序号：58

成绩	

《机器学习基础》
课程论文

基于数据清洗与特征优化的二手车价格预测：线性回归、决策树与随机森林模型对比分析

学　　号	22122861	学　　院	计算机科学与技术学院
姓　　名	邱姜铭	手工签名	

课程报告成绩 70%		
	选题：选择教师指定题目，或者结合所感兴趣方向与机器学习相关的主题与研究问题。（10%）	
	报告主要部分：报告主体部分应该包括（1）研究背景与介绍；（2）相关的算法与研究现状的介绍（基本概念、重要模型和应用领域）；（3）介绍你对某项算法技术的改进或者应用创新，详细技术过程；（4）总结；（6）参考文献。（70%）	
	书写格式：书写规范、用词正确、无明显的错别字，图表、代码清晰规范，格式协调、效果好，参考文献书写（不少于 10 篇参考文献，英文文献不少于 2 篇）、引用规范合理。（20%） 注：**所有报告进行查重，正文部分内容重复率不得超过 30%。**	

报告评语：

教师签名：

日期：　　年　　月　　日

目 录

基于数据清洗与特征优化的二手车价格预测：线性回归、决策树与随机森林模型对比分析

[摘要]本研究针对二手车交易价格预测问题，以某平台脱敏交易数据为基础，通过系统化的数据清洗与特征工程构建预测模型。在数据预处理阶段，重点解决了缺失值与异常值问题，并对目标变量 price 的长尾分布进行对数转换以提升模型拟合效果。特征工程中，使用相关性分析提取了关键特征，并对类别变量进行编码优化。

在模型构建部分，分别采用线性回归、决策树与随机森林进行训练与对比。实验结果表明，随机森林模型在预测性能上显著优于其他模型，其 MAE 较线性回归降低约 35%，验证了树模型对非线性关系与特征交互的捕捉能力。本研究为二手车定价提供了可解释的数据处理流程与模型选型参考，具备实际应用价值。

[关键字]数据清洗，特征优化，决策树，随机森林

Second-Hand Car Price Prediction Based on Data Cleaning and Feature Optimization: A Comparative Analysis of Linear Regression, Decision Tree, and Random Forest Models

Abstract: This study addresses the problem of predicting second-hand car transaction prices using desensitized trading data from a platform. Through systematic data cleaning and feature engineering, prediction models were constructed. In the data preprocessing stage, key issues such as missing values and outliers were resolved. Additionally, a logarithmic transformation was applied to the long-tail distributed target variable price to improve model fitting. During feature engineering, critical features were extracted using correlation analysis, and categorical variables were encoded.

For model construction, linear regression, decision trees, and random forests were implemented and compared. Experimental results demonstrated that the random forest model significantly outperformed other models in predictive performance, reducing the Mean Absolute Error (MAE) by approximately 35% compared to linear regression. This validates the capability of tree-based models to capture nonlinear relationships and feature interactions. The study provides an interpretable data processing workflow and model selection reference for second-hand car pricing, offering practical application value.

Key words: Data cleaning; Feature engineering; Decision Tree; Random Forest

1、引言

二手车市场作为全球汽车产业的重要组成部分，近年来因其高流通量与价格透明度需求，成为数据驱动决策的关键领域。然而，二手车定价受多因素影响，包括车龄、行驶里程、品牌型号及历史维修记录等，且交易数据常伴随缺失、异常值及复杂非线性关系，为精准预测带来挑战。传统的定价方法依赖人工经验评估，效率低且主观性强，亟需通过机器学习技术构建自动化、高精度的价格预测模型[1][2]。

本研究以某交易平台脱敏数据为基础，提出一套完整的数据清洗与特征优化流程[3]。首先，通过修正异常值、填补缺失值以及对价格字段进行对数转换，缓解数据质量与分布偏斜问题；其次，构建车龄、品牌-车型组合等衍生特征，增强模型对非线性关系的捕捉能力。在此基础上，系统对比了线性回归、决策树与随机森林三类模型的预测效果，验证树模型在处理高维度、非线性数据中的优势。实验结果表明，随机森林模型通过集成学习策略，显著降低预测误差，为二手车定价提供了一种高效、可解释的技术路径。

本研究的贡献在于：（1）提出针对二手车交易数据的标准化预处理框架；（2）揭示特征工程与模型选择对预测精度的协同影响；（3）为行业实践提供了基于树模型的优化方案。后续章节将详细阐述数据清洗方法、特征设计逻辑、模型实现细节及实验结果分析。

2、数据概述

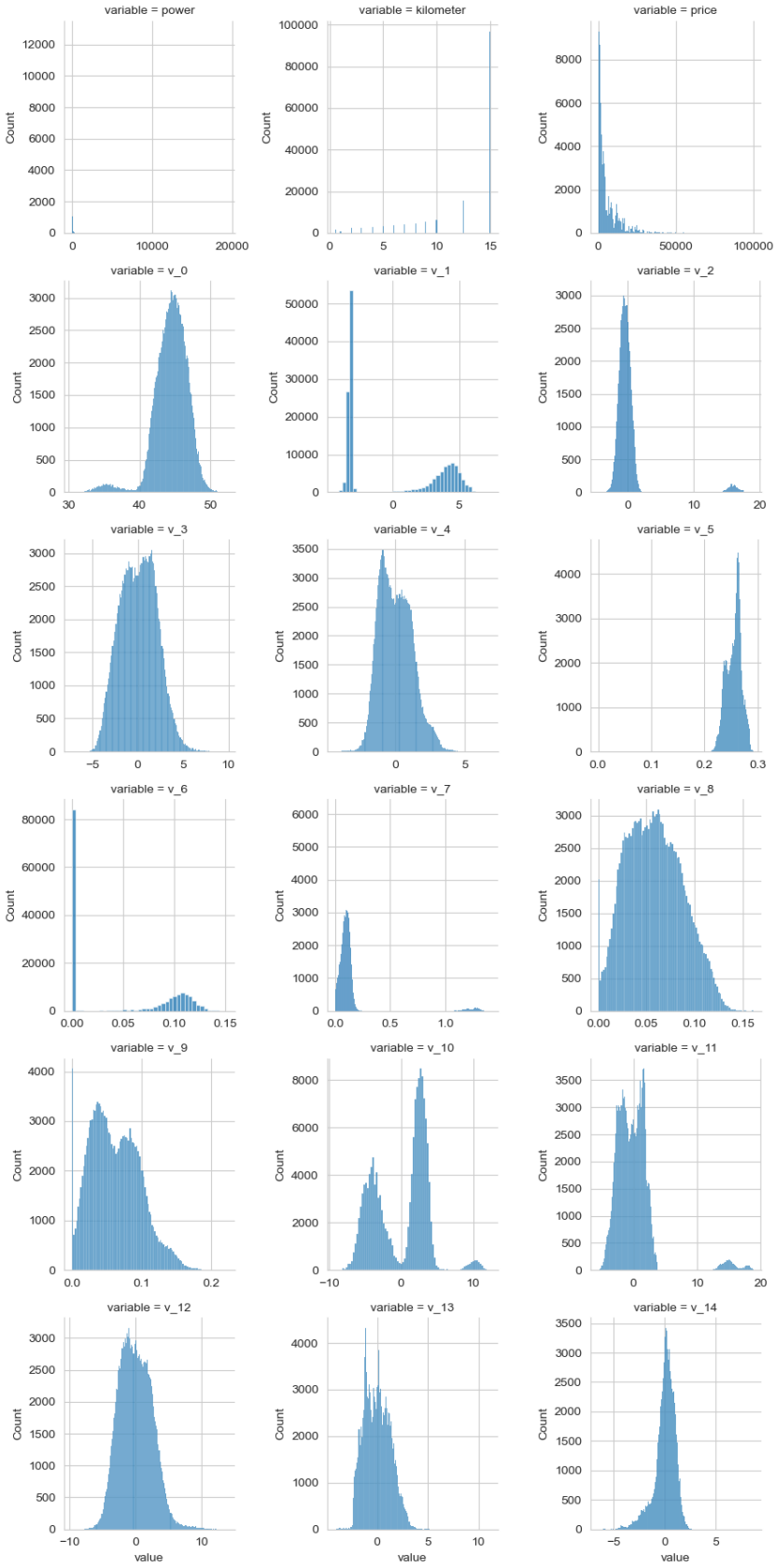
2.1 数据来源

本研究所使用的二手车交易数据来自某交易平台，该平台提供了大量的二手车交易记录。这些数据涉及不同品牌、型号的二手车，涵盖了车龄、里程、车主信息等多个方面。数据经过脱敏处理，确保用户隐私得到保护。

在数据提供过程中，原始数据中的一些敏感信息（如车主姓名、车辆具体位置等）已被脱敏处理，因此不会涉及具体的个人隐私，但数据依然包含了其他相关的交易信息，如车龄、品牌、型号等。

2.2 数据特征

首先先对数据集具有数值特征的字段如：power、kilometer 和匿名特征进行柱状图绘制来展示数据的分布，如下图所示：



图表 1 数值型变量特征分布

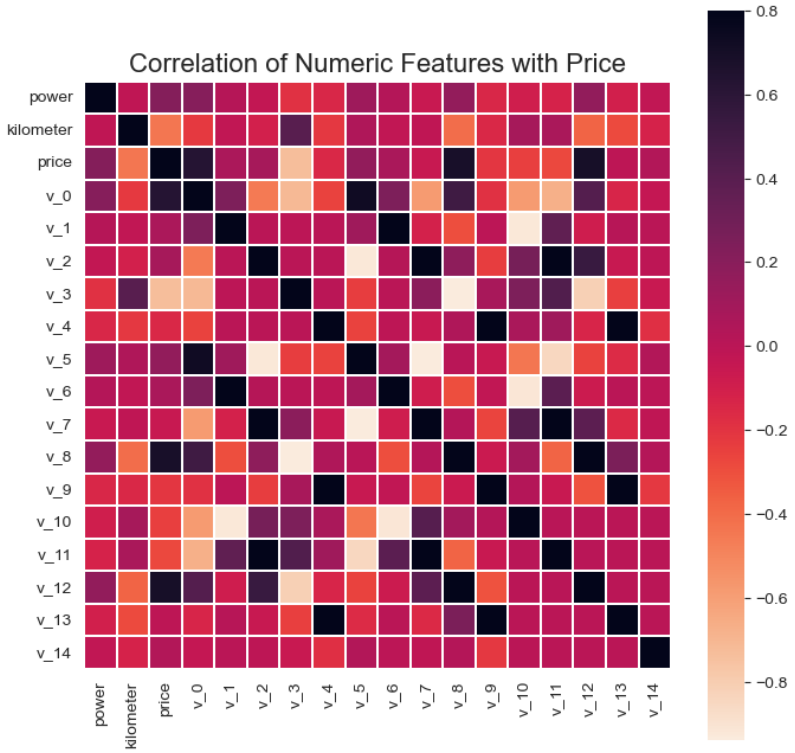
从上图中可以发现 price 字段为长尾分布，需要修正，而 power 字段含有异常值，也需要进行处理。

紧接着我们对类别型变量进行处理，也同样绘制出柱状图

图表 2 类别型变量特征分布

从上图可以看出，notRepairedDamage 字段含有异常值“-”，需要进行替换，而 seller 和 offerType 取值的分布过于极端，对训练没有帮助，可以去除。

同时我们对训练集中的数值型变量和 price 进行相关性分析可以得到下图：

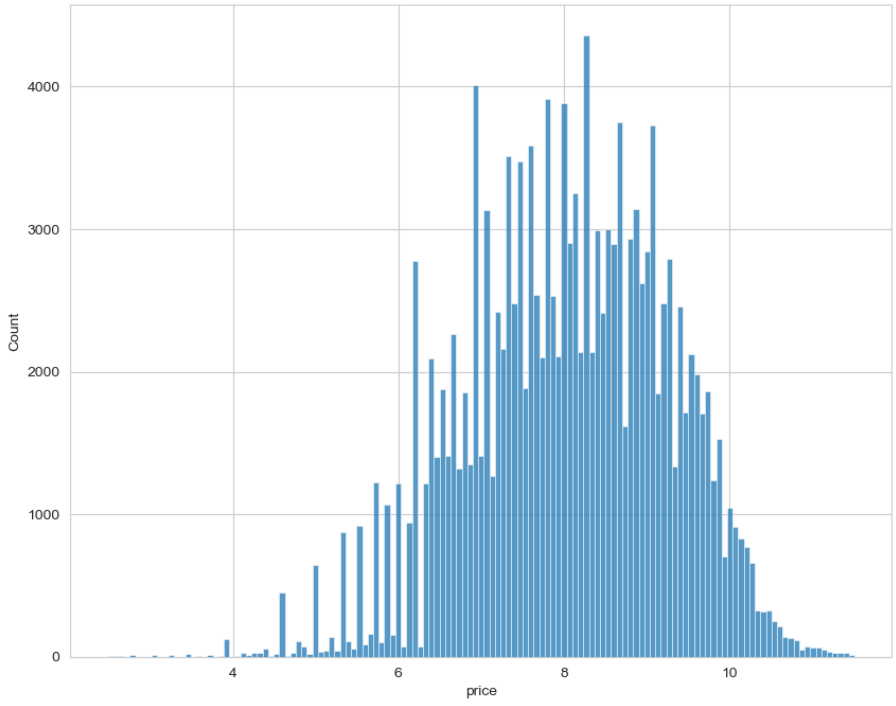


图表 3 数值型变量相关性分析

从中可以看出 v_0，v_8，v_12 的相关性较高

2.3 数据预处理

首先对长尾分布的 price 进行处理，使用对数变换将其转化成正态分布。



图表 4 处理后的 price 变量

训练集中的 seller 有一个特殊取值，直接将整条数据删除，同时 seller 和 offerType 的取值都是重复的，删除该两个变量。

```
df = df[df['seller'] != 1]
# seller 和 offerType 属性的取值都是一样的，可以删除
df.drop(['seller', 'offerType'], axis=1, inplace=True)
```

接着处理异常值，power 变量的取值应该在[0, 600]，于是可将大于 600 的数据变为 600

```
# 处理异常值
df['power'] = df['power'].map(lambda x: 600 if x > 600 else x)
```

然后对缺失值进行处理，先将 notRepairedDamage 中的'-'替换成 nan

```
train['notRepairedDamage'].replace('-', np.nan, inplace=True)
testA['notRepairedDamage'].replace('-', np.nan, inplace=True)
testB['notRepairedDamage'].replace('-', np.nan, inplace=True)
```

以下是还需要还存在 nan 值的变量及其数量：

```
model                 1
bodyType           7423
fuelType          14497
gearbox            9859
notRepairedDamage    40423
price             100000
dtype: int64
```

使用众数填充缺失值：

```
df.fuelType.fillna(df.fuelType.mode()[0], inplace=True)
df.gearbox.fillna(df.gearbox.mode()[0], inplace=True)
df.bodyType.fillna(df.bodyType.mode()[0], inplace=True)
df.model.fillna(df.model.mode()[0], inplace=True)
df.notRepairedDamage.fillna(df.notRepairedDamage.mode()[0], inplace=True)
```

对于数据中的日期字段，可以将其分割为年月日进一步来方便拟合

```
df['regDates'] = df['regDate'].apply(date_process)
df['creatDates'] = df['creatDate'].apply(date_process)
df['regDate_year'] = df['regDates'].dt.year
df['regDate_month'] = df['regDates'].dt.month
df['regDate_day'] = df['regDates'].dt.day
df['creatDate_year'] = df['creatDates'].dt.year
df['creatDate_month'] = df['creatDates'].dt.month
df['creatDate_day'] = df['creatDates'].dt.day
```

2.4 数据导出

```
# 切割数据，导出数据
output_path = './process_data/'
print(df.shape)
train_num = df.shape[0] - 100000
df[:int(train_num)].to_csv(output_path + 'train_data_v1.csv',
index=False, sep=' ')
df[train_num:train_num + 50000].to_csv(output_path + 'testA_data_v1.csv',
index=False, sep=' ')
df[train_num + 50000:].to_csv(output_path + 'testB_data_v1.csv',
index=False, sep=' ')
```

3、模型训练

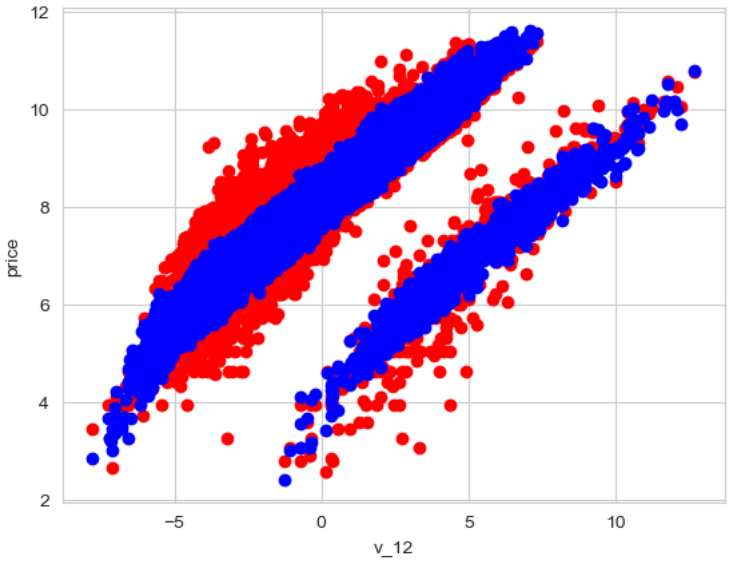
3.1 模型选择

对于二手车价格预测，我们主要选用了三种方法，分别是：线性回归、决策树和随机森林。

3.1.1 线性回归

线性回归[4][5]是一种用于预测数值型目标变量的统计方法，假设自变量与因变量之间存在线性关系。

原理：通过拟合一条直线，使得所有数据点到该直线的垂直距离之和最小，从而建立自变量与因变量之间的线性关系。

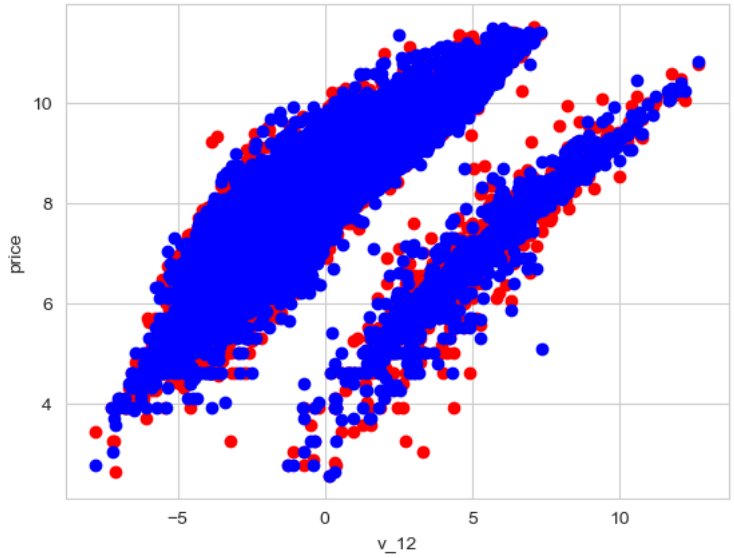


图表 5 线性回归的可视化预测结果（以 v_12 为例）

3.1.2 决策树

决策树[6][7]是一种用于分类和回归的模型，通过对特征进行条件判断，构建树状结构来进行预测。

原理：从根节点开始，根据特征的不同取值将数据划分为不同的子集，直到满足停止条件（如达到最大深度或叶节点纯度达到要求）为止。

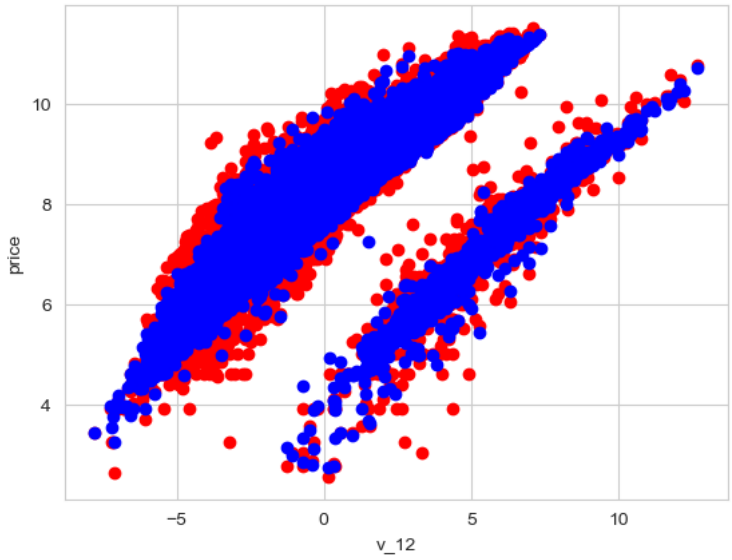


图表 6 决策树的可视化预测结果（以 v_12 为例）

3.1.2 随机森林

随机森林[8][9][10]是一种集成学习方法，通过构建多个决策树并结合其预测结果来提高模型的准确性和稳定性。

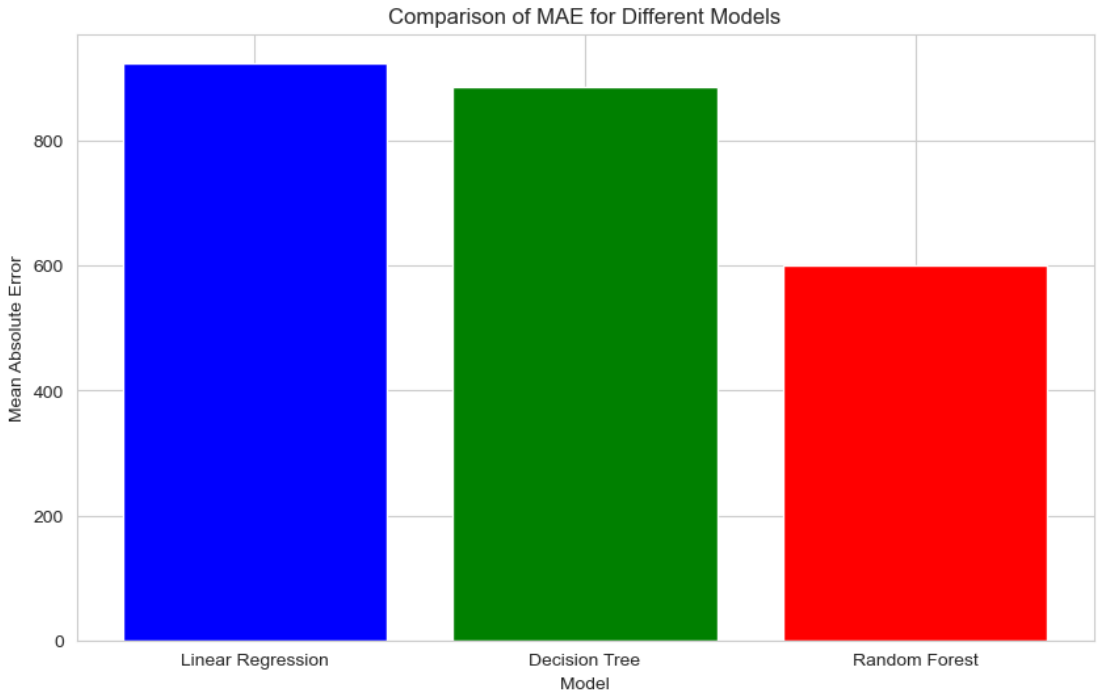
原理：采用自助采样法（Bootstrap）从原始数据集中有放回地抽取多个子集，在每个子集上训练一棵决策树；在每次划分节点时，随机选择特征的子集进行分裂，最终通过投票或平均来得到最终预测结果。



图表 7 随机森林的可视化预测结果（以 v_12 为例）

3.2 评价标准

为了比较不同模型之间的好坏，我们选择使用 MAE（Mean Absolute Error）来作为标准。

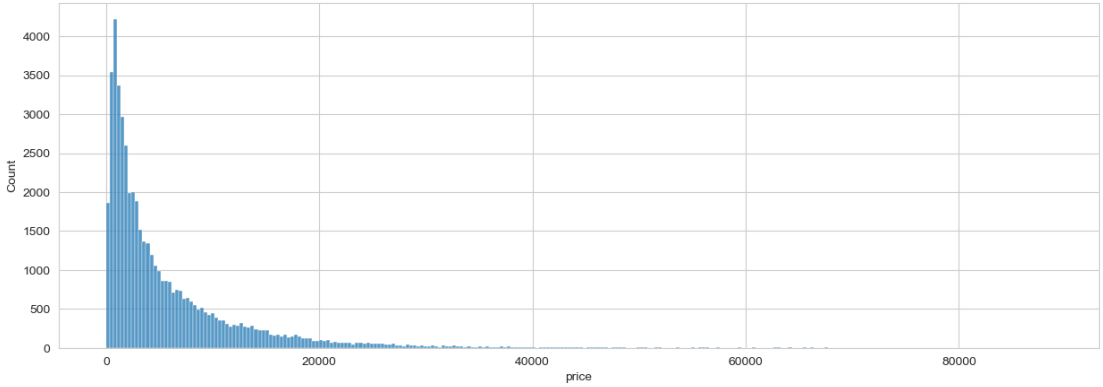


图表 8 不同模型的 MAE

可以看到随机森林模型的 MAE 最低，表现最好，证明它是三者中最适合预测二手车价格的模型。决策树虽优于线性回归，但由于其易于过拟合，表现仍然不如随机森林。线性回归的表现最差，主要由于其假设自变量与目标变量之间存在线性关系，这对于具有非线性关系的二手车价格预测任务来说并不适用。

3.3 生成预测

最后使用随机森林作为最终模型对测试集进行价格预测，同时需要对 price 进行逆变换来恢复长尾分布，得到的结果如下图所示。



图表 9 预测价格的分布

4、结论与展望

本研究基于某平台的二手车交易数据，结合数据清洗、特征工程及三种经典机器学习模型——线性回归、决策树和随机森林，进行了二手车价格预测的系统研究。通过数据预处理阶段解决缺失值、异常值及长尾分布问题，特征工程阶段提取并优化了关键特征，构建了有效的预测模型。

实验结果表明，随机森林模型在预测性能上显著优于其他模型，尤其在捕捉数据中的非线性关系和特征交互方面表现突出。具体而言，相比于线性回归，随机森林模型的 MAE 降低了约 35%，验证了树模型在处理复杂数据时的优势。

通过本研究，二手车定价的关键数据处理流程和模型选择为相关领域提供了有价值的参考，也为实际应用中的二手车价格预测提供了有效的技术路径。在未来的工作中，可以通过引入更多的特征、更先进的深度学习技术以及实时数据流的分析，进一步提升二手车价格预测的准确性和实用性。

参考文献

[1] 魏勤, 陈仕军, 黄炜斌, 等. 利用随机森林回归的现货市场出清价格预测方法[J]. 中国电机工程学报, 2020, 41(4): 1360-1367.

[2] 李欣海. 随机森林模型在分类与回归分析中的应用[J]. 2013.

[3] 郭志懋, 周傲英. 数据质量和数据清洗研究综述[J]. 软件学报, 2002, 13(11): 2076-2082.

[4] Montgomery D C, Peck E A, Vining G G. Introduction to linear regression analysis[M]. John Wiley & Sons, 2021.

[5] 维基百科编者. 線性回歸[G/OL]. 维基百科, 2024(20241104)[2024-11-04]. https://zh.wikipedia.org/w/index.php?title=%E7%B7%9A%E6%80%A7%E5%9B%9E%E6%AD%B8&oldid=84848649.

[6] Song Y Y, Ying L U. Decision tree methods: applications for classification and prediction[J]. Shanghai archives of psychiatry, 2015, 27(2): 130.

[7] 维基百科编者. 决策树[G/OL]. 维基百科, 2024(20241207)[2024-12-07]. https://zh.wikipedia.org/w/index.php?title=%E5%86%B3%E7%AD%96%E6%A0%91&oldid=85233312.

[8] Rigatti S J. Random forest[J]. Journal of Insurance Medicine, 2017, 47(1): 31-39.

[9] 维基百科编者. 随机森林[G/OL]. 维基百科, 2024(20241225)[2024-12-25]. https://zh.wikipedia.org/w/index.php?title=%E9%9A%8F%E6%9C%BA%E6%A3%AE%E6%9E%97&oldid=85439116.

[10] 方匡南, 吴见彬, 朱建平, 等. 随机森林方法研究综述[J]. 统计与信息论坛, 2011, 26(3): 32-38.