上海科技大学 本科生

（2018 年-2019 第 1 学期）

《数学建模》课程小论文

姓　　名	郑钧仁 念中林 蒋承越 曾昊 李屏天
学　　号	96904603 24911192 98917846 18245234 15451961
论文题目	基于弹性网络回归等多个机器学习模型对比分析的爱荷华州房价预测
参考文献题目	[1]如何在 Kaggle 首战中进入前 10% [2]Kaggle 首战拿银总结
最便捷联系电话	18800295897 18800295753 15335193100 18001623375 15002127692
Email	zhengjr@shanghaitech.edu.cn nianzhl@shanghaitech.edu.cn jiangchy@shanghaitech.edu.cn zenghao@shanghaitech.edu.cn lipt@shanghaitech.edu.cn
指导教师	肖柳青

数学建模课程项目报告

论文题目:
基于弹性网络回归等多个机器学习模型对比分析的爱荷华州房价预测

组员信息(姓名-学号-专业):

郑钧仁 96904603 计算机科学
念中林 24911192 计算机科学
蒋承越 98917846 计算机科学
曾 昊 18245234 计算机科学
李屏天 15451961 计算机科学

摘要:

本次数据来源于 kaggle 比赛，已经为我们提供了美国衣阿华州 xxx 份房价样本资料，其中包括房子的价格，面积，游泳池面积等已经为参赛者提供好的 79 个特征信息。由于原始数据存在许多 NaN 以及不利于机器学习的特征，所以需要在原始数据的基础上进行特征工程发掘深层信息，并利用修改后的特征进行建模与预测，在经过对比 Ridge, Lasso, Random Forest, SVM, ensamble learning 以及 Elastic net 等模型的结果之后，确定 Elastic net 拥有最好的预测效果。

关键词: 数据清洗，特征工程，回归，机器学习，模型选择

正文

一、介绍

Kaggle 主要是为开发商和数据科学家提供举办机器学习竞赛、托管数据库、编写和分享代码的平台。这次比赛我们选择了美国衣阿华州房屋价格预测作为挑战。在所提供的数据中包括了 79 个已经为参赛者设置好的特征信息，房屋价格，室内面积，与路面的距离和地下室面积等，期中房屋价格是比赛所要预测的目标。房屋价格的预测可以帮助房地产经纪人制定计划和确定价格，所以也帮助了他们的客户寻找最合适的出售价格，除此之外，房地产公司也会利用房屋价格与房屋特征之间的关系来确定用户需求并实现自身利益的最大化。

这次项目的流程主要分为三部分：

1. 特征工程
2. 模型选择
3. 模型评估

由于收集到的数据并不是完全有效的，在进行学习之前需要对数据进行清洗，去除或填补缺失值（NaN），但是由于信息缺失是由多种原因造成的，简单的使用中位数或者众数对所有的信息进行填补会影响信息的有效性以及机器学习的效果，所以针对不同类型的数值缺失，我们提出了不同的解决方案，例如线性回归预测以及众数填补等。除此之外，并非所有提供的原始特征都适合于机器学习模型，例如游泳池面积这一特征，面积为 0 代表房屋不配备游泳池，并且从常识上去理解，一个房屋是否有游泳池会比有游泳池的情况下泳池面积大小更能影响房屋价格，所以如果把泳池面积特征替换为是否有游泳池，那么在机器学习的时候会产生更好的结果。

在数据处理完成之后，需要将数据带入模型中进行训练，由于我们使用了 python 的 sklearn 库，这降低了我们建立模型所要求的代码量，同时也允许我们使用多种模型进行对比并获得最优模型。对于每个模型中超参的调整我们采用了网格搜索，避免了手动调参来提升参数的全面于可靠性。

二、数据预处理与特征工程

2.1 NaN 统计数据

经过对数据的统计，在提供的原始数据中有较多的 NaN 出现，在进行进一步处理之前，需要先将 NaN 替换为可处理的数值或者直接剔除无效样本。

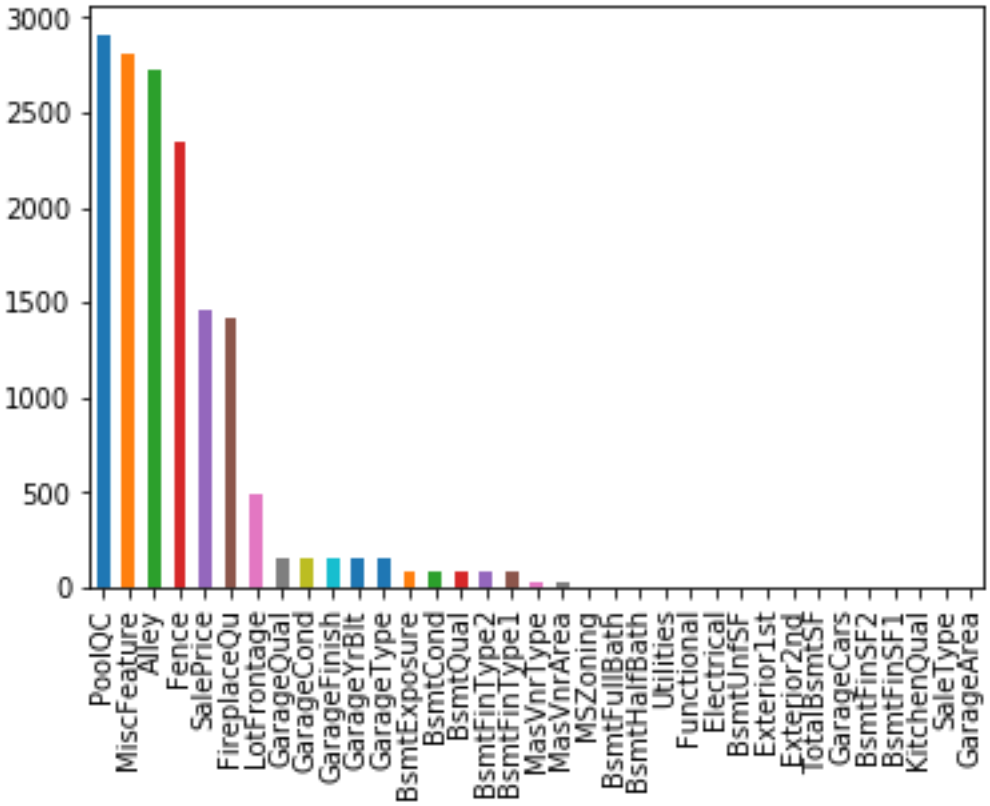


Figure 2.1: 对不同 Feature 样本数据的 NaN 数量统计

在 Figure2.1 中可以看到 PoolQZ，MiscFeature，Alley 等特征有较多 NaN 出现，如果不加处理将会对模型最后的效果造成很大的影响。

2.2 有实际意义的 NaN 处理

在数据描述文件中，可以发现并非所有的 NaN 都是由于数值缺失，部分的 NaN 代表房屋不存在该种特征。比如车库特征中的 NaN 代表房屋不配备车库，而非缺失车库信息，单一的处理方式会导致训练数据更片面，因此对于范畴特征，我们将 NaN 替换为“None”字符串，对于数值特征，我们将 NaN 替换为 0。

在所有的特征中，LotFrontage（房屋空地面积）特征的处理最为复杂，当然如果为了简化处理过程，可以选择用均值或者中值进行替换，但是为了保证预测效果，我们选择简单的调用 sklearn 库并利用除了价格以外的相关特征对 LotFrontage 进行预

测。从常识上去理解，房屋占地的形状和大小是与房屋空地面积是较为相关的特征，所以我们选择这两个特征对 LotFrontage 进行回归预测，以填补缺失信息。

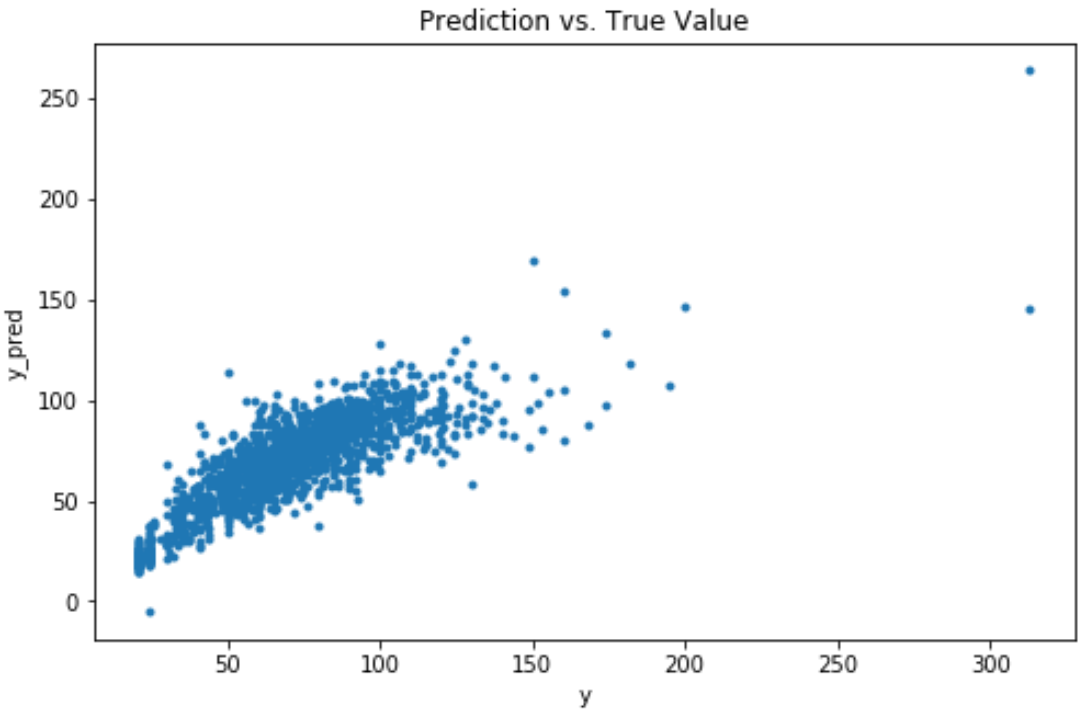


Fig 2.2: 基于 Ridge 模型对房屋空地面积的预测和真实值的对比

在这张图中数据点越多分布在对角线上代表预测效果越好，在图 2.2 中，我们可以发现大多数点都分布在对角线上，存在少量误差，但整体趋势处于可接受的范围内，因此我们选择用预测结果填补缺失数值。

2.3 特征工程

在处理完所有 NaN 之后，我们开始着手准备探索特征更深一层的信息并对其进行分析，以使其转换为适合于模型拟合的格式。

2.3.1 Basement Finish Type（地下室完成类型）

在地下室特征中有两种区域表示某种完成类型。一种是“高质量生活区域”，另一种是“未完成区域”。然而地下室的完成类型和它的面积在不同的列中，并且对于一个普通的模型来说它并不一定会意识到这两列特征之间的相关性，因此，我们删除了原始的特征列，并创建了新的特征列，期中每一列表示某一种区域所占的面积，如果不存在这一完成类型的区域则将其数值设为 0。

进一步分析，在日常生活中，买家并不会关注地下室的绝对面积大小，而是更

对关注在相对于房屋面积，地下室所占的比例，也就是 $\frac{\text{BasementArea}}{\text{LotArea}}$，因此对于不同完成类型的地下室，我们又增添了其所占房屋面积的比例为新特征列加入我们的特征集合。

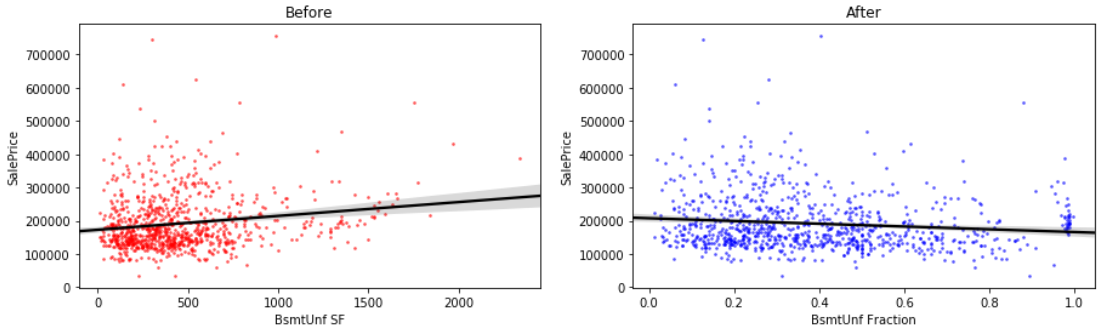


Fig 2.3.1: Before 展示了地下室面积于房价之间的关系
After 展示了地下室面积与房屋面积的比例和房价之间的关系

在图 2.3.1 的左图中可以观察到，有很多数据偏离了拟合的直线，也就说明地下室面积与房价之间并没有非常明显的相关性，但是在将数据换位地下室占房屋总面积的比例后，可以观察到地下室所占的比例与房价之间有较为明显的负相关性，这也可以理解，因为占总面积更多的地下室更有可能会有未完工，而未完工的部分会导致房价的下降。

2.3.2 一层二层面积

如同地下室，我们加入了每一层面积占总面积的比例作为新的特征输入进行训练。同时我们也增加了（整个资产面积-室内面积和生活用面积）等可能会更为被购买者关注的特征。

2.3.3 范畴特征

一部分的范畴特征可以直接转化为数值特征，比如说表示某一事物质量的范畴特征，它们一般包括从 “Poor” 到 “Excellent”，我们将其对应到相应的数值上，我们提供如下的映射方式：

Ex	Gd	TA	Fa	Po	None
5	4	3	2	1	0

2.4 处理零值

之前我们用 0 值填充过并不关键的特征值，现在我们需要对它们进行更深一步的分析，如果在一个特征中只有少量的非 0 值样本，那么在模型拟合时将会使得模型产生偏向性，即将含有大量 0 值样本的特征替换为“01”类型的样本会使得模型更加灵活的处理特征，这种新加入的变量我们称为“哑变量”。如果非 0 值的样本数量并不可以被忽略，那么我们仅仅加入新的“哑变量”来处理，并不删除原始特征列。

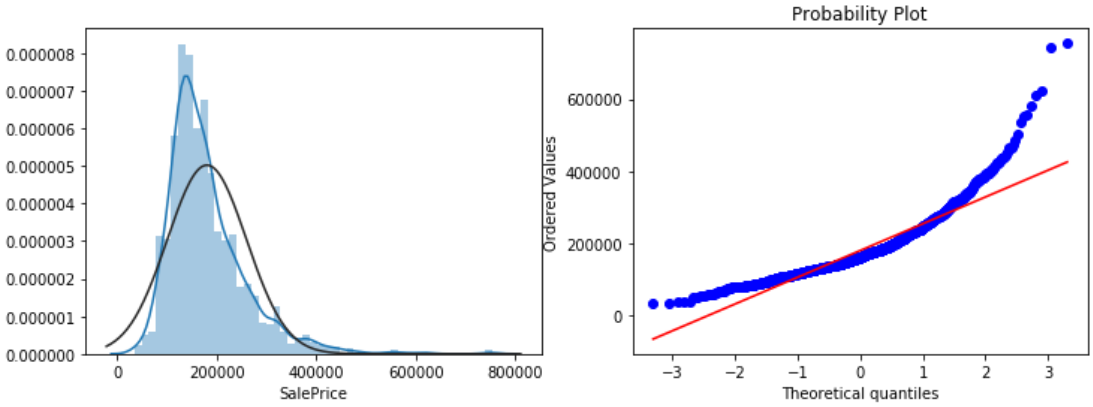
在我们的原始特征中，PoolArea（泳池面积）有 0.995546 的部分都是 0 值，我们在此直接将 PoolArea 替换为是“否有泳池”，并以 bool 类型对其进行赋值。而像 LowQualFinFrac 的特征，有 0.986297 的部分为 0 值，我们在此只加入了新的呀变量，并没有删除原始特征列。

	frac_zeros	n_unique	xs_zeros
LowQualFinFrac	0.986297	39	0.960656
LowQualFinSF	0.986297	36	0.958519
3SsnPorch	0.987324	31	0.955066
MiscVal	0.964714	38	0.938398
PoolArea	0.995546	14	0.924118
BsmtLwQFrac	0.917437	225	0.912993
BsmtBLQFrac	0.885235	322	0.882129
BsmtBLQSF	0.885235	283	0.881701
BsmtRecFrac	0.865365	362	0.862602
BsmtRecSF	0.865365	302	0.862054

Table 2.4: 部分特征 0 值所占比例

2.5 房价分布分析

由于我们需要预测的特征就是房价，下面我们先简单的可视化房价信息，并试图从中提取可用信息。从左图中可以看出房价分布服从了长尾分布，对原始数据进行 Log 处理会使其更接近正态分布，而由于正态分布我们较为熟悉，所以在此之后的预



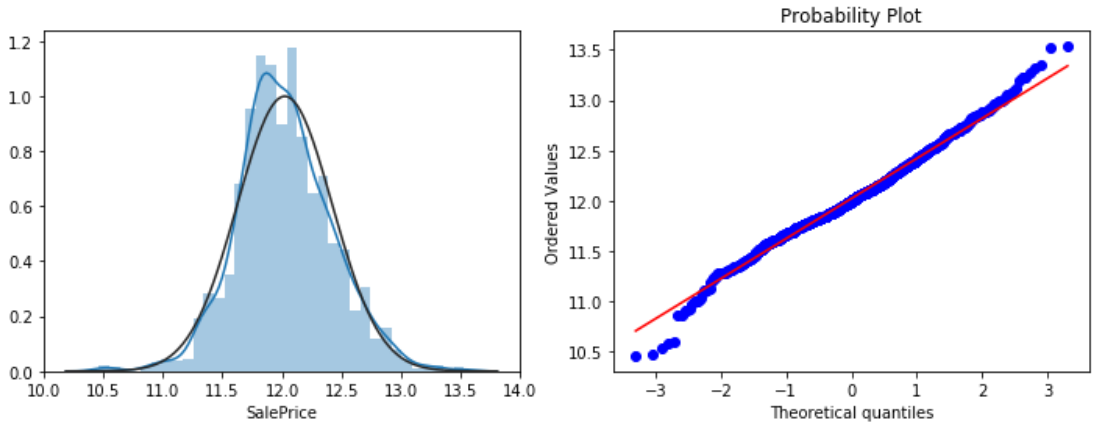


Fig 2.5-2：左图 log(Property Price) 右图 是否符合正态

测并非是直接对原始价格数据的预测，而是对价格信息 Log 处理之后的预测。

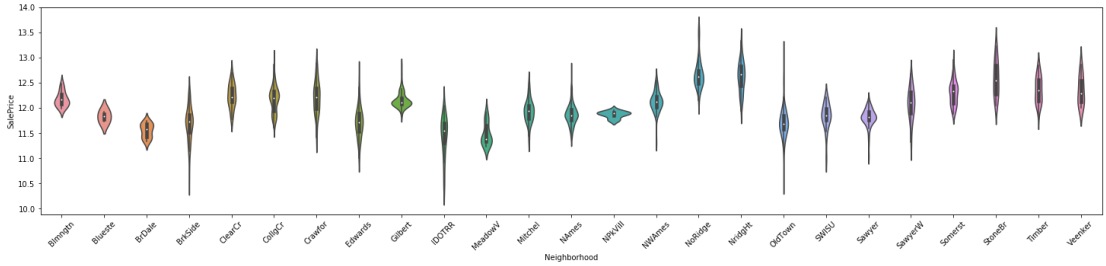
从 图 2.5-2 中可以看出，在对原始价格数据进行 Log Transform 之后，其分布更加接近正态分布。接下来所有的模型都将基于正态分布去建模与分析。

2.6 特征类型

对于不同的特征类型，在分析数据时应该有不同的处理方式，我们将所有的特征分为三种：

1. 数值离散特征：拥有少于 13 种数值的数值特征，如月份只有 12 个值。
2. 数值连续特征：拥有超过 13 种数值的数值特征，如年份，年龄等。
3. 范畴特征：剩余的非数值范畴特征。

首先我们分析了所有的范畴特征与房屋售价之间的关系，在这里我们使用提琴图来表示关联性。在提琴图中，提琴的轮廓代表了可能的数值，宽度代表了数值的常见程度，所以一个宽的提琴代表某一数值出现的次数较为集中，长的提琴代表这一特征中包含许多的数值。除此之外，提琴内层有两部分轮廓，最内层的轮廓表示了占总数值 50%的特征数值，外层轮廓代表占总数值 95%的特征数值。提琴图最中央的点代表了该特征的均值。



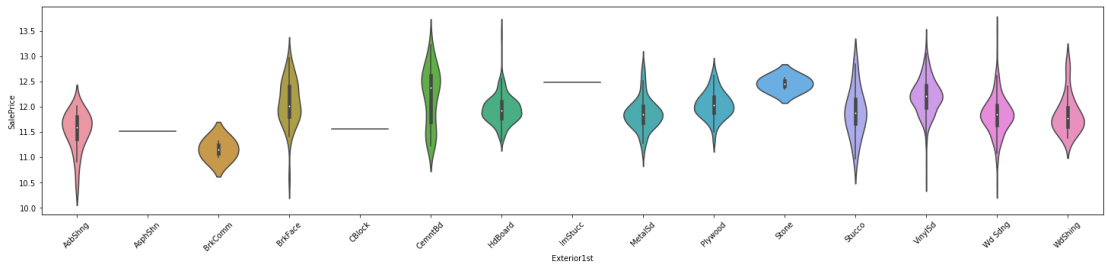


Fig 2.6.1：范畴特征提琴图表示

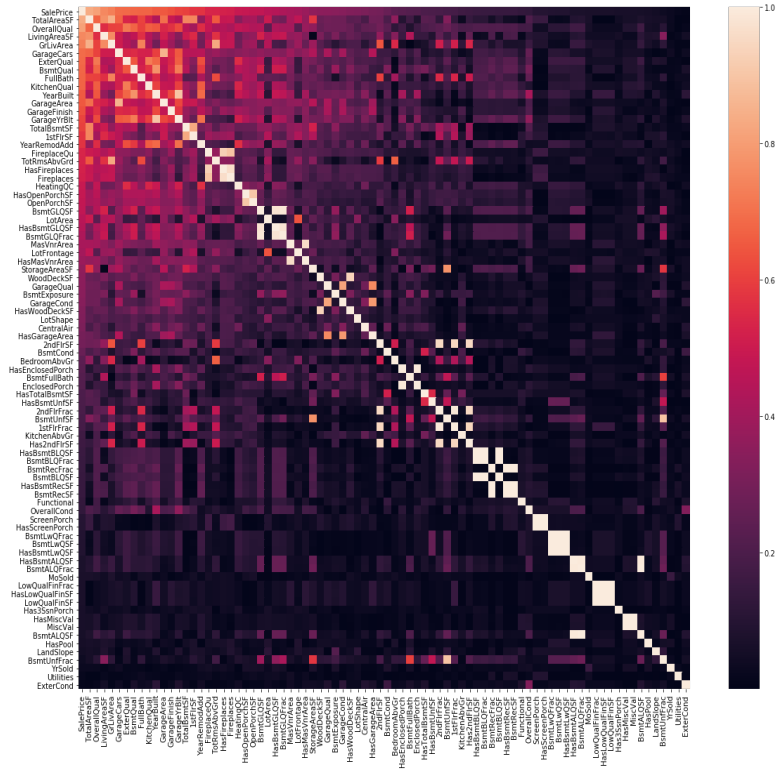
为了进一步确定每个范畴特征与房屋价格的关系，我们采用方差分析来量化其重要性。

Neighborhood 4.043304e-243	Fence 6.560319e-13
GarageType 8.427845e-125	RoofStyle 1.705740e-12
MSSubClass 5.096023e-115	BldgType 3.436794e-12
Foundation 1.350671e-111	Condition1 1.173444e-10
MasVnrType 1.670061e-65	Alley 1.358108e-09
MSZoning 1.021343e-59	Heating 2.484312e-09
Exterior1st 1.053128e-53	LandContou 3.086224e-08
Exterior2nd 2.429938e-49	LotConfig 6.214575e-07
SaleCondition 1.689708e-41	RoofMatl 4.504239e-04
SaleType 5.497893e-36	MiscFeature 6.059928e-03
PavedDrive 1.090995e-31	Condition2 1.382042e-02
Electrical 6.081144e-31	Street 2.837931e-02
HouseStyle 1.636077e-30	dtype: float64

在对范畴特征分析完之后，我们同样对离散数值变量分别做了方差分析与提琴图。接下来我们分析了数值特征之间的相关性，在此我们使用较为常用的热力图来可视化。

可以看到，在图中颜色越鲜艳的部分相关性越明显，可以看出来有很多特征都具有明显的相关性，比如 SalePrice，TotalAreaSF，OverallQual 等特征。我们之前创建的哑变量与其对应的原始特征有着明显的相关性。然而具有共线性特征的数据会使

得我们的模型难以确定哪些参数是真正重要的。理论上，将部分只能提供少量信息的



特征剔除会为我们的模型带来一定的增益，但是在这次项目中我们并没有进一步探索什么样的特征应当被忽略，在有闲暇时间的时候我们可能会填补在这里留下的问题。

Fig 2.6.2：数值特征相关性热力图

在观察完所有的数据特征之后，我们将所有的数值特征单位化，防止模型偏向于数值较大的特征而非正真重要的特征。

由于我们的模型建立在所有的数据都在一定程度上服从正态分布的基础上，对于其他的变量我们试图也通过寻找类似于 Log Transform 的方法使得特征信息在变换之后更加符合正态分布。在这里我们提出了除 Log Transform 以外的另一种变换方：Box-Cox Transform。Box-Cox Transform 可以一定程度上减小不可观测的误差和预测变量的相关性。具体细节不在这里赘述，通过比较 Log Transform 与正态分布的接近程度来选择最优的变换方式。

2.7 去除 Outlier

到现在我们基本上已经可以进行模型拟合了，只需要将数据中的 outlier 去除即可。这一步我们先拟合了一个线性 Ridge 回归模型，它已经为我们提供了足够好的结果，足以用来识别偏离常规的数据点。

我们将收集数据点与预测点之间的残差，并计算其标准差，从数据点中去除残

差超过 3 倍残差标准差的数据，以避免其对数据正确性与合理性的影响。

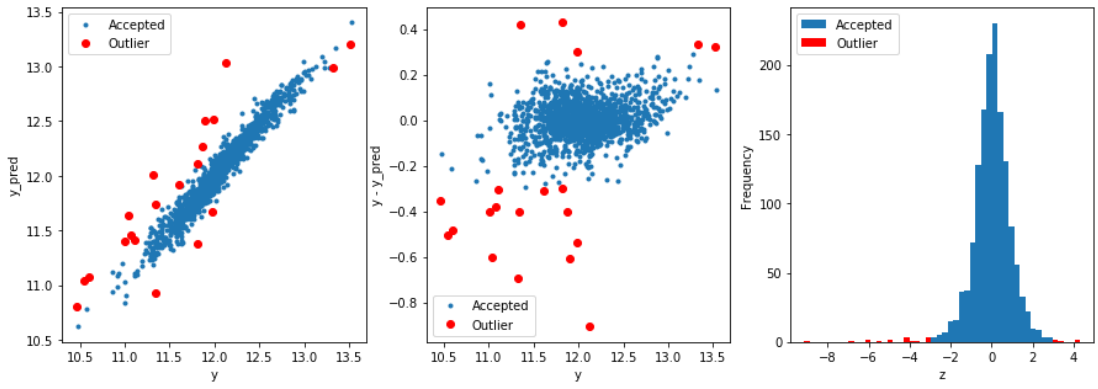


Fig 2.7: 图中红点为 Outlier

三、模型拟合与优化

为了防止模型过拟合，我们使用了 5 层交叉验证来进行参数优化以及模型选择。也就是说，我们每次会用数据中百分之 80 的数据作为训练集，另外百分之 20 的数据作为验证集，而最后我们会使用 Kaggle 的官方测试集作为我们最终结果的评估。由于我们有过多模型去拟合，调参的工作我们并没有手动处理，而是用较为常见的网格搜索去自适应的寻找最优参数，我们所需要做的只是为参数提供一个范围。

3.1 Linear Regression

由于预测的值是在连续空间中，所以很自然地我们决定将其建模为回归模型，而其中最常见也是最常用的就是线性回归模型。我们的假设集如下所示，我们认为经过特征工程之后提取到的特征，通过线性组合影响了最终的房价结果，所以我们把线性回归模型及其改进和变体作为我们的首要尝试。如下公式所示：

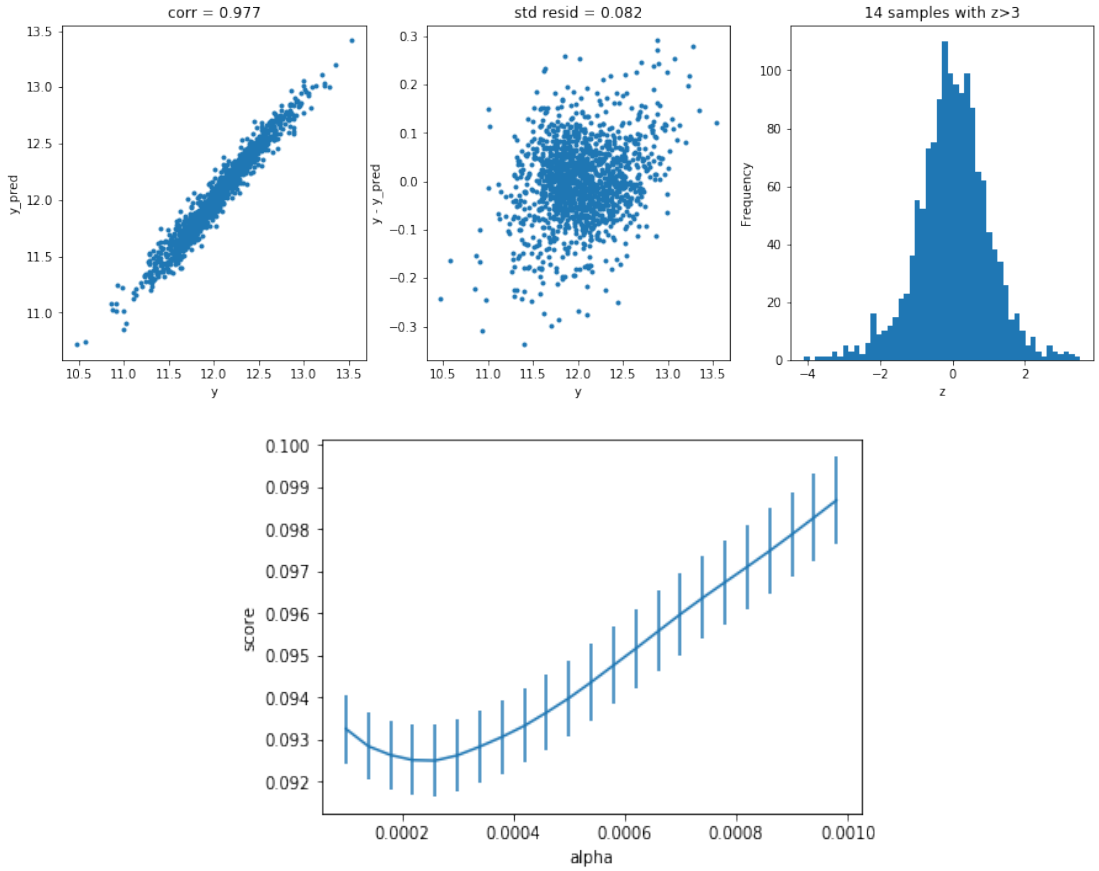
$$h_\theta(x) = \theta^T X = \theta_0 x_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_n x_n$$

我们使用了并且评估了三种线性回归模型变体的表现，他们分别是 Ridge、Lasso 和 ElasticNet。因为所有的模型都可以通过调用强大的 sklearn 实现，在此我们省去模型具体的分析，而是只提供其数学表达。

3.1.1 Ridge

使用正则化系数 w 来惩罚权重，以此控制方差和偏差。此处对于正则化系数，Ridge 模型中将使用二范数。

$$\mathrm{Err}_{\mathrm{ridge}}\left(w, w_0\right)=\sum_{i=1}^{n}\left(wx^{(i)}+w_0-y^{(i)}\right)^2+\lambda\left\|w\right\|_2^2$$



$$w_{\mathrm{ridge}}, w_{0\mathrm{ridge}}=\arg\min_{w,w0}\sum_{i=1}^{n}\left(wx^{(i)}+w_0-y^{(i)}\right)^2,\text{subject to}\left\|w\right\|_2^2<\eta$$

Fig 3.1.1: 上方三张图代表数据的拟合效果

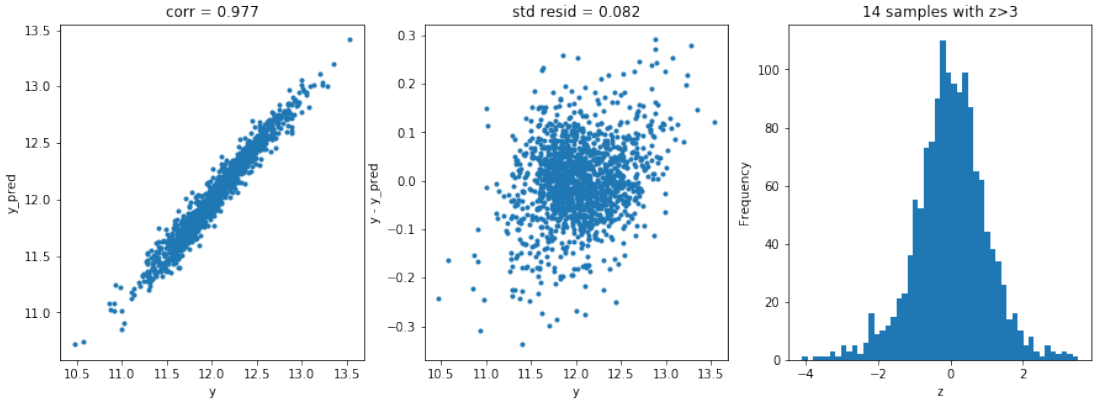
下方图片 代表不同参数条件下的拟合效果

3.1.2 Lasso

Lasso 模型中使用一范数。不同于 Ridge 模型，Lasso 会使得部分 w 值趋于 0，因此会导致结果的稀疏性。

$$\mathrm{Err}_{\mathrm{lasso}}\left(w, w_0\right)=\sum_{i=1}^{n}\left(wx^{(i)}+w_0-y^{(i)}\right)^2+\lambda\left\|w\right\|_1$$

$$w_{\mathrm{lasso}}, w_{0\mathrm{lasso}}=\arg\min_{w,w0}\sum_{i=1}^{n}\left(wx^{(i)}+w_0-y^{(i)}\right)^2,\text{subject to}\left\|w\right\|_1<\eta$$



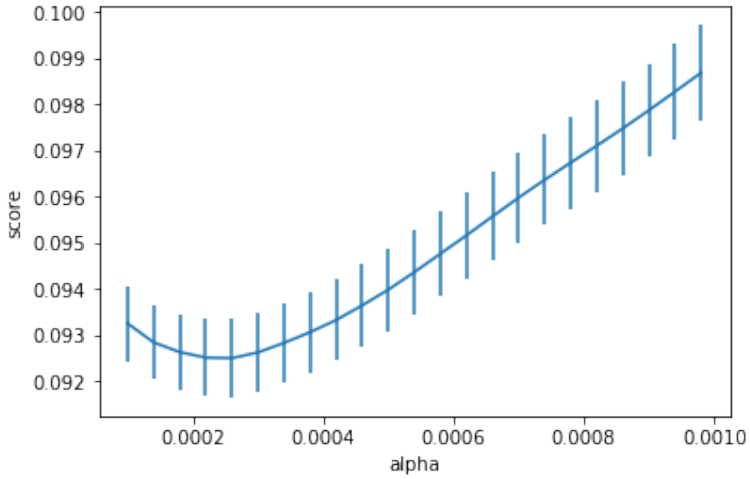


Fig 3.1.2: 上方三张图代表数据的拟合效果

下方图片 代表不同参数条件下的拟合效果

3.1.3 Elastic Net

公式：

$$\hat{\beta} = \mathrm{argmin}_{\beta}((y - X\beta)^2 + \lambda_2\beta^2 + \lambda_1\beta)$$

结合 L1，L2 的惩罚，我们就得到了弹性网络模型。它是 Lasso 和 Ridge 的结合，当遇到高相关性的特征时，Lasso 会随机在其中选择，ElasticNet 会在其中选择至少两个特征。我们一般在有多组相关特征时选择使用 ElasticNet。

3.1.4 回归结果

Model	**Cross Validation Score**
Lasso	0.092497

Ridge	0.094613
ElasticNet	0.091744

由于我们使用残差作为评分标准，所以分数越低，回归效果越优，在以上的三个模型中弹性网络模型获得了最优的回归效果。

3.2 模型对比

在将弹性网络的回归效果与其他回归模型对比过后，我们得到了一下结果。在图中可以看出最基本的 Ridge，Lasso 与 ElasticNet 有较理想的回归效果，普遍比基于树的回归模型有着更低的误差，Ensemble Learning 和 Random Forest 等复杂的算法在这里也无法媲美简单的线性回归模型，而 KNeighbors 由于是分类模型，用在预测问题上难免会有些力不从心，所以结果上看差强人意。

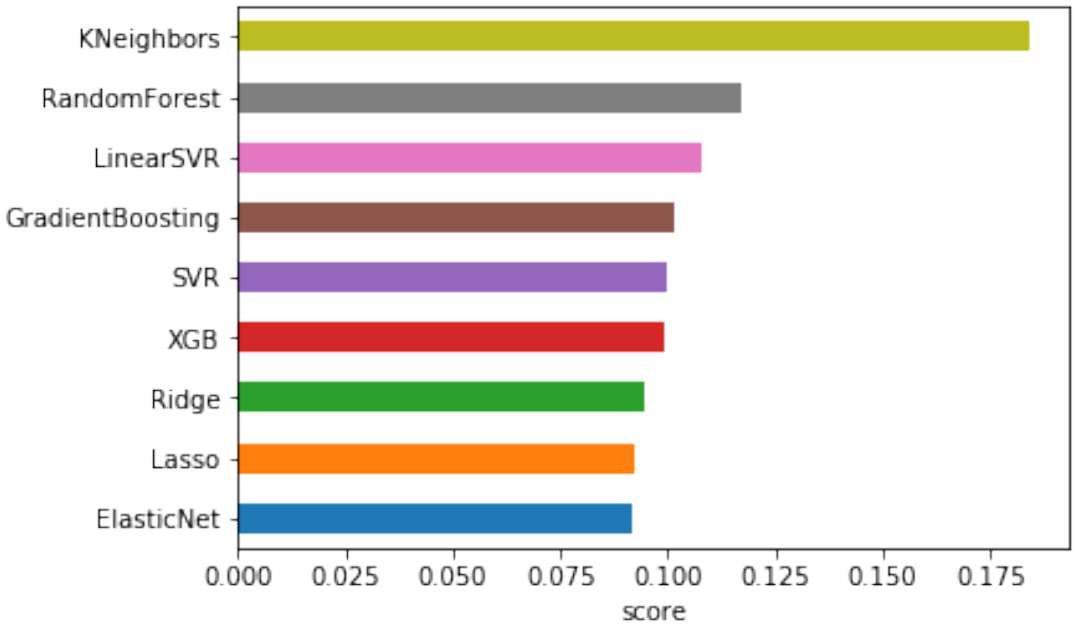


Fig 3.2: 模型预测结果对比

四、结论

在这次的房价预测比赛中，由于通过机器学习预测数据需要有可靠和精确的特征描述，我们的团队通过数据清洗，特征工程等途径对官方提供的特征数据进行深入的探索和优化改造，使得其更能体现问题的本质，在特征工程部分我们的分析基于生活中人们对房屋的直观感受，进一步间接的提升了机器学习的预测效果。整个模型都建立在收集的数据在一定程度上符合正态分布的基础上，所以对于原始数据我们并没

有直接进行处理，而是在对比不同变换方式后，选择可以将原始数据映射到正态分布的变换方式，并进一步探索。在评估不同的特征时，我们使用了提琴图，热力图等方式直观的获取数据特征并以此为基础进一步探索数据信息。

在数据处理完成后，我们对比了多个回归模型，并进一步确定线性回归比 SVM 和基于树结构的回归模型有更优的预测效果，其中 Elastic Net 拥有者最佳的预测效果。

在所有项目工程都完工之后，我们将代码提交到 kaggle，并用官方的测试集进行测试，获得了前 7%的优异成绩。

参考：

[1] Wille，如何在 Kaggle 首战中进入前 10%，
https://dnc1994.com/2016/04/rank-10-percent-in-first-kaggle-competition/
[2] Scarlet Pan，Kaggle 首战拿银总结 | 入门指导（长文、干货），
https://blog.csdn.net/jdbc/article/details/72468001