Análise de Modelos para Precificação de Imóveis em São Paulo

Júlia Ronquetti Rodrigues [1]
Rogério de Oliveira [1]
[1]Faculdade de Computação e Informática (FCI)
Universidade Presbiteriana Mackenzie São Paulo, SP – Brasil

10389531@mackenzista.com.br, rogerio.oliveira@mackenzie.br

Resumo

Este trabalho avalia diferentes modelos para a de predição de preço de venda de imóveis anunciados em plataformas online. Particularmente, compara o uso de modelos de Regressão Linear, indicados como modelo padrão pela ABNT (Associação Brasileira de Normas Técnicas), com modelos de Aprendizado de Máquina adicionando textos descritivos dos imóveis nos anúncios e suas imagens. No uso das imagens como preditoras do preço, adota uma abordagem alternativa ao uso direto das imagens com redes convolucionais, empregando objetos de interesse detectados e quantificados, para criar um vetor de objetos das imagens como preditor.

Palavras-chave: Modelo preditivo; Aprendizado de Máquina; Imóveis.

Abstract

This work evaluates different models for predicting the sale price of properties advertised on online platforms. In particular, it compares the use of linear regression models, indicated as a standard model by ABNT (Brazilian Association of Technical Standards), with machine learning models adding descriptive texts of the properties in the advertisements and their images. In using images as price predictors, it adopts an alternative approach to the direct use of images with convolutional networks, employing detected and quantified objects of interest, to create a vector of objects from the images as a predictor.

Keywords: Predictive Model; Machine Learning; Real Estate.

1 Introdução

Determinar o preço dos imóveis desempenha um papel importante no mercado imobiliário, afetando as opções de compra, venda e investimento. Esse trabalho avalia diferentes modelos para predição do preço de venda de imóveis anunciados em plataformas online para a cidade de São Paulo, envolvendo dados quantitativos e imagens dos anúncios.

Em São Paulo o preço médio do metro quadrado chegou a 7.157 reais no primeiro trimestre de 2024, segundo o Relatório QuintoAndar sobre Compra e Venda, quanto pelos impactos das estimativas imprecisas no setor imobiliário (QUINTOANDAR, 2024). A falta de definição de preços justos afeta as transações comerciais, estendendo o tempo médio de venda de imóveis no Brasil para cerca de 16 meses, enquanto nos Estados Unidos, esse período é de cerca de 3 meses, segundo a Associação Brasileira de Incorporadoras Imobiliárias (Abrainc) e a plataforma Zillow (UMBRASAS, 2019).

Os dados desse estudo, foram coletados de uma plataforma de *maketplace* de grande penetração no mercado. Foram selecionados anúncios da cidade de São Paulo, no período de abril a junho de 2024, num total de 23.433 imóveis e 605K imagens. Esses dados encontram-se publicamente disponíveis em (RODRIGUES; OLIVEIRA, 2024), assim como todos os códigos e resultados em (RODRIGUES, 2024).

Esse trabalho traz, assim, três principais contribuições:

1. Fornece uma grande base de dados aberta de preços de anúncios de imóveis em plataformas online, incluindo dados dos imóveis e imagens, para exploração e desenvolvimento de outros modelos na área.

2. Compara diferentes modelos, incluindo modelos de regressão e de Aprendizado de Máquina, na previsão de preços.

3. Explora o uso conjunto de texto dos anúncios e de imagens, a partir de objetos nelas detectados, como variáveis preditoras para os preços.

2 Referencial Teórico

A avaliação de preços de imóveis no Brasil é orientada pela Associação Brasileira de Normas Técnicas (ABNT), que recomenda o uso de modelos econométricos para a avaliação de propriedades urbanas, conforme a NBR 14653-2 – "Avaliação de imóveis urbanos" (Associação Brasileira de Normas Técnicas (ABNT), 2011). Os avaliadores coletam dados de imóveis semelhantes, considerando, fatores como área, localização e número de quartos, para estimar o valor de mercado de uma propriedade. No entanto, esses métodos econométricos tradicionais possuem limitações, especialmente em relação à capacidade de capturar a complexidade do mercado imobiliário em cenários mais dinâmicos.

Com o avanço das técnicas de Aprendizado de Máquina, diversos estudos têm buscado soluções inovadoras para aprimorar a avaliação de preços de imóveis, superando as limitações dos métodos econométricos tradicionais recomendados pela ABNT. Esses trabalhos exploram o potencial dessas novas abordagens aplicadas a diferentes mercados imobiliários, utilizando conjuntos de dados diversificados e testando diversos modelos para alcançar maior precisão nas estimativas. Modelos como Random Forest, XGBoost e Deep Learning são capazes de identificar padrões complexos em grandes volumes de dados (WOHLWEND, 2023), permitindo uma estimativa mais concisa dos preços de imóveis. Esses

algoritmos são particularmente eficazes ao lidar com variáveis que apresentam relações não lineares.

A Tabela 1 apresenta alguns dos estudos recentes da aplicação de diferentes modelos na predição preços de imóveis juntamente com as respectivas métricas de desempenho obtidas nesses trabalhos.

Tabela 1 – Estudos recentes de predição de preços de imóveis, com a predominância de modelos de aprendizado de máquina.

Autor(es) | Ano	Quantidade de Imóveis	Modelo Ganhador	MAE	MAPE (%)	R^2
(CHOU; FLESHMAN; TRUONG, 2022)	13.220	PSO-Bagging-ANNs	2.273.866	11,59	0,97
(MOHAMED; IBRAHIM; HAGRAS, 2023)	1.990	Deep Learning	85.460	9,50	0,90
(NERI, 2020)	18.796	Random Forest	40.287,09	8,21	0,93
(MARZAGãO; FERREIRA; SALES, 2021)	15.552	Random Forest	35.463,24	8,16	0,93
(TEKIN; SARı, 2022)	22.176	XGBoost	21.881	21,.81	0,91
(ALENCAR; CAURIN, 2022)	5.510	Random Forest	73.346,39	13,34	0,88
(HANSSON; HOLMQVIST; ANDERSSON, 2023)	55.636	XGBoost	440.463	8,42	0,92

Esses estudos evidenciam o uso de *ensemble models* (Random Forest e XGBoost), também empregados aqui. O número médio de casos, 18.982 imóveis, sugere um número adequado de instâncias empregadas neste trabalho (23.433), sendo os resultados também compatíveis (MAPE médio de 11.5%, para 6% obtido no modelo final deste estudo). Esses trabalhos, entretanto, não empregam imagens, exceto (MARZAGãO; FERREIRA; SALES, 2021) que apresenta resultados preliminares e sugestões para uso de imagens nos modelos.

O modelo de avaliação imobiliária mais amplamente utilizado no mercado é o Zestimate, desenvolvido pela plataforma Zillow. Com um erro absoluto mediano de aproximadamente 7.9%, o Zestimate é baseado em um grande volume de dados, abrangendo mais de 110 milhões de propriedades dos Estados Unidos (FONTINELLE, 2024). Esse modelo utiliza muitas variáveis para suas previsões, como preços de transações reais, histórico de vendas, dados fiscais detalhados (como os valores de impostos pagos), além de características das propriedades e informações demográficas locais. A quantidade de dados e a complexidade das variáveis incluídas permitem que o Zestimate forneça estimativas mais próximas do valor de mercado das propriedades em comparação com modelos que se baseiam apenas em preços de oferta ou dados limitados.

2.1 Modelos lineares

Modelos de regressão são amplamente utilizados para solucionar diversos problemas de estimativa de valores, como a predição de preços, vendas, consumo de energia, etc. (KUMARI; YADAV, 2018).

A Regressão Linear é uma técnica estatística que prevê uma variável dependente (ou alvo) com base em uma ou mais variáveis independentes (preditoras) a partir de um modelo linear:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \epsilon$$

onde y é a variável dependente que se quer prever, β_i são os coeficientes da regressão, x_i as variáveis independentes e ϵ representa o erro do modelo.

Homocedasticidade. A aplicação desse modelo tem alguns requisitos. Uma das mais importantes é a homocedasticidade, que requer que a dispersão dos dados permaneça

aproximadamente igual para todo o conjunto (YANG; TU; CHEN, 2019). Esse requisito muitas vezes é apenas parcialmente verificado e para reduzir a heterocedasticidade podem ser aplicadas transformações às variáveis alvo e preditoras do modelo (ex., $\ln y$, $1/y$) (LOPES, 2024), o que é bastante aplicado em modelos de preços. A qualidade do modelo também pressupõe que os resíduos (os erros do modelo) tenham uma distribuição normal.

Coeficiente de Determinação. O Coeficiente de Determinação (R^2) é medida estatística usada para avaliar a qualidade do ajuste de um modelo aos dados. Um valor próximo a 1 indica que o modelo explica quase toda a variabilidade da variável dependente, enquanto um valor próximo a 0 indica que o modelo explica muito pouco da variabilidade observada. O R^2 elevado, entretanto, não garante que o modelo seja adequado para predições, podendo apresentar sobreajuste ou variáveis preditoras não significativas.

Termos de interação. Termos de interação permitem examinar como a relação entre duas ou mais variáveis independentes influencia a variável alvo, o que permite avaliar o impacto de uma variável preditora dependente do nível de outra. Os termos de interação podem ser incluídos a partir do produto das variáveis preditoras que se espera interação:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon$$

Neste estudo, diferentes modelos de Regressão Linear são empregados, incluindo-se modelos com variáveis transformadas e termos de interação.

2.2 Modelos de Aprendizado de Máquina

Modelos de Aprendizado de Máquina vêm sendo utilizados com sucesso em uma série de problemas, do reconhecimento de imagens, recomendação de produtos à personalização e análises preditivas (JANIESCH; ZSCHECH; HEINRICH, 2021).

Aprendizado Supervisionado. Predições de preços podem ser realizadas com modelos de Aprendizado Supervisionado, que utilizam dados rotulados. Nesses modelos, características dos imóveis, ou dados históricos de vendas, variáveis de entrada do modelo (preditoras) são mapeadas para os preços (rótulos) que se quer prever.

Modelos de Aprendizado Supervisionado podem realizar tanto tarefas de classificação (quando as saídas ou rótulos assumem um número discreto de valores) como de regressão (quando os rótulos assumem valores contínuos) (ZHOU, 2018) como no caso do preço dos imóveis.

O esquema geral de funcionamento dos modelos supervisionados é representado na Figura 1. Um conjunto de amostras de dados rotulados (conjunto de treinamento) é empregado como entrada para uma classe de modelos (por exemplo, os modelos de Regressão Linear). Esse modelo tem então seus parâmetros (os coeficientes, no caso da Regressão Linear) ajustados por algum método de otimização (método de mínimos quadrados, OLS, para a Regressão Linear) para buscar os melhores parâmetros que forneçam o menor erro entre as saídas obtida pelo modelo e as desejadas (o rótulo dos dados).

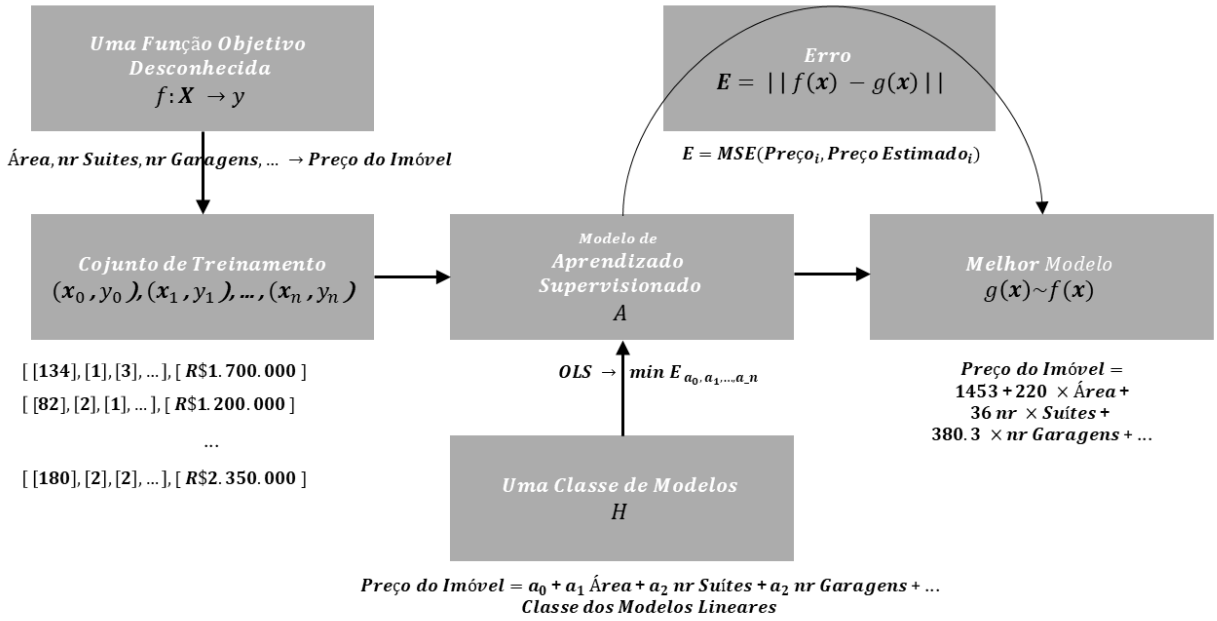

Figura 1 – Esquema Geral do Aprendizado Máquina, exemplificado através de um modelo de Regressão Linear para predição de Preços de Imóveis.

Modelos Supervisionados e *ensemble*. Há vários modelos de Aprendizado de Máquina Supervisionado, como Árvores de Decisão, K-Vizinhos mais Próximos, Máquinas de Vetores de Suporte, incluindo diferentes modelos de Regressão Linear, cada um diferindo nas técnicas empregadas para se obter o melhor mapeamento das entradas e saídas dos dados e seus parâmetros. Esses algoritmos são amplamente conhecidos e podem ser encontrados em diversos livros texto como em (BISHOP, 2006) e (MITCHELL; MITCHELL, 1997). Modelos *ensemble*, como Random Forest e XGboost, são métodos em Aprendizado de Máquina que combinam múltiplos modelos individuais para fazer previsões. Eles agregam vários modelos (por exemplo, várias Árvores de Decisão no caso dos Random Forest) e, em geral, resultam apresentam uma melhor performance do modelo em termos de precisão, generalização e robustez. Esses modelos encontram-se entre os de melhores resultados para predição de preços (Tabela 1) e também são os modelos que desempenham melhor neste estudo. Para mais detalhes sobre modelos *ensemble* e suas técnicas de combinação empregadas ver (BISHOP, 2006).

2.3 Avaliação dos Modelos

Existem vários aspectos e diferentes métricas para a avaliação de modelos de regressão. As métricas mais comuns e de interesse residem nas métricas de erro (a diferença entre os valores estimados e os esperados que constam do conjunto de exemplos, seja para um modelo de Regressão ou algum outro modelo de Aprendizado de Máquina mais elaborado).

Métricas. As métricas mais comuns são o Coeficiente de Determinação (R^2), o Erro Médio Quadrático (MSE) e sua raiz (RMSE), o Erro Médio Absoluto (MAE),

o Erro Mediano Absoluto (MedAE) e o Erro Percentual Absoluto Médio (MAPE). O Coeficiente de Determinação (^{2}) é mais importante para modelos de Regressão Linear, o MSE e RMSE são medidas úteis na comparação de resultados de diferentes modelos, enquanto o MAPE e MedAE, fornecem medidas mais práticas apresentando o percentual de erro (independente, portanto, de escala e da quantidade de casos) e o valor do erro mais frequentemente esperado (no caso de preços, o valor absoluto em reais). Essas métricas são amplamente conhecidas, mas uma definição detalhada pode ser encontrada em (CHICCO; WARRENS; JURMAN, 2021). A depender do objetivo de cada etapa, diferentes métricas são empregadas ao longo deste estudo.

Conjuntos de treinamento e Teste. No contexto do Aprendizado de Máquina, a prática geral é dividir o conjunto de dados em dois subconjuntos aleatórios dos dados, um de treinamento e outro teste (em geral $20 - 30\%$ dos dados). O conjunto de treinamento é usado para se ajustar os parâmetros do modelo, permitindo que ele aprenda os padrões e as relações entre as variáveis preditoras e a variável-alvo. O modelo ajustado (treinado) é então aplicado ao conjunto de teste, que inclui novos dados que não foram empregados no treinamento. Isso permite avaliar o erro, mas também a capacidade de generalização do modelo, evitando assum problemas de sobreajuste (*overfitting*) do modelo.

Validação Cruzada.A validação cruzada (*cross Validation*) é ainda utilizada para maior assertividade das avaliações. Uma única divisão entre conjuntos de treinamento e teste pode levar a resultados diferentes para outros conjuntos de teste. Na Validação Cruzada o conjunto de dados é dividido em k partições (conjuntos de validação). O modelo é então treinado k vezes, utilizando a cada iteração, uma partição diferente como conjunto de teste e os outras como treinamento. Ao final das k iterações, calcula-se a média das métricas de erro obtidas em cada partição, o que fornece uma medida de erro mais próxima da que se pode esperar, do que o erro para um conjunto de teste específico (BATES; HASTIE; TIBSHIRANI, 2021).

Assim, no contexto do Aprendizado de Máquina, diferentes modelos são selecionados a partir do menor erro médio sobre as partições do conjunto de treinamento, que representa o ajuste dos dados ao modelo. Por outro lado, modelos estatísticos em geral, como modelos lineares, são avaliados mais comumente sobre todo conjunto de dados, pois se assume que o dado se ajusta ao modelo e o erro, medido sobre todo o conjunto de dados, representa o melhor ajuste do modelo. Assim, diferentes métricas de erro podem ser obtidas, seja sobre o conjunto total dos dados, sobre os dados de validação ou sobre os dados de teste. Neste estudo, sendo particularmente de interesse o modelo que melhor se ajusta aos dados, como também por concisão, são apresentadas aqui as **métricas de erro sobre todo o conjunto de dados**. Entretanto, as demais métricas, sobre dados de teste e de validação, podem ser consultadas em detalhe em (RODRIGUES, 2024), e foram consideradas ao longo da análise para a seleção de modelos (por exemplo, o modelo de Árvore de Decisão, embora com menor erro de treinamento, é sobreajustado, e descartado nos resultados a seguir).

3 Metodologia

Esta seção apresenta os principais aspectos da coleta e preparação dos dados, bem como os modelos e técnicas de ajuste empregadas.

3.1 Dados

O desenvolvimento dos modelos de previsão de preços de imóveis exige a extração e preparação de dados utilizados como entrada. Os dados finais para treino e teste passam por diversas etapas, conforme ilustrado na Figura 2.

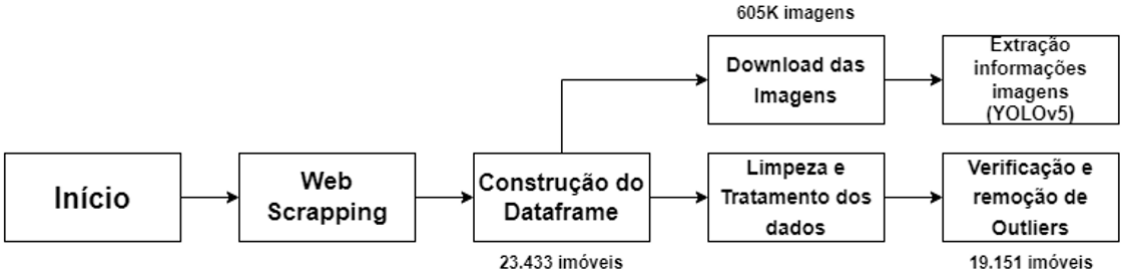

Figura 2 – Diagrama de extração e preparação dos dados para uso nos modelos preditivos.

Web Scrapping. Os dados utilizados neste estudo são extraídos de um site (um *marketplace*) com anúncios de venda de imóveis. Para o *Web Scraping* (MITCHELL, 2018) de extração são selecionados anúncios somente da cidade de São Paulo. O processo emprega a bibliotecas BeautifulSoup, para processar o HTML das páginas, e Selenium, que lida com os elementos dinâmicos das páginas. A extração completa dos dados levou aproximadamente 15 horas. As informações extraídas incluem:

- Título
- Preço
- Publicação destaque?
- Categoria
- Tipo
- Preço condomínio
- Área (m2)
- Número de quartos
- Número de banheiros
- Vagas na garagem
- IPTU
- Zona da cidade
- Bairro
- Academia?
- Elevador?
- Permitido animais?
- Piscina?
- Portaria?
- Salão de festas?
- Condomínio fechado?
- Segurança 24h?
- Portão eletrônico?
- Área murada?
- Área de serviço?
- Armários na cozinha?
- Armários no quarto?
- Churrasqueira?
- Mobiliado?
- Quarto de serviço?
- Ar condicionado?
- Porteiro 24h?
- Varanda?

Após a extração dos dados, o dataframe resultante apresentou 23.433 imóveis da cidade de São Paulo e 33 atributos.

Imagens. Também foram extraídas 605 mil imagens dos anúncios dos imóveis. As imagens foram processadas com o modelo pré-treinado YOLOv5, para detectar 17 categorias de objetos de interesse, dentre as 80 classes padrão reconhecidas pelo YOLO, como TV, mesa, sofá, poltrona, mesa, pia, pessoa, etc. Um vetor binário, indicando a presença ou não desses objetos no conjunto de imagens, é adicionado aos atributos de cada imóvel para uso nos modelos preditivos.

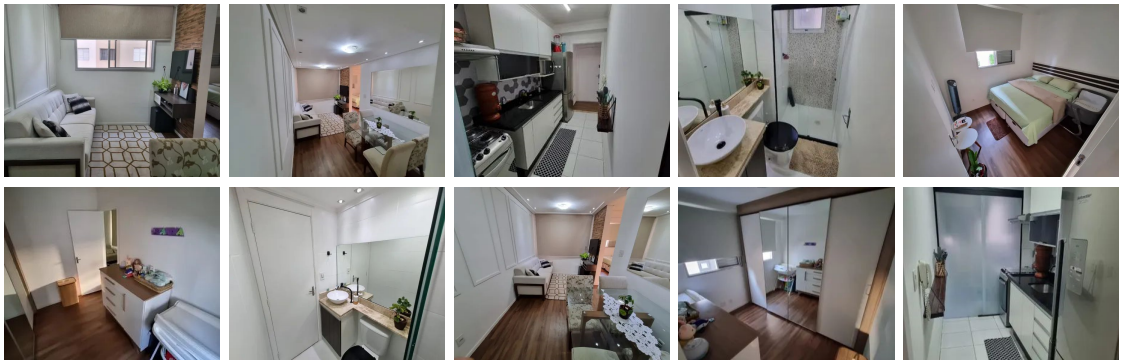

Figura 3 – Imagens do anúncio de um imóvel. A partir do conjunto de imagens é criado um vetor binário dos objetos identificados.

	price	location	condominio	area – util	quartos	...	person	bicycle	car	motorcycle	chair	couch	pottedplant	bed	...
0	540000	SãoPaulo, BelaVista	831	80	2	...	0	0	0	0	1	0	0	0	...
1	2700000	SãoPaulo, JardimPaulista	1760	180	3	...	1	0	1	0	1	1	1	1	...
2	3192212	SãoPaulo, Santana	0	224	4	...	0	1	1	1	1	1	1	1	...
3	1829980	SãoPaulo, SantaCecília	2883	234	3	...	0	0	0	0	1	0	1	1	...
4	393617	SãoPaulo, VilaBancáriaMunhoz	220	60	2	...	0	0	0	0	1	1	1	1	...

Figura 4 – Dados dos imóveis e o respectivo vetor binário identificando a presença dos objetos selecionados no conjunto de imagens do anúncio.

Limpeza e tratamento dos dados. Dados inconsistentes, incompletos ou duplicados são eliminados para maior qualidade das informações. São excluídos dados ausentes para *Condomínio, IPTU e Banheiros* (prefazem menos 1% dos dados). 3.884 imóveis apresentam outliers nos preços e são removidos (método IQR, Interquartile Range). São também excluídos dados inconsistentes: 20 imóveis com preço do m^2 dez vezes menor que o preço médio, preços menores que R$ 50mil (20 imóveis) e com área útil menor que $18m^2$ (128). A final, do conjunto de dados inicial de 23.433 imóveis, 19.151 imóveis foram selecionados.

Para as variáveis categóricas, *Categoria* e *Zona da Cidade*, é feito *hot encode*, com a exclusão da primeira categoria para evitar multicolinearidade. A normalização z-score é aplicada às variáveis preditoras nos modelos de Aprendizado de Máquina, mas não é necessária para os modelos de Regressão Linear.

price	location	condominio	area – util	quartos	banheiros	garagem	iptu	academia	zona	...
540000	SãoPaulo, BelaVista	831	80	2	2	0	110	0	CENTRO	...
2700000	SãoPaulo, JardimPaulista	1760	180	3	2	2	61	0	SUL	...
3192212	SãoPaulo, Santana	0	224	4	5	4	0	0	NORTE	...
1829980	SãoPaulo, SantaCecília	2883	234	3	2	2	883	0	CENTRO	...
393617	SãoPaulo, VilaBancáriaMunhoz	220	60	2	2	1	0	0	NORTE	...

Figura 5 – Amostra dos dados finais (19.151 imóveis) a serem empregados nos modelos, após o pré-processamento dos dados.

Resumo dos dados. A tabela 2 apresenta a distribuição de imóveis por região da cidade e indica uma boa representatividade das diferentes regiões. A Zona Sul é ligeiramente sub-representada, correspondendo a 17,62% do total. A Zona Oeste apresenta os preços mais altos, influenciado pela presença de imóveis com áreas maiores na região.

Tabela 2 – Características dos imóveis por região: Quantidade, área e preço médio.

ZONA	Qtde. Imóveis	Área Média	Preço Médio (M)	Distribuição (%)
CENTRO	3917	110.37	1.11	20.45
LESTE	3907	104.76	0.77	20.40
NORTE	4067	103.92	0.68	21.24
OESTE	3886	114.99	1.32	20.29
SUL	3374	103.91	1.16	17.62

A distribuição de preços dos imóveis está concentrada em faixas de valores mais acessíveis (Figura 6). A maioria dos imóveis tem preços entre 300 mil e 600 mil reais, enquanto uma pequena fração atinge valores mais elevados. Essa distribuição assimétrica reflete a predominância de imóveis de menor custo no conjunto de dados.

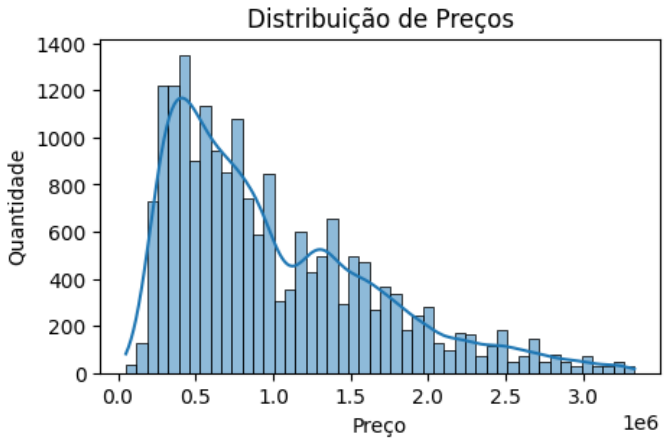

Figura 6 – Distribuição de Preços dos Imóveis

3.2 Treinamento e Seleção dos modelos

Neste estudo diversas abordagens para a previsão dos preços dos imóveis são exploradas, incluindo modelos de Regressão Linear e algoritmos de Aprendizado de Máquina, com incorporação de características das imagens e textos dos anúncios. As abordagens empregadas são descritas a seguir.

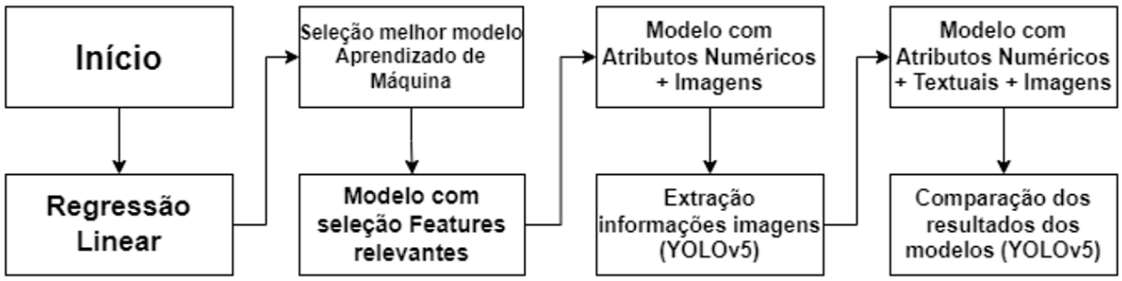

Figura 7 – Diagrama de seleção e aplicação dos modelos preditivos.

Modelos de Regressão Linear. Os modelos de regressão linear são ajustados com todas as variáveis preditoras contínuas e categóricas, com exceção do vetor de objeto das imagens e o texto do título dos anúncios. No ajuste do modelo é considerado a análise da significância dos coeficientes, mantendo-se apenas os atributos com coeficientes significativos ($p < 0.05$). Para reduzir a heterocedasticidade associada aos preços, avalia-se um modelo com a transformação logarítmica Preço $\times$ Área Útil. A capacidade do modelo é ainda melhorada com a adição de variáveis de interação, Quartos $\times$ Área útil, Área útil $\times$ banheiros e Área útil $\times$ Categoria Casas.

Modelos de Aprendizado de Máquina. A aplicação dos Modelos de Aprendizado de Máquina emprega como preditores, além das variáveis preditoras contínuas e categóricas empregadas antes nos modelos de Regressão Linear, o vetor dos objetos detectados das imagens e o texto de título dos anúncios (um texto em geral mais longo com características sobre o imóvel ofertado) em diferentes combinações.

Para os atributos numéricos são avaliados os seguintes modelos: Decision Tree; K-Nearest Neighbors; Ridge e Lasso Regressor; os modelos lineares robustos HuberRegressor, TheilSenRegressor e RANSACRegressor; e os modelos ensemble Random Forest e Gradient Boosting. Para o Random Forest e Gradient Boosting, o número de estimadores é ajustado para 1000. Os modelos são avaliados em uma validação cruzada de 5 partições e com 20% de dados de teste. A métrica empregada para a seleção é o RMSE. As variáveis com ganho de informação (*mutual information*) próximo ≈ 0 são excluídas e o melhor modelo selecionado nessa fase (o Random Forest, ver na seção de resultados) é empregado para a predição envolvendo os demais atributos de imagens e título dos anúncios.

São explorados assim, com os modelos de Aprendizado de Máquina, os seguintes diferentes conjuntos de atributos preditores:

- **Preditores originais, convertidos para numéricos**
- **Preditores numéricos + vetor de objetos (YOLO)**
- **Preditores numéricos + título (Texto)**
- **Preditores numéricos + título (Texto) + vetor de objetos (YOLO)**

Texto e encode. O título do anúncio traz frequentemente informações complementares que parecem ser significativas, como em *Apartamento MOBILIADO com ÓTIMA Localização*. Esse texto é, portanto, incorporado às variáveis preditoras do modelo. Um vetor de 1024 é empregado para fazer o encode dos textos empregando o método TfidfVec-

torizer (método do scikit-learn para encode TF-IDF, *Term Frequency-Inverse Document Frequency*).

4 Resultados e discussão

A seguir são apresentados e discutidos os resultados dos diferentes modelos, da Regressão Linear aos Modelos de Aprendizado de Máquina, e das diferentes combinações de preditores.

4.1 Regressão linear

Os resultados dos modelos de Regressão Linear estão apresentados na Tabela 3. Dos atributos iniciais, embora sem impacto significativo nos resultados, foram removidos os preditores em coeficientes significativos (p-value > 0.05, Tabela 4).

Tabela 3 – Resultados dos modelos de Regressão Linear.

Modelo Regressão linear	R2	RMSE	MAE	MedAE	MAPE
Todos os atributos	0.68	372598.11	270493.26	200519.62	0.34
Atributos significativos	0.68	372674.25	270638.24	201149.62	0.34
Transformação log	0.65	390198.80	260643.52	162704.30	0.28
Log com Var. de Interação	**0.63**	**402047.49**	**210913.13**	**127497.35**	**0.22**

Tabela 4 – Atributos não significativos removidos do modelo.

Atributo	p-value
Área Murada	0.68
Armários no Quarto	0.53
Portão Eletrônico	0.25
Mobiliado	0.19
Quarto de Serviço	0.068

Um melhor resultado é obtido com a aplicação da transformação logarítmica dos atributos Área Útil e Preço, por reduzir a heterocedasticidade característica em muitos modelos de preços. Uma melhora adicional ainda é obtida ao se adicionar variáveis de interação. Dentre outras interações avaliadas são ao final empregadas as interações Quartos $\times$ Área útil, Área útil $\times$ banheiros e Área útil $\times$ Categoria Casas. Ao final, um modelo de MAPE=0.22 é obtido, sendo um modelo útil para predição de preços.

4.2 Modelos de Aprendizado de Máquina

A Tabela 5 sumariza o resultado dos melhores modelos aplicados aos atributos preditores numéricos (incluindo os encodes) originais, isto é, sendo o título e o vetor construído a partir das imagens.

Tabela 5 – Melhores Resultados da Seleção dos Modelos de Aprendizado de Máquina.

Modelo	R2	RMSE	MAE	MedAE	MAPE
Gradient Boosting (all)	0.99	**71283.11**	50316.73	35095.12	0.07
Random Forest (all)	0.96	130296.62	88421.16	57367.88	0.10
K-Nearest Neighbors (all)	0.67	375741.02	260570.99	172230.00	0.31
Random Forest (CV)	0.66	**373052.16**	255488.21	158293.09	0.30
Gradient Boosting (CV)	0.65	379349.45	260305.51	165807.51	0.30
K-Nearest Neighbors (CV)	0.45	478450.23	333341.20	218894.80	0.41

O menor erro é o obtido pelo Gradient Boosting quando considerado o ajuste do modelo ao conjunto completo de dados (itens *(all)* na Tabela 5). O Random Forest, entretanto, foi selecionado como o melhor modelo por apresentar um resultado muito próximo para o conjunto completo de dados, e ainda o melhor resultado sobre os dados de teste (itens *(CV)*), indicando uma melhor generalização. Os demais modelos apresentaram erros muito maiores ou foram descartados, como o Decision Tree, que apresentou menor erro (MAPE = 0.0) sobre o conjunto total dos dados, mas é sobrejustado e tem erro muito maior sobre os dados de teste (MAPE = 0.41). O resultado detalhado de todos os modelos pode ser acessado em (RODRIGUES, 2024).

Esse modelo selecionado, Random Forest, é o empregado em todos os experimentos seguintes com os diferentes conjuntos preditores.

Random Forest, atributos númericos. Nos experimentos a seguir é empregado o modelo Random Forest. Das variáveis preditoras originais (numéricas e encodes), são excluídos os atributos *porteiro 24h, portao eletronico, tipo Casa de vila, quarto de servico* e *destaque* (dos anúncios), por apresentarem ganho de informação ≈ 0 (Tabela 6) com relação à predição dos preços.

Tabela 6 – Atributos com Maior e Menor Ganho de Informação para a Predição dos Preços.

Maiores Ganho de Informação		**Menores Ganho de Informação**	
Atributo	**Valor**	**Atributo**	**Valor**
condominio	0,95	porteiro_24h	0,0036
area_util	0,93	tipo_Casa de vila	0,0020
iptu	0,80	armarios_no_quarto	0,0011
location	0,64	tipo_Duplex ou triplex	0,0000
banheiros	0,39	portaria	0,0000

Os maiores e menores ganhos de informação são úteis não apenas para o ajuste dos modelos, mas também por indicarem a importância de cada variável na formação dos preços. O resultado final do modelo ajustado encontra-se na Tabela 7.

Tabela 7 – Modelo Random Forest, atributos numéricos

Modelo Random Forest	**R^2**	**RMSE**	**MAE**	**MedAE**	**MAPE**
Atributos Numéricos (all)	0.98	95300.42	57290.74	29912.10	0.07

Random Forest, atributos numéricos + vetor de objetos (YOLO). Ao modelo anterior é acrescido o vetor de objetos detectados das imagens dos anúncios às variáveis preditoras. O resultado desse encontra-se na Tabela 8. Não há uma diferença significativa nos resultados com relação ao modelo anterior sem os dados de imagem (o MAPE, por exemplo, permanece em 0.07). Porém, reavaliado o ganho de informação, dentre os 15 atributos com maior ganho de informação, 7 são dos atributos relativos ao vetor de imagens: *'sink', 'chair', 'toilet', 'potted plant', 'bed', 'tv' e 'couch'*, sendo, portanto, importantes na predição dos preços.

Tabela 8 – Modelo Random Forest, atributos numéricos + vetor de objetos (YOLO)

Modelo Random Forest	**R^2**	**RMSE**	**MAE**	**MedAE**	**MAPE**
Atributos Númericos + Imagens (all)	0.98	95801.03	57979.46	30782.84	0.07

Random Forest, atributos numéricos + título (Texto). A Tabela 9 traz os resultados do modelo Random Forest com as variáveis preditoras originais (numéricas e hot encodes), acrescido do encode TF-IDF do título dos anúncios.

Tabela 9 – Modelo Random Forest, atributos numéricos + título (Texto)

Modelo Random Forest	**R^2**	**RMSE**	**MAE**	**MedAE**	**MAPE**
Atributos Numéricos + Texto (all)	0.98	93045.62	55591.42	29443.23	0.06

A diferença é bastante pequena com relação ao modelo que só emprega as variáveis originais numéricas, mas não exibe um resultado pior, o que sugere que codificações melhores do texto (vetor de encode > 1024, maior número de *n-grams* ou uso de *embeddings* pode levar a resultados ainda melhores.

Random Forest, atributos numéricos + título (Texto) + vetor de objetos (YOLO). Por fim, todas as variáveis preditoras, o título dos anúncios e os dados derivados das imagens são empregados. Os resultados são apresentados na Tabela 10.

Tabela 10 – Modelo Random Forest, atributos numéricos + título (Texto) + vetor de objetos (YOLO).

Modelo Random Forest	**R^2**	**RMSE**	**MAE**	**MedAE**	**MAPE**
Atributos Núm. + Texto + Imgs. (all)	0.98	92842.89	55535.63	29494.47	0.06

O modelo apresenta resultados melhores do que os que empregam o uso exclusivo dos preditores numéricos originais e, embora a diferença seja pequena e mesmo não significativa, é o melhor modelo obtido. A Figura 8 apresenta de forma gráfica os principais resultados desse modelo.

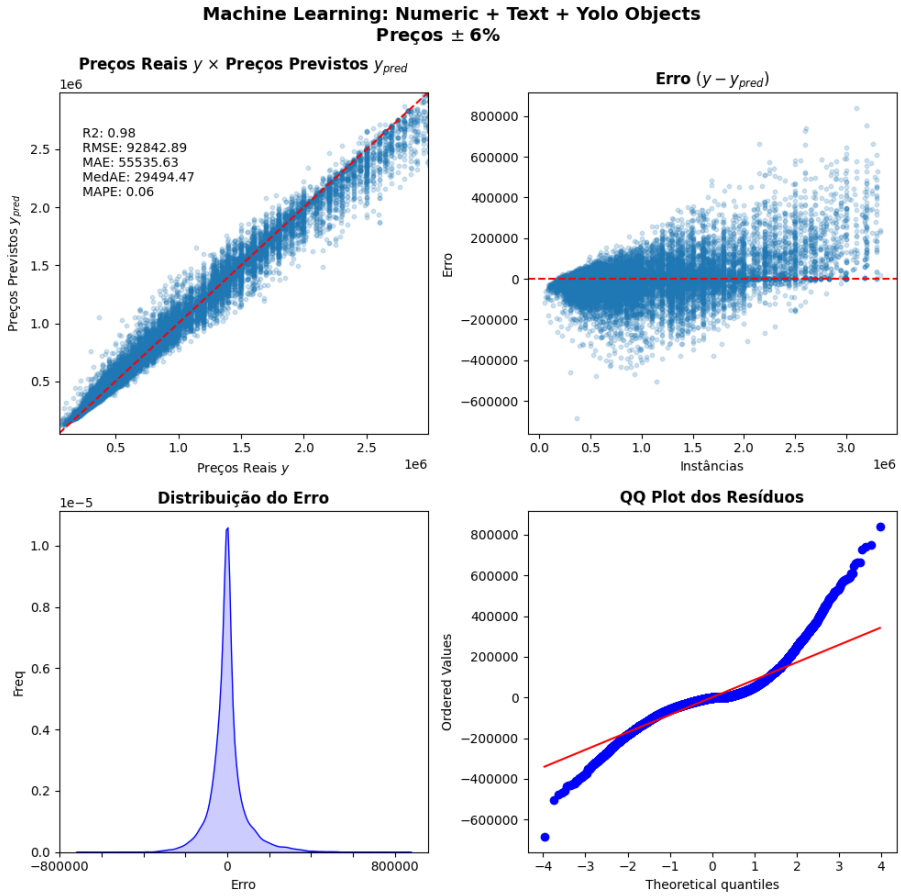

Figura 8 – Resultados do Modelo Random Forest Final.

5 Conclusão

Este estudo conclui que os modelos de Aprendizado de Máquina têm um potencial significativo na previsão de imóveis apresentando resultados melhores que os modelos de regressão linear tipicamente empregados, incluídos aqui os modelos com transformações de variáveis, variáveis de interações e modelos lineares robustos. Em particular, modelos *ensemble*, como o Random Forest e o Gradient Boosting, destacam-se pelos melhores resultados entre os modelos testados.

A inclusão dos objetos detectados nas imagens dos anúncios e seus títulos (texto) como variáveis preditoras nos modelos de aprendizado de máquina traz, neste estudo, uma melhoria dos resultados, mas que não chega a ser expressiva, mas indica a viabilidade dessa abordagem. Trata-se, portanto, de um campo de exploração promissor e para alcançar melhores resultados, pode ser necessário detectar-se objetos mais relevantes nas imagens ou aprimorar a codificação dos dados textuais com técnicas mais elaboradas de *embedding*.

Outros trabalhos futuros, para a melhor predição dos preços, podem envolver o uso direto das imagens em modelos de redes convolucionais pré-treinados, como VGG16, ResNet e Inception, aproveitando as capacidades avançadas desses modelos na extração de características visuais, ou ainda de modelos generativos. Além disso, do ponto de vista prático, parece ser útil a construção de modelos hierárquicos que segmentem a predição por características específicas, como zona geográfica ou número de quartos. Todas essas

abordagens podem contribuir para melhores predições de preços, e o aprimoramento dos resultados obtidos neste estudo.

6 Referências bibliográficas

ALENCAR, S. R. R.; CAURIN, G. *Precificação de imóveis com machine leaning.* 2022.

Associação Brasileira de Normas Técnicas (ABNT). *Avaliação de bens Parte 2: Imóveis urbanos.* Rio de Janeiro: [s.n.], 2011. Norma Técnica NBR 14653-2. Disponível em: https://www.prefeitura.sp.gov.br/cidade/secretarias/upload/infraestrutura/arquivos/ACESSO

BATES, S.; HASTIE, T.; TIBSHIRANI, R. Cross-validation: what does it estimate and how well does it do it? *Journal of the American Statistical Association*, 2021.

BISHOP, C. M. Pattern recognition and machine learning. *Springer google schola*, v. 2, p. 1122–1128, 2006. Acesso: Novembro, 2024.

CHICCO, D.; WARRENS, M.; JURMAN, G. The coefficient of determination r-squared is more informative than smape, mae, mape, mse and rmse in regression analysis evaluation. *PeerJ Computer Science*, v. 7, 2021.

CHOU, J.-S.; FLESHMAN, D.-B.; TRUONG, D.-N. Comparison of machine learning models for early estimates of real estate price. *Journal of Housing and the Built Environment*, 03 2022.

FONTINELLE, A. *What Are Zestimates and How Are They Calculated?* [S.l.]: Investopedia, 2024. Https://www.investopedia.com/articles/personal-finance/111115/zillow-estimates-not-accurate-you-think.asp. Accessed: 11 September 2024.

HANSSON, M.; HOLMQVIST, M.; ANDERSSON, P. Improving house price prediction models: Exploring the impact of macroeconomic features. In: . [s.n.], 2023. Disponível em: <https://api.semanticscholar.org/CorpusID:268753931>.

JANIESCH, C.; ZSCHECH, P.; HEINRICH, K. Machine learning and deep learning. *Electronic Markets*, v. 31, p. 685 – 695, 2021.

KUMARI, K.; YADAV, S. Linear regression analysis study. *Journal of the Practice of Cardiovascular Sciences*, v. 4, p. 33 – 36, 2018.

LOPES, A. *Regressão Linear — Interpretando os dados após a transformação logarítmica.* [S.l.]: medium, 2024. Https://medium.com/@datalopes1/regressAccessed: 15 October 2024.

MARZAGãO, T.; FERREIRA, R.; SALES, L. A note on real estate appraisal in brazil. *Revista Brasileira de Economia*, Fundação Getúlio Vargas, v. 75, n. 1, p. 29–36, Jan 2021. ISSN 0034-7140. Disponível em: <https://doi.org/10.5935/0034-7140.20210003>.

MITCHELL, R. *Web Scraping with Python.* USA: O'Reilly Media, 2018.

MITCHELL, T. M.; MITCHELL, T. M. *Machine learning*. [S.l.]: McGraw-hill New York, 1997. v. 1.

MOHAMED, H.; IBRAHIM, A.; HAGRAS, O. Forecasting the real estate housing prices using a novel deep learning machine model. *Civil Engineering Journal*, v. 9, p. 46–64, 03 2023.

NERI, E. H. M. Modelo preditivo do preço de venda de apartamentos em belo horizonte utilizando random forest. *Universidade Federal de Minas Gerais, Curso de Especialização em Estatística*, 2020. Trabalho de Conclusão de Curso, Orientadora: Ilka Afonso Reis. Disponível em: <https://repositorio.ufmg.br/bitstream/1843/34610/1/Monografia_Evandro_Neri_Versao_Final_20201130.pdf>.

QUINTOANDAR. *Preço de imóveis à venda em São Paulo sobe 3,7[Accessed 10-05-2024]*.

RODRIGUES, J. R. *Modelo_predicao_SP*. [S.l.]: GitHub, 2024. <https://github.com/juliaronquetti/Modelo_predicao_SP>.

RODRIGUES, J. R.; OLIVEIRA, R. de. *Image-Based Property Price Prediction in São Paulo*. Kaggle, 2024. Disponível em: <https://www.kaggle.com/ds/5665290>.

TEKIN, M.; SARı, U. Real estate market price prediction model of istanbul. *Real Estate Management and Valuation*, v. 30, p. 1–16, 12 2022.

UMBRASAS, K. *What Is the Average Time to Sell a House? — zillow.com*. 2019. Https://www.zillow.com/learn/average-time-to-sell-a-house/. [Accessed 10-05-2024].

WOHLWEND, B. *Decision Tree, Random Forest, and XGBoost: An Exploration into the Heart of Machine Learning*. 2023. Accessed: 2024-09-13. Disponível em: <https://medium.com/@brandon93.w/decision-tree-random-forest-and-xgboost-an-exploration-into-the-heart-of-machine-learning-90dc212f4948>.

YANG, K.; TU, J.; CHEN, T. Homoscedasticity: an overlooked critical assumption for linear regression. *General Psychiatry*, v. 32, 2019.

ZHOU, Z.-H. A brief introduction to weakly supervised learning. *National Science Review*, v. 5, p. 44–53, 2018.