International Journal of Trend in Scientific Research and Development (IJTSRD)
Volume 5 Issue 4, May-June 2021 Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

Prediction of Car Price using Linear Regression

Ravi Shastri[1], Dr. A Rengarajan[2]

[1]Student, [2]Professor,
[1,2]School of CS & IT, Department of MCA, Jain University, Bangalore, Karnataka, India

ABSTRACT

In this paper, we look at how supervised machine learning techniques can be used to forecast car prices in India. Data from the online marketplace quikr was used to make the predictions. The predictions were made using a variety of methods, including multiple linear regression analysis, Random forest regressor and Randomized search CV. The predictions are then analyzed and compared to determine which ones provide the best results. A seemingly simple problem turned out to be extremely difficult to solve accurately. All of the strategies yielded similar results. In the future, we want to make predictions using more advanced technologies.

***KEYWORDS**: Multiple linear regression, Random forest, Randomized search CV and Supervised learning*

***How to cite this paper**:* Ravi Shastri | Dr. A Rengarajan "Prediction of Car Price using Linear Regression" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-5 | Issue-4, June 2021, pp.866-869, URL: www.ijtsrd.com/papers/ijtsrd42421.pdf

IJTSRD42421

INTRODUCTION

Given the demand for cars, the second-hand car market has been growing in popularity, providing opportunities for both buyers and sellers. Buying a used car is the best option for customers in several countries because the price is fair and affordable. After a few years of use, it might be possible to resell them for a profit. However, many factors affect the price of a used car, including its age and current condition. In most cases, the price of a used car on the market fluctuates. As a result, a model for evaluating car prices is needed to assist in trading.

In this paper, we used multiple linear regression, random forest regression to build a price model for the car. Each algorithm relied on information gathered from a website. The main goal of this paper is to find the best predictive model for car price prediction. Predicting a car's resale value is not an easy job. The fact that the value of used cars is determined by a variety of variables. The most significant ones are typically the car's age, model, origin (the manufacturer's original country), mileage (the number of kilometer's it has travelled), and horsepower.

The fuel economy is also important because of rising fuel prices. Unfortunately, most people may not realise how much fuel their car consumes per km driven in reality. Other factors include the type of fuel it uses, the interior style, the braking system, acceleration, the volume of its cylinders (measured in cc), safety index, the car's size, number of doors, paint colour, weight, consumer reviews, prestigious awards won by the car manufacturer, the car's physical condition, whether it is a sports car, whether it has cruise control, and whether it is automatic or manual transmission, whether it belonged to a person or a business, as well as other features like air conditioning, sound system, power steering, cosmic wheels, and GPS navigator, may all affect the price.

The following is the outline for this research paper. In section II the segment looked at some prior studies that were close to this one. We have discussed our methodology in section III. We analysed and compared the results of our algorithms in section IV. Section V concludes with a conclusion and a potential opportunity.

Literature Review- Richardson [1] worked on the theory that car manufacturers are more likely to produce cars that do not depreciate rapidly in another university study. He demonstrated that hybrid cars (cars that use two separate power sources to drive the vehicle, i.e. they have both an internal combustion engine and an electric motor) are more able to keep their value than conventional vehicles by using a multiple regression study. This is most likely due to increased environmental concerns regarding climate change and higher fuel efficiency. Other variables such as age, mileage, make, and MPG (miles per gallon) were also taken into account in this report. He gathered all of his information from various website.

To estimate the price of a vehicle, Noor and Jan [2] used multiple linear regression. They used a variable selection method to find the variables that had the greatest influence and then eliminated the rest. Just a few variables are included in the data, which were used to create the linear regression model. With an R-square of 98 per cent, the result was remarkable.

Peerun [3] et al. researched to assess the neural network's success in predicting used car prices. However, particularly on higher-priced vehicles, the predicted value is not very similar to the actual price. In predicting the price of a used car, they found that support vector machine regression outperformed neural networks and linear regression.

To forecast the residual value of privately used vehicles, Gonggi [4] suggested a new model focused on artificial neural networks. The mileage, maker, and estimated useful life were the three key features used in this analysis. The model was tweaked to accommodate nonlinear relationships, which are difficult to analyze using traditional linear regression approaches. This model was found to be fairly effective at estimating the residual value of used vehicles.

Sun et al. [5] suggested using the optimized BP neural network algorithm to develop an online used car price assessment model. To maximize secret neurons, they developed a new optimization method called Like Block-Monte Carlo Method (LB-MCM). As compared to the non-optimized model, the result showed that the optimised model produced higher accuracy. Based on previous related works, we discovered that no one had yet used the random forest regression model to estimate the price of a used vehicle. As a result, we chose to use a random forest regression model to build a model for evaluating used car prices.

METHODOLOGY:

This section presents the research methodology

The car dataset for this study was obtained from www.quikr.com. For each vehicle, the following information was gathered: make, model, seller type, kilometre's driven, year of manufacture, fuel type, and price. A sample of the collected data is shown below in Table 1.

Table I. Sample Data collection

SI. no	Car Name	Year	Selling Price	Kms Driven	Fuel Type	Seller Type
1.	Ritz	2014	3.35	27000	Petrol	Dealer
2.	Sx4	2013	4.75	43000	Diesel	Dealer
3.	Ciaz	2017	7.25	6900	Petrol	Dealer
4.	Wagon r	2011	2.85	5200	Petrol	Dealer
5.	Swift	2014	4.60	42450	Diesel	Dealer
6.	Vitara brezza	2018	9.25	2071	Diesel	Dealer
7.	S cross	2015	6.50	33429	Diesel	Dealer
8.	Ciaz	2016	8.75	20273	Diesel	Dealer
9.	City	2016	9.50	33988	Diesel	Dealer
10.	Brio	2015	4.00	600000	Petrol	Dealer

# Selling Price: In Lakhs

Table II. DESCRIPTIVE STATISTIC OF NUMERICAL VARIABLES

Attributes	Mean	Std	Min	Max
Selling Price	4.661296	5.082812	0.100000	35.000000
Present Price	7.628472	8.644115	0.320000	35.000000
Kms Driven	36947.205980	38886.883882	500.000000	500000.000000
Owner	0.043189	0.247915	0.000000	3.000000

These datasets will contain a large amount of used car data, so they will most likely need some tuning and engineering. Duplicated, for example, the model output can be affected by observations, so they must be excluded beforehand.

Each attribute requires some tweaking, according to the statistical details in Table II. The average price, in particular, was 4.661296, with a standard deviation of 5.082812. This suggested that the price values in the dataset are widely dispersed.

In predictive statistics and machine learning, attributes with a high correlation coefficient have a greater effect on the prediction variable, although this is not always the case. The correlation coefficient is a statistical measure that defines the relationship between variables, as its name suggests. The correlation coefficient between two attributes is always in the range of 1 (Positive relationship) to -1 (Negative relationship), while 0 indicates that there is no correlation at all.

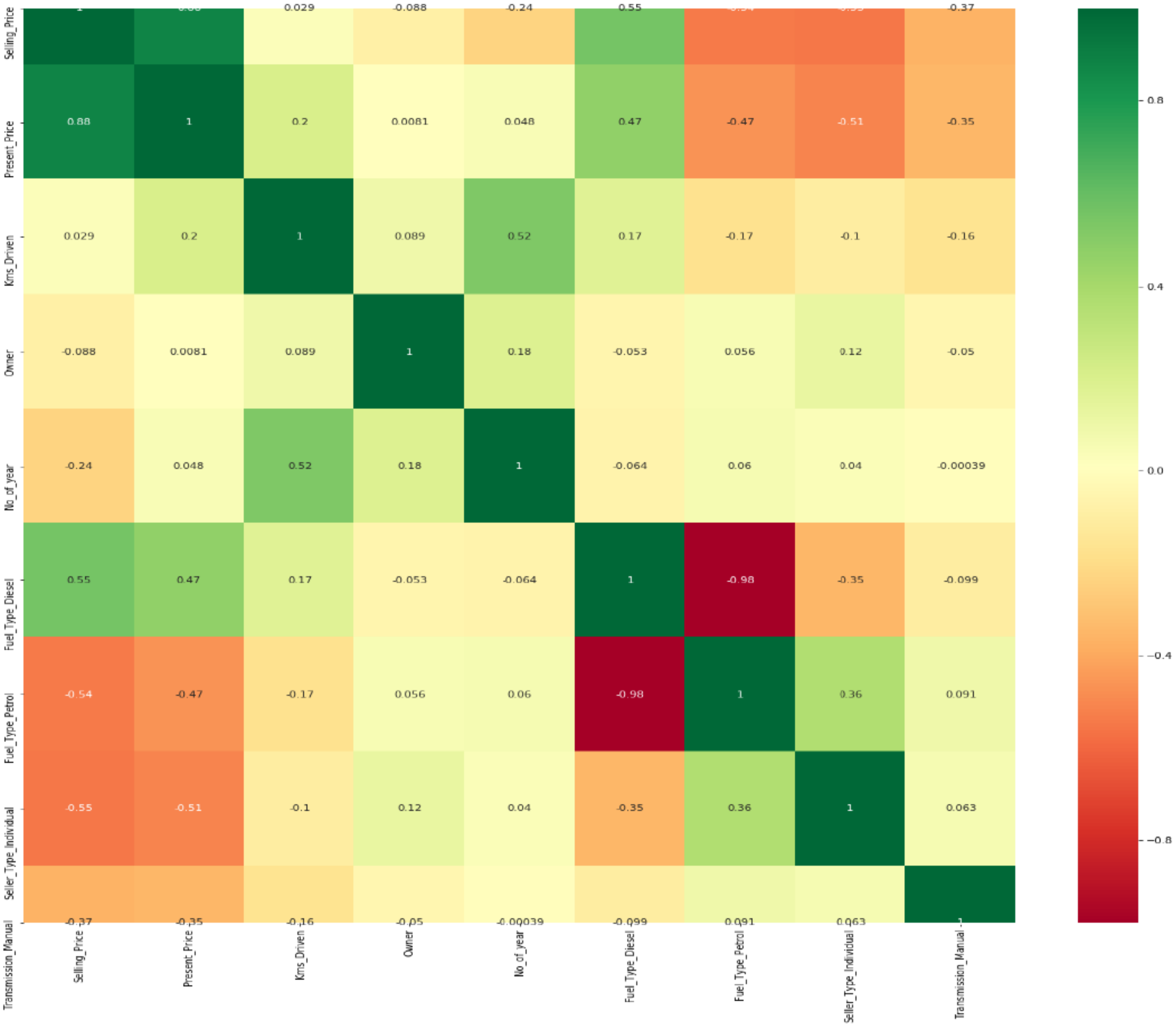

Fig I A correlation matrix of every attribute

Price prediction: a comparative review

This study uses the Scikit-learn machine learning library to implement multiple machine learning algorithms. The same training data is used to train each model, and the same testing data is used to evaluate it. In the following section, the results are compared and defined.

The regression-based approach is reliable in predicting continuous variables in supervised machine learning. A single linear regression model, as shown in, is sufficient to predict Y, where Y is the dependent variable and X is the independent variable. The model will forecast the future value of Y by determining the Y-intercept and slope of the regression line plus noise.

RESULT AND DISCUSSION:

Using testing data as input to multiple linear regression and random forest regression, the following results are evaluated. The mean absolute error of multiple linear regression and random forest regression was compared using mean absolute error as a criterion. With an MAE of = 0.72, random forest regression produces the best results.

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |x_i - x|$$

It should be noted that MAE is a negative focused ranking, meaning the closer the value is to zero, the better the model prediction.

Conclusion:

The authors of this study performed a comparison of regression-based model results. The data for this study was scraped from a popular e-commerce site called Quikr and then processed using the Python programming language. As a consequence, there are 240 rows and 8 attributes in the final data. On that particular dataset, we used multiple linear regression and random forest regression to test the results. The same testing data was used to assess and model. The mean absolute error is then used as a criterion for comparing the outcomes. With only MAE =0.72, random forest regression generated the best results. As a result, we came to the conclusion that using random forest regression trees to create the price evaluation model.

This research can be used to improve future work by fine-tuning each model parameter. To generate better training data, more appropriate data engineering can be used. The models can also be used in real-life situations.

References:

[1] RICHARDSON, "Determinants of Used Car Resale Value," 2009.

[2] Sadaqat, N. Kanwal and a. J., "Vehicle Price Prediction System using Machine Learning Techniques," *International Journal of Computer Applications,* pp. 27-31, 2017.

[3] S. Peerun, N. H. Chummun and a. S. Pudaruth, "Predicting the Price of Second-hand Cars using Artificial Neural Networks," *The Second International Conference on Data Mining, Internet Computing, and Big Data,* pp. 17-21, 2015.

[4] GONGGI, "New model for residual value prediction of used cars based on BP neural network and non-linear curve fit," *International Conference on Measuring Technology and Mechatronics Automation (ICMTMA),,* pp. 682-685, 2011.

[5] N. Sun, H. Bai, Y. Geng and a. H. Shi, "Price evaluation model in second-hand car system based on BP neural network theory," *International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD),* pp. 431-36, 2017.

[6] Monburinon, N. a. Chertchom, P. a. Kaewkiriya, T. a. Rungpheung, S. a. Buya, S. a. Boonpou and Pitchayakit, "Prediction of prices for used car by using regression models," *International Conference on Business and Industrial Research,* no. IEEE, pp. 115-119, 2018.

[7] Gegic, E. a. Isakovic, B. a. Keco, D. a. Masetic, Z. a. Kevric and Jasmin, "Car price prediction using machine learning techniques," *TEM Journal,* vol. 8, p. 113, 2019.

[8] Sinha, S. a. Azim, R. a. Das and Sourav, "Linear Regression on Car Price Prediction," 2020.

[9] Yang, R. R. a. Chen, S. a. Chou and Edward, Vehicle price prediction using visual features, 2018.

[10] Kiran and S, "Prediction of Resale Value of the Car Using Linear Regression Algorithm," *International Journal of Innovative Science and Research Technology,* vol. 5, no. 7, pp. 382-386, 2020.