Used Car Price Analysis using Machine Learning

Iftehaz Newaz
Dept. of Computer Science and Engineering,
International Islamic University Chittagong
Chittagong, Bangladesh

***Abstract*—Throughout the previous ten years, the number of vehicles produced has gradually increased. The used car market, which has grown into a thriving business on its own, has resulted from this. The recent rise of online platforms has enabled the requirement for both the buyer and the seller to be more aware of the trends and patterns that define a used car's market value. An accurate used car price assessment acts as a stimulus for the market's healthy evolution. In various studies, data mining has been utilized to forecast used car prices. Unfortunately, there is limited research on how different algorithms perform when used to predict used car prices. This paper offers analysis of various supervised machine learning model to predict the price of used cars. In this paper, the Decision Tree model fitted as the optimum model with an accuracy of around 93.73%.**

***Index Terms*—Regression, Supervised Learning, Machine Learning.**

I. Introduction

As a result of the high demand for private cars globally, there is a growing market for used cars that offers opportunities for both buyers and sellers. In many nations, purchasing a used car is the ideal option for the client because of the car's reasonable pricing. After a few years of use, they may be resold for a profit. Nevertheless, a variety of factors, including a used car's age, make, origin (the nation from where it was originally manufactured), mileage (the distance it has traveled), and horsepower, affect its price. Fuel economy is especially crucial due to the increasing cost of fuel. Additional factors include the fuel type, design, braking system, cylinder volume (in cubic centimeters), acceleration, number of doors, safety rating, size, weight, and height, paint color, customer feedback, and significant awards won by the manufacturer. The price may also be affected by other accessories such a music system, air conditioner, power steering, cosmic wheels, and a GPS navigator. [1], [2]

In this study, we propose a supervised learning method for predicting the price of used cars. Here, we have a dataset from Kaggle that includes used cars. The model was trained using the data. Using the K- Fold approach, which is simple to comprehend and use, the same model is cross-validated in order to evaluate the model's performance.

The paper structured in the following manner: Section II contains the Problems and the Dataset related to the field of used cars price prediction. In section III the methodology of the study was proposed. Section IV elaborates the Experimental Setup. The Section V specifies the Result and Discussion. Finally, Section VI contains the conclusion.

II. Problem and The dataset

For the used car market to grow healthily, an accurate evaluation of used car prices is necessary. Customers can purchase or sell old cars with confidence if they are aware of the fair price; car rental firms can price their services more profitably by estimating the residual value; and banks and other financial institutions can manage their lending quotas by assessing the worth of a lender's car. Thus, this is a significant and intriguing subject. [3] To shade light in this topic we have used machine learning techniques on the data which have been obtained from kaggle. The dataset contains 8128 instances and 9 features. Mileage feature contains 221 missing values. Similarly, engine 221, max_power 216, and seats 221.

TABLE I
Summary of dataset

Features	Type	Range of Values
Name	String	Arbitary
Year	Integer	1983 - 2020
Selling_price	Integer	29999 - 10000000
Km$_{driven}$	Integer	1 - 2360457
Fuel	String(Diesel-0, Petrol-1, LPG-2, CNG-3)	0-3
Seller_type	String (Individual-0, Dealer-1, and Trustmark Dealer-2)	0-2
Transmission	String (Manual-0, Automatic-1)	0,1
Owner	String (First-0, Second-1, Third-2, Fourth and above-3, and Test Drive Car-4)	0-4
Mileage	Float	0.0 - 42.0
Engine	Float	624.0 - 3604.0
Max_power	float	0.0 - 108495.0
Seats	float	2.0 - 14.0

III. Methodology

We selected the best-performing classifiers for this domain, those are Linear Regression, Elastic-net Regression, K-Nearest Neighbor Regression, and Decision Tree.

A. Linear Regression

A statistical technique called "linear regression" is used to regress data in which the independent variables might have either continuous or categorical values while the dependent variables have continuous values. To put it another way, "Linear Regression" is a technique for predicting the values of the dependent variable (Y) based on the values of the independent variables (X). When we want to anticipate

some continuous quantity, we may utilize it. – for example, Predicting traffic in a retail store.
Our goal while performing linear regression is to find the line that best fits the distribution and is closest to the majority of the points. Hence, the distance (error term) between the data points and the fitted line is reduced.

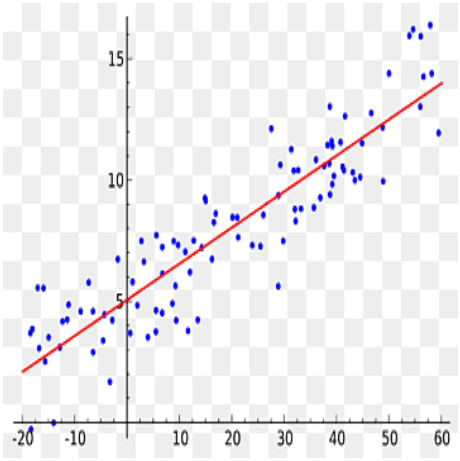

Fig. 1. Linear Regression

For instance, in the picture above, the line depicts an approximation of a line that may be used to illustrate the relationship between the 'x' and 'y' axes, while the dots indicate various data points. Such a line is what we aim to discover using linear regression. For instance, the relationship between one independent variable (X) and one dependent variable (Y) can be shown in the form of the following equation:

$$y = \beta_0 + \beta_1 . X \tag{1}$$

Where,

- Y = Dependent Variable
- X = Independent Variable
- β_0 = Constant term a.k.a Intercept
- β_1 = Coefficient of relationship between X & Y

B. ElasticNet Regression

For training simpler models with smaller coefficient values, penalties are added to the loss function as an extension of linear regression. These modifications are known as penalized linear regression or regularized linear regression. Elastic net is a popular type of regularized linear regression. By adding a penalty to these coefficients, ElasticNet Regression seeks for the coefficients that minimize the sum of error squares. L1 and L2 (Lasso and Ridge) techniques are combined in ElasticNet. As a result, it performs a more efficient smoothing process.
Features of ElasticNet Regression:

- It combines the L1 and L2 approaches.
- It performs a more efficient regularization process.
- It has two parameters to be set, λ and α.

Elastic Net aims at minimizing the following loss function:

$$L_{enet}(\hat{\beta}) = \frac{\sum_{i=1}^{n}(y_i - x_i^j \hat{\beta})^2}{2n} + \lambda \left(\frac{1-\alpha}{2} \sum_{j=1}^{m} \hat{\beta}_j^2 + \alpha \sum_{j=1}^{m} |\hat{\beta}_j| \right) \tag{2}$$

C. K-Nearest Neighbor Regression

KNN regression is a non-parametric technique that, by averaging the data in the same neighborhood, intuitively approximations the relationship between independent variables and the continuous result.

Euclidean distance is measured by:

$$\sqrt{\sum_{i=1}^{k}(x_t - y_t)^2} \tag{3}$$

Manhattan Distance is measured by:

$$\sum_{i=1}^{k} |x_t - y_t| \tag{4}$$

Minkowski Distance is measured by:

$$\left(\sum_{i=1}^{k} (|x_t - y_t|)^q \right)^{1/q} \tag{5}$$

Only continuous variables can be used with the three distance metrics mentioned above. When dealing with categorical variables, you must calculate the Hamming distance, which counts the instances in which similar symbols in two strings of equal length deviate from one another.

We examined the error rate for 1 to 10 k values and we can see that for the first 3 k values the score rate increases, after that it gradually decreases. looking at the Fig. 2 we can pick the best k for KNN regression. Thus we picked 3 as our k value.

D. Decision Tree

There are three different sorts of nodes in this tree-structured classifier. The initial node, known as the Root Node, represents the entire sample and may be divided into other nodes. The branches represent for the decision rules, while the interior nodes reflect the characteristics of a data set. The result is represented by the Leaf Nodes in the end.

A specific data point is traversed entirely through the tree by responding to True/False questions until it reaches the leaf node. The average of the dependent variable's value at that specific leaf node serves as the final prediction. The Tree can predict a suitable value for the data point after several iterations. Fig. 3 represents the diagram of Decision Tree principle.

E. Our approach

Before feeding the data to the machine learning model, we load the dataset and perform some preprocessing. After dealing with all the missing data, we examine whether the performance improves when the missing values are removed or replaced. The outcome shown that performance improves when the missing value is removed. Second, we are aware that models cannot handle data of the string type. As a result, we had to change the data's type to numeric.

The dataset is first preprocessed before being fed to models like linear regression, elastic net regression, k-nearest neighbor regression, and decision tree regression. The performance is then evaluated, and the results are compared. The bar plot of the r2 score in the result section depicts which one performs better. Fig. 4 represents the diagram of our approach.

IV. Experimental Setup

A. Data Preprocessing

To input the data into the models, we had to preprocess the data and give it a shape. Thus, we had to do missing value handling and label encoding to convert the string data type to numeric one,

Missing Value: Mileage feature contains 221 missing values. Similarly, engine 221, maxpower 216, and seats 221. We evaluated the effectiveness by dropping it and substituting the median for those variables. The experiment shown that it performs better when the missing data are removed.

Label Encoding: Just a few features have the string data type. We have to use the pandas library function pd.factorize() to transform the string value into a numeric type because the models cannot handle it. Here, it recognizes all distinct values and has the ability to switch out the string value for a numeric one afterwards.

Attribute reshaping The year attribute represents the time it was being purchased. To find out how many years it had been used, we replaced the year attribute with the difference between the purchasing year and 2022.

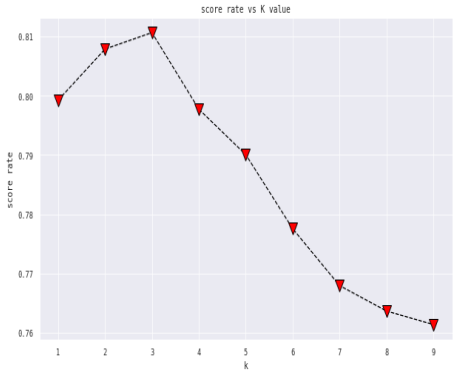

Fig. 2. K value vs Score Rate

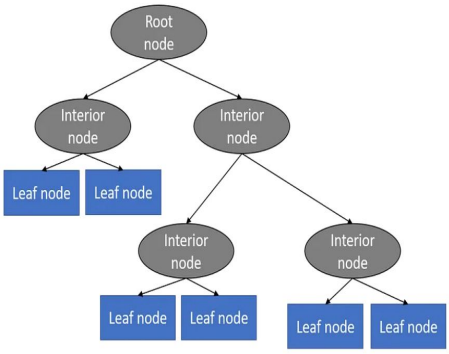

Fig. 3. Diagram of Decision Tree

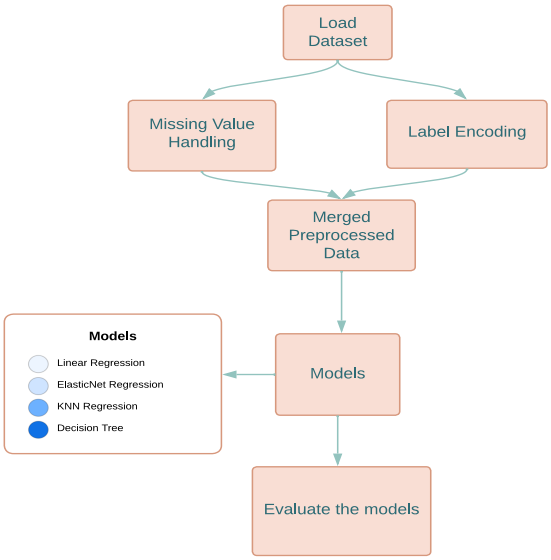

Fig. 4. Diagram of Our Approach

1) Heatmap: Multivariate data that is graphically represented as a matrix of columns and rows is known as a heat map. The association between various numerical variables may be described using heat maps, which can help to highlight patterns and abnormalities. Fig. 5 represents the heatmap of the features.

B. Histogram

The quickest approach to understand how each attribute in a dataset is distributed is to use a histogram, which groups the data into bins. The following are some histogram characteristics:

- It gives us a count of the observations contained in each bin that was made for visualization.
- We can quickly determine the distribution's type—whether it is Gaussian, skewed, or exponential—from the shape of the bin.

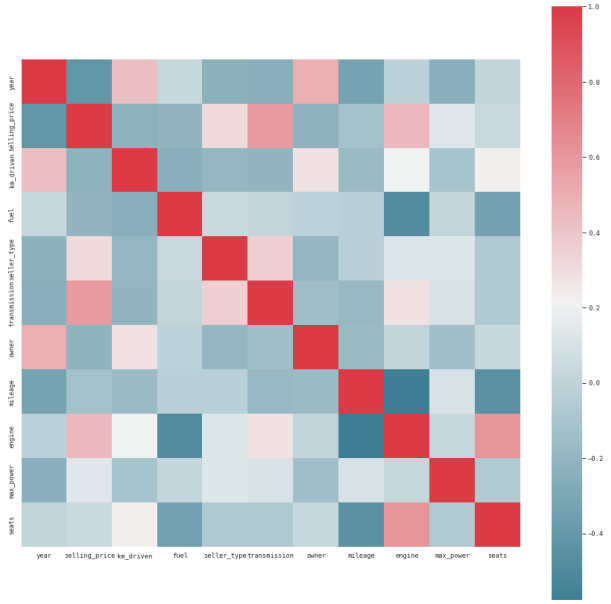

Fig. 5. Diagram of Heatmap

- Using histograms, we can also spot potential outliers.

C. Scatter

Using dots in two dimensions, a scatter plot illustrates the connection between two variables or how much one affects the other. In that they employ horizontal and vertical axes to depict data points, scatter plots are quite similar to line graphs in this regard.

V. Results and Discussion

After preprocessing the dataset, we did 5 fold cross validation of models and R2 score of the models are:

TABLE II
R2 Score of Models with 5 Fold CV

Models Name	R2 Score
Linear Regression	55.83%
Elastic Regression	48.22%
K-Nearest Neighbor	81.74%
Decision Tree	93.73%

From table II, we can see the R2 score of Linear Regression, ElasticNet Regression, KNN Regression, Decision Tree Regression. the R2 score is 93.73% for Decision Tree which is the highest one. the R2 score is 48.22% for ElasticNet Regression which is the lowest one.

TABLE III
Parameters

Models Name	Parameters
Linear Regression	No parameter used
ElasicNet Regression	random$_state = 0$
K-Nearest Neighbor	neighbor = 3
Decision Tree	random$_state = 0, max_depth = 13$

VI. Conclusion

In conclusion we have implemented several machinelearning regression models on the used car dataset which was obtained fromkaggle. The experiment showed Decision Tree Regressor performs best and ElasticNet Regressor performs worst among all the models. In future work we plan to explore more methodology as well as deep learning techniques for better performance.

References

[1] Monburinon, N., Chertchom, P., Kaewkiriya, T., Rungpheung, S., Buya, S., Boonpou, P. (2018, May). Prediction of prices for used car by using regression models. In 2018 5th International Conference on Business and Industrial Research (ICBIR) (pp. 115-119). IEEE.

[2] Pal, N., Arora, P., Kohli, P., Sundararaman, D., Palakurthy, S. S. (2019). How much is my car worth? A methodology for predicting used cars' prices using random forest. In Advances in Information and Communication Networks: Proceedings of the 2018 Future of Information and Communication Conference (FICC), Vol. 1 (pp. 413-422). Springer International Publishing.

[3] N. Kanwal and J. Sadaqat, "Vehicle Price Prediction System using Machine Learning Techniques," International Jounal of Computer Applications, vol. 167, no. 9, pp. 27–31, 2017.

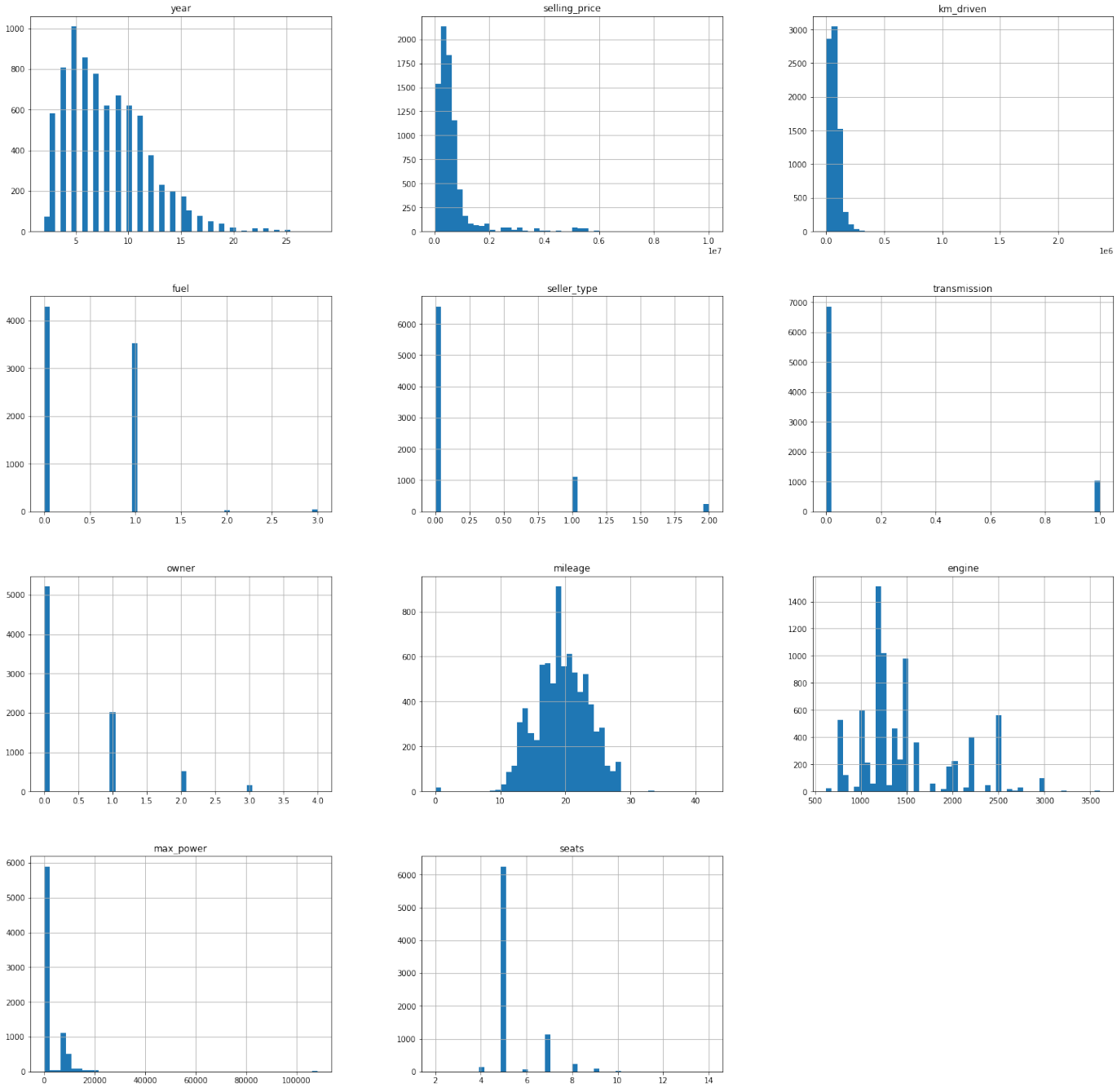

Fig. 6. Histograms

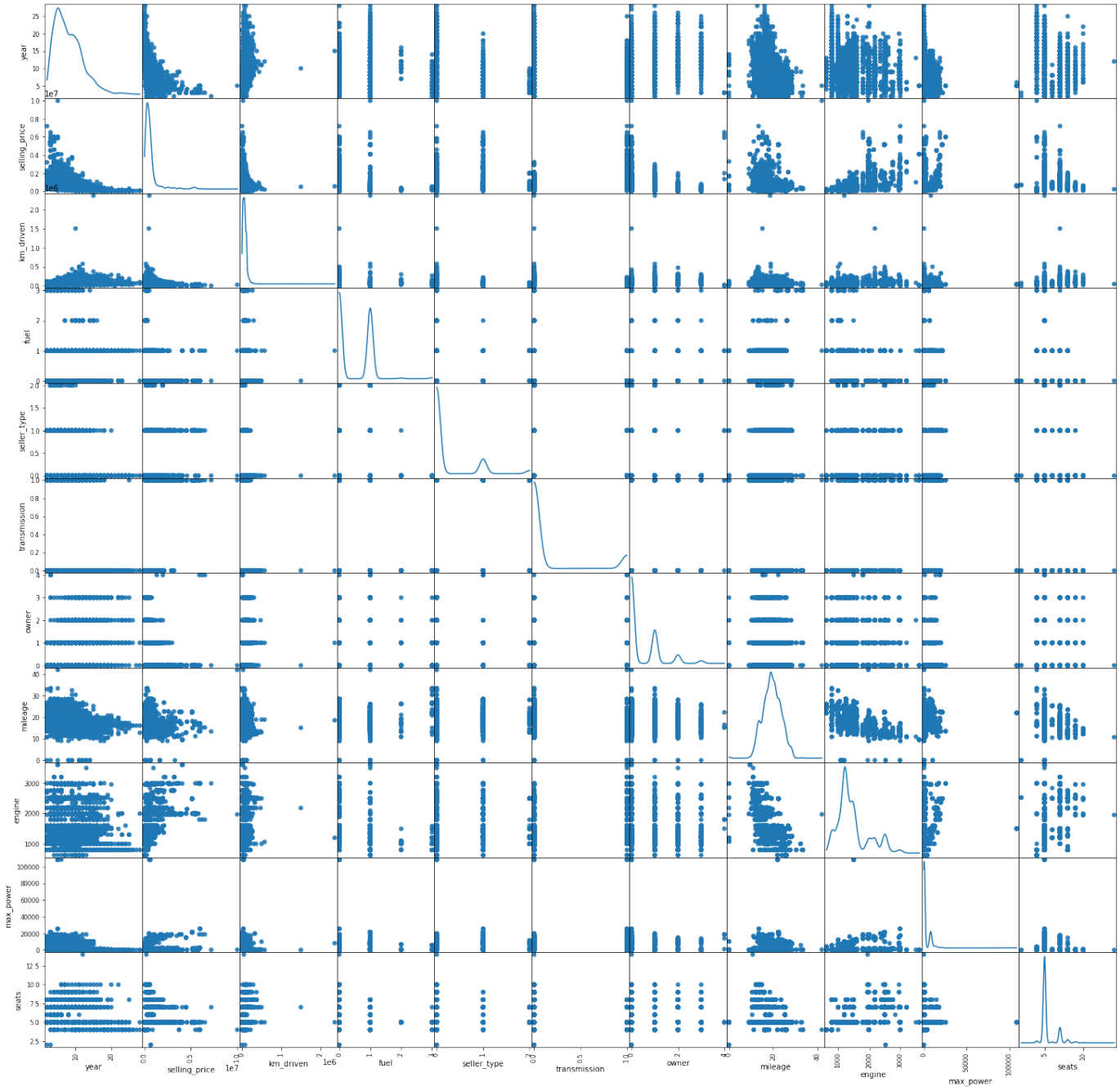

Fig. 7. Scatter plots