JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Used-Car Price Prediction Using Machine Learning Technique

[1]Siddiqui Mohd. Shakeeb, [2]Prof. Sarwesh Site
[1]M.Tech. Scholar, [2]Asst. Professor
Department of Computer Science & Engineering
All Saints' College of Technology, Bhopal. India

Abstract:
In this study, we investigated the feasibility of forecasting car prices in India using supervised machine learning methods. To produce the forecasts, data from the internet marketplace Olx.com was used. The forecasts were made utilising numerous techniques, such as Random forest and multiple linear regression analysis and CV of a random search. The forecasts are then examined and compared them to see which ones deliver the best outcomes. A seemingly straightforward issue proved to be very challenging to correctly tackle. All of the methods produced comparable outcomes. Future improvements we would like to make predictions made with more cutting-edge technology.

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

1. INTRODUCTION:

The used car industry has become more well-known as a result of the increase in demand for automobiles, presenting opportunities for both buyers and sellers. For clients in many nations, purchasing a used car is the greatest choice because the cost is reasonable and accessible. It could be possible to sell them again after using them for a while and making a profit. However, a used car's price is influenced by a variety of factors, including the vehicle's age and present state. The cost of a used car on the market typically changes. As a result, trading would benefit from a model for assessing car prices.

In this study, we developed a price model for the car using multiple linear regression and random forest regression. Every algorithm was dependent on data acquired from a website.
The main objective of this study is to identify the most effective predictive model for predicting car prices. Calculating a car's resale value is a difficult task. the reality that a number of factors affect how much used cars are worth. The most important ones are usually the vehicle's make, model, origin (the nation of manufacturing), mileage (the distance covered) and horsepower.
Because of the rising cost of fuel, the fuel efficiency is especially crucial. Unfortunately, the majority of individuals might not be aware of how much fuel their automobile actually uses each mile travelled. Other

The fuel it consumes, the interior design, and other aspects are among the acceleration, braking system, and cylinder volume (in cc), safety index, the car's size, and the quantity of doors, weight, colour of the paint, client feedback, and prestigious honours received by the automaker, the vehicle's physical attributes condition, sports vehicle status, cruise control, and more
control, and whether the transmission is automated or manual, as well as whether it belonged to a person, a firm, additional amenities, such as a sound system, power steering
the effects of the GPS navigator, cosmic wheels, and air conditioning.

2. Literature Review:

Review of the Literature Richardson developed the hypothesis that automakers are more likely to produce cars that do not depreciate quickly in a university research. Using a multiple regression study, he showed that hybrid cars (automobiles that have both an internal combustion engine and an electric motor to power the vehicle) are better able to maintain their value than traditional vehicles. This is most likely a result of growing environmental worries about climate change and improved fuel economy. In addition, factors including age, mileage, make, and MPG (miles per gallon) were considered in this research. All of his information was taken from several websites.

Multiple linear regression was used by Noor and Jan to predict the cost of a car. They employed a variable selection strategy to determine which factors had the biggest impact before removing the remainder. The data, which were used to build the linear regression model, only contain a small number of variables. The result was astounding, with an R-square of 98%.

Peerun et al. conducted study to evaluate the neural network's performance in used-car price prediction. The anticipated value is not very close to the real price, especially for more expensive cars. They discovered that support vector machine regression beat neural networks and linear regression in forecasting the cost of a second-hand car. A new artificial neural network-based methodology was proposed by Gonggi to predict the residual value of privately owned vehicles. The three main criteria employed in this analysis were mileage, manufacturer, and predicted useful life. Nonlinear relationships, which are challenging to examine using conventional linear regression procedures, were modified into the model. It was discovered that this model can reasonably predict the residual value of old cars.

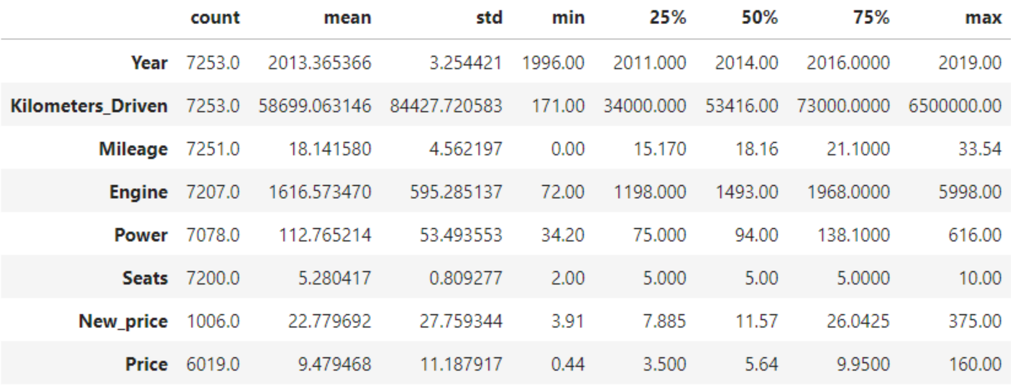
Sun et al.'s suggestion was to create an online used automobile pricing evaluation model utilising the improved BP neural network method. They created a novel optimization technique dubbed the Like Block Monte Carlo Method to optimise secret neurons (LB-MCM). The outcome shown that the optimised model delivered higher accuracy when compared to the non-optimized model. To estimate privately-owned vehicles' residual value,

We found that no one had previously used the random forest regression model to estimate the cost of a second-hand car based on prior relevant works. As a result, we decided to create a model for assessing used automobile prices using a random forest regression model.

3. Methodology:

The research technique is presented in this section. The olx.com was used to collect the automobile dataset for this investigation. Each vehicle's make, model, seller type, number of miles driven, year of manufacture, fuel type, and price were recorded. In Table 1, a sample of the information gathered is displayed.

Table I. Collected Test Data



	count	mean	std	min	25%	50%	75%	max
Year	7253.0	2013.365366	3.254421	1996.00	2011.000	2014.00	2016.0000	2019.00
Kilometers_Driven	7253.0	58699.063146	84427.720583	171.00	34000.000	53416.00	73000.0000	6500000.00
Mileage	7251.0	18.141580	4.562197	0.00	15.170	18.16	21.1000	33.54
Engine	7207.0	1616.573470	595.285137	72.00	1198.000	1493.00	1968.0000	5998.00
Power	7078.0	112.765214	53.493553	34.20	75.000	94.00	138.1000	616.00
Seats	7200.0	5.280417	0.809277	2.00	5.000	5.00	5.0000	10.00
New_price	1006.0	22.779692	27.759344	3.91	7.885	11.57	26.0425	375.00
Price	6019.0	9.479468	11.187917	0.44	3.500	5.64	9.9500	160.00

These datasets will likely need some engineering and adjusting because they will likely contain a lot of used automobile data. For instance, duplicated observations may have an impact on model output; as a result, they must first be eliminated. Power and engine are important predictors of price. New_price is also a significant predictor of price.

Although this is not always the case, in predictive statistics and machine learning, qualities having a high correlation coefficient have a stronger impact on the prediction variable. As its name implies, the correlation coefficient is a statistical metric that describes the relationship between variables. The range of the correlation coefficient between two parameters is always between 1 and -1 (positive to negative), with 0 denoting no association at all.

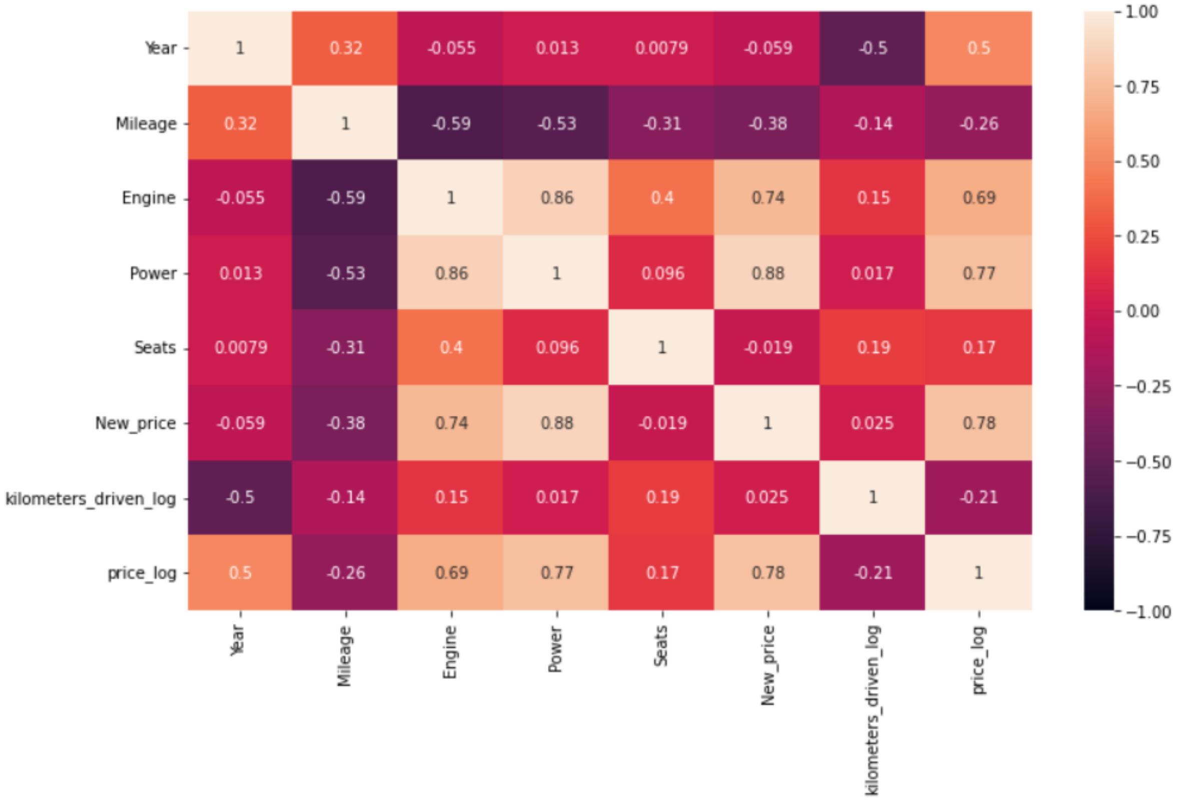


Fig.1 Heat Map

4. Price forecasting:

The Scikit-learn machine learning package is used in this work to develop a number of machine learning algorithms. Each model is trained using the same training data, and it is tested using the same testing data. The results are contrasted and defined in the section that follows.

In supervised machine learning, the regression-based approach is trustworthy for predicting continuous variables. As can be seen in where Y is the dependent variable and X is the independent variable, a single linear regression model is adequate to predict Y. The Y-intercept, slope, and noise of the regression line, combined with noise, will be used by the model to predict the future value of Y.

5. Results and Outcomes:

The following findings are assessed using testing data as input to multiple linear regression and random forest regression. Using mean absolute error as a criteria, the mean absolute error of multiple linear regression and random forest regression were evaluated. Random forest regression yields the best results with an MAE of 0.72.

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

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

6. Conclusion:

We analysed and compared results of a model based on regression. Olx, a well-known online platform, provided the data for this study, and the data was processed with the help of the Python programming language. Several linear regressions were performed on that particular dataset. Linear Regression and random forest regression were employed to test the results. The same test data were used in both the evaluation and modelling phases. The Following that, mean absolute error is used as a benchmark for comparing the outcomes of Random with just MAE =0.72, The best outcomes were produced via random forest regression. As a result, we decided to build the price evaluation model utilising random forest regression trees.

This research can be used to improve future work by finetuning 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. Gongqi, S., Yansong, W., & Qiang, Z. (2011). A New Model for Residual Value Prediction of the Used Car Based on BP Neural. Third International Conference on Measuring Technology and Mechatronics Automation (pp. 682-685). Shanghai: IEEE. doi:10.1109/ICMTMA.2011.45
2. Nabarun Pal, P. A. (2018). How much is my car worth? A methodology for predicting used cars prices using Random Forest. Future of Information and Communications Conference (FICC) 2018 , 1-6.
3. Singh, A., Thakur, N., & Sharma, A. (2016). A review of supervised machine learning algorithms. 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)
4. Used Vehicle Value Index. (2021, April). Retrieved from manheim: https://publish.manheim.com/en/services/consulting/used-vehicle-value-index.html
5. Richardson, M., 2009. Determinants of Used Car Resale Value. Thesis (BSc). The Colorado College.
6. Noor et al,2017. International Journal of Computer Applications (0975 – 8887) Volume 167 – No.9, June 2017
7. Saamiyah Peerun, 2015. Proceedings of the Second International Conference on Data Mining, Internet Computing, and Big Data, Reduit, Mauritius 2015Predicting the Price of Second-hand Cars using Artificial Neural Networks
8. 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.
9. Sinha, S. a. Azim, R. a. Das and Sourav, "Linear Regression on Car Price Prediction," 2020.