房价回归预测

摘要

随着城市化进程的加速和房地产市场的发展，房价预测已成为经济研究和决策支持的重要课题。本文利用 Kaggle 房价数据集，基于多元线性回归模型（OLS）构建预测模型，探讨了多种变量处理和模型优化方法。研究通过对连续型变量进行标准化、离散型变量进行编码，并引入多项式特征和主成分分析（PCA）等技术，逐步改善模型效果。最终模型的均方误差（MSE）与平均绝对误差（MAE）均达到较低水平，显示出较好的预测精度与稳定性。本研究不仅验证了回归模型在房价预测中的适用性，还为进一步应用其他机器学习模型提供了思路。

关键词：最小二乘；逐步回归；因子筛选；假设检验

一、引言

1.1 背景介绍

房价作为影响民生和经济的重要指标，一直是政府、房地产企业以及购房者关注的焦点。房价的波动不仅影响城市经济发展，还直接关系到居民的生活质量与投资决策。因此，如何通过数据分析和模型构建准确预测房价，具有重要的现实意义。

近年来，随着大数据技术和机器学习模型的发展，房价预测逐渐从传统的统计分析向智能化方向转变。线性回归模型作为一种经典的统计方法，因其可解释性强、建模简单而被广泛应用。然而，房价数据通常包含非线性关系和多重共线性等问题，使得模型效果难以满足实际需求。因此，如何对变量进行合理预处理，并通过优化建模方法来提高预测精度，是当前研究的关键。

本文基于 Kaggle 房价预测数据集，选取具有代表性的连续型和离散型变量，经过标准化、编码及特征工程处理，构建多元线性回归模型（OLS）进行房价预测。此外，通过残差分析与回归假设检验，探讨了模型的不足，并引入多项式特征和主成分分析（PCA）对模型进行优化，最终实现了对房价的准确预测。本研究不仅为房价建模提供了实践案例，也为后续模型的改进与扩展提供了参考依据。

1.2 数据集说明

数据集来自 Kaggle [错误!未找到引用源。]，任务是基于一系列房屋的物理特征和地理位置，构建回归模型预测美国爱荷华州艾姆斯市（Ames）房屋的最终销售价格。训练集共 1460 条数据，测试集共 1459 条数据；包括有数值型变量如房屋面积、房间数量、建造年份等和分类变量如房屋质量、外墙材质、街道类型等。

1.3 方法介绍

1.3.1 多元回归模型基本概念

多元回归模型假设因变量Y与多个自变量 $X_1, X_2, \ldots, X_p$ 存在一定的线性关系。其基本形式为：

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon$$

其中：

Y是因变量，$X_1, X_2, \ldots, X_p$ 是自变量。

β_0 是截距项。

$\beta_1, \beta_2, \ldots, \beta_p$是回归系数，表示每个自变量对因变量的影响程度。

ϵ是误差项，表示所有未被模型捕捉到的随机因素。

1.3.2 多元回归模型的假设

为了确保回归分析结果的有效性和可靠性，多元回归模型通常假设以下几点：

1、线性关系：因变量与自变量之间应该存在某种线性关系。

2、独立性：自变量之间应相互独立。

3、同方差性：误差项的方差应该是恒定的，即对于所有的自变量组合误差的波动幅度是相同的。

4、正态分布的误差项：误差项应近似符合正态分布，尤其是在样本量较小的情况下。

5、无多重共线性：自变量之间不能存在高度的线性相关性，否则回归系数的估计会不稳定。

1.3.3 前向逐步回归

前向逐步回归从一个简单的模型开始，只包含一个常数项，然后，通过逐步添加新的变量来优化模型，每一步添加一个新的变量，以最能改善模型的某种信息准则如 AIC 或 BIC 等信息来确定当前步骤可添加的最佳变量。

基本步骤如下步骤：

1、从一个包含常数项的模型开始。

2、在所有可选的自变量中，选择最能减少模型误差的自变量。

3、将选中的自变量添加到模型中。

4、重复步骤 2，直到添加更多变量无法显著提高模型性能为止。

1.3.4 后向逐步回归

后向逐步回归的过程与前向逐步回归相反，它从包含所有候选变量的模型开始，然后逐步去除最不重要的变量。

基本步骤如下步骤：

1、从一个包含所有自变量的模型开始。

2、根据 AIC 或 BIC 等标准删除一个最不显著的变量。

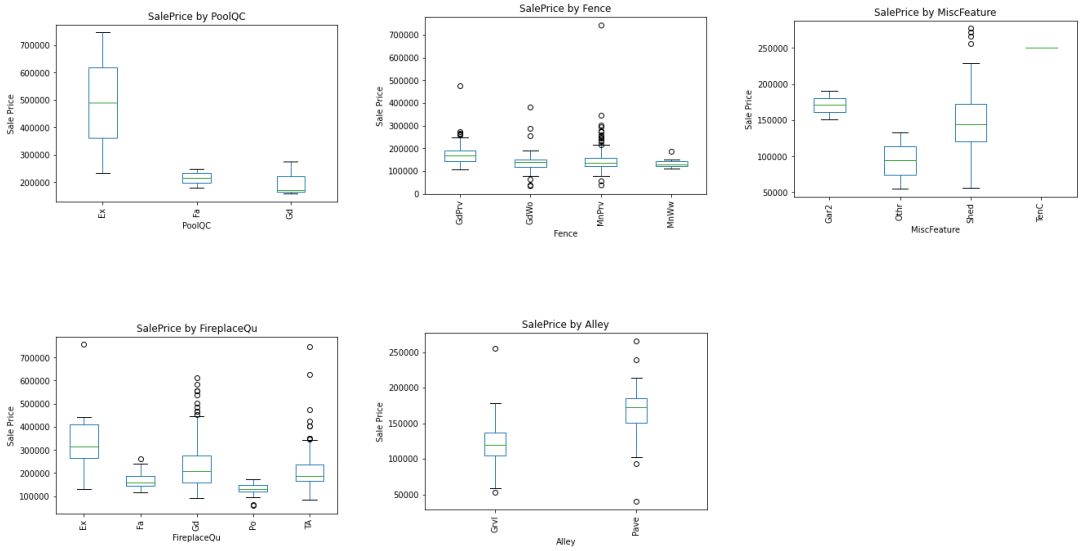
3、重新拟合模型并进行显著性检验。

4、重复步骤 2 和 3，直到移除更多变量无法显著提升模型的拟合效果为止。

二、数据预处理

2.1 数据描述统计与连续型变量初步分析

因子总计有 80 个，其中包括浮点型数据 12 个(float64)，整数型数据 22 个(int64)，离散型数据 46 个(object)。粗略统计发现，’PoolQC’, ‘Fence’, ‘MiscFeature’, ‘FireplaceQu’, ‘Alley’ 五个因子的有效数据量分别为：10，571，105，1499，198，远小于训练集与测试集总样本数 2919；分别画出这五个因子关于价格的箱线图如下所示：

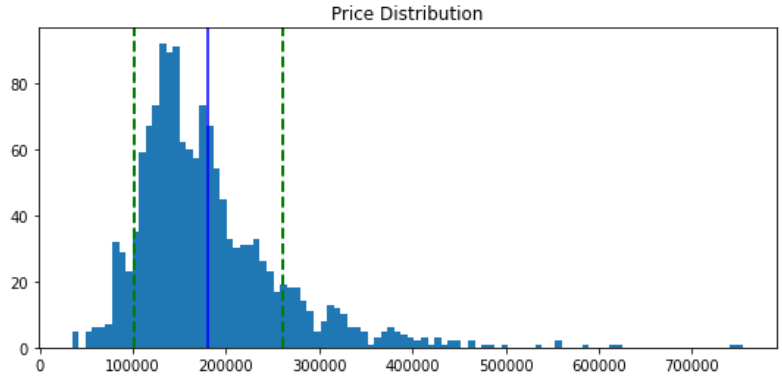


并且经过单因素方差分析，结果如下：

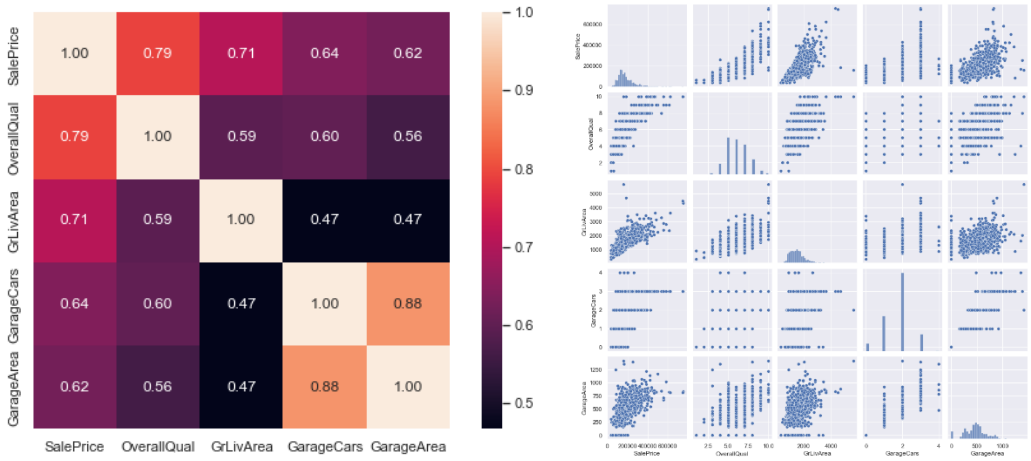
	PoolQC	Fence	MiscFeature	FireplaceQu	Alley
F-statistic	1.6275	4.9482	2.1573	24.3989	35.5621
p-value	0.3040	0.0023	0.1047	0.0000	0.0000

ANOVA 结果表明 PoolQC、MiscFeature 对价格没有显著影响，Fence 有一定影响，FireplaceQu 和 Alley 对价格有较大影响。考虑到有效数据较少，选择先将这五个因子去除，在模型建立完成后考虑再加入因子 FireplaceQu 和 Alley 进行进一步研究。其余因子缺失值分别依据各个因子的分布情况采取均值或者中值进行填充。

绘制出房屋价格的分布图如下，可以看到，价格呈右偏分布，即大部分房屋价格集中在均值(180921)到减一倍样本方差之间(180921-79442=101479)；数据异常值主要集中在房屋价格较高的部分。

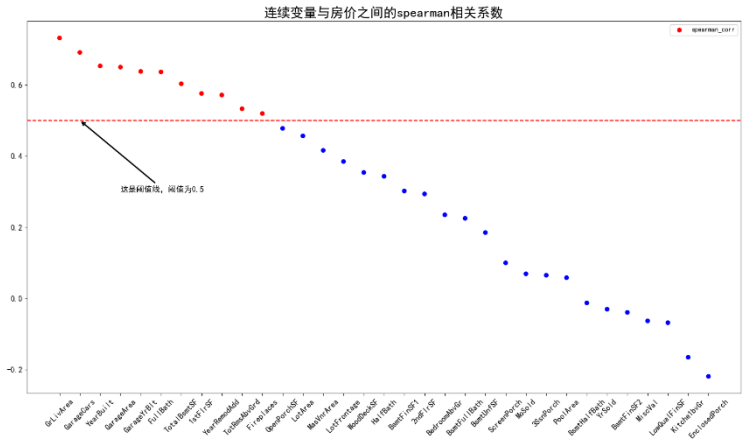


计算数值型数据与价格的相关系数，并绘制相关系数较高的前 5 个变量的热力图以及散点图结果如下：



热力图显示变量之间可能存在相关性，需要在后面进行具体检验并处理；散点图显示变量与价格之间明显存在线性关系。

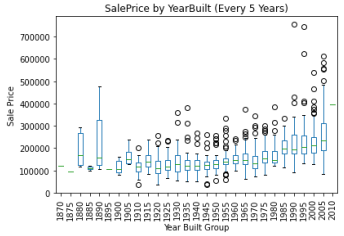
绘制所有数值型数据的相关系数如下图所示。

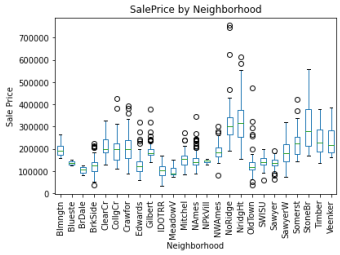


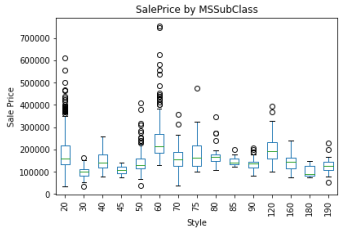
2.2. 离散型变量处理

离散型变量缺失值采用众数，即出现次数最多的情况进行填充。

特别的，考虑价格与房屋建成年份、房屋所在位置和房屋类型三个因子，绘制箱线图如下所示，



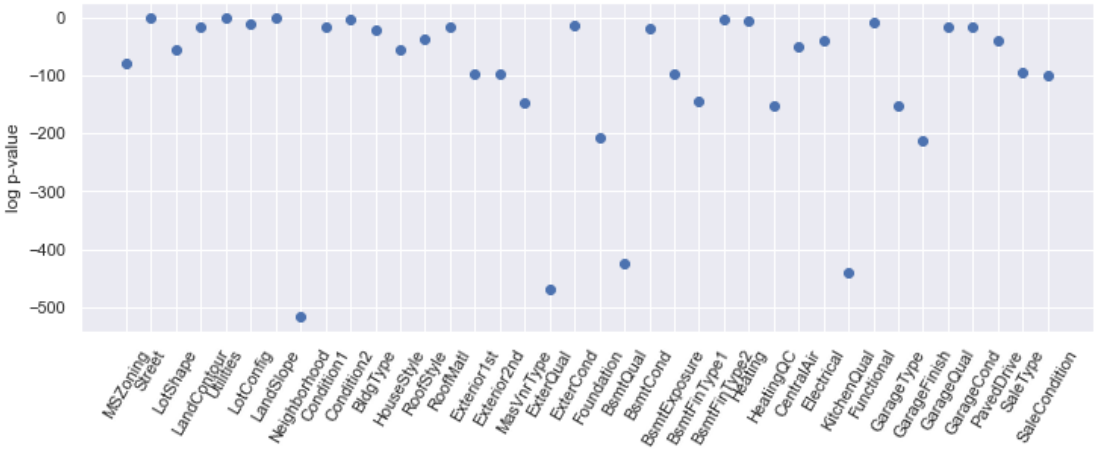




显示房屋价格随时间而增加，不同类别的价格分布也有明显区别。进行单因素方差分析结果如下表所示，显示所选三个因子对房屋价格均有显著影响。

因子	df	SSE	MSE	F-statistic	p-value
YearBuilt	1	$2.5176e^{12}$	$2.5176e^{12}$	548.6658	$2.99e^{-103}$
Residual	1458	$6.6903e^{12}$	$4.5887e^{9}$		
Neighborhood	24	$5.0236e^{12}$	$2.0932e^{11}$	71.7848	$1.56e^{-225}$
Residual	1435	$4.1843e^{12}$	$2.9159e^{9}$		
MSSubClass	1	$6.5411e^{10}$	$6.5411e^{10}$	10.4315	$1.3e^{-3}$
Residual	1458	$9.1425e^{12}$	$6.2708e^{9}$		

对于其他离散变量采用相同的方法进行单因素方差分析，结果如下图，显示因子 Utilities、Street、LandSlope 对价格没有显著影响，需要剔除这三个因子。



2.3. 因子筛选与处理

由 1.1 和 1.2 章节初步分析，对于数值型数据，依据因子与房屋价格的相关系数大小顺序选取较大的前 10 个因子进行分析建模；对于离散型数据，根据因子的单因素方差分析 p 值结果顺序选取较小的后 6 个因子进行分析建模。具体因子如下：

Continuous variable		Discrete variable	
OverallQual	房屋评分	Neighborhood	街道位置
GrLivArea	地上总面积	ExterQual	外墙质量
GarageCars	车库大小（辆）	KitchenQual	厨房质量
GarageArea	车库面积	BsmtQual	地下室高度
TotalBsmtSF	地下室总面积	GarageFinish	车库装修情况
1stFlrSF	一楼面积	Foundation	地基类型
FullBath	浴室情况		
TotRmsAbvGrd	房间总数		
YearBuilt	原始建造日期		
YearRemodAdd	改建日期		

对于离散型数据，需要进行数值化编码转化为数值型数据，可以采取的方法有：虚拟变量法、顺序编码、频率编码和目标编码等多种方法，观察数据的分布情况，并且考虑分类的数量，分别决定对 Neighborhood、ExterQual、KitchenQual 和 BsmtQual 采用各类均值排序进行编码；对于 GarageFinish、Foundation 采用频率编码。

对于连续型变量，使用 z-score 进行标准化处理。记录数值型数据标准化的期望、方差和离散型数据编码的映射，对于测试集采用相同处理方法进行数据处理。

2.4 异常值处理

由前文数据预处理与分析可以观察到，样本中存在许多异常值，表现为低房价特征但是对应于高房价，或者中等房价特征对应特高房价，由于 OLS 模型对异常值非常敏感，所有需要考虑这些异常值对模型的影响程度，考虑是保留还是去除。

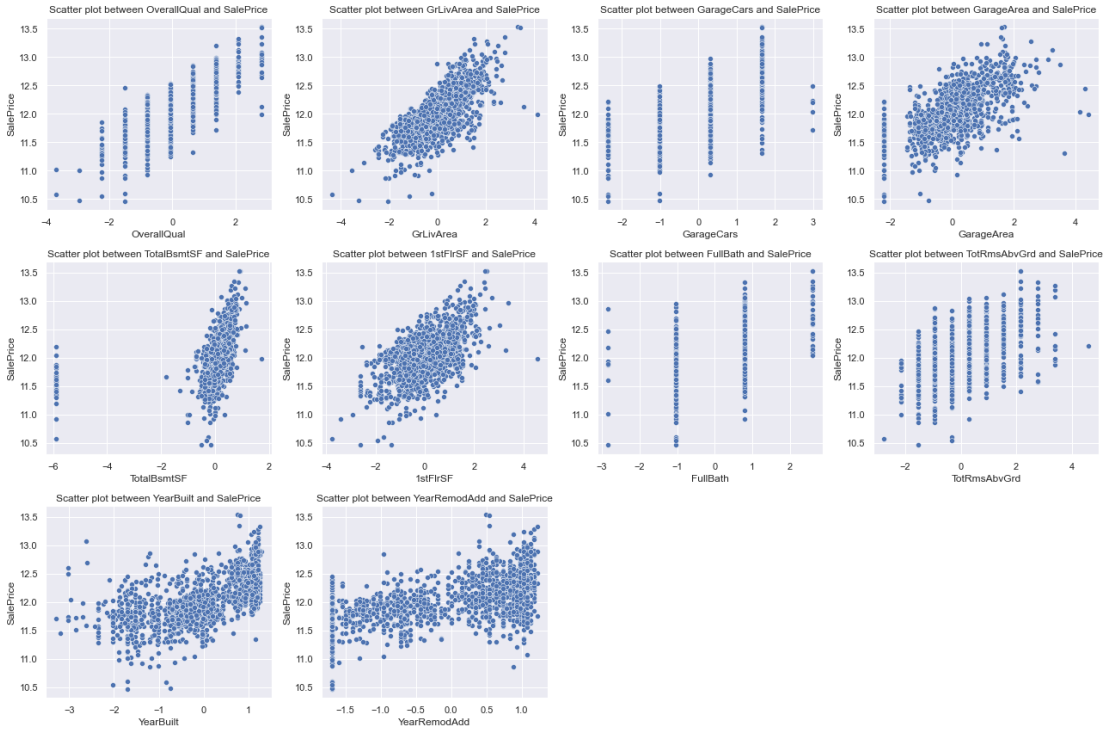
本实验中选择采用 Cook 距离以及 DFFITS 值判断异常点。Cook 距离用于衡量样本点对整体的影响程度，较大的 Cook 距离表明从回归统计量和计算中排除个案之后，系数会发生根本变化，代表该点对模型有显著较大的影响，本例中选取$\frac{4}{n}$（其中n为样本容量）作为判定样本点是否为异常值的标准；DFFITS 衡量去掉单个样本点对回归预测结果影响，较大的

DFFITS 值代表去掉该点会显著影响回归预测结果，本例中选取$2\sqrt{\frac{p+1}{n}}$（其中p为变量个数，n为样本容量）作为判定样本是否为异常值的标准。将两个标准取并集作为需要去掉的异常点集。

三、回归假设验证

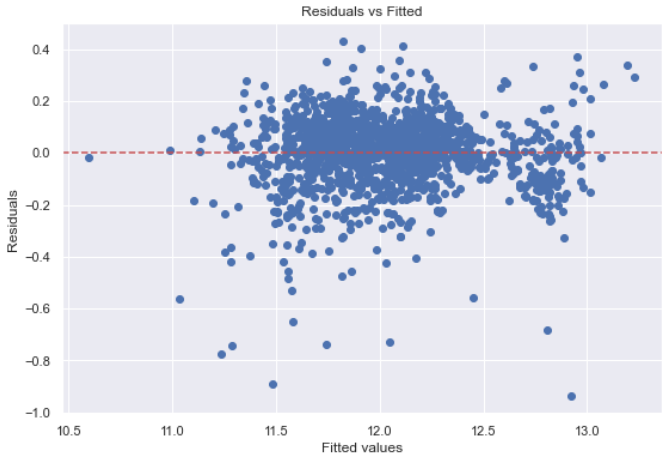
3.1 线性性

绘制训练集数据与价格的散点图如下所示，可以明显看到每一个变量与价格都有或多或少的趋势关系，即自变量与因变量之间存在显著线性关系，认为满足线性性。

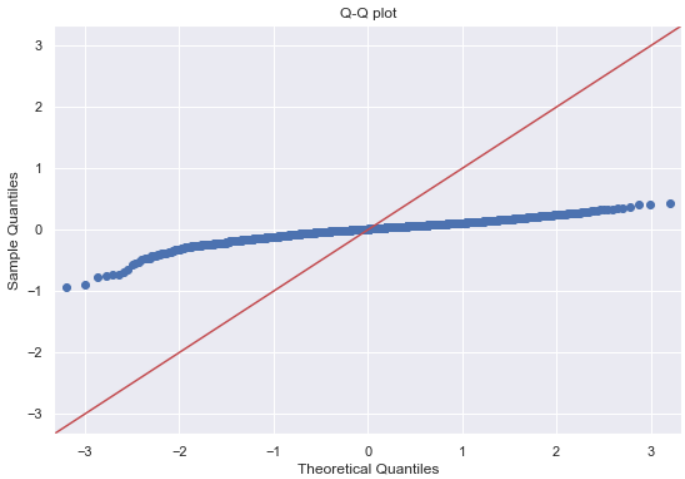


3.2 方差齐性、正态性

进行简单多元最小二乘回归，绘制残差图如下所示，可以发现随着价格升高，残差的方差越来越大，认为模型不满足方差齐性；观察残差散点图，残差均匀分布于 0 附近，认为可能满足零条件均值。

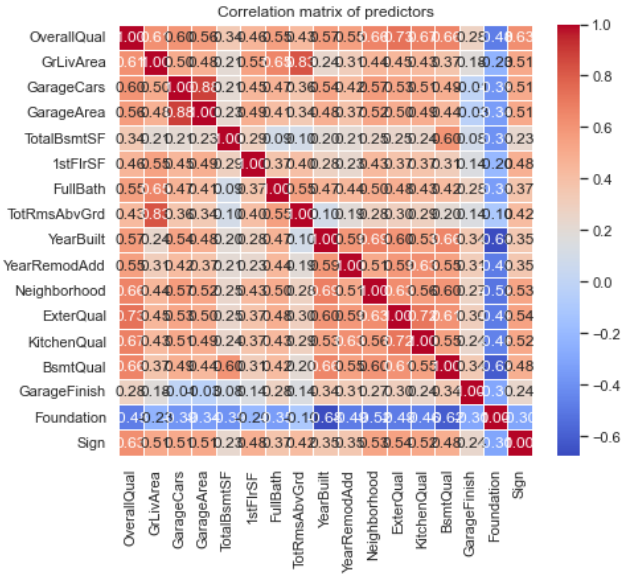


绘制残差的 Q-Q 图如下所示，认为残差不满足正态性假设；进行 Shapiro-Wilk 检验结果 p 值为$1.5924e^{-26}$，认为残差显著不满足正态分布。



3.3 独立性与多重共线性

计算并绘制各个自变量之间的相关系数热力图如下所示，其中 GarageCars 与 GarageArea 之间的相关系数为 0.88，TotalBsmtSF 与 1stFlrSF 之间的相关系数为 0.82，综合考虑认为自变量之间可能存在一定相关性与多重共线性。



进一步进行 VIF 检验，计算 VIF 值如下表，所有自变量的 VIF 值均小于 10，认为不存在显著的多重共线性。

const	63.0844	1stFlrSF	1.7960	ExterQual	2.9656
OverallQual	3.7225	FullBath	2.3205	KitchenQual	2.6242
GrLivArea	5.2371	TotRmsAbvGrd	3.4355	BsmtQual	3.9491
GarageCars	5.3870	YearBuilt	3.8824	GarageFinish	1.4373
GarageArea	4.9377	YearRemodAdd	2.0775	Foundation	2.2201
TotalBsmtSF	2.1220	Neighborhood	2.6808	Sign	2.1188

（注：Sign 变量为人工变量，于 3.1 介绍）

综合以上分析，认为模型具有异方差性，考虑使用加权最小二乘回归等方法；模型可能存在高阶项，考虑添加高次项或者进行 Box-Cox 变换等操作。

四、多元回归模型建立

4.1 模型初探索

考虑将如上所有变量一起进行简单 OLS 回归得到对房价训练集的预测结果如下所示，可以观察到预测的误差也就是波动非常高，说明直接用一条回归线拟合的效果不好，考虑对价格进行分段添加人工变量即惩罚项 Sign，将房价按价格分布划分为 7 类（10%、20%，40%，60%，80%，90%为分割），分别赋值-27、-8，-1，0，1，32，243 以体现价格在极高价区间和极低价区间的异常情况，并且使用 0-1 标准化处理。

其中训练集预测误差为 MSE: 0.02247，MAE: 0.1076。

添加人工变量 Sign 后再次进行第二章的检验，除残差正态性假设未通过，其余结论与第二章一致。

对 17 个自变量再次拟合简单 OLS 回归，并进行 t 检验，结果如下表所示，$R^2_{adj} = 0.926$，有七个自变量没有通过 t 检验，p 值大于 0.05，考虑留下系数显著不为零的 9 个变量进行下一步。

OLS Regression Results

Dep. Variable:	SalePrice	R-squared:	0.927
Model:	OLS	Adj. R-squared:	0.926
Method:	Least Squares	F-statistic:	802.5
Date:	Mon, 16 Dec 2024	Prob (F-statistic):	0.00
Time:	22:49:29	Log-Likelihood:	974.62
No. Observations:	1089	AIC:	-1913.
Df Residuals:	1071	BIC:	-1823.
Df Model:	17		
Covariance Type:	nonrobust		

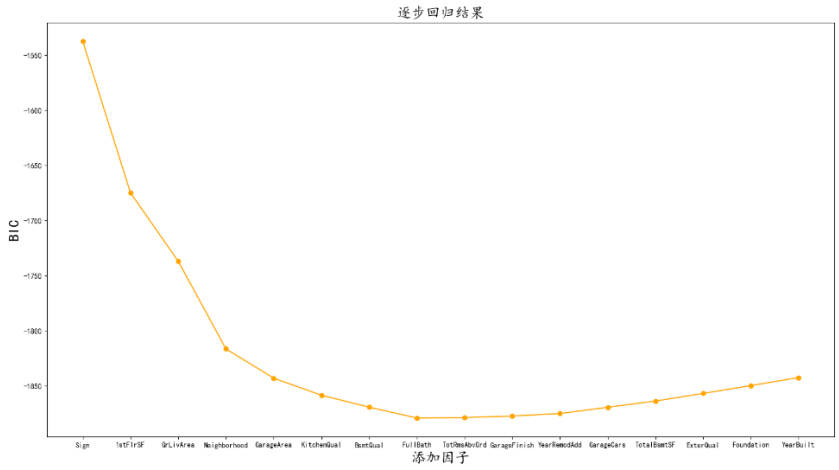
	coef	std err	t	P>|t|	[0.025	0.975]
const	11.8563	0.025	477.493	0.000	11.808	11.905
OverallQual	0.0634	0.006	10.443	0.000	0.051	0.075
GrLivArea	0.1120	0.007	15.580	0.000	0.098	0.126
GarageCars	0.0065	0.007	0.878	0.380	-0.008	0.021
GarageArea	0.0286	0.007	4.016	0.000	0.015	0.043
TotalBsmtSF	0.0348	0.004	7.820	0.000	0.026	0.044
1stFlrSF	0.0399	0.004	9.881	0.000	0.032	0.048
FullBath	-0.0068	0.005	-1.409	0.159	-0.016	0.003
TotRmsAbvGrd	0.0030	0.006	0.511	0.609	-0.008	0.014
YearBuilt	0.0334	0.006	5.328	0.000	0.021	0.046
YearRemodAdd	0.0351	0.004	7.822	0.000	0.026	0.044
Neighborhood	0.2159	0.019	11.373	0.000	0.179	0.253
ExterQual	-0.0085	0.009	-0.899	0.369	-0.027	0.010
KitchenQual	0.0281	0.008	3.718	0.000	0.013	0.043
BsmtQual	-0.0011	0.008	-0.149	0.882	-0.016	0.014
GarageFinish	0.0020	0.004	0.502	0.616	-0.006	0.010
Foundation	-0.0089	0.006	-1.461	0.144	-0.021	0.003
Sign	0.2553	0.016	15.989	0.000	0.224	0.287

Omnibus:	20.315	Durbin-Watson:	2.014
Prob(Omnibus):	0.000	Jarque-Bera (JB):	21.433
Skew:	-0.305	Prob(JB):	2.22e-05
Kurtosis:	3.315	Cond. No.	34.1

4.2 变量贡献度研究

为探究每一个变量对模型效果的影响，采用向前逐步回归，以相关系数最高的变量 OverallQual 为起点，计算每增加一个变量对模型 BIC 值的影响，BIC 值用于评估不同模型的拟合优度，每一步增加选择使得模型 BIC 值最小的变量，直到 BIC 不显著减小。

绘制 BIC 变化折线图如下所示，最终选取 OverallQual，Sign，1stFlrSF，GrLivArea，Neighborhood，GarageArea，KitchenQual，BsmtQual 八个变量作为最终模型的自变量，此时 BIC 值达到-1869 且不再显著减小。



4.3 构建多元回归模型

根据如上分析，选取如上共 9 个自变量建立多元回归模型，使用最小二乘法估计参数，得到回归结果如下表所示，$R^2_{adj} = 0.915$，基本检验均通过，但是 Jarque-Bera 检验 p 值为 0.000164<0.05，拒绝原假设，即认为样本不来自正态总体。

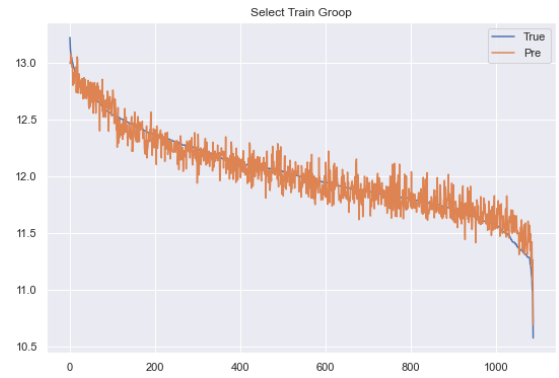
OLS Regression Results

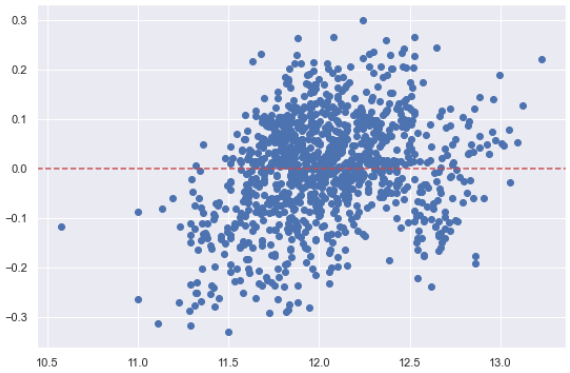
Dep. Variable:	SalePrice	R-squared:	0.915
Model:	OLS	Adj. R-squared:	0.915
Method:	Least Squares	F-statistic:	1456.
Date:	Mon, 16 Dec 2024	Prob (F-statistic):	0.00
Time:	23:23:00	Log-Likelihood:	891.23
No. Observations:	1089	AIC:	-1764.
Df Residuals:	1080	BIC:	-1720.
Df Model:	8		
Covariance Type:	nonrobust		

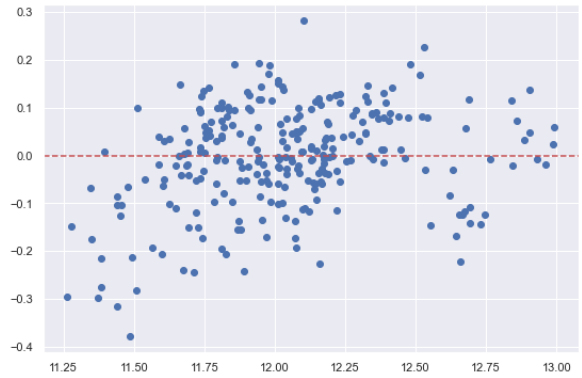
	coef	std err	t	P>|t|	[0.025	0.975]
const	11.6398	0.017	678.728	0.000	11.606	11.673
OverallQual	0.0711	0.006	11.418	0.000	0.059	0.083
Sign	0.2194	0.016	13.400	0.000	0.187	0.252
1stFlrSF	0.0437	0.004	10.449	0.000	0.036	0.052
GrLivArea	0.1063	0.005	23.143	0.000	0.097	0.115
Neighborhood	0.2573	0.018	14.266	0.000	0.222	0.293
GarageArea	0.0382	0.005	8.478	0.000	0.029	0.047
KitchenQual	0.0499	0.007	7.073	0.000	0.036	0.064
BsmtQual	0.0603	0.006	10.197	0.000	0.049	0.072

Omnibus:	16.990	Durbin-Watson:	2.039
Prob(Omnibus):	0.000	Jarque-Bera (JB):	17.427
Skew:	-0.309	Prob(JB):	0.000164
Kurtosis:	3.057	Cond. No.	20.2

绘制训练集预测图与残差图结果以及测试集预测图与残差图结果如下，可以观察到，预测波动性即误差小很多，说明模型稳定性提升较为明显。







其中训练集预测误差为 MSE: 0.01139，MAE: 0.08423；测试集预测误差为 MSE: 0.01143，MAE: 0.08313。对训练集残差进行 Shapiro-Wilk 检验结果 p 值为 0.00027，认为残差仍然存在非正态性。

由第一章分析可知，猜测残差中存在未被解释的高次项信息，在 Stata 中进行 ovtest（Overidentification Test）过度识别检验，发现求得 F 值为 19.88，对应 p 值显著小于 0.05，认为模型存在未解释的遗漏项。增加所选变量的二次项以及三次项进行向后逐步回归，保留 t 检验更显著的变量，最终留下 overallqual、sign、1stflrsf、grlivarea、neighborhood、garagearea、

kitchenqual、bsmtqual、sign_2、sign_3、neighborhood_2、1stflrsf_3、neighborhood_3 共计 13 个变量，但是对结果进行 VIF 检验发现 sign、sign_2、sign_3 三个变量存在高度多重共线性，检验结果如下。

	sign	sign_2	sign_3
VIF	1800.70	1191.65	130.32

使用主成分分析对这三个变量进行降维，第一主成分$\mathrm{Comp1} = 0.5750\mathrm{sign} + 0.5777\mathrm{sign}_2 + 0.5794\mathrm{sign}_3$的贡献达$99.11\%$，选取第一主成分作为 sign、sign_2、sign_3 三个变量的替代特征 sign_pca。

Component	Eigenvalue	Difference	Proportion	Cumulative
Comp1	2.97331	2.94699	0.9911	0.9911
Comp2	.026315	.025935	0.0088	0.9999
Comp3	.000379956	.	0.0001	1.0000

使用带高次项的多元线性回归模型进行最终拟合，结果如下，$\mathrm{R}^2_{adj} = 0.942$，各项检验均已通过。

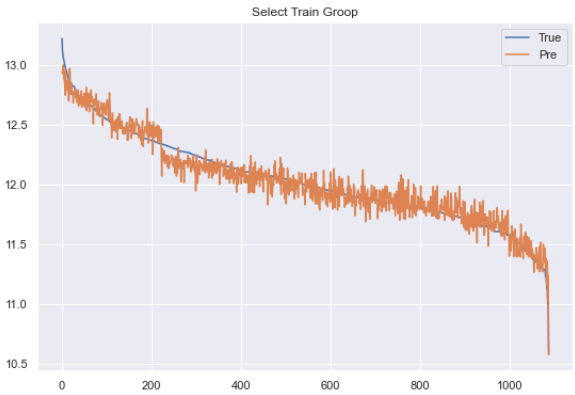
OLS Regression Results

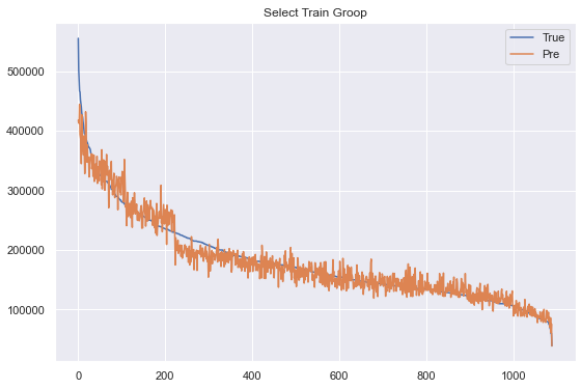
Dep. Variable:	SalePrice	R-squared:	0.943
Model:	OLS	Adj. R-squared:	0.942
Method:	Least Squares	F-statistic:	1472.
Date:	Mon, 16 Dec 2024	Prob (F-statistic):	0.00
Time:	23:23:52	Log-Likelihood:	1103.8
No. Observations:	1089	AIC:	-2182.
Df Residuals:	1076	BIC:	-2117.
Df Model:	12		
Covariance Type:	nonrobust		

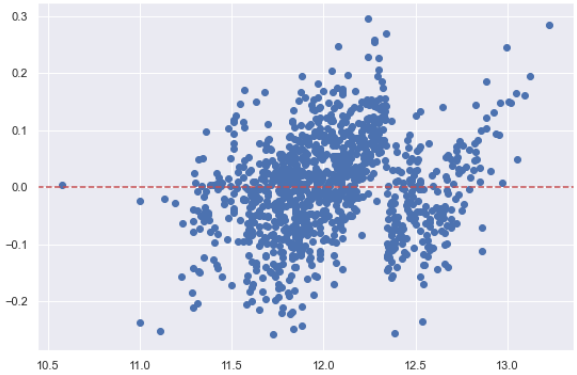
	coef	std err	t	P>|t|	[0.025	0.975]
const	11.5249	0.018	637.058	0.000	11.489	11.560
OverallQual	0.0539	0.005	10.302	0.000	0.044	0.064
Sign	2.9043	0.132	22.019	0.000	2.645	3.163
1stFlrSF	0.0150	0.005	3.211	0.001	0.006	0.024
GrLivArea	0.0820	0.004	20.709	0.000	0.074	0.090
Neighborhood	0.2797	0.028	9.996	0.000	0.225	0.335
GarageArea	0.0258	0.004	6.813	0.000	0.018	0.033
KitchenQual	0.0398	0.006	6.795	0.000	0.028	0.051
BsmtQual	0.0474	0.005	9.535	0.000	0.038	0.057
Neighborhood_2	-0.2238	0.040	-5.602	0.000	-0.302	-0.145
1stFlrSF_3	0.0053	0.001	4.647	0.000	0.003	0.008
Neighborhood_3	-0.5208	0.153	-3.394	0.001	-0.822	-0.220
Sign_pca	-2.1775	0.108	-20.121	0.000	-2.390	-1.965

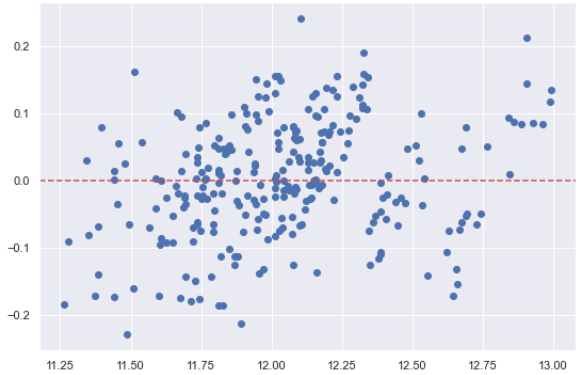
Omnibus:	1.051	Durbin-Watson:	2.084
Prob(Omnibus):	0.591	Jarque-Bera (JB):	0.928
Skew:	-0.034	Prob(JB):	0.629
Kurtosis:	3.126	Cond. No.	242.

绘制训练集预测图与残差图结果以及测试集预测图与残差图结果如下，









其中训练集预测误差为 MSE: 0.00771，MAE: 0.06920；测试集预测误差为 MSE: 0.0075，MAE: 0.0692。对训练集残差进行 Shapiro-Wilk 检验结果 p 值为 0.4278，成功解决残差非正态性问题。

五、总结与展望

5.1 总结

本研究通过对房价数据进行详细的预处理、变量筛选、模型构建与优化，成功建立了多元回归模型以实现房价的精准预测。研究主要结论如下：

1. 数据预处理与变量筛选

通过对连续型变量进行相关性分析与标准化处理，对离散型变量进行均值排序编码、频率编码等方法，有效筛选出对房价影响显著的变量。此外，基于单因素方差分析剔除了无显著影响的变量，确保了模型的精简性与有效性。

2. 回归假设验证与优化

在初步 OLS 回归建模后，通过残差分析发现模型存在异方差性与残差非正态性问题。针对这些问题，研究通过添加惩罚项 Sign、构建高次项、主成分分析降维等技术，逐步优化了模型，使其满足回归假设并显著提升了预测效果。

3. 模型效果与评价

最终模型引入高次项与 PCA 处理，显著改善了预测精度。训练集与测试集的误差均降至较低水平（MSE: 0.00771，MAE: 0.06920），残差正态性假设也成功通过检验，表明模型具备较好的泛化能力与稳定性。

5.2 展望

尽管本研究在房价预测方面取得了一定成效，但仍存在一些未尽之处，例如：进一步探索非线性模型如树模型、集成学习方法，以提升模型精度；如何进行更有效的特征工程以提高模型的学习效率；深入分析残差的空间分布，例如探讨地理位置因素对房价的复杂影响；考虑引入其他外部数据（如经济指数、市场趋势）以丰富模型的解释能力。

总体而言，本研究通过系统的数据处理与模型优化，为房价预测提供了一个可靠的分析框架，并为后续相关研究奠定了基础。

参考文献

[1] Anna Montoya and DataCanary. House Prices - Advanced Regression Techniques. https://kaggle.com/competitions/house-prices-advanced-regression-techniques, 2016. Kaggle.