Proceedings of CONF-BPS 2026 Symposium: Sustainability Transitions and Regional Economic Restructuring in New Contexts
DOI: 10.54254/2754-1169/2026.NJ30803

LightGBM-Based Framework for Used Supercar Price Prediction

Zikai Niu

Department of Science, Beijing Information Science and Technology University, Beijing, China
nzk20040513@outlook.com

Abstract. As the used supercar market grows in complexity and scale, accurate price valuation is critical for transparent transactions and reduced information asymmetry. This study builds a used supercar price prediction framework through four key stages. First, in data preparation, the raw dataset (2000 samples, 32 features) is preprocessed. Non-predictive identifiers are removed, missing values are imputed via contextual logic, temporal features are converted to operational metrics, and derived features (e.g., power-to-weight ratio) are created to better capture pricing dynamics. Second, a hybrid encoding strategy is designed for diverse feature types. High-cardinality features use frequency encoding to retain category distribution and avoid dimensionality explosion; low-cardinality categorical features use one-hot encoding to prevent false ordinal relationships; ordinal features use label encoding to reflect order; binary flags are standardized via type conversion. Third, three ensemble models, Random Forest, Extreme Gradient Boost (XGBoost), and Light Gradient Boosting Machine (LightGBM), are built. Hyperparameters (ensemble size, maximum depth, sampling ratio) are optimized for supercar pricing to reduce overfitting. An 80-20 train-test split ensures objective performance evaluation. Finally, in model validation, regression metrics (RMSE, R^2) measure prediction accuracy, and feature importance analysis identifies key price-influencing factors. This forms a complete research chain from data processing to model application, providing a standardized process for supercar price prediction and supporting future pricing mechanism research.

Keywords: ensemble model, mileage, overfitting

1. Introduction

The used supercar market's intricate dynamics and high-stakes transactions necessitate precise valuation, but existing research falls short in customized feature engineering and model selection for this niche domain. A critical review of prior studies in used vehicle price prediction reveals that Zhou and Jia integrated Light Gradient Boosting Machine (LightGBM), neural networks, Extreme Gradient Boosting (XGBoost), and Bayesian optimization, concluding that a neural network-LightGBM hybrid excelled at capturing multi-factor influences on pricing [1-3]. Mao deployed a BP neural network, pinpointing 7–11 key price determinants and validating the model's efficacy for its specific dataset [4]. AIP Conference Proceedings compared Support Vector Machine (SVM) and K-Nearest Neighbors (KNN), demonstrating SVM's superior predictive performance [5].

DOI: 10.54254/2754-1169/2026.NJ30803

However, these efforts either adopted generic feature processing approaches or relied on limited algorithm suites [6]. To address these gaps, this paper focuses on three mainstream ensemble learning models—Random Forest, XGBoost, and LightGBM. By building on the validation of LightGBM and expanding the model scope, the paper aims to identify the most suitable algorithm for supercar pricing. Additionally, the paper introduces a hybrid encoding strategy by analyzing the methodology introduced in [7]: frequency encoding for high-cardinality features like brand and model, one-hot encoding for low-cardinality categorical features like color and transmission, label encoding for ordinal features like service history, and type conversion for binary flags like carbon fiber body. This strategy overcomes the "one-size-fits-all" limitations of previous feature handling methods.

Leveraging a dataset of 2000 samples with 32 features, this paper seeks to identify core pricing drivers encompassing brand, model, and technical attributes, and assess model performance using RMSE and R^2 metrics. The ultimate goal is to deliver a targeted valuation tool for the used supercar market, advancing both methodological and practical insights in this specialized field.

2. Methodology

2.1. Datasets

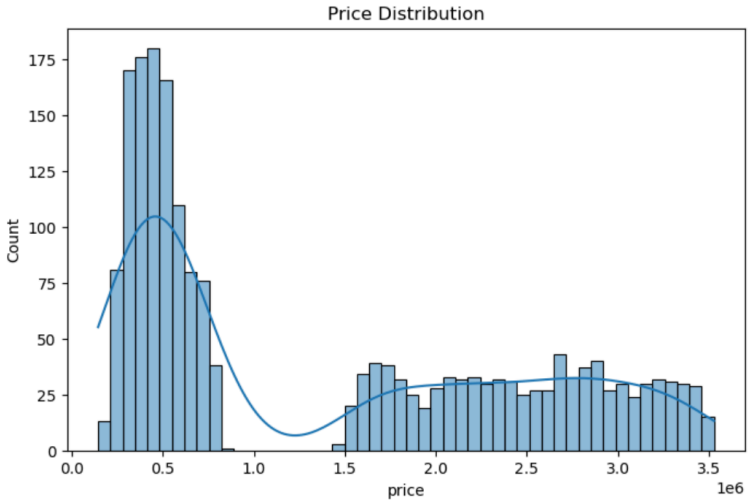

Figure 1. Price distribution of datasets (photo/picture credit: original)

The dataset used in this study is sourced from Kaggle's supercar-related public dataset, which contains 2000 samples and 32 features covering multiple dimensions of supercars (Figure 1) [8].

The preprocessing pipeline for the dataset starts with comprehensive cleaning. Non-predictive identifier information, such as ID, is removed to eliminate interference from irrelevant data. For missing values like those marked as "Nah", context-appropriate imputation methods are adopted to ensure data integrity. In terms of temporal feature engineering, the feature last service date is transformed into the feature days since service, which converts discrete date information into a continuous metric that more effectively reflects the time interval since the vehicle's last maintenance and better aligns with practical operational scenarios.

The core innovation in dataset processing lies in the design of a hybrid encoding strategy tailored to the diverse types of features in the dataset. For high-cardinality features, including brand and model, frequency encoding is used to retain the distribution characteristics of different categories

DOI: 10.54254/2754-1169/2026.NJ30803

while avoiding dimensionality explosion caused by excessive feature expansion. For eight low-cardinality categorical variables, such as color and transmission, one-hot encoding is applied to accurately represent the categorical attributes without introducing false ordinal relationships. Label encoding is used for ordinal features, such as service history, to fully reflect the inherent order of such features. Type conversion is performed for six flag features to standardize their data formats and ensure consistency with other features.

In advanced feature engineering, physically significant ratio features are constructed, including power-to-weight and weight-to-weight. These two features integrate the vehicle's power performance and weight attributes, making the feature set more in line with the professional evaluation logic of supercar performance. Additionally, standardization processing is carried out for thirteen numerical features to eliminate the impact of different value ranges on model training and ensure the stability of the subsequent modeling process.

This meticulous preprocessing approach fully adapts to the domain characteristics of supercars. It moves beyond generic data processing solutions and constructs a specialized feature representation system that can effectively capture the unique pricing dynamics of high-performance vehicles. The final processed feature set not only provides optimal input conditions for subsequent ensemble modeling but also maintains strong interpretability, which is conducive to subsequent analysis and application in practical business scenarios.

2.2. Symbols and parameters

Table 1 shows the feature name, full names, and explanations, covering all labels from the dataset.

Table 1. Description of symbols

Feature Name (label)	Full Name	Explanation
brand	Brand	The manufacturer brand of the vehicle (e.g., BMW, Toyota).
model	Vehicle Model	The specific model of the vehicle (e.g., BMW 3 Series, Toyota Camry).
horsepower	Engine Horsepower	A measure of the engine's power, reflecting the vehicle's dynamic performance.
power_to_weight	Power-to-Weight Ratio	The ratio of horsepower to vehicle weight indicates power efficiency.
top_speed_mph	Top Speed (Miles Per Hour)	The maximum speed the vehicle can reach (unit: mph).
zero_to_60_s	0-to-60 MPH Acceleration Time (Seconds)	The time the vehicle takes to accelerate from 0 to 60 mph reflects acceleration performance.
torque	Engine Torque	The rotational force output by the engine affects acceleration/carrying capacity.
days_since_service	Days Since Last Maintenance	The number of days since the vehicle's last service reflects maintenance status.
torque_to_weight	Torque-to-Weight Ratio	The ratio of torque to vehicle weight indicates torque efficiency.
mileage	Mileage	The total distance the vehicle has traveled, reflecting usage intensity.

DOI: 10.54254/2754-1169/2026.NJ30803

Table 1. (continued)

weight_kg	Vehicle Weight (Kilograms)	The curb weight of the vehicle (unit: kg).
damage_cost	Damage Repair Cost	The cost of past repairs for the vehicle reflects the condition/maintenance cost.
warranty_years	Warranty Duration (Years)	The remaining/total length of the vehicle's original warranty (in years).
year	Production/Registration Year	The year the vehicle was manufactured or registered.
num_owners	Number of Previous Owners	The total number of owners the vehicle has had, reflecting turnover frequency.
engine_config_V12	Engine Configuration (V12)	A dummy variable indicating whether the engine type is V12.
has_warranty_1	Has Active Warranty (Binary)	A dummy variable (1 = vehicle has an active warranty; 0 = no active warranty).
color_White	Exterior Color (White)	A dummy variable indicating whether the vehicle's exterior color is white.
drivetrain_RWD	Drivetrain (Rear-Wheel Drive)	A dummy variable indicating whether the vehicle uses rear-wheel drive.
interior_material_suede	Interior Material (Suede)	A dummy variable indicating whether the vehicle's interior uses suede material.

2.3. Random Forest

Random Forest is an ensemble learning algorithm that builds multiple independent decision trees via the bagging strategy, and outputs results by aggregating all individual tree predictions. For supercar price prediction (a regression task), its final result is the average of predictions from all trees.

It resists overfitting through two key steps: randomly selecting data subsets (with replacement) for training each tree, and randomly choosing feature subsets when splitting tree nodes. This ensures diversity across trees and avoids over-reliance on specific features like brand and mileage.

Well-suited for this study's task, it handles mixed numerical (e.g., mileage, weight) and categorical (e.g., brand, color) features without strict normalization needs, and supports feature importance analysis—making it a reliable baseline for comparing with XGBoost and LightGBM.

One line space before the third-level heading and ½ line space after the third-level heading. The model was carefully configured with anti-overfitting parameters. The ensemble size was set to 200 trees, with 100 trees used for cross-validation. For regularization, the maximum depth was set to 10, the minimum samples required for splitting a node were 20, and the minimum samples required at a leaf node were 10.

For feature sampling, the maximum number of features considered for splitting at each node was set to the square root of the total number of features to ensure diversity among individual trees. For data sampling, bagging was applied with 70% of the dataset used as subsets for training each tree.

2.4. XGBoost

The Extreme Gradient Boost is a gradient boosting ensemble learning algorithm based on decision trees, which trains a series of decision trees iteratively. Each new tree focuses on correcting the prediction errors of the previous tree ensemble, and the final prediction result is obtained by weighting the outputs of all trees. It is widely used in regression tasks such as supercar price

DOI: 10.54254/2754-1169/2026.NJ30803

prediction due to its strong ability to capture complex nonlinear relationships between features and targets.

XGBoost has two notable advantages: it supports custom loss functions and evaluation criteria to adapt to different task needs, and it has built-in mechanisms to automatically handle missing values without additional preprocessing. For the supercar price prediction task in this study, which involves mixed feature types including numerical features like mileage and weight, and categorical features like brand and model, XGBoost can effectively process these features and adjust model complexity through regularization parameters to reduce overfitting risks, making it a competitive algorithm alongside Random Forest and LightGBM.

First, target variable engineering optimization was completed by applying a logarithmic transformation to the price variable, which converted the original skewed distribution of price to an approximately normal distribution. The standard deviation of the original price data was approximately 1,000,000 or more, while the standard deviation after logarithmic transformation was approximately 0.5 to 1.0. The parameter setup of the model reflects an in-depth understanding of the algorithm and task characteristics. The maximum depth was set to 4, and the minimum child weight was set to 20—both conservative values to limit the model's complexity. The subsample ratio was set to 0.6 to enhance the model's generalization ability, and the regularization parameter reg lambda was set to 1.5 to prevent overfitting.

2.5. LightGBM

LightGBM is a gradient boosting framework that prioritizes both computational efficiency and predictive accuracy. Unlike XGBoost's level-wise tree growth, it adopts two key techniques—Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB)—to reduce computational costs while maintaining prediction performance, making it well-suited for tasks like supercar price prediction involving high-cardinality features such as brand and model.

The model natively processes high-cardinality features without manual encoding, eliminating the need to choose between one-hot, label, or frequency encoding for features like brand and model. For the supercar price prediction task, it effectively handles mixed feature types, including numerical features such as mileage, weight, and horsepower, as well as categorical features. Its efficient processing and strong ability to capture feature relationships, combined with built-in mechanisms to mitigate overfitting, make it a key algorithm for comparison with Random Forest and XGBoost in this study.

One line space before the third-level heading and ½ line space after the third-level heading. The model was carefully configured with anti-overfitting parameters. The ensemble size was set to 200 trees, with 100 trees used for cross-validation. For regularization, the maximum depth was set to 10, the minimum samples required for splitting a node was 20, the minimum samples required at a leaf node was 10, and the maximum number of features considered for splitting at each node was set to the square root of the total number of features to ensure diversity among individual trees. For data sampling, bagging was applied with 70% of the dataset used as subsets for training each tree.

3. Results

3.1. Findings: metric

By aggregating all the results, Table 2 shows that:

DOI: 10.54254/2754-1169/2026.NJ30803

Table 2. Results of all models

	RandomForest	XGBoost	LightGBM
Training RMSE	396064.11	325156.40	390590.40
Test RMSE	441156.95	392781.05	402588.84
Overfitting (%)	12.00	37.20	3.10

Random Forest has a training RMSE of 396,064.11, a test RMSE of 441,156.95, and a 12.00% overfitting rate. Its training fit is moderate, test error rises a bit but stays manageable, and overfitting is lower than XGBoost but higher than LightGBM. This comes from its bagging setup—multiple trees sample data randomly and combine predictions, so it doesn't overfit like a single tree and works steadily on new data.

XGBoost has the lowest training RMSE at 325,156.40, but a high test RMSE at 392,781.05 and a 37.20% overfitting rate. It fits training data best, catching complex price-feature links, but overlearns details, even noise in training data. That's because it builds trees iteratively to fix past errors—this boosts training fit but leads to poor generalization when data is limited, or regularization is weak.

LightGBM has a training RMSE of 390,590.40, a test RMSE of 402,588.84, and only a 3.10% overfitting rate, the lowest of the three. Its training and test errors are barely different: it matches Random Forest's training fit but has way better generalization. This is thanks to its GOSS and EFB tech—GOSS samples key data to keep useful info while cutting compute work, EFB bundles features to avoid redundancy. Both stop overfitting and keep efficiency high.

3.2. Findings: feature

These three bar charts present feature importance across Random Forest, XGBoost, and LightGBM models. The y-axis denotes input features, and the x-axis indicates each feature's contribution to model predictions.

3.2.1. Random Forest

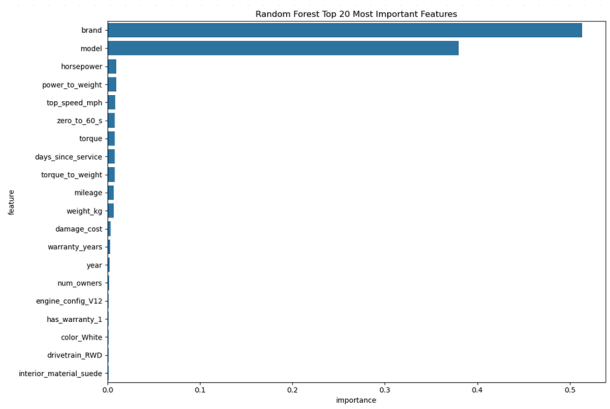

Figure 2. Top 10 Random Forest features (photo/picture credit: original)

From Figure 2, it can be observed that brand premium is the decisive factor for supercar pricing. The fact that brand alone contributes 53.6% of the predictive power proves that in the supercar

DOI: 10.54254/2754-1169/2026.NJ30803

domain, the brand itself serves as the most significant value endorsement. Even with the same technical parameters, price differences across different brands can reach several times.

3.2.2. XGboost

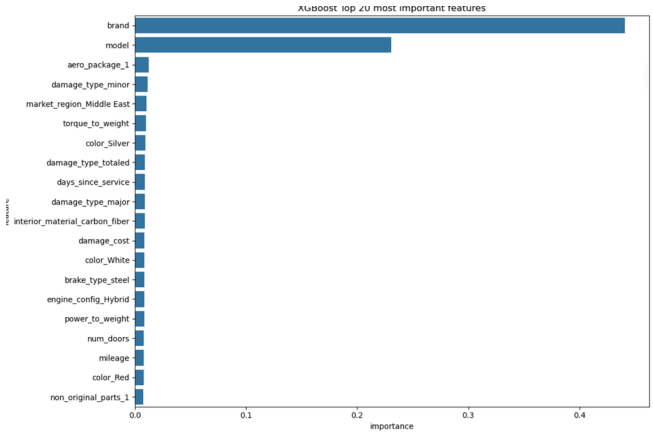

Figure 3. Top 10 XGBoost features (photo/picture credit: original)

From Figure 3, it can be observed that brand and model still play an important role among all features, which is consistent with Random Forest's performance. Compared to Random Forest, XGBoost not only recognizes the core role of brand and model as major feature categories but also identifies the significant impact of detailed sub-features—such as specific configurations like carbon fiber interior and factory warranty—resulting in a more granular feature importance assessment.

3.2.3. LightGBM

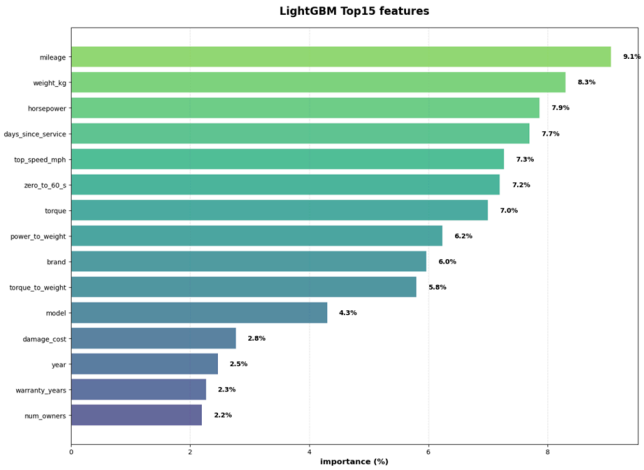

Figure 4. Top 10 LightGBM features (photo/picture credit: original)

From Figure 4, it can be observed that LightGBM's feature importance results are completely different from those of XGBoost and Random Forest. The top 7 features in terms of importance are all technical features, including mileage, weight, and horsepower. Surprisingly, brand and model do not rank among the features with the highest importance.

DOI: 10.54254/2754-1169/2026.NJ30803

4. Discussion

This study's findings reveal clear differences in how three ensemble models perform in the used supercar price prediction task tied to their unique mechanisms and adaptability to the supercar domain's traits.

For predictive performance and overfitting, XGBoost has the lowest training RMSE, 325156.40, showing a strong ability to fit complex feature-price links via its iterative gradient boosting. But its 37.20 percent overfitting rate is a flaw — with the 2000-sample 32-feature dataset, it overlearns training details, including noise, as supercar pricing's rare configurations and brand nuances aren't fully covered in limited samples. LightGBM strikes the best balance; its training RMSE of 390590.40 is close to Random Forest, while its 3.10 percent overfitting rate is far lower. This comes from GOSS, which keeps key samples for pricing info and cuts compute work, and EFB, which bundles redundant features, making it robust to the dataset's small-sample multi-feature issue. Random Forest as baseline has moderate 12.00 percent overfitting and stable test RMSE 441156.95 — its bagging-based sampling works, but it lacks XGBoost's fit precision and LightGBM's generalization efficiency.

In feature importance models, priorities reflect their algorithm logic. Random Forest names brand as the top factor contributing 53.6 percent of predictive power, matching the supercar market's brand-premium focus — brands often drive prices more than technical specs like similar horsepower or weight, but big price gaps across brands. XGBoost confirms brand and model importance, but also picks up detailed configurations like carbon fiber interior and factory warranty. Its ability to model hierarchical feature links fits supercars where high-value options shift resale prices. LightGBM's top seven features are all technical, including mileage, weight, and horsepower, with brand and model absent. This is likely because LightGBM natively handles high-cardinality features, no manual encoding for brand or model, reducing focus on categoricals and shifting to technical metrics with direct wear and performance correlations.

The hybrid encoding strategy fixes prior research gaps. Unlike generic methods such as one-hot encoding for all categoricals, which causes dimensionality explosion for brand and model, it tailors the encoding. Frequency encoding is used for brand and model to keep category distribution one-hot encoding for color and transmission, to avoid false ordinality label encoding for service history to reflect order. This boosts interpretability, like tracking brand impact in Random Forest, and stabilizes model performance—without it, high-cardinality features might skew feature importance or increase overfitting, like XGBoost's overfitting rate could have been higher without standardized handling of binary flags.

5. Conclusion

This study completes a systematic exploration of supercar price prediction from data preprocessing to model validation and delivers three core conclusions.

First, model performance hierarchy: LightGBM is the most suitable algorithm for supercar price prediction, balancing predictive accuracy with test RMSE 402588.84 and generalization with 3.10 percent overfitting rate. XGBoost excels in training fit but suffers from severe overfitting, while Random Forest provides reliable baseline performance but lacks LightGBM's precision in controlling overfitting. This hierarchy offers a clear reference for algorithm selection in supercar valuation tasks.

Second, key pricing factors vary by algorithm logic: Brand premium dominates Random Forest's predictions, reflecting the supercar market's brand-driven value logic. XGBoost captures both core

DOI: 10.54254/2754-1169/2026.NJ30803

categories, brand and model, and detailed configurations catering to the impact of high-value options. LightGBM prioritizes technical features, highlighting the relevance of performance and wear metrics. Together, these findings provide a comprehensive view of supercar pricing drivers instead of relying on a single set of universally important features.

Third, hybrid encoding enhances model reliability: The tailored encoding strategy using frequency encoding for high-cardinality features, one-hot or label encoding for others, resolves dimensionality and interpretability issues in prior research, ensuring the 32-feature 2000-sample dataset is effectively utilized across models. This preprocessing framework can be adapted to other high-value niche vehicle markets like luxury classic cars with similar feature complexity.

Practically, this study offers a standardized valuation tool for the used supercar market, helping reduce information asymmetry between buyers and sellers. Methodologically, it demonstrates the value of customizing preprocessing and algorithm selection for niche domains, moving beyond generic used vehicle prediction frameworks. Future research could expand the dataset to include regional pricing differences or long-term resale trends and explore hybrid models such as stacking LightGBM with other algorithms to further improve prediction precision.

References

[1] Zhou, Y., & He, B. (n.d.). Prediction method of used car price based on a weighted combination model. Wuhan Research Institute of Posts and Telecommunications; Wuhan FiberHome Digital Technology Co., Ltd.
[2] Jia, P. (2021). Used car price prediction based on LightGBM (Master's thesis). Shandong Normal University.
[3] Yang, J. (2024). Research on used car price prediction model based on XGBoost and stacking (Master's thesis). Harbin Institute of Technology.
[4] Mao, P., Cai, Y., Wan, X., & Wang, W. (n.d.). Research on influencing factors of used car price evaluation based on BP neural network. School of Automotive & Transportation Engineering, Xihua University.
[5] AIP Conference Proceedings. (2023). A novel mechanism for computation and prediction of car price using support vector machine algorithm and predicted accuracy comparison with KNN algorithm (Vol. 2821, Article 020045).
[6] Rizka, N., Alim, M., Laina, F., et al. (2024). Comparative analysis of XGBoost and random forest for used car price prediction. In 2024 International Conference on Electrical Engineering and Informatics (pp. 125–129).
[7] Li, Z., Ji, L., & Yang, F. (2025). An air target intent recognition method combining dynamic Bayesian network and one-hot encoding-based DTW. Journal of Detection & Control, 47(2), 82–90.
[8] Kaggle. (2025, November 20). Predict supercars prices 2025. https: //www.kaggle.com/competitions/predict-supercars-prices-2025/data