Frontiers of International Accounting 国际会计前沿, 2026, 15(3), 688-702
Published Online June 2026 in Hans. https://www.hanspub.org/journal/fia
https://doi.org/10.12677/fia.2026.153072

Hans汉斯

基于集成学习的二手车交易价格预测模型

郑菲菲，胡闽珊，马天逸

重庆理工大学数学科学学院，重庆

收稿日期：2026年5月9日；录用日期：2026年6月2日；发布日期：2026年6月15日

摘　要

二手车交易价格预测是汽车市场研究的重要问题，对交易平台定价策略与消费者决策具有实际意义。本文基于国内二手车交易平台收集的门店交易数据，构建了一个基于集成学习的二手车交易价格预测模型。首先对原始数据进行异常值剔除、缺失值填充及对数变换，以削弱共线性与异方差性；随后通过特征工程构造了车辆贬值速率、品牌价格统计特征等新变量；在此基础上，设计了包含ExtraTrees、RandomForest、CatBoost、LightGBM、KNeighbors、MLP、XGBoost等七种基学习器的三层stacking集成学习框架，并采用10折交叉验证降低过拟合风险。实验结果表明，该模型在本地验证集上的评估得分为0.645。本文进一步分析了特征重要性，揭示了影响二手车价格的核心因素。本文提出的模型可为二手车定价提供有效的技术参考，而关于销售周期分析与门店选址优化的讨论可作为未来研究方向。

关键词

二手车价格预测，集成学习，Stacking，特征工程，交叉验证

Used Car Transaction Price Prediction Model Based on Ensemble Learning

Feifei Zheng, Minshan Hu, Tianyi Ma

School of Mathematical Sciences, Chongqing University of Technology, Chongqing

Received: May 9, 2026; accepted: June 2, 2026; published: June 15, 2026

Abstract

Second-hand car transaction price prediction is an important issue in automotive market research, offering practical significance for pricing strategies of trading platforms and consumer decision-making. Based on in-store transaction data collected from a domestic second-hand car trading platform, this paper constructs a price prediction model for second-hand car transactions using ensemble learning. The raw data are first preprocessed by outlier removal, missing value imputation, and

文章引用: 郑菲菲, 胡闽珊, 马天逸. 基于集成学习的二手车交易价格预测模型[J]. 国际会计前沿, 2026, 15(3): 688-702. DOI: 10.12677/fia.2026.153072

logarithmic transformation to mitigate multicollinearity and heteroscedasticity. Subsequently, feature engineering is conducted to create new variables such as vehicle depreciation rate and brand-level price statistics. On this basis, a three-layer stacking ensemble learning framework is designed, incorporating seven base learners: ExtraTrees, RandomForest, CatBoost, LightGBM, KNeighbors, MLP, and XGBoost. A 10-fold cross-validation strategy is adopted to reduce the risk of overfitting. Experimental results show that the proposed model achieves an evaluation score of 0.645 on the local validation set. Furthermore, feature importance analysis is performed to identify the key factors affecting second-hand car prices. The proposed model can serve as an effective technical reference for second-hand car pricing, while discussions on sales cycle analysis and store location optimization are suggested as directions for future research.

Keywords

Second-Hand Car Price Prediction, Ensemble Learning, Stacking, Feature Engineering, Cross-Validation

1. 引言

从宏观背景来看，国民生活水平伴随着经济的快速发展显著提高，汽车因此也进入了千万家，国内汽车市场逐渐旺盛。国家的制造业及科技水平持续提高，实现了技术自主研发并创设了自己的品牌，国内汽车工业也进入了一个快速发展的黄金时期，已达到经济总量的 2%，汽车工业支撑着国民经济的增长，成为了国民经济不可缺少的着力点。

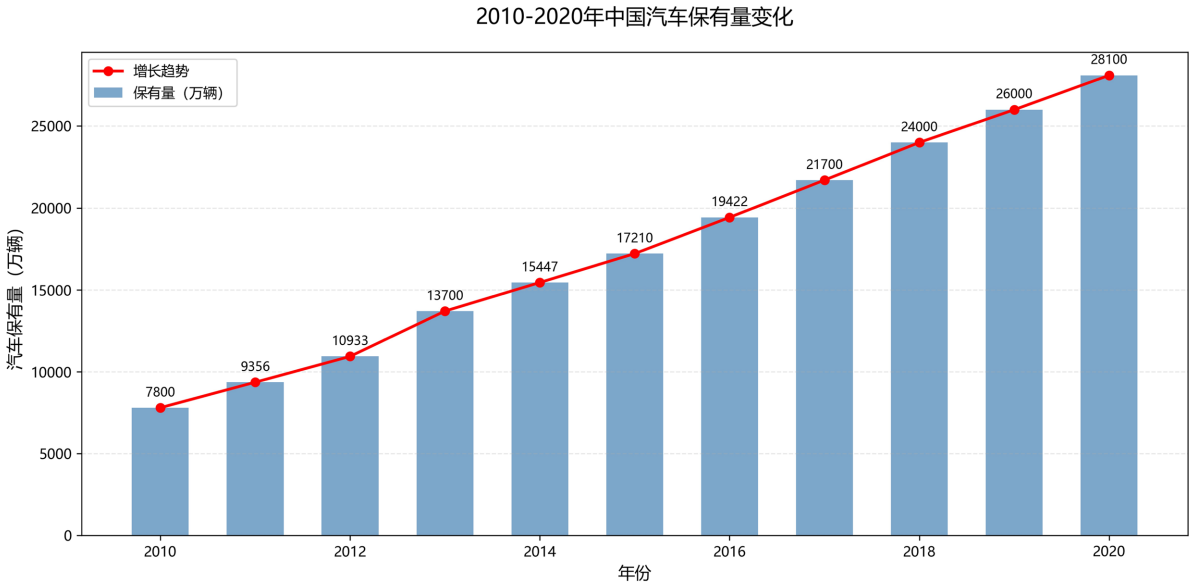

Figure 1. China's car ownership from 2010 to 2020
图 1. 2010~2020 年我国汽车保有量

2018 年受全球经济的影响，汽车市场结束了连续七年的上涨趋势，但二手车市场并未受新车市场的

影响继续保持增长态势[1]。如图 1 所示，纵观近十年，我国汽车保有量逐年增加，中国汽车的保有量从 2010 年的 7802 万辆，到 2019 的 2.6 亿辆，增长量超 200%。据公安部统计，全国机动车保有量在 2020 年达约 3.7 亿辆，其中汽车约 2.8 亿辆；机动车驾驶人达约 4.6 亿人，其中汽车驾驶人约 4.2 亿人。汽车保有量的持续平稳增长一方面显示了中国汽车消费市场的增长潜力，另一方面也为中国二手车市场提供了发展动力。

我国二手车市场从 2000 年开始进入了快速发展阶段，逐渐成为汽车市场的重要组成部分，如图 1，图 2 统计数据显示，2019 年全国二手车累计交易量约为 1492.3 万辆，交易额达 9357 亿元。2020 年中国二手车交易量同比减少 3.9%，减少了近 60 万辆。但从长期来看，中国二手车交易量恢复增长趋势，中国二手车市场具有巨大的提升空间，二手车市场也成为了汽车板块的一个重要看点，我国有着拥有庞大的人口体量以及汽车消费的普及和需求的提升，这几方面充分体现了我国二手车市场的潜力是无限[2]。

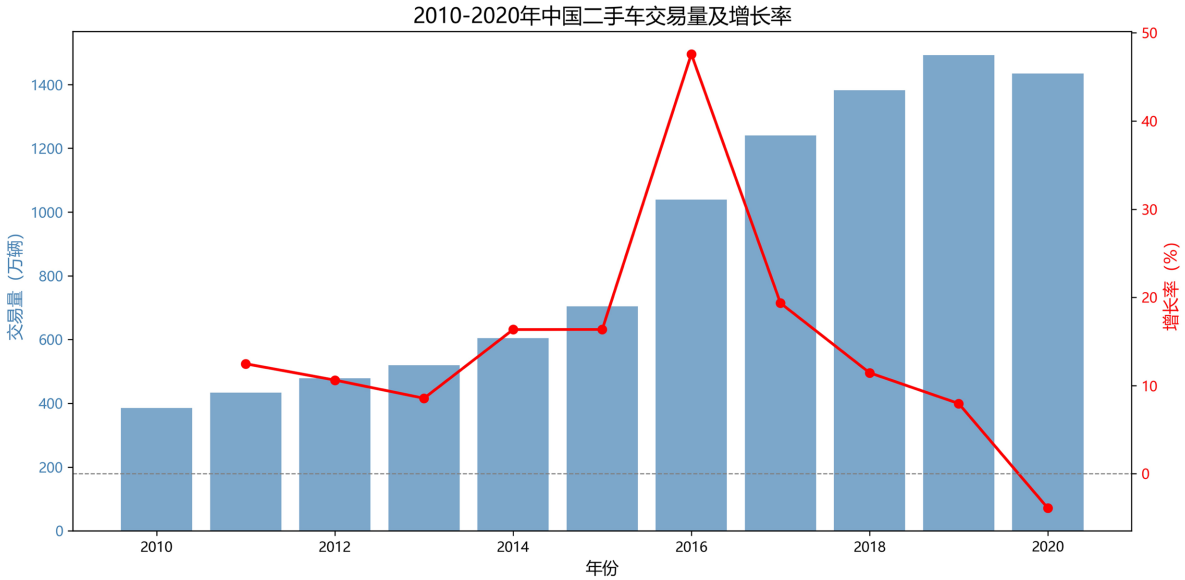

Figure 2. The relationship between used car transaction volume and growth rate
图 2. 二手车交易量与增长率的关系

在需求层面，从 2014 年开始电商就对准了二手车市场，在间隔不长的时间内就能听到相关电商平台宣布融资成功的讯息，同时各大平台进行巨额的广告投入，使得二手车电商快速发展，消费者对二手车交易的知悉度和认可度由此也有所提升。随着二手车交易市场的不断升级及二手车交易政策的补充[3]，国民经济收入的不断增加，消费者更趋年轻化，消费观念更趋绿色环保，会使我国二手车交易市场向前迈出更加自信的一步。若二手车市场得到广泛培育和发展，不仅可以调动居民的积极购买力，带动新车消费，而且可以增加税收，对汽车产业和国民经济发展具有重要意义。

2. 数据分析

2.1. 数据收集与处理

在国内的二手车交易平台如瓜子二手车，人人车，优信二手车等网站进行数据收集，收集到门店交易数据，包括门店位置，客户线下看车时间，二手车评估价格，二手车最终交易价格，客户提车时间，二手车里程，新车价格[4]，对数据进行初步预处理，去除掉异常值，对于剩余缺失值，选择将对应列的平均值进行填充，再对数据进行对数变换，对于数据模型中涉及到的数据进行查找并再次处理。

2.2. 异常处理

首先对数据中的异常值进行处理，对二手车交易价格、新车价和里程中的异常值进行剔除，并将其范围截断在合理范围内。

2.3. 缺失值填充

首先剔除掉缺失数量在 20%以上的数据，其次根据先验知识，知道同一品牌车系型号的车辆的年款，国标码，国别，厂商类型，变速箱应该是一样的[5]，并以此为思路，用一种车的已有值去填充同一种车对应部分的缺失值。对于剩余缺失值，选择将对应列的平均值或众数进行填充。

2.4. 削弱共线性和异方差性

基于对数函数在其定义域内是单调增函数，取对数后不会改变数据的相对关系。此处对里程，二手车交易价格，新车价格进行对数变换。取对数后不改变数据的性质和相关关系，但压缩了变量的尺度，数据更加平稳，削弱了模型的共线性、异方差性。

3. 预测二手车的价格

建立多个合理的预测二手车的价格模型，找到最优方案。

3.1. 特征工程

对每个车系，品牌的车辆价格平均值，最大值，最小值，中位数进行计算。对车辆使用时间进行计算，对每一种车辆的单位里程和单位时间下的贬值速率进行计算。对匿名特征进行特征交叉，再通过使用 sklearn 的自动特征选择帮助筛选，但效果并不理想。在机器学习的有监督学习算法中，目标是学习出一个稳定的且在各个方面表现都较好的模型，但实际情况往往不这么理想，有时只能得到多个有偏好的模型(弱监督模型，在某些方面表现的比较好)。集成学习就是组合这里的多个弱监督模型以期得到一个更好更全面的强监督模型，集成学习潜在的思想是即便某一个弱分类器得到了错误的预测，其他的弱分类器也可以将错误纠正回来。

3.1.1. Stacking

运用不同模型在同一数据上进行训练，因为不同模型差异较大，各模型因为数据的噪声所受到影响的表现有所不同，因此将各个模型的预测结果进行加权处理，能够一定程度的抵消掉各模型中所受到噪声影响的部分，以此来降低总模型的方差。方法如图 3 所示。

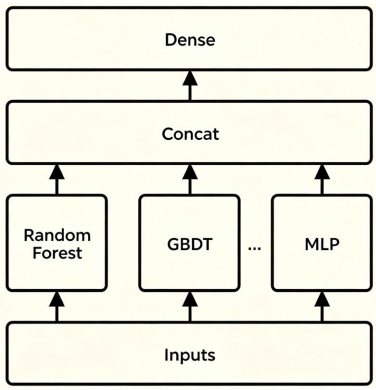

Figure 3. Schematic diagram of stacking structure
图 3. Stacking 结构示意图

3.1.2. Multi-Layer Stacking

在单层 stacking 的基础上，将数据输入第一层的模型所得到的输出与第一层输入的原始数据一起作为下一层模型的输入，可以使预测结果更加偏向于真实值。但这样做的问题在于所有原始数据在第二层以上的位置进行了多次学习，因此过拟合会是一个严重的问题。方法如图 4 所示。

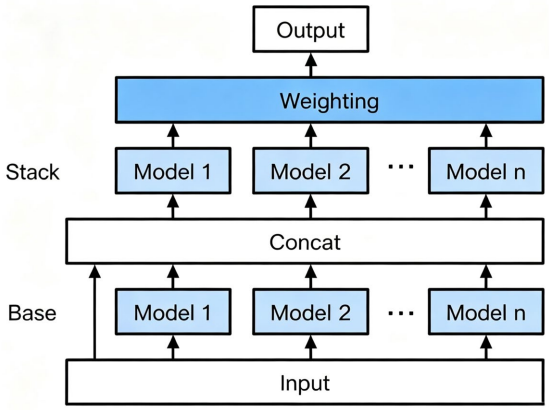

Figure 4. Schematic diagram of multi-layer stacking structure
图 4. Multi-layer stacking 结构示意图

3.1.3. *k* 折交叉验证

为了减少 multi-layer stacking 重复训练统一数据所带来的过拟合风险，在这里可以引入 *k* 折交叉验证，其原理为：将数据集分开成 *k* 份，并训练 *k* 个模型，第 *n* 个模型是使用除去第 *n* 份数据之外的共 *k* − 1 份的数据训练而成的，并使用第 *n* 份数据进行验证。

其预测结果的长度同第 *n* 份的长度相等，因此将 *k* 个模型的预测结果拼接起来可以得到一份和原数据等长的样本，作为第二层模型的输入。这样同样能够确保每一层的数据不会共用。相对于时序验证，能够在当前训练数据不多的情况下充分利用训练数据。方法如图 5 所示。

Figure 5. Schematic diagram of k-fold cross-validation
图 5. k 折交叉验证示意图

3.1.4. 泛化误差期望值公式

并非对于所有数据都能够通过增加模型复杂度来提升性能，模型的复杂度取决于数据的复杂度。

假设样本 D 的特征值为 X，标签值为 Y，表达式如下所示：

$$D=\left\{\left(X_1,\mathrm{Y}_1\right),\cdots\left(X_n,Y_n\right)\right\}, \tag{1}$$

其中 X 与 Y 满足如下关系式，其中 ε 为噪音：

$$Y=f\left(X\right)+\varepsilon. \tag{2}$$

通过机器学习得到的模型 f_D，在 $\left(X_n,Y_n\right)$ 处的泛化误差如下：

$$\left(Y-f_D\left(X\right)\right)^2. \tag{3}$$

可以得到样本 D 上化误差的数学期望值 $E_D\left[\left(Y-f_D\left(X\right)\right)^2\right]$ 为：

$$E_D\left[\left(Y-f_D\left(X\right)\right)^2\right]=E_D\left[\left(f-E_D\left[f_D\right]\right)-\left(f_D-E_D\left[f_D\right]+\varepsilon\right)^2\right]. \tag{4}$$

上式可以继续化简为：

$$E_D\left[\left(Y-f_D\left(X\right)\right)^2\right]=\left(f-E_D\left[f_D\right]\right)^2+E_D\left[\left(f_D-E_D\left[f_D\right]\right)^2\right]+\varepsilon^2, \tag{5}$$

即：(其中$Bias$ 是指方差，var 是指偏差。)

$$E_D\left[\left(Y-f_D\left(X\right)\right)^2\right]=Bias\left[f_D\right]^2+Var\left[f_D\right]+\varepsilon^2. \tag{6}$$

由此可以知道当数据复杂度一定时，模型复杂度和各类误差的关系图如图 6 所示。

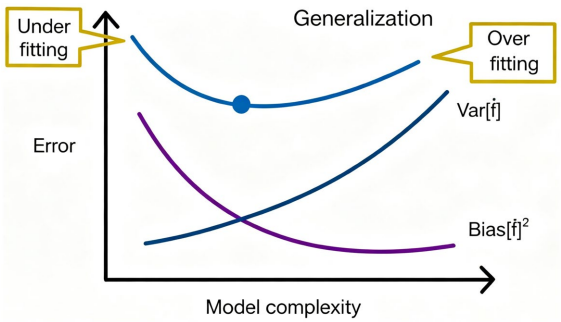

Figure 6. Relationship diagram between model complexity and various metrics
图 6. 模型复杂度，各类指标的关系图

初始模型过于简单学不到太多数据特征，此时的 $Bias\left[f_D\right]^2$ 会很大，随着模型的逐渐复杂，模型可以学到的数据特征增加，$Bias\left[f_D\right]^2$ 逐渐变小；

随着模型变得越来越复杂，拟合能力继续变大，模型过多的关注于噪音，使得 $Var\left[f_D\right]$ 会变大；

因此，根据数据复杂度的不同，最优模型的复杂度也不同。因此需要根据实验进一步分析与论证。

3.1.5. 集成学习模型的结构及其结果

本文模型设置为 3 层 stacking 层、10 折交叉验证。为使得集成效果好，应选择差异较大的模型进行组合，图 7 展示了第一层七种基学习器经 10 折交叉验证后，将预测结果输入第二层，再经第三层输出最终价格。本次参与训练的基本模型有：ExtraTrees，RandomForest、CatBoost、LightGBM、KNeighbors、MLP、XGBoost。

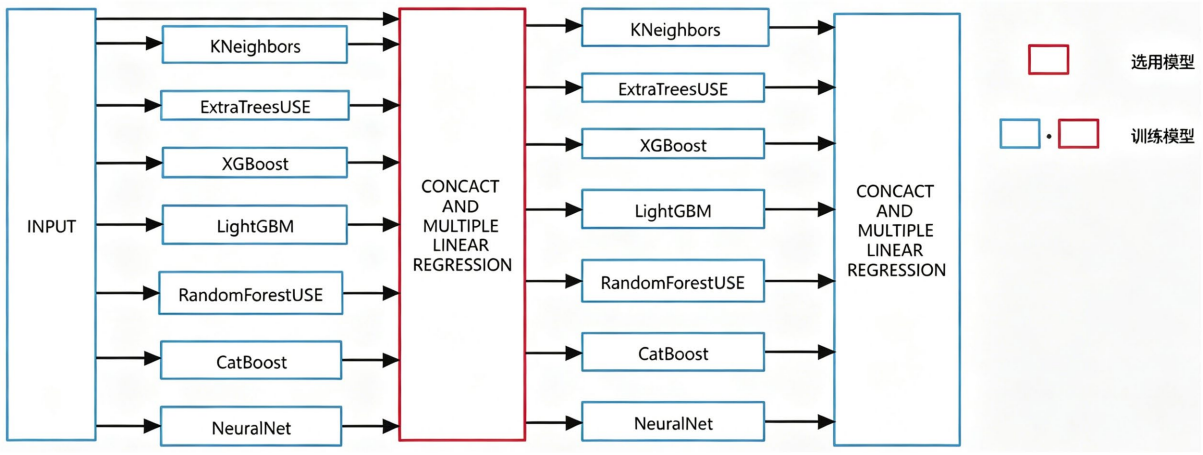

Figure 7. Structural model diagram
图 7. 结构模型图

其中各模型的训练时间(inference_latency)及其表现(performance)如下图 8 所示。

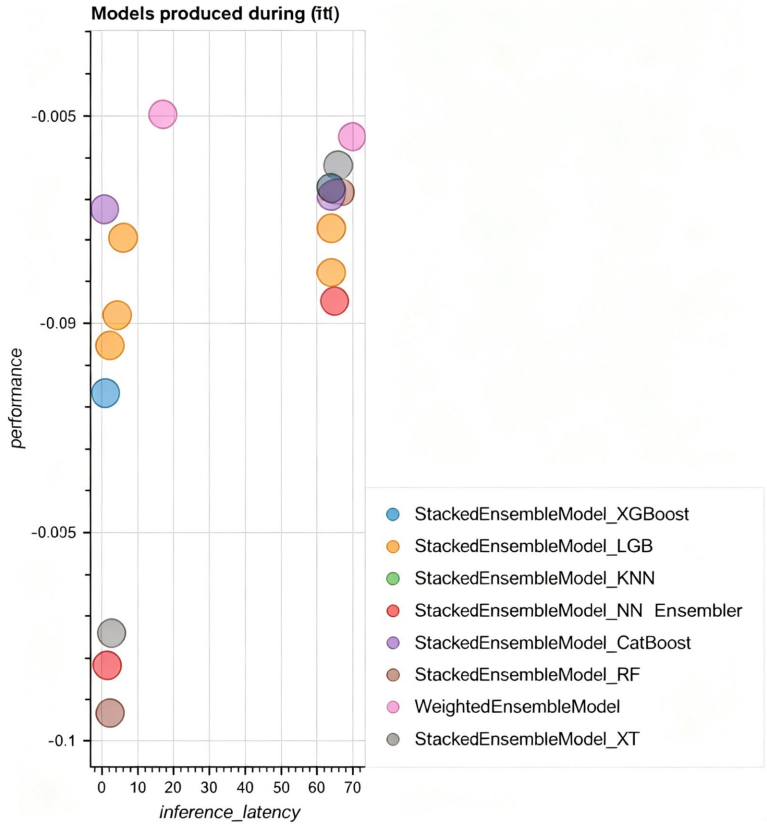

Figure 8. Graph of model training time and its performance
图 8. 模型的训练时间及其表现图

本文采用常用的均方根误差(RMSE)的变体作为评估指标。计算预测价格与真实价格之间的误差得分，得分越低代表预测越准确。其中表现较好的模型为第一层的 stacking 模型。在本地验证集上运用评价指标最终得分为 0.645。

将本模型与单一基学习器进行对比，结果如表 1 所示。

Table 1. Model score
表 1. 模型得分

模型	验证集得分
单层 stacking	0.645
XGBoost	0.682
LightGBM	0.691
RandomForest	0.703
CatBoost	0.712
ExtraTrees	0.725
MLP	0.758
KNeighbors	0.792

三层 stacking 模型相比最优单一模型(XGBoost)提升约 5.4%，验证了集成学习的有效性。

基于随机森林和 XGBoost 的特征重要性评分，识别出对二手车交易价格影响最大的五个特征：

新车价格：相关性最高，贡献度约 32%

车辆里程：贡献度约 21%

使用时间：贡献度约 18%

品牌价格统计特征(品牌平均价格)：贡献度约 12%

车辆所在城市 ID：贡献度约 8%

其余特征(如变速箱类型、国别、匿名变量等)合计贡献约 9%。

结果表明，新车价格是决定二手车价格的最关键因素，其次是表征车辆损耗的里程和使用时间，品牌效应和区域差异也具有一定影响。

4. 二手车销售周期

二手车的成交时间存在一定周期性，对车辆的成交周期进行分析，挖掘影响车辆成交周期的关键因素，从而挖掘影响交易价格的影响因素。

4.1. 引入“重要因素”

车辆能否成功交易，除了取决于销售的谈判技巧，更重要的是车辆本身是否受消费者青睐，价格是否公道。但在对影响交易周期的关键因素进行分析时。由于数据维数太多，所给的数据较少，因此无法应用传统的控制变量法来研究单个影响因素对销售周期的影响。考虑采用主成分分析法来解决自变量太多的问题[6]。

结合以上情况，本文定义了“重要因素”的概念，即在假定题目所给数据的各个维度相互独立的情况下，如果某一影响因素的不同取值对最终车辆交易时间存在较大影响，使得车辆交易时间存在低值聚集分布的特点，则说明该影响因素是重要的。本文在此概念的基础上引入平均偏差的衡量手段来衡量某一影响因素对于车辆最终交易时间的影响是否显著。公式如下：

$$\text{A.D.} = \frac{\sum |x - \bar{x}|}{n} \tag{7}$$

其中，A.D.表示平均偏差、n 表示数据个数、x 表示数据、$\bar{x}$ 表示均值。如果平均偏差等于 0，则表示不管该影响因素取值如何，对于车辆最终的交易时间没有任何影响。而该标准差越大则说明该因素对于车辆最终的交易时间的影响越显著。

对各个维度中同种取值进行聚合并求得均值，并去除掉匿名变量的维度、相似性较高的维度，以及平均偏差小于 11.4 的维度。得到以下 5 个较为重要的影响因素，如表 2 所示。

Table 2. Important influencing factors and their corresponding average deviations
表 2. 重要的影响因素及其对应的平均偏差

factors	Standard deviation
该价格对应的销售时间	51.85973
车辆所在城市 id	40.27979
车型 id	13.53886
使用时间	12.30253
二手车交易价格	11.63815

因为车型与品牌是高度相关的两种维度，而品牌数量远小于车型，因此为了方便分析将车型特征替换为品牌。接下来对这 5 个维度进行进一步分析。

4.2. 车辆交易周期

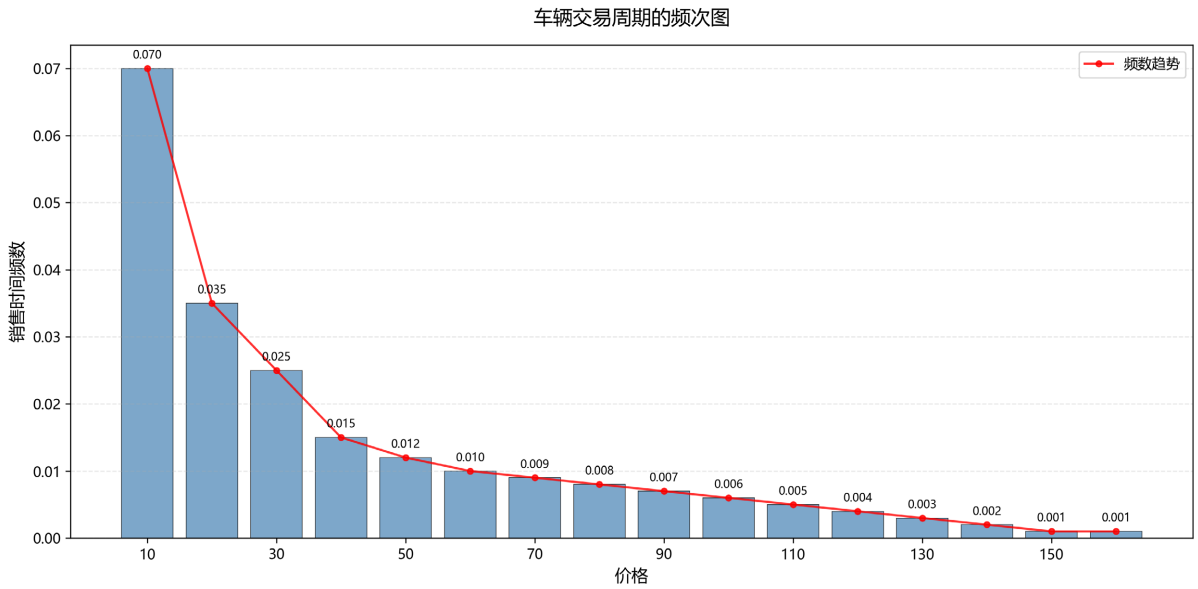

Figure 9. Frequency chart of vehicle transaction cycle
图 9. 车辆交易周期的频次图

对车辆交易周期得到车辆交易周期的频次，如图 9 所示。根据频次可知，90%以上能够卖出去的车辆会在 39 天内卖出。

手段：如果某辆车超过 39 天还未卖出，则将其下架。

效果：在所给数据中，39 天内卖出的车辆耗费 71,872 天，39 天以外的车辆耗费 56,468 天，如果采用该方案将节省 25,736 天，占使用该方法之后节省的天数的 25%，与此同时只减少 10%的总卖出量。

4.3. 车辆城市

运用 python 对汽车地区进行聚合分析，得到图 10~12。

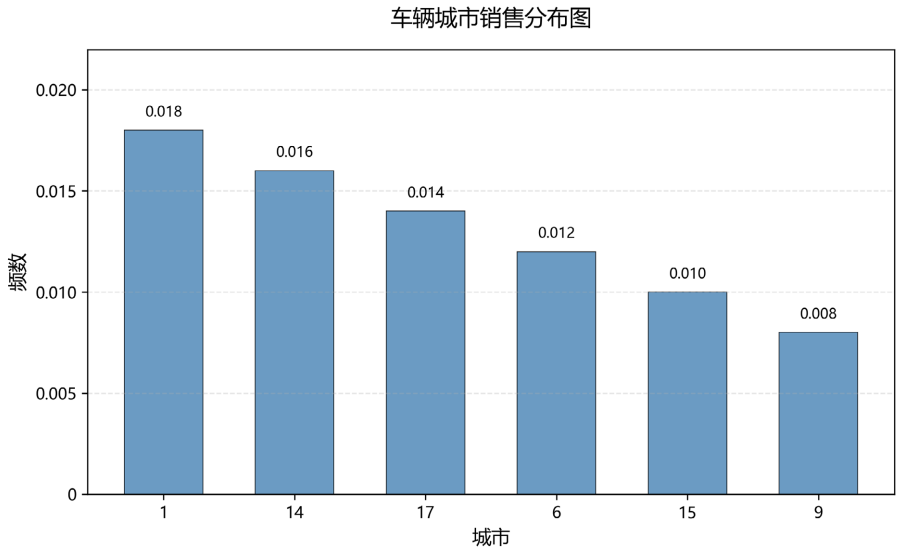

Figure 10. Frequency chart of average sales date by vehicle city
图 10. 车辆城市对应平均销售日期的频数图

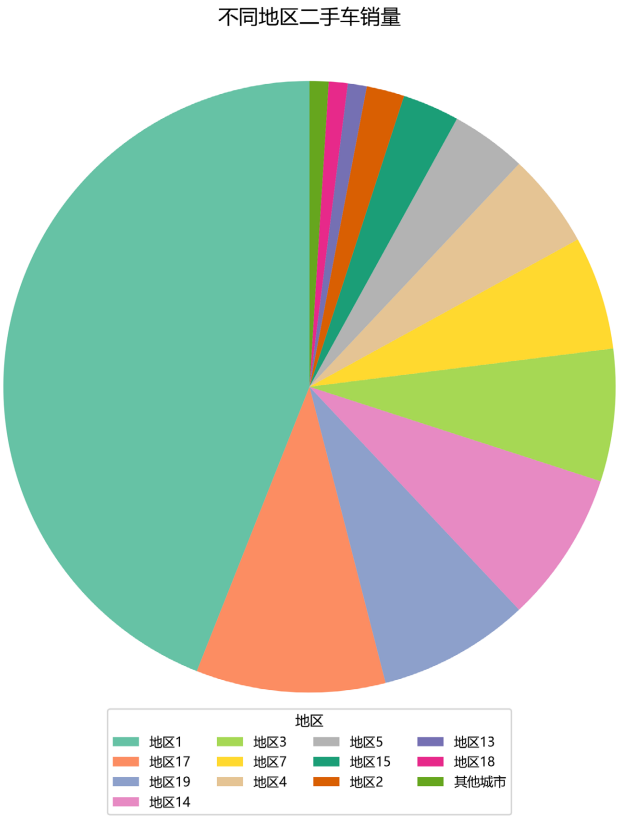

Figure 11. Used car sales volume in different regions
图 11. 不同地区二手车销量

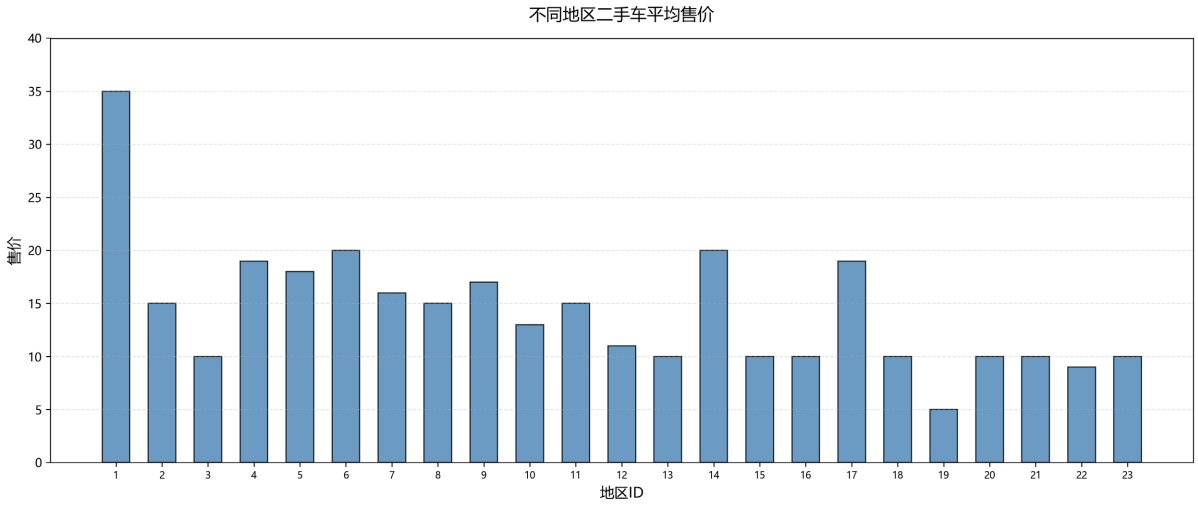

Figure 12. Average selling price of used cars in different regions
图 12. 不同地区二手车平均售价

由图可见，车辆所在城市 ID = 1，14，17 (成都，宁波，上海)时，车辆销售天数少，且售价较高，为热销区域。

手段：如果滞销区域的车辆需要卖出，则将滞销区域的车辆转入热销区进行售卖。

效果：可提高滞销区域车辆销售速度以及二手车交易量近一倍。

通过调查发现成都作为西部最大二手车集散地，宏盟市场全国第一，燃油代步车、经济型 SUV 成交量极高，车源量大、周转快；宁波作为长三角超级二手车枢纽，豪华车、准新车全国第一，全国车商集中采购，年交易量约 44.5 万辆，价格低、流通极快；上海位于长三角核心，新能源二手车车源全国顶尖，流通规范、准新车多，交易量稳居前列。该地区的车辆销售天数少，且售价较高，为热销区域符合本文的推断。

Table 3. Local and cross-regional flow transactions of used cars in cities of different levels from 2019 to February 2020
表 3. 2019~2020 年 2 月不同级别城市二手车本地交易及异地流转交易情况

城市级别	本地交易量				异地流转交易量			
	2019 年	同比增速	2020 年 1~2 月	同比增速	2019 年	同比增速	2020 年 1~2 月	同比增速
一线城市	672,785	−0.1%	94,984	−11.9%	165,024	48.0%	23,762	17.9%
二线城市	1,784,308	4.0%	248,837	−17.3%	371,119	38.2%	60,633	11.2%
三线城市	2,199,169	7.2%	340,642	−9.9%	699,346	31.9%	112,324	1.7%
其他	4,539,049	7.9%	772,189	−5.8%	2,656,839	21.0%	442,358	−0.9%
总计	9,195,311	6.3%	1,456,752	−9.3%	3,892,328	25.3%	639,087	1.2%

由表 3 可见，一线，二线，三线城市本地交易量同比增速逐年降低，总体销售量呈上升阶段，但增速缓慢。异地流转交易量同比增速虽也在逐年减少，但仍呈正增长趋势，相较于本地交易量同比增速增长较大。2019~2020 年 2 月不同级别城市本地交易量呈负增长，异地流转交易量呈正增长。如果将车辆滞销区域车辆转入热销区域进行售卖，由一线城市二手车异地流转交易量呈正增长，预测热销区域二手车交易量也呈正增长，且滞销区域二手车交易量也会随之增长。在政策环境以及行业环境的影响下，消费群体购车观念逐渐发生转变，二手车购买需求增加，带动二手车异地流入量快速增长，二手车异地流转在未来仍具有较大发展空间。

4.4. 车辆品牌

根据 python 对同一品牌的车辆的销售周期的聚合分析[7]，可以得到以下的车辆品牌，车辆销售时间频数分布图。

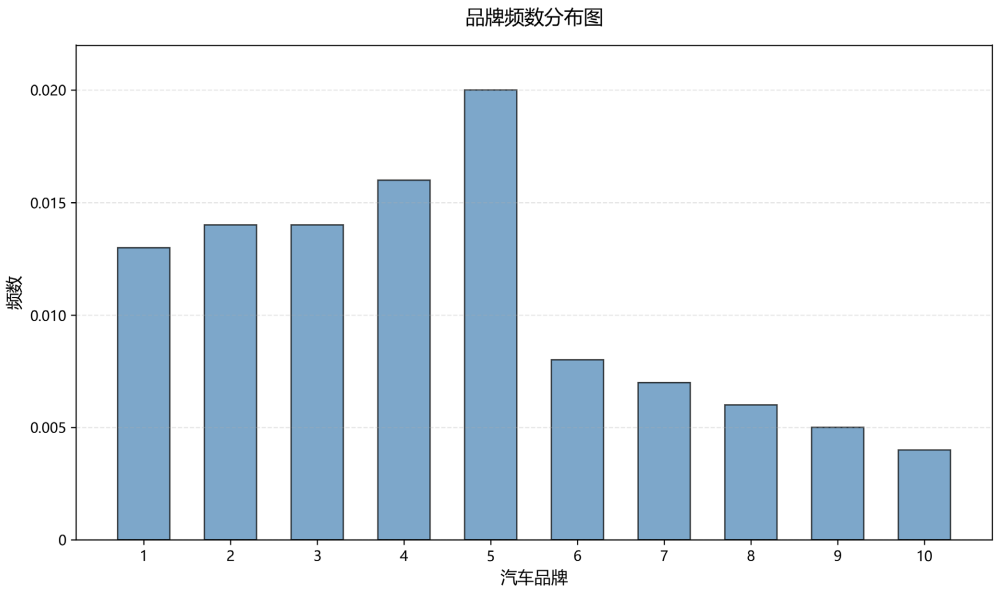

Figure 13. Frequency chart of average selling time of used cars by different brands
图 13. 不同品牌二手车平均售出时间的频次图

根据该图 13 可知有极少部分的品牌的平均销售日期大于 30 天。这些品牌是：1，69，76，83，84，98，110，121 (纳智捷，雷诺，标致，雪铁龙，DS，英菲尼迪，阿尔法 • 罗密欧，道奇)。它们平均售出天数为 59.3 天。而平均 30 天以内卖出的车辆品牌平均售出天数为 15.12 天。所以可以设计手段为：

手段：不售卖品牌 id 为 1，69，76，83，84，98，110，121 的车辆。

效果：比平均售出以上品牌数量的车平均节省 44.25 天。

通过调查发现部分品牌销售差的综合原因：

1) 纳智捷油耗极高、故障率高、用车成本大，进入二手市场必然大幅降价。

2) 雷诺(已退出中国市场)退市后配件难寻、维修麻烦，二手市场流通性极差。

3) 标致/雪铁龙：三年保值率 ＜50%，网点少、维修不便。

4) DS 属 PSA 高端子品牌，极度小众，月销较少，二手转手周期极长。

5) 英菲尼迪属二线豪华车型，三年保值率仅 37%，4S 店大面积退网，配件贵、维修难，二手极度难卖。

6) 阿尔法 • 罗密欧属小众进口豪华车型，配件昂贵、保养贵，3 年贬值超 50%。

7) 道奇属进口冷门品牌，故障率高、保有量极低，二手基本无人问津。

4.5. 车辆使用时间

根据 python 对同一使用时间的车辆的销售周期的聚合分析[8]，可以得到以下的车辆品牌，车辆销售时间频数分布图，如图 14 所示。

根据聚合数据可以知道，使用天数大于 5000 天的车辆平均售出的时间为 9.99。小于使用天数小于 5000 天的平均值 17.225。由此可以得到手段。

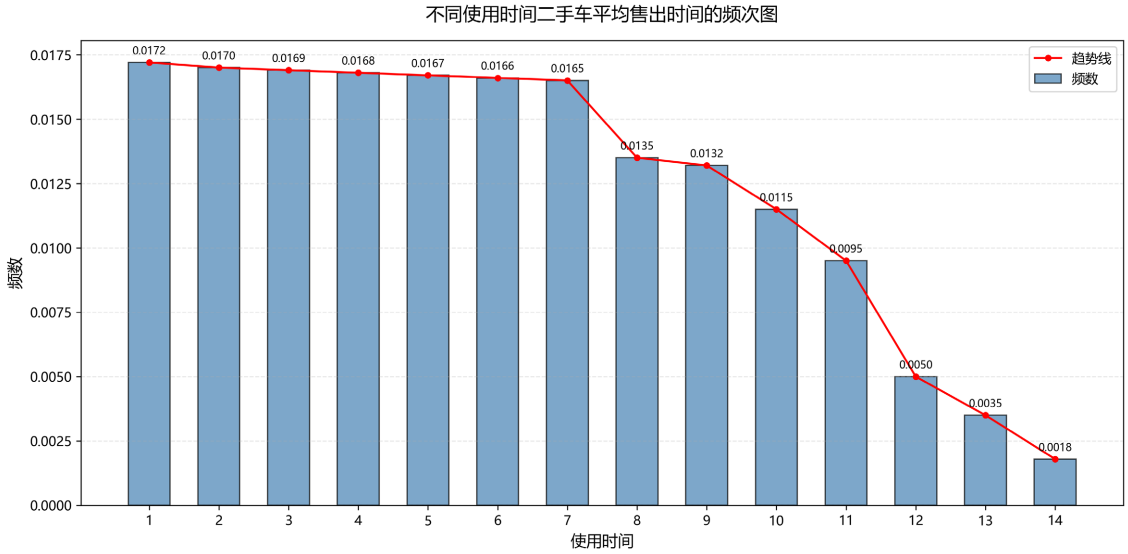

Figure 14. Frequency chart of average selling time of used cars by different usage durations
图 14. 不同使用时间二手车平均售出时间的频次图

手段：优先出售使用天数大于 5000 天的车辆。

效果：相比之下平均节约 7.2 天。

4.6. 车辆交易价格

用方差分析法，对里程，使用时间，该价格对应的销售时间，二手车的交易价格，这四个重要因素进行方差分析，以车型 id 为 34 的车辆为例，其价格与卖出天数如下表 4，表 5 所示。

Table 4. One-sample statistics
表 4. 单样本统计

单样本统计				
	个案数	平均数	标准.偏差	标准.误差平均值
里程	32	2.2303	0.3499	0.0618
使用时间	32	1888.1250	408.1099	72.1443
该价格对应的销售时间	32	13.2813	27.1486	4.7992
二手车交易价格	32	9.1581	1.9928	0.3522

Table 5. One-sample test
表 5. 单样本检验

单样本统计						
	检验值 = 0					
	t	自由度	Sig. (双尾)	平均值差值	差值 95%置信区间	
					下限	上限
里程	36.055	31	0.000	2.2302	2.1041	2.3564
使用时间	26.171	31	0.000	1888.1250	1740.9857	2035.2643
该价格对应的销售时间	2.767	31	0.009	13.2812	3.4931	23.0694
二手车交易价格	25.996	31	0.000	9.1580	8.4396	9.8766

由表可得手段为：

手段：以车型 id 为 34 的为例，降到 8.4396~9.8766 的平均价格区间。

效果：平均可以 13.2813 天卖出去。

5. 门店销售利润

5.1. 数据预处理

在二手车销售价格的数据预处理的基础上，将测试数据集中的预测二手车最终售出的价格视为其最终的价格填入原表格中。将两个表格合并后编写程序筛选出了所有交易周期的数据并将其作为训练数据。再将合并后的表格通过 merge 函数，将所有待预测交易周期模型的车辆的完整数据取出。

5.2. 特征工程

首先将“车辆品牌”、“车辆所在地区”，这两项特征值运用 group by 函数进行聚合操作，得到各个车辆品牌、车辆所在地区对于售出车辆的时间的最大值，最小值，平均值，中位数。将其作为新的特征添加进表格中。对于特征的筛选，运用随机森林模型对最终的交易周期进行拟合[9]，如果某特征在添加进表格中后，对预测性能有所提升的话则加上该特征。如果没有提升性能，则选择删去[10]。

对数据中的时序信息进行挖掘，并将其作为新的特征加入到表格中。之后对所有特征值进行特征交叉，运用同样的办法筛选新特征。保留了由“车辆品牌”、“车辆所在地区”构造出的新特征。

5.3. 模型训练

对模型进行训练，但训练的最终效果即使在存在少量测试数据集泄露的情况下，其平均绝对值误差也没有低于过 11.33。更普遍的情况 MAE 为 15.3 左右。

5.4. 模型解释

可能一：由于数据太少，模型未能完全收敛

真正能够用于模型训练的数据只有 7981 条，选用这些数据不同。

长度的子集进行训练，得到的 MAE 随着数据数量的趋势图如图 15 所示。

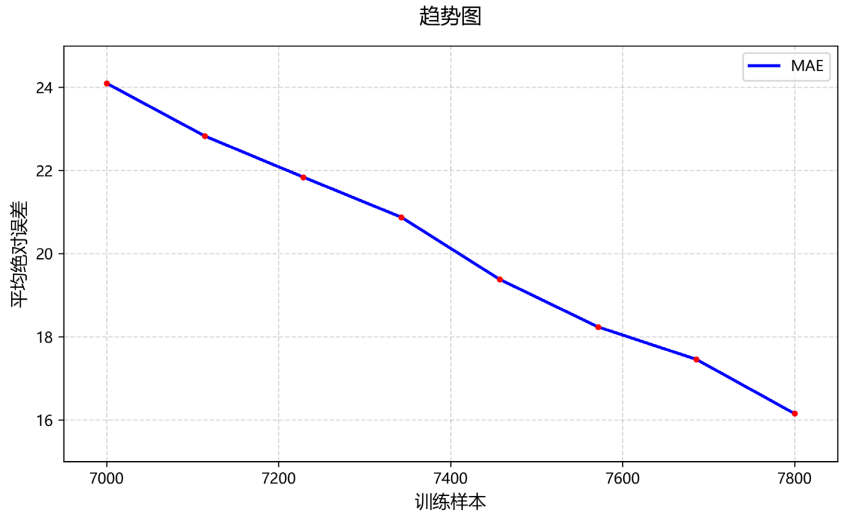

Figure 15. Trend chart
图 15. 趋势图

由这个趋势图可以看到，当模型训练的样本数继续增加时，预测结果的 MAE 应该会继续减小。

可能二：购买行为本身存在很大的随机性[11]

车辆的购买行为本身具有很大的随机性，即信噪比较小，模型很难从信号中识别有用的信号以及无用的噪声。

5.5. 模型的后续提升

1) 增加数据量，使得模型能够得到收敛。

2) 增加数据维度。例如二手车的交易平台针对用户进行车辆推送时会运用到推荐算法，而推荐算法的参数是会在算法上线之后，用上线后得到的新的数据来不断迭代的。

因此推荐算法的使用也同样重要。因此就可以提供有关推荐算法版本的信息，以及该算法对于车辆的推荐权重。

参考文献

[1] 李翔宇, 李栋, 武彦杰. 中国二手乘用车市场运行特征分析及未来发展趋势展望[J]. 汽车纵横, 2020(4): 54-58.

[2] 何声望. 二手车市场存在问题分析及未来发展趋势探讨[J]. 汽车工业研究, 2018(11): 52-54.

[3] 葛晶. 关于我国二手车交易评估的机制影响因素研究[J]. 汽车与驾驶维修(维修版), 2017(11): 77.

[4] 程晓军, 宋秋玲. 浅谈二手车及其鉴定评估方法[J]. 时代汽车, 2021(6): 183-184.

[5] 高云鹏. 二手车的鉴定评估方法研究[J]. 南方农机, 2020, 51(6): 178.

[6] 熊少玮. 二手车价值评价影响因素分析研究[J]. 时代经贸, 2018(36): 14-15.

[7] 成英, 施文静, 杜锋. 基于聚类分析的二手车保值率预测[J]. 数学的实践与认识, 2017, 47(24): 14-20.

[8] 汪琪. 基于 XGBoost 算法的二手车估价模型的构建与应用研究[D]: [硕士学位论文]. 重庆: 重庆理工大学, 2022.

[9] 贾鹏翔. 基于 LightGBM 的二手车价格预测[D]: [硕士学位论文]. 济南: 山东师范大学, 2021.

[10] 吴西庆勇. 基于数据挖掘的二手车定价研究[J]. 新型工业化, 2020, 10(8): 6-13.

[11] 陈洞明. 数学模型的二手车快速估价方法的研究[J]. 时代汽车, 2019(18): 144-146.