IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 11, Issue 4, April 2023

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 4, April 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1104020 |

Forecasting Used Car Price Using Machine Learning

Naresh Lohar, Yash Rathod, Asif Sayyed, Rohit Yadav, Prof. Amol Dhumal

Student, Dept. of Computer Engineering, Shah & Anchor Kutchhi Engineering College, Chembur, Mumbai, India

Professor, Dept. of Computer Engineering, Shah & Anchor Kutchhi Engineering College, Chembur, Mumbai, India

ABSTRACT: For this study, we compared the effectiveness of regression using supervised machine learning models for this research. E-commerce website's(Cars24.com & Cardekho.com) car industry data, available on kaggle, is used to train each model. Last but not least, the gradient-enhanced regression tree performed the best with an absolute error (MSE) of 0.28. The next two are, correspondingly, multiple linear regression with MSE = 0.55 and random forest regression with MSE = 0.35.

I. INTRODUCTION

As there is such a high demand for personal vehicles globally, there is a growing market for used cars that offers opportunities for both buyers and sellers. In many nations, purchasing a used vehicle is the best option for the customer because of the car's affordable price. After a few years of use, they might be sold again for a profit. However, a number of variables, including the age of the vehicles and their current state, affect the price of a used car. The cost of used vehicles on the market is typically not constant. Consequently, a model for evaluating vehicle prices is necessary to aid in trading.

In order to develop a price model for used cars, we did a comparative study using multiple linear regression, random forest regression, and gradient boosted regression trees. Each programme made use of data that was taken from an online store. Finding the best predictive model for used vehicle price prediction is the main goal of this essay.

This study paper has the following structure. The research looked at a few of the earlier, comparable works in section II. We discussed the use of machine learning models in computation in part III. We assessed and compared the output of our algorithms in section IV. Section V concludes with a statement about the upcoming opportunity.

II. RELATED WORK

The topic of used vehicle price prediction has been covered in a number of related works in the past. Pudaruth [1] used multiple linear regression, k-nearest neighbors, naive Bayes, and decision trees to forecast the cost of used cars in Mauritius. Despite the fact that there were fewer cars observed, their findings were poor for prediction. In his article, Pudaruth came to the conclusion that variables with continuous values cannot be used with decision trees or naive Bayes. Multiple linear regression was used by Noor and Jan [2] to forecast automobile price. They used the variable selection method to identify and then eliminate the variables that had the greatest influence. Only a few of the variables used to build the linear regression model are present in the data. R-square was impressively high at 98%. In order to assess the effectiveness of the neural network in predicting used car prices, Peerun et al. [3] conducted study. However, particularly for more expensive cars, the predicted value is not very close to the final cost. They came to the conclusion that support vector machine regression performed marginally better at used vehicle price prediction than neural network and linear regression. Sun et al.'s [4] application of the improved BP neural network algorithm-based online used vehicle price evaluation model was suggested. To optimise hidden neurons, they presented a brand-new optimisation technique dubbed Like Block-Monte Carlo Method (LB-MCM). When compared to the non-optimized model, the results showed that the optimised model produced better accuracy. We discovered that none of the earlier related works had yet used the gradient boosting technique to predict the price of a used car. As a result, we made the decision to use gradient boosted regression trees to create a model for evaluating used vehicle prices.

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 4, April 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1104020 |

III. PROPOSED ALGORITHM

A. :Data understanding and Data preparation

The used vehicle information used in this study was gathered from www.kaggle.com, where Orges Leka had posted it under a public domain licence. The attributes of used cars in this dataset, which includes 371,528 vehicle observations, are taken from the German e-commerce site eBay- Kleinanzeigen, as shown in Tables I and II.

TABLE I. DESCRIPTIVE STATISTIC CATEGORICAL VARIABLES

Attributes	Count	Unique	Top	Freq.
dataCrawled	371,528	280500	2016-03-24	7
name	371,528	233531	Ford Flesta	657
seller	371,528	2	pivat	371525
offerType	371,528	2	Angebot	371516
abtest	371,528	2	test	192585
vehicleType	333,659	8	limousine	95894
gearbox	351,319	2	manuell	274214
model	351,044	251	golf	30070
fuelType	338,142	7	benzin	223857
brand	371,528	40	volkswagen	79640
notReparedDamage	299,468	2	nein	263182
dateCreated	371,528	114	2016-04-03	14450
lastSeen	371,528	182806	2016-04-07	17

TABLE II. DESCRIPTIVE STATISTIC OF NUMERICAL VARIABLES

Attributes	Mean	Std.	Min	Max
price	17,295.14187	3.59E+06	0	2.15E+09
Year Of Registration	2004.577997	9.29E+01	1000	1.00E+04
powerPS	115.549477	1.92E+02	0	2.00E+04
kilometer	125618.6882	4.01E+04	5000	1.50E+05
MonthOf Registration	5.734445	3.71E+00	0	1.20E+01
Nr Of Pictures	0	0.00E+00	0	0.00E+00
postalCode	50820.66764	2.58E+04	1067	1.00E+05

These datasets may include a sizable amount of used vehicle information, so they probably need some engineering and tinkering. For instance, duplicated data must be removed beforehand [5] because they may affect model performance. For this action, the research used the Python programming language. [6]

A descriptive statistic for categorical factors is shown in Table I. Technically speaking, characteristics like dateCrawled, lastSeen, postal-Code, and dateCreated don't affect price prediction at all; as a result, they can be removed to enhance model performance.[7] As part of the data preparation procedure, attributes like seller, offerType, abtest, and nrOfPicture were also eliminated due to their wildly unbalanced values. Last but not least, name was eliminated as well because it has too many unique entries.

According to statistical information of attributes shown in Table II, each attributes require some tweaking. Especially on price, the average of price was 17,295.14, with a standard deviation of 3,587,954. This demonstrated that the dataset's price values are widely dispersed. Price has a right-skewed distribution, as seen in Fig. 1. Log change can be used to resolve this issue [8]. Price now has a bell-shaped distribution in Fig. 2. Observe that the price ranges from a minimum value of zero, which is mathematically impossible, to a highest value that is an outlier with a value of over 2.2 billion. By choosing the right range for analysis, 19% of the data from the total dataset were eliminated.

Regression using a machine learning method is not appropriate for categorical variables like gearbox, notRepairedDamage, model, brand, fuelType, and vehicleType. In order to help normalize these attributes, the label encoding algorithm was developed. A straightforward method for managing categorical variables that converts each value in a property is label encoding.

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 4, April 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1104020 |

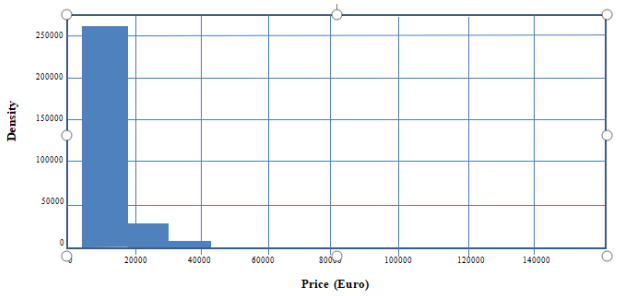

Fig. 1. A right-skewed distribution of price before log transformation

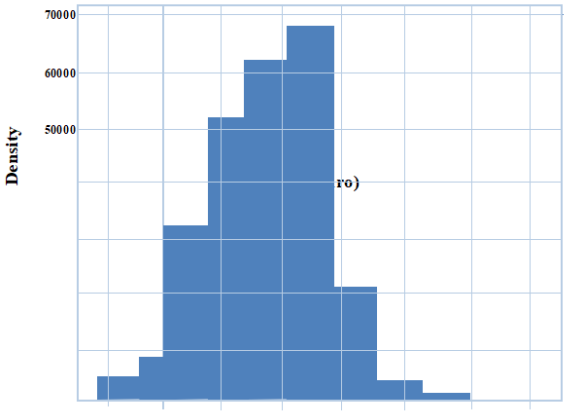

Fig. 2. A bell curve distribution of price after log transformation

Since there are a range of possible numerical values from 0 to n-1, some algorithms may perceive lower values as having less weight and higher values as having more weight. One hot encoding is a different approach to this particular issue that changes each category value into a new attribute with a 1 or 0 value showing whether or not an observation includes this value. This seems like a better choice for more realistic data interpretation. However, due to our limitation on computational resources, label encoding method is preferred for now.

An attribute with a high correlation coefficient frequently, but not always, has more impact on the prediction variable in predictive statistics and machine learning [9]. As its name suggests, the correlation coefficient is a statistical measure that depicts the connection between variables. The range of the correlation coefficient between any two attributes is always between 1 and 1, with 0 denoting no connection at all.

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \qquad (1)$$

The correlation matrix of every attribute is visualized in Fig. 3. We hypothesized that attributes such as powerPS, kilome-ter, yearOfRegistration, and gearbox feat which have high correlation coefficient value with the price of 0.573037, −0.444440, 0.385264, and −0.297746 respectively should have more impact on price prediction compared to others.

Finally, we splinted the data to create training and testing data with ratios of 0.67 and 0.33 respectively. Training data will be used to fit our predictive model, and testing data will be used to evaluate model performance

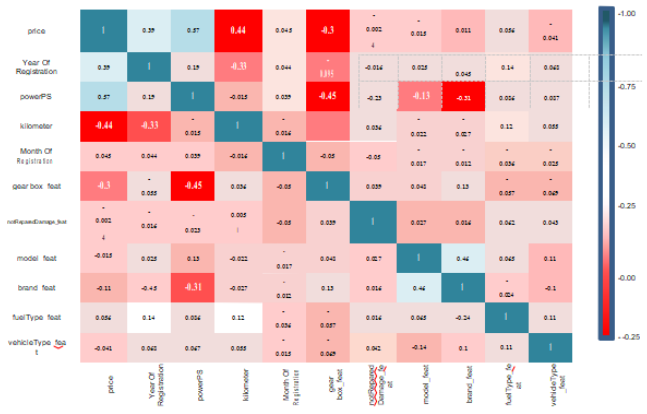

Fig. 3. A correlation matrix of every attribute

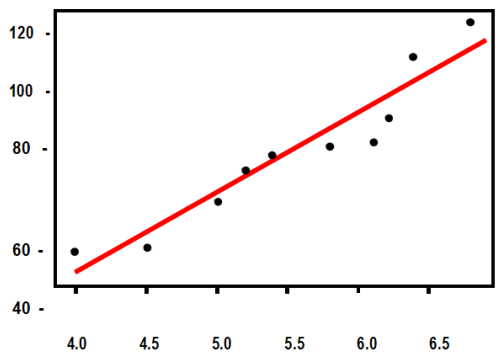

Fig. 4. An example plot of linear regression line over the entire dataset

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 4, April 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1104020 |

Comparative study on price prediction

Several machine learning methods from the Scikit-learn machine learning library are used in this study [10]. Every model is evaluated using the same testing data and learned using the same training data. The following part compares and describes the outcome.The regression-based approach has been shown to accurately forecast continuous variables in supervised machine learning [2]. For basic predictive modeling, single linear regression model as expressed in (2) is enough to predict Y where Y is dependent variable and X is the independent variable. By finding the Y-intercept and slope of regression line plus noise, the model can estimate the future value of Y

$$\hat{Y}_i = \hat{\beta}_o + \hat{\beta}_i X_i + c_i \qquad (2)$$

Multiple linear regression models expressed in are another popular substitute for simple linear regression when data includes multiple attributes. (3). It has identical qualities to its predecessor above. simply using a number of independent factors.

$$\hat{Y}_i = \hat{\beta}_o + \hat{\beta}_i X_i + \hat{\beta}_2 X_i \ldots + \hat{\beta}_n X_n + c_i \qquad (3)$$

A single regression line is used for the complete dataset in the linear regression method, as shown in Fig. 4. Therefore, it takes time to solve a more challenging problem with numerous attributes and strong nonlinearities. With the regression tree model, that is not the situation. A regression tree is a type of predictive tree that effectively applies the idea of recursive partitioning to handle nonlinear regression problems [11]. The complete dataset is divided into subdivisions, which are then divided again and again until the data in each subdivision are sufficiently simple that a learner can fit on them. [12] Each partition is represented by a regression tree as its leave, or terminal, node, and each terminal node has a basic model that was developed using that node's local data. Although a regression tree partition can be used to apply a number of straightforward models, the most recommended approach is to simply use the sample mean of the dependent variables in that partition as expressed in (4). An illustration of a simple regression tree is shown in Fig. 5.

$$\bar{Y} = \frac{1}{\square} \sum_{i=1}^{\square} y_i \qquad (4)$$

Although a single model can already be used to predict the target variable, ensemble methods usually yield better performance by combining several models to give a final prediction. [13] Bagging and boosting are two different kinds of assembling. Using testing data as input to multiple linear regression, random forest regression, and gradient boosted regression trees, many independent models are combined in bagging ensemble to produce the findings shown below. The findings are then compared using a standard called mean absolute error. Table III displays the mean absolute error (5) of gradient-boosted regression trees, random forest regression, and multiple linear regression in that sequence. The best result is produced by gradient-boosted regression, which has a mean absolute error of just 0.28. Second place goes to random forest regression, which has a mean absolute error of 0.35. When compared to the other averaged using some averaging techniques, multiple linear regression has a mean absolute error that is comparatively large at 0.55. An example of bagging ensemble is random forest, which use collection of classification or regression tree to help predict the outcome.

model	mean absolute error
Multiple linear regression	0.55
Random forest regression	0.35

TABLE III. THE PERFORMANCE OF EACH MODEL IN COMPARISON

Contrarily, the boosting method trained the data for each model sequentially rather than independently. The errors discovered in earlier models are used to inform succeeding models through iteration-based learning. The prediction value moves noticeably closer to the actual number with each iteration. Gradient boosting [14] is an illustration of a boosting algorithm. In order to determine which model is the best when it comes to solving regression problems, we did a comparative study on multiple linear regression, random forest regression, and gradient boosted regression trees. In this instance, a model for used car price prediction constitutes our regression issue.

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 4, April 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1104020 |

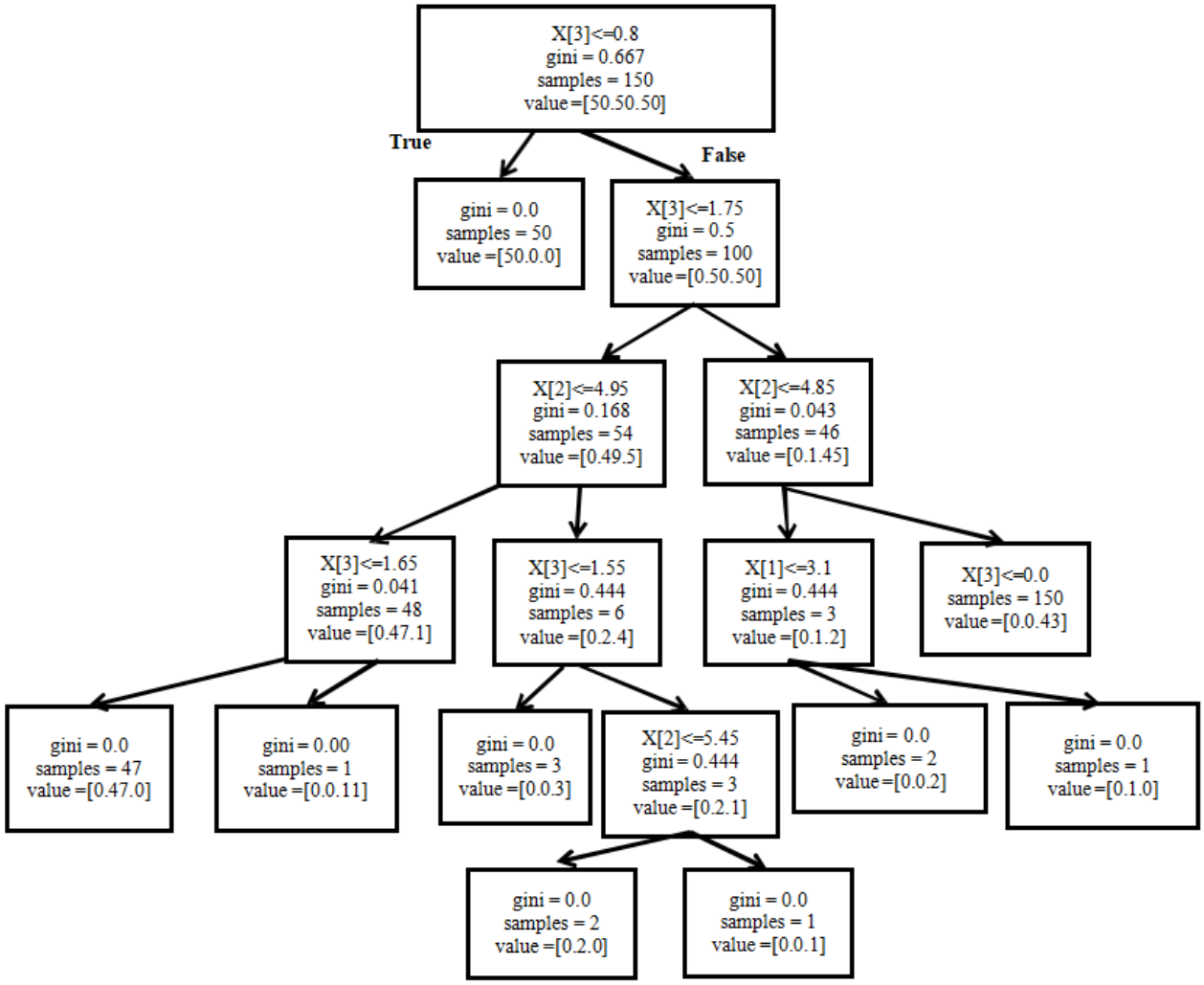

Fig. 5. An example visualization of regression tree using iris data

IV. RESULTS

The following findings compare the performance of gradient-boosted regression trees, multiple linear regression, and regression using a random forest. The findings are then compared using a standard called mean absolute error. Table III lists the mean absolute error (5) for gradient-boosted regression trees, random forest regression, and multiple linear regression in that sequence. The greatest results come from gradient-boosted regression, which only has an MSE of 0.28. With MSE = 0.35, random forest regression comes in second. When compared to other regression methods, multiple linear regression has an MSE that is comparatively high (0.55).

$$\square = \frac{1}{\square} \sum_{i=1}^{\square} |y_i - \hat{y}_i| \qquad (5)$$

Noted that MAE is a negative oriented score which means that the closer the value is to zero, the better the model prediction.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 4, April 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1104020 |

V. CONCLUSION

In this investigation, authors compared the effectiveness of regression-based models. Data for this study was scraped from a German e-commerce website and processed using the Python computer language. The final data consist of 304,133 rows and 11 attributes as a consequence. On that specific dataset, we evaluated the data using multiple linear regression, random forest regression, and gradient boosted regression trees. The same test data were used to assess each model. The outcomes are then contrasted using the mean absolute error as a standard. Regression trees that had been gradient-boosted performed best with an MAE of only 0.28. Then came multiple linear regression, which had 0.55 errors, and random forest regression, which had 0.35 errors. As a result, we came to the conclusion that gradient boosted regression trees are suggested for building price assessment models. This study can be used to inform future work by fine-tuning each model parameter. More appropriate data engineering can be utilize to create the better training data. For a more accurate data interpretation on categorical data, one hot encoding can be used as an alternative to label encoding, as was stated in Section III. The models can also be used in practical applications, but this requires further refinement.

REFERENCES

1. N. Kanwal and J. Sadaqat, "Vehicle Price Prediction System using Machine Learning Techniques," International Jounal of Computer Ap-plications, vol. 167, no. 9, pp. 27–31, 2017.
2. S. Peerun, N. H. Chummun, and 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, no. August, pp. 17-21, 2015.
3. N.Sun, H. Bai, Y. Geng, and H. Shi, "Price evaluation model in second-hand car system based on BP neural network theory," in 2017 18th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), jun 2017, pp. 431–436.
4. G.Rossum, "Python Reference Manual," Amsterdam, The Netherlands, The Netherlands, Tech. Rep., 1995.
5. A. K. Elmagarmid, P. G. Ipeirotis, and V. S. Verykios, "Duplicate Record Detection: A Survey," IEEE Transactions on Knowledge and Data Engineering, vol. 19, no. 1, pp. 1–16, jan 2007.
6. G.Chandrashekar and F. Sahin, "A survey on featureselection methods," Computers & Electrical Engineering, vol.40,no.1,pp.16–28,2014.[Online].Available: http://www.sciencedirect.com/science/article/pii/S0045790613003066
7. M.C.Newman,"Regression analysis of log-transformed data: Statistical bias and its correction," Environmental Toxicology and Chemistry, vol. 12, no. 6, pp. 1129–1133, 1993. [Online]. Available: http://dx.doi.org/10.1002/etc.5620120618
8. R.Taylor, "Interpretation of the Correlation Coefficient: A Basic Re- view," Journal of Diagnostic Medical Sonography, vol. 6, no. 1, pp. 35–39, 1990.
9. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vander- (5)plas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duches-nay, "Scikit-learn: Machine Learning in fPgython," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.
10. J. Morgan, "Classification and Regression Tree Analy-sis," Bu.Edu, no. 1, p. 16, 2014. [Online]. Available: http://www.bu.edu/sph/files/2014/05/MorganCART.pdf

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor: 8.379

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details