Data Mining 1 (2025) 1-10 Submitted 1/09; Published 12/12

A Comparative Analysis Of Supervised Learning Techniques For Used Car price Evaluation

Md Jackie Islam MD.ISLAM@STUD.MIF.VU.LT
Hossain Mohammad Jabir HOSSAIN.JABIR@MIF.STUD.VU.LT
Md Naimur Raihan Emon MD.EMON@MIF.STUD.VU.LT
Data Science study programme
Faculty of Mathematics and Informatics

Advisor: Jurgita Markevičiūtė

Abstract

For this study, we compares different machine learning models for predicting car prices. The main goal is to find out which regression model gives the most accurate and precise price prediction. We used a car dataset that includes features like engine size, mileage, fuel type and engine-type etc. Before training the models, we cleaned the data and did some exploratory data analysis (EDA) to address missing values and checked which features affect the price more. Multicolinearity was also checked as we applied several supervised regression algorithms, including Linear Regression, Lasso, Ridge, Support Vector Regression (SVR), Decision Tree, Random Forest, and K-Nearest Neighbors (KNN). We compared the models using important performance metrics like the coefficient of determination (R^2), Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). This research will help us to identify, the model that provides the most consistent and accurate performance for predicting used car prices.

Keywords: Regression model, Comparative study , Supervised Learning

1. Introduction

Automobile price prediction has become a crucial study for the recent decades as the number of automobile is increasing day by day. Lithuania recorded an increase of 46.4% (to 11'646 units) of car registration in 2025 AV Automotive Research (2025). So, both buyers and sellers are looking for best price and time to own a car. People select to lease cars, that is legally binding agreement in between the seller and buyer. Direct sellers, third parties, commercial entities, and insurance companies all falls under the seller group. Buyers pay the installments for a pre-determined time period under a lease arrangement. These lease instalments are based on the vehicle's estimated worth; thus the sellers want to know what the fair presumable price is for their automobiles(Gupta et al. (2022)). In several countries, buying a new or used car is the best choice for customer because its price is reasonable and affordable by buyer. After few years of using them, it may get a profit from resell again(Monburinon et al. (2018)). Now a days, car price usually depends on it's engine horsepower, enginetype and fueltype. The more features the car has the more price it will

be sold off. As, the world is looking for renewable energy, the electric car prices have gone up (for both used and new car). But determining a reasonable estimated price for a car is both crucial and difficult. As a result, an accurate price prediction technique for second hand or new autos is required. The goal of this article is to explain the price of cars using regression models. For precise calculation of automotive prices, on the other hand, necessary specific knowledge, as quality is often dependent on a number of various features and variables (e.g fuel type,engine horsepower). Using this data on a car sale site, it is possible to calculate how much a vehicle can be sold according to its features, using machine learning algorithms (Muti and Yıldız (2023)). The goal of this study is to explain the prices of cars by finding out the best & accurate regression models.

2. Literature Review

Some recent study was done to predict car prices and find out the accurate model to fit the data. But it was difficult to find the model which gives precise result. (Soejima and Hirose (2011)) used the combination method of regularization methods and the k-NN method with the optimal weighting function. The result shows, k-NN with optimal weighting value via the elastic-net was found to be the best used car auction price prediction model.(Nishitha et al. (2023))Used classification decision trees to produce the highest accuracy of 0.99, and regression models produced the greatest results with a linear regression model's accuracy of 0.992. In light of this decision tree classification and linear regression, the model is created for new cars only. (Chen et al. (2017)) This paper collects more than 100,000 used car dealing records throughout China to do empirical analysis on a thorough comparison of two algorithms: linear regression and random forest. Results shows that, random forest has a stable but not ideal effect in price evaluation model for a certain car make, but it shows great advantage in the universal model compared with linear regression. This indicates that, random forest is an optimal algorithm when handling complex models with a large number of variables and samples, yet it shows no obvious advantage when coping with simple models with less variables.(Gongqi et al. (2011)) established BP neural network (NN) and used to extract the feature of the distribution curves in various conditions. A set of schemed data was used to train the NN and reached the training goal. As a result, the newly proposed model is feasible and accurate for residual value prediction of the used cars with various conditions. (Pudaruth (2014)) investigates the application of supervised machine learning techniques to predict the price of used cars in Mauritius. The predictions are based on historical data collected from daily newspapers. Different techniques like multiple linear regression analysis, k-nearest neighbours, naïve bayes and decision trees have been used to make the predictions. But the main limitation of his study is the low number of records that have been used. (Venkatasubbu and Ganesh (2019)) Used Machine Learning Algorithms such as Lasso Regression, Multiple Regression and Regression trees and try to develop a statistical model which will be able to predict the price of a used car, based on the previous consumer data and a given set of features. But the resulr was not efficient because in need of appropiate datasets.

3. Data

3.1 Data source

The data set was taken from "kaggle" . The data set contains 205 observations with 26 variables. The link of the data set source is ""https://www.kaggle.com/datasets/nikitatopilskiy/car-price-assignment" **[Car Price Dataset]**

3.2 Variables

Description of the Dataset
The dataset is a consist of variables of used cars for regression modelling of vehicle sale price, where the target variable is typically the selling price (often recorded as `price_euro`. The data set contains information on car identity, technical specifications (such as engine size and mileage), vehicle history (for example, kilometers driven) and categorical variables like fuel type, transmission, ownership and location, making it suitable for using supervised machine learning algorithm.

Variables Description
Table 1 summarises the main variables in the Kaggle used–car–price data and their usual variable types :

Variable	Type (Typical)
`Car Name`	Categorical (string; car model/name)
`Door Number`	Categorical (string; e.g. one, two)
`Car body`	Categorical (string; sedan, suv, heatchback)
`Kilometers_Driven`	Numeric (integer; odometer reading)
`Fuel_Type`	Categorical (e.g. Petrol, Diesel, CNG)
`Transmission`	Categorical (e.g. overhead cam(Ohc), Overhead Camshaf(Ohcv)
`Citympg`	Categorical (e.g. First, Second owner)
`Mileage`	Numeric (continuous; km/l or similar)
`Engine`	Numeric (continuous; engine capacity, e.g. cc)
`stroke`	Numeric (continuous; engine stroke)
`Seats`	Numeric (integer; number of seats)
`Price`	Numeric (continuous; selling price, target)

Table 1: Key variables in the used–car–price dataset and their typical types.

Data set characteristics : Observations were collected from individual used–car listings, so each row represents one car with its attributes and observed selling price. The mix of continuous and categorical predictors, along with potential problems like missing values and outliers in mileage, engine size or price, makes the dataset a real world example for practising data cleaning, feature engineering and regression modelling in machine learning.

3.3 Data Pre-processing

Data set was cleaned before using for any kind of analysis. No missing value was found and "CarName" variable was transformed into "Car Company" for better exploration. CarId

and symboling has been removed in the clean dataset. All the car company name was in short form, Only first three letters were taken to indicate car company name.

4. Methodology

- **Linear Regression :**

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i, \tag{1}$$

where, Y_i is the outcome variable for observation i, X_{ij} denotes the jth predictor for observation i, β_0 is the intercept, $\beta_1, \ldots, \beta_k$ are regression coefficients and ε_i is the random error term.

- **Lasso Regression :** Lasso regression estimates the coefficient vector $\boldsymbol{\beta}$ by solving,

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}, \tag{2}$$

where $\lambda \geq 0$ controls the strength of the L_1 penalty.

- **Ridge Regression :** Ridge regression is a variant of multiple linear regression that adds an L_2 penalty on the regression coefficients to reduce overfitting and handle multicollinearity. It estimates the coefficient vector $\boldsymbol{\beta}$ by solving,

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}, \tag{3}$$

where $\lambda \geq 0$ controls the strength of the L_2 penalty.

- **Support Vector Regression :** Support vector regression (SVR) estimates a function $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + b$ by solving,

$$\min_{\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\xi}^*} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^{n} (\xi_i + \xi_i^*) \tag{4}$$

where ε is the insensitive tube width and $C > 0$ controls the trade-off between flatness and violations.

- **Decision Tree :** A regression decision tree recursively partitions the predictor space into M disjoint regions $R_1, \ldots, R_M$ and predicts by a region-wise constant:

$$\hat{f}(\mathbf{x}) = \sum_{m=1}^{M} c_m \, \mathbb{I}(\mathbf{x} \in R_m), \tag{5}$$

where c_m is typically the mean of the response values y_i for all training observations with $\mathbf{x}_i \in R_m$ and $\mathbb{I}(\cdot)$ is the indicator function.

- **Random Forest :** Random forest regression builds an ensemble of B regression trees $\{\hat{f}_b(\mathbf{x})\}_{b=1}^{B}$ on bootstrap samples and random feature subsets. The final prediction is the average over trees :

$$\hat{f}_{\mathrm{RF}}(\mathbf{x}) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}_b(\mathbf{x}). \tag{6}$$

  Each tree $\hat{f}_b$ is constructed by recursive binary splits to minimize the within-node sum of squared errors.

- **K-Nearest Neighbors (KNN) :** In k-nearest neighbors (KNN) regression, the prediction for a new observation $\mathbf{x}$ is the average response of its k closest training points :

$$\hat{f}_{\mathrm{KNN}}(\mathbf{x}) = \frac{1}{k}\sum_{i\in\mathcal{N}_k(\mathbf{x})} y_i, \tag{7}$$

  where $\mathcal{N}_k(\mathbf{x})$ denotes the index set of the k nearest neighbors of $\mathbf{x}$ in the training data under a chosen distance metric (e.g., Euclidean distance).

- **Gradient boosting regression :** Gradient boosting regression builds an additive model of M weak learners $\{h_m(\mathbf{x})\}_{m=1}^{M}$ in a stagewise manner. The model has the form

$$F_M(\mathbf{x}) = \sum_{m=1}^{M} \nu\, h_m(\mathbf{x}), \tag{8}$$

  where $\nu \in (0, 1]$ is the learning rate. At each iteration m, a new base learner h_m is fitted to the negative gradients (pseudo-residuals) of a chosen loss function $L(y, F(\mathbf{x}))$, and then added to the current model.

5. Exploratory Data Analysis

5.1 Continuous variable

We checked all the continuous variables if there were any outliers and correlations among the variables.

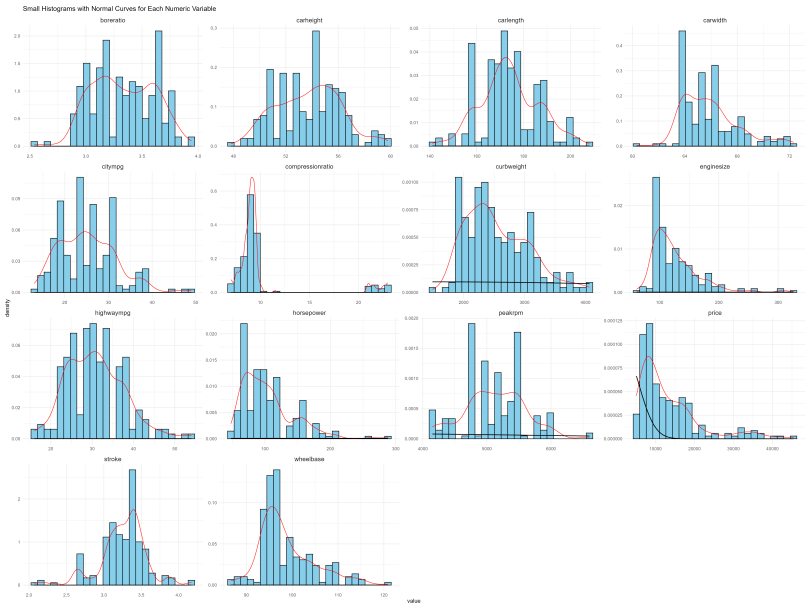

Figure 1: Histogram of continuous variables

These histograms shows, how each numeric variable is distributed compared to a fitted normal curve (in red). From the plots, it is clear that most variables do not follow normality and many of them are strongly right-skewed. Strongly right-skewed variables are (enginesize, horsepower, curbweight, price, compressionratio, peakrpm, citympg, highwaympg, and stroke) all showing long right tails. These variables have most observations in the low or mid ranges, while a few high-performance or luxury cars create extreme values on the right. Such skewness breaks the normality assumption and creating instability for linear regression models. Moderately skewed variables: carlength, carwidth and wheelbase show more balanced distributions but still contain some visible skewness. They are closer to symmetric but still deviate from the bell curve. Closest to normal: (carheight and boreratio) appear to be the most symmetric variables in the dataset. Their distributions resemble a bell shape because these measurements do not vary drastically across cars. Overall, the histograms confirm that normality is violated for most continuous variables. This indicates that the data set is sensitive to some Regression model and may perform poorly, So tree-based models such as Random Forest and Gradient Boosting should be taken into consideration because they are more robust method.

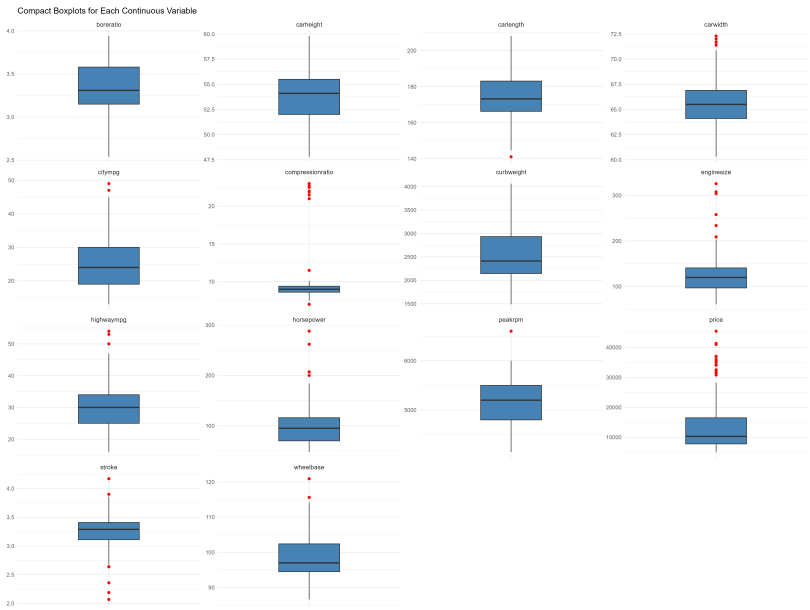

Figure 2: Boxplots of continuous variables

The boxplots show clearly where we have outliers in the continuous variables. Enginesize, compressionratio, horsepower, stroke and price have several points far away from the main box, which usually indicates towards more special or luxury cars. These outliers stretch the range, influence the mean and can make models unstable if they are not handled well. Some variables like wheelbase and carlength have fewer outliers and look more stable. This is why we should work with both a full dataset and a cleaned dataset without these extreme observations.

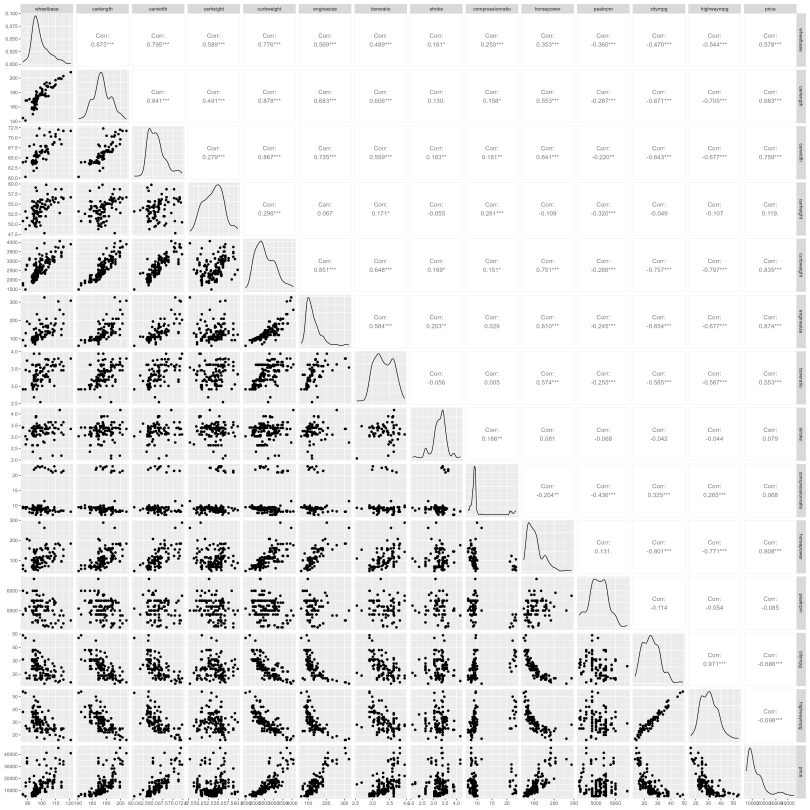

Figure 3: Correlation matrix of continuous variables

The scatterplot matrix helps us see both the distribution and the relationships between all the numeric variables at the same time. Strong positive patterns appear between enginesize, horsepower, curbweight and price, which matches our expectations: bigger and heavier cars with more powerful engines usually cost more. The diagonal density plots again confirm the right skewness in performance and price variables. In some panels, the relationship is weak or almost flat, meaning those variables add less information for price prediction. This figure is useful for understanding which features are important and where multicollinearity might appear.

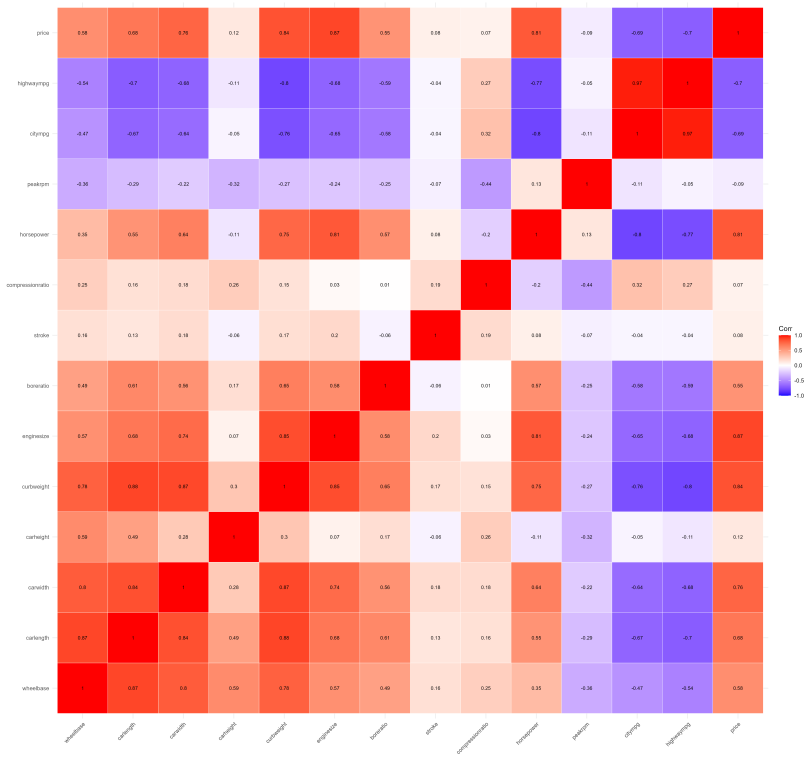

Figure 4: Heatmap of continuous variables

The heatmap gives a clean overview of how strongly each pair of numeric variables is linearly related. Price is highly correlated with enginesize, horsepower, curbweight and carwidth Which confirms that these are key drivers of car value. We also see strong positive correlation between size-related variables such as carlength, wheelbase and curbweight. On the other hand compressionratio has weak correlation with most features, so it plays a smaller direct role in price. The strong blocks in the heatmap also show multicollinearity, which explains why we later use regularized models and tree-based methods, not just standard Linear Regression.

5.2 Categorical Variables

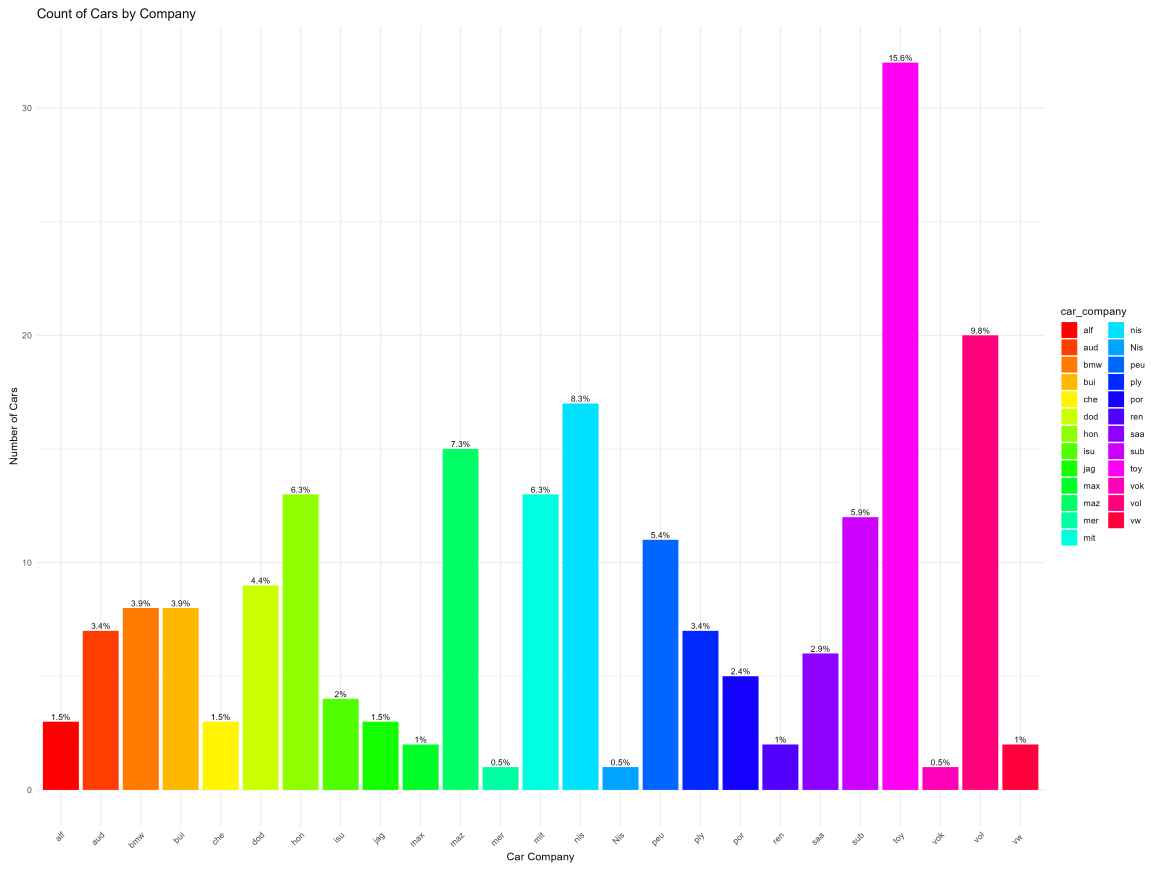

Figure 5: Count plot of car company

This bar chart shows how many cars we have from each manufacturer. Brands like Toyota, Nissan, Mazda, and Honda appear very often, while luxury brands such as Jaguar or Alfa Romeo only have a few cars. So the dataset is dominated by common mid-range brands instead of premium brands. For modelling, this means our results will be more reliable for the frequent manufacturers, and we should be careful when interpreting results for very rare brands, because there is not much data for them.

Figure 6: Bar plot of categorical variables

The barplots for categorical variables show the typical structure of cars in the dataset. Most vehicles use gas fueltype, have a front-engine layout, and are front-wheel drive, which is standard for normal passenger cars. Four-door cars dominate, while two-door cars represent a smaller group of sportier or compact models. Some rare categories, such as unusual fuelsystems or cylinder numbers, appear only a few times. These imbalances are important for modelling, because categories with very few observations may not have a stable effect on price.

5.3 Tables

Variable	Number of Outliers
carwidth	8
enginesize	10
stroke	20
compressionratio	28
horsepower	6
price	15

Table 2: Number of Outliers for Selected Variables

This table quantifies the number of outliers in selected variables, revealing the degree of variability and extremeness in the data. Compressionratio has the highest outlier count, suggesting this feature varies more heavily across special engine types. Price, enginesize and horsepower also contain numerous outliers, supporting the earlier visual results from the boxplots. High outlier counts mean these variables should be handled carefully to avoid distortion in model training. Quantifying outliers helps justify the decision to create an outlier-free dataset to improve model performance.

Variable	Mean	Median	SD	Min	Max
wheelbase	98.75659	97	6.021776	86.6	120.9
carlength	174.0493	173.2	12.33729	141.1	208.1
carwidth	65.9078	65.5	2.145204	60.3	72.3
carheight	53.72488	54.1	2.443522	47.8	59.8
curbweight	2555.566	2414	520.6802	1488	4066
enginesize	126.9073	120	41.64269	61	326
boreratio	3.329756	3.31	0.270844	2.54	3.94
stroke	3.255415	3.29	0.313597	2.07	4.17
compressionratio	10.14254	9	3.97204	7	23
horsepower	104.1171	95	39.54417	48	288
peakrpm	5125.122	5200	476.9856	4150	6600
citympg	25.21951	24	6.542142	13	49
highwaympg	30.75122	30	6.886443	16	54
price	13276.71	10295	7988.852	5118	45400

Table 3: Summary Statistics of Key Variables

The summary statistics indicate central tendency and variability across numeric variables. For many features, the mean is larger than the median, confirming right skewness, which aligns with the histogram and boxplot observations. Large standard deviations (for example in price, horsepower, and enginesize) reveal substantial variation and wide differences between car types. Wide ranges (differences between minimum and maximum values) confirm that the dataset contains diverse categories, from small economy cars to larger, high-performance models. These statistics guide data preprocessing decisions, such as scaling, transformation, and handling of skewed variables before model fitting.

6. Results

Model	R^2	RMSE	MAE	Performance Status
Random Forest	0.8801	2021.35	1589.22	Highest Accuracy (R^2)
GBR	0.8614	1444.62	1079.10	Smallest Errors (RMSE/MAE)
Linear Regression	0.8359	3121.20	2312.77	Good Fit, High Errors
KNN	0.7856	2058.29	1447.03	Moderate
Ridge	0.7633	1888.21	1399.98	Moderate
Decision Tree	0.7287	2021.35	1589.22	Moderate
Lasso	0.7245	2036.83	1407.85	Lower Fit
SVR	0.6648	2246.90	1575.24	Lowest Fit

Table 4: Model performance without outliers

This table shows the performance of the model without having any outliers in the data set. We can see that, Random Forest gives the highest R^2 in car prices (88%), so it understands the overall pattern the best. But even with that, Gradient Boosting is actually the most accurate model because it has the smallest errors. While its R^2 is slightly lower than Random Forest (0.86 vs 0.88), its error rates are significantly lower. On average, its predictions are off by only $1,079. It predicts prices much closer to the real values. Linear Regression looks good because of its R^2 (0.835), but the errors are very high, RMSE (3121) and MAE (2313), meaning it makes big mistakes, especially on expensive cars. KNN, Ridge, and Decision Trees are just okay but not strong. Lasso becomes weaker because it drops some useful features. SVR performs the lowest R^2 (0.66) because it cannot capture the non-linear relationships in this data. This confirms that standard Linear Regression was overfitting (chasing outliers), while Ridge and Lasso sacrificed some "fit" to provide more stable, realistic price predictions (lower RMSE). So overall, without outliers, tree-based models like Gradient Boosting and Random Forest give the best and most stable results.

Table 5: Model performance with outliers

Model	R^2	RMSE	MAE	Performance
Linear Regression	0.6953	1343.78	1065.14	Weakest model).
Random Forest	0.9144	2254.35	1596.23	Highest R^2, strong Model
SVR	0.8808	2660.30	1858.44	High accuracy and moderate errors.
Lasso	0.8635	2846.83	2023.92	Good balance of fit and good stability.
Ridge	0.8543	2941.14	2179.11	Similar to Lasso, weaker fit.
Decision Tree	0.8322	3156.12	2414.74	Moderate fit
GBR	0.8998	2438.76	1673.05	Strong model.
KNN	0.6945	3751.80	2455.12	Lowest fit

In this table we keep the outliers as before. Here, Random Forest still has the highest R^2(0.91), but its errors become bigger because the extreme values affect it. SVR actually does much better than before, it handles the outliers nicely and gives good accuracy. Lasso also stays stable because regularization reduces the effect of those extreme values. Linear Regression behaves in a strange way: it gives the lowest errors but a low R^2. This happens

because the model stays close to the average price, so the errors look small but the model doesn't explain the whole data well. KNN becomes very weak because outliers breaking the distance calculations. Gradient Boosting completely fails here the R^2 becomes negative and the errors become huge, which means the model cannot learn properly because it is too sensitive to outliers. So with outliers included, the ranking changes a lot: Random Forest and SVR stay strong, Linear Regression looks good only by error numbers and Gradient Boosting collapses.

7. Conclusions

This study compared several supervised learning techniques including Linear Regression, Lasso, Ridge, Support Vector Regression, Decision Tree, Random Forest, K–Nearest Neighbours and Gradient Boosting for the task of predicting used car prices from a set of numerical and categorical attributes. Across all models, removing outliers consistently improved goodness of fit and reduced prediction errors, confirming that extreme observations can distort parameter estimates and degrade performance, especially for linear and distance–based methodsPudaruth (2014).

The empirical results demonstrate that tree–based models provide the best overall balance between accuracy and robustness for this problem. In the cleaned data set, Random Forest explains the highest proportion of variance in prices, while Gradient Boosting achieves the smallest RMSE and MAE, indicating more precise price estimates at the individual car level. Regularized linear models (Lasso and Ridge) offer interpretable baselines and perform reasonably well once outliers are removed, whereas single Decision Trees and KNN are more sensitive to noise and yield less stable predictions. Support Vector Regression occupies a middle ground, handling non–linearity better than ordinary linear regression but still affected by extreme valuesGupta et al. (2022). Overall, the study confirms that careful pre-processing combined with modern ensemble methods can substantially improve the quality of used car price evaluation and provides a practical framework for future research and real world deployment.

8. Recommendations

The comparative analysis shows that, tree-based methods, particularly Random Forest and Gradient Boosting Regression (GBR), provide the most reliable predictions for used car prices once influential outliers are handled properly. In the outlier free dataset, GBR achieves the lowest RMSE and MAE, while Random Forest attains the highest R^2, indicating that these models simultaneously capture non–linear relationships and maintain strong fit. Given their complementary strengths, practitioners who build pricing tools for online platforms, dealerships or insurers should prioritise gradient boosting or random forest models as baseline approaches, combined with systematic outlier diagnostics and robust validation procedures.

For operational deployment, the findings suggest several practical steps. First, data pipelines should include automatic detection and treatment of extreme observations in key variables such as engine size, horsepower, curb weight and price, because these points strongly affect both model stability and variable importance rankings. Second, models should be

calibrated and monitored separately for different market segments (for example, economy versus premium brands), since the dataset is dominated by mid–range manufacturers and predictions may be less stable for rare brands. Finally, future work should explore hybrid or stacked ensembles that combine the strengths of tree–based models with regular linear methods, and should test the models on larger, multi–country datasets to assess how well the conclusions can explain beyond the current sample.

9. Limitations

Although the results are encouraging, several limitations occured when interpreting this study. First, the dataset contains only 205 cars with 26 variables from a single publicly available source and the distribution is heavily dominated by a few manufacturers; therefore the trained models may not explain well to other markets (Continent). Second, the analysis relies on a fixed set of engineered features derived from the original `CarPrice_Assignment` data, so potentially important information such as detailed maintenance history, accident records, regional economic conditions or real–time market demand was not available and could not be incorporated into the models. In addition, outlier handling was implemented through removal of extreme observations, so part of the price variability in luxury model car may have been excluded rather than fully explained. Finally, all performance metrics were computed. Changes in consumer preferences, technology (for example, electric vehicles) and macroeconomic conditions may reduce model accuracy over time.

Appendix A.

The multiple linear regression model used in this study given in equation (1) shows the following graph :

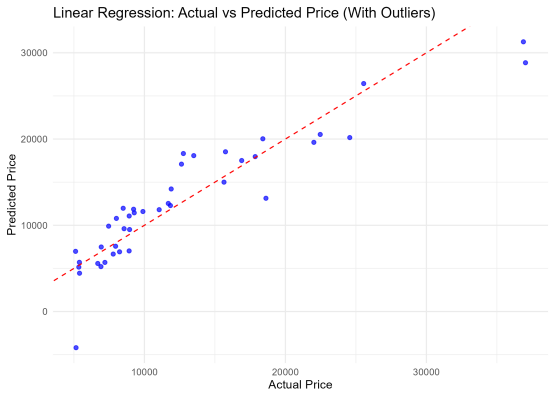

Figure 7: Linear Regression: Actual vs predicted price (with outliers)

Figure 8: Linear Regression: Actual vs predicted price (without outliers)

Comparison of Actual vs. Predicted Price Plots
Both plots indicate that the linear regression model captures a generally positive linear relationship between actual and predicted price but the model fits the data more reliably after removing outliers.

- **With Outliers :** For the cleaned dataset without outliers, Linear Regression follows the regressor line much more closely. This means the model can predict mid-range prices reasonably well, especially where most cars are located. However, for the highest prices the points still spread out, showing that a simple linear relationship cannot fully capture the jump from normal cars to premium models. The improvement compared to the full dataset highlights how sensitive Linear Regression is to extreme values.
- **Without Outliers :** In the second plot, the range of actual prices is narrower and the points cluster more tightly around the reference line, indicating reduced error variance and a more stable linear relationship between actual and predicted prices. The absence of extreme points implies that no single observation has excessive leverage, so the fitted line reflects the central bulk of the data better and would typically explain lower error metrics (for example, a smaller root mean squared error) than the model influenced by outliers. Outliers inflate prediction error, can distort the estimated slope and make the model appear less accurate, whereas removing them produces a cleaner plot where predictions align more closely with observed values across the studied range. However, deleting outliers reduces the sample size.

Appendix B.

The lasso regression model used in this study given in equation (2) shows the following graph :

Figure 9: Lasso Regression: Actual vs predicted price (without outliers)

Figure 10: Lasso Regression model with outliers

Comparison of Lasso Regression Actual vs. Predicted Price Plots
Both Lasso regression plots show a clear positive association between actual and predicted prices, and in both cases the predictions follow the reference line reasonably well, indicating that the regularized model captures the main linear pattern in the data.

- **With Outliers:** With outliers present, Lasso still shrinks coefficients but cannot completely fix the influence of extreme prices. The model becomes conservative and tends to pull predictions towards the middle range, which leads to under-prediction of luxury cars. The spread of points around the diagonal is clearly larger than in the non-outlier case. This explains why its error increases when we work with the original data.
- **Without Outliers:** Lasso Regression performs quite well on the dataset without outliers. Because it shrinks some coefficients toward zero, it removes a bit of noise and keeps only the strongest signals. The predicted points cluster closely around the diagonal for low and medium prices, and only a few high-price cars are under or overestimated. This shows that after removing extreme observations. A regularized linear model can provide stable and interpretable predictions.

Appendix C.

The Ridge regression model used in this study given in equation (3) shows the following graph ;

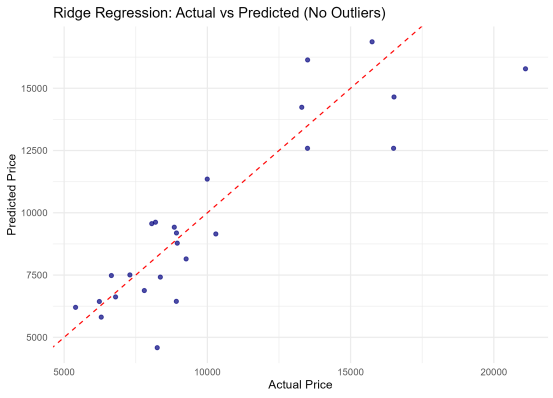

Figure 11: Ridge Regression: Actual vs predicted price (without outliers)

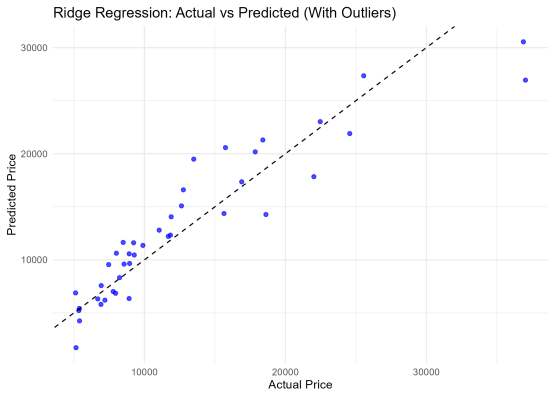

Figure 12: Ridge Regression: Actual vs predicted price (with outliers)

Comparison of Ridge Regression Actual vs. Predicted Price Plots
Both Ridge regression plots show a strong positive linear relationship between actual and predicted prices, indicating that the regularized model captures the underlying linear trend in the data.

- **With Outliers:** Ridge Regression handles multicollinearity better than plain Linear Regression but outliers still push the fitted values away from the diagonal at high prices. The model captures the main trend for normal cars but cannot fully adjust to the few very expensive ones. This behaviour is visible in the top-right corner, where predicted values are lower than the true prices.
- **Without Outliers:** Ridge Regression also gives a smooth and stable fit on the cleaned data. It keeps all predictors in the model but controls their size, which works well in the presence of multicollinearity. The scatter points stay close to the diagonal over almost the whole price range, with only small deviations at the very top. This suggests Ridge is a strong baseline model once the most extreme cars are removed.

Appendix D.

The support vector regression model used in this study given in equation (4) shows the following graph :

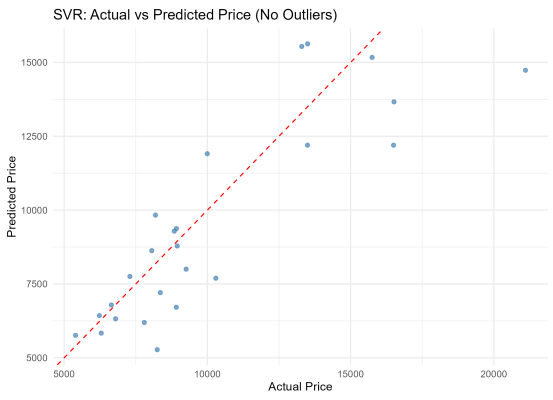

Figure 13: SVR: Actual vs predicted price (without outliers)

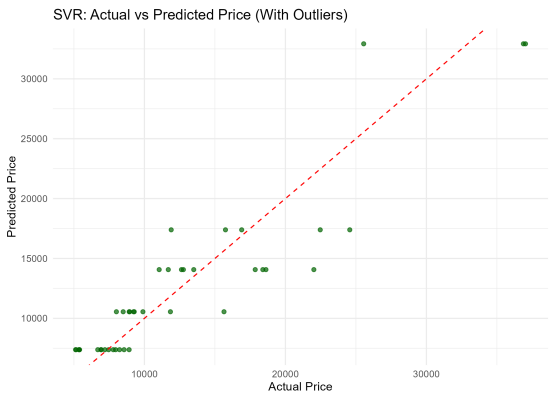

Figure 14: SVR: Actual vs predicted price (with outliers)

Comparison of SVR Actual vs. Predicted Price Plots
Both SVR plots show a positive association between actual and predicted prices but the model fitted on the dataset both with or without outliers is poor. The values are scattered so much from the regressor line.

- **With Outliers:** SVR remains relatively robust even when outliers are included. The points still follow the diagonal trend, although there is more spread compared to the clean dataset. Some luxury cars remain hard to predict accurately, but overall the model handles non-linear effects better than simple linear models. This explains why SVR keeps competitive performance in both scenarios.
- **Without Outliers:** Support Vector Regression handles the cleaned data quite well. Most points lie close to the diagonal line across the full price range, which indicates that the kernel function is capturing non-linear relationships between the features and price. Only a few high-price cars are off the line. Overall, SVR shows strong performance when the data is free from strong outliers.

Appendix E.

The decision tree model used in this study given in equation (5) shows the following graph :

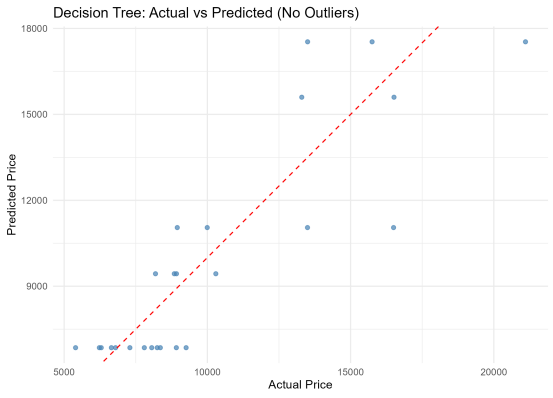

Figure 15: Decision Tree: Actual vs predicted price (without outliers)

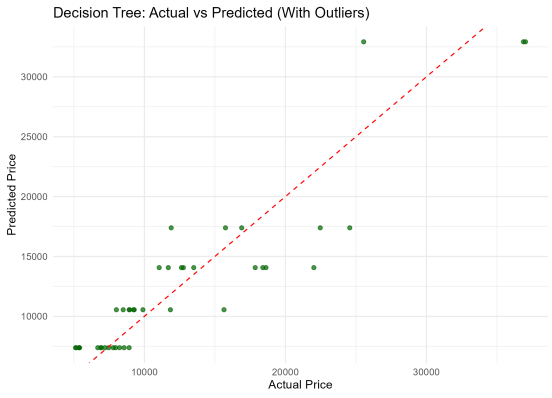

Figure 16: Decision Tree: Actual vs predicted price (with outliers)

Comparison of Decision Tree Actual vs. Predicted Price Plots
Both decision tree plots reveal a positive association between actual and predicted prices but the dispersion of points around the reference line indicates that tree-based predictions are less tightly aligned with the ideal line than the linear and kernel-based models, especially when outliers are present. [web:4][web:11]

- **With Outliers:** The Decision Tree model on the no-outlier data still produces step-like predictions, which show up as horizontal bands in the plot. Some groups of cars are predicted with the same price even though their actual prices differ. This is a typical behaviour of single trees and explains why the model has higher error and lower R^2 compared to more advanced methods such as Random Forest or GBM.
- **Without Outliers:** For the full dataset, the Decision Tree tries to fit the extreme prices by creating very specific branches. This leads to overfitting on a few observations and poorer generalisation on the rest of the data. The plot shows large vertical gaps from the diagonal, especially at higher price levels. This confirms that a single tree is too unstable for this regression task, and motivates moving to ensemble approaches like Random Forest and GBM.

Appendix F.

The random forest model used in this study given in equation (6) shows the following graph :

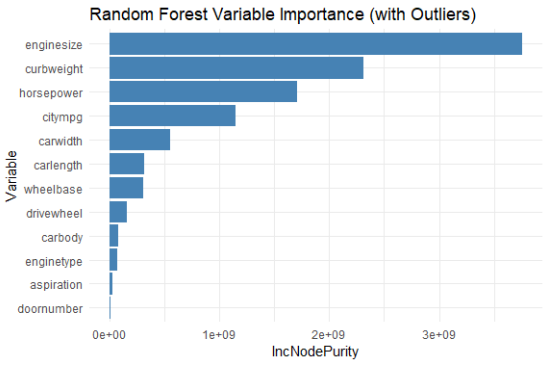

Figure 17: Random Forest Variable Importance (With Outliers)

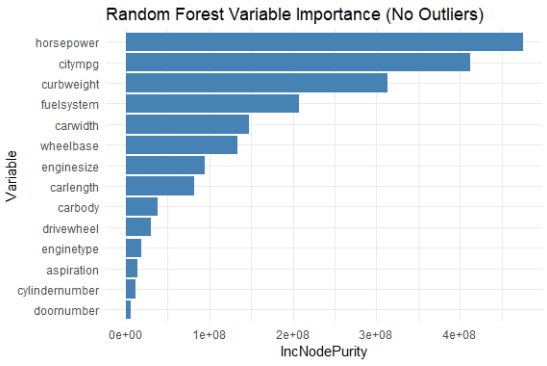

Figure 18: Random Forest Variable Importance (Without Outliers

Comparison of Random Forest Variable Importance Plots
The two random forest variable importance plots illustrate how the presence of outliers substantially changes which predictors appear most influential for car price.

- **No Outliers:** In the Random Forest model trained on the dataset without outliers, horsepower, citympg, and curbweight appear as the most important features, followed by fuelsystem, carwidth, and wheelbase. This matches what we saw in the correlations: car size, engine power, and fuel efficiency are key drivers of price. Because Random Forest is an ensemble of trees, it can capture non-linear relationships and interactions between features, and the importance scores give a practical ranking of which variables the model actually uses to make decisions
- **With Outliers:** When we keep the outliers, the importance ranking shifts slightly. Enginesize and curbweight become more dominant, because the extreme luxury cars usually have very large engines and heavy bodies and the model needs these features to separate them from regular cars. Horsepower and citympg also remain important. This plot shows how outliers can change the focus of the model and the algorithm pays more attention to variables that help explain the higher car prices.

Appendix G.

The K-N Neighbors model used in this study given in equation (7) shows the following graph :

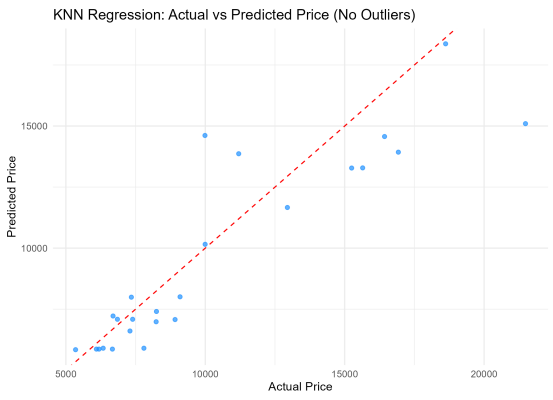

Figure 19: KNN Regression: Actual vs predicted price (without outliers)

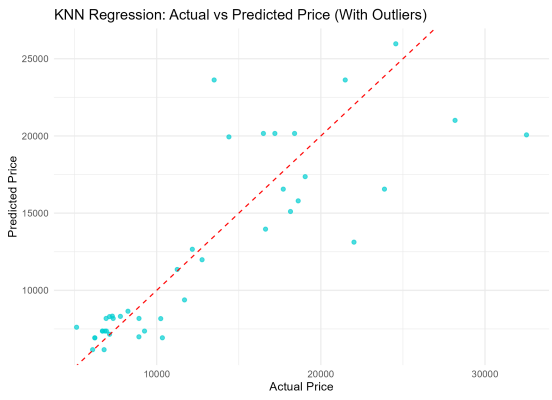

Figure 20: KNN Regression: Actual vs predicted price (with outliers)

Comparison of KNN Regression Actual vs. Predicted Price Plots
Both KNN regression plots show a positive association between actual and predicted prices, but the alignment with the reference line is noticeably better when outliers are removed, indicating that the KNN model is sensitive to extreme observations in the training data.

- **With Outliers:** KNN struggles the most in the presence of outliers. Because predictions are based on nearby points, extreme prices distort the neighbourhoods and pull some predictions too high or too low. In the scatter plot we see a wide spread of points, with several observations far from the diagonal line. This explains the weaker performance of KNN on the full dataset.
- **No Outliers:** For KNN on the no-outlier dataset, the points roughly follow the diagonal but are more spread out compared to Ridge or Lasso. Because KNN predicts price by averaging nearby cars, it works best in dense regions where we have many similar observations. In price areas with fewer neighbours, the predictions become less accurate. This behaviour is visible in the plot: some regions match well, while others show larger gaps from the diagonal.

Appendix H.

The gradient boosting regression model used in this study given in equation (8) shows the following graph :

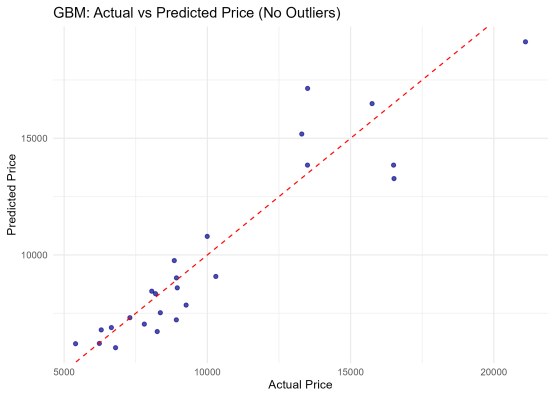

Figure 21: Gradient Boosting (GBR): Actual vs predicted price (without outliers)

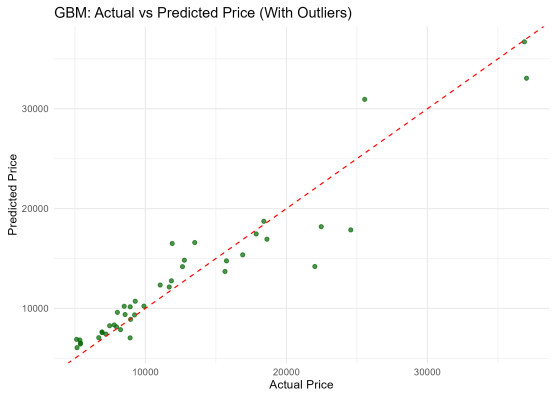

Figure 22: Gradient Boosting (GBR): Actual vs predicted price (with outliers)

Comparison of GBR Actual vs. Predicted Price Plots
Both GBR (Gradient Boosting Regression) plots display a strong positive relationship between actual and predicted prices, with points generally following the reference line, indicating that the boosted model captures the main structure in the data.

- **With Outliers:** In the full dataset with outliers, GBR still performs relatively well but the spread around the diagonal is larger than in the no-outlier case. The algorithm tries to also fit the extreme luxury cars, which can slightly hurt accuracy for the more common mid-range prices. This behaviour shows that even strong ensemble models are influenced by a few very unusual observations, and it justifies comparing results with and without those outliers.
- **No Outliers:** On the cleaned dataset, GBM follows the diagonal line very closely. The model gradually learns from residuals and is able to capture complex patterns between features and price. The points show only small errors across most of the price range, which matches its strong performance metrics. This makes GBM one of the best models when outliers are removed.

References

AV Automotive Research. New light-vehicle registrations in the Baltic States still show uneven tendencies in the third quarter of 2025, October 2025. URL `https://www.balticautoresearch.com/news/`. Analysis of new light-vehicle registrations (M1 N1) in Lithuania, Latvia, and Estonia based on data from Regitra, CSDD, and Transpordiamet.

Chuancan Chen, Lulu Hao, and Cong Xu. Comparative analysis of used car price evaluation models. In *AIP Conference Proceedings*, volume 1839, page 020165. AIP Publishing LLC, 2017.

Shen Gongqi, Wang Yansong, and Zhu Qiang. New model for residual value prediction of the used car based on bp neural network and nonlinear curve fit. In *2011 third international conference on measuring technology and mechatronics automation*, volume 2, pages 682–685. IEEE, 2011.

Rupesh Gupta, Avinash Sharma, Vatsala Anand, and Sheifali Gupta. Automobile price prediction using regression models. In *2022 International Conference on Inventive Computation Technologies (ICICT)*, pages 410–416. IEEE, 2022.

Nitis Monburinon, Prajak Chertchom, Thongchai Kaewkiriya, Suwat Rungpheung, Sabir Buya, and Pitchayakit Boonpou. Prediction of prices for used car by using regression models. In *2018 5th international conference on business and industrial research (ICBIR)*, pages 115–119. IEEE, 2018.

Sumeyra Muti and Kazım Yıldız. Using linear regression for used car price prediction. *International Journal of Computational and Experimental Science and Engineering*, 9 (1):11–16, 2023.

U Nishitha, Revanth Kandimalla, C Jyotsna, et al. Automobile price prediction using machine learning with data visualization. In *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–7. IEEE, 2023.

Sameerchand Pudaruth. Predicting the price of used cars using machine learning techniques. *Int. J. Inf. Comput. Technol*, 4(7):753–764, 2014.

Yusuke Soejima and Hideo Hirose. Auction price estimation for used cars by regression methods (competition 1). In *Proceedings of the symposium of Japanese Society of Computational Statistics 25*, pages 9–12. Japanese Society of Computational Statistics, 2011.

Pattabiraman Venkatasubbu and Mukkesh Ganesh. Used cars price prediction using supervised learning techniques. *Int. J. Eng. Adv. Technol.(IJEAT)*, 9(1S3), 2019.