# 基于Python实现xgboost回归模型(XGBRegressor)项目实战

## 1.项目背景

随着大数据时代的到来，具备大数据思想至关重要，人工智能技术在各行各业的应用已是随处可见。在生产制造业，人工智能技术可以极大地提高生产效率，节省劳动成本，提升产品质量；在服务业，可以优化行业现有产品和服务，提升其质量和劳动生产率；金融、医疗等领域，也因人工智能技术的加入而愈发繁荣，人们的生活也因为其更加便利。

房屋作为每个公民的必需品，在生活中的地位非常重要，买房已成为人们谈论较多的话题，如何在合适的时间买房卖房也成为了人们关注焦点，因此在这样的背景下，产生了本次的房价预测相关问题。目前在房价预测领域主要体现在两个问题上：一是选择合适的数学模型来预测房价走向，用以评估房价的变化；二是寻找引起房价变化的原因，国家可借此来帮助市场协调房价变化，公民可以根据时事来判断入手时机。本项目主要分析第一个问题，即选择合适的数学模型来帮助预测房价。

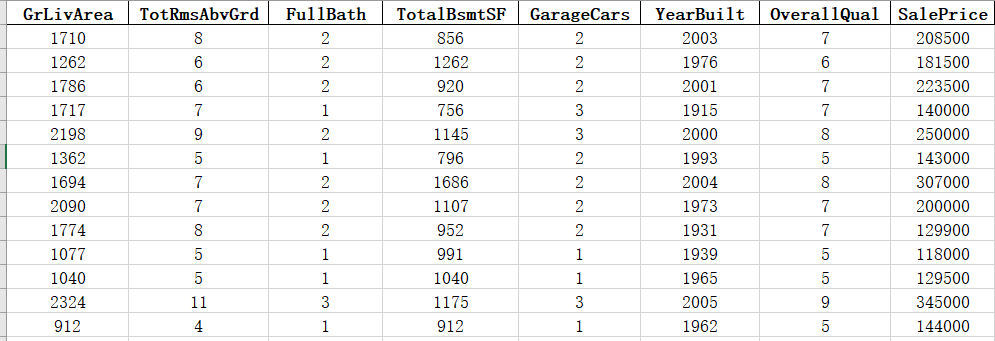
本项目将从某地区的房价数据为着手点，以该地区的房屋的相关属性来作为特征，筛选重要信息，并且将一些信息做适当处理，最终用以预测该地区的其他房屋价格。

## 2.数据获取

本次建模数据来源于网络(本项目撰写人整理而成)，数据项统计如下：

|  |  |  |
| --- | --- | --- |
| **字段名** | **释义** | **备注** |
| GrLivArea | 生活面积 |  |
| TotRmsAbvGrd | 总房间数 |  |
| FullBath | 浴室数量 |  |
| TotalBsmtSF | 地下室总面积 |  |
| GarageCars | 车库 |  |
| YearBuilt | 建造年份 |  |
| OverallQual | 总体评价 |  |
| SalePrice | 房屋价格 | 标签 |

数据详情如下：

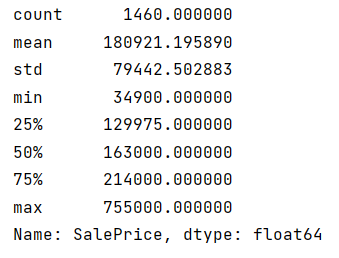


## 3.探索性数据分析

### 3.1房屋价格描述性统计

关键代码：



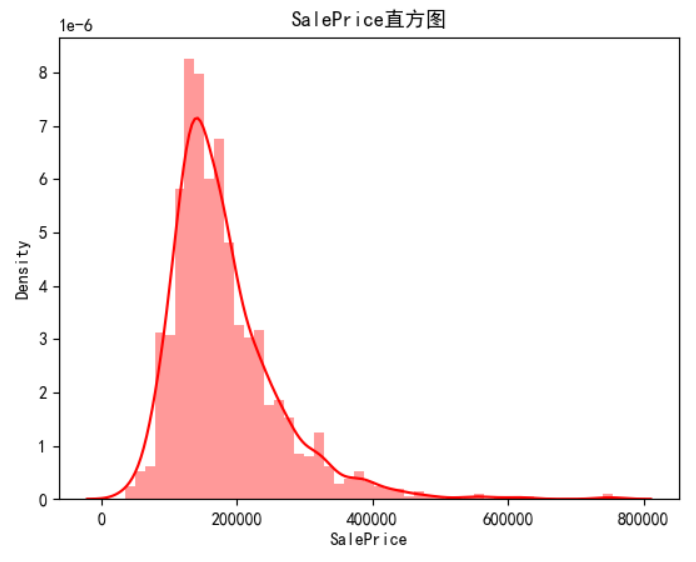


从上图可以看出，SalePrice 总数据量1460，平均值为180921.1959、标准差为79442.5029以及最小值、最大值和分位数。

### 3.2房屋价格直方图

关键代码：





由上图可以知道房价呈现正态分布。

### 3.3房屋价格偏度峰度

关键代码：



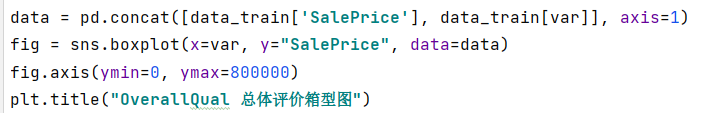
结果如下表所示：

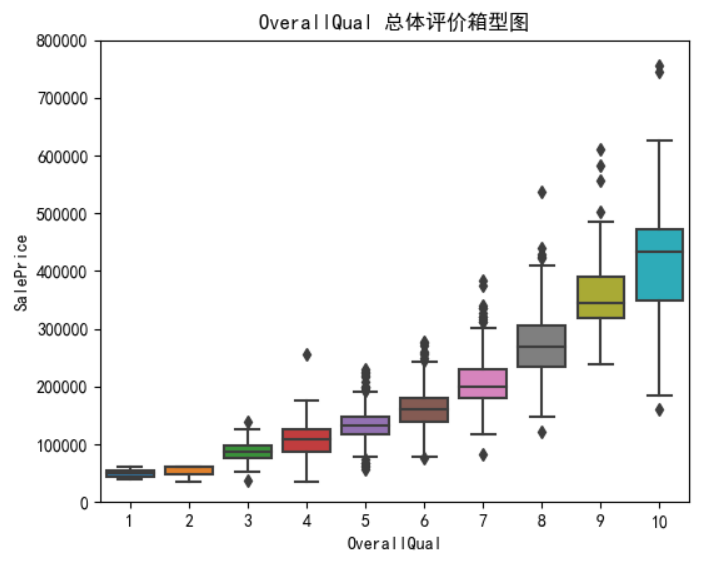
|  |  |
| --- | --- |
| 偏度 | 1.882876 |
| 峰度 | 6.536282 |

从上表可以看到，偏度大于0 说明比正态分布的高峰更加陡峭——尖顶峰；峰度大于0 说明正偏差数值较大，为正偏或右偏。长尾巴拖在右边。那么，结合房屋价格直方图可以确定长尾巴确实拖在右边，而且高峰陡峭。

### 3.4 OverallQual 总体评价箱型图

关键代码：

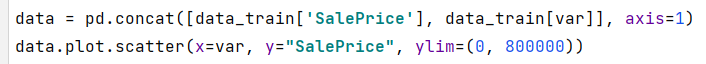


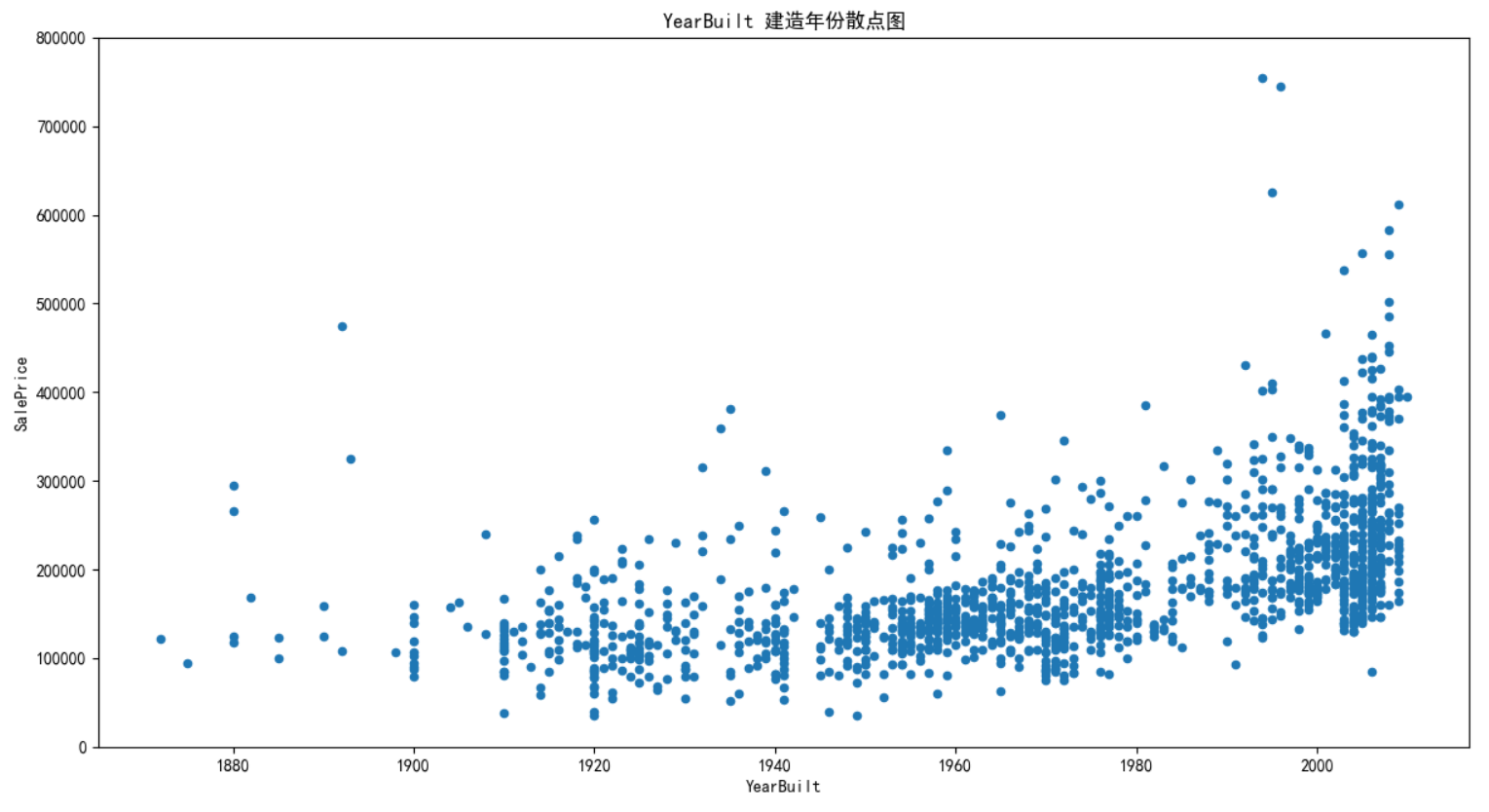


从上图可以看出，整体评价越高 房屋价格越高。

### 3.5 YearBuilt 建造年份散点图

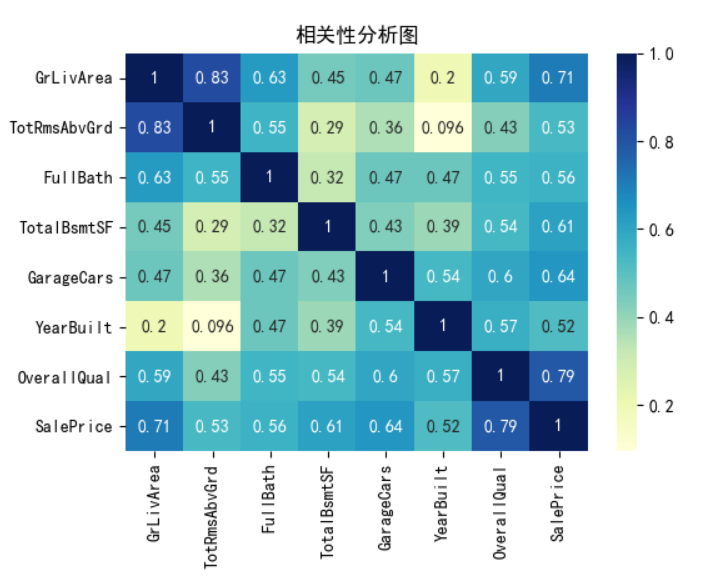
关键代码：





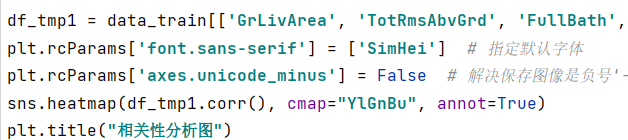
从上图可以看到，房屋价格随建造年份成线性增长。

### 3.6相关性分析



从上面相关性热力图可以看出，房屋价格与这些特征相关性都是正相关，而且相关性比较强。

关键代码：

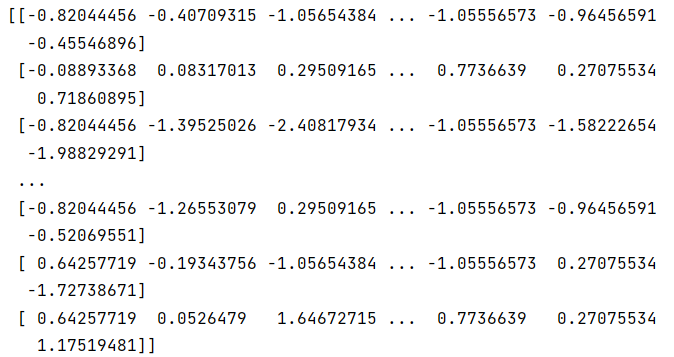


## 4.数据预处理

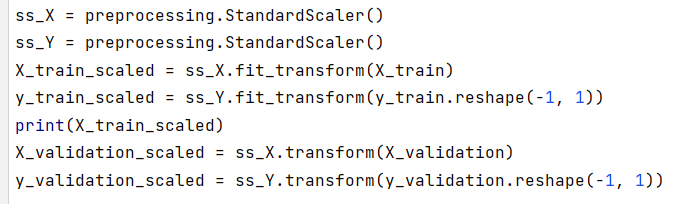
真实数据中可能包含了大量的缺失值和噪音数据或人工录入错误导致有异常点存在，非常不利于算法模型的训练。数据清洗的结果是对各种脏数据进行对应方式的处理，得到标准的、干净的、连续的数据，提供给数据统计、数据挖掘等使用。数据预处理通常包含数据清洗、归约、聚合、转换、抽样等方式，数据预处理质量决定了后续数据分析挖掘及建模工作的精度和泛化价值。以下简要介绍数据预处理工作中主要的预处理方法：

### 4.1数据标准化

标准化后的数据如下：



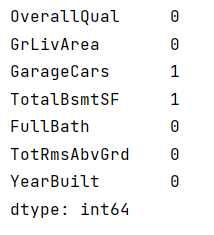
关键代码：



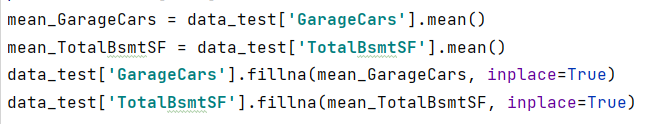
### 4.2测试集数据缺失值处理

关键代码：





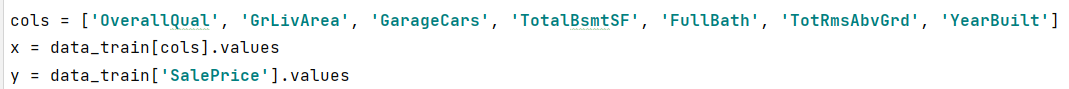
从上图可以看出，测试数据集GarageCars、TotalBsmtSF存在缺失值；这里针对缺失值进行均值填充：



## 5.特征工程

### 5.1 建立特征数据和标签数据

SalePrice 为标签数据，除 SalePrice 之外的为特征数据。关键代码如下：



### 5.2数据集拆分

训练集拆分，分为训练集和验证集，80%训练集和20%验证集。关键代码如下：



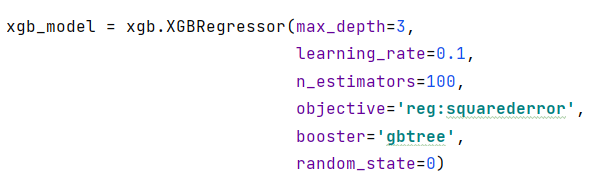
## 6.构建xgboost回归模型

根据数据中“生活面积”、“总房间数”、“浴室数量”、“地下室总面积”、“车库”、“建造年份”、“总体评价”7种变量的特征，预测“房屋价格”。使用XGBRegressor算法，用于目标回归。

### 6.1模型参数

|  |  |
| --- | --- |
| **编号** | **参数** |
| 1 | max\_depth=3 |
| 2 | learning\_rate=0.1 |
| 3 | n\_estimators=100 |
| 4 | objective='reg:squarederror' |
| 5 | booster='gbtree' |
| 6 | random\_state=0 |

关键代码如下：



## 7.模型评估

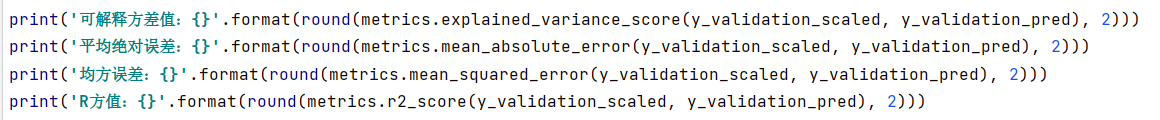
### 7.1评估指标及结果

评估指标主要包括可解释方差、平均绝对误差、均方误差、R方等等。

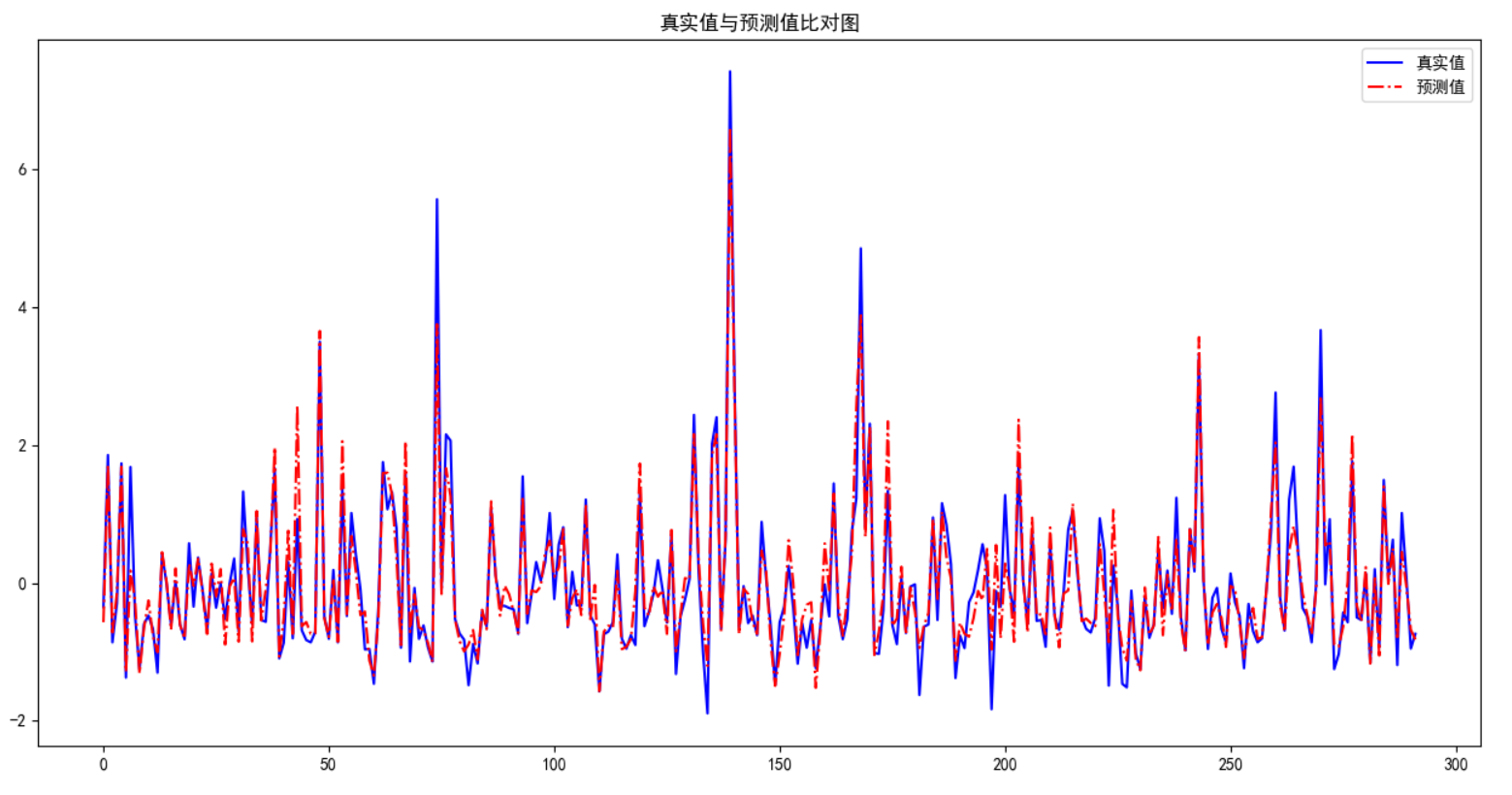
|  |  |  |
| --- | --- | --- |
| **模型名称** | **指标名称** | **指标值** |
| 验证集 | | |
| 支持向量回归模型 | 可解释方差 | 0.82 |
| 平均绝对误差 | 0.26 |
| 均方误差 | 0.24 |
| R方 | 0.82 |
| xgboost回归模型 | 可解释方差 | 0.90 |
| 平均绝对误差 | 0.24 |
| 均方误差 | 0.13 |
| R方 | 0.9 |

从上表可以看出，模型拟合效果较好，R方为0.9，非常接近1.

关键代码如下：

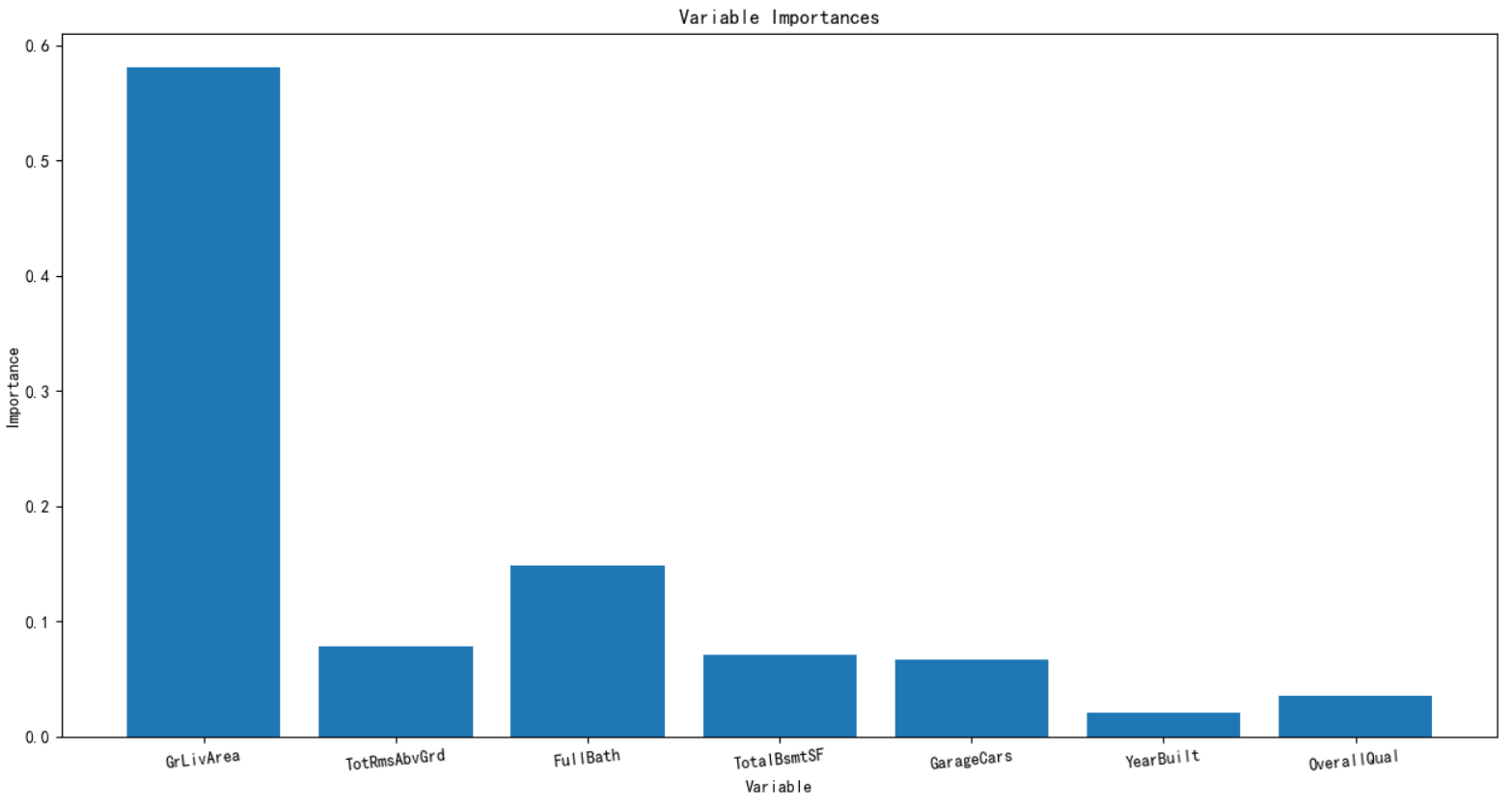


### 7.2真实值与预测值对比图



从上图可以看出，预测值和真实值比较一致，说明模型拟合效果较好。

### 7.2模型特征重要性



从上图可以看到，特征重要性依次为：GrLivArea、FullBath、TotRmsAbvGrd等等。

## 8.结论与建议

综上所述，本文采用了xgboost模型，最终证明了我们提出的模型效果良好。通过本次房屋预测模型的建立，能够得到影响房屋总价值的不仅仅是房屋居住总面积这一个重量级特征，还有总房间数、浴室数量等其他影响特征。数据帮助我们筛选出重要的特征，排除了一些“想当然”的结果，能够更好地抓住事情的本质，这就是大数据辅助我们解决生活中的事情的实际例子。。

预测结果数据如下：

