Vehicles Price Prediction with Multiple Linear Regression

Murpys Mendez

Abstract

This paper explores the underlying principles of linear regression and demonstrates its application using a real-world dataset. Multiple linear regression(MLR) is a linear regression model encompassing several predictors.

We use the dataset “Vehicles Sales Data”, from kaggle.com to fit an MLR for predicting car prices based on various attributes, such as make, model, year, mileage, and transmission type. Using R, including libraries such as `stats` (Base R), `base` (Base R), caret, ggplot2, and dplyr; we implement a multiple linear regression model to analyze the relationships between these features and retail price. The dataset is preprocessed to handle missing values and categorical variables, and we assess model performance through statistical metrics, including R-squared and root mean squared error (RMSE). The model provides insights into the factors influencing car prices and helps us understand the dynamics of the application of MLR in Machine Learning.

1 Introduction

The used car industry is a massive and rapidly growing sector, representing billions of dollars in annual sales worldwide. As one of the largest segments of the automotive market, it plays a critical role in meeting consumer demand for affordable and reliable transportation

With the vast amount of data available, Machine Learning (ML) is being used as an effective way to build predictive models that can estimate the Retail Price of a used car.

The vehicle sales industry widely uses machine learning to enhance decision-making and optimize processes. Accurate price predictions are critical for manufacturers, dealerships, and consumers, influencing inventory management, marketing strategies, and customer satisfaction. Factors such as vehicle age, mileage, brand reputation, and market trends contribute to the complexity of pricing decisions, making data-driven approaches essential. By analyzing large datasets, stakeholders can uncover patterns and gain insights that help set competitive prices, forecast demand, and improve operational efficiency in a rapidly evolving market.

This work uses multiple linear regression to examine the relationship between car features and their influence on pricing. This approach helps quantify the impact of attributes such as mileage,

age, brand, and condition on a vehicle's market value. This study focuses on understanding the application of linear regression and exploring the general steps involved in building and evaluating machine learning models.

2 Methodology

The implementation procedure can be understood as five basic steps: data collection, exploratory data analysis(EDA) and pre-processing, model training, prediction, and evaluation.

2.1 The data

The dataset contains 558837 observations of 14 variables, compiling thousands of real-world transactions in the used vehicle market and describing characteristics of the vehicles sold. Table 1 contains a description of the features.

Feature	Description	Type
year	vehicle manufacturing year	numeric int
make	vehicle brand	character
model	specific model within a brand	character
trim	version of a model	character
body	portion of the vehicle mounted on the frame	character
transmission	mechanical system that transfers power to the wheels	character
VIN	vehicle identification number	character
state	state where the sale takes place	character
condition	quantified status of the vehicle	numeric int
odometer	number of miles driven	numeric
color	exterior color	character
interior	interior color	character
seller	seller	character
mmr	manheim market report index	numeric
sellingprice	retail price	numeric
saledate	date of the sale	character

Table 1: Features in the dataset “Vehicles Sales Data”

The data was imported from Kaggle.com in CSV format.

2.2 EDA

During Exploratory Data Analysis, the missing values in their different variants were removed. At this stage in ML is also important to process duplicated values. The categorical variables were found to have more levels than necessary due to different font case. These levels were regularized and merged.

2.2.1 Feature Selection

EDA is a necessary step in ML to gain a better understanding of the dataset. The impact of the features on the target variable was assessed aided by graphics built with functions from the ggplopt2 library. Graphic 1 is an example of the variation of selling price with the year as an input.

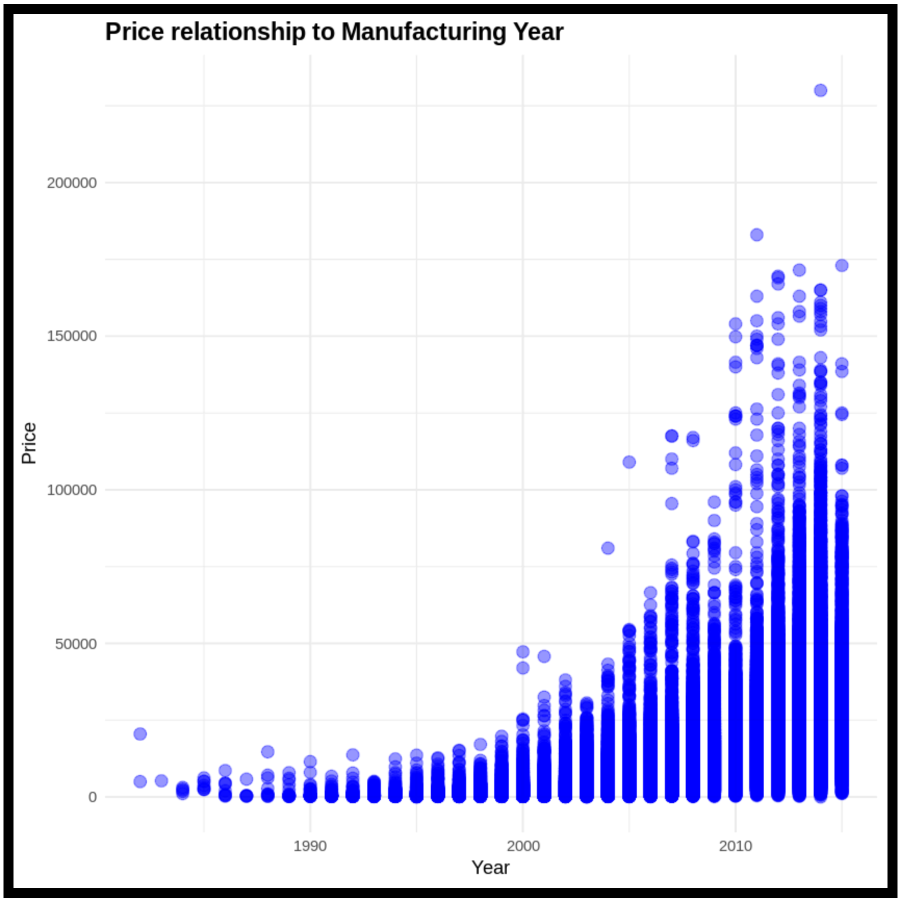


Figure 1: Price vs Year Plot

Features that do not impact the behavior of the target variable, were removed from the dataset. Unrealistic observations, or observations with incorrect values were also removed.

2.2.2 Encoding

The categorical variables were encoded before training the model. We used frequency encoding, a machine learning technique that replaces categorical data with a numerical value representing how often the category appears in the dataset (Florian Pargent, et al. 2022)

$$Frequency\_Encoding(level) = \frac{Count\ of\ observations\ per\ \ level}{Total\ number\ of\ observations} \qquad (1)$$

2.3 Modeling

MLR is a statistical technique to predict the result of an answer variable, using a number of explanatory variables. The object of (MLR) is to model the linear relationship between the independent variables x and dependent variable y (Dastan Hussen Maulud1 & Adnan Mohsin Abdulazeez. 2020)

The equation for Linear Regression is as shown below:

$$y = b_0 + b_1 x_1 + b_2 x_2 +...+ b_n x_n \qquad (2)$$

where
y : dependent variable
x_i: independent variables or predictors
b_0: intercept or bias
b_i: coefficients or weights

The equation can be written in matrix form as:

$$y = X\beta + \epsilon \qquad (3)$$

Where:

y is the vector of observed values (the dependent variable).

X is the matrix of input features, where each row represents an observation and each column represents a feature.

β is the vector of coefficients (parameters) to be estimated.

ϵ is the vector of errors (residuals).

The loss function for multiple linear regression is typically the **Mean Squared Error (MSE)**, which measures the average squared difference between the observed and predicted values:

$$J(\beta) = \frac{1}{2m}\sum_{i=1}^{m} (y_i - \widehat{y}_i)^2 \tag{4}$$

where:

m is the number of observations.

y_i is the actual value for the i-th observation.

$\widehat{y}_i$ is the predicted value for the i-th observation.

2.3.1 Ordinary Least Squares Solution (OLS)

The OLS solution is used to find the optimal values of β,

$$\widehat{\beta} = (X^T X)^{-1} X^T y \tag{5}$$

This formula provides a closed-form solution to the linear regression problem.

Where:

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_m \end{bmatrix}, X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1m} \\ 1 & x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix}, Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

2.3.2 Predictions

Once the model is trained (coefficients β are estimated), predictions for new data are made using the following equation:

$$\widehat{y} = X\widehat{\beta} \tag{6}$$

Where:

- $\hat{y}$ is the predicted values.
- X is the new data for which predictions are needed.

3 Experiments

The experiments were run in Google Colab, with an R runtime backed by a System RAM of 12.7 GB and a Disk size of 107.7 GB.

After the data cleaning and exploration, various model versions were created using different combinations of features. The performance was assessed in every case using R square, Mean Absolute Error (MAE), and Root Mean Square Error (RMSE)

R square indicates how well the model explains the variability in the target variable. Specifically the proportion of the variance in the dependent variable that is explained by the predictors. R square is calculated as shown below:

$$R^2 = 1 - \frac{SS_{residual}}{SS_{total}} \tag{7}$$

RMSE is a metric used to measure the accuracy of a regression model. Its formula is written as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_{actual,i} - y_{predicted,i})^2} \tag{8}$$

MAE, is a common metric that measures the average magnitude of errors. Equation 9 shows the formula of MAE.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right| \tag{9}$$

During EDA we noticed a high correlation between the feature “mmr” and the target variable (see Figure 2) consequently the model was evaluated both including and excluding the feature, obtaining the results illustrated in Table 2.

Model	MAE	RMSE	R^2
model 1	5455.8220	7440.9561	0.4309
model 2	1080.7439	1703.1344	0.9682
model 3	1041.4026	1634.4755	0.9707

Table 2: Metrics Results

Model 1:

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4 + b_5x_5 + b_6x_6 + b_7x_7 + b_8x_8 + b_9x_9 + b_{10}x_{10} + b_{11}x_{11} + b_{12}x_{12}$$

Model 2:

$$\hat{y} = b_0 + b_{13}x_{13}$$

Model 3

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4 + b_5x_5 + b_6x_6 + b_7x_7 + b_8x_8 + b_9x_9 + b_{10}x_{10} + b_{11}x_{11} + b_{12}x_{12} + b_{13}x_{13}$$

Where:

x_1:year, x_2:transmission, x_3:condition, x_4: odometer, x_5:make, x_6:model, x_7:trim, x_8:state, x_9:color, x_{10}:interior, x_{11}:seller, x_{12}:body, x_{13}:mmr

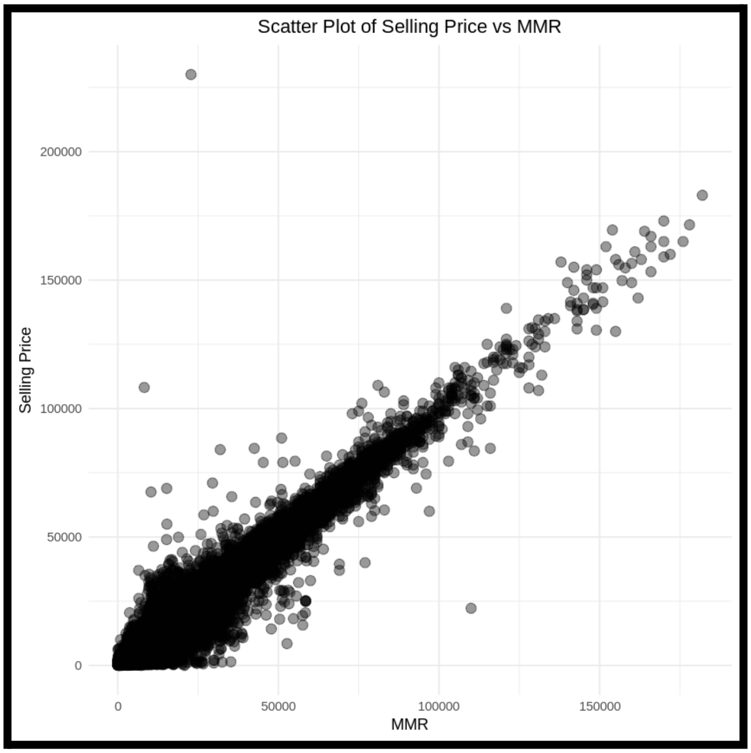


Figure 2: Plot of Selling Price vs mmr

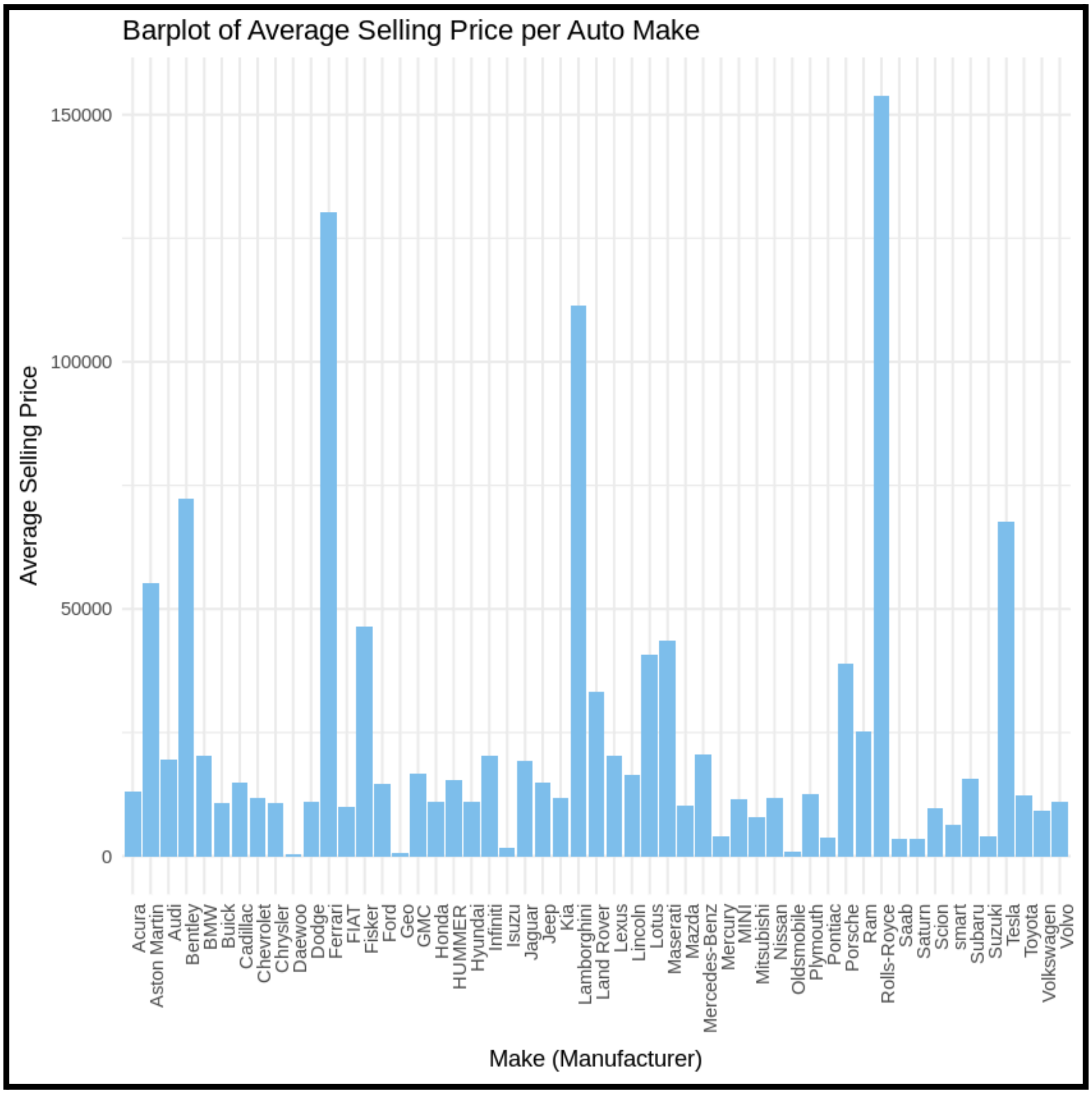
4 Conclusions

In this project we have applied Multiple Linear Regression to an automobile dataset in order to predict selling price. The model was studied in different variants, and the performance was evaluated and presented using RMSE, MAE, and R square. In this case, we learned that most of the variability on the target variable can be attributed to one predictor. As an improvement of the data, we suggest revisiting the indicators included in the dataset, as in real world scenarios the market index (mmr) is not always available. Including predictors that have a greater effect on the target variable, would lead to more balanced and universal models. The project served to increase our grasp on the general process of machine learning workflow and Linear Regression in particular. In the future, we would consider applying other types of regression models

References

[1]	Maulud, Dastan, and Mohsin Abdulazeez, Adnan. “ Review on Linear Regression Comprehensive in Machine Learning”. Journal of Applied Science and Technology Trends, vol. 1, no. 2, Dec. 2020, pp. 140-7
[2]	S. Pathak, I. Mishra and A. Swetapadma. "An Assessment of Decision Tree based Classification and Regression Algorithms". *2018 3rd International Conference on Inventive Computation Technologies (ICICT)*, Coimbatore, India, 2018, pp. 92-95, doi: 10.1109/ICICT43934.2018.9034296.
[3]	Hankar, M., Birjali, M., Beni-Hssane, A. “Machine Learning Modeling to Estimate Used Car Prices.” Innovations in Smart Cities Applications Volume 6. SCA 2022. Lecture Notes in Networks and Systems, vol 629. Springer, Cham. (2023).https://doi.org/10.1007/978-3-031-26852-6_49
[4]	S. Shaprapawad, P. Borugadda and N. Koshika, "Car Price Prediction:An Application of Machine Learning," *2023 International Conference on Inventive Computation Technologies (ICICT)*, Lalitpur, Nepal, 2023, pp. 242-248, doi: 10.1109/ICICT57646.2023.10134142.
[5]	Claudio Gambella, et al. "Optimization problems for machine learning: A survey". European Journal of Operational Research. Volume 290, Issue 3. 2021. Pages 807-828, ISSN 0377-2217. https://doi.org/10.1016/j.ejor.2020.08.045.
[6]	Pargent, F., Pfisterer, F., Thomas, J. *et al.* Regularized target encoding outperforms traditional methods in supervised machine learning with high cardinality features. *Comput Stat* **37**, 2671–2692 (2022). https://doi.org/10.1007/s00180-022-01207-6

Appendix A
EDA Graph sample and Code link



Google Colab

Car_prices.ipynb