关于房价的信息分析与预测

[摘要] 近年来房地产行业的快速发展带动了我国经济的高速发展，但也面临着房价泡沫的问题。随着国民经济水平的提高，人们对于生活质量和生活环境的要求不断提高。住房成为人们关注的重点，交通是否便利、是否为学区房、周围环境质量和气候情况等问题成为购房者考虑的主要因素，而房价则是综合这个影响因素的一个主要体现分析房地产市场的走势，预测房价的变化规律对保持房地产市场稳定、促进房地产业健康发展有重大意义。本文将使用python语言对该房价相关数据进行数据分析与预测，并将其可视化。

[关键词] 房价预测； 信息分析

目录

[第1章 引言 1](#_Toc1646)

[1.1背景介绍与研究意义 1](#_Toc30653)

[1.2研究内容与目标 1](#_Toc11911)

[第2章 数据分析 2](#_Toc13560)

[2.1分析数据 2](#_Toc105)

[（1） 描述性数据总结 2](#_Toc9270)

[（2） 数据类型统计 2](#_Toc32546)

[（3） 绘制直方图 3](#_Toc20817)

[（4） 数据偏度和峰度度量 3](#_Toc20511)

[（5） 与数字型变量的关系 4](#_Toc22806)

[（6） 与类别型变量的关系 5](#_Toc15808)

[（7） 相关系数矩阵 7](#_Toc8708)

[（8） SalePrice相关系数矩阵 7](#_Toc19362)

[（9） SalePrice和相关变量之间的散点图 8](#_Toc25591)

[第3章 数据预处理 9](#_Toc21023)

[3.1数据降维 9](#_Toc15908)

[（1） PCA算法简介 9](#_Toc716)

[（2） PCA算法流程 9](#_Toc22043)

[（3） PCA算法总结 10](#_Toc1545)

[3.2 线性回归模型 11](#_Toc4003)

[3.3 算法 11](#_Toc8660)

[(1) SVM 11](#_Toc11439)

[(2) KNN 12](#_Toc7302)

[(3) 决策树 14](#_Toc23995)

[(4) 随机森林 15](#_Toc25120)

[3.4 结果可视化 17](#_Toc20049)

[结 论 18](#_Toc26960)

[参考文献 19](#_Toc22157)

# 第1章 引言

## 1.1背景介绍与研究意义

随着国民经济水平的提高，人们对于生活质量和生活环境的要求不断提高。住房成为人们关注的重点，而房价则是综合许多影响因素的一个主要体现分析房地产市场的走势，预测房价的变化规律对保持房地产市场稳定、促进房地产业健康发展有重大意义。

## 1.2研究内容与目标

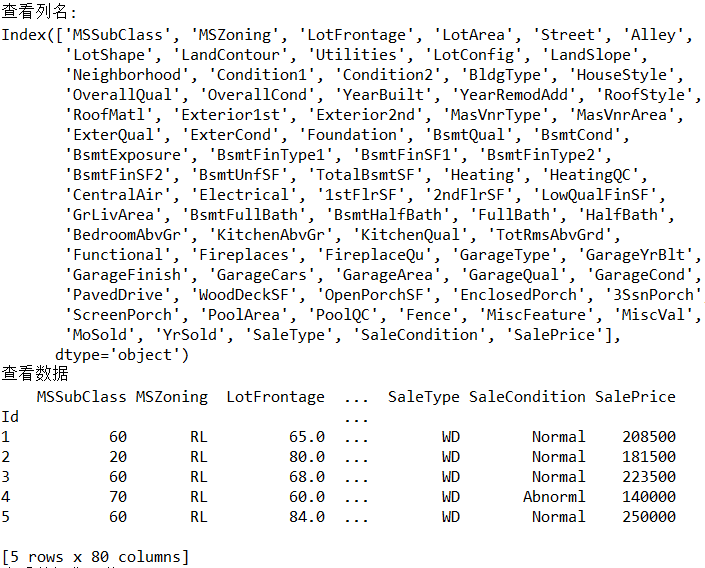
本文将利用python语言对原始数据进行分析及可视化；对数据进行降维，比较其降维前后模型效果；使用相关算法svm、决策树、随机森林等对模型进行训练；并将预测值和真实值的曲线的结果可视化。

# 第2章 数据分析

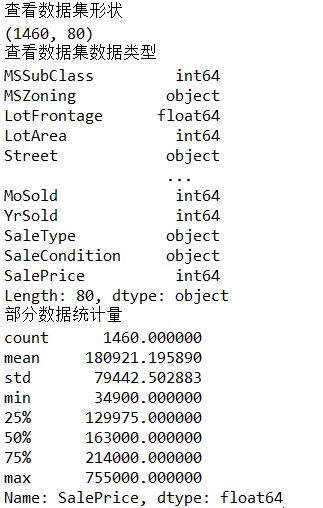
## 2.1分析数据

基于给出的 79 个特征，对房屋的售价等相关数据进行分析。

1. 描述性数据总结

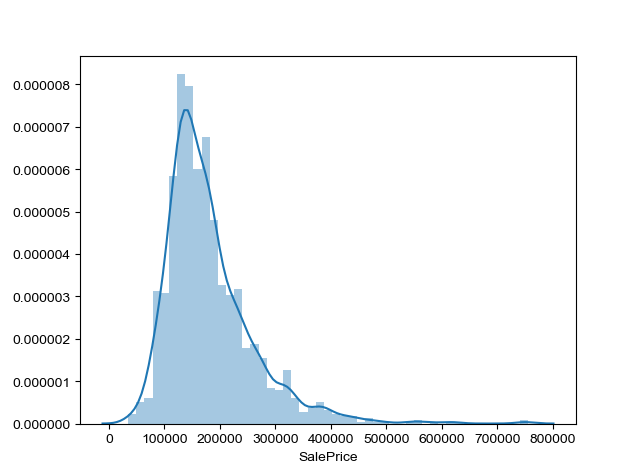


1. 数据类型统计



1. 绘制直方图

从直方图可以看出，数据偏离正态分布，数据正偏，有峰值。



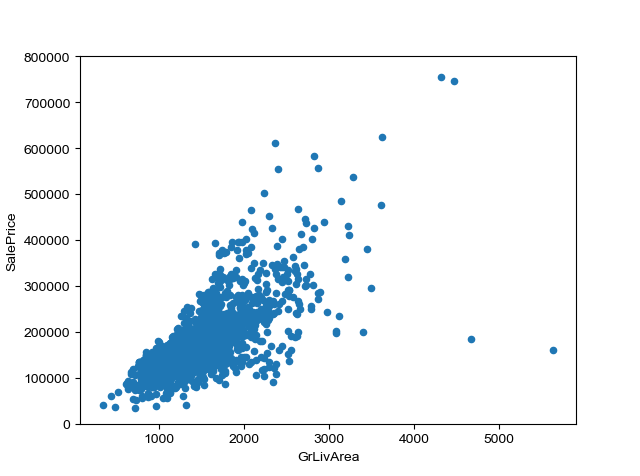
1. 数据偏度和峰度度量

偏度，是统计数据分布偏斜方向和程度的变量，是统计数据分布非对称程度的数字特征。

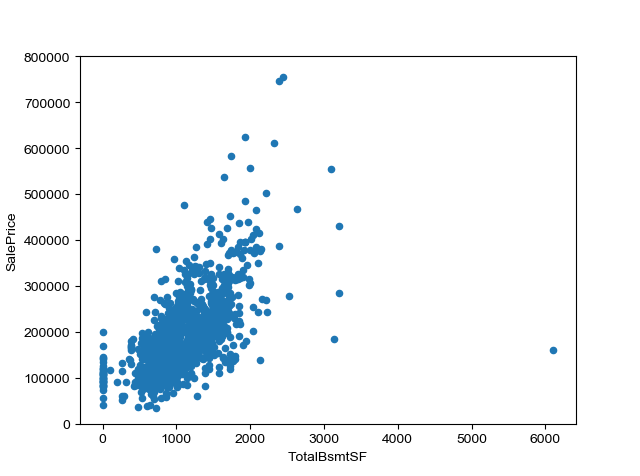
峰度又称峰态系数。表征概率密度分布曲线在平均值处峰值高低的特征。直观看来，峰度反映了峰部的尖度。



1. 与数字型变量的关系
2. **Grlivarea（居住面积）与 SalePrice（售价）散点图**

****

1. **TotalBsmtSF（地下室总面积）与 SalePrice散点图**

****

1. **与类别型变量的关系**

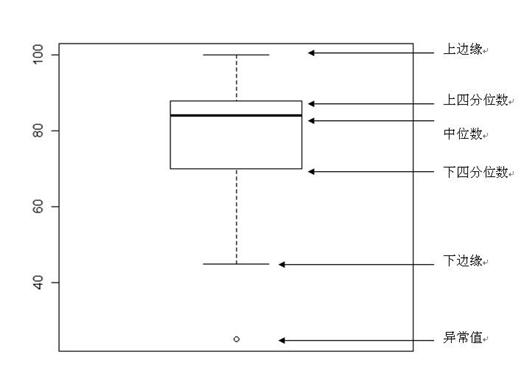
箱**型**图（Boxplot），是利用数据中的五个[统计量](https://wiki.mbalib.com/wiki/%E7%BB%9F%E8%AE%A1%E9%87%8F" \o "统计量)：最小值、下[四分位数](https://wiki.mbalib.com/wiki/%E5%9B%9B%E5%88%86%E4%BD%8D%E6%95%B0" \o "四分位数)、[中位数](https://wiki.mbalib.com/wiki/%E4%B8%AD%E4%BD%8D%E6%95%B0" \o "中位数)、上四分位数与最大值来描述数据的一种方法，它也可以粗略地看出数据是否具有有对称性，分布的分散程度等信息，特别可以用于对几个样本的比较。

异常值：大于上四分位数1.5倍四分位数差的值，或者小于下四分位数1.5倍四分位数差的值。

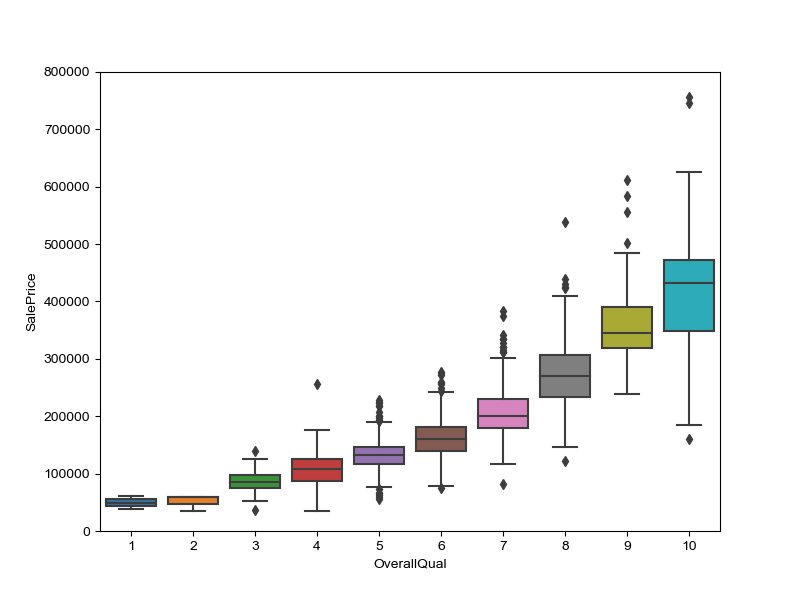
极端异常值：即超出四分位数差3倍距离的异常值，用实心点表示；较为温和的异常值，即处于1.5倍-3倍四分位数差之间的异常值，用空心点表示。

下分位数 =（n+1）/4

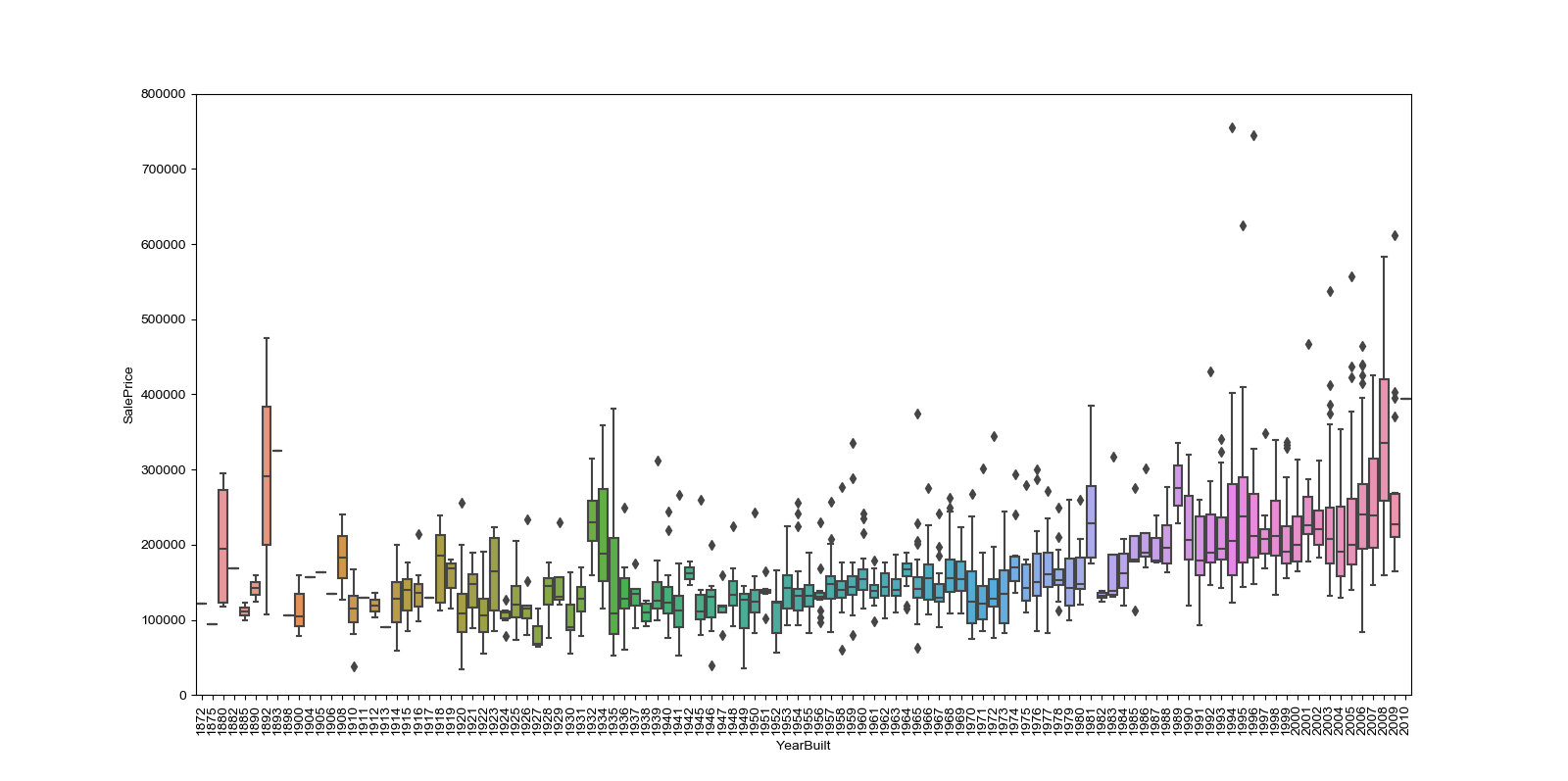
上分位数 = 3（n+1）/4



1. **OverallQual（整体材料和表面质量）与SalePrice箱型图**



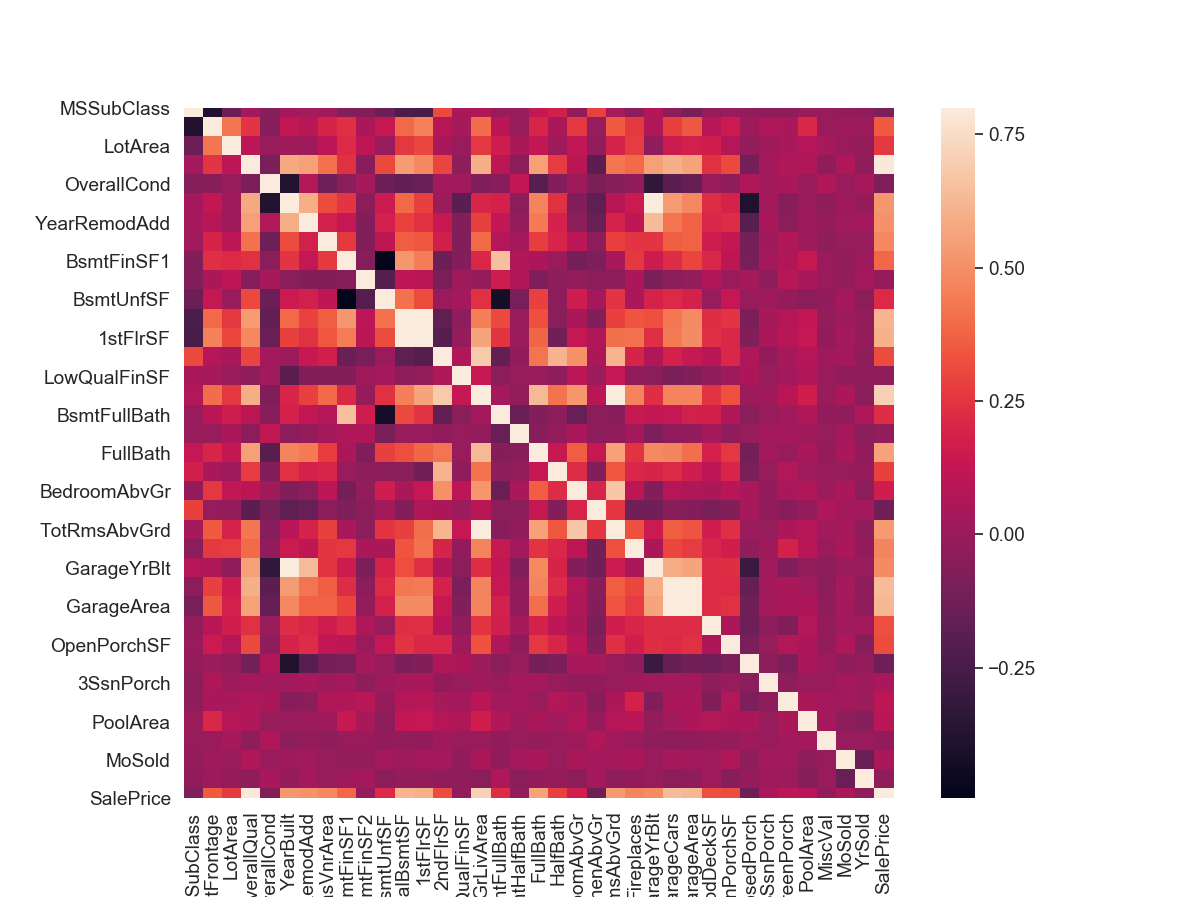
1. **YearBuilt（原始施工日期）与 SalePrice箱型图**



1. 相关系数矩阵

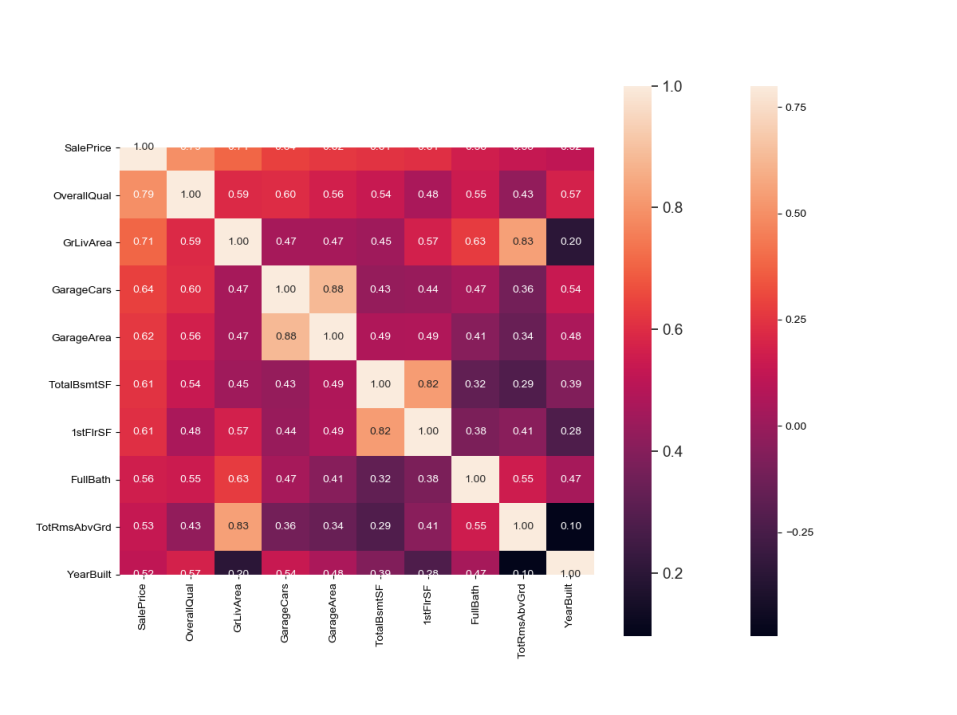
**有两个地方值得注意，第一个是 TotalBsmtSF 和 1stFlrSF 变量的相关系数，第二个是 GarageX 变量群。这两个示例都显示了这些变量之间很强的相关性。实际上，相关性的程度达到了一种多重共线性的情况。可以总结出这些变量几乎包含相同的信息，所以确实出现了多重共线性。**

**另一个引起注意的地方是SalePrice的相关性。可以看出之前分析的GrLivArea，TotalBsmtSF和OverallQual 的相关性很强，除此之外也有很多其他的变量应该进行考虑。**

****

1. **SalePrice**相关系数矩阵

**从图中可以看出：**OverallQual，GrLivArea以及TotalBsmtSF与SalePrice有很强的相关性。GarageCars和GarageArea 也是相关性比较强的变量。车库中存储的车的数量是由车库的面积决定的，不需要专门区分GarageCars和GarageAre，所以我们只需要其中的一个变量。这里我们选择了GarageCars，因为它与SalePrice 的相关性更高一些。TotalBsmtSF 和1stFloor 与上述情况相同，我们选择TotalBsmtS。FullBath 几乎不需要考虑。TotRmsAbvGrd 和 GrLivArea 也是不需要专门区分。YearBuilt 和 SalePrice 相关性似乎不强。



1. **SalePrice和**相关变量之间的散点图

从上至下依次为：SalePrice,，OverallQuGrLival，Area，GarageCars，TotalBsmtSF，FullBath，YearBuilt。



# 第3章 数据预处理

## 3.1数据降维

本文中采用主成分分析（PCA）算法对数据进行降维处理。下面对PCA算法进行简单介绍。

1. PCA算法简介

PCA是一种无监督降维算法，它是最常用的降维算法之一，可以很好的解决因变量太多而复杂性，计算量增大的弊端。他在尽可能减少信息损失的前提下，找到某种方式降低数据的维度的方法。PCA通常用于高维数据集的探索与可视化，还可以用于数据压缩，数据预处理。

通常来说，我们期望得到的结果，是把原始数据的特征空间（n个d维样本）投影到一个小一点的子空间里去，并尽可能表达的很好（损失信息最少）。常见的应用在于模式识别中，我们可以通过减少特征空间的维度，抽取子空间的数据来最好的表达我们的数据，从而减少参数估计的误差。注意，主成分分析通常会得到协方差矩阵和相关矩阵。这些矩阵可以通过原始数据计算出来。协方差矩阵包含平方和向量积的和。相关矩阵与协方差矩阵类似，但是第一个变量，也就是第一列，是标准化后的数据。如果变量之间的方差很大，或者变量的量纲不统一，我们必须先标准化再进行主成分分析。

1. PCA算法流程
2. 去掉数据的类别特征（label），将去掉后的 d 维数据作为样本
3. 计算 d 维的均值向量（即所有数据的每一维向量的均值）
4. 计算所有数据的散布矩阵（或者协方差矩阵）
5. 计算特征值（e1 , e2 , e3 , .... ed）以及相应的特征向量（lambda1，lambda2，...lambda d）
6. 按照特征值的大小对特征向量降序排序，选择前 k 个最大的特征向量，组成 d\*k 维的矩阵W（其中每一列代表一个特征向量）
7. 运行 d\*K 的特征向量矩阵W将样本数据变换成新的子空间。

注意1：虽然PCA有降维的效果，也许对避免过拟合有作用，但是最好不要用PCA去作用于过拟合。

注意2：在训练集中找出PCA的主成分，（可以看做为映射mapping），然后应用到测试集和交叉验证集中，而不是对所有数据集使用PCA然后再划分训练集，测试集和交叉验证集。

1. PCA算法总结

PCA算法作为一个非监督学习的降维方法，它只需要特征值分解，就可以对数据进行压缩，去噪。

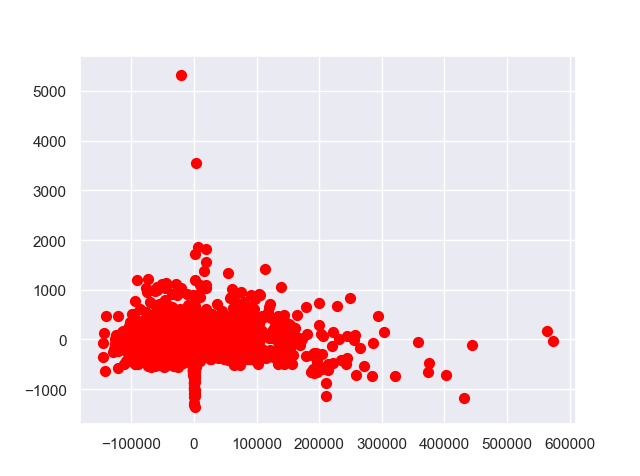
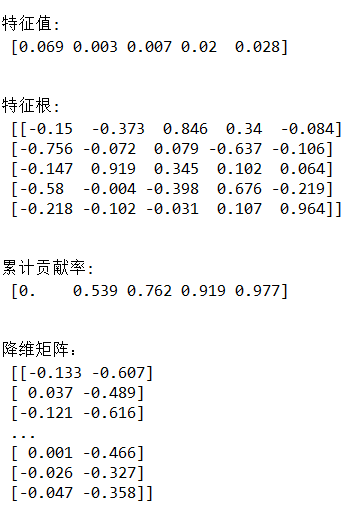
优点：

1. 仅仅需要以方差衡量信息量，不受数据集以外的因素影响。
2. 各主成分之间正交，可消除原始数据成分间的相互影响的因素。
3. 计算方法简单，主要运算时特征值分解，易于实现。

缺点：

1. 主成分各个特征维度的含义具有一定的模糊性，不如原始样本特征的解释性强。
2. 方差小的非主成分也可能含有对样本差异的重要信息，因降维丢弃可能对后续数据处理有影响。

在本次实验中，使用scikit-learn PCA类，即sklearn.decomposition.PCA处理数据。如下图所示。

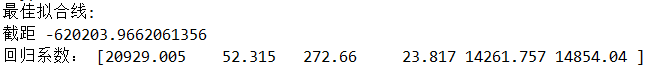
 

## 3.2 线性回归模型

1. 多元线性回归模型

这次实验中数据共有测试集和训练集两个文件，而多元回归模型则是根据训练集建立模型。

计算出的相关数据如下。



## 3.3 算法

1. SVM

SVM既可应用于线性（回归问题）分类，也可应用于非线性分类。

通过调节核函数参数的设置，可将数据集映射到多维平面上，对其细粒度化，从而使它的特征从二维变成多维，将在二维上线性不可分的问题转化为在多维上线性可分的问题，最后再寻找一个最优切割平面（相当于在决策数基础上再寻找一个最优解），因此svm的分类效果是优于大多数的机器学习分类方法的。

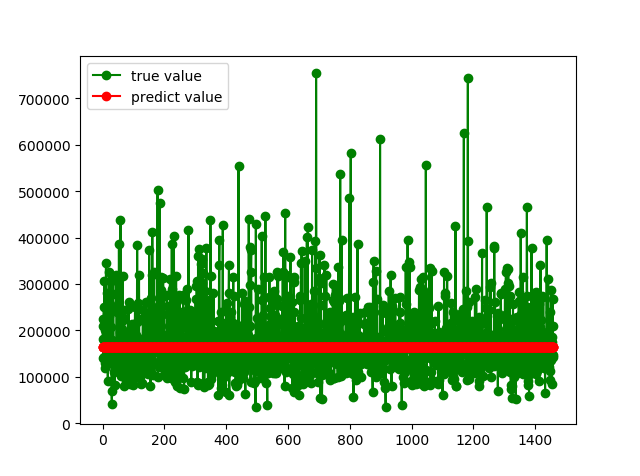
通过其它参数的设置，svm还可以防止过拟合的问题。

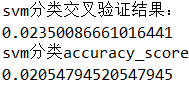
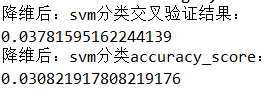
优点：

1. 非线性间隔映射是SVM方法的理论基础，SVM利用内积核函数代替向高维空间的非线性映射。
2. SVM的目标是找到对特征空间划分的最优超平面，SVM方法的核心是最大化分类边际的思想。
3. SVM的训练结果是支持向量，在分类决策中起到决定性作用。
4. SVM是一种小样本的学习方法，从本质上看，它避开了从归纳到演绎的传统过程，实现了高效的从训练样本到预测样本的“转导推理”，简化了通常的分类和回归问题。
5. SVM最终的决策函数只由少数的支持向量决定，计算的复杂性取决于支持向量的数目，而不是样本空间的维数，在某种意义上避免了“维数灾难”。
6. 有较好的鲁棒性：增删非支持向量样本对模型没有影响，SVM方法对核的选取不敏感。

缺点：

1. SVM算法对大规模训练样本难以实施。
2. SVM对解决多分类问题存在困难。



1. KNN

邻近算法，或者说K最近邻（KNN，K-NearestNeighbor）分类算法是[数据挖掘](https://baike.baidu.com/item/%E6%95%B0%E6%8D%AE%E6%8C%96%E6%8E%98/216477" \t "https://baike.baidu.com/item/%E9%82%BB%E8%BF%91%E7%AE%97%E6%B3%95/_blank)分类技术中最简单的方法之一。所谓K最近邻，就是K个最近的邻居的意思，说的是每个样本都可以用它最接近的K个邻近值来代表。近邻算法就是将数据集合中每一个记录进行分类的方法。该方法的不足之处是计算量较大，因为对每一个待分类的文本都要计算它到全体已知样本的距离，才能求得它的K个最邻近点。

KNN是一种**非参的**，**惰性**的算法模型。**非参**的意思并不是说这个算法不需要参数，而是意味着这个模型不会对数据做出任何的假设，与之相对的是线性回归。也就是说KNN建立的模型结构是根据数据来决定的，这也比较符合现实的情况，毕竟在现实中的情况往往与理论上的假设是不相符的。同样是分类算法，逻辑回归需要先对数据进行大量训练（tranning），最后才会得到一个算法模型，而KNN算法却不需要，它没有明确的训练数据的过程，或者说这个过程很快。

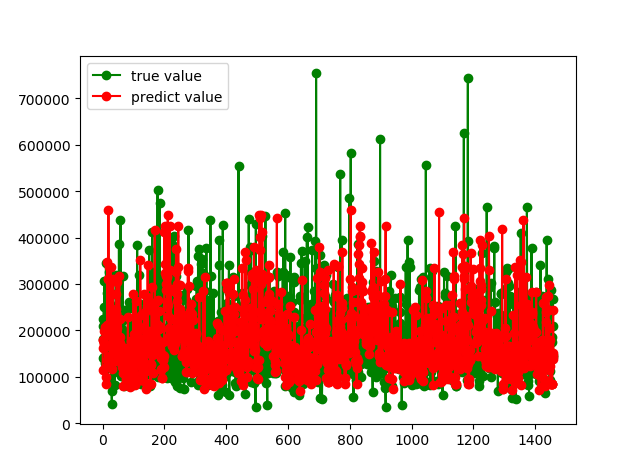
算法流程：

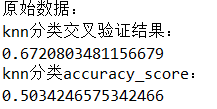
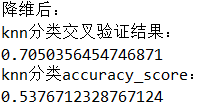
1. 准备数据，对数据进行预处理；
2. 计算测试样本点（也就是待分类点）到其他每个样本点的距离；
3. 对每个距离进行排序，然后选择出距离最小的K个点；
4. 对K个点所属的类别进行比较，根据少数服从多数的原则，将测试样本点归入在K个点中占比最高的那一类。

优点：

1. 简单易用，相比其他算法，KNN算是比较简洁明了的算法。即使没有很高的数学基础也能搞清楚它的原理。
2. 模型训练时间快，上面说到KNN算法是惰性的，这里也就不再过多讲述。
3. 预测效果好。
4. 对异常值不敏感。

缺点：

1. 对内存要求较高，因为该算法存储了所有训练数据。
2. 预测阶段可能很慢。
3. 对不相关的功能和数据规模敏感。

1. 决策树

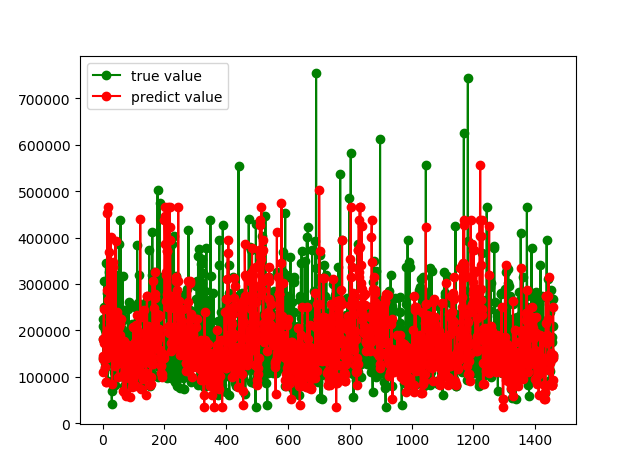
决策树是无参数的监督学习算法，可用于分类和回归，它的目标是通过学习从数据特征推断得到的决策规则，构建一个可以预测目标变量的决策模型。决策树是一种树形结构，每个内部节点表示一个属性上的测试，每个分支表示一个测试的输出，每个叶子节点代表一种类别。决策树学习的目的是为了产生一棵泛化能力强，即处理未见实例能力强的决策树，其基本流程遵循简单且直观的“分而治之”的策略。决策树学习最关键的在于**如何选择最优划分属性**。

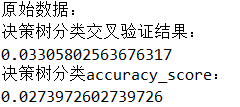
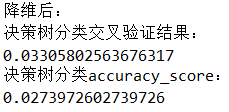
优点：

1. 决策树易于理解和实现，通过解释人们都有能力去理解决策树表达的意义。
2. 数据处理前应当去掉多余的或者空白的属性。
3. 能够同时处理数据型和常规型的属性，可以接受数据属性的多样性。
4. 对缺失值不敏感。
5. 可以处理不相关的特征数据。
6. 只需构建一次，便可以反复使用，每一次预测的最大计算次数不超过决策树的深度。

缺点：

1. 容易造成过拟合，即将训练集自身的一些特点作为所有数据的一般性质，导致过拟合，需要进行剪枝处理。
2. 对于有时间顺序的数据，需要很多的预处理的工作。
3. 当类别太多时，错误可能就会增加的比较快。
4. 只能根据一个字段进行分类。
5. 在处理特征关联性比较强的数据时，表现不好。



1. 随机森林

为了防止过拟合的问题，随机森林相当于多颗决策树。

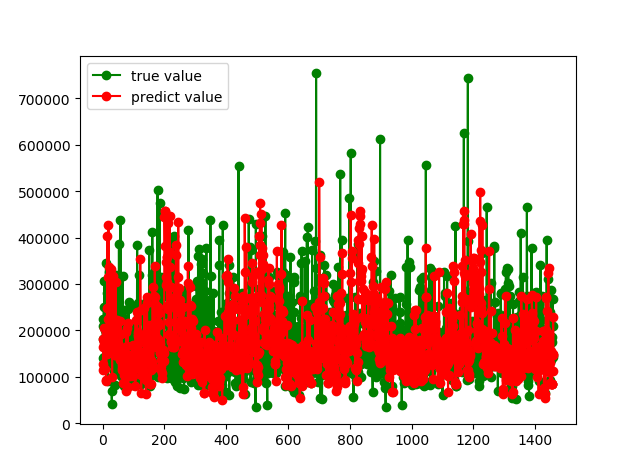
随机森林是一个用随机方式建立的，包含多个决策树的分类器。其随机性主要体现在两个方面：(1)训练每棵树时，从全部训练样本(样本数为N)中选取一个可能有重复的大小同样为N的数据集进行训练(即BootStrap取样)；(2)在每个节点，随机选取所有特征的一个子集，用来计算最佳的分割方式。

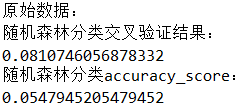
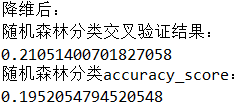
优点：

1. 能够处理高维(即特征很多)的数据，并且不用进行特征选择，是随机选择的。
2. 训练结束后，能够给出哪些特征比较重要。
3. 模型的泛化能力较强。
4. 训练速度快，容易做成并行化方法，训练时树与树之间是相互独立的。
5. 在训练过程中，能够检测到特征间的相互影响。
6. 对于不平衡的数据集来说，可以平衡误差。
7. 具有鲁棒性，即使有特征遗失，仍可以维持准确度。

缺点：

1. 在噪音较大的分类或回归问题上会出现过拟合。
2. 对于有不同取值的属性的数据，取值划分较多的属性会对随机森林产生很大的影响，因此随机森林在这种数据上产生的属性权值是不可信的。



1. 小结

各算法score如下表所示：

|  |  |  |
| --- | --- | --- |
|  | 原始数据 | 降维后 |
| SVM | 0.021 | 0.031 |
| KNN | 0.503 | 0.538 |
| 决策树 | 0.027 | 0.027 |
| 随机森林 | 0.055 | 0.195 |

从表中可以看出，数据降维后算法的效果比降维前更好。

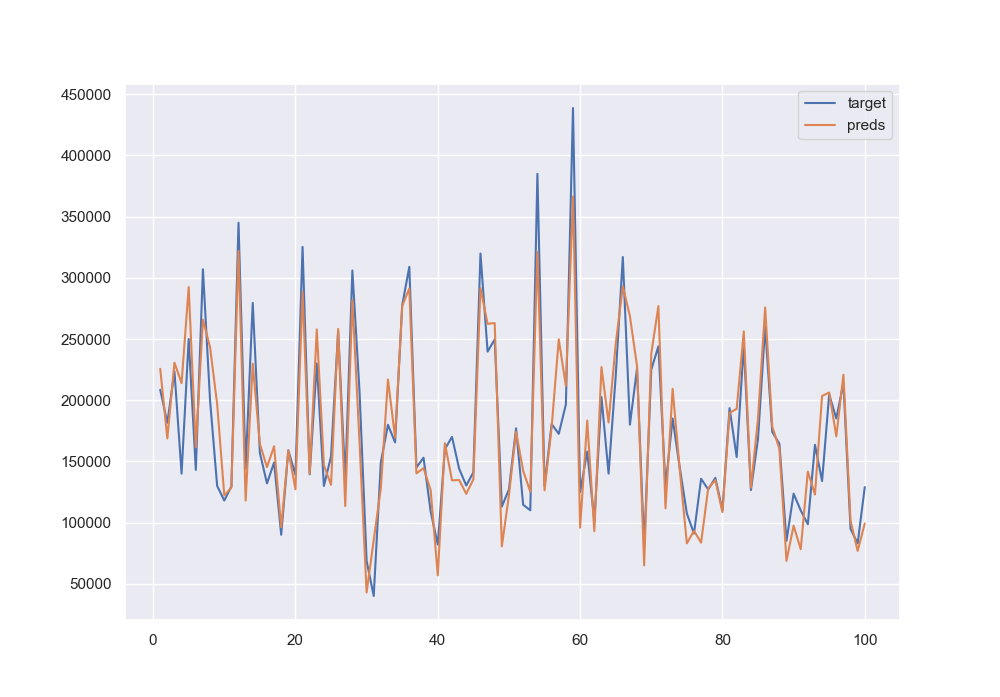
在四个算法中，KNN算法的计算效能虽然较差，但其算法效果最好。

## 3.4 结果可视化

基于梯度下降的线性回归模型

主要步骤如下：

1. 先确定向下一步的步伐大小，我们称为学习率α；
2. 任意给定一个初始值；
3. 确定一个向下的方向，并向下走预先规定的步伐，并更新θ值；
4. 当下降的高度小于某个定义的值ε，则停止下降。

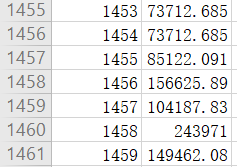
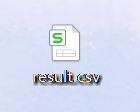


使用均方误差来评价模型好坏

均方误差（mean-square error, MSE）是反映[估计量](https://baike.baidu.com/item/%E4%BC%B0%E8%AE%A1%E9%87%8F/6395750" \t "https://baike.baidu.com/item/%E5%9D%87%E6%96%B9%E8%AF%AF%E5%B7%AE/_blank)与被估计量之间差异程度的一种度量。



将预测出的房价输出到result.csv中

# 结 论

房价综合了许多影响因素，主要体现分析了市场的走势，预测房价的变化规律对保持房地产市场稳定、促进房地产业健康发展有重大意义。本文中数据集中有７９个特征，实验中选取了与SalePrice有关的６个主要特征，对其房价进行分析与预测。其中，使用PCA算法对数据进行降维，使用SVM、KNN、决策树、随机森林等算法对数据进行预测，还通过基于梯度下降的线性回归模型对数据进行预测和可视化，并通过MSE和RMSE来评价模型。

通过上述分析与预测可以较为方便的看出一些数据与房价之间的关系，以及较好的预测未来的房价。

# 参考文献

[1]陈泽坤,程晓荣.基于梯度下降算法的房价回归分析与预测[J].信息技术与信息化,2020(05):10-13.

[2]张砚博.基于多元线性回归分析的西安市房价预测分析[J].西部皮革,2020,42(10):71.

[3]任鹏宇,王林山.基于主成分分析的随机遗传BP网络及其房价预测[J].滨州学院学报,2019,35(04):58-64.

[4]黄琪.基于Python的数据可视化方法和系统实现[J].信息与电脑(理论版),2019(14):137-140.