Statistics and Application 统计学与应用, 2026, 15(3), 153-164
Published Online March 2026 in Hans. https://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2026.153064

Hans汉斯

基于LightGBM模型的二手车价格预测研究

杨　越

重庆移通学院数字经济商学院，重庆

收稿日期：2026年2月25日；录用日期：2026年3月16日；发布日期：2026年3月26日

摘　要

针对我国二手车交易价格评估困难、市场透明度不足的问题，本文提出一种融合特征工程与集成学习的优化预测方法。首先对来自某交易平台的15万条二手车数据进行预处理，包括异常值修正、缺失值填充和特征衍生；接着通过相关性分析筛选出关键特征；然后对比XGBoost、随机森林、CatBoost和LightGBM四种模型的性能，发现经参数优化的LightGBM模型表现最佳，其平均绝对误差(MAE)为487.03元，决定系数(R^2)达0.9708，平均绝对百分比误差(MAPE)为13.20%；最后将该模型应用于5万条测试数据，生成价格预测及区间预测。实验表明，本文方法能有效提升二手车价格评估的准确性与可靠性。

关键词

二手车价格预测，特征工程，LightGBM，XGBoost，模型过拟合诊断

A Research Study on Used Car Price Prediction Based on the LightGBM Model

Yue Yang

Digital Economy Business School, Chongqing Yitong University, Chongqing

Received: February 25, 2026; accepted: March 16, 2026; published: March 26, 2026

Abstract

To address the difficulties in second-hand car price evaluation and the lack of market transparency in China, this paper proposes an optimized prediction method that integrates feature engineering with ensemble learning. First, 150,000 second-hand car data entries from a trading platform are preprocessed, including outlier correction, missing value imputation, and feature derivation. Next, key features are selected through correlation analysis. Then, the performance of four models—XGBoost, Random Forest, CatBoost and LightGBM—is compared, and it is found that the parameter-optimized LightGBM model performs the best, with a mean absolute error (MAE) of 487.03 yuan, a

文章引用: 杨越. 基于 LightGBM 模型的二手车价格预测研究[J]. 统计学与应用, 2026, 15(3): 153-164.
DOI: 10.12677/sa.2026.153064

coefficient of determination (R^2) of 0.9708, and a mean absolute percentage error (MAPE) of 13.20%. Finally, this model is applied to 50,000 test data entries to generate price predictions and confidence intervals. Experiments show that the proposed method can effectively improve the accuracy and reliability of second-hand car price evaluation.

Keywords

Used Car Price Prediction, Feature Engineering, LightGBM, XGBoost, Model Overfitting Diagnosis

1. 引言

近年来，我国汽车保有量持续增长，二手车交易市场日益活跃。根据中国汽车流通协会数据显示，2023 年我国二手车交易量已突破 1800 万辆，市场规模达万亿元级别。然而，与传统新车市场相比，二手车定价问题始终是制约市场健康发展的瓶颈[1]。

当前二手车定价主要面临三大挑战：第一，传统估价方法过度依赖评估师个人经验，主观性强且缺乏统一标准；第二，车辆信息不对称现象普遍存在，买方难以获取真实车况；第三，各交易平台采用不同的定价算法，导致同一车辆在不同平台的估值差异显著。这些因素不仅影响交易的公平性，也增加了市场摩擦成本。

在学术研究领域，二手车价格预测方法已从传统的重置成本法、年限折旧法逐步发展到基于机器学习的数据驱动方法。早期研究如吕劲(2019) [2]采用支持向量机(Support Vector Machine, SVM)模型，李富强等(2021) [3]使用深度神经网络(Deep Neural Network, DNN)，均取得了优于传统方法的预测精度。近期研究趋势更倾向于集成学习算法，郑婕(2021) [4]比较了 XGBoost (Extreme Gradient Boosting)、梯度提升决策树(Gradient Boosting Decision Tree, GBDT)和 LightGBM (Light Gradient Boosting Machine)三种模型，发现 XGBoost 在特定数据集上表现最优；崔四帅(2021) [5]则通过 Stacking 集成方法进一步提升了预测精度。

然而，现有研究仍存在以下不足：第一，对二手车特征的工程化处理不够深入，多数研究仅使用原始特征；第二，模型的可解释性较差，难以分析关键影响因素；第三，缺乏对大数据量场景下模型效率的考量。

本文针对上述问题，提出一套完整的二手车价格预测解决方案。主要贡献包括：

(1) 系统化的特征工程：通过对车辆使用特征(功率、里程、车龄等)进行非线性变换和交互特征构建，提取具有物理意义的衍生特征，增强了模型的表达能力。

(2) 多模型对比与优化：在统一实验框架下对比了 XGBoost、随机森林、CatBoost 和 LightGBM 四种主流集成算法，并通过参数调优和正则化技术获得了最优的 LightGBM 模型。

(3) 大规模数据验证：使用包含 15 万条训练样本和 5 万条测试样本的真实交易数据，验证了模型在大数据场景下的稳定性和实用性。

(4) 完整的应用体系：不仅提供预测模型，还建立了从数据预处理、特征工程到模型部署的完整流程，并输出包含区间的预测结果，为实际应用提供了可靠参考。

本文研究不仅为二手车交易提供了量化的价格评估工具，也为其他非标商品的价格预测问题提供了

可借鉴的方法论框架。后续章节将详细介绍数据处理方法、模型构建过程、实验结果分析以及实际应用价值。

2. 数据预处理与特征工程

2.1. 数据来源

数据来源于 Datawhale 与天池联合举办的二手车交易价格预测大赛，包含训练集 15 万条、测试集 5 万条，共 31 个特征变量，其中 15 个为匿名特征(v_0 至 v_14)。

来源网址：https://aistudio.baidu.com/aistudio/datasetdetail/25091。

2.2. 变量描述与数据清洗

本文训练数据包含 31 列变量信息，其中 15 列为匿名变量。这 31 个变量信息如表 1。

Table 1. Training set variable information
表 1. 训练集变量信息

变量名称	变量描述	变量名称	变量描述
SaleID	交易 ID，唯一编码	kilometer	已行驶公里 km，范围[0.5, 15]
Name	汽车交易名称，已脱敏	price	价格(预测目标)范围[11, 99999]
regDate	注册日期	regionCode	地区编码，已脱敏
model	车型编码，已脱敏	seller	销售方：个体 0，非个体 1
brand	汽车品牌，已脱敏	offerType	报价类型：提供 0，请求 1
bodyType	车身类型：八种	creatDate	汽车开始售卖日期
fuelType	燃油类型：七种	V 簇特征	V0~V14，15 个匿名特征
gearbox	变速箱：手动 0，自动 1	notRepairedDamage	是否有损坏：是 0，否 1
power	功率：范围[0, 19312]		

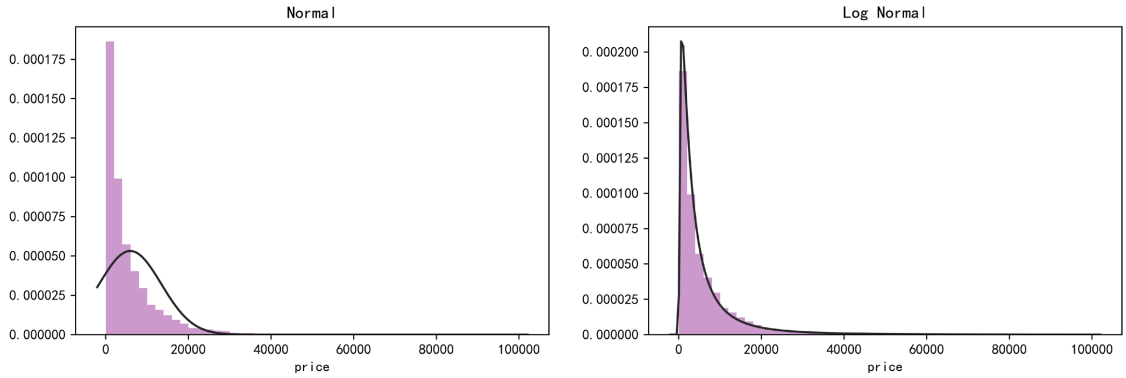
注册日期中存在“199910001”等不合理日期。汽车交易名称、车型编码、汽车品牌和地区编码都已脱敏处理转换为了对应的虚构数据。车身类型共有 0~7 八种。0 代表豪华轿车，1 代表微型车，2 代表厢型车，3 代表大巴车，4 代表敞篷车，5 代表双门汽车，6 代表商务车，7 代表搅拌车。燃油类型共有 0~6 七种，0 代表汽油，1 代表柴油，2 代表液化石油气，3 代表天然气，4 代表混合动力，5 代表其他，6 代表电动。

通过统计分析，发现 notRepairedDamage 变量存在三种类型值，150,000 辆二手车中有 111,361 辆二手车有损坏，约占总汽车的 74%，有 14,315 辆汽车没有损坏，约占总汽车的 10%，有 24,324 辆二手车卖家没有填写，占总汽车 16%。本文直接删除没有填写的信息。删除后数据形状：(125,676, 31)。

对于预测变量 price 总体分布概况如图 1。

从图中可以看出 price 存在异常值，对 price 进行箱线图截断。并且 price 不服从正态分布，所以在进行回归之前，进行对数转换。

训练集和测试集数据中变量 bodyType、fuelType 和 gearbox 都存在几千的缺失，本文对数值型特征按偏度选择均值或中位数填充，分类特征用众数填充。异常值处理中针对功率，参考文献一般车辆的功率在 100 千瓦至 200 千瓦，混合车在 300 至 400 千瓦之间，只有豪华跑车才能在 500 至 600 KW 之间，因此本文将超过 600 KW 的位置视为异常点。power 大于 600 KW 的记录，用中位数替换。



(a) (b)

Figure 1. Price distribution situation
图 1. Price 分布概况

2.3. 特征衍生

本文主要根据 regDate、power、kilometer 以及 15 个匿名特征衍生以下新特征(表 2)。

Table 2. New derived features
表 2. 衍生新特征

变量名称	变量描述
核心特征	
car_age	车龄 = 当前年份 − 注册年份。反映车辆的自然折旧规律。
power_per_km	单位里程功率(功率/里程)。衡量车辆在使用过程中的“动力效率”。
power_density	功率密度，用于衡量车辆的动力充沛程度。
value_score	基于功率、里程、车龄的加权综合评分。
power_km_score	功率与里程的综合评分。
power_age_interaction	功率与车龄的交互项。用于分析动力性能随车龄变化的规律。
对数变换特征	(用于处理偏态分布，降低极端值影响)
car_age_log	车龄取自然对数。
power_log	功率取自然对数。
kilometer_log	行驶里程取自然对数。
平方特征	(用于捕捉非线性关系)
car_age_squared	车龄的平方值。常用于捕捉车龄对价格的加速折旧效应。
power_squared	功率的平方值。用于捕捉功率对车辆价值的非线性增益效应。
主成分特征	(用于降维和信息提取)
v_pca_0, v_pca_1, v_pca_2, v_pca_3, v_pca_4	由 15 个匿名 V 系列特征经主成分分析(PCA)生成的前 5 个主成分，解释了 99.4%的方差。

对于 15 个匿名 V 系列特征，在缺乏官方业务文档的前提下，任何对匿名特征物理意义的推断均具有不确定性，若据此进行特征筛选可能导致信息损失或人为偏差。基于上述认识，本研究对 15 个匿名特征采取“完整性保留 + 降维补充 + 相关性选择”处理策略，即所有匿名特征全部保留，不基于主观判

断进行删除。为缓解多重共线性问题，本研究额外构造 5 个 PCA 主成分特征，与原始匿名特征共同构成增强特征集。PCA 累计方差贡献率达 99.4%，有效捕获了原始特征的主要变异信息。本研究不尝试对匿名特征进行业务语义标签化，将其作为纯数值特征在相关性特征选择后输入树模型，由模型自主学习特征与目标的关系。

2.4. 特征选择

计算各数值特征变量与变量 price 的 Pearson 相关系数，结果如图 2。

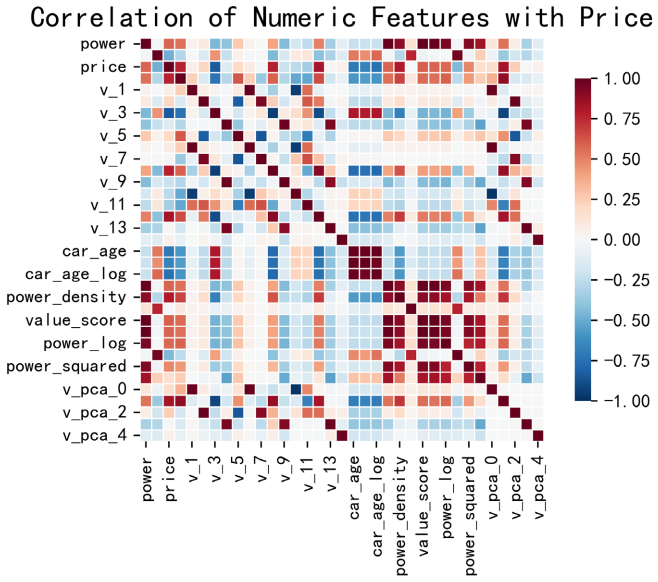


Figure 2. Correlation analysis
图 2. 相关性分析

从图中可以看出颜色越浅表示相关性越弱，颜色越深表示相关性越强。本文保留与 price 这个预测变量 Pearson 相关系数绝对值大于 0.1 的 26 个特征，包括：power、kilometer、v_0、v_2、v_3、v_4、v_5、v_8、v_9、v_10、v_11、v_12、car_age、car_age_squared、car_age_log、power_per_km、power_density、value_score、power_km_score、power_log、kilometer_log、power_squared、power_age_interaction、v_pca_0、v_pca_1 和 v_pca_2。

对于分类变量只提取 bodyType、fuelType、gearbox、notRepairedDamage、model 这 5 个变量。

3. 模型构建

3.1. 模型参数

本文运用 XGBoost、随机森林、CatBoost 和 LightGBM 四种模型。其中 XGBoost：采用 hist 树方法，n_estimators=1000；随机森林：n_estimators=300，max_depth=12；CatBoost：iterations=2000，depth=8；LightGBM：经参数调优(num_leaves=137, max_depth=9, learning_rate=0.02)，使用早停法(early_stopping)防止过拟合并添加 L1/L2 正则化(reg_alpha=0.05, reg_lambda=0.27)。CatBoost 对于分类特征原生支持，是 LightGBM 的直接竞争对手，但在本文中尽管全部分类特征入样进行训练，LightGBM 也要优于 CatBoost。

为了保证训练模型的可用性和预测的真实性，本文将在原训练集上训练模型，在原测试集上进行预测。按 7:1.5:1.5 的比例将原训练集随机划分为训练集、验证集与测试集。划分后的数据集信息如图 3。

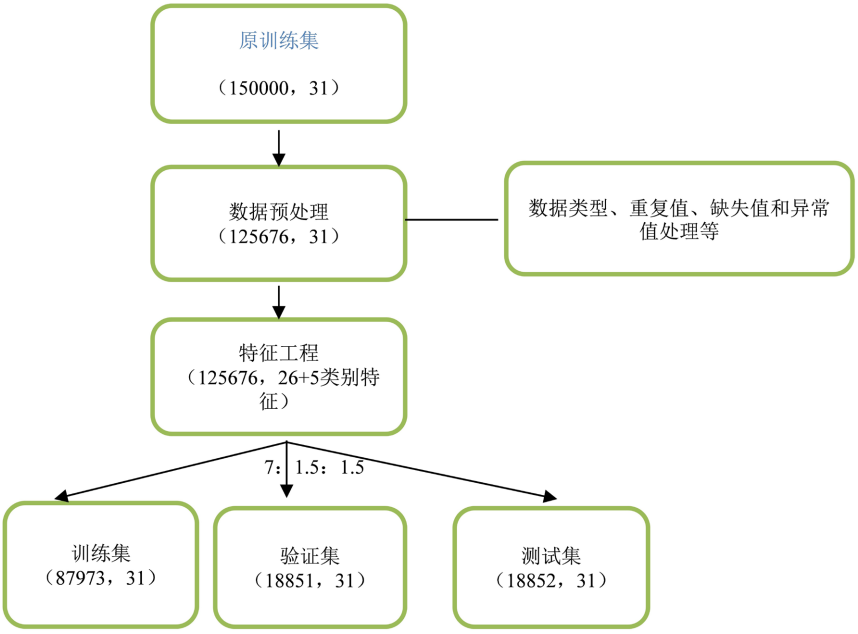


Figure 3. Data partitioning information
图 3. 数据划分信息

3.2. 评价指标

本文采用平均绝对误差(MAE)、均方根误差(RMSE)、决定系数(R^2)和平均绝对百分比误差(MAPE)进行模型评估。评估指标公式及意义如下：

(1) 平均绝对误差(MAE)

$$\mathrm{MAE}=\frac{1}{n}\sum_{i=1}^{n}\left|y_i-\hat{y}_i\right|,$$

平均绝对误差表示预测值与实际观测值之间的平均绝对差异。

(2) 均方根误差(RMSE)

$$\mathrm{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2},$$

均方根误差对异常值敏感，关注极端错误预测的代价较大的场景。

(3) 决定系数(R^2)

$$R^2=1-\frac{\sum_{i=1}^{n}\left(y_i-\hat{y}i\right)^2}{\sum i=1^{n}\left(y_i-\bar{y}\right)^2},$$

决定系数取值范围通常在[0, 1]或[−∞, 1]，值越大越好。等于 1 表示模型完美拟合数据。

(4) 平均绝对百分比误差(MAPE)

$$\mathrm{MAPE}=\frac{100\%}{n}\sum_{\mathrm{i}=1}^{n}\left|\frac{y_i-\hat{y}_i}{y_i}\right|,$$

平均绝对百分比误差是平均误差占真实值的百分比。值越小越好。

4. 实验结果与分析

4.1. 模型性能对比

对于 XGBoost、随机森林、CatBoost 和 LightGBM 这四种模型在训练集上进行模型训练，在验证集上进行参数调优，在测试集上模型表现效果如表 3。

Table 3. Model performance comparison
表 3. 模型性能对比

模型	MAE	RMSE	R^2	MAPE
XGBoost	585.71	1049.19	0.9640	14.91%
随机森林	615.51	1082.27	0.9617	16.01%
CatBoost	783.32	1309.08	0.9440	18.36%
LightGBM	487.03	945.46	0.9708	13.20%

从表中信息可以得到，XGBoost 模型在测试集上的平均绝对误差 MAE 为 585.71 元，这表示平均每个预测值偏离真实值 585.71 元，XGBoost 模型在测试集上的均方根误差 RMSE 为 1049.19 元，XGBoost 模型在测试集上的决定系数 R^2 为 0.9640，这表示可以解释约 96%的价格变异，XGBoost 模型在测试集上的平均绝对百分比误差 MAPE 为 14.91%，这表示平均误差占真实价格的 14.91%；随机森林模型在测试集上的平均绝对误差 MAE 为 615.51 元，这表示平均每个预测值偏离真实值 615.51 元，随机森林模型在测试集上的均方根误差 RMSE 为 1082.27 元，随机森林模型在测试集上的决定系数 R^2 为 0.9617，这表示可以解释约 96%的价格变异，随机森林模型在测试集上的平均绝对百分比误差 MAPE 为 16.01%，这表示平均误差占真实价格的 16.01%；CatBoost 模型在测试集上的平均绝对误差 MAE 为 783.32 元，这表示平均每个预测值偏离真实值 783.32 元，CatBoost 模型在测试集上的均方根误差 RMSE 为 1309.08 元，CatBoost 模型在测试集上的决定系数 R^2 为 0.9440，这表示可以解释约 96%的价格变异，CatBoost 模型在测试集上的平均绝对百分比误差 MAPE 为 18.36%，这表示平均误差占真实价格的 18.36%；LightGBM 模型在测试集上的平均绝对误差 MAE 为 487.03 元，这表示平均每个预测值偏离真实值 487.03 元，LightGBM 模型在测试集上的均方根误差 RMSE 为 945.46 元，LightGBM 模型在测试集上的决定系数 R^2 为 0.9708，这表示可以解释约 97%的价格变异，LightGBM 模型在测试集上的平均绝对百分比误差 MAPE 为 13.20%，这表示平均误差占真实价格的 13.20%，在二手车这类非标品定价中，属于可接受范围(一般商业场景 MAPE < 15%即为可用)。这三个模型的均方根误差 RMSE 都大于平均绝对误差 MAE，这表明测试集数据中存在较大的预测误差样本，但 LightGBM 的 RMSE 最低，说明它对极端误差的控制最好。从这四个指标都可以看出 XGBoost 和 LightGBM 在各项指标上均表现较好，其中 LightGBM 最优。

下面对于该 LightGBM 模型进行模型性能诊断，检查是否存在过拟合情况。87,973 个训练集样本的平均绝对误差 MAE 为 386.24，平均绝对百分比误差 MAPE 为 8.99%；18,852 个测试集样本的平均绝对误差 MAE 为 487.03，平均绝对百分比误差 MAPE 为 13.20%，残差均值为 86.64，残差标准差为 940.26。测试/训练 MAE 比值为 1.259，测试/训练 MAPE 比值为 1.467，MAE 比值 > 1.1，略有差异，但可接受，模型不存在过拟合情况。LightGBM 模型诊断报告情况如图 4。

残差 vs 预测值图用于检查异方差性即误差是否随预测值变化。从图中可以看出残差随机分布在 0 线上下，无明显趋势。实际 vs 预测图用于判断系统偏差，点紧密围绕在 y = x 对角线附近。误差分布图用于检验残差正态性。从图中可以看出效果比较理想，残差呈近似对称的钟形分布，均值为 0。百分比误差

分布图用于评估相对误差的稳定性。百分比误差集中在 0 附近。以上结果表明 LightGBM 模型性能良好且没有过拟合。

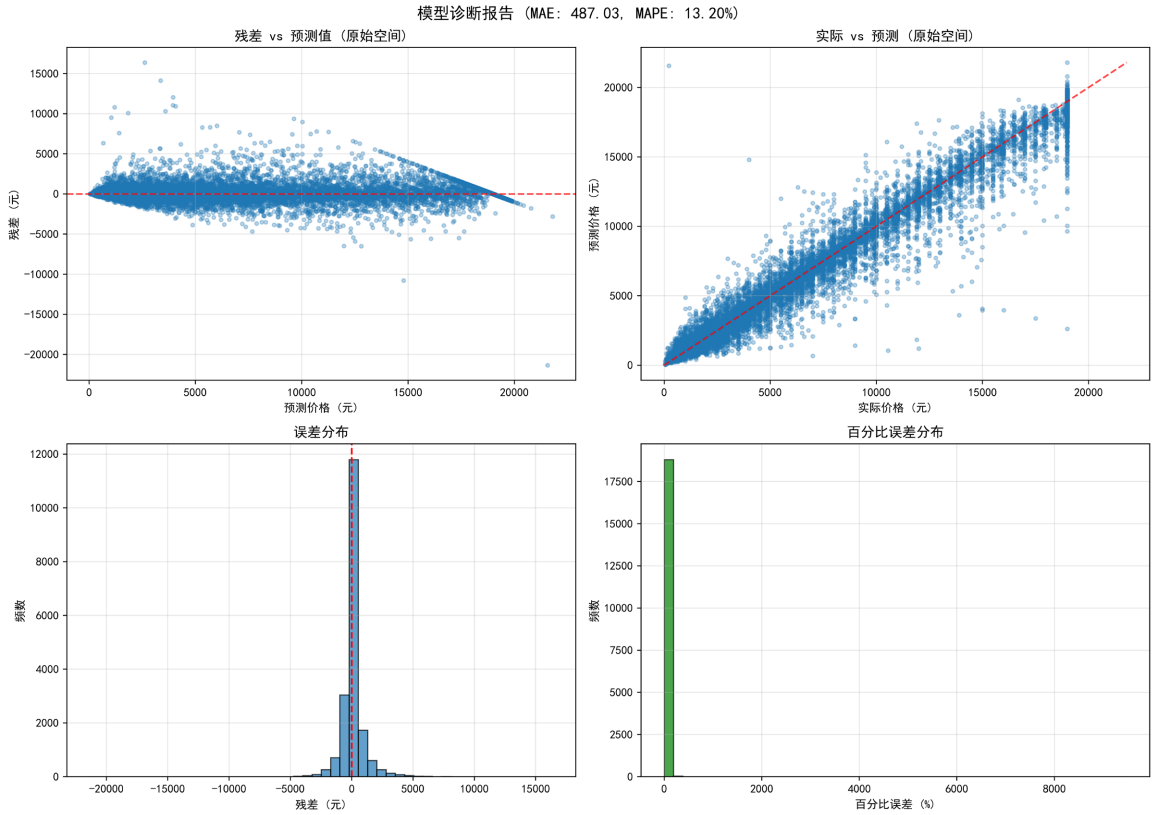


Figure 4. LightGBM model diagnostic report
图 4. LightGBM 模型诊断报告

4.2. 预测结果

本文仅展示前 20 辆二手车预测价格情况(表 4)。

Table 4. LightGBM model prediction results
表 4. LightGBM 模型预测结果

SaleID	predicted_price	price_category	price_lower_bound	price_upper_bound
150000	16789.58	90%~95%	14575.08	19004.08
150001	365.80	最低 25%	317.55	414.05
150002	3571.16	50%~75%	3100.13	4042.18
150003	8908.26	75%~90%	7733.28	10083.23
150004	612.87	最低 25%	532.04	693.71
150005	1161.71	最低 25%	1008.48	1314.93
150006	4542.83	50%~75%	3943.64	5142.01
150007	6723.13	50%~75%	5836.37	7609.89

续表

150008	1750.29	25%~50%	1519.43	1981.15
150009	2117.62	25%~50%	1838.31	2396.92
150010	1425.49	25%~50%	1237.47	1613.50
150011	5824.42	50%~75%	5056.19	6592.64
150013	11102.51	75%~90%	9638.11	12566.90
150014	10720.29	75%~90%	9306.31	12134.26
150015	414.60	最低 25%	359.92	469.29
150016	560.13	最低 25%	486.25	634.01
150017	3687.23	50%~75%	3200.89	4173.56
150018	10291.83	75%~90%	8934.37	11649.29
150019	1345.76	25%~50%	1168.26	1523.27
150020	3179.22	50%~75%	2759.89	3598.55

从图中可以看出，对于预测通过 SaleID 锁定车辆，本文不仅预测了车辆的价格，并且给出了车辆价格属于哪个分位数类别以及价格的区间。本文采用基于平均绝对百分比误差(MAPE)的经验区间方法，为二手车价格预测提供参考范围。尽管该方法在统计严谨性上有所不足，但其在本研究场景下的应用具有以下合理性：

(1) 业务可解释性：MAPE 是二手车价格评估中广泛使用的误差指标，“预测值 ±MAPE%”的表示方式直观易懂，便于业务人员理解和使用。

(2) 计算效率：相较于分位数回归或基于残差正态假设的方法，本方法计算简单，适合快速迭代的实验环境。

(3) 保守性估计：本方法构建的区间宽度(2 × MAPE)通常大于基于正态假设的 95%置信区间，提供了更为保守的价格波动范围，在风险控制场景下具有实用价值。

对于 41,969 条测试数据(原本 5 万条测试数据，删除变量 notRepairedDamage 中没有填写的数据)预测显示结果见链接：https://pan.baidu.com/s/1FsWFDBpd_76NqIPEPMKgdg?pwd=ikuh。

对于整个原测试集二手车价格预测结果分析图如图 5。

本文最终预测数量为 41,969 辆二手车，价格预测范围在[39.49, 24,663.66]，平均价格为 5019.63 元，价格中位数为 3041.60 元，价格标准差为 5029.58 元。从图中也可以看出车辆大多数在低价位。价格区间使用分位数自动划分，划分结果为最低 25%：价格区间在¥40~¥1237，该区间有 10,493 辆车；25%~50%：价格区间在¥1237~¥3046，该区间 10,492 辆车；50%~75%：价格区间在¥3046~¥6993，该区间有 10,492 辆车；75%~90%：价格区间在¥6993~¥13,486，该区间有 6295 辆车；90%~95%：价格区间在¥13,486~¥17,236，该区间有 2098 辆车和最高 5%：价格区间在¥17,236~¥22,741，该区间有 2099 辆车。

本研究采用 shap (Shapley Additive Explanations)方法对 LightGBM 二手车价格预测模型进行可解释性分析。shap 摘要图(Summary Plot)能够直观展示各特征对模型预测结果的贡献程度及影响方向，特征重要性和特征效应图，展示了每个特征的 shap 值的分布情况。由于测试集近 5 万条数据，本文按价格区间分层抽样，每个区间抽取 200 个样本即一共 1200 个样本进行 shap 值计算。得到测试集上摘要图如图 6。(这里只展示前 20 个特征情况)：

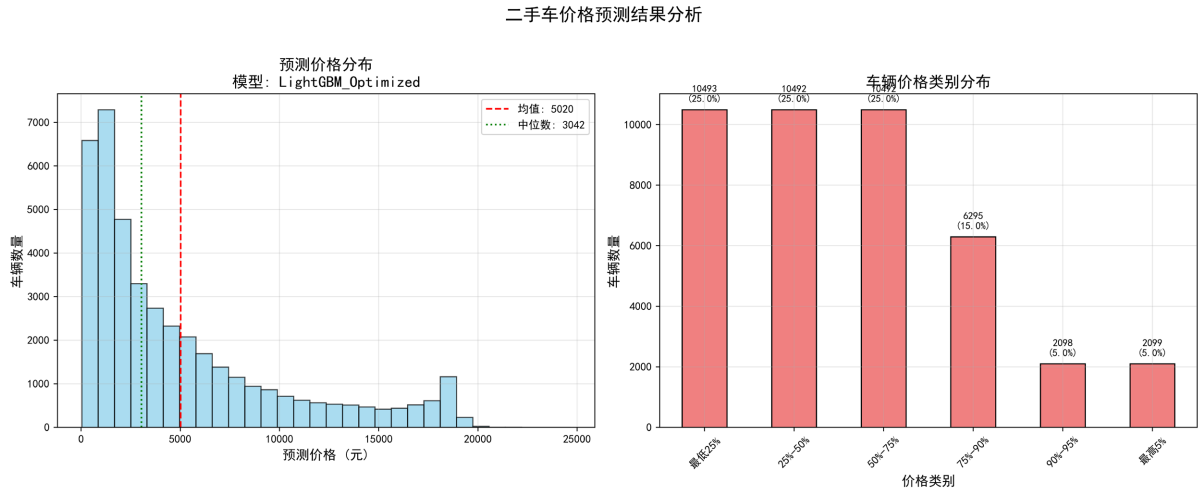


Figure 5. Used car price prediction results
图 5. 二手车价格预测结果

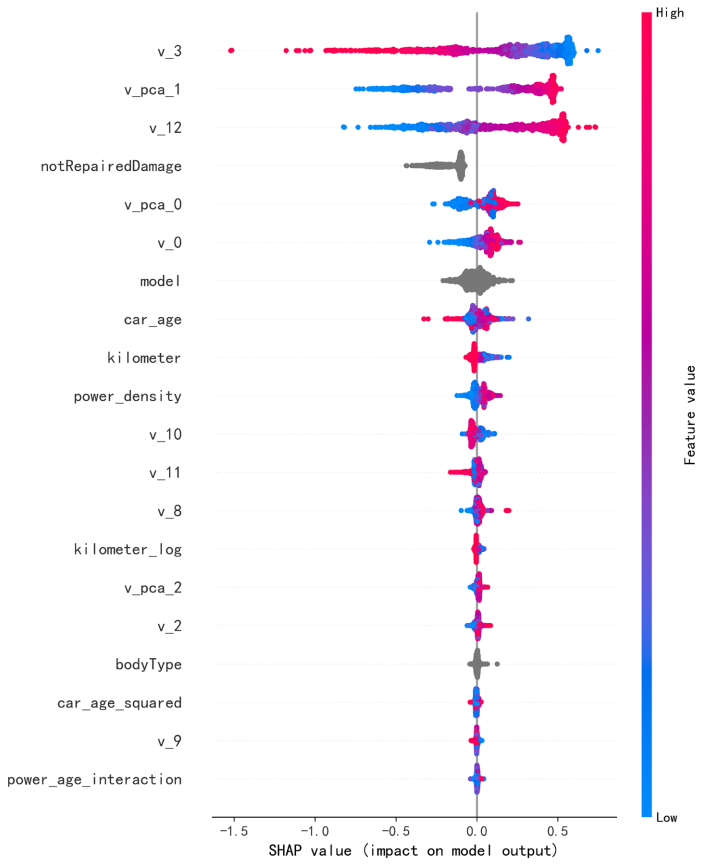


Figure 6. Summary plot
图 6. 摘要图

如图所示，纵轴按特征重要性从上到下排列，特征值的大小由颜色表示，粉红色点表示该特征值较大，蓝色点表示该特征值较小。横轴 shap 值显示每个特征对预测结果的影响大小，点越远离中心线(零点)，表示该特征对模型输出的影响越大，正的 shap 值表示正面影响即拉高价格，负的 shap 值表示负面

影响，拉低价格。

由图可知，模型预测能力主要由前 5 个特征贡献，按重要性排序依次为：v_3、v_pca_1、v_12、notRepairedDamage、v_pca_0。其中，notRepairedDamage 作为分类变量位列第 4，表明车辆损伤状况对二手车价格具有显著影响，有损伤的点全部集中在左侧，对价格造成负面影响，损伤拉低价格，这与二手车市场的业务常识高度吻合；v_3 作为重要性最高的特征，呈现出独特的非线性影响模式：v_3 几乎粉红色点都集中在中心线左侧蓝色点都集中在中心线右侧，特征值较小时，当特征值较小时，shap 值为较大的正值，对价格产生显著的推高作用；随着特征值增大，正面影响逐渐减弱，当特征值达到较高水平时，shap 值转为负值且绝对值较大，对价格产生强烈的拉低效应。这种“低值推高、高值拉低”的非单调关系表明，v_3 可能代表了某个具有最优区间的特征变量；v_pca_1、v_12、v_pca_0 等特征呈现出一致的线性影响规律：蓝色点(低特征值)集中在 shap 负值区域，粉红色点(高特征值)集中在 shap 正值区域。这表明这些特征的取值与价格呈正相关关系—特征值越低，越拉低价格；特征值越高，越推高价格。其他降维特征如 v_pca_2 也遵循相同的规律，反映了 PCA 主成分所代表的综合因子对价格的线性正向驱动作用；kilometer 特征的 shap 分布显示：粉红色点(高里程)主要集中在负值区域，蓝色点(低里程)主要集中在正值区域。这表明里程越高，对价格的负面影响越大，越拉低价格；反之，低里程车辆对价格产生正向贡献。模型准确捕捉了“里程越高、贬值越严重”的市场规律；车龄特征(car_age)的 shap 分布呈现以下特点：低车龄样本(蓝色点)在零点两侧均有分布，表明新车并不必然推高价格——当与其他负面特征(如高里程、损伤记录等)共同作用时，低车龄也可能产生负向贡献；高车龄样本(粉红色点)则主要集中在负 shap 区域，反映车龄增长对价格的普遍拉低效应。这一分布特征印证了车龄作为核心定价因素的复杂性，其影响受其他特征调节，存在明显的交互效应。车龄平方项(car_age_squared)的 shap 分布范围较原始车龄特征明显收窄，点集更紧密地聚集于零点附近。这表明平方项主要用于捕捉车龄对价格的加速贬值效应——即在车龄较高阶段，单位时间贬值幅度加大。但其对整体预测的贡献相对有限(特征重要性排名第 18)，说明原始车龄特征已能较好地刻画车龄与价格的主要线性关系，平方项仅作为非线性补偿。

综合上述分析，模型整体合理：关键业务特征(行驶里程、损伤状况)的影响方向符合市场常识，验证了模型学习的有效性。

力图(Force Plot)用于直观地展示单个样本的 shap 值及其对模型预测结果的影响，通过力图，可以清晰地看到每个特征对该样本预测值的贡献。下面对具体样本进行力图分析，如图 7。图中的起点 base valuc 表示模型的基线值，终点表示模型对该样本的最终预测值，这是基线值加上所有特征贡献的总和，在这里为 8.85 (价格对数化后结果)，相对于基线值最终预测值变大表明所有特征整体上产生了正向贡献。每个特征的贡献通过带颜色的条表示，条的长度表示该特征对最终预测值的影响大小，红色条表示正向贡献，即该特征使预测值增加，蓝色条表示负向贡献，即该特征使预测值减少。从图中可以看出特征 kilometer 取值为 8.0，v_0 取值为 48.42，v_12 取值为 1.96，v_3 取值为-0.86，v_pca1 取值为 3.50 时产生较大正向贡献，而有损坏的情况负向贡献最大，是主要拉低价格因素。

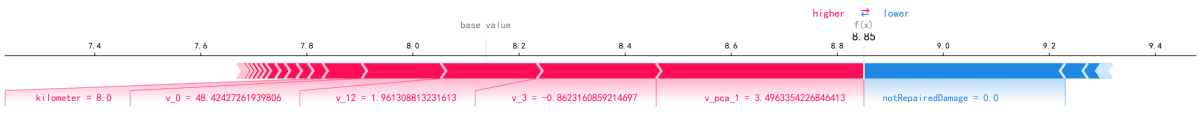


Figure 7. Force plot
图 7. 力图

5. 结论与展望

本研究针对二手车价格预测问题，构建并比较了多种机器学习模型。实证结果表明，基于梯度提升

框架的 LightGBM 模型综合性能最优。该模型在测试集上取得了平均绝对误差(MAE) 487.03 元、平均绝对百分比误差(MAPE) 13.20%的预测精度。这意味着，对于一辆价值 10 万元的二手车，模型的平均预测偏差约为 1.32 万元。在高度非标准化的二手车交易场景中，MAPE 低于 15%通常被视为具有较高的商业实用价值，这表明本文所构建的模型已能满足初步估价、市场分析等实际业务需求。

与随机森林、XGBoost 等主流模型相比，LightGBM 不仅在预测精度(MAE, RMSE)上全面领先，其高达 0.9708 的决定系数(R^2)也证明其能够解释车辆价格 97.08%的变异，具备出色的数据拟合与特征挖掘能力。本研究验证了通过微观车辆属性(如车龄、里程、功率及衍生特征)进行价格预测的有效性，为自动化、数据驱动的二手车估值提供了一套可行且高效的解决方案。

尽管本研究取得了阶段性成果，但仍有若干方向值得在未来工作中深入探索：

(1) 引入宏观与动态因素：本研究主要基于静态的微观车辆属性进行预测，其基本假设是宏观市场环境稳定。未来研究可纳入时序特征(如月份、季节性波动)、区域经济指标(如人均 GDP、二手车周转率)以及实时市场行情(如供需关系、政策变动)等宏观变量，构建更贴近动态市场的价格预测模型。

(2) 探索更先进的模型架构：深度学习模型在处理复杂非线性关系与高维特征交互方面具有潜力。未来可尝试引入注意力机制(如 Transformer)以捕捉不同特征间的重要性差异，或利用图神经网络(GNN)对车辆配置、品牌关联等结构化关系进行建模，以期进一步提升预测精度与模型的可解释性。

(3) 提升模型的可解释性与公平性：在商业应用中，模型决策的透明度和公平性至关重要。后续工作可结合 SHAP (Shapley Additive Explanations)等可解释性人工智能技术，量化各特征对具体价格预测的贡献，并检验模型在不同车型、地域及价格区间上是否存在系统性偏差，确保估价的公正与可靠。

(4) 构建端到端的估价系统：未来的实践方向是将预测模型集成至完整的业务系统中，实现从车辆信息录入、特征自动工程、价格预测到估价报告生成的端到端自动化流程，从而真正赋能二手车行业的数字化转型。

总而言之，二手车价格预测是一个多因素交织的复杂问题。本研究奠定了基于微观特征的机器学习方法的有效性基础，而通过融合宏观动态信息与更先进的建模技术，有望构建出更精准、更稳健、更智能的新一代二手车价格预测体系。

参考文献

[1] 刘岳阳, 何彦廷, 李瑜, 方健荣, 史佳硕. 互联网+背景下国内二手车市场模式创新[J]. 时代汽车, 2022(3): 183-185

[2] 吕劲. 基于特征优化组合 SVM 的二手车价格预测研究[D]: [硕士学位论文]. 武汉: 中南财经政法大学, 2019.

[3] 李富强, 彭海丽, 杨熙, 张文静. 基于深度学习的二手车价格预测模型及影响分析[J]. 汽车工程学报, 2021, 11(5): 379-385.

[4] 郑婕. 基于随机森林和 XGBoost 算法的二手车价格预测[J]. 数字技术与应用, 2021, 39(6): 90-93+188.

[5] 崔四帅. 基于集成学习的国内二手车价格预测分析[D]: [硕士学位论文]. 大连: 大连理工大学, 2021.