UNIVERSIDADE DE SÃO PAULO
ESCOLA POLITÉCNICA

CELSO GABRIEL DE AZEVEDO RIBEIRO

Estudo sobre o impacto das políticas públicas através da modelagem de preços aplicada ao mercado de *Real Estate* na cidade de São Paulo

São Paulo
2023

RIBEIRO, CELSO	**Estudo sobre o impacto das políticas públicas através da modelagem de preços aplicada ao mercado de *Real Estate* na cidade de São Paulo**		São Paulo 2023

CELSO GABRIEL DE AZEVEDO RIBEIRO

Estudo sobre o impacto das políticas públicas através da modelagem de preços aplicada ao mercado de *Real Estate* na cidade de São Paulo

VERSÃO ORIGINAL

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

Orientador:

Prof. Dr. Flávio Almeida de Magalhães Cipparrone

SÃO PAULO

2023

Catalogação-na-publicação

Ribeiro, Celso
Estudo Sobre o Impacto das Políticas Públicas através da Modelagem de Preços Aplicada ao Mercado de Real Estate na Cidade de São Paulo / C. Ribeiro -- São Paulo, 2023.
85 p.

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

1.Real Estate 2.Machine Learning 3.Investimento Público I.Universidade de São Paulo. Escola Politécnica. Departamento de Engenharia de Sistemas Eletrônicos II.t.

Nome: RIBEIRO, Celso Gabriel de Azevedo

Título: Estudo sobre o impacto das políticas públicas através da modelagem de preços aplicada ao mercado de *Real Estate* na cidade de São Paulo

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

Aprovado em:

Banca Examinadora

Prof. Dr. _____

Instituição: _____

Julgamento: _____

Prof. Dr. _____

Instituição: _____

Julgamento: _____

Prof. Dr. _____

Instituição: _____

Julgamento: _____

AGRADECIMENTOS

Agradeço primeiramente ao professor Flavio Cipparone, que com gentileza e atenção aceitou me orientar nesse trabalho e proporcionou conselhos e pontos de vista valiosos para seu desenvolvimento e plenitude.

Agradeço também a meu grande amigo Flavio de Falcão que ajudou em diversas partes deste trabalho, ao escrever em conjunto artigo científico que está sendo publicado e, dessa forma, originado diversas das discussões aqui apresentadas.

Finalmente, agradeço à minha família e aos nossos amigos, que proporcionaram a calma e tranquilidade durante o desenvolvimento desse trabalho.

"My great concern is not whether you have failed, but whether you are content with your failure."

(Abraham Lincoln)

RESUMO

O orçamento público dos entes federativos brasileiros comumente é visto como fonte de controvérsia, seja pela escolha das prioridades para alocação dos recursos, seja pela crise fiscal que o país passa. Este trabalho tem como objetivo principal proporcionar uma estimativa do impacto gerado por algumas políticas públicas quando implementadas no ambiente construído da cidade de São Paulo, sendo mensurado através da previsão do preço dos imóveis da cidade e consequente avaliação dos fatores que impactaram para a precificação. Esta predição será feita usando aprendizado de máquina, especificamente o modelo baseado em árvore XGBoost, comparado com um modelo hedônico de previsão de preços (*Semi-Log Regression*). Na parte de análise será utilizado o SHAP, que é um método *Tree Explainer*, para estimar o impacto das variáveis do modelo de predição de preços e avaliar suas importâncias. A modelagem feita considera variáveis intrínsecas (número de quartos, área construída, ano de construção, etc.), bem como variáveis extrínsecas (qualidade do pavimento, transportes públicos, arborização, etc.) e tem como parâmetro de treinamento o preço de mercado, obtido de anúncios online. Finalmente, concluída a análise, é possível identificar as variáveis extrínsecas, relacionadas ao investimento público, que trouxeram maior impacto positivo no preço dos imóveis, sugerindo sua escolha para o administrador público responsável.

***Palavras-chave**: Real Estate, investimento público, precificação, XGBoost*

ABSTRACT

The public budget of Brazilian federative entities is commonly seen as a source of controversy, either because of the choice of priorities for resource allocation, or because of the fiscal crisis that the country is going through. The main objective of this work is to provide an estimate of the impact generated by some public policies when implemented in the built environment of the city of São Paulo, being measured by predicting the price of real estate in the city and consequent evaluation of the factors that impacted on pricing. This prediction will be made using machine learning, specifically the XGBoost tree-based model, compared with a hedonic price prediction model (Semi-Log Regression). In the analysis part, SHAP will be used, which is a Tree Explainer method, to estimate the impact of the variables of the pricing model and evaluate their importance. The modeling that was made considers intrinsic variables (number of rooms, built area, year of construction, etc.), as well as extrinsic variables (pavement quality, public transport, afforestation, etc.) and has the market price as a training parameter, obtained from online advertisements. Finally, once the analysis is concluded, it is possible to identify the extrinsic variables, related to public investment, which had the greatest positive impact on property prices, suggesting their choice to the responsible public administrator.

Keywords: *Real Estate, public investment, pricing, XGBoost*

SUMÁRIO

1 Introdução

É de grande relevância notar como o investimento público pode ser usado para fortalecer certos pontos fracos na economia e, assim, atuar para o crescimento econômico e consequente redução da pobreza, com melhoria na qualidade de vida da população. O direcionamento mais eficiente dos investimentos públicos tornou-se cada vez mais importante em uma era de reformas macroeconômicas nas quais os governos estão sob pressão para reduzir os orçamentos. (FAN ET AL, 2004).

Por outro lado, imóveis representam um dos bens mais importantes que uma pessoa pode possuir ao longo de sua vida. De fato, as casas não são apenas um lugar para as pessoas viverem, mas geralmente representam o componente mais relevante da riqueza privada. Portanto, a modelagem de preços imobiliários é importante para diferentes atores do mercado: para políticas governamentais, sendo fundamental para previsões econômicas de curto prazo e bem-estar social; para os negócios, uma vez que as construtoras se deparam com a difícil questão sobre a viabilidade de construir ou não e precisam de informações para tomar uma decisão correta (RAFIEI & 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.

Este trabalho buscará unir esses dois temas, de forma a entender como o investimento público, realizado especificamente no ambiente construído das cidades, pode impactar nos preços dos ativos imobiliários e como esse investimento pode ser feito de forma inteligente a maximizar o impacto positivo, trazendo valorização econômica, maiores preços de imóveis e consequentemente uma maior arrecadação de impostos.

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[1], no qual o autor deste trabalho também figura entre a lista de autores. Além disso, outro artigo, dos mesmos autores, está em processo de publicação, na revista *Real Estate Finance*, também baseado neste trabalho. Assim sendo, cabe destacar que este referido capítulo foi uma construção conjunta entre este autor e Flavio F. Helena.

1.1. Objetivos

O presente trabalho é composto por três principais objetivos aos quais se pretende responder: **1 –** treinar e testar um modelo de predição de preços de ativos imobiliários, **2 –** avaliar quais são as variáveis mais impactantes desse modelo e quais delas são relacionadas ao investimento público, mensurando seu impacto e, finalmente, **3 –** a partir dessa avaliação, encontrar aqueles investimentos que possuem os maiores impactos positivos (valorização) no preço dos imóveis e na arrecadação de impostos relacionados, sugerindo sua escolha para o administrador público.

1.2. Justificativa

O tema da eficiência do investimento público e a análise de seu custo-benefício já vem sendo abordado pela academia há algum tempo. Desde a década de 60 Maass (1966) já defende sua importância, assim como as limitações dessa análise devido aos ganhos muitas vezes intangíveis de certas políticas públicas. Nessa linha, pesquisas mais recentes já identificam através de modelos hedônicos que alguns investimentos públicos, como em transporte, aumentam o preço da habitação dos bairros beneficiados direta ou indiretamente. (CHERNOFF & CRAIG, 2022). Finalmente, com os avanços vivenciados recentemente na área de pesquisa do *machine learning*, diversos estudos vêm surgindo usando métodos de aprendizado de máquina na precificação de imóveis, com resultados superiores aos modelos hedônicos. (SELIM, 2009)

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

1.3. Metodologia

No primeiro objetivo, a implementação de um modelo de predição dos preços dos imóveis, serão considerados 4.019 imóveis situados em seis diferentes bairros da cidade de São Paulo, Brasil, nos quais serão relacionados diferentes conjuntos de dados, resultando em uma base de dados com 42 variáveis de modelo (entrada) e uma variável de objetivo (saída), para a qual o modelo será treinado e testado.

Já no segundo objetivo, a mensuração do impacto de cada uma das variáveis no modelo, será utilizado o SHAP, que, como dito, é um método *Tree Explainer*, que conta com diversas ferramentas gráficas e matemáticas para discutir o impacto de cada uma das variáveis na variação dos preços dos imóveis estimados pelo modelo desenvolvido.

Por fim, no terceiro objetivo, que tentará responder a pergunta: "*Qual é o investimento público que ao mesmo tempo possui o maior impacto positivo nos preços dos imóveis (segundo o modelo desenvolvido)?*", serão, ainda através do uso do SHAP, mensurados os impactos estimados no preço previsto pelo modelo, de cada um dos investimentos públicos considerados inicialmente. Para esses investimentos também serão consultados seus impactos diretos na arrecadação de impostos urbanos, especificamente do Imposto Predial e Territorial Urbano (IPTU), considerando a forma de cálculo do imposto, e propondo uma estimativa de impacto nesta receita pública.

1.4. Estrutura

Este trabalho está dividido em 5 capítulos, como é explicado a seguir:

No **Capítulo 2** é realizada a revisão bibliográfica dos principais temas abordados no trabalho. São revisados e avaliados trabalhos recentes que discorrem sobre os temas de precificação imobiliária, tais como: impacto de variáveis intrínsecas e extrínsecas em preços de imóveis, aplicação de *machine learning* na precificação, utilização de métodos explanatórios para modelos de *machine learning.* Além disso também serão apresentados trabalhos que tratam do investimento público no ambiente construído da cidade e seus impactos no mercado imobiliário.

No **Capítulo 3** é explicada a metodologia utilizada para desenvolvimento e construção da base de dados que será utilizada na modelagem. Neste capítulo, a origem de cada dado é evidenciada assim como todas as premissas adotadas para junção dos

dados. Também se explica sobre o motivo das escolhas das variáveis e como os imóveis objeto de estudo foram selecionados para fazer parte do Modelo.

No **Capítulo 4** é iniciada uma análise exploratória de dados e alguns dos principais pontos encontrados a partir dela são discutidos. Algumas hipóteses são levantadas e relacionadas com outros resultados conhecidos da literatura.

No **Capítulo 5** as modelagens (aprendizado de máquina e hedônica) são implementadas, seus resultados são comparados e várias análises são realizadas para comparar a eficiência de cada modelo. Feito isso, o modelo de *machine learning* é submetido à um método explanatório (SHAP), no qual suas variáveis de entradas podem ter seu impacto mensurado no resultado final e as suas relações intrínsecas com o modelo são analisadas.

No **Capítulo 6** há uma discussão e modelagem para entender como o cálculo do Imposto Predial e Territorial Urbano (IPTU) depende diretamente do preço dos imóveis. Em sequência é estimado, a partir de todos os resultados obtidos até este momento, o impacto na arrecadação deste imposto, avaliando as variáveis relacionadas ao investimento público que afetam os preços dos imóveis.

No **Capítulo 7** são feitas as considerações finais do trabalho, resumindo os resultados mais relevantes encontrados até aqui. Além disso, são explicitadas as limitações do presente trabalho e são colocadas sugestões para trabalhos futuros no sentido de explorar o desenvolvimento do tema em novas aplicações.

2 Revisão Bibliográfica

Avaliações imobiliárias são fornecidas por diferentes atores do mercado, como avaliadores, imobiliárias, corretores, investidores e gestores de fundos, além de outros especialistas (SELIM, 2009). Geralmente, o valor dos imóveis está diretamente ligado a uma avaliação de mercado livre, regida por mecanismos de oferta e demanda. Segundo Pagourtzi et. al (2003), o valor de um imóvel é uma estimativa de quanto o imóvel valeria se fosse vendido no mercado livre. As suposições incluem não apenas a condição física do edifício, mas também o momento do mercado e a consideração de outros compradores nesse mercado.

Pode-se dizer que a precificação de habitações é influenciada por duas amplas categorias: características intrínsecas e extrínsecas. As características intrínsecas englobam todos os elementos estritamente relacionados à própria residência, como a quantidade de quartos e banheiros, a presença de janelas, varandas, a idade da edificação, entre outros. Em contraste, as características extrínsecas compreendem todos os fatores geográficos que não mantêm uma ligação direta com a propriedade em si, tais como a proximidade a instalações públicas como parques e praças urbanas, a qualidade e disponibilidade do transporte público, as condições das vias e calçadas, a qualidade ambiental, a distância a importantes centros de trabalho, e assim por diante. (D'ACCI, 2019). Estas duas categorias de características desempenham papéis fundamentais na determinação dos preços das habitações, refletindo a complexidade subjacente na avaliação do mercado imobiliário.

Através de pesquisa bibliográfica, é possível encontrar vários trabalhos que evidenciam o impacto de características intrínsecas nos preços dos imóveis, em seu valor percebido e em sua solvência no mercado. Troy et al. (2008) mostraram que quando a taxa de criminalidade é relativamente baixa, os parques têm um impacto positivo nos valores das propriedades. Cervero & Kang (2011) mostraram que o transporte público, especificamente Bus Rapid Transit (BRT), oferece prêmios de até 10% para residências dentro de 300 metros de paradas de BRT e mais de 25% para varejo e outros usos não residenciais, dentro de um raio de 150 metros de distância. Chernoff & Craig (2022) também confirmam essa evidência ao mostrar que a expansão do Vancouver Rapid Transit aumentou o preço dos imóveis nos bairros para os quais ele foi expandido, assim como nos diversos bairros onde a rede de transporte já estava presente, uma vez que agora ela passaria a ligar mais lugares. Hui et al. (2007) mostraram que, para Hong Kong, o tempo de viagem do apartamento ao distrito central de negócios está negativamente correlacionado com os preços da habitação. Além disso, as pessoas estão dispostas a pagar mais por apartamentos com vista para o mar e melhor qualidade do ar. Inesperadamente, verificou-se que o nível de ruído estava positivamente correlacionado com o preço.

Nestes trabalhos mostrados anteriormente, fica evidente que as características extrínsecas afetam o preço dos imóveis, sendo certo que dessas características, várias delas como Transporte Público, Ciclovias, Parques, Arborização, entre outras estão ligadas ao investimento público, ou seja, ao ato da administração pública em direcionar recursos para construir um corredor de ônibus ou um novo parque por exemplo. Whachter & Gillen (2006) mostraram que dentre diversos investimentos públicos, sendo eles: melhorias em corredores comerciais, gestão de terrenos baldios, estratégias de esverdeamento de bairros, como ruas verdes e “distritos de melhorias comerciais”, intervenções de qualidade de vida, melhorias escolares e acesso ao trânsito na cidade de Filadélfia, um dos maiores impactos no preço dos imóveis foi verificado nas propriedades que se situavam próximas aos parques, com preços cerca de 28% superiores aos de outros imóveis similares.

Barreca et. al (2020) vai além, referindo ao conceito de “*Urban Vibrancy*” que é comumente reconhecido a estar associado à atração, diversidade e acessibilidade de um lugar, de modo que a intensidade dessas características (em termos de serviços que um bairro oferece aos seus moradores) pode ser usada como um *proxy* para medir a *Urban Vibrancy*. Os autores reconhecem, no entanto, que embora a *Urban Vibrancy* seja

amplamente estudada na literatura com diversas finalidades, em sua pesquisa não foram encontrados estudos, que forneçam evidências deste conceito sendo relacionado ao mercado imobiliário, especificamente, estudos que o relacione aos preços dos imóveis. Finalmente, neste estudo, os autores conseguem demonstrar o impacto da presença desses serviços públicos no preço dos imóveis, com a valorização dos preços, concluindo que a partir da análise desta valorização "As políticas municipais poderão ser efetivamente orientadas para o desenvolvimento sustentável das áreas urbanas, promovendo o bem-estar social e econômico integrado, passando da valorização do ambiente físico urbano à melhoria da qualidade de vida dos atuais e futuros habitantes." (Tradução livre).

O estudo anterior mostra que há um campo de estudo em aberto, que consiste em precificar as intervenções extrínsecas, especialmente aquelas relacionadas ao investimento público na infraestrutura urbana, por meio da análise de seus impactos nos preços do mercado imobiliário. Cabe ressaltar que, para proceder com uma avaliação imobiliária consistente, é necessária a escolha de um modelo adequado que possa incluir, no processo de precificação, variáveis intrínsecas e extrínsecas.

Rosen (1974) definiu de forma pioneira um grupo de modelos chamados de "Modelos de precificação hedônica", que têm características específicas, características essas que consistem na decomposição dos valores das casas "em grupos de atributos de utilidade que contribuem para a heterogeneidade observada nos preços. Os preços das casas observados podem então ser considerado como a soma composta de elementos que representam preços estruturais e locacionais implícitos". (Tradução livre). Assim, os preços dos imóveis serão compostos por adições incrementais de múltiplas características, incluindo as intrínsecas e extrínsecas.

Nesse sentido, Malpezzi (2003) demonstra que o método das equações hedônicas é uma maneira de decompor os gastos com habitação em preços mensuráveis e quantidades, possibilitando a previsão e comparação desses preços de diferentes habitações ou de habitações idênticas em diferentes locais. Em sua forma mais simples, uma equação hedônica é uma regressão dos preços (seja de aluguéis ou de valores dos imóveis) com características habitacionais. As variáveis independentes representam as características individuais da habitação, e os coeficientes de regressão podem ser convertidos em estimativas dos preços implícitos dessas características. Desde então, o HPM é um dos métodos mais conhecidos e utilizados para a precificação de imóveis e vem sendo estudado por pelo menos cinco décadas. (HERATH & MAIER, 2010).

No entanto, embora o HPM tenha sido amplamente aplicado na literatura, ele apresenta limitações que precisam ser levadas em consideração no processo de precificação. Selim (2009) aponta que nos casos em que a meta de precificação é ampliada para incluir aspectos como outliers, não linearidade, dependência espacial e outros tipos de dependência entre observações, descontinuidade e imprecisão, os métodos hedônicos podem produzir resultados ruins. Além disso, Shiller (2008) argumentou que a abordagem hedônica pode gerar efeitos de regressão espúrios. Como a relação precisa entre informações hedônicas e preços de venda é desconhecida e provavelmente complexa, há um grande espaço para modelagem conduzida pelo acaso.

Na literatura, existem muitas alternativas possíveis ao HPM quando se trata de modelar preços imobiliários. Liu et al. (2006) aplicou um modelo de predição de preços de rede neural *fuzzy*, que pode aprender com dados históricos sobre as correlações entre vários fatores que influenciam os preços dos imóveis e os preços reais de venda. O modelo foi então utilizado para estimar o nível de preço mais adequado para um imóvel. Ele encontrou uma forte capacidade de aproximação, mostrando que essa metodologia pode ser adequada para a previsão de preços de imóveis dependendo da qualidade dos dados disponíveis.

Selim (2009) realizou uma comparação entre o desempenho preditivo dos modelos de regressão hedônica e das redes neurais artificiais, revelando que as Redes Neurais Artificiais (ANN) podem emergir como uma alternativa mais precisa e robusta na previsão dos preços das casas na Turquia. Por sua vez, McCluskey et al. (2013) evidenciaram que a ANN apresenta um desempenho notável em termos de capacidade preditiva e, consequentemente, precisão na avaliação, superando a abordagem tradicional de análise de regressão múltipla (ARM) no contexto da valoração imobiliária.

Chen & Guestrin. (2016) afirmam que métodos baseados em modelos de árvore, como *Random Forest* e *XGBoost*, geralmente são o estado da arte em muitos domínios, especialmente em conjuntos de dados de estilo tabular, onde cada recurso individual é significativo por si só e não possui dependências que são principalmente temporais ou espaciais, como séries de preços de ações ou imagens. Eles definem o XGBoost como um sistema de aprendizado de máquina escalável.

Usando dados da competição Kaggle, os autores mostram que 17 das 29 soluções vencedoras usaram o XGBoost em 2015. Exemplos de problemas nessas soluções

vencedoras incluem: previsão de vendas na loja; classificação de eventos físicos de alta energia; classificação de texto da web; previsão do comportamento do cliente; detector de movimento; previsão de taxa de cliques de anúncios; classificação de malware; categorização de produtos; predição de risco; previsão da taxa de abandono de cursos on-line. Portanto, considerando a natureza linear das variáveis imobiliárias, pode-se supor que os modelos baseados em árvore alcançariam desempenho de ponta quando se trata de avaliação imobiliária.

No entanto, a aplicação de modelos baseados em árvores ainda não foi totalmente desenvolvida para a precificação de imóveis. Dimopoulos & Bakas (2019) realizaram uma revisão bibliométrica dos artigos mais relevantes sobre precificação imobiliária e, entre 486 trabalhos de pesquisa encontrados, nenhum dos artigos usou o modelo de *Random Forest*. Mesmo não sendo uma busca exaustiva, ela reafirma a possibilidade de uma exploração mais aprofundada.

Mais recentemente, Zhao et al. (2019) utilizou *deep learning* com um modelo *XGBoost* para avaliação imobiliária, analisando não apenas registros históricos de venda, mas também conteúdo visual, com fotos de casas online. O experimento mostrou uma melhoria significativa no desempenho da precisão da previsão do preço da casa. O artigo escrito por Zhao et al. (2019) e outros trabalhos publicados recentemente apenas acrescentam a possibilidade de explorar os campos da modelagem de aprendizado de máquina para avaliação imobiliária.

Como os modelos baseados em árvores são considerados modelos flexíveis, eles são altamente precisos e capazes de funcionar bem em diferentes cenários. No entanto, eles permanecem em muitas análises como uma caixa-preta, sem explicação para os resultados finais. Nesse sentido, uma análise hedônica não pode ser realizada, pois o preço final previsto não é representado como uma soma de incrementos positivos ou negativos de cada uma das características e o próprio modelo contém características completamente não lineares.

Os modelos de melhor desempenho em muitos domínios são complicados modelos de caixa-preta cujas previsões geralmente parecem difíceis de explicar (KOH & LIANG, 2017). Em situações de alto risco, é importante que um modelo tenha transparência e responsabilidade. Quando a explicabilidade não é alcançada, Rudin (2019) aponta consequências graves em muitas áreas importantes, por exemplo, negação de liberdade

condicional, decisões de fiança ruins que levam à libertação de criminosos perigosos, entre outros casos em medicina, finanças e confiabilidade energética.

Lundberg & Lee. (2017) apontam que há uma rica história de métodos de interpretação para modelos baseados em árvores que resumem o impacto dos recursos de entrada no modelo como um todo, mas muito menos atenção tem sido dada às explicações locais que respondem pelo impacto das variáveis de entrada no modelo e em previsões individuais (ou seja, para uma única amostra). A solução é um método desenvolvido pelos autores, denominado SHApley Additive exPlanations (SHAP), que propõe uma abordagem baseada em propriedades da teoria dos jogos para permitir uma explicabilidade ótima, dependente de várias propriedades naturais da teoria dos jogos cooperativos e representativa dos efeitos de interação.

O aspecto mais interessante do SHAP trata-se da sua capacidade de explicar, para cada amostra, o impacto de cada uma das variáveis de entrada na variável resultante do modelo (variável de saída). Nesse sentido, é possível prever o preço dos imóveis e, caso seja necessário ou desejado, o resultado final da modelagem pode ser analisado em função dos incrementos positivos e negativos gerados por cada uma das variáveis de entrada. Isso é útil tanto para a análise hedônica quanto para uma compreensão profunda da relação do preço com as características extrínsecas e intrínsecas da propriedade.

SHAP foi criado e aplicado em muitos campos por Lundberg & Lee. (2017). Alguns exemplos incluem i) identificação de fatores de risco de mortalidade não lineares de alta magnitude, mas de baixa frequência na população geral dos EUA, ii) destacando subgrupos populacionais distintos com características de risco compartilhadas, iii) identificação de efeitos de interação não linear entre fatores de risco para doença renal crônica, e iv) monitorar um modelo de aprendizado de máquina implantado em um hospital, identificando quais recursos estão degradando o desempenho do modelo ao longo do tempo.

Além disso, Mokhtari et al. (2019) também aplicou o SHAP especificamente para explicar como os modelos não lineares se comportam em dados de séries temporais financeiras e mostrou que os valores do SHAP podem ser úteis para melhorar a precisão da previsão e alavancar o agrupamento de importância de recursos.

Por sua vez, no campo da precificação imobiliária, trabalhos recentes aplicaram o método SHAP para avaliar o desempenho dos modelos de aprendizado de máquina. Chen et. al

(2020) mostrou que ao combinar o modelo de aprendizado de máquina e o SHAP, foram calculadas e mapeadas as contribuições das características ambientais derivadas de dados de visualização de ruas e dados de sensoriamento remoto na cidade de Xangai. Os resultados experimentais revelam que todas as características ambientais urbanas contribuem com 16% dos preços das habitações em Xangai. As relações entre os preços das habitações e duas características verdes (índice de vista verde a partir de dados de visualização de ruas e taxa de cobertura verde urbana a partir de sensoriamento remoto) são ambas não lineares.

Recentemente um dos primeiros trabalhos combinando o uso do modelo XGBoost e o método SHAP na precificação e avaliação das características que afetam os preços dos imóveis foi publicado, por Dou et al. (2023) que, utilizando dados detalhados de transações imobiliárias em Xangai, aplicaram os modelos XGBoost e SHAP para investigar os efeitos não lineares entre bairros e preços de habitações. Fatores relacionados aos bairros, incluindo densidade populacional, amenidades de serviços públicos, amenidades de serviços privados e visualização de ruas, juntos representam 45,48% do efeito geral na previsão de preços de habitações.

Neste trabalho, a precificação de imóveis será realizada na cidade de São Paulo, Brasil, utilizando os dados públicos disponíveis, bem como dados coletados em um site de anúncios imobiliários. O modelo XGBoost será aplicado nos dados e comparado com os resultados de um modelo hedônico, com o método SHAP para entender os impactos que cada variável tem no resultado final, sendo avaliadas principalmente as variáveis relacionadas ao investimento público em infraestrutura urbana. Toda a pesquisa feita por este autor não encontrou qualquer evidência de trabalho semelhante feito, especificamente no Brasil usando esta abordagem.

3 Dados e Metodologia

Cabe aqui ressaltar uma importante peculiaridade do Brasil quando falamos de transações imobiliárias. O Brasil utiliza um sistema de registro de transações por meio Cartorial, onde as informações são pertencentes ao poder público as quais não são disponibilizadas a ninguém (TIERNO et al., 2007). Sendo assim, diferentemente de vários países desenvolvidos, que não possuem esta limitação de acesso às informações por terem privatizado completamente os registros imobiliários, no Brasil não há uma base de dados oficial que contenha os preços reais negociados pelos imóveis.

Diante desse detalhe, foi necessário utilizar uma forma engenhosa de obter essas informações, obtendo-as por meio de anúncios imobiliários online. Além disso, outros dados sobre o "ambiente construído" das cidades (qualidade do pavimento, calçadas, arborização) são escassos na maior parte do país, sendo a cidade de São Paulo uma das únicas que possuem uma vasta base de dados com estas informações.

Pelas razões discutidas acima, 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. Utilizando os dados de *PMSP, 2007 – Índice de Desenvolvimento Humano por Distrito (IDH-D)*, entre os 50 primeiros colocados no ranking do município, foram escolhidos seis Distritos, totalizando 5% da

população do município (PMSP, 2010). Os distritos escolhidos e seus dados demográficos são apresentados na Tabela 3.1.

Tabela 3.1

Bairros escolhidos da Cidade, escopo do estudo.

Distrito	IDH-D (2007)	Posição no Ranking (IDH-D)	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 deles, a escolha das variáveis e fontes de dados também foi realizada. Aqui, partindo dos diversos trabalhos já abordados no Capítulo 2, foram selecionadas variáveis comumente associadas ao preço e à qualidade de um imóvel, tanto intrínsecas quanto extrínsecas. A Tabela 3.2 mostra um breve resumo das principais fontes de dados onde essas variáveis foram obtidas.

Tabela 3.2

Bases de dados utilizadas neste estudo.

Bases de dados com variáveis intrínsecas	Bases de dados com variáveis extrínsecas
◊ Viva Real (Webscrape) ◊ IPTU ◊ Cadastro 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 a entrada para os modelos de precificação adotados.

3.1. Bases de Dados com Variáveis Intrínsecas

Características intrínsecas são aquelas que estão contidas ou diretamente associadas a um imóvel, bons exemplos são o número de quartos, número de banheiros, área construída, vagas de garagem, dentre outras. Estas características são comumente ligadas ao preço do imóvel pois afetam diretamente sua usabilidade.

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 Viva Real[2]. Foi desenvolvido um código do tipo *webscrape* em linguagem Python para obter todos os anúncios das 100 primeiras páginas do site, repetindo o processo para cada um dos seis bairros escolhidos, este processo foi realizado com dados extraídos em 17 de janeiro de 2021. Concluída a execução, foram extraídos 5.479 anúncios, formando o que será chamado a partir daqui de "Base de dados *Webscrape*". A Tabela 3.3 mostra as variáveis intrínsecas que foram extraídas de cada um dos anúncios, contidas nessa base de dados.

Tabela 3.3

Variáveis contidas na Base de dados *Webscrape*.

Variável	Descrição
Área Útil	Área privativa do imóvel (metros quadrados).
Garagens	Número de vagas de garagem do imóvel.
Quartos	Número de quartos.
Banheiros	Número de banheiros.
Preço	Valor total do imóvel anunciado.
Bairro	Bairro onde o imóvel está localizado.
Endereço Completo	Endereço do imóvel com número.

A segunda fonte é baseada nos dados sobre cobrança do Imposto de Propriedades Urbanas (IPTU), disponibilizadas pela prefeitura na plataforma *GeoSampa*[3]. São dados de acesso público, atualizados anualmente, que reúnem diversas informações das propriedades imobiliárias da cidade, assim como de seus proprietários. Esta informação obtida será chamada de "Base de dados IPTU". A Tabela 3.4 mostra as variáveis contidas nessa base de dados, relativas ao ano de 2020:

Tabela 3.4

Variáveis contidas na Base de dados IPTU.

Variável	Descrição
Ano de Construção	Ano de Construção do imóvel.
Número de pavimentos	Número de pavimentos do imóvel.
Número de Esquinas	Número de esquinas do terreno onde o imóvel está contido.
Fração Ideal	Percentual da área que cada unidade possui do edifício. Para casas o valor é sempre 1.
Área do Terreno	Área do terreno (milhares de metros quadrados).

[2] O Viva Real é o maior portal online de anúncios imobiliários do Brasil, contendo anúncios de imóveis de diversas cidades do país, inclusive da cidade de São Paulo. Disponível em <https://www.vivareal.com.br/>.
[3] O GeoSampa é o portal cartográfico oficial da Cidade de São Paulo e reflete a infraestrutura municipal em dados geográficos. A plataforma traz mais de 240 tipos de informações, como fotos aéreas, dados de equipamentos públicos, rede de transporte, sistema viário, dados ambientais, zoneamento, sítios históricos, entre outros. É a maior coleção de dados geoespaciais da cidade. Disponível em <http://geosampa.prefeitura.sp.gov.br/PaginasPublicas/_SBC.aspx>.

Área Construída	Área construída do imóvel, considerando toda a área habitável que possui benfeitorias (milhares de metros quadrados).
Área Útil	Área correspondente à área construída, menos a parte ocupada pela alvenaria e outras intervenções. Área "desimpedida" do imóvel (milhares de metros quadrados).
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 estado de conservação. Quanto mais próximo de 0, mais depreciado.
Testada	Comprimento da frente do terreno (metros).
Endereço do Imóvel	Endereço do imóvel com número.
CodLog	Código único relativo ao nome da rua do imóvel.
CEP	Código postal do imóvel.
Número de Contribuinte	Número de registro único do imóvel na prefeitura.

A terceira e última fonte também foi obtida na plataforma GeoSampa e possui a geolocalização do imóvel por meio de seu respectivo polígono geolocalizado. Este dado está disponível para cada um dos imóveis da Cidade. Nesta base de dados, que será chamada de "Base de dados Geo/Polígono", cada polígono está associado à um único Número de Contribuinte e não possui outros metadados (Tabela 3.5). Esta será uma fonte importante para relacionar todas as características intrínsecas com seu respectivo posicionamento geográfico na cidade.

Tabela 3.5

Variáveis contidas na Base de dados Geo/Polígono.

Variável	Descrição
Número de Contribuinte do Terreno	Número de registro único para cada terreno na prefeitura.
Polígono	Polígono geolocalizado contendo o terreno do imóvel com suas dimensões.

3.2. Bases de Dados com Variáveis Extrínsecas

No campo das características extrínsecas, situam-se as peculiaridades da vizinhança à qual cada imóvel pertence. Dados como ciclovias, pontos de ônibus, metrô, arborização, entre outros, podem ser caracterizados como extrínsecos ao imóvel analisado. Diversos trabalhos na academia têm relacionado este tipo de característica como impactante na determinação do preço dos imóveis afetados por elas. (D'ACCI, 2019).

Na cidade de São Paulo, estas características estão disponíveis para serem acessadas por duas grandes fontes públicas. A primeira delas é disponibilizada pela "Fundação Sistema

Estadual de Análise de Dados Estatísticos" (SEADE[4]), que reúne dados de diversas esferas da administração pública.

A partir dessa fonte primária, objetivou-se obter informações referentes aos equipamentos de uso comum disponibilizados à população, mais especificamente sua localização na cidade, com latitude e longitude, compilados em uma base de dados que será chamada de "Base de dados SEADE". Esses dados foram obtidos para as seguintes classes listadas na Tabela 3.6.

Tabela 3.6

Classes usadas 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, geridos pelo governo.
Escolas Públicas (Federais, estaduais, municipais)	Latitude e longitude das escolas públicas de ensino médio da cidade, divididas por tipo de administração.
Escolas privadas e outras escolas.	Latitude e longitude das escolas de ensino médio privadas.
Universidade 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 Médicos	Latitude e longitude dos consultórios 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 de faculdades técnicas administradas pelo governo.
Poupatempo	Latitude e longitude das repartições públicas que emitem documentos e oferecem 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.
Pontos de 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 e os dados obtidos a partir desta fonte formaram a "Base de dados *GeoSampa*". Aqui, também, o foco foi em equipamentos de uso comum que não estavam disponíveis

[4] SEADE (Fundação Sistema Estadual de Análise de Dados Estatísticos) vinculada ao Governo do Estado de São Paulo, é referência nacional na produção e divulgação de análises e estatísticas socioeconômicas e demográficas. Para isso, realiza pesquisas diretas e levantamentos de informações produzidas por outras fontes, compondo um amplo acervo, disponibilizado gratuitamente, que permite a caracterização de diferentes aspectos da realidade socioeconômica do estado de São Paulo, suas regiões e cidades e sua evolução histórica. Disponível em <https://www.seade.gov.br/>.

nas bases de dados da SEADE. Os dados de localização foram obtidos para as categorias mostradas na Tabela 3.7.

Tabela 3.7

Classes usadas na Base de dados *GeoSampa*.

Variable	Description
Bicicletários Públicos	Latitude e longitude of the public bike racks located in the City.
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 ciclofaixas 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 sinalização horizontal.
Bus Rapid Transit (BRT)	Figura geométrica geolocalizada (linha) contendo a posição da infraestrutura de BRT da cidade.
Faixas de ônibus arteriais	Figura geométrica geolocalizada (linha) contendo a posição das faixas de ônibus arteriais da cidade, aquelas que cortam a cidade ao longo de grandes eixos e têm grande fluxo de passageiros.
Faixas de ônibus coletoras	Figura geométrica geolocalizada (linha) contendo a posição das faixas de ônibus coletoras da cidade, aquelas que ligam regiões residenciais e comerciais.
Faixas de ônibus locais	Figura geométrica geolocalizada (linha) contendo a posição das faixas de ônibus locais da cidade, aquelas que geralmente ligam regiões distintas dentro de um mesmo bairro ou de bairros próximos.
Árvores	Latitude e longitude das árvores na área urbana.
Imóveis com Acessibilidade reconhecida pela Prefeitura	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.
Atendimentos para reparo em malha viária (pavimento)	Latitude e longitude das reparações de pavimento das ruas que foram constatadas e efetuadas (concluídas) pelo município.

Conforme mencionado anteriormente, todos esses dados de características extrínsecas são geolocalizados, e serão considerados pela sua proximidade a cada um dos imóveis estudados. 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 Obtidas

Nota-se que todas as bases de dados citadas até aqui são desconectadas e não possuem formas fáceis ou naturais de serem relacionadas. A partir da análise das Tabelas 3.3 a 3.7, apresentadas nos itens 3.1 e 3.2, observa-se a existência de alguns pontos de convergência que podem permitir a junção das bases de dados. No entanto, esses pontos devem ser avaliados com cuidado.

A ideia principal na lógica da metodologia aqui adotada para a junção das bases de dados consiste em dividir o processo em etapas de um fluxo, na qual cada uma das questões é resolvida, com a adoção de um critério conjunto. Após isso, obtém-se um novo banco de dados com o resultado, passando para a próxima etapa do fluxo de trabalho. A Figura 3.1 ilustra a ordem de cada uma dessas etapas.

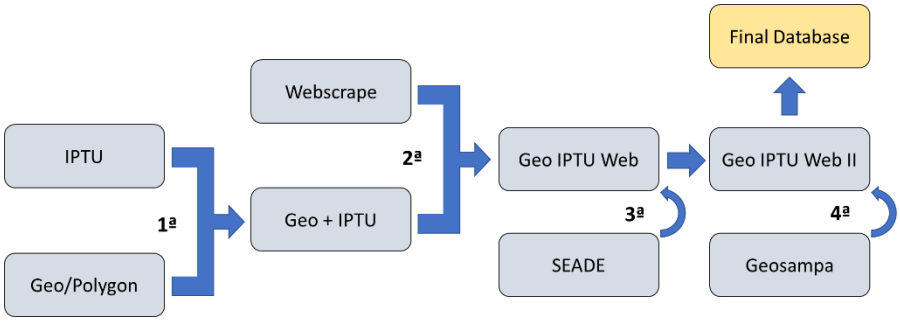

Figura 3.1: Diagrama de construção da Base de Dados completa

A Primeira Etapa da junção inicia-se com a "Base de dados IPTU", no qual foram adicionadas as geolocalizações dos polígonos representativos dos imóveis, através da junção com a "Base de dados Geo/Polígono". O critério de junção aqui adotado foi o uso do "número de contribuinte", campo que está presente nas duas bases de dados e pode ser usado, com restrições, como chave indexadora. Há uma particularidade a destacar: como já referido, existe um número de contribuinte relativo ao terreno e outro relativo ao imóvel. Quando um terreno contém apenas uma propriedade, ambos os números são iguais, mas se o terreno contém mais de uma propriedade, a situação muda. Por exemplo, prédios com vários apartamentos terão um número de contribuinte para cada unidade individual (presente na "Base de dados IPTU"). Já na respectiva base contendo os polígonos ("Base de dados Geo/Polígono"), 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 falamos de terrenos ou residências, juntar as duas bases de dados (IPTU e Geo/Polígono) é simples, pois o número de contribuinte será igual e terá a mesma formação em ambas. Ele será composto pela junção do Código do Setor Administrativo[5] + Código da Quadra + Código do Lote + Código da Unidade. Porém, no caso de apartamentos, o número presente na "Base de dados IPTU" será formado pelo Código do Setor Administrativo + Código da Quadra + Código do Lote + Código da Unidade + Número do Condomínio dentro da Quadra. Para um mesmo imóvel, o número do terreno na base Geo/Polygon, será formado apenas pela agregação do Código do Setor + Código da Quadra + Código do Lote + Número do Condomínio dentro da Quadra. Ou seja, no segundo caso não há Código de Unidade formando o número.

Consequentemente, para unir as duas bases de dados, o número do contribuinte é modificado na "Base de dados IPTU", removendo o código da unidade toda vez que for identificado que o registro é um apartamento. Em resumo, a mudança feita dentro da "Base de dados IPTU", nas chaves, ocorre da seguinte forma:

◊ ***Se o registro*** é ***um apartamento, então***:

$$Nova\_Chave = CodSetor + CodQuadra + CodLote + CodCondomínio$$

◊ ***Caso contrário***: # casa, terreno

$$Chave = CodSetor + CodQuadra + CodLote + CodUnidade$$

Por fim, por meio de um *left join* utilizando esta 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. Sendo nesse caso, evidente que no banco de dados resultante, todos os apartamentos de um mesmo edifício terão o mesmo polígono associado a cada um. Esta nova base de dados criada será chamada de "Base de dados Geo + IPTU".

Após esta Primeira Etapa, o próximo desafio é agregar as informações da "Base de dados *Webscrape*". Inicialmente, é importante ter em mente que essas informações, por serem coletadas diretamente dos anúncios online, apresentam diversos problemas, pois dependem das informações inseridas pelo usuário final no momento da criação do

[5] A cidade de São Paulo é dividida em diversos Setores administrativos. Por exemplo, pinheiros é o setor de número 15.

anúncio. Ou seja, muitas vezes os dados apresentam problemas de preenchimento, erros de digitação ou até mesmo informações incorretas. Portanto, é fundamental que essas informações sejam tratadas e uma das primeiras ações nesse sentido é retirar do banco de dados todos os registros que possuem endereço incompleto, ou seja, quando falta o nome da rua ou o número do imóvel.

A Segunda Etapa da junção começa após este processo de limpeza. Partindo do endereço completo disponível em cada anúncio e utilizando a biblioteca *googlemaps*[6] 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 construção dos Códigos de Endereçamento Postais (CEPs) apresenta problemas e limitações que podem levar a inconsistências. Por exemplo, existem ruas segmentadas em várias seções, tendo cada uma dessas seções (da mesma rua) um CEP diferente. (Aranha, 1997). Este problema estava presente na biblioteca usada para obter códigos postais. Para essas ruas (com múltiplos CEPs) a biblioteca havia atribuído um "CEP padrão", desconsiderando o fato de que demais trechos dela possuíam outro CEP.

Este fato poderia 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, ou seja, nestas ruas com múltiplos CEP's os imóveis possuíam diferentes CEPs de acordo com o segmento da rua em que se encontravam e não um "CEP padrão" para toda a rua, como identificado na biblioteca. Uma nova solução engenhosa era necessária para superar esse problema.

A partir da "Base de dados Geo + IPTU", foi criado um dicionário contendo todos os CEPs e seus respectivos "CodLog" (conforme Tabela 3.4) do município. Embora uma mesma rua possa ter dois ou mais CEPs, ela nunca terá mais de um CodLog. O próximo passo foi associar, na "Base de dados *Webscrape*", as informações desse dicionário. Ao inserir, para cada CEP o seu respectivo CodLog, foi possível criar uma chave comum

[6] O Python Client for Google Maps Services é uma biblioteca Python Client para uso com as APIs do Google Maps, fornecendo acesso a diferentes ferramentas como Directions, Distance Matrix, Elevation, Geocoding, Geolocation, Time Zone, Roads, Places, Maps Statics, entre outros. Está disponível em https://github.com/googlemaps/google-maps-services-python.

para ambas as bases de dados, composta pelo ***CodLog + Número do imóvel***. A partir desta chave, foi realizado o *inner join* entre os dois bancos de dados.

Há mais uma questão aqui que deve ser destacada: Imagine dois apartamentos no mesmo prédio, uma grande cobertura e outro um pequeno Studio. Ambas as propriedades terão a mesma chave descrita acima, pois estão localizadas no mesmo endereço (*CodLog + número do imóvel*). Se o anúncio da base de dados Webscrape fosse para o Studio, como seria possível associar as informações desse anúncio ao cadastro correto contido na "Base de dados Geo + IPTU" já que a chave não é única? O anúncio do Studio poderia ficar associado ao cadastro da Cobertura. A resposta para essa pergunta está em adotar uma condição de contorno utilizando a área do imóvel, assim será possível escolher, dentre todos os registros da "Base de dados Geo + IPTU" com esta mesma chave, aquele que possui a área que mais se aproxima da área do o 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 3.4 – Geo IPTU) e "Área Construída" (conforme Tabela 3.3 – *Webscrape*). Então é mantido apenas um registro por chave (*CodLog + número do imóvel*), que é escolhido como aquele com a menor diferença (absoluta) entre as duas variáveis.

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, são calculados todos os centróides dos respectivos polígonos com suas latitudes e longitudes. Eles podem ser vistos como os pontos amarelos na Figura 3.2. Em seguida, as distâncias entre esses pontos e os correspondentes à latitude e longitude obtidos do endereço completo (dos anúncios da "Base de dados *Webscrape"*) usando a biblioteca *googlemaps* – os pontos vermelhos na Figura 3.2 – são calculadas. 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.

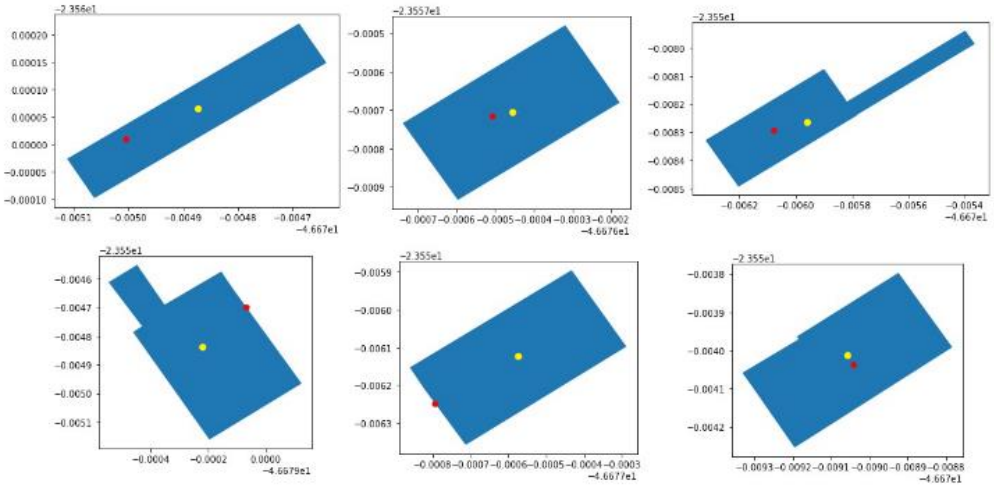

Figura 3.2: Exemplos de Polígonos representando os terrenos de alguns imóveis. Os pontos em amarelo são os centróides das figuras geométricas e os em vermelho são a geolocalização (lat/long) obtida com a biblioteca *googlemaps* a partir do endereço completo.

Então, o banco de dados contendo todas as características intrínsecas já mencionadas está pronto. Esse novo banco de dados será chamado de "Base de dados de 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 SEADE". Essa agregação é feita para cada uma das classes da "Base de dados SEADE" (conforme Tabela 3.6). 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 centróide 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 *haversine*[7]. Nesse caso, o objetivo de contar a quantidade de pontos de cada classe que estão nas proximidades de um

[7] A fórmula de Haversine determina a distância do superficial circular entre dois pontos em uma esfera, dadas suas longitudes e latitudes. Definida como:

$$d_{hav} = 2\,r \arcsin\left(\sqrt{\sin^2\left(\frac{lat_2 - lat_1}{2}\right) + \cos(lat_1)\cos(lat_2)\sin^2\left(\frac{long_2 - long_1}{2}\right)}\right)$$

Na qual "r" é o raio da esfera.

determinado imóvel será um proxy, ou seja, uma forma de mensurar o impacto que a presença de unidades de uma determinada classe possui no preço desse imóvel.

Essa mesma lógica também foi adotada para incluir informações sobre características extrínsecas da "Base de dados *GeoSampa*", conforme Tabela 3.7 (Quarta Etapa), listando quantos itens de suas classes estão nas proximidades dos imóveis. Existem apenas duas particularidades nesta base de dados. A primeira diz respeito ao fato de que determinadas classes, como as relacionadas à Infraestrutura Cicloviária ou à Infraestrutura de Ônibus, 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 um determinado imóvel, ao invés do cálculo de pontos, feito anteriormente. Além disso, uma vez que a geometria da linha está disponível na "Base de dados *GeoSampa*", seu comprimento total foi calculado, e essas informações foram agregadas ao cruzar os dois 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 3.8 contém os valores dessas distâncias.

Tabela 3.8

Distância de Contagem Adotada para cada classe na Base de Dados GeoSampa.

Classe	Distância de Contagem	Justificativa
Árvores	Até 500 m	Alta densidade na cidade.
Bicicletários Ciclovias Ciclofaixas Ciclorrotas Reparos em Pavimento	Até 1000 m	Densidade média na cidade.
BRT Faixas de ônibus Arteriais Faixas de ônibus Coletoras Faixas de ônibus Locais Imóveis com acessibilidade	Até 2000 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 dos preços.

3.4. Variáveis Criadas

Todas as variáveis descritas nas Tabelas 3.3 a 3.7 dos itens anteriores, foram obtidas diretamente de classes ou características contidas nas bases de dados iniciais. Nenhuma delas sofreu conversão de escala ou foi submetida a operações matemáticas. No entanto, para melhorar a forma de retratar e mensurar o impacto dessas variáveis nos imóveis, assim como tornar o processo de modelagem mais efetivo, algumas operações matemáticas são necessárias com os dados obtidos até aqui.

O primeiro exemplo é a **variável objeto** 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 adotada aqui 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 ampla, ao atingir mais regiões da cidade.

Por outro lado, modelos de Machine Learning podem ter seu desempenho consideravelmente melhorado quando as variáveis (características) passam por operação de escala 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 variáveis que possuem ordem de grandeza cem vezes o valor médio global foram escalonadas pela média 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. Ao calcular este racional entre estas duas variáveis existe uma forma de quantificar o impacto de um imóvel possuir quartos com banheiros equivalentes, penalizando os imóveis que têm muitos quartos, mas não têm banheiros suficientes. (HEIDARI et al., 2021). A Tabela 3.9 mostra um resumo das variáveis criadas no banco de dados.

Tabela 3.9

Variáveis criadas, baseadas em variáveis anteriormente obtidas.

Variável	Cálculo (equação)	Descrição
Preço/m²	$Preço\ /\ Área\ útil$	Variável objetivo do modelo.
Número ponderado de Ciclovias, Ciclofaixas, Ciclorrotas, BRT, Faixas de ônibus Arteriais, Coletoras e Locais.	$N°_i * Comp\_tot_i\ /\ Média\_global(comp)$ Por exemplo: $N°_{ciclovias\ na\ vizinhança\ do\ imóvel} * Comprimento\_tot_{\ dessas\ ciclovias}\ /\ média\_global(Comp._{Todas\ as\ Civlovias})$	Número ponderado de cada infraestrutura por seu respectivo comprimento, relativo à média global desse comprimento no banco de dados.
Número de Árvores *Scaled*, Consultórios Médicos *Scaled*, Clínicas Médicas *Scaled*, Reparos no Pavimento *Scaled.*	$Variável_i\ /\ média(Variável)$	Variáveis escalonadas por sua respectiva média global no banco de dados.
Quartos por Banheiros	$N°_{quartos}\ /\ N°_{banheiros}$	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 explicada a base de dados resultante e quais as variáveis consideradas no processo de modelagem.

3.5. Base de Dados Resultante

Após realizar todas as etapas descritas na Figura 3.1, 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 do Modelo Hedónico de Preços e do Modelo Tree-Based. Contém 42 variáveis que representam características do modelo (entradas) e 1 variável objetivo (saída). Um resumo dessas variáveis é apresentado na Tabela 3.10.

Tabela 3.10

Variáveis no Banco de Dados Resultante, para serem utilizadas no Modelo.

Variável	Tipo
Preço/m²	Variável Objetivo
Área Construída	Variável de Modelo
Quartos por Banheiros	Variável de Modelo
Número de Garagens	Variável de Modelo
Número de Esquinas	Variável de Modelo
Fração Ideal	Variável de Modelo
Área do Terreno	Variável de Modelo
Área útil	Variável de Modelo
Ano de Construção	Variável de Modelo
Fator de obsolescência	Variável de Modelo
Número de andares	Variável de Modelo
Testada	Variável de Modelo
Número de UBS	Variável de Modelo
Número de CREAS	Variável de Modelo
Número de CRAS	Variável de Modelo
Número de Escolas Privadas	Variável de Modelo
Número de Escolas Públicas Estaduais	Variável de Modelo
Número de Escolas Públicas Municipais	Variável de Modelo
Número de Escolas Públicas Federais	Variável de Modelo
Número de Escolas (outras)	Variável de Modelo
Número de FATECs	Variável de Modelo
Número de Universidades Particulares	Variável de Modelo
Número de Universidades Públicas	Variável de Modelo
Número de Museus	Variável de Modelo
Número de unidades do *Poupatempo*	Variável de Modelo
Número de Centros Populares	Variável de Modelo
Número de Hospitais Públicos	Variável de Modelo
Número de Hospitais Privados	Variável de Modelo
Número Escalado de Consultórios Médicos	Variável de Modelo
Número Escalado de Clínicas Médicas	Variável de Modelo
Número Escalado de Reparos no Pavimento	Variável de Modelo
Número Escalado de Árvores	Variável de Modelo
Número de imóveis com acessibilidade	Variável de Modelo
Número de Bicicletários Públicos	Variável de Modelo
Número de Estações de metrô	Variável de Modelo
Número de pontos de ônibus	Variável de Modelo
Número ponderado de Ciclovias	Variável de Modelo
Número ponderado de Ciclofaixas	Variável de Modelo
Número ponderado de Ciclorrotas	Variável de Modelo
Número ponderado de Faixas de ônibus Locais	Variável de Modelo
Número ponderado de Faixas de ônibus Coletoras	Variável de Modelo
Número ponderado de Faixas de ônibus Arteriais	Variável de Modelo

A Base de dados resultante contém 4.019 registros, cada um correspondendo a um único Imóvel e reunindo todas as informações descritas na Tabela 3.10, para ele. A Figura 3.3 mostra a distribuição desses imóveis pelos bairros da cidade, discutidos anteriormente na Tabela 3.1.

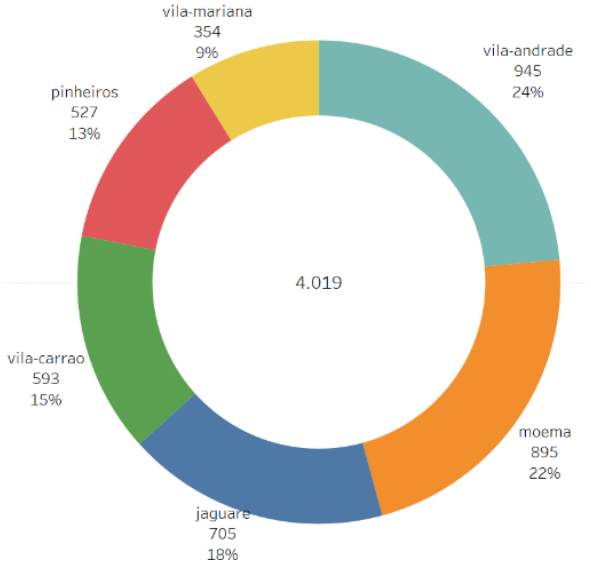

Figura 3.3: Propriedades Imobiliárias no Banco de Dados Resultante de acordo com os bairros a que elas pertencem.

Vale ressaltar que o processo de *webscraping* do site VivaReal reuniu 5.479 registros na base de dados inicial (Base de dados *Webscrape*) 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 junção de cerca de 73%, além disso, todos os registros que possuíam valores nulos em todas as variáveis foram removidos, por serem considerados outliers.

4 Análise Exploratória dos Dados

Nos próximos capítulos, a discussão focará em atender aos objetivos pretendidos, definidos anteriormente, os quais:

i. Treinar e testar um modelo de predição de preços de ativos imobiliários, utilizando o algoritmo XGBoost e comparar seus resultados com um modelo hedônico.
ii. Avaliar quais são as variáveis mais impactantes desse modelo de aprendizado de máquina (XGBoost) por meio do uso do SHAP Tree Explainer, relacionando aquelas que são diretamente ligadas ao investimento público na infraestrutura urbana.
iii. A partir dessa avaliação, encontrar aqueles investimentos que possuem os maiores impactos positivos (valorização) no preço dos imóveis e mensurar – ou estimar – este impacto no Imposto Predial Territorial Urbano (IPTU), sugerindo sua escolha para o administrador público.

Até este momento, o trabalho focou em criar as bases, tanto através da revisão bibliográfica, quanto da análise prática para que estes objetivos fossem alcançados. Todo o esforço de engenharia de dados feito para consolidar as informações será fundamental para a implementação do modelo, seu treino e teste. Esses objetivos definidos serão desenvolvidos na sequência desse trabalho, no entanto, antes de abordar seus aspectos, cabe fazer uma breve introdução sobre o comportamento dos dados aqui coletados, por meio de uma análise exploratória.

4.1. Análise Espacial dos Dados

Analisando a Figura 3.3, é possível perceber a distribuição numérica dos imóveis analisados pelos bairros da cidade. A Figura 4.1 representa visualmente essas mesmas propriedades geolocalizadas no mapa da cidade. As cores do mapa representam a variação do preço por metro quadrado em todos os imóveis, em uma escala verde-vermelha, onde o verde representa os preços mais baixos e o vermelho os mais altos

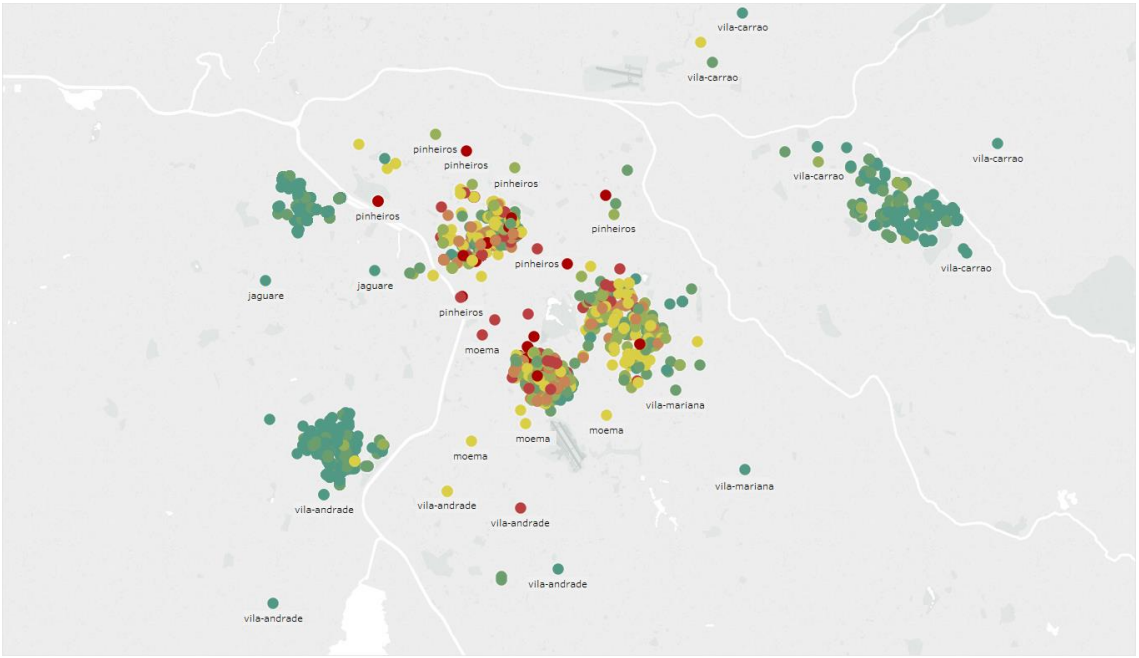

Figura 4.1: Propriedades Imobiliárias do banco de dados, geolocalizadas no mapa da cidade de São Paulo. A cor representa a variação do preço por metro quadrado (vermelho é alto, verde é baixo).

Nota-se, ao observar a Figura 4.1, que os bairros mais centralizados da cidade possuem preços/metro quadrado mais elevados, enquanto os bairros mais distantes possuem preços mais acessíveis. Essa tendência de variação de preços em função da distância do centro da cidade já foi abordada por D'acci (2019), citado anteriormente neste trabalho.

Nesse sentido, o leitor pode se perguntar: Por que a distância entre o centro da cidade e o imóvel não foi considerada como característica (variável de entrada), ou mesmo, se o modelo não poderia ser aplicado em cada bairro? A resposta está baseada no fato de que, como um dos principais objetivos deste trabalho é medir o impacto de variáveis extrínsecas nos preços dos Imóveis, ao adotar uma variável (distância) altamente correlacionada com as mesmas características extrínsecas medidas, o modelo poderia

descartá-las. É o caso das Ciclovias, por exemplo, já que as Ciclovias estão mais presentes nas áreas centrais da cidade e, talvez, esse seja um dos motivos pelos quais essas áreas centrais possuem preços mais elevados, o que deve ser avaliado pelo modelo.

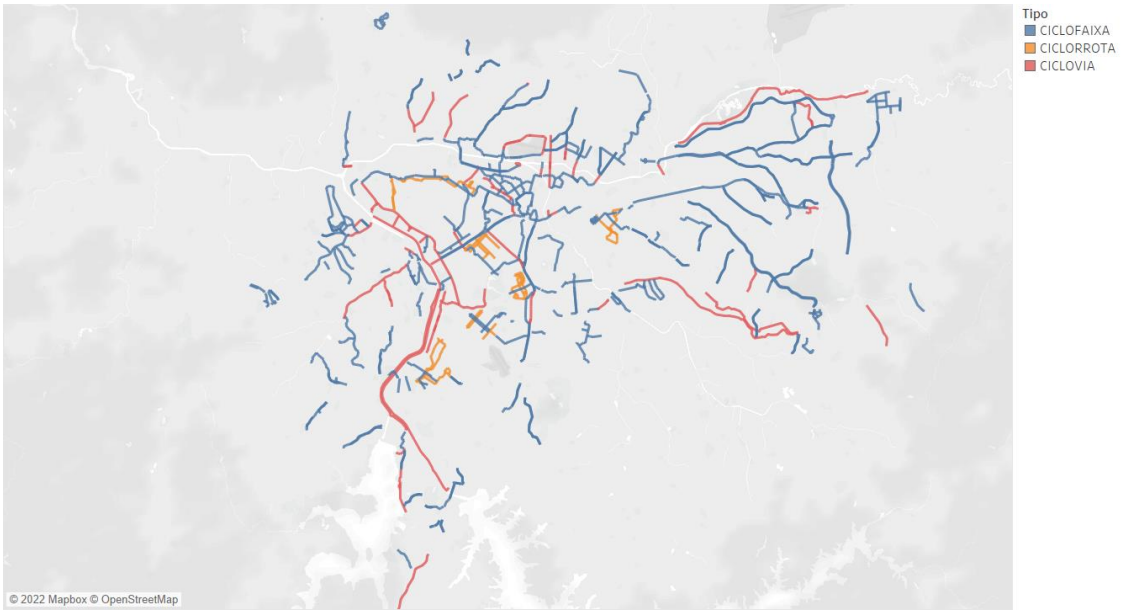

Figura 4.2: Infraestrutura Cicloviária Geolocalizada no mapa da cidade de São Paulo. Nota-se como as Ciclovias (em vermelho) estão mais presentes no eixo central da cidade (próxima das marginais)

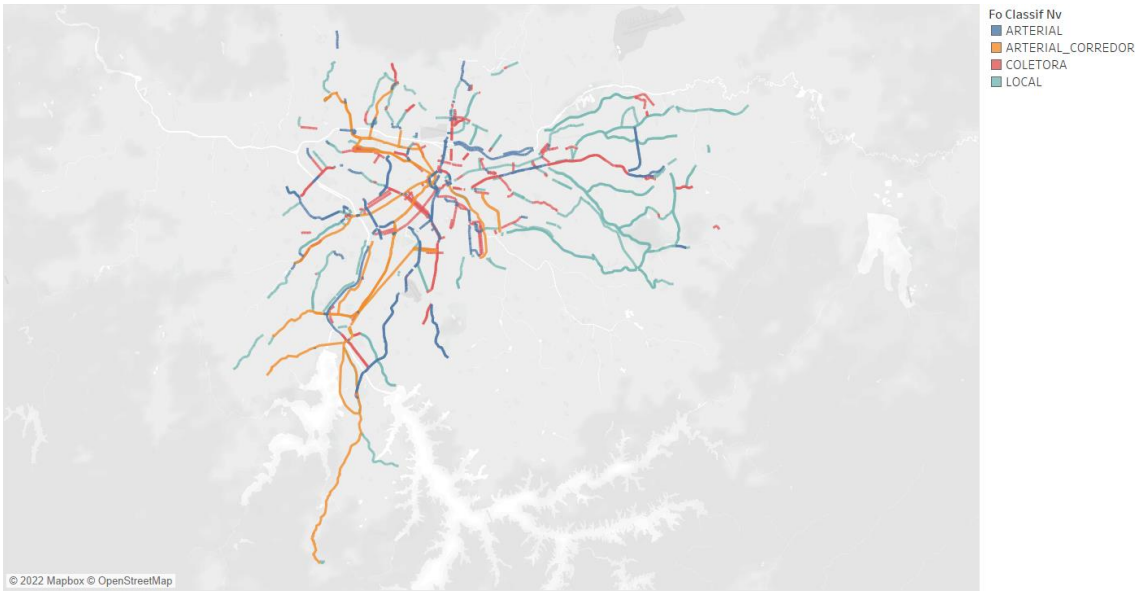

Figura 4.3: Infraestrutura de ônibus Geolocalizada no mapa da cidade de São Paulo. Nesse caso, assim como nas ciclovias, vemos o BRT (Arterial_Corredor em laranja) mais concentrado no eixo central.

Pela análise espacial e inicial dos dados, de algumas classes, seguindo o exemplo da infraestrutura de ônibus e cicloviária (Figuras 4.2 e 4.3), já é possível notar que as áreas centrais da cidade possuem uma maior concentração desses serviços públicos. Da mesma

forma, também possuem os imóveis com maior preço, como visto na Figura 4.1. Se esta relação é apenas correlação ou se há realmente uma relação de causa-efeito, isto será avaliado após o processo de modelagem, com os resultados obtidos, mas neste ponto já é possível notar que este comportamento está presente.

4.2. Análise Gráfica dos Dados

As Figuras 4.4 e 4.5 mostram a distribuição dos valores de cada uma das variáveis de modelagem (entradas), extrínsecas e intrínsecas, respectivamente, em todo o banco de dados, ou seja, para cada uma das propriedades.

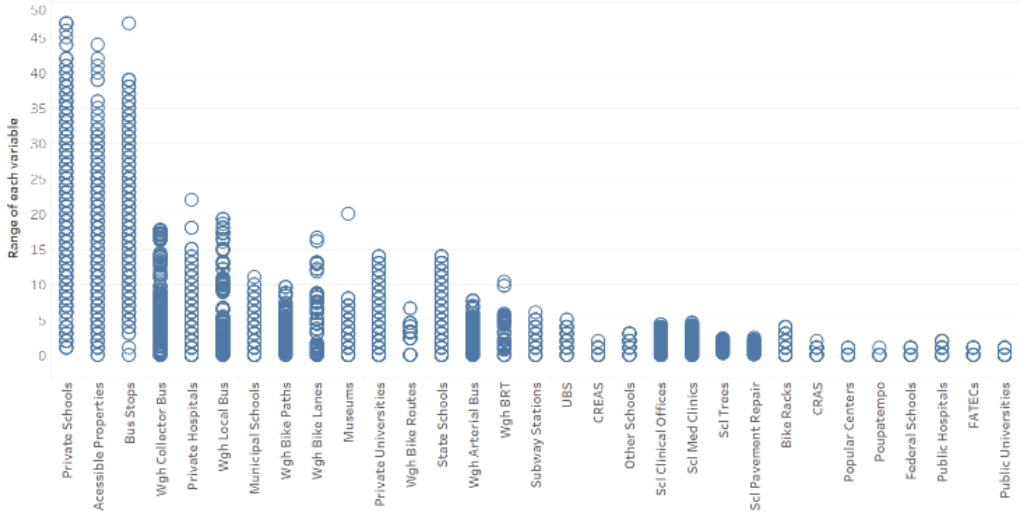

Figura 4.4: Valores para as variáveis Extrínsecas na Base de Dados Resultante.

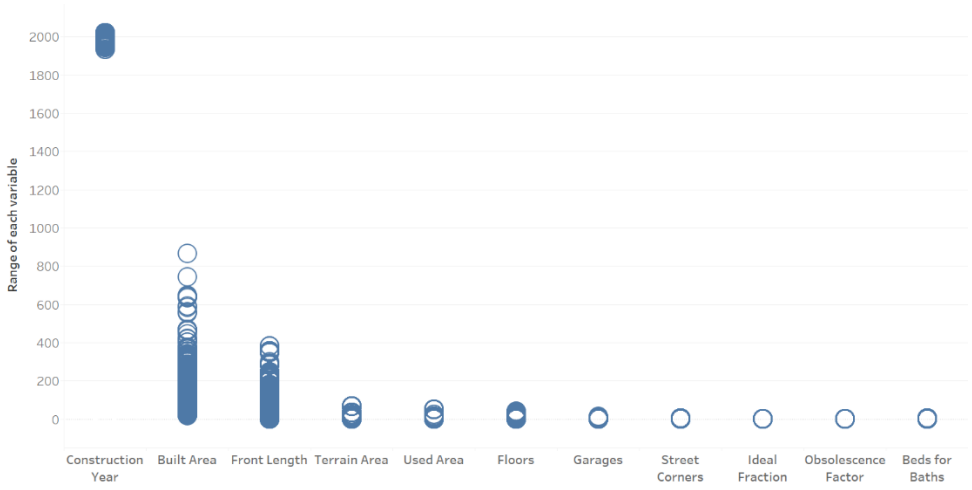

Figura 4.5: Valores para as variáveis Intrínsecas na Base de Dados Resultante.

Uma vez reunidos os valores das variáveis na base de dados, a seguinte análise realizada foi o gráfico representado na Figura 4.6. Esta figura mostra um gráfico de dispersão múltipla envolvendo todas as variáveis do modelo (entradas) e a variável objetiva (saída). Além disso, há uma linha de tendência linear desenhada para identificar o comportamento de cada interação entre as respectivas variáveis.

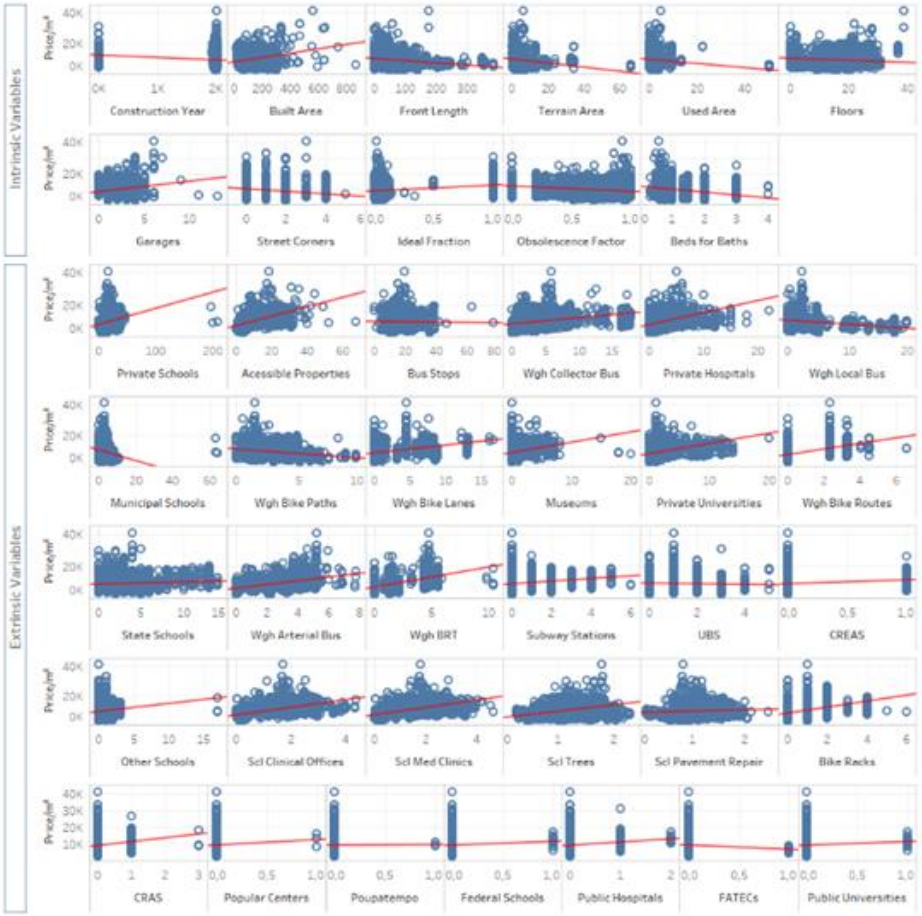

Figura 4.6: Gráfico de dispersão Múltipla para cada Variável (entrada) versus Preço/m².

A partir desse gráfico, percebe-se que nenhuma das variáveis do modelo apresenta claramente um comportamento linearmente relacionado ao preço por metro quadrado. Outra conclusão é que, em relação às variáveis intrínsecas, a maioria delas apresenta uma tendência inversamente proporcional ao Preço/m². Já nas variáveis extrínsecas, a situação é oposta, onde a maioria delas apresenta uma tendência diretamente proporcional à variável objetivo.

Esta última conclusão também pode ser verificada pela matriz de correlação de Pearson, exposta na Figura 4.7. Aqui é aplicada a mesma lógica da análise anterior, calculando a

correlação entre cada uma das variáveis de modelo (entradas) e a variável objetivo (saída). Observe que a maioria das variáveis intrínsecas tem correlação negativa, e o inverso ocorre para as variáveis extrínsecas, que apresentam correlação positiva na maioria dos casos.

Figura 4.7: Matriz de correlação de Pearson para cada Variável (entrada) versus Preço/m². Escala de temperatura (Vermelho – Azul) onde vermelho representa os menores valores e azul os maiores.

Para analisar mais detalhadamente o comportamento das variáveis com as maiores correlações absolutas, foram selecionadas as quatro maiores correlações positivas e as quatro menores correlações negativas, conforme visto na Figura 4.7. As escolhidas com os maiores valores positivos foram:

- Número Escalado de Consultórios Médicos (0,680)
- Número Escalado de Clínicas Médicas (0,659)
- Número ponderado de ciclovias (0,636)
- Número ponderado de BRT (0,624)

Os que apresentaram os menores valores negativos foram:

- Número ponderado de barramento local (-0,324)
- Número de quartos/banheiros (-0,298)
- Número ponderado de ciclovias (-0,278)
- Comprimento frontal do terreno – Testada (-0,243)

Cada uma dessas oito variáveis foi plotada em um gráfico de dispersão, sempre versus o a variável preço por metro quadrado. Os gráficos podem ser vistos nas Figuras 4.8 e 4.9.

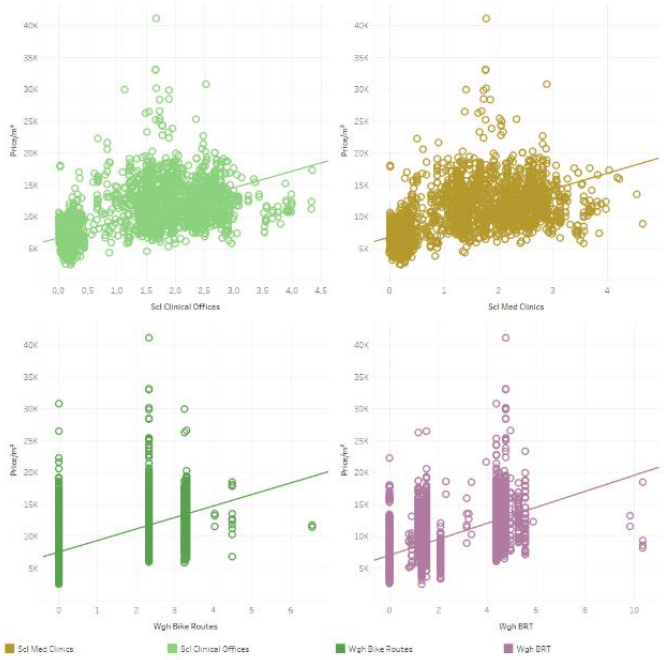

Figura 4.8: Variáveis com maiores correlações positivas versus preço/m².

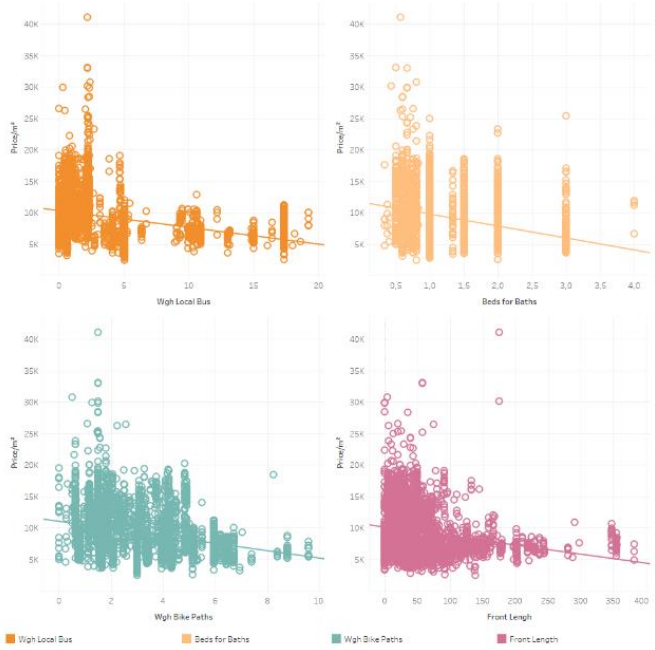

Figura 4.9: Variáveis com menores correlações negativas versus preço/m².

Na Figura 4.8 pode-se observar que as duas primeiras variáveis (Consultórios e Clínicas Médicas), que possuem maior correlação com o preço por metro quadrado, apresentam uma dispersão de pontos muito similar, indicando que apresentam alta correlação entre si. Isso faz sentido, pois estão relacionados ao mesmo tema: serviços de saúde. Nesse caso, há duas interpretações possíveis: ou a presença de imóveis caros aumenta a oferta de serviços de saúde ou a grande oferta desses serviços eleva o preço dos imóveis ao seu redor.

Tabela 4.1

Valores médios para as variáveis Número escalado de consultórios médicos e Preço/m² no Banco de Dados Resultante, para serem utilizadas no Modelo.

Bairro	Med (Número escalado de consultórios médicos)	Med (Price/m²) - decrescente
Pinheiros	1.497	R$ 12,943.26
Moema	2.209	R$ 12,247.51
Vila Mariana	2.310	R$ 11,302.80
Vila Carrão	0.211	R$ 6,846.98
Vila Andrade	0.241	R$ 6,846.88
Jaguaré	0.143	R$ 6,528.76

De fato, pelo que pode ser notado na Tabela 4.1, os bairros que possuem maiores preços por metro quadrado, também possuem imóveis com mais consultórios médicos em sua vizinhança, no entanto não é uma relação completamente linear, sendo o bairro de pinheiros um exemplo de “descontinuidade”.

Adicionalmente, as outras duas variáveis (BRT e Ciclovias) também fazem parte das características extrínsecas dos imóveis. No entanto, essas variáveis estão relacionadas ao investimento público e às políticas públicas, mais especificamente à infraestrutura de transporte público, levando a uma associação inicial entre a disponibilidade desse tipo de transporte e o aumento dos preços dos imóveis, hipótese já levantada anteriormente a partir da análise das Figuras 4.1 a 4.3.

No lado oposto, em relação às variáveis com as menores correlações versus preço por metro quadrado, há duas pertencentes ao grupo de características intrínsecas (quartos/banheiros e comprimento frontal – Testada) e duas que fazem parte das características extrínsecas (faixas de ônibus locais e ciclorrotas).

Algumas hipóteses podem ser levantadas para esse comportamento. No campo das características extrínsecas, tanto as faixas de ônibus locais quanto as ciclorrotas estão

posicionadas em locais periféricos da cidade, com o objetivo de dar acesso a bairros mais afastados que, por sua vez, possuem imóveis mais baratos. Esse fato mostra que podem ser variáveis com comportamento explicativo, apenas correlacional (proxy) diferentemente de uma relação causa-efeito.

Quanto às características intrínsecas, a explicação é mais lógica. Um maior número de leitos/banheiros indica que o imóvel possui poucas suítes, ou que possui muitos quartos sem banheiro, o que poderia justificar um preço por metro quadrado menor. Da mesma forma, um grande comprimento de frente geralmente ocorre em imóveis baldios, que ainda serão construídos/modificados por incorporadoras e, portanto, têm preços mais baixos.

A partir dessa análise inicial dos dados, todas as hipóteses levantadas serão abordadas durante o processo de modelagem, que será realizado por meio de um modelo XGBoost. Nesse sentido, a utilização do SHAP desempenhará um papel fundamental ao oferecer uma interpretação numérica do impacto de cada variável no modelo final, permitindo a identificação das variáveis mais relevantes para a formação do preço.

5 Modelagem e Interpretabilidade

Até esta etapa do trabalho, os objetivos a serem atingidos direcionaram o andamento e construção das análises feitas. Essas análises resultaram em hipóteses que estão intrinsecamente relacionadas ao terceiro objetivo definido no Capítulo 4 – *"iii. A partir dessa avaliação, encontrar aqueles investimentos que possuem os maiores impactos positivos (valorização) no preço dos imóveis e mensurar – ou estimar – este impacto no Imposto Predial Territorial Urbano (IPTU), sugerindo sua escolha para o administrador público."*

Com a análise espacial e gráfica dos dados, já foi possível identificar uma correlação positiva entre a presença de Ciclovias e BRT's e os preços por metro quadrado dos imóveis, em linha com o que pode ser visto na literatura. De outro lado, outras variáveis que demonstraram correlação com o preço (presença de consultórios, por exemplo) trazem questionamentos se este comportamento identificado é apenas uma correlação sem causa-efeito associada. Nesse sentido, o próximo passo será a implementação da modelagem de preços para endereçar essas questões, permitindo que as variáveis consideradas forneçam explicabilidade e interpretabilidade para a análise de seus impactos individuais no resultado final.

5.1.Implementação dos Modelos: XGBoost vs. Hedônico.

O principal modelo a ser implementado para a modelagem do preço, como mencionado anteriormente, será o XGBoost, que se trata de um algoritmo de *machine learning*

baseado em árvores de decisão, paralelamente ao uso da função gradiente para otimizar as diversas árvores de decisão implementadas em cadeia. Conforme discutido no Capítulo 2, modelos baseados em árvores, como *Random Forest* e *XGBoost*, geralmente são o estado da arte em muitos domínios, especialmente em conjuntos de dados de estilo tabular, onde cada recurso individual é significativo por si só e não possui dependências que são principalmente temporais ou espaciais, como seria o caso de séries de preços de ações ou imagens. (CHEN & GUESTRIN, 2016). Este é exatamente o caso em análise aqui, no qual uma base de dados tabular contém todas as características individuais e independentes entre si associadas a cada imóvel. Um exemplo de sua implementação matemática pode ser visualizado na Figura 5.1:

Input: training set $\{(x_i, y_i)\}_{i=1}^{N}$, a differentiable loss function $L(y, F(x))$, a number of weak learners M and a learning rate α.

Algorithm:

1. Initialize model with a constant value:

$$\hat{f}_{(0)}(x) = \arg\min_{\theta} \sum_{i=1}^{N} L(y_i, \theta).$$

2. For m = 1 to M:

   1. Compute the 'gradients' and 'hessians':

$$\hat{g}_m(x_i) = \left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f(x)=\hat{f}_{(m-1)}(x)}.$$

$$\hat{h}_m(x_i) = \left[\frac{\partial^2 L(y_i, f(x_i))}{\partial f(x_i)^2}\right]_{f(x)=\hat{f}_{(m-1)}(x)}.$$

   2. Fit a base learner (or weak learner, e.g. tree) using the training set $\left\{x_i, -\frac{\hat{g}_m(x_i)}{\hat{h}_m(x_i)}\right\}_{i=1}^{N}$ by solving the optimization problem below:

$$\hat{\phi}_m = \arg\min_{\phi \in \Phi} \sum_{i=1}^{N} \frac{1}{2} \hat{h}_m(x_i) \left[-\frac{\hat{g}_m(x_i)}{\hat{h}_m(x_i)} - \phi(x_i)\right]^2.$$

$$\hat{f}_m(x) = \alpha \hat{\phi}_m(x).$$

   3. Update the model:

$$\hat{f}_{(m)}(x) = \hat{f}_{(m-1)}(x) + \hat{f}_m(x).$$

3. Output $\hat{f}(x) = \hat{f}_{(M)}(x) = \sum_{m=0}^{M} \hat{f}_m(x).$

Figura 5.1: Exemplo da implementação matemática do algoritmo XGBoost (CHEN & GUESTRIN, 2016)

Além da implementação do modelo de *machine learning*, haverá um método hedônico escolhido para ser aplicado aos mesmos dados, fornecendo assim um benchmark em relação aos resultados do XGBoost na previsão de preços de imóveis. Dentre os modelos hedônicos, as regressões de diversos tipos são comumente usadas em dados tabulares, por serem fáceis de aplicar e fornecerem resultados simples de interpretar. Sua forma funcional pode ser linear ou logarítmica. No caso logarítmico, pode-se usar um logarítmico comum, de base 10, ou um logarítmico natural, de base e. Também, neste

caso, o modelo pode ser semi-log, onde apenas um lado da equação tem uma transformada logarítmica aplicada, ou log-log em que o logarítmico está aplicado em ambos os lados.

Nas análises de alto nível, o semi-log é uma das formas mais usuais, com muitas vantagens. Ele efetivamente resolve ou reduz o problema de heterocedasticidade, com estimativas de coeficientes proporcionais ao preço, sendo diretamente atribuíveis à respectiva característica. Além disso, pode ser descrito em uma expressão simples (Herath et al. 2010). Assim, optou-se pelo modelo semi-log com base logarítmica natural, em detrimento do modelo de regressão linear. O modelo HPM na forma semi-logarítmica é apresentado na Equação 5.1.

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

Equação 5.1: Variáveis com menores correlações negativas versus preço/m².

Nessa equação, $\log(P_i)$ é o logaritmo natural da variável objetivo (preço por metro quadrado) de cada imóvel. S_{ij} representa os valores das variáveis de modelo da Tabela 3.10 para cada [j] características e [i] propriedades imobiliárias. β_0 é a constante de interceptação e β_j são os coeficientes estimados para as relacionadas as variáveis da tabela anteriormente citada. Para consistência na comparação, XGBoost será também implementado com a variável objetivo (preço/m²) transformada pela função de log natural.

Seguindo para a implementação dos modelos, o primeiro passo foi dividir o Banco de Dados Resultante (da Tabela 3.10) em dois subconjuntos: um para treinamento e outro para teste. Essa divisão foi realizada aleatoriamente, com o auxílio da biblioteca *sklearn* do *Python*, na qual foi estabelecido que 80% dos registros iriam para o banco de dados de treinamento e 20% para o banco de dados de teste. Ambos os modelos que serão implementados utilizarão os mesmos bancos de dados de teste e treinamento, eliminando qualquer influência da variabilidade na seleção dos dados.

No contexto da aplicação do modelo hedônico, iniciou-se com o cálculo do logaritmo natural da variável objetivo. Em seguida, é ajustada uma regressão multivariável aos dados de treinamento, incorporando tanto as variáveis do modelo quanto a variável

objetivo transformada. Para realizar esse ajuste, a biblioteca *sklearn* é utilizada, com foco no módulo *'linear regressor'* do *python*.

Já em relação à implementação do modelo XGBoost, o regressor foi ajustado ao banco de dados de treinamento utilizando a biblioteca *xgboost* do *Python*. Na construção do regressor utilizado para treinar o modelo, a maioria dos parâmetros foi mantida como "padrão". Apenas três parâmetros foram modificados para aprimorar o desempenho do regressor, utilizando-se para tanto, a técnica k-fold de validação cruzada na determinação desses parâmetros.

Após selecionar o valor dos parâmetros do regressor que estão sendo avaliados, esta técnica consiste em dividir o conjunto de dados em k subconjuntos (ou "folds") aproximadamente iguais, na qual cada subconjunto é usado como conjunto de teste uma vez, enquanto os outros (k-1) subconjuntos são usados como conjunto de treinamento. Esse processo é repetido k vezes, com cada subconjunto atuando como o conjunto de teste em uma iteração. Ao final das k iterações, as métricas de avaliação (Score. Coeficiente de determinação e erro quadrático médio) são calculadas. Então, escolhe-se novos valores para os parâmetros do regressor e a técnica é aplicada novamente. Após esse processo, seleciona-se os valores dos parâmetros que resultaram no menor erro quadrático médio, com score e coeficiente de determinação mais próximos entre si.

No caso em questão, após esse processo, os parâmetros avaliados, juntamente com seus respectivos valores escolhidos, foram:

- *Learning Rate = 0.05*
- *Number of Estimators = 200*
- *Maximum Depth = 7*

Com isso, ambos os modelos foram implementados seguindo as especificações mencionadas. Seus resultados, suas saídas e desempenhos serão discutidos na próxima seção.

5.2. Resultados e Métricas dos Modelos.

Após a aplicação do Método de Precificação Hedônica (HPM) aos dados, obteve-se um coeficiente de determinação (R-quadrado) de 0,692 e uma Raiz Quadrada do Erro Médio

(RMSE) de 0,202. No entanto, o XGBoost, quando aplicado ao mesmo conjunto de teste e treinamento, demonstrou, em geral, um desempenho superior, apresentando um R-quadrado de 0,859 e um RMSE de 0,143. Isso indica uma capacidade de explicação de 86% em relação à variação do modelo, sugerindo uma melhor adaptação aos dados.

Além disso, outros indicadores, como a porcentagem de RMSE (%RMSE) e o Erro Médio Absoluto (MAE), também foram utilizados para avaliar o desempenho dos modelos. Todos esses detalhes estatísticos do XGBoost e do HPM estão disponíveis na Tabela 5.1 para análise mais detalhada. Os resultados destacam claramente que o XGBoost superou o HPM em termos de desempenho e capacidade de explicação do modelo, proporcionando uma visão mais precisa e confiável para os dados em questão.

Tabela 5.1

Comparação das Métricas Estatísticas entre o Modelo Hedônico Semi-Log (HPM) e o XGBoost.

Métrica	XGBoost	Semi-Log (HPM)
R-Square	0.859	0.692
RMSE	0.143	0.202
%RMSE	1.578%	2.238%
MAE	0.099	0.150

Considerando o erro mediano no preço por metro quadrado, o XGBoost apresenta o melhor resultado de previsão, com um erro mediano de 6,6%, em comparação com um erro mediano de 11,4% no HPM. A Figura 5.2 exibe uma amostra aleatória de 20 propriedades e seus respectivos valores de preço por metro quadrado, comparando o preço previsto pelos dois modelos com os valores reais (y em vermelho) no eixo y. Esses resultados corroboram com a maior precisão do modelo XGBoost na estimativa do preço por metro quadrado em relação ao HPM.

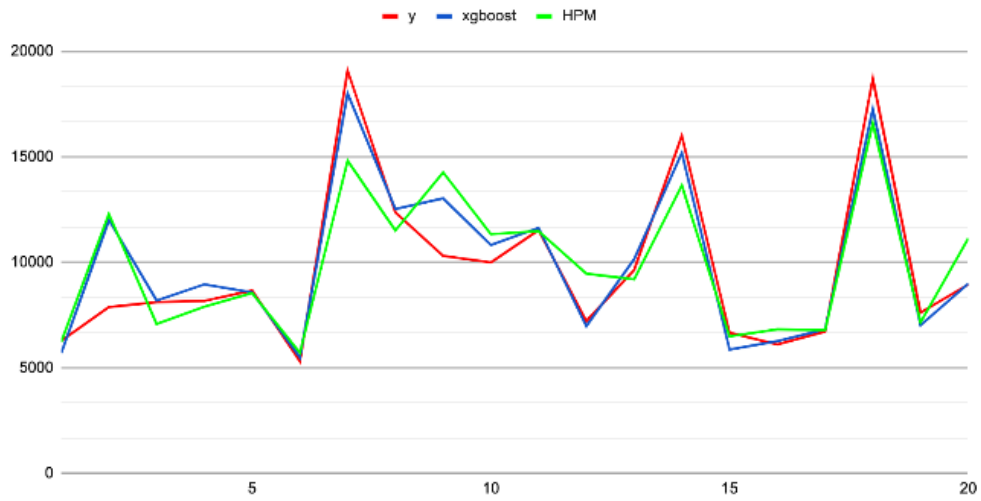

Figura 5.2: Valores Reais (y), Valores Previstos pelo XGBoost e Valores Previstos pelo HPM para o Preço por Metro Quadrado de 20 propriedades selecionadas aleatoriamente.

Com este gráfico, é possível observar que o HPM apresenta, de fato, uma variância mais alta em comparação com o XGBoost, que se aproxima melhor dos valores reais do que o HPM na maioria das propriedades selecionadas. Isso indica que o XGBoost tende a produzir previsões mais consistentes e próximas dos valores reais em comparação com o HPM, que mostra maior dispersão em suas estimativas, em linha com as estatísticas elencadas na Tabela 5.1. Essa análise reforça a superioridade do XGBoost na tarefa de prever os preços de imóveis, fato que já é conhecido na literatura específica, como foi abordado por Peng et al. (2019): “Comparado com modelos de regressão linear, o algoritmo XGBoost possui uma melhor capacidade de generalização e robustez na previsão de dados, além de prevenir o fenômeno de *overfitting*, estabelecendo uma base sólida para a subsequente previsão de preços de casas usadas”. (Tradução livre).

Após a implementação dos modelos e a clara percepção de que o XGBoost supera o HPM em termos de desempenho na previsão de preços por metro quadrado das propriedades selecionadas, a próxima etapa focará na interpretabilidade do modelo XGBoost. Nesse contexto, o uso do método SHAP (SHapley Additive exPlanations) será explorado. O SHAP é uma técnica que visa fornecer interpretabilidade aos modelos de aprendizado de máquina, ajudando a entender como as variáveis de entrada influenciam as previsões do modelo.

5.3. Interpretabilidade do Modelo com SHAP.

Como apresentado anteriormente neste trabalho, o método SHAP (SHapley Additive exPlanations) é uma ferramenta que auxilia na compreensão do impacto de cada característica, ou variável, no resultado final, especialmente para modelos considerados como "caixas-pretas". (LUNDBERG, 2017). Embora o XGBoost demonstre, em geral, uma melhor capacidade de previsão em comparação com o método hedônico, ele não fornece uma maneira tão simples como o HPM para compreender o impacto de cada variável no modelo final. Com o SHAP, podemos analisar individualmente o impacto de cada variável nas previsões do XGBoost, identificar quais características são mais significativas na determinação dos preços por metro quadrado e entender como as interações entre as variáveis afetam as previsões. Isso não apenas fornece insights valiosos para entender o comportamento do modelo, mas também pode ajudar na tomada de decisões mais informadas, seja na identificação de fatores-chave que impulsionam os preços imobiliários ou na detecção de possíveis vieses no modelo.

Semelhante ao modelo hedônico, o SHAP permite a análise da variância de cada variável no preço da propriedade previsto, exibindo-a em visualizações de fácil interpretação, como pode ser observado na Figura 5.3. O SHAP retorna as características que contribuem com a maior variância nos valores previstos do modelo em uma ordem decrescente de importância, representando no eixo x o impacto individual de cada observação, no eixo y a concentração de observações com o mesmo impacto para cada característica, e, finalmente, a cor azul representando um ponto no qual a variável em questão tem um valor baixo e a cor vermelha representando os casos em que a variável possui um valor alto. Essa análise detalhada fornece uma visão profunda das características mais influentes e como elas afetam as previsões do modelo, tornando-o mais interpretável e informativo.

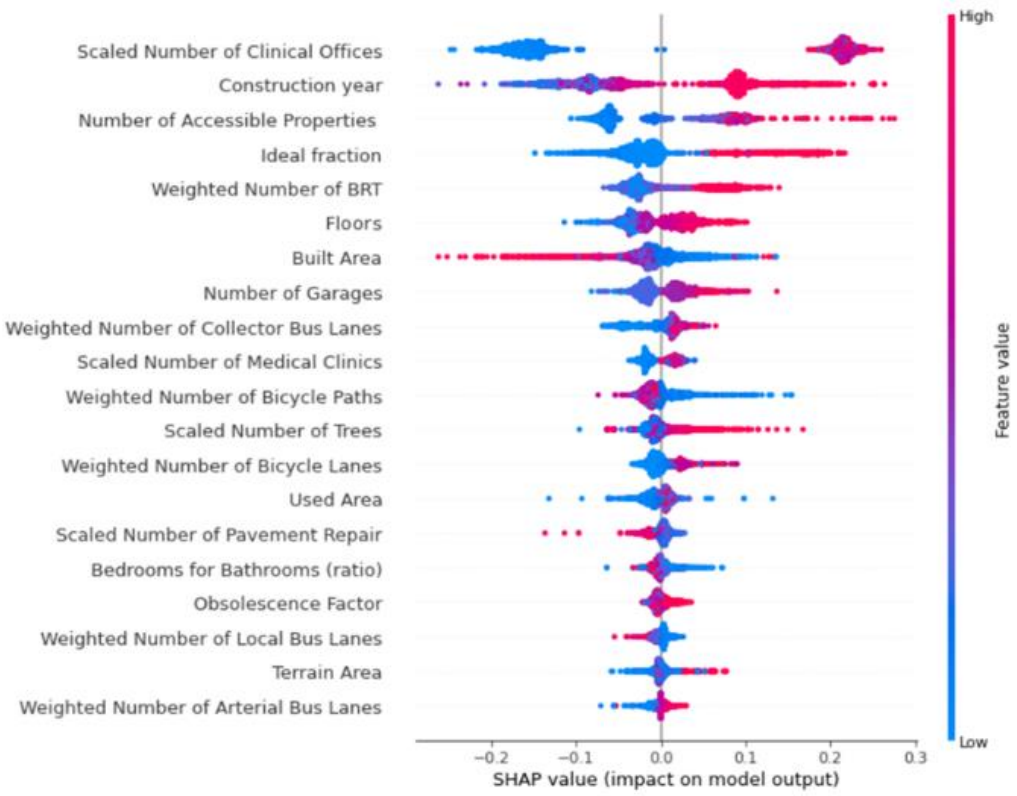

Figura 5.3: Impacto individual, medido pelo SHAP, de cada característica no preço final por metro.

Tomando como exemplo, na Figura 5.3, a variável “Ano de Construção”, fica claro como o método SHAP identifica seu impacto: Há uma concentração de pontos à esquerda do gráfico. Essa concentração possui cor azul, mostrando que são anos com valores baixos, inferiores à média da variável, ou seja, inferiores a 1990. O impacto médio desses pontos, no preço por metro quadrado, visto no eixo x, é negativo (cerca de -0.2 em escala logarítmica). Por outro lado, há outra concentração, de pontos em vermelho (anos após 2000), que impactam positivamente no valor por metro quadrado. Sendo assim, isto leva a conclusão de que o ano de construção está diretamente relacionado com o preço do imóvel, com imóveis novos sendo mais caros do que os antigos, sendo essa uma das variáveis com maior impacto. Fato já verificado em outros trabalhos que identificaram que o fator idade do imóvel exerce um dos maiores níveis de influência nas previsões dos preços dos imóveis. (GEREK, 2014).

Para a importância absoluta de cada variável no modelo, a Figura 5.4 apresenta um gráfico representando as 20 variáveis com maior impacto. Por exemplo, a variável com o maior impacto médio no preço por metro quadrado, medido por sua variância, é “Número

Escalado de Consultórios Médicos" e seu impacto é cerca de 0.2, o que representa, uma vez que o eixo x está em forma logarítmica[8], aproximadamente R$ 2.000 no preço por metro quadrado. Além disso, o gráfico mostra que cinco variáveis têm um impacto médio superior a 0.03, o que equivale a R$ 290 no preço por metro quadrado.

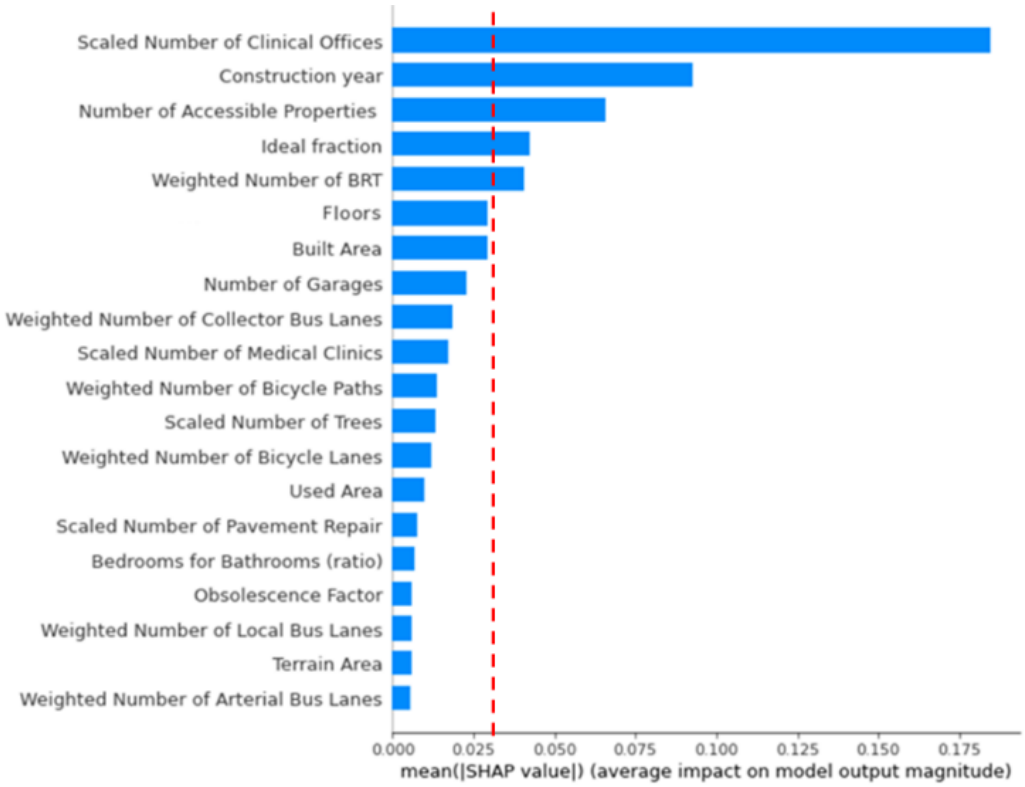

Figura 5.4: Média do impacto absoluto de cada variável do modelo no preço por metro quadrado. O impacto foi medido em uma transformação logarítmica natural no eixo x.

Uma vez que o SHAP foi aplicado ao modelo, diversas análises se tornaram possíveis para aprofundar a compreensão do comportamento do XGBoost e identificar quais características desempenham um papel mais significativo. Para tentar aprofundar a interpretabilidade fornecida pelo SHAP, foram elencadas as cinco variáveis mais impactantes, ressaltadas à direita da linha vermelha na Figura 5.4, a partir das quais uma investigação mais detalhada foi realizada. Para isso, foram utilizados gráficos individuais de explicação gerados pelo SHAP, permitindo uma análise minuciosa dessas variáveis.

[8] Cabe lembrar, conforme explicado na seção 5.1, a variável preço por metro quadrado sofreu uma transformação logarítmica. Sendo assim, o eixo x representa o logarítmico do delta no preço por metro quadrado em comparação com a média global dessa variável. A equação que converte os valores logarítmicos (x) no eixo x (Figuras 5.3 e 5.4) em valores lineares é: $\Delta preço/m^2 = avg(price/m^2) * (e^x - 1)$, onde a média do preço/m² é calculada em toda a base.

Esses gráficos permitem visualizar como cada variável contribui ou atenua o valor previsto, destacando tendências e comportamentos específicos. Essa abordagem ajuda a conectar os resultados do modelo com a realidade e a compreender de forma mais precisa como essas variáveis afetam o preço das propriedades ao longo dos seus valores

Número Escalado de Consultórios Médicos: Esta variável foi considerada a mais relevante. Analisando mais de perto os dados, é possível observar que a maioria dos consultórios clínicos em São Paulo está localizada em bairros com o Índice de Desenvolvimento Humano (IDH) mais alto, como pode ser visto na Tabela 5.2, que mostra o número de consultórios clínicos em cada um dos bairros e seu preço médio por metro quadrado.

Tabela 5.2

Valores médios das variáveis Número Escalado de Consultórios Médicos e do Preço/m², por distrito.

Distrito	Avg (Número Escalado de Consultórios Médicos)	Avg (Preço/m²) - Decrescente
Pinheiros	1.497	R$ 12.833
Moema	2.209	R$ 12.695
Vila Mariana	2.310	R$ 11.476
Vila Carrão	0.211	R$ 6.966
Vila Andrade	0.241	R$ 6.802
Jaguaré	0.143	R$ 6.616

O impacto dessa variável nos preços dos imóveis pode ser explicado por duas hipóteses diferentes: em primeiro lugar, quando o bairro é rico e tem uma alta concentração de casas de alto valor, os médicos são atraídos para abrir seus consultórios na região; em segundo lugar, quando bons médicos abrem muitos consultórios em um bairro específico, juntamente com outros fatores, mais pessoas são atraídas para procurar propriedades na região, aumentando a demanda e os preços das casas. Isso reforça a hipótese anteriormente apresentada na seção 4.2, quando notou-se que essa variável era uma das que possuía alta correlação com o preço/m².

A Figura 5.5 representa um gráfico com o efeito da variável "Número Escalado de Consultórios Médicos " nos preços dos imóveis para diferentes valores da característica. Este gráfico mostra em seu eixo x o valor da variável analisada e no eixo y o preço das propriedades imobiliárias previsto pelo modelo XGBoost, versus a sua média no banco de dados (linha central). Como pode ser observado, abaixo de 0,5, essa característica tem um impacto negativo (representado pela cor azul) nos preços, abaixando o preço em cerca de R$ 2.000 versus a média. Já acima de 0,5 o impacto se torna positivo, com alguma

variação, mas alta consistência, aumentando entre R$ 2.000 e R$ 3.000 o preço dos imóveis em relação à média. É relevante notar que essas magnitudes de variação são coerentes com o valor de impacto médio mensurado e apresentado anteriormente na Figura 5.4.

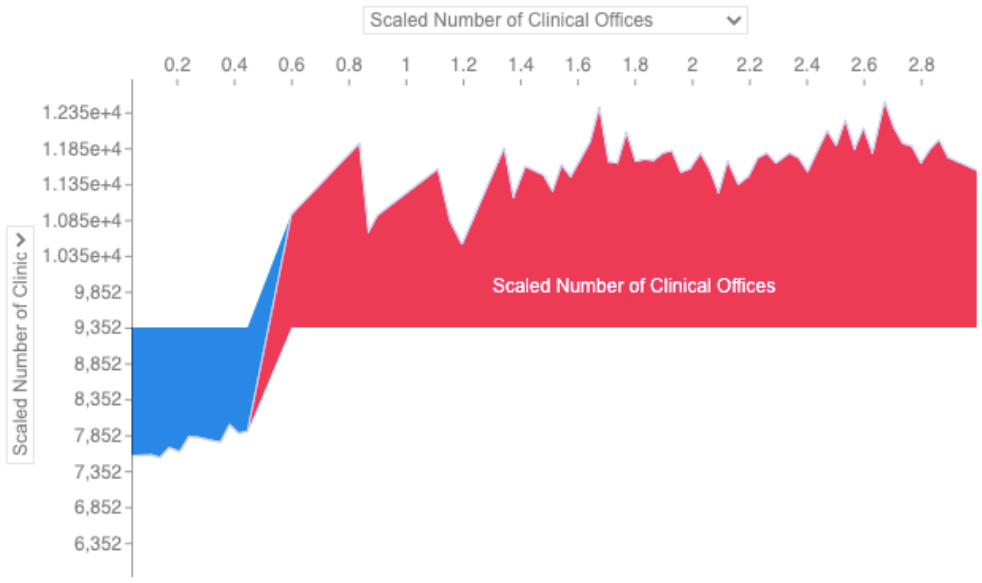

Figura 5.5: Impacto individual dos valores da variável "Número Escalado de Consultórios Médicos" no preço por metro quadrado dos imóveis analisados. A linha central representa a média de preços previstos pelo Modelo XGBoost ao longo da base de dados.

Ano de Construção: É uma conclusão intuitiva que o envelhecimento de uma propriedade impacta negativamente em seu preço, com muitos outros estudos chegando a resultados semelhantes ao modelar os preços das casas (CASE et al., 2004; GEREK, 2014). De fato, como já discutido anteriormente ao analisar a Figura 5.3, é possível observar que quase exclusivamente as propriedades mais recentes tiveram um incremento positivo na idade em relação ao preço final por metro quadrado, como evidenciado pela alta densidade de observações em vermelho no lado positivo do eixo x para esta característica. Em contrapartida, a maioria das propriedades com impacto negativo da idade no preço eram as mais antigas, também demonstrado pela alta densidade de observações em azul no eixo x à direita. O gráfico da Figura 5.6 segue a mesma lógica do gráfico anterior (Figura 5.5) e indica que, para anos acima de 2002, é possível identificar o impacto desta variável levando à preços mais altos em relação à média e para anos

anteriores a 2002, preços mais baixos. No entanto, percebe-se também que a magnitude desses impactos em relação à média é menor do que o visto na variável anteriormente analisada.

Figura 5.6: Impacto individual dos valores da variável "Ano de Construção" no preço por metro quadrado dos imóveis analisados.

Número de Imóveis com Acessibilidade: Esta métrica, descrita como a concentração (nas proximidades dos imóveis) de locais que atendem a todos os critérios de acessibilidade para pessoas com deficiência, especialmente em suas calçadas, é o terceiro fator mais impactante. Uma das hipóteses, levantadas pelos resultados de Adair et al. (2000), é que, em nível de sub-mercado, especialmente em áreas com menor renda, a acessibilidade pode ser uma influência importante, o que não necessariamente acontece em larga escala na cidade. Na Figura 5.3, o método SHAP mostrou que, para valores baixos dessa variável, existe uma alta concentração de imóveis que são impactados negativamente no preço. No entanto, quando se trata de valores altos, o impacto é mais difundido, como pode ser observado na Figura 5.7, onde para os imóveis com valores dessa variável acima de 6, há diferentes magnitudes no impacto do preço por metro quadrado, variando de R$ 500 até R$ 2.000 no seu extremo.

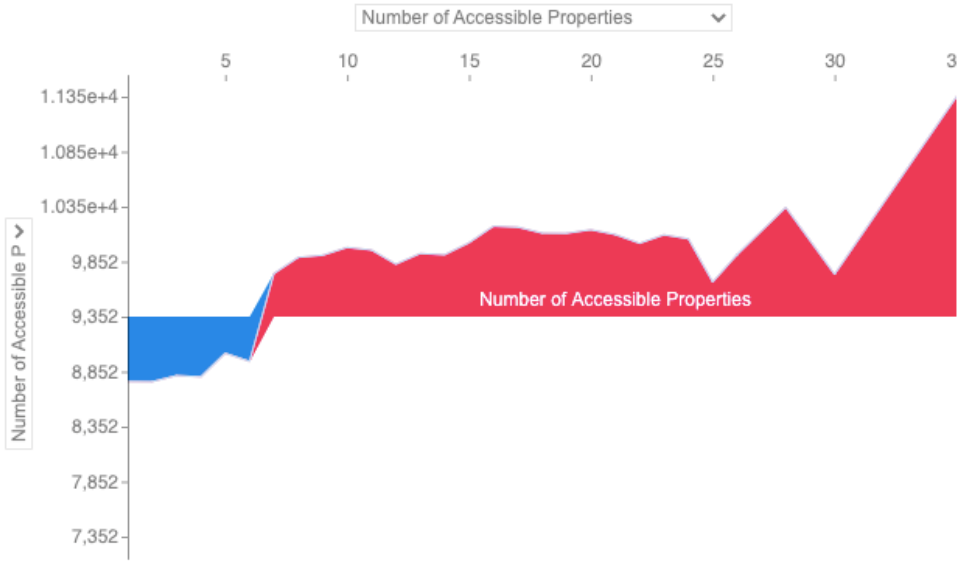

Figura 5.7: Impacto individual dos valores da variável "Número de Imóveis com Acessibilidade" no preço por metro quadrado dos imóveis analisados.

Fração Ideal: Esta característica, descrita como a porcentagem equivalente de um edifício que é "possuída" pela propriedade imobiliária em análise, foi a quarta característica mais impactante. Uma das hipóteses para essa relação está relacionada com a precificação de coberturas e os melhores apartamentos em um edifício, que normalmente apresentam um preço premium e uma área maior em comparação com outras propriedades no mesmo edifício. Consequentemente, esses imóveis possuem uma maior fração ideal, pois são os maiores apartamentos de um edifício.

É importante mencionar que, como a quarta variável mais relevante, o limiar entre o impacto negativo e positivo não é tão claramente definido como nas três variáveis anteriormente expostas, um fato que pode ser observado na Figura 5.8. Considerando o comportamento mostrado na Figura 5.8, nota-se um alto impacto positivo quando a variável tem um valor de 1, já que nesses casos (fração ideal igual à 1), o imóvel em questão trata-se de uma casa. Na cidade de São Paulo, casas costumam ter preços por metro quadrado mais elevados que apartamentos.

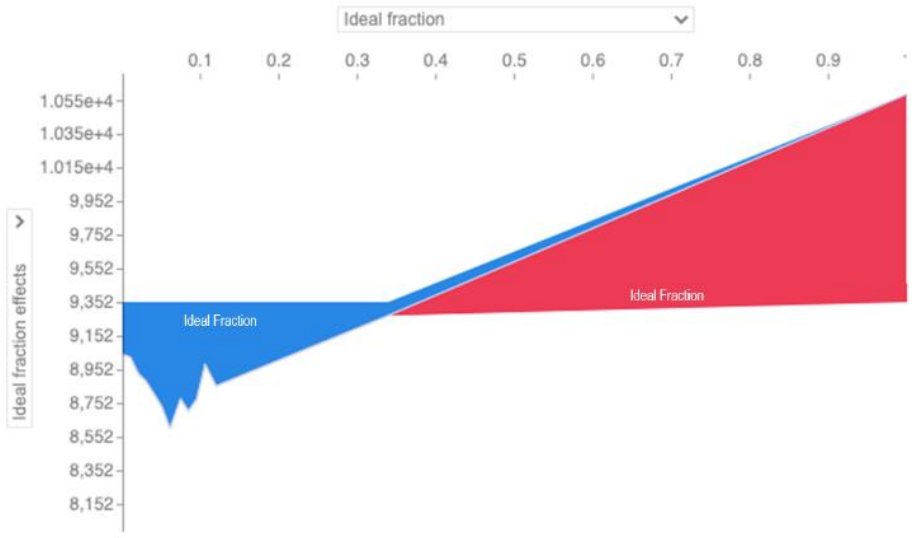

Figura 5.8: Impacto individual dos valores da variável "Fração Ideal" no preço por metro quadrado dos imóveis analisados.

Número Ponderado de BRTs: A proximidade com a infraestrutura de transporte rápido e eficiente no meio urbano está correlacionada com preços de casas mais altos, como muitos outros estudos já demonstraram (MCMILLEN et al., 2004; FILLIPOVA et al., 2020). A Figura 5.9 mostra consistência com essa tendência, onde valores elevados desta característica estão relacionados a um impacto positivo nos preços das propriedades no modelo, enquanto valores mais baixos estão relacionados a um impacto negativo nos preços das propriedades.

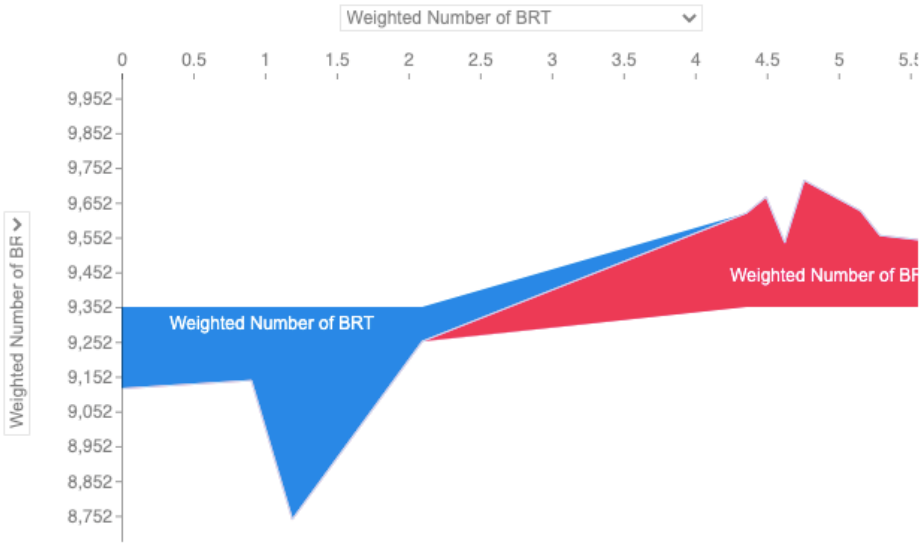

Figura 5.9: Impacto individual dos valores da variável "Número Ponderado de BRTs" no preço por metro quadrado dos imóveis analisados.

De fato, a presença de um transporte eficiente em uma cidade desempenha um papel crucial na valorização dos imóveis. Isso se deve à conveniência e acessibilidade que um sistema de transporte público de alta qualidade oferece aos residentes, reduzindo o tempo de deslocamento e as despesas relacionadas a veículos. A valorização imobiliária é frequentemente impulsionada pela busca de moradias próximas a estações e rotas de transporte público, resultando em um desenvolvimento urbano mais amplo e melhorias na infraestrutura local. Nesse sentido, esse impacto na valorização dos imóveis é algo que deve ser levado em conta também pelo poder público, sobre o qual recai a responsabilidade de planejar a rede de transporte da cidade.

A seguir foi feito um resumo da Figura 5.4, para as variáveis elencadas, em formato de tabela e com os valores de impactos relativos já convertidos tanto para valores absolutos, quanto no impacto relativo em relação à média do preço por metro quadrado. Esses resultados estão elencados na Tabela 5.3

Tabela 5.3

Impacto Logarítmico, Absoluto e Relativo em relação ao preço/m² das 20 principais variáveis do modelo.

Variável	Impacto Médio (SHAP-Log)	Impacto Absoluto (Linear)	% Var (Preço/ m²)	Acum % Var (Preço/ m²)
Número Escalado de Consultórios Médicos	0,2014	R$ 2.086,70	22,3%	22%
Ano de Construção	0,1030	R$ 1.014,68	10,8%	33%
Número de imóveis com acessibilidade	0,0650	R$ 628,11	6,7%	40%
Fração Ideal	0,0401	R$ 382,66	4,1%	44%
Número ponderado de BRT	0,0385	R$ 367,10	3,9%	48%
Número de andares	0,0306	R$ 290,61	3,1%	51%
Área Construída	0,0299	R$ 283,87	3,0%	54%
Número de Garagens	0,0254	R$ 240,60	2,6%	57%
Número ponderado de Faixas de ônibus Coletoras	0,0213	R$ 201,35	2,2%	59%
Número Escalado de Clínicas Médicas	0,0208	R$ 196,57	2,1%	61%
Número ponderado de Ciclofaixas	0,0191	R$ 180,35	1,9%	63%
Número Escalado de Árvores	0,0185	R$ 174,63	1,9%	65%
Número ponderado de Ciclovias	0,0167	R$ 157,50	1,7%	66%
Área útil	0,0158	R$ 148,94	1,6%	68%
Número Escalado de Reparos no Pavimento	0,0152	R$ 143,25	1,5%	69%
Quartos por Banheiros	0,0133	R$ 125,22	1,3%	71%
Fator de obsolescência	0,0128	R$ 120,48	1,3%	72%
Número ponderado de Faixas de ônibus Locais	0,0120	R$ 112,91	1,2%	73%
Área do Terreno	0,0117	R$ 110,07	1,2%	74%
Número ponderado de Faixas de ônibus Arteriais	0,0105	R$ 98,72	1,1%	76%

Na Tabela 5.3 vemos o impacto quantitativo médio, em reais, no preço por metro quadrado dos imóveis, segundo o modelo XGBoost. É interessante observar que, das 10 características mais relevantes no modelo, 5 são intrínsecas (Ano de Construção, Fração

Ideal, Número de Pavimentos, Área Construída e Número de Garagens), que se relacionam diretamente com a própria propriedade; e 5 são extrínsecas (Número de Consultórios Clínicos Padronizados, Número de Propriedades Acessíveis, Número Ponderado de BRT, Número de Árvores Padronizadas e Número Ponderado de Faixas de Ônibus Coletores), que se relacionam com a localização e as facilidades ao redor da propriedade. No entanto, uma vez que foram utilizadas 30 variáveis extrínsecas e apenas 11 variáveis intrínsecas, o estudo sugere uma maior importância das variáveis intrínsecas na previsão dos preços das propriedades.

Além disso, percebe-se que dessas 20 variáveis selecionadas, sete delas (em vermelho na Tabela 5.3) são relacionadas ao investimento público no ambiente construído da cidade, sendo que o número ponderado de BRTs é a variável mais impactante dentre essas 7, representando quase 4% do valor do preço por metro quadrado do imóvel. Por outro lado, olhando na coluna do impacto acumulado, vemos que essas 20 primeiras variáveis possuem um impacto acumulado de 76%, o que é muito próximo da assertividade do modelo (Coeficiente de Determinação), que ficou em aproximadamente 86%. Esse fato demonstra que as outras 22 variáveis, que não estão na tabela, tem impacto marginal, dentro do desvio padrão do modelo, podendo serem desconsideradas numa análise mais aprofundada.

Na próxima seção, a análise se aprofundará nessas sete variáveis que estão diretamente relacionadas aos investimentos públicos no ambiente construído da cidade. O objetivo será entender melhor o impacto dessas variáveis no preço das propriedades e, igualmente importante, examinar seu retorno para a prefeitura por meio de impostos. Será explorado como essas variáveis contribuem para a valorização das propriedades e, por extensão, para a arrecadação de receitas públicas, proporcionando uma visão mais abrangente das implicações econômicas desses investimentos urbanos

6 Análise dos Impactos e Resultados

À medida que o trabalho se aproxima de sua conclusão, é importante ressaltar que os objetivos estabelecidos no capítulo introdutório foram abordados em etapas planejadas e executadas. Primeiramente, foi construída uma base de dados sólida e abrangente que representa o mercado imobiliário da cidade de São Paulo, garantindo a confiabilidade dos dados utilizados na análise. Em seguida, procedeu-se com a implementação de um modelo de aprendizado de máquina, preditivo de preços de imóveis, que fosse não apenas preciso, mas também representativo das complexas interações que moldam o mercado imobiliário, visando capturar os múltiplos fatores que influenciam os preços das propriedades em São Paulo e realizando um *benchmark* com outro modelo mais simples, de regressão múltipla, comprovando a superioridade do modelo de *machine learning*.

Por fim, adotou-se uma abordagem explanatória, baseada no método SHAP, para identificar as variáveis mais impactantes nos preços das propriedades e quantificar esses impactos com base em evidências concretas. Nesse sentido, foram identificadas sete variáveis, que possuem impacto relevante no preço dos imóveis e são relacionadas ao investimento público no ambiente construído da cidade. Sendo assim, para finalmente atingir o objetivo final deste trabalho, resta quantificar esse impacto de forma mensurável, no preço dos imóveis e a influência desse impacto na arrecadação municipal, especificamente com relação ao imposto sobre propriedades imobiliárias.

6.1. Cálculo do Imposto Territorial e Predial Urbano (IPTU)

Nesta seção se faz necessária uma análise de como é apurado o Imposto Predial Territorial Urbano (IPTU) na cidade de São Paulo, tendo em mente que é necessário avaliar como o impacto previsto no preço dos imóveis, através do modelo aqui implementado, afeta a arrecadação municipal, especificamente do IPTU, que é um imposto baseado no valor de mercado do imóvel (valor venal), no entanto com particularidades em sua apuração.

A Base de Cálculo do Imposto Predial e Territorial Urbano (IPTU) corresponde ao valor venal do imóvel. A determinação desse valor é realizada com base nas informações do imóvel registradas no cadastro da Secretaria da Fazenda da Cidade de São Paulo, incluindo a área do terreno, área construída, idade da construção, entre outros dados relevantes, se aproximando do valor de mercado do imóvel. Essa avaliação segue os critérios e parâmetros estabelecidos pela Lei 10.235/1986 e suas atualizações, garantindo uma metodologia consistente e transparente para a determinação do valor venal utilizado como base de cálculo do IPTU. (PMSP, 2023)

Sobre o Valor Venal, a depender da faixa em que se encontra, incidem alíquotas, com descontos relativos. A Tabela 6.1 mostra como essas alíquotas e descontos são aplicados sobre o Valor Venal, resultando no valor final do imposto (IPTU) .

Tabela 6.1

Forma de Cálculo do IPTU, baseado no Valor Venal do imóvel.

Faixa de Valor Venal	Multiplicar	Subtrair
Até R$ 150.000,00	0,007	R$ 0,00
De R$ 150.001,00 a R$ 300.000,00	0,009	R$ 300,00
De R$ 300.001,00 a R$ 600.000,00	0,011	R$ 900,00
De R$ 600.001,00 a R$ 1.200.000,00	0,013	R$ 2.100,00
Acima de R$ 1.200.000,00	0,015	R$ 4.500,00

Como o próprio cálculo do imposto depende especificamente do valor absoluto do imóvel, o cálculo do impacto, nesse imposto, de uma valorização no preço por metro quadrado de um imóvel seria bastante complexo, pois não haveria como saber se o imóvel passaria a pertencer a outra faixa de tributação, na qual haveria outra alíquota. Pensando no impacto global na cidade, a situação seria ainda mais complexa, sendo necessário avaliar cada imóvel individualmente, qual faixa se encontra e como diferentes impactos no seu preço poderiam alterar a faixa, alíquota e desconto.

Nesse sentido, encontrar um modelo para aproximar o cálculo do IPTU, seria muito mais eficiente e permitiria realizar uma análise contínua e global, encontrando uma função que relacionasse diretamente o impacto no preço do imóvel, com o impacto no IPTU.

Seguindo esta linha de raciocínio, o primeiro passo foi realizar a plotagem dos dados da Tabela 6.1, para diferentes preços de imóveis simulados. Os resultados dessa plotagem podem ser vistos na Figura 6.1.

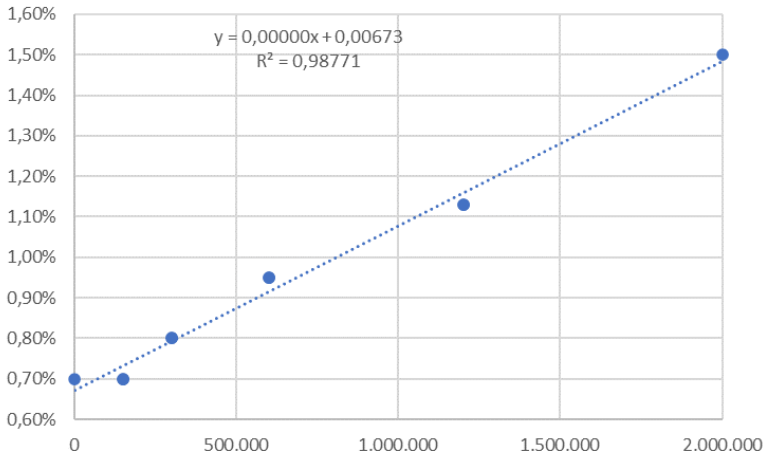

Figura 6.1: Simulação da alíquota efetiva do IPTU, para diferentes valores de imóveis.

Na Figura 6.1, é possível ver que a reta ajustada aos pontos, pelo método dos mínimos quadrados, tem um coeficiente de determinação (R^2) de 0,99, ou seja, com essa assertividade é plenamente possível adotar a equação dessa reta, dependente do valor venal do imóvel, como forma de cálculo da Alíquota Efetiva do IPTU, em uma forma de aproximar e simplificar os dados da Tabela 6.1. Essa reta ajustada, tem o coeficiente angular (α) de $4{,}05008 * 10^{-9}$ e coeficiente linear (β) $6{,}72741 * 10^{-3}$.

Portanto, para cálculo do IPTU de um determinado imóvel, tem-se o sistema descrito a seguir, na Equação 6.1.

$$\begin{cases} Alq = \alpha * Vv + \beta \\ IPTU = Vv * Alq \end{cases}$$

Equação 6.1: Fórmula de Cálculo do IPTU, a partir da reta encontrada.

Com a Equação 6.1, é possível obter o valor do IPTU, baseado no Valor Venal (Vv.) do imóvel. Seguindo na dedução, o próximo passo será simplificar o sistema da Equação 6.1, e adotar dois períodos de tempo, t_1 e t_2, supondo uma variação desse Valor Venal nesses entre esses períodos. A Equação 6.2 demonstra essa operação algébrica.

$$IPTU_1 = Vv_1 * (\alpha * Vv_1 + \beta) \quad (1)$$

$$IPTU_2 = Vv_2 * (\alpha * Vv_2 + \beta) \quad (2)$$

Equação 6.2: IPTU em diferentes períodos de tempo, baseado na variação do Valor Venal.

Como o objetivo é encontrar a variação do IPTU, dependente da variação do Valor Venal do imóvel, o próximo passo será dividir (2) por (1) e seguir na dedução algébrica. A Equação 6.3 demonstra esse cálculo.

$$\frac{IPTU_2}{IPTU_1} = \Delta \boldsymbol{IPTU} = \frac{Vv_2 * (\alpha * Vv_2 + \beta)}{Vv_1 * (\alpha * Vv_1 + \beta)}.$$

$$No\ entanto, Vv_2 = \delta * Vv_1,\ \ onde\ \delta\ é\ uma\ constante\ \in \mathbb{R}.\ Então:$$

$$\Delta \boldsymbol{IPTU} = \frac{Vv_2 * (\alpha * Vv_2 + \beta)}{Vv_1 * (\alpha * Vv_1 + \beta)} = (\delta) * \frac{(\alpha * (\delta * Vv_1) + \beta)}{(\alpha * Vv_1 + \beta)}$$

$$= \delta * \frac{\delta * \alpha * Vv_1 + \beta}{\alpha * Vv_1 + \beta}$$

Equação 6.3: Variação do IPTU dependente do valor venal inicial e sua variação.

Nota-se, a partir da Equação 6.3, que a variação do IPTU, agora está resumida em uma equação que depende apenas da variação do valor venal (δ) e do próprio valor venal inicial (Vv_1). No entanto, depender do valor venal, ainda é um complicador, pois este estudo até o momento direcionou todos os esforços em entender o impacto de diversas variáveis no preço por metro quadrado de cada imóvel e não em seu valor absoluto, correspondente ao valor venal. Sendo assim, um próximo passo será estudar como a função encontrada na Equação 6.3 se comporta nos limites extremos, quando o valor venal tende a zero ou a infinito. Esse estudo dos limites da função está desenvolvido na Equação 6.4.

$$\lim_{Vv_1 \to 0} \Delta \boldsymbol{IPTU} = \lim_{Vv_1 \to 0} \delta * \frac{\delta * \alpha * Vv_1 + \beta}{\alpha * Vv_1 + \beta} = \delta$$

$$\Delta \lim_{Vv_1 \to +\infty} \Delta \boldsymbol{IPTU} = \lim_{Vv_1 \to +\infty} \delta * \frac{\delta * \alpha * Vv_1 + \beta}{\alpha * Vv_1 + \beta} = \delta^2$$

Equação 6.4: Limites da função.

O resultado visto na Equação 6.4 é extremamente interessante, pois demonstra que a função que define a variação do IPTU está limitada, em seu domínio, entre a variação do próprio valor venal do imóvel (δ) e seu valor ao quadrado. De fato, realizando uma simulação de valorização de 10% no valor venal (e consequentemente no seu preço por metro quadrado) em um dado imóvel, vemos que a depender do valor venal do imóvel, o ganho no IPTU seria algo entre 10% e 21% (equivalente a 1.1^2). O resultado da simulação está ilustrado na Figura 6.2.

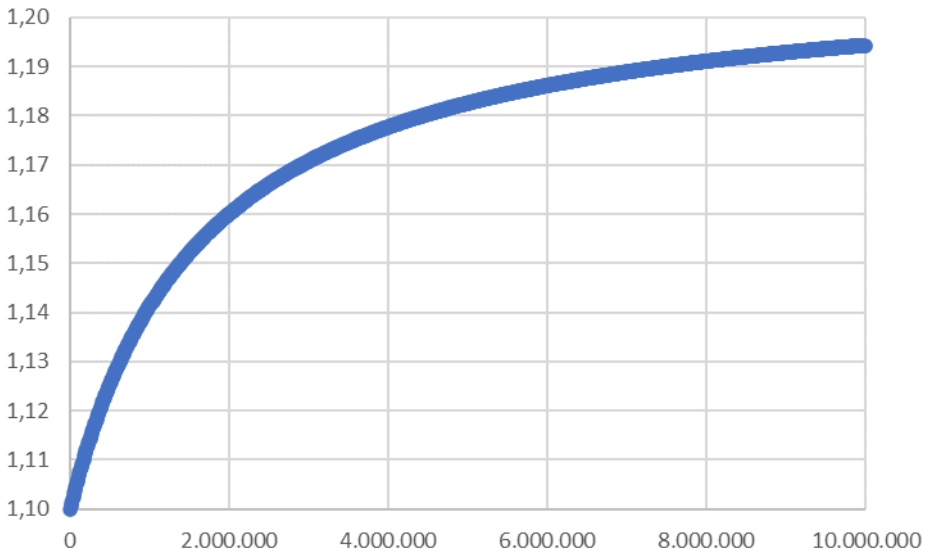

Figura 6.2: Simulação do ganho no IPTU de um imóvel, quando ele se valoriza 10%, considerando valores para esse imóvel entre R$ 0 e R$ 10 Milhões.

Sendo assim, com os resultados obtidos até aqui, pode-se concluir que a valorização de um imóvel desempenha um papel significativo na arrecadação de Imposto Predial e Territorial Urbano (IPTU). A partir dessa análise também foi possível concluir que,

aproximadamente, através da modelagem de uma alíquota linear para esse imposto, o ganho na arrecadação do IPTU, pode ser aproximado pelo valor da valorização do preço por metro quadrado de um imóvel, variando entre o percentual dessa valorização e este mesmo percentual ao quadrado, a depender do valor absoluto do imóvel.

Com essa conclusão, o próximo passo será aplicar essa faixa de ganho mensurada, nas variáveis anteriormente selecionadas com maior impacto – e ligadas ao investimento público no ambiente construído da cidade – buscando evidenciar qual fornece o maior retorno e deveria ser priorizada pelo administrador público responsável.

6.2. Sugestão de Escolha de Investimentos Baseado em Seu Impacto no Modelo.

Neste ponto é preciso retornar ao objetivo final que foi exposto anteriormente, no início do Capítulo 4, o qual consiste em encontrar os investimentos públicos que possuem os maiores impactos positivos (valorização) no preço dos imóveis e mensurar – ou estimar – este impacto no Imposto Predial Territorial Urbano (IPTU), sugerindo sua escolha para o administrador público.

Dividindo este objetivo em partes, deve-se atentar para dois detalhes importantes, os quais são fundamentais de serem garantidos na análise a seguir: busca-se encontrar investimentos que levem à valorização (ganho) dos imóveis avaliados e, somente então, mensurar o impacto desta valorização na arrecadação do IPTU destes imóveis. Seguindo nesta linha, a primeira etapa a ser feita é realizar uma análise para cada uma das sete variáveis selecionadas anteriormente, nos moldes do que foi feito na seção 5.3, utilizando os gráficos de impacto específico gerados pelo método SHAP. As variáveis avaliadas serão:

- *Número ponderado de BRT*
- *Número ponderado de Faixas de ônibus Coletoras*
- *Número ponderado de Ciclofaixas*
- *Número ponderado de Ciclovias*
- *Número Escalado de Reparos no Pavimento*
- *Número ponderado de Faixas de ônibus Locais*
- *Número ponderado de Faixas de ônibus Arteriais*

Os gráficos da análise realizada podem ser vistos na Figura 6.3

Figura 6.3: Gráficos de Impacto SHAP, para as sete variáveis elencadas.

Através dos gráficos, a primeira consideração a ser feita é a de que a variável "número escalado de reparos no pavimento" não apresenta uma tendência clara no impacto no preço dos imóveis avaliados. Esta variável alterna entre impactos negativos e positivos, evidenciando que não é possível inferir um significado para tentar entender qual seria seu valor ideal, que impactaria positivamente o modelo. Sendo assim ela será descartada aqui na análise.

Por outro lado, as variáveis "número ponderado de ciclofaixas" e "número ponderado de faixas de ônibus locais" apresentaram um comportamento inverso ao esperado. Um aumento de seus valores causa um impacto negativo, ou seja, desvalorização no preço dos imóveis. Esse comportamento, inclusive, já foi discutido anteriormente na seção 4.3, quando foi levantada a hipótese de que essas infraestruturas públicas (faixas de ônibus locais e ciclofaixas) estão mais presentes na periferia do que no centro urbano da cidade de São Paulo, o que pode demonstrar que há uma correlação dessas variáveis com preços mais baixos de imóveis, mas não uma causalidade, ou seja, a presença desse tipo de infraestrutura não prejudica o preço, apenas é um *proxy* para se identificar áreas mais periféricas e menos valorizadas. No entanto, mesmo considerando isso, como no modelo essas variáveis apresentam impactos negativos, elas também serão desconsideradas aqui na escolha, por incapacidade de mensurar seu real impacto.

Temos então quatro variáveis restantes que claramente impactam positivamente o preço dos imóveis, em maior ou menor medida, de acordo com o modelo. São elas: "Número ponderado de BRT", "Número ponderado de faixas de ônibus Coletoras", "Número ponderado de Ciclovias" e "Número ponderado de faixas de ônibus Arteriais". Todas as variáveis são relacionadas à infraestrutura de transporte na cidade. Relembrando os resultados discutidos na seção anterior, teremos um impacto previsto na arrecadação de IPTU, conforme elencado na Tabela 6.2.

Tabela 6.2

Impacto mínimo e máximo na arrecadação de IPTU, para as quatro variáveis selecionadas.

Variável	% Var (Preço/m²)	ΔIPTU min	ΔIPTU max
Número ponderado de BRT	3,9%	3,9%	8,0%
Número ponderado de Faixas de ônibus Coletoras	2,2%	2,2%	4,4%
Número ponderado de Ciclovias	1,7%	1,7%	3,4%
Número ponderado de Faixas de ônibus Arteriais	1,1%	1,1%	2,1%

Percebe-se que a partir dos dados da Tabela 6.2, a variável “Número Ponderado de BRT”, possui a maior expectativa de impacto, podendo aumentar o valor do IPTU arrecadado nos imóveis afetados pela construção de uma nova linha de BRT, por exemplo, em até 8%. Lembrando aqui que neste estudo, foi considerado, especificamente para essa variável, os imóveis impactados num raio de 2 km desta infraestrutura de transporte. Informação importante que demonstra que seu raio de influência é alto, afetando grande quantidade de imóveis. De fato, a literatura já demonstra há anos, em diversos trabalhos, que uma estrutura de transporte público eficiente (exatamente como são os BRTs na cidade de São Paulo) é um dos principais fatores de valorização dos imóveis em uma metrópole. (Zhou et al., 2022).

Finalmente, é interessante comentar como o “Número ponderado de Ciclovias” também exerce grande influência nos preços, podendo impactar em até 3.4% a arrecadação. Ciclovias são uma das estruturas de transporte com melhor custo-benefício na sua implantação, uma vez que são baratas de construir, quando comparadas com outros meios de transporte de alta demanda e ainda trazem benefícios para a saúde e estilo de vida dos habitantes da cidade. (GU & MUENNIG, 2017).

Chega-se então à conclusão desse trabalho, sugerindo que a implantação de corredores de Ônibus (BRTs) é, dos investimentos aqui avaliados, o que traz maior impacto nos preços das propriedades. Por outro lado, o investimento em Ciclovias é o que possui maior custo-benefício por sua facilidade de implantação, custo e impacto financeiro nos imóveis.

7 Conclusão e Trabalhos Futuros

Este trabalho chega ao seu final com algumas conclusões a serem ressaltadas aqui, focando principalmente em entender quais respostas foram dadas aos objetivos inicialmente endereçados no capítulo introdutório e reforçados no início do Capítulo 4.

O primeiro objetivo consistia no treino e teste de um modelo de aprendizado de máquina baseado em árvores de decisão, sendo escolhido o modelo XGBoost, e sua comparação com um modelo hedônico, no caso o Semi-Log, avaliando o desempenho de ambos os modelos e seus resultados na previsão dos preços a partir de sua aplicação em uma base de dados contendo 42 variáveis de entrada, relacionadas aos 4.019 imóveis avaliados, e uma variável de saída a ser modelada, o preço por metro quadrado. Nesse sentido, confirmado a partir da ampla literatura discutida anteriormente, o XGBoost mostrou-se superior ao modelo hedônico nesta aplicação, resultando em um R-quadrado de 0.86, que é 25% maior ao ser comparado com o coeficiente de determinação resultante no modelo Semi-Log. Além disso, o XGBoost apresentou melhores valores em todas as outras métricas estatísticas consideradas, como apresentado na Tabela 5.1. Esse resultado pode indicar que o XGBoost é um modelo capaz de fornecer resultados superiores no segmento de precificação imobiliária, eventualmente sendo utilizado por empresas imobiliárias para melhorar a precisão dos preços negociados.

Já o segundo objetivo consistia na aplicação do método SHAP, que se trata de um método explanatório para modelos baseados em árvore, exatamente o caso do XGBoost. Esse método permite fornecer para um modelo *caixa-preta* a interpretabilidade das variáveis de entrada, explicando como cada uma delas impacta nos resultados da modelagem e

como o modelo se comporta a partir dos diferentes valores delas ao longo da base de dados. Com a identificação das variáveis mais impactantes e de seus respectivos impactos, para concluir este objetivo, o próximo passo seria a identificação de quais delas estavam relacionadas ao investimento público.

Seguindo esta linha de raciocínio, primeiramente o trabalho aplicou o método SHAP ao modelo XGBoost previamente treinado, encontrando as 20 variáveis com maiores impactos médios no preço por metro quadrado modelado, seus impactos individuais foram calculados e o resultado combinado mostrou que essas 20 variáveis representavam cerca de 76% da variabilidade do preço do imóvel, sendo que as cinco primeiras (número escalado de consultórios médicos, ano de construção, número de imóveis com acessibilidade, fração ideal, número ponderado de BRT) representam 48% dessa mesma variabilidade. Finalmente, dentre essas 20 variáveis, 7 delas são relacionadas ao investimento público no ambiente construído da cidade, das quais 6 estão interligadas especificamente à infraestrutura de transporte urbano. A mais importante delas, o número ponderado de BRTs (corredores de ônibus de alta capacidade), mostram uma conclusão já bastante conhecida na literatura, de que uma rede de transporte eficiente afeta positivamente o preço dos imóveis na sua área de influência. Os BRTs na cidade de São Paulo são confortáveis e confiáveis, além de garantirem um acesso relativamente rápido a várias partes da cidade. Esses resultados podem ser um indicativo de que investimentos na infraestrutura de transporte podem trazer um ganho de valor imobiliário para a cidade.

Finalmente o objetivo final deste trabalho consistia em realizar uma análise das variáveis com maior impacto positivo no preço por metro quadrado dos imóveis e quantificar seu impacto tanto nos preços quanto na arrecadação de impostos decorrentes dessa valorização imobiliária, sugerindo a escolha dos melhores investimentos para o administrador público. Buscando atingir esse objetivo, este trabalho demonstrou que o aumento na arrecadação de impostos imobiliários, especificamente o IPTU (Imposto predial e Territorial Urbano) está diretamente relacionado ao ganho de valor imobiliário de um imóvel, sendo este aumento arrecadatório limitado entre o percentual do ganho e o valor deste ganho ao quadrado, dependendo do valor absoluto do imóvel. Além disso, das sete variáveis anteriormente identificadas no objetivo anterior, utilizando a quantificação estatística fornecida pelo método SHAP, três dessas variáveis foram descartadas por não terem uma relação causal comprovada com a alteração no preço dos imóveis. As quatro variáveis restantes identificadas (número ponderado de BRT, número

ponderado de faixas de ônibus coletoras, número ponderado de ciclovias, número ponderado de faixas de ônibus arteriais) tiveram então seus impactos, tanto no preço dos imóveis, quanto na arrecadação do IPTU, quantificados.

Ficou demonstrado que essas variáveis podem resultar, em conjunto, em um aumento entre 8% e 18% na arrecadação do imposto, nos imóveis em sua zona de influência. Além disso, a terceira mais importante, relacionada ao número de ciclovias, demonstrou uma capacidade de impactar em até 3,4% a arrecadação, podendo indicar ser uma das melhores escolhas em se tratando de custo-benefício, já que ciclovias tem implantação menos custosa e mais rápida em comparação a outros meios de transporte de passageiros de alta demanda.

Por fim, como conclusão deste trabalho, vale ressaltar que existem limitações para extrapolar esses resultados, especialmente considerando que a amostra analisada (4.019 propriedades) ainda é relativamente pequena para a cidade de São Paulo. Além disso, os preços modelados para essas propriedades foram obtidos a partir de anúncios online e não dos registros governamentais de transações imobiliárias (que são de código aberto). Em futuros trabalhos, a adoção de uma amostra maior pode trazer resultados melhores na identificação do impacto do investimento público nos preços das propriedades, permitindo avaliar este impacto dentro de cada bairro, por exemplo. Possibilidades para trabalhos futuros incluem a expansão dos bairros estudados em São Paulo, oferecendo uma perspectiva mais ampla sobre o impacto de características extrínsecas nos preços das habitações; a aplicação de outros modelos baseados em árvores ao banco de dados em busca de resultados mais otimizados e a expansão do banco de dados para incluir outros dados extrínsecos, como os fornecidos pelo Google Maps (restaurantes, farmácias, entre outros.).

Referências Bibliográficas

ADAIR, A., MCGREAL, S., SMYTH, A., COOPER, J., & RYLEY, T. (2000). House prices and accessibility: The testing of relationships within the Belfast urban area. Housing studies, 15(5), 699-716.

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

BARRECA, A., CURTO, R., & ROLANDO, D. (2020). Urban vibrancy: An emerging factor that spatially influences the real estate market. *Sustainability*, *12*(1), 346.

CASE, B., CLAPP, J., DUBIN, R., & RODRIGUEZ, M. (2004). Modeling spatial and temporal house price patterns: A comparison of four models. The Journal of Real Estate Finance and Economics, 29, 167-191.

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.

CHEN, L., YAO, X., LIU, Y., ZHU, Y., CHEN, W., ZHAO, X., & CHI, T. (2020). Measuring impacts of urban environmental elements on housing prices based on

multisource data—a case study of Shanghai, China. *ISPRS International Journal of Geo-Information*, *9*(2), 106.

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.

CHERNOFF, A., & CRAIG, A. N. (2022). Distributional and Housing Price Effects from Public Transit Investment: Evidence from Vancouver. *International Economic Review*, *63*(1), 475-509.

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.

DIMOPOULOS, T., & BAKAS, N. (2019). An artificial intelligence algorithm analyzing 30 years of research in mass appraisals. *RELAND: International Journal of Real Estate & Land Planning*, *2*, 10-27.

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

DOU, M., Gu, Y., & FAN, H. (2023). Incorporating neighborhoods with explainable artificial intelligence for modeling fine-scale housing prices. *Applied Geography*, *158*, 103032.

FAN, S., JITSUCHON, S., & METHAKUNNAVUT, N. (2004). *The importance of public investment for reducing rural poverty in middle-income countries: The case of Thailand* (No. 580-2016-39362).

FILIPPOVA, O., & SHENG, M. (2020). Impact of bus rapid transit on residential property prices in Auckland, New Zealand. Journal of transport geography, 86, 102780.

GEREK, I. H. (2014). House selling price assessment using two different adaptive neuro-fuzzy techniques. *Automation in Construction*, *41*, 33-39.

GU, J., MOHIT, B., & MUENNIG, P. A. (2017). The cost-effectiveness of bike lanes in New York City. *Injury prevention*, *23*(4), 239-243.

HEIDARI, M., ZAD, S., & RAFATIRAD, S. (2021). 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. (2018). Impact of distance to school on housing price: Evidence from a quantile regression. The Empirical Economics Letters, 17(2), 149-156.

JUSZCZAK, P., TAX, D., & DUIN, R. P. (2002). Feature scaling in support vector data description. In Proc. asci (pp. 95-102). Citeseer.

KOH, P. W., & LIANG, P. (2017). Understanding black-box predictions via influence functions. In International conference on machine learning (pp. 1885-1894). PMLR.

KOMAGOME-TOWNE, A. (2016). Models and visualizations for housing price prediction. *Faculty of California State Polytechnic University, Pomona*.

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.

LUNDBERG, S. M., & LEE, S. I. (2017). A unified approach to interpreting model predictions. Advances in neural information processing systems, 30.

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.

MAASS, A. (1966). Benefit-cost analysis: Its relevance to public investment decisions. *The Quarterly Journal of Economics*, *80*(2), 208-226.

MALPEZZI, S. (2003). Hedonic pricing models: a selective and applied review. *Housing economics and public policy*, *1*, 67-89.

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. In *Selection at the AMSE Conferences* (p. 97).

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.

MCMILLEN, D. P., & MCDONALD, J. (2004). Reaction of house prices to a new rapid transit line: Chicago's midway line, 1983–1999. Real Estate Economics, 32(3), 463-486.

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.

MOKHTARI, K. E., HIGDON, B. P., & BAŞAR, A. (2019). Interpreting financial time series with SHAP values. In Proceedings of the 29th Annual International Conference on Computer Science and Software Engineering (pp. 166-172).

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.

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.

PENG, Z., HUANG, Q., & HAN, Y. (2019). Model Research on Forecast of Second-Hand House Price in Chengdu Based on XGboost Algorithm. 2019 IEEE 11th International Conference on Advanced Infocomm Technology (ICAIT), 168-172. https://doi.org/10.1109/ICAIT.2019.8935894.

PEREZ, R. A., & KIMURA, D. S. (2015). Análise de mercado como ferramenta para a abertura de novos loteamentos. *Revista Brasileira de Planejamento e Desenvolvimento*, *3*(1), 129-141.

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.

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.

RUDIN, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206-215.

SCHULZ, R., & WERWATZ, A. (2004). A state space model for Berlin house prices: Estimation and economic interpretation. *The Journal of Real Estate Finance and Economics, 28*, 37-57.

SELIM, H. (2009). Determinants of house prices in Turkey: Hedonic regression versus artificial neural network. Expert Systems with Applications, 36(2), 2843–2852. https://doi.org/10.1016/j.eswa.2008.01.044.

SHILLER, R. J. (2008). Derivatives markets for home prices (No. w13962). National Bureau of Economic Research.

SHIM, J., BIN, O., & HWANG, C. (2014). Semiparametric spatial effects kernel minimum squared error model for predicting housing sales prices. Neurocomputing, 124, 81–88. https://doi.org/10.1016/j.neucom.2013.07.035.

TIERNO, R., CARVALHO, P. A., & MINISTÉRIO DAS CIDADES. (2007). O registro imobiliário: Conceitos e Bases Legais. PINHEIRO, OM et al. Acesso à terra urbanizada: implementação de planos diretores e regularização fundiária plena. Florianópolis: UFSC, 239-278.

TROY, A., & GROVE, J. M. (2008). Property values, parks, and crime: A hedonic analysis in Baltimore, MD. Landscape and Urban Planning, 87(3), 233–245. https://doi.org/10.1016/j.landurbplan.2008.06.005.

TSUTSUMI, M., SHIMADA, A., & MURAKAMI, D. (2011). Land price maps of Tokyo Metropolitan Area. Procedia - Social and Behavioral Sciences, 21, 193–202. https://doi.org/10.1016/j.sbspro.2011.07.046.

TYRVÄINEN, L. (1997). The amenity value of the urban forest: An application of the hedonic pricing method. Landscape and Urban Planning, 37(3–4), 211–222. https://doi.org/10.1016/S0169-2046(97)80005-9.

WACHTER, S. M., & GILLEN, K. C. (2006). Public investment strategies: How they matter for neighborhoods in Philadelphia. *unpublished report of the Wharton School of the University of Pennsylvania*.

WANG, C., LI, J., & GUO, P. (2015). The normalized interval regression model with outlier detection and its real-world application to house pricing problems. Fuzzy Sets and Systems, 274, 109–123. https://doi.org/10.1016/j.fss.2014.06.009.

WANG, W. K. (2005). A knowledge-based decision support system for measuring the performance of government real estate investment. Expert Systems with Applications, 29(4), 901–912. https://doi.org/10.1016/j.eswa.2005.06.017.

WOLVERTON, M. L. (1997). Empirical study of the relationship between residential lot price, size and view. *Journal of Property Valuation and Investment*, *15*(1), 48-57.

WU, H., JIAO, H., YU, Y., LI, Z., PENG, Z., LIU, L., & ZENG, Z. (2018). Influence Factors and Regression Model of Urban Housing Prices Based on Internet Open Access Data. Sustainability, 10(5), 1676. https://doi.org/10.3390/su10051676.

XU, T. (2017). The Relationship between Interest Rates, Income, GDP Growth and House Prices. Research in Economics and Management, 2(1), 30. https://doi.org/10.22158/rem.v2n1p30.

ZHANG, R., DU, Q., GENG, J., LIU, B., & HUANG, Y. (2015). An improved spatial error model for the mass appraisal of commercial real estate based on spatial analysis: Shenzhen as a case study. Habitat International, 46, 196–205. https://doi.org/10.1016/j.habitatint.2014.12.001.

ZHAO, Y., CHETTY, G., & TRAN, D. (2019). Deep learning with XGBoost for real estate appraisal. In 2019 IEEE symposium series on computational intelligence (SSCI) (pp. 1396-1401). IEEE.

ZHOU, Y., TIAN, Y., JIM, C. Y., LIU, X., LUAN, J., & YAN, M. (2022). Effects of public transport accessibility and property attributes on housing prices in Polycentric Beijing. *Sustainability*, *14*(22), 14743.

Anexos

Anexo 1: Artigo científico publicado nos Anais do XXXVIII Simpósio Brasileiro de Bancos de Dados.

Construção de Banco de Dados do Mercado Imobiliário. Um estudo na Cidade de São Paulo

Celso G. A. Ribeiro[1], Flavio F. Helena[1], Flavio A. M. Cipparrone[1]

[1]Universidade de Sao Paulo, Departamento de Sistemas Eletrônicos. Av. Prof. Luciano Gualberto, trav. 3, n 158, CEP 05508-900, São Paulo/SP, Brasil

celso.ribeiro@usp.br, flavio.helena@usp.br, prof.cipparrone@gmail.com

Abstract. *Obtaining and organizing reliable data plays a key role in understanding and analyzing the real estate market. In this study, a specific methodology for obtaining and processing data applied in the city of São Paulo is proposed, with the aim of overcoming existing data limitations and providing comprehensive and scalable information for an in-depth analysis of this constantly evolving sector. The availability of this complete and reliable database provides valuable insights for sellers and buyers, facilitating informed decision-making and enriching understanding of the real estate market through benchmarking and other relevant analysis.*

Resumo. *A obtenção e organização de dados confiáveis desempenham um papel fundamental na compreensão e análise do mercado imobiliário. Neste estudo, propõe-se uma metodologia específica de obtenção e tratamento de dados aplicada na cidade de São Paulo, com o objetivo de superar as limitações de dados existentes e fornecer informações abrangentes e escaláveis para uma análise aprofundada desse setor em constante evolução. A disponibilidade dessa base de dados completa e confiável proporciona insights valiosos para vendedores e compradores, facilitando a tomada de decisões embasadas e enriquecendo a compreensão do mercado imobiliário por meio de análises comparativas de preços e outras análises relevantes.*

A obtenção de dados confiáveis no mercado imobiliário é um desafio complexo e relevante. A escassez e precariedade das informações registradas nas prefeituras municipais, bem como as dificuldades de acesso aos Cartórios de Registros de Imóveis ou Receita Federal, contribuem para a falta de transparência e confiança nos dados desse setor (Abreu & Amorim, 2014).

Em alguns casos, os trabalhos de precificação enfrentam dificuldades na coleta de dados, recorrendo a pesquisas de campo e formulários manuais (Abreu & Amorim, 2014), o que limita o escopo da pesquisa. Outras vezes, usam bases virtuais, mas com restrições de tamanho e variáveis (Pinto & Fernandes, 2019; Paz & Nobre, 2020), diminuindo a assertividade. Segundo Mullainathan & Spiess (2017), overfit em análise de precificação é maior com amostras menores, reforçando a importância de bases de dados com número suficiente de amostras.

Diante dessas limitações, é crucial desenvolver uma abordagem sistemática para obter informações imobiliárias confiáveis. O objetivo deste estudo é estabelecer uma metodologia que viabilize a coleta de dados atualizados e abrangentes por meio de listagens imobiliárias, de forma coesa. Essa abordagem não se limita a dados intrínsecos, como características dos imóveis, mas inclui também dados extrínsecos, como elementos do ambiente em que o imóvel se encontra.

Para construir uma base de dados sólida, é necessário integrar diversas fontes de informação, selecionadas por aspectos relevantes. Essa metodologia envolve mineração, filtragem e modelagem de dados, garantindo informações confiáveis, essenciais para análises do mercado imobiliário. Nesse sentido, São Paulo destaca-se como exemplo ideal para consolidar o estudo proposto, usando anúncios online, informações sobre amenidades públicas, registros de imóveis e outras fontes na construção de uma base de dados abrangente.

Uma base de dados completa e confiável do mercado imobiliário fornece insights valiosos, facilitando decisões embasadas e permitindo análises comparativas das características dos imóveis e dos padrões de preço. Um exemplo é o estudo de caso de D'acci (2019), que investigou o impacto das variáveis extrínsecas no preço utilizando uma base de dados abrangente em Torino, ilustrando a importância do uso de base de dados confiável na compreensão do mercado imobiliário.

No contexto da pesquisa sobre obtenção de dados confiáveis no mercado imobiliário, a metodologia se destaca em relação aos estudos previamente explorados. Enquanto muitos lidam com restrições na coleta de dados, esta abordagem abrange várias fontes, integrando dados intrínsecos e extrínsecos por meio de listagens imobiliárias e informações públicas. Especificamente no cenário brasileiro, onde esta metodologia ainda não é amplamente usada, este estudo busca preencher essa lacuna, introduzindo uma perspectiva original e eficaz para servir de subsídio a análises do setor imobiliário.

Como discutido anteriormente, a existência de um banco de dados contendo as informações de características intrínsecas e extrínsecas de imóveis é um passo primordial para a construção de modelos que possam ajudar a analisar e explicar o comportamento do mercado imobiliário, com diversas aplicações para tais modelagens. O Brasil utiliza um sistema de registro de transações por meio Cartorial, onde as informações são

pertencentes ao poder público. (Tierno et al., 2007). Dessa forma, não há uma base de dados oficial disponibilizada de forma sistemática e com abrangência nacional que contenha os preços reais negociados pelos imóveis.

Dessa forma propõe-se a metodologia exemplificada a seguir, utilizando uma forma engenhosa de obter essas informações, por meio de anúncios imobiliários online. Nesse sentido, a cidade de São Paulo possui uma vasta base de dados com informações e dados sobre o "ambiente construído" (qualidade do pavimento, calçadas, arborização), sendo, portanto, o local escolhido para o desenvolvimento deste trabalho. Nesse sentido, seis distritos da cidade foram escolhidos, garantindo tanto a variabilidade da realidade socioeconômica quanto a representatividade estatística da amostra. Os distritos escolhidos foram: Moema, Pinheiros, Vila Mariana, Vila Carrão, Vila Andrade e Jaguaré. Cabe destacar que, neste estudo a Cidade de São Paulo foi escolhida, no entanto tal metodologia é replicável para outras localidades, na medida da disponibilidade de bases de dados com informações similares nestes locais.

Como próximo passo, a escolha das variáveis e fontes de dados também foi realizada. Aqui, objetivando uma representatividade do contexto em que o imóvel está inserido, foram selecionadas variáveis comumente associadas ao preço e à qualidade de um imóvel, tanto intrínsecas quanto extrínsecas. As variáveis foram provenientes de bases de dados distintas, onde diversas premissas para suas uniões foram adotadas. A Figura 1 ilustra essas bases e como foram consolidadas no banco de dados final.

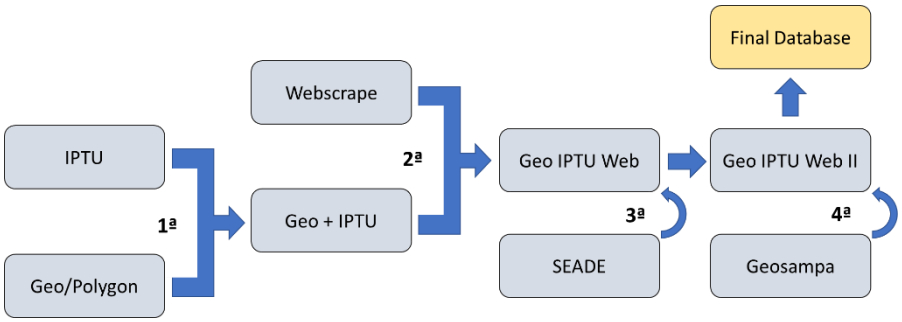

Figura 1. Processo de Junção das bases de dados

A Primeira Etapa da junção inicia-se com a "Base de dados IPTU", que reúne diversas informações das propriedades imobiliárias da cidade (ano de construção, número de pavimentos, área construída, etc.), disponibilizadas pela prefeitura na plataforma

GeoSampa[9]. À esta base são adicionadas as geolocalizações dos polígonos representativos dos imóveis (figuras geolocalizadas dos terrenos), através da junção com a "Base de dados Geo/Polígono". O critério de junção aqui adotado foi o uso do "número de contribuinte", campo que está presente nas duas bases de dados e pode ser usado, com restrições, como chave indexadora. Há uma particularidade a destacar: como já referido, existe um número de contribuinte relativo ao terreno e outro relativo ao imóvel. Quando um terreno contém apenas uma propriedade, ambos os números são iguais, mas se o terreno contém mais de um imóvel (edifícios, por exemplo), a situação muda. Consequentemente, para unir as duas bases de dados, o número do contribuinte é modificado na "Base de dados IPTU", removendo o código da unidade toda vez que for identificado que o registro é um apartamento, e substituindo pelo código representativo do terreno. Por fim, por meio de um *left join* utilizando esta 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.

A Segunda Etapa da junção envolve agregar à esta base de dados gerada anteriormente as informações de anúncios de imóveis online do site Viva Real[10]. Foi desenvolvido um código do tipo *webscrape* em linguagem *Python* para obter todos os anúncios das 100 primeiras páginas do site, repetindo o processo para cada um dos seis bairros escolhidos, reunindo todas as informações relevantes de cada anúncio (área, preço, número de quartos, etc.). Neste código foi adotado um filtro que seleciona apenas imóveis com endereços completos (Nome do Logradouro e Número do imóvel) para registro no banco de dados. Este filtro foi aplicado pois, utilizando este endereço completo e também a biblioteca *googlemaps* no *Python*, é possível obter o CEP do imóvel, o código do logradouro (CodLog), bem como sua Latitude e Longitude. Assim, com o CEP, CodLog e o Número do imóvel, usados como chave indexadora, pode-se fazer a composição dos dois bancos de dados obtidos até o momento.

Dessa forma, o banco de dados contendo todas as características intrínsecas do imóvel está pronto. Esse novo banco de dados será chamado de "Base de dados de Geo IPTU Web", passando para a Terceira Etapa do fluxo: agregar as informações sobre características extrínsecas disponíveis na "Base de dados SEADE". Esta base de dados

[9] O GeoSampa é o portal cartográfico oficial da Cidade de São Paulo e reflete a infraestrutura municipal em dados geográficos. A plataforma traz mais de 240 tipos de informações, como fotos aéreas, dados de equipamentos públicos, rede de transporte, etc. É a maior coleção de dados geoespaciais da cidade de SP.
[10] Viva Real é o maior portal online de anúncios imobiliários do Brasil. <https://www.vivareal.com.br/>.

proveniente da SEADE[11] reúne dados e informações referentes aos equipamentos de uso comum disponibilizados à população (pontos de ônibus, estações de metrô, escolas, faculdades, etc.), mais especificamente sua localização na cidade, com latitude e longitude. O processo de agregação dessas informações consiste em contar quantos pontos de cada variável existem nas proximidades de cada imóvel, considerando uma distância de até 500 metros do centróide do polígono do imóvel. Por exemplo, contar quantas Universidades Públicas estão a 500 metros de cada imóvel. Neste processo foi adotada a distância de *haversine*[12].

Essa mesma lógica também foi adotada para incluir informações sobre características extrínsecas da "Base de dados GeoSampa" (Quarta Etapa), listando quantos itens de suas variáveis estão nas proximidades dos imóveis. Existe apenas uma particularidade nesta base de dados que consiste no fato de que determinadas variáveis, como as relacionadas à Infraestrutura Cicloviária ou à Infraestrutura de Ônibus, são linhas geolocalizadas, contendo a latitude e longitude de seu contorno. Isso leva ao cálculo de quantas linhas existem nas proximidades de um determinado imóvel, em vez do cálculo de pontos. Além disso, uma vez que a geometria da linha está disponível na "Base de dados GeoSampa", seu comprimento total foi calculado, e essas informações foram agregadas ao cruzar os dois bancos de dados, finalizando a junção.

Após realizar todas as etapas descritas no Capítulo anterior, para junção de todas as informações selecionadas, o banco de dados resultante está completo. Contém 43 variáveis que representam as características de cada imóvel. São elas: *Preço/m², Área Construída, Quartos por Banheiros, Número de Garagens, Número de Esquinas, Fração Ideal, Área do Terreno, Área útil, Ano de Construção, Fator de obsolescência, Número de andares, Testada do terreno, Número de UBS, Número de CREAS, Número de CRAS, Número de Escolas Privadas, Número de Escolas Públicas Estaduais, Número de Escolas Públicas Municipais, Número de Escolas Públicas Federais, Número de Escolas (outras), Número de FATECs, Número de Universidades Particulares, Número de Universidades Públicas, Número de Museus, Número de unidades do Poupatempo,*

[11] SEADE (Fundação Sistema Estadual de Análise de Dados Estatísticos) vinculada ao Governo do Estado de SP, é referência nacional na produção e divulgação de análises e estatísticas socioeconômicas.
[12] A fórmula de Haversine determina a distância do superficial circular entre dois pontos em uma esfera, dadas suas longitudes e latitudes.

Número de Centros Populares, Número de Hospitais Públicos, Número de Hospitais Privados, Número Escalado de Consultórios Médicos, Número Escalado de Clínicas Médicas, Número Escalado de Reparos no Pavimento, Número Escalado de Árvores, Número de imóveis com acessibilidade nas proximidades, Número de Bicicletários Públicos, Número de Estações de metrô, Número de pontos de ônibus, Número ponderado de Ciclovias, Número ponderado de Ciclofaixas, Número ponderado de Ciclorrotas, Número ponderado de Faixas de ônibus Locais, Número ponderado de Faixas de ônibus Coletoras, Número ponderado de Faixas de ônibus Arteriais, Número ponderado de BRT.

A Base de dados resultante contém 4.019 registros, cada um correspondendo a um único Imóvel e reunindo todas estas informações descritas. Vale ressaltar que o processo de *webscraping* do site VivaReal reuniu 5.479 registros na base de dados inicial (Base de dados Webscrape) e, após todas as regras, validações e limpeza dos dados para junção dessas bases totalmente distintas, restaram 4.019 registros, resultando em um índice de assertividade para o processo de junção de cerca de 73%.

3.1. Breve Análise Exploratória de Dados.

A partir do banco de dados resultante, abre-se um leque para realização de diversas análises ligadas ao mercado imobiliário que não são o escopo deste trabalho. No entanto, apenas para ilustrar parte dessas aplicações, realizou-se aqui uma breve análise exploratória, ilustrada com a Figura 2, que é uma representação visual desses 4.019 imóveis, geolocalizados no mapa da cidade de São Paulo. As cores do mapa representam a variação do preço por metro quadrado em todos os imóveis, em uma escala verde-vermelha, onde o verde representa os preços mais baixos e o vermelho os mais altos

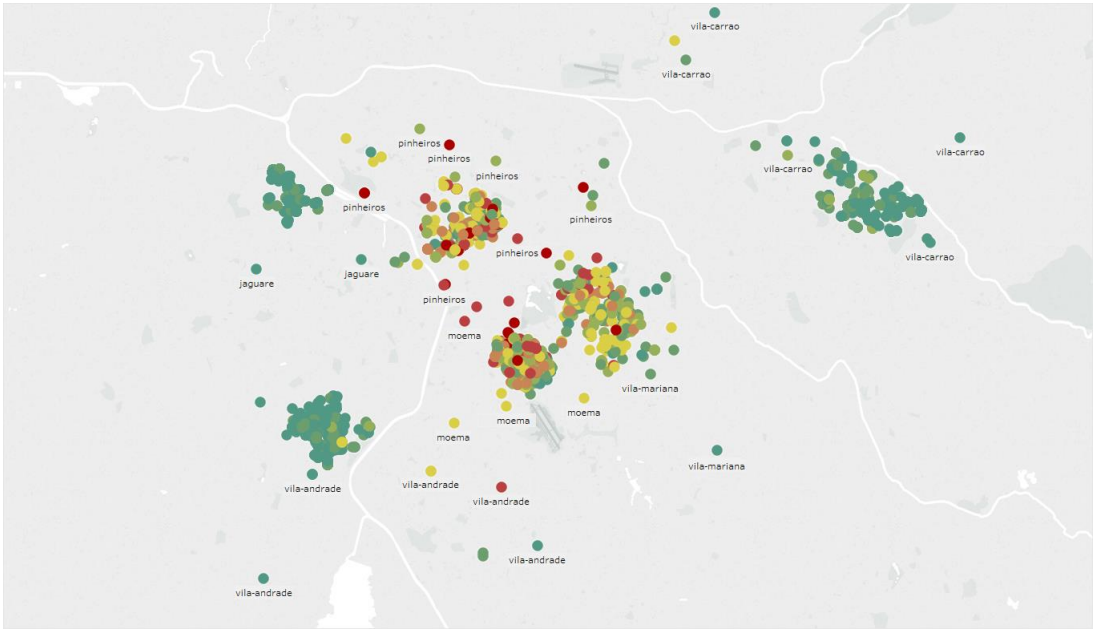

Figura 2. Propriedades Imobiliárias do banco de dados, geolocalizadas no mapa da cidade de São Paulo. A cor representa a variação do preço por metro quadrado (vermelho é alto, verde é baixo).

Nota-se, ao observar a Figura 2, que os bairros mais centralizados da cidade possuem preços/metro quadrado mais elevados, enquanto os bairros mais distantes possuem preços mais acessíveis. Essa tendência de variação de preços em função da distância do centro da cidade já foi abordada por D'acci (2019), citado anteriormente neste trabalho. Na literatura, é possível encontrar uma grande quantidade de trabalhos mostrando os efeitos de uma gama de características extrínsecas na avaliação de imóveis. Cervero & Kang (2011), por exemplo, mostraram que o transporte público, especificamente Bus Rapid Transit (BRT), oferece prêmios de até 10% para residências dentro de 300 metros de paradas de BRT e mais de 25% para varejo e outros usos não residenciais, dentro de um raio de 150 metros de distância. Troy et al. (2008) mostraram que quando a taxa de criminalidade é relativamente baixa, os parques têm um impacto positivo nos valores das propriedades. Estes trabalhos demonstram a utilidade de um banco de dados reunindo características relacionadas a imóveis como um ponto de partida para a identificação dos comportamentos do mercado imobiliário, assim como o banco de dados consolidado neste presente estudo.

A partir do que foi discutido neste artigo, conclui-se que é possível criar uma base de dados unindo diversas características intrínsecas e extrínsecas de propriedades imobiliárias, a partir de anúncios disponibilizados em sites de compra e venda imobiliária e de informações geolocalizadas de características do ambiente urbano. Como limitação

deste trabalho, cabe ressaltar que este processo foi feito apenas para a cidade de São Paulo, com dados de 2021, sendo dependente dos valores inseridos pelos usuários dos sites de anúncio, sujeitos a erros. Trabalhos futuros podem tanto expandir a metodologia usada para a construção do banco de dados aqui proposta, assim como utilizar a base de dados na modelagem de preço regional de propriedades imobiliárias.

Abreu, M. A., & Amorim, W. V. (2014). O estudo do mercado imobiliário em cidades médias: procedimentos para coleta e sistematização dos dados. Geo UERJ, 2(25), 297-323.

Boulic, R. and Renault, O. (1991) “3D Hierarchies for Animation”, In: New Trends in Animation and Visualization, Edited by Nadia Magnenat-Thalmann and Daniel Thalmann, John Wiley & Sons ltd., England.

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.

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.

Mullainathan, S., & Spiess, J. (2017). Machine learning: an applied econometric approach. Journal of Economic Perspectives, 31(2), 87-106.Pinto, V. H. L., & Fernandes, R. A. S. (2019). Análise de preços hedônicos no mercado imobiliário residencial de Conselheiro Lafaiete, MG. Interações (Campo Grande), 20, 627-643.

Paz, R.R., Nobre, L.H., & Nobre, F.C. (2020). Determinantes De Preços No Mercado Imobiliário À Luz Do Modelo Hedônico. Revista Gestão em Análise, 9, 60.

Tierno, R., Carvalho, P. A., & MINISTÉRIO DAS CIDADES. (2007). O registro imobiliário: Conceitos e Bases Legais. PINHEIRO, OM et al. Acesso à terra urbanizada: implementação de planos diretores e regularização fundiária plena. Florianópolis: UFSC, 239-278.

Troy, A., & Grove, J. M. (2008a). Property values, parks, and crime: A hedonic analysis in Baltimore, MD. Landscape and Urban Planning, 87(3), 233–245.