Proceedings of the 3rd International Conference on Mathematical Physics and Computational Simulation
DOI: 10.54254/2753-8818/2025.22553

# *Unveiling Factors Affecting Price of Second-hand Cars Using Multiple Linear Regression Models and Random Forest*

**Wanning Luo**

*Adcote School Shanghai, Shanghai, China*
*wanning_luo213@outlook.com*

***Abstract:*** With the development of global economy and evolution of consumer concepts, automobiles, as an important means of transportation and consumer product, have seen a continuous increase in market demand. In recent years, with the growing number of vehicles in use, the second-hand car market has gradually become an important part of the automotive circulation sector. This study investigates the factors influencing second-hand car pricing in the current market environment. By combining multiple linear regression and Random Forest analysis, the author examines the significance and impact of various factors on the final selling price of second-hand vehicles. The data were collected from Kaggle and supplemented by relevant academic literature, covering variables such as power, transmission, mileage, engine type, and fuel type. The author seeks to reveal the characteristics and trends of current second-hand car pricing, provide guidance for marketing strategies, offer a basis for relevant policy-making, and explore the long-term impact of second-hand car pricing on the entire consumer market and social-economic development. As the results, the vehicle's power has the greatest impact on the final transaction price, which is the pricing in the second-hand car market, and it is positively correlated.

***Keywords:*** Second-hand cars, vehicle's power, multiple linear regression, random forest.

## 1. Introduction

In the second decade of the 21st century, China's automotive market has undergone tremendous changes. As of the end of September 2023, the number of vehicles in China has exceeded 330 million, among which the number of new energy vehicles is 18.21 million. Meanwhile, only about one-third of people holding driver's licenses own a car, which indicates that China's automotive market still has huge potential for growth [1]. As a consumer product, automobiles need to be constantly updated and replaced to stimulate consumption and maintain market vitality. Many old vehicles are being phased out and enter the second-hand car market. In recent years, the scale of the second-hand car market has been expanding continuously, and the transaction volume has been increasing year by year. However, the volatility and uncertainty of second-hand car prices have also brought many challenges to buyers and sellers. As a core element of second-hand car transactions, the pricing mechanism and influencing factors of second-hand car prices have always been a focus for market participants. Therefore, an in-depth study of the factors affecting second-hand car pricing is of great significance

DOI: 10.54254/2753-8818/2025.22553

for understanding current market dynamics, predicting future development trends, and guiding the marketing strategies of second-hand car companies.

There is trend reflects the increased activity in the second-hand car market and the growing consumer demand for used vehicles [2]. Regarding the valuation of second-hand cars, researchers have explored the issue from multiple perspectives. By using the ABC analysis method, Zhao categorized these factors (physical, functional, and economic) were weighted and valued to establish a more rational second-hand car valuation system [3]. Ming and Ouyang compared the current market value method and the replacement cost method, ultimately selecting the latter as the evaluation approach and modifying the model based on the technical conditions of pure electric second-hand cars [4].

In terms of evaluation models, Li and Guan utilized a dataset of nearly 1,000 entries, combining hedonic pricing theory, they employed the Random Forest algorithm to select important features and used the Gradient Boosting Tree model for prediction. The results showed a model fit of 0.98 and a mean absolute error of 0.37, demonstrating the model's significant effectiveness in valuing pure electric second-hand cars [5]. Additionally, Zhou and He proposed a weighted combination model of neural networks and improved LightGBM algorithms, which exhibited stronger predictive capabilities compared to single models [6]. Ning et al. constructed second-hand car transaction samples using RStudio software based on hedonic pricing theory. They calculated the importance values of each feature, built a Random Forest regression model for price evaluation, and compared it with the replacement cost method [7]. Jin employed three machine learning algorithms—XGBoost, LightGBM, and Random Forest—to model second-hand car pricing. The results showed that the Random Forest-based pricing strategy had significant advantages in precision and effectiveness [8]. However, the current second-hand car market still faces challenges. Issues such as information asymmetry, trust crises, financing difficulties, and consumer rights protection have led to an overall pessimistic market environment [9].

Overall, China's second-hand car market has shown great potential for development under the backdrop of policy support and consumption upgrades. In the future, with the progress of technology and market standardization, second-hand car valuation and transactions will become more scientific and efficient, providing consumers with more cost-effective choices.

## 2. Methodology

### 2.1. Data source and description

In the current economic environment, it is crucial to have a deep understanding of the pricing models for second-hand cars. Prices are highly likely to determine the vibrancy of the second-hand car market. By studying the factors that influence prices, one can promote the healthy development of the second-hand car market, encourage more car owners to dispose of their old vehicles reasonably, and allow cars to circulate among owners with different needs. This extends the lifespan of vehicles and reduces the consumption of resources.

This paper primarily relies on data from Kaggle for analysis. The platform offers a wealth of datasets, making it highly suitable for this study. Additionally, the author has referenced some data from academic literature, which, though limited in quantity, provides valuable supplementary information.

This study employs multiple linear regression to investigate the relationship between second-hand car pricing and its influencing factors. This is because, according to several papers the author reviewed, this method is widely used in many studies. For example, in 2023, Zheng, Li, and Guo, based on data from the second-hand car trading platform 58.com, analyzed the meaning and correlation of relevant data. They cleaned and processed the data, established a multiple linear

DOI: 10.54254/2753-8818/2025.22553

regression model for predicting second-hand car transaction prices, and evaluated the model's performance using indicators such as average relative error and accuracy [10].

### 2.2. Data collection

The following variables are part of the data obtained for this study. These data were collected from the Kaggle website and are used in this paper, such as seats, power, transmissions, owner type, etc. Since the author aims to study the factors influencing second-hand car pricing, the paper sets the final selling price of the second-hand car as the dependent variable Y, and the other factors influencing it as independent variables X1, X2, ⋯, X10. The variables used in this study and the description of each variable are included in the following Table 1.

Table 1: The definition of variables

| Symbol | Variable | Description |
|---|---|---|
| Y | sales_price | The selling price of the car, which is the target variable to predict. |
| X1 | seats | The number of seat available in the car |
| X2 | power | The maximum power output of the car |
| X3 | transmission | The transmission type of the car (e.r., Manual, Automatic) |
| X4 | owner_type | The number of previous owners of the car (e.g., First, Second) |
| X5 | mileage | The fuel efficiency of the car in kilometers per liter |
| X6 | engine | The engine capacity of the car |
| X7 | year | The manufacturing year of the car |
| X8 | fuel_type | The type of fuel used by the car (e.g., Petrol, Diesel, Electric, etc.) |
| X9 | kilometers_driven | The total kilometers driven by the car |
| X10 | model | The model of the car (e.g., Camry, Civic, Mustang, etc.) |

### 2.3. Introduction of methods

A sort of mathematical regression model called multiple linear regression examines how a dependent variable varies in response to two or more independent factors [11]. Because in reality, the occurrence of a phenomenon is often the result of multiple factors, it is meaningless and unrealistic to estimate the occurrence of this phenomenon using only one variable. Therefore, a joint prediction or estimation by multiple independent variables will make the results more valid and realistic. In the general case, assuming n different predictor variables, the formula for a multiple linear regression model is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon \quad (1)$$

where Y is the dependent variable; $X_1, \cdots, X_n$ is an independent variable; $\beta_0$ is the intercept term; $\beta_1 ... \beta_n$ is the regression coefficient, which represents the effect of each independent variable on the dependent variable; $\varepsilon$ is an error term that represents random variation that cannot be explained by the model [12,13].

## 3. Results and discussion

### 3.1. Correlation analysis

Table 2 represents the correlations among the ten variables the author used: seats, power, transmission, owner_type, mileage, engine, year, fuel_type, kilometers_driven, and model. This article will use correlation coefficients to represent the correlations between the final selling price of second-hand cars and these variables. The specific analysis is shown in Table 2.

DOI: 10.54254/2753-8818/2025.22553

The standardized coefficients for seats and power are positive, which indicates that these two independent variables have a positive correlation with the price of second-hand cars. In contrast, the other variables have negative standardized coefficients. Therefore, transmission, owner_type, mileage, engine, year, fuel_type, kilometers_driven, and model have a negative correlation with the value of second-hand cars at the time of sale.

Table 2: Standardized coefficients

| | Standardized Coefficients | | Standardized Coefficients |
|---|---|---|---|
| seats | 0.053 | engine | -0.099 |
| power | 0.627 | year | -0.077 |
| transmission | -0.232 | fuel_type | -0.188 |
| owner_type | -0.077 | kilometers_driven | -0.165 |
| mileage | -0.156 | model | -0.102 |

The Table 3 above shows that selected seats, power, transmission, owner_type, mileage, engine, year, fuel_type, kilometers_driven, and model as the independent factors and price as the dependent variable for the linear regression analysis. The table also includes the model equation:

$$\begin{aligned}\text{Price} = {} & +136239567.774 + 70906.023 * \text{Seats} + 8144.062 * \text{Power} \\ & -467370.383 * \text{Transmission} - 111151.443 * \text{Owner}_{\text{T}}\text{ype} \\ & -47285.246 * \text{Mileage} - 156.578 * \text{Engine} - 65892.573 * \text{Year} \\ & -374434.247 * \text{Fuel}_{\text{T}}\text{ype} - 18602.282 * \text{Brand} \\ & -18.144 * \text{Kilometers}_{\text{D}}\text{riven} - 5871.645 * \text{Model}\end{aligned} \quad (2)$$

With an R-squared value of 0.852, the model can explain 85.2% of the variation in price.

Table 3: Unstandardized coefficient B and standard error

| | Std. Error | B |
|---|---|---|
| Constant | 124436654.602 | 136239567.774 |
| seats | 69019.212 | 70906.023 |
| power | 1275.243 | 8144.062 |
| transmission | 107400.372 | -467370.383 |
| owner_type | 69234.500 | -111151.443 |
| mileage | 19461.384 | -47285.246 |
| engine | 146.333 | -156.578 |
| year | 61517.916 | -65892.573 |
| fuel_type | 110017.597 | -374434.247 |
| kilometers_driven | 7.883 | -18.144 |
| model | 2561.693 | -5871.645 |

## 3.2. Linear regression

After conducting an F-test on the model, it was found that the model is significant ($F = 46.068$, $p = 0.000 < 0.05$). This indicates that the price is influenced by at least one of the following variables: power, transmission, mileage, kilometers_driven. It supports part of the hypothesis that these factors will affect the changes in the final transaction price of second-hand cars.

Finally, as shown in Table 4, the specific analysis of the regression model indicates that the p-values of certain variables are less than 0.05. The regression coefficient for power is 8144.062 (t =

DOI: 10.54254/2753-8818/2025.22553

6.386, p = 0.000 < 0.05). This positive value shows a significant positive correlation with price. Another variable with a positive regression coefficient is seats. However, its p-value is not less than 0.05, indicating that the factor seats does not affect the final selling price of second-hand vehicles.

Table 4: Result of linear regression

| | t | p | VIF | tolerance |
|---|---|---|---|---|
| Constant | 1.095 | 0.002** | - | - |
| seats | 1.027 | 0.307 | 1.593 | 0.628 |
| power | 6.386 | 0.000** | 5.727 | 0.175 |
| transmission | -4.352 | 0.000** | 1.698 | 0.589 |
| owner_type | -1.605 | 0.112 | 1.364 | 0.733 |
| mileage | -2.430 | 0.001** | 2.466 | 0.405 |
| engine | -1.070 | 0.288 | 5.073 | 0.197 |
| year | -1.071 | 0.287 | 3.086 | 0.324 |
| fuel_type | -3.403 | 0.017* | 1.814 | 0.551 |
| kilometers_driven | -2.302 | 0.004** | 3.073 | 0.325 |
| model | -2.292 | 0.024* | 1.179 | 0.848 |

It is worth noting that transmission is one of the factors that shows a negative correlation with price. Its regression coefficient is -467,370.383, indicating that in the second-hand car market, vehicles with automatic transmission tend to be priced higher than those with manual transmission. In this context, the regression coefficient for mileage is -47,285.246, suggesting that a lower fuel consumption per kilometer for a second-hand car makes it more likely to be sold at a higher price. Additionally, the p-values for fuel_type, kilometers_driven, and model are all less than 0.05, with regression coefficients of -374,434.247, -18.144, and -5,871.645, respectively. This indicates that these three factors are negatively correlated with the price of second-hand cars.

Through the multiple linear regression model, it is found that the p-values for owner_type, engine, and year are all greater than 0.05. This suggests that these three variables are not related to pricing. It appears that sellers do not consider the number of previous owners, the type of engine, or the year of manufacture to be factors that affect the sale price of a second-hand car. In summary, power has a significant positive impact on price. Meanwhile, transmission, mileage, fuel_type, kilometers_driven, and model have significant negative impacts on price. However, seats, owner_type, engine, and year do not have any impact on price.

Table 5: The value of the weights of each independent variable

| variables(x) | weighted value | variables(x) | weighted value |
|---|---|---|---|
| seats | 0.013 | engine | 0.035 |
| power | 0.625 | year | 0.020 |
| transmission | 0.119 | fuel_type | 0.020 |
| owner_type | 0.011 | kilometers_driven | 0.021 |
| mileage | 0.100 | model | 0.033 |

### 3.3. Random forest

In addition to linear regression, an analytical method known as Random Forest was also employed. Through Random Forest, it is possible to identify the independent variables that have the most significant impact on the target variable. Table 5 shows the importance weights of each independent

DOI: 10.54254/2753-8818/2025.22553

variable. The feature weights indicate the relative importance of each variable in contributing to the model, with their sum equaling 1. As shown in the Table 5, Power accounts for 62.52% of the total weight, making it the most significant feature and a key contributor to the model. Transmission accounts for 11.94%, and Mileage accounts for 10.05%. Together, these three features account for a combined 84.51% of the total weight. The remaining seven features—Engine, Model, Kilometers_Driven, Year, Fuel_Type, Seats, and Owner_Type—account for 3.54%, 3.33%, 2.13%, 2.03%, 2.02%, 1.33%, and 1.12% of the total weight, respectively.

### 3.4. Comparison

From the combination of these two methods, the vehicle's power is positively correlated with the pricing of second-hand cars and has the most significant impact on the final price. The power of a vehicle directly affects its dynamic performance, including acceleration, hill-climbing ability, and high-speed stability. A high-power engine can provide stronger acceleration, allowing the vehicle to reach higher speeds in a shorter time. This is highly attractive to consumers who seek driving pleasure and high performance. Additionally, high-power vehicles perform better in hill-climbing and high-speed driving scenarios, offering better stability and handling.

Transmission ranks second in terms of its impact on pricing and has a negative correlation with second-hand car prices, which is not surprising. Automatic transmissions are favored for their ease of operation. Mileage, which ranks third in terms of weight, does not have as significant an impact as one might expect. Vehicles with high fuel efficiency can significantly reduce fuel consumption over the long term, thereby lowering the operating costs for car owners. For consumers, lower fuel consumption means lower operating costs, making high-fuel-efficiency vehicles more attractive in the market and thus reflecting a higher value in second-hand car pricing.

Among the remaining independent variables, fuel_type, kilometers_driven, and model all have negative correlations with second-hand car prices. However, their respective weights are 0.02, 0.02, and 0.03, indicating that their impact on price is minimal. Meanwhile, seats, owner_type, engine, and year do not have a significant impact on the price. Overall, the vehicle's power has the greatest impact on the final transaction price, which is the pricing in the second-hand car market, and it is positively correlated.

## 4. Conclusion

This paper investigates the factors influencing second-hand car pricing in China under the context of consumption trend upgrades in 2024, based on information from 100 vehicles sold in the second-hand market in 2024. By utilizing multiple linear regression and random forest models to analyze the factors influencing second-hand car prices. The findings reveal that several aspects of a vehicle significantly affect its price, including its power, transmission type, mileage, fuel type, total kilometers driven, specific model, seating capacity, previous ownership type, engine specifications, and the year of manufacture. Within the multiple linear regression framework, vehicles with higher power tend to command a higher price. Conversely, factors such as transmission type, mileage, fuel type, total kilometers driven, and the specific model of the car are found to have a negative correlation with price, meaning that changes in these aspects are associated with a decrease in the vehicle's value. However, the analysis also shows that seating capacity, previous ownership type, engine specifications, and the year of manufacture do not have a significant impact on the price of second-hand cars. The random forest analysis further highlights that power is the most influential factor in determining the selling price of a second-hand car. Transmission type follows in importance, with its influence being comparable to that of mileage.

DOI: 10.54254/2753-8818/2025.22553

Based on this paper, it can provide better guidance for both sellers and buyers in the second-hand car market. Sellers can use these factors to more effectively set vehicle prices, attract more customers, and maximize profits. Consumers can refer to these factors to purchase second-hand cars that ensure good value for money. Combining the results of both methods, power is positively correlated with vehicle price and has the highest weight (0.63). Transmission follows with a weight of 0.12, slightly lower than the weight of competitor's price but negatively correlated with total sales volume. The remaining independent variables account for only 0.15 and have a limited impact on second-hand car pricing. As the results, power has the most significant impact on the final transaction price, which is the pricing in the second-hand car market, and it is positively correlated. Overall, this study has provided a valuable perspective on understanding the pricing mechanisms of second-hand cars, but there is still significant room for improvement. By introducing more sophisticated methods, expanding the range of variables, and increasing the volume of data, future research is expected to offer more precise and comprehensive theoretical support and practical guidance for the pricing of second-hand cars.

## References


[1] *Wang Zhishan. (2024). The second-hand car market is poised to enter a rapid growth track. Business Observer (20), 6-9.*

[2] *Tian Meng & Chang Kaidi. (2025). Analysis of the used passenger car market from January to November 2024. Auto Vertical and Horizontal (01), 110-113.*

[3] *Zhao Qiuyuan. (2024). Analysis of influencing factors of used car value and appraisal value case. Automobile Repair & Maintenance (12),106-108.*

[4] *Ming Jianping & Ouyang Ronghua. (2024). Research on the value evaluation of used cars of pure electric vehicles based on the improved replacement cost method. Automotive Practical Technology (20), 124-128.*

[5] *Li Chen & Guan Yongjun. (2024). Value evaluation of new energy used vehicles based on gradient boosting tree model. Business Observer (29), 21-25.*

[6] *Zhou Yuan & He Botao. (2024). Used car price prediction method based on weighted combination model. Computer and Digital Engineering (05), 1449-1452 1458.*

[7] *Ning Yuantian, Zhu Qin, Wei Jinming, Yin Pengfei & Zhu Honglin. (2024). Research on used car price evaluation based on random forest regression model. Journal of Guilin Institute of Aerospace Technology (01), 82-93.*

[8] *Jin Zhirong. (2023). Research on Used Car Pricing Strategy Based on Random Forest. Science & Technology Economic Market (06), 89-91.*

[9] *Bai Weizhong. (2023). Research on Business Model Innovation in Domestic Second-hand Car Market under the Background of "Internet". Times Auto (23), 166-168.*

[10] *Zheng Aiping, Li Binbin and Guo Chuanhao. (2023). "Used car transaction price prediction analysis based on linear regression and neural network model". Intelligent Computers and Applications (09),103-110.*

[11] *Soffritti, G., & Galimberti, G. (2010). Multivariate linear regression with non-normal errors: A solution based on mixture models. Statistics and Computing, 21(4), 523–536.*

[12] *Sheather, S. J. (2009). A modern approach to regression with R. Springer.*

[13] *Sangeetha, J. Margaret, & K. Joy Alfia. (2024). Financial stock market forecast using evaluated linear regression based machine learning technique. Measurement: Sensors, 31, 100950.*