文章编号:2096-1472(2024)01-0049-06
DOI:10.19644/j.cnki.issn2096-1472.2024.001.012

基于 5CV-Optuna-LightGBM 回归模型的数据预测方法

顾 靓,谈子楠,荣 静

(扬州大学广陵学院,江苏 扬州 225000)

✉ 1649790465@qq.com; 2318147480@qq.com; 060096@yzu.edu.cn

摘 要:为解决各类复杂的数据预测问题,文章提出以五折交叉验证(5CV)、Optuna 超参数优化和 LightGBM 回归预测模型为基础的 5CV-Optuna-LightGBM 混合回归预测模型。采用影响二手车价格的因素数据集,首先进行数据预处理与 Pearson 相关性分析,确定 37 个特征指标。其次通过 L1 正则化对模型进行降噪处理,并利用交叉验证和 Optuna 算法不断优化模型,最终得到在 5CV-Optuna-LightGBM 回归预测模型下的数据预测结果。从准确率、花费时间等多个评价指标出发,开展实验分析模型的预测效果,得到准确率为 99.433%、花费时间为 15 s、平均绝对误差为 0.306%的结果,与其他模型对比,其预测值更加准确、建模效率更高、拟合度更高。

关键词:Pearson;五折交叉验证;Optuna;LightGBM;正则化

中图分类号:TP391 **文献标志码**:A

Data Prediction Method Based on 5CV-Optuna-LightGBM Regression Model

GU Liang, TAN Zinan, RONG Jing

(*Guangling College, Yangzhou University, Yangzhou 225000, China*)

✉ 1649790465@qq.com; 2318147480@qq.com; 060096@yzu.edu.cn

Abstract: In order to solve various complex data prediction problems, this paper proposes a 5CV-Optuna-LightGBM mixed regression prediction model based on five-fold cross validation (5CV), Optuna hyper-parameter optimization, and LightGBM regression prediction model. Data preprocessing and Pearson correlation analysis are first conducted on a dataset of factors that affect used car prices to determine 37 feature indicators. Next, the model is denoised through L1 regularization, and the cross-validation and Optuna algorithm are used to continuously optimize the model. Finally, the data prediction results under the 5CV-Optuna-LightGBM regression prediction model are obtained. Experiments based on multiple evaluation indicators of accuracy, time consumption, and average absolute error, are conducted to analyze the predictive performance of the model. The results show that the accuracy is 99.433%, the time spent is 15 seconds, and the average absolute error is 0.306%. Compared with other models, the proposed method has more accurate predicted values, higher modeling efficiency and fitting degree.

Key words: Pearson; five-fold cross-validation; Optuna; LightGBM; regularization

0 引言(Introduction)

机器学习模型相对于传统模型来说,其学习能力和泛化能力更强,占用的内存更低,训练速度更快而且效率也更高。随着机器学习的兴起,大量学者尝试建立相关的机器学习模型预测各种类型的数据,达到节约时间和经费的目的。与依赖于已知模式推导的线性回归等传统模型不同,机器学习模型不需要推导参数化详细的模型方程[1]。

本文以五折交叉验证、Optuna 超参数优化和 LightGBM

收稿日期:2023-05-09

基金项目:教育部产学合作协同育人项目(202002320005);江苏省高校哲学与社会科学专题项目(2022SJSZ1130)

回归预测模型为基础，建立 LightGBM 混合模型，命名为 5CV-Optuna-LightGBM 回归预测模型。为探讨 5CV-Optuna-LightGBM 回归预测模型的适用性，本文采用影响二手车价格的因素数据集对二手车价格进行预测。对比常用的 5CV-LightGBM 回归预测模型和最优尺度回归模型，本文采用的 5CV-Optuna-LightGBM 回归预测模型占用的内存更低、预测值更加准确、建模效率更高及拟合度更高。

1 预备知识(Preliminary knowledge)

LightGBM 是一种机器学习算法，自 2017 年被首次提出以来，得到了广泛的研究和应用。KE 等[2]提出了一种高效的基于决策树的梯度提升算法，采用多种优化技术提高了算法效率和泛化性能；SHEHADEH 等[3]建议使用修正决策树(MDT)、LightGBM 和 XGBoost 回归模型预测建筑设备的残值，提高了准确性并激发机器学习的潜力。除此之外，SRINIVAS 等[4]利用优化的 XGBoost 分类器，使用超参数优化技术(OPTUNA)对超参数进行适当的调整，并使用五种指标评估系统的效率，证明该模型的预测结果更好。

1.1 LightGBM 回归预测模型

1.1.1 LightGBM 算法

LightGBM 是一个实现梯度提升决策树(Gradient Boosting Decision Tree，GBDT)算法的框架[5]，基于直方图的决策树算法，通过基于梯度的单边采样算法 GOSS(Gradient-based One-Side Sampling)和互斥特征捆绑算法 EFB(Exclusive Feature Bundling)改进，支持高效率并行训练，具有速度快、节省内存、泛化能力较好等优点[6]。直方图算法把连续的浮点特征值离散化成 k 个整数，由此寻找最优切分点并取得最大增益，k 越小，其拟合准确度越低。利用直方图可以进行差分加速提高运算速度。在此基础上，LightGBM 使用按叶生长(Leaf-Wise)的算法[7]降低了模型损失。此外，LightGBM 需要设置一个决策树的最大深度用于分裂增益最大的结点，避免分裂的次数增加而发生过拟合的情况。在优化改进中，LightGBM 采用 GOSS 算法保留大梯度样本，对小梯度样本随机采样，减少训练误差；采用 EFB 算法将互斥特征进行合并，降低特征维度。

LightGBM 中最重要的是模型训练。模型训练主要通过以下几个步骤进行参数设置[8]。(1)数据收集：收集影响二手车价格的因素数据；(2)特征工程：寻找能最大限度地反映因变量本质的自变量分类原始数据；(3)模型训练：不断训练数据，只有达到规定的迭代次数或者迭代过程，收敛才能停止；(4)交叉验证和模型评估：待 LightGBM 模型达到最优后，通过模型评测对样本集和测试集进行模型检验，观察预测结果是否符合真实值。若数据不符合，则重新分类原始数据，重复上述步骤直到得出预期结果。

1.1.2 L1 正则化

LightGBM 在每次迭代时，会根据结果对样本进行权重调整，随着迭代次数的增加，模型偏差不断降低，导致模型对噪声越来越敏感。$L1$ 正则项将回归模型(regression_$L1$)作为目标函数(objective)，通过参数稀疏化进行特征选择、降低噪声。随着正则项不断增大，相应变量系数不断缩减，直至为 0。剔除零值特征，减少 LightGBM 预测误差，损失函数达到全局最小值。损失函数 L 的计算公式如下：

$$L=\min\frac{1}{2m}\sum_{i=1}^{m}(f(x)-y^{(i)})^2+\mu\|\boldsymbol{w}\|_1 \tag{1}$$

其中，$L1$ 正则项为 $\mu\|\boldsymbol{w}\|_1$，即权重向量中各元素的绝对值之和。在 $L1$ 正则项回归模型中，μ 直接决定进入模型的变量个数，影响模型回归的准确性。

1.2 五折交叉验证

模型应用于验证数据中的评估常用的是交叉验证，又称循环验证[9]。特征交叉通过合成特征在多维特征数据集上进行非线性特征拟合，从而提高模型的准确性，防止过拟合。原始数据分成 k 组不相交的子集，每个子集数据抽出 m 个训练样例。在训练样例中随机抽取 1 组子集作为一次验证集，剩下的 $k-1$ 组子集数据作为训练集，每组子集都经过一次验证，得到 k 个模型。

假设数据集有特征 x_1 和 x_2，那么引入交叉特征值 x_3，使 $x_3=x_1x_2$，最终表达式如下：

$$y=b+w_1x_1+w_2x_2+w_3x_3 \tag{2}$$

1.3 Optuna 超参数自动优化

Optuna 是一种自动化软件框架，能对模型超参数进行优化[10]。Optuna 可以通过选择多种优化方式确定最佳超参数，例如网格搜索、随机搜索和贝叶斯优化等。在 Optuna 的优化程序中，三个核心的概念分别为目标函数(objective)、单次试验(trial)和研究(study)。在机器学习中，为了找到最优的模型参数，研究人员需要定义一个待优化的函数 objective，这个函数的输入是模型参数(也就是参/超参数)，输出是针对这些模型参数的模型效果评估指标。对于每组参数，研究人员需要进行一次 trial，通过贝叶斯优化或网格搜索等优化算法，探索参/超参数的范围。优化算法需要 study 对象进行管理和控制试验的次数、参数的探索范围等并记录下来，从而确定最优的模型参数组合。在优化过程中，Optuna 利用修剪算法删除对分类作用小的过程，并通过反复调用和评估不同参数值的目标函数降低过拟合概率，获得最优解，降低误差。

2 数据处理(Data processing)

本文实验选取影响二手车价格的因素数据集，数据集样本共有 30 000 个。为方便处理数据，需要进行如下操作：提取 tradeTime、registerDate、licenseDate 等日期指标的年月日数值；anonymousFeature11 数据表现为 1+2、2+3 等六种字符串，按顺序用数字 1～6 对这六种字符串数据进行标签编码。对其他数据按照本文 2.1 至 2.4 章节的步骤进行处理。

2.1 填补缺失值

条形密度图(图 1)显示主要特征的缺失率，图中的空白越多，代表缺失的数据越多。部分二手车价格数据变量缺失率高于 80%，直接去除会造成数据的严重浪费等问题。对高于 80%的部分特征，如匿名特征(anonymousFeature4)等，则直接删除；余下的缺失率低于 80%的部分特征，如 cityid(车辆所在城市 id)等，根据样本之间的相似性进行众数填充[11]。

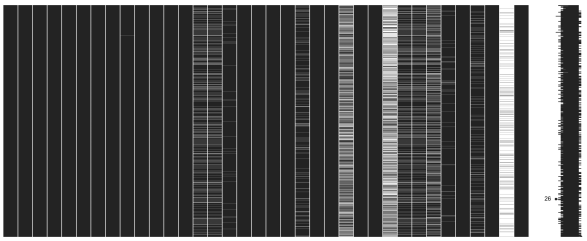
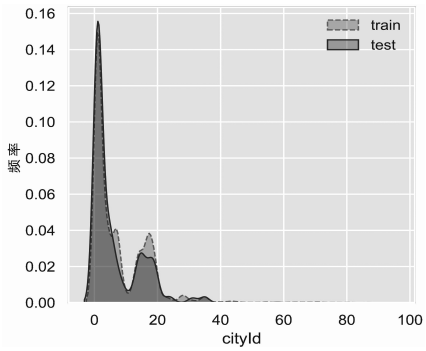


图 1　条形密度图

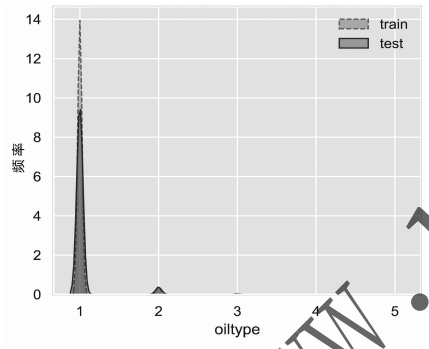
Fig. 1 Bar density map

2.2　长尾特征处理

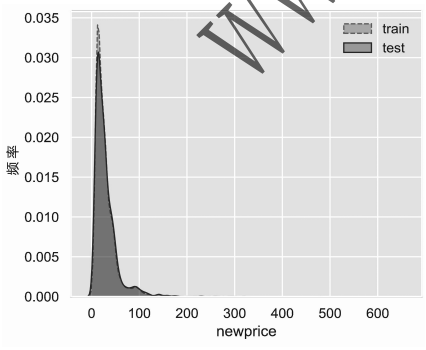
长尾是指某个或某几个连续变量的数值分布差别很大,呈现长尾图样式(图 2)。



(a)车辆所在城市 id



(b)燃油类型



(c)新车价

图 2　二手车价格指标长尾特征

Fig. 2 Long-tail characteristics of used car price indicators

对数变换的目标是帮助稳定方差,始终保持数据分布接近于正态分布,使得数据与分布的平均值无关。通过对数变换对 cityid(车辆所在城市 id)、oiltype(燃油类型)、newprice(新车价)等长尾特征进行处理[12]。

2.3　去除异常值

为了确保后续预测结果的准确性,训练数据必须符合实际情况。通过 matplotlib(2D 绘图库)的箱线图(图 3)分析指标,去除数值大于或小于其整体数值(超出箱线图边距)的异常变量中的数据。

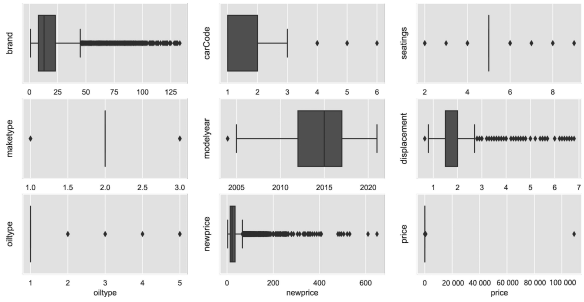


图 3　二手车价格指标箱线图

Fig. 3 Box plot of used car price indicators

2.4　相关性分析

通过 Pearson 相关系数对时间特征进行相关性分析。Pearson 相关系数是描述两个定距变量间联系的紧密程度和线性相关关系的参数[13]。通过 Pearson 相关系数可探求影响二手车价格变量之间的相关性,其计算公式如下:

$$r=\frac{N\sum x_i y_i-\sum x_i\sum y_i}{\sqrt{N\sum {x_i}^2-\left(\sum x_i\right)^2}\sqrt{N\sum {y_i}^2-\left(\sum y_i\right)^2}},i=1,2,\cdots,N \tag{3}$$

其中:N 表示变量个数,x、y 表示变量的观测值。相关系数 r 的绝对值越大,其相关性越强;当 $r\in(0,1]$时,表示 x 与 y 呈正相关,当 $r\in[-1,0)$时,表示 x 与 y 呈负相关,当 $r=0$ 时,x 与 y 无线性关系。

通过公式(3)对指标进行相关性计算,得到二手车价格指标热力图(图 4),可知 carid(车辆 id)和 model、brand(品牌 id)和 anonymousFeature5(匿名特征)、modelyear(年款)和 registerDate_year(注册日期)等变量之间的相关系数均大于 0.8,说明这些指标之间高度正相关,可以用其中一个指标代替其他指标[14]。

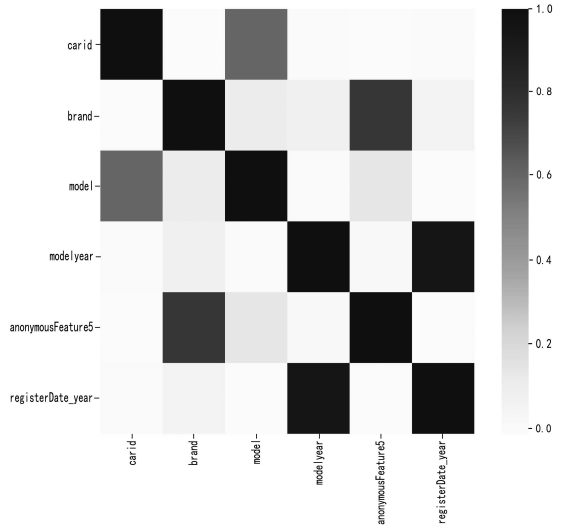


图 4　二手车价格指标热力图

Fig. 4 Heat map of used car price indicators

3 5CV-Optuna-LightGBM 回归预测模型(5CV-Optuna-LightGBM regression prediction model)

为了提升预测结果的精度,本文在原始 LightGBM 回归预测模型的基础上引入五折交叉验证和 Optuna 超参数自动优化,形成 5CV-Optuna-LightGBM 回归预测模型。

3.1 模型训练及参数优化

为提高 LightGBM 模型在默认参数下的诊断准确率,本文采用 split 方法将训练集和测试集划分为 5 份,每次迭代随机选取 4 份数据作为训练集,并对 LighGBM 模型中的 10 个特定超参数进行寻优,参数值在给定范围内随机生成,迭代 1 000 000 次。其中:参数 num_leaves 代表叶子节点数;参数 max_depth 代表树的深度,合适的树深度在一定程度上可以避免过拟合;参数 feature_fraction 代表子特征处理列采样;参数 bagging_fraction 代表建树的采样比例,具有泛化数据的能力;参数 bagging_freq 代表每 k 次迭代进行子采样;参数 learning_rate 代表学习率,如果设定过小,会导致梯度下降很慢,而设定过大又会跨过最优值,产生振荡;参数 min_child_weight 代表叶子节点中样本数目;参数 min_child_samples 代表叶子节点最小记录数;参数 seed 代表指定随机种子数。Optuna 高纬度参数关系图如图 5 所示。

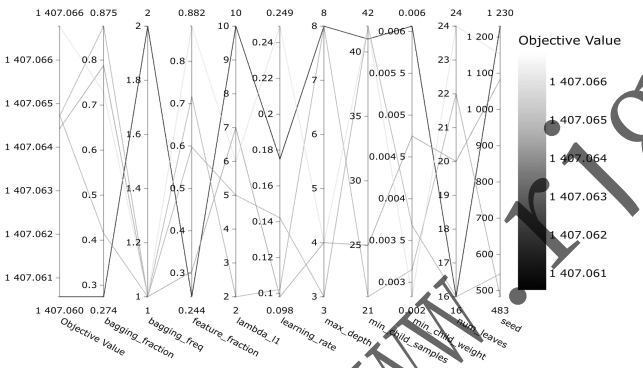


图 5 Optuna 高纬度参数关系图

Fig. 5 Map of Optuna high-latitude parameter relationships

图 5 中 Objective Value 最高值对应的各参数值为最优参数。

3.2 交叉验证 *ROC* 曲线

将训练集中 price 指标中低于 10 万元的标为 0,高于 10 万元的标为 1,利用受试者工作特征曲线 *ROC*(Receiver Operating Characteristic)判断交叉验证拟合度。其中,真阳性率(所有实际为阳性的样本被正确地判断为阳性的个数与所有实际为阳性的样本个数之比)为纵坐标,假阳性率(所有实际为阴性的样本被错误地判断为阳性的个数与所有实际为阴性的样本个数之比)为横坐标绘制的曲线,曲线越靠近左上方,则代表拟合程度越高。同时,二手车样本的检测数据变化较大,使用 *ROC* 曲线可以使数据分析更加稳定,交叉验证 *ROC* 如图 6 所示。

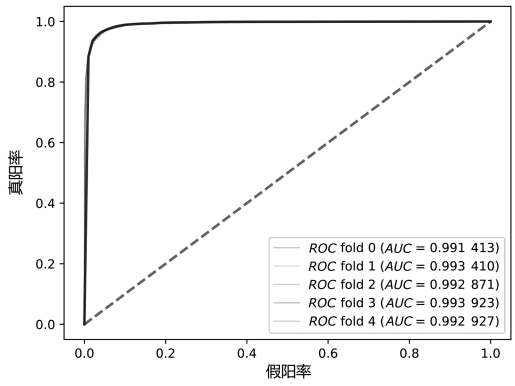


图 6 交叉验证 *ROC*

Fig. 6 Cross-validation *ROC*

通过 *ROC* 曲线分析,表明交叉验证下 *ROC* 均大于 0.99,模型性能优良。

3.3 评价指标

本文所采用的模型评价指标有平均绝对百分误差(*MAPE*)、对称平均绝对百分误差(*SMAPE*)、平均绝对值误差(*MAE*)、均方误差(*MSE*)、均方根误差(*RMSE*)和准确率(*Accuracy*),代表对样本的整体预测的准确程度,其中真实值 $y=(y_1,y_2,\cdots,y_m)$,模型预测为 $\hat{y}=(\hat{y}_1,\hat{y}_2,\cdots,\hat{y}_m)$,其计算公式分别如下:

$$MAPE=\frac{1}{m}\sum_{i=1}^{m}\left|\frac{\hat{y}_i-y_i}{y_i}\right| \tag{4}$$

$$SMAPE=\frac{1}{m}\sum_{i=1}^{m}\frac{|\hat{y}_i-y_i|}{(|\hat{y}_i|+|y_i|)/2} \tag{5}$$

$$MAE=\frac{1}{m}\sum_{i=1}^{m}|\hat{y}_i-y_i| \tag{6}$$

$$MSE=\frac{1}{m}\sum_{i=1}^{m}(y_i-\hat{y}_i)^2 \tag{7}$$

$$RMSE=\sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i-\hat{y}_i)^2} \tag{8}$$

$$\begin{cases}Accuracy=0.2\times(1-Mape)+0.8\times Accuracy_5\\ Ape=\dfrac{|\hat{y}-y|}{y}\\ Accuracy_5=\dfrac{count(Ape\leqslant 0.05)}{count(total)}\end{cases} \tag{9}$$

Accuracy 采用相对误差在 5%以内[$count(Ape\leqslant 0.05)$]的样本数量,其中 Ape 为相对误差。

3.4 模型性能对比

利用公式(4)至公式(9)分别求出 5CV-Optuna-LightGBM 回归预测模型、5CV-LightGBM 回归预测模型、Optuna-LightGBM 回归预测模型和最优尺度回归预测模型的评价指标,并进行模型对比,模型对比结果如表 1 所示。

表1　模型对比结果

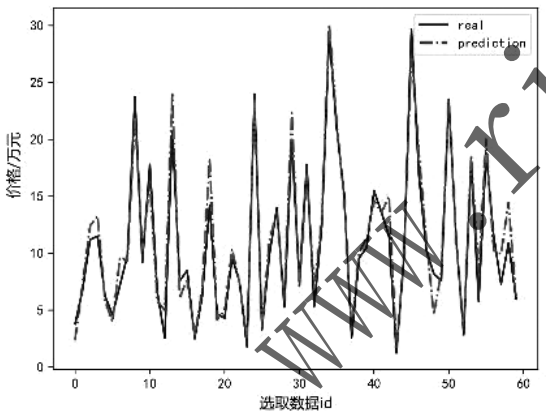
Tab.1 Model comparison

预测模型	评估指标						
	Accuracy/%	*MAPE*	*SMAPE*	*MAE*	*MSE*	*RMSE*	花费时间/s
5CV-Optuna-LightGBM 回归预测模型	99.433	0.003	0.003	0.005	0.000	0.011	15
5CV-LightGBM 回归预测模型	76.173	0.047	0.047	0.091	0.016	0.127	10
Optuna-LightGBM 回归预测模型	80.136	0.157	0.157	2.398	106.140	10.302	15
最优尺度回归预测模型	50.283	0.153	0.153	2.235	97.337	0.886	55

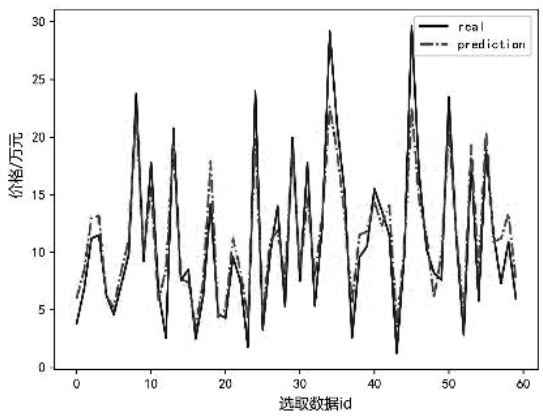
对比四个模型的准确率、平均决定值误差等误差指标和花费时间可以看出(表 1),5CV-Optuna-LighGBM 回归预测模型的准确率最高,达到了 99.433%。在预测值和实际值之间的差距方面,5CV-Optuna-LighGBM 回归预测模型的 *MAE* 等误差指标最小,预测最准确。在建模效率方面,5CV-Optuna-LightGBM 回归预测模型花费的时间最少、效率最高。

4　预测结果(Predicted results)

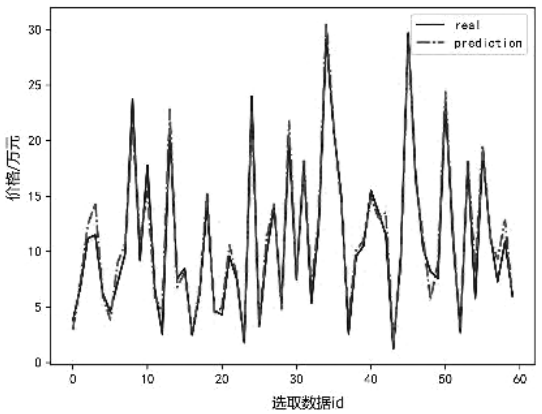
对二手车价格的预测值和真实值进行三次实验,提高结果的可靠性和泛化性,5CV-Optuna-LightGBM 回归预测模型预测值和真实值三次实验对比结果如图 7 所示。对比二手车价格的预测值和真实值分析得到数据基本一致,进一步验证模型的准确性。



(a)第一次实验



(b)第二次实验



(c)第三次实验

图 7　5CV-Optuna-LightGBM 回归预测模型预测值和真实值三次实验对比结果

Fig. 7 Comparison of the predicted and true values of the 5CV-Optuna-LightGBM regression prediction model over three experiments

利用 5CV-Optuna-LightGBM 回归预测模型求出二手车价格预测结果,二手车价格预测密集图如图 8 所示。

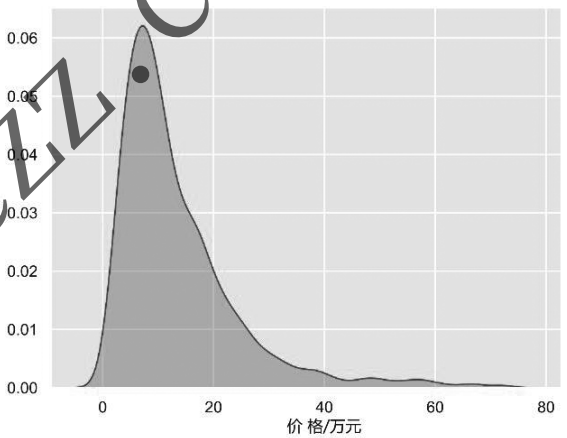


图 8　二手车价格预测密集图

Fig. 8 Intensive chart of used car price forecasts

图 8 中,预测价格在 10 万元左右的二手车成交频率最高,意味着大多数二手车的里程适中、座位数较少、车龄较老。通常,一些热门品牌和车型的二手车比较保价,而其他二手车的价格可能较低。10 万元左右的二手车大多为中档车型或一些比较受欢迎的品牌,主要分布在二线和三线城市。此外,过户次数、车辆生产国家、排量等因素也在一定程度上影响二手车价格。

5　结论(Conclusion)

本文在影响二手车价格因素数据集上研究 5CV-Optuna-LightGBM 回归预测模型对于预测类问题的优势,并对该模型进行有效性检验。从实验结果来看,基于 5CV-Optuna-LightGBM 回归预测模型可将预测精度提高到 99.433%,而预测时间降低到 15 s,平均绝对值误差(*MAE*)、均方误差(*MSE*)、均方根误差(*RMSE*)、平均绝对百分误差(*MAPE*)、对称平均绝对百分误差(*SMAPE*)分别减少到 0.005、0.000、0.011、0.003、0.003,预测结果更准确。此模型可以在其他经济市场中为产品估价提供一定参考意见。

但是，本研究仍存在一些不足。数据处理解释了可能的误差来源，总体误差是可控的，但即使对几种方法的结果进行了比较，误差也依然存在。每种方法都有其优缺点，因此在不同的背景下，评估哪种方法是适宜的，具有一定的挑战性。此外，相关拟合方法仍有改进和完善的空间，可以添加拟合方法进行拟合度对比，获取更高的拟合度。今后，值得探索的一个领域是研究多组平行对照组。

参考文献(References)

[1] DEY A. Machine learning algorithms: a review[J]. International journal of computer science and information technologies, 2016, 7(3): 1174-1179.

[2] KE G L, MENG Q, FINLEY T, et al. LightGBM: a highly efficient gradient boosting decision tree[C]//ACM. Proceedings of the 31st International Conference on Neural Information Processing Systems, New York: ACM, 2017: 3149-3157.

[3] SHEHADEH A, ALSHBOUL O, MAMLOOK R E A, et al. Machine learning models for predicting the residual value of heavy construction equipment: an evaluation of modified decision tree, LightGBM, and XGBoost regression[J]. Automation in construction, 2021, 129: 103827.

[4] SRINIVAS P, KATARYA R. HyOPTXg: OPTUNA hyper-parameter optimization framework for predicting cardiovascular disease using XGBoost[J]. Biomedical signal processing and control, 2022, 73: 103456.

[5] 黄伟，李阳. 基于 MCS-MIFS 与 LightGBM 的燃气轮机功率预测方法[J]. 电力科学与工程，2020，36(5)：23-31.

[6] 朱晓晨，尹奇志，赵福芹，等. 基于 LightGBM 的船舶航速预测模型[J]. 大连海事大学学报，2023，49(1)：56-65.

[7] 俞国燕，李少伟，董晔弘. 基于 XGBoost-LightGBM-LSTM 的风机齿轮箱轴承故障预警[J]. 轴承，2023(6)：140-145.

[8] 刘新伟，黄武斌，蒋盈沙，等. 基于 LightGBM 算法的强对流天气分类识别研究[J]. 高原气象，2021，40(4)：909-918.

[9] 杨雪宁，张永强，张选泽，等. 基于留一交叉验证法的 APSIM-Maize 产量模拟[J]. 作物学报，2023，49(10)：2854-2860.

[10] 方志，余粟. 基于 IGA-Optuna-LightGBM 的民航潜在旅客预测[J]. 国外电子测量技术，2022，41(10)：142-147.

[11] 曹幸运，曾鑫，吴刘仓. 偏正态数据下众数回归模型的统计诊断[J]. 高校应用数学学报 A 辑，2021，36(1)：9-20.

[12] FU Y, XIANG L Y, ZAHID Y, et al. Long-tailed visual recognition with deep models: a methodological survey and evaluation[J]. Neurocomputing, 2022, 509: 290-309.

[13] 陶洋，祝小钧，杨柳. 基于皮尔逊相关系数和信息熵的多传感器数据融合[J]. 小型微型计算机系统，2023，44(5)：1075-1080.

[14] ZHAO Y, FEDERICO A, PAITS T, et al. Animalcules: interactive microbiome analytics and visualization in R[J]. Microbiome, 2021, 9(1): 76.

作者简介：

顾　靓(2002-)，女，本科生。研究领域：计算机科学与技术。

谈子楠(2002-)，男，本科生。研究领域：计算机科学与技术。

荣　静(1990-)，女，硕士，讲师。研究领域：信息安全。本文通信作者。

(上接第 48 页)

受限于项目初期样本量，本文仅尝试了基于机器学习的分类方法，在后期样本量得到扩充后，将进一步研究深度学习的应用场景，以期进一步提高分类模型的泛化能力。对于建立的分类模型，将结合摄影测量技术，对采集到的摄影图像进行分类判别，实现对儿童颅骨畸形的早筛。

参考文献(References)

[1] FOSTER J, AHLUWALIA R, SHERBURN M, et al. Pediatric cranial deformations: demographic associations[J]. Journal of neurosurgery pediatrics, 2020, 26(4): 415-420.

[2] YILMAZ E, MIHCI E, NUR B, et al. Recent advances in craniosynostosis[J]. Pediatric neurology, 2019, 99: 7-15.

[3] BARBERO-GARCÍA I, LERMA J L, MIRANDA P, et al. Smartphone-based photogrammetric 3D modelling assessment by comparison with radiological medical imaging for cranial deformation analysis[J]. Measurement, 2019, 131: 372-379.

[4] YOU L, ZHANG G M, ZHAO W L, et al. Automated sagittal craniosynostosis classification from CT images using transfer learning[J]. Clinics in surgery, 2020, 5: 2746.

[5] SABETI M, BOOSTANI R, MORADI E, et al. Machine learning-based identification of craniosynostosis in newborns[J]. Machine learning with applications, 2022, 8: 100292.

[6] QUANG D, CHEN Y F, XIE X H. DANN: a deep learning approach for annotating the pathogenicity of genetic variants[J]. Bioinformatics, 2015, 31(5): 761-763.

[7] LEE M J, HONG H, SHIM K W, et al. Quantitative analysis and automatic classification of skull deformity based on combined two-and three-dimensional shape indices[C]// IEEE. Proceedings of the 2018 IEEE 15th International Symposium on Biomedical Imaging. Piscataway: IEEE, 2018: 994-997.

[8] FELZENSZWALB P F, HUTTENLOCHER D P. Efficient graph-based image segmentation[J]. International journal of computer vision, 2004, 59: 167-181.

[9] ZHU B H, JIAO J T, STEINHARDT J. When does the tukey median work? [C]//IEEE. Proceedings of the 2020 IEEE International Symposium on Information Theory. Piscataway: IEEE, 2020: 1201-1206.

[10] GANAPATHI RAJU V N, LAKSHMI K P, JAIN V M, et al. Study the influence of normalization/transformation process on the accuracy of supervised classification[C]// IEEE. Proceedings of the 2020 Third International Conference on Smart Systems and Inventive Technology. Piscataway: IEEE, 2020: 729-735.

[11] 陈婉琦，林勇. 基于集成学习的骨质疏松性骨折预测研究[J]. 中国医学物理学杂志，2021，38(2)：254-258.

作者简介：

张顺雨(1997-)，男，硕士生。研究领域：智能医学影像分析。

胡　骏(2001-)，男，本科生。研究领域：智能医学影像分析。

林　勇(1978-)，男，博士，副教授。研究领域：统计遗传学和智能医学信息处理。本文通信作者。