题目：

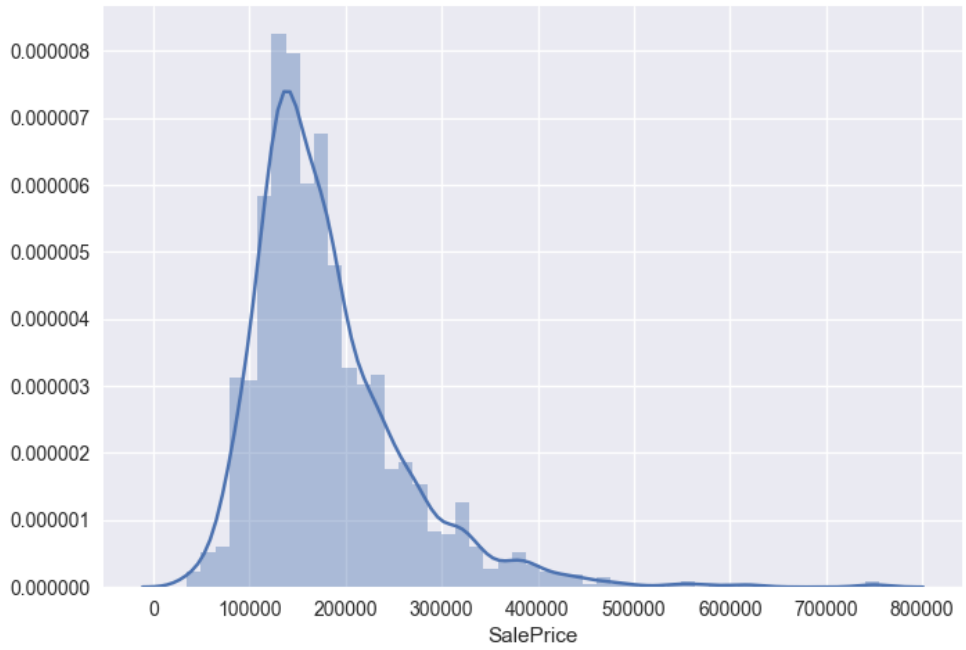
通过 79 个描述美国爱荷华州埃姆斯住宅房屋的不同变量，预测每栋房屋最终的售卖价格。

By 徐伟玲 2018-7-17 于 西安交通大学

一．可视化探索

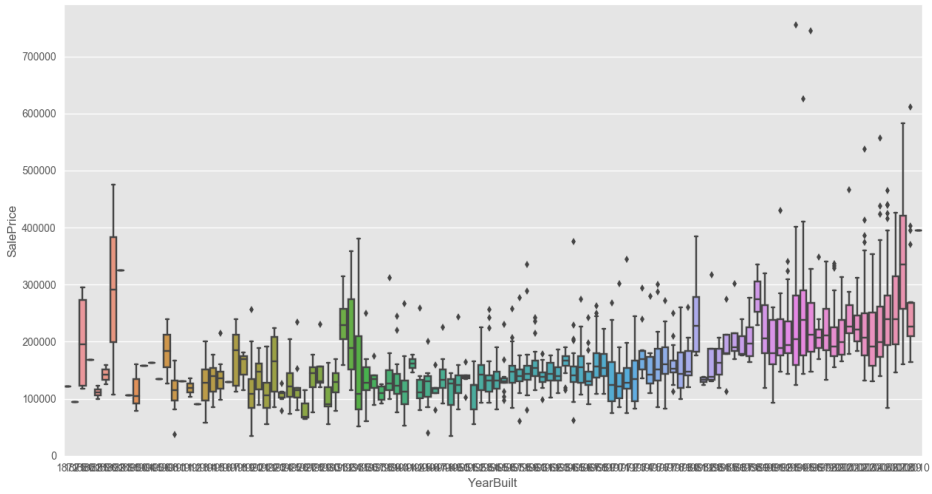
通过对给出的房屋特征进行了解，选择认为和 SalePrice 相关性大的特征进行探索。

1、 SalePrice 直方图



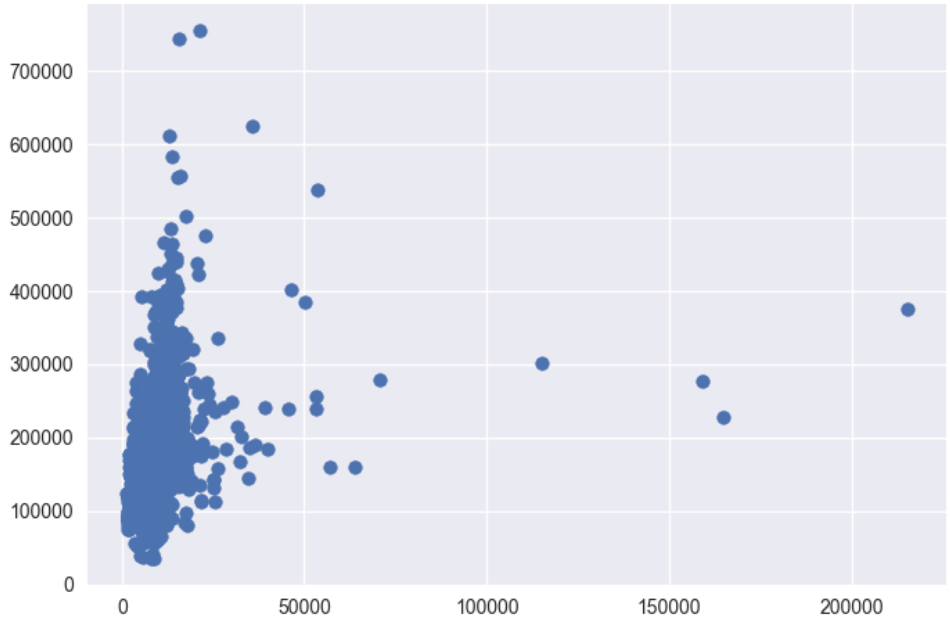
房价偏离正太分布（正偏），为了更好地在模型中拟合，需要对房价进行对数转换。其他的几个数值型特征也有这种情况，都可以进行相应的处理。

2、 YearBuilt-SalePrice 箱型图



一般认为新房子比较贵，老房子比较便宜，从图上看大致也是这个趋势，由于建造年份(YearBuilt) 这个特征存在较多的取值（从 1872 年到 2010 年），直接 one hot encoding 会造成过于稀疏的数据，并且年份和 SalePrice 有正比关系，因此在特征工程中会将其进行数字化编码（LabelEncoder）。

3、 LotArea-SalePrice 散点图

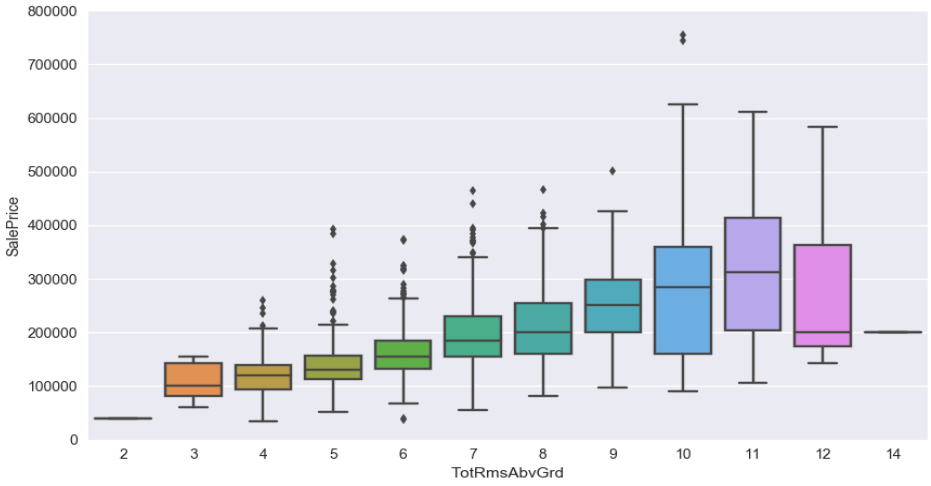


可以看出，SalePrice 和 LotArea 关系比较密切，基本呈线性关系，并且正相关。有几个点，当 LotArea 很大时，SalePrice 并没有很大，因为这个点对应的地块面积虽大，但是所属的区域不是昂贵的区域，因此房价偏低。

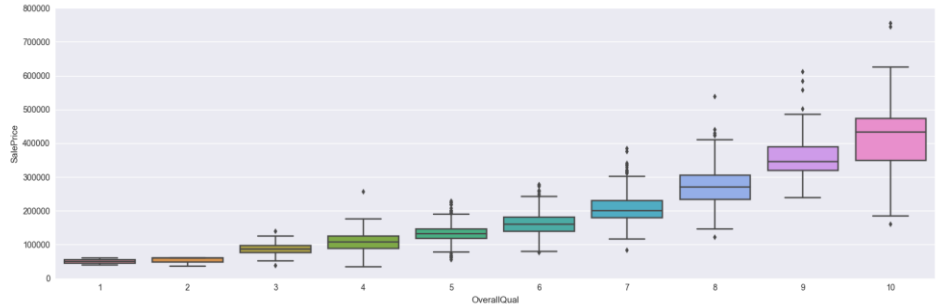
GrLivArea-SalePrice 散点图

由散点图可以看出，GrLivArea 和 SalePrice 之间也呈现正相关，但是趋势没有 LotArea 明显。在右下角有两个点，基本上不在两个变量的正相关辐射相关范围，认定它们是异常值，需要剔除。

4、 TotRmsAbvGrd-SalePrice（满足等级要求的房间总数）箱型图

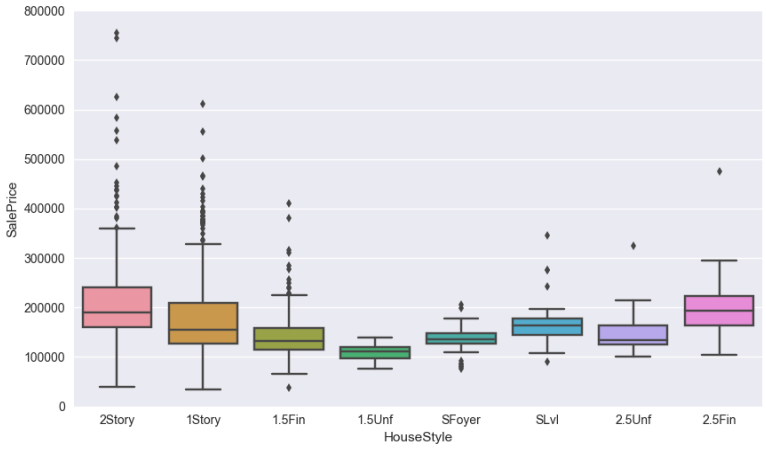


在房间数少于 11 时，房价和房间数成正比，TotRmsAbvGrd 值在 11 附近房价达到最高。



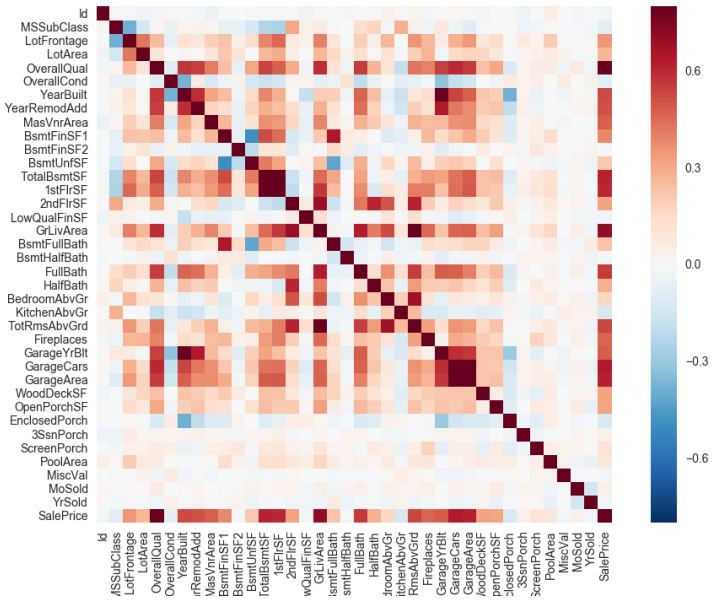
OverallQual 和 SalePrice 之间也存在类似的趋势，即整体质量越高，房价越高。此类离散型特征，数值大小和 SalePrice 有紧密的联系。

5、 HouseStyle-SalePrice 箱型图



不同的房子类型，房价水平不同。这类特征可以根据每组数据，包含的房价均值、中位数等，设置不同的等级，使变换后的等级数值和最终房价有线性相关性。

6、 相关性分析



图中不同变量之间，颜色越强的相关性越强，可以看出相关系数最大的有， TotalBsmtSF（地下室总面积）和 1stFlrSF（一楼平方英尺）； Garage 变量群； YearBuilt 和 GarageYearBuilt;
和 SalePrice 相关性较大的有（相关性依次递减）： OverallQual; GrLivArea（（地面）生活区平方英尺）; TotalBsmtSF、lstFlrSF、GarageCars、GarageArea。
由图中可以看出，原始特征中存在多重相关性，这种相关性会使模型不稳定，因此后期需要去除变量的多重相关性。

7、接下来，将训练数据和测试数据，连到一起，进行数据的预处理。

二．数据清洗

1、查看有缺失值的特征以及缺失数量：

PoolQC	2908	MasVnrArea	23
MiscFeature	2812	MSZoning	4
Alley	2719	BsmtFullBath	2
Fence	2346	BsmtHalfBath	2
SalePrice	1459	Utilities	2
FireplaceQu	1420	Functional	2
LotFrontage	486	Electrical	1
GarageQual	159	BsmtUnfSF	1
GarageCond	159	Exterior1st	1
GarageFinish	159	Exterior2nd	1
GarageYrBlt	159	TotalBsmtSF	1
GarageType	157	GarageCars	1
BsmtExposure	82	BsmtFinSF2	1
BsmtCond	82	BsmtFinSF1	1
BsmtQual	81	KitchenQual	1
BsmtFinType2	80	SaleType	1
BsmtFinType1	79	GarageArea	1
MasVnrType	24	SalePrice	1459

2、其中，LotFrontage 和 LotArea（无缺失）有很大的相关性，所以参考不同 LotArea 对应的 LotFrontage 得出缺失的 LotFrontage。在这里，我们将 LotArea 进行分组，计算每个组内 LotFrontage 的中位数（为了减少异常值的影响，我们选取了中位数而不是均值），来填充缺失值。

3、平方英尺的砌体饰面区域、地下室未完成的平方英尺、地下室总面积、 车库容纳的车数、2 型成品平方英尺、类型 1 完成平方英尺、车库面积，这些变量都是数值型变量，或者是有大小关系的数值类别。根据特征描述文档可知，这些值缺失，则表示没有相关的配置，因此用 0 填充。

4、泳池质量、其他类别未涵盖的其他功能、通往房产的胡同类型、栅栏质量、壁炉质量、车库质量、车库状况、车库的完成情况、车库建成年份、车库类型（位置）、指花园层墙、地下室的评估、评估地下室的高度、地下室成品区域的评级、地下室完工区域的评级、砌体贴面类型，属于类别类型，这些值缺失，可能因为主体并不存在（与 3 情况类似），因此用 None 填充。

5、标识销售的一般分区分类、有齐全浴室的地下室、半浴室的地下室、公用设施、家庭功能、电气系统、厨房质量、销售类型、房屋外墙、房屋外墙（如果有多种材料），这些值是每栋房子所必有的，缺失可能因为统计遗漏，因此用众数填充。
6、现在只剩下需要预测的 SalePrice 包含缺失值。

三．特征工程

1、为了在一些特征上使用 labelEncoder 和 get_dummies，把一些数值型离散特征转化为字符型特征。这些特征有：

识别销售涉及的住宅类型、有齐全浴室的地下室、半浴室的地下室、半个上等级的浴室、楼层以上的卧室、厨房、已售出月份（MM）、已售出年份（YYYY）、原始施工日期、重塑日期、低质量的平方英尺、车库建成年份

2、做值映射表

现在要创建尽可能多的特征，然后再用模型选择好的特征。所以对 SalePrice 根据其他特征进行分组，并根据每个组中 SalePrice 的均值和中位数分类，映射成新的特征（这样类别的数字大小就是有意义的）。

如：根据住宅类型对 SalePrice 进行分类，计算均值、中位数和计数。

	SalePrice		
	mean	median	count
MSSubClass			
120	200779.080460	192000.0	87
160	138647.380952	146000.0	63
180	102300.000000	88500.0	10
190	129613.333333	128250.0	30
20	185224.811567	159250.0	536
30	95829.724638	99900.0	69
40	156125.000000	142500.0	4
45	108591.666667	107500.0	12
50	143302.972222	132000.0	144
60	240403.542088	216000.0	297
70	166772.416667	156000.0	60
75	192437.500000	163500.0	16
80	169736.551724	166500.0	58
85	147810.000000	140750.0	20
90	133541.076923	135980.0	52

根据计算出的中位数大小分等级，并进行映射：

'180' : 1
'30' : 2　　'45' : 2
'190' : 3, '50' : 3, '90' : 3,
'85' : 4, '40' : 4, '160' : 4

'70' : 5, '20' : 5, '75' : 5, '80' : 5
'120': 6, '60' : 6

用同样的方法处理的类别特征有：MSZoning、Neighborhood、Condition1、BldgType、HouseStyle、Exterior1st、MasVnrType、ExterQual、Foundation、BsmtQual、BsmtExposure、Heating、HeatingQC、KitchenQual、Functional、FireplaceQu、GarageType、GarageFinish、PavedDrive、SaleType、SaleCondition。

3、对三个年份特征（创建年份、重塑日期、车库建成年份）使用标签编码

从之前的可视化探索中可知，年份和 SalePrice 是有关系的，越近的年份，房价越高，且如果采用独热编码，易导致特征稀疏，因此使用 labelEncoder 对其进行编码。

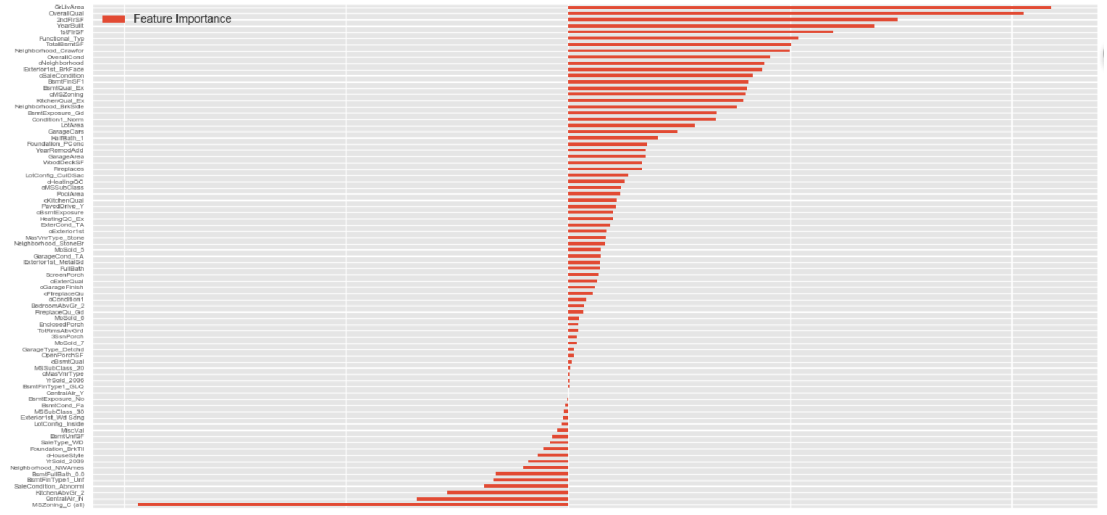
4、对数值特征中，分布倾斜的特征，使用 log 函数进行转化（经过试验，发现 log1p 效果较好），之后通过 get_dummies()函数进行独热编码。

因为很多模型都是基于数据分布符合正太分布的假设，通过观察知，大部分数据呈现右偏分布，即有大值拉高了数据的均值，因此使用 log 转化对其进行处理。

5、使用 RobustScaler 对训练数据和测试数据进行转化（根据四分位数来缩放数据。我们没有过多的对异常值的处理，对于数据有较多异常值的情况，使用均值和方差来标准化显然不合适，按中位数，一、四分位数缩放效果要好）。

6、特征扩展

使用 Lasso 选择特征，得出特征的重要性。根据重要性绘制水平直方图：



根据特征的重要性，以及自己对各种变量的认知，增加一些其他的特征：

房子总面积 = 地下室总面积 + 一楼平方英尺 + 二楼平方英尺

全部的面积=地下室总面积 + 一楼平方英尺 + 二楼平方英尺 + 车库面积

房子总面积*整体质量　　（如果是加法，整体质量数值太小，没有什么意义）

地面以上生活区面积*整体质量

标识销售的一般分区分类*全部面积

标识销售的一般分区分类 + 整体质量

标识销售的一般分区分类 + 创建年份

Ames 城市范围内的物理位置*房子总面积

Ames 城市范围内的物理位置 + 整体质量
Ames 城市范围内的物理位置 + 创建年份
类型 1 完成平方英尺*整体质量
家庭功能*房子总面积
家庭功能 + 整体质量
地块面积*整体质量
房子总面积 + 地块面积
接近各种条件*房子面积
接近各种条件 + 整体质量
地下室面积=类型 1 完成面积 + 类型 2 完成面积 + 地下室未完成的面积
房间数=全浴室数目 + 不包括浴室的房间总数
门廊面积= 开放式门廊面积 + 封闭式门廊面积 + 三季门廊面积 + 屏幕门廊面积
总体面积=地下室总面积 + 一楼平方英尺 + 二楼平方英尺 + 车库面积 + 开放式门廊面积 + 屏幕门廊面积 + 三季门廊面积 + 屏幕门廊面积

7、原始特征，以及构建的特征，部分存在高度相关性，使模型不稳定，效果降低， PCA 可以去除共线性。因为这儿的目的不是降维，所以 PCA 的参数大致与特征个数相同。

四．基本建模&评估

1、用交叉检验策略（K 折交叉验证（k-fold）把初始训练样本分成 k 份，其中（k-1）份被用作训练集，剩下一份被用作评估集，这样一共可以对分类器做 k 次训练，并且得到 k 个训练结果），计算均方根误差（kaggle 上评分用的是均方根误差），最小的模型效果最好。

总共选择了 13 种模型，用 5 折交叉验证来评估模型，使用的模型包括：

LinearRegression 线性回归
Ridge 岭回归
Lasso Lasso
Random Forrest 随机森林
Gradient Boosting Tree 渐变提升树
Support Vector Regression 支持向量回归
Linear Support Vector Regression 线性支持向量回归
ElasticNet 弹性网络
Stochastic Gradient Descent 随机梯度下降
BayesianRidge 贝叶斯岭
KernelRidge 内核岭
ExtraTreesRegressor 极端随机森林回归
XgBoost

先使用默认参数或常用参数，对各个模型的打分进行分析，从中选择效果较好的几个模型：Lasso、Ridge、SVR、KernelRidge、ElasticNet。

五．参数调整

使用 gridsearch 方法，进行参数调整。
对选出的每种模型选择几个合适的参数组合，并计算其运算的 score(均方根误差)，选出效果最好的参数组合：
lasso = Lasso(alpha=0.0005, max_iter=10000)
ridge = Ridge(alpha=60)
svr = SVR(gamma=0.0004, kernel='rbf', C=13, epsilon=0.009)
ker = KernelRidge(alpha=0.2, kernel='polynomial', degree=3, coef0=0.8)
ela = ElasticNet(alpha=0.005, l1_ratio=0.08, max_iter=10000)
bay = BayesianRidge()

六．集成方法

1、加权平均数
首先基于每个模型的 score（均方根误差），分配对应的权重。

Model	Lasso	Ridge	SVR	KernelRidge	ElasticNet	BayesianRidge
mean_score	0.111297	0.110202	0.108232	0.108270	0.111171	0.110577
std_ score	0.001339	0.001205	0.001621	0.001209	0.001312	0.0060
weight	0.02	0.2	0.25	0.3	0.03	0.2

取各个模型预测结果的加权平均值，计算 score 为 0.10768459878
尝试一下使用两个最模型的平均，其交叉验证分数为：0.106683495872（说明其他模型拖后腿了）
2、stacking 方法
除了正常的 stacking，还增加了 get_oof 方法，因为之后会将由 stacking 产生的特征和原始特征结合起来。
用 Lasso，Ridge，SVR，Kernel Ridge，ElasticNet，BayesianRidge 作为第一层模型，Kernel Ridge 作为第二层模型。
接下来提取从 stacking 产生的特征，之后和原始特征组合起来。
stacking 方法得出的 score 为：0.101824683704，想比加权平均，有了很大的提升。
最终使用 stacking 方法进行预测，得出结果。