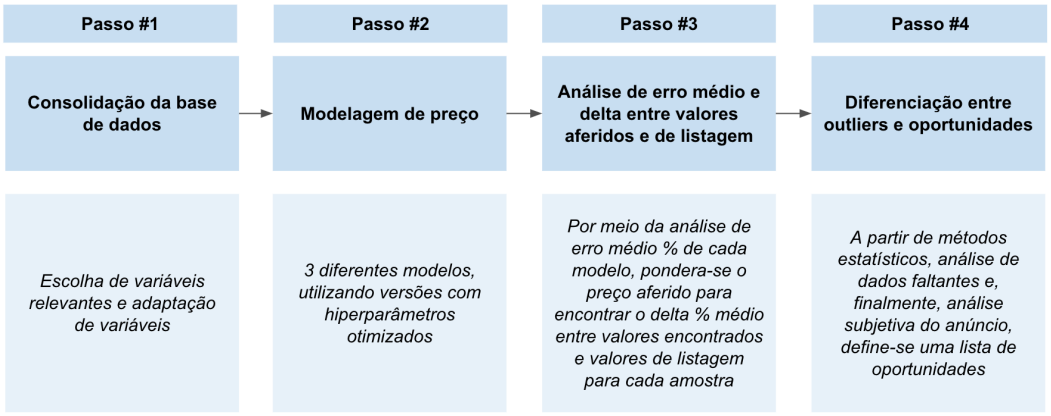


Fig. 10 – Metodologia proposta para encontrar oportunidades no mercado imobiliário

Em suma, 4 passos são executados com objetivo de encontrar oportunidades na amostra apresentada no item 3 deste trabalho.

Em primeiro lugar, a base de dados do item 3 é selecionada, com suas 3900 amostras, eliminando todas as variáveis não numéricas e segmentando a base em variáveis independentes e variável dependente.

Em segundo lugar, os modelos utilizados na avaliação de preço são escolhidos, segundo a literatura trazida no capítulo 2 deste trabalho. O primeiro modelo a ser utilizado é um método hedônico. A forma funcional da equação de regressão hedônica pode ser linear ou logarítmica. Esse trabalho usará a versão linear da regressão, como pode ser visto na equação a seguir:

$$P_i = \beta_0 + \sum_{i,j} \beta_j S_{ij} + \varepsilon$$

Eq. 1 – Forma Linear do MPH.

Nessa equação, Pi é o preço por metro quadrado de cada imóvel. Sij representa os valores para cada [j] características e [i] propriedades imobiliárias. β0 é o termo de interceptação e βj são os coeficientes estimados para as variáveis descritas na Tabela 10.

O segundo modelo proposto é o XGBoost, uma biblioteca open-source que se baseia no framework de gradient boosting. O gradient boosting é uma técnica de aprendizado de máquina que tem recebido destaque pela eficácia na construção de modelos preditivos

precisos e robustos. Nesse processo, um modelo inicial é ajustado aos dados e, subsequentemente, outros modelos são criados para corrigir erros anteriores, refinando gradualmente o modelo final capaz de capturar nuances nos dados e lidar com problemas de alta dimensionalidade. O XGBoost aprimora essa técnica permitindo a otimização paralela de grupos de árvores, o que intensifica a capacidade de generalização do modelo e acelera o treinamento. No âmbito deste trabalho, a aplicação do XGBoost é cuidadosamente considerada como parte da metodologia adotada para encontrar oportunidades de precificação de propriedades imobiliárias, destacando sua contribuição para o aprimoramento da análise e identificação de oportunidades.

No contexto do terceiro modelo, o KNN (K-Nearest Neighbors), a abordagem é inspirada na simplicidade e eficácia na proximidade entre os pontos de dados. Ele classifica ou faz previsões atribuindo uma observação à classe mais frequente entre seus k vizinhos mais próximos, onde k é um parâmetro determinado. Cada ponto de dados é tratado como um ponto em um espaço multidimensional, e a distância entre esses pontos é calculada, geralmente usando a distância Euclidiana. Sua simplicidade é uma de suas principais características, tornando-o eficaz para tarefas de regressão, especialmente em cenários com dados não-lineares ou sem distribuição específica. No contexto deste trabalho, a aplicação do algoritmo KNN contribui para a análise de preços das propriedades imobiliárias em questão, particularmente útil para identificar tendências em dados de distribuição complexa.

Em uma etapa subsequente da modelagem, **em terceiro lugar**, será executada a comparação dos preços listados com os valores estimados pelos modelos. Nesse contexto, o foco será identificar as disparidades mais significativas entre esses valores, **as quais serão ponderadas inversamente em relação ao erro percentual absoluto médio (MAPE na sigla em inglês)**, resultando na determinação de um delta. A equação subjacente ao cálculo desse erro médio, para cada modelo, é a seguinte.

$$\text{MAPE}(y, \widehat{y}) = \frac{1}{n} \sum_{i=i}^{n} \frac{|y_i - \widehat{y_i}|}{max(\boldsymbol{\varepsilon}, |y_i|)}$$

Eq. 2 – Cálculo do Erro médio quadrático percentual.

O cálculo do delta médio percentual, chamado de métrica de desvio e avaliado pelo algoritmo, será realizado segundo a equação a seguir, uma variação da equação de média ponderada.

$$\Delta_{Médio\ \%} = \frac{\sum_{i=i}^{N} \left(|P_{listagem} - P_{modelo\ i}| * MAPE_i^{-1} \right)}{P_{listagem} * \sum_{i=i}^{N} MAPE_i^{-1}}$$

Eq. 3 – Cálculo do delta médio percentual (métrica de desvio), sendo i cada um dos modelos utilizados.

A premissa para classificação de cada amostra será:

Se $\Delta_{Médio\ \%}$ > Threshold definido:

Possível oportunidade

Caso contrário:

Descartar

A tabela a seguir traz um exemplo ilustrativo da metodologia aplicada a 4 amostras hipotéticas, em que, uma vez definido um threshold de erro dos modelos, encontram-se 2 possíveis oportunidades.

Tabela 14

Simulação do algoritmo de procura de oportunidades

Amostras	Preço de listagem	Modelo semi-log Hedônico	RNA	Xgboost	Delta médio	Delta % médio (Módulo)	Threshold	Possível oportuni-dade?
1	R$10,000	R$11,885	R$13,685	R$14,176	-R$3,249	0	0	Sim
2	R$12,000	R$13,337	R$14,485	R$14,877	-R$2,233	0	0	Não
3	R$13,400	R$14,584	R$14,960	R$14,938	-R$1,427	0	0	Não
4	R$11,000	R$13,432	R$14,733	R$13,350	-R$2,838	0	0	Sim

Em quarto lugar, as amostras selecionadas são analisadas por métodos estatísticos e manuais para entender se são outliers, no caso sendo descartadas, ou oportunidades, gerando uma lista para análise baseada em outros critérios, objetivos e subjetivos, por um avaliador ou agente imobiliário.

5.2 Passo #1: Base de dados

Esse passo foi executado no item 3.

5.3 Passo #2: Execução dos modelos na amostra

5.3.1 Descrição geral dos modelos utilizados

Para uma compreensão abrangente da metodologia, é essencial garantir um entendimento completo de cada um dos modelos utilizados. Neste sentido, a seguir, cada modelo será detalhado, abordando a perspectiva metodológica estatística, os parâmetros empregados, a validação dos dados e a análise da importância das variáveis.

- **Modelo Hedônico Linear Multivariável**

Modelo Hedônico Linear Multivariável é uma técnica estatística amplamente empregada na avaliação de preços de bens, como imóveis, que leva em consideração diversas características intrínsecas e extrínsecas desses bens para explicar suas variações de preço. Na metodologia estatística do modelo, utiliza-se uma abordagem de regressão linear multivariável, na qual o preço do bem é a variável dependente, enquanto as características relevantes intrínsecas e extrínsecas, como área, número de quartos e outros atributos, são as variáveis independentes. O modelo pressupõe que o preço de um bem pode ser explicado como uma soma ponderada dessas características, com cada peso (coeficiente) representando o impacto relativo daquela característica no preço final. A partir disso, a análise estatística avalia como essas variáveis independentes influenciam o preço e em que magnitude.

Os parâmetros utilizados no Modelo Hedônico Linear Multivariável são os coeficientes de regressão estimados a partir da análise dos dados. Cada coeficiente corresponde a uma variável independente e reflete a relação entre essa variável e o preço do bem. A estimação desses coeficientes é geralmente realizada por meio de técnicas de mínimos quadrados, buscando encontrar os valores que melhor ajustam o modelo aos dados observados. Esses coeficientes são essenciais para compreender a contribuição relativa de cada característica na determinação do preço, permitindo a análise de quais fatores são mais significativos na precificação do bem, o que será feito no próximo item.

A validação dos dados no Modelo Hedônico Linear Multivariável é um passo crítico para garantir a robustez e confiabilidade das conclusões. Isso envolve a verificação da qualidade e integridade dos dados utilizados no modelo, incluindo a detecção e correção de possíveis erros ou outliers. Além disso, a validação também pode incluir a análise de pressupostos fundamentais do modelo, como a normalidade dos resíduos e a independência dos erros, a fim de verificar se o modelo está bem especificado e se as estimativas são confiáveis, análises que também são realizadas no próximo item.

A análise da importância das variáveis neste modelo permite identificar quais características têm o maior impacto nos preços dos bens. Isso é feito por meio da análise dos coeficientes estimados para cada variável independente, bem como da realização de testes de significância estatística. Variáveis com coeficientes significativos e de grande magnitude são consideradas as mais importantes na explicação das variações de preço. Essa análise ajuda a fornecer insights valiosos para tomadores de decisão, como investidores imobiliários e governos, ao destacar as características que mais influenciam o valor de mercado dos bens e orientar estratégias de investimento e políticas públicas.

- **Modelo XGBoost**

O modelo XGBoost é uma técnica de aprendizado de máquina que se baseia em uma metodologia estatística chamada de "Boosting". Ele trabalha de forma iterativa, treinando uma série de modelos fracos (geralmente árvores de decisão) e os combina para formar um modelo forte. A metodologia estatística por trás do XGBoost visa minimizar a função de perda, ajustando os pesos dos erros cometidos nas iterações anteriores, o que resulta em uma melhoria gradual na precisão do modelo. Isso permite que o XGBoost lide eficazmente com problemas de classificação e regressão, além de ser menos suscetível ao overfitting.

No que diz respeito aos parâmetros utilizados, o XGBoost oferece uma ampla gama de configurações para controle do comportamento do modelo. Alguns dos parâmetros mais comuns incluem a taxa de aprendizado (learning rate), a profundidade máxima das árvores (max_depth), o número de estimadores (n_estimators), entre outros. A escolha adequada dos parâmetros desempenha um papel crítico no desempenho e na generalização do modelo, sendo neste trabalho utilizada a função GridSearch do sklearn para executar a otimização descrita.

A validação dos dados é um aspecto fundamental na utilização do XGBoost. Geralmente, a técnica de validação cruzada, como a validação cruzada k-fold, é empregada para avaliar o desempenho do modelo e garantir sua capacidade de generalização. A divisão dos dados em conjuntos de treinamento, validação e teste é essencial para medir a eficácia do modelo em dados não vistos durante o treinamento e evitar o overfitting, que será mostrado no próximo item deste trabalho.

A análise de importância das variáveis é um recurso valioso do XGBoost. Este modelo permite calcular a importância relativa de cada variável em relação à previsão final. Isso é feito medindo o quanto cada variável contribui para a redução da função de perda durante o processo de treinamento.

- **Modelo KNN**

O modelo KNN (K-Nearest Neighbors) é um algoritmo de aprendizado de máquina utilizado para tarefas de regressão, que visa prever valores numéricos com base em observações vizinhas no espaço de características. A ideia fundamental por trás do KNN é que os pontos de dados semelhantes tendem a ter valores de saída semelhantes. O algoritmo funciona identificando os k pontos de treinamento mais próximos a um ponto de teste, onde a proximidade é geralmente definida por uma métrica de distância, como a distância Euclidiana. Em seguida, ele calcula a média (ou outra medida) dos valores de saída desses k vizinhos para prever o valor de saída do ponto de teste. O KNN é um modelo não paramétrico, o que significa que não faz suposições sobre a distribuição subjacente dos dados.

Apesar de não ser paramétrico, o modelo KNN possui dois parâmetros principais: o número de vizinhos k e a métrica de distância. A escolha adequada de k é fundamental, pois um valor muito baixo pode levar a um modelo instável e suscetível a ruído, enquanto um valor muito alto pode levar a um modelo subajustado. A seleção da métrica de distância também é

relevante, pois diferentes métricas podem ser mais apropriadas para diferentes tipos de dados e problemas. A escolha desses parâmetros geralmente envolve técnicas de validação cruzada, como a busca em grade, para determinar os valores ótimos que maximizam o desempenho do modelo, sendo feita de forma semelhante ao modelo XGBoost, citado anteriormente.

A validação dos dados desempenha um papel crucial no uso do modelo KNN para regressão e é feito de maneira semelhante aos modelos citados anteriormente. Uma vantagem do modelo KNN é sua capacidade inerente de considerar todas as variáveis presentes nos dados, uma vez que se baseia na proximidade entre pontos de dados. No entanto, a relevância relativa das variáveis pode variar dependendo do valor de k escolhido e da métrica de distância utilizada. A análise da importância das variáveis pode ser realizada calculando-se a contribuição de cada variável na previsão dos valores de saída. Isso pode ser útil para selecionar um subconjunto de características mais relevantes e reduzir a dimensionalidade dos dados, se necessário.

5.3.2 Comparação entre modelos executados - Métricas gerais

A qualidade de um modelo de machine learning depende do equilíbrio entre viés e variância, em que um alto viés pode resultar em sub ajuste e uma alta variância, em sobreajuste. Encontrar o ponto intermediário é crucial para garantir que o modelo generalize bem e mantenha o desempenho consistente.

No contexto da análise estatística de modelos, métricas são medidas quantitativas que permitem avaliar o desempenho e a precisão desses modelos em relação aos dados observados, desempenhando um papel fundamental na validação e na seleção dos melhores modelos, contribuindo assim para uma compreensão mais profunda das relações entre as variáveis de interesse.

Entre as métricas mais amplamente utilizadas na avaliação de modelos, destacam-se seis: o Erro Quadrático Médio (MSE), o Erro Quadrático Médio com Raiz (RMSE), o Coeficiente de Determinação (R^2), o Erro Médio Absoluto (MAE), o Erro Percentual Médio Absoluto (MAPE), e o RMSPE (Erro Percentual Quadrático Médio). Cada uma dessas métricas oferece uma perspectiva específica do desempenho do modelo e é frequentemente empregada em diferentes contextos e tipos de análises.

Tanto o **Erro Quadrático Médio (MSE)** quanto o **Erro Quadrático Médio com Raiz (RMSE)** desempenham funções cruciais na avaliação da qualidade das previsões de um modelo, porém, apresentam perspectivas diferentes. O MSE fornece uma medida da variabilidade dos erros, mas não possui uma interpretação direta em termos das unidades originais dos dados, ao passo que o RMSE calcula a raiz quadrada da média dos quadrados dos erros entre as previsões do modelo e os valores observados. O RMSE, por sua vez, atribui maior peso aos erros maiores e é sensível a valores discrepantes, proporcionando uma visão mais precisa da dispersão dos erros nas unidades originais. Essa distinção é particularmente relevante para avaliar a magnitude dos erros e seu impacto nas previsões, destacando a complementaridade entre essas duas métricas na análise do desempenho do modelo.

O **Coeficiente de Determinação (R^2)**, por outro lado, mede a proporção da variabilidade total nos dados que é explicada pelo modelo. Ele varia de 0 a 1 e fornece uma medida da qualidade do ajuste do modelo aos dados. Quanto mais próximo de 1, melhor o modelo se ajusta aos dados, indicando que uma porcentagem maior da variabilidade é explicada pelas variáveis independentes.

O **Erro Médio Absoluto (MAE)** é uma métrica simples que calcula a média dos valores absolutos dos erros entre as previsões do modelo e os valores observados. O MAE é menos sensível a valores discrepantes do que o RMSE, sendo útil para avaliar a magnitude média dos erros sem considerar sua direção.

O **Erro Percentual Médio Absoluto (MAPE)** é uma métrica que calcula a média das porcentagens dos erros em relação aos valores observados. É especialmente útil quando se deseja avaliar o desempenho do modelo em termos de porcentagem de erro em relação aos valores reais, sendo relevante em cenários de previsão em que a escala das variáveis é significativa.

O **RMSPE (Erro Percentual Quadrático Médio)** é uma métrica de avaliação de desempenho de modelos de previsão que calcula a raiz quadrada da média dos quadrados dos erros percentuais entre as previsões do modelo e os valores observados. Isso é particularmente relevante quando os erros percentuais têm maior impacto do que os erros absolutos. Essa será uma métrica porcentual acessória ao MAPE para comparação.

No contexto deste trabalho de comparação entre modelos lineares e não lineares (hedônico vs. XGBoost) para prever variações nos preços imobiliários no mercado turco, essas métricas desempenham papéis distintos. O RMSE, MSE e o MAE permitem avaliar a precisão das

previsões e a magnitude dos erros, enquanto o R^2 indica o grau de ajuste do modelo aos dados.

Por fim, o uso do MAPE no trabalho é justificado devido à sua simplicidade e capacidade de expressar os erros de previsão em termos de porcentagem, facilitando a compreensão e a comparação das métricas de desempenho dos modelos de forma direta e intuitiva. Como afirma Hyndman e Koehler (2006), o MAPE é mais fácil de interpretar porque é expresso em termos de porcentagem.

No entanto, Armstrong (1985) foi o pioneiro em destacar a falta de simetria no Erro Percentual Médio Absoluto (MAPE), afirmando que ele tende a favorecer estimativas que subestimam os valores reais. Algum tempo depois, Armstrong e Collopy (1992) argumentaram que o MAPE aplica uma penalização mais severa às previsões que ultrapassam os valores reais do que às que ficam aquém desses valores. Dentro do contexto do modelo atual, apesar dessa falta de simetria, a utilização do MAPE é apropriada, uma vez que essa assimetria acaba por realçar os erros nas previsões que ficam aquém do esperado, tornando o modelo mais sensível a esses valores, que são os de interesse. Essa, portanto, será a métrica principal utilizada para ponderação dos dados.

5.3.3 Resultados

- **Modelo Hedônico Linear Multivariável**

No contexto do modelo linear, é essencial analisar os resultados apresentados a seguir, os quais foram obtidos por meio da aplicação da função OLS da biblioteca statsmod. É importante ressaltar que essas métricas se aplicam exclusivamente aos dados de treinamento, o que nos leva a considerar a possibilidade de overfitting, um aspecto crítico no aprendizado supervisionado. O overfitting é conhecido por dificultar o ajuste ideal dos modelos aos dados de treinamento e, ao mesmo tempo, prejudicar a capacidade de generalização para dados não vistos no conjunto de teste. Essa complexidade surge devido ao ruído nos dados, às limitações do tamanho do conjunto de treinamento e à complexidade do modelo, como discutido por Ying (2019). No entanto, ao compararmos esses resultados com as métricas obtidas no

conjunto de teste, observamos uma notável semelhança, sugerindo que, apesar da possível presença de viés, o modelo demonstra uma capacidade considerável de generalização.

Tabela 15

Resultados de R-squared para o modelo hedônico linear

Métrica	Valor Treino	Valor Teste
R-squared:	0.725	0.726
Adj. R-squared:	0.721	0.709

As métricas de performance podem ser vistas na tabela a seguir.

Tabela 16

Métricas de performance do modelo hedônico linear

Métrica	Método Hedônico Linear
R^2	0.726
MSE	3,687,655.97
MAE	1,340.42
RMSE	1,920.33
RMSPE	19.79%
MAPE	14.64%
Custo Computacional	0.25 segundos

O mesmo modelo anteriormente citado também revela os coeficientes de cada variável, inclusive da constante. Cada variável possui um p-valor e um intervalo de confiança. No âmbito deste estudo, o p-valor desempenha um papel essencial ao determinar a significância estatística das variáveis independentes. Ele permite avaliar se cada variável contribui de maneira estatisticamente relevante para a variável dependente. Essa análise ajuda a selecionar as variáveis mais relevantes e a construir modelos de regressão confiáveis. O p-valor é crucial para distinguir relações estatisticamente significativas de resultados ao acaso, fortalecendo a interpretação dos resultados desta pesquisa. O p-valor é frequentemente utilizado com um

limiar de 0,05 para avaliar a significância estatística das relações entre variáveis em pesquisas científicas. No entanto, a interpretação do p-valor deve levar em conta o contexto da pesquisa, uma vez que não é uma regra universal e pode variar entre áreas de estudo. É uma ferramenta valiosa, mas sua interpretação requer consideração cuidadosa.

Tabela 17

Resultado da regressão linear multivariada

	coef	P>|t|	[0.025	0.975]
const	-393200	0.00	-457000.0	-329000.0
Número de Garagens	436.5	0.00	326.6	546.3
Número de UBS	381.9	0.00	277.0	486.7
Número de Escolas (Outros)	279.5	0.00	127.7	431.3
Número de Universidades Privadas	-264.1	0.00	-347.9	-180.2
Número de Universidades Públicas	-2532.2	0.00	-3588.7	-1475.7
Número de Estações de Metrô	-603.5	0.00	-776.2	-430.8
Fração Ideal	4619.9	0.00	4265.1	4974.7
Valor do Terreno	0.65	0.00	0.5	0.8
Ano de Construção	200.9	0.00	167.8	233.9
Número de Pavimentos	81.9	0.00	65.1	98.6
Fator de obsolescência	-9874.6	0.00	-12500.0	-7267.3
Número de Acessibilidade	64.2	0.00	36.3	92.1
Número Escalado de Atendimentos Relativos à Asfalto	536.5	0.00	239.3	833.6
Número Escalado de Clínicas	1694.0	0.00	1049.4	2338.6
Número Ponderado de Linhas de Ônibus Arteriais	517.0	0.00	412.8	621.2
Número Ponderado de Linhas de Ônibus Locais	65.2	0.00	32.6	97.7
Número Ponderado de Linhas de Ônibus Coletoras	151.4	0.00	95.7	207.1
Número Ponderado de Ciclofaixas	-97.5	0.00	-159.9	-35.0
Número Ponderado de Ciclovias	73.0	0.00	26.9	119.1
Número Escalado de Árvores	398.5	0.01	109.7	687.3
Número Ponderado de Ciclorrotas	193.3	0.01	44.3	342.2

Número de Hospitais Públicos	-686.1	0.03	-1292.5	-79.7
Área Construída	-1.9	0.03	-3.5	-0.2
Número de Esquinas	92.8	0.04	3.4	182.2
Número Escalado de Clínicas Médicas	-609.2	0.06	-1232.7	14.4
Número Ponderado de BRTs	96.1	0.06	-4.8	197.0
Número de Escolas Públicas Municipais	-44.9	0.07	-93.4	3.6
Número de Poupatempos	-2838.2	0.09	-6093.7	417.3
Número de Centros Populares	-2028.7	0.10	-4458.6	401.3
Área Ocupada	0.03	0.12	0.0	0.1
Valor do Condomínio/m2	-0.0015	0.21	-0.004	0.001
Número de Paradas de Ônibus	9.36	0.21	-5.2	23.9
Número de CREAS	463.18	0.24	-303.0	1229.4
Número de Escolas Públicas Federais	716.22	0.29	-603.8	2036.2
Número de Escolas Públicas Estaduais	23.00	0.37	-27.7	73.7
Número de Hospitais Particulares	32.11	0.38	-39.4	103.6
Área do Terreno	0.01	0.46	-0.02	0.04
Número de Escolas Privadas	-4.66	0.48	-17.6	8.3
Número de Museus	19.28	0.69	-75.6	114.2
Testada	-0.29	0.70	-1.8	1.2
Número de Estacionamentos de Bicicleta	-32.62	0.72	-207.9	142.6
Número de Quartos por Banheiros	18.72	0.79	-118.2	155.6
Número de FATECs	-92.49	0.87	-1171.4	986.5
Número de CRAS	24.00	0.95	-771.4	819.4

É notável que, entre as variáveis analisadas, 24 delas apresentam um valor de p-valor inferior a 0,05, sendo que 19 destas variáveis demonstram um p-valor praticamente nulo, evidenciando sua alta relevância estatística. Algumas destas variáveis merecem um aprofundamento adicional.

No que diz respeito às características intrínsecas da propriedade, observamos, primeiramente, o número de garagens, cujo coeficiente é de 436,5. Isso corrobora com a análise realizada no capítulo 4 deste trabalho, na qual se identificou a relação de que um maior número de garagens está associado a um aumento no preço por metro quadrado da propriedade. Neste contexto, o coeficiente indica um aumento de aproximadamente 436 reais

por metro quadrado para cada garagem adicional. Um efeito semelhante foi encontrado para o ano de construção, no qual a adição de cada ano sugere um aumento de 200,9 reais no valor do preço por metro quadrado, reforçando a conclusão de que propriedades mais recentes são mais valorizadas. Além disso, o número de pavimentos do edifício também parece influenciar positivamente no preço por m2, com um coeficiente de 81,9.

No que se refere às características extrínsecas, observamos que o número de estações de metrô parece ter um efeito negativo, reduzindo o valor da propriedade em 603,5 reais por metro quadrado para cada estação adicional. Esse valor pode estar muito relacionado à ausência de estações de metrô em Moema, um dos bairros mais valorizados. No entanto, todas as linhas de ônibus (arteriais, locais e coletoras) exercem um efeito positivo sobre o preço por m2. Além disso, notamos que ciclovias e ciclorrotas também afetam positivamente o preço, o que não ocorre com as ciclofaixas. Outras conclusões podem ser observadas na tabela 17.

Ao analisar a dispersão dos valores previstos em relação aos valores reais, é evidente que o modelo tem a capacidade de capturar uma parte considerável da variabilidade nos dados, como indicado pelo valor significativo do coeficiente de determinação (R^2) no teste, que atinge 0,726. Essa medida de ajuste é um indicativo positivo do modelo de regressão linear, embora seja importante ressaltar que a análise por bairro revelou variações significativas no desempenho do modelo em diferentes regiões geográficas.

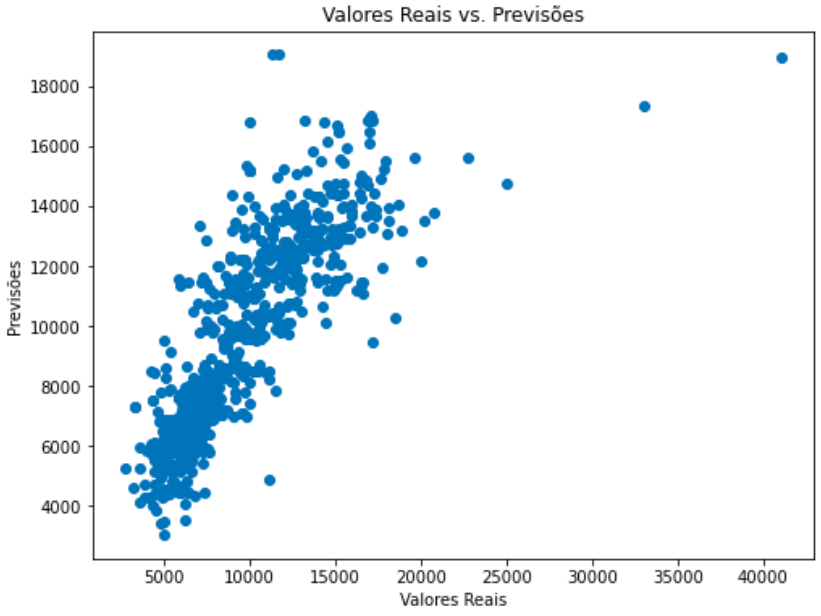


Fig. 11 – Dispersão dos valores reais em relação aos modelados

Ao analisar as previsões geradas pelo modelo linear, desagregando os resultados por bairros, fica evidente que a acurácia da previsão não é uniforme em todas as regiões. A variação na precisão da previsão entre diferentes bairros é claramente ilustrada por meio das dispersões observadas na figura 12. Essa observação ressalta a complexidade e a heterogeneidade das relações subjacentes entre as variáveis de entrada e a variável alvo, bem como a influência de fatores geográficos e contextuais nas previsões do modelo. Essa variação na acurácia por bairro é um aspecto crucial a ser considerado ao interpretar e aplicar os resultados da metodologia proposta neste trabalho de encontrar oportunidades.

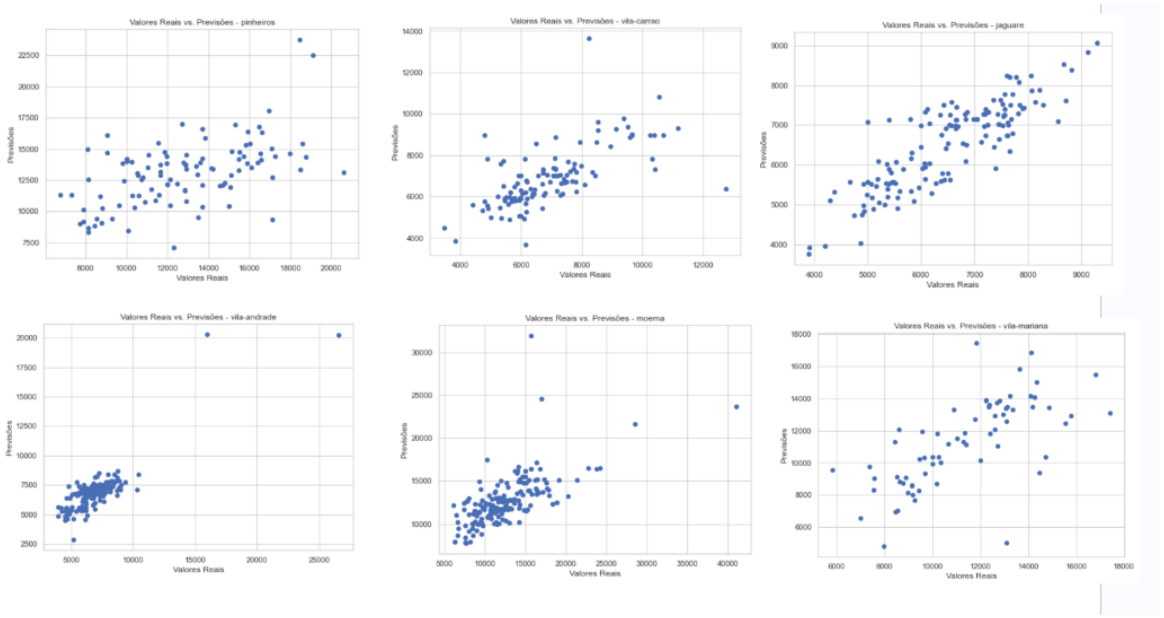


Fig. 12 – Dispersão dos resultados do modelo por bairro

Ao avaliar o desempenho do modelo por métricas de erro, a observação visual é reforçada, e torna-se evidente uma disparidade substancial nos erros de previsão entre diferentes bairros. Essa disparidade é quantificada por meio de quatro métricas de avaliação: RMSE, R^2, RMSPE e MAPE. A análise dessas métricas revela que o modelo apresenta variações significativas em sua precisão ao prever os valores-alvo em bairros distintos

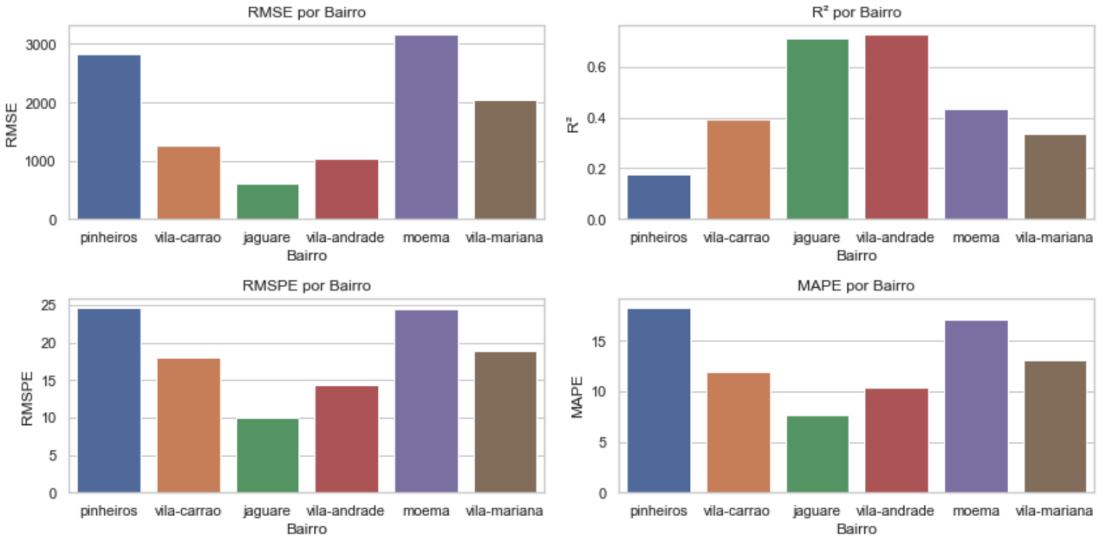


Fig. 13 – Performance do modelo por bairro

Observamos que, em todas as métricas de avaliação, com exceção do coeficiente de determinação (R^2), o bairro Jaguaré apresenta o melhor desempenho, seguido por Vila Andrade, Vila Carrão e Vila Mariana. Por outro lado, Moema e Pinheiros são os bairros com o desempenho mais fraco, com o valor de R^2 de Pinheiros abaixo de 0,2. Essas descobertas são cruciais para a avaliação manual, pois revelam que diferentes bairros exibem níveis distintos de precisão nas previsões, mesmo quando o modelo é aplicado de forma conjunta. É fundamental ter em mente que esses resultados são específicos para o modelo de regressão linear utilizado. Essas informações destacam a importância de considerar a heterogeneidade geográfica ao avaliar e interpretar o desempenho do modelo, bem como ao tomar decisões baseadas nas previsões em áreas urbanas diversificadas.

Podemos também examinar os resíduos a seguir. Observa-se uma regularidade importante, corroborada pelo resultado do teste de Durbin-Watson com um valor de 2,043, que sugere uma distribuição normal dos resíduos. Essa constatação é relevante, pois indica que o modelo de regressão linear aplicado às previsões possui um bom ajuste aos dados, com os resíduos distribuídos de forma homogênea ao longo das observações. Isso é um indicativo positivo da qualidade do modelo e da sua capacidade de capturar as variações nos preços por metro quadrado em diferentes bairros.

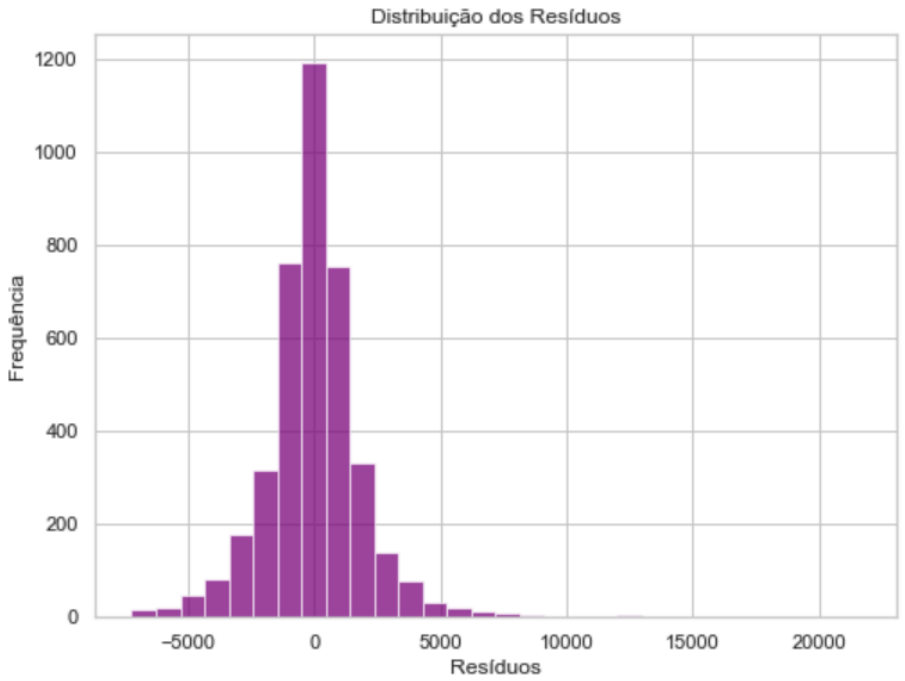


Fig. 14 – Histograma dos resíduos para o modelo linear

- **Modelo XGBoost**

No contexto da otimização do modelo XGBoost, foi conduzida uma análise abrangente dos hiperparâmetros por meio de uma técnica conhecida como "grid search". O grid search é uma abordagem sistemática e essencialmente exploratória que tem ampla aplicação no campo de aprendizado de máquina. Seu propósito é identificar a configuração de hiperparâmetros que resulta no desempenho mais eficaz do modelo, medido por métricas de avaliação específicas, como o erro quadrático médio ou o coeficiente de determinação (R^2).

Neste estudo específico, concentrou-se a atenção em três hiperparâmetros cruciais do XGBoost:

1. **Número de Estimadores ('n_estimators'):** Esta variável controla o número de árvores de decisão a serem construídas pelo modelo. Avaliamos três valores distintos: 100, 200 e 300. O objetivo foi entender como a variação desse parâmetro impactaria o desempenho geral.

2. **Taxa de Aprendizado ('learning_rate'):** A taxa de aprendizado é fundamental, pois influencia o tamanho dos passos dados durante o processo de treinamento. Três valores diferentes foram testados: 0,01, 0,1 e 0,2. Essa pesquisa visa descobrir como a taxa de aprendizado afeta o equilíbrio entre a convergência eficiente e a ocorrência de overfitting.

3. **Profundidade Máxima da Árvore ('max_depth'):** A profundidade máxima das árvores de decisão é controlada por esse hiperparâmetro. Foram investigados três valores: 3, 4 e 5, cada um representando o número máximo de níveis nas árvores e permitiu compreender como a complexidade das árvores impacta o desempenho do modelo.

O grid search explorou meticulosamente todas as combinações possíveis desses hiperparâmetros, registrando o desempenho do modelo para cada configuração. Os resultados finais revelaram os valores ótimos para 'learning_rate, max_depth e n_estimators como sendo 0,2, 5 e 200, respectivamente. Essa abordagem rigorosa de otimização é fundamental para garantir que o modelo XGBoost esteja adequadamente calibrado e pronto para fornecer previsões precisas.

Após a definição dos hiperparâmetros otimizados, o modelo de regressão XGBoost foi treinado e avaliado, gerando uma série de métricas para avaliar seu desempenho.

O coeficiente de determinação R^2, calculado no conjunto de treinamento, apresentou um valor de 0.9819. Esse valor indica que o modelo de regressão XGBoost captura aproximadamente 98.19% da variabilidade nos dados de treinamento, sugerindo um ajuste muito bom aos dados de treinamento.

Além disso, no conjunto de teste, o modelo de regressão XGBoost obteve os seguintes resultados:

Tabela 18

Resultado da regressão linear multivariada

Métrica	Modelo XGBoost
R^2	0.898
MSE	1,370,011
MAE	773.67
RMSE	1,170.47
RMSPE	12.46%
MAPE	8.60%
Custo Computacional	11.6 segundos

Essas métricas proporcionam uma visão abrangente do desempenho do modelo. O alto valor de R^2 no conjunto de teste (0.8983) indica que o modelo é capaz de capturar uma

parcela significativa da variabilidade nos dados de teste. O valor de mais destaque aqui é o MAPE, sendo inferior a 10% de erro.

É também relevante analisar a importância de cada uma das variáveis do modelo, como pode ser observado na figura 15.

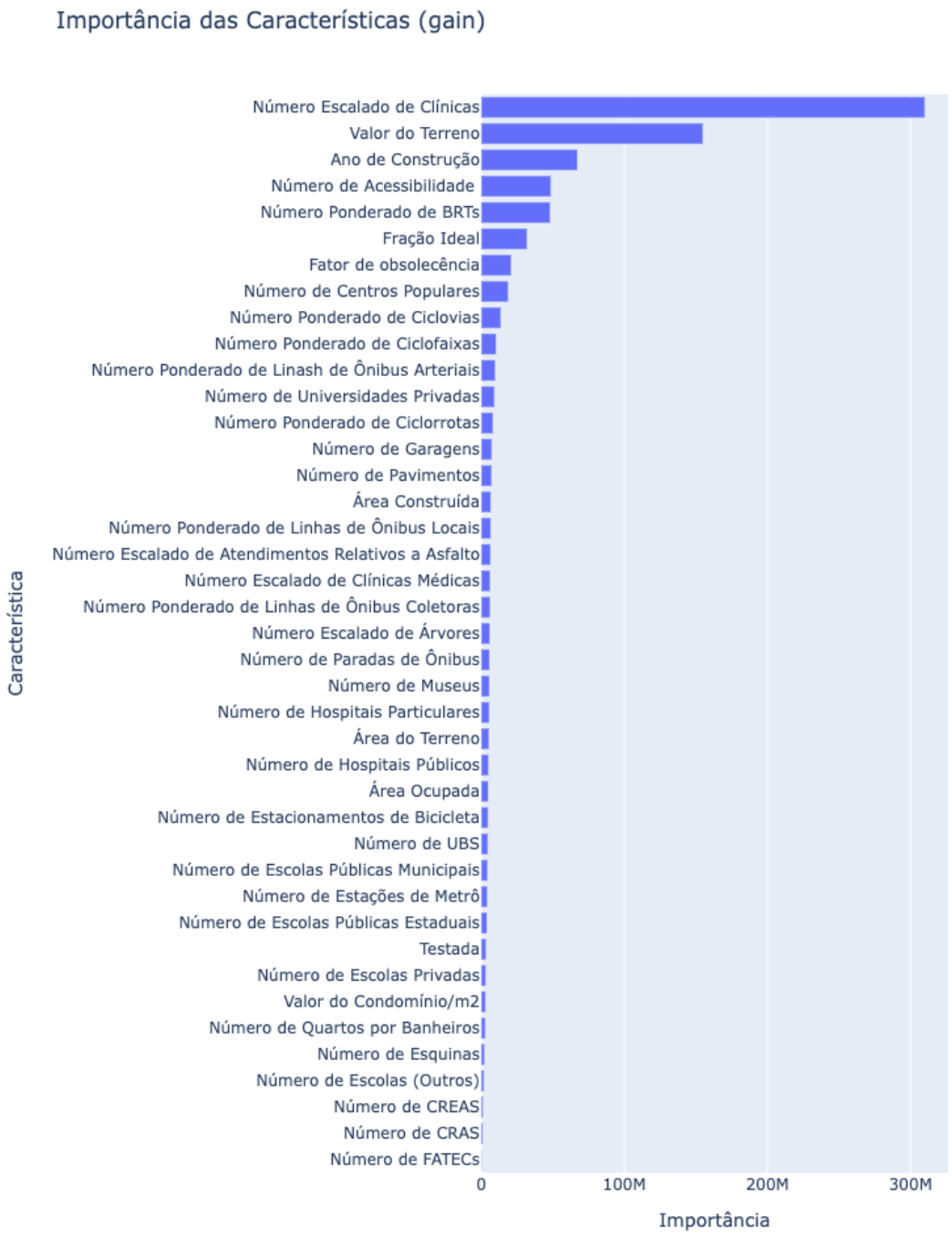


Fig. 15 – Análise de relevância das variáveis do modelo xgboost

O método gain é uma métrica usada para avaliar a importância das variáveis em um modelo XGBoost, e ele pode oferecer insights valiosos no contexto do mercado imobiliário. No topo da lista das variáveis mais importantes, encontramos Número Escalado de Clínicas, Valor do Terreno e Ano de Construção, que se destacam como as três características mais influentes na tomada de decisões do modelo.

A interpretação do "gain" pode ser um pouco mais desafiadora em comparação com outras métricas, como contagem de uso ("weight"). No entanto, ele fornece uma perspectiva essencial sobre como as variáveis afetam as decisões do modelo, sendo uma medida do quanto uma variável contribui para melhorar a pureza das divisões nas árvores de decisão do modelo. Quanto maior o ganho, mais informativa é a variável para a separação dos dados em grupos mais homogêneos.

É relevante observar que algumas variáveis que o modelo linear considerou como importantes coincidem com as três principais variáveis de importância para o XGBoost, e todas elas são estatisticamente significativas para o modelo linear. No entanto, é interessante notar uma diferença notável no caso da variável número de garagens. Enquanto esta variável foi considerada altamente relevante no modelo linear, seu impacto no modelo XGBoost parece ser menos pronunciado.

Vale mencionar que existem ferramentas mais abrangentes e específicas para analisar a importância das variáveis em modelos não lineares, como o LIME e o SHAP. No entanto, é importante ressaltar que essas ferramentas não fazem parte do escopo deste trabalho e não foram exploradas aqui. Elas oferecem abordagens mais detalhadas para entender como as variáveis afetam modelos complexos, como o XGBoost, permitindo uma análise mais profunda da contribuição de cada característica para as previsões do modelo.

- **Modelo KNN**

No contexto da otimização do modelo KNN (K-Nearest Neighbors), foi conduzida uma análise abrangente dos hiperparâmetros também pela técnica grid search. Dentre os resultados do grid search, os hiperparâmetros selecionados para o modelo KNN foram os seguintes:

- 'n_neighbors': 13,
- 'weights': 'distance'.

Essa seleção foi baseada em uma análise cuidadosa das combinações de hiperparâmetros testadas, visando otimizar o desempenho do modelo em termos de métricas de avaliação relevantes.

As métricas de avaliação do modelo KNN (K-Nearest Neighbors) revelaram resultados significativos para a previsão de preços de imóveis.

Tabela 19

Resultado do modelo KNN

Métrica	KNN
R^2	0.733
MSE	3,592,673
MAE	1,038.13
RMSE	1,895.43
RMSPE	19.25%
MAPE	11.16%
Custo Computacional	0.72 segundos

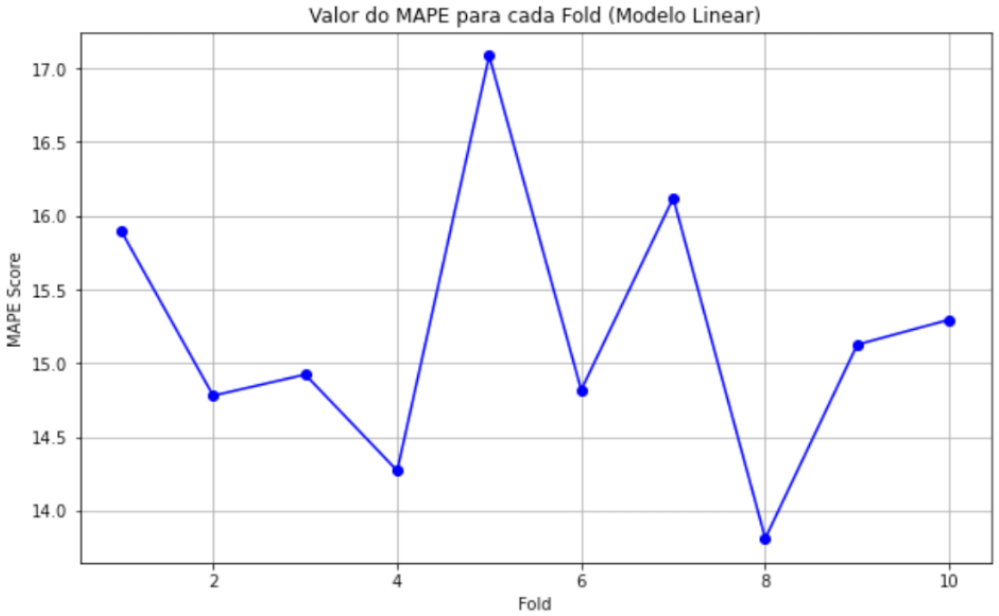
Essas métricas coletivas ressaltam a capacidade do modelo KNN em prever preços de imóveis de forma precisa e eficaz, considerando tanto a precisão das previsões quanto o custo computacional associado. No entanto, é notável que os resultados são semelhantes aos da regressão linear, com custo computacional mais alto. A discussão sobre esses pontos será feita posteriormente.

5.3.4 Validação cruzada

A validação cruzada, no contexto de machine learning, é uma técnica fundamental para avaliar a performance de um modelo de forma robusta e confiável. Ela consiste em dividir o conjunto de dados em múltiplos subconjuntos, geralmente chamados de "folds," onde o modelo é treinado em uma parte dos dados e testado em outra, repetindo esse processo várias vezes, de modo que cada subconjunto seja utilizado tanto para treinamento quanto para teste.

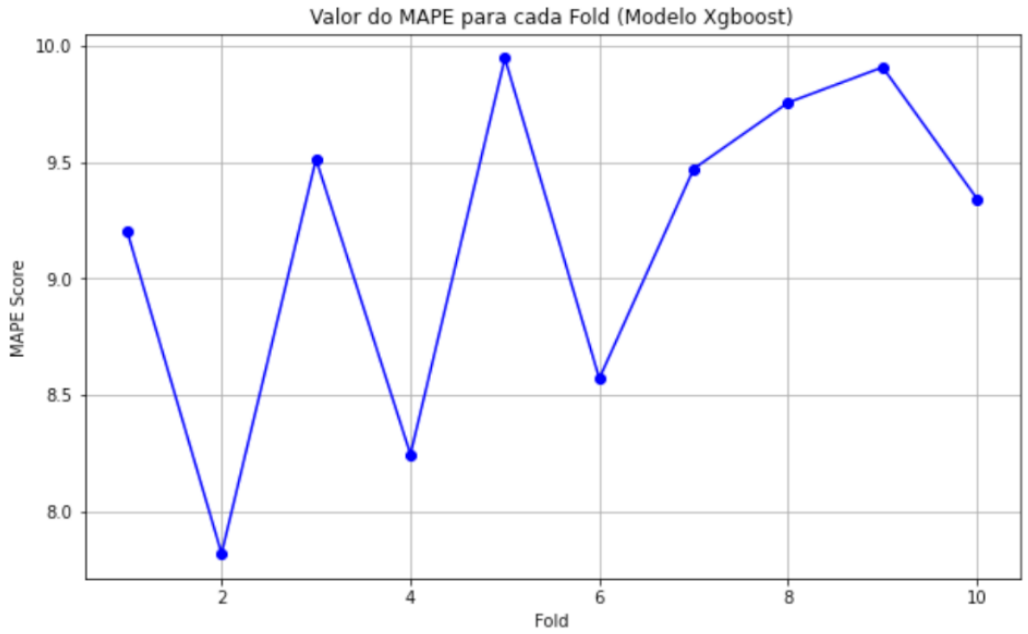
A principal importância da validação cruzada reside em sua capacidade de fornecer uma estimativa mais precisa do desempenho do modelo, uma vez que evita o viés que pode ocorrer com uma única divisão entre conjuntos de treinamento e teste. Isso ajuda a identificar se o modelo está sofrendo de overfitting (ajuste excessivo aos dados de treinamento) ou underfitting (falta de ajuste aos dados). Além disso, a validação cruzada permite uma avaliação mais completa da capacidade de generalização do modelo, uma vez que todos os dados são utilizados tanto para treinamento quanto para teste em algum momento, o que é particularmente crucial em trabalhos científicos, pois proporciona maior confiabilidade nas conclusões e na escolha do melhor modelo para determinada tarefa de aprendizado de máquina.

Para tal, foi executado para cada modelo uma validação de dados com 10 folds, avaliando os valores resultantes de MAPE para entendimento. Segue para os 3 modelos, nas figuras 16 a 18.



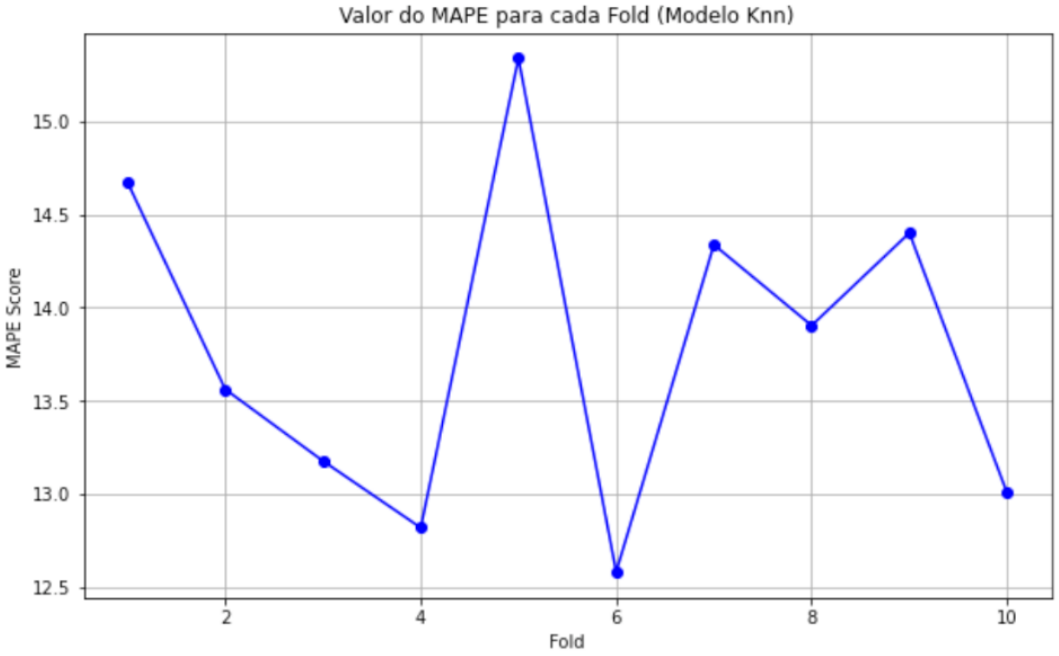

```
Média MAPE (Linear do Modelo): 15.21
Desvio Padrão MAPE (Linear do Modelo): 0.9
```

Fig. 16 – Validação cruzada para o modelo linear




```
Média MAPE (Xgboost do Modelo): 9.18
Desvio Padrão MAPE (Xgboost do Modelo): 0.69
```

Fig. 17 – Validação cruzada para o modelo xgboost




```
Média MAPE (Knn do Modelo): 13.78
Desvio Padrão MAPE (Knn do Modelo): 0.85
```

Fig. 18 – Validação cruzada para o modelo KNN

Consolidando todas as métricas em um gráfico de boxplot, temos:

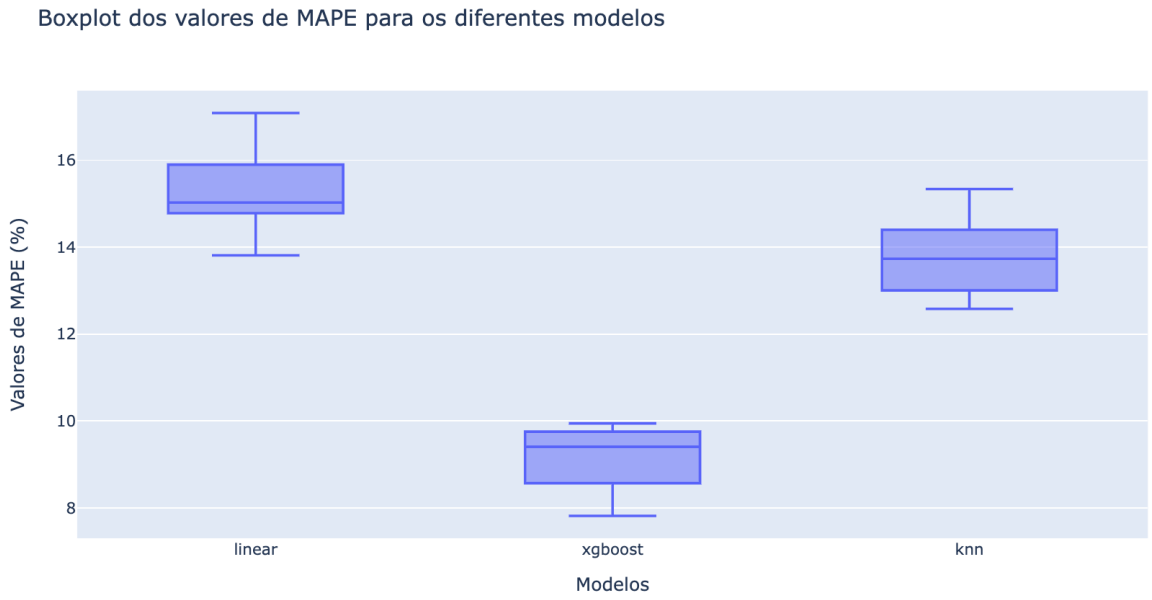


Fig. 19 – Boxplot dos valores de MAPE para cada modelo na validação cruzada

Observa-se de maneira nítida, pela figura 19, que o modelo XGBoost é consistentemente superior aos outros dois modelos, que pouco se diferenciam, existindo uma pequena vantagem do modelo KNN. O item a seguir traz uma visão mais aprofundada dessa comparação.

5.3.5 Comparação dos modelos

A comparação entre os modelos é um passo crucial na avaliação de seu desempenho, e essa análise envolve a utilização das métricas descritas anteriormente que oferecem perspectivas distintas sobre sua eficácia. Essa abordagem abrangente permite determinar, de forma mais abrangente, qual dos modelos se destaca em relação aos outros. A comparação pode ser observada na figura 20 a seguir:

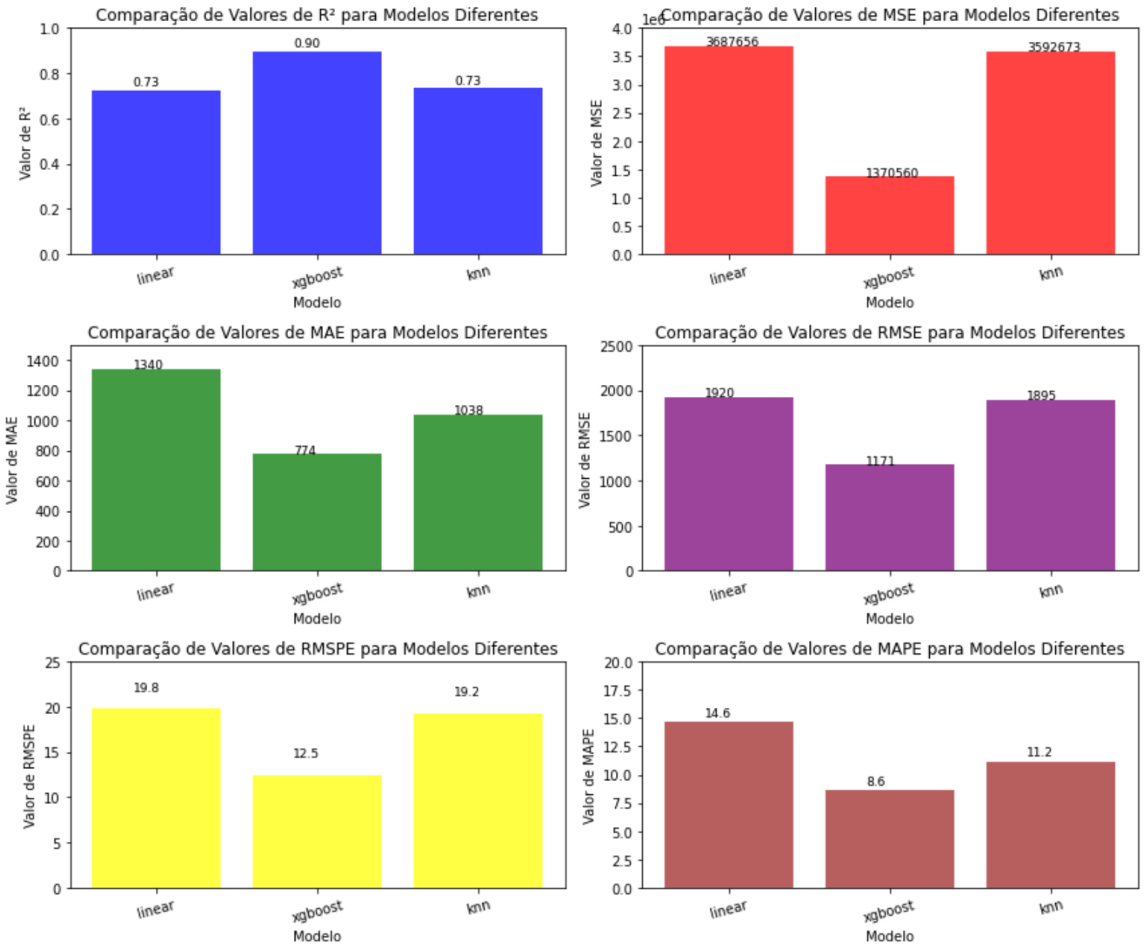


Fig. 20 – Comparação das métricas para os diferentes modelos

De fato, ao analisar as seis métricas de erro, podemos inferir que o modelo XGBoost apresenta um desempenho superior em comparação com os outros dois modelos (Linear Regression e K-Nearest Neighbors). Embora o modelo KNN tenha uma demanda computacional mais elevada, tendo tempo de rodagem de 0.75 segundos comparado a 0.25 segundos do modelo linear, não demonstrou uma superioridade significativa em termos de precisão preditiva. Durante a validação cruzada, o KNN mostrou um desempenho consistente em comparação com o modelo Linear, porém ainda consideravelmente abaixo do XGBoost em termos de acurácia.

Essa análise sugere que, embora o KNN possa ser uma opção viável em termos de consistência, o XGBoost se destaca como a escolha preferencial devido ao seu melhor desempenho geral em previsões, mesmo que tenha um custo computacional superior (11.6 segundos).

A literatura corrobora esses resultados. Segundo Jha *et al.* (2020) em seu estudo sobre a previsão do mercado imobiliário com algoritmos de aprendizado de máquina, os resultados empíricos ilustram que, com base no desempenho do modelo de previsão, incluindo o Coeficiente de Determinação (R2), Erro Quadrático Médio (MSE), Erro Médio Absoluto (MAE) e tempo computacional, o algoritmo XGBoost apresenta um desempenho superior em relação aos outros modelos na previsão de preços imobiliários.

5.4 Análise das oportunidades

Concluída a modelagem das três opções e determinados seus respectivos coeficientes, métricas e erros, é realizada a análise das oportunidades. No contexto deste estudo, uma consideração essencial é a seleção de uma métrica adequada para avaliar as previsões geradas por diferentes modelos. Seis métricas distintas foram previamente discutidas, cada uma delas apresentando suas próprias vantagens e limitações. No entanto, optou-se por adotar o Erro Percentual Absoluto Médio (MAPE), devido à sua natureza mais intuitiva, uma vez que permite interpretar o erro dos modelos de forma relativa, em oposição a métricas absolutas. Vale destacar que o coeficiente de determinação (R2) também poderia ser utilizado; no entanto, ele apresenta a limitação de possuir uma escala que varia de 0 a 1, o que dificulta a normalização e a interpretação direta do erro, tornando desafiador afirmar se um modelo é proporcionalmente melhor ou pior ajustado.

Ao calcular o valor da métrica de desvio, definido na equação 3 do item 5 deste trabalho, encontramos a seguinte distribuição na amostra treino da base, com um total de 780 propriedades para análise:

Tabela 20

Métrica de desvio e sua preponderância na amostra treino

Métrica de desvio	% da amostra
5%	49%
10%	81%
15%	94%
20%	96%

Na figura a seguir pode-se ver o histograma acumulado dessa distribuição.

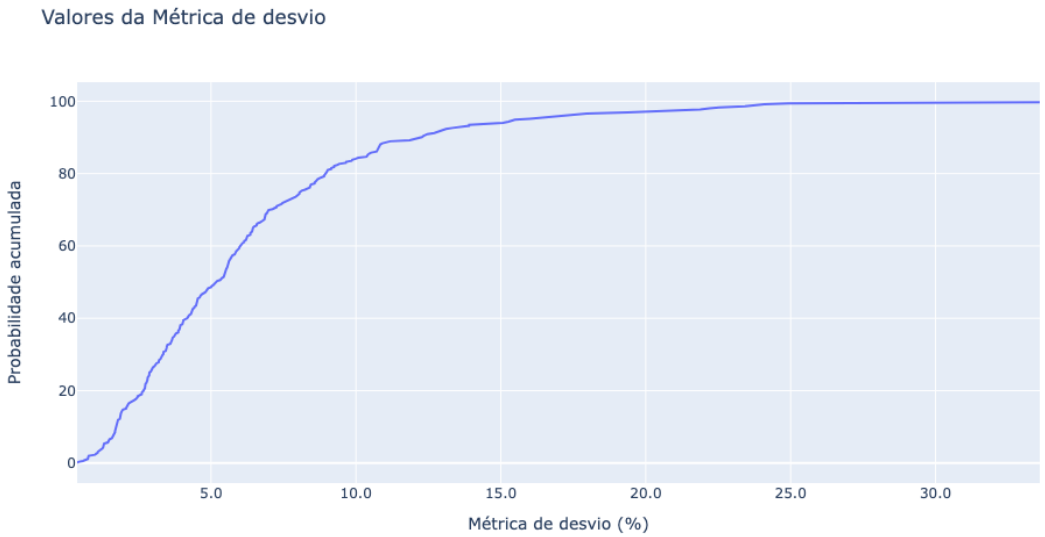


Fig. 21 – Histograma acumulado da métrica de desvio na amostra

Pode-se observar que apenas 3,8% das amostras se enquadram no critério de possível oportunidade, tendo um desvio de 20% ou mais quando comparados os valores previstos e anunciados da propriedade e tendo valor previsto maior que o anunciado. Esse número reduzido de amostras reafirma a assertividade de usar 20% como limite de análise, possibilitando efetivamente encontrar apenas um pequeno número de propriedades com preços considerados extremos.

Na tabela a seguir é possível observar as 30 propriedades que são consideradas possíveis oportunidades, com seus valores de previsão dos 3 modelos considerados e o valor anunciado, além da métrica de desvio considerada.

Tabela 21

Possíveis oportunidades selecionadas, ordenadas decrescentemente pela métrica de desvio

Amostra	Previsão Linear	Previsão XGBoost	Previsão KNN	Valor anunciado	Métrica de Desvio
1	R$10,629	R$8,649	R$14,412	R$5,287	67%
2	R$10,545	R$7,363	R$10,465	R$4,556	64%

3	R$10,102	R$9,546	R$10,845	R$6,077	40%
4	R$7,052	R$5,218	R$5,924	R$3,627	40%
5	R$7,939	R$4,152	R$4,249	R$3,297	39%
6	R$2,739	R$5,126	R$7,368	R$3,333	38%
7	R$6,686	R$4,593	R$5,880	R$3,487	38%
8	R$6,311	R$4,888	R$5,226	R$3,500	34%
9	R$10,414	R$10,144	R$13,421	R$7,377	32%
10	R$11,613	R$13,488	R$11,722	R$8,017	32%
11	R$11,798	R$6,857	R$6,535	R$5,488	32%
12	R$12,025	R$9,898	R$11,372	R$7,328	31%
13	R$13,011	R$11,432	R$11,732	R$7,996	30%
14	R$10,943	R$11,086	R$10,490	R$7,194	30%
15	R$8,314	R$5,813	R$6,895	R$4,800	27%
16	R$11,000	R$10,434	R$15,066	R$8,442	26%
17	R$7,513	R$8,210	R$9,769	R$5,926	26%
18	R$12,644	R$13,465	R$13,056	R$9,167	25%
19	R$10,419	R$8,749	R$13,557	R$7,690	25%
20	R$10,861	R$9,440	R$11,321	R$7,500	24%
21	R$11,049	R$8,356	R$10,535	R$7,160	23%
22	R$10,808	R$10,196	R$6,667	R$6,667	23%
23	R$6,455	R$4,717	R$5,209	R$3,947	23%
24	R$10,341	R$9,622	R$13,624	R$8,214	22%
25	R$12,254	R$9,121	R$10,209	R$7,727	22%
26	R$5,735	R$5,629	R$6,961	R$4,528	21%
27	R$7,258	R$6,100	R$6,861	R$5,000	21%
28	R$5,289	R$4,091	R$6,823	R$4,036	20%
29	R$11,917	R$10,892	R$12,788	R$8,868	20%
30	R$6,242	R$8,278	R$7,945	R$5,600	20%

Um ponto relevante de ser ressaltado inicialmente é que, ao analisar de forma visual a tabela, fica claro que nem todas as previsões são consistentes para os 3 modelos. Há algumas em que o desvio padrão entre as três previsões é bastante baixo, indicando consistência mesmo para modelagem distintas, mas em outros casos a diferença entre as previsões sugere uma inconsistência que deve ser verificada.

Outro ponto que se destaca nas possíveis oportunidades é que os preços são, em vários casos, muito baixos quando se considera a realidade. Possivelmente isso pode ter acontecido como uma forma de atrair clientes para a propriedade, uma prática comum do mercado imobiliário chamada de "phishing", sendo indicado um preço real ou a inexistência da propriedade no momento do contato. Isso ressalta a importância da verificação manual dessa lista.

5.4.1 Análise manual de algumas propriedades

Uma vez que as propriedades foram coletadas em 2021, a maior parte delas não estão mais disponíveis. Soma-se ao fato do tempo de coleta o preço baixo, que indica uma possível boa oportunidade que poderia ser falsa (para atrair clientes, uma prática comum no mercado) ou realmente uma boa oportunidade, que venderia rapidamente em um mercado dinâmico. Dessa forma, para enriquecer o entendimento de cada propriedade, fontes complementares serão consultadas, por meio de outros websites, de modo a explorar se existe algum erro na propriedade (dados incorretos, como por exemplo área, endereço) ou se realmente a oportunidade é promissora.

- **Oportunidade em análise #1: Alameda dos Maracatins, 185 - Moema**

A tabela a seguir resume a oportunidade #1, mostrando algumas das suas informações.

Tabela 22

Resumo da oportunidade em análise #1

Exemplo 1			
Endereço	Alameda dos Maracatins, 185 - Moema		
Área anunciada	181 m2		
Link	id = 2507142431		
Métrica de desvio	40%		
	Linear	**Xgboost**	**KNN**
Valor estimado	10102	9546	10845
Valor do anúncio	6077		

Fig. 22 – Fachada do edifício Juliana

Ao analisar o primeiro exemplo de possível oportunidade, percebemos a consistência na previsão de preço pelos 3 modelos propostos. Aparentemente, o valor de anúncio próximo de R$6000 por m2 parece estar bastante atrativo. No entanto, quando é feito um escrutínio nas informações deste condomínio, observa-se que não existem apartamentos de 181m2, como descrito no anúncio. A figura a seguir mostra as informações gerais do condomínio.

Fig. 23 – Informações sobre o edifício Juliana, com valores de área (Loft, 2023)

Ao considerar o valor de área sugerido, de 110m2, e o valor de anúncio, de R$1.099.900, encontramos um valor de preço por m2 próximo de R$10000, muito próximo aos sugeridos pela modelagem, excluindo a propriedade como oportunidade. No entanto, se for averiguado que a área é efetivamente 181m2, talvez falando de uma cobertura ou duplex, a oportunidade pareceria ser promissora.

- **Oportunidade em análise #2: Rua Eça de Queiroz, 58 - Vila Mariana**

A tabela a seguir resume a oportunidade #2, mostrando algumas das suas informações.

Tabela 23

Resumo da oportunidade em análise #2

Exemplo 2			
Endereço	Rua Eça de Queiroz, 58 - Vila Mariana		
Área anunciada	77 m2		
Link	id = 2494226941		
Métrica de desvio	32%		
	Linear	**Xgboost**	**KNN**
Valor estimado	10414	10144	13421
Valor do anúncio	7377		

Fig. 24 – Fachada do edifício Eça de Queiroz

No segundo caso, ao consultar o anúncio, também foi identificada uma inconsistência na área da propriedade em análise. Observou-se uma ambiguidade entre dois possíveis valores de área: 77m2 e 69m2, um fenômeno frequentemente associado à inclusão de vagas de garagem como parte da área total do apartamento no Brasil. Ao observar os valores de venda no mesmo prédio na figura 25, observa-se também que há um preço extremamente baixo, de R$4783/m2, indicando, no caso de ser um dado coerente, que a possível oportunidade não é tão atrativa quanto se imaginava.

Histórico de vendas em Condomínio Edificio Eca de Queiroz

Confira os dados da prefeitura sobre os imóveis vendidos neste condomínio.

Valor de venda	Características	Data de venda
R$ 610.000 R$ 7.886/m²	77.35 m² • 2 quartos • 1 vaga • Andar baixo	fev/2020
R$ 370.000 R$ 4.783/m²	77.35 m² • 2 quartos • 1 vaga • Andar baixo	mai/2019

Fig. 25 – Informações de vendas no edifício Eça de Queiroz (Loft, 2023)

- **Oportunidade em análise #3: Rua João Moura, 328 - Pinheiros**

A tabela a seguir resume a oportunidade #3, mostrando algumas das suas informações.

Tabela 24

Resumo da oportunidade em análise #3

Exemplo 3			
Endereço	Rua João Moura, 328 - Pinheiros		
Área anunciada	119 m2		
Link	id = 2489309113		
Métrica de desvio	32%		
	Linear	**Xgboost**	**KNN**
Valor estimado	11613	13488	11722
Valor do anúncio	8017		

Fig. 26 – Fachada do edifício Araguaia

No terceiro caso, ao examinar o anúncio, aparenta haver consistência na área informada. Adicionalmente, ao analisar outras vendas similares na figura 27, torna-se evidente que o preço de venda anunciado está consideravelmente abaixo das transações reais, que variam de 9287 e até mesmo acima de 11000. Portanto, essa oferta merece uma avaliação manual mais minuciosa e uma investigação local mais aprofundada para possível aquisição.

Histórico de vendas em Condomínio Edificio Araguaia

Confira os dados da prefeitura sobre os imóveis vendidos neste condomínio.

Valor de venda	Características	Data de venda
R$ 940.000 R$ 9.287/m²	101.21 m² • 3 quartos • 1 vaga • Andar baixo	out/2022
R$ 940.000 R$ 9.287/m²	101.21 m² • 3 quartos • 1 vaga • Andar baixo	ago/2022
R$ 940.000 R$ 11.928/m²	78.8 m² • 2 quartos • 1 vaga • Andar baixo	mar/2021
R$ 1.140.000 R$ 11.287/m²	101 m² • 2 quartos • 1 vaga • Andar baixo	nov/2020
R$ 1.280.000 R$ 16.243/m²	78.8 m² • 2 quartos • 1 vaga • Andar alto	nov/2020

Fig. 27 – Informações de vendas no edifício Araguaia (Loft, 2023)

- **Oportunidade em análise #4: Avenida Jamaris, 100 - Moema**

A tabela a seguir resume a oportunidade #4, mostrando algumas das suas informações.

Tabela 25

Resumo da oportunidade em análise #4

Exemplo 4			
Endereço	Avenida Jamaris, 100 - Moema		
Área anunciada	57 m2		
Link	id = 2507081929		
Métrica de desvio	30%		
	Linear	**Xgboost**	**KNN**
Valor estimado	13011	11432	11732
Valor do anúncio	7996		

Fig. 28 – Fachada do edifício Cosmopolitan Mix (Loft, 2023)

No quarto caso, ao verificar o anúncio, foi identificada uma aparente inconsistência na área do imóvel. O anúncio menciona uma área de 57m2, enquanto as informações do condomínio indicam que os apartamentos possuem uma área máxima de 45m2, conforme evidenciado na Figura 29. O preço de 455 mil nesse contexto sugere um valor por metro quadrado próximo a 10 mil, o que torna essa oportunidade consideravelmente menos atrativa do que inicialmente suposto.

Fig. 29 – Informações sobre o edifício Cosmopolitan Mix, com valores de área (Loft, 2023)

Fig. 30 – Informações de vendas no edifício Cosmopolitan Mix (Loft, 2023)

- **Oportunidade em análise #5: Avenida Conselheiro Rodrigues Alves, 793 - Vila Mariana**

A tabela a seguir resume a oportunidade #5, mostrando algumas das suas informações.

Tabela 26

Resumo da oportunidade em análise #5

Exemplo 5			
Endereço	Avenida Conselheiro Rodrigues Alves, 793		
Área anunciada	180 m2		
Link	id = 2457773326		
Métrica de desvio	30%		
	Linear	**Xgboost**	**KNN**
Valor estimado	10943	11086	10490
Valor do anúncio	7194		

Fig. 31 – Fachada do edifício Ilha de Mont Serrat

No quinto caso, ao analisar o anúncio, percebe-se que a área informada parece estar em conformidade com as expectativas. Além disso, o preço também parece estar alinhado com as transações recentes no mesmo edifício, conforme evidenciado na Figura 32, onde os valores variam entre 5875 e 6060 por metro quadrado.

Histórico de vendas em Condomínio Ilha de Mont Serrat

Confira os dados da prefeitura sobre os imóveis vendidos neste condomínio.

Valor de venda	Características	Data de venda
R$ 990.000 R$ 5.875/m²	168.5 m² • 4 quartos • 3 vagas • Andar baixo	dez/2021
R$ 2.000.000 R$ 6.060/m²	330 m² • 4 quartos • 4 vagas • Andar baixo	abr/2021

Fig. 32 – Informações de vendas no edifício Ilha de Mont Serrat (Loft, 2023)

5.5 Discussão

Em um contexto de análise de oportunidades de investimento imobiliário, dentre as 30 oportunidades inicialmente selecionadas, optou-se por uma análise mais aprofundada de apenas 5 delas. É importante ressaltar que essa seleção não abrange toda a amostra, mas oferece insights valiosos sobre alguns padrões cruciais. Entre esses padrões, destacam-se os três seguintes:

1. Informações não confiáveis: Uma constatação relevante é que nem sempre as informações disponíveis nos anúncios estão corretas, o que pode afetar drasticamente a análise. Por exemplo, quando a área anunciada de uma propriedade é superior à realidade, o preço aparenta ser mais baixo por metro quadrado, mas essa percepção é enganosa. Isso ocorreu nas oportunidades 1 e 4, invalidando esses casos como possíveis oportunidades.

2. Tendência de Superestimação do Preço na Modelagem: Mesmo quando as informações são precisas, o modelo de avaliação pode, em certos casos, tender a superestimar o preço da propriedade. Isso pode ser observado quando comparamos o preço previsto com os preços de vendas de unidades semelhantes no mesmo prédio. Essa superestimação pode ocorrer devido a características específicas da propriedade que confundem o modelo, como no caso das oportunidades 2 e 5.

3. Possíveis Casos Genuínos de Oportunidade: Em contrapartida, em alguns casos, uma oportunidade de investimento parece genuína. Por exemplo, na oportunidade 3, uma análise de campo detalhada, investigação da situação legal da propriedade, avaliação visual e da região, bem como uma análise minuciosa realizada por especialistas, podem confirmar a viabilidade da oportunidade.

É importante ressaltar que, mesmo com uma taxa de oportunidades de apenas 3% em uma amostra, das quais 10-20% podem se mostrar efetivamente viáveis, esse percentual pode representar um número significativo de propriedades com potencial de investimento. Tomando como exemplo a cidade de São Paulo, que conta com aproximadamente 1 milhão de anúncios de propriedades apenas no website Zap Imóveis (ZAP, 2023), esse cenário pode se traduzir em um intervalo de 30.000 a 60.000 propriedades com potencial de compra. Mesmo após uma análise rigorosa que venha a resultar em apenas 0.1% das propriedades inequivocamente abaixo do valor de mercado, ainda restariam mais de 30 propriedades que poderiam ser consideradas para compra e posterior revenda com ágio ou para moradia.

Além disso, é crucial considerar a dinâmica do mercado, onde novas oportunidades surgem constantemente. Esse dinamismo motiva a busca diária ou semanal de propriedades por parte de tomadores de decisão interessados, visando aproveitar as flutuações e identificar oportunidades lucrativas no mercado imobiliário.

6. Conclusão

O objetivo deste trabalho é desenvolver uma metodologia sistemática para identificar propriedades imobiliárias que possam estar subvalorizadas, tornando o processo de busca por apartamentos para investimento ou oportunidades de negócios mais consistente e dinâmico. A abordagem metodológica empregou três modelos distintos de aprendizado de máquina, otimizados com hiperparâmetros cuidadosamente ajustados e avaliados com uma ampla variedade de métricas para garantir consistência e robustez. O modelo XGBoost se destacou como o mais eficaz, com um Erro Quadrático Médio (RMSE) de 1170, superando os modelos KNN e Hedônico, com valores de respectivamente 1895 e 1920. O erro percentual médio absoluto (MAPE) alcançou 8.6%, em contraste com os valores 11.2% e 14.6% obtidos nos modelos KNN e Hedônico, respectivamente.

Nesse contexto, a metodologia de identificação de oportunidades demonstrou sua capacidade de distinguir casos atípicos e gerar uma lista de oportunidades que abrange cerca de 3% da base de dados total. Uma análise estatística e revisão manual complementar permitiu identificar possíveis oportunidades de forma mais precisa.

Contudo, é fundamental destacar que a metodologia apresenta limitações importantes. Em primeiro lugar, a base de dados utilizada neste estudo não contempla todas as variáveis relevantes que podem afetar os preços das propriedades, limitando-se às informações disponíveis. Variáveis não incluídas, como o estado de conservação, questões de ruído ou problemas judiciais, podem impactar em casos específicos, potencialmente influenciando a percepção de oportunidades.

Em segundo lugar, erros de inserção de dados, intencionais ou não, podem introduzir ruído nos resultados e levar à identificação de oportunidades que, na prática, podem não ser vantajosas.

No entanto, é importante ressaltar que o propósito deste estudo é fornecer uma lista de oportunidades que possa ser aplicada em diversos contextos imobiliários e servir como um ponto de partida em análises mais abrangentes para tomada de decisões de compra. A metodologia aqui apresentada oferece uma abordagem científica para avaliação, direcionada a investidores, compradores e proprietários de imóveis, e pode ser enriquecida com informações adicionais que não foram contempladas neste estudo.

Para futuras pesquisas, é recomendável ampliar esta metodologia considerando a aplicação de outros modelos de Machine Learning, como Redes Neurais, para explorar diferentes abordagens de modelagem. Além disso, seria interessante estender a análise para abranger outras cidades e distritos na região de São Paulo, utilizando conjuntos de dados mais atualizados. Isso permitiria a generalização dos resultados e uma compreensão mais abrangente das dinâmicas imobiliárias em diferentes localidades. Por fim, sugere-se considerar outras métricas de avaliação em futuros trabalhos, a fim de enriquecer a metodologia e proporcionar uma visão mais completa da performance dos modelos utilizados.

7. Referências bibliográficas

Alhodiry, A., Rjoub, H., & Samour, A. (2021). Impact of oil prices, the U.S interest rates on Turkey's real estate market. New evidence from combined co-integration and bootstrap ARDL tests. PloS one, 16(1), e0242672. https://doi.org/10.1371/journal.pone.0242672.

ANAIS DO SIMPÓSIO BRASILEIRO DE BANCO DE DADOS (SBBD): https://sol.sbc.org.br/index.php/sbbd/index.

Aranha, F. (1997). Atlas dos setores postais: uma nova geografia a serviço da empresa. Revista de Administração de Empresas, 37(3), 20-27.

Armstrong, J Scott, and Fred Collopy. 1992. “Error Measures for Generalizing about Forecasting Methods: Empirical Comparisons.” International Journal of Forecasting, 8(1), 69–80. https://doi.org/10.1016/S0169-2070(98)00007-7.

Armstrong, J Scott. 1985. Long-Range Forecasting: From Crystal Ball to Computer. Wiley.

Baldominos, A., Blanco, I., Moreno, A. J., Iturrarte, R., Bernárdez, Ó., & Afonso, C. (2018). Identifying real estate opportunities using machine learning. Applied sciences, 8(11), 2321.

Benson, E. D., Hansen, J. L., Schwartz, A. L., & Smersh, G. T. (1998). Pricing residential amenities: the value of a view. The Journal of Real Estate Finance and Economics, 16, 55-73.

Cazzolato, J. D. (2005). Os bairros como instância territorial local-contribuição metodológica para o caso de São Paulo (Doctoral dissertation, Universidade de São Paulo).

Cervero, R., & Kang, C. D. (2011). Bus rapid transit impacts on land uses and land values in Seoul, Korea. Transport Policy, 18(1), 102–116. https://doi.org/10.1016/j.tranpol.2010.06.005.

Chau, K. W., & Chin, T. L. (2003). A critical review of literature on the hedonic price model. International Journal for Housing Science and its applications, 27(2), 145-165.

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. https://doi.org/10.1145/2939672.2939785.

Choy, L. H., & Ho, W. K. (2023). The Use of Machine Learning in Real Estate Research. Land, 12(4), 740.

Cordera, R., Coppola, P., dell'Olio, L., & Ibeas, Á. (2019). The impact of accessibility by public transport on real estate values: A comparison between the cities of Rome and Santander. Transportation Research Part A: Policy and Practice, 125, 308-319.

D'Acci, L. (2019). Quality of urban area, distance from city centre, and housing value. Case study on real estate values in Turin. Cities, 91, 71–92. https://doi.org/10.1016/j.cities.2018.11.008.

De Myttenaere, A., Golden, B., Le Grand, B., & Rossi, F. (2016). Mean absolute percentage error for regression models. Neurocomputing, 192, 38-48.

Dimopoulos, T., & Bakas, N. (2019). An artificial intelligence algorithm analyzing 30 years of research in mass appraisals.

Dimopoulos, T., Tyralis, H., Bakas, N. P., & Hadjimitsis, D. (2018). Accuracy measurement of Random Forests and Linear Regression for mass appraisal models that estimate the prices of residential apartments in Nicosia, Cyprus. Advances in Geosciences, 45, 377–382. https://doi.org/10.5194/adgeo-45-377-2018.

Guliker, E., Folmer, E., & Sinderen, M. (2022). Spatial Determinants of Real Estate Appraisals in The Netherlands: A Machine Learning Approach. ISPRS Int. J. Geo Inf., 11, 125. https://doi.org/10.3390/ijgi11020125.

Haurin, D. R., Haurin, J. L., Nadauld, T., & Sanders, A. (2010). List prices, sale prices and marketing time: an application to US housing markets. Real Estate Economics, 38(4), 659-685.

Heidari, M., Zad, S., & Rafatirad, S. (2021, April). Ensemble of supervised and unsupervised learning models to predict a profitable business decision. In 2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS) (pp. 1-6). IEEE.

Herath, S., & Maier, G. (2010). The hedonic price method in real estate and housing market research: A review of the literature.

Huang, P., & Hess, T. (2018). Impact of Distance to School on Housing Price: Evidence from a Quantile Regression.

Hui, E. C., Chau, C. K., Pun, L., & Law, M. Y. (2007). Measuring the neighboring and environmental effects on residential property value: Using spatial weighting matrix. Building and environment, 42(6), 2333-2343.

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. International Journal of Forecasting, 22(4), 679-688.

Jha, S., Babiceanu, R., Pandey, V., & Jha, R. (2020). Housing Market Prediction Problem using Different Machine Learning Algorithms: A Case Study. ArXiv, abs/2006.10092.

Jim, C. Y., & Chen, W. Y. (2009). Value of scenic views: Hedonic assessment of private housing in Hong Kong. Landscape and urban planning, 91(4), 226-234.

Jin, T., Cheng, L., Liu, Z., Cao, J., Huang, H., & Witlox, F. (2022). Nonlinear public transit accessibility effects on housing prices: Heterogeneity across price segments. Transport Policy, 117, 48-59.

Juszczak, P., Tax, D., & Duin, R. P. (2002, May). Feature scaling in support vector data description. In Proc. asci (pp. 95-102). Citeseer.

Kang, H., & Gardner, M. (1989). Selling price and marketing time in the residential real estate market. Journal of Real Estate Research, 4(1), 21-35.

Koh, P. W., & Liang, P. (2017). Understanding Black-box Predictions via Influence Functions.

Komagome-Towne, A. (2016). Models and Visualizations for Housing Price Prediction.

Kontrimas, V., & Verikas, A. (2011). The mass appraisal of the real estate by computational intelligence. Applied Soft Computing, 11(1), 443–448. https://doi.org/10.1016/j.asoc.2009.12.003.

Kuşan, H., Aytekin, O., & Özdemir, İ. (2010). The use of fuzzy logic in predicting house selling price. Expert Systems with Applications, 37(3), 1808–1813. https://doi.org/10.1016/j.eswa.2009.07.031.

Liu, J.-G., Zhang, X.-L., & Wu, W.-P. (2006). Application of Fuzzy Neural Network for Real Estate Prediction. In J. Wang, Z. Yi, J. M. Zurada, B.-L. Lu, & H. Yin (Orgs.), Advances in Neural Networks—ISNN 2006 (Vol. 3973, p. 1187–1191). Springer Berlin Heidelberg. https://doi.org/10.1007/11760191_173.

Liu, X., Deng, Z., & Wang, T. (2011). Real estate appraisal system based on GIS and BP neural network. Transactions of Nonferrous Metals Society of China, 21, s626–s630. https://doi.org/10.1016/S1003-6326(12)61652-5.

LOFT. Cosmopolitan Mix - Moema Índios, São Paulo, SP. Disponível em: https://loft.com.br/condominio/cosmopolitan-mix-moema-indios-sao-paulo-sp/1wmfycb. Acesso em: 7 out. 2023.

LOFT. Edifício Araguaia - Pinheiros, São Paulo, SP. Disponível em: https://loft.com.br/condominio/edificio-araguaia-pinheiros-sao-paulo-sp/oqa2bd. Acesso em: 7 out. 2023.

LOFT. Edifício Eça de Queiroz - Vila Mariana, São Paulo, SP. Disponível em: https://loft.com.br/condominio/edificio-eca-de-queiroz-vila-mariana-sao-paulo-sp/v2imtm. Acesso em: 7 out. 2023.

LOFT. Edifício Juliana - Moema Índios, São Paulo, SP. Disponível em: https://loft.com.br/condominio/edificio-juliana-moema-indios-sao-paulo-sp/6pxfrw. Acesso em: 7 out. 2023.

LOFT. Ilha de Mont Serrat - Vila Mariana, São Paulo, SP. Disponível em: https://loft.com.br/condominio/ilha-de-mont-serrat-vila-mariana-sao-paulo-sp/15t4u3l. Acesso em: 7 out. 2023.

Lughofer, E., Trawiński, B., Trawiński, K., Kempa, O., & Lasota T. (2011). On employing fuzzy modeling algorithms for the valuation of residential premises. Information Sciences, 181(23), 5123–5142. https://doi.org/10.1016/j.ins.2011.07.012.

Ma, D., Lv, B., Li, X., Li, X., & Liu, S. (2023). Heterogeneous Impacts of Policy Sentiment with Different Themes on Real Estate Market: Evidence from China. Sustainability. https://doi.org/10.3390/su15021690.

Masías, V. H., Valle, M. A., Crespo, F., Crespo, R., Vargas, A., & Laengle, S. (2016). Property Valuation using Machine Learning Algorithms: A Study in a Metropolitan-Area of Chile.

McCluskey, W. J., McCord, M., Davis, P. T., Haran, M., & McIlhatton, D. (2013). Prediction accuracy in mass appraisal: a comparison of modern approaches. Journal of Property Research, 30(4), 239-265.

Miller, N., Sah, V., & Sklarz, M. (2018). Estimating Property Condition Effect on Residential Property Value: Evidence from U.S. Home Sales Data. Journal of Real Estate Research, 40(2), 179–198. https://doi.org/10.1080/10835547.2018.12091497.

Morano, P., De Mare, G., & Tajani, F. (2013). LMS for Outliers Detection in the Analysis of a Real Estate Segment of Bari. In B. Murgante, S. Misra, M. Carlini, C. M. Torre, H.-Q. Nguyen, D. Taniar, B. O. Apduhan, & O. Gervasi (Orgs.), Computational Science and Its Applications – ICCSA 2013 (Vol. 7974, p. 457–472). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-39649-6_33.

Osborne, J., & Overbay, A. (2004). The power of outliers (and why researchers should ALWAYS check for them). Practical Assessment, Research and Evaluation, 9, 6. https://doi.org/10.7275/QF69-7K43.

Pagourtzi, E., Assimakopoulos, V., Hatzichristos, T., & French, N. (2003). Real estate appraisal: a review of valuation methods. Journal of Property Investment & Finance, 21(4), 383-401.

Park, B., & Bae, J. K. (2015). Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data. Expert Systems with Applications, 42(6), 2928–2934. https://doi.org/10.1016/j.eswa.2014.11.040.

Perez, R. A., & Kimura, D. S. (2014). ANÁLISE DE MERCADO COMO FERRAMENTA PARA A ABERTURA DE NOVOS LOTEAMENTOS MARKET ANALYSIS AS TOLL FOR NEW DEVELOPMENTS STARTUP. 3(1), 13.

PREFEITURA MUNICIPAL DE SÃO PAULO - PMSP. "Atlas do Trabalho de Desenvolvimento do Município de São Paulo" (2007). Available in <http://atlasmunicipal.prefeitura.sp.gov.br/Login/Login.aspx >. Accessed on October 15, 2020.

PREFEITURA MUNICIPAL DE SÃO PAULO - PMSP. “Dados demográficos dos distritos pertencentes às Subprefeituras” (2010). Available in: <https://www.prefeitura.sp.gov.br/cidade/secretarias/subprefeituras/subprefeituras/dados_demograficos/index.php?p=12758 >. Accessed on October 15, 2020.

PREFEITURA MUNICIPAL DE SÃO PAULO - PMSP. “Ranking de violência da cidade de São Paulo por distrito” (2019). Available in: <https://www.prefeitura.sp.gov.br/cidade/secretarias/upload/SITE/Ranking%20-%20TODOS%20(atualizado%20com%20viol%C3%AAncia)_compressed.pdf >. Acesso em 8 de Dezembro de 2022.

Rafiei, M. H., & Adeli, H. (2016). A Novel Machine Learning Model for Estimation of Sale Prices of Real Estate Units. Journal of Construction Engineering and Management, 142(2), 04015066. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001047.

Renigier-Biłozor, M., Janowski, A., & d’Amato, M. (2019). Automated Valuation Model based on fuzzy and rough set theory for real estate market with insufficient source data. Land Use Policy, 87, 104021. https://doi.org/10.1016/j.landusepol.2019.104021.

Rosen, S. (1974). Hedonic prices and implicit markets: product differentiation in pure competition. Journal of political economy, 82(1), 34-55.

Schober, P., Boer, C., & Schwarte, L. A. (2018). Correlation coefficients: appropriate use and interpretation. Anesthesia & Analgesia, 126(5), 1763-1768.

Schulz, R., & Werwatz, A. (2004). A State Space Model for Berlin House Prices: Estimation and Economic Interpretation. 21.

Selim, H. (2009). Determinants of house prices in Turkey: A quantile regression analysis. ERES 2009, 17.

Shi, J., & Cheung, K. K. (2017). Land supply and real estate prices: Empirical evidence from Hong Kong. Habitat International, 60, 1-10. https://doi.org/10.1016/j.habitatint.2016.12.004.

Simar, L., & Zelenyuk, V. (2017). Quantile regression for dynamic panel data with fixed effects. Journal of Econometrics, 197(2), 111–126. https://doi.org/10.1016/j.jeconom.2016.06.002.

Su, Y. C., & Wang, T. C. (2018). Impact of proximity to public transportation on housing prices in metropolitan districts of Taipei. Sustainable Cities and Society, 42, 224-230.

Sun, X., Zhu, Y., Lin, Z., He, Y., & Zhang, C. (2022). Understanding Housing Price Formation Mechanism in Megacities: Evidence from Shanghai. Sustainability, 14(1), 41.

Sun, Z., Wu, W., & Liu, H. (2014). A new approach to real estate valuation using neuro-fuzzy model. Expert Systems with Applications, 41(6), 3080–3086. https://doi.org/10.1016/j.eswa.2013.10.032.

Swinburn, T. K., & Teh, B. T. (2003). Application of artificial neural networks to the prediction of residential property values. Journal of Property Investment & Finance, 21(4), 328–346. https://doi.org/10.1108/14635780310486072.

Tang, G., & Nara, Y. (2021). Understanding housing prices in a hyper-centralized city: A case study of Beijing. Land Use Policy, 107, 105514. https://doi.org/10.1016/j.landusepol.2021.105514.

Tang, L., & Li, S. (2021). Identifying neighborhood determinants of housing prices: A new method and evidence from Shanghai. Journal of Housing and the Built Environment, 36(4), 1127-1150. https://doi.org/10.1007/s10901-021-09820-6.

Tian, L., Zhao, Y., Li, J., Wang, H., & Niu, D. (2019). A land use regression model for estimating the PM2. 5 concentration in Beijing. Chemosphere, 218, 908-915.

Tofallis, C. (2015). A better measure of relative prediction accuracy for model selection and model estimation. Computational Statistics & Data Analysis, 89, 190–207. https://doi.org/10.1016/j.csda.2015.03.019.

Tofallis, C. (2019). Rethinking the Mean Absolute Percentage Error. International Journal of Forecasting, 35(4), 1448-1456. https://doi.org/10.1016/j.ijforecast.2019.01.004.

Trivedi, K. S. (2002). Regression analysis of count data. Cambridge University Press.

Tsaousis, K., Kitsakis, D., Meletiou, G., & Karkanis, I. (2019). Fuzzy logic in real estate valuation: a critical review. Journal of Property Investment & Finance, 37(2), 196–225. https://doi.org/10.1108/JPIF-09-2018-0090.

Vargas, S. R. (2005). Avaliação Imobiliária: métodos estatísticos para estimação de preços. São Paulo: Nobel.

Verikas, A., & Bacauskiene, M. (2006). Forestalling building maintenance of multi-apartment houses. Fuzzy sets and systems, 157(5), 663–680. https://doi.org/10.1016/j.fss.2005.09.003.

Vieira, D. M., Veríssimo, L., & Antunes, P. (2022). Housing Prices Prediction: A Machine Learning Approach. Sustainability, 14(5), 2079. https://doi.org/10.3390/su14052079.

Wan, X., & Kim, S. (2021). Automated Valuation Models for Real Estate Appraisal: A Review of the State of the Art. Sustainability, 13(17), 9486. https://doi.org/10.3390/su13179486.

Wang, L., He, G., Zhang, Z., & Ma, S. (2022). The impact of land use regulation on housing prices in China: A heterogeneity perspective. Cities, 124, 103617. https://doi.org/10.1016/j.cities.2022.103617.

Wang, T. C., Wu, C. C., & Chang, P. Y. (2010). The impact of mass rapid transit on population density: the case of Taipei metropolitan area. Cities, 27(5), 335-343.

Wu, J., & Peng, Y. (2018). Application of a Fuzzy Logic Model for Real Estate Valuation. In D. Li, Z. Zhao, & M. Kim (Orgs.), Communications and Networking (p. 163–177). Springer International Publishing. https://doi.org/10.1007/978-3-030-00713-3_14.

Wu, J., & Tang, L. (2015). Assessing the impact of urban public transportation expansions on housing prices: The case of the Los Angeles Orange Line. Transport Policy, 44, 117-125. https://doi.org/10.1016/j.tranpol.2015.08.001.

Wu, J., & Tang, L. (2019). Estimating the impact of public transportation expansion on real estate prices in Los Angeles: A spatial difference-in-differences approach. Journal of Transport Geography, 74, 199-208. https://doi.org/10.1016/j.jtrangeo.2018.12.013.

Wu, J., & Tang, L. (2020). Impact of public transportation on home prices: a spatial–temporal study of Los Angeles bus rapid transit. Journal of Geographical Systems, 22(2), 179-201. https://doi.org/10.1007/s10109-019-00301-4.

Wu, J., & Yin, H. (2020). The impacts of public transportation on residential property values in Chicago. Journal of Transport Geography, 82, 102606. https://doi.org/10.1016/j.jtrangeo.2019.102606.

Wu, J., Yin, H., & Tang, L. (2019). Assessing the impact of public transportation on housing prices: A longitudinal analysis in Los Angeles from 1996 to 2011. Urban Studies, 56(6), 1232-1251. https://doi.org/10.1177/0042098017728495.

Xie, W., Wang, X., & Zhao, X. (2014). Impact of metro accessibility on residential property value: Hedonic price analysis in Hangzhou, China. Habitat International, 44, 240-246. https://doi.org/10.1016/j.habitatint.2014.09.018.

Xu, Z., Li, W., Wang, X., & Huang, L. (2017). Effects of public transit on housing development: Evidence from Nanjing. Habitat International, 68, 59-68. https://doi.org/10.1016/j.habitatint.2017.08.002.

Yang, C. (2008). Residential property assessment using support vector machines. Expert Systems with Applications, 34(3), 1822–1832. https://doi.org/10.1016/j.eswa.2007.01.030.

Yang, T., & Mao, L. (2021). The Impacts of Urban Rail Transit on Housing Prices in Medium-Sized Cities: A Quasi-Natural Experiment in Suzhou. Sustainability, 13(11), 5826. https://doi.org/10.3390/su13115826.

Zhang, D., Ding, X., & Kim, Y. S. (2021). Housing Price Prediction Using LSTM Neural Networks. Sustainability, 13(11), 6156. https://doi.org/10.3390/su13116156.

Zhao, X., Zhao, X., Wu, J., & Wei, F. (2020). Housing prices around transit stations: Impact of the subway expansion in Nanjing, China. Land Use Policy, 94, 104522. https://doi.org/10.1016/j.landusepol.2020.104522.

Zheng, H., Shen, Y., & Wang, L. (2021). Forecasting the Impact of Urban Rail Transit on Housing Prices Based on Multi-Scale Spatially Weighted Regression. Sustainability, 13(19), 10977. https://doi.org/10.3390/su131910977.

Zhu, H., Sun, M., & Zhang, S. (2020). Understanding housing prices in urban Beijing: A mixed geographically weighted regression approach. Applied Geography, 119, 102241. https://doi.org/10.1016/j.apgeog.2020.102241.