National Higher School Of Mathematics

Used Car Price Prediction:
A Multiple Linear Regression Approach

Regression Models

Marouf Haider

2023/2024

Contents

Abstract

This study aims to build a multiple linear regression model for used cars price prediction based on various features such as age of car, kilometer driven, engine etc. The model aims to provide accurate price predictions, facilitating informed decision-making for buyers and sellers in the used cars market.

Keywords

Used cars, Car price prediction, Multiple linear regression, Regression model, data analysis.

1 Introduction

With more and more people buying and selling cars, there's a big interest in knowing how much a used car might cost, this is especially true in places where people might not have a lot of money to spend on a new car, so they look to buy used ones instead. The main aim of this study is using multiple linear regression to figure out the price of a used car by looking at things like how old it is, how much it's been driven or how many persons owned this car before, alongside dimensions and other power metrics, in order to help help make informed purchasing decisions.

2 The Data

2.1 Dataset description

The dataset used in this study contains information about used cars, according to the publisher, it was gathered from CarDekho[1] website. It is publicaly available on kaggle, there is no information about to which period of time the data belongs. Dataset size is: 2059 rows × 20 columns.

For more information about the dataset follow this link.
The following table shows details of features within this dataset:

Column	Data Type	Unique Values	Description
Make	object	33	The manufacturer of the car
Model	object	1050	Model of the vehicle
Price	int64	619	Price in rupees
Year	int64	22	Year of manufacture
Kilometer	int64	847	Kilometers driven in KM
Fuel Type	object	9	
Transmission	object	2	Manual or Automatic
Location	object	77	Location of sale
Color	object	17	
Owner	object	6	Number of previous owners
Seller Type	object	3	individual, dealer or coroporate
Engine	object	108	Engine capacity in cc
Max Power	object	335	
Max Torque	object	290	
Drivetrain	object	3	FWD, RWD or AWD
Length	float64	248	
Width	float64	170	
Height	float64	196	
Seating Capacity	float64	6	Number of seats
Fuel Tank Capacity	float64	55	capacity in liters

Table 1: Description of Columns in the Dataset

2.2 Data processing

Before starting, we made sure to clean up the data by getting rid of any mistakes, missing or irrelevant information. We're using some data mining techniques to help us with this and since this study main goal is not about data mining and analysis we are going to display only major changes to perform on the dataset:

- **Handling Missing Values:** Any missing values in the dataset were handled by removal.
- **Encoding Categorical Variables:** The categorical variables 'Drivetrain', 'Fuel Type', 'Transmission', 'Owner' was encoded using one-hot encoding or ordinal encoding.
- **Normalization:** All variables were standardized to ensure they are on a comparable scale.

[1] CarDekho is India's leading car search venture

- **First Feature Selection**: The following features were considered irrelevant to this study: 'Color', 'Make', Model' and 'Location','Seller Type' ,.
- **Removing Outliers**: Outliers in the '**Price**', '**Year**', **Kilometer** features were removed using the Interquartile Range (IQR) method.

Summary:

- **Data size**: 1123 rows x 19 columns.
- '**Seller Type**' = Individual , '**Fuel Type**' = 'Petrol' or 'Diesel'.
- '**Owner**' = 1, 2 or 3 , '**Seating Capacity** = 4, 5, 6, 7, 8.
- '**Drivetrain**': FWD, RWD, AWD. '**Transsmission**' = 'Manual' or 'Automatic'.
- '**Price**' range:
- '**Year**' range: [2008 : 2022].
- '**Kilometer**' range: [600 : 130000].
- '**Fuel Tank Capacity**' range: [27 : 67].

Note: Python's library dtale was used to process data.

2.3 Data Visualisation & Analysis

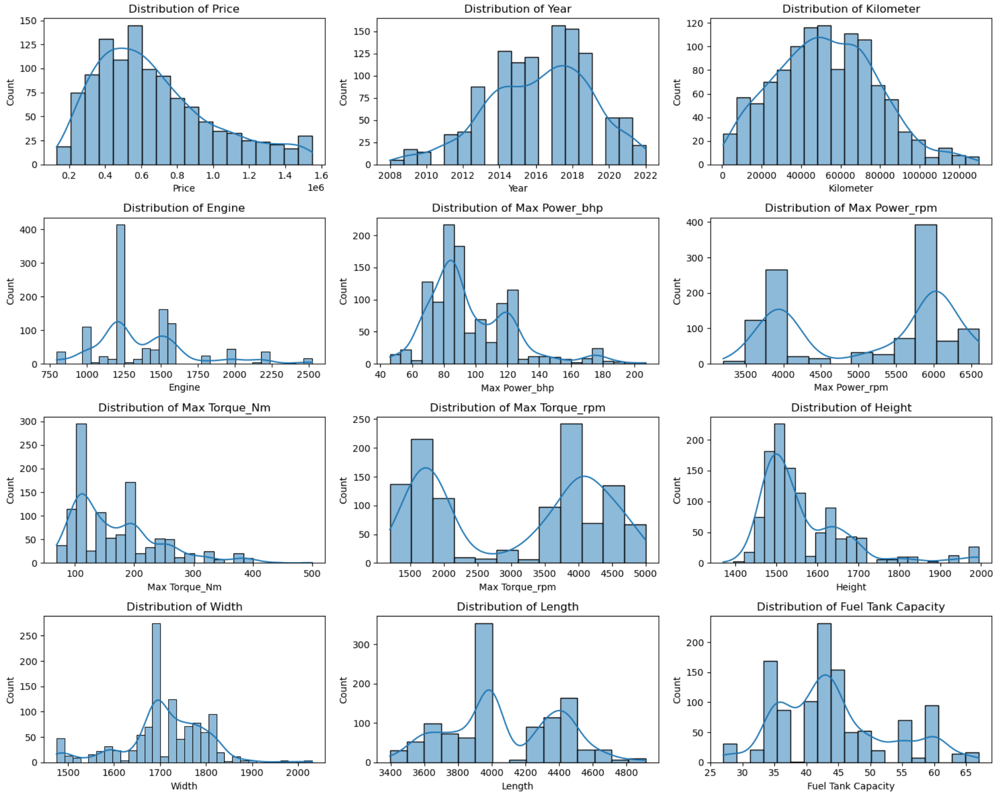


Figure 1: Numerical Features Distributions

Observations: Figure 1 shows that:

- The following features are right skewed : '**Price**', '**Max Power (bhp)**', '**Max Torque (Nm)**', '**Height**', '**Fuel Tank Capacity**'.

- '**Year**' is left skewed .

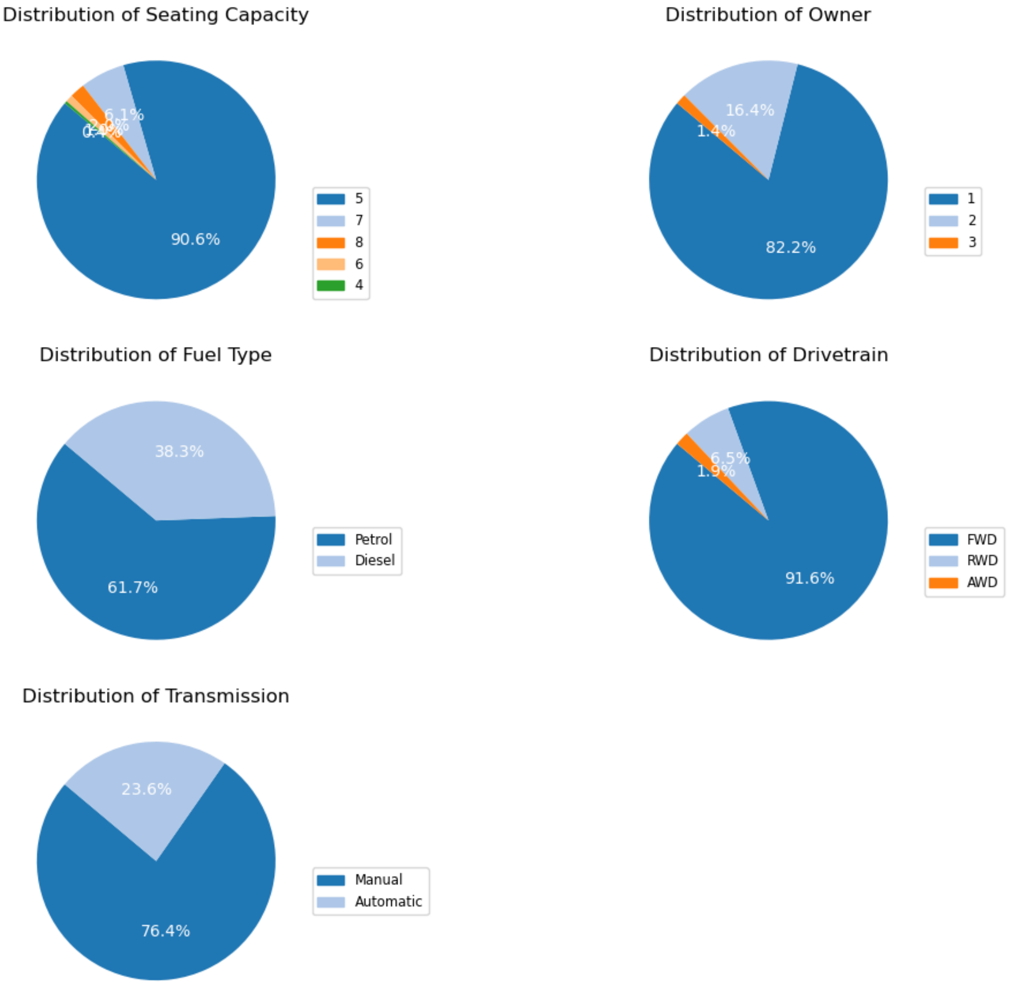


Figure 2: Categorical Features Distributions

Observations: From Figure 2 it can be concluded that:

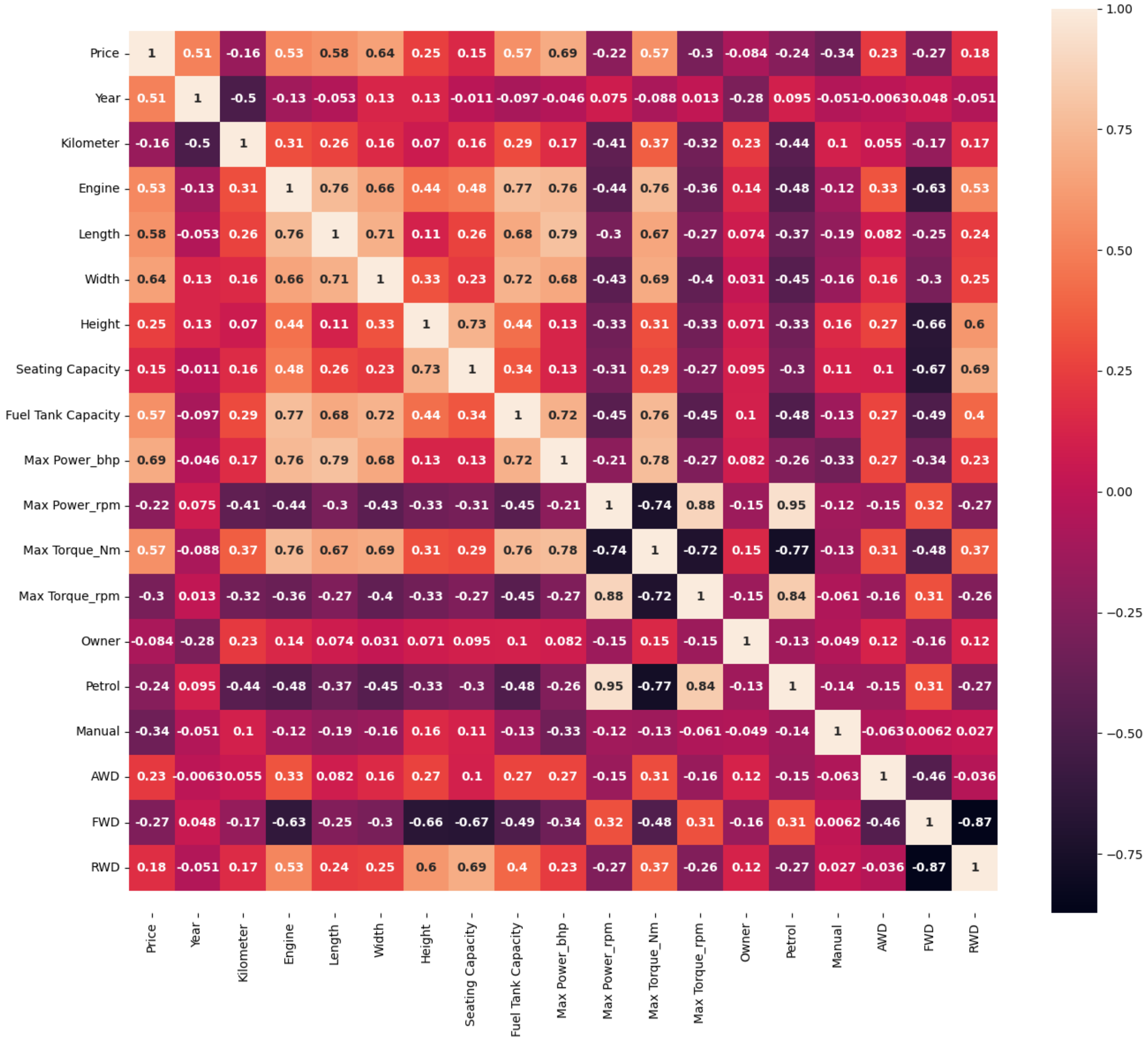
- Most of cars are '**Front Wheels Driving**' and '**Manual**' transmission .

- **'Fuel type'** of cars present in data is balanced between '**Petrol**' and '**Diesel**'.

- 5 seats cars are dominant within the dataset as well as cars with unique previous owner.

Summary

The dataset shows unbalanced distribution of some features which necessarily will affect the overall accuracy of the final model, equivalently, it causes some limitations when using the model in predictions. This problem can be solved by using some resampling techniques.



	Price	Year	Kilometer	Engine	Length	Width	Height	Seating Capacity	Fuel Tank Capacity	Max Power_bhp	Max Power_rpm	Max Torque_Nm	Max Torque_rpm	Owner	Petrol	Manual	AWD	FWD	RWD
Price	1	0.51	-0.16	0.53	0.58	0.64	0.25	0.15	0.57	0.69	-0.22	0.57	-0.3	-0.084	-0.24	-0.34	0.23	-0.27	0.18
Year	0.51	1	-0.5	-0.13	-0.053	0.13	0.13	-0.011	-0.097	-0.046	0.075	-0.088	0.013	-0.28	0.095	-0.051	-0.0063	0.048	-0.051
Kilometer	-0.16	-0.5	1	0.31	0.26	0.16	0.07	0.16	0.29	0.17	-0.41	0.37	-0.32	0.23	-0.44	0.1	0.055	-0.17	0.17
Engine	0.53	-0.13	0.31	1	0.76	0.66	0.44	0.48	0.77	0.76	-0.44	0.76	-0.36	0.14	-0.48	-0.12	0.33	-0.63	0.53
Length	0.58	-0.053	0.26	0.76	1	0.71	0.11	0.26	0.68	0.79	-0.3	0.67	-0.27	0.074	-0.37	-0.19	0.082	-0.25	0.24
Width	0.64	0.13	0.16	0.66	0.71	1	0.33	0.23	0.72	0.68	-0.43	0.69	-0.4	0.031	-0.45	-0.16	0.16	-0.3	0.25
Height	0.25	0.13	0.07	0.44	0.11	0.33	1	0.73	0.44	0.13	-0.33	0.31	-0.33	0.071	-0.33	0.16	0.27	-0.66	0.6
Seating Capacity	0.15	-0.011	0.16	0.48	0.26	0.23	0.73	1	0.34	0.13	-0.31	0.29	-0.27	0.095	-0.3	0.11	0.1	-0.67	0.69
Fuel Tank Capacity	0.57	-0.097	0.29	0.77	0.68	0.72	0.44	0.34	1	0.72	-0.45	0.76	-0.45	0.1	-0.48	-0.13	0.27	-0.49	0.4
Max Power_bhp	0.69	-0.046	0.17	0.76	0.79	0.68	0.13	0.13	0.72	1	-0.21	0.78	-0.27	0.082	-0.26	-0.33	0.27	-0.34	0.23
Max Power_rpm	-0.22	0.075	-0.41	-0.44	-0.3	-0.43	-0.33	-0.31	-0.45	-0.21	1	-0.74	0.88	-0.15	0.95	-0.12	-0.15	0.32	-0.27
Max Torque_Nm	0.57	-0.088	0.37	0.76	0.67	0.69	0.31	0.29	0.76	0.78	-0.74	1	-0.72	0.15	-0.77	-0.13	0.31	-0.48	0.37
Max Torque_rpm	-0.3	0.013	-0.32	-0.36	-0.27	-0.4	-0.33	-0.27	-0.45	-0.27	0.88	-0.72	1	-0.15	0.84	-0.061	-0.16	0.31	-0.26
Owner	-0.084	-0.28	0.23	0.14	0.074	0.031	0.071	0.095	0.1	0.082	-0.15	0.15	-0.15	1	-0.13	-0.049	0.12	-0.16	0.12
Petrol	-0.24	0.095	-0.44	-0.48	-0.37	-0.45	-0.33	-0.3	-0.48	-0.26	0.95	-0.77	0.84	-0.13	1	-0.14	-0.15	0.31	-0.27
Manual	-0.34	-0.051	0.1	-0.12	-0.19	-0.16	0.16	0.11	-0.13	-0.33	-0.12	-0.13	-0.061	-0.049	-0.14	1	-0.063	0.0062	0.027
AWD	0.23	-0.0063	0.055	0.33	0.082	0.16	0.27	0.1	0.27	0.27	-0.15	0.31	-0.16	0.12	-0.15	-0.063	1	-0.46	-0.036
FWD	-0.27	0.048	-0.17	-0.63	-0.25	-0.3	-0.66	-0.67	-0.49	-0.34	0.32	-0.48	0.31	-0.16	0.31	0.0062	-0.46	1	-0.87
RWD	0.18	-0.051	0.17	0.53	0.24	0.25	0.6	0.69	0.4	0.23	-0.27	0.37	-0.26	0.12	-0.27	0.027	-0.036	-0.87	1

Figure 3: Heat-map correlations matrix

From the next Figure 3 it can be seen that:

- **Price** is strongly correlated with vehicle dimensions (length, width), engine size, and performance metrics (horsepower, torque).
- **Engine size** is a central factor, strongly linked to several performance characteristics such as fuel tank capacity, horsepower, and torque.
- **Vehicle dimensions** (length and width) are positively correlated with price and engine size, while width and height have an inverse relationship.

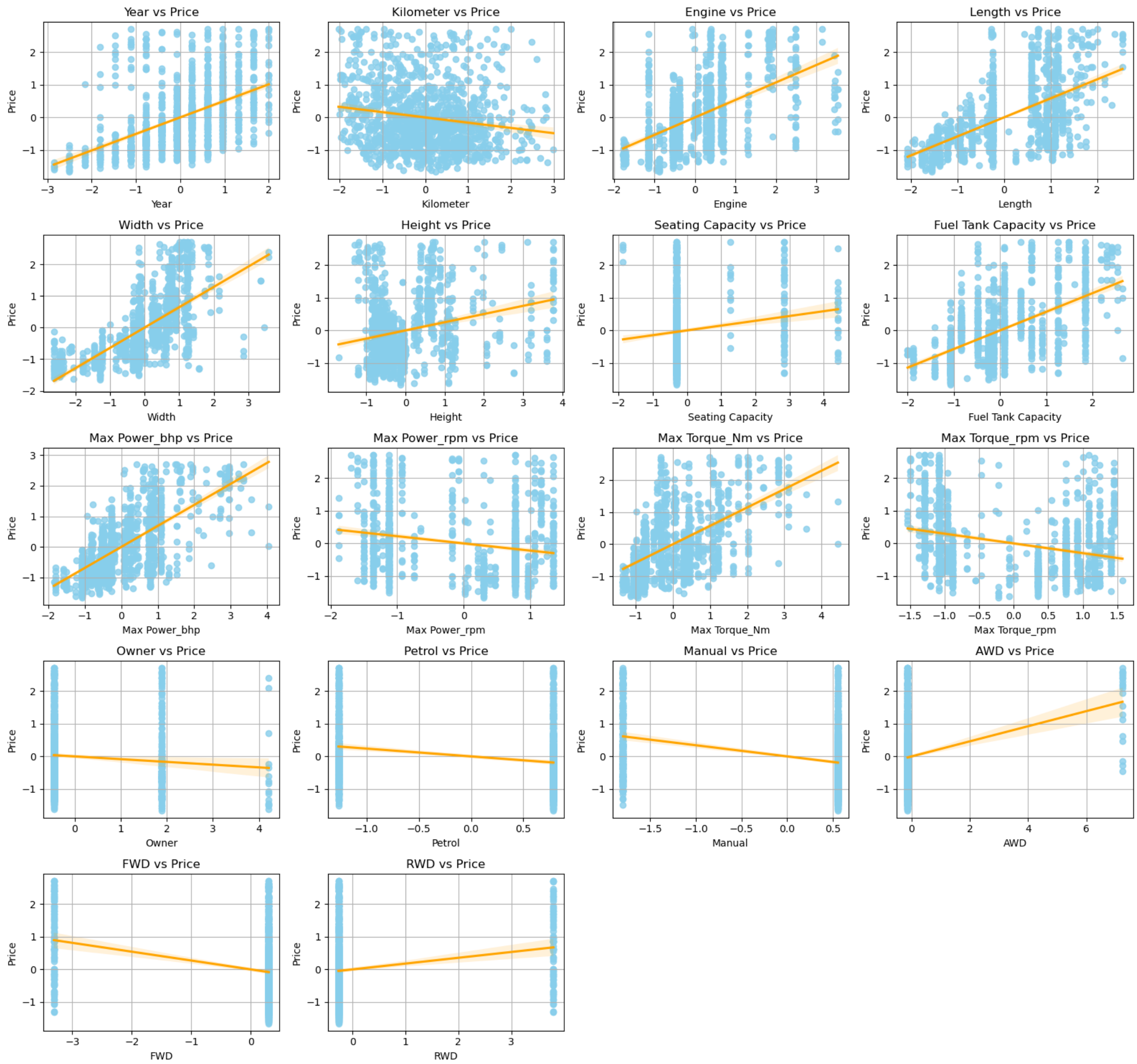


Figure 4: Scatter plots of each feature against the price + regression line

3 Model building

3.1 General Model

Under the theme of used car price prediction and using the dataset presented at the beginning, the general form of the multiple linear regression model is given by:

$\text{Price} = \beta_0 + \beta_1\textbf{Year} + \beta_2\textbf{Kilometer} + \beta_3\textbf{Engine} + \beta_4\textbf{Max Power} + \beta_5\textbf{Max Torque} + \beta_6\textbf{Length} + \beta_7\textbf{Width} + \beta_8\textbf{Height} + \beta_9\textbf{Fuel Tank Capacity} + \beta_{10}\textbf{Petrol} + \beta_{11}\textbf{FWD} + \beta_{12}\textbf{RWD} + \beta_{13}\textbf{Manual} + \beta_{14}\textbf{Seating Capacity} + \epsilon$

where β_0 is the intercept, $\beta_1, \beta_2, \ldots, \beta_{16}$ are the coefficients of the features, and ϵ is the error term.

Software Output:

Mean Squared Error: 0.24044511679937985

R-squared: 0.7835218769245862

Feature	Coefficient
Year	0.476305
Kilometer	-0.091514
Engine	0.078275
Length	0.091746
Width	0.054981
Height	0.042429
Seating Capacity	-0.058567
Fuel Tank Capacity	0.131817

Feature	Coefficient
Max Power_rpm	0.039954
Max Torque_Nm	-0.033661
Max Torque_rpm	-0.135396
Owner	-0.007342
Petrol	-0.008547
Manual	-0.119359
AWD	0.023468
FWD	-0.005175
RWD	-0.007082

Table 2: Initial model coefficients

3.2 Forward stepwise regression

This part focuses on constructing a reliable regression model by employing the forward stepwise regression method. Forward stepwise regression is a variable selection technique where predictors are added to the model one at a time based on their ability to improve the model's fit. Method main steps are :

1. **Initial Model**: We start with a null model containing no predictors.
2. **Forward Selection**
   (a) **Select Predictor:** We evaluate each predictor individually and choose the one that yields the highest increase in model fit, as measured by a chosen criterion such as adjusted R-squared, AIC (Akaike Information Criterion), or BIC (Bayesian Information Criterion).
   (b) **Add Predictor:** The selected predictor is added to the model.
   (c) **Model Fit Assessment:** We assess the overall fit of the model using the chosen criterion.
   (d) **Iterate:** Steps a-c are repeated iteratively until adding more predictors does not significantly improve the chosen criterion.
3. **Final Model Selection**
   (a) **Final Model Evaluation:** After the addition of each predictor, we evaluate the model's performance using appropriate metrics (e.g., R-squared, RMSE, etc.) and diagnostic plots to ensure the model assumptions are met.
   (b) **Model Simplification:** If necessary, we may simplify the model by removing non-significant predictors or those causing multicollinearity issues.
   (c) **Cross-Validation:** To validate the final model's performance, we use techniques such as k-fold cross-validation or holdout validation on a separate test dataset.

OLS Regression Results

```
Selected feature indices: (0, 1, 3, 7, 8, 11, 14, 15)
Selected feature names: (’Year’, ’Kilometer’, ’Length’, ’Fuel Tank Capacity’,
 ’Max Power_bhp’, ’Max Torque_rpm’, ’Manual’, ’AWD’)
R^2 score on test set: 0.7891455346052128
Mean Squared Error on test set: 0.23419884577370711
                            OLS Regression Results
==============================================================================
Dep. Variable:                  Price   R-squared:                       0.818
Model:                            OLS   Adj. R-squared:                  0.817
Method:                 Least Squares   F-statistic:                     501.1
Date:                Wed, 22 May 2024   Prob (F-statistic):          2.96e-323
Time:                        14:32:53   Log-Likelihood:                -492.89
No. Observations:                 898   AIC:                             1004.
Df Residuals:                     889   BIC:                             1047.
Df Model:                           8
Covariance Type:            nonrobust
======================================================================================
                         coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------
const                 -0.0118      0.014     -0.842      0.400      -0.039       0.016
Year                   0.4915      0.017     29.197      0.000       0.458       0.525
Kilometer             -0.0936      0.018     -5.205      0.000      -0.129      -0.058
Length                 0.1088      0.025      4.297      0.000       0.059       0.159
Fuel Tank Capacity     0.1788      0.023      7.794      0.000       0.134       0.224
Max Power_bhp          0.4467      0.027     16.280      0.000       0.393       0.501
Max Torque_rpm        -0.0986      0.017     -5.957      0.000      -0.131      -0.066
Manual                -0.1168      0.015     -7.636      0.000      -0.147      -0.087
AWD                    0.0307      0.015      2.001      0.046       0.001       0.061
==============================================================================
Omnibus:                       39.041   Durbin-Watson:                   1.940
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               92.379
Skew:                           0.208   Prob(JB):                     8.71e-21
Kurtosis:                       4.515   Cond. No.                         4.20
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

4 Model fitness

The provided OLS regression results can be interpreted as follows:

Coefficients and Statistical Significance

- **const**: -0.0118 (p=0.400). The constant term is not statistically significant, indicating that when all other variables are zero, the price prediction is not meaningfully different from zero.

- **Year**: 0.4915 (p=0.000). This positive and significant coefficient indicates that newer cars are associated with higher prices.

- **Kilometer**: -0.0936 (p=0.000). This negative and significant coefficient suggests that higher mileage reduces the car’s price.

- **Length**: 0.1088 (p=0.000). This positive and significant coefficient indicates that longer cars tend to be more expensive.

- **Fuel Tank Capacity**: 0.1788 (p=0.000). This positive and significant coefficient suggests that cars with larger fuel tanks are priced higher.

- **Max Power_bhp**: 0.4467 (p=0.000). This positive and significant coefficient indicates that cars with higher horsepower have higher prices.

- **Max Torque_rpm**: -0.0986 (p=0.000). This negative and significant coefficient suggests that cars with higher torque RPMs have lower prices.

- **Manual**: -0.1168 (p=0.000). This negative and significant coefficient indicates that manual transmission cars are less expensive compared to automatic ones.

- **AWD**: 0.0307 (p=0.046). This positive and significant coefficient indicates that all-wheel drive (AWD) cars are slightly more expensive than their two-wheel drive counterparts.

4.1 Coefficients interpretation

The coefficients represent the change in the dependent variable (Price) for a one-unit change in the predictor variable, holding all other variables constant.

- **Max_Power_bhp**: For each additional unit of max power in bhp, the price increases by 0.4064 units, holding other factors constant.

- **Year**: For each additional year, the price increases by 0.4854 units, holding other factors constant.

- **Fuel_Tank_Capacity**: For each additional unit of fuel tank capacity, the price increases by 0.1698 units, holding other factors constant.

- **Manual Transmission (Manual)**: Cars with manual transmission have prices lower by 0.1251 units compared to cars with automatic transmission, holding other factors constant.

- **Max_Torque_rpm**: For each additional unit of max torque in rpm, the price decreases by 0.1034 units, holding other factors constant.

- **Kilometer**: For each additional kilometer, the price decreases by 0.0976 units, holding other factors constant.

- **Length**: For each additional unit of length, the price increases by 0.1056 units, holding other factors constant.

- **All-Wheel Drive (AWD)**: Cars with AWD have prices higher by 0.0398 units compared to cars without AWD, holding other factors constant.

- **Width**: For each additional unit of width, the price increases by 0.0518 units, holding other factors constant.

Overall Model Fit

- R-squared: 0.818, meaning that approximately 81.8% of the variability in the dependent variable (Price) is explained by the independent variables in the model.
- Adjusted R-squared: 0.817, which adjusts for the number of predictors in the model, showing a similar explanatory power as the R-squared.
- F-statistic: 501.1 with a very low p-value (2.96e-323), indicating that the model is statistically significant and that the independent variables together reliably predict the dependent variable.

4.2 Final model

Based on results above, the considered final model is given by:

$$\begin{aligned}\textbf{Price} &= \alpha_0 + \alpha_1\textbf{Year} + \alpha_2\textbf{Kilometer} + \alpha_3\textbf{Length} + \alpha_4\textbf{Fuel Tank Capacity}\\ &+ \alpha_5\textbf{Max power\_bhp} + \alpha_6\textbf{Max Torque\_rpm} + \alpha_7\textbf{Manual} + \alpha_8\textbf{AWD}\end{aligned}$$

where $\alpha_0, \alpha_1, \ldots \alpha_8$ are the model coefficients provided in Section 3.2

The model used for predecting is given by:

$$\begin{aligned}\textbf{Price} = -1.125 \times 10^8 + (5.563 \times 10^4)\textbf{Year} + (-1.1853)\textbf{Kilometer} + (107.8898)\textbf{Length}\\ + (6785.8427)\textbf{Fuel Tank Capacity} + (5334.8607)\textbf{Max power\_bhp}+\\ (-26.3223)\textbf{Max Torque\_rpm} + (-8.975 \times 10^4)\textbf{Manual} + (7.385 \times 10^4)\textbf{AWD}\end{aligned}$$

4.3 VIF Analysis

VIF (Variance Inflation Factor) measures the amount of multicollinearity in regression analysis. Table 3 below provide the VIF for each feature and its interpretation:

Feature	VIF	Interpretation
Year	1.41	Very low multicollinearity
Kilometer	1.67	Low multicollinearity
Length	3.26	Moderate multicollinearity
Fuel Tank Capacity	2.62	Low to moderate multicollinearity
Max Power_bhp	3.71	Moderate multicollinearity
Max Torque_rpm	1.36	Very low multicollinearity
Manual	1.16	Very low multicollinearity
AWD	1.20	Very low multicollinearity

Table 3

- The highest VIF is for the **Max Power_bhp** feature (VIF: 3.707702). This indicates some level of multicollinearity but is still within a generally acceptable range (below 5).
- All features show VIF values well below the threshold of 10, suggesting that multicollinearity is not a severe issue in this model.
- Given these VIF values, there is no immediate need to remove or combine features to address multicollinearity.

4.4 Assumptions of multiple linear regression

- **Independence of errors**:OLS results shows that Durbin-Watson statistic is 1.940, it suggests that there is little to no autocorrelation in the residuals, which is a good indication that the assumption of independence of errors is not violated.
- **Errors mean = 0**: the mean of error in this case is: 0.059 which is almost null.
- **Constant variation of errors**:

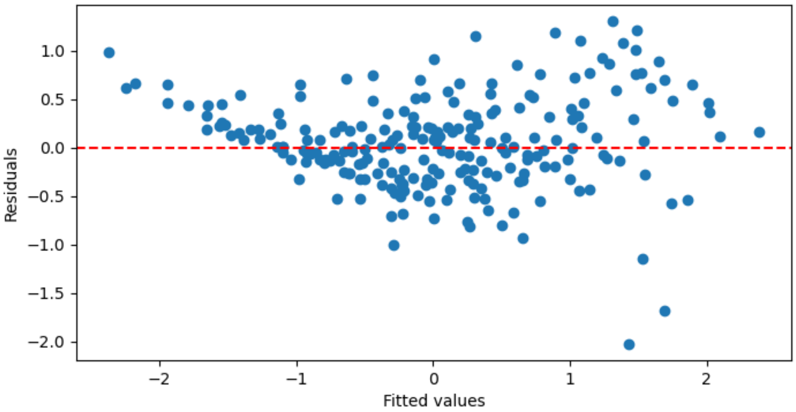


Figure 5: residuals vs fitted values

According to figure 5 errors does not realy have constant variance.

- **Normality of errors**: the next figure display the distribution of errors on the test set: Figure 6 is showing that errors are normaly distributed.

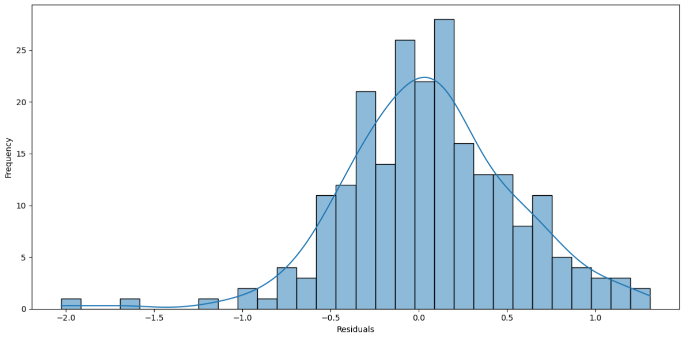


Figure 6: residuals distribution

5 Predictions

This section shows the application of the final model on the test set in order to make some predictions and testing how accurate this model is. The following figure shows a plot of the testing data points against the predicted values:

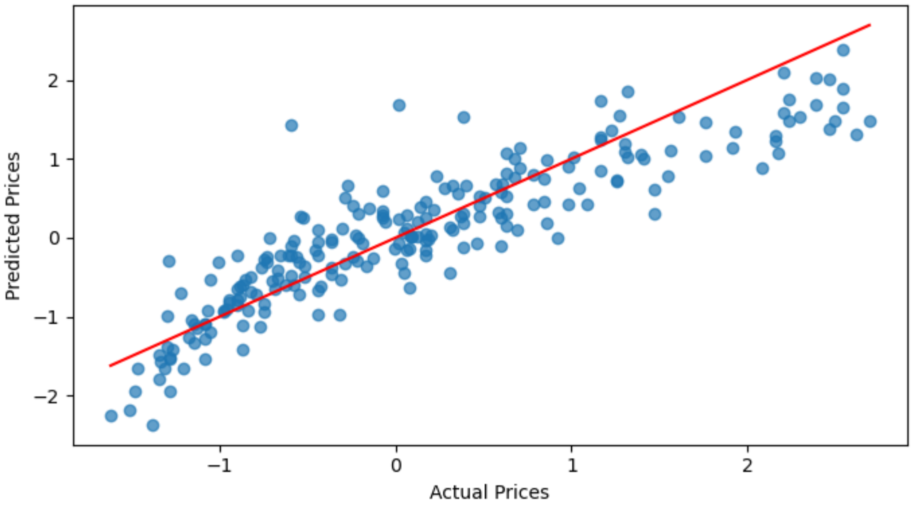


Figure 7: Predictions using the final model

Real time prediction:
The car showed in this link was picked to predict its price using the built model:

$$\begin{aligned}\textbf{Price} &= -1.125\times10^{8} + (5.563\times10^{4})\mathbf{2017} + (-1.1853)\mathbf{51000} + (107.8898)\mathbf{3995}\\ &\quad + (6785.8427)\mathbf{37} + (5334.8607)\mathbf{81} + (-26.3223)\mathbf{4200} + (-8.975\times10^{4})\mathbf{1} + (7.385\times10^{4})\mathbf{0}\\ &= 559175.6876\end{aligned}$$

the model predict this car to be priced around: 559175 rupees = 5,5 lakh[2]while the real price appearing on the website is 5.05 Lakh rupee which is quit good approximation of the real pricing.

6 Limitations

1. The results of this study are strongly dependent on how the data is distributed thus working more on data resampling may result in better model that is better predictions.

2. Relations discussed along this report between features may not have a sense in real life or my not represent the real relations between features (ex: price and max power_rpm, max torque_rpm) but they holds in the data subject to study.

3. For example while processing data some categories were droped from some features as in case with seller type where only cars selled by individuals were considered, which makes the model irrelevant and not accurate when predicting price of cars selled by dealers or coroperate, the same apply for cars with other fuel types.

4. All considered cars are from cardekho which is also a limitation.

All models are not perfect but there are some reliable ones.

[2] A lakh is a unit in the Indian numbering system equal to one hundred thousand

7 Conclusion

The regression model demonstrates a strong ability to predict used car prices based on the selected features, with Year, Kilometer, Length, Fuel Tank Capacity, Max Power_bhp, Max Torque_rpm, Manual transmission, and AWD as significant predictors. The residuals show some deviation from normality, which should be considered when making inferences from this model.

References

[1] Scribbr. (n.d.). *Multiple Linear Regression.* Retrieved from `https://www.scribbr.com/statistics/multiple-linear-regression/`

[2] CarDekho. (n.d.). *CarDekho Website.* Retrieved from `https://www.cardekho.com`

[3] Kaggle. (n.d.). *Vehicle Dataset from CarDekho.* Retrieved from `https://www.kaggle.com/datasets/nehalbirla/vehicle-dataset-from-cardekho/data`

[4] AlShared, A. (2021). Used Cars Price Prediction and Valuation using Data Mining Techniques. Rochester Institute of Technology. Retrieved from `https://repository.rit.edu/theses/11086`