Proceedings of CONF-BPS 2026 Symposium: GenAI, Labour Markets, and the Economics of Human and Financial Capital
DOI: 10.54254/2754-1169/2026.LD30669

Key Influencing Factors of Used Car Prices Driven by Machine Learning and Price Prediction

Meihan Yu

School of Economics and Management, Jiangxi Agricultural University, Nanchang, China
enddme0706@outlook.com

Abstract. Research on Price Prediction and Influencing Factors of Used Car Prices Based on Machine Learning Algorithms: Addressing the core issue of opaque pricing in the current used car market, this study utilized real car data, linear regression, and random forest among other machine learning methods to achieve precise price prediction and analysis of related influencing factors. During the research process, many distribution charts were used to more intuitively compare the correlations among multiple variables. The results show that the random forest model has superior predictive performance, with an R^2 coefficient reaching 0.9264. Through the analysis of feature importance, it was found that vehicle age and mileage are the most significant factors influencing the price of used cars, and their importance is significantly higher than that of other vehicle characteristics. This study provides a scientific pricing basis for used car transactions, effectively enhancing the fairness and transparency of the transactions, but there are still some shortcomings: the diversity of data needs to be strengthened, and the stability of the model's prediction needs to be addressed. The absence of these models' explainability is the crucial factor that leads to the discrepancy between theoretical research and practical operation analysis.

Keywords: Statistical Model, Linear Regression, Random Forest, Price Prediction

1. Introduction

The development of the used car market helps to promote economic circulation and consumption growth [1]. The market price is affected by factors such as vehicle age and mileage, and the pricing process is complex and relies on experience-based judgment, which is prone to causing transaction disputes. Utilizing big data and machine learning to build a pricing model has become a feasible solution to promote price standardization.

In recent years, machine learning has become the mainstream method for predicting the prices of used cars. Muhammed Ali YILMAZ achieved nearly perfect prediction accuracy ($R^2 = 1.000$) by using the CatBoost model, and quantified the influence of key features such as engine displacement, vehicle age, and mileage through SHAP analysis [2]. Bharambe based on empirical research in the Indian market found that random forests (RFs) significantly outperformed traditional models in terms of prediction accuracy and confirmed that the factory year and mileage were core pricing factors [3].

DOI: 10.54254/2754-1169/2026.LD30669

This study has developed a highly accurate and interpretable used car pricing model. By employing feature engineering and the RF algorithm, it quantitatively analyzes the influence of key features such as vehicle age and mileage, providing scientific pricing references for market transactions.

2. Methodology

This study is based on a dataset of 50,000 used car transactions from Kaggle, using data mining and machine learning techniques to build a scientific frame of vehicle price prediction and analysis [4]. Machine learning techniques can accurately predict used car prices, aiding buyers and sellers in informed decision-making in the automotive industry [5]. The following will depict a detailed account of the research exploration process.

2.1. Data source and preprocessing

This study conducted a rigorous cleaning of the initial dataset, which contained multiple dimensions such as manufacturer, vehicle model, engine displacement, fuel type, production year, mileage, and price. The process involved handling missing values, outliers, and duplicate values to ensure the reliability of the analysis foundation. For non-numeric features such as manufacturer and vehicle model, one-hot encoding (OneHotEncoding) was employed to convert them into numerical vectors, thereby avoiding errors in subsequent processing steps.

2.1.1. Feature engineering

In fact, before the research conducted the research, it was easy to observe that the original data merely presented a direct record of the facts, and there might be some complex correlations among the features. The study executed the feature engineering process that the research built, creatively derived two features and the formulas are:

$$Car_{Age} = Current\ Year - Year\ of\ Manufacture \tag{1}$$

$$MIleage_{per\ year} = \frac{Mileage}{Car_{Age}+1} \tag{2}$$

Based on the original features, there is a high degree of correlation (i.e., multicollinearity) between the year of manufacture and mileage. Generally, the mileage of new cars is lower, while that of older cars is higher. So the study created the variables car age and miles per year, which helps to some extent to break down variables finality and visualize the correlations between the features. Car age mainly captures the depreciation caused by time, while miles per year mainly captures the depreciation resulting from usage. For instance, using the manufacturing year and the current year to calculate the vehicle age is a more intuitive way to reflect the depreciation of a vehicle compared to a single absolute year figure. Similarly, the total mileage makes sense only when considered in conjunction with the vehicle age, a car with a 10-year age that has traveled 200,000 kilometers has a completely different level of engine wear compared to a car with a 20-year age that has also traveled 200,000 kilometers. Moreover, the total mileage range varies significantly among different brands and models. Mileage per year can unify the usage intensity of all vehicles to a relatively comparable standard (kilometers per year). In short, the features more accurately reflect the actual usage wear and tear of the vehicle [6].

DOI: 10.54254/2754-1169/2026.LD30669

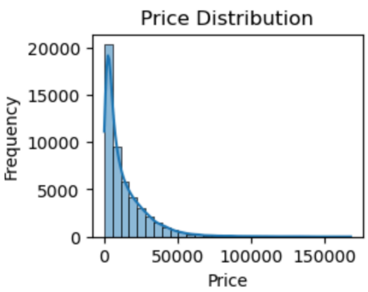
For the linear regression (LR) model, the study standardized (StandardScaler) all numerical features, which is a crucial and correct step, as it ensures the stability and accuracy of model convergence and it eliminates the differences in scale among the features. And transformed them into a distribution with a mean of 0 and a standard deviation of 1.

2.1.2. EDA: statistic summaries and correlation analysis

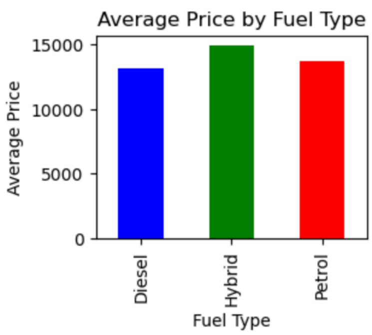
During the EDA phase, this study conducted a comprehensive exploratory analysis to reveal the inherent structure and patterns of the data. By drawing the histogram and kernel density estimation of the price distribution (Figure 1(a)), the overall distribution characteristics of used car prices were visually presented [7]. The data showed that the prices were significantly right-skewed, with most vehicle prices concentrated in the range of 0 - 25,000. As the price increased, the number of vehicles decreased sharply, and the peak of the kernel density curve was in the low-price range, confirming the concentration trend of the price distribution and reflecting that high-priced cars accounted for a minority in the market.

The grouped bar charts (Figure 1(b) and Figure 2(a)) further demonstrated the impact of different manufacturers (the top five brands: Porsche, BMW, Toyota, Ford, Volkswagen) and fuel types (diesel, hybrid, gasoline) on average price. The average price of new energy vehicles is the highest, while the prices of traditional fuel types are relatively more affordable [8].

Furthermore, Figure 2(b) shows a clear negative correlation between the two, meaning that the longer the mileage, the lower the price, which is in line with the basic laws of the used car market. In the figure, outliers can also be observed: in the low-mileage range, there are some high-priced vehicles (possibly new or near-new models), while in the high-mileage range, the prices tend to be close to zero, indicating that the residual value of the vehicles is extremely low.



(a)



(b)

Figure 1: (a) Histogram of price distribution; (b) Comparison of average prices for different fuel types (original)

DOI: 10.54254/2754-1169/2026.LD30669

(a) (b)

Figure 2: (a) Scatter plot of price and mileage; (b) Comparison of average prices among different manufacturers (original)

For numerical variables, this study employs the Pearson correlation coefficient matrix for quantitative analysis.

The correlation intensities between engine displacement, manufacturing year, mileage and price were quantitatively calculated through Figure 3. The analysis shows that manufacturing year has the strongest positive correlation with price (0.71), having the greatest impact on price; while mileage has the strongest negative correlation with price.

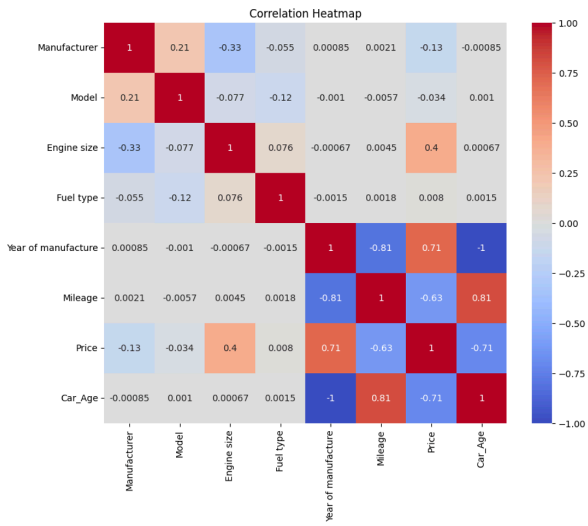


Figure 3. Correlation heatmap (original)

2.2. Model and evaluation

This research utilized two representative algorithms, namely RF regression and LR, for a systematic and comparison. The focus was on the performance and applicability differences in the task of predicting the prices of used cars.

DOI: 10.54254/2754-1169/2026.LD30669

2.2.1. RF regression algorithm

The RF algorithm, first put forward by Breiman in 2001, is an advanced regression method based on ensemble learning, which is regarded as a powerful statistical classifier, becoming more popular in many fields [9]. The RF algorithm enhances prediction performance by constructing multiple decision trees and integrating their results. This algorithm employs two randomization methods: using Bootstrap sampling to generate different training subsets and randomly selecting some features during node splitting. This design increases model diversity and reduces the risk of overfitting.

During the prediction stage, the RF averages the output results of all the decision trees to obtain the final prediction value:

$$y^{\hat{}} = \frac{1}{\mathrm{B}} \sum_{\mathrm{b}=1}^{\mathrm{B}} \mathrm{T}_{\mathrm{b}}(\mathrm{x}) \tag{3}$$

Among them, B represents the total number of trees, and Tb(x) is the prediction result of the b-th tree.

2.2.2. LR model

LR, as a classic statistical learning method, is renowned for its transparent model structure and intuitive parameter interpretation [10]. This method assumes that there is a linear relationship between the dependent variable and the independent variable. Establishing a mathematical model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon \tag{4}$$

Here, β_0 represents the interception, β_1 represents the slope, and ϵ represents the error. Using $y\ \hat{} = \beta_0\ \hat{} + \beta_1\ \hat{}\ x$ to predict Y, when the error is infinitely close to 0.

Based on the 50,000-group observation data of used car transactions {(x_1 , y_1), (x_2 , y_2),..., (x_n , y_n)}, it needs to estimate the unknown parameters β_0 and β_1 . This is a mathematical optimization problem: finding the optimal parameter estimates $\beta\hat{}_0$ and $\beta\hat{}_1$ to ensure that the regression line minimizes the overall deviation from all the data points.

Estimating the parameter model using the least squares method, it minimizes the sum of squared residuals between the predicted values and the actual values.

2.2.3. Experimental design and evaluation framework

To ensure the reliability of the model evaluation results, this study employed a stratified random sampling strategy. The complete dataset was divided into a training set and a test set at an ration of 8:2. The training set was used for learning and tuning of model parameters, while the test set served as an independent dataset for evaluating the model. In addition, the random seed was fixed at 42 to ensure consistency in the results of each experimental partition.

The model performance is evaluated using a double-index system: the coefficient of determination (R^2) , which is used to measure the model's ability to explain the changes in used car prices, with a value closer to 1 indicating a higher goodness of fit; the root mean square error (RMSE) reflects the average deviation between the predicted values and the actual transaction prices, and a smaller value indicates higher prediction accuracy. By comprehensively examining the performance of the two algorithms in these indicators, and find the best evaluation model.

DOI: 10.54254/2754-1169/2026.LD30669

3. Results

The performance evaluation of the two models on the test set is shown in Table 1. The R^2 value of LR is 0.720, and the RMSE value is 8712.267. At the same time, the R^2 value of RF is 0.940, and the RMSE value is 4034.715.

Table 1. The performance value of the two models on the test set

	R^2	RMSE
LR	0.720	8712.267
RF	0.940	4034.715

According to the RF model, the ranking of the top five important features can be obtained in a single run, which is presented in Table 2. The results of the feature importance analysis for price prediction are presented as follows: engine size (0.350), model (0.304), manufacturer (0.289), mileage (0.048), year of manufacture (0.007).

Table 2. Feature importance

Feature	Engine size	Model	Manufacturer	Mileage	Year of manufacture
Importance	0.350	0.304	0.289	0.048	0.007

4. Discussion

This study verified the effectiveness of machine learning in predicting the prices of used cars through systematic modeling. The RF regression model achieved the best prediction performance on the test set (R^2= 0.940). The feature importance analysis revealed that engine displacement, vehicle model, and manufacture are the three core factors influencing the price.

Compared with the existing studies, the feature of engineering in this research is more comprehensive. By constructing derived variables such as "annual average mileage", the model performance has been further enhanced. The study confirmed that RFs outperform LR in capturing complex nonlinear relationships, which is consistent with the earlier findings of Pal [11]. However, although SHAP analysis was introduced to enhance model interpretability, there is still room for improvement in inferring causal relationships between features.

In view of the limitations of the current research, future work can be carried out in the following aspects: draw on the multi-platform data integration method of Hemendiran to enhance data diversity [12]. Adopt the cross-domain pattern matching technology of Shraga to improve the integration effect of heterogeneous data sources [13]. Introduce the SMAT attention mechanism of Zhang et al to achieve automated pattern matching [14].

This study has the following limitations: The data source is limited, making it difficult to fully reflect the actual market situation; the feature dimensions are incomplete, lacking key information such as accident records and maintenance histories; the stability of predictions for extreme cases is insufficient.

Given the limitations of the existing research, the future development direction can refer to the neural network and machine learning integration scheme proposed by Sruneethi, establishing a dynamic update mechanism to achieve the collaborative optimization of price prediction and depreciation curve analysis [15].

DOI: 10.54254/2754-1169/2026.LD30669

5. Conclusion

This study successfully developed and validated a used car price prediction model based on the RF algorithm. The empirical results show that this model performs well in the quantitative analysis of key pricing factors: vehicle age has a significant negative impact on the price; engine size has a positive impact on the price, but this impact gradually weakens; manufacturers and fuel types exhibit clear brand premiums. Not only does it immediately provide reliable and sustainable suggestions for the dealer platform, and identify overpriced or underpriced inventories for rebalancing, but it also provides important theoretical and technical support for improving the transparency of the used car market.

Although the model has achieved the expected results, it must be acknowledged that it still has deficiencies in terms of comprehensiveness and feature completeness. Therefore, future research will focus on integrating multi-dimensional data and establishing a more complete and stable evaluation system, providing more accurate quantitative basis for industry decision-making.

References

[1] Abdelouahed, S., Abla, R., Asmae, E., & Abdellah, A. (2024). Harnessing feature engineering to improve machine learning: A review of different data processing techniques. 2024 International Conference on Intelligent Systems and Computer Vision (ISCV), 1–6.
[2] Anju, O. M., Kruthika, M. M. K., S. (2024). Predicting used car prices using machine learning: A comparative analysis of regression and ensemble models. International Research Journal on Advanced Engineering Hub (IRJAEH).
[3] Barron, J., Taylor, B., & Umbeck, J. R. (2004). Number of sellers, average prices, and price dispersion. International Journal of Industrial Organization, 22(8–9), 1041–1066.
[4] Bharambe, P., Bagul, B., Dandekar, S., & Ingle, P. (2022). Used car price prediction using different machine learning algorithms. International Journal for Research in Applied Science and Engineering Technology.
[5] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.
[6] Feng, B. (2024). Multiple contradictions intertwined in the international oil market in 2025. China Oil & Gas, 2024(6).
[7] Hemendiran, B., & Renjith, P. (2023). Predicting the prices of the used cars using machine learning for resale. 2023 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS), 1–5.
[8] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2023). Linear regression. In An introduction to statistical learning with applications in Python (pp. 69–127). Springer Nature.
[9] Mandys, F. (2023). Search for an affordable clean car: Pricing of conventional and clean automobiles. Transportation Research Part D: Transport and Environment, 118, 103738.
[10] Mohsen Behdani. (2025). Car sales dataset: Model, features, and pricing [Data set]. Kaggle. https: //www.kaggle.com/datasets/msnbehdani/mock-dataset-of-second-hand-car-sales/data
[11] Muhammed Ali YILMAZ. (2025). Predicting used car prices with ML insights [Notebook]. Kaggle. https: //www.kaggle.com/code/muhammedaliyilmazz/predicting-used-car-prices-with-ml-insights
[12] Pal, N., Arora, P., Palakurthy, S. S., Sundararaman, D., & Kohli, P. (2017). How much is my car worth? A methodology for predicting used cars' prices using random forest. arXiv.
[13] Shraga, R., Gal, A., & Roitman, H. (2020). ADnEV: Cross-domain schema matching using deep similarity matrix adjustment and evaluation. Proceedings of the VLDB Endowment, 13(9), 1401–1415.
[14] Sruneethi, C., & Lokeswari, E. (2025). Prediction of used car prices using artificial neural networks and machine learning. International Journal of Research Publication and Reviews.
[15] Zhang, J., Shin, B., Choi, J. D., & Ho, J. C. (2021). SMAT: An attention-based deep learning solution to the automation of schema matching. In L. Bellatreche, M. Dumas, P. Karras, & R. Matulevičius (Eds.), Advances in Databases and Information Systems (Vol. 12843, pp. 260–274). Springer.