可视计算基础大作业报告

17042127 陶逸群

选题

本题来自于kaggle平台，题目给出训练集和测试集，训练集给出与房屋价格有关的一系列特征以及房屋的实际价格，测试集仅包含与房屋价格有关的一系列特征，要求训练模型预测测试集中的房屋价格，预测结果以.csv的格式提交平台评分。

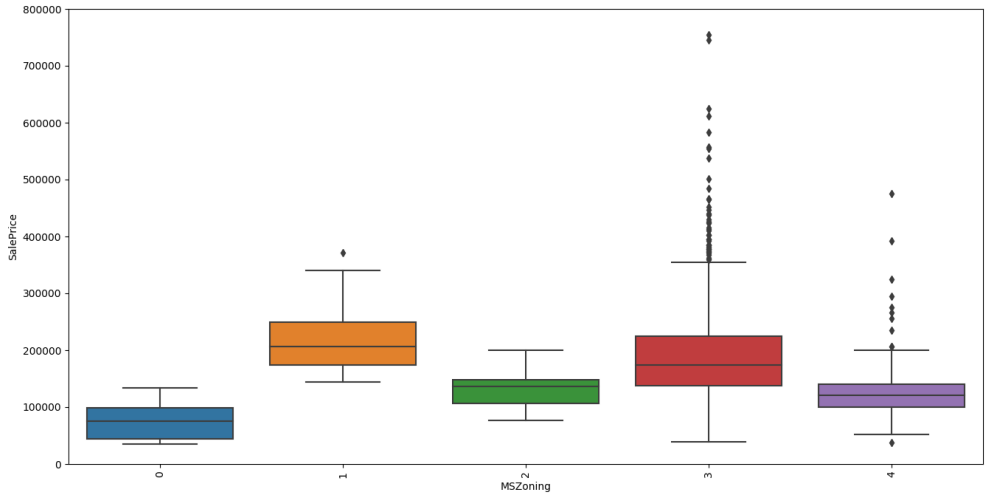
概述

这是一个典型的回归问题，为了很好的评估房屋价格，我采用了集成模型来进行房价预测，具体包括六个模型：岭回归模型、袋装法模型、核函数岭回归模型、贝叶斯回归模型、增强树模型以及深度神经网络模型，并且根据其预测好坏进行了加权平均。

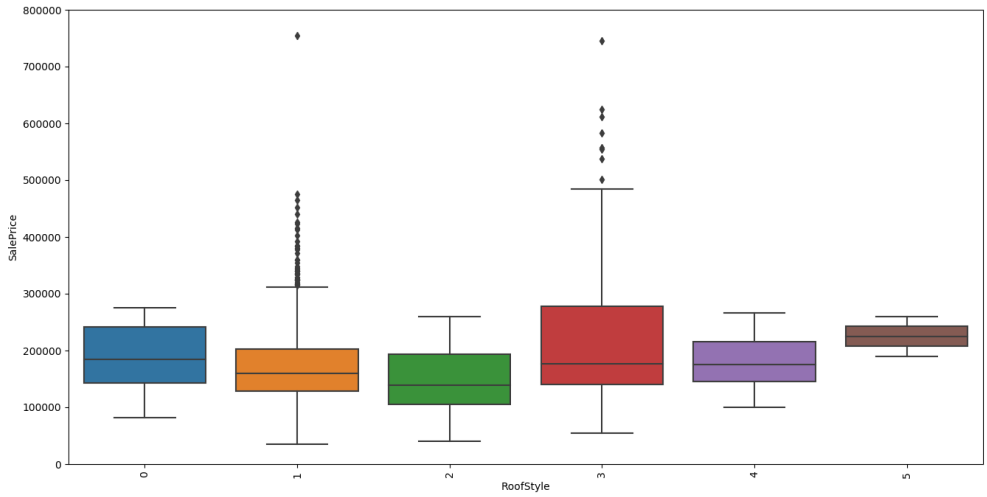
在这些模型的基础上，我具体做了数据处理，参数调优，权重分析等工作。经过处理，我得到的集成模型的预测结果在平台上排名2136名。

数据观察

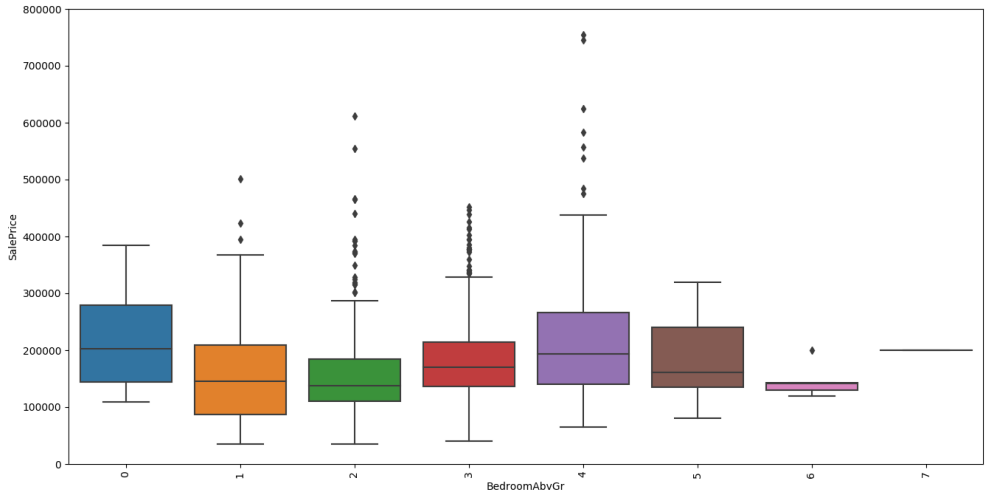
为了预测房价，我做出了每一个特征与房价之间的关联散点分布图。有些特征明显呈现分类的特征，便做出这些特征与房价之间的关联线箱图。由于特征太多下面仅举例几张。



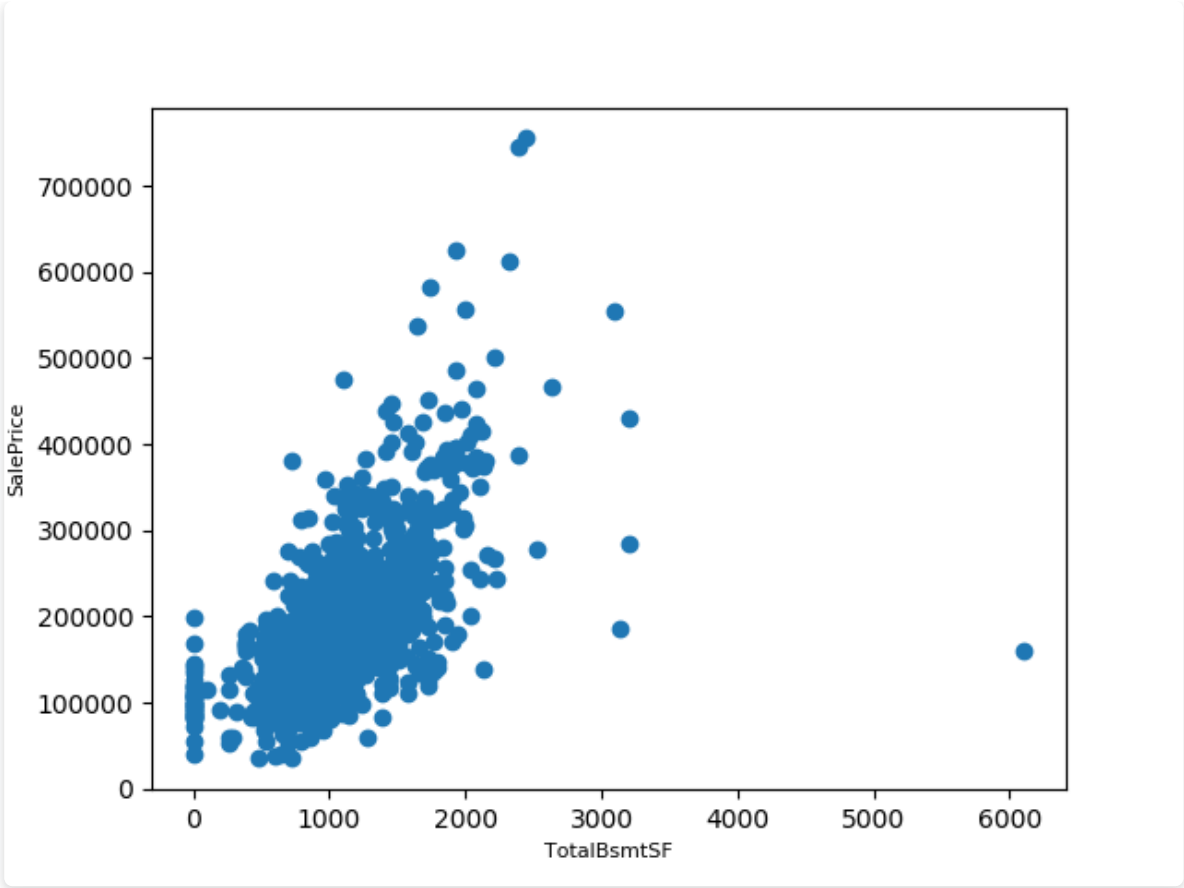
建筑物类型与房价间关联线箱图



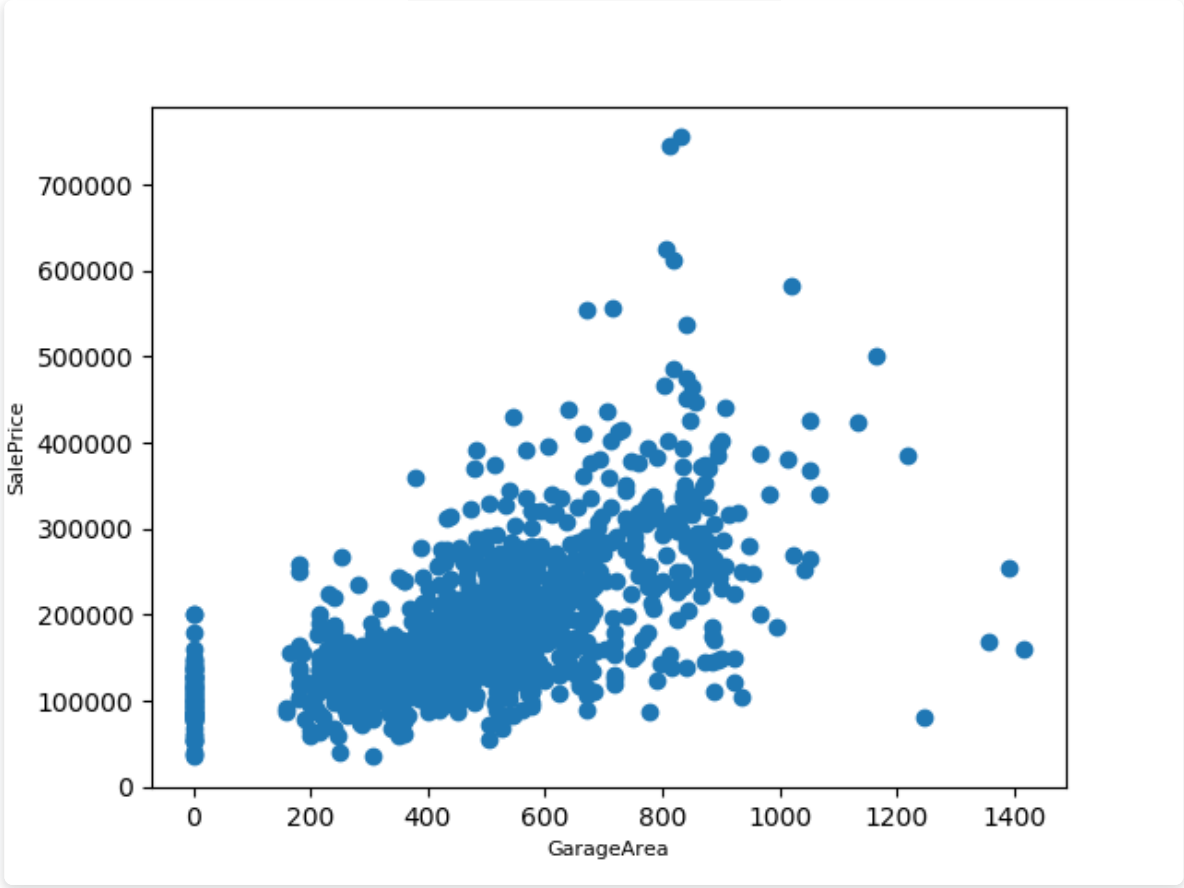
房顶类型与房价间关联线箱图



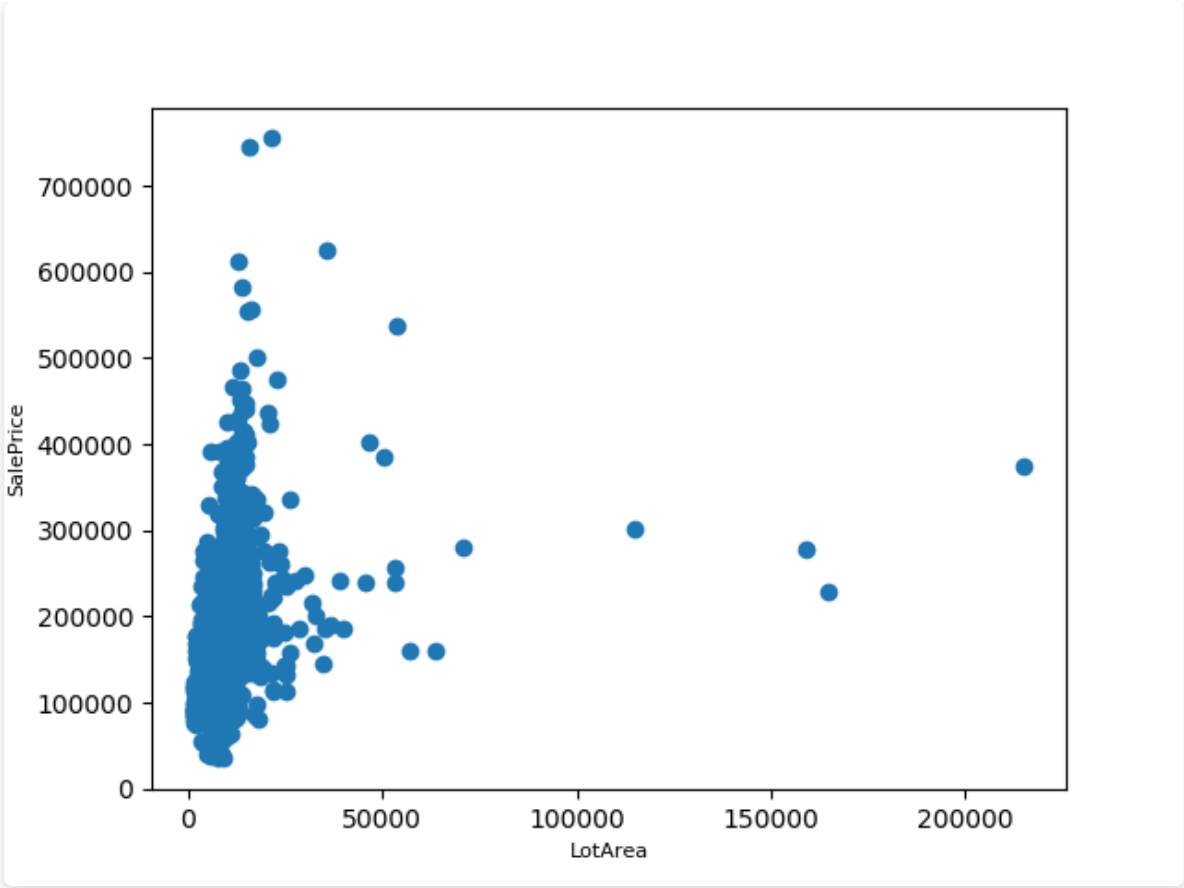
卧室个数与房价间关联线箱图



地下室面积与房价间关联散点图

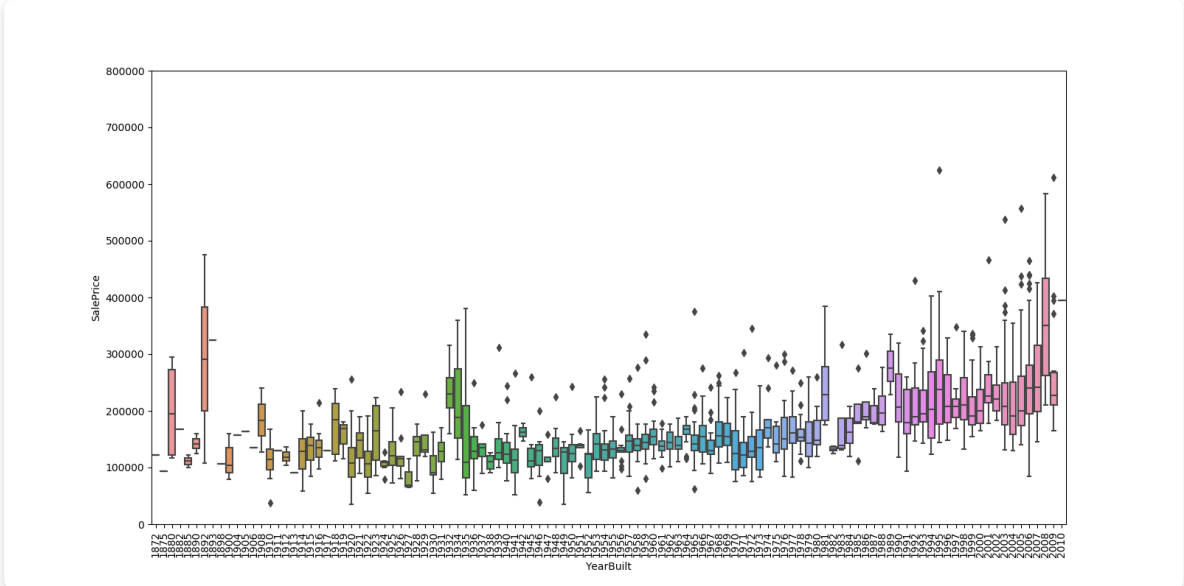


车库面积与房价间关联散点图



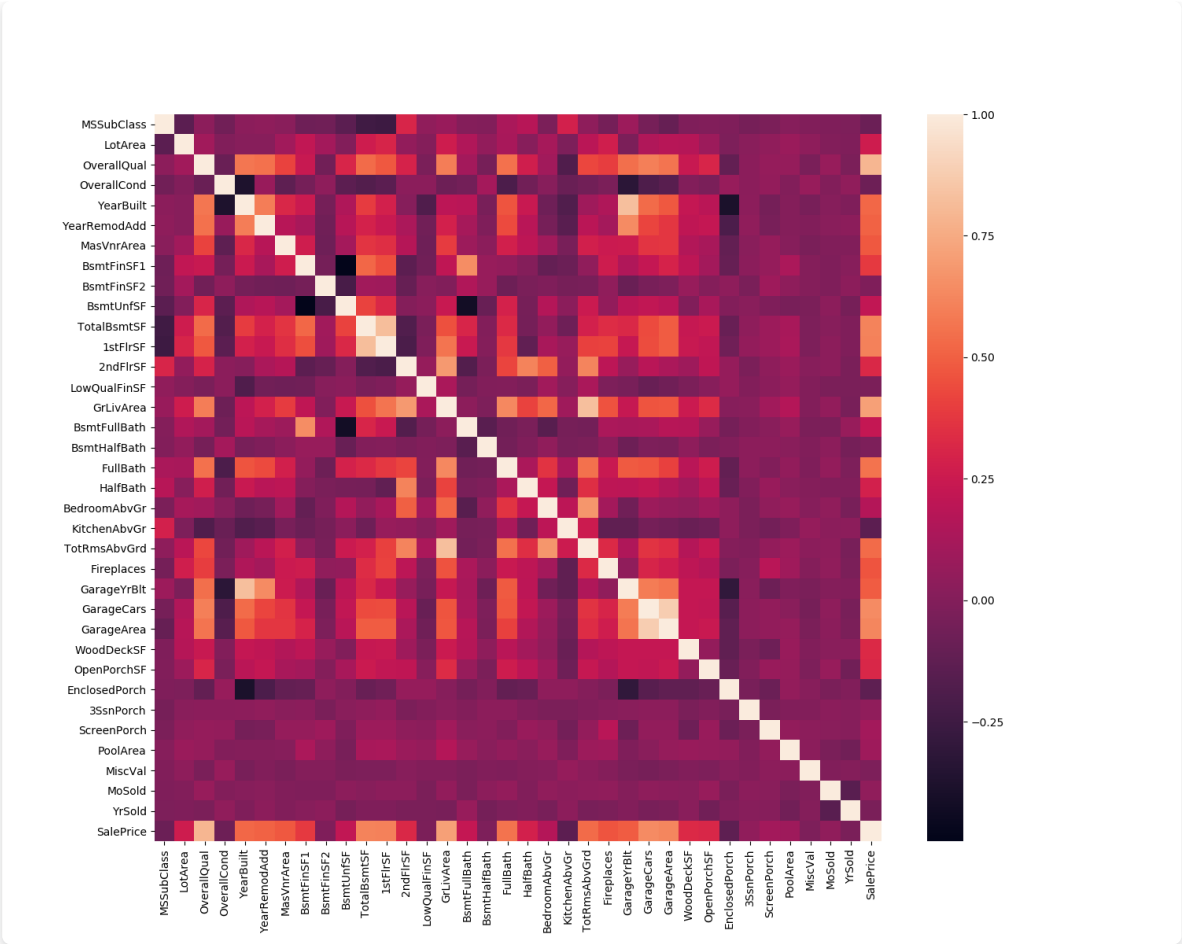
房屋占地面积与房价间关联散点图

做出房价关于修建年代的分布图。



房价关于修建年代的分布图

做出各个特征间的相关系数图。



各个特征间的相关系数图

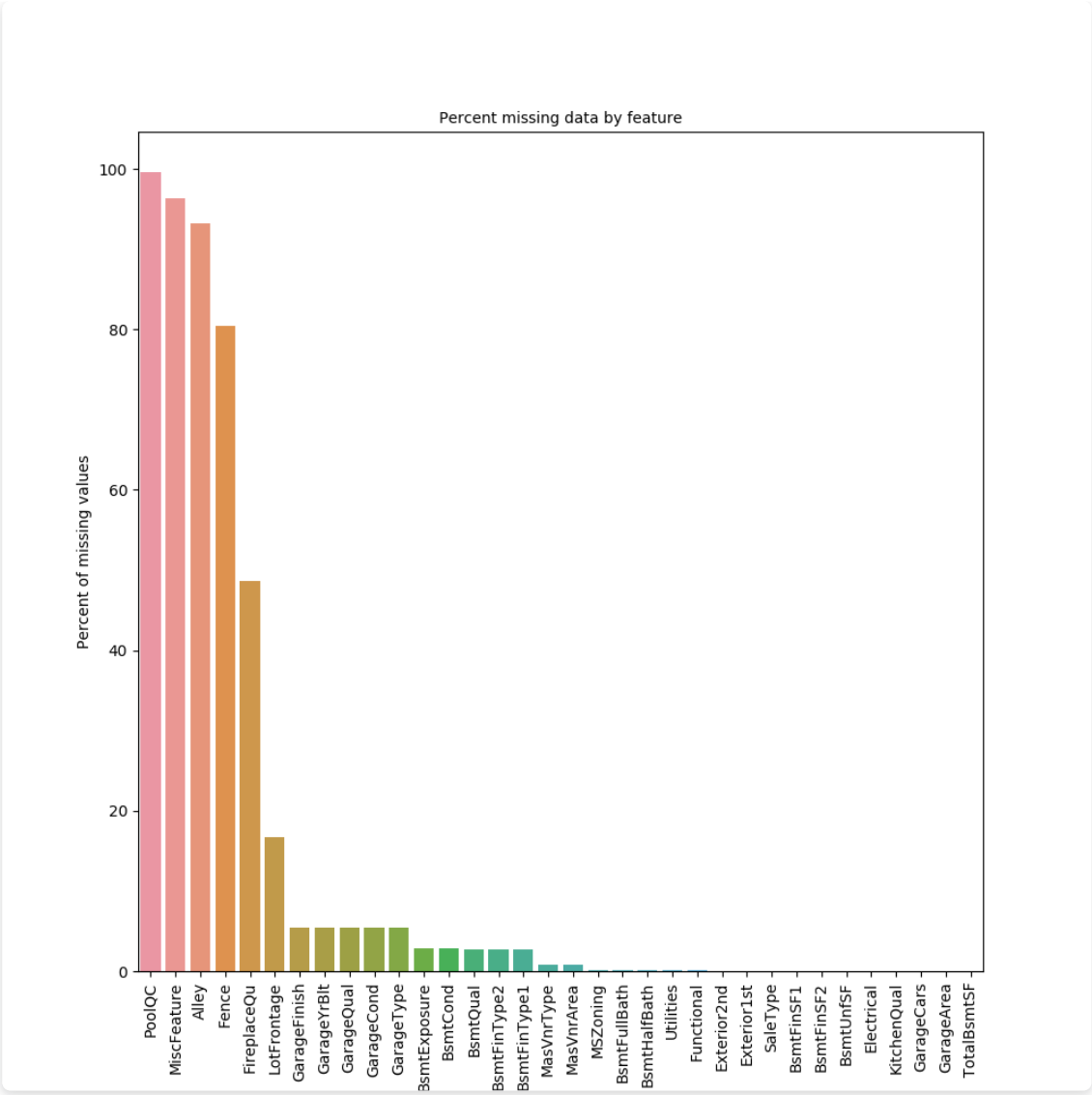
数据处理

数据来自于kaggle平台，数据包含与房价有关的79个特征，其中训练集还包含房价的价格，训练集包含1460个数据，测试集包含1459个数据。

这79个特征中既有数值型数据也有字符串描述的类型值。并且数据中存在缺失值。

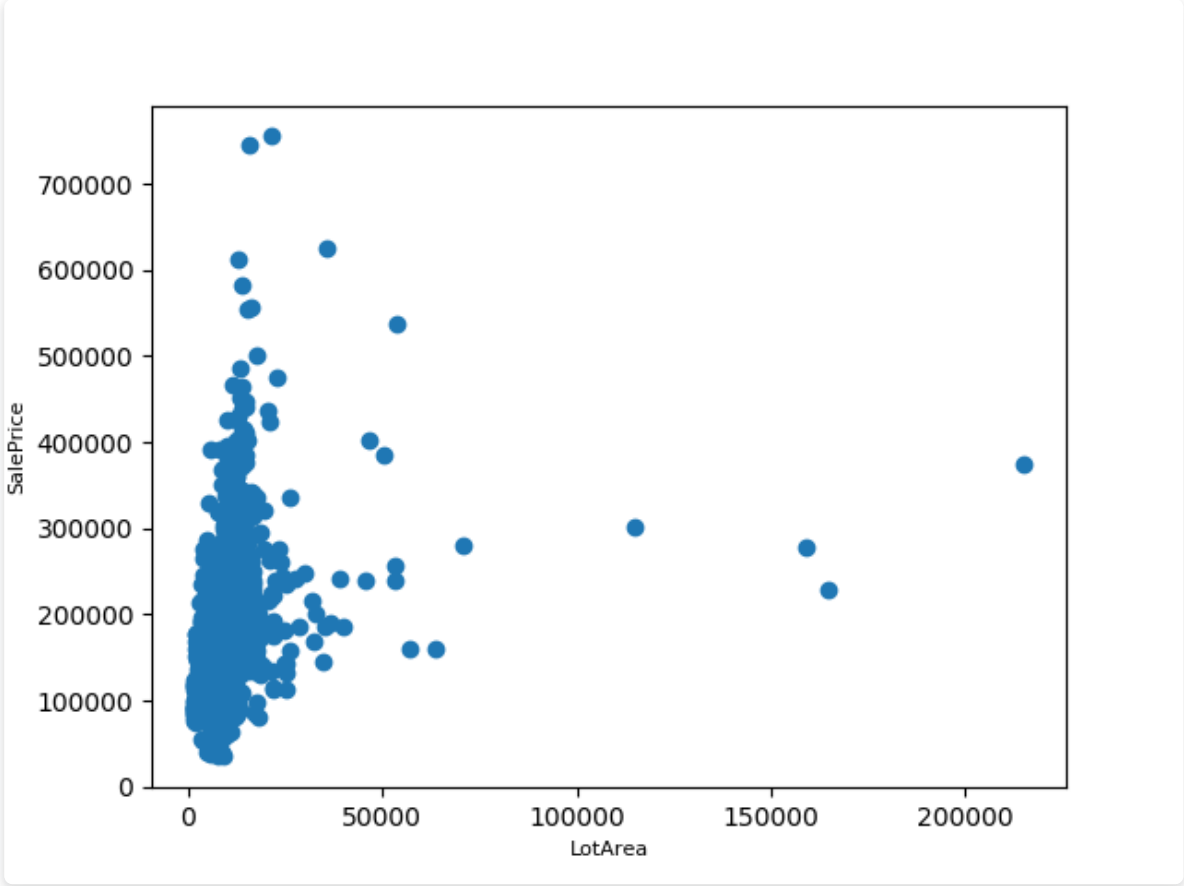
对数据进行预处理的步骤如下：

- 去除缺失值多的特征（缺失率），具体为去除缺失率大于10%的特征。

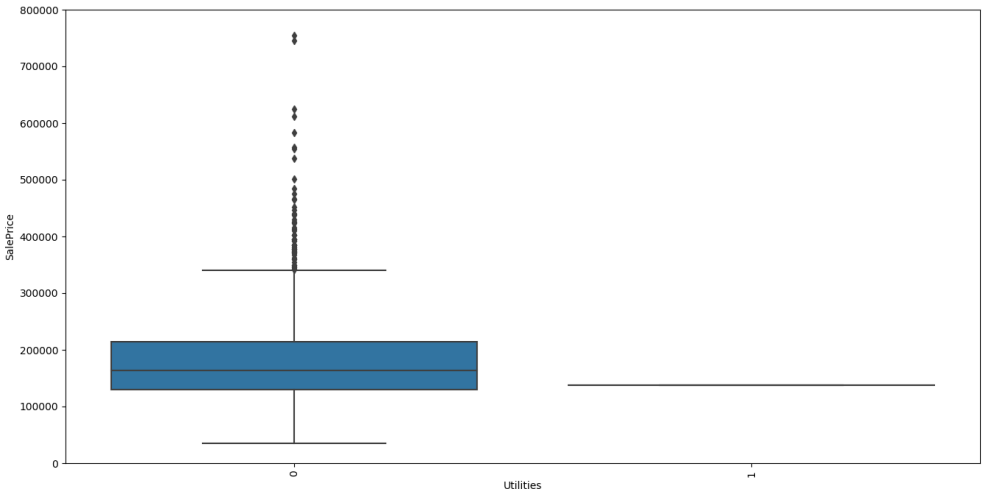


缺失值统计

- 对读入数据进行类型转换。具体为将读入为数字型的却描述类别的特征转换为字符串型，（如MSSubClass建筑物类型特征等），为后续类型值编码做准备；以及将读入int64型数值转化为float64。
- 去除异常值，孤立点。下图展示了索要去除的部分异常值。

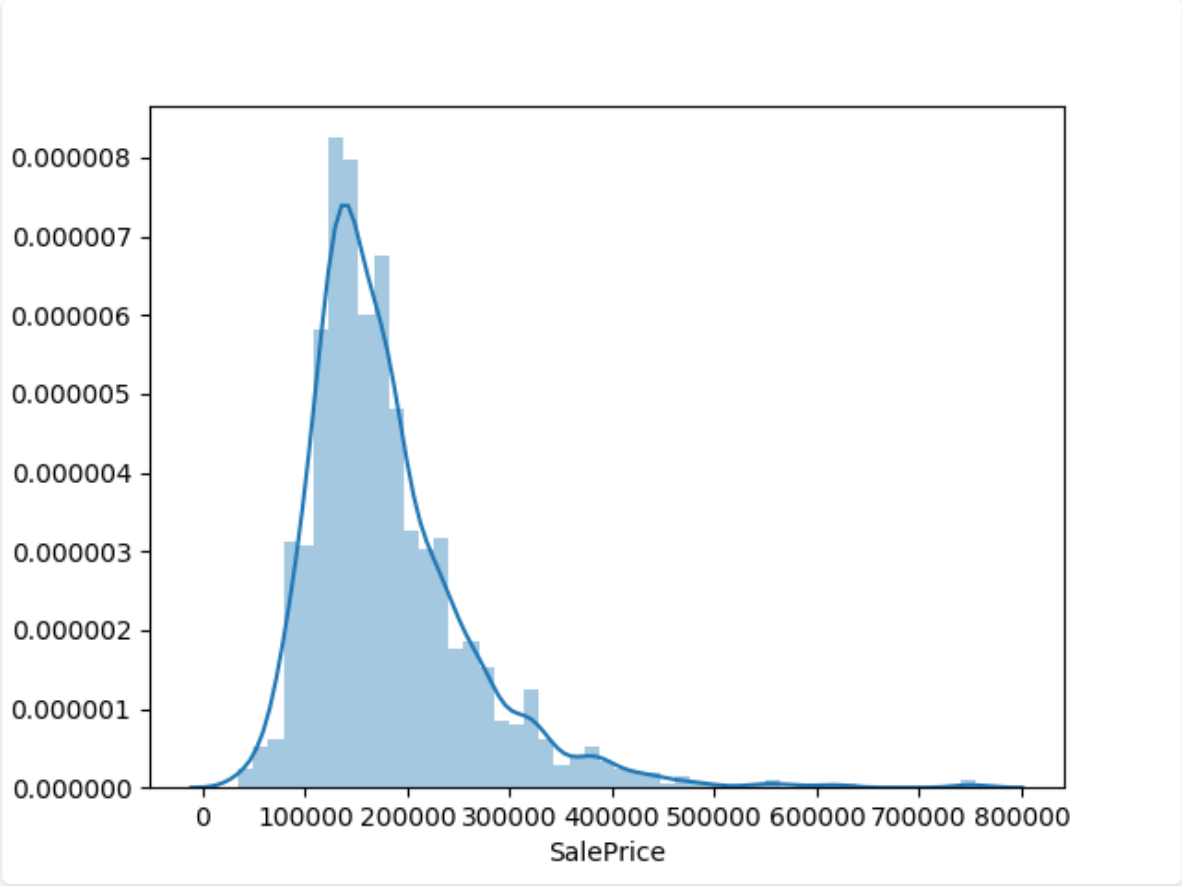


LotArea异常值(大于60000部分)



Utilities异常性质(仅有一个样本与其他样本不同)

- 对缺失值进行填充，类型值缺失填充众数，数字值缺失填充平均数。
- 对类型值采用one - hot编码。
- 对数字值进行归一化。
- 由于价格的差值比较大，使用函数np.log1p()对训练集价格进行数据平滑处理。使其更好的服从高斯分布，便于后续训练和预测。这样最后的预测结果使用函数np.expm1()进行还原。



价格分布情况

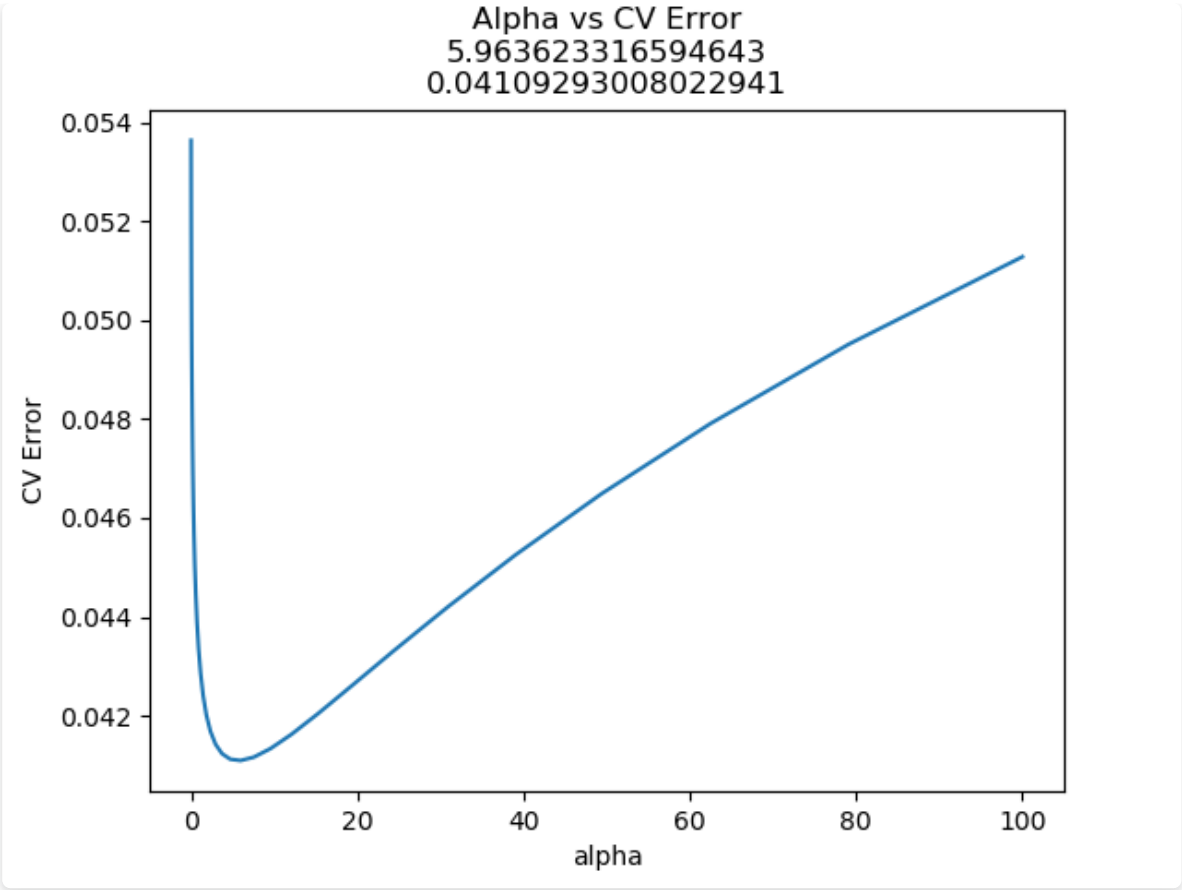
一些试探

做出相关系数矩阵，本意想去除与房价关联度低的特征，后来发现取出后训练出来的模型不如没有训练，故放弃。

在没有采用集成学习前，我仅仅使用了岭回归模型，最后训练集上的均方误差在0.04左右，后来采用六个模型的集成学习在训练集上的均方误差在0.03左右。

模型调参

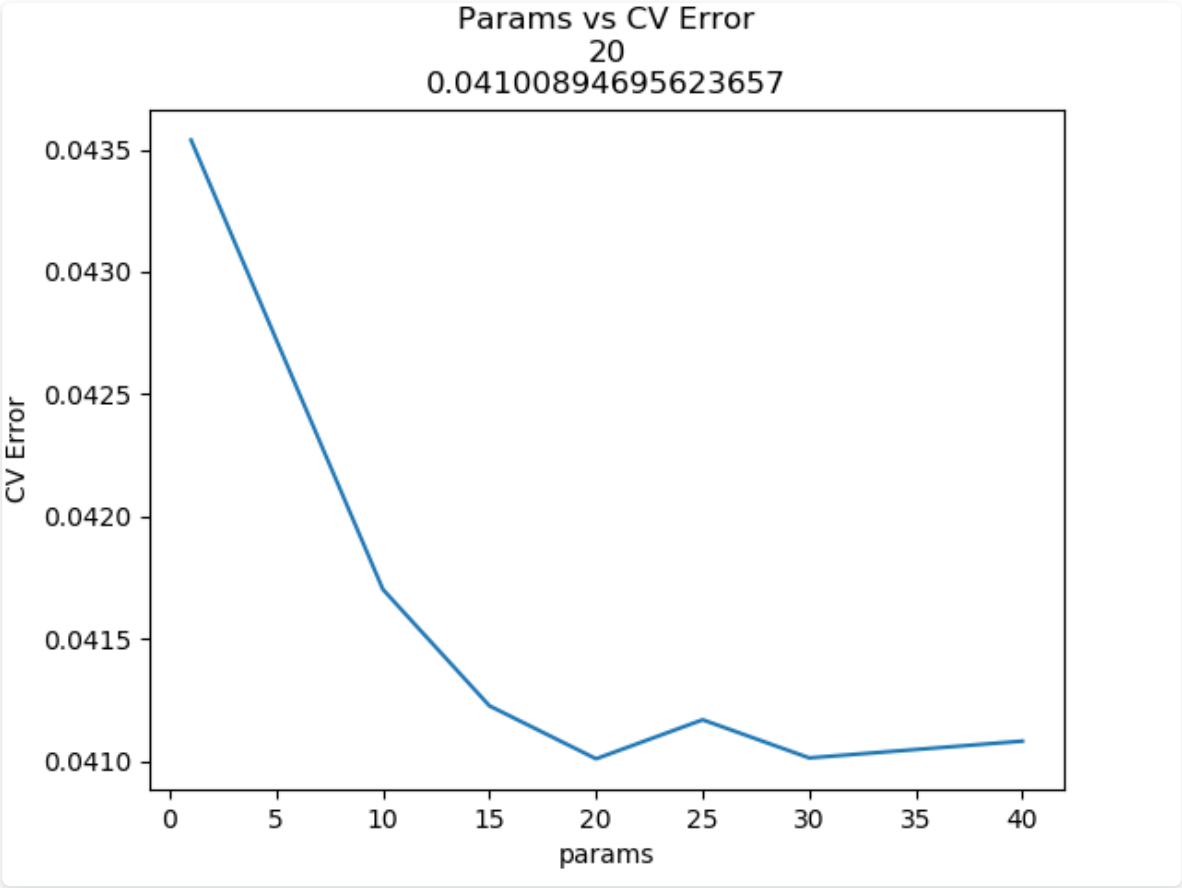
岭回归模型



岭回归模型调参

由上图发现，采用岭回归模型的最优的alpha值为5.96。

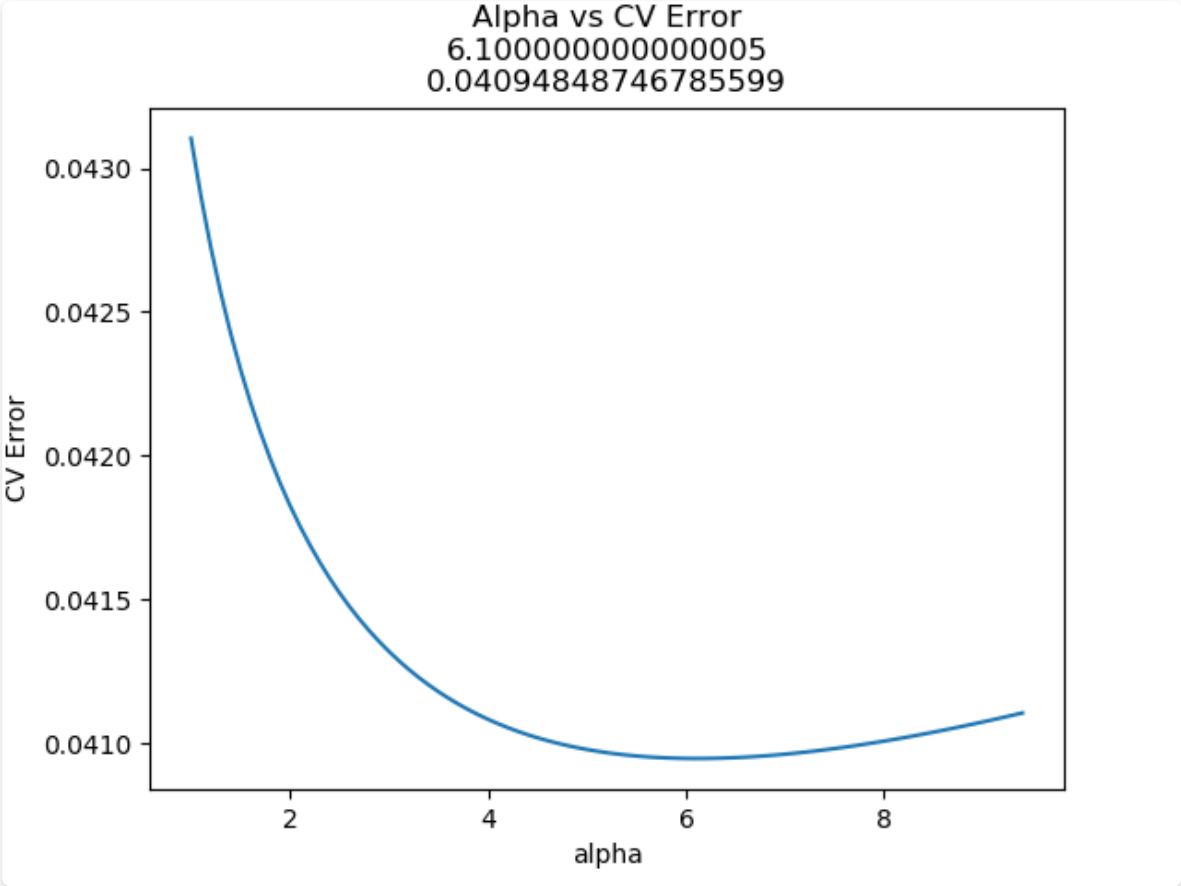
袋装法



袋装法调参

由上图发现，袋装法最优的n_estimator值为20。

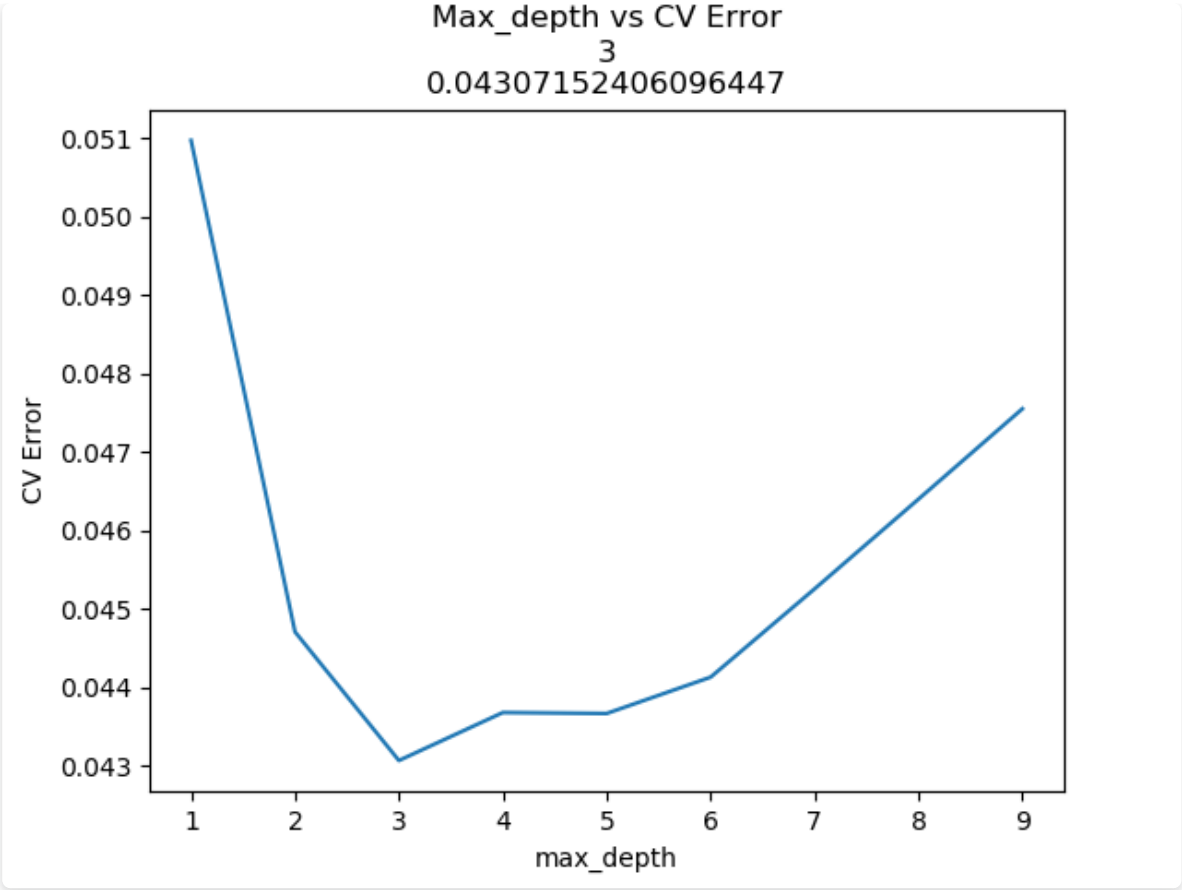
核函数岭回归模型



核函数岭回归模型调参

由上图发现，采用核函数岭回归模型的最优的alpha值为6.1。

XGB增强树模型

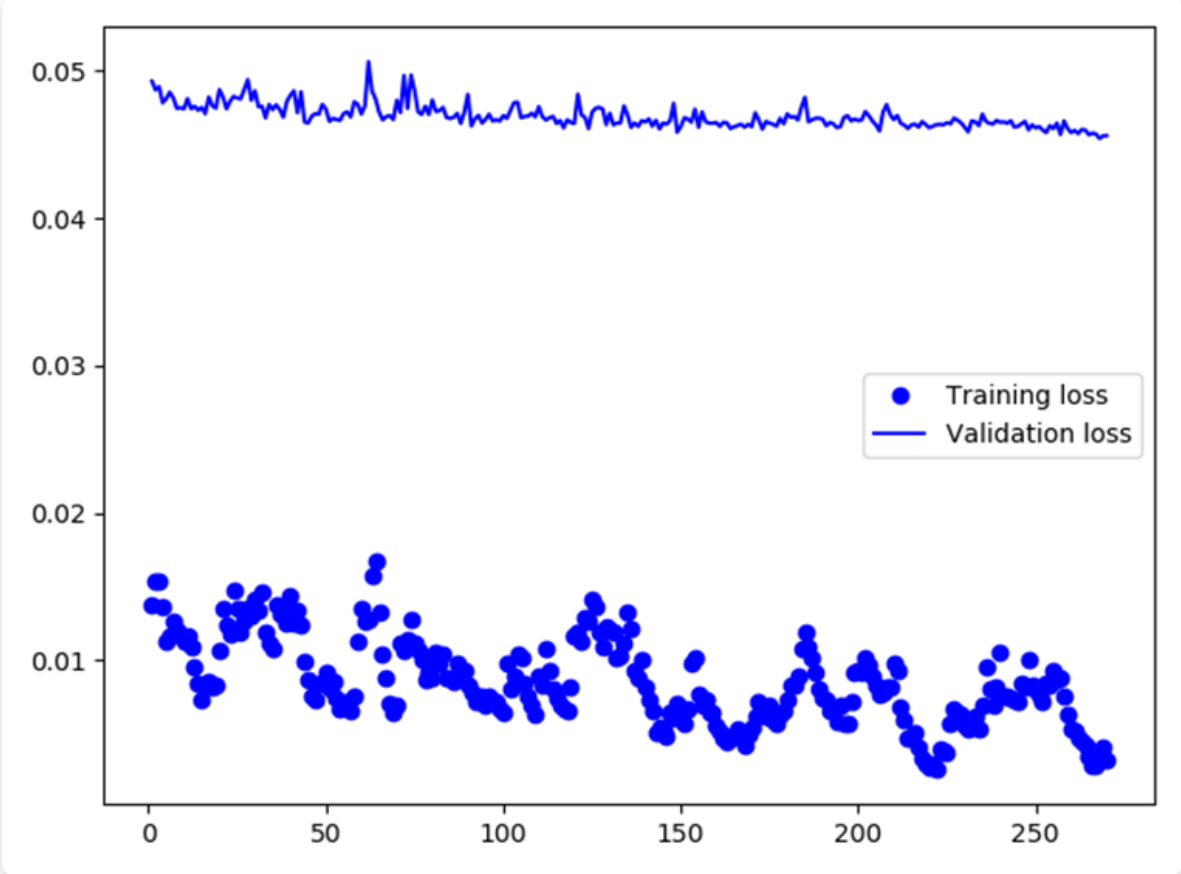


增强树模型调参

由上图发现，采用增强树模型的最优的max_depth值为3。

深度神经网络模型

采用序贯模型（Sequential）搭建7层神经网络，网络与网络之间均采用全连接，激活函数采用relu，采用均方误差作为损失函数。



训练轮数与损失函数变化趋势

模型训练300轮趋于稳定。

集合模型

除了深度神经网络模型均方误差在0.048左右，其他模型的均方误差均为0.041，所以在集合模型中神经网络模型的权重要小一些。前5个模型权重取0.18，神经网络模型权重取0.1进行模型融合。

结果

挖掘结果以.csv文件附后。

排名情况：

#	Team Name	Notebook	Team Members	Score	Entries	Last
1	DiegoJohnson			0.00287	35	43m
2	ajitrajurkar			0.07347	13	1mo
3	Joselyn			0.07728	10	1mo
4	servietsky			0.09919	181	9h
5	wmy_kaggle			0.09935	48	1mo
6	Patrick Bruecker			0.09953	1	18d
7	bradds			0.10069	41	1mo
8	Uncle Red			0.10115	65	3d
9	Güneş Evitan			0.10159	11	1d
2137	train toy			0.13329	5	-10s

Your Best Entry

Your submission scored 0.13352, which is not an improvement of your best score. Keep trying!