SemeAd 2024
XXVII Seminários em Administração

05, 06, 07 e 08 de novembro de 2024
ISSN 2177-3866

Machine Learning e a Precificação de Imóveis: um Estudo Comparativo entre Modelos de Preços Hedônicos

WILLIAM VIANA BORGES
UNIVERSIDADE FEDERAL DO PARANÁ (UFPR)

RAFAEL BUTTINI SALVIATO
UNIVERSIDADE FEDERAL DO PARANÁ (UFPR)

THIAGO HENRIQUE MOREIRA GOES
UNIVERSIDADE FEDERAL DO PARANÁ (UFPR)

Machine Learning e a Precificação de Imóveis: um Estudo Comparativo entre Modelos de Preços Hedônicos

Resumo
O presente trabalho tem como objetivo avaliação e a comparação de ferramentas de auxílio na precificação de ativos imobiliários. A partir de dados referentes aos valores recolhidos do Imposto sobre Transmissão de Bens Imóveis (ITBI), foram elaborados seis modelos hedônicos que se utilizam de métodos de *machine learnig* (XGBoost, SVM Kernel Linear, *Random Forest*, Rede Neural, Regressão Linear e GLM Gama). Estes modelos foram avaliados posteriormente pelos métodos MAE, MAPE, RMSE e BIAS, entre os modelos que apresentaram os melhores ajustes estão o XGBoost e SVM Kernel Linear, já o pior ajuste identificado foi o do modelo GLM Gama.

Palavras-chave: precificação de imóveis, modelos de preços hedônicos, *machine learning*, ITBI.

Abstract

This study aims to evaluate and compare tools for pricing real estate assets. Using data from the Imposto sobre Transmissão de Bens Imóveis (ITBI), six hedonic models incorporating machine learning methods (XGBoost, Linear SVM, Random Forest, Neural Network, Linear Regression, and GLM Gamma) were developed. These models were assessed using MAE, MAPE, RMSE, and BIAS metrics. The XGBoost and Linear SVM models demonstrated the best performance, while the GLM Gamma model showed the poorest results.

Keywords: real estate pricing, hedonic pricing models, machine learning, ITBI.

1. Introdução

A avaliação imobiliária é um fenómeno complexo e multifacetado, influenciado por princípios microeconómicos básicos. A análise dos determinantes individuais do mercado imobiliário enfatiza a importância dos conceitos de oferta e demanda na determinação dos preços (VARIAN, 2003). A escassez relativa aliada às flutuações na demanda por imóveis em diferentes localidades e com diferentes atributos é essencial para a compreensão da formação de preços no setor imobiliário. Além disso, teorias como a utilidade marginal fornecem uma estrutura conceitual para compreender como os consumidores valorizam características específicas de uma propriedade como tamanho, localização e comodidades, o que afeta diretamente no seu preço (PINDYCK; RUBINFELD, 2013). A interação entre escolhas individuais, custos de oportunidade e maximização da utilidade pessoal desempenha um papel fundamental na dinâmica de precificação do mercado imobiliário, aspectos que se baseiam em grande parte em princípios e teorias microeconômicas.

A precificação de ativos imobiliários urbanos é objeto de estudo há algum tempo, desde a discussão teórica à pesquisa empírica (SHEPPARD, 1999). A precificação de imóveis é de grande importância para a sociedade como um todo, pois permite às famílias escolherem com maior grau de assertividade a hora de comprar ou vender o seu imóvel. Auxilia incorporadoras e construtoras a avaliar com maior precisão a viabilidade econômico-financeira de empreendimentos imobiliários. Ajuda na determinação da base de cálculo de impostos patrimoniais, como o Imposto sobre Transmissão de Bens Imóveis (ITBI) e o Imposto Predial

e Territorial Urbano (IPTU), bem como a permitir a elaboração de estudos econômicos aos entes governamentais.

Na literatura, existe uma grande diversidade de modelos para precificação de imóveis, muitos destes modelos utilizam-se da técnica de *web scraping*, técnica que pode ser definida como uma "raspagem" de dados da *web* a partir do uso de *bots* (BARBOSA; CAVALCANTI, 2020). Os dados dos modelos que se utilizam da técnica de *web scraping* são geralmente provenientes de sites de anúncios de imóveis, o que pode gerar inconsistência em relação ao seu valor real de compra/venda, pois muitos dos preços anunciados podem estar defasados ou simplesmente não correspondem ao valor de fechamento do negócio. Uma alternativa para contornar esta problemática consiste no uso de dados referentes ao ITBI, pois o valor deste imposto é calculado com base nos valores de compra/venda registrados no cartório de registro de imóveis.

Dos modelos propostos com dados do ITBI, observou-se que estes trabalhos aplicam diversos métodos na geração de suas inferências, alguns dos métodos explorados: regressão linear, meta algoritmo *bagging*, meta algoritmo *boosting*, redes neurais artificiais, *general linear model* (GLM). Contudo, não foram identificados trabalhos que aplicam os métodos de *machine learning* XGBoost e SVM Kernel Linear, métodos estes que foram apontados em alguns estudos como detentores de uma alta acurácia para a precificação de imóveis (i.e: ZAKI; NAYYAR.; DALAL; ALI, 2022; YOO; IM; WAGNER, 2012).

Este trabalho tem como objetivo a avaliação e a comparação de modelos para precificação de imóveis a partir de métodos de *machine learning*. Para atingir este objetivo, foram elaborados seis modelos hedônicos que se utilizam dos métodos XGBoost, SVM Kernel Linear, Regressão Linear, *Random Forest*, Rede Neural e GLM Gama. A aplicação destes modelos se deu em um contexto real, cujos dados foram fornecidos no Relatório do ITBI, disponibilizado mensalmente pela Prefeitura Municipal de Belo Horizonte. A avaliação e a comparação foram feitas seguindo quatro métricas de desempenho (MAE, MAPE, RMSE e BIAS). Os resultados apontam uma alta acurácia para os modelos XGBoost, SVM Kernel Linear, e uma baixa acurácia para o modelo GLM Gama.

2. Referencial Teórico

A literatura que trata sobre abordagens e modelos para a precificação de imóveis é vasta e bastante diversificada, contudo, conforme algumas características que estas abordagens e modelos possuem em comum, é possível subdividi-la e classificá-la em alguns grupos. Adetiloye e Eke (2014) classificaram estes modelos e abordagens da seguinte forma: (1) Abordagem comparativa de mercado ou vendas; (2) Abordagem do custo; (3) Abordagem da capitalização de renda; (4) Modelos de preços hedônicos; (5) Preços de transação médios ou medianos; (6) Método de repetição de vendas; (7) Métodos híbridos; (8) Modelo de opções; (9) Modelo de rendimento equivalente; (10) Crescimento periódico ou Modelo de rendimento equacionado.

Dentre os modelos e abordagens apresentados por Adetiloye e Eke (2014), destacam-se os modelos de preços hedônicos (HPM). Propostos por Court (1939), os HPM são de natureza estatística e baseados na regressão multivariada. O modelo relaciona a variável dependente “preço de venda” e as variáveis independentes "com base nos atributos de quantidade e qualidade de uma grande amostra de imóveis semelhantes. As amostras são frequentemente selecionadas usando uma combinação de distância geográfica e alguns atributos-chave, como o tamanho do bem" (MILLER; GELTNER, 2005, p. 1886). Na Equação 1, é demonstrada a função que define a abordagem clássica do HPM segundo Herath e Maier (2010).

$$P \approx F(A, L, B, C, t) \quad (1)$$

Em que a variável "P" corresponde ao preço do imóvel, já a variável explicativa "A" diz respeito aos atributos do imóvel (tamanho, ano de construção, entre outros), "L" é a sua localização, "C" e "B" são as características da vizinhança e do bairro respectivamente, e "t" é um indicador de tempo.

O estimador de Mínimos Quadrados Ordinários (MQO) é o mais explorado nos HPM, os modelos elaborados com base no MQO são conhecidos também como regressão hedônica clássica. Com o avanço da capacidade computacional e o surgimento de métodos de *machine learning*, novos estudos foram publicados empregando tais métodos. Alguns estudos internacionais que aplicaram algoritmos de *machine learning*, apresentaram uma alta acurácia se comparados à regressão hedônica clássica. Entre estes estudos, pode-se mencionar o trabalho de Zaki *et. al.* (2022), em que os autores empregaram o método XGBoost e da regressão hedônica clássica, na comparação dos resultados foi exposto que o modelo que incorpora o XGBoost apresentou uma acurácia de 84,1%, já o modelo hedônico clássico 42%. Na pesquisa de Yoo, Im e Wagner (2012), os autores aplicaram os métodos *Cubist, Random Forest* (RF) e a regressão hedônica clássica na elaboração de modelos de precificação de imóveis, na comparação entre os modelos foi identificado uma maior acurácia do RF, seguido pelo *Cubist* e por último a regressão hedônica clássica.

A pesquisa bibliográfica que guiou este artigo, foi elaborada a partir da análise de artigos científicos nacionais expostos na ferramenta *Google Scholar* e ordenados por relevância, foram realizadas combinações dos seguintes termos de busca: "ITBI", "precificação de imóveis", "preços médios de imóveis" e "modelos de preços hedônicos". Dentre os modelos identificados nesta busca, cujos dados advém de relatórios do ITBI, pode-se constatar uma vasta gama de trabalhos que empregam a regressão hedônica clássica. Entre estes trabalhos está o estudo de Furtado (2011), o autor buscou analisar a influência espacial do tecido urbano na formação dos preços dos imóveis. Para isso, foi aplicado o método dos Mínimos Quadrados Ordinários (MQO) para construir um modelo de regressão, do qual o log do preço do imóvel é a variável dependente, as características do imóvel e de sua localização correspondem às variáveis independentes.

Sobre os trabalhos em que os autores aplicam algum método de *machine learning* para a construção de modelos de precificação de imóveis, consta o trabalho de Florencio (2010), em que o autor se utilizou do método *Generalized Additive Model for Location, Scale and Shape* (GAMLSS) para explicar a variabilidade de preços de imóveis, cuja variável dependente corresponde ao preço unitário do terreno e as variáveis independentes dizem respeito à característica estruturais, econômicas e locacionais. Oliveira, Bandeira e Silva (2022) conduziram uma pesquisa comparativa entre os métodos *Decision Tree* (DT) e *Bagging*, *Random Forest* (RF) e Gradient Boosting Regression empregados na valoração do solo urbano. Já Silva (2019), aplicando metodologias de aprendizagem computacional assistida, propôs diversos modelos e os comparou. Entre os modelos propostos, foram utilizados os métodos de regressão linear, regressão gaussiana, *random forest, gradient boosting machine* e redes neurais artificiais. Utilizando-se também de redes neurais, Veras (2019) elaborou um modelo univariado com base em rede neural recorrente do tipo *Long Short Term Memory* (LSTM), o autor pode projetar os preços médios dos imóveis nos anos seguintes a partir da variável explicativa "preço dos imóveis".

3. Metodologia

3.1 Coleta e tratamento dos dados

Os dados empregados neste estudo, juntamente com o dicionário das variáveis, são disponibilizados pela Secretaria Municipal de Fazenda de Belo Horizonte (SFMA) no portal

Dados Abertos. Na busca por dados de outros municípios, foi identificado que apenas o município de Belo Horizonte pública os dados referentes ao ITBI. Foram feitas solicitações à alguns municípios para que disponibilizassem os seus dados, a resposta mais recorrente é a de que o acesso ao banco de dados é vedado por conta de dispositivos da Lei 5172/1966 - CTN e Lei 13709/2018, estes dispositivos tratam sobre o sigilo tributário e a proteção de dados.

Para a realização deste estudo, a robustez de um conjunto de dados é requisito para a condução dos testes de avaliação dos modelos (MAE, MAPE, RMSE e BIAS), a base de dados do Município de Belo Horizonte satisfaz este requisito. Por conta disso, os dados utilizados neste estudo correspondem somente ao município de Belo Horizonte, tais dados estão tabulados em uma planilha eletrônica com as seguintes informações sobre os imóveis: endereço; bairro; ano de construção; área total do terreno; área construída adquirida; área adquirida (unidades somadas); fração ideal adquirida; padrão de acabamento (unidade); tipo construtivo preponderante; descrição do tipo de ocupação (unidade); valor base cálculo; zona urbana; e data da inclusão da transação.

Para este trabalho foram coletados os dados mensais entre os meses de janeiro de 2022 a fevereiro de 2023, para cada mês estava disponível uma planilha eletrônica. Por meio do software R Studio estas planilhas foram agrupadas em um uma única base de dados, gerando um total de 27.985 observações. Este estudo é voltado apenas para os imóveis do tipo "apartamento", portanto, aplicou-se um filtro para esta variável, resultando em 19.632 observações após a exclusão de NA 's.

3.2 Modelo teórico e variáveis

Seguindo a abordagem clássica do HPM apresentada na Equação 1, foi elaborado o modelo teórico que serve de base na construção dos modelos empíricos deste artigo, a Equação 2 apresenta a função dos modelos em questão.

$$\frac{P}{m^2} \approx F(B, I, A, Z) \tag{2}$$

Onde $\frac{P}{m^2}$ corresponde ao preço comercializado do imóvel por metro quadrado, *B* é o bairro onde está localizado o imóvel, *I* é o ano de construção da unidade, *A* é o padrão de acabamento e *Z* é a zona de uso. A variável resposta $\frac{P}{m^2}$ foi atualizada utilizando o Índice Nacional de Custo da Construção (INCC) para o mês base agosto/2023. Na Tabela 1, são apresentadas as variáveis que incorporam os modelos deste trabalho.

Tabela 1 – Variáveis dos modelos

Variável	Descrição
$\boldsymbol{\frac{P}{m^2}}$	Variável contínua, expressa o valor do metro quadrado
B	Variável nominal, corresponde ao bairro onde o imóvel está localizado
I	Variável discreta, ano de construção do imóvel
A	Variável ordinal, define o padrão de acabamento do imóvel (P1, P2, P3, P4 e P5), sendo que P1 é o padrão mais baixo e P5 o mais alto
Z	Variável nominal, diz respeito à zona de uso do imóvel, definida conforme o planejamento urbano do município

Fonte: Elaborado pelos autores, baseado no dicionário das variáveis (SFMA, 2023).

A construção das variáveis expressas na Tabela 1 se deu a partir das variáveis contidas nos dados publicados pela SFMA, a variável $"\frac{P}{m^2}"$ é resultado da divisão entre as variáveis "valor base cálculo" e "área construída adquirida". Já a variável "*B"* foi trabalhada com *dummies*, assim como as variáveis "*A"* e "Z". A variável "I" não sofreu qualquer tipo de intervenção. Na próxima seção são apresentados os métodos utilizados na elaboração dos modelos empíricos.

3.3 Métodos empregados

Nesta seção serão descritos os métodos empregados no presente estudo, bem como as suas origens e recomendações de uso. Todos os métodos foram implementados por meio do *software* estatístico R Studio. No que diz respeito ao processo de modelagem, para evitar sobreajuste do modelo nos dados foi realizado um processo de validação cruzada (REFAEILZADEH, 2016, p. 978). Este processo é muito comum quando se busca selecionar o melhor modelo de aprendizado de máquina dentre uma gama de modelos possíveis (e.g: LITTLE; VAROQUAUX; SAEB; LONINI, 2017; RODRIGUEZ; PEREZ; LOZANO, 2010).

O primeiro método denominado XGBoost, proposto por Chen (2016), é um algoritmo de aprendizado de máquina muito conhecido no âmbito da ciência de dados. Ele é inspirado em um algoritmo chamado *gradient boosting*, sua utilização compreende tanto a regressão quanto classificação, sendo o mais indicado para conjunto de dados que apresentam alta dimensionalidade (i.e: um grande conjunto de covariáveis) e discrepância. Além dos exemplos citados na introdução deste artigo, Zhao, Chetty e Tran (2019) também traz um exemplo onde o XGBoost enriquece a acurácia na precificação de imóveis quando acoplado em um modelo de aprendizagem profunda, que representa um contexto de alta dimensionalidade.

O segundo método é a máquina de vetor suporte, ou como geralmente é conhecido, *support vector machine* (SVM). Introduzido por Cortes e Vapnik (1995), o SVM é sugerido na redução da dimensionalidade de um conjunto de dados, em um contexto de regressão é o mais indicado em pequenas amostras e alta dimensionalidade. Wang, Li e Zhao (2008), além de Cao, Zhan e Wu (2009) mostram o potencial desse tipo de modelo tanto para predição quanto para velocidade de ajuste em contextos de mercado imobiliário, de crédito e mercado acionário. Existem também algumas adaptações do algoritmo SVM, incluindo algoritmos de otimização para ajustar os seus parâmetros de modo a minimizar o erro das predições. Pode-se encontrar o uso destes modelos no mercado imobiliário em Gu, Zhu e Jiang (2011) e Wang, Wen; Zhang e Wang (2014).

O terceiro método, chamado de *random forest*, foi desenvolvido por Breiman (2001) e baseado no modelo de regressão de árvore denominado *classification and regression tree* (CART). O *random forest* também é indicado para dados que apresentam valores discrepantes e com forte ruído (i.e: variabilidade). Embora haja evidências que não favoreçam o uso do modelo *random forest* para precificação de imóveis como em Zeng (2021), ainda assim é um método de aprendizado de máquina muito utilizado para realizar a precificação de imóveis (e.g: ZHANG; HUANG; ZHANG; LIU; SHORMAN, 2022; SRI; REDDY; KUMAR; VINOD; REDDY, 2023; HU; CHUN; GRIFFITH, 2022).

A regressão linear talvez seja o método estatístico mais difundido nas mais diversas áreas das ciências. É um método intrinsecamente probabilístico, que descreve a variável resposta como sendo uma variável aleatória condicionada a outra(s) variável(eis) aleatória(s) ou não. A sua origem data do século XIX e foi concebida por vários estudiosos da época. Merecem destaque Galton (1886) e Pearson (1892). Por conta de sua difusão e praticidade de uso, o modelo de regressão continua sendo muito utilizado na atualidade, e com exemplos de uso em estudos relacionados ao escopo do presente trabalho (e.g: GHOSALKAR; DHAGE,

2018; AMARESH; SINGH; KAMAL; KULKARNI, 2022; WANG; CHEN; FAKIEH; ALHAMAMI, 2021).

O quinto método é a rede neural artificial, um modelo amplamente difundido na área de ciência de dados, com inspiração na biologia do sistema nervoso humano. O modelo consiste em uma sequência de camadas por onde passam os dados de entrada até chegar numa camada de saída com a resposta predita pelo modelo. A primeira aparição foi através do trabalho de Rosenblatt (1958) e o uso do modelo é recomendado para contextos em que a relação entre a variável resposta e as covariáveis não é linear ou é muito complexa. De forma mais específica, o presente trabalho fez o uso de uma rede neural artificial *feedforward*. São inúmeros os trabalhos que fizeram uso de redes neurais artificiais no contexto do mercado imobiliário (e.g: AL-SHAYEA, 2012; LEE; RYU, 2021; ĆETKOVIĆ; LAKIĆ; LAZAREVSKA; ŽARKOVIĆ; VUJOŠEVIĆ; CVIJOVIĆ; GOGIĆ, 2018). Em alguns contextos a rede neural pode trazer resultados superiores se comparados a outras metodologias como o SVM e o ARIMA (e.g: ABIDOYE; CHAN; ABIDOYE; OSHODI, 2019).

Por último, tem-se os modelos lineares generalizados (GLM) propostos por Nelder e Wedderburn (1972). Os GLM consistem em uma extensão do modelo de regressão linear, que permite modelar dados cuja resposta condicionada às covariáveis não segue necessariamente uma distribuição normal (mas que ainda assim, a distribuição da variável resposta condicionada às covariáveis segue uma distribuição que pertence à família exponencial de distribuição de probabilidades). No presente trabalho, aplicou-se o GLM com resposta Gama, similar ao que foi empregado por Zewotir, Bax e North (2019).

3.4 Métodos de avaliação dos modelos

A seguir, são apresentadas as métricas utilizadas para avaliar os modelos do presente trabalho. No total são quatro, amplamente difundidas no universo da ciência de dados. Conforme estas métricas, é possível avaliar a distância entre o real e o previsto de cada modelo, bem como se o modelo em si apresenta viés ou não. O erro absoluto médio (MAE) representa a média do valor absoluto dos erros do modelo. Quanto menor o MAE, melhor é o ajuste do modelo. Na Equação 3 é apresentada a sua fórmula.

$$MAE = \frac{\sum_{i=1}^{n} |p_i - r_i|}{n} \quad (3)$$

Onde *n* é o número de observações, p_i é o valor i-ésimo predito e r_i é o i-ésimo valor realizado. Já o erro percentual absoluto médio (MAPE) é muito similar ao MAE, porém, ao invés da média dos erros absolutos, o MAPE representa a média dos erros expressos em percentuais relativos ao valor realizado conforme a Equação 4.

$$MAPE = \frac{\sum_{i=1}^{n} \left| \frac{p_i - r_i}{r_i} \right|}{n} \quad (4)$$

Em que *n* é o número de observações, p_i é o valor i-ésimo predito e r_i é o i-ésimo valor realizado. Assim como o MAE, quanto menor o MAPE, melhor o ajuste do modelo. O erro quadrático médio (RMSE) também é uma medida de erro, só que ao invés de realizar a média dos erros absolutos, ele corresponde a raiz da média do quadrado dos erros. A sua fórmula é expressa na Equação 5.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (p_i - r_i)^2}{n}} \quad (5)$$

Onde *n* é corresponde ao número de observações, p_i é o valor i-ésimo predito e r_i é o i-ésimo valor realizado. A medida de viés (ou como é mundialmente conhecida, *bias*) mede a diferença entre a média dos valores reais e a média dos valores previstos pelo modelo. Ela pode ser representada pela Equação 6.

$$BIAS = \underline{r_i} - \underline{p_i} \tag{6}$$

Em que $\underline{r_i}$ é a média dos valores reais e $\underline{p_i}$ é a média dos valores previstos.

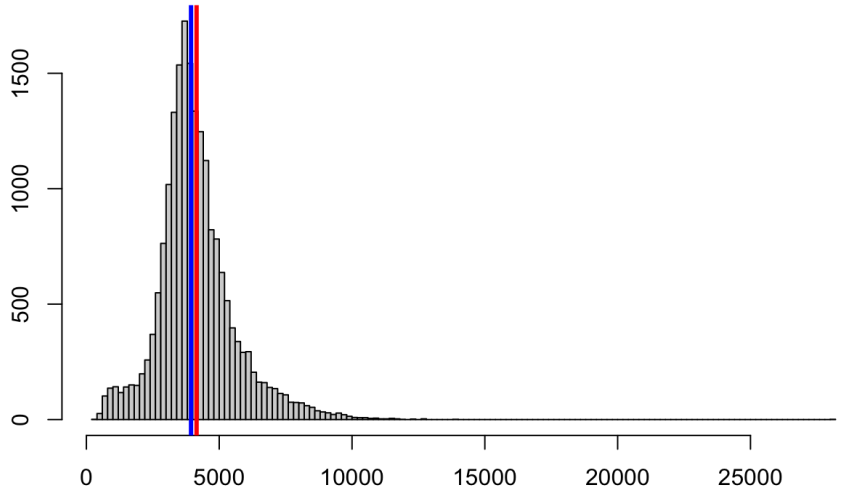
4. Resultados e Discussões

A seguir, são expostos os resultados obtidos no presente estudo. O conjunto de dados possui cerca de dezenove mil observações, cujos imóveis foram transacionados entre janeiro de 2022 e fevereiro de 2023 no município de Belo Horizonte, MG.

4.1 Análise exploratória

Na Figura 1, tem-se a distribuição do valor do metro quadrado, a linha vermelha representa a média, já a linha azul representa a mediana. Nota-se que, apesar dos valores extremos, os dados possuem certa simetria em sua distribuição.

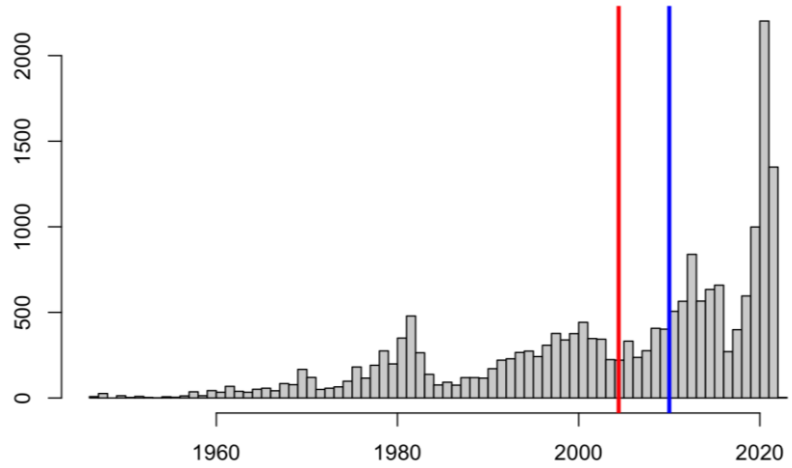
Figura 1 - Histograma do valor corrigido por metro quadrado



Fonte: resultados da pesquisa.

No que diz respeito à covariável "ano de construção da unidade", o conjunto de dados expõe imóveis construídos entre os anos 1946 e 2023. A maior concentração de imóveis transacionados da base foi construída nos anos de 2020, 2021 e 2022. A Figura 2 permite visualizar a distribuição dos dados, onde é possível ver a assimetria presente no ano de construção da unidade. Assim como no Figura 1, a linha vermelha representa a média e a linha azul representa a mediana.

Figura 2 - Histograma do ano de construção da unidade



Fonte: resultados da pesquisa.

Sobre as variáveis padrão de acabamento e zona de uso, a Figura 3 ilustra um mapa de calor, que descreve a relação entre as duas variáveis, e a variável resposta. Sobre o padrão de acabamento, é possível identificar que a média do valor do metro quadrado aumenta ao longo das categorias P1, P2, P3, P4 e P5. Também é possível identificar que essa mesma média do valor por metro quadrado diminui na zona de uso, do grupo de categorias ZPAM, ZCBH e ZP3, para o grupo ZAR1, ZA, ZP2, e para o restante dos outros tipos de zona. Por último, o mapa de calor mostra, ainda que a olho nu e sem nenhum teste formal, que existe uma certa relação entre padrão de acabamento e zona de uso.

Figura 3 - Mapa de calor do valor corrigido por metro quadrado, por zona de uso e padrão de acabamento da unidade.

		Padrão de acabamento da unidade					
		P1	P2	P3	P4	P5	Totals
Zona de uso (ITBI)	ZE		R$ 2.866,25	R$ 3.569,58	R$ 4.883,65	R$ 5.153,97	R$ 3.471,13
	ZCVN		R$ 2.807,43	R$ 3.526,24			R$ 3.501,46
	ZP1		R$ 3.162,93	R$ 3.601,81	R$ 3.436,32		R$ 3.531,96
	ZESFR		R$ 3.051,10	R$ 4.463,23			R$ 3.656,30
	ZAP	R$ 3.326,14	R$ 3.130,37	R$ 3.703,64	R$ 4.078,78	R$ 4.994,12	R$ 3.735,08
	ZCBA			R$ 3.871,65	R$ 3.372,31		R$ 3.778,02
	ZHIP		R$ 3.494,02	R$ 3.828,41	R$ 4.523,40		R$ 3.822,39
	ZAR2	R$ 2.746,01	R$ 3.085,41	R$ 3.726,57	R$ 4.278,32	R$ 5.201,89	R$ 3.855,71
	ZAR1		R$ 3.702,90	R$ 3.959,04	R$ 4.813,83	R$ 5.888,62	R$ 4.461,22
	ZA	R$ 3.375,54	R$ 3.679,33	R$ 3.938,14	R$ 4.870,66	R$ 6.524,28	R$ 4.495,91
	ZP2		R$ 4.222,68	R$ 4.833,71			R$ 4.735,16
	ZPAM				R$ 6.821,71	R$ 5.622,67	R$ 6.102,28
	ZCBH		R$ 4.440,79	R$ 5.245,14	R$ 6.345,69	R$ 7.554,90	R$ 6.256,48
	ZP3				R$ 5.537,30	R$ 6.765,92	R$ 6.321,07
	Total	R$ 3.124,58	R$ 3.238,93	R$ 3.813,68	R$ 4.718,78	R$ 6.396,50	R$ 4.147,42

Fonte: resultados da pesquisa.

Sobre a variável dos bairros, cabe mencionar que o conjunto de dados apresentou cerca de 231 bairros. Por isso, em vários modelos utilizados no estudo, a dimensionalidade ficou consideravelmente alta, pois se tem pelo menos 230 variáveis dicotômicas a serem imputadas no modelo. Dos bairros com as maiores médias de valor por metro quadrado, merecem destaque os bairros: Jardim Atlântico, Savassi, Belvedere, Funcionários e Santo Agostinho.

4.2 Resultados da modelagem

Conforme exposto anteriormente, foram ajustados seis modelos de aprendizado de máquina ao conjunto de dados, e seus respectivos desempenhos foram avaliados sob a ótica de quatro métricas de avaliação de modelos, amplamente difundidas no universo da ciência de dados. A Tabela 2 traz alguns resultados deste estudo, ordenando por linha do melhor MAPE até o pior.

Tabela 2 - Métricas dos modelos calibrados.

Modelo	MAPE (%)	RMSE (R$)	MAE (R$)	BIAS (R$)
XGBoost	0,1336	97,9468	4,9284	-0,5507
SVM Kernel Linear	0,7948	47,7806	24,2267	24,0169
Regressão Linear	24,9498	1.009,9592	686,7192	-1,1411
Random Forest	26,1194	1.064,1411	739,9533	-72,1434
Rede Neural	27,3023	1.103,4192	775,8395	-43,2840
GLM Gama	100,0000	4.106,9136	3.873,0755	-3.873,0755

Fonte: resultados da pesquisa.

Sobre a interpretação dos resultados (lembrando que a nossa variável resposta é o valor do metro quadrado), as unidades de medida estão em reais com exceção do MAPE que está em percentual. O RMSE nos mostra a média das distâncias entre o predito e o realizado, o MAPE e o MAE nos mostram a média do erro entre o previsto e o realizado, tanto em percentual quanto em valores absolutos (respectivamente). O BIAS representa em média, quanto que a previsão está acima ou abaixo do valor real.

Entre os principais resultados, o XGBoost e o SVM de Kernel Linear apresentaram os melhores ajustes de acordo com as métricas expostas. De fato, tais modelos são indicados para esse tipo de conjunto de dados, haja vista a alta dimensionalidade dos dados por conta do grande número de bairros e zonas de uso. Estes resultados estão em consonância com as descobertas documentadas por Chen (2016), Zhao et al. (2019) e Cortes e Vapnik (1995).

O desempenho negativo do GLM com resposta Gama também chama a atenção, pois este modelo respondeu como sendo zero praticamente todas as predições de valor do metro quadrado. Vale ressaltar que, nenhum modelo passou por uma otimização de hiperparâmetros. Logo, é natural que o próximo passo deste estudo seja calibrar os parâmetros de cada modelo para extrair a melhor performance de cada ferramenta.

5. Conclusões

Sobre a comparação entre o modelo teórico proposto e o modelo hedônico clássico, cabe ressaltar que foram utilizadas *proxies* para as variáveis do modelo hedônico clássico na construção do modelo teórico proposto. A única variável que não foi contemplada diz respeito às "características da vizinha", esta variável é abordada em outros modelos como o nível de segurança ou violência da região em que o imóvel está localizado. Como sugestão para futuras pesquisas, pode-se trabalhar como esta variável em cima dos modelos propostos.

Entre os resultados da aplicação dos modelos para o município de Belo Horizonte, observou-se que os bairros que possuem o maior valor do m² são os bairros Jardim Atlântico, Savassi, Belvedere, Funcionários e Santo Agostinho. Sobre a variação do valor médio do m² referente ao padrão de acabamento do imóvel, foi possível identificar que ele varia de 2.746,01 reais/m² (P1) a 7.554,90 reais/m² (P5). Foi observado a grande influência do zoneamento nos

preços médios do m², apresentando variação em torno de 50% nos preços médios para o mesmo padrão de acabamento, estes casos também podem servir de objeto de um estudo futuro.

Uma questão importante a se frisar é a falta de publicidade dos dados do ITBI, muitos municípios justificam a não publicação de seus dados com base na Lei 5172/1966 - CTN e na Lei 13709/2018. Vale a pena ressaltar que é possível publicar os dados do ITBI e respeitar os mecanismos impostos por estas leis, vide o caso do município de Belo Horizonte. A disponibilidade destes dados é importante pois permite a elaboração e a replicabilidade de estudos como este.

Referente às limitações deste estudo, vale ressaltar que nenhum modelo passou por uma otimização de hiperparâmetros, ou seja, a calibração poderia extrair uma melhor performance de cada ferramenta. Outra limitação é a de que não foram atribuídas variáveis de ordem macroeconômicas aos modelos, é sabido da influência destas variáveis no mercado imobiliário. Tais limitações poderiam constituir o objeto para estudos futuros.

Por fim, neste trabalho foram propostos seis modelos utilizados na estimação do preço do metro quadrado de imóveis situados no município de Belo Horizonte. Destes seis modelos, dois apresentaram um bom ajuste (XGBoost e SVM Kernel Linear), o que vai de encontro com o que foi constatado no referencial teórico. Estes resultados divergem em relação ao modelo GLM Gama, o ajuste insatisfatório vai na contramão do que foi exposto em outras pesquisas. Portanto, a partir dos resultados do presente trabalho, sugere-se a aplicação dos modelos XGBoost e SVM Kernel Linear como ferramenta de auxílio na precificação de imóveis.

Referências

ADETILOYE, K. A.; EKE, P. O. A Review of Real Estate Valuation and Optimal Pricing Techniques. **Asian Economic and Financial Review**, v. 4, n. 12, p. 1878–1893, 2014.

ABIDOYE, R. B.; CHAN, A. P. C.; ABIDOYE, F. A.; OSHODI, O. Predicting property price index using artificial intelligence techniques: Evidence from Hong Kong. **International journal of housing markets and analysis**, v. 12, n. 6, p. 1072-1092, 2019.

AL-SHAYEA, Q. K. Estate market forecast using artificial neural networks. **Journal of Advanced Research,** v. 2, n. 1, p. 1073-1080, 2012.

AMARESH, V.; SINGH, R. R.; KAMAL, R.; KULKARNI, A. Linear Regression Models based Housing Price Forecasting. In: **2022 International Conference on Industry 4.0 Technology (I4Tech)**. IEEE, p. 1-5, 2022.

BARBOSA, A. B. G.; CAVALCANTI, A. B. Web Scraping e Análise de dados. In: **Congresso Nacional de Pesquisa e Ensino em Ciências – CONAPESC**, 2020.

BELO HORIZONTE. Secretaria Municipal de Fazenda - SFMA. **Relatório ITBI**. Belo Horizonte, 2023. https://dados.pbh.gov.br/dataset/relatorio-itbi

BREIMAN, Leo. Random forests. Machine learning, v. 45, p. 5-32, 2001.

CAO, B.; ZHAN, D.; WU, X. Application of svm in financial research. In: **2009 International joint conference on computational sciences and optimization**. IEEE, 2009. p. 507-511.

ĆETKOVIĆ, J.; LAKIĆ, S.; LAZAREVSKA, M.; ŽARKOVIĆ, M.; VUJOŠEVIĆ, S.; CVIJOVIĆ, J.; GOGIĆ, M. Assessment of the real estate market value in the European market by artificial neural networks application. **Complexity**, v. 2018, 2018.

CHEN, T; GUESTRIN, C. Xgboost: A scalable tree boosting system. In: **Proceedings of the 22nd International conference on knowledge discovery and data mining**. p. 785-794. 2016.

CORTES, C; VAPNIK, V. Support-vector networks. **Machine learning**, v. 20, p. 273-297, 1995.

COURT, A. T. Hedonic price indexes with automotive examples. In: **The Dynamics of Automobile Demand**. General Motors, New York, p. 98-119, 1939.
FLORENCIO, L. A. Engenharia de Avaliações com Base em Modelos GAMLSS. Dissertação (Mestrado). **Programa de Pós-Graduação em Estatística - Universidade Federal de Pernambuco**, Recife, 2010.
FURTADO, B. A. Análise quantílica-espacial de determinantes de preços de imóveis urbanos com matriz de bairros: Evidências do mercado de Belo Horizonte. **Instituto de Pesquisa Econômica Aplicada (IPEA)**, n. 1570, 2011.
GALTON, F. Regression towards mediocrity in hereditary stature. **Journal of the Anthropological Institute of Great Britain and Ireland**, v. 15, p. 246-263, 1886.
GHOSALKAR, N. N.; DHAGE, S. N. Real estate value prediction using linear regression. In: **2018 fourth international conference on computing communication control and automation (ICCUBEA)**. IEEE, p. 1-5, 2018.
GU, J.; ZHU, M.; JIANG, L. Housing price forecasting based on genetic algorithm and support vector machine. **Expert Systems with Applications**, v. 38, n. 4, p. 3383-3386, 2011.
HERATH, S. K.; MAIER, G. The hedonic price method in real estate and housing market research. A review of the literature. **Institute for Regional Development and Environment** (pp. 1-21). Vienna, Austria: University of Economics and Business, 2010.
HU, L.; CHUN, Y.; GRIFFITH, D. A. Incorporating spatial autocorrelation into house sale price prediction using random forest model. **Transactions in GIS**, v. 26, n. 5, p. 2123-2144, 2022.
LEE, J.; RYU, J. P. Prediction of housing price index using artificial neural network. **Journal of the Korea Academia-Industrial Cooperation Society**, v. 22, n. 4, p. 228-234, 2021.
LITTLE, M. A.; VAROQUAUX, G.; SAEB, S.; LONINI, L. Using and understanding cross-validation strategies. **Perspectives on Saeb et al. GigaScience**, v. 6, n. 5, 2017.
MILLER, N. G.; GELTNER, D. M. **Real estate principles for the new economy**. Mason, Ohio: Thomson, South-Western, 251-330, 2005.
MOHD, T.; JAMIL, N. S.; JOHARI, N.; ABDULLAH, L.; MASROM, S. An Overview of Real Estate Modelling Techniques for House Price Prediction. In: **Kaur, N., Ahmad, M. (eds) Charting a Sustainable Future of ASEAN in Business and Social Sciences**. Springer, Singapore, 2020.
NELDER, J. A.; WEDDERBURN, R. W. M. Generalized linear models. **Journal of the Royal Statistical Society Series A: Statistics in Society**, v. 135, n. 3, p. 370-384, 1972.
OLIVEIRA, A. A. F.; BANDEIRA, S. R. V.; SILVA, C. V. A. Estimativa de Desempenho de Métodos de Aprendizado de Máquina Baseados em Árvores de Decisão na valoração do Solo no Município de Fortaleza, Brasil. **Revista da Sociedade Brasileira de Engenharia de Avaliações**, v. 1, n.1, 2022.
PEARSON, K. The Grammar of Science. **A. & C. Black**, Londres, 1892.
PINDYCK, R.; RUBINFELD, D. **Microeconomia.** 8 Ed. São Paulo: Pearson Education do Brasil, 2013
R CORE TEAM. R: **A Language and Environment for Statistical Computing. R Foundation for Statistical Computing**, Vienna, Austria, 2023.
REFAEILZADEH, P.; TANG, L.; LIU, H. **Cross-Validation, 1–7**. Springer New York, New York, NY. v. 10, 2016.
RODRIGUEZ, J. D.; PEREZ, A.; LOZANO, J. A. Sensitivity analysis of k-fold cross validation in prediction error estimation. **IEEE transactions on pattern analysis and machine intelligence**, v. 32, n. 3, p. 569-575, 2009.
ROSENBLATT, F. The perceptron: a probabilistic model for information storage and organization in the brain. **Psychological review**, v. 65, n. 6, p. 386, 1958.

SHEPPARD, S. Hedonic analysis of housing markets. In: **Handbook of regional and urban economics.** Amsterdam: North-Holland, p. 1.595-1.636, 1999.
SILVA, G. H. P. Modelos de aprendizagem de máquina para precificação de imóveis na cidade de Fortaleza. Monografia **(**Bacharelado). **Curso de Engenharia Civil - Universidade Federal do Ceará**, Fortaleza, 2019.
SRI, B. U.; REDDY, C. S. K.; KUMAR, C. R.; VINOD, A. V. B.; REDDY, B. K. Random Forest-based House Price Prediction. In: **2023 Second International Conference on Augmented Intelligence and Sustainable Systems (ICAISS)**. IEEE, 2023. p. 954-959.
VARIAN, H. R. **Microeconomia: Princípios Básico**s. Campus/Elsevier, Rio de Janeiro, 2003.
VERAS, A. D. Uma Proposta de Utilização de Redes Neurais Recorrentes na Previsão de Preços de Imóveis no Distrito Federal. Monografia **(**Bacharelado). **Curso de Administração – Universidade de Brasília**, Brasília, 2019.
WANG, T.; LI, Y.; ZHAO, S. Application of SVM Based on Rough Set in Real Estate Prices Prediction. **2008 4th International Conference on Wireless Communications, Networking and Mobile Computin**g, 1-4, 2008.
WANG, F.; CHEN, W.; FAKIEH, B.; ALHAMAMI, M. A. Stock price analysis based on the research of multiple linear regression macroeconomic variables. **Applied Mathematics and Nonlinear Sciences**, v. 7, n. 1, p. 267-274, 2021.
WANG, X.; WEN, J.; ZHANG, Y.; WANG, Y. Real estate price forecasting based on SVM optimized by PSO. **Optik**, v. 125, n. 3, p. 1439-1443, 2014.
ZAKI, J.; NAYYAR, A.; DALAL, S.; ALI, Z. H. House price prediction using hedonic pricing model and machine learning techniques. **Concurrency and Computation – Practice and Experience**, v. 34, n. 27, 2022.
ZENG, L. Research on Batch Evaluation of Real Estate Price Based on XGBoost. ***BCP Business & Management***, 2021.
ZEWOTIR, T.; BAX, D.; NORTH, D. A gamma generalized linear model as an alternative to log linear real estate price functions. **Journal of Economic and Financial Sciences**, v. 12, n. 1, p. 1-11, 2019.
ZHANG, Y.; HUANG, J.; ZHANG, J.; LIU, S.; SHORMAN, S. Analysis and prediction of second-hand house price based on random forest. **Applied Mathematics and Nonlinear Sciences**, v. 7, p. 27-42, 2022.
ZHAO, Y., CHETTY, G.; TRAN, D. Deep Learning with XGBoost for Real Estate Appraisal, **2019 IEEE Symposium Series on Computational Intelligence (SSCI)**, 1396-1401, 2019.