

Predicting Car Prices Using Machine Learning

Model evaluation, feature availability and browser deployment

Abdirahman Jibril

Original project: 2025

Research extension and interactive demo: October 2026

Abstract

This project, started in 2025, investigates used-vehicle price prediction using regression methods. Its 2026 extension adds an auditable Python pipeline, a feature-availability study and an interactive browser deployment. The 301-row public vehicle dataset supports a comparison of baseline and revised protocols. The revised attributes-only experiment compares a median baseline, linear regression, ridge, a decision tree and a random forest using development-set cross-validation. The selected forest achieves a held-out MAE of 1.495 and $R^2 = 0.740$ on 60 observations. A separate exploratory nested cross-validation study, restricted to the 239 development observations, investigates access to a reference-price feature. Forest outer-fold MAE changes from 1.600 without that feature to 0.743 with it. The result illustrates the importance of the prediction-time input contract; it does not establish availability of the feature, Finnish-market performance or insurance repair-cost accuracy. All errors are reported in dataset price units because the currency scale is insufficiently documented in the inspected primary metadata.

1 Problem and research questions

The task is to estimate a vehicle's recorded selling price from its observed attributes. It is a supervised regression problem. The study treats the target as a recorded listing value; a completed transaction is not established by the source documentation. Prediction of current listing values, future market prices and collision repair costs are different tasks.

The research questions are:

1. How do linear and tree baselines perform under the original input contract?
2. Which baseline performs best under an attributes-only input contract?
3. How does the availability of a reference-price feature affect development-set performance?
4. What do error analysis and data limitations imply for further development?

The contribution is an engineering study of reproducibility, evaluation and feature availability. It is not a new learning algorithm or a production valuation system.

2 Focused literature review

Searches conducted on 25 October 2026 used Google Scholar, IEEE Xplore discovery records, Aaltodoc and primary methodology publications. This is a focused review, not an exhaustive

systematic review. The accompanying source matrix records queries, access status and the implementation decision supported by each source.

2.1 Tabular modelling and pricing context

Grinsztajn et al. [2] provide a broad benchmark supporting tree-based methods as serious tabular baselines. Their medium-sized datasets do not establish a winner for this much smaller sample; they motivate including forest and boosting comparisons rather than assuming a neural network is necessary.

The IEEE case study by Hankar et al. [4] and the NetCIT pricing paper [5] describe comparisons of used-car regression methods. Only their abstracts and discovery metadata were examined. Their reported outcomes are not imported into our results, and differently sampled studies cannot be ranked by comparing headline R^2 values.

Guan’s Aalto thesis [6], examined at metadata and abstract level, connects electric-vehicle valuation to additional sentiment and insurance indicators. Those fields do not exist in our dataset. Mäkinen’s Aalto thesis [7], for which selected full-text methods and results sections were examined, offers a model-comparison example in car-sales forecasting. Its target is sales volume, not per-vehicle price.

2.2 Evaluation and uncertainty

Cawley and Talbot [3] show why model selection can overfit a finite evaluation sample. We therefore place hyperparameter selection inside the outer folds of the exploratory extension. Historical inspection still prevents describing this dataset as an untouched external benchmark.

Meinshausen [8] distinguishes conditional-distribution estimation from ordinary point prediction. Our bootstrap interval estimates uncertainty in MAE conditional on a fixed model and split. It is not an individual vehicle-price interval, and no quantile forest is claimed as implemented. Official scikit-learn guidance [9, 10] informs fold-local preprocessing.

3 Dataset provenance

The data were re-downloaded from Nehal Birla and collaborators’ Kaggle *Vehicle dataset*, version 4 [1]. The selected archive member is `car_data.csv`, not the larger similarly named files in the same archive. It contains 301 rows and nine columns. The SHA-256 is recorded in `data/PROVENANCE.json`.

The primary Kaggle metadata labels the licence “Database: Open Database, Contents: Database Contents”. The package preserves source attribution, metadata and licence links. Some vehicle names identify motorcycle models; an exclusively passenger-car population is not assumed. Year ranges from 2003 to 2018. The target ranges from 0.1 to 35.0 in dataset units. Currency scale, observation dates, repeated-vehicle identifiers and transaction status are not established by the inspected primary metadata.

3.1 Original-protocol baseline experiment

The baseline experiment manually encodes fuel type, creates dummy variables before splitting, retains `Present_Price`, and uses a 50/25/25 split with seed 42. The resulting counts are 150 training, 75 validation and 76 test observations.

Table 1: Baseline verification results under the original 50/25/25 protocol, rerun in 2026.

Model	Val. MAE	Val. MSE	Test MAE	Test R^2
Linear regression	1.742	14.598	1.218	0.812
Decision tree	0.876	2.647	0.624	0.933

The baseline results are reported in dataset units. The revised experiment below changes the input contract and evaluation procedure, so its scores are not interpreted as a direct improvement percentage over this baseline.

4 Attributes-only experiment

4.1 Input contract and preparation

The primary input contract uses year, mileage, prior-owner count, fuel type, seller type and transmission. Vehicle name is excluded. Reference price is excluded until its meaning and prediction-time availability are established. Exclusion is a conservative contract choice; the feature is not declared target leakage merely because it is predictive.

The pipeline validates the schema, removes invalid targets and impossible numerical values, and drops exact duplicate model rows. This removes two duplicates and leaves 299 observations. It does not prove that all repeated vehicles or near-duplicates have been identified. Median imputation, scaling and categorical imputation/one-hot encoding are fitted inside each training fold. Unknown categories are handled explicitly.

4.2 Selection and frozen holdout

A seed-42 split reserves 60 observations for the frozen test and 239 for development. Five-fold development CV selects hyperparameters and model family by MAE. Every candidate uses the same folds. The selected configuration is refitted on all development rows and evaluated once on the test set. The later exploratory study never uses that holdout to select its models.

Table 2: Development CV in the primary experiment. Fold SD is not a confidence interval.

Model	CV MAE	Fold SD
Random forest	1.577	0.389
Decision tree	1.710	0.369
Ridge regression	2.060	0.434
Linear regression	2.090	0.406
Median baseline	3.270	0.526

The random forest uses 200 trees, maximum depth 6 and minimum leaf size 3. Its development CV MAE is approximately 52% lower than the median baseline. This is a within-protocol CV comparison, not a claimed improvement over the baseline experiment.

4.3 Metrics

$$\text{MAE} = \frac{1}{n} \sum_i |y_i - \hat{y}_i|, \quad \text{RMSE} = \sqrt{\frac{1}{n} \sum_i (y_i - \hat{y}_i)^2},$$

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}.$$

MAE and RMSE have target units; MSE has squared units. R^2 is not classification accuracy and can be negative. We do not label dataset units as euros.

4.4 Test results and errors

Selected model: Random forest. Held-out MAE: 1.495; RMSE: 2.590; R^2 : 0.740. Units: dataset price units (currency scale unverified). Conditional bootstrap 95% MAE interval: [1.042, 2.074].

The 95% percentile bootstrap interval for MAE is approximately [1.042, 2.074], based on 1,000 resamples of fixed test errors. It omits uncertainty from data collection, model selection and retraining. The test results differ from the historical experiment because the input contract, split and deduplication differ. They are not a like-for-like performance comparison.

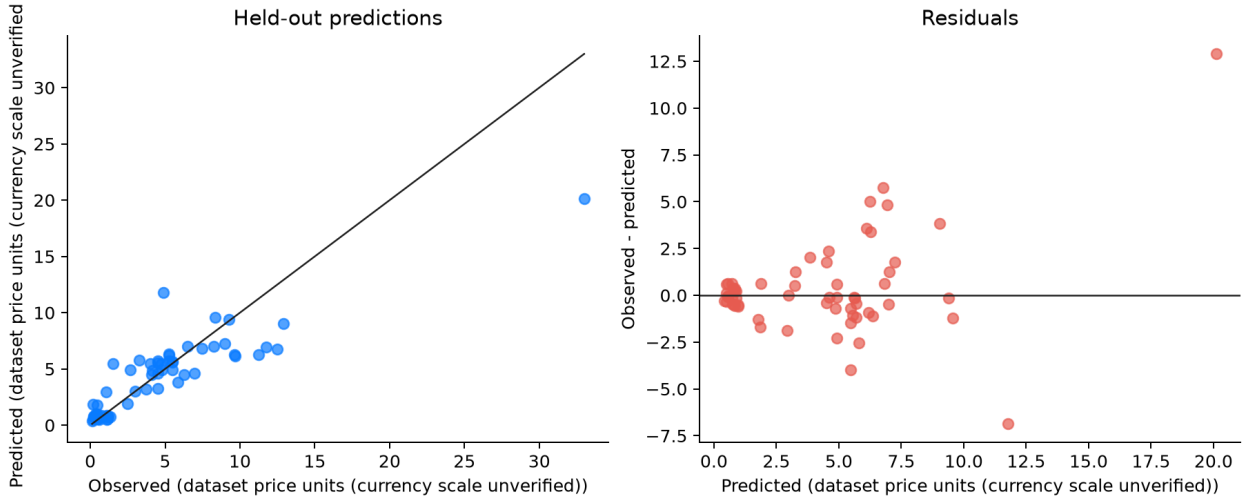


Figure 1: Frozen holdout predictions and residuals. Large errors remain; a global score does not describe every vehicle equally well.

The largest absolute test error is 12.892 dataset units at source row 64 (observed 33.000, predicted 20.108). The median absolute test error is 0.621; the 90th percentile is 3.864. These summaries describe this fixed test set, not a population guarantee. Permutation importance is exported as exploratory, non-causal analysis. Correlated features can share importance. Further model changes motivated by holdout errors would require a new evaluation protocol rather than repeated test-set tuning.

5 Exploratory nested CV and feature availability

This extension was designed after inspection of the initial holdout results and is labelled exploratory. It uses only the original 239 development rows. The outer five-fold partition (seed 2026) is identical across both input contracts. Three inner folds (seed 42) select hyperparameters separately within each outer training fold.

The procedures are median prediction, ridge, a decision tree, a 150-tree forest and histogram gradient boosting. The boosting comparison is new to the exploratory extension; it is not selected using frozen-test results. Two scenarios are evaluated: attributes only, and the same attributes plus `Present_Price`. The original development IDs are retained so differences do not arise from changing the row population.

Table 3: Exploratory outer-fold MAE, mean and descriptive SD over five folds.

Procedure	Attributes only	With reference price
Median	3.284 ± 0.511	3.284 ± 0.511
Ridge	2.118 ± 0.346	1.105 ± 0.098
Decision tree	1.806 ± 0.428	0.827 ± 0.269
Random forest	1.600 ± 0.244	0.743 ± 0.211
Histogram gradient boosting	1.886 ± 0.276	0.993 ± 0.198

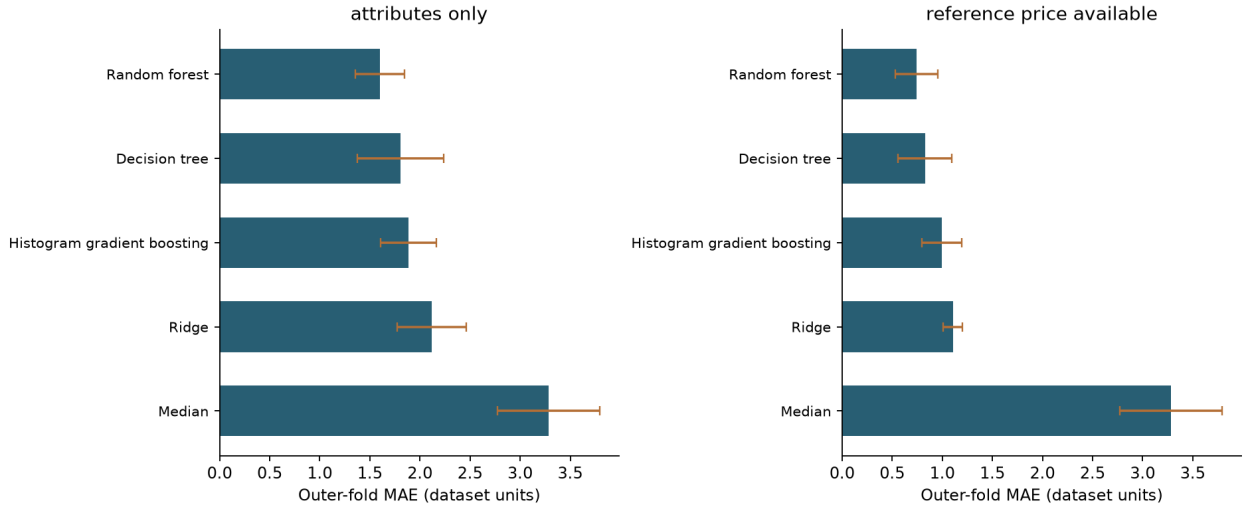


Figure 2: Development-only nested comparison. Error bars show fold SD, not confidence intervals.

For the forest, mean outer-fold MAE is 1.600 without reference price and 0.743 with it. The feature has substantial predictive value under this experimental contract. This does not establish that a trustworthy reference price is available at deployment time. Adding it to a real system requires a documented source, timestamp and retrieval process.

With five dependent folds and a small historical sample, we do not claim a statistically established universal ranking. We do not rerank candidate models by the frozen holdout or claim that adding a more complex algorithm must improve generalisation.

6 Engineering and reproducibility

The package contains two notebooks, shared Python modules, dependency versions, source meta-data, model-selection records, row IDs, predictions, plots and a model card. Notebook 00 evaluates the original baseline protocol; Notebook 01 explains the attributes-only and exploratory experiments. Each fresh training run writes to a new output folder.

The Jupyter version uses the same implementation as the command-line experiment. Recorded verified outputs can be inspected without retraining; optional cells rerun the explicit protocol. The selected 200-tree forest is exported with its preprocessing parameters for browser inference. A JavaScript implementation reproduces numerical scaling, one-hot encoding and tree traversal. Its predictions are verified against Python on all 301 source observations with an absolute tolerance of 10^{-10} . This is deployment-equivalence verification, not an additional generalisation test.

7 Limitations and practical next steps

- The sample is small, old and heterogeneous, with uncertain sampling and price units.
- A random split does not evaluate market drift or new vehicle populations. Grouped or chronological evaluation needs reliable identifiers or timestamps.
- Historical inspection means this is an exploratory portfolio study, not an independent external benchmark.
- Reference-price availability is an input-contract question, not settled by a lower CV error.
- Source support and reproducible numbers do not establish commercial or insurance suitability.

A stronger next dataset would contain documented units, observation dates, vehicle identifiers and a relevant target population. Insurance repair-cost modelling additionally needs actual repair-cost labels and features available before assessment. A listing-price dataset cannot establish collision repair costs. For insurance workflow engineering, the separate IntakeAI prototype investigates document extraction and human review; it is a different task.

8 Conclusion

The project brings together baseline model comparison, a frozen attributes-only benchmark, a development-only feature-availability study and a working browser demo. Its main finding is that the data contract and evaluation design matter at least as much as the choice of model family. The work demonstrates an auditable engineering process while retaining clear limits on deployment claims.

References

- [1] Nehal Birla and collaborators. *Vehicle dataset*. Kaggle, version 4, retrieved 25 October 2026. <https://www.kaggle.com/datasets/nehalbirla/vehicle-dataset-from-cardekho>.
- [2] L. Grinsztajn, E. Oyallon and G. Varoquaux. *Why do tree-based models still outperform deep learning on typical tabular data?* NeurIPS 2022. <https://arxiv.org/abs/2207.08815>.
- [3] G. C. Cawley and N. L. C. Talbot. *On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation*. JMLR 11:2079–2107, 2010. <https://jmlr.org/papers/v11/cawley10a.html>.
- [4] M. Hankar, M. Birjali and A. Beni-Hssane. *Used car price prediction using machine learning: A case study*. IEEE, 2022. Abstract examined. <https://ieeexplore.ieee.org/document/9800719/>.
- [5] *Prediction of Used Car Price Based on Supervised Learning Algorithm*. NetCIT 2021. DOI: 10.1109/NetCIT54147.2021.00036. Abstract examined.
- [6] X. Guan. *Predicting prices of used electric vehicles based on sentiment and insurance indicators*. Aalto University master’s thesis, 2025. Abstract examined. <https://urn.fi/URN:NBN:fi:aalto-202601151262>.
- [7] D. Mäkinen. *Comparison of machine learning methods for social media and open data to predict sales*. Aalto University master’s thesis, 2015. Selected full-text sections examined.
- [8] N. Meinshausen. *Quantile Regression Forests*. JMLR 7:983–999, 2006. <https://www.jmlr.org/papers/v7/meinshausen06a.html>.
- [9] scikit-learn developers. *Common pitfalls and recommended practices*. Accessed 25 October 2026. https://scikit-learn.org/stable/common_pitfalls.html.
- [10] scikit-learn developers. *Cross-validation: evaluating estimator performance*. Accessed 25 October 2026. https://scikit-learn.org/stable/modules/cross_validation.html.