Frontiers of International Accounting 国际会计前沿, 2026, 15(4), 1042-1051
Published Online August 2026 in Hans. https://www.hanspub.org/journal/fia
https://doi.org/10.12677/fia.2026.154110

Hans汉斯

基于Ames数据集的房价预测

——模型比较与超参数优化

周　侥

重庆理工大学数学科学学院，重庆

收稿日期：2026年7月12日；录用日期：2026年8月11日；发布日期：2026年8月25日

摘　要

房价受到多个因素的影响，对其的预测存在非线性、高维度及数据异方差性问题，本文基于Kaggle House Prices数据集，构建了一套包含线性模型、集成学习与核方法的多元回归预测框架。本研究选取支持向量回归(SVR)、随机森林(RandomForest)、XGBoost、LightGBM及弹性网络(ElasticNet)五种模型对房价进行预测。实验对每种模型采用5折交叉验证，以对数均方根误差(log-RMSE)、对数平均绝对误差(log-MAE)及决定系数(R^2)作为评价指标。研究结果表明这五种模型中最优的是SVR，三个评价指标都是最优的。基于前面五种模型的对比，选取SVR模型进行调优，通过Optuna贝叶斯优化对SVR中的三个超参数进行调优后，模型性能进一步提升，log-RMSE降至0.1249，R^2为0.8985，显著优于其他对比模型。

关键词

房价预测，支持向量回归，贝叶斯优化

House Price Prediction Based on the Ames Dataset

—Model Comparison and Hyperparameter Optimization

Yao Zhou

School of Mathematical Sciences, Chongqing University of Technology, Chongqing

Received: July 12, 2026; accepted: August 11, 2026; published: August 25, 2026

Abstract

House Prices are influenced by multiple factors, and there are problems of nonlinearity, high dimensionality and data heteroscedasticity in their prediction. Based on the Kaggle House Prices dataset,

文章引用: 周侥. 基于 Ames 数据集的房价预测[J]. 国际会计前沿, 2026, 15(4): 1042-1051.
DOI: 10.12677/fia.2026.154110

this paper constructs a multiple regression prediction framework including linear models, ensemble learning and kernel methods. This study selects five models, namely Support Vector Regression (SVR), RandomForest, XGBoost, LightGBM and ElasticNet, to predict housing prices. For each model in the experiment, 5-fold cross-validation was adopted, and the logarithmic root mean square error (log-RMSE), logarithmic mean absolute error (log-MAE), and coefficient of determination (R^2) were used as evaluation indicators. The research results show that among these five models, SVR is the best, and all three evaluation indicators are optimal. Based on the comparison of the previous five models, the SVR model was selected for tuning. After tuning the three hyperparameters in SVR through Optuna Bayesian optimization, the model performance was further improved. log-RMSE dropped to 0.1249 and R^2 was 0.8985, which was significantly better than other comparison models.

Keywords

House Price Prediction, Support Vector Regression, Bayesian Optimization

1. 引言

随着机器学习回归方法的快速发展，数据驱动的智能化决策已成为解决复杂现实问题的重要手段。在众多回归预测任务中，房价受到多种因素的影响，因此房价的预测涉及因素繁多、数据非线性强、经济价值高等特点，成为检验模型性能的“试金石”。传统的线性回归模型在处理高维稀疏特征时具有一定的解释性，但往往难以捕捉房价数据中复杂的非线性交互关系与非平稳性，导致预测精度受限。

本文选取弹性网络回归(ElasticNet)、随机森林回归(RandomForest)、XGBoost、LightGBM 与支持向量基回归(SVR)五类回归模型，统一采用对数变换处理目标变量(房价)，构建标准化评估体系。通过 5 折交叉验证对各模型在 log-RMSE、log-MAE 与 R^2 三项指标上的表现进行量化对比，揭示不同模型在房价预测任务中的优势与局限。针对 SVR 模型，本文实施网格搜索与贝叶斯优化相结合的调参策略，获得优化后的模型。

2. 数据集与探索数据分析

2.1. 数据质量分析

为了探究响应变量(房价)的分布，绘制了房价在原始状态与对数变换后的分布对比，如图 1 所示。原始房价数据呈现出显著的偏态分布(右偏)，对房价进行了对数变换，变换后的数据分布趋近于对称的正态分布，偏度显著降低，均值与中位数趋于重合，有效改善了数据的统计特性，从而为后续构建高精度回归模型提供了更稳定的数据基础。

2.2. 协变量分析

为了深入了解数据集的内在特征，我们对关键数值型变量进行了分布可视化分析，同时为了探究数值型特征对房价的线性影响，本研究绘制了关键特征与房价对数的散点图，如图 2。OverallQual 和 GarageCars 等离散特征则表现出明显的多峰分布，反映了常见住房配置的市场集中度。

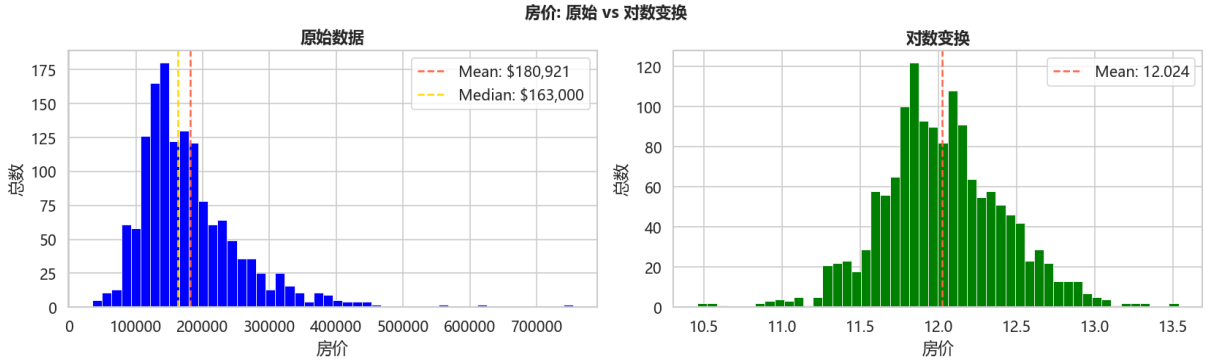


Figure 1. Distribution of sale prices: Original data vs. log-transformed data
图 1. 售卖价格分布：原始 vs 对数变换后的数据

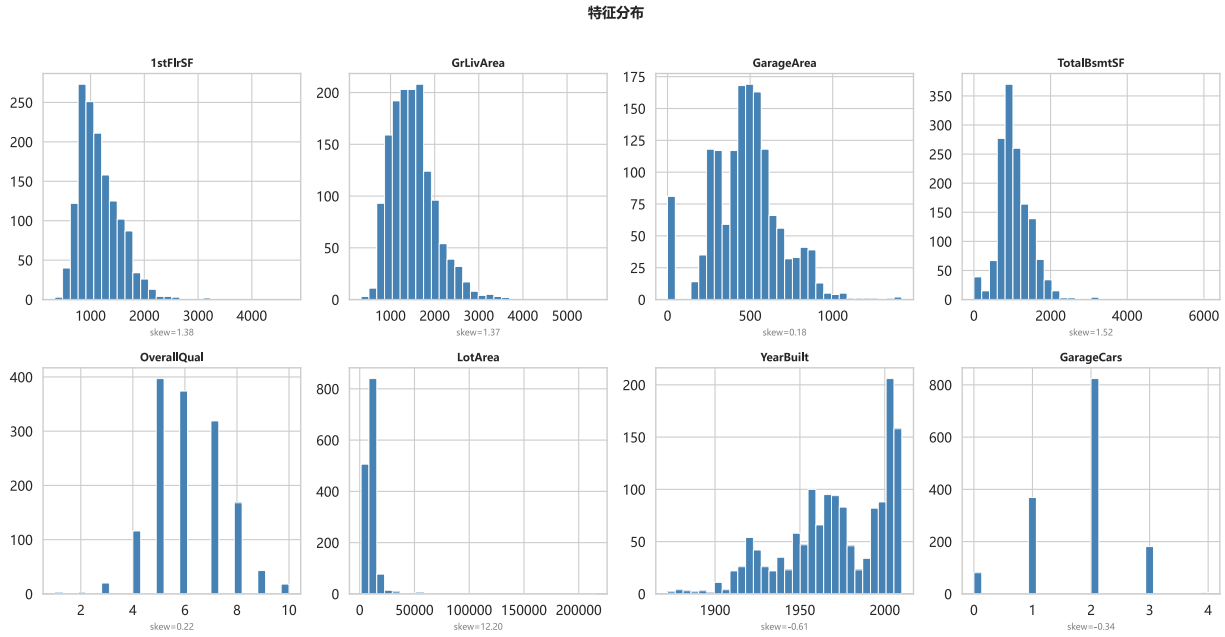


Figure 2. Distribution of key features
图 2. 重要特征的分布图

2.3. 缺失值填补策略对比实验

1) 填补方式一：本文基于房地产领域的专业知识对缺失特征做了区分处理，分为三类：结构性缺失、数值型缺失以及其他缺失，不同的缺失有不同的处理方法。结构性缺失填补：对于那些表示房屋“该房屋的某一种设施的质量”的分类特征，其缺失值解释为“该房屋不具有这种设施”，因此这类特征的缺失值统一填充为字符串“None”。数量型缺失填补：对于与它的结构关联的数值型特征，如果结构的本身是不存在的，则其对应的面积、数量等数值应该为 0，因此这类特征的缺失值应填充为 0。基于中位数填补：经过上述两种基于业务规则的填补后，数据中仍存在少量无法通过上述逻辑推断的缺失值。为确保数据完整性，这部分缺失值在后续的建模流程中统一用中位数或众数填充。

2) 填补方式二：将缺失值超过 50%的变量删除，低于 50%的变量用中位数填充。

3. 模型原理与对比实验

3.1. 模型的基本原理

3.1.1. 支持向量机回归

支持向量回归(Support Vector Regression, SVR)是支持向量机(SVM)在回归问题中的扩展。与传统的回归方法不同，SVR 假设能容忍 $g(\boldsymbol{x}_i)$ 与 y_i 之间最多有 ε 的偏差，即仅当 $g(\boldsymbol{x}_i)$ 与 y_i 之间差别的绝对值大于 ε 时计算损失，也就是说 SVR 不直接最小化预测误差，而是试图找到一个函数，使得预测值与真实值之间的偏差在一个可控范围内，同时保证模型的复杂度尽可能低[1]。

存在一个训练集 $D=\{(\boldsymbol{x}_i,y_i),i=1,2,\cdots,n\}$，其中 $\boldsymbol{x}_i=(x_{i1},x_{i2},\cdots,x_{ip})^{\mathrm{T}}\in\mathbb{R}^p$ 为特征向量，$y_i\in\mathbb{R}$，回归模型如下：

$$g(\boldsymbol{x})=\beta_0+\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x} \tag{1}$$

其中：$\boldsymbol{\beta}=(\beta_1,\beta_2,\cdots,\beta_p)^{\mathrm{T}}$ 与 β_0 是待估的模型参数。

与传统的回归不同，SVR 回归引入了 ε-不敏感损失函数，定义为：

$$\ell_\epsilon(z)=\begin{cases}|z|-\epsilon, & |z|>\epsilon\\ 0, & \text{其他}\end{cases} \tag{2}$$

其中 $\varepsilon>0$ 为调节参数，由式子(2)得到，如果残差 $z_i=y_i-g(\boldsymbol{x}_i)$ 的绝对值小于或者等于 ε，则损失为 0。

关于模型参数 β_0 和 $\boldsymbol{\beta}$ 的极小化问题：

$$\min_{\beta_0,\boldsymbol{\beta}}\left\{\frac{1}{2}\|\boldsymbol{\beta}\|_2^2+\alpha\sum_{i=1}^{n}\ell_\varepsilon\left(y_i-g(\boldsymbol{x}_i)\right)\right\} \tag{3}$$

引入非负的松弛变量 ξ_i 和 ξ_i^*，可将式(3)重新写为：

$$\begin{cases}\min\limits_{\beta_0,\boldsymbol{\beta},\xi_i,\xi_i^*}\left\{\frac{1}{2}\|\boldsymbol{\beta}\|_2^2+\alpha\sum\limits_{i=1}^{n}(\xi_i+\xi_i^*)\right\},\\ \text{s.t.}\ \begin{cases}g(\boldsymbol{x}_i)-y_i\le\epsilon+\xi_i,\\ y_i-g(\boldsymbol{x}_i)\le\epsilon+\xi_i^*,\\ \xi_i\ge0,\\ \xi_i^*\ge0,\end{cases} & i=1,\cdots,n.\end{cases} \tag{4}$$

引入非负的 Lagrange 乘子 $\lambda_i\ge0,\lambda_i^*\ge0,\eta_i\ge0$ 和 $\eta_i^*\ge0$，利用 Lagrange 乘子法可得如下的 Lagrange 乘子函数：

$$\begin{aligned}L(\beta_0,\boldsymbol{\beta},\xi_i,\xi_i^*)=&\frac{1}{2}\|\boldsymbol{\beta}\|_2^2+\alpha\sum_{i=1}^{n}(\xi_i+\xi_i^*)-\sum_{i=1}^{n}\lambda_i\xi_i-\sum_{i=1}^{n}\lambda_i^*\xi_i^*\\&+\sum_{i=1}^{n}\eta_i(g(\boldsymbol{x}_i)-y_i-\epsilon-\xi_i)+\sum_{i=1}^{n}\eta_i^*(y_i-g(\boldsymbol{x}_i)-\epsilon-\xi_i^*),\end{aligned} \tag{5}$$

其中 $g(\boldsymbol{x}_i)=\beta_0+\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_i$。令 $L(\beta_0,\boldsymbol{\beta},\xi_i,\xi_i^*)$ 分别对这四个参数求偏导数，并令其为 0，可得如下等式：

$$\begin{cases}0=\sum\limits_{i=1}^{n}(\eta_i^*-\eta_i),\\ \boldsymbol{\beta}=\sum\limits_{i=1}^{n}(\eta_i^*-\eta_i)\boldsymbol{x}_i,\\ \alpha=\lambda_i+\eta_i,\\ \alpha=\lambda_i^*+\eta_i^*.\end{cases} \tag{6}$$

将上式代入，可得到 SVR 的对偶问题为：

$$\begin{cases}\min\limits_{\boldsymbol{\eta},\boldsymbol{\eta}^*}\left\{\sum\limits_{i=1}^{n}y_i\left(\eta_i^*-\eta_i\right)-\epsilon\sum\limits_{i=1}^{n}\left(\eta_i^*+\eta_i\right)-\frac{1}{2}\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{n}\left(\eta_i^*-\eta_i\right)\left(\eta_j^*-\eta_j\right)\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_j\right\}, \\ \text{s.t.}\quad\begin{cases}\sum\limits_{i=1}^{n}\left(\eta_i^*-\eta_i\right)=0, \\ 0\le\eta_i,\eta_i^*\le\alpha,\end{cases}\quad i=1,\cdots,n.\end{cases}$$

其中：$\boldsymbol{\eta}=\left(\eta_1,\cdots,\eta_n\right)^{\mathrm{T}}$ 和 $\boldsymbol{\eta}^*=\left(\eta_1^*,\cdots,\eta_n^*\right)$，上述过程需要满足 KKT 条件，对 $i=1,\cdots,n$，有：

$$\begin{cases}\eta_i\left(g\left(\boldsymbol{x}_i\right)-y_i-\epsilon-\xi_i\right)=0, \\ \eta_i^*\left(y_i-g\left(\boldsymbol{x}_i\right)-\epsilon-\xi_i^*\right)=0, \\ \eta_i\eta_i^*=0, \\ \xi_i\xi_i^*=0, \\ \left(\alpha-\eta_i\right)\xi_i=0, \\ \left(\alpha-\eta_i^*\right)\xi_i^*=0.\end{cases}\tag{7}$$

根据上述公式可以得到最终的预测函数如下：

$$\begin{aligned}\hat{g}\left(\boldsymbol{x}\right)&=\hat{\beta}_0+\hat{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{x}=\hat{\beta}_0+\sum_{i=1}^{n}\left(\hat{\eta}_i^*-\hat{\eta}_i\right)\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x} \\ &=\hat{\beta}_0+\sum_{i=1}^{n}\left(\hat{\eta}_i^*-\hat{\eta}_i\right)\left\langle\boldsymbol{x}_i,\boldsymbol{x}\right\rangle\end{aligned}\tag{8}$$

3.1.2. 随机森林回归

随机森林(RandomForest)是通过集成学习的思想将多棵决策树集成的一种算法，他的基本单元是决策树，它的本质是机器学习的一大分支集成学习，由 Breiman 在 2001 年提出。

给定训练集 $D=\left\{\left(\boldsymbol{x}_i,y_i\right),i=1,2,\cdots,n\right\}$，其中 $\boldsymbol{x}_i=\left(x_{i1},x_{i2},\cdots,x_{ip}\right)^{\mathrm{T}}\in\mathbb{R}^p$ 为协变量或者特征变量向量的观测数据，$y_i\in\mathbb{R}$ 为响应变量的观测数据，本文的数据响应变量为定量变量，所以考虑随机变量回归。随机森林回归的具体步骤如下[1]：

步骤 1 对训练数据集 D 进行 B 次有放回的抽样，得到 B 个 Bootstrap 样本，每个样本的容量都是 n，其中第 b 个 Bootstrap 样本为：

$$\left\{\left(\boldsymbol{x}_i^{*b},y_i^{*b}\right),i=1,\cdots,n\right\},\ b=1,\cdots,B$$

步骤 2 对于 $b=1,\cdots,B$，根据 Bootstrap 样本构建 B 棵不同的决策树，对于第 b 棵决策树 T_b，在每次节点分裂时，随机选择 m 个变量作为候选的分裂变量，其余 $p-m$ 个变量不用，在下一次分裂时再随机选取 m 个变量作为分裂变量，对于回归问题一般选取 $m=p/3$ 个变量。记其决策树的预测值为 $\hat{T}_b\left(x\right)$。

步骤 3 对于回归树，则将 B 棵树的预测结果进行平均，得到估计：

$$\hat{y}_{RF}\left(x\right)=\frac{1}{B}\sum_{b=1}^{B}\hat{T}_b\left(x\right)\tag{9}$$

根据上述步骤完成随机森林回归估计问题。

3.1.3. XGBoost 回归

XGBoost 回归是一种基于梯度提升树结构的集成学习模型，因其预测精度高、鲁棒性强及训练效率

高等优势，在高维非线性回归问题中得到广泛应用。XGBoost 通过迭 a 代优化损失函数，并引入正则项以抑制过拟合，提升模型泛化能力[2]。

给定训练集 $D=\{(\boldsymbol{x}_i,y_i),i=1,2,\cdots,n\}$，其中 $\boldsymbol{x}_i=(x_{i1},x_{i2},\cdots,x_{ip})^{\mathrm{T}}\in\mathbb{R}^p$ 为协变量或者特征变量向量的观测数据，$y_i\in\mathbb{R}$ 为响应变量的观测数据，XGBoost 的预测函数为 K 棵回归树的和：

$$\hat{y}_i^{(K)}=\sum_{k=1}^{K}f_k(\boldsymbol{x}_i),\quad f_k\in\mathcal{F} \tag{10}$$

XGBoost 的目标函数定义如下：

$$\mathcal{L}^{(t)}=\sum_{i=1}^{n}L\left(y_i,\hat{y}_i^{(t-1)}+f_t(\boldsymbol{x}_i)\right)+\Omega(f_t) \tag{11}$$

其中：损失函数定义为平方损失 $L(y_i,\hat{y}_i)=(y_i-\hat{y}_i)^2$；正则项为 $\Omega(f_t)=\gamma T+\frac{1}{2}\lambda\sum_{j=1}^{T}w_j^2$，用于控制模型复杂度、防止过拟合；$f_k$ 表示第 k 棵决策树；t 为树的叶节点数量；w 为输出权重；γ,λ 为正则化系数。该目标函数兼顾模型拟合误差与复杂度，有效避免过拟合。

3.1.4. LightGBM 回归

LightGBM 是基于梯度提升决策树(Gradient Boosting Decision Tree, GBDT)框架的一种高效机器学习算法[3] [4]。与传统 GBDT 模型相比，LightGBM 通过基于直方图决策树学习算法，减少内存使用并提升训练速率。该算法核心思想是将特征按梯度方向实施最优分割从而不断构建弱学习器，并最终通过加权投票机制得到强预测模型。LightGBM 算法通过构建一系列决策树来提升模型性能[3]。

给定训练集 $D=\{(\boldsymbol{x}_i,y_i),i=1,2,\cdots,n\}$，其中 $\boldsymbol{x}_i=(x_{i1},x_{i2},\cdots,x_{ip})^{\mathrm{T}}\in\mathbb{R}^p$ 为协变量或者特征变量向量的观测数据，$y_i\in\mathbb{R}$ 为响应变量的观测数据，模型训练的目的是找到一个函数 $f(x)$ 使得预测值 $\hat{y}_i=f(x_i)$ 最接近真实值 y_i，即最小化损失函数：

$$L(f)=\sum_{i=1}^{n}l(y_i,f(x_i)) \tag{12}$$

其中：l 是损失函数，常选用均方误差 $l(y_i,\hat{y}_i)=\frac{1}{2}(y_i-\hat{y}_i)^2$。

3.1.5. 弹性网络回归

弹性网络(Elastic Net)回归算法，融合了岭回归算法和 LASSO 回归算法，结合了 L_1-范数和 L_2-范数正则化优势。其中，L_1 正则化能够有效地进行变量选择，识别对系统输出影响较大的输入变量；L_2 正则化可以控制模型的复杂度，防止过拟合，使得模型在训练数据和测试数据上都能有较好的性能[5]。

给定训练集 $D=\{(\boldsymbol{x}_i,y_i),i=1,2,\cdots,n\}$，其中 $\boldsymbol{x}_i=(x_{i1},x_{i2},\cdots,x_{ip})^{\mathrm{T}}\in\mathbb{R}^p$ 为协变量或者特征变量向量的观测数据，$y_i\in\mathbb{R}$ 为响应变量的观测数据，普通的多元线性回归模型为：

$$\boldsymbol{y}=\beta_0+\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon} \tag{13}$$

其中：β_0 是截距项，$\boldsymbol{\beta}$ 是回归系数向量。

弹性网络回归的目标函数如下：

$$\min_{\beta_0,\beta}\left\{\frac{1}{2n}\sum_{i=1}^{n}\left(y_i-\beta_0-\boldsymbol{x}_i^{\top}\boldsymbol{\beta}\right)^2+\lambda\left[\alpha\|\boldsymbol{\beta}\|_1+\frac{1-\alpha}{2}\|\boldsymbol{\beta}\|_2^2\right]\right\} \tag{14}$$

其中：α 为超参数，$\alpha\in[0,1]$。当 α 为 0 或 1 时，式(14)分别对应岭回归或 Lasso 回归。当 $\alpha\in(0,1)$ 时，

弹性网络回归同时兼具 L_1 和 L_2 正则化的优点。一方面，L_1 正则化项可将某些特征的回归系数压缩为 0，从而筛选出显著相关的特征；另一方面，L_2 正则化项可有效解决多重共线性问题，保留更多特征，避免信息丢失。

3.2. 评估指标定义

为全面衡量模型在房价预测任务中的拟合精度与泛化能力，本文采用对数尺度下的三类评价指标，分别为对数均方根误差(log-RMSE)、对数平均绝对误差(log-MAE)以及决定系数(R^2)。所有指标均基于 $\log(1+SalePrice)$ 计算，以降低房价长尾分布带来的异方差性影响。

1) 对数均方根误差(log-RMSE)计算公式：

$$\text{log-RMSE}=\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\log(1+y_i)-\log(1+\hat{y}_i)\right)^2} \tag{15}$$

其中：y_i 是真实的房价；$\hat{y}_i$ 是预测得到的房价，N 是训练集中的样本总数。

2) 对数平均绝对误差(log-MAE)计算公式：

$$\text{log-RMSE}=\frac{1}{N}\sum_{i=1}^{N}\left|\log(1+y_i)-\log(1+\hat{y}_i)\right| \tag{16}$$

3) 决定系数(R^2)计算公式：

$$R^2=1-\frac{\sum_{i=1}^{N}\left(\log(1+y_i)-\log(1+\hat{y}_i)\right)^2}{\sum_{i=1}^{N}\left(\log(1+y_i)-\overline{\log(1+\hat{y}_i)}\right)^2} \tag{17}$$

上述三项指标均在 5 折交叉验证框架下进行计算，其中 log-RMSE 作为主优化目标，log-MAE 用于辅助评估预测稳定性，R^2 用于衡量模型的整体解释能力。

3.3. 实验结果与分析

根据式(15)~(17)和五折交叉验证方法计算出来评估指标和标准差，如表 1 所示。SVR 模型表现出最优的基线性能，其 log-RMSE 为 0.1256，R^2 达到 0.8985。这一结果得益于 SVR 通过核函数对高维特征空间的非线性拟合能力，使其在面对房价数据的复杂分布时优于传统的线性模型。

相比之下，基于树的集成模型(LightGBM, XGBoost, RandomForest)在默认参数下表现稍逊于 SVR。值得注意的是，LightGBM 在此阶段领先于 XGBoost，显示出其在处理该类特征时的初步优势。然而，树模型通常对超参数(如学习率、最大深度、正则化系数)极为敏感，其较低的表现暗示了后续进行系统性超参数调优的巨大潜力。这一基准测试为后续模型优化提供了明确的参照系。

Table 1. Comparison of evaluation metrics across five models
表 1. 五种模型的评估指标对比

	SVR	RandomForest	XGBoost	LightGBM	ElasticNet
Log-RMSE	0.1256	0.1421	0.1459	0.1337	0.3988
Log-MAE	0.0824	0.0946	0.0978	0.0893	0.3100
R^2	0.8985	0.8687	0.8624	0.8840	−0.0035

实验结果表明，填补方式一在多数模型(如 SVR、LightGBM)上表现出更优的预测性能，尤其在 SVR 模型中优势更为明显。相比之下，填补方式二因删除高缺失变量，导致部分关键信息丢失，模型在复杂

非线性任务中表现相对较弱。进一步观察发现，结合房地产领域知识的填补方式能够更有效地保留数据结构信息，提升模型对房价非线性关系的捕捉能力。因此，本文最终采用填补方式一作为正式预处理方案，并在此基础上开展后续 SVR 模型的超参数调优。

4. 超参数调优实验

本文采用 Optuna 框架对 SVR 的超参数进行优化，其内部使用 TPE 算法进行贝叶斯优化，以 5 折交叉验证的负均方根误差作为优化目标。

4.1. 优化算法

本研究采用贝叶斯优化(Bayesian Optimization)框架[4]对模型超参数 $\theta=(\theta_1,\theta_2,\cdots,\theta_k)$ 进行寻优，具体使用 Tree-structured Parzen Estimator (TPE)算法。给定超参数空间 Θ 与 K 折交叉验证损失函数 L，优化目标为：

$$\boldsymbol{\theta}^{*}=\arg\min_{\boldsymbol{\theta}\in\Theta}\mathbb{E}_{\left(X_{\text{train}}^{(k)},X_{\text{val}}^{(k)}\right)}\left[\mathcal{L}\left(y_{\text{val}}^{(k)},\hat{y}^{(k)}\left(X_{\text{train}}^{(k)};\boldsymbol{\theta}\right)\right)\right] \tag{18}$$

其中：$\mathcal{L}(y,\hat{y})=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i-\hat{y}_i)^2}$。

TPE 算法将观测到的超参数损失对 $\left\{\left(\theta_j,\ell_j\right)\right\}_{j=1}^{t}$ 按照损失划分为较好集合($loss\le\gamma$)：$D_{good}=\theta_j\mid l_j\le l_\gamma$ 和较差集合：$D_{bad}=\theta_j\mid l_j>l_\gamma$；TPE 用这两个独立的和密度估计分别建立模型：

$$p(\theta\mid good)\propto\sum_{\theta_j\in D_good}K(\theta,\theta_j) \tag{19}$$

$$p(\theta\mid bad)\propto\sum_{\theta_j\in D_bad}K(\theta,\theta_j) \tag{20}$$

其中：$K(\cdot,\cdot)$ 为核函数。TPE 不会直接最大化损失函数，而是等价地最大化如下比值：

$$\theta_{(t+1)}=\arg\max_{\theta}\left[p(\theta\mid good)/p(\theta\mid bad)\right] \tag{21}$$

式(21)表明：优先在好参数分布中出现概率高、坏参数分布中出现概率低的区域采样，从而以较少试验逼近全局最优。

4.2. 参数空间设计

SVR 的性能主要受正则化系数 C、核函数带宽 γ 以及不敏感损失带 ε 的共同影响。正则化系数控制模型复杂度与训练误差之间的权衡，核函数带宽决定样本在高维空间的相似度衰减速度，不敏感损失带定义预测值与真实值之间可容忍的误差边界。本文选取 RBF 核作为映射函数，其参数搜索空间如表 2 所示。

Table 2. Hyperparameter search space design for SVR
表 2. SVR 超参数搜索空间设计

参数	C	γ	ε
取值范围	$[2,20]$	$\left[3\times10^{-4},3\times10^{-3}\right]$	$\left[1\times10^{-5},2\times10^{-3}\right]$

4.3. 结果分析

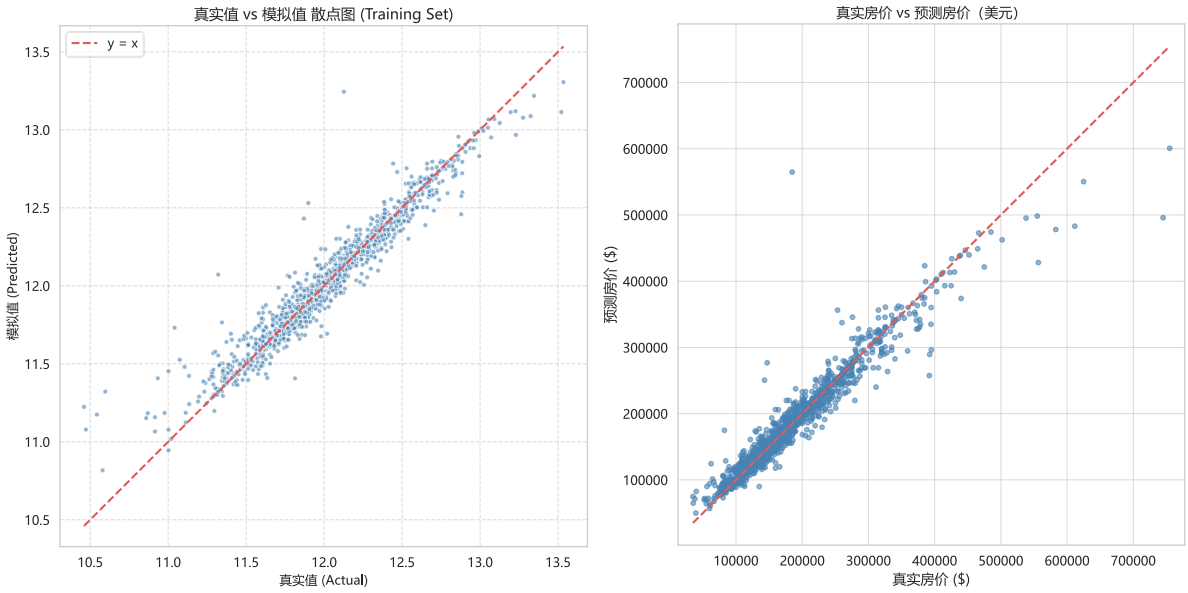
根据优化算法得到 SVR 的超参数的最优值，如表 3 所示，根据这个参数我们计算得到的 log-RMSE、

log-MAE、R^2 分别为：0.1249、0.079 和 0.8985。将这个数值与表 1 对比得到，通过网格搜索与贝叶斯优化对 SVR 的超参数进行调优，模型在对数尺度下的预测误差(log-RMSE 与 log-MAE)显著降低，同时决定系数略有提升，验证了超参数调优在提升模型泛化能力方面的有效性。经过系统调参的 SVR 模型在本实验的五个评估模型中表现最优，显著优于未调优模型及 ElasticNet、XGBoost 等对比模型，证明了超参数优化与特征工程协同作用的重要性。

Table 3. Optimal hyperparameter values for SVR
表 3. SVR 超参数的最优值

参数	C	γ	ε
最优值	2.0211	0.0013	0.0019

对于这个优化后的 SVR 模型，本文对于真实值和预测值进行了可视化分析，如图 3 所示。左图是对房价取了对数的值，右图是没有取对数的原始数据，由图可知，大部分的点都在 $y = x$ 这条线的附近，说明拟合效果较好。



(a) (b)

Figure 3. Scatter plots of predicted vs. actual values based on the optimized SVR model
图 3. 基于优化后的 SVR 模型的预测值和真实值的散点图

基于这个优化后的模型在测试集进行了预测，部分样本预测值如表 4 所示。

Table 4. Predicted values for partial samples in the test set
表 4. 测试集上部分样本的预测值

ID	1461	1462	1463	1464	1465
预测值	119598.6508	160208.2514	183572.5654	201195.1616	192727.0179

5. 结论与建议

本章基于前文构建的特征工程策略与超参数优化方法，对模型的预测性能进行综合评估，并从误差来源、模型对比与泛化能力三个角度展开深入讨论。经过贝叶斯优化调参后的 SVR 模型取得了最好的效果，显著优于随机森林、XGBoost 和 LightGBM 等集成学习方法。这一结果表明，在房价预测这一典型的高维、非线性回归任务中，核方法(Kernel Method)能够通过 RBF 核隐式地将特征映射到高维空间，从而捕捉到房屋面积、建造年份与房价之间复杂的非线性交互关系。相比之下，线性模型(ElasticNet)由于假设限制，无法拟合复杂的市场定价机制，导致误差显著偏高。

尽管当前模型取得了较好的效果，但仍存在以下局限性：房价受宏观经济周期影响较大，但模型未引入外部经济指标(如贷款利率、CPI 等)；对于泳池、壁炉等低频特征，本文将其设置为 2 分类问题，模型仍主要依赖二值标志，缺乏对稀缺资源价值的精细化量化。针对以上问题，未来的工作可以考虑引入地理加权回归(GWR)或融合深度学习模型，以进一步提升复杂特征交互的建模能力。

参考文献

[1] 李高荣. 统计学习(R 语言版) [M]. 北京: 高等教育出版社, 2024.

[2] 牟凯文. 一种基于 XGBoost 回归与粒子群优化算法的大位移井钻速提升方法研究[J]. 石油地质与工程, 2026, 40(3): 115-121.

[3] 赵鹏. 基于 LightGBM 回归算法的山东半岛的降水量空间分布建模[J]. 水利技术监督, 2026(6): 212-215.

[4] 张鲁川, 李一博, 张雷, 等. 基于贝叶斯优化 LightGBM 算法的深层页岩储层分级评价[J]. 天然气地球科学, 2026, 37(4): 801-815.

[5] 王俊峰, 张宝雷, 刘金海, 等. 基于弹性网络回归的集输管线流体状态兰姆波测量辨识方法[J]. 天津科技, 2025, 52(8): 19-23+31.