

A Heterogeneous Ensemble Learning Framework for Used Car Price Prediction

Zhidong Zhao^{1,a,#}, Xindong Bai^{1,b,#}, Bo'en Zhao^{1,c,#}, Haoran Su^{1,d,#}, Sirui Zhao^{1,e,#}

¹Beijing 21st Century School, Beijing, China

^azzdfox@outlook.com, ^b2822411860@qq.com, ^cboen1234562022@163.com, ^d1713273550@qq.com, ^ezsr319@outlook.com

[#] Co-first author

Abstract: Accurate used car valuation remains challenging, as transaction prices are jointly determined by multiple interacting factors including vehicle attributes, condition and market conditions, while raw listing data suffer from noise, incompleteness and severe right-skewed distribution. This paper proposes an end-to-end prediction pipeline for used car price estimation, which integrates robust data preprocessing, systematic feature engineering and a heterogeneous ensemble composed of four base models: LightGBM, XGBoost, CatBoost and a multilayer perceptron. Gradient-boosted tree models are trained on log-transformed targets with L1-type loss functions theoretically consistent with the mean absolute error (MAE) metric. Predictions from all base learners are aggregated via a non-negative weighted average, where the weights are optimized on out-of-fold predictions using the Nelder-Mead algorithm. Experimental results demonstrate that the ensemble model achieves superior predictive performance compared with individual models and traditional linear regression baselines. Error analysis indicates that residual errors are mainly concentrated in the sparse high-price vehicle segment, and the model implicitly captures duplicate vehicle records through fine-grained partitioning of embedding features. The findings verify that metric-aligned objective function design and the complementarity of heterogeneous learners, rather than the performance of any single model, are the primary drivers of the final predictive performance.

Keywords: Used Car Price Prediction, Ensemble Learning, Gradient Boosting Decision Tree, Neural Network, Feature Engineering

1. Introduction

China's used car market has entered a stage of rapid expansion: in the first half of 2026, transaction volume reached 9.71 million units and, for the first time in a half-year period, surpassed new car sales. While this growth releases substantial consumption potential, it also places much higher demands on the accuracy, efficiency, and scientific rigor of used car price appraisal.

Pricing a used vehicle is intrinsically hard because its value is determined by highly multidimensional and nonlinear mechanisms. Unlike new cars, which carry clear manufacturer suggested retail prices, the final transaction price of a used car is governed by a host of heterogeneous factors: vehicle-specific parameters (brand, model category, age, odometer reading, engine power, number of prior transfers), condition-related attributes (maintenance records, accident history, exterior condition), and external variables such as regional supply-demand fluctuations, new car pricing adjustments, and seasonal cycles. These factors interact in ways that simple linear relationships cannot capture. The problem is compounded by information asymmetry between buyers and sellers: sellers possess far more vehicle-history information than buyers, which can compromise buyer interests and impede transactions. Traditional appraisal methods, chiefly the replacement cost approach, the current market price approach, and the income capitalization approach, rely heavily on the appraiser's subjective judgment and therefore struggle to deliver objective and consistent outcomes. Empirical evidence shows that machine learning approaches can substantially reduce prediction error relative to traditional regression modeling [1], which motivates a data-driven treatment of the valuation problem.

A review of the existing literature and of practical applications reveals three challenges that still confront used car valuation modeling. First, insufficient model diversity: most studies focus on a single model or on narrow cross-model comparisons, and lack a systematic exploration of integrating algorithms of different types. Second, inadequate feature extraction: raw records contain substantial

latent information, and there remains considerable room for systematic feature construction and selection pipelines that turn it into high-value signals. Third, uneven data quality: raw data are commonly plagued by noise, outliers, and missing values, which severely undermine prediction stability when fed directly into models.

To address these challenges, this study develops a valuation framework that spans the entire chain of data preprocessing, feature engineering, model ensembling, and prediction output. Specifically: (1) robust preprocessing strategies, including missing-value handling, outlier clipping, and a logarithmic target transformation, ensure the quality and stability of the model inputs; (2) a discriminative feature set is constructed from multiple perspectives, covering temporal, regional, statistical, and interaction features, to excavate latent value signals; (3) four heterogeneous algorithms, LightGBM, XGBoost, CatBoost, and a multi-layer perceptron (MLP), are combined through a weighted fusion whose weights are learned on out-of-fold predictions, which mitigates individual model bias and variance; and (4) the experimental analysis, conducted under a strict holdout protocol, quantifies where the remaining error resides, furnishing interpretable evidence for valuation practice.

The contributions of this article are as follows:

- **Robust Preprocessing & Metric Alignment:** A complete preprocessing pipeline covering outlier clipping, missing value handling and logarithmic target transformation is constructed, which theoretically aligns the training objective with the MAE evaluation metric and guarantees the quality and stability of model inputs.
- **Heterogeneous Ensemble Fusion Framework:** Three gradient-boosted tree models and a multi-layer perceptron are fused via non-negative weighted averaging optimized by the Nelder-Mead algorithm on out-of-fold predictions, leveraging error complementarity across heterogeneous learners to reduce single-model bias and boost overall generalization performance.
- **Interpretable Pricing Analysis:** Core drivers of used car pricing are extracted through feature importance ranking, and the distribution patterns and underlying causes of residual errors are systematically uncovered, providing interpretable decision support for practical valuation practices.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 details the dataset, the feature engineering pipeline, the base learners, and the fusion strategy. Section 4 reports the experimental results, compares them against baselines, and conducts the error analysis. Section 5 concludes.

2. Related Work

Used car price prediction has been studied extensively with statistical, machine learning, and deep learning methods. On the side of classical models, Lessmann and Voss [1] conducted an empirical study of resale price forecasting and showed that random forest regression outperforms traditional linear regression, and that the choice of regression method has a material impact on forecast accuracy. Bukvic et al. [2] adopted a supervised learning framework in which linear regression is used to predict used vehicle prices and the regression task is contrasted with a vehicle classification task to analyze market price trends. Amik et al. [3] compared linear regression, LASSO, decision trees, random forests, and XGBoost on a used car dataset, screening the best model with MAE, MSE, and RMSE error metrics. Collectively, these studies confirm that classical regression models are simple to implement but exhibit a clear ceiling when confronted with the multidimensional, heterogeneous features of used cars.

Ensemble learning algorithms represented by random forests, XGBoost, and LightGBM have since become the mainstream in this field. Cui et al. [4] proposed an iterative framework that couples XGBoost and LightGBM to improve prediction accuracy under large feature sets and large data volumes. Huang et al. [5] further fused XGBoost, CatBoost, LightGBM, and an artificial neural network into a stacked ensemble, reaching $R^2 = 0.9845$ on 30,000 Chinese used-car transactions. Fayyaz et al. [6] provided a comparative analysis of machine learning models for used car price prediction with a strong emphasis on feature engineering, showing that carefully constructed features contribute more to accuracy than the choice among modern learners. Deep learning models have also been explored on large-scale transaction data — Dutulescu et al. [7] trained multiple neural architectures on over 300,000 German listings and 15,000 Romanian quotes, achieving $R^2 > 0.95$, yet they are hampered in practice by complex hyperparameter tuning and high training cost.

In summary, machine learning and ensemble learning are widely applied to used car price prediction,

yet existing models still leave room for improvement in the treatment of feature noise and in multi-model fusion strategies. Most prior work either compares individual learners or stacks homogeneous tree models [4], [5]; systematic fusion of heterogeneous learners whose objectives are explicitly aligned with the evaluation metric remains under-explored, and it is the focus of this paper.

3. Methodology

The overall framework of the proposed approach is illustrated in Figure 1. Raw records are cleaned and transformed, two feature sets are engineered for tree models and for the MLP set respectively, four heterogeneous base learners are trained under K-fold cross-validation, and their out-of-fold (OOF) predictions are fused by a learned weighted average that produces the final test-set prediction.

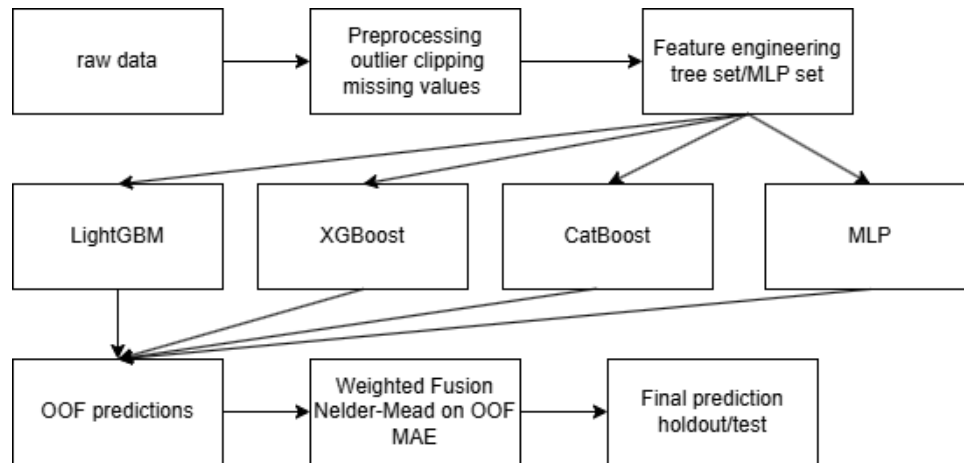


Figure 1: Overall structure of the proposed ensemble framework.

3.1. Dataset

The dataset consists of 200,000 de-identified used car transaction records released by the Alibaba Tianchi data mining competition, of which 150,000 constitute the training set (with ground-truth prices) and 50,000 constitute the test set. Each record contains 31 attributes, including basic vehicle information (brand, model, body type, fuel type, engine power, mileage, and repair status), temporal information (registration and listing dates), and 15 anonymized features v0 to v14, which are embedding vectors derived from a large amount of textual and tagging information about each vehicle.

3.2. Preprocessing and Target Transformation

Raw records are first cleaned: implausible engine power values are clipped to the official range [0, 600], the placeholder symbol in the repair-status field is converted to a missing value and left to the tree models' native missing-value handling, and invalid dates are repaired conservatively. The price distribution is heavily right-skewed: the median is 3,250 CNY, the mean is 5,923 CNY, and the maximum is 99,999 CNY. Training directly on the raw price would let a small fraction of luxury vehicles dominate the loss, so the target is compressed by a logarithmic transformation

$$y' = \ln(1 + y), \quad (1)$$

and the original scale is recovered at inference time by the inverse mapping

$$y = \exp(y') - 1. \quad (2)$$

This choice is not merely a variance-stabilization trick; it is aligned with the evaluation metric in theory. The Bayes-optimal prediction under the MAE metric is the conditional median, and since the logarithm is a monotone transformation it preserves the median. Consequently, the modeling route of a log-transformed target, an L1-type loss, and an exponential back-transformation is consistent with the MAE objective, a property that squared-error training does not share on skewed data.

3.3. Feature Engineering

Based on the raw attributes, a comprehensive set of engineered features is constructed, as summarized in Table 1. Temporal features capture vehicle age, one of the most predictive variables, at both day and year granularity. Frequency encodings and grouped statistics of the target transform categorical variables into informative numeric signals, and pairwise interactions among the anonymized embedding dimensions capture higher-order structure. After deduplication of redundant columns, 195 informative dimensions are retained. Two parallel input sets are then produced: the tree models consume the full feature set directly, whereas the MLP consumes a separately normalized version, since neural networks are sensitive to input scale.

Table 1: Summary of the feature engineering pipeline.

Feature category	Description
Temporal features	Year/month/day of registration and listing; vehicle age in days and years
Regional features	Count of vehicles per region; city derived from the region code prefix
Statistical encodings	Frequency encoding and grouped statistics (mean, median, std) of price by brand/model/region/mileage/body/fuel
Binning features	Equal-width binning of power and model
Interaction features	Pairwise products of anonymized features v0-v14 and interactions with power and age

3.4. Gradient-Boosted Tree Learners

LightGBM, XGBoost, and CatBoost are all implementations of gradient boosting decision trees (GBDT). GBDT builds an additive ensemble of regression trees in which each new tree is fitted to the pseudo-residuals of the current ensemble. Let $F_{m-1}(x)$ denote the ensemble after $m - 1$ iterations; the residual of a sample (x_i, y_i) is

$$r_i = y_i - F_{m-1}(x_i) \quad (3)$$

and the ensemble is updated with a learning rate η that shrinks the contribution of each new tree h_m :

$$F_m(x) = F_{m-1}(x) + \eta \cdot h_m(x) . \quad (4)$$

The procedure is gradient descent in function space: the residuals correspond to the negative gradient of the loss. Because the evaluation metric is the MAE, all three tree learners adopt L_1 -type losses built around the absolute error

$$L_1(y, \hat{y}) = |y - \hat{y}|, \quad (5)$$

whose negative gradient is the sign of the residual and whose optimal leaf value is a median rather than a mean. Concretely, LightGBM is trained with the Huber loss ($\alpha = 0.6$), a smooth robust variant of the absolute error; XGBoost with the adaptive absolute-error objective introduced in XGBoost 2.0, which uses median leaf values and therefore matches the MAE metric exactly; and CatBoost with its native MAE loss. LightGBM accelerates GBDT with leaf-wise growth, histogram-based split finding, and gradient-based one-side sampling; the gain it maximizes at each split is

$$Gain = \frac{1}{2} \left[\frac{GL^2}{(HL+\lambda)} + \frac{GR^2}{(HR+\lambda)} - \frac{(GL+GR)^2}{(HL+HR+\lambda)} \right], \quad (6)$$

where G and H are the sums of first- and second-order gradients in the left and right children and λ is the L_2 regularization coefficient. XGBoost formulates the objective at iteration t by a second-order Taylor expansion with an explicit complexity penalty:

$$L^{(t)} \approx \sum_i [g_i f_t(x_i) + \frac{1}{2} h_i f_t(x_i)^2] + \gamma T + \frac{1}{2} \lambda \sum_j w_j^2, \quad (7)$$

where T is the number of leaves and w_j the leaf weights. CatBoost additionally addresses the target-leakage problem of conventional GBDT through ordered boosting, in which the residual of each sample is estimated using only the samples preceding it in a random permutation, which confers strong robustness against noise.

3.5. Multi-Layer Perceptron

The neural member of the ensemble is an MLP with three hidden layers of sizes 256, 128, and 64, as illustrated in Figure 2. The forward computation of layer l is

$$a^{(l)} = \text{ReLU}(W^{(l)}a^{(l-1)} + b^{(l)}), \quad (8)$$

where $W^{(l)}$ and $b^{(l)}$ are the layer's weight matrix and bias. The network consumes the separately normalized feature set and regresses the log-transformed target with a squared-error criterion, trained by the Adam optimizer (initial learning rate 0.001, batch size 512, at most 120 epochs) with an $L2$ weight penalty of 0.0001 and early stopping on an internal 10% validation split (patience of 12 epochs). Although a squared-error criterion targets the conditional mean rather than the median, the log transformation renders the target distribution approximately symmetric, so the mean-median gap is small; the role of the MLP in the ensemble is to contribute a different inductive bias rather than the lowest individual error.

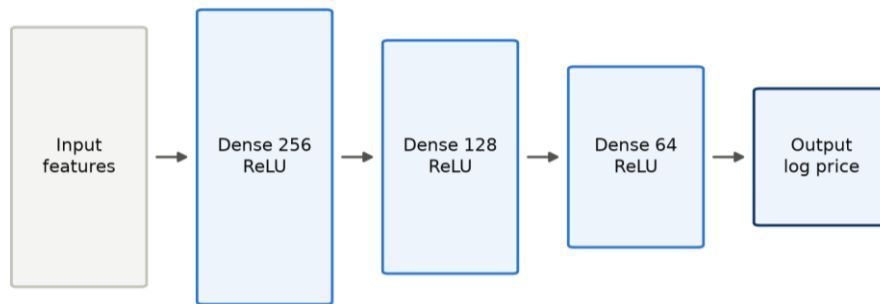


Figure 2: Architecture of the multi-layer perceptron.

3.6. Ensemble Fusion

Because the four base learners rely on different inductive biases, they commit different errors, and their predictions are combined by a constrained weighted average

$$\hat{y} = \sum_k w_k \hat{y}_k, \sum_k w_k = 1, w_k \geq 0, \quad (9)$$

where w_k denotes the non-negative fusion weight assigned to the k -th based learner and $\hat{y}_k$ is the prediction of the k -th based learner.

All member predictions are first mapped back to the original price space. The weights are then learned on the OOF predictions by directly minimizing the price-space MAE with the Nelder-Mead simplex method, and the resulting weights are applied unchanged to the test-set predictions of the members. Fitting the weights on OOF predictions rather than on training-set fits is essential: it prevents the fusion from rewarding members that overfit their own training folds.

4. Experiments and Analysis

4.1. Experimental Setup

The evaluation metric is the mean absolute error computed in the original price space:

$$MAE = \left(\frac{1}{n}\right) \sum_i |y_i - \hat{y}_i|, \quad (10)$$

where y_i and $\hat{y}_i$ represents the ground truth and prediction of i -th sample.

Compared with MSE and RMSE, the MAE is robust to extreme values, which is desirable for the long-tailed price distribution of this task; as noted in Section 3.2, its optimal predictor is the conditional median. Model validation follows a two-level protocol with fixed random seeds. A random 10% of the labeled records (15,000 rows) is set aside as a holdout set that never participates in training, hyperparameter selection, or fusion-weight fitting, and is used only to report the final test error. The remaining 135,000 records are used for 5-fold cross-validation, whose OOF predictions serve both for performance estimation and for fitting the fusion weights; the holdout and test-set predictions of each member are averaged over its fold models. All tree learners use early stopping (patience of 500 rounds

for LightGBM and 400 rounds for XGBoost and CatBoost), and LightGBM additionally monitors a custom evaluation function that reports the MAE after back-transformation to the price space, so that stopping is aligned with the target metric. Hyperparameter configuration of the models is shown in Table 2.

Table2: Hyperparameter configuration of the models

Model	Hyperparameter configuration
LightGBM	Huber loss $\alpha = 0.6$, leaves = 63, learning rate = 0.05, feature subsampling = 0.65, bagging fraction = 0.8, L2 regularization=2
XGBoost	maximum depth = 9, learning rate = 0.05, row subsampling = 0.8, column subsampling = 0.65, minimum child weight = 5, L2 regularization = 2
CatBoost	depth = 8, learning rate=0.08, Bernoulli subsampling = 0.8, L2 regularization = 3
MLP	initial learning rate= 0.001, batch size=512, L2 penalty=0.0001

4.2. Descriptive Statistics of the Target

Figure 3 visualizes the price distribution of the 150,000 training records. The raw price is heavily right-skewed with a skewness of 3.35: half of the vehicles cost no more than 3,250 CNY, while the top decile starts at 14,500 CNY and stretches to 99,999 CNY. After the log transformation the distribution becomes approximately unimodal and symmetric, which stabilizes gradient-based training. The scale of the task is set by two naive references evaluated under the same protocol: always predicting the median price of the development set (3,250 CNY) yields a holdout MAE of 4,371.20 CNY, and a ridge-regularized linear regression on the numeric attributes, trained on the log target, reaches an OOF MAE of 939.22 CNY and a holdout MAE of 916.15 CNY.

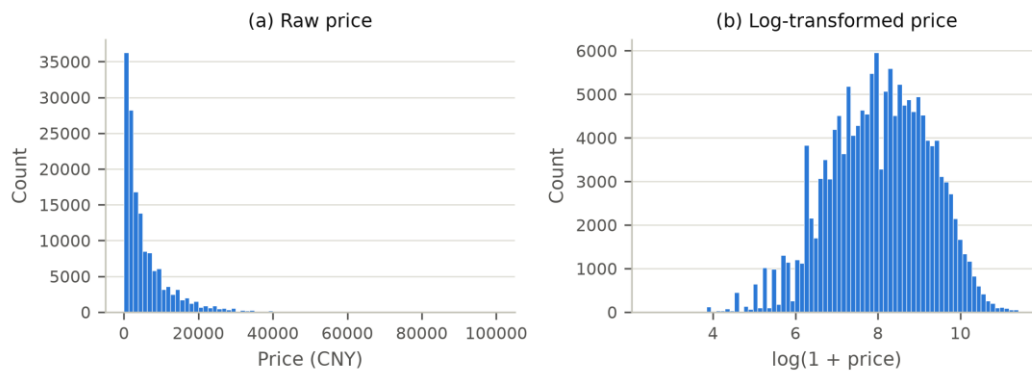


Figure 3: Distribution of the transaction price before (a) and after (b) the logarithmic transformation.

4.3. Main Results

Table 3 reports the OOF and holdout MAE of the two baselines, the four base learners, and the ensemble; Figure 4 visualizes the holdout comparison. Every learner improves dramatically over the linear baseline, confirming the strongly nonlinear nature of the pricing function. Among the single models LightGBM (holdout MAE 429.31 CNY) and XGBoost (438.59 CNY) are the best, CatBoost trails slightly (445.51 CNY), and the MLP is the weakest (494.95 CNY). The error correlations between the MLP and the tree models (0.74-0.77 on the OOF predictions) are, however, markedly lower than the correlations among the tree models themselves (0.86-0.93), and this heterogeneity makes the MLP valuable in the fusion despite its accuracy gap.

Table 3: OOF and holdout MAE of baselines, single models, and the ensemble (original price space).

Model	OOF MAE (CNY)	Holdout MAE (CNY)	Holdout reduction vs. linear
Constant (median price)	4,488.05	4,371.20	-
Linear regression (ridge, log target)	939.22	916.15	0.0%
MLP	603.45	494.95	46.0%
CatBoost (MAE loss)	497.12	445.51	51.4%
XGBoost (absolute-error objective)	472.43	438.59	52.1%
LightGBM (Huber loss)	470.52	429.31	53.1%
Ensemble (weighted fusion)	449.47	421.03	54.0%

The Nelder-Mead search assigns fusion weights of 0.41 to LightGBM, 0.30 to XGBoost, 0.23 to CatBoost, and a small but non-zero 0.06 to the MLP. The ensemble reaches an OOF MAE of 449.47 CNY and a holdout MAE of 421.03 CNY, the final test error: 21.05 CNY (4.5%) below the best single model on the OOF estimate, 8.28 CNY (1.9%) below it on the holdout set, and 54.0% below the linear baseline. Notably, the weakest member retains a non-zero weight precisely because of its low error correlation with the trees, which confirms that the fusion gain stems from the complementarity of heterogeneous members rather than from single-model accuracy alone.

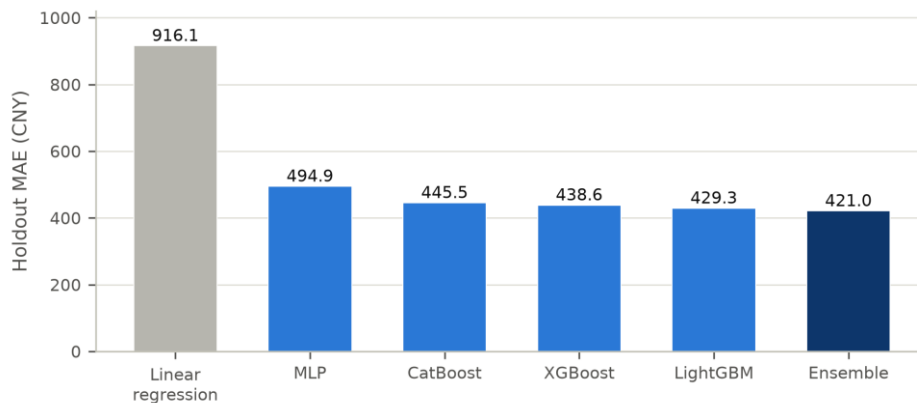


Figure 4: Holdout MAE comparison of the linear baseline, the four base learners, and the ensemble.

4.4. Error Analysis

Two observations, both measured on the ensemble's holdout predictions, characterize the remaining error. First, it concentrates in the high-price segment: the top decile of true prices (above 14,299 CNY) contributes 28.9% of the total absolute error, with a segment MAE of 1,215.59 CNY against 332.75 CNY for the remaining 90% of vehicles. The logarithmic transformation compresses the resolution of the high-price range, and such vehicles are sparse in the training data, so this segment is the dominant error source. Second, 6.56% of the holdout samples (984 of 15,000) have anonymized embedding vectors that reappear identically in the development set, indicating the same physical vehicle listed more than once. On these samples the ensemble achieves an MAE of only 164.31 CNY, versus 439.05 CNY on the remaining samples, which shows that the gradient-boosted trees implicitly memorize duplicated vehicle records through fine-grained splits on the embedding features. Together these findings suggest that further headway on this task lies less in stronger learners than in a better treatment of the sparse high-price regime.

5. Conclusion

This paper presented a heterogeneous ensemble framework for used car price prediction that integrates three gradient-boosting libraries and a neural network through an OOF-weighted fusion. The experiments support three conclusions. First, metric-aligned objective design, namely a log-transformed target, L1-type losses, and cross-validated early stopping, is the primary factor behind single-model performance. Second, fusing heterogeneous base learners with weights fitted on out-of-fold predictions reduces the error beyond the best single model by 4.5% on the cross-validation estimate and 1.9% on the holdout set, and the gain is driven by error complementarity rather than member accuracy. Third, the final ensemble attains a holdout MAE of 421.03 CNY, a 54.0% reduction relative to a linear baseline, with the residual error concentrated in the sparse high-price segment. Future work will explore richer target encodings, principled feature selection, additional diversity in the model pool, and dedicated treatment of high-price vehicles, for example through segment-aware objectives or sample reweighting.

References

- [1] Kim, M., Choi, J., Kim, J., Kim, W., Baek, Y., Bang, G., Son, K., Ryou, Y., & Kim, K. (2023). Trustworthy residual vehicle value prediction for auto finance. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(13), 15480–15488.
- [2] Bukvic, L., Pasagic Skrinjar, J., Fratovic, T., & Abramovic, B. (2022). Price prediction and classification of used-vehicles using supervised machine learning. *Sustainability*, 14(24), 17034.

- [3] Amik, F. R., Lanard, A., Ismat, A., & Momen, S. (2021). *Application of machine learning techniques to predict the price of pre-owned cars in Bangladesh*. *Information*, 12(12), 514.
- [4] Cui, B., Ye, Z., Zhao, H., Renqing, Z., Meng, L., & Yang, Y. (2022). *Used car price prediction based on the iterative framework of XGBoost+LightGBM*. *Electronics*, 11(18), 2932.
- [5] Huang, J., Yu, Z., Ning, Z., & Hu, D. (2023). *Used car price prediction analysis based on machine learning*. In N. Radojević et al. (Eds.), *Proceedings of the 2022 International Conference on Artificial Intelligence, Internet and Digital Economy (ICAID 2022)* (pp. 356–364). Atlantis Press. https://doi.org/10.2991/978-94-6463-010-7_37
- [6] Fayyaz, I., Ali, G. G. M. N., & Khairunnesa, S. S. (2025). *Advanced feature engineering and machine learning techniques for high accurate price prediction of heterogeneous pre-own cars*. *Vehicles*, 7(3), 94.
- [7] Dutulescu, A. , Catruna, A. , Ruseti, S. , Iorga, D. , Ghita, V. , & Neagu, L. M. , et al. (2023). *Car price quotes driven by data-comprehensive predictions grounded in deep learning techniques*. *Electronics*, 12(14), 25.