Report

1.摘要

随着人工智能技术的快速发展，机器学习被广泛应用于各行各业，并在多个领域都取得了较好的成果。房价是一个影响因素复杂的热点问题，难以对其做出全面准确的预测。因此，本文尝试将相关的机器学习算法应用于房价预测中，在一个相对稳定条件下，建立一个符合房价复杂特性的模型，对房价进行分析和预测。本文的主要工作如下：

1、对数据进行预处理：包括缺失值填充，可视化探索，离群值处理

2、特征工程：包括新特征筛选，feature dummy code，特征规范化

3、模型选择：包括对 the least square 在内的 5 中模型进行训练和验证，并对每个模型进行合理评估

4、参数调优：使用网格搜索办法，对模型进行参数组合寻优，以此来找到最好的参数组合模型

最后通过实验得出，在本文尝试的 5 种回归模型种，RandomForestRegressor 模型的表现结果最后，R2 指标可以达到 0.8556，其 RMSE 评价指标值为 31574.2035，较 kaggle 种其他作品都有较第的 RMSE 值和较高的模型的表现力。

2.介绍

机器学习是属于人工智能科学的一门多领域交叉学科，其涉及到概率论、最优化理论、统计学、逼近论等学科的相关知识。随着大数据和计算机硬件技术的发展，机器学习有了前所未有的机遇，机器学习的应用也越来越广泛。例如，天气预报利用机器学习技术对收集到的天气数据进行分析，去预测后面日期的天气；企业通过机器学习对客户行为习惯进行分析，从而得到客户画像，根据客户画像来制定有针对性的营销策略;网络搜索引擎通过机器学习技术来检索相关内容并进行重要度排序，优先显示重要度高的内容，让使用者能够快速准确地找到需要的内容。除此之外，在人脸识别、环境监测、推荐系统、疾病诊断、故障诊断、自动驾驶等许多方面都有着机器学习技术的身影。许多事实表明，机器学习被广泛应用于医学、金融、工业、农业等多个领域，且取得了一定成就，因此机器学习技术 具有重要的应用价值。

无论是在任何时期，住房都是人们生活最基本的需求，与人们的日常生活息息相关。飞速发展的房地产市场已成为推动经济增长、拉动内需的基础性支柱产业，在国民经济发展中扮演着重要角色。所以，住房问题不仅是民生问题，也是经济问题，关系着国家和社会的稳定。房价是房产的市场价值，其对人们的生活水平和国民经济发展有着很大的影响。房价的研究已受到统计学、管理学、计算机科学等多个领域的重点关注，房价的预测也成为许多学者研究探讨的问题。顺应大数据和机器学习的发展趋势，结合网络数据，利用机器学习算法分析预测房价问题更具科学性。

从整体上来看，对房价的预测研究可以归结为两类，一类是对房价进行定性估价预测，更多的是倾向于经济学分析，主要关注市场信息，很少使用数学模型。另一类侧重于定量分析，利用数学模型对房价进行量化预测。而定性分析很容易受到主观性因素的影响，因此，对房价进行分析预测时，采用定量分析要比采用

定性的方法更科学、更合理。对于定量分析的房价预测，目前国内外学者大致有两种思路：一是把房价的变化看作是一个时间序列来预测房价。二是分析房价的影响因素，利用影响因素建立指标体系来预测房价。针对影响因素构建预测模型对房价进行预测分为两个角度：一个是从宏观经济角度出发，利用 GDP、贷款利率等宏观指标对平均房价进行预测。另一个是从房屋自身角度出发，根据房屋自身特征因素对具体房屋的房价进行预测。本文即采样最后一种方法，从房屋自身角度出发来对房屋价格进行预测。

根据房屋的自身特征，例如户型、修建年份、房屋面积、装修情况等影响因素建立具体的房屋价格。Rosen 等首次将特征价格(Hedonic)理论引入到房价预测中，并提出了住宅特征模型，该模型第一次研究了住宅价格与居住环境的关系。Hasan Selim 对土耳其房价的影响因素进行分析，利用人工神经网络模型和 Hedonic 模型对房价进行了预测，对比发现人工神经的预测效果优于 Hedonic 模型。陈世鹏以襄阳房贷数据建立随机森林模型，并与 ARIMA 模型及多元线性回归模型预测结果进行对比，实验证明随机森林模型预测效果较好。Chia-Chen Fan 等建立了房价预测和房屋信息销售的网络服务系统，该系统结合分析方法和预测模型来预测房价。Stephen Law 等使用深度神经网络模型和住房特征结合，以估算英国伦敦房价。Naalla Vineeth 等通过使用机器学习算法中的简单线性回归、多元线性回归和神经网络建立房价预测模型，用于帮助买卖双方找到房子的最佳价格。

本文以 kaggle 的房价数据为着手点，通过对数据特征进行构建筛选，找出和 sale price 相关性较高的特征，然后分别对最小二乘法回归，lasso 回归，ridge 回归，MLPregreesion 和集成学习的 RandomForestRegressor 模型进行训练和预测，得出 RandomForestRegressor 模型相对于其他模型有较好的表现，随后通过参数网格搜索方法，对模型进行调优处理，以此得出最好的模型参数组合和模型结果。通过实验环节，得出本次回归模型的最好 R2 值为 84.83%。

The rest of the paper is organized as follows: Section II is an overview of the data; Section III details our methodology for this work; Section IV covers our experiments and analyzes their results; and in Section V we draw our conclusions.

3.数据描述

本论文的数据来源于 kaggle 中的房价预测比赛， 数据一共包含 79 个特征，表一展示了部分特征的名字及其描述信息。其特征主要包括了几个方面，一个是房屋的自身特征，比如房屋类型，建造年份，房屋面积，装修类型等，涉及和房屋相关的方方面面，相对比较全面；另外一个是房屋的空间位置，比如房屋所在的街道，房屋是否靠近主干道，街道的类型，离物业的距离等等。

Table 1 part feature name and it's description

Feature	Description
HeatingQC	Heating quality and condition
CentralAir	Central air conditioning
Electrical	Electrical system
1stFlrSF	First Floor square feet
2ndFlrSF	Second floor square feet

LowQualFinSF	Low quality finished square feet (all floors)
GrLivArea	Above grade (ground) living area square feet
BsmtFullBath	Basement full bathrooms
BsmtHalfBath	Basement half bathrooms
FullBath	Full bathrooms above grade
HalfBath	Half baths above grade

数据中既有定性数据比如特征 MSZoning ， Street，Utilities 等，其中的内容都是字符串类型，也有数值类型的数据，比如特征 MSSubClass， LotFrontage， OverallQual。同时，数据中数值性数据的 scope 也有很多区别，比如特征的 LotFrontage 的最大值和最小值分别为 313 和 21， 而特征 GrLivArea 的最大值是 5642，其最小值为 334。除了上述的情况之外，本次数据还存在诸多的缺失值。

通过对数据的基础了解，可以看出原始数据不能直接用于模型之中，因此本文将通过数据探索，数据清洗，数据可视化，特征工程等一系列处理，将当前数据变成清洗成适用于模型的数据。

4.方法

在本节，我们将根据第二部分提到的数据特点情况和问题进行针对的处理，并且阐述为什么要处理和这样处理操作的理由，并在最后对最后清洗完成数据的可视化展示。需要注意的是，由于训练集和测试集都存在相同的问题，因此本文在做数据操作之前，已经将测试集和训练集进行了合并，这样方便我们对两个数据进行同步的处理，因此下文于数据相关的所有描述和处理，都是针对测试集和训练集合并之后的整体数据。

4.1 缺失值处理

Table 2 feature with null value

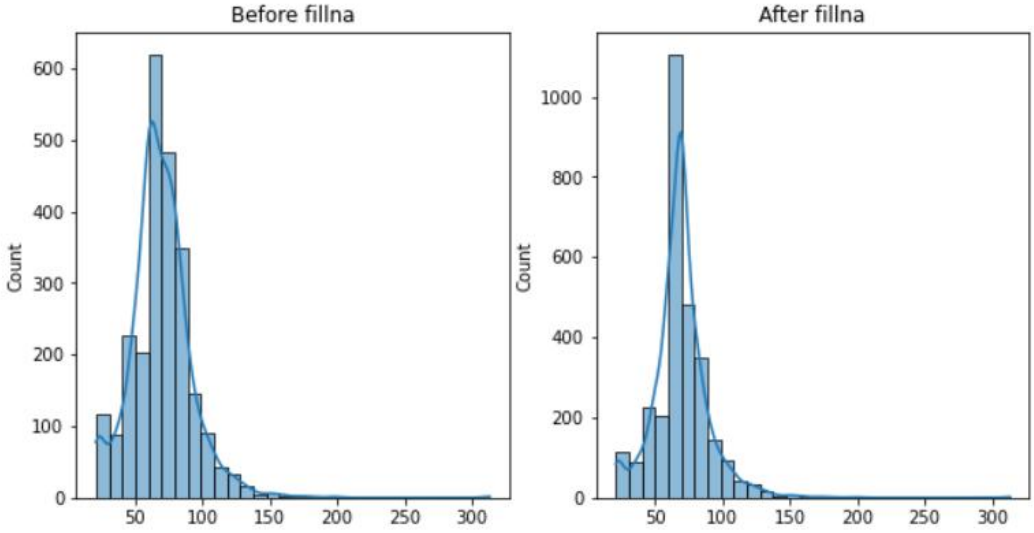
Feature	percent	feature	percent
LotFrontage	0.17739726	Electrical	0.000684932
Alley	0.937671233	FireplaceQu	0.47260274
MasVnrType	0.005479452	GarageType	0.055479452
MasVnrArea	0.005479452	GarageYrBlt	0.055479452
BsmtQual	0.025342466	GarageFinish	0.055479452
BsmtCond	0.025342466	GarageQual	0.055479452
BsmtExposure	0.026027397	GarageCond	0.055479452
BsmtFinType1	0.025342466	PoolQC	0.995205479
BsmtFinType2	0.026027397	Fence	0.807534247
MiscFeature	0.963013699		

数据中很多特征都存在缺失值，其中 MiscFeature， Fence，PoolQC 等缺失值超过 80%以上，可以直接判定这几个特征为无效特征， 虽然填充缺失值的技术有很多，比如有基于已有数据的简单填充方法，如平均值填充，中位数填充； 也有基于模型拟合的填充方法等，但是针对缺失值比例很大的特征，我们可以有理由认为该特征不重要或无意义。本文将特征值缺失比例的阈值设置到 0.2，即当某个特征有超过 20%的数值确实的时候，我们即将该特征剔除掉。通过该操作， 一共剔除 5 个特征。

对于缺失值比例较小的特征， 最简单的办法是删除所有带有缺失值的样本，但是由于本次数据样本量本身就不多，直接删除会损失到很多信息，因此本文进行填充处理。考虑到存在缺失值的特征中即包括数值型数据和字符型数据，分别采取两种不同的策略来填充。针

对数值型数据，我们使用已有数据的平均值来填充，针对字符型数据，我们采用众数值来进行填充，这样可以在很大程度上，保持特征的分布形态，对后期的模型拟合有较好的作用。

图 1 展示了特征 LotFrontage 在缺失值填充前后整体分布的变化情况。



Picture 1 feature distribution before and after fillna

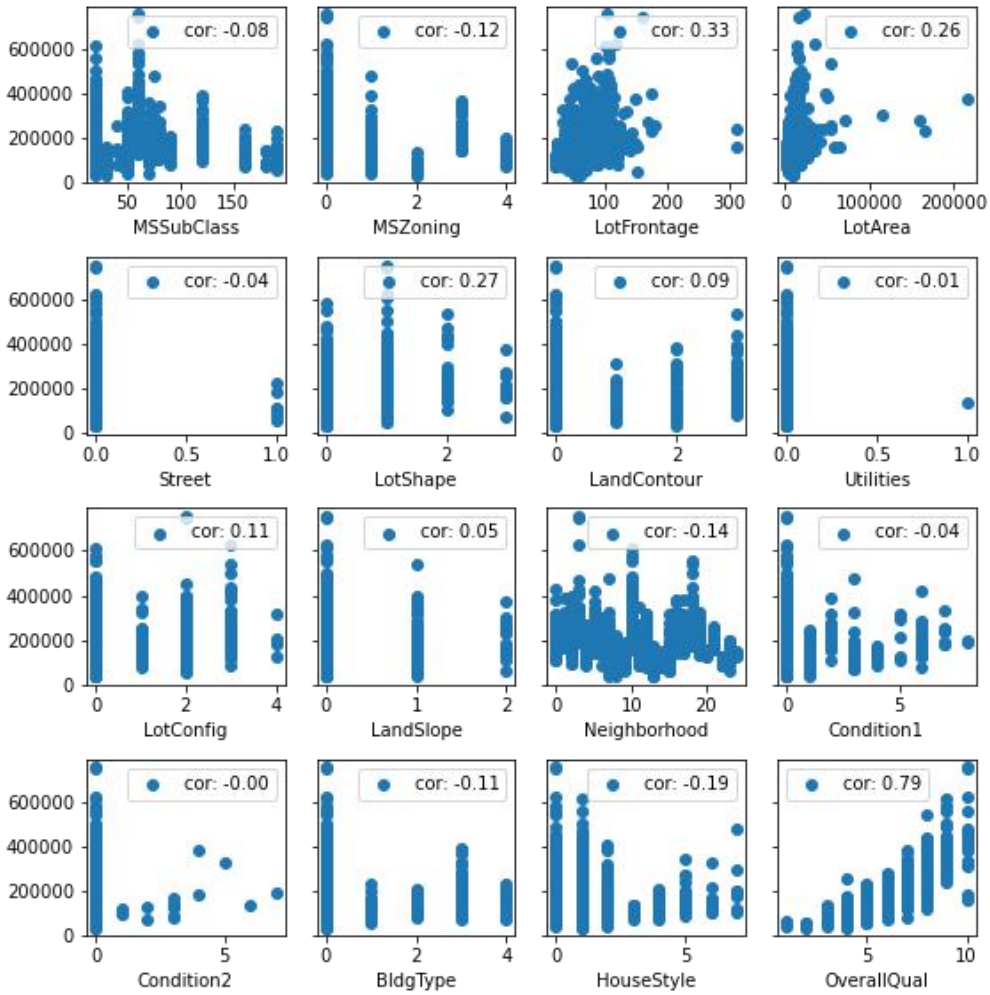
4.2 特征选择

特征选择(Feature Selection)也称特征子集选择(Feature Subset Selection , FSS)，或属性选择(Attribute Selection)。是指从已有的 M 个特征(Feature)中选择 N 个特征使得系统的特定指标最优化，是从原始特征中选择出一些最有效特征以降低数据集维度的过程,是提高学习算法性能的一个重要手段,也是模式识别中关键的数据预处理步骤。对于一个学习算法来说,好的学习样本是训练模型的关键。

考虑到本文的模型是一个回归模型，因此考虑使用相关系数来判断特征和 label 之间的相关关系的强度。由于当前部分特征是字符类型，不能直接计算相关系数，因此考虑将这些字符型特征转化成数值型。将字符型转化成数值型的操作主要有：one-hot 编码，dummy code 编码或者直接将不同的分类值用数值代替。本文采用的是最后一种方法，比如特征 Street 中有两个值，分别是 Pave, Grval， 本文将用 0，1 来表示这两个值。但是如果某个特征的值中，存在明显的大小关系，比如针对包含 郊区公寓，普通公寓，高级公寓这样带有好坏之分的特征的时候，我们应该将将等级高的值设置较大的数值来表示。

(1) 散点图-特征剔除

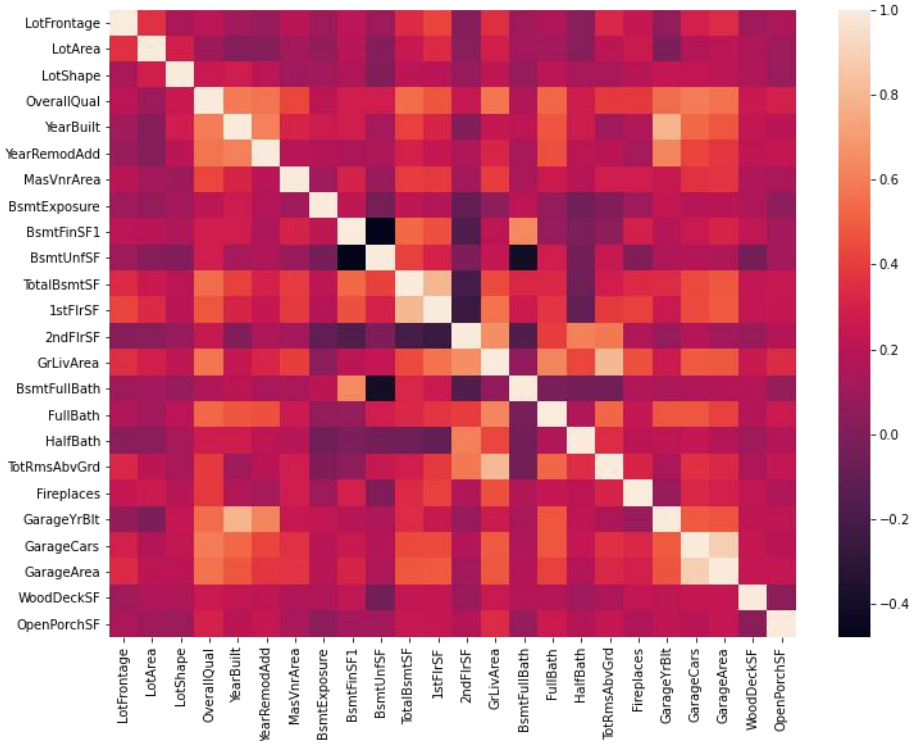
本文将所有过滤留下来的特征和 y(即 Sale price)值做散点图，来查看特征与 y 之间的关系。图 2 展示了部分特征与 y 值的散点图情况，同时各个特征和 y 之间的相关系数描述在每个图的图例中，这样可以方便的查看不同特征和 y 之间的相关关系，比如图 2 中左上角的散点图，特征 MSSubClass 和 y 之间的相关系数为-0.08， 说明两者基本没有相关关系；同时查看图中右下角最后一幅图的散点图，可以从散点图中的趋势明显的看出两者有较强的相关性，同时查看图例可以发现两者的相关系数为 0.79，属于较强的相关性。



Picture 2 feature and y scatter plots

我们一般认为相关系数的绝对值小于 0.2 的时候，两个变量的不具备有相关关系；当相关系数的绝对值在 0.2 -0.6 之间的时候，属于弱相关，当数值大于 0.6 的时候，可以是强相关。因此本文首先通过相关系数进行过滤，将特征和 y 的相关系数小于 0.2 的所有特征都剔除掉，因为相关性比较低的特征不会对回归模型做出明显的贡献。经过这步操作之后，数据还剩 24 个特征。

然后我们对剩余的所有特征的相关性绘制热图，如图 3 所示，从图中可以看出有较多的特征之间存在比较强的相关性，如 GrLivArea(Above grade (ground) living area square feet) 和 TotRmsAbvGrd(Total rooms above grade (does not include bathrooms)) 两个变量从定义上看，确实具有较强的相关性。因此特征之间是存在共线性的问题，面对这个问题，我们可以考虑使用降维技术，如 PCA,Lasso 等来消除共线性的存在，让模型可以更好的拟合数据，或者使用对共线性不敏感的模型，如决策树，随机森岭等这样的技术。



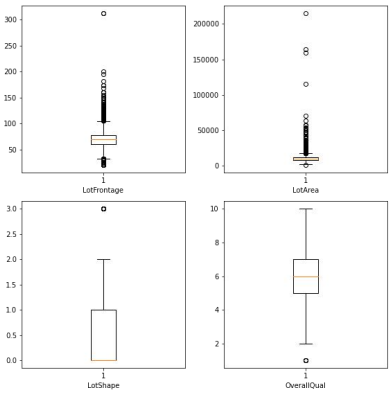
Picture 3 feature correlation hot map

(2) 箱型图-消除离群值

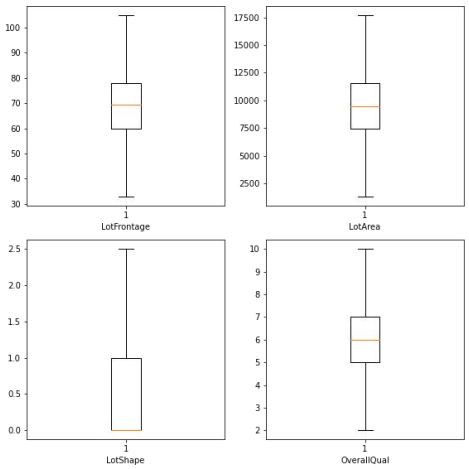
（https://en.wikipedia.org/wiki/Outlier 维基百科的地址，下面关于离群值的定义来自维基）

In statistics, an outlier is a data point that differs significantly from other observations.An outlier may be due to variability in the measurement or it may indicate experimental error; the latter are sometimes excluded from the data set.An outlier can cause serious problems in statistical analyses.

箱型图是一种可以非常有效检测离群值的方式，因此本文对所有特征绘制箱型图，如图 4 所示，从图四中可以非常明显的看出有很多特征都有离群值，比如特征 LotFrontage, LotArea 等等， 且部分特征离群值的数量占了总数很大一部分比例。考虑到本次数据总体样本量较少，对离群值直接进行删除会严重影响样本的数量，因此本文对所有离群值进行截断处理，当某值大于其所属特征箱型图的 top value 的时候，则将对其进行裁剪使其值为 top value, 对于小于 bottom value 的值也做同样的处理，这样就可以保证所有值都在箱型图的 top 和 bottom value 之间，同时保留了所有的样本。图 5 展示了对数据执行上述操作之后的箱型图，可以看出所有值都已经在合理取值范围之内。



Picture 4 some feature boxplot



Picture 5 some feature boxplot after delete outlier

(3) 特征缩减

在第二章节，对数据的描述部分我们可以知道，当前数据的特征有 scale 有较大的差别，这不利于一些回归模型的拟合，因为模型会偏向较大尺度特征的影响。因此我们需要将所有特征归一化到同一个取值范围。本文采用 min-max scale 方式，让所有特征的取值范围都在 0-1 之间，如公式(1)所示：

$$X_{new} = \frac{X - X_{min}}{X_{min}} \qquad (1)$$

5. Experiment

经过章节 4 的数据预测里之后，我们已经得到了比较完备可靠的干净数据，接下来我们需要选择合适的模型，对数据进行拟合，并进行预测结果。

5.1 模型选择

本文将分别使用 LinearRegression, LassoCV, RidgeCV, MLPRegressor,RandomForestRegressor 这 5 种模型对数据进行拟合，选择这些模型的原因是：

1.LinearRegression 作为经典的回归模型，其模型具有较好的解释性，并且易于理解。

2.Lasso regression 和 Ridge regression 都可以较好的解决特征共线性的问题

3.MLPRegressor 可以像神经网络一下，提取深层的特征信息。

4.RandomForestRegressor 考虑到特征种有很多分类变量，而决策树比较擅长拟合分类变量，因此也尝试使用以决策树为弱分类器的继承学习模型

以上几种模型都有自己的优势和劣势，因此本章节，将先对每个模型进行拟合（在 sklearn 的默认参数下），然后选出表现较好的模型，再对其进行参数调优，以求得更好的拟合效果。

5.2 初步实验结果

将整体数据集按照 80：20 的比例拆分成训练集和验证集，然后用基础模型(即模型种所有参数使用 sklearn 种的默认参数)对数据进行拟合和评价（使用验证集），用 r square 和 root mean square error 来对模型进行评价.

Table 3 model performance

Model	R square	rmse
LinearRegression	0.7727	39617.9016
Lasso regression	0.7650	40279.8869
Ridge regression	0.7635	40409.1121
MLPRegressor	-4.6249	197092.2343
RandomForestRegressor	0.8506	32109.9905

从表三种可以看出，LinearRegression， Lasso regression 和 Ridge regression 不管在 R square and rmse 都具有相近的解决，且都较差。MLPRegressor 的 r square 值小于 0，这说明该模型的拟合效果十分的差，其预测值和真实值具有非常打的差距。在这 5 个模型中，RandomForestRegressor 的结果最好， r square 可以达到 0.8574 ，说明模型可以在很大程度上去解释数据。

5.3 参数调优

通过 5.2 的实验数据，我们已经得出 RandomForestRegressor 是 5 个模型中最好的模型，因此本小节， 将对该模型进行参数调整，以让模型得到更好的效果。这里使用网格搜索法，即对所有预设置好的参数进行全排列，然后对每一种参数组合的模型进行训练和评估，最后选出评价指标最高的参数组合。

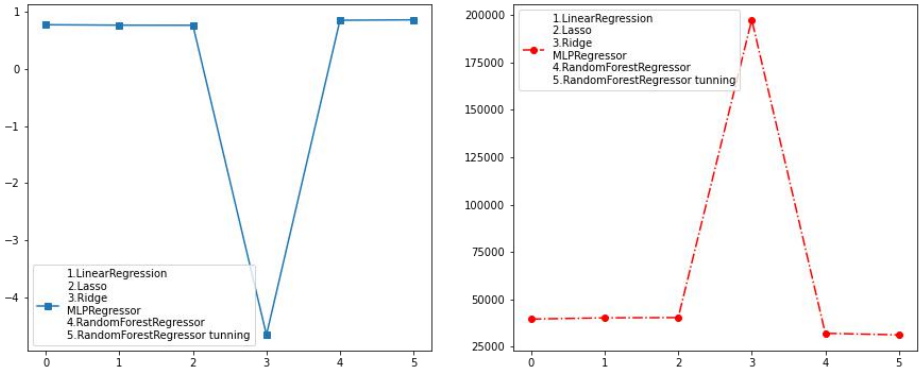
Table 4 parameter

参数名字	参数定义	参数值列表
n_estimators	The number of trees in the forest.	100,300, 500
max_depth	The maximum depth of the tree	None, 1, 2, 5, 8
bootstrap	Whether bootstrap samples are used when building trees	True, False
criterion	The function to measure the quality of a split	mae, mse

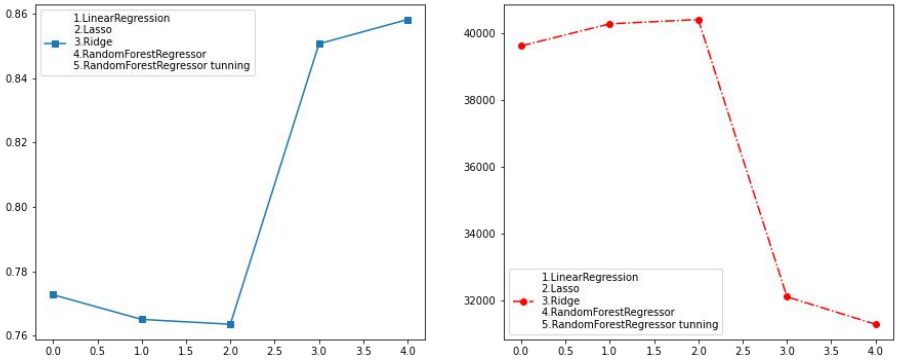
通过对以上参数及其参数值进行组合和训练评价，最后得出最好的参数组合为，n_estimators = 300, max_depth=None, bootstrap = True, criterion=mae。最好的模型评价结果，R square = 0.8582, rmse=31287.7824。

5.4 结果可视化

从下面图 6 和图 7 可以轻松对比不同模型之间的性能区别，mlpregression 的表现解决最差，进行参数调整的随机森林模型结果最好。

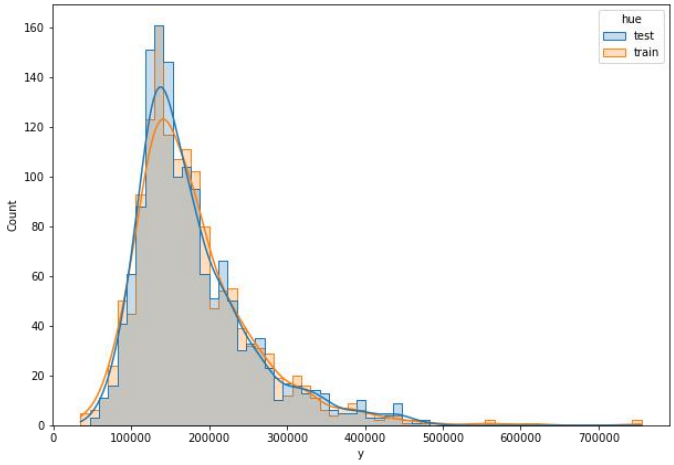


Picture 6 model compare with mlpregression



Picture 7 model compare without mlpregression

然后我们用选出的最好模型对测试及的结果进行预测，并将最后的预测结果和训练集中的 y 值做直方图。可以看出两者的分布基本吻合，这可以说明本次预测的结果可信度较高，如果两个分布的结果截然不同，那就可以很大概率的认为模型的预测结果很差。



Picture 7 true and prediction y distribution

6. 结论

通过本次建模预测，可以得出表 5， 该表展示不同特征的重要性，可以发现房屋的定价只受部分房屋自身特征的影响，如特征 OverallQual 和 GrLivArea 等对价格起了决定性的影响。

Table 5 features importance

特征	重要性
OverallQual	0.3594
GrLivArea	0.1144
TotalBsmtSF	0.0660

本文运用多种数据科学技术，从数据清洗到模型训练预测，得出了一个较为满意的模型，其预测结果也交有说服力，但是整体的模型 rmse 值却较大，说明依然有很大的改进空间。后期将更加深入了解特征构建和神经网络等相关技术，希望能在此当前基础上，搭建更加强大的预测模型。

[23]Rosen S . Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition[J]. Journal of Political Economy, 1974, 82(1):34-55.
[24]Selim H. Determinants of house prices in Turkey: Hedonic regression versus artificial
neural network[J]. Expert Systems with Applications, 2009, 36(2):2843-2852.
[25]陈世鹏. 基于随机森林模型的房价预测[J]. 科技创新与应用, 2016(4):52-52.
[26]Fan C C , Yuan S M , Zhang X , et al. A House Price Prediction for Integrated Web Service
System of Taiwan Districts[C]// International Conference on Genetic & Evolutionary Computing. Springer, Singapore, 2017.
[27]Stephen Law, Brooks Paige, Chris Russell. Take a Look Around: Using Street View and
Satellite Images to Estimate House Prices[J]. Papers, 2018.

[28]Vineeth N , Ayyappa M , Bharathi B . House Price Prediction Using Machine Learning
Algorithms: Second International Conference, ICSCS 2018, Kollam, India, April 19–20,
2018, Revised Selected Papers[M]// Soft Computing Systems. 2018,425-433.