Modelo Hedônico para Estimação do Valor de Imóveis: Aplicação em Nova Friburgo-RJ☆

Hedonic Model for Real Estate Prices: Application to Nova Friburgo-RJ

Leonardo de Almeida Teodoro, Marco André Abud Kappel†

Centro Federal de Educação Tecnológica Celso Suckow da Fonseca – CEFET/RJ – Campus Nova Friburgo, Nova Friburgo, RJ, Brasil

†**Autor correspondente:** marco.kappel@cefet-rj.br

Resumo

Com a expansão do mercado imobiliário na Região Serrana do estado do Rio de Janeiro, uma parcela cada vez maior da população precisa tratar com a compra e venda de imóveis. Porém, a avaliação justa de uma unidade imobiliária não é uma tarefa simples e pode ser influenciada por diferentes atributos da edificação. Com o propósito de auxiliar nessa incumbência, o presente trabalho tem como objetivo identificar as características mais importantes na avaliação de um imóvel nessa região e, em seguida, propor um modelo matemático simples capaz de estimar o seu valor de mercado. Para isso, informações sobre valores de comercialização e detalhes construtivos de casas e apartamentos à venda na cidade de Nova Friburgo foram extraídos de portais de anúncios online, formando uma base única de dados imobiliários, sobre a qual foram, posteriormente, aplicadas técnicas de seleção de variáveis e regressão linear múltipla para a obtenção do modelo pretendido. Os resultados obtidos revelaram que a característica de maior influência na determinação do preço de compra e venda de um imóvel na região é a sua área construída. Por outro lado, o modelo matemático construído foi capaz de estimar os preços de comercialização de uma propriedade com aproximadamente 25% de desvio percentual médio da base de testes.

Palavras-chave

Aprendizado de Máquina • Regressão Linear Múltipla • Avaliação de Imóveis

Abstract

With the expansion of the real estate market in the mountainous region of Rio de Janeiro State, an increasing number of people need to deal with real estate purchasing and sale. However, to fair evaluate, a real estate unit is not a simple task and can be influenced by different characteristics. To assist in this task, the present work's objective is to identify the most critical attributes for evaluating a property and build a simple mathematical model that can be used to estimate the value of the property in this region. Data from properties for sale in the city of Nova Friburgo were extracted from online ad portals to build a unique database of real estate data. In these data, variable selection techniques and a multiple linear regression process were applied to obtain a mathematical model that describes prices based on the property's essential characteristics. The obtained results revealed that the most crucial aspect of the evaluation is the property's total area. The developed model could also predict prices with a mean percentage deviation of approximately 25% on the test database.

Keywords

Machine Learning • Multi-linear Regression • Real Estate Price Prediction

☆ Este artigo é uma versão estendida do trabalho apresentado no XXIII Encontro Nacional de Modelagem Computacional e XI Encontro de Ciência e Tecnologia de Materiais, realizado em Palmas/TO, 2020.

1 Introdução

O mercado imobiliário é um dos setores mais importantes da economia, não apenas por estar relacionado com moradia e qualidade de vida, mas igualmente com desenvolvimento urbano, investimento financeiro e impostos gerados nas transações comerciais. Além disso, quanto mais populosa e desenvolvida é uma cidade ou região, maior é o número de imóveis ofertados e demandados, o que torna esse um mercado bastante complexo. Na Região Serrana do estado do Rio de Janeiro, um fator que vem influenciando o crescimento da comercialização de imóveis é a violência existente na Região Metropolitana da capital. Segundo o Cenário do Mercado Imobiliário da SECOVI RIO [1], apesar da Região Serrana ser mais conhecida por seu potencial turístico, ela tem atraído cada vez mais moradores fixos devido a aspectos como segurança e qualidade de vida, quando comparada a outras zonas do estado. Como consequência, o preço dos imóveis nessa área tem sofrido significativos acréscimos, enquanto, na cidade do Rio de Janeiro, observa-se uma desvalorização ano a ano. Ainda segundo a SECOVI RIO, quase todos os bairros pesquisados nos municípios serranos apresentaram valorização no período de julho de 2017 a julho de 2018.

Como consequência dessa expansão regional do mercado imobiliário, um número cada vez maior de indivíduos precisa lidar com a compra e venda de edificações. A sua disposição existem hoje inúmeros portais de anúncios de imóveis disponíveis na internet, como OLX e Vivareal, que facilitam o acesso rápido às características dos mais variados tipos de construção e seu valor para comercialização.

No município de Nova Friburgo, por exemplo, a consulta por anúncios de imóveis na internet aumentou 24% na comparação do primeiro semestre de 2018 com o primeiro de 2017 [1]. Porém, seja quando o interesse é pela venda ou aquisição de um imóvel, um dos maiores obstáculos encontrados é a sua justa avaliação. Devido à grande heterogeneidade das características construtivas e estado de manutenção de cada propriedade, a atribuição de um valor para ela é uma das tarefas mais complexas do mercado [2]. Nessa operação, muitos fatores normalmente precisam ser considerados, fazendo com que a estimativa dos preços dos imóveis nem sempre seja a mais adequada ou coerente com seus reais atributos.

O objetivo do presente trabalho é identificar as características mais importantes para a avaliação de um imóvel na Região Serrana do estado do Rio de Janeiro e, em seguida, construir um modelo matemático simples que possa ser usado para estimar o seu valor.

2 Revisão Bibliográfica

Existem diversas abordagens para a tarefa de modelagem do preço de imóveis e identificação de suas características mais importantes para essa atividade. A técnica mais comumente aplicada para alcançar esse objetivo é a regressão linear múltipla, que torna possível a construção de um modelo hedônico para o valor dos imóveis de acordo com seus atributos [3], [4], [5], [6], [7]. Nesses estudos, realizados em geral em regiões específicas do planeta como uma cidade ou um conjunto de aglomerados urbanos, fica evidente que o preço estimado está relacionado com fatores locacionais, econômicos e estruturais. Os estudos disponíveis utilizam bases de dados revelam o comportamento do preço em regiões específicas do mundo, geralmente uma cidade ou conjunto de cidades. Não se trata, no entanto, de uma iniciativa fácil, muito menos em nosso país com grande desigualdade social e vasta e diversificada área territorial. Por isso, alguns autores optam pelo levantamento de dados via plataformas de anúncios online [6], [4]. Por outro lado, encontramos na literatura trabalhos que comparam diferentes métodos para a predição dos valores dos imóveis como, por exemplo, [5], [8]e [9].

A abordagem [5], em particular, propõe dois modelos obtidos por regressão linear múltipla aplicados na cidade de João Monlevade, estado de Minas Gerais. O primeiro modelo foi ajustado a partir de uma base de dados com 749 imóveis à venda e o segundo com 271 disponíveis para locação. Os autores concluíram que a quantidade de quartos e vagas de garagem foram as variáveis mais importantes em ambos os casos. Também reportaram que, apesar de os coeficientes de determinação registrados serem baixos (R^2 = 0,43 para venda e R^2 = 0,34 para aluguel), os resultados representaram a realidade do mercado imobiliário local. Outro estudo [3], realizado para avaliar o valor de mercado de apenas apartamentos residenciais em Fortaleza, Ceará, aplicando regressão linear múltipla, obteve um coeficiente de determinação R^2 igual a 0,91 e erro percentual médio de 21,10%. Segundo os autores, os bons resultados obtidos resultaram da delimitação da categoria de imóveis estudados. Já a pesquisa realizada na cidade de Sorocaba, estado de São Paulo [4], empregou o processo de regressão linear múltipla sobre dados obtidos a partir de *websites* de imobiliárias locais. Entretanto, foram aplicadas técnicas de seleção de variáveis para identificar as mais relevantes na determinação do valor dos imóveis. O modelo obtido alcançou apenas 13% de erro percentual médio. O mesmo processo foi aplicado na cidade de Kuala Lumpur, Malásia alcançando um R^2 de aproximadamente 0,84, como pode ser visto em [7]. Cabe citar ainda a abordagem descrita em [8], que teve como finalidade comparar diferentes técnicas para a estimação dos preços dos imóveis em Pequim, China. Nesse caso, os autores fizeram uso da base de dados da maior agência de imóveis do país,

envolvendo 14.758 propriedades com distintos aspectos estruturais e locacionais, obtendo um coeficiente de determinação R^2 de apenas 0,42.

Fica evidente, portanto, que embora a literatura destaque atributos mais significativos para a predição do valor de uma propriedade, a heterogeneidade dos mercados imobiliários requer a realização de estudos para localidades específicas onde distintos aspectos das construções podem influenciar de forma diferente o valor estimado para a comercialização de imóveis.

3 Metodologia

O procedimento adotado no presente trabalho envolveu o processamento de dados imobiliários e a aplicação de técnicas computacionais para a previsão do valor de imóveis de acordo com suas características.

Com a finalidade de obter um número elevado de dados sobre imóveis foi aplicado um processo de *web scraping* automatizado a partir do site vivareal.com.br durante o mês de agosto de 2020, portal com maior número de imóveis anunciados para venda na cidade de Nova Friburgo, sendo extraídos um total de 905 cadastros cuja localização espacial pode ser vista na Fig. 1.

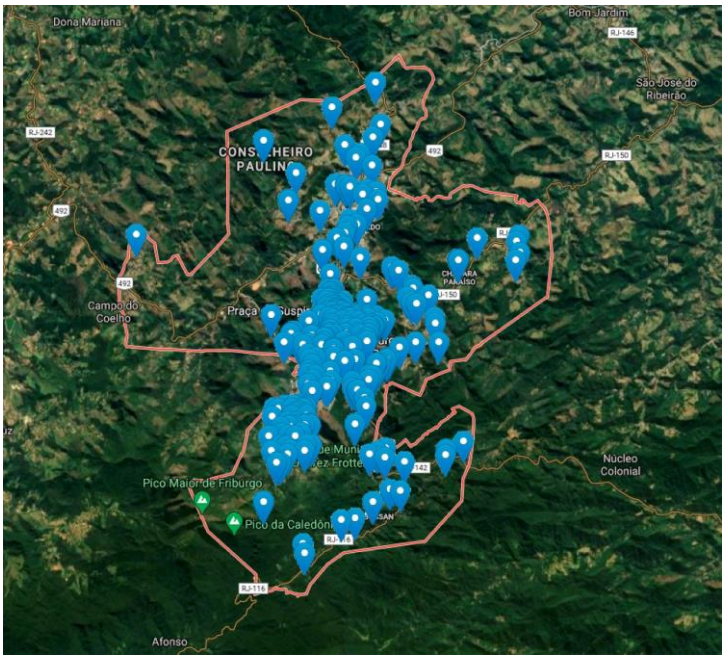

Figura 1: Imagem de satélite da cidade de Nova Friburgo com a localização dos 905 imóveis considerados inicialmente no presente trabalho. Fonte: Google Maps

A partir dessa base de dados, foi possível identificar as seguintes características de cada imóvel, que constituíram 16 variáveis consideradas no estudo: valor de comercialização em reais, categoria (casa isolada, casa em condomínio, apartamento ou cobertura), bairro, área total construída, quantidade de dormitórios, vagas de estacionamento, banheiros, e a presença das seguintes estruturas: varanda, área de serviço, lareira, dependência completa, elevador, churrasqueira, piscina e sauna.

A variável bairro, por ser nominal categórica, apresentou 33 categorias e durante a fase de pré-processamento foi substituída por variáveis *dummy*, cada uma responsável por indicar se o imóvel pertencia ou não a um determinado bairro. Com isso, o total de variáveis previsoras disponíveis elevou-se para 47. Inicialmente, eram 15 características, sendo uma o bairro. Porém, a variável bairro foi dividida em 33 variáveis booleanas, uma para cada bairro. Assim, no total, consideramos o total de 47 variáveis (14 + 33).

Devido à elevada volatilidade do mercado imobiliário, a ausência de um modelo local para o estabelecimento de uma estimativa de preços, a possível inexperiência de alguns corretores imobiliários ou ainda a inserção no portal pesquisado de valores arbitrados por pessoas leigas no setor, que podem resultar em informações subestimadas ou superestimadas, foi necessário considerar a existência de *outliers* na base de dados gerada. Da mesma forma, de acordo com [18], imóveis que se localizam em uma mesma região possuem valores com distribuições similares, sendo importante também verificar a existência de *outliers* por zoneamento geográfico.

Para solucionar essas questões, foram considerados *outliers* todos os imóveis cujos valores estimados estavam além dos limites definidos pela regra 1,5×IQR [10] em cada bairro, sendo o IQR o intervalo interquartil.

A Figura 2a mostra *boxplots* dos valores de venda dos imóveis de acordo com o bairro em que ele está localizado. Analisando a figura, é possível perceber a real presença de *outliers*. A Figura 2b mostra o resultado obtido após a remoção dos outliers. Ao final desse processo, restaram 853 imóveis na base de dados. Três imóveis estavam abaixo de 1,5×IQR e 49 acima.

a) b)

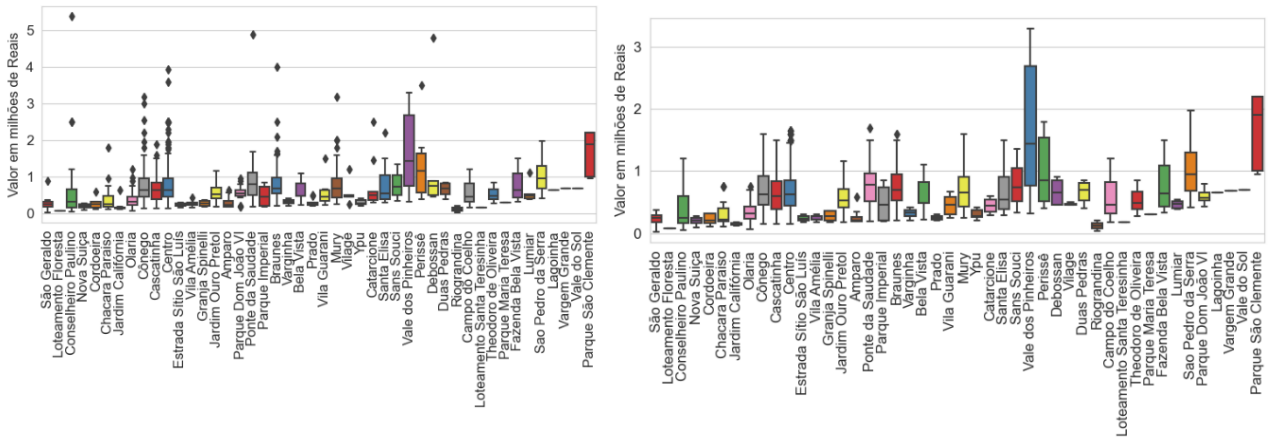

Figura 2: *Boxplots* dos valores dos imóveis expressos por bairro: com *outliers* (a) e sem *outliers* (b).

Em seguida, a técnica *Random Forest* [11] foi aplicada para a identificação das variáveis mais importantes para a estimação do valor dos imóveis. O objetivo, aqui, foi reduzir o número total de variáveis a serem usadas na regressão linear múltipla a apenas aquelas com maior importância na estimativa do preço. Métodos estatísticos como o *Random Forest* podem ser utilizados na seleção de variáveis [12], pois executam internamente um processo de avaliação de importância de cada entrada, utilizando métricas como entropia ou impureza de Gini [13]. Nestes casos, a seleção de variáveis é feita de maneira "mergulhada", o que faz esses métodos serem conhecidos como *embedded* [14].

Após esta etapa, a base foi separada em duas partes: dados para treinamento e para testes. A base de treinamento, que representa os dados efetivamente usados no processo de aprendizagem, foi composta por 75% dos dados, selecionados aleatoriamente. A base de testes foi composta pelos 25% restantes. Ela foi usada como um conjunto de dados inéditos, nunca testados pelo sistema.

Posteriormente, a regressão linear múltipla expressa pela Eq. (1) foi aplicada na base de treinamento, para que o modelo desejado fosse construído, mas considerando apenas as 20 principais variáveis obtidas pela seleção anterior.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon \tag{1}$$

onde y é a variável objetivo, no caso, o valor do imóvel, $\beta_n, n = 0{,}1, \ldots, k$ são os coeficientes de regressão a serem definidos, $x_n, n = 1, \ldots, k$ são as variáveis previsoras do modelo, que representam, no caso, as características do imóvel, e ε é uma variável aleatória que representa o erro do modelo. É importante observar que esse tipo de modelo considera algumas suposições estatísticas: o erro deve ser uma variável aleatória que segue uma distribuição normal com média igual a zero e variância constante [15]. Além disso, supõe-se que as variáveis previsoras sejam independentes, caso contrário, pode-se caracterizar uma multicolinearidade, e, por consequência, o modelo pode ser prejudicado [16]. Para avaliar a qualidade do modelo, o desvio relativo médio foi calculado, assim como o coeficiente de determinação R^2. Todas as implementações neste estudo foram desenvolvidas com o uso da linguagem Python e das ferramentas disponíveis no pacote scikit-learn [17].

Importante destacar que o ajuste dos coeficientes foi realizado com o método *Ridge Regression* implementado na biblioteca scikit-learn.

O modelo treinado foi, então, testado e avaliado quanto a sua eficácia a partir das informações disponíveis na base de testes, cujos valores não foram utilizados na fase de treinamento e, portanto, eram desconhecidos.

A Figura 3 apresenta um esquema da metodologia aplicada.

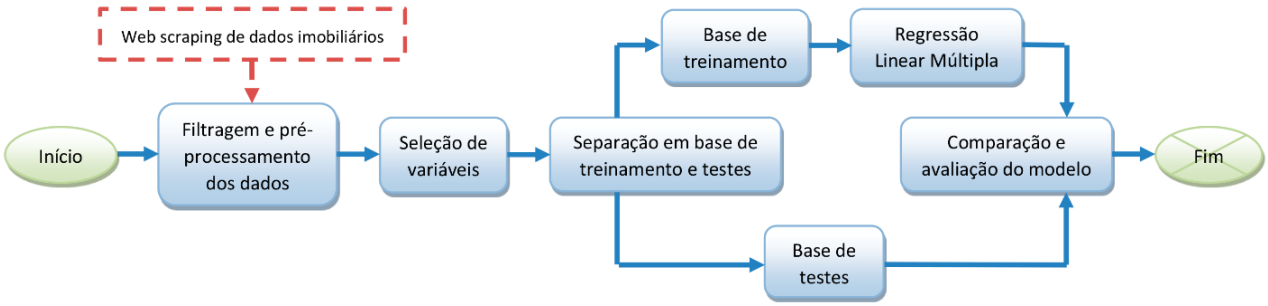

Figura 3: Resumo da metodologia utilizada.

4 Resultados e Discussão

A partir da aplicação da metodologia descrita no presente trabalho, é possível visualizar - na Fig. 4 - as 20 variáveis mais importantes identificadas para a determinação do valor de um imóvel no município de Nova Friburgo, cuja descrição mais detalhada pode ser encontrada na Tabela 1. Essas, portanto, foram as variáveis empregadas como entrada para a construção do modelo por regressão linear múltipla. Ainda de acordo com a Figura 4, a área total do imóvel foi a variável que se revelou como mais importante, enquanto a quantidade de quartos e banheiros vêm em seguida. A quantidade de vagas disponíveis para veículos mostra também significativa influência, seguida pelo tipo de imóvel e a presença de lareira.

Com referência à localização do imóvel, aparece na sétima posição da referida listagem a variável "Centro", demonstrando a importância na valorização do preço da propriedade pelo fato dela estar localizada na área central da cidade, ou seja, no centro urbano, região com maior número de atividades comerciais e financeiras. Esse fato é corroborado pelas estatísticas recentes do mercado imobiliário, que mostram que o centro é a área mais valorizada em Nova Friburgo nos últimos anos [1].

A Figura 5, por sua vez, mostra os dados obtidos pela correlação de Pearson entre as variáveis previsoras e a variável objetivo. Na figura, é possível observar que a maior parte das variáveis possuem forte correlação com a variável objetivo. Porém, algumas das variáveis possuem correlação entre si, o que pode demonstrar a existência de multicolinearidade, um fato comum em problemas de estimação de valor de imóveis [3]. Por exemplo, verifica-se que a presença de piscina no imóvel está relacionada linearmente com a presença de churrasqueira e sauna. Neste estudo, para contornar o problema de multicolinearidade, a regressão linear múltipla foi realizada utilizando o *Ridge Regression* [19]. Sugere-se, no entanto, que futuros estudos abordem mais detalhadamente o problema da combinação e transformação de variáveis com a finalidade de se aprimorar o modelo proposto.

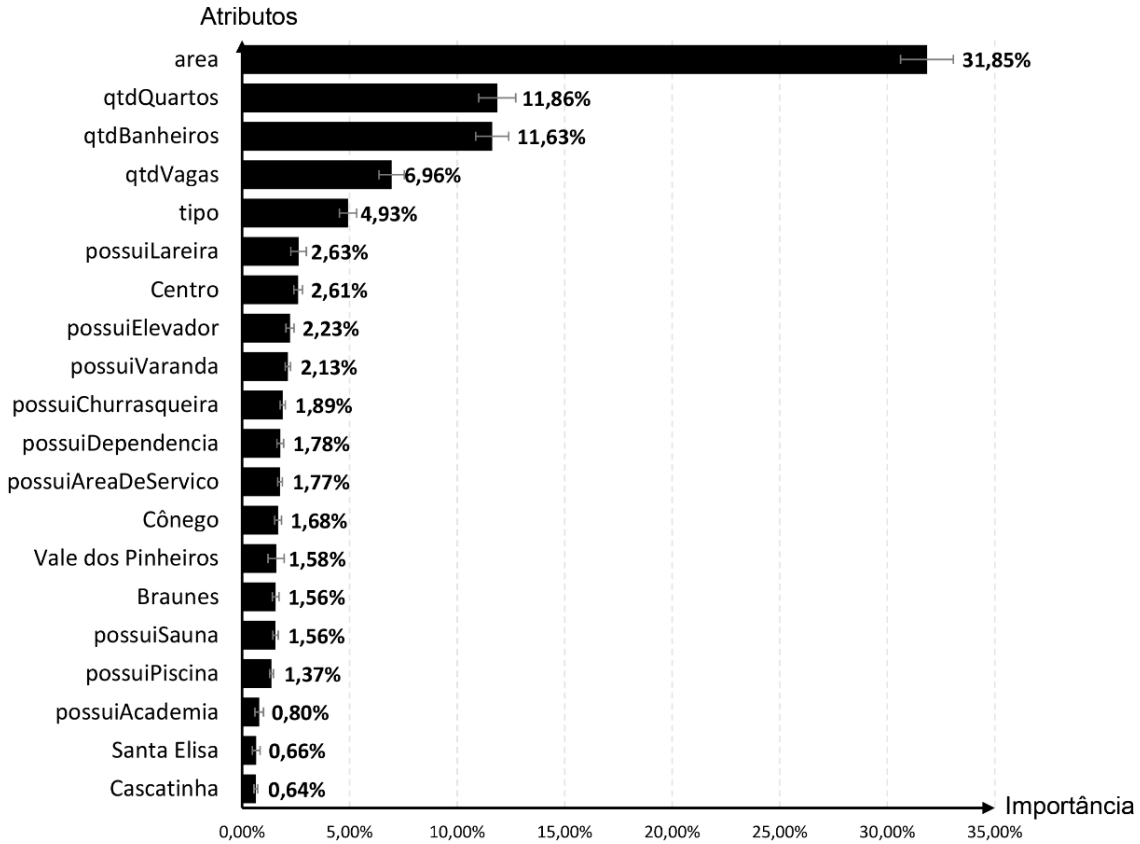

Figura 4: Lista ordenada de importância das 20 variáveis/características consideradas no valor de um imóvel em Nova Friburgo, RJ, obtidas pelo método *Random Forest*.

Tabela 1: 20 características mais importantes no preço de um imóvel, segundo o método *Random Forest*, em ordem decrescente de importância.

x	Nome da Variável	Descrição	Tipo
1	area	Área total	real
2	qtdQuartos	Quantidade de Quartos	inteiro
3	qtdBanheiros	Quantidade de Banheiros	inteiro
4	qtdVagas	Quantidade de Vagas	inteiro
5	tipo	Tipo de Imóvel	Categórico nominal: 0 – Casa 1 – Casa em condomínio 2 – Apartamento 3 – Cobertura
6	possuiLareira	Possui lareira?	lógico (0 ou 1)
7	Centro	Bairro: Centro	lógico (0 ou 1)
8	possuiElevador	Possui elevador?	lógico (0 ou 1)
9	possuiVaranda	Possui varanda?	lógico (0 ou 1)
10	possuiChurrasqueira	Possui churrasqueira?	lógico (0 ou 1)
11	possuiDependencia	Possui dependência completa?	lógico (0 ou 1)
12	possuiAreaDeServico	Possui área de serviço?	lógico (0 ou 1)

13	Cônego	Bairro: Cônego	lógico (0 ou 1)
14	Vale dos Pinheiros	Bairro: Vale dos Pinheiros	lógico (0 ou 1)
15	Braunes	Bairro: Braunes	lógico (0 ou 1)
16	possuiSauna	Possui sauna?	lógico (0 ou 1)
17	possuiPiscina	Possui piscina?	lógico (0 ou 1)
18	possuiAcademia	Possui academia?	lógico (0 ou 1)
19	Santa Elisa	Bairro: Santa Elisa	lógico (0 ou 1)
20	Cascatinha	Bairro: Cascatinha	lógico (0 ou 1)

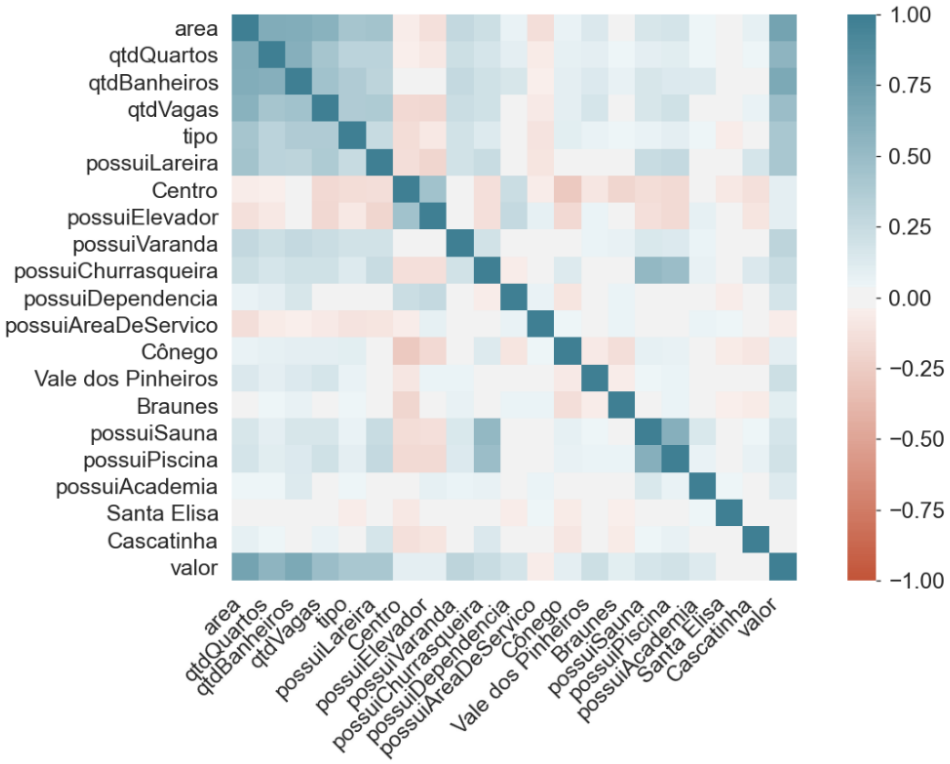

Figura 5: Correlação de Pearson entre as variáveis consideradas para a construção do modelo.

O modelo matemático resultante para a estimativa dos valores dos imóveis em Nova Friburgo pode ser visto na Eq. (2) onde y corresponde ao preço estimado do imóvel e $x_n, n = 1, \ldots, 20$ às características listadas na Tabela 1.

$$\begin{aligned} y = -128030 &+ 1628{,}77x_1 + 23145x_2 + 70539x_3 + 18364{,}3x_4 \\ &+ 46711{,}6x_5 + 155475x_6 + 177501x_7 + 99719{,}8x_8 \\ &+ 35095{,}5x_9 + 35759{,}6x_{10} + 65556x_{11} + 36459{,}2x_{12} \\ &+ 129585x_{13} + 264647x_{14} + 152649x_{15} + 10303{,}6x_{16} \\ &+ 27483{,}2x_{17} + 126190x_{18} + 151610x_{19} - 25718{,}6x_{20} \end{aligned} \quad (2)$$

Para utilizar o modelo, basta substituir os valores de cada caraterística do imóvel e obter seu valor estimado. Para avaliar a eficácia do modelo, ele foi aplicado a todos os imóveis da base de testes, cujos valores não foram utilizados na fase de treinamento e, portanto, são desconhecidos. Os valores do erro relativo do modelo tanto na base de treinamento, quanto na base de testes, podem ser vistos na Fig. 6. O valor do desvio relativo médio obtido na base de testes foi de 25,22%, valor compatível com outros modelos encontrados na literatura. Além disso, o

modelo alcançou um R^2 score de 0,74, também nos dados de teste. A Figura 7 mostra um histograma com a distribuição estatística dos erros encontrados. Analisando essa figura, é possível perceber que tanto na base de treinamento, quanto na base de testes, o erro se aproxima de uma distribuição normal com média 0, conforme a suposição adotada para o desenvolvimento da regressão linear múltipla. Fatores não considerados, mas que certamente influenciam são o estado de conservação do imóvel e as condições circunvizinhas. Esses fatores, se disponíveis nos anúncios, poderiam ser usados no ajuste do modelo e, provavelmente, melhorariam a taxa de acerto.

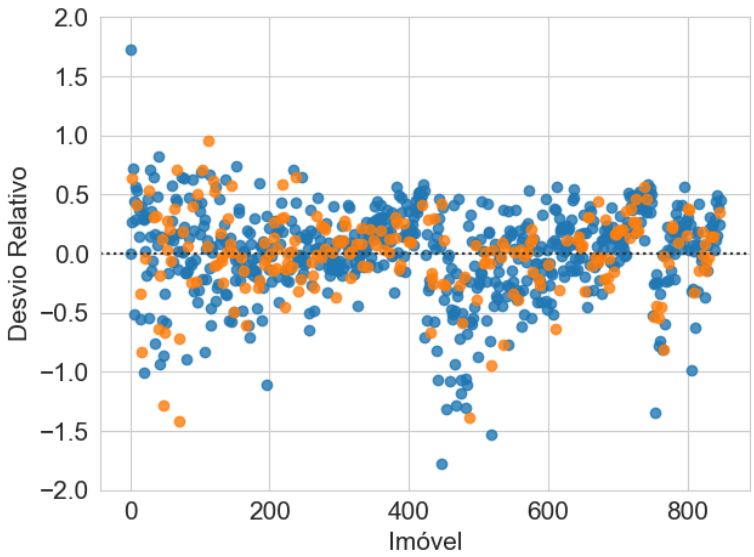

Figura 6: Desvio relativo obtido pelo modelo quando aplicado às bases de treinamento e de testes. A cor azul refere-se a dados de treinamento, e a laranja a teste.

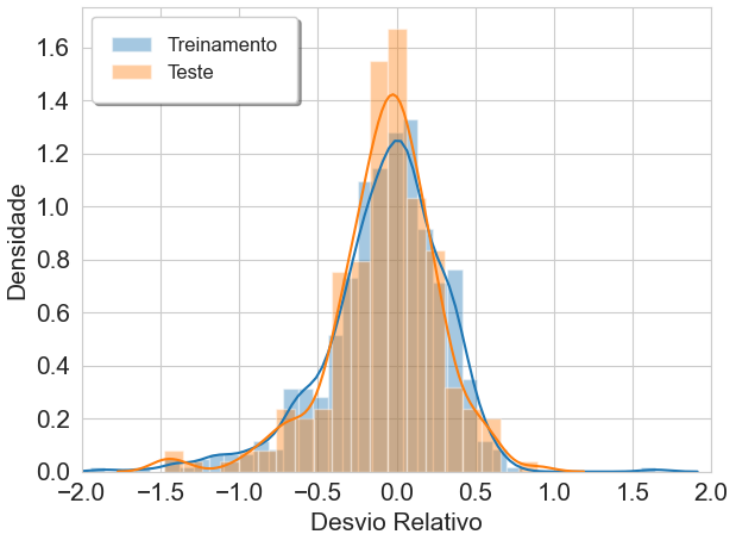

Figura 7: Histograma do desvio relativo obtido pelo modelo quando aplicado nas bases de treinamento e de testes.

5 Conclusões

A atual disponibilização de plataformas online com anúncios referentes ao mercado imobiliário possibilitou que fossem coletados de forma ágil dados reais e em grande quantidade, não apenas sobre os valores de comercialização dos imóveis, como também sobre características básicas de cada unidade imobiliária. Após uma fase de filtragem de dados e remoção de *outliers*, o presente trabalho utilizou a base obtida para a construção de um modelo matemático simples, que auxiliasse na estimação de preços de imóveis na cidade de Nova Friburgo,

principalmente para aqueles que não possuem experiência em corretagem imobiliária, com base nos anúncios existentes em agosto de 2020.

As variáveis previsoras consideradas no modelo desenvolvido foram capazes de descrever, com aproximadamente 75% de acerto, os valores adequados de compra e venda de imóveis da cidade de Nova Friburgo. Os resultados obtidos demonstraram-se compatíveis com modelos de precificação de imóveis encontrados na literatura para as cidades relacionadas na literatura utilizada. É provável que informações não disponíveis sobre os imóveis considerados, como, por exemplo, a idade da construção, o número de andares, o nível de poluição sonora, a obstrução na vista de janelas e varandas, sejam responsáveis por um percentual de incerteza do modelo desenvolvido. Ainda assim, o modelo matemático aplicado atingiu com sucesso o objetivo de ser um guia auxiliar na estimativa de preços de casas e apartamentos em Nova Friburgo.

Agradecimentos

O desenvolvimento desta pesquisa teve o apoio do CEFET/RJ.

Referências

[1] SECOVI RIO, "Cenário do Mercado Imobiliário da Região Serrana do Rio de Janeiro - 2018," 2018. [Online]. [Acesso em 11 setembro 2020].

[2] M. T. A. C. N. A. B. S. N. &. A. V. Steiner, "Métodos estatísticos multivariados aplicados à engenharia de avaliações," *Gestão & Produção,* vol. 15, no. 1, pp. 23-32, 2008. Disponível em: https://www.scielo.br/scielo.php?pid=S0104-530X2008000100004&script=sci_arttext&tlng=pt

[3] D. B. Nunes, J. D. P. B. Neto e S. M. d. Freitas, "Modelo de regressão linear múltipla para avaliação do valor de mercado de apartamentos residenciais em Fortaleza, CE," *Ambiente Construído,* vol. 19, no. 1, pp. 89-104, 2019. Disponível em: https://www.scielo.br/scielo.php?script=sci_arttext&pid=S1678-86212019000100089&lng=pt&tlng=pt

[4] J. C. Pereira, S. Garson e E. G. d. Araújo, "Construção de um modelo para o preço de venda de casas residenciais na cidade de Sorocaba-SP," *Gestão da Produção, Operações e Sistemas,* vol. 7, pp. 153-167, 2012. Disponível em: https://revista.feb.unesp.br/index.php/gepros/article/view/861

[5] V. S. Rosa, P. B. d. Oliveira e R. L. Pinto, "Modelos de precificação para locação e venda de imóveis residenciais na cidade de João Monlevade-MG via regressão linear multivariada," *Gestão da produção, operações e sistemas.,* vol. 14, no. 3, pp. 151-167, 2019. Disponível em: https://revista.feb.unesp.br/index.php/gepros/article/view/2614

[6] V. Pinto e R. A. S. Fernandes, "Análise de preços hedônicos no mercado imobiliário residencial de Conselheiro Lafaiete, MG," *Interações,* vol. 20, no. 2, pp. 627-643. Disponível em: https://interacoesucdb.emnuvens.com.br/interacoes/article/view/1788

[7] A. M. Yusof e S. Ismail, "Multiple Regressions in Analysing House Price Variations," *Communications of the IBIMA,* vol. 2012, pp. 1-9. Disponível em: https://ibimapublishing.com/articles/CIBIMA/2012/383101/

[8] Z. Yan e L. Zong, "Spatial Prediction of Housing Prices in Beijing Using Machine Learning Algorithms," em *HPCCT & BDAI 2020: Proceedings of the 2020 4th High Performance Computing and Cluster Technologies Conference & 2020 3rd International Conference on Big Data and Artificial Intelligence*, Nova Iorque, 2020. Disponível em: https://dl.acm.org/doi/10.1145/3409501.3409543

[9] H. Wu, H. Jiao, Y. Yu, Z. Li, Z. Peng, L. Liu e Z. Zeng, "Influence Factors and Regression Model of Urban Housing Prices Based on Internet Open Access Data," *Sustainability,* vol. 10, pp. 1-17. Disponível em: https://www.mdpi.com/2071-1050/10/5/1676

[10] S. Walfish, "A review of statistical outlier methods," *Pharmaceutical technology,* vol. 30, no. 11, p. 82. Disponível em: https://www.pharmtech.com/view/review-statistical-outlier-methods

[11] L. Breiman, "Random Forests," *Machine Learning,* vol. 45, pp. 5-32. Disponível em: https://link.springer.com/article/10.1023/A:1010933404324

[12] U. Grömping, "Variable Importance Assessment in Regression: Linear Regression versus Random Forest," *The American Statistician,* vol. 63, no. 4, pp. 308-319. Disponível em: https://www.tandfonline.com/doi/abs/10.1198/tast.2009.08199

[13] K. J. Archer e R. V. Kimes, "Empirical characterization of random forest variable importance measures," *Computational Statistics & Data Analysis,* vol. 52, no. 4, pp. 2249-2260, 10 Janeiro 2008. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0167947307003076?via%3Dihub

[14] T. N. Lal, O. Chapelle, J. Weston e A. Elisseeff, "Embedded Methods," em *Feature Extraction: Foundations and Applications*, 1 ed., I. Guyon, S. Gunn, M. Nikravesh e L. A. Zadeh, Eds., Berlim, Springer-Verlag Berlin Heidelberg, 2006, pp. 137-165. Disponível em: https://link.springer.com/book/10.1007%2F978-3-540-35488-8

[15] R. A. Jonhson e D. W. Wichern, Applied Multivariate Statistical Analysis, 6 ed., Upper Saddle River, Nova Jersey: Pearson, 2007, p. 360. Disponível em: https://www.pearson.com/us/higher-education/product/Johnson-Applied-Multivariate-Statistical-Analysis-6th-Edition/9780131877153.html?

[16] E. R. Mansfield e B. P. Helms, "Detecting Multicollinearity," *The American Statistician,* vol. 36, no. 3a, pp. 158-160, 1982. Disponível em: https://www.tandfonline.com/doi/abs/10.1080/00031305.1982.10482818

[17] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot e É. Duchesnay, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research,* vol. 12, no. 85, pp. 2825–2830, 2011. Disponível em: https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html

[18] X. Liang, Y. Liu, T. Qiu, Y. Jing e F. Fang, "The effects of locational factors on the housing prices of residential communities: The case of Ningbo, China," *Habitat International,* vol. 81, pp. 1-11, 16 Setembro 2018. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0197397517311797?via%3Dihub

[19] M. El-Denery e N. I. Rashwan, "Solving Multicollinearity Problem Using Ridge Regression Models," *International Journal of Contemporary Mathematical Sciences,* vol. 6, pp. 585-600, 2011. Disponível em: http://www.m-hikari.com/ijcms-2011/9-12-2011/rashwanIJCMS9-12-2011.pdf