Finance 金融, 2024, 14(4), 1552-1562
Published Online July 2024 in Hans. https://www.hanspub.org/journal/fin
https://doi.org/10.12677/fin.2024.144160

Hans汉斯

基于机器学习的房价预测研究

王玉洁

曲阜师范大学统计与数据科学学院，山东 曲阜

收稿日期：2024年6月17日；录用日期：2024年6月27日；发布日期：2024年7月26日

摘　要

房地产行业是我国国民经济的重要组成部分，关乎国计民生，而房价的走势直接影响到社会的金融稳定和整体宏观社会的长期发展，因此对房价进行预测研究对个人消费者、房地产开发商以及国家宏观调控部门都至关重要。本文基于Kaggle在线平台上2020年5月至2021年5月美国King County的房屋销售价格以及房屋的基本信息数据，分别利用支持向量机和XGBoost模型对房屋价格进行预测，采用平均绝对误差、均方根误差和拟合优度作为评价标准将各个预测模型对房价的预测效果进行评价与比较，得出结论：XGBoost模型拟合和预测的效果最好。整体而言，本研究为房价预测提供了科学的模型和方法，为房屋出售者和房屋购买者提供科学的参考依据。

关键词

房价预测，支持向量机，XGBoost

House Price Prediction Based on Machine Learning

Yujie Wang

School of Statistics and Data Science, Qufu Normal University, Qufu Shandong

Received: Jun. 17th, 2024; accepted: Jun. 27th, 2024; published: Jul. 26th, 2024

Abstract

The real estate industry is an important component of China's national economy, affecting livelihoods and national economic planning. Trends in housing prices directly impact financial stability and overall macroeconomic development. Therefore, researching and predicting housing prices are crucial for individual consumers, real estate developers, and national macroeconomic regulators. This study is based on housing sales data and basic property information from King County, USA, collected from May 2020 to May 2021 via the Kaggle platform. Support Vector Machine (SVM)

文章引用: 王玉洁. 基于机器学习的房价预测研究[J]. 金融, 2024, 14(4): 1552-1562.
DOI: 10.12677/fin.2024.144160

and XGBoost models were employed to predict housing prices. Evaluation criteria including Mean Absolute Error, Root Mean Square Error, and coefficient of determination were used to assess and compare the predictive performance of these models. The conclusion drawn was that the XGBoost model demonstrated the best fitting and predictive performance. Overall, this research provides a scientific approach to housing price prediction, offering valuable insights for both sellers and buyers in the housing market.

Keywords

Housing Price Prediction, Support Vector Machine, XGBoost

1. 引言

1.1. 研究背景

近些年来，我国的商品房价格总体呈现上涨趋势，尽管最近有所回落，但人们对购房的需求依然高涨，同时房地产市场也出现了倒买倒卖等问题，导致行业发展逐渐偏离正常轨道。因此，准确了解房价发展趋势，找到较为合适的房价预测方法对于有购房需求的人们来说就显得尤为重要。

随着人工智能技术的迅速发展，机器学习已经深入到各行各业，并取得了显著的成就。机器学习利用计算机处理人类思考和归纳经验的过程，能够解决很多复杂的问题，利用机器学习提供的算法分析数据成为了当前数据分析与建模的主流方法。所以，顺应大数据和机器学习的发展趋势，利用机器学习算法分析预测房价问题更具科学性和前瞻性。

房屋价格受多种因素影响，如地理位置、房屋年龄和状态、周边环境和设施和经济市场等。过去，大多数房价预测模型采用回归方法，但其精度有限，泛化能力不强，特别是在高维度数据和共线性特征存在时表现更为明显。本文针对这些问题，提出了基于特征选择和机器学习的房价预测模型，对美国 King County 房屋销售价格的影响因素进行实证研究，所建立模型能够在一定条件下较为准确地对房价进行预测。

1.2. 研究目的意义

房价不仅仅是经济状况的晴雨表，它更直接关系到房地产开发商和购房者的切身利益。因此，建立精准的房价预测模型不仅对金融市场至关重要，也对人们的日常生活和社会经济稳定具有深远意义。然而，有太多影响房价的因素，这给房价预测带来了巨大的挑战。房价与其他宏观经济因素之间存在着微妙的相互作用，使得预测过程非常复杂。

SVR 模型在处理非线性数据和噪声较多的情况下表现优异，而 XGBoost 模型在处理高维数据和追求高准确性的问题上具有明显优势。因此，可以充分发挥它们各自的优势，提高房价预测模型的准确性和鲁棒性。

1.3. 文献综述

从整体上来看，对房价的预测研究可以归结为两类，一类是对房价进行定性估价预测，更多的是倾

向于经济学分析，主要关注市场信息，很少使用数学模型。另一类侧重于定量分析，利用数学模型对房价进行量化预测。

1.3.1. 房价预测国外研究现状

国外与房价的相关研究起步较早，因为国外的住房商品化早于国内，且房地产交易数据相对完整。

最开始的研究大多数将房价作为时间序列数据，通过历史价格数据的训练学习建立模型，并对未来进行预测。例如，Rangan Gupta [1]通过动态因子分析和贝叶斯收缩估计的时间序列模型来预测美国四个区域的房价增长率。Miles [2]采用广义自回归模型(GAR)对美国五大洲的房价指数进行预测，该研究的结论表明，其所提出的模型在实证样本外的预测时明显优于自回归移动平均模型(ARMA)和 GARCH 模型。

现如今，机器学习渗入到科学的各行各业中，随机森林是一个包含多个决策树的分类器，预测方法后来也渐渐发展为基于树的模型方法。例如，Gu 等[3]采用 SVM 建立房价预测模型，针对 SVM 模型的参数选择问题，采用遗传算法(GA)进行寻优，通过实证分析证明了该模型具有较好的预测效果。进一步，将神经网络应用于房价预测，可以改善过度拟合的问题，例如，Zaheer 等[4]提出了一种基于 LSTM 的混合深度学习预测模型对股票数据进行预测。Limsombunchai [5]基于 ANN 模型，对新西兰克赖斯特彻奇的 200 个房屋信息进行预测，并将结果与经典房价预测模型——Hedonic 价格模型作对比，结果表明他们的 ANN 模型预测更精准。Serrano [6]基于 RNN 模型对时间序列数据，特别是价格进行预测。从房地产、股票和金融科技市场的领域进行了验证，实验结果表明，该方法能够对不同的投资组合做出准确的预测。

1.3.2. 房价预测国内研究现状

目前国内外学者大致有两种思路：一是把房价的变化看作是一个时间序列来预测房价。二是分析房价的影响因素，利用影响因素建立指标体系来预测房价。

与国外研究过程类似，侯普光和乔泽群[7]将小波分析与 ARIMA 模型相结合，通过对房价的数据进行分解和重构进行降噪，以及平稳性检验，估计参数而后建立相应的 ARIMA 模型进行预测。刘丽泽[8]则基于多元线性回归模型及 ARIMA 模型进行分析，对北京市的未来房地产走势进行预测，也针对房地产的行业发展提出了建议。王兆娟[9]基于多元线性回归模型、BP 神经网络模型和 ARIMA 模型对山东省商品房价格进行预测，结果发现 BP 神经网络模型预测值最准确，但从模型的拟合效果看不太适合用于长期预测。

为了谋求算法思想上的创新以及构建新的模型框架，提出了使用回归支持向量机模型(SVR)，它具有高水平的小样本学习能力，申瑞娜等[10]先用主成分分析的方法对初始数据降维处理，而后建立 SVR 模型，对上海的房价进行预测。结果发现预测精度较高，泛化能力较强。何卓[11]等基于 Stacking 集成学习的 Lasso-GBDT 组合对短时间的区域房价进行预测，发现组合模型比单一模型具有更高的准确性和稳定性。朱海煜等[12]以南京江北新区为例，用于 XGBoost 算法进行预测。高玉明等[13]采用遗传算法(GA)优化的 BP 神经网络成功预测了贵阳市的房价，结果显示，GA 优化的 BP 神经网络不仅显著提升了网络训练速度，还显著增强了对房价预测的准确性。曾婷婷[14]通过 Python 爬取了房屋基本信息，构建随机森林、支持向量机、BP 神经网络及 LSTM 模型，通过对比发现，LSTM 模型预测模型效果更好。

综上所述，国内外学者在房价预测研究中有着很多贡献。但目前对房价预测的研究依然存在一定的局限性和缺陷。由于国内房地产业发展时间短，基于时间序列的样本数据量少、不完整，所以利用时间序列模型预测我国的房价效果并不好，而且房价具有波动性、非线性、易受外界因素干扰等特性，线性模型可能对数据中的异常值非常敏感，这些异常值可能会对模型参数产生显著影响，因此通过线性求解房价，可能会造成预测不准确、忽略关键因素等问题，因此，使用更为复杂和灵活的机器学习模型通常能够提供更准确的预测和分析结果，帮助更好地理解和把握房价趋势。

2. 房价预测模型的建立与评价原理

2.1. SVR 模型

支持向量回归(SVR)是一种用于连续型输出变量预测的方法。SVR 的核心理念在于通过定义一个最优边界，即超平面或曲线，来最大化拟合数据点与该边界之间的边际距离，以最小化预测误差，如图 1 所示。

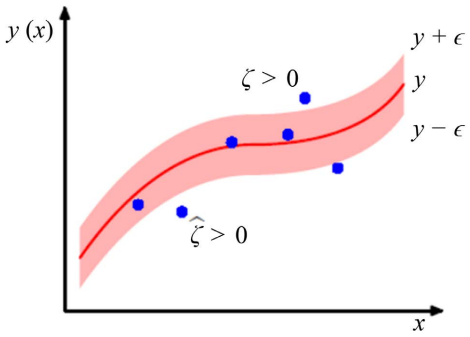

Figure 1. Schematic diagram of support vector machine
图 1. 支持向量机结构示意图

在支持向量机中，核函数的选择非常重要，它决定了模型能够学习的函数空间。常见的核函数包括：

1) 线性核函数(Linear Kernel)

$$K(X,y)=X^T*y$$

线性核函数主要适用于线性可分的情况下，在输入的样本特征维数很大的时候使用。

2) 多项式核函数(Polynomial Kernel)

$$K(X,y)=\left(X^T*y+c\right)^d$$

其中 c 为常数，d 为多项式的阶数。通过多项式函数将数据映射到高维空间，可以处理一定程度的非线性关系。

3) 高斯核函数(Gaussian Kernel)

$$K\left(x_i,x_j\right)=\exp\left(\frac{x_i-x_j}{\sigma}\right)$$

也称为径向基函数(RBF)，通过高斯分布将数据映射到无穷维的特征空间，可以处理更复杂的非线性关系。

4) sigmoid 核函数(Sigmoid Kernel)

$$K\left(x_i,x_j\right)=\tanh\left(\beta x_i^T x_j+\theta\right)$$

通过 sigmoid 函数将数据映射到高维空间，适用于二分类问题。

2.2. XGBoost 模型

XGBoost 预测模型是一种典型的基于集成思想的有监督的加法线性组合模型，里面的每一个基学习器均采用回归树，其训练形式是采用前向分步算法逐棵优化基学习器，当模型训练到第 t 棵回归树时，就需要拟合前 $t-1$ 棵树对训练样本分类回归预测所形成的残差，这样在原有模型的基础上就会增添一棵

新的树，每次增添的是使目标函数值最小的树，与 GBDT 不同的是，XGBoost 模型迭代的目标函数中除了包含常见的损失函数之外，还考虑了模型的复杂度，用正则项表示。

XGBoost 模型作为 boosting 模型中目前最为出色的算法，线性与非线性分类器统筹兼顾，对目标函数引入泰勒公式二阶展开提高了模型的精度，灵活性也更强了。

2.3. 模型评价标准

在回归预测领域中，常有的模型评价指标有：平均绝对误差，简写为 MAE，该指标能够不受正负相抵的影响；均方根误差，简写为 RMSE，该指标适合于不同模型对同一数据的预测效果比较；拟合优度 R 方值，该指标能够忽略量纲的因素影响，用回归平方和与总体平方和之比来表示，其中，RMSE 与 MAE 的值体现误差大小，值越小，说明模型的效果越好，R 方的范围为[0~1]，其值越靠近 1，说明模型的拟合预测效果越好，预测的准确度越高，评价指标公式如下：

$$\mathrm{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2}$$

$$\mathrm{MAE}=\frac{1}{n}\sum_{i=1}^{n}\left|y_i-\hat{y}_i\right|$$

$$R^2=1-\frac{\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2}{\sum_{i=1}^{n}\left(\bar{y}_i-y_i\right)^2}$$

3. 实证分析

3.1. 数据预处理

3.1.1. 数据来源

本文通过 Kaggle 在线平台，获取了美国 King County 从 2020 年 5 月至 2021 年 5 月的房屋销售价格以及房屋的基本信息数据，其中包含卧室数、浴室数和房屋面积等 12 个特征变量。

3.1.2. 数据变换

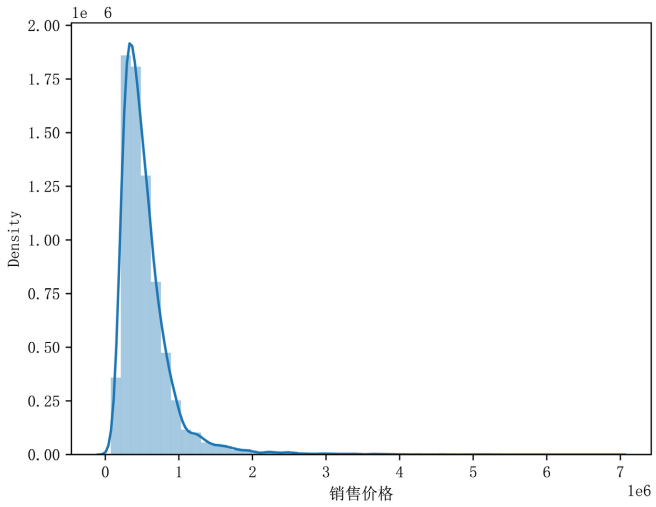

Figure 2. Distribution of sales prices
图 2. 销售价格分布图

绘制目标变量销售价格的分布图，结果如图 2 所示，房价明显右偏，有极大值，而右偏分布，对后续模型训练会有一定影响，故需要对右偏分布进行处理。

采用对目标变量取对数的方法，使其右偏分布近似转化成正态分布，以便于后续的模型构建，取对数后的分布图如图 3 所示，销售价格基本上符合正态分布。

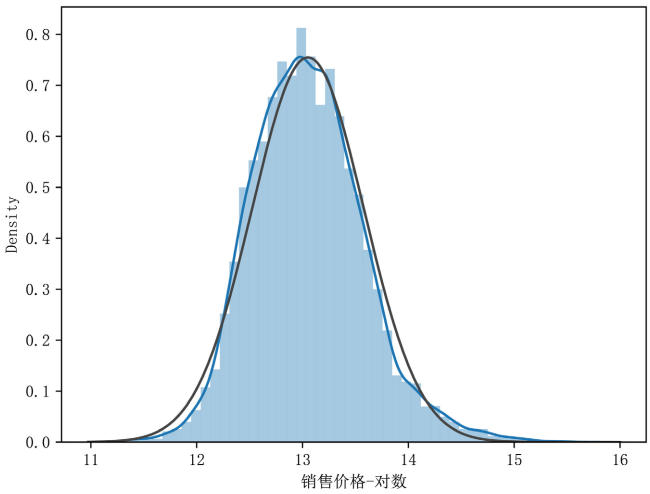

Figure 3. Distribution of sales prices-logarithmic scale
图 3. 销售价格－对数分布图

3.2. 特征选取

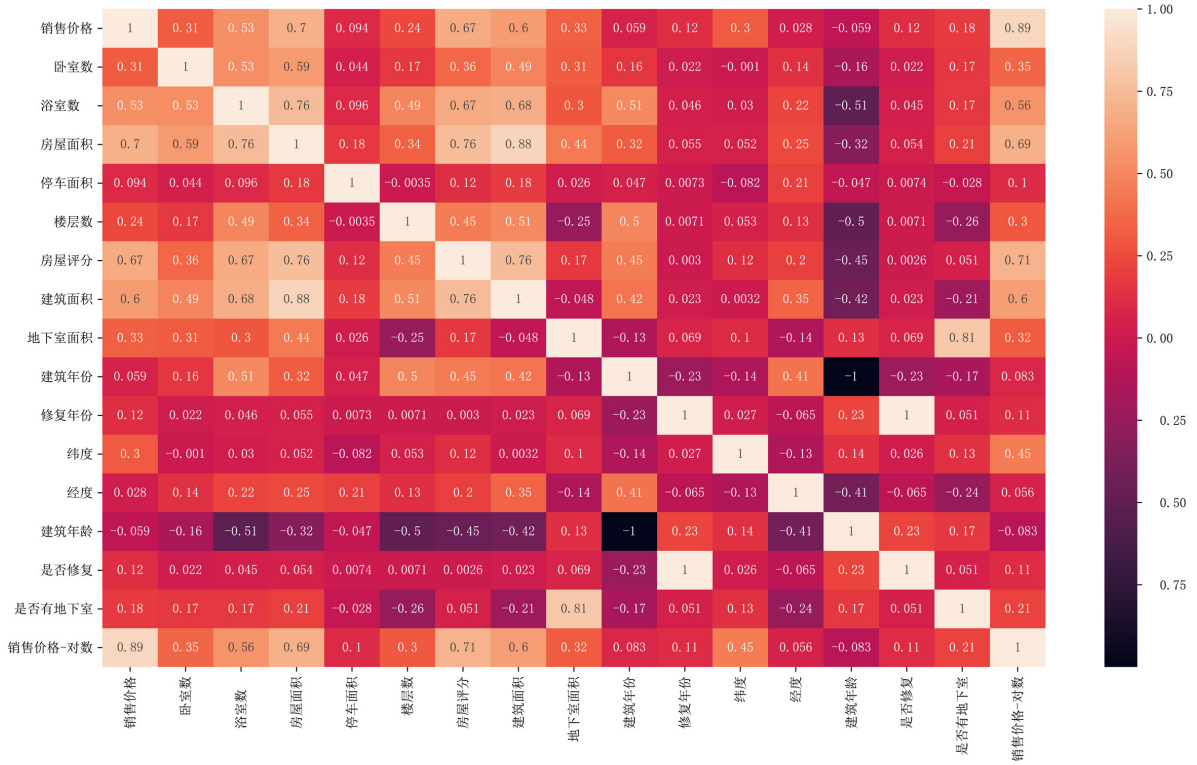

Figure 4. Heatmap
图 4. 热力图

我们通过绘制热力图进行特征选择，根据图 4 我们可以看出，销售价格－对数除了与停车面积、建筑年份、修复年份、经度、是否修复、是否有地下室外，与其他特征相关性都比较明显。故选取相关性高于 0.3 的作为本次建模特征，最终所选数据特征如表 1 所示。

Table 1. Data features table
表 1. 数据特征表

变量类型	变量名称
目标量	销售价格－对数
特征量	卧室数
	浴室数
	房屋面积
	楼层数
	房屋评分
	建筑面积
	地下室面积
	纬度

3.3. SVR 预测模型建立

SVR 预测模型流程如图 5 所示，主要分五步：导入输入、输出变量并划分训练集和测试集；将数据进行标准化操作；选择最佳函数和 SVR 参数；使用建好的 SVR 模型对测试集验证，绘制真实值与预测值的对比图；借助模型评价准则对模型进行评价和误差分析。

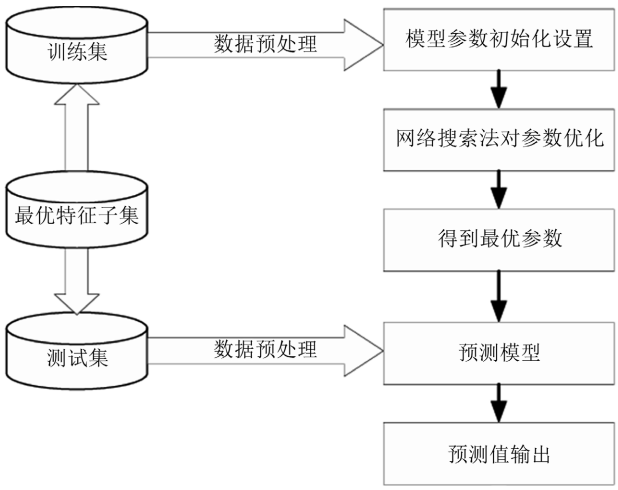

Figure 5. SVR Modeling workflow diagram
图 5. SVR 建模流程图

3.3.1. 数据处理

使用 sklearn 的 Standard Scaler 类，将训练集和测试集中的所有特征变量进行标准化，即：

$$X_{\text{standardized}} = \frac{X - \mu}{\sigma}$$

3.3.2. SVR 参数确定

在支持向量机预测模型中，需要将数据从低维空间通过核函数映射到高维，故首先确定核函数，紧接着是确定核函数的系数 gamma，为防止过拟合加入正则化系数 C，还需要确定间隔带的宽度 d，Python 中用 epsilon 表示。

利用网络搜索法确定 SVR 预测模型最优参数组合，如表 2 所示。

Table 2. Network search method SVR model parameter settings
表 2. 网络搜索法 SVR 模型参数设置

参数	数值
核函数	rbf
gamma	0.1
C	10
epsilon	0.1

3.3.3. SVR 预测结果

使用最佳模型对测试集预测失业率，绘制出 SVR 模型下测试集的预测值和真实值对比图，如图 6 所示，观察图像，所建立的支持向量回归预测模型将失业率测试集中的基本波动都预测的比较吻合，但波峰和波谷处预测效果不是特别理想。

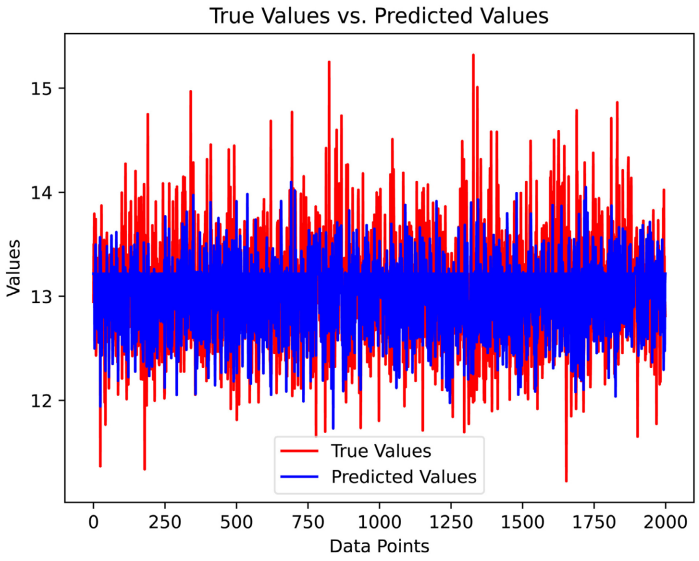

Figure 6. Comparison between SVR predictions and actual values
图 6. SVR 预测值与真实值对比图

Table 3. Evaluation of the optimal SVR model
表 3. 最优 SVR 模型评价

评价指标	数值
R^2	0.6842
RMSE	0.426502
MAE	0.317856

下面计算测试集的均方根误差(RMSE)、平均绝对误差(MAE)和拟合优度(R^2)，如表 3 所示。NMSE 值越小，说明模型拟合度越好；MSE 值越小，说明模型稳定性越好。

3.4. XGboost 预测模型建立

XGBoost 模型预测流程主要分为四步：1) 导入输入变量和输出变量并划分前后段训练集与测试集；2) 使用随机搜索法对训练集确定模型最优参数；3) 使用建好的 XGBoost 模型对测试集验证，绘制真实值与预测值对比图；4) 借助模型评价准则对模型进行评价和误差分析。

3.4.1. XGBoost 参数确定

经过 5 重交叉验证的 Randomized Search CV 确定出 XGBoost 预测模型最优参数如表 4 所示：

Table 4. XGBoost model parameter settings
表 4. XGBoost 模型参数设置

参数	数值
max_depth	8
learning_rate	0.01
n_estimators	1000
objective	reg: squarederror
booster	gbtree
random_state	0

3.4.2. XGBoost 模型预测结果

根据预测值和真实值对比图，如图 7 所示，观察图像，所建立的预测模型将房价测试集中的基本波动都预测的相当吻合，波峰和波谷处也被 XGBoost 预测模型较理想的预测出来，整体预测效果比 SVR 预测模型的强。

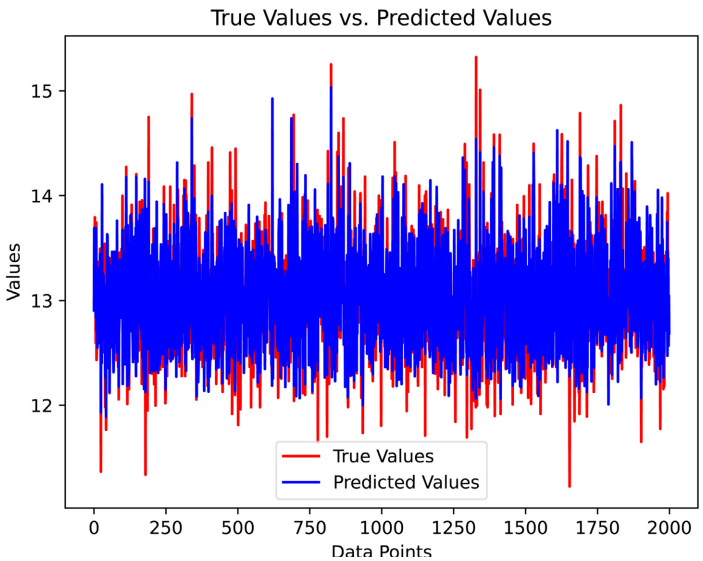

Figure 7. Comparison between XGBoost predictions and actual values
图 7. XGBoost 预测值与真实值对比图

下面计算 XGBoost 模型测试集的 RMSE、MAE 和 R^2，如表 5 所示。

Table 5. Evaluation of the XGBoost prediction model
表 5. XGBoost 预测模型评价

评价指标	数值
R^2	0.8221
RMSE	0.219554
MAE	0.161956

3.4.3. 模型预测效果比较

如表 6 所示，XGBoost 预测模型的 R^2 = 0.8221，比 SVR 模型精度提高了 0.1379，RMSE 与 MAE 分别为 0.219554 和 0.161956，比 SVR 模型分别降低了 0.206948 和 0.1559。综上，无论是图像直观而言，还是从拟合优度值和误差值来看，XGBoost 预测模型对房价的预测效果均优于 SVR 预测模型。

Table 6. Comparison of predictive results from two models
表 6. 两个模型预测结果对比表

模型名称	R^2	RMSE	MAE
SVR	0.6842	0.426502	0.317856
XGBoost	0.8221	0.219554	0.161956

4. 总结与展望

4.1. 总结

随着机器学习技术的高速发展，基于大数据的机器学习预测模型在各行各业得到了广泛的认可和应用。变幻莫测的房价成为了当前社会的焦点话题，如果能够较为准确地预测房价，对购房者、售房者都有很大的辅助作用。本文分析了国内外房价预测的研究现状，拟探索一种基于机器学习的房价预测模型，来实现房价的预测，并根据预测模型建立房价预测系统，实现其应用价值。本文选择 SVR 模型和 XGBoost 模型作为预测房价的基础模型，通过网格搜索法优化参数，最终得到 XGBoost 模型的预测性能优于 SVR 模型。

4.2. 展望

4.2.1. 构建组合模型

基于已有文献的发现，构建组合模型是一个具有潜力和前景的研究方向。通过整合多种模型的预测能力，组合模型往往能够取得比单一模型更优秀的预测效果。例如，陈绵旺[15]等人在商品住宅价格预测中使用了 RS-SVM 模型，其拟合优度明显高于使用单一模型进行预测。

4.2.2. 增加特征变量

房价的影响因素不仅仅是 Kaggle 平台提供特征变量，还有包括文化、政策，甚至是周边的配套情况，如学区房、地铁等都需要获取。

参考文献

[1] Gupta, R. (2013) Forecasting House Prices for the Four Census Regions and the Aggregate US Economy in a Da-

ta-Rich Environment. *Applied Economics*, **45**, 4677-4697. https://doi.org/10.1080/00036846.2013.797561

[2] Miles, W. (2007) Boom-Bust Cycles and the Forecasting Performance of Linear and Non-Linear Models of House Prices. *The Journal of Real Estate Finance and Economics*, **36**, 249-264. https://doi.org/10.1007/s11146-007-9067-1

[3] Gu, J., Zhu, M. and Jiang, L. (2011) Housing Price Forecasting Based on Genetic Algorithm and Support Vector Machine. *Expert Systems with Applications*, **38**, 3383-3386. https://doi.org/10.1016/j.eswa.2010.08.123

[4] Zaheer, S., Anjum, N., Hussain, S., Algarni, A.D., Iqbal, J., Bourouis, S., *et al*. (2023) A Multi Parameter Forecasting for Stock Time Series Data Using LSTM and Deep Learning Model. *Mathematics*, **11**, Article 590. https://doi.org/10.3390/math11030590

[5] Visit, L. (2004) House Price Prediction: Hedonic Price Model vs. Artificial Neural Network. *New Zealand Agricultural and Resource Economics Society Conference*, Blenheim, 25-26 June 2004, 25-26.

[6] Serrano, W. (2020) The Random Neural Network in Price Predictions. *Artificial Intelligence Applications and Innovations*, Neos Marmaras, 5-7 June 2020, 303-314. https://doi.org/10.1007/978-3-030-49161-1_26

[7] 侯普光, 乔泽群. 基于小波分析和 ARMA 模型的房价预测研究[J]. 统计与决策, 2014(15): 20-23.

[8] 刘丽泽. 基于多元线性回归模型及 ARIMA 模型的北京市房价预测[J]. 科技经济导刊, 2018, 26(29): 182-183.

[9] 王兆娟. 山东省商品房价格预测研究[J]. 合作经济与科技, 2023(17): 60-65.

[10] 申瑞娜, 曹昶, 樊重俊. 基于主成分分析的支持向量机模型对上海房价的预测研究[J]. 数学的实践与认识, 2013, 43(23): 13-18.

[11] 何卓, 马少娟, 陈泓霖. 基于 Stacking 集成学习的 Lasso-GBDT 组合房价预测模型研究[J]. 江苏商论, 2023(6): 75-77.

[12] 朱海煜. 基于 XGBoost 算法的城市热点区域房价预测——以南京江北新区为例[J]. 建筑经济, 2022, 43(z2): 433-437.

[13] 高玉明, 张仁津. 基于遗传算法和 BP 神经网络的房价预测分析[J]. 计算机工程, 2014, 40(4): 187-191.

[14] 曾婷婷. 基于机器学习的房价预测模型研究[D]: [硕士学位论文]. 绵阳: 西南科技大学, 2020.

[15] 陈绵旺. 基于 RS-SVM 的商品住宅价格预测研究[D]: [硕士学位论文]. 南昌: 华东交通大学, 2016.