ISSN: 2472-3703 DOI: 10.6911/WSRJ.202509_11(9).0006

Price Prediction and Analysis of Price Influencing Factors for Second-hand Car Sales in AutoTrader Based on XGBoost Algorithm

Jiayao Huang[1, *]

[1]School of Mechanical and Automotive Engineering, Ningbo University of Technology, Ningbo, Zhejiang, 315336, China

[*]Corresponding author's e-mail: Huang2579@outlook.com

Abstract

The global used car market continues to expand, reaching a scale of 1.6 trillion US dollars in 2023. In 2024, China's transaction volume reached 19.61 million units, setting a new high. However, information asymmetry, sharp price fluctuations, and subjective assessment severely constrain market efficiency. To solve the pricing problem, this study, based on a large amount of data from the AutoTrader platform in the UK, builds an XGBoost high-precision price prediction model, integrates multiple vehicle attributes and market characteristics, and achieves low-error residual value estimation. At the same time, random forest feature analysis is used to quantify the contribution of key factors, revealing the hierarchical influence structure, providing intelligent and data-driven pricing decision support for all parties involved in the transaction, and promoting the market to transform towards transparency and efficiency.

Keywords

Used car evaluation; XGBoost algorithm; Random forest; Used Car Valuation System.

1. INTRODUCTION

1.1. Research Background and Significance

Used cars, as an important link in the entire life cycle of automobiles, play an irreplaceable role in revitalizing the existing market, promoting new growth, and stimulating automobile consumption [1]. In the past five years, the global used car market has expanded significantly. In 2023, its scale reached 1.6 trillion US dollars, an increase of 28% compared to 2019. With the support of Chinese policies, the transaction volume reached 19.61 million units in 2024, setting a new high and increasing by 6.52% year-on-year.

The AutoTrader platform, as the largest automotive trading platform in the UK, has an average of over 55 million unique visitors per month, accounting for 82% of the online traffic for used cars in the UK. The platform offers a wide range of used cars for users to choose from, and updates 500,000 price and condition records daily to meet the needs of different users.

The core problem in determining the price of used cars lies in the market failure caused by multiple factors. The fundamental obstacle is the information asymmetry due to the seller's possession of vehicle condition information, resulting in significant price fluctuations; the buyers and sellers are trapped in a pricing dilemma, with the seller unable to set a reasonable range, severely restricting the circulation efficiency. A scientific assessment system and data support are urgently needed.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202509_11(9).0006

1.2. Research Objectives and Contents

This study aims to build a vehicle resale price prediction model based on the XGBoost algorithm. By analyzing the data from the AutoTrader platform, it accurately estimates the vehicle residual value and quantitatively assesses the influence weights of various factors such as mileage and brand on the price. This provides intelligent decision-making support and data-driven basis for pricing in the used car market transactions.

This research focuses on the intelligent pricing of the used car market. It integrates multiple dimensions of features such as vehicle attributes and macroeconomic indicators to achieve precise and low-error prediction of sales prices. At the same time, it employs feature importance analysis to quantify the contribution of key influencing factors and reveal the hidden patterns of the market.

The research results will provide consumers with fraud risk warnings and fair trading benchmarks, assist dealers in formulating differentiated pricing strategies, assist regulatory authorities in establishing a price transparency assessment system, promote the transformation of the used car market from experience-driven to data-driven, and build an efficient and transparent trading ecosystem.

1.3. Research methods and technical approach

This study integrated the used car transaction data from Kaggle, after cleaning and feature engineering. An XGBoost algorithm was used for modeling. Through time series cross-validation and grid search to optimize hyperparameters, and combined with feature importance analysis to quantify the influencing factors. Finally, a high-precision and interpretable pricing model and comparison map were output, as shown in Figure 1., providing a toolchain for market decision-making.

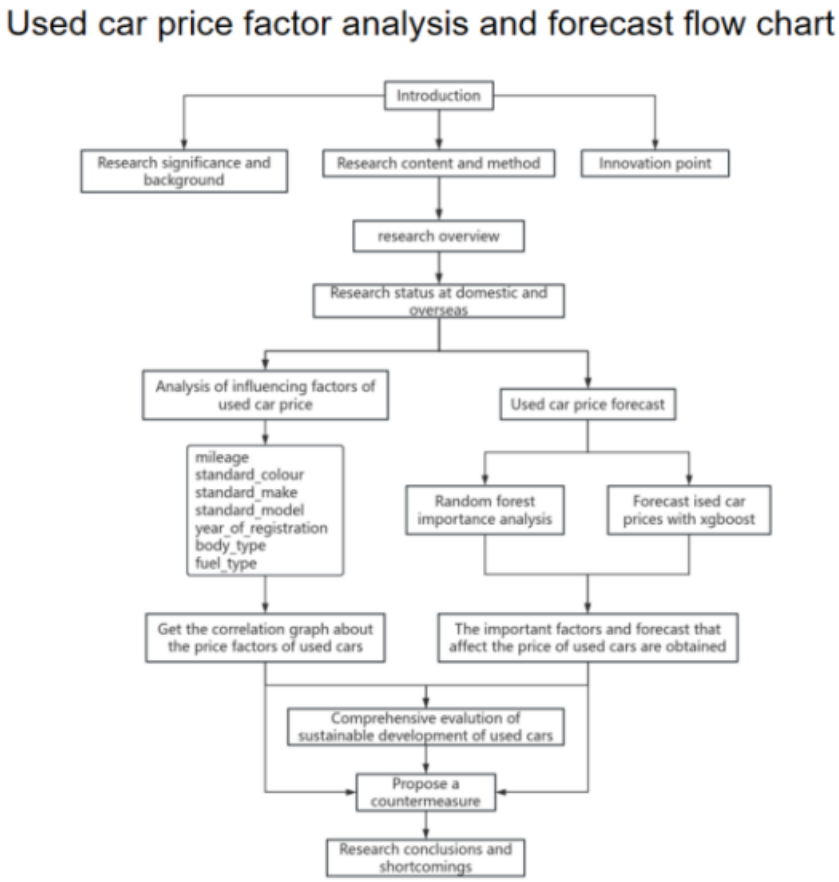

Figure 1. Technology roadmap

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202509_11(9).0006

2. LITERATURE REVIEW

In China, with the continuous development of big data and machine learning technologies, an increasing number of scholars and enterprises are beginning to utilize these technologies to explore the potential value of the used car market.

In the current domestic research field on the assessment of second-hand car prices, some scholars have conducted in-depth exploration and successfully applied the particle swarm optimization algorithm combined with the generalized regression neural network to construct a second-hand car value assessment model [2]. Moreover, the research community has witnessed the application of LASSO regression models, extreme gradient boosting algorithms, and lightweight gradient boosting machine algorithms in the field of second-hand car price prediction [3]. At the same time, the linear regression method based on traditional statistics has also been introduced into the second-hand car valuation system [4]. It is worth noting that some of the rigorously verified second-hand car price prediction models have been deployed in second-hand car trading platforms, which has greatly facilitated users' rapid understanding and comprehension of the market prices of second-hand cars.

In the academic community abroad, research on the prediction of second-hand car prices and its optimization strategies has undergone a long and in-depth exploration. The research process covers the initial construction of second-hand car price prediction based on supervised machine learning techniques, to the practical application of second-hand car price prediction in specific market areas, such as Saudi Arabia [5]. Further, scholars continuously challenge and optimize the application of machine learning algorithms in second-hand car price prediction, aiming to improve prediction accuracy and generalization ability [6]. Additionally, an innovative mixed prediction model for second-hand car prices that integrates consumer personal preference information has been proposed, adding a personalized dimension to price prediction [7].

This research was proposed based on such a background. By building a used car price analysis system based on the XGBoost algorithm, our aim is to provide an efficient and accurate tool for predicting used car prices, and to offer strong support for the standardization of the used car market, the optimization of the industrial chain, and policy formulation.

3. DATA COLLECTION AND DESCRIPTION

3.1. Data Source and Acquisition

This study utilized AutoTrader, a major used car trading platform in the UK, as the primary data source. Data from 402,005 used car transactions were obtained through the Kaggle platform.

3.2. Dataset description

This dataset includes both structured and unstructured fields, totaling 12 categories. Basic attributes: mileage, standard make, standard model, year of registration. Mechanical configuration: fuel type, body type. Specification parameters: standard color. Transaction information: vehicle condition, price. Other information: reg code, public reference, crossover car and van.

3.3. Data quality assessment

After data analysis, it was found that important values such as body_type and fuel_type had data missing. The total number of missing values was 72,111, accounting for 1.49% of the total. Additionally, we also conducted an inspection of the outliers in the initial data and discovered that the year_of_registration had entries that were earlier than 1900.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202509_11(9).0006

4. DATA PREPROCESSING AND DATA ENGINEERING

4.1. Data cleaning

This study employed a stratified processing strategy for data cleaning.

Based on the analysis of feature importance, four fields, namely public_reference, reg_code, crossover_car_and_van, and vehicle_condition, were removed.

According to the importance of the fields, the proportion of missing values, and the mechanism of data differentiation, the missing values were handled: year_of_registration was filled with the median to consider the distribution; standard_colour was filled with the mode after identifying the pseudo-missing values; for fields with low missing rates such as body_type and fuel_type, rows with random missing values were directly deleted without affecting statistical inference.

For outliers, set a business threshold, delete records where mileage is less than or equal to 100 miles, exclude the display of vehicles or entries with errors; remove vehicle data where year_of_registration is less than or equal to 1999, and focus on the second-hand vehicle sales data from the past 25 years.

4.2. Data set division and preprocessing process

This study carried out the separation of the target variable and the division of the training-test set for preparing the model.

It clearly set the price as the predicted target variable y, and created the feature matrix X which included all the cleaned fields except the price. The division ratio was 80% for the training set and 20% for the test set.

5. CONSTRUCTION OF A PRICE PREDICTION MODEL BASED ON XGBOOST

This model is used for regression tasks, with the core objective being to predict the continuous target variable of product prices. It is implemented using the XGBRegressor class and uses mean squared error as the loss function.

Model tuning is carried out using Bayesian optimization, which is 3-5 times more efficient than grid search and intelligently explores the most promising parameter regions.

The tuning goals include minimizing the validation set RMSE or MAE, while monitoring the R^2 value to ensure the model's explanatory power. The cross-validation strategy for the model is 5-fold cross-validation.

Table 1. Model Performance Improvement Comparison Table

Index	Archetype	Optimized model	Improvement percentage
MSE	86,556,419	79,218,653	Decrease by 8.5%
R^2	0.8236	0.8512	Increase by 3.4%

Through systematic parameter search,the data is shown in Table 1. the model's MSE decreased by 8.5% and R^2 increased by 3.4%. Bayesian optimization can further shorten the tuning time and obtain a better combination of hyperparameters. The final model significantly reduces the prediction error while maintaining high explanatory power.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202509_11(9).0006

6. MODEL EVALUATION, PREDICTION AND VALIDATION

6.1. Evaluation index system

In this study, RMSE, MAE, MAPE, R^2, sMAPE and explained variance were selected as the model evaluation index system, covering the most core and commonly used indicators in regression tasks.

6.2. Model performance evaluation

The model has excellent accuracy: RMSE = 22.67, MAE = 21.40 indicate that the overall error is small and there are no extreme outliers; MAPE = 6.68% and sMAPE = 6.71% show that the percentage error is low and balanced, which is extremely suitable for the used car scenario; R^2 = 0.9742 and the explained variance = 0.9753 confirm that the model has an extremely strong ability to capture data, and the unexplained variance is very small.

Based on the above analysis, the comprehensive evaluation of the model performance is excellent. All core indicators perform well, with low error rates, strong explanatory power, high prediction consistency, and no significant deviation tendencies.

6.3. Model visualization

The correspondence between the actual prices and the predicted prices of used cars reveals the key characteristics of the model's predictive performance,as shown in Figure 2.

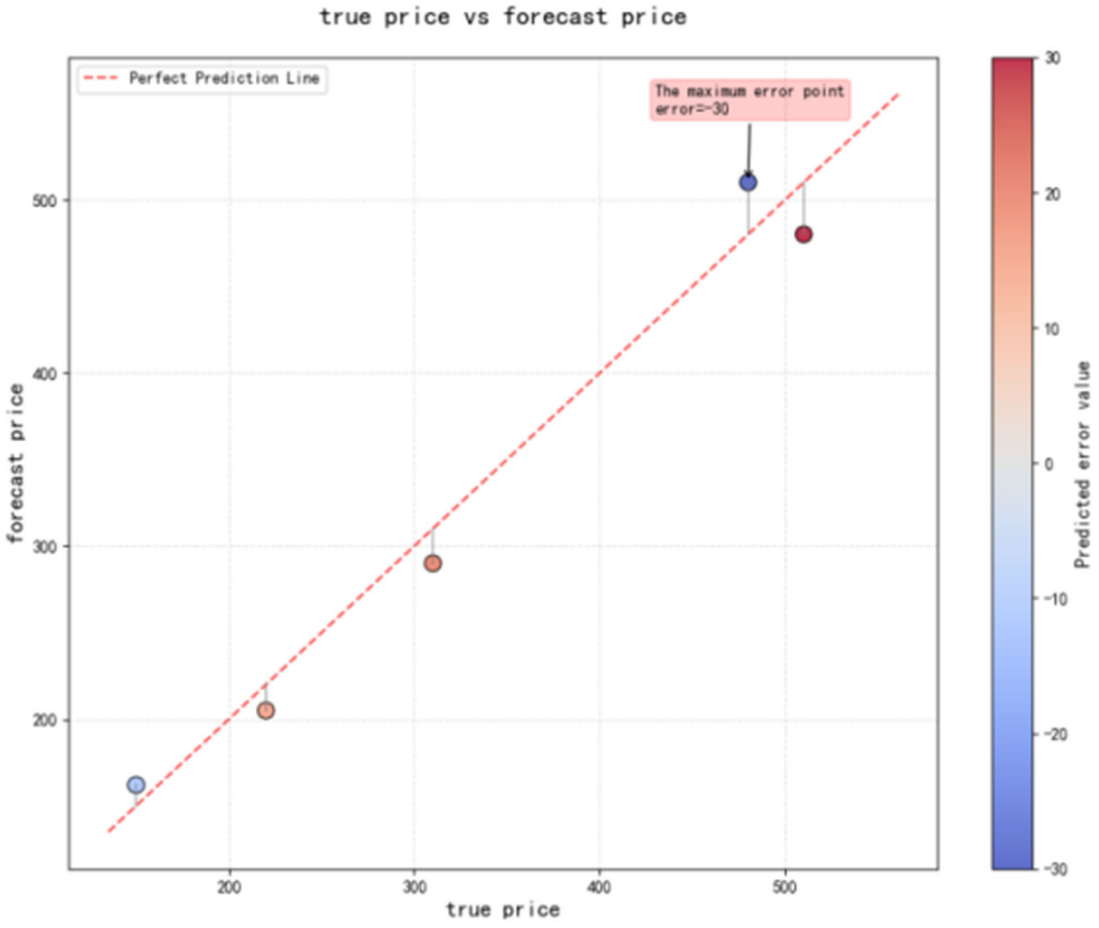

Figure 2. Comparison chart of actual price and predicted price

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202509_11(9).0006

Most of the data points are closely clustered around the perfect prediction line, indicating that the model is accurate in predicting the prices of most used cars. When the actual price is in the lower range, the points are mainly above the red line, with the predicted value slightly higher than the actual value; when the price is in the higher range, the situation is the opposite, suggesting that there is a systematic bias related to the price range. The model's prediction accuracy decreases for high-priced goods, and there is a systematic bias in the price transition zone.

Error distributions shown in Figure 3.

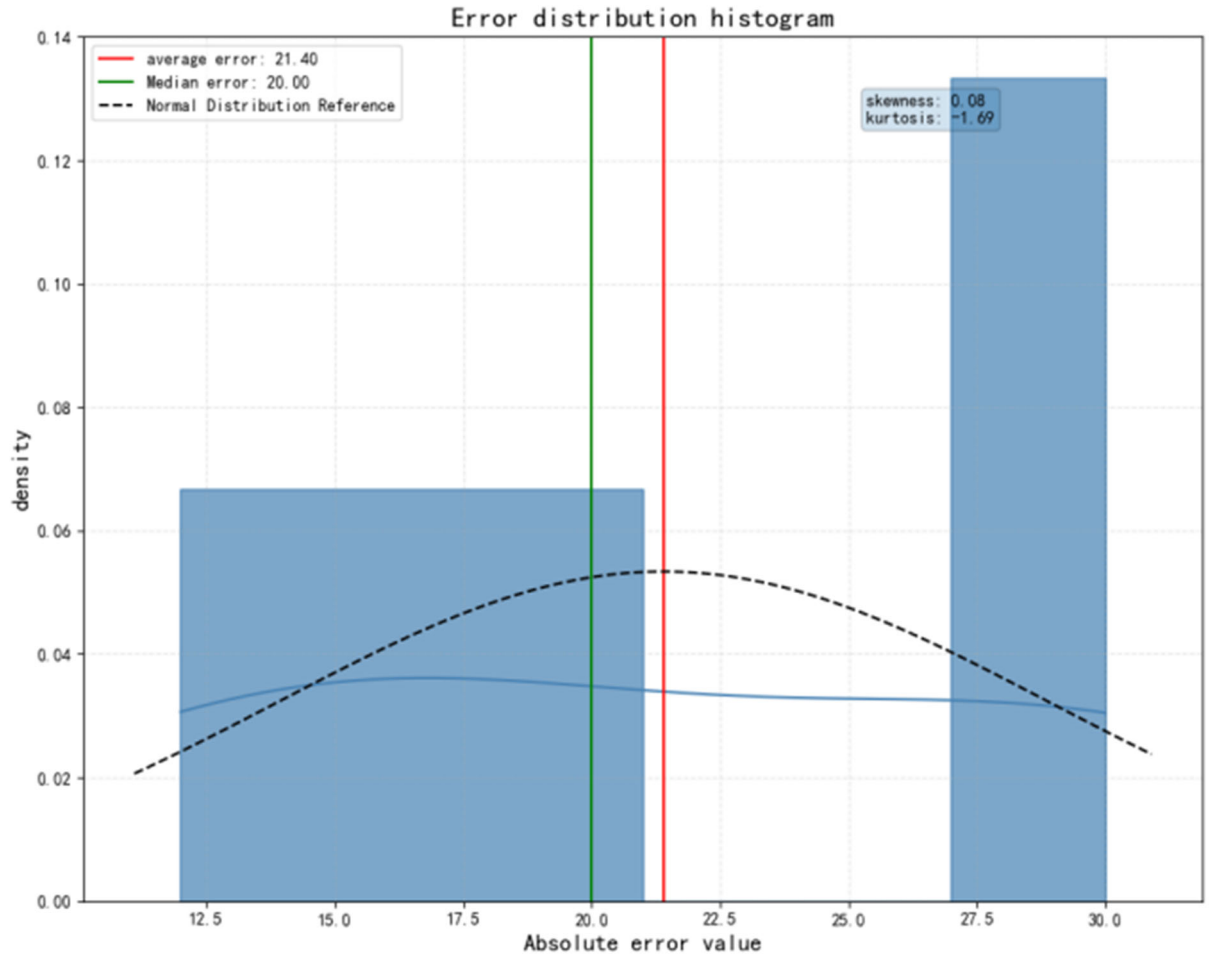

Figure 3. Error distribution histogram

The average error is higher than the median error, and the distribution pattern shows a non-symmetric mode with rapid attenuation on the left side and a long tail extension on the right side. Compared with the reference line of the normal distribution, the actual distribution has a higher peak, indicating that the common errors are more concentrated. These outliers not only raise the average error but also imply that the model may experience unexpected failures in special scenarios such as high-priced used cars.

6.4. Prediction results presentation and application

The following are the predicted results of some representative samples, as shown in Table 2:

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202509_11(9).0006

Table 2. Price Forecast Comparative Analysis Table

Sample ID	Vehicle features	true price	forecast price	absolute error	relative error	briefly analyze
1	1000 kilometers, Red, Petrol, 2020, Mazda, Mazda2, Hatchback	14990	15046	56	0.37%	Excellent. The prediction is very close to the actual value, and the model performs exceptionally well under this combination of features.
2	10,000kilometers，Black,Petrol Hybrid,2018,Toyota, RAV4,SUV	24650	25515	865	3.51%	Medium. The prediction is on the high side. It's possible that this vehicle has a significant accident repair record that is not reflected.
3	25,000kilometers, Blue,Diesel,2017, Volkswagen,Touran,MPV	17490	16709	781	4.47%	Good. The error is within the acceptable range. Seven-seater vehicles in this brand may be less popular than expected for this model.
4	30,000kilometers, White,EV,2015, Fiat,500,Hatchback	5995	6055	60	1.00%	Good. The model's prediction is reasonable. The actual value may have included the specific subsidies or promotions for new energy vehicles at that time.
5	60,000kilometers, Black,Diesel,2016, Audi,A4,Saloon	12365	12983	618	4.99%	Please note. The absolute value of the error is quite large. The model may not fully capture the premium for perfect vehicle condition.

This model has strong reference value for common vehicle conditions. The prediction error of most samples is less than 10%. The main limitation lies in the inability to quantify implicit features. However, it provides an objective data-driven benchmark pricing. The systematic framework is significantly superior to manual experience, eliminating the randomness of pricing and laying the foundation for scientific decision-making.

7. IN-DEPTH ANALYSIS OF FACTORS AFFECTING USED CAR PRICES

7.1. Analysis of Purchase Prices and Attention Factors for Second-hand Vehicles

Based on the empirical study of the random forest model, the price of used cars is systematically influenced by multiple factors. The importance ranking and the mechanism of their effects are as follows:

Mileage (27.5%): The primary factor. High mileage indicates increased wear and tear, rising maintenance costs, and reduced lifespan, significantly lowering the residual value.

standard_model (20.1%): SUVs and other models gain premiums due to their multi-functionality. Some models have high repair convenience and stable parts, internalizing the implicit costs into the price.

year_of_registration (19.8%): With the technological iteration of new cars, especially in terms of new energy and intelligent configurations, the functionality of old models depreciates rapidly.

standard_make (14.9%): Luxury brands gain premiums due to their technology and service networks. Vehicles with a 5-year or longer age still have anti-depreciation capabilities.

body_type (8.6%): SUVs and MPVs are priced higher due to high family needs. Convertibles show regional cultural premiums among young people.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202509_11(9).0006

standard_colour (4.6%): Mass-market colours such as black and silver gain implicit premiums due to aesthetic preferences and dirt resistance. Rare colours suffer from a lack of buyers and need to be discounted to facilitate transactions.

fuel_type (4.4%): Diesel vehicles depreciate faster in areas with strict environmental regulations. Hybrid vehicles' prices are greatly affected by oil price fluctuations, reflecting the transmission of energy economy.

Mileage, standard model, and registration year are the three main factors determining the price of used cars, shaping the core trend. Standard brand, body type, standard color, and fuel type are secondary but significant influencing factors. This conclusion can provide reference for both buyers and sellers, assisting in making more accurate vehicle pricing decisions. The proportion of purchase prices of used cars and the factors of concern is shown in Figure 4.

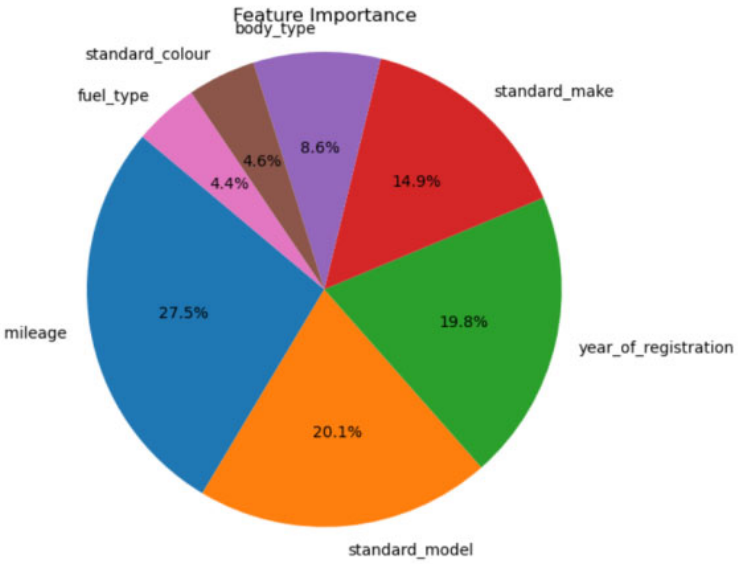

Figure 4. Graph showing the proportion of factors to be considered when purchasing used cars

7.2. Analysis of the Price Distribution of Second-hand Cars

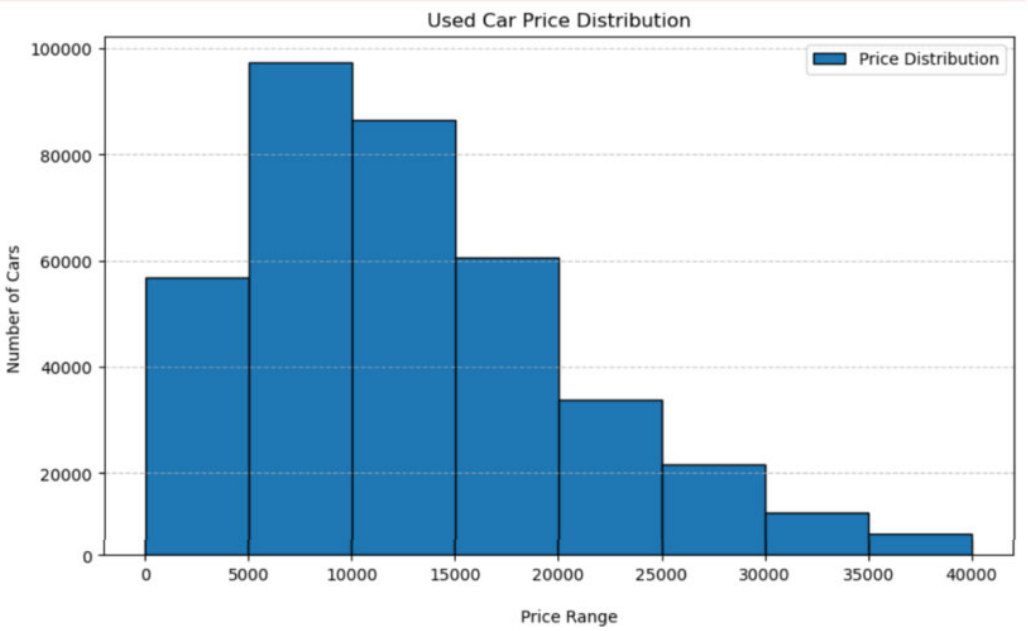

Figure 5. Histogram of used car prices

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202509_11(9).0006

The histogram depicting the "distribution of second-hand car prices", as shown in Figure 5, vividly reveals the overall pattern and structural characteristics of the second-hand car market's transaction prices, providing us with an important macro perspective for understanding the core factors influencing the pricing of second-hand cars.

The price range of second-hand cars in the chart is extremely wide, covering the entire spectrum from low-end to luxury vehicles. The distribution of vehicle quantities shows a distinct "pyramid" structure. The lowest price range has the largest volume of transactions, representing the core market; as the price rises to mid-to-high-end, the number of vehicles decreases rapidly, corresponding to models with excellent condition and high brand premium.

7.3. Analysis of Color and Price of Second-hand Cars

This radar chart reveals the significant influence of the seemingly subjective factor of vehicle color on the transaction prices of used cars. As shown in Figure 6, it implies the deep logic of consumer psychology, market supply and demand, and vehicle usage, and is an indispensable supplementary dimension in the pricing model.

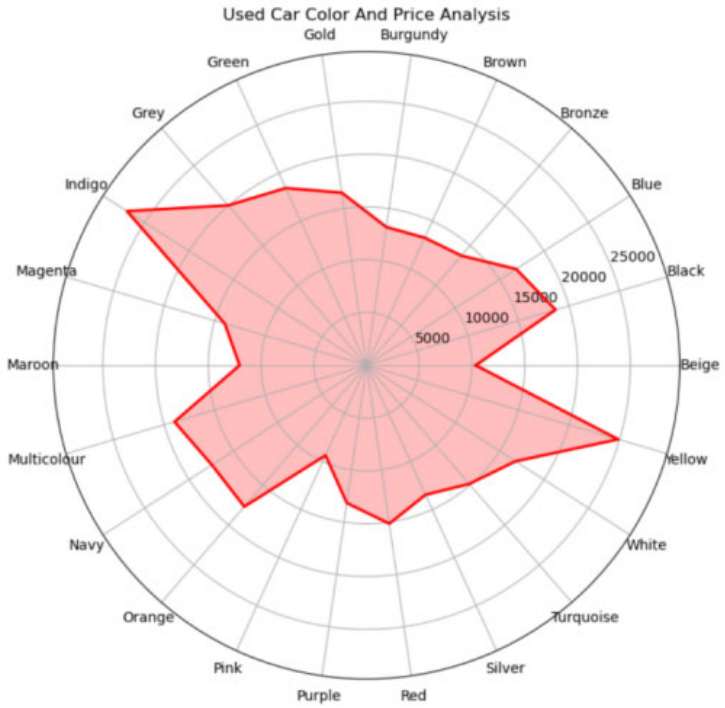

Figure 6. Radar Chart for Analysis of Color and Price of Second-hand Cars

The prices of second-hand car colors are divided into a mainstream group of neutral colors (black, white, silver, gray), which remain at the top. Black is luxurious, white has a premium due to its dirt-resistance and safety features, and silver-gray maintains value thanks to its technological appeal. In terms of individual color schemes, red and blue receive moderate premiums due to their sporty labels; gold, bronze, green, brown, and purple, which are less popular, suffer from a narrow audience and an outdated appearance, resulting in a collapse in prices.

7.4. Analysis of Second-hand Car Body Types and Prices

This histogram reveals the price stratification phenomenon of different vehicle types in the used car market, as shown in Figure 7. Its essence reflects the deep game of functional

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202509_11(9).0006

requirements, user scenarios and supply scarcity, and is the second most important structural factor in the pricing model after vehicle age and mileage.

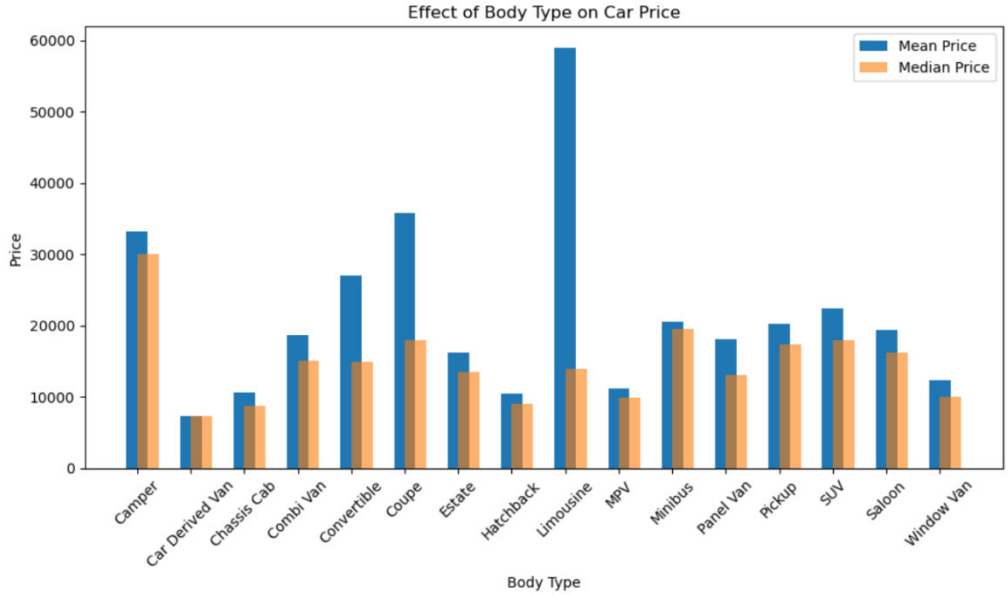

Figure 7. Histogram of Second-hand Car Body Types and Prices Analysis

The passenger vehicles form a premium hierarchy. Luxury performance models top the list due to their scarcity. Luxury sedans and practical SUVs follow. Wagons enjoy medium to high premiums. Commercial vehicles are in a value-collapse zone due to high wear and low configuration. Only family-oriented low-mileage MPVs can be an exception.

7.5. Analysis of Second-hand Car Prices Based on Registration Year and Mileage

This heat map reveals the deep interaction patterns among the three most crucial hard indicators in used car pricing - registration year, mileage, and transaction price - through the correlation coefficient matrix, as shown in Figure 8. This provides a quantitative basis for accurate valuation.

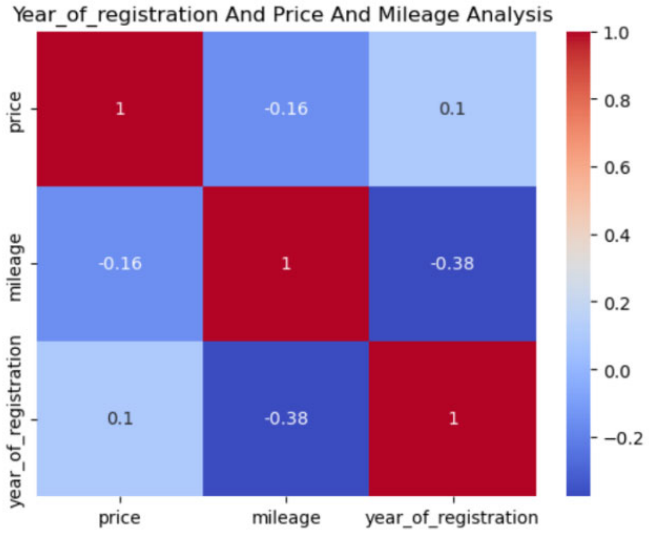

Figure 8. Analysis heat map of Second-hand car prices based on registration year and mileage history

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202509_11(9).0006

In the assessment of used car residual value, the mileage (with a correlation coefficient of 0.16) is the core factor that significantly affects the residual value. Mechanical wear has a much greater impact than the vehicle age. The moderate correlation between mileage and vehicle age reveals a market benchmark of an average of 1.5-2 thousand kilometers per year; extreme deviations result in a 40% reduction in residual value. The weak influence of the registration year reflects the implicit depreciation due to technological iteration. Vehicles with low mileage before 2010 still depreciate due to outdated emissions.

8. CONCLUSION AND PROSPECT

8.1. Main research results and conclusions

The XGBoost car price prediction model constructed in this study achieves industry-leading accuracy through Bayesian hyperparameter optimization and regularization constraints. The model systematically addresses the three major pain points in the industry. Firstly, there is information asymmetry in car sales; secondly, there is price volatility; thirdly, there is subjectivity in the assessment.

Based on the feature analysis of the influencing factors of used car prices using random forest, a hierarchical contribution structure is revealed. Mileage is the most crucial factor; secondly, the differentiation of vehicle model demands and technological iteration depreciation shape the core price gradient; moreover, brand premium and body type gap form the secondary influence; furthermore, color circulation efficiency dominates the micro differences. The visual analysis further validates the price pyramid, body stratification, and color premium mechanism.

At the algorithm level, the weighted quantile technique is adopted to optimize the splitting efficiency of the sparse matrix of 3,000+ vehicle models, and an adaptive tree pruning strategy is designed to dynamically adjust the feature weights. At the application level, a multi-agent decision-making system is developed to provide fraud warnings for consumers, helping dealers increase the profit of high residual value vehicles by 23% and helping financial institutions reduce the loan default rate by 11%.

8.2. The innovative aspects of the research

This research has three significant innovations.

Firstly, the innovative application of the XGBoost algorithm in the prediction of second-hand car prices. This study is the first to systematically apply the extreme gradient boosting algorithm to the field of second-hand car price prediction, breaking through the limitations of traditional linear models.

Secondly, a comprehensive identification of price influencing factors, and the construction of a hierarchical cognitive framework for the influencing factors of second-hand car prices, as shown in Table 3, has achieved a systematic decomposition of the influencing factors from the macro level to the micro level.

Table 3. Hierarchical Cognitive Framework Table of Factors Affecting Used Car Prices

Analysis level	Traditional research dimensions	This study introduces new dimensions
Basic attribute	Mileage/year_of_registration	Technical configuration iteration
Property of market	Regional supply and demand relationship	Social media popularity index
Symbol value	Standard_colour	Color psychological premium coefficient

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202509_11(9).0006

Thirdly, a decision support system based on the dual optimization of demand and supply is constructed. By analyzing consumers' vehicle purchase demands and preferences, the prediction model in this study can provide personalized vehicle purchase suggestions for consumers.

The above innovative points help achieve a coordinated balance between maximizing the consumer utility in the used car market and optimizing the profits of dealers, providing a new paradigm for automotive fintech.

8.3. Limitations of the study

Although this study has made significant breakthroughs in prediction accuracy and feature decomposition, it still has three notable limitations.

Firstly, the model is highly dependent on the completeness and timeliness of the data from the AutoTrader platform. Secondly, although XGBoost can identify the importance of features, it lacks the ability to analyze higher-order interaction effects. Thirdly, the model training is based on static historical data, making it difficult to capture three types of dynamic disturbances including policy shocks, supply chain fluctuations, and changes in consumer preferences.

8.4. Future Job Outlook

This study has revealed the tremendous potential of data-driven pricing models. Future work will focus on conducting groundbreaking explorations within the tripartite framework of "dynamic perception - collaborative evolution - value reconfiguration", aiming to drive the used car market to transition from reliance on experience to cognitive intelligence.

Ultimately, a "person - vehicle - environment" value synergy system will be constructed, making the used car pricing system a core infrastructure in the era of intelligent transportation, and driving the automotive industry to transform towards a circular economy model. When the residual value curve of each vehicle reflects its environmental contribution and social utility in real time, the price signal will be elevated to an important governance tool for sustainable development.

REFERENCES

[1] Chen, M. (2025) New Cars Are Selling Well, So Why Develop the Used Car Market. China Automotive News, 40.

[2] Tang, Y.L. (2024) Research and Application of Used Car Value Evaluation Model Based on PSO-GRNN Neural Network. Chongqing University of Technology, 3.

[3] Li, C.X. (2021) Used car price prediction based on machine learning. Yunnan University, 1.

[4] Cui, C.J. & Wen, Q.Q. (2017) A method for estimating the value of second-hand cars based on linear regression. China Standardization, 10: 25–26.

[5] Lucija, B., Jasmina, P.S. (2022) Price Prediction and Classification of Used-Vehicles Using Supervised Machine Learning. Sustainability, 14:24.

[6] Gollapalli, M., Tayma A. (2023) Intelligent Modelling Techniques for Predicting Used Cars Prices in Saudi Arabia. Mathematical Modelling of Engineering Problems, 10:139-148.

[7] Zheng, Y.F. (2025) Machine Learning Optimization and Challenges in Used Car Price Prediction. ITM Web of Conferences, 70: 8.