

Predictive Modeling of Real Estate Prices Using Machine Learning

Atílio Cardoso Azevedo¹, Reginaldo Cordeiro dos Santos Filho¹

¹Laboratório de Inteligência de Dados (LID)
Programa de Pós-Graduação em Computação (PPGCOMP)
Universidade Federal do Pará (UFPA)

atilio1999@gmail.com, regicsf@ufpa.br

Abstract. *Determining the value of a property is a complex task due to the wide range of factors that influence its pricing, such as location, size, number of rooms, and other relevant attributes. With the objective of simplifying the pricing process, this article proposes the use of Machine Learning Algorithms capable of estimating a property's value based on its physical and locational characteristics. The algorithms were trained on a dataset containing online listings of properties for sale in the city of Belém, Pará, during the period from December 2023 to January 2025. The results of this proof of concept indicate that the Random Forest Regressor algorithm achieved the best performance, with an R^2 of 0.902 and a MAPE of 26%.*

Resumo. *Determinar o valor de um imóvel é uma tarefa complexa devido à ampla variedade de fatores que influenciam essa precificação, como localização, tamanho, número de compartimentos e demais atributos relevantes. Com o objetivo de simplificar o processo de precificação, este artigo propõe o uso de Algoritmos de Aprendizado de Máquina capazes de estimar o valor de um imóvel com base em suas características físicas e de localidade. Os algoritmos foram treinados em uma base de dados contendo anúncios online de venda de imóveis na cidade de Belém do Pará, entre dezembro de 2023 a janeiro de 2025. Os resultados desta prova de conceito indicam que o algoritmo Random Forest Regressor obteve o melhor desempenho, com um R^2 de 0.902 e um MAPE de 26%.*

1. Introdução

A Inteligência Artificial (IA) vem ganhando espaço e se tornando cada vez mais presente em diversas áreas do conhecimento, e o mercado imobiliário não é exceção [Firpo 2024]. Algumas empresas do setor já utilizam algoritmos de aprendizado de máquina para precificar imóveis, tanto para venda quanto para aluguel, com base em suas características físicas e de localização. Essa abordagem permite que vendedores e compradores tomem decisões mais embasadas e realistas. No entanto, no Brasil, ainda há uma carência de estudos acadêmicos que explorem a aplicação desses modelos de forma regionalizada, considerando as particularidades do mercado local. Sabe-se, por exemplo, que até 2014 não havia registros do uso de modelos de aprendizado de máquina com foco em vendas imobiliárias no Brasil [Afonso et al. 2019]. Além disso, até os dias atuais, a maioria das plataformas virtuais ainda depende de uma avaliação empírica por parte do vendedor ou

de corretores para definir o preço do imóvel anúncio. No caso específico de Belém, capital do Pará, observa-se uma ausência de estudos sobre a aplicação de aprendizado de máquina no mercado imobiliário local.

No Pará, a demanda por imóveis é especialmente concentrada em suas metrópoles. Na plataforma de vendas online OLX¹, Belém registra um número de anúncios duas vezes maior que Ananindeua, a segunda cidade mais procurada que, por sua vez, apresenta uma quantidade significativamente superior à de Santarém, a terceira cidade mais populosa do estado. Esse padrão também é observado em outras plataformas de vendas, como o VivaReal².

Este artigo avalia o desempenho de algoritmos de aprendizado de máquina na tarefa de regressão, estimando o preço de venda de imóveis em Belém do Pará com base em atributos como área, número de quartos, vagas na garagem, banheiros, localização (bairro), tipo de imóvel (apartamento ou casa), valor do condomínio e IPTU. Esses atributos são comumente cadastrados em plataformas de anúncios online. Os resultados mostram que a precificação automatizada com base nesses dados pode facilitar o trabalho de vendedores e corretores, além de garantir preços mais justos para os compradores.

Este artigo está organizado da seguinte forma: a Seção 2 apresenta os trabalhos relacionados, a Seção 3 detalha a metodologia, a Seção 4 discute os resultados e a Seção 5 traz as considerações finais e propostas futuras.

2. Trabalhos correlatos

Esta seção apresenta os trabalhos correlatos à estimação de preços de imóveis. A Tabela 1 resume as principais informações e resultados de cada estudo analisado, facilitando a condução das discussões.

Autores	Melhor AAM	Atributos mais relevantes	Resultado
[Zaki et al. 2022]	XGBoost	Quartos	84.1% acurácia
[Dabreo et al. 2021]	XGBoost	Quartos	88.0% acurácia
[Panchal et al. 2022]	XGBoost	Quartos, Gasconnection	-
[Truong et al. 2020]	Regressão híbrida	Localização, Tempo	0.163 RMSLE
[Wang et al. 2019]	ARIMA	-	78.0% acurácia
[Mysore et al. 2022]	SVM	Quartos	0.89 R ²

Tabela 1. Informações relevantes sobre cada trabalho correlato.

Diversos estudos recentes apontam o XGBoost como o modelo com melhor desempenho na predição de preços de imóveis. Em [Zaki et al. 2022], foi utilizada uma base do Kaggle com 13 variáveis, incluindo métricas como criminalidade e taxa de impostos, alcançando uma acurácia de 84,1% com o XGBoost. De forma semelhante, [Dabreo et al. 2021] utilizou dados de vendas de imóveis em Boston, identificando que o número de quartos era a variável com maior correlação com o preço final, enquanto a distância até centros comerciais foi a de menor impacto. Também empregando o XGBoost, [Panchal et al. 2022] propuseram o sistema MakanSETU, que prioriza atributos

¹OLX: <https://www.olx.com.br/imoveis/venda/estado-pa>.

²VivaReal: <https://www.vivareal.com.br/venda/para/>

relevantes com base na matriz de correlação de Pearson, destacando variáveis como área, número de quartos, latitude e conexão de gás como valores que tiveram uma alta correlação com a variável alvo. Os autores mencionaram que o modelo de *Random Forest Regressor* também apresentou um bom desempenho.

Outras abordagens buscaram explorar modelos híbridos. Em [Truong et al. 2020], foi aplicada uma regressão híbrida com 19 variáveis referentes à cidade de Pequim, incluindo localização em relação ao centro urbano, atingindo um RMSLE de 0.16372. Já [Wang et al. 2019] exploraram a não-linearidade dos preços por meio de uma rede neural com função de ativação ReLU, combinada com o modelo estatístico ARIMA para prever tendências temporais, embora tenham apontado limitações em sua escalabilidade ao longo dos anos. Por fim, [Mysore et al. 2022] obteve os melhores resultados com o SVM, mas também destacou o bom desempenho do CatBoost, além de aplicar técnicas robustas de pré-processamento como o *Robust Scaler* e o *Yeo-Johnson Transformer*.

3. Metodologia

O modelo *Cross-Industry Standard Process for Data Mining (CRISP-DM)* foi adotado neste artigo por ser um dos *frameworks* mais amplamente utilizados em projetos de ciência de dados e por ser recomendado para iniciativas voltadas ao mercado, tal como a temática tratada neste trabalho [Chapman et al. 2000]. O CRISP-DM destaca-se, especialmente, por sua abordagem iterativa, que permite que projetos de mineração de dados sejam conduzidos de maneira a promover revisões contínuas a cada etapa, garantindo uma avaliação constante das decisões tomadas. Nas subseções seguintes é explicado como a base de dados de anúncios foi obtida e o pré-processamento aplicado na mesma.

3.1. Construção da base de dados

A base de dados utilizada nesta pesquisa foi extraída da plataforma de vendas OLX, considerada a maior do Brasil no segmento de comércio circular. Dentre suas categorias, encontram-se anúncios de imóveis, que podem ser publicados e editados diretamente pelos usuários. Os dados foram extraídos quinzenalmente por meio de um algoritmo de *web scraping* desenvolvido para este trabalho. Para garantir uma amostragem adequada, foram aplicadas técnicas de *data scraping*. A Figura 1 apresenta a interface de um anúncio na plataforma.

A base de dados foi construída a partir de anúncios coletados entre dezembro de 2023 e janeiro de 2025, totalizando 14 meses. Para maximizar a captação de informações, foram realizadas duas coletas mensais: uma na primeira quinzena e outra no final do mês. Essa estratégia foi adotada devido à limitação da plataforma OLX, que exibe no máximo 100 páginas de anúncios em seu sistema de pesquisa. Observou-se empiricamente que, após 15 dias, essas páginas já estavam preenchidas com novos anúncios, justificando o intervalo definido para a coleta de dados.

Os atributos relevantes para a construção da base de dados estão todos contidos dentro da página web de um anúncio e são obtidos a partir da leitura do código HTML. O algoritmo primeiramente obtém as URLs resultantes de uma busca por imóveis na cidade de Belém. Nesta etapa, aproximadamente 5000 URLs (100 páginas e 50 anúncios por página) são coletadas. Em seguida, todos esses links são analisados para que os atributos de interesse sejam extraídos e, por fim, sejam guardados dentro de um arquivo em formato “.csv”.

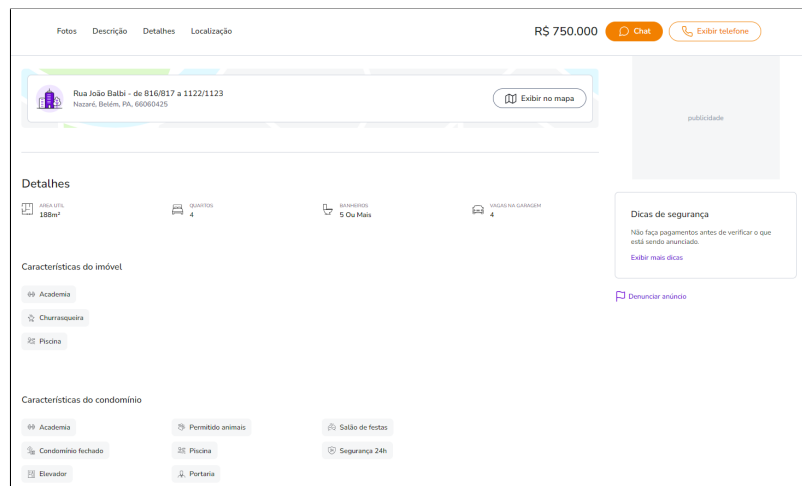


Figura 1. Interface de um anúncio de imóvel na plataforma OLX.

3.2. Atributos da base de dados

Os atributos de cada anúncio estão listados na Tabela 2. Anúncios que não informam o número de quartos ou banheiros são removidos da base. Para anúncios sem especificação de garagem, foi imputado o valor 0 (zero). O imóvel é categorizado de forma binária: 0 para casa e 1 para apartamento. Da mesma forma, os valores de condomínio e IPTU são binarizados, onde 0 indica a ausência desses custos e 1 representa sua presença.

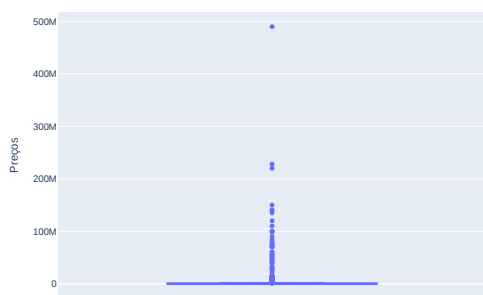
Tabela 2. Atributos dos dados extraídos dos anúncios.

Nome	Descrição	Tipo
Área	Metro quadrado do imóvel.	Discreto
Quarto	Nº de compartimentos.	Discreto
Garagem	Nº de carros possíveis.	Discreto
Banheiro	Nº de banheiros.	Discreto
Bairro	Localização do imóvel.	Texto
Imóvel	Apartamento ou Casa	Texto
Condomínio	Valor do espaço comunitário.	Contínuo
IPTU	Valor do imposto.	Contínuo
Preço	Valor do imóvel.	Contínuo

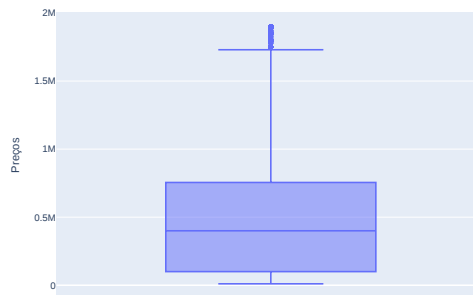
3.3. Pré-processamento de dados

Inicialmente, foram removidos anúncios duplicados. Para isso, foram utilizadas categorias contendo identificadores únicos, como a URL e o ID do anúncio. Essa etapa foi necessária, pois anúncios promovidos pela plataforma podem aparecer em múltiplas páginas. Já a detecção de *outliers* foi realizada utilizando a técnica de *boxplot*. Conforme ilustram as Figuras 2 e 3, tanto os preços quanto as áreas dos imóveis apresentaram valores discrepantes. Após a remoção desses *outliers*, a base de dados resultante passou a conter 51.000 amostras.

Para lidar com a coluna de bairros, foi aplicada a técnica de *one-hot encoding*, que transforma dados categóricos em valores booleanos, aumentando a quantidade de

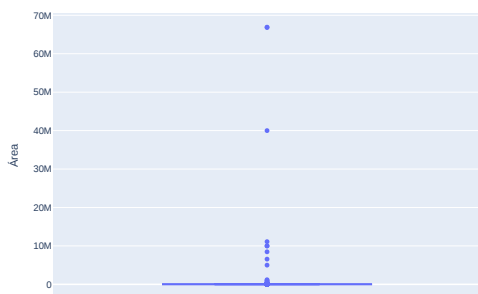


(a) Distribuição dos preços (antes).

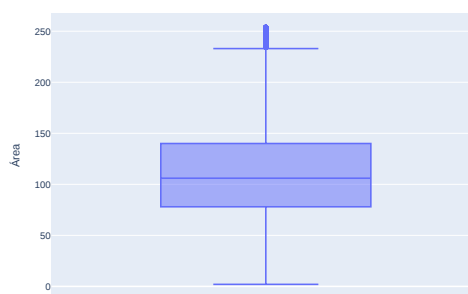


(b) Distribuição dos preços (depois).

Figura 2. Distribuição dos preços dos imóveis (antes à esquerda e depois à direita).



(a) Distribuição das áreas (antes).



(b) Distribuição das áreas (depois).

Figura 3. Distribuição das áreas dos imóveis (antes à esquerda e depois à direita).

colunas proporcionalmente ao número de bairros considerados. Cada amostra recebe 1 para o bairro correspondente e 0 para os demais, permitindo que o algoritmo de regressão interprete essa informação numericamente e reduzindo a variância nos valores preditos. Além disso, os atributos valores de IPTU e condomínio foram convertidos em booleanos. No caso do IPTU, o atributo é 1 se o valor informado ultrapassar o mínimo de isenção indicado pela Secretaria de Finanças de Belém. Já para o atributo condomínio recebe 1 quando o imóvel possui características de condomínio ou com valores declarados acima de zero. Essa decisão foi baseada em uma análise interna, que revelou inconsistências nos anúncios, como imóveis que apresentavam características de condomínio sem informar um valor correspondente e vice-versa.

3.4. Divisão da base de dados

A base de dados foi dividida em conjuntos de treino e teste, onde 70% das amostras foram reservadas para o treino e 30% para o teste. O conjunto de treino é utilizado para ensinar o modelo a mapear os valores de saída (*output*) a partir das entradas (*input*), enquanto o conjunto de teste contém amostras não vistas anteriormente, sendo usado para avaliar a acurácia do modelo. Essa divisão foi aplicada de forma consistente aos três modelos

analisados, garantindo comparabilidade entre os resultados. Na próxima seção, são apresentadas visualizações dos dados pós-processamento para permitir inferências sobre os valores extraídos da plataforma OLX, seguidas da descrição dos três modelos escolhidos para o precificador e da análise dos resultados obtidos, incluindo o comportamento aprendido por cada modelo.

4. Análise Exploratória dos Dados

Uma das maiores exclusividades que surgem com a análise de regiões específicas é a inclusão dos bairros como ponto focal para o comportamento dos preços. Por isso, é importante fazermos uma análise destas regiões separadamente. Diante desse contexto, a Figura 4 apresenta a frequência de anúncios por bairro de Belém e adjacências publicados na plataforma OLX no período considerado. Bairros com menos de 100 anúncios foram removidos da visualização. O gráfico de barras mostra que a maioria dos bairros que possuem mais anúncios também são aqueles que possuem a maior número de habitantes [Wilson 2022]. A lista, obtida em 2022, revela os seguintes número de habitantes por bairro: Guamá com 81.227, Tapanã com 69.295, Pedreira com 64.465, Marco com 59.847 e Marambaia com 58.870.

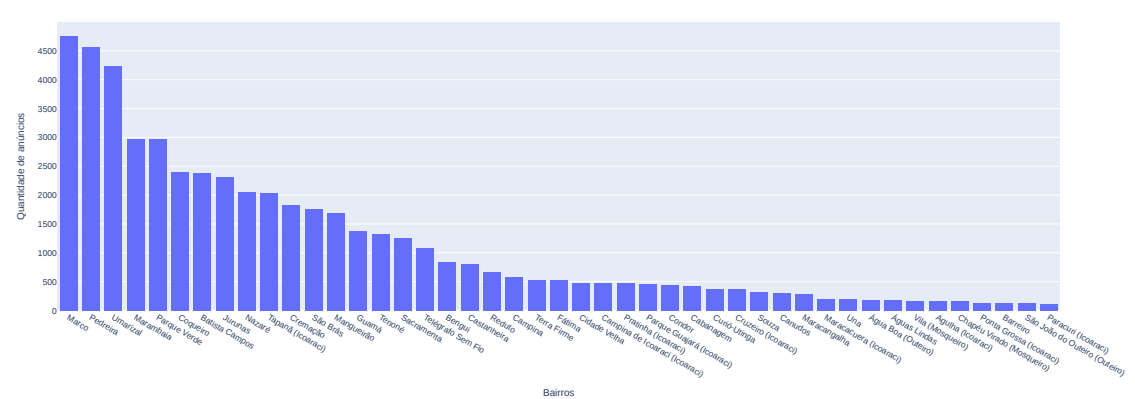


Figura 4. Frequência de anúncios por bairro na grande Belém e adjacências.

Portanto, é possível perceber que há uma correlação entre os bairros mais populosos e o número de anúncios na plataforma. Entretanto, há alguns bairros que, mesmo não aparecendo entre os bairros com mais habitantes, também possuem uma grande quantidade de anúncios, como os bairros Umarizal, São Brás, Batista Campos e Nazaré. Estes bairros, de acordo com a escola de Corretoria [IBRESP 2024] estão na lista dos bairros mais caros de Belém. Por conseguinte, dentre os bairros com mais anúncios, existem dois grupos que possuem uma grande quantidade de anúncios na plataforma. O primeiro conjunto é formado por bairros populosos, onde o grande número de residentes estimula uma maior procura e venda de anúncios de imóveis, devido ao intenso fluxo de cidadãos que ali residem. O segundo conjunto, por outro lado, é formado por bairros considerados nobres e que possuem boa procura por serem regiões comerciais.

Para dar suporte a esta hipótese, foi analisado com mais detalhes a divisão entre os dois conjuntos mencionados. Foi feita, então, uma discriminação dos cinco bairros com maior número de anúncios, buscando relacioná-los com seus preços. A Figura 5 apresenta um gráfico de barras que ilustra a relação entre a distribuição dos preços dos imóveis em

quartis (eixo das abscissas) e a frequência de imóveis à venda em cada quartil (eixo das ordenadas) para os cinco bairros mais populosos identificados na Figura 4.

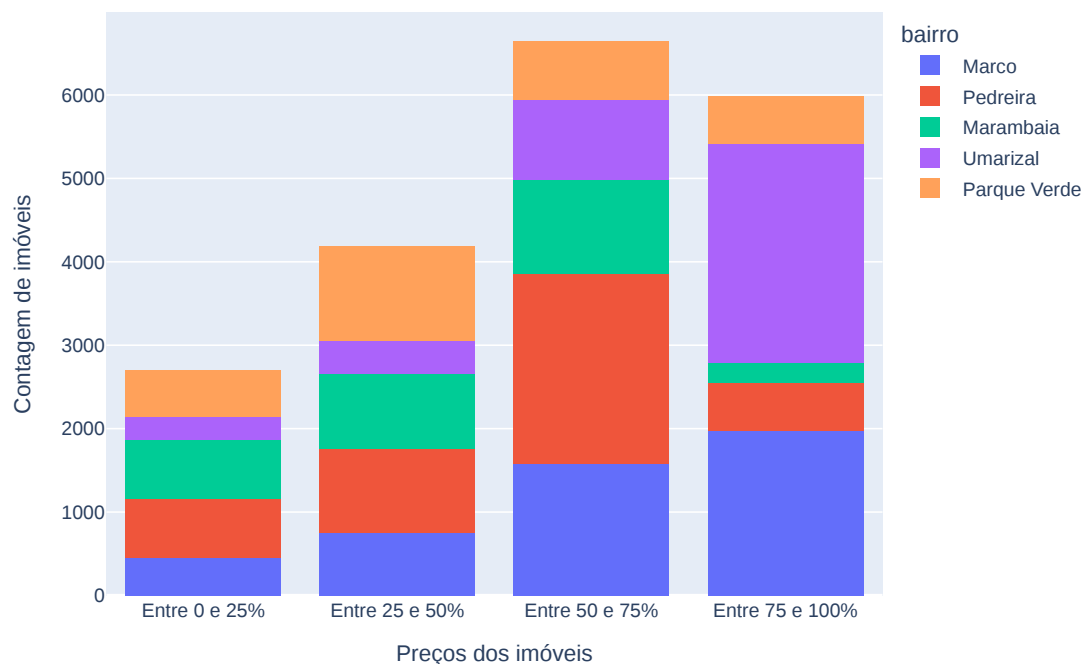


Figura 5. Relação entre os preços e a frequência de anúncios por quartil.

A visualização permite inferir que, no primeiro quartil, onde os imóveis apresentam os menores preços, predominam os bairros Marambaia e Pedreira, que também estão entre os mais populosos. Por outro lado, no último quartil, os bairros Umarizal e Marco concentram os imóveis de maior valor. O gráfico reforça a tendência mencionada anteriormente: bairros com maior população tendem a apresentar preços mais baixos. No entanto, o bairro do Marco se destaca como uma exceção, combinando preços elevados com uma grande densidade populacional.

Por fim, a fim de analisar melhor a relação entre os bairros e sua valorização dentro dos anúncios da OLX, a Figura 6 apresenta um *treemap* da base de dados, onde os tamanhos dos retângulos representam a média de preço dos imóveis à venda e a cor é referente à sua área média (tamanho).

A imagem permite uma análise coerente com o custo de vida em Belém. Os bairros com maior média de preço e menor área tendem a ser os mais caros para se residir, como é o caso dos bairros considerados nobres, tais como Umarizal e Nazaré. Além disso, dois aspectos do *treemap* chamam atenção: a distribuição de imóveis no Parque Guajará e em Água Boa, no distrito de Outeiro. Esses dois locais possuem áreas relativamente maiores em comparação com outros bairros, sendo que Água Boa se destaca pelo melhor custo-benefício em termos de preço por metro quadrado.

Treemap dos bairros pela média de preço (tamanho) e área (cor)



Figura 6. Relação entre a média de preço e a média de área (tamanho) do imóvel.

5. Resultados e discussão

Esta seção apresenta os resultados obtidos a partir do treinamento de três algoritmos de regressão: regressão linear múltipla, XGBoost e *Random Forest*. A decisão sob a escolha dos três algoritmos levou em consideração a proeminência dos mesmos nos artigos apresentados durante o levantamento da metodologia. Por isso, foram selecionados os modelos de XGBoost e *Random Forest*. Além deles, o modelo de regressão linear múltipla foi escolhido como baseline para que a escolha dos modelos mais robustos fosse justificada. As discussões são feitas sobre as métricas R^2 , MAPE e RMSE de cada um dos três modelos.

Para a avaliação dos modelos, a busca pelos melhores hiperparâmetros foi realizada por meio de *Grid Search* com validação cruzada com cinco dobras (*5-fold cross-validation*), utilizando a mesma divisão dos dados para todos os algoritmos. No caso do modelo de regressão *Random Forest*, os melhores hiperparâmetros encontrados foram: *bootstrap* foi habilitado, as árvores não possuem profundidade máxima, porém a divisão mínima para um nó é de cinco amostras (caso este número seja menor que cinco, o nó se torna a última folha e não possui filhos) e utilização de 200 árvores na floresta. Para o modelo XGBoost, os melhores resultados foram obtidos ao utilizar a função objetivo voltada para regressão com erro quadrático, uma taxa de aprendizado de 0,1, profundidade máxima de nove níveis e um total de 325 árvores. Esses hiperparâmetros foram definidos após uma busca sistemática entre outras combinações dos hiperparâmetros citados acima, e os modelos foram avaliados com base nas métricas R^2 , RMSE e MAPE.

Levando isso em consideração, a Tabela 3 apresenta a performance de cada modelo. Estes resultados foram obtidos através de uma média de 50 iterações, onde a base de treinamento e teste mudaram a partir de cada uma das iterações. Os valores correspondem a base de teste, contendo 30% da quantidade de dados total que foram reservados somente para aqueles não vistos durante a fase de treinamento.

Modelo	R^2	MAPE	RMSE
Regressão Linear Múltipla	0.703	84.0%	234.773
XGBoost	0.903	27.8%	133.704
<i>Random Forest Regressor</i>	0.902	26.0%	134.785

Tabela 3. Comparação de performance dos modelos.

Iniciando com o algoritmo de menor custo computacional, a regressão linear múltipla apresentou um R^2 score de 0.7, mostrando uma baixa variância dentro desse escopo. Entretanto, a porcentagem MAPE apresentada pelo algoritmo foi de 84%, significando que os valores preditos estavam quase duas vezes mais altos, evidenciando uma baixa resistência para possíveis *outliers* e ruídos. Já o algoritmo XGBoost obteve resultados relativamente melhores, obtendo uma menor variância e um valor MAPE menor que aquele apresentado sob a regressão múltipla. Por fim, o algoritmo de *Random Forest* obteve o melhor resultado, trazendo um valor de R^2 de 0.9 e um MAPE de 26.0%.

Para fazer uma melhor representação visual dos resultados obtidos entre os três algoritmos, a Figura 7 apresenta a evolução das métricas R^2 , RMSE e MAPE ao longo de 50 iterações. É possível observar, portanto, que os modelos de *Random Forest* e *XGBoost* se mantiveram consistentemente superior nas métricas em relação ao modelo de Regressão Linear Múltipla.

Com base no desempenho dos modelos testados, e levando em consideração que os algoritmos de *Random Forest* e *XGBoost* apresentaram resultados similares, os resultados destes algoritmos foram selecionados para a análise dos erros residuais nas predições dos valores dos imóveis. A Figura 8 ilustra a relação entre o erro residual e o preço dos imóveis. No eixo das abscissas, está o preço real de cada imóvel listado na plataforma, enquanto no eixo das ordenadas, encontra-se a diferença entre o valor predito e o valor real.

O gráfico da Figura 8a revela que a variância do modelo acompanha o aumento do preço, mantendo um padrão consistente. Observa-se que, para imóveis de maior valor, o modelo treinado tende a superestimar o preço, enquanto, para aqueles com preços mais baixos, ocorre uma subestimação. Ademais, a distribuição dos erros residuais tem o formato retangular em perspectiva, o que indica que a média de erros está proporcionalmente aumentando conforme o preço dos imóveis aumenta. Um dos motivos para o desempenho inferior em imóveis mais caros pode ser atribuído à menor quantidade de amostras em residências de alto valor. Por conseguinte, o algoritmo de *RF* faz generalizações mais inconsistentes do que aquelas realizadas nos imóveis mais baratos.

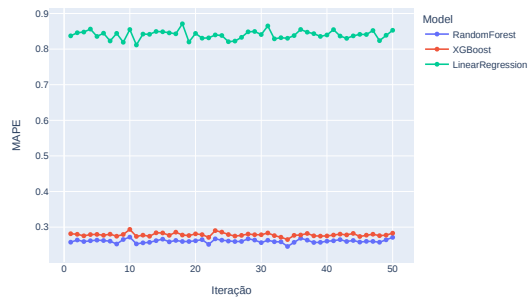
Por outro lado, o algoritmo *XGBoost*, da Figura 8b apresenta uma performance similar ao algoritmo apresentado anteriormente, porém com generalizações menos precisas nos preços mais baixos, enquanto faz predições mais próximas do preço real para os imóveis mais caros.

Variação do R^2 por Iteração



(a) Performance em R^2 por execução.

Variação do MAPE por Iteração



(b) Performance em MAPE por execução.

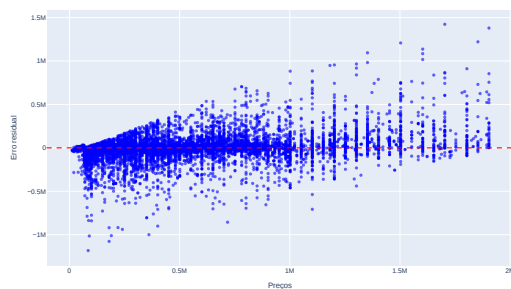
Variação do RMSE por Iteração



(c) Performance em RMSE por execução.

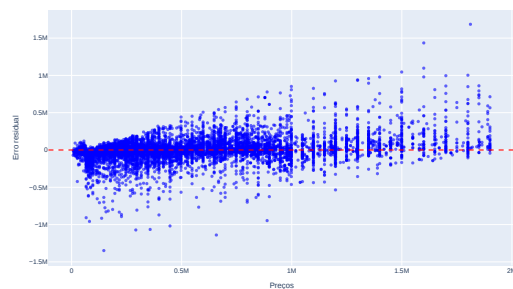
Figura 7. Gráfico de linha das métricas por iteração.

Erro residual vs. Preços com Random Forest Regressor



(a) Random Forest.

Erro residual vs. Preços com XGBoost Regressor



(b) XGBoost.

Figura 8. Relação entre os erros residuais e preço dos imóveis à venda.

6. Conclusão

Este trabalho explorou a precificação de imóveis à venda em Belém do Pará com base nos atributos físicos e de localização descritos nos anúncios da plataforma OLX. A análise exploratória dos dados permitiu avaliar diferentes aspectos do mercado imobiliário, enquanto a aplicação de algoritmos de aprendizado de máquina possibilitou a criação de modelos de regressão para estimar os preços dos imóveis. Entre os modelos testados, o *Random Forest* apresentou o melhor desempenho na previsão dos valores.

No entanto, a precificação de imóveis envolve múltiplas variáveis que vão além

das características diretamente descritas nos anúncios. Observou-se que o modelo tende a superestimar o valor de imóveis mais caros e a subestimar o preço daqueles de menor valor, indicando que fatores não considerados na modelagem podem influenciar os resultados. Aspectos como proximidade a escolas, supermercados e segurança pública podem ter um impacto significativo na valorização ou desvalorização de um imóvel, tornando essencial uma abordagem mais abrangente.

Dessa forma, futuras investigações podem explorar esses fatores adicionais, promovendo uma análise interdisciplinar que relacione a precificação de imóveis ao impacto social e urbano. Ao incorporar novas fontes de dados e perspectivas, este estudo pode ser expandido para oferecer uma compreensão mais precisa e aprofundada do mercado imobiliário e de suas dinâmicas.

Referências

- Afonso, B., Melo, L., Oliveira, W., Sousa, S., and Berton, L. (2019). Housing prices prediction with a deep learning and random forest ensemble. In *Anais do XVI Encontro Nacional de Inteligência Artificial e Computacional*, pages 389–400, Porto Alegre, RS, Brasil. SBC.
- Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., and Wirth, R. (2000). *CRISP-DM 1.0: Step-by-step Data Mining Guide*. SPSS Inc., Chicago, IL. <https://www.the-modeling-agency.com/crisp-dm.pdf>.
- Dabreo, S., Rodrigues, S., Rodrigues, V., Shah, P., Student, F., and Conceicao (2021). Real estate price prediction.
- Firpo, R. (2024). Como a inteligência artificial começou a revolucionar o mercado de imóveis. <https://veja.abril.com.br/coluna/real-estate/como-a-inteligencia-artificial-comecou-a-revolucionar-o-mercado-de-imoveis>. Acessado em 30-03-2025.
- IBRESP (2024). Confira os bairros com imóveis mais valorizados em belém. <https://www.ibresp.com.br/blogs/2023/confira-bairros-com-imoveis-mais-valorizados-em-belem/>. Acessado em: 07-02-2025.
- Mysore, S., Muthineni, A., Nandikandi, V., and Behera, S. (2022). Prediction of house prices using machine learning. *International Journal for Research*, 10.
- Panchal, Y., Mer, M., and Ghosh, A. (2022). Residential property price prediction using machine learning: Makansetu. In *2022 International Seminar on Application for Technology of Information and Communication (iSemantic)*, pages 257–262.
- Truong, Q. H., Nguyen, M. V., Dang, H., and Mei, B. (2020). Housing price prediction via improved machine learning techniques. *Miscellaneous*, 174:433–442.
- Wang, F., Zou, Y., Zhang, H., and Shi, H. (2019). House price prediction approach based on deep learning and arima model. In *2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT)*, pages 303–307.
- Wilson, I. (2022). Censo 2022 - guamá é o bairro mais populoso de belém. <https://www.oliberal.com/para/censo-2022-guama-e-o-bairro-mais-populoso-de-belem-confira-os-dez-primeiros-1.887339>. Acessado em: 07-02-2025.
- Zaki, J., Nayyar, A., Dalal, S., and Ali, Z. H. (2022). House price prediction using hedonic pricing model and machine learning techniques. *Wiley Online Library*, 34(27).