**基于神经网络的未来房价趋势预测**

**组 员：**

李本源3120180995

李博闻3120180996

李营3120181003

卢是男3220190724

**指导教师：** 汤世平

目录

[一、任务概述 3](#_Toc9948880)

[二、国内外研究综述 4](#_Toc9948881)

[三 数据集介绍 5](#_Toc9948882)

[四.实验方法 5](#_Toc9948883)

[4.1 数据处理与特征分析 5](#_Toc9948884)

[4.2 数据特征的选择 6](#_Toc9948885)

[4.3 模型的选择与优化 8](#_Toc9948886)

[为了实现房价的预测，我们首先将数据拆分为训练数据及测试数据。 8](#_Toc9948887)

[4.3.2 SVR模型预测房价 8](#_Toc9948888)

[我们尝试使使用三种核，'linear', 'poly', 'rbf'分别来对房价进行预测 8](#_Toc9948889)

[4.3.3 KNN模型预测房价 9](#_Toc9948890)

[4.3.4 决策树预测房价 10](#_Toc9948891)

[4.4 模型评估 11](#_Toc9948892)

[4.5 模型优化 13](#_Toc9948893)

[4. 6最优模型确定 18](#_Toc9948894)

[4.7 模型预测 18](#_Toc9948895)

# 一、任务概述

近年来，全国各大小城市的房价一直在上涨，房价上涨的原因是多样的，房地产供需不平衡、购 房者的投机性需求、人口上涨以及住房的改善性需求、 GDP 增长的需求等均为房价上涨的原因。综合考虑，影响房地产 价格上涨的因素有很多并且也比较复杂，各个因素之间相互 影响，相互作用，如何对房价走势有效地进行预测，控制房价 过快增长成为当今社会人们关心并且迫切想知道的热点 问题。

通过翻阅相关资料获知，影响房价的因素较多且复杂， 比如政府设定的地价，房屋所处的地理位置，住房供需关系， 国家出台的关于房地产的政策，人口状态，投资额，家庭平均 收入，建房建筑面积、公摊面积以及使用面积等。由于影响 因素较多，以及众影响因素之间的相互关系，考虑的方面比 较多，因此对于房价的预测是比较困难的，再加上很多影响 因素的数据难以获得，只能将其作为影响因素的参考，而不 便于放入模型中，所以对于房价的预测只能尽可能地去做多次模拟以提高预测的精度。

传统的计量经济模型，大多数模型根据设定的变量可以 很好地解释模型运行的结果，比如线性模型。而房价的影响因素错综复杂，走势多变，使用传统的模型去预测房价走势， 其模型效果检验可能不会太好，误差较大。而BP神经网络 模型不用明确输入输出之间的函数关系，虽然对于输出结果 不太好解释，但对房价的预测却比较合适。因为使用BP神 经网络模型所考虑的函数关系越复杂，其非线性逼近能力就 越强。综合考虑影响房价的众多因素以及其复杂性，本文在美国某经济学杂志上发布的历年数据的基础上，利用 BP 神经 网络模型，对20世纪70年代波士顿房屋价格进行预测并验证其准确性。

# 二、国内外研究综述

从古至今，“吃穿住行”构成人们生活的主要部分，而住 与人们的生活休戚相关，住是提供人们一个休憩之所，住房 是人们口口相传的家的象征。而现如今，房价波动上升的趋 势以及人们对房价的关心，使得房价成为一大热点问题。广 大学者对住宅房价问题进行了诸多研究。在城市商品住宅 价格影响方面，高霞、蒋立红和厉文平通过因子分析将影响 因素划分为城市基础设施与环境因子、城市规模因子、城市 经济因子和城市区位因子。之后通过主成分回归法分析了 以上因子与房价的关联度，分析发现: 城市经济因子、城市规 模因子、城市基础设施和环境因子、城市区位因子这四类因 子对房价的影响重要性逐渐减弱。杨贵中通过成都 1997— 2005 年的房价诸多影响因素的相关数据对房价进行预测，通过多元回归建立单方程模型，研究发现: 成都市非农人口的 增长是影响成都商品住宅价格最重要的因素之一。乔林、孔淑红利用2000—2009年4个一线城市以及14个二三线城市 的房地产数据分别对两类城市房价因素做出分析，研究发现:一线城市与二三线城市的影响因素存在较大差异。一线城市房价受居民收入的影响较为显著，二三线城市房价受人 口因素的影响较为显著。该研究结果与杨贵中的“成都市房 价最重要的影响因素为非农人口的增长”也是比较吻合的。随着社会经济的发展，影响房价的因素越来越复杂，根据传统计量模型对房价进行预测局限性越来越大。李东月对房价预测模型进行了比较，发现5次多项式模型比灰色－马尔可夫预测模型的预测精度高。Limsombunchai 和 Gan 等 根据夏威夷 Christchurch 市 200 个交易数据分别建立人工神 经网络模型和 Hedonic 模型，并对这两个模型进行比较分析， 发现人工神经网络模型的预测效果比 Hedonic 模型好。周学君和陈文秀根据 BP 神经网络模型对黄冈市房价进行预测， 结果发现，预测结果符合预期误差。

从学术研究现状来看，预测房价的方法较多，在计量经 济学的基础上建立起来的模型大多数都是线性模型，现如 今，房价走势复杂多变，房价影响因素错综复杂，房价的变化 呈现一种非线性的态势，仅通过线性模型预测房价存在较大 的误差，无法给出较好的预测结果。传统的计量模型已不适 用于房价的预测。根据参考众多文献发现，BP神经网络无须事先确定输入与输出之间映射关系的数学方程，仅通过自身的训练，学习某种规则，在给出输出结果的情况下，最大限度地逼近输出结果，由此根据输入和输出逼近一个函数使得预测结果精度最大化。这是一种智能信息处理系统，是技术发展的新产品。因此本文对南京市房价预测使用 BP 神经网 络模型，之后对南京市房价进行验证，评估其预测结果的准确性。 三、相关技术

# 三 数据集介绍

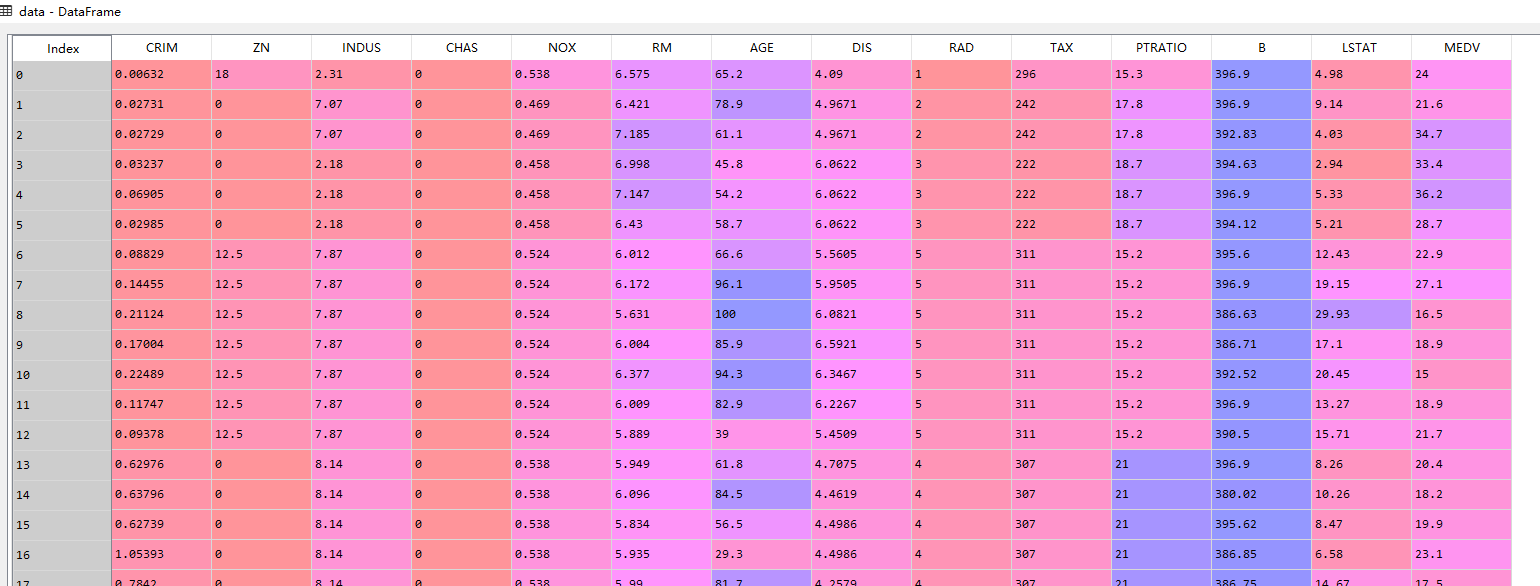
要预测的是是20世纪70年代波士顿房屋价格的中位数。这里给出的数据包括犯罪率、当期房产税率等。本次，我们有的数据点共有506个，分为404个训练样本和102个测试样本。输入数据的每个特征都有不同的取值范围。有些特征是比例，取值范围为0-1，有的特征取值范围为1-12；还有的特征取值范围为0-100等。这里每个样本有13个特征，比如犯罪率、每个住宅平均房屋间数、告诉公路的可达性等。我们的目标（或者说希望的测试结果）是房屋价格的中位数，单位是千美元。

数据集中的每一行数据都是对波士顿周边或城镇房价的描述：  
CRIM: 城镇人均犯罪率  
ZN: 住宅用地所占比例  
INDUS: 城镇中非住宅用地所占比例  
CHAS: CHAS 虚拟变量,用于回归分析  
NOX: 环保指数  
RM: 每栋住宅的房间数  
AGE: 1940 年以前建成的自住单位的比例  
DIS: 距离 5 个波士顿的就业中心的加权距离。  
RAD: 距离高速公路的便利指数  
TAX: 每一万美元的不动产税率  
PRTATIO: 城镇中的教师学生比例  
B: 城镇中的黑人比例  
LSTAT: 地区中有多少房东属于低收入人群  
MEDV: 自住房屋房价中位数（也就是均价）

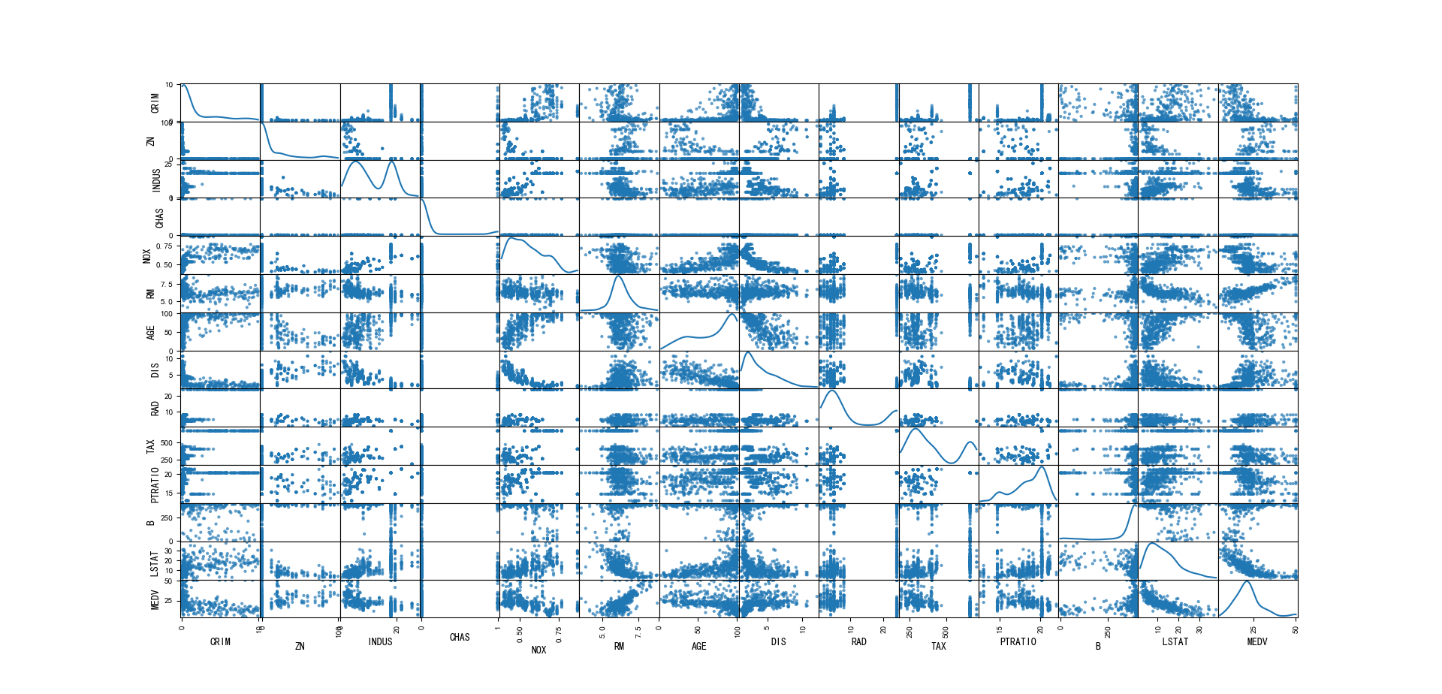
# 四.实验方法

## 4.1 数据处理与特征分析

下图是我们得到的数据的截图，每一列都是一个属性的值。



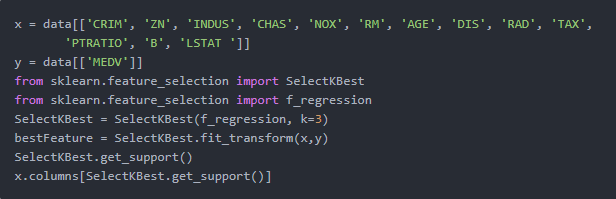
首先我们检查数据中有没有空值，我们将数据中的空值剔除，然后查看各个特征的散点分布，下图是各个变量的矩阵三点图，根据矩阵散点图形状，看变量间的相互关系。

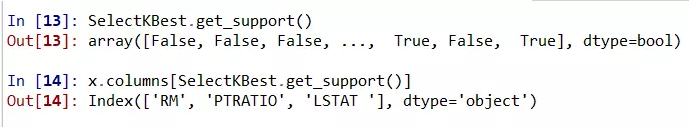


## 4.2 数据特征的选择

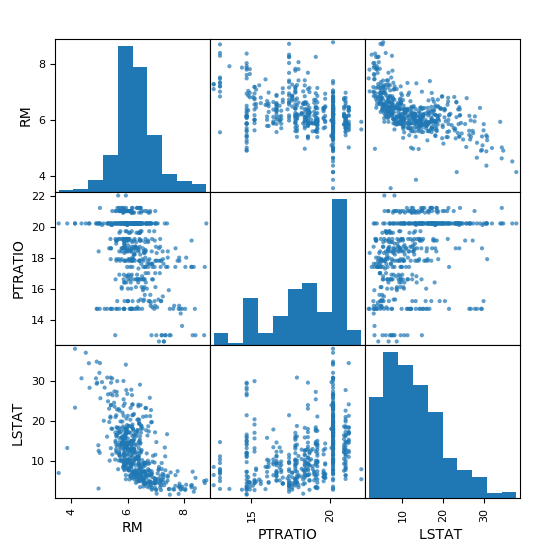
特征维度较大，为了保证模型的高效预测，需要进行特征选择。每种特征都有自己含义和数据量级，单纯地依靠方差来判断可能效果不好，直接使用与目标变量的相关性强的变量作为最终的特征变量。

我们通过相关系数法进行特征选择





通过分析，我们看出和波士顿房价相关性最强的三个因素，分别是，RM(每栋住宅的房间数)，PTRATIO(城镇中的教师学生比例)，LSTAT(地区中有多少房东属于低收入人群)。  
还是具备一定逻辑性的，首先，房子越大房价自然高（不管在哪个地域），其次，师生比与房价成反比，教育的重视，教育资源越是富裕的地方，生源就会大，师生比自然会降低，周边的房价会升高，这就是所谓的“学区房”概念吧。关于有多少房东属于低收入人群和房价的负相关关系，这个也比较好理解，各种原因吧。



## 4.3 模型的选择与优化

### 为了实现房价的预测，我们首先将数据拆分为训练数据及测试数据。

from sklearn.cross\_validation import train\_test\_split

x\_train, x\_test, y\_train, y\_test = train\_test\_split(features[['标准化RM', '标准化PTRATIO', '标准化LSTAT ']], y, test\_size=0.3,random\_state=33)

房价预测方面，打算尝试以下方法：

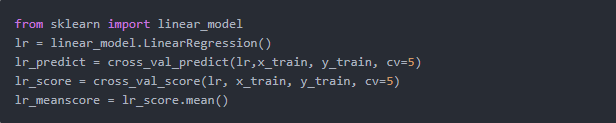
1.线性回归（这应该是最简单也是最好理解的一种方法。）

2.使用支持向量回归模型SVR

3.KNN模型

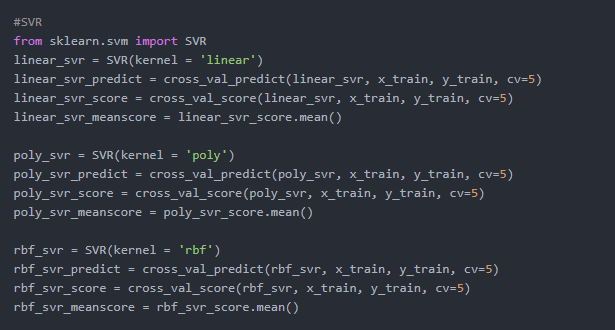
4.决策树

* 4.3.1 线性回归模型预测房价
* 使用线性回归的方法来预测房价的方法很简单，实现的代码如下：



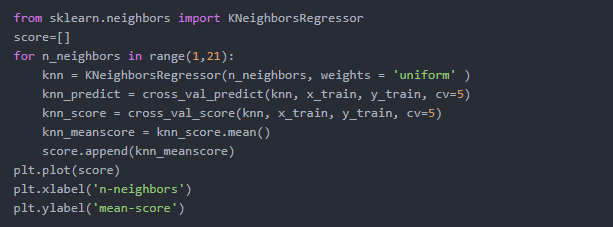
### 4.3.2 SVR模型预测房价

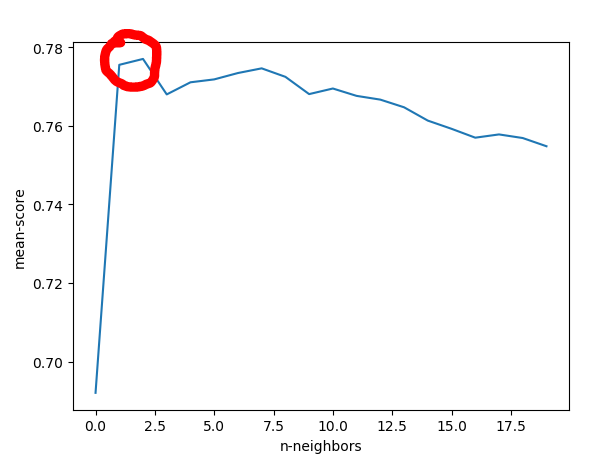
### 我们尝试使使用三种核，'linear', 'poly', 'rbf'分别来对房价进行预测



### 4.3.3 KNN模型预测房价

在KNN的回归模型当中，我们没法确定n\_neighbors，因此需要最优化这个参数。威威QQ11分别计算n\_neighbors=[1,2,...,19,20]。

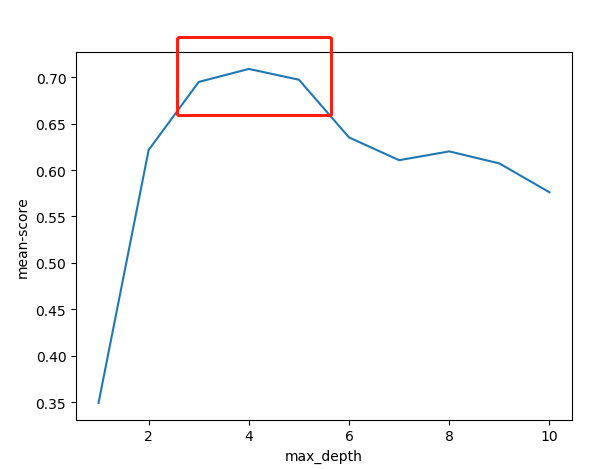




从上图可以发现，随着n\_neighbors的逐渐增大，模型预测能力逐渐增强，但是当n\_neighbors大于2以后，模型评分趋于下降。因此我们选取n\_neighbors=2。

### 4.3.4 决策树预测房价

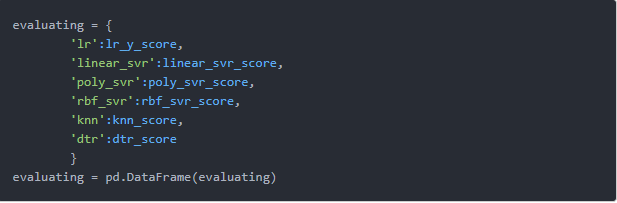
和KNN类似，我们没法确定决策树的深度，因此令最大深度分别是1至10。然后我们选取一个最优的深度。

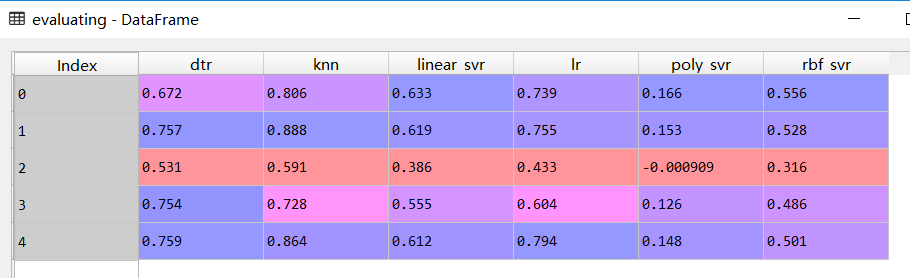


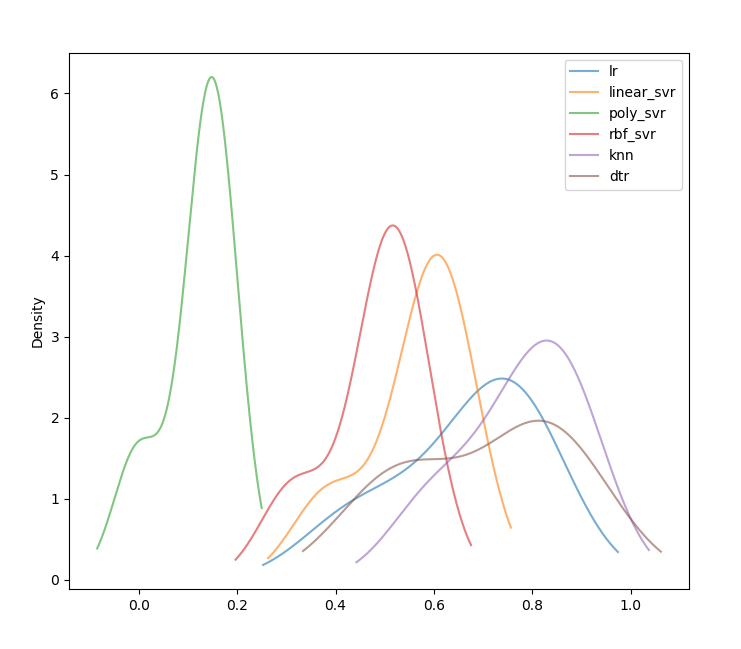
根据上图我们发现，发现当max\_depth为[3, 4, 5]时，决策时模型评分处于极值的样子。因此取max\_depth为4。

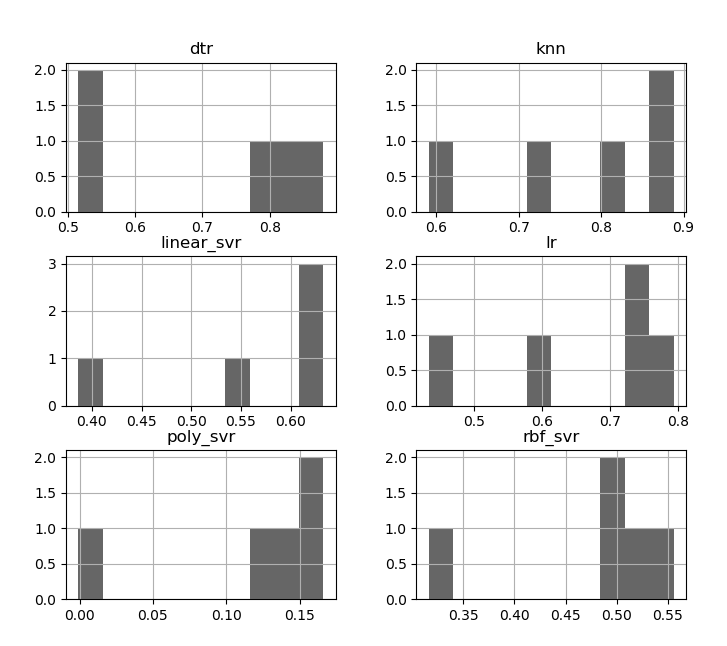
## 4.4 模型评估

介绍完模型之后，我们接下来，汇总一下各个模型的评分。









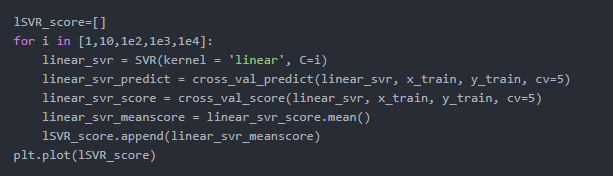
从以上两张图发现，kernerl为poly的SVR得分受数据影响明显，而且得分偏低，其他的几个模型类似linear/rbf的SVR，dtr，都呈现出相同的趋势，KNN模型应该是算是截至在现在得分最高的模型。

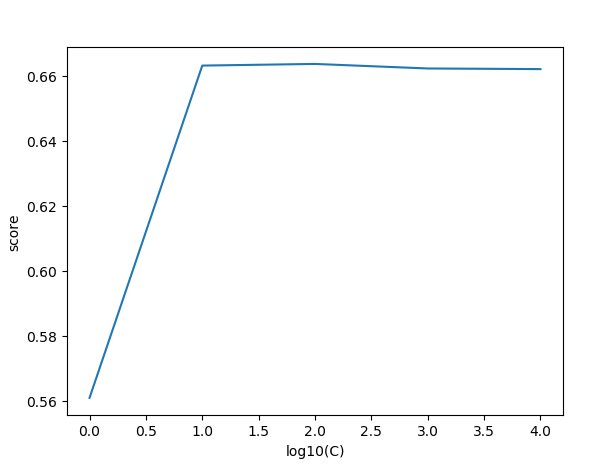
## 4.5 模型优化

接下来想想办法，看看SVR还能不能被拯救。

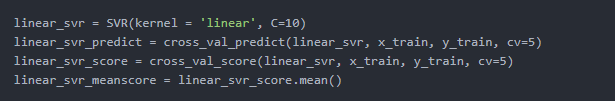
首先对线性核进行最优化。

线性核我们通过更改惩罚系数C来查看对模型的影响。

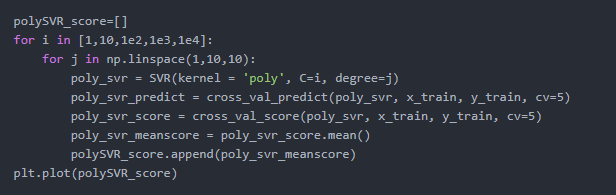


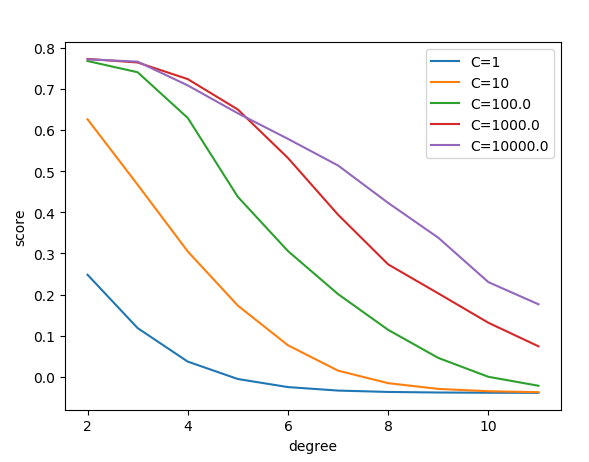


观察C为[1,10,100,1000]时对模型的影响，发现当C为10时，模型评分处于极值状态，随着C的增大，模型得分趋于稳定变化不大，因此认为C为10时模型最优。而sklearn关于线性核的默认状态C为1。



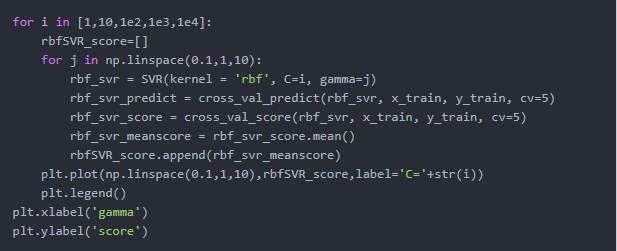
'poly'核优化，通过尝试修改惩罚系数C和degree。

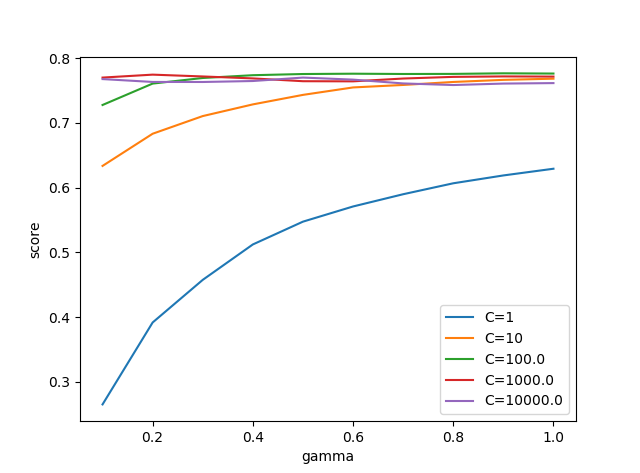




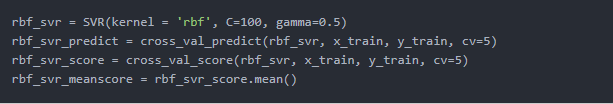
从上图看出，对变量C来说，当C>10时，poly核模型得分普遍较高，当degree=2时，模型得分最高。我们查看SVR的API，发现，poly核默认的degree为3，C为1，可以从图中看出，优化后的模型得分增加不少。

‘rbf'核优化，优化惩罚系数C和gamma。

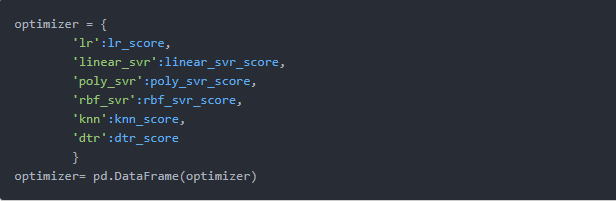


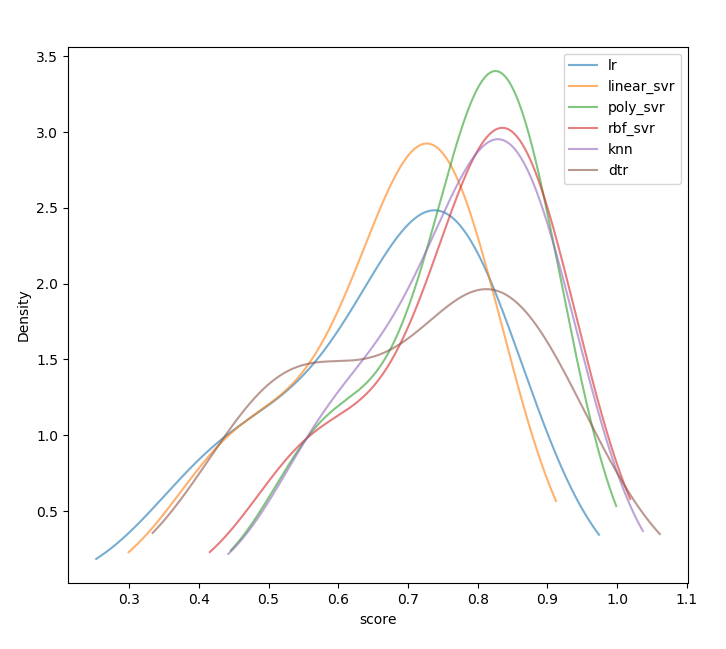


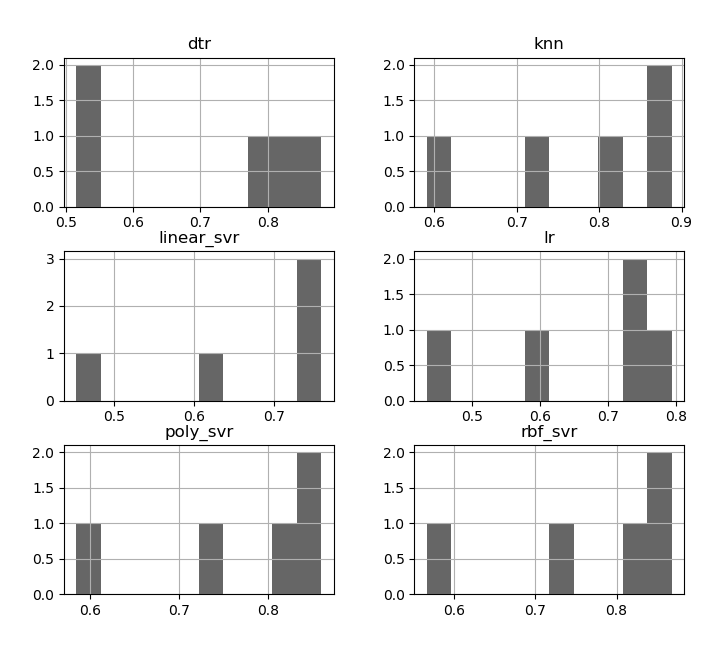
从上图发现，gamma从0.1递增到1.0，步长为0.1，模型得分有递增的趋势，当C>=10时，gamma对模型的影响较小，当C=100时，gamma=0.5时，’rbf‘模型评分更高。



接下来，对最优化的模型进行二次归类。

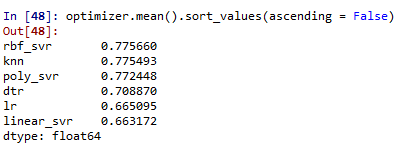






## 4. 6最优模型确定

对每个经过优化后的模型进行得分对比

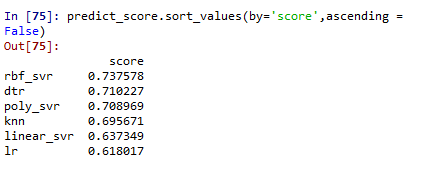


此时发现，rbf核的SVR模型在优化后变成了最好的模型。线性核的SVR和线性回归因为策略的局限性，模型能力排在最后。

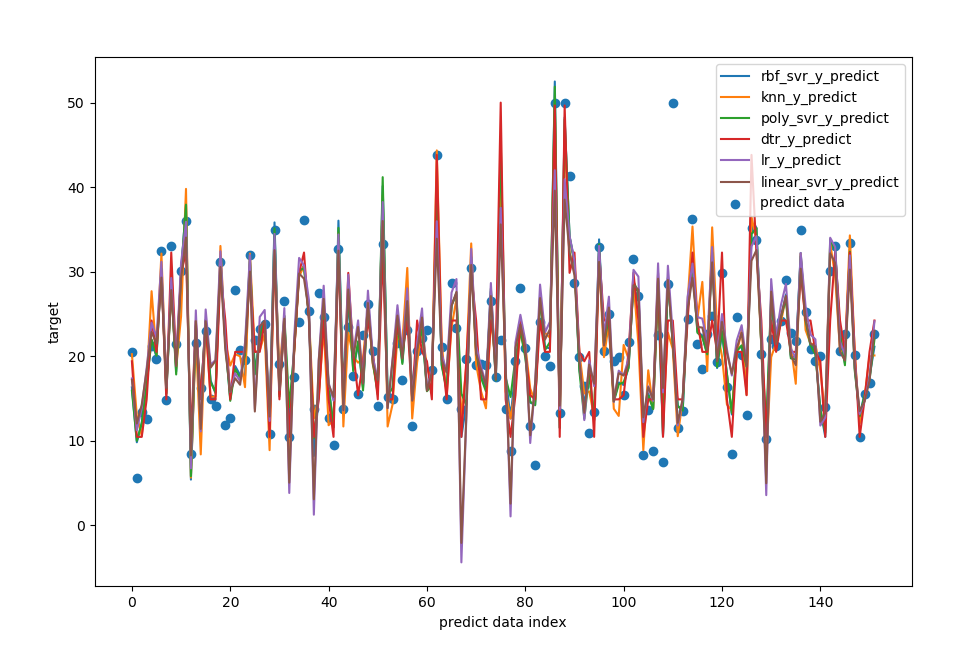
## 4.7 模型预测

接下来对测试数据集进行预测。

预测结果排名



对各个模型的预测值整理



综合表现来看，rbf\_svr的鲁棒性更强一点。