湖南大學

HUNAN UNIVERSITY

数据挖掘与商务智能决策

期末课程设计

论文题目：	**房价影响因素分析与房价预测**
学生姓名：	杨 超 然
学生学号：	202106060220
专业班级：	电子商务 2021 级 2 班
学院名称：	工商管理学院
指导老师：	江资斌

2023 年 6 月 10 日

房价影响因素分析与房价预测

摘　要

房价影响因素分析是理解房产定价逻辑的一个重要方法，也是预测未来房价变动趋势的一个重要判标。但是在实际的分析中，由于研究人员对数据挖掘技术的理解不深刻，以至于他们在选择模型的过程中存在不严谨性。本文将在进行数据预处理与特征工程的基础上，进行特征重要性分析并通过 matplotlib 展现各个特征的影响程度与特征间的交互作用，再运用 K 折交叉验证与随机网格搜索进行算法比较与模型选择，从而实现对房价影响因素的准确分析与预测。

关键词：房价分析与预测；特征重要性分析；K 折交叉验证；随机网格搜索；算法比较与模型选择

Analysis of influencing factors and forecast of housing price

Abstract

analysis of influencing factors of housing price is an important method to understand the logic of real estate pricing, and also an important judgment to predict the trend of future housing price changes. However, in the actual analysis, because researchers do not have a deep understanding of data mining technology, they are not rigorous in the selection of models. On the basis of data preprocessing and feature engineering, this paper will analyze the importance of features, show the influence degree of each feature and the interaction between features through matplotlib, and then use K-fold cross-validation and random grid search for algorithm comparison and model selection, so as to achieve accurate analysis and prediction of housing price influencing factors.

Key words: Housing price analysis and forecast; Feature importance analysis; K-fold cross verification; Random grid search; Algorithm comparison and model selection

目录

一、绪 论

（一）研究背景

“为什么这个房子这么贵？”、“到底值不值得这么贵？”应该是许多消费者在买房过程中会发自内心问出的问题。作为可能会影响消费者甚至其后代的重大消费决策，我们在进行房屋的购置时，总是会考虑再三。房价，作为房屋价值在考虑多种影响因素后的货币表现，面对着这个看起来有些冰冷的数字，我们是否会发出疑问：这个数字究竟从何而来，受到哪些因素的影响，这些因素之间是否存在着千丝万缕的联系，又会导致房屋价格呈现何种变化趋势呢？

进行房价分析与预测，可以很好的解答这一疑问。在此次实验中，我们将从房屋的常见特征入手（同样也是被消费者和销售者纳入考量频率较高的房屋特征），依据各个特征指标和对应的房屋销售价格，建立合适的模型，进行特征重要性分析与房价预测与回归分析，从而探究房价主要受哪些特征影响，并尝试着在给定这些特征值的条件下，对房价进行预测。

（二）国内外研究文献综述

目前国内对于房价影响因素分析的研究文章和著作都比较多。由文献[5，7-9]可知，基本上每一本新闻学原理著作，在论述新闻真实性时都会提到整体真实理论，对于整体真实的概念也是各有说法[9，11，13-15]。

目前国内外对于房价影响因素分析的研究文章和著作都比较多，但研究方法各有差异。如文献[1-2，8-9，12]侧重于建立纯粹的计量模型进行分析，文献[3-4，10]运用统计学方法，而文献[5-7，11]则使用了相关算法于电脑分析软件。研究结果方面，基本上每一篇文献都提到了 GDP 对房价的影响，文献[7]给出了 GDP 与房价的精确变动比例。但当前研究仍存在不足，如缺少对机器学习的运用，缺少方法间的比较与择优过程，且各个文献都注重于分析房屋外因素对房屋的影响，对房屋本身的特征分析有所忽略。基于以上，本文提出如下改进方案：一是尝试运用机器学习搭建模型进行分析，二是进行多算法之间的比较与选择，三是重点研究房屋特征。

（三）研究目的与意义

进行房价影响因素分析与房价预测，对于消费者而言，该分析得到的重要特征可以为经验不足的消费者提供一定的考量依据，得到的价格预测也可以作为其购置房屋时的参考；对于营销人员而言，特征重要性分析可以其找到消费者最看重房屋

的哪些“硬软件配置”，从而抓住要点进行精准营销与宣传；对于政府相关部门而言，进行房价分析与预测能够为国家房价调控提供参考。综上所述，开展房价影响因素分析与预测，对于多方而言都是一大利好，具有深刻的实验分析意义。

二、问题描述

（一）数据来源

本次实验所用到的相关数据来源于 Kaggle 入门预测竞赛中的“House Prices - Advanced Regression Techniques”房屋价格预测竞赛数据集，其中包含艾姆斯市（Ames city）的房屋原始数据与训练集 train.csv、用于预测的测试集 test.csv。相关重要数据字段包括售价 (salePrice)、住宅类型 (MSSubClass)、住宅分区 (MSZoning) 等，完整的数据字段描述都存储见附录：

需要特别说明的是，并非所有数据都是数值类型（很好理解，如“类型”类数据，以房屋类型为例，如此多的房屋种类，全部进行 labelcoding 编号处理不太直观、也不太现实），而我们的目标变量 SalePrice，则是数值类型。此外，还有已经进行了编号处理的变量（如 OverallQual，以 1-10 表示房屋的质量评级，1 为最好，10 为最差），因此本数据集中的数据肯定不能直接拿去分析，需要先进行数据格式转换。

（二）研究内容

本实验将依据数据集中的房屋特征 (如平均居住面积、住宅类型、装修风格等) 与房屋销售价格之间的关系展开探究与分析，以解决如下问题：

1.探究房价各个特征的影响程度：基于房屋本身特征，通过量化方法，选择适当的评估标准展现各个因素的影响程度并进行重要性排名。

2.探究各个特征变量间的联系：将各个特征两两结合，找到对模型具有重要影响的交互特征

3.房价预测：在给定房屋特征条件下，对房屋销售价格实现较为准确可信的预测

三、数据预处理

（一）数据读取

首先，对下载的数据进行分类，创建文件夹 input,将数据训练集和测试集放在‘house-prices-advanced-regression-techniques’文件夹中，所有原始数据放在‘ames-housing-dataset’文件夹中，路径为：‘/User/yangchaoran/Desktop/input’。

接着，导入相关库，通过 pandas 中的 read_csv 方法读取原始数据与训练集，使用 replace()函数替换数据中的空白值，并通过 info()和 head()展示样本数据的相关信息

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2930 entries, 0 to 2929
Data columns (total 82 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Order           2930 non-null   int64
 1   PID             2930 non-null   int64
 2   MSSubClass      2930 non-null   int64
 3   MSZoning        2930 non-null   object
 4   LotFrontage     2440 non-null   float64
 5   LotArea         2930 non-null   int64
 6   Street          2930 non-null   object
 7   Alley           198 non-null    object
 8   LotShape        2930 non-null   object
 9   LandContour     2930 non-null   object
 10  Utilities       2930 non-null   object
 11  LotConfig       2930 non-null   object
 12  LandSlope       2930 non-null   object
 13  Neighborhood    2930 non-null   object
 14  Condition1      2930 non-null   object
 15  Condition2      2930 non-null   object
 16  BldgType        2930 non-null   object
 17  HouseStyle      2930 non-null   object
 18  OverallQual     2930 non-null   int64
 19  OverallCond     2930 non-null   int64
 20  YearBuilt       2930 non-null   int64
 21  YearRemod/Add   2930 non-null   int64
 22  RoofStyle       2930 non-null   object
 23  RoofMatl        2930 non-null   object
 24  Exterior1st     2930 non-null   object
 25  Exterior2nd     2930 non-null   object
 26  MasVnrType      2907 non-null   object
 27  MasVnrArea      2907 non-null   float64
 28  ExterQual       2930 non-null   object
```

	Order	PID	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	...	PoolArea	Po
0	1	526301100	20	RL	141.0	31770	Pave	NaN	IR1	Lvl	...	0	
1	2	526350040	20	RH	80.0	11622	Pave	NaN	Reg	Lvl	...	0	
2	3	526351010	20	RL	81.0	14267	Pave	NaN	IR1	Lvl	...	0	
3	4	526353030	20	RL	93.0	11160	Pave	NaN	Reg	Lvl	...	0	
4	5	527105010	60	RL	74.0	13830	Pave	NaN	IR1	Lvl	...	0	

5 rows × 82 columns

图 1 表样本数据相关信息

（二）数据划分与异常处理

在进行探索性分析与算法比较，我们需要对数据的特征和其中的数据情况有了一定的了解，因此需要进行变量提取、数据集划分与处理其中的缺失值、重复值等异常数据。

需要注意的是，在真实的项目中，测试数据直到最后才可用。因此，测试数据应该包含与训练集相同类型的数据，以相同的方式进行预处理。在这里，测试集是可用的。它包含一些在训练数据集中没有出现的观察结果，并且使用虚拟编码可能会引发几个问题。解决这个问题的最简单的方法(如果没有可用的测试数据，则不适用)是将训练集和测试集连接起来，预处理并再次划分它们。

1．提取目标变量与特征变量

首先，将目标变量与特征变量通过 df.drop()函数单独提取出来，再使用 pd.concat()函数将训练集与测试集进行连接。

2．划分测试集与训练集

由于 Kaggle 直接给我们了“train”数据集，因此只要根据下载的训练集路径直接进行读取即可，而 house_data 中的剩余部分就会被归位测试集。

```
target = data_w['SalePrice']
test_id = test['Id']
test = test.drop(['Id'],axis = 1)
data_w2 = data_w.drop(['SalePrice','Order','PID'], axis = 1)
```

```
train_test = pd.concat([data_w2,test], axis=0, sort=False)
```

图 2 变量提取与数据集合并

3．数据清洗

在收集到的数据中，通常会存在不同程度的数据缺失情况，因此我们需要查找不同类型数据的缺失情况，并进行相应的处理。

（1）数据缺失情况探究

查找各个特征的缺失值数量，加总储存在 Nan_sum 中，并用缺失值数量除以数据集样本数量，得到数据缺失比重。最后进行可视化展示，查看哪些变量数据缺失情况较为严重，则需要我们优先进行处理。

	NaN_sum	feat	Perc(%)	Usability
Exterior1st	1	Exterior1st	0.068493	Keep
Exterior2nd	1	Exterior2nd	0.068493	Keep
KitchenQual	1	KitchenQual	0.068493	Keep
Electrical	1	Electrical	0.068493	Keep
SaleType	1	SaleType	0.068493	Keep
TotalBsmtSF	2	TotalBsmtSF	0.136986	Keep
GarageArea	2	GarageArea	0.136986	Keep
GarageCars	2	GarageCars	0.136986	Keep
Utilities	2	Utilities	0.136986	Keep
Functional	2	Functional	0.136986	Keep
BsmtUnfSF	2	BsmtUnfSF	0.136986	Keep
BsmtFinSF1	2	BsmtFinSF1	0.136986	Keep

图 3 数据缺失值 DataFrame(部分)

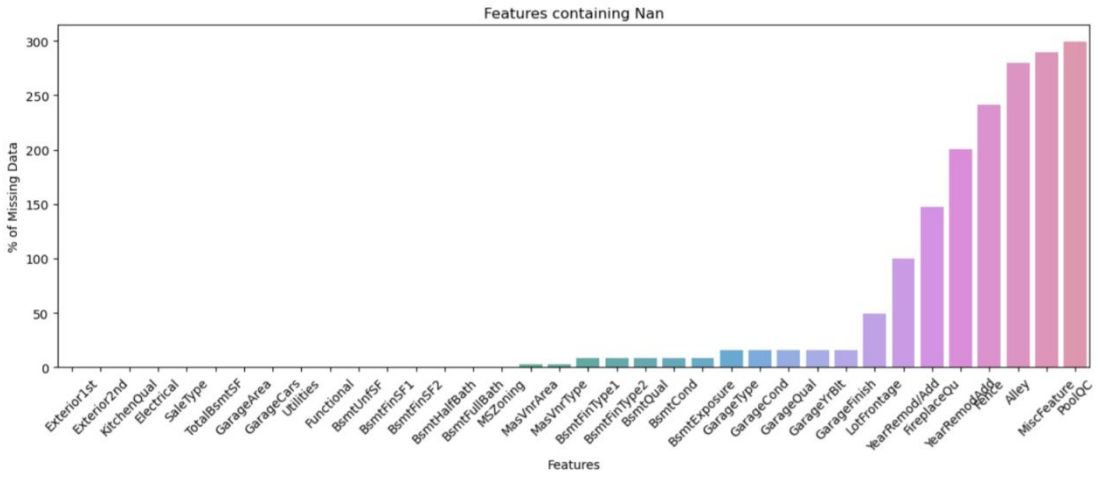

图 4 数据缺失率柱状图

可见：例如泳池质量（PoolQC）、其他类别中未包含的杂项功能（MiscFeature）等含有 NaN 的特征变量有着近乎夸张的数据丢失率，依据上述可视化结果就对相关特征妄下论断显然是有失偏颇的。

通过查看数据集中这些特征变量的内容与变量描述文件,需要进一步确认的是：所有这些 NaN 都是真正的缺失值吗?因为我们可以看到这些 NaN 中的大多数是如何反映某些东西的缺失的，因此，它们并不是真正的 NaN 值或是缺失值。我们可以计算它们(对于数值特征)或用文件中的数据替换它们。

（2）缺失数据处理

对于变量中的缺失数据，我们需要有针对性地进行处理：幸运的是，对于训练集中的大部分含 NaN 值特征，因为有对应的变量描述，因此我们可以依据特征含义填充 NaN 值,而对于销售量等本身为数值类型的数据，则需要将转化为字符串。

```
train_test['MSSubClass'] = train_test['MSSubClass'].apply(str)
train_test['YrSold'] = train_test['YrSold'].apply(str)
train_test['MoSold'] = train_test['MoSold'].apply(str)
```

```
train_test['Functional'] = train_test['Functional'].fillna('Typ')
train_test['Electrical'] = train_test['Electrical'].fillna("SBrkr")
train_test['KitchenQual'] = train_test['KitchenQual'].fillna("TA")
train_test['Exterior1st'] = train_test['Exterior1st'].fillna(train_test['Exterior1st'].mode()
train_test['Exterior2nd'] = train_test['Exterior2nd'].fillna(train_test['Exterior2nd'].mode()
train_test['SaleType'] = train_test['SaleType'].fillna(train_test['SaleType'].mode()[0])
train_test["PoolQC"] = train_test["PoolQC"].fillna("None")
train_test["Alley"] = train_test["Alley"].fillna("None")
train_test['FireplaceQu'] = train_test['FireplaceQu'].fillna("None")
train_test['Fence'] = train_test['Fence'].fillna("None")
train_test['MiscFeature'] = train_test['MiscFeature'].fillna("None")

for col in ('GarageArea', 'GarageCars'):
    train_test[col] = train_test[col].fillna(0)

for col in ['GarageType', 'GarageFinish', 'GarageQual', 'GarageCond']:
    train_test[col] = train_test[col].fillna('None')

for col in ('BsmtQual', 'BsmtCond', 'BsmtExposure', 'BsmtFinType1', 'BsmtFinType2'):
    train_test[col] = train_test[col].fillna('None')
```

图 5 数据转换、填充过程

处理完毕后，核对相关特征的类型：利用遍历的方法，将含有数值类型的数据进行打印，看看其中是否还有未处理干净的特征，处理结果如图 6:

```
0    RL
0    RH
Name: MSZoning, dtype: object
0    141.0
0     80.0
Name: LotFrontage, dtype: float64
0    AllPub
0    AllPub
Name: Utilities, dtype: object
0    1960.0
0       NaN
Name: YearRemod/Add, dtype: float64
0    Stone
0     None
Name: MasVnrType, dtype: object
0    112.0
0      0.0
Name: MasVnrArea, dtype: float64
0    639.0
0    468.0
Name: BsmtFinSF1, dtype: float64
0      0.0
0    144.0
Name: BsmtFinSF2, dtype: float64
0    441.0
0    270.0
Name: BsmtUnfSF, dtype: float64
0    1080.0
```

```
0    270.0
Name: BsmtUnfSF, dtype: float64
0    1080.0
0     882.0
Name: TotalBsmtSF, dtype: float64
0    1.0
0    0.0
Name: BsmtFullBath, dtype: float64
0    0.0
0    0.0
Name: BsmtHalfBath, dtype: float64
0    1960.0
0    1961.0
Name: GarageYrBlt, dtype: float64
0       NaN
0    1961.0
Name: YearRemodAdd, dtype: float64
```

图 6 数据处理结果（部分）

（三）数据估计与更新

K 近邻算法，是在已有数据中寻找与它最相似的 K 个数据，或者说“离它最近”的 K 个数据，如果这 K 个点大多数属于某一个类别，则该样本也属于这个类别。

在实验过程中，我们搭建 KNN 回归模型，进行数据估算预测，构建函数对训练集和测试集进行划分，其中缺少数据的索引将成为我们的测试集。再使用 KNR 无监督方法进行回归训练。

完成模型搭建后，使用 update()对数据进行更新即可，具体见源代码。

四、数据探索性分析

在进行后续算法比较模型搭建之前，理解数据是很重要的。实现这一目标的关键一步是探索性数据分析(EDA)、可视化和统计分析(单一、双向和多元)的结合，

帮助我们更好地理解我们正在处理的数据，并深入了解它们之间的关系。在这一部分中，将从目标变量以及其他特征出发进行探索性分析与可视化展现，看看相关特征变量是如何影响到整体数据的。

（一）目标变量分布情况探究

通过可视化方法，使用 matplotlib 绘制目标变量 SalePrice 的样本分布，结果如图 7:

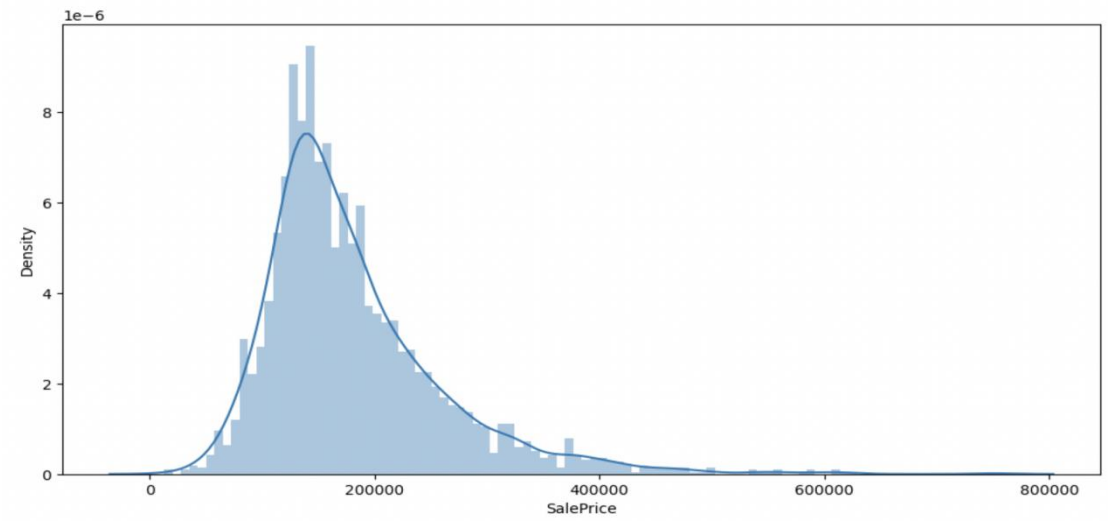

图 7 SalePrice 分布情况

由于 SalePrice 分布情况与正太分布类似，我们需要进行判断，首先通过绘制多图，直观展现目标变量分布与正态分布的关系，如图 8:

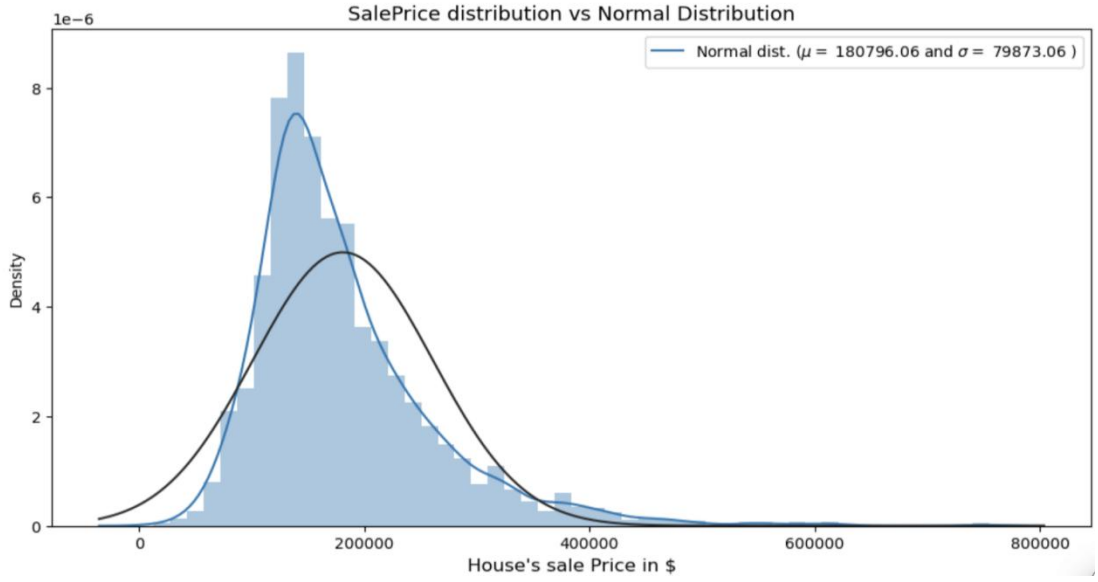

图 8 SalePrice 分布与正态分布比对

可见，SalePrice 的分布由于高度右偏，似乎不服从正太分布，因此需要进一步的探究。查阅资料可知：偏度的可接受范围为[-0.5,0.5]之间，峰度的可接受范围在[-2,2]之间[①]，因此，借助 Shapiro 正态性检验[②]，通过 python 数学计算库 scipy 计算其峰度（kurtosis）、偏度（skewness）和 Shapiro 检验的 p 值等相关参数，输出结果如表 1：

表 1 SalePrice Shapiro 正态性检验参数结果

	输出值	服从正太分布下的可接受范围
Skewness	1.743500	[-0.5,0.5]
Kurtosis	5.118900	[-2,2]
Shapiro_Test	0.000000	—

可见，输出结果显然远远大于了这个范围，此外，Shapiro 检验的 p 值输出为 0，也说明要拒绝“参数是正态分布”的原假设，因此我们可以做出 SalePrice 是不服从正态分布的判断。但尽管如此，我们在此还是先将结果放在这里，在特征工程部分，我们将对目标分布进行标准化处理以将其转化为正太分布。

（二）特征相关性分析

对已知变量进行相关性分析，可以衡量两个变量因素的相关密切程度。由于本次分析的特征较多，通过 python 绘制相关系数矩阵热力图，可以相对直观全面地展现两个变量间的相关性强弱，如图 9:

[①]崔恒建,成平.PP 偏度,峰度正态性检验的 P—VALUES[J].自然科学进展：国家重点实验室通讯, 1995, 5(6):7.DOI:CNKI:SUN:ZKJZ.0.1995-06-007.
[②]孙玉芝,李春禄.介绍两种正态性检验方法[J].天津师大学报：自然科学版, 1992, 000(001):30-34.

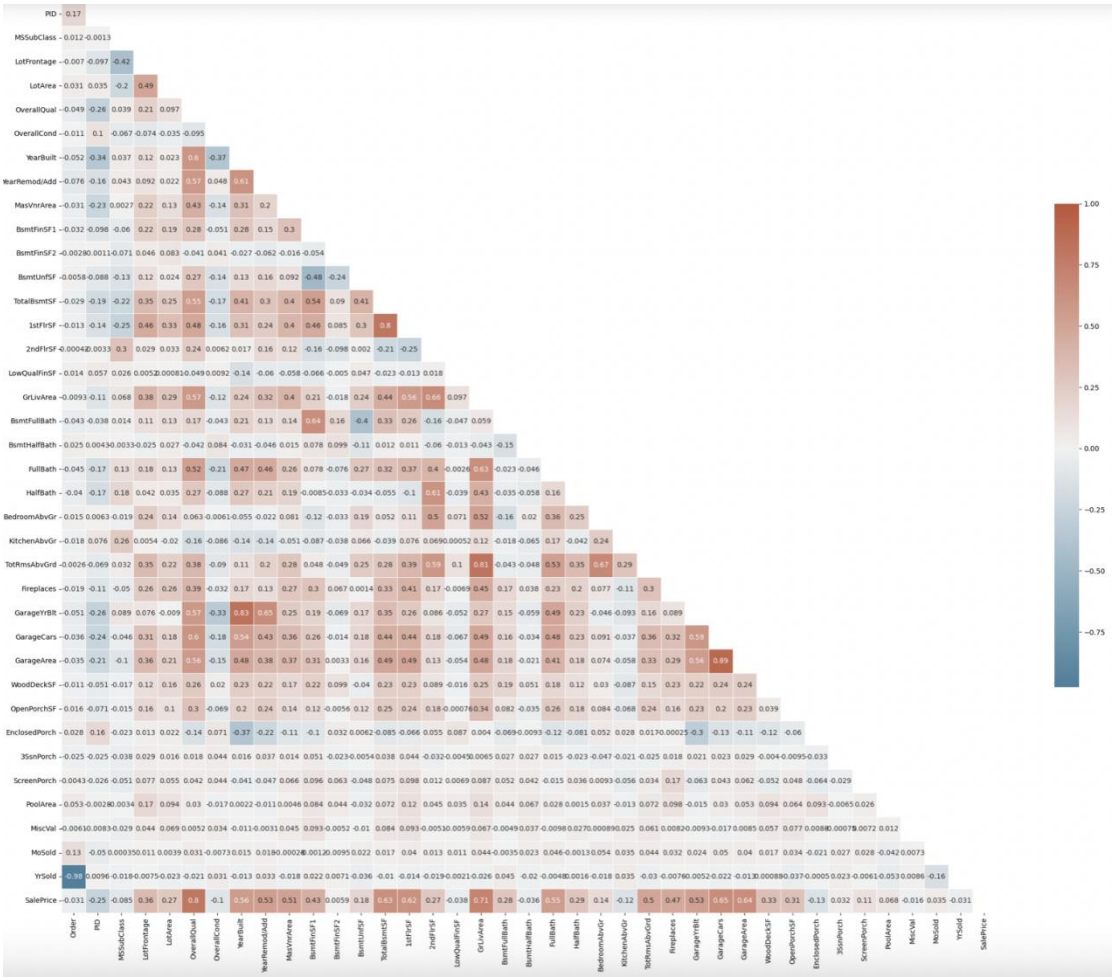

图 9 样本特征相关系数矩阵热力图

其中，矩阵取值均在-1 到 1 之间，且绝对值越接近于 1，说明两者越相关。在热力图中，特征与目标变量越相关，对应区块的颜色就越深。依据输出的相关矩阵热力图，以主要研究的销售价格 SalePrice 为例，与该特征最为相关的变量为整体质量 OverallQual，除此之外，还有地面居住面积 GriLivArea、小区可容纳车辆能力 GarageCars 等变量，如图 10：

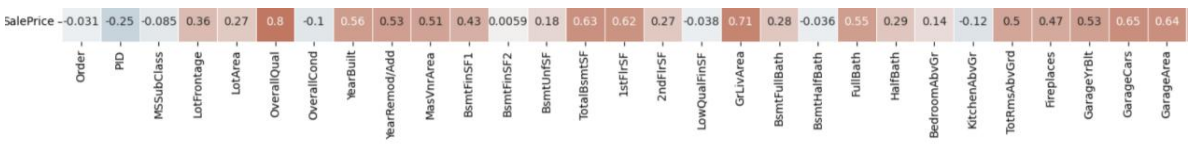

图 10 SalePrice 与其他特征的相关性情况

（三）特目标变量与重要特征的关系探究

基于上述相关矩阵分析，我们得到与 SalePrice 最为相关的特征为 OverallQual、TotRmsAbvGrd 等。依据我们的相关常识，我们可以做出假设：如房屋住宅区的整体质量越好，房屋的销售价格就会越高。因此在此部分我们将通过绘制不同的图形，以可视化探究的方式对 SalePrice 的分布区间与同特征间的关系（如是否为线性关系等）进行分析。

展示方面，散点图、箱型图和提琴图等图形直观展现了特征变量的分布情况与分布趋势，故将作为我们的可视化结果

1.房屋价格与房屋整体质量间的关系

通过 matplotlib，绘制 SalePrice 与 OverallQual 间的散点图、箱形图与提琴图，如图 11:

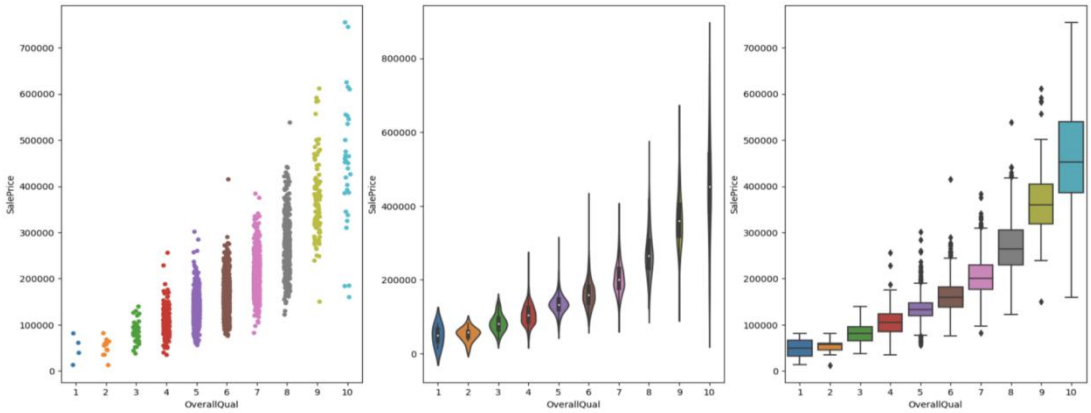

图 11 SalePrice 与 OverallQual 间的散点图、提琴图与箱形图

可见 SalePrice 与 OverallQual 的关系符合我们的预期，即当房屋的整体质量水平较高时，房屋价格也分布于较高的范围内。

2.房屋价格与房屋总体房间数间的关系

同理，修改参数即可绘制相关图形，输出结果如图 12:

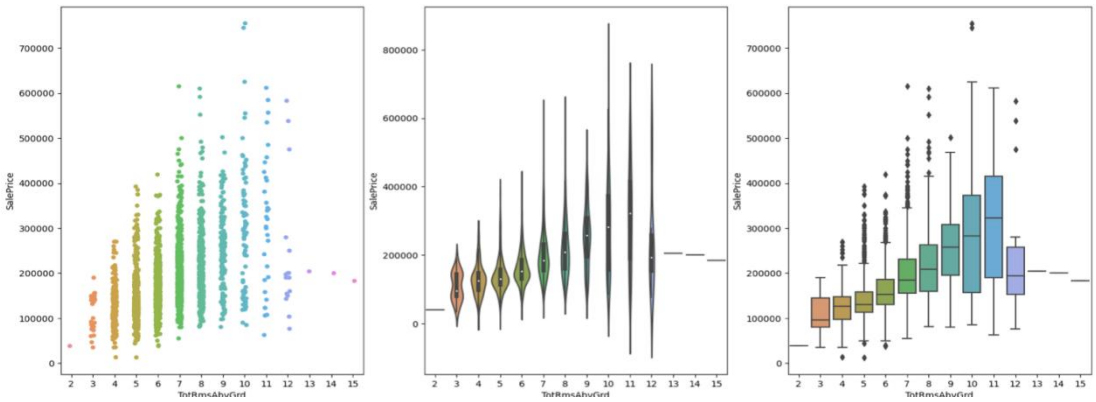

图 12 SalePrice 与 TotRmsAbvGrd 间的散点图、提琴图与箱形图

可见，如果对于楼房销售而言，平均房间数在合理范围内增加在一定程度上是有助于拉高房价的，但绝对不能称之为重要因素，因为从箱型图来看，该特征并未明显地改变房价数据的分布情况，且值得注意的是，当平均房间数过高（大于 9）时，反而可能会对房屋价格带来负效应。所以，房屋质量需越高越好，但房间数量可不是越多越好的。

3.房屋价格与房屋居住面积间的关系

通过 python 绘制散点回归图，探究房屋价格 SalePrice 与房屋居住面积 GrLivArea 间的关系，输出结果如图 13:

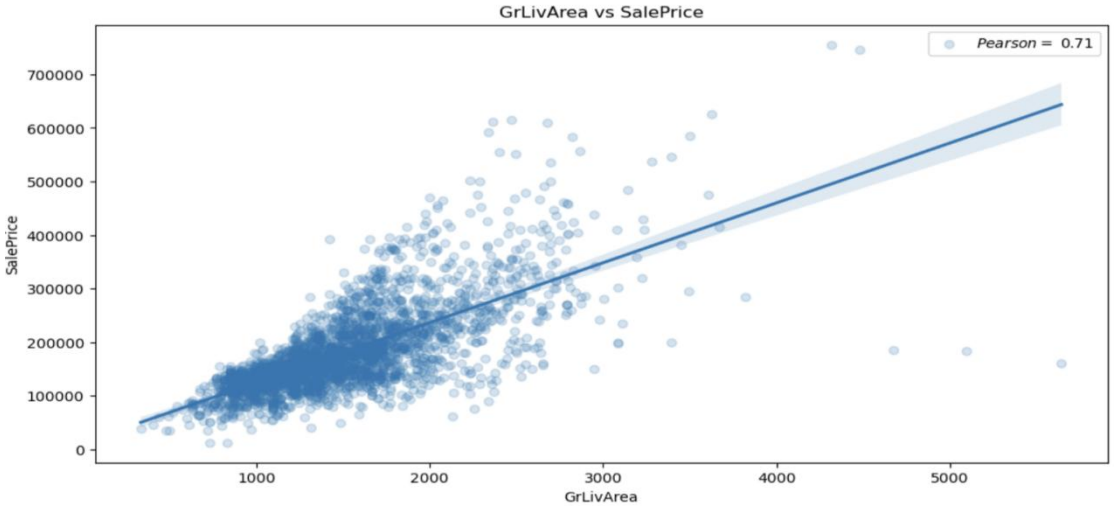

图 13 SalePrice 与 GrLivArea 的散点回归图

可见，平均居住面积和房价呈正相关的线性关系，即一般居住面积较大的房屋，价格也越高。

4.房屋价格与地下室总平方数间的关系

同理，绘制两个变量间的散点回归图如图 14:

图 14 SalePrice 与 TotalBsmtSF 的散点回归图

可见，地下室总平方数和房价呈正相关的线性关系，即一般地下室面积较大的房屋，价格也越高，与上述的居住面积一起分析，不难看出若其他特征都处于正常水平，房屋的整体面积越大，卖的就越贵。

5.房屋价格与建造年份间的关系

绘制两个变量间的散点回归图如图 15:

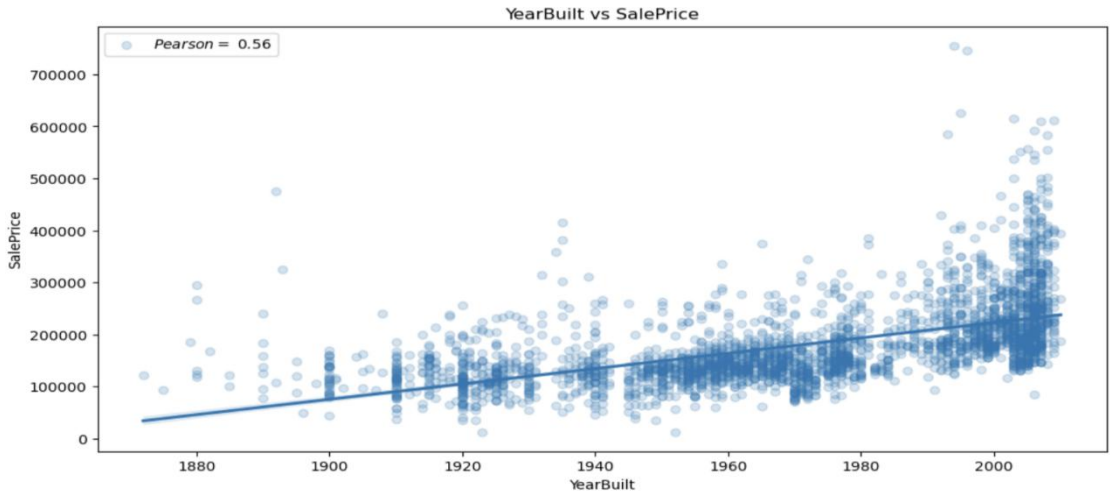

图 15 SalePrice 与 YearBulit 的散点回归图

可见，可见建造年份越靠后，即越新建造的房屋，销售价格相对旧房屋而言会更高。

6.房屋价格与售量年份间的关系

考虑房屋销售年份 YrSold 与销售价格 SalePrice，求得房屋年销售的中位数，并通过柱状图进行可视化展示，可直观反映房价随销售量随时间的波动情况：

图 16 基于不同年份下的 SalePrice 中位数柱状图

可见，从中位数角度考量，柱状图高度差异并不明显，即房价的波动并不大，销售价格趋于平稳。

五、特征工程

在进行数据预处理后，现在的数据就已经是“准备好”了的，接下来，将通过特征工程对前述特征进行进一步的分析与处理，构造一些新的特征，进行数据转化和哑变量处理，以提高决策性能。

（一）特征使用方案

由于数据集中的数据较为完备，且在前述过程中已经进行了初步的数据预处理，故无需考虑数据获取的难度与数据覆盖率等问题。我们只需将房价 SalePrice 作为因变量，将其余房屋特征作为自变量即可。

除此之外，基于探索性数据分析，我们还构造了一些新特征来评估房价：

SqFtPerRoom -房屋居住空间占比，测度一个房屋中有多少空间是可供人们居住，而非用作其他功能用途，通过居住面积/房间面积（不包括浴室厨房）+功能性房间面积得到

Total_Home_Quality-整体房屋质量，测度一个房屋的整体的硬软件设施情况由整体质量与整体居住环境两个指标相加得到。

Total_Bathrooms-浴室总和，测度房屋中浴室的占地情况，通过赋予完全的浴室（bath）1 的权重，半浴室半其他房间（halfbath）0.5 的权重，各个楼层的 bath 和 half 相加得到。

HighQualSF-精装面积，用于测度房屋的装修水平，由各个楼层的被划分为“HighQual”英尺数相加得到。

（二）特征获取方案

对于非新增特征，通过 python 读取数据集即可实现获取，而新增特征，需要通过代码进行构造与存储

```
train_test["SqFtPerRoom"] = train_test["GrLivArea"] / (train_test["TotRmsAbvGrd"] +
                                                       train_test["FullBath"] +
                                                       train_test["HalfBath"] +
                                                       train_test["KitchenAbvGr"])

train_test['Total_Home_Quality'] = train_test['OverallQual'] + train_test['OverallCond']

train_test['Total_Bathrooms'] = (train_test['FullBath'] + (0.5 * train_test['HalfBath']) +
                                 train_test['BsmtFullBath'] + (0.5 * train_test['BsmtHalfBath'

train_test["HighQualSF"] = train_test["1stFlrSF"] + train_test["2ndFlrSF"]
```

图 17 新特征构造

（三）进一步的特征处理

在加入新构造的变量后，需要做进一步的特征处理，在进行非数值类型数据的处理和 dummy 哑变量处理的基础上，还需要进行数据标准化解决将数据探索性分析中提到的 SalePrice 分布问题。

这里使用对数转换缩小 SalePrice 的绝对范围，从而将其变换到正态分布中，消除数据之间的量纲问题，使数据看起来更加规整。

$$X_{log} = log(X)$$

在完成标准化前后，分别绘制 SalePrice 的分位数-分位数图 Quantile-Quantile Plot 和样本分布图查看转化效果，如图 18 与图 19 所示：

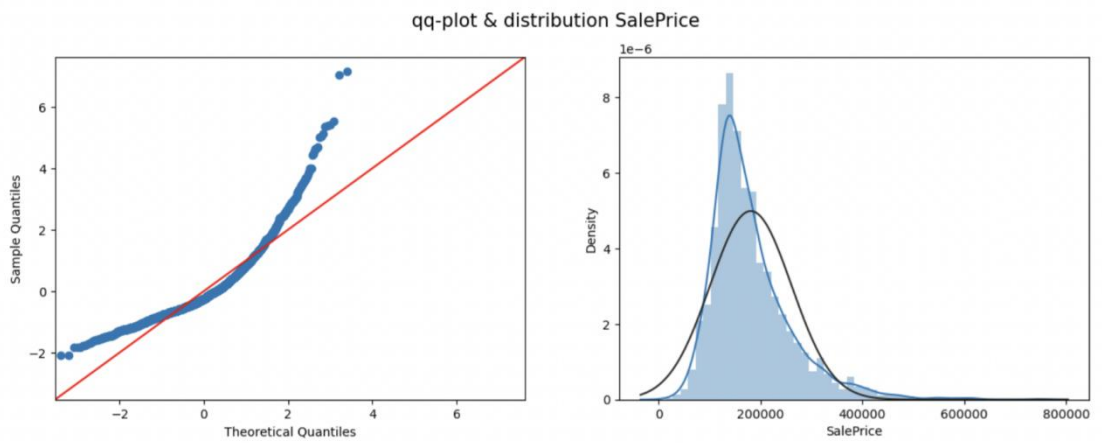

图 18 标准化前的 qq-plot 与数据分布图

可见：此时的 SalePrice 分布和探索性分析中得到的结果是一样的，且在 qq-plot 中，样本点显然不呈直线分布，说明其与猜测的正态分布并不相同，也符合我们之前得到的结论。

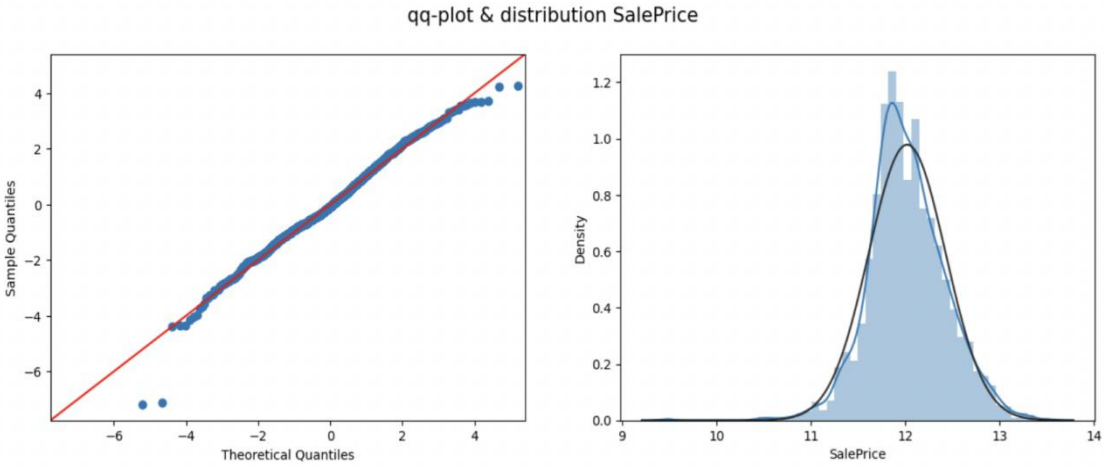

图 19 标准化后的 qq-plot 与数据分布图

可见，此时的 qq-plot 已经近乎为一条直线，分布图也与正态分布一致，说明我们成功完成了标准化，接下来便可以利用数据搭建模型，进行特征评估与预测了。

六、算法比较与模型选择

（一）模型训练与评估方法

模型训练阶段，采用 K 折交叉验证的方法。原理为：将数据集随机等分为 K 份，每次选取 K-1 份为训练集训练模型，然后用剩下的 1 份作为测试集，得到 K 个模型后将这 K 个模型的平均测试效果作为最终的模型效果

模型效果评估指标，选择均方根误差：均方根误差 (RMSE) 作为均方误差 (MSE) 的平方根形式，其本质上是反应测量数据与真是数据的偏离程度，RMSE 越小，则测量精度越高，模型表现越好，其公式为：

$$RMSE = \sqrt{MSE} = \sqrt{\frac{SSE}{n}} = \sqrt{\frac{1}{n}\sum_{i-1}^{n}(y_i - y_i^g)}$$

（二）模型搭建与训练

令 K=10，搭建线性回归、LightGBM、支持向量机回归、决策树回归、CatBoost 回归、随机森林回归等 10 个回归模型进行训练，并将其 RMSE 值储存在 DataFrame 中。

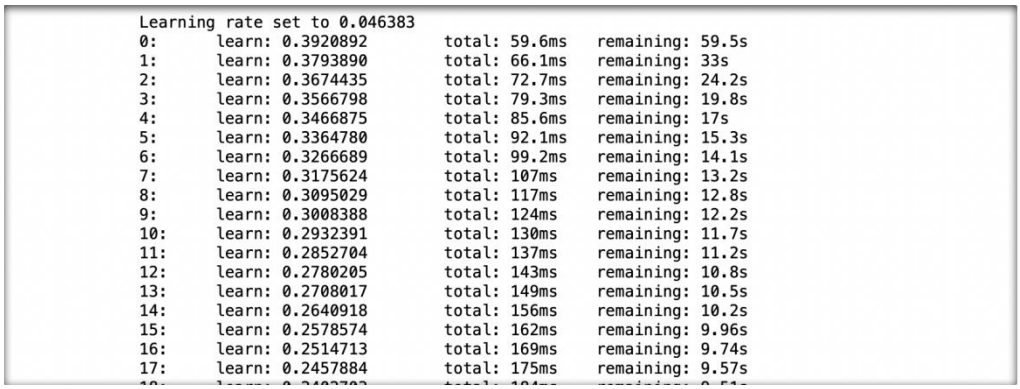

图 20 模型训练过程

（三）算法比较与模型选择

模型训练完成后，查看基础模型选择，并绘制模型得分柱状图：

表 2 基础模型得分

	Regressors	RMSE_mean	RMSE_std
0	Linear_Reg.	0.139996	0.021668

1	Bayesian_Ridge_Reg.	0.121452	0.018995
2	LGBM_Reg.	0.123800	0.013532
3	SVR	0.269460	0.019030
4	Dec_Tree_Reg.	0.197853	0.023202
5	Random_Forest_Reg.	0.135913	0.012990
6	XGB_Reg.	0.131712	0.015318
7	Grad_Boost_Reg.	0.123772	0.011377
8	Cat_Boost_Reg.	0.113339	0.016694
9	Stacked_Reg.	0.115086	0.014753

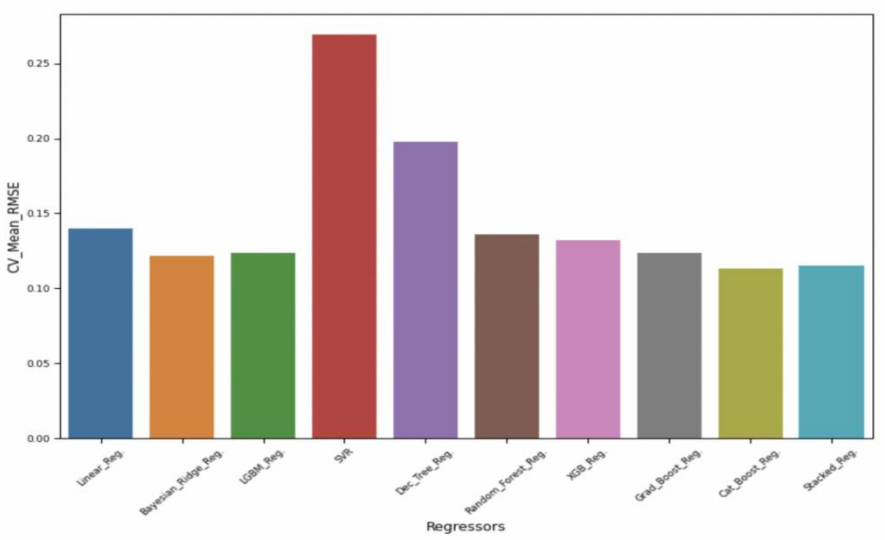

图 21 基础模型得分柱状图

几个 GBDT 算法的 RMSE 均值和误差都不大，且 Cat Boost 回归模型有着最小的均方根误差平均值，因此在初步认定 GBDT 算法在此次数据集回归分析表现较好的基础上，我们选择 Cat Boost 算法搭建最终的模型，进行最终的模型搭建。

七、模型搭建

基于算法比较研究结论，搭建 Cat Boost 回归模型。在对数据集 Ameshousing.csv 进行划分后，投喂数据进行预测，并通过计算 RMSE 值查看模型得分

```
In [49]: # 划分训练集和测试集

         X_train,X_val,y_train,y_val = train_test_split(train,target_log,test_size = 0.1,random_state

         # Cat Boost 回归模型搭建

         cat = CatBoostRegressor()
         cat_model = cat.fit(X_train,y_train,
                             eval_set = (X_val,y_val),
                             plot=True,
                             verbose = 0)

         MetricVisualizer(layout=Layout(align_self='stretch', height='500px'))

In [50]: # 查看模型得分
         cat_pred = cat_model.predict(X_val)
         cat_score = rmse(y_val, cat_pred)
         cat_score

Out[50]: 0.095326965478548
```

图 22 模型搭建与预测过程

可见，当前模型的 RMSE 值已经很低了，但仍有优化空间，我们将在进行特征重要性分析后，进行超参数调优。

八、特征重要性分析

特征重要性分析，是指探究特征对目标变量的影响程度，也即找到其在模型中的重要性程度。我们将从单个特征与特征交互等方面，探究变量对模型的预测与解释效用大小，从而得到影响房价的重要因素。

（一）特征重要性展示

在进行了前述的模型搭建与训练后，按降序查看各个特征的重要性，并将前 20 个重要特征绘制成柱状图，如图 23 和图 24：

	Feature Id	Importances
0	OverallQual	15.961307
1	GrLivArea	7.052773
2	HighQualSF	6.864981
3	Total_Home_Quality	6.860706
4	TotalBsmtSF	5.270604
...	...	...
349	SaleType_CWD	0.000000
350	SaleType_Con	0.000000
351	SaleType_ConLI	0.000000
352	SaleType_VWD	0.000000
353	SaleType_WD	0.000000

354 rows × 2 columns

图 23 特征重要性表格

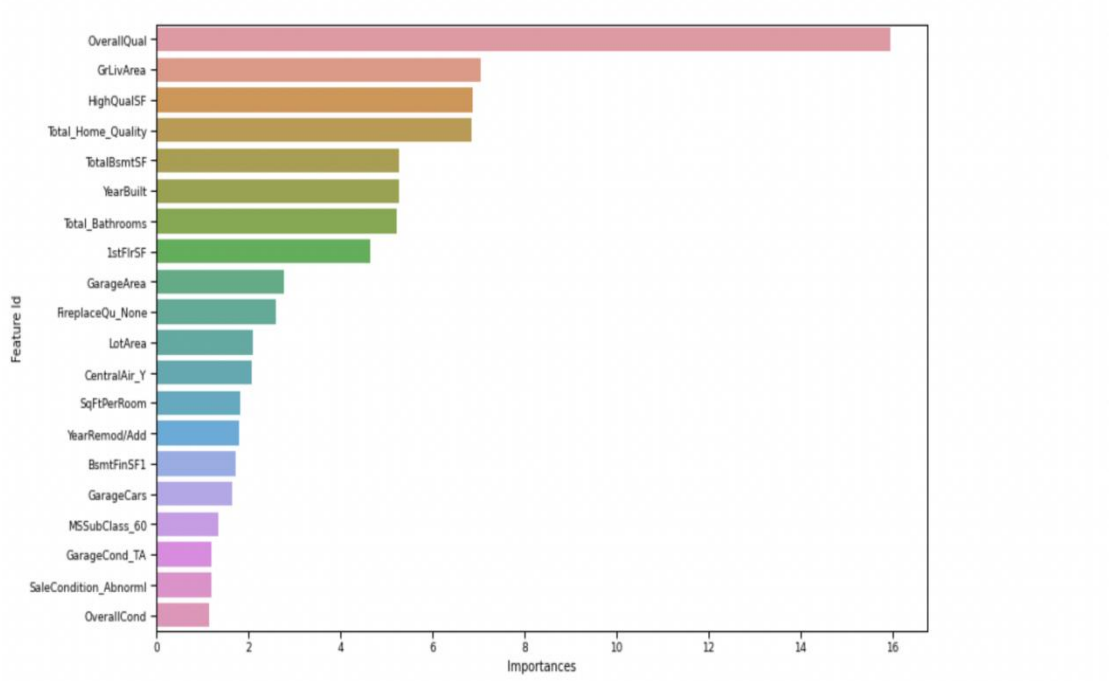

图 24 特征重要性柱状图

（二）多样本预测解释贡献

如要查看多个样本数据的预测值解释，我们可以通过绘制交互图（froce plot）的方式，查看各个原始数据样本中，每个样本值中各个特征变量的贡献与各个变量

交互（即一起作用）时带来的效果，相关方法为：将之前建立的回归模型放进解释器中，并用对象池来计算相应的贡献值。

需要说明的是，交互图所反映的重要性特征其实不如绘制简单图形进行排列直观易懂，但相较之下则能更为细致地反映各个特征在样本中的贡献程度。因此在得到特征重要性后，我们仅取样本中的前 200 条数据进行交互图的绘制，在之前的基础上做一个简单的的验证。

1. 基于房价样本的交互图

以样本数据为交互图的主体，可以直观的展现一个样本中有哪些重要特征起到了作用，贡献程度为多少。

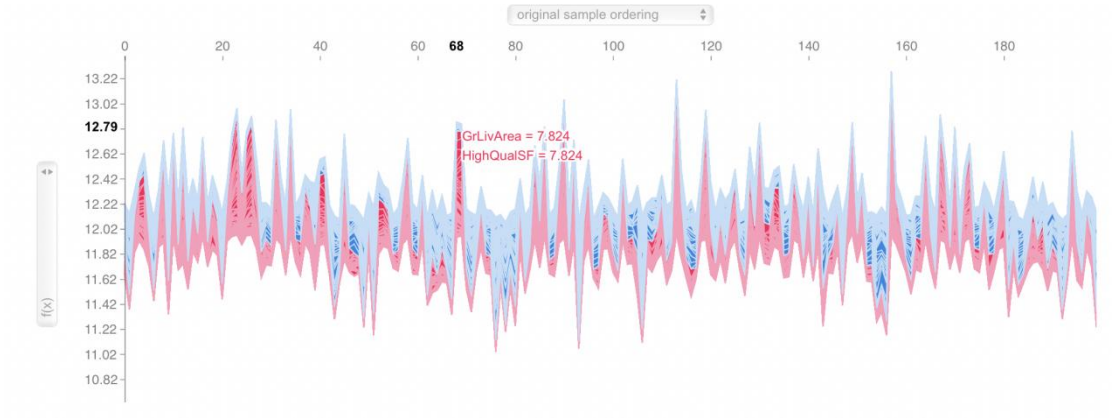

图 25 按样本顺序排列的交互图

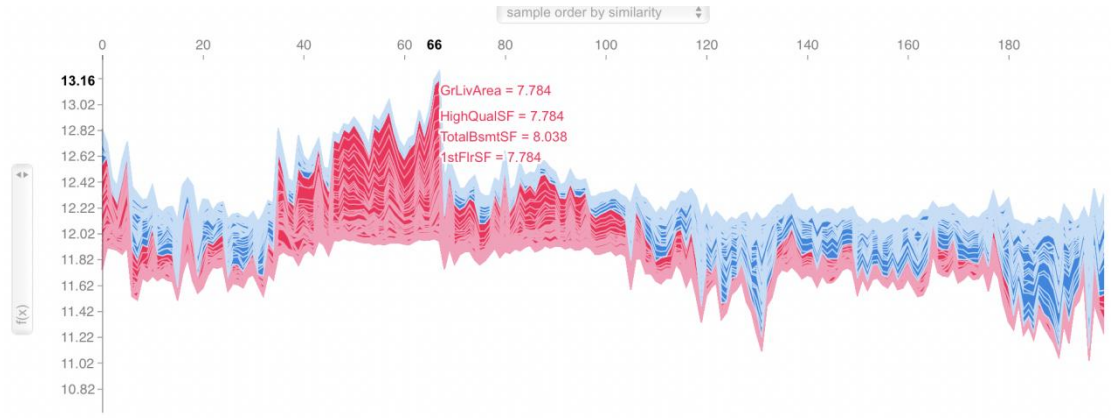

图 26 按样本相似程度排列的交互图

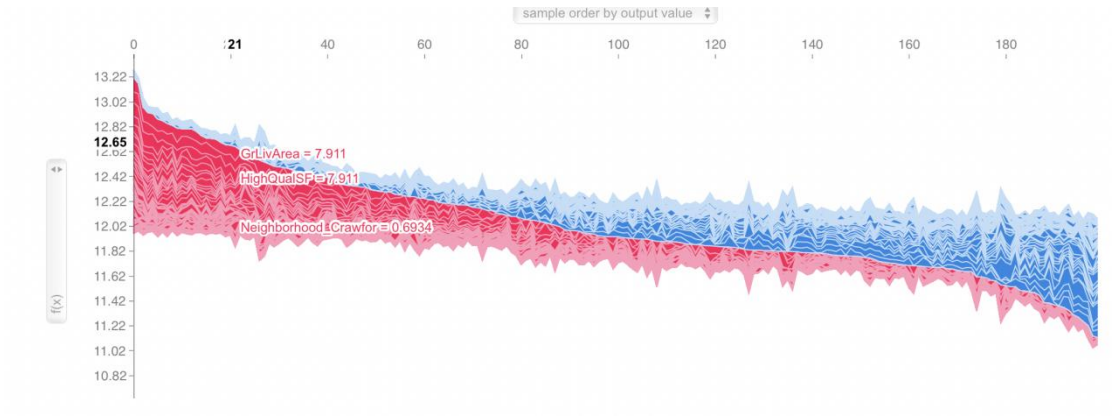

图 27 按样本输出值排列的交互图

2. 基于房屋特征的交互图

以特征为交互图的主体，可以更直观地展现某一特征对模型的影响程度。对于较为重要的特征，在交互图上表现为较高的“峰值”；而对于重要性不高的特征，其峰值就较低。更极端的，当特征重要性为 0，在交互图中则不会有输出结果，如图 28、29、30 所示。

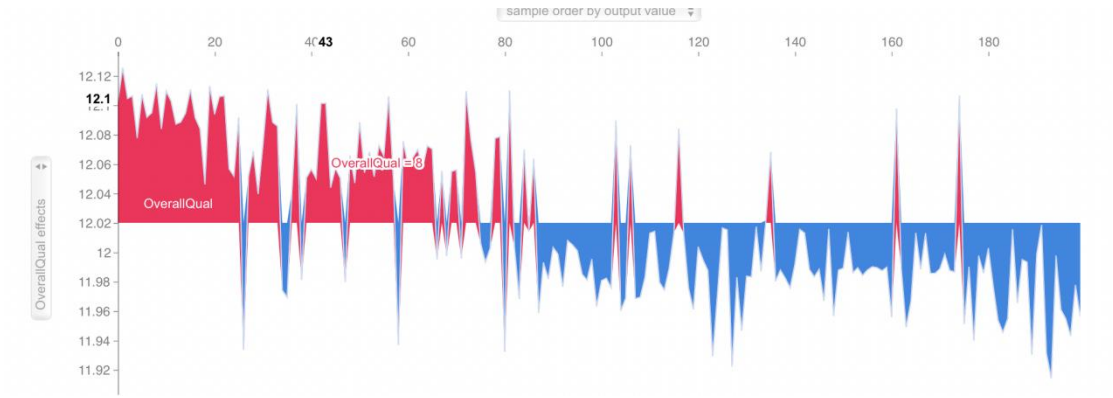

图 28 重要特征的交互图 （以 OverallQual 为例）

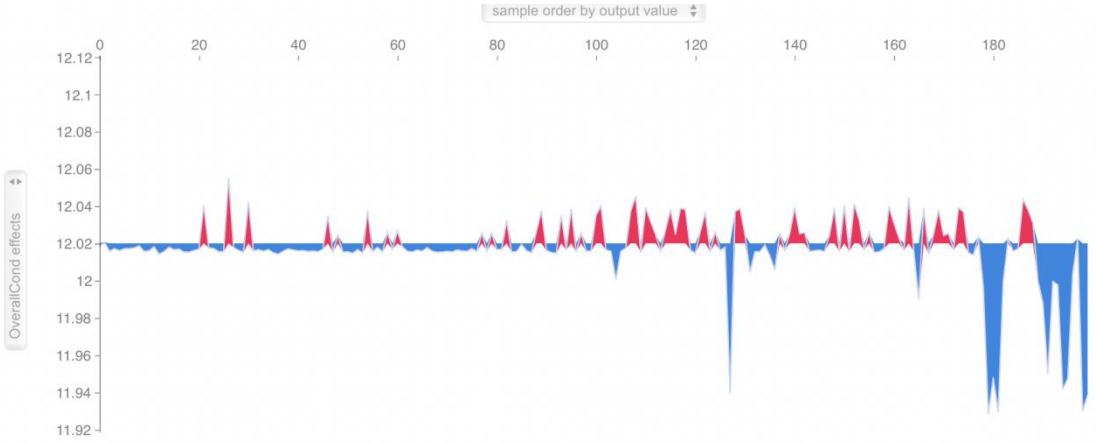

图 29 非重要特征的交互图 （以 OverallCond 为例）

图 30 重要性为 0 特征的交互图 （以 SaleType_WD 为例）

（三）特征对样本正负值预测的贡献探究

绘制概括图（summary plot）以查看全部样本特征的 shaple 的求和值，从而反映出特征重要性与每个特征对样本正负值预测的贡献，如图 31：

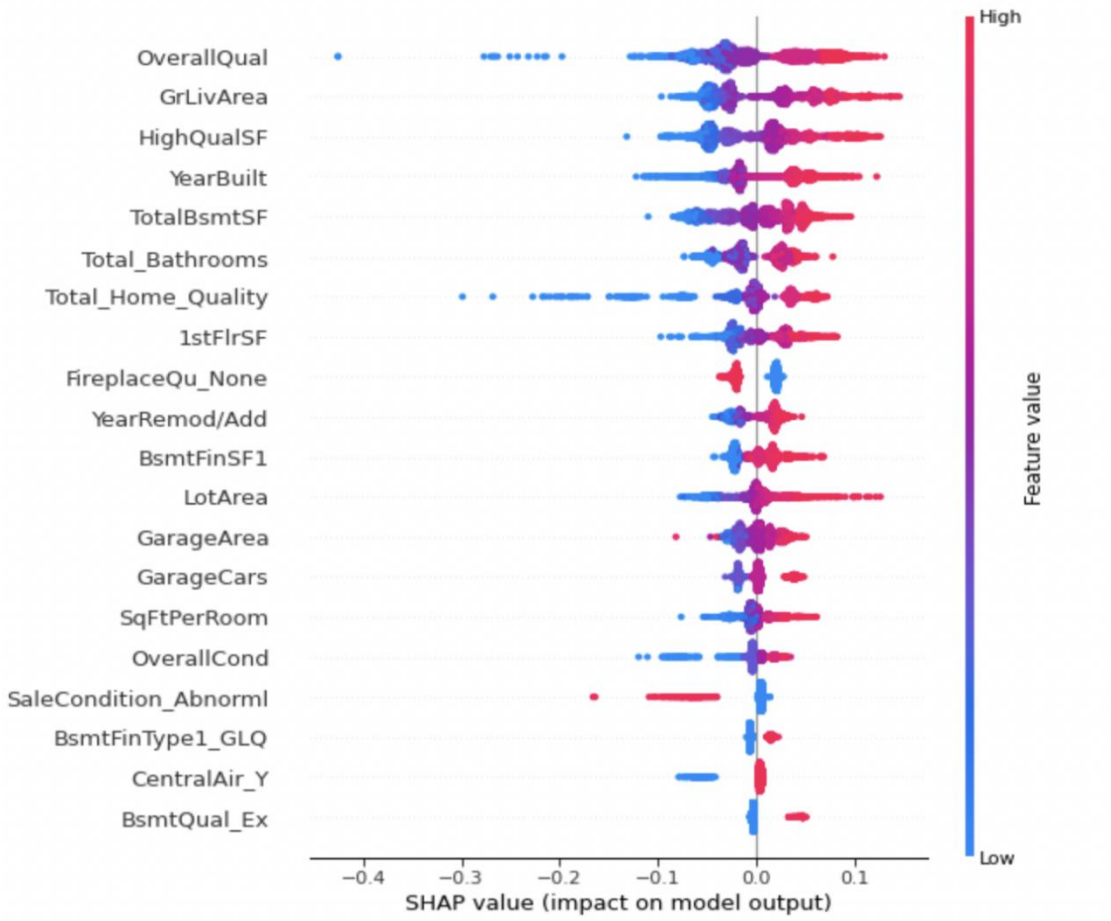

图 31 特征预测贡献概括图

上图表示所呈现特征(y 轴)的每个观测值(x 轴)。每个点在 x 轴上的 x 位置反映了该特征对模型预测的影响，而点的颜色表示该特征对该精确观测的值。在直线上堆积的点表示密度。在这里，我们可以看到，与柱状图中反映的结果一致，仍是 OverallQual 等特征对模型预测输出有较高贡献。而“BsmtFinType1_GLQ”或“BsmtQual_Ex”等特征在产生最终预测方面没有显著贡献。

（四）特征交互探究

查看了单个特征对模型的贡献与多个特征交互对预测值的影响，我们不禁思考，除了探究单个特征对模型预测的影响外，将特征进行两两组合，对模型带来的影响是否可观呢?

Catboost 附带了一个很好的方法:get_feature_importance。该方法可用于发现特征之间的重要交互。这是一个巨大的优势，因为它可以让我们了解可能创建的新功能，从而提高性能。

我们采用 catboost 的 get_feature_importance 方法来查看如有两个变量交互，其

对模型会产生怎样的影响，并通过降序排列，展示前二十条重要性较高的重要特征交互，输出结果如表 3。

表 3 交互特征的重要性

	feature1	feature2	importance
0	OverallQual	GarageArea	0.671071
1	OverallQual	GrLivArea	0.482861
2	Total_Home_Quality	HighQualSF	0.425536
3	TotalBsmtSF	HighQualSF	0.422249
4	Total_Home_Quality	Total_Bathrooms	0.371956
5	LotArea	Total_Home_Quality	0.346599
6	LotArea	YearBuilt	0.344178
7	OverallQual	BsmtFinSF1	0.342282
8	TotalBsmtSF	Neighborhood_Edwards	0.341279
9	LotArea	OverallQual	0.329355
10	KitchenAbvGr	Total_Home_Quality	0.323395
11	OverallQual	YearBuilt	0.322466
12	YearBuilt	SqFtPerRoom	0.320906
13	YearRemod/Add	HighQualSF	0.315036
14	OverallQual	CentralAir_Y	0.309563
15	GarageCars	Total_Home_Quality	0.308153
16	OverallQual	1stFlrSF	0.301750
17	YearBuilt	2ndFlrSF	0.290301
18	OverallQual	YearRemod/Add	0.289749
19	LotArea	1stFlrSF	0.289086

房屋的各个质量指标在交互变量组中占据了半壁江山，可见质量对于房屋的重要性之大。

九、模型优化与结果预测

（一）超参数调优

超参数，在训练中一般是固定数值或者以预设规则变化，比如批大小（batch size）、学习率（learning rate）、正则化项系数（weight decay）、核函数中的 gamma 等。

我们使用网格搜索，遍历所有可能的超参数组合，找到能得到最佳性能（比如最小化泛化误差）的超参数组合，先创造参数空间，并将已经进行交叉验证完毕的最终模型 final_model 作为回归器，然后再定义随机搜索。

```
MetricVisualizer(layout=Layout(align_self='stretch', height='500px'))

bestTest = 7.312869812
bestIteration = 999

bestTest = 7.335527218
bestIteration = 999

bestTest = 0.2238880003
bestIteration = 999

bestTest = 0.1337460046
bestIteration = 5997

bestTest = 0.1493668278
bestIteration = 5550

bestTest = 0.1760261086
bestIteration = 998

bestTest = 0.2559708205
bestIteration = 734

bestTest = 0.2598507338
bestIteration = 994
```

图 32 网格搜索输出（部分）

找到最佳参数组合后，便可以搭建优化模型，进行最终的房价预测了。

（二）结果预测与保存

我们将对 train.csv 中给出的数据进行房价预测，并将房价识别编号和预测房价作为结果保存为 csv 文件，命名为 result.csv，具体见附件。

In [61]:
```
# 进行房价预测，并使用head()查看前几行数据
test_pred = cat_f.predict(test)
submission = pd.DataFrame(test_id, columns = ['Id'])
test_pred = np.expm1(test_pred)
submission['SalePrice'] = test_pred
submission.head()
```

Out[61]:

	Id	SalePrice
0	1461	119113.833599
1	1462	162308.354766
2	1463	183938.720301
3	1464	195256.950289
4	1465	187733.652959

In [62]:
```
# 保存预测结果为csv文件

submission.to_csv("/Users/yangchaoran/Desktop/result.csv", index = False, header = True)
```

图 34 预测结果保存

	A	B	C	D	E	F	G
1	Id	SalePrice					
2	1461	119113.834					
3	1462	162308.355					
4	1463	183938.72					
5	1464	195256.95					
6	1465	187733.653					
7	1466	174039.665					
8	1467	179676.777					
9	1468	168095.03					
10	1469	198816.149					
11	1470	121773.658					
12	1471	189166.587					
13	1472	98257.07					
14	1473	96879.4201					

图 34 结果文件

十、实验结论与建议

（一）实验结论

1.〝平均〞与〝稳定〞才是房屋价格的常态

在读取数据的时候，我们便已经对房屋价格的分布做了一个大致的了解：大多数的房屋价格都分布在一个价位区间中，很少出现房价特别高或者特别低的情况，且按照时间序列来看，房屋售价的中位数变化也并不明显。

房屋售价如此“平稳”，是因为房屋这一商品拥有庞大的购买者群体，即无论当下一个消费者是否有足够的资金去购置一套房屋，都会或多或少的考虑房屋购买问题，或者尝试通过贷款、分期等方式购买房屋，因此销售决策者肯定会在保证自身有一定盈利空间的情况下，针对目标客户群体制定一个较为“平均”的价格，以保证不会因为价格太高而一套都卖不出去，或价格太低而血本无归。

当然，在各个方面对房屋的硬件软件进行优化，自然能够提升房产的售卖价格，但这样做是否值得，当下广大消费者对你的“优化”是否买账，也是需要销售者再三衡量的问题。当然，存在只为富豪精英打造的豪宅府邸，或为追求极致性价比的消费者打造的简房，但那毕竟是少数。绝大多数消费者都是处在“中间的大多数”，其带来的市场也是最大的，所以大多数销售者，即便是优化房屋设施，也会控制成本，因为需要将房屋价格定在大多数消费者的平均消费能力上。

综上所述，房价也可以看作的是一段时间，一个地域内消费者平均消费水平的体现。因此，抛开炒房、租房等其他非直接购买行为不谈，当地域等其他因素固定时，“平均”与“均衡”，才是房价分布的常态。

2.〝质量〞与〝面积〞决定房屋价格

在对数据集中的特征进行重要性分析后，不难看出，“质量”（Qual 变量）与“面积”（Area 变量）是决定房屋价格。

质量因素，自然不用多说。依据常识我们也可以知道，没有人会想要住进一栋看起来破破烂烂的楼房，甚至是一栋“危楼”，所以建材与最终成品房屋的质量也就自然而然的被纳入了我们的首要考量范围内。

从分析结果来看，在以特征重要性降序排列的柱状图中，房屋整体质量 QverallQual 几乎“一骑绝尘”，以巨大的差距成为影响房价的最重要特征，且排名前五的特征变量中，质量指数（即 Quality 变量）就占了三席；在交互图中，输出的 QverallQual 影响效果图也是普遍的“高峰值”；而在交互变量重要性排序中，也几乎是清一色的“Qual 变量”，可见在保证质量这件事上，是很愿意花费心思

和钱财的。

除此之外，面积因素也是消费者购置房屋的考虑重心，因为一个合适的房屋大小，对于居住者而言就代表者无限的可能性，他们可以利用这些面积进行改造和创意设计，从而带来额外的价值。所以，在考虑周围配套设施和其他个性化要求之前，房屋的面积是否恰好合适，能否满足消费者的基本生存需求，甚至于能否满足消费者之后天马行空的“设想”，向较之下就显得更加重要。

从分析结果来看，平均居住面积 GrLivArea 是仅次于房屋整体质量的第二大重要特征，在影响效果交互图中也有着很高的峰值。除此之外，车库面积 GarageArea 和地块面积 LotArea 也榜上有名，这些结果都足以说明面积指标对房价的显著影响。

综上所述，无论是质量指标还是面积指标，其背后都反映着消费者想要“住的更舒适”，“住的更安心”的需求，而这种需求是相对基本的，也即消费者普遍需要满足的（虽然可能不同的消费者对“舒适”、“安心”等词语的定义也不尽相同，但不可否认的是大家都存在对这些词对考量）。这些需求，最直观的体现就是对房屋的质量与面积有所要求，而诸如街道、铁路之类的配套设施，更像是一个附加选项。因此，房价几乎决定于质量与面积，就也在情理之中了。

（二）管理启示与建议

1.坚持〝房住不炒〞，促进房地产市场健康稳定发展

正如分析房价整体趋势得到的结论，房价应以稳定与合理为常态，而现实是，由于不同群体对自身利益的考量，这种“常态”是很难在自然条件或是脱离管控的情况下实现的。房地产市场的健康稳定发展，需要相关部门的指引。

就我国房地产市场而言，针对某些地区房价居高不下，炒房现象泛滥而导致的房源供应紧张等问题，我国政府提出了“房住不炒”，强调房屋是用来住的而不是炒的，在稳定预期、防范风险、促进转型三个方面发力，增强政策的精确性与协调性，保障房地产供需基本平衡、结构基本合理、价格基本稳定，同经济社会与住宅产业发展相协调。

实践证明，坚持“房住不炒”，能够使房地产火热的炒房行为得以抑制，减少经济泡沫的产生；能够推动房产企业的结构优化，使其企业根据市场规律调整业务结构和开发方向，更多的关注民众的刚需；能够助力保障性住房等民生住房工程，推动房地产企业从“卖房”向“服务”转变；更重要的是，能够维护房地产市场的长期稳定，促使其更加规范健康、推进城市化进程。

综上，房价的稳定需要房地产市场这个宏观环境的稳定，而房地产市场的稳定离不开相关部门的调控与身为公民的每一个我们的自觉遵守，故坚持“房住不炒”，对促进房地产市场健康稳定发展而言，是十分必要的。

2.坚持质量导向，推动质量与房价良性交互

无论是依据实验结论，还是依据生活常识，我们都不难判断，房屋的质量是影响房价的主要因素，也是广大消费者做出买房决策的首要依据，因此，无论是政府还是房地产商，都应坚持以房屋质量为导向，营造良好的购房住房环境。

政府方面，要进行正确的方向引导与精准的政策帮扶。首先，要制定相应的质量检验标准，同时形成房屋安全长效机制，研究建立房屋体验、养老、保险等制度。其次，要赏罚分明，对于肯在房屋质量上下功夫的优质房地产企业，要给予相应的支持与补贴，鼓励其发挥带头作用；而对于不符合质量要求的房屋和存在风险的险房企，该责令整改的整改，该破产的破产，该惩罚的惩罚，做房屋高质量建设的“督导员”，把握好质量与价格之间的尺度，既不能放任不管，任由房地产企业野蛮生长，也不能一味压低价格，打压企业的生产建设积极性，做好保楼交楼工作，

房地产企业方面，要以建设高质量房地产为己任，积极承担企业的社会责任，大力提升房屋品质，以合理的价格，为人民群众建设好房子；同时，大力提升物业相关配套服务，让人民群众生活更方便，更安心。

综上，只要凝聚起推动房产高质量发展的伟大合力，将房地产工作融入党和国家事业的大棋局，锚定新时代新征程下党的使命任务和当前的中心工作，推动房屋质量与房价的良性交互，就一定能实现老百姓安心放心，企业兼顾名利，产业链稳定发展的共赢局面。

3.房屋调控，应〝因地制宜〞而非〝一刀切〞

即使通过对房价影响因素与房价整体趋势的分析，我们得出了要稳定房价，注重质量等结论，但要真正落实，上述结论也只能做为“大纲”进行一个方向的指引，房屋调控，还是应具体问题具体分析，做到因地制宜。

所谓因地制宜，指的是在围绕稳地价、稳房价、稳预期调控大目标的前提下，坚持因城施策，一城一策，夯实城市主体责任，处理好中央和地方的关系，不搞一刀切，鼓励地方政府因地制宜进行房地产调控。

每个城市都有自己特有的经济状况和产业链状况，因此只有因地制宜，实行差异化的调控措施，才能避免政策冲击对宏观经济形成干扰，更加有效地进行房屋调控，从而实现让整个市场和宏观经济更加平稳运行的目标。

参考文献

[1]吴国杰. 房价影响因素分析[J]. 现代商业, 2011(12):1.

[2]王明. 基于非参数可加模型的房价影响因素分析[D]. 浙江工商大学.

[3]张金芳. 基于因子分析法的中国房价影响因素分析[J]. 辽宁经济统计, 2012(12):3.

[4]罗玉波. 房价影响因素分析:分位数回归方法[J]. 统计与决策, 2011(6):2.

[5]王玲刘平清王梅张小敏秦裕蕾. 基于 VAR 模型对贵州省房价影响因素分析[J]. 电脑知识与技术:学术版, 2022, 18(2):122-125.

[6]何威, 薛小荣. 基于灰色关联分析的西安市房价影响因素分析[C]// International Conference on Computational Intelligence & Industrial Application. 0.

[7]刘美辰. 基于卡尔曼滤波法的房价波动影响因素分析[J]. 现代经济信息, 2017.

[8]刘爱芳, 任晓宇. 房地产价格变动的影响因素分析[J]. 商业时代, 2011, 000(032):117-118.

[9]徐春娥. 中国房价影响因素分析[J]. 合作经济与科技, 2013(4):2.

[10]续云丰, 孔亚仙, 徐仁旭. 统计模型和博弈论视角下的杭州房价影响因素分析[J]. 长沙大学学报, 2014, 28(2):3.

[11] Sheng J , Pan D . factor analysis of housing price based on boosting regression tree -taking boston as an example[J]. 2018.

[12]储亚伟, 黄贤峰, 郑语欣. 房价影响因素的研究及预测——以阜阳市为例[J]. 山东省农业管理干部学院学报, 2019, 036(001):63-67.

[13]崔恒建,成平.PP 偏度,峰度正态性检验的 P—VALUES[J].自然科学进展: 国家重点实验室通讯, 1995, 5(6):7.DOI:CNKI:SUN:ZKJZ.0.1995-06-007.

[14]崔恒建,陈广雷.带有误差变量的偏度和峰度正态性检验[J].北京师范大学学报(自然科学版), 2000.DOI:CNKI:SUN:BSDZ.0.2000-01-001.

[15]孙玉芝, 李春禄.介绍两种正态性检验方法[J].天津师大学报: 自然科学版, 1992, 000(001):30-34.

论文附录 A

数据描述

MSSubClass: Identifies the type of dwelling involved in the sale.

20	1-STORY 1946 & NEWER ALL STYLES
30	1-STORY 1945 & OLDER
40	1-STORY W/FINISHED ATTIC ALL AGES
45	1-1/2 STORY - UNFINISHED ALL AGES
50	1-1/2 STORY FINISHED ALL AGES
60	2-STORY 1946 & NEWER
70	2-STORY 1945 & OLDER
75	2-1/2 STORY ALL AGES
80	SPLIT OR MULTI-LEVEL
85	SPLIT FOYER
90	DUPLEX - ALL STYLES AND AGES
120	1-STORY PUD (Planned Unit Development) - 1946 & NEWER
150	1-1/2 STORY PUD - ALL AGES
160	2-STORY PUD - 1946 & NEWER
180	PUD - MULTILEVEL - INCL SPLIT LEV/FOYER
190	2 FAMILY CONVERSION - ALL STYLES AND AGES

MSZoning: Identifies the general zoning classification of the sale.

A	Agriculture
C	Commercial
FV	Floating Village Residential
I	Industrial
RH	Residential High Density
RL	Residential Low Density
RP	Residential Low Density Park
RM	Residential Medium Density

LotFrontage: Linear feet of street connected to property

LotArea: Lot size in square feet

Street: Type of road access to property

Grvl	Gravel
Pave	Paved

Alley: Type of alley access to property

Grvl	Gravel
Pave	Paved
NA	No alley access

LotShape: General shape of property

Reg Regular
IR1 Slightly irregular
IR2 Moderately Irregular
IR3 Irregular

LandContour: Flatness of the property

Lvl Near Flat/Level
Bnk Banked - Quick and significant rise from street grade to building
HLS Hillside - Significant slope from side to side
Low Depression

Utilities: Type of utilities available

AllPub All public Utilities (E,G,W,& S)
NoSewr Electricity, Gas, and Water (Septic Tank)
NoSeWa Electricity and Gas Only
ELO Electricity only

LotConfig: Lot configuration

Inside Inside lot
Corner Corner lot
CulDSac Cul-de-sac
FR2 Frontage on 2 sides of property
FR3 Frontage on 3 sides of property

LandSlope: Slope of property

Gtl Gentle slope
Mod Moderate Slope
Sev Severe Slope

Neighborhood: Physical locations within Ames city limits

Blmngtn Bloomington Heights
Blueste Bluestem
BrDale Briardale
BrkSide Brookside
ClearCr Clear Creek
CollgCr College Creek
Crawfor Crawford
Edwards Edwards
Gilbert Gilbert
IDOTRR Iowa DOT and Rail Road
MeadowV Meadow Village
Mitchel Mitchell
Names North Ames

NoRidge Northridge
NPkVill Northpark Villa
NridgHt Northridge Heights
NWAmes Northwest Ames
OldTown Old Town
SWISU South & West of Iowa State University
Sawyer Sawyer
SawyerW Sawyer West
Somerst Somerset
StoneBr Stone Brook
TimberTimberland
Veenker Veenker

Condition1: Proximity to various conditions

Artery Adjacent to arterial street
Feedr Adjacent to feeder street
Norm Normal
RRNn Within 200' of North-South Railroad
RRAn Adjacent to North-South Railroad
PosN Near positive off-site feature--park, greenbelt, etc.
PosA Adjacent to postive off-site feature
RRNe Within 200' of East-West Railroad
RRAe Adjacent to East-West Railroad

Condition2: Proximity to various conditions (if more than one is present)

Artery Adjacent to arterial street
Feedr Adjacent to feeder street
Norm Normal
RRNn Within 200' of North-South Railroad
RRAn Adjacent to North-South Railroad
PosN Near positive off-site feature--park, greenbelt, etc.
PosA Adjacent to postive off-site feature
RRNe Within 200' of East-West Railroad
RRAe Adjacent to East-West Railroad

BldgType: Type of dwelling

1Fam Single-family Detached
2FmCon Two-family Conversion; originally built as one-family dwelling
Duplx Duplex
TwnhsE Townhouse End Unit
TwnhsI Townhouse Inside Unit

HouseStyle: Style of dwelling

1Story One story
1.5Fin One and one-half story: 2nd level finished
1.5Unf One and one-half story: 2nd level unfinished
2Story Two story

2.5Fin Two and one-half story: 2nd level finished
2.5Unf Two and one-half story: 2nd level unfinished
SFoyer Split Foyer
SLvl Split Level

OverallQual: Rates the overall material and finish of the house

10 Very Excellent
9 Excellent
8 Very Good
7 Good
6 Above Average
5 Average
4 Below Average
3 Fair
2 Poor
1 Very Poor

OverallCond: Rates the overall condition of the house

10 Very Excellent
9 Excellent
8 Very Good
7 Good
6 Above Average
5 Average
4 Below Average
3 Fair
2 Poor
1 Very Poor

YearBuilt: Original construction date

YearRemodAdd: Remodel date (same as construction date if no remodeling or additions)

RoofStyle: Type of roof

Flat Flat
Gable Gable
Gambrel Gabrel (Barn)
Hip Hip
Mansard Mansard
Shed Shed

RoofMatl: Roof material

ClyTile Clay or Tile
CompShg Standard (Composite) Shingle
Membran Membrane
Metal Metal

Roll Roll
Tar&Grv Gravel & Tar
WdShake Wood Shakes
WdShngl Wood Shingles

Exterior1st: Exterior covering on house

AsbShng Asbestos Shingles
AsphShn Asphalt Shingles
BrkComm Brick Common
BrkFace Brick Face
CBlock Cinder Block
CemntBd Cement Board
HdBoard Hard Board
ImStucc Imitation Stucco
MetalSd Metal Siding
Other Other
Plywood Plywood
PreCast PreCast
Stone Stone
Stucco Stucco
VinylSd Vinyl Siding
Wd Sdng Wood Siding
WdShing Wood Shingles

Exterior2nd: Exterior covering on house (if more than one material)

AsbShng Asbestos Shingles
AsphShn Asphalt Shingles
BrkComm Brick Common
BrkFace Brick Face
CBlock Cinder Block
CemntBd Cement Board
HdBoard Hard Board
ImStucc Imitation Stucco
MetalSd Metal Siding
Other Other
Plywood Plywood
PreCast PreCast
Stone Stone
Stucco Stucco
VinylSd Vinyl Siding
Wd Sdng Wood Siding
WdShing Wood Shingles

MasVnrType: Masonry veneer type

BrkCmn Brick Common
BrkFace Brick Face
CBlock Cinder Block
None None

Stone Stone

MasVnrArea: Masonry veneer area in square feet

ExterQual: Evaluates the quality of the material on the exterior

Ex Excellent
Gd Good
TA Average/Typical
Fa Fair
Po Poor

ExterCond: Evaluates the present condition of the material on the exterior

Ex Excellent
Gd Good
TA Average/Typical
Fa Fair
Po Poor

Foundation: Type of foundation

BrkTil Brick & Tile
CBlock Cinder Block
PConc Poured Contrete
Slab Slab
Stone Stone
Wood Wood

BsmtQual: Evaluates the height of the basement

Ex Excellent (100+ inches)
Gd Good (90-99 inches)
TA Typical (80-89 inches)
Fa Fair (70-79 inches)
Po Poor (<70 inches
NA No Basement

BsmtCond: Evaluates the general condition of the basement

Ex Excellent
Gd Good
TA Typical - slight dampness allowed
Fa Fair - dampness or some cracking or settling
Po Poor - Severe cracking, settling, or wetness
NA No Basement

BsmtExposure: Refers to walkout or garden level walls

Gd Good Exposure
Av Average Exposure (split levels or foyers typically score average or above)

MnMimimum Exposure
No No Exposure
NA No Basement

BsmtFinType1: Rating of basement finished area

GLQ Good Living Quarters
ALQ Average Living Quarters
BLQ Below Average Living Quarters
Rec Average Rec Room
LwQ Low Quality
Unf Unfinshed
NA No Basement

BsmtFinSF1: Type 1 finished square feet

BsmtFinType2: Rating of basement finished area (if multiple types)

GLQ Good Living Quarters
ALQ Average Living Quarters
BLQ Below Average Living Quarters
Rec Average Rec Room
LwQ Low Quality
Unf Unfinshed
NA No Basement

BsmtFinSF2: Type 2 finished square feet

BsmtUnfSF: Unfinished square feet of basement area

TotalBsmtSF: Total square feet of basement area

Heating: Type of heating

Floor Floor Furnace
GasA Gas forced warm air furnace
GasW Gas hot water or steam heat
Grav Gravity furnace
OthW Hot water or steam heat other than gas
Wall Wall furnace

HeatingQC: Heating quality and condition

Ex Excellent
Gd Good
TA Average/Typical
Fa Fair
Po Poor

CentralAir: Central air conditioning

N No
Y Yes

Electrical: Electrical system

SBrkr Standard Circuit Breakers & Romex
FuseA Fuse Box over 60 AMP and all Romex wiring (Average)
FuseF 60 AMP Fuse Box and mostly Romex wiring (Fair)
FuseP 60 AMP Fuse Box and mostly knob & tube wiring (poor)
Mix Mixed

1stFlrSF: First Floor square feet

2ndFlrSF: Second floor square feet

LowQualFinSF: Low quality finished square feet (all floors)

GrLivArea: Above grade (ground) living area square feet

BsmtFullBath: Basement full bathrooms

BsmtHalfBath: Basement half bathrooms

FullBath: Full bathrooms above grade

HalfBath: Half baths above grade

Bedroom: Bedrooms above grade (does NOT include basement bedrooms)

Kitchen: Kitchens above grade

KitchenQual: Kitchen quality

Ex Excellent
Gd Good
TA Typical/Average
Fa Fair
Po Poor

TotRmsAbvGrd: Total rooms above grade (does not include bathrooms)

Functional: Home functionality (Assume typical unless deductions are warranted)

Typ Typical Functionality
Min1 Minor Deductions 1
Min2 Minor Deductions 2
Mod Moderate Deductions
Maj1 Major Deductions 1
Maj2 Major Deductions 2
Sev Severely Damaged

Sal Salvage only

Fireplaces: Number of fireplaces

FireplaceQu: Fireplace quality

Ex Excellent - Exceptional Masonry Fireplace
Gd Good - Masonry Fireplace in main level
TA Average - Prefabricated Fireplace in main living area or Masonry Fireplace in basement
Fa Fair - Prefabricated Fireplace in basement
Po Poor - Ben Franklin Stove
NA No Fireplace

GarageType: Garage location

2Types More than one type of garage
Attchd Attached to home
Basment Basement Garage
BuiltIn Built-In (Garage part of house - typically has room above garage)
CarPort Car Port
Detchd Detached from home
NA No Garage

GarageYrBlt: Year garage was built

GarageFinish: Interior finish of the garage

Fin Finished
RFn Rough Finished
Unf Unfinished
NA No Garage

GarageCars: Size of garage in car capacity

GarageArea: Size of garage in square feet

GarageQual: Garage quality

Ex Excellent
Gd Good
TA Typical/Average
Fa Fair
Po Poor
NA No Garage

GarageCond: Garage condition

Ex Excellent
Gd Good
TA Typical/Average

Fa Fair
Po Poor
NA No Garage

PavedDrive: Paved driveway

Y Paved
P Partial Pavement
N Dirt/Gravel

WoodDeckSF: Wood deck area in square feet

OpenPorchSF: Open porch area in square feet

EnclosedPorch: Enclosed porch area in square feet

3SsnPorch: Three season porch area in square feet

ScreenPorch: Screen porch area in square feet

PoolArea: Pool area in square feet

PoolQC: Pool quality

Ex Excellent
Gd Good
TA Average/Typical
Fa Fair
NA No Pool

Fence: Fence quality

GdPrv Good Privacy
MnPrv Minimum Privacy
GdWo Good Wood
MnWw Minimum Wood/Wire
NA No Fence

MiscFeature: Miscellaneous feature not covered in other categories

Elev Elevator
Gar2 2nd Garage (if not described in garage section)
Othr Other
Shed Shed (over 100 SF)
TenC Tennis Court
NA None

MiscVal: $Value of miscellaneous feature

MoSold: Month Sold (MM)

YrSold: Year Sold (YYYY)

SaleType: Type of sale

WD Warranty Deed - Conventional
CWD Warranty Deed - Cash
VWD Warranty Deed - VA Loan
New Home just constructed and sold
COD Court Officer Deed/Estate
Con Contract 15% Down payment regular terms
ConLw Contract Low Down payment and low interest
ConLI Contract Low Interest
ConLD Contract Low Down
Oth Other

SaleCondition: Condition of sale

Normal Normal Sale
Abnorml Abnormal Sale - trade, foreclosure, short sale
AdjLand Adjoining Land Purchase
Alloca Allocation - two linked properties with separate deeds, typically condo with a garage unit
Family Sale between family members
Partial Home was not completed when last assessed (associated with New Homes)