算法分析

DOI:10.19695/j.cnki.cn12-1369.2021.06.29

基于随机森林和XGBoost算法的二手车价格预测*

郑婕
(北方工业大学,北京 100144)

摘要:为更好的预测二手车价格,根据二手车数据集,将随机森林和XGBoost算法相结合进行预测。首先对原始数据进行数据预处理,对奇异值与缺失值进行预处理以及数据的结构化处理。再利用随机森林算法进行变量选择,根据随机森林输出的得分排序选择分数不为零的变量作为预测价格的特征变量。再将XGBoost,GBDT和lightGBM三种算法进行网格搜索寻到最优参数,对比后选用XGBoost算法进行二手车价格预测得到最优的二手车价格预测结果。

关键词:二手车价预测;随机森林;网格搜索;XGBoost
中图分类号:TP393 **文献标识码**:A **文章编号**:1007-9416(2021)06-0090-04

0 引言

随着机器学习的广泛发展与应用,简单的学习器或模型已经满足不了需求,集成算法应用而生。所谓集成算法,需要构建多个学习器,然后用一些方法巧妙的将它们结合在一起,再来完成学习任务的,这样可以获得比单一学习效果更好的学习器。周志华[1]指出个体学习器的“准确性”和“多样性”本身就存在冲突,一般准确性很高之后,要增加多样性就需牺牲准确性。产生并结合‘好而不同’的个体学习器,正是集成学习研究的核心。按照个体学习器之间的关系,分为Bagging、Boosting、Stacking三大类。

Bagging的原理首先是基于自助采样法(bootstrap sampling)一些样本被随机的得到来训练出不同的基学习器,然后对这些不同的基学习器进行投票,得出分类结果,随机森林就是这个算法的典型代表[2]。随机森林具有广泛的应用,宋欠欠[3]在运用随机森林对高维数据变量筛选的研究中指出利用随机森林算法进行变量筛选结果稳定,并能够保证良好的预测效果。

Boosting,提升算法,它通过反复学习得到一系列弱分类器,一个强分类器由这些弱分类器组合得到,此时,弱学习器是强学习器提升的过程[1]。总体而言,Boosting的效果要比Bagging好,但是这个算法中新模型是在旧模型的基础上生成的,就不能用并行的方法去训练,并且由于对错误样本的关注,也可能造成过拟合。Boosting的算法族中有很多有名的算法,比如Adaboost、GBM、XGBoost[4]。陈天奇[5]在对XGBoost的算法研究中指出稀疏数据和加权分位数草图提供了一种新的稀疏感知算法,用于近似数学习算法上的优化使得在利用其进行预测计算时可以得到更加准确的结果。最后一类Stacking训练一个模型用于组合其他各个基模型。具体方法是把数据分成两部分,用其中一部分训练几个基模型A1,A2,A3,用另一部分数据测试这几个基模型,把A1,A2,A3的输出作为输入,训练组合模型B,Stacking可以组织任何模型,实际中常使用单层logistic回归作为模型[6]。

在用算法进行预测时,常用到的算法有XGBoost、GBDT、LightGBM、神经网络算法等,魏长亮在对岩柱稳定性的预测研究中使用了这三种算法进行对比,并使用五重交叉验证寻求每个模型的最优参数配置再进行预测[7]。谢勇对每月住房租金进行预测时使用了XGBoost和LightGBM两种算法进行预测对比,发现 XGBoost的表现更好[8]。所以通过文献综合本文在进行二手车价格预测上选择了比较经典的XGBoost算法,同时将它和GBDT、LightGBM两个算法进行对比,得出XGBoost预测误差最小,性能最好。

我国有一个庞大的二手车需求市场,二手车的销售对市场经济有很大的作用,市场潜力很大,但是目前它的潜力还没有完全发挥出来。二手车的交易价格受许多因素的影响,车的型号,行驶里程,车的配置,实用年限,包括车的车系、颜色、品牌溢价这些因素都会影响二手车的交易价格。当然,对于不同的二手车交易市场,所处的地理

收稿日期:2021-04-12
★基金项目:北方工业大学毓优人才项目(107051360021XN083/058)和北方工业大学新教师科研启动经费(110051360002)。
作者简介:郑婕(1996—),女,山西忻州人,研究生,研究方向:应用统计。

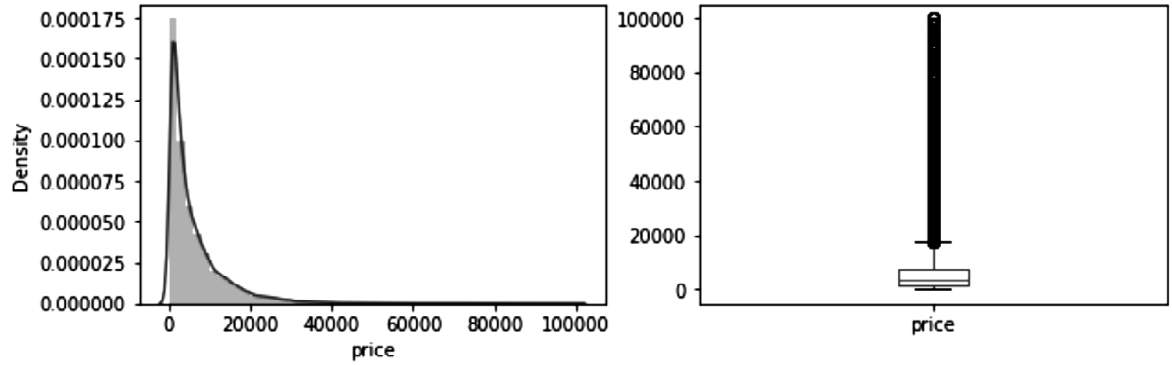

图1 Price分布图

Fig.1 Price distribution

表1 数据集表

Tab.1 Data set table

数据集	样本数
训练集（training_set）	120000
测试集（testing_set）	30000

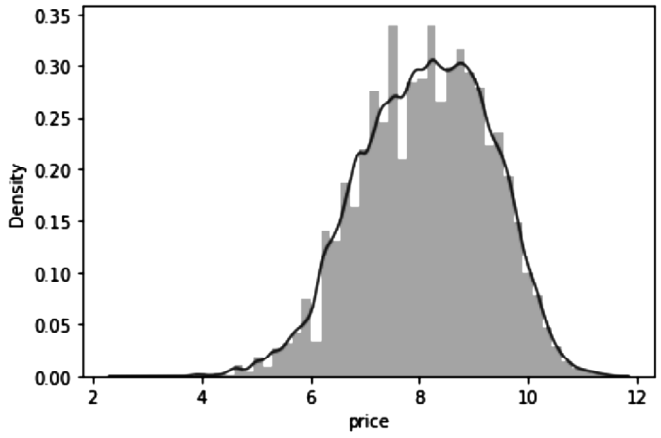

图2 对数变换后Price分布图

Fig.2 Price distribution graph after logarithmic transformation

位置的消费水平，人均可支配收入等，也影响着二手车的交易价格[9]。

1 数据预处理

二手车的数据来自于阿里云天池大赛，共有3个数据集，这里用到的是Car train数据集，由于原始数据集并非结构化数据，因此数据预处理首先对数据集进行结构化处理。用众数填充缺失值较多的样本。原始数据包含31列变量信息，其中15列为匿名变量，共150000个样本。本文运用scikit-learn中的train_test_split函数对数据集进行拆分，测试集的比例设为20%，最终训练集和测试集的划分如表1。

在对特征变量进行观测时发现变量“seller”和“offerType”存在严重的偏态分布，特征倾斜严重的变量直接删除。同时对预测变量的分布观测发现不符合正态分布如图1。

所以对预测变量price采取对数变换得到如图2。

表2 变量名称表

Tab.2 Variable name table

序号	变量名称	重要度
1	v_12	0.0766
2	v_8	0.0442
3	汽车使用时间	0.0375
4	v_3	0.0303
5	v_10	0.0082
6	发动机功率	0.0068
7	城市信息	0.0014
8	品牌销售额	0.0005
9	v_1	0.0002

数据预处理这个过程运用了python中的numpy、pandas等模块，结合excel等办公软件，大大提高了数据预处理的工作效率。

2 变量选择

经过数据预处理后对数据变量进行特征工程，对原有的变量进行特征构建得到新变量，例如汽车使用时间为creatDate- regDate，城市信息从regionCode也就是邮编中提取，通过brand和price计算品牌的销售额，构造新的变量进行预测。

二手车的数据集中有150000个样本，每个样本都有31个变量，1个观测变量，30个特征变量，为找到可操作变量中的主要因素，更好地预测二手车的价格，找到影响二手车价格的最高因素，本文选择使用随机森林算法进行变量选择。

随机森林是从数据表中随机选择K个特征建立决策树，重复n次。这K个特征经过不同随机组合建立起n棵决策树，对每个决策树都传递随机变量来预测结果，从n棵决策树中得到n种结果。决策树会预测输出值，通过随机森林中所有决策树预测值的平均值计算得出最终预测值，最终得到筛选出的K个变量的重要度。在随机森林中某个特征X的重要性的计算方法如下：

(1)在随机森林中的每一棵决策树，都用它对应的袋外

数据即OOB，去计算它的袋外数据误差即errOOB1。

(2)在袋外数据OOB的每一个样本的特征X加入噪声干扰这样可以随机改变样本在特征X处的值，然后再计算袋外数据误差，记为errOOB2。

(3)当我们假设随机森林有N棵树时，用Σ(errOOB2−errOOB1)/N来表示X的重要性，特征被随机加入噪声后，如果袋外准确率下降很多，就表明此特征对样本分类结果影响很大，这表明它的重要程度高[10]。

通过随机森林对变量重要度进行排序，变量包括汽车交易ID、汽车交易名称、汽车注册日期、车型编码、汽车品牌、车身类型、燃油类型、变速箱、发动机功率、汽车已行驶公里、汽车有尚未修复的损坏、地区编码、汽车上线时间、匿名特征，以及包含v0−14在内15个匿名特征。本文选择分数不为零的变量作为预测价格的特征变量，共9个变量，如表2所示。

3 建立模型

运用随机森林筛选出9个变量后，用XGBoost进行回归任务，建立XGBoost模型对二手车的价格进行预测，使用训练集进行参数训练，测试集进行预测[11]。

3.1 XGBoost原理

对于集成方法之一的XGBoost算法其预测模型可以表示为：

$$\hat{y}_i = \sum_{k=1}^{K} f(x_i) \tag{1}$$

其中K为分类器的总个数，f(x)表示第k分类器，$\hat{y}_i$表示集成K个分类器后对样本x_i的预测结果，损失函数表示为：

$$obj(\theta)=\sum_{i=1}^{n} l(y_i,\hat{y}_i)+\sum_{k-1}^{K}\Omega(f_k) \tag{2}$$

其中$l(y_i,\hat{y}_i)$为样本x_i的训练误差，$\Omega(f_k)$表示第k棵树的正则项，$\Omega(f_k)=\gamma T+\frac{1}{2}||w||^2$，其中Ω为惩罚项控制模型的复杂度，同时使w更光滑从而避免过拟合。通过最小化正则化损失函数$\Omega(f_k)$得到最优参数P^*，继而可以得到既简单又精确的模型$f^*(x)$即：

$$P^*=\arg\max \phi(x) \tag{3}$$

显然无法直接通过计算得出$f^*(x)$的解，所以我们要考虑优化的方法，f(x)为决策树模型，给出权重w和树结构q即可确定一棵决策树，而树结构q实质上就是划分分裂节点的问题，所以$f^*(x)$可以转化为找最优权重w和划分分裂节点的问题[6]。

$y_i^{(0)}=0$，初始化模型中没有树时，预测结果为0；

$y_i^{(1)}=y_i^{(0)}+f_1(x_i)$加入第一棵树；

…

$y_i^{(t)}=\sum_{k=1}^{t} f_k(x_i)=y_i^{(t-1)}+f_t(x_i)$加入第n棵树。

此时利用泰勒展开式近似目标函数：

$$obj^t(\theta)=\sum_{j=1}^{\nu}\{g_i f_t(x_i)+\frac{1}{2}h_i f_t(x_i)\}+\gamma T+\frac{1}{2}\lambda\sum_{j=1}^{T}w_j^2 \tag{4}$$

因为对于$\sum_{j=1}^{\gamma}\{g_i f_t(x_i)+\frac{1}{2}h_i f_t(x_i)\}$可以看作是每个样本在第t棵树的叶子节点得分值相关函数的结果之和，所以可以从第t棵树的叶子节点上来表示[9]。

$$obj^t(\theta)=\sum_{j=1}^{\gamma}\{(\Sigma g_i)w_j+\frac{1}{2}(\Sigma h_i+\lambda)w_j^2\}+\gamma T \tag{5}$$

此时令$G_i=\Sigma g_i$；$H_j=\Sigma h_i$，代入上式w_j求偏导，使其导函数等于0得到：

$$G_i+(H_j+\lambda)w_j=0 \tag{6}$$

解得：$w_j^*=-\frac{G_j}{H_j+\lambda}$

所以目标函数最优解为$obj^*=-\frac{1}{2}\sum_{j=1}^{\gamma}\frac{G_j}{H_j+\lambda}$。再根据目标函数，分裂样本数据。

3.2 网格搜索

网格搜索法是一种寻找最优参数的方法，它是将估计函数的参数用交叉验证的方法得到的一种算法，将每个参数的可能取值进行排列组合，把所有可能的形式用“网格”的形式表示出来，再对它进行评估，在计算机上对每种可能的参数形式进行计算训练狗，得到一个最优的参数组合。网格搜索虽然比较耗时，但是它有很广泛的搜索范围，有很大可能找到最优参数组合[12]。使用GBDT、LightGBM和XGBoost模型建模分析时，参数的选择对模型的预测结果有着较大的影响，故需要对若干参数进行调优使用网格搜索对上述模型参数进行自动寻优，首先确定学习率，把learning_rate设置成0.1，其他参数使用默认参数，使用GridSearchCV函数进行网格搜索确定合适的迭代次数，找到合适的迭代次数后使用GridSearchCV函数对模型的其他两个主要参数进行网格搜索自动寻优，减小(增大)学习率，同时增大(减小)迭代次数，找到合适的学习率是使得在误差最小时迭代次数最少，找到最优参数[13]。

3.3 实证分析

网格搜索的结果分别如表3所示。

由网格搜索结果可以得出，max_depth=9，它表示树的最大深度为9，main_child_weight=7，决定最小叶子节点样本权重和，它是为了防止过拟合。n_estimators=300，是弱学习器的最大迭代次数，或者说最大的弱学习器的个数。一般来说n_estimators太小，容易欠拟合，

表3 XGBoost网格搜索结果
Tab.3 XGBoost grid search results

参数	结果
max_depth	8
main_child_weight	5
n_estimators	300
scores	0.9345858224796677

表4 LightGBM网格搜索结果
Tab.4 LightGBM grid search results

参数	结果
max_depth	7
main_child_weight	1
n_estimators	95
scores	0.9333038265163888

表5 GBDT网格搜索结果
Tab.5 GBDT grid search result

参数	结果
max_depth	8
main_samples_split	1
n_estimators	100
scores	0.9349200615330541

n_estimators太大，计算量会太大，并且n_estimators到一定的数量后，再增大n_estimators获得的模型提升会很小，该模型中选择的值是300。Scores=0.9345858224796677这个参数状态下的模型打分约为0.935，效果较好。

由网格搜索结果表4可以得出，树的深度为7，main_child_weight=1，n_estimators=95时Scores=0.9333038265163888这个参数状态下的模型打分约为0.933，效果较好。

由网格搜索结果表5可以得出，树的深度为7，main_samples_split=1，n_estimators=100时Scores=0.9349200615330541这个参数状态下的模型打分约为0.935，效果较好。

将网格搜索出的最优参数分别带入到GBDT、LightGBM和XGBoost中进行预测，得到如表6所示结果。

由表6对比结果所示，XGBoost模型在预测值与实际值的拟合度上表现较好。其预测性能高于GBDT模型的预测性能，与lightGBM模型进行比较也具有相对优势。表6显示XGBoost模型在MSE具有出色的表现。

表6 对比结果
Tab.6 Compare results

算法	r^2-scores	Mean-square
XGBoost	0.9311598159996348	0.3052431978601984
GBDT	0.9389792143013034	0.30565530213922515
lightGBM	0.9332016127012778	0.30968224409639405

4 结论与建议

本文使用3种机器学习模型对二手车价格进行预测，XGBoost和lightGBM作为机器学习近年提出的新方法，比传统GBDT能达到更好的预测精度，同时XGBoost在模型拟合程度和均方误差上的表现都远超LightGBM和GBDT。本文的不足之处在于XGBoost虽然能够得到较好的预测精度，但是XGBoost是基于启发式算法，寻找的解为局部最优并非全局最优。

参考文献

[1] 杨剑锋,乔佩蕊,李永梅,等.机器学习分类问题及算法研究综述[J].统计与决策,2019,35(6):36-40.

[2] 吕红燕,冯倩.随机森林算法研究综述[J].河北省科学院学报,2019,36(3):37-41.

[3] 宋欠欠,李轶群,侯艳,等.随机森林的变量捕获方法在高维数据变量筛选中的应用[J].中国卫生统计,2015,32(1):49-53.

[4] 何清,李宁,罗文娟,等.大数据下的机器学习算法综述[J].模式识别与人工智能,2014,27(4):327-336.

[5] Chen T,Guestrin C.XGBoost:A scalable tree boosting system[C]//Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining.2016:785-794.

[6] 徐慧丽.Stacking算法的研究及改进[D].广州:华南理工大学,2018.

[7] GB/T 7714Liang W,Luo S,Zhao G,et al.Predicting hard rock pillar stability using GBDT,XGBoost,and LightGBM algorithms[J].Mathematics,2020,8(5):765.

[8] 谢勇,项薇,季孟忠,等.基于XGBoost和LightGBM算法预测住房月租金的应用分析[J].计算机应用与软件,2019,36(9):151-155+191.

[9] 张远森.基于神经网络的二手车价格评估模型[D].天津:天津大学,2018.

[10] Bernard S,Adam S,Heutte L.Dynamic random forests[J].Pattern Recognition Letters,2012,33(12):1580-1586.

[11] 王晓晖,张亮,李俊清,等.基于遗传算法与随机森林的XGBoost改进方法研究[J].计算机科学,2020,47(S2):454-458+463.

[12] 王健峰,张磊,陈国兴,等.基于改进的网格搜索法的SVM参数优化[J].应用科技,2012,39(3):28-31.

[13] Li Guoqing,Wang Wanliang,Zhang Weiwei,Wang Zheng,Tu Hangyao,You Wenbo. Grid search based multi-population particle swarm optimization algorithm for multimodal multi-objective optimization[J].Swarm and Evolutionary Computation,2021,62(prepublish).

……下转第188页

R1、R、R3，进入Prepare阶段；3)Prepare：R0、R1、R2、R3互相广播，告知其他节点已收到投票请求，并进入了Prepare阶段；4)Commit：处在Prepare阶段的节点收到2/3节点的返回后表示都在Prepare阶段，则广播自己进入Commit阶段；5)Reply：在Commit阶段的节点收到其他2/3节点都进入Commit阶段了，则对C进行反馈，宣布自己的投票结果。

2.2 Move编程语言

Move是为数字资产而生的智能合约平台型语言，是一种新的编程语言，用于在Libra区块链中实现自定义交易逻辑和智能合约。Move语言总结为以下特点：每个资源有且只有一个唯一的所有者，不允许移动或复制资源，目的是防止数字资产重复和丢失，以此来确保交易的安全性[6]。

3 总结

2008年，比特币的横空出世，开启了人类对数字加密货币的探索之旅。经历了十余年的演变，到现如今各国央行第一次集体接受严肃的全球化数字货币申请，也是全球第一次集体从实践角度系统评估数字加密货币带来的影响。经历了数年的数字加密货币的演变，各国央行推行由国家信用做背书的数字货币的热潮汹涌而至，新一轮的货币战争正在进行之中。

参考文献

[1] 蒋鸥翔,张磊磊,刘德政.比特币、Libra、央行数字货币综述[J].金融科技时代,2020(2):57-68.

[2] 杨晓晨,张明.Libra:概念原理、潜在影响及其与中国版数字货币的比较[J].金融评论,2019(4):54-66+125.

[3] 李彬.浅谈非对称加密方式及其应用[J].信息记录材料,2021(1):214-215.

[4] 贺元香,夏甜,史宝明.区块链核心技术探究[J].兰州文理学院学报(自然科学版),2020,34(6):92-98.

[5] 郭亚宁,宋佳明.区块链在数字货币应用的可行性浅析[J].互联网天地,2020(12):27-31.

[6] 刘琴,王德军,王潇潇,等.法律合约与智能合约一致性综述[J].计算机应用研究,2021(1):1-8.

Analysis of the Two Cryptocurrencies of Bitcoin and Libra

GUO Ya-ning,YIN Ya-li

(Data Research Center of China Academy of Information and Communication, Beijing 100000)

Abstract: This article first sorts out the general characteristics of the two cryptocurrencies of Bitcoin and Libra, and strives to make readers have a deeper understanding and understanding of cryptocurrencies. Based on the existing public academic papers, a summary of the development process of cryptocurrency, this paper explores the technology and application characteristics of bitcoin and Libra, in the hope of helping readers to further understand cryptocurrency, looking forward to the earth shaking changes it brings to the digital economy.

Key words: Digital economy; Cryptocurrency; Bitcoin; Blockchain; Libra

……上接第93页

Prediction of Used Car Price Based on Random Forest and XGBoost

ZHENG Jie

(North China University of Technology, Beijing 100144)

Abstract: In order to better predict the price of second-hand cars, according to the second-hand car data set, the random forest and XGBoost algorithm are combined to make predictions. First, the original data is preprocessed, singular values and missing values are preprocessed, and the data is structured. Then use the random forest algorithm for variable selection, and select the variable with a score that is not zero according to the score output of the random forest as the feature variable of the predicted price. The three algorithms of XGBoost, GBDT and lightGBM are searched to find the optimal parameters. After comparison, the XGBoost algorithm is selected to predict the used car price to obtain the optimal used car price prediction result.

Key words: Used car price forecast; Random Forest; Grid Search; XGBoost