Predição de Valores de Imóveis da Região de Toledo com Aprendizado de Máquina

Predicting Real Estate in the Toledo Region with Machine Learning

Martín Ávila Buitrón[1], Gustavo Henrique Paetzold[2]

RESUMO

O presente trabalho aborda o desenvolvimento de um preditor de preços imobiliários em Toledo por meio de aprendizado de máquina. Inicialmente, procedeu-se à coleta de um conjunto de dados de propriedades imobiliárias disponíveis gratuitamente na internet. Posteriormente, foram treinados diversos modelos de aprendizado de máquina com o objetivo de prever o valor por metro quadrado (m²) das propriedades. No geral, os resultados indicaram que o modelo Hist Gradient Boosting Regressor apresentou o melhor desempenho, demonstrado pelo coeficiente de Pearson, com um resultado de 0.819. Essa taxa indica uma forte correlação entre as previsões do modelo e os valores reais dos imóveis.
PALAVRAS-CHAVE: Aprendizado de máquina; Preços imobiliários em Toledo; Modelos de aprendizado de maquina .

ABSTRACT

The present work addresses the development of a real estate price predictor in Toledo through machine learning. Initially, a dataset of real estate properties available for free on the internet was collected. Subsequently, several machine learning models were trained with the aim of predicting the price per square meter (m²) of the properties. Overall, the results indicated that the Hist Gradient Boosting Regressor model showed the best performance, as demonstrated by the Pearson coefficient, with a result of 0.819. This value suggests a strong correlation between the model's predictions and the actual property values.
KEYWORDS: Machine learning; Real estate price; Machine learning models

INTRODUÇÃO

A avaliação de imóveis, que envolve calcular o valor de propriedades, é fundamental tanto para compradores quanto para vendedores, pois serve como base para negociações. (RANG PANG, 2017). O uso de aprendizado de máquina no contexto da previsão de bens e ativos é importante devido aos bons resultados em comparação aos cálculos manuais presentes na maioria dos estudos. De acordo com Mora (2004), é observado que os modelos de aprendizado de máquina tendem a apresentar, em média, uma melhoria de desempenho de aproximadamente 5 a 10% em comparação com modelos de regressão mais simples (MORA, 2004).

Portanto, a problemática deste trabalho refere-se a como predizer o valor de imóveis da região usando modelos modernos de aprendizado de máquina a partir de um conjunto de dados não explorado. As próximas seções descrevem como o trabalho foi estruturado e conduzido para obter o melhor modelo que se adapte ao nosso objetivo.

[1] Bolsista voluntário. Universidade Tecnológica Federal do Paraná, Toledo, Paraná, Brasil. E-mail: martinavila@alunos.utfpr.edu.br.
[2] Docente no curso Eng. Computação. Universidade Tecnológica Federal do Paraná, Toledo, Paraná, Brasil. E-mail: ghpaetzold@utfpr.edu.br. ID Lattes: 3576463426605379.

MATERIAIS E MÉTODOS

O projeto, em todas as suas etapas, foi desenvolvido utilizando a linguagem Python (MULLER, 2017). As bibliotecas do scikit-learn foram empregadas tanto no pré-processamento de dados quanto na otimização e modelagem. (PEDREGOSA, 2011). Posteriormente foi testada a biblioteca do modelo Xtreme Gradient Boosting (MACHINE LEARNING MASTERY, 2023).

Inicialmente, coletaram-se dados do mercado imobiliário de Toledo utilizando plataformas abertas. Posteriormente, esses dados passaram por um processo de limpeza e formatação adequada, incluindo a remoção de valores ausentes e a aplicação de outras técnicas de pré-processamento de dados. Por fim, realizou-se o treinamento de diferentes modelos de aprendizado de máquina com os dados coletados, seguido pela análise de desempenho dos modelos criados.

COLETA DE DADOS

O conjunto de dados utilizado para treinamento e análise do modelo foi construído a partir de informações coletadas gratuitamente disponíveis na web. O tamanho do conjunto de dados é de 2058 registros e 26 características, o que o torna robusto. As características mais relevantes incluem quartos, banheiros, garagens, área do terreno em metros quadrados, área construída em metros quadrados e preço (target).

Assim, após a coleta dos dados, definiu-se uma estratégia para dividir o conjunto de dados em conjuntos de teste e treinamento, preservando as informações e buscando avaliar os resultados de maneira precisa. Portanto, foram alocados 85% dos dados para o conjunto de treinamento e 15% para o conjunto de teste, verificando que os dados de treinamento não continham valores nulos.

A partir desse ponto, utilizou-se o OneHotEncoder da biblioteca scikit-learn para lidar com as variáveis categóricas. Com o conjunto de dados expandido de 24 para 96 características, optou-se por empregar a técnica Iterative Imputer juntamente com o Random Forest Regressor para preencher os valores NAN e não ter perda de amostras. Por fim, as características foram normalizadas usando a biblioteca MinMaxScale de scikit-learn. Desta forma foram rescalados os valores para um intervalo específico, garantindo que todas as características tenham a mesma escala e não intervenham desproporcionalmente nos resultados dos modelos de aprendizado de máquina.

MODELOS DE APRENDIZADO DE MÁQUINA

Após a conclusão do pré-processamento de dados, procedemos à seleção dos modelos para avaliação. Entre os modelos escolhidos foram regressão linear, máquinas de vetores de suporte, Gradient Boosting Regressor e suas respectivas variantes. Nos parágrafos seguintes, é apresentado cada modelo e são justificados os motivos que fundamentaram sua escolha.

O primeiro modelo é o mais fundamental por trás dos modelos de regressão.“O termo "regressão linear" é uma área que engloba um conjunto amplo de técnicas estatísticas utilizadas para modelar relações entre variáveis e prever o valor de uma ou mais variáveis dependentes (ou de resposta) a partir de um conjunto de variáveis independentes (ou preditoras)” (MAROCO, 2003). Estas relações entre variáveis que a

regressão linear utiliza foram importantes para escolhê-la como modelo a ser testado. Sendo a base de problemas de regressão, foi importante começar por ele e entender os hiperparâmetros que utiliza, para então testar outros modelos.

As Máquinas de Vetores de Suporte (SVM) são amplamente utilizadas devido à sua versatilidade, sendo capaz de lidar não apenas com problemas de classificação, mas também de regressão linear. “As Máquinas de Vetores de Suporte são uma extensão que permite modelos mais complexos que não são definidos apenas por hiperplanos no espaço de entrada. Conceitos semelhantes se aplicam a regressão com Máquinas de Vetores de Suporte, conforme implementado em Regressão de Vetor de Suporte (SVR).” (MULLER, 2018, pg 92).

Foi escolhido o algoritmo SVR, derivado do SVM, pensando na flexibilidade que ele oferece resolvendo problemas lineares e não lineares e no possível resultado que pode aportar no problema de um data set multi variável.

Para compreender melhor porque foi escolhida a técnica Gradient Boosting (GB) Regressor e variante Hist, é importante definir as árvores de decisão. “As árvores de decisão são modelos amplamente utilizados em tarefas de classificação e regressão.”(GÉRON,2019). Essencialmente, as árvores de decisão aprendem uma estrutura hierárquica de perguntas condicionais, resultando em uma tomada de decisão.

De acordo com Muller (2018), as árvores de decisão desempenham um papel fundamental no algoritmo GB, pois são os componentes individuais usados para construir o modelo em etapas sequenciais. No processo, várias árvores de decisão são criadas de forma iterativa, onde cada árvore tenta corrigir os erros da árvore anterior. Isso permite que o modelo capture relações complexas nos dados e melhore gradualmente a precisão das previsões.(MULLER, 2018, pg 88).

Tanto o Gradient Boosting Regressor quanto a variante Hist Gradient Boosting (Hist GB) foram implementados. No entanto, o Hist GB, de acordo com a documentação oficial do scikit-learn em 2023, é mais rápido para conjuntos de dados maiores e, além disso, ele suporta valores ausentes. Isso o torna único e é o motivo pelo qual foi escolhido para esta pesquisa.

Por último, foi escolhido o Xtreme Gradient Boost, que é uma variação do Gradient Boosting e também é construído sobre modelos de árvores de decisão. O XGBoost foi escolhido devido à sua reputação como o método preferido e frequentemente vencedor na maioria das competições do site Kaggle, sendo um modelo rápido e robusto. “(KAGGLE, 2023).

TREINAMENTO E OTIMIZAÇÃO .

Os hiperparâmetros dos modelos escolhidos foram escolhidos por meio da técnica Grid Search. Hiperparâmetros são parâmetros que precisam ser ajustados para cada modelo, a fim de gerar respostas distintas. A ideia por trás do uso do grid search é encontrar a melhor combinação de parâmetros para cada cenário de acordo com o nosso conjunto de dados. A Tabela 1 mostra os parâmetros escolhidos para cada modelo por meio do grid search.

Tabela 1 – Parâmetros

Modelos	Parâmetros
Linear	[fit_intercept, n_jobs, positive]
SVR	[svr__C, svr__kernel, svr__gamma]
XGBoost	[learning_rate, n_estimators, max_depth, gamma]
GBRegressor	[learning_rate, n_estimators, max_depth, min_impurity_decrease]
Hist GB Regressor	[learning_rate, max_iter, max_depth]

Fonte: Autoria Própria

Após realizar o pré-processamento de dados e a seleção de modelos, o próximo passo envolveu a avaliação. As métricas utilizadas para avaliação são:

a. Erro Médio Absoluto (MAE), mede a média das diferenças absolutas entre as previsões do modelo e os valores reais;
b. Erro Quadrático Médio (RMSE), calcula a raiz quadrada da média dos erros quadrados entre as previsões do modelo e os valores reais;
c. Correlação de Pearson, avalia a relação linear entre as previsões do modelo e os valores reais;
d. Coeficiente de determinação (R^2), indicando o quão bem as previsões se ajustam aos valores reais.

Essas métricas nos proporcionaram uma visão geral do desempenho dos modelos, permitindo-nos identificar aquele que melhor se ajusta às características do nosso conjunto de dados.

ANÁLISE E RESULTADOS

A Tabela 2 apresenta as métricas de desempenho de todos os modelos treinados. Na Figura 1, esses resultados são exibidos em forma de histograma. Como pode ser observado, o modelo com melhor desempenho foi o Hist GB Regressor. Ao avaliar o MAE, obteve um resultado de 0.0092, onde um valor menor indica maior precisão. No caso do RMSE, registrou um resultado de 0.0163, e quanto menor o valor, maior a precisão, o que também indica o modelo mais preciso. Além disso, o coeficiente de Pearson, com um valor de 0.8190, próximo a 1, indica que o modelo estabelece uma boa relação linear entre as predições feitas e os valores reais. Por fim, no R^2 com um valor de 0.5696, próximo a 1, indica uma boa adaptação do modelo ao conjunto de dados.

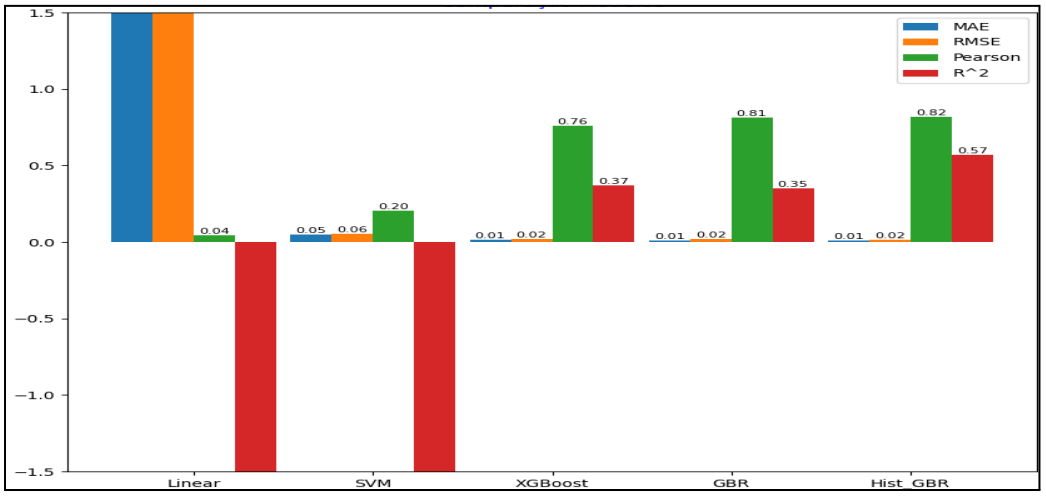
Tabela 2 – Resultados

Modelos	MAE	RMSE	Pearson	R^2
Linear	110386819.35	1122121237.16	0.0434	-2.0365
SVM	0.0494	0.0559	0.2044	-4.0601
XGBoost	0.0123	0.0197	0.7610	0.3681
GBRegressor	0.0117	0.0200	0.8147	0.3530
Hist GB Regressor	0.0092	0.0163	0.8190	0.5696

Fonte: Autoria Própria

É interessante também analisar os outros modelos. Na Figura 1, o modelo Linear possui MAE e RMSE extremamente elevados, indicando um desempenho muito inferior. O modelo SVM apresenta valores baixos de MAE e RMSE,porém possui um coeficiente de Pearson e R^2 negativos, sugerindo que não está estabelecendo uma boa relação linear com os dados. Em contrapartida, os modelos XGBoost e GBRegressor apresentam resultados promissores, com valores muito baixos de MAE e RMSE, além de coeficientes de Pearson e R^2 acima de 0.75, indicando um bom ajuste aos dados.

Figura 1 – Comparação dos resultados



Fonte: Autoria Própria

CONCLUSÕES

Foram coletados dados do mercado imobiliário de Toledo, com um tamanho do conjunto de dados de 2058 registros e 26 características. Esses dados passaram por um pré-processamento, com a ideia de limpar, corrigir e reescalar seus valores. Foram escolhidos os modelos regressão linear, Support Vector Machine, Extreme Gradient Boosting, Gradient Boosting Regressor e Hist Gradient Boosting, os quais tiveram seus hiperparâmetros otimizados por meio do Grid Search. Por fim, o último passo envolveu a avaliação dos resultados dos modelos, usando métricas estatísticas.

Foi observado que o modelo com melhor desempenho foi o Hist GB Regressor. Este modelo alcançou um MAE de 0.0092 e um RMSE de 0.0163, indicando alta precisão

na previsão dos valores. Além disso, o coeficiente de Pearson com um resultado de 0.819 demonstra uma forte relação linear entre as variáveis de entrada, enquanto o R^2 com um resultado de 0.5696 indica uma excelente adaptação do modelo aos dados. Em conjunto, esses resultados ressaltam a eficácia do Hist GB Regressor na tarefa de análise de dados imobiliários.

Pode-se concluir que o modelo de regressão tradicional, embora seja uma base de referência, está longe de ser confiável para decisões no mercado imobiliário. Os modelos alternativos, como SVM, GBR e XG Boost, mostraram-se robustos, mas não atingiram plenamente os objetivos desta pesquisa. Por último, a identificação do Hist Gradient Boosting como o modelo mais adequado para decisões no mercado imobiliário possui implicações de grande relevância. Essas implicações poderiam ser exploradas em futuras pesquisas e análises, considerando que a implementação de redes neurais artificiais poderia ser uma abordagem promissora para melhorar ainda mais a precisão.

Agradecimentos

Agradeço a Deus, à minha família e ao Prof. Gustavo Henrique Paetzold pelo apoio.

Disponibilidade de código

https://github.com/mab0205/Toledo-Regression-Housing-predictor-IC
https://www.kaggle.com/datasets/mab0205/housing-prices-in-toledobr-2023

Conflito de interesse

Não há conflito de interesse.

REFERÊNCIAS

PEDREGOSA, F. et al. Scikit-learn: machine learning in Python. **Journal of Machine Learning Research**, v.12, p.2825–2830, 2011.

GÉRON, A. **Hands-on machine learning with Scikit-Learn, Keras and TensorFlow:** concepts, tools, and techniques to build intelligent systems. 2. ed. O'Reilly Media, 2019

MULLER, A. C.; GUIDO, S. **Introduction to machine learning with Python: a guide for data scientists**. 1. ed. O'Reilly Media, 2017.

RANG PANG,Liangliang Cao Jiebo Luo Quanzeng You. **Image-based appraisal of real estate properties**, IEEE Transactions on Multimedia, v. 19, p. 2751-2759, 2017. doi: 10.1109/TMM.2017.2710804.

MORA, Esperanza. **La Inteligencia Artificial Aplicada a La Valoración de Inmuebles: Un Ejemplo Para Valorar Madrid**,Catastro Web, Madrid, p. 51–68, 2004.

KAGGLE COMPETITTIONS. Kaggle site, 2023. Disponível em: <https://www.kaggle.com/competitions>. Acesso em: 06 de jun. de 2023