Prediction of Prices for Used Car by Using Regression Models

Nitis Monburinon
Faculty of Information Technology
Thai-Nichi Institute of Technology
Bangkok, Thailand
boonitis@gmail.com

Prajak Chertchom
Faculty of Information Technology
Thai-Nichi Institute of Technology
Bangkok, Thailand
prajak@tni.ac.th

Thongchai Kaewkiriya
Faculty of Engineering and Technology
Panyapiwat Institute of Management
Nontaburi, Thailand
tkaewkiriya@gmail.com

Suwat Rungpheung
Faculty of Information Technology
Thai-Nichi Institute of Technology
Bangkok, Thailand
ru.suwat_st@tni.ac.t

Sabir Buya
Faculty of Information Technology
Thai-Nichi Institute of Technology
Bangkok, Thailand
bu.sabir_st@tni.ac.th

Pitchayakit Boonpou
Faculty of Information Technology
Thai-Nichi Institute of Technology
Bangkok, Thailand
bo.pitchayakit_st@tni.ac.th

***Abstract*—For this research, we conducted a comparative study on performance of regression based on supervised machine learning models. Each model is trained using data of used car market collected from German e-commerce website. As a result, gradient boosted regression trees gives the best performance with mean absolute error (MSE) = 0.28. . Followed by random forest regression with MSE = 0.35 and multiple linear regression with MSE = 0.55 respectively.**

Keywords—comparative study, multiple linear regression, random forest, gradient boosting, supervised learning

I. Introduction

Considering the demand for private car all around the world, the demand of second-hand car market has been rising and creating a chance in business for both buyer and seller. In several countries, buying a used car is the best choice for customer because its price is reasonable and affordable by buyer. After few years of using them, it may get a profit from resell again. However, various factors influence the price of a used car such as how old of those vehicles and the condition in current scenario of them. Normally, the price of used cars in the market is not constant. Thus, car price evaluation model is required for helping in trading.

In this paper, we conducted a comparative study using multiple linear regression, random forest regression and gradient boosted regression trees to build a price model of used car. Each algorithm used data scraped from e-commerce website. The primary objective of this paper is to find the best predictive model for predicting used car price.

The structure of this research paper is as follows. In section II, the study reviewed some of the similar works that have been done previously. In section III, we described machine learning model using in computation. In section IV, we evaluated and compared the result of our algorithms. Finally, conclusion and future opportunity are stated in section V.

II. Literature Reviews

Several related works have been done previously on the subject of used car price prediction. Pudaruth [1] predicted the price of used cars in Mauritius using multiple linear regression, k-nearest neighbors, naive Bayes and decision trees.

Although their results was not good for prediction due to a less number of car observation. Pudaruth concluded in his paper that the decision tree and naive Bayes are unable to use for variable with a continuous value.

Noor and Jan [2] used multiple linear regression to predict vehicle car price. They performed variable selection technique to find the most influencing variables then eliminate the rest. The data contain only selected variable that used to form the linear regression model. The result was impressive with R-square = 98%.

Peerun et al. [3] did a research to evaluate the performance of the neural network in used car price prediction. The predicted value, however, are not very close to the actual price, especially on cars with a higher price. They concluded that support vector machine regression slightly outperform neural network and linear regression in predicting used car price.

Sun et al. [4] proposed the application of online used car price evaluation model using the optimized BP neural network algorithm. They introduced a new optimization method called Like Block-Monte Carlo Method (LB-MCM) to optimize hidden neurons. The result shown that the optimized model yielded higher accuracy when it compared to the non-optimized model Based on the previous related works, we realized that none of them had implemented gradient boosting technique in the prediction of used car price yet. Thus, we decided to build a used car price evaluation model using gradient boosted regression trees.

978-1-5386-5254-1/18/$31.00 ©2018 IEEE

III. METHODOLOGY

This section presents the research methodology.

A. Data understanding and Data preparation

The used car data used in this research were collected from www.kaggle.com which uploaded by Orges Leka under the public domain license. This dataset consists of 371,528 car observations and the attributes of used car are from eBay-Kleinanzeigen, a German e-commerce site as shown in Table 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 contain a significant number of used car information with several presumably, they require some tweaking and engineering. For example, duplicated observations may effect on model performance and they must be removed beforehand [5]. The study used python programing language for this action. [6]

Table I shows a descriptive statistic of categorical variables. Technically, attributes such as dateCrawled, lastSeen, postal-Code, and dateCreated have no impact on price prediction at all, thus, they can be removed to improve model performance.[7] With data preparation process by inspection of more detail on dataset, attributes such as seller, offerType, abtest, and nrOfPicture were also removed because their values are extremely unbalanced. Lastly, name was also removed because it contains too many unique values.

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 indicated that the price values are considerably spread across the dataset. Fig. 1 shows that price has a right-skewed distribution. This problem can be solved by using log transformation [8]. In Fig. 2, price now has a bell curve distribution. Also, notice that minimum value of price is zero, which is technically impossible and the maximum value is an outlier with a value over 2.2 billion. By selecting appropriate range for analysis, 19% of data were removed from the entire dataset.

Categorical such as gearbox, notRepairedDamage, model, brand, fuelType, and vehicleType are not suitable for regression based on machine learning algorithm. Thus, label encoding algorithm was implemented to help normalize these attributes. Label encoding is just a simple approach for handling categorical variables which convert each value in an attribute.

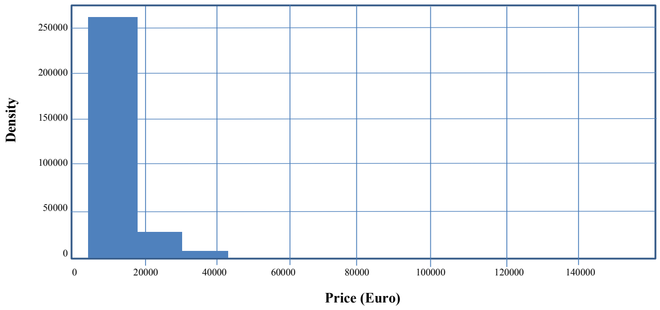

Fig. 1. A right-skewed distribution of price before log transformation

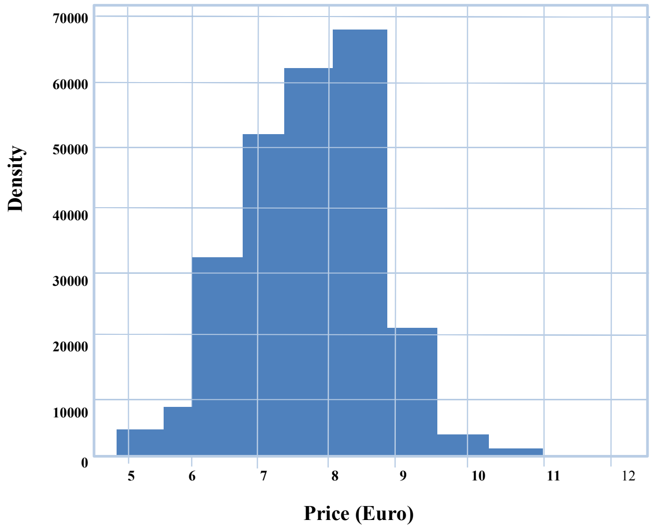

Fig. 2. A bell curve distribution of price after log transformation

Since numerical values may range from 0 to n-1, some algorithm may interpret lesser value with lower weight and higher value with greater weight. Another alternative approach to this specific problem is known as one hot encoding which converts each category value into a new attribute with 1 or 0 value indicating whether an observation contains this value or not. 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.

In predictive statistic and machine learning, an attributes with high correlation coefficient often, but not always, have more influence on prediction variable [9]. The correlation coefficient, as its name implies, is a statistical measure that describes the relationship between variables. The correlation coefficient of two attributes is always range between 1 (Positive relationship) to −1 (Negative relationship) whereas 0 implies no correlation at all.

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

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

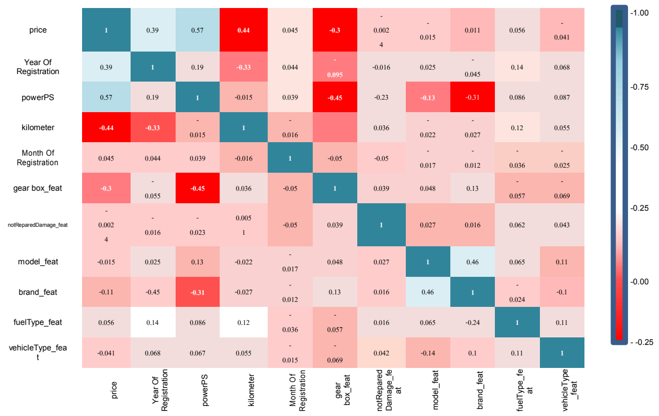

Fig. 3. A correlation matrix of every attribute

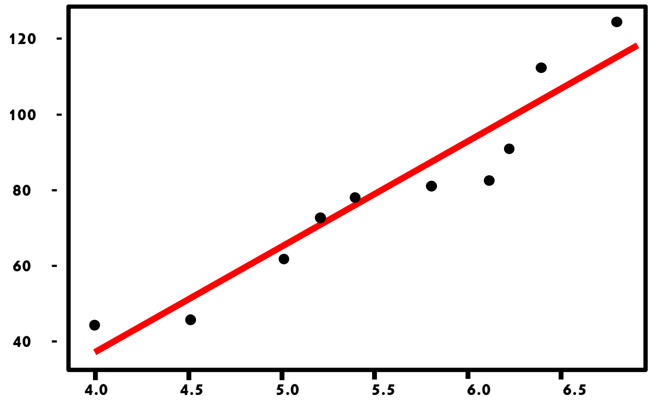

Fig. 4. An example plot of linear regression line over the entire dataset

B. Comparative study on price prediction

This research implements several machine learning algorithms available in Scikit-learn machine learning library [10]. Each model is trained using same training data and evaluate using same testing data. The result then compared and described in the next section.

In supervised machine learning, the regression-based method has been proven to be reliable in predicting continuous variable [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

$$\hat{Y}_i = \hat{\beta}_{\mathrm{o}} + \hat{\beta}_{\mathrm{i}} X_i + \hat{\epsilon}_i \tag{2}$$

Another common alternative to simple linear regression when data contains multiple attributes is, unsurprisingly, multiple linear regression models expressed in (3). It shares similar attribute with its precedence above. Just with multiple independent variables.

$$\hat{Y}_i = \hat{\beta}_{\mathrm{o}} + \hat{\beta}_{\mathrm{i}} X_i + \hat{\beta}_2 X_i \ldots + \hat{\beta}_{\mathrm{n}} X_n + \hat{\epsilon}_i \tag{3}$$

Linear regression method, as visualized in Fig. 4, use single regression line for the entire dataset. Thus, solving a more complicated problem involving multiple attributes that have strong nonlinearities is tedious. That is not the case with regression tree model.

A regression tree is a variation of a predictive tree that can solve nonlinear regression problem efficiently using the concept of recursive partitioning [11]. Where entire dataset is partition into subdivisions, then subdivisions are partitioned again and again until it reaches the point that data in that subdivision are so simple that a learner can comfortably fit on them. [12]

Regression tree represents each partition as its leave, or terminal node, where each node has its simple model that is trained using its local data. While several simple models can be implemented on regression tree partition, the most encouraged method is merely using the sample mean of dependent variables in that partition as expressed in (4). Fig. 5 shows an example of basic regression tree.

$$\overline{Y} = \frac{1}{n} \sum_{i=1}^{n} y_i \tag{4}$$

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] There are two types of assembling called Bagging and Boosting. In bagging ensemble, many independent models are combined, and the results are the following results are evaluating using testing data as input to multiple linear regression, random forest regression, and gradient boosted regression trees. The results are then compared using mean absolute error as a criterion. Table III show mean absolute error (5) of multiple linear regression, random forest regression, and gradient boosted regression trees, in order. Gradient boosted regression yield the best performance with only 0.28 mean absolute error. Random forest regression is in second place with 0.35 mean absolute error. Multiple linear regression has relatively large mean absolute error of 0.55 when compared with the other averaged using some averaging techniques. An example of bagging ensemble is random forest, which use collection of classification or regression tree to help predict the outcome.

TABLE III. THE PERFORMANCE OF EACH MODEL IN COMPARISON

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

On the other hand, boosting technique did not perform the training of data for each model independently, but did them sequentially. Using iteration based learning, subsequent models learn from the errors found in precedent models. As the iteration increasing, prediction value becomes significantly closer and closer to the actual value. An example of boosting algorithm is gradient boosting [14].

In this research paper, we conducted a comparative study on multiple linear regression, random forest regression, and gradient boosted regression trees to find out which model are the best when it will be used to solve regression problem. In this case, our regression problem is a model for used car price prediction.

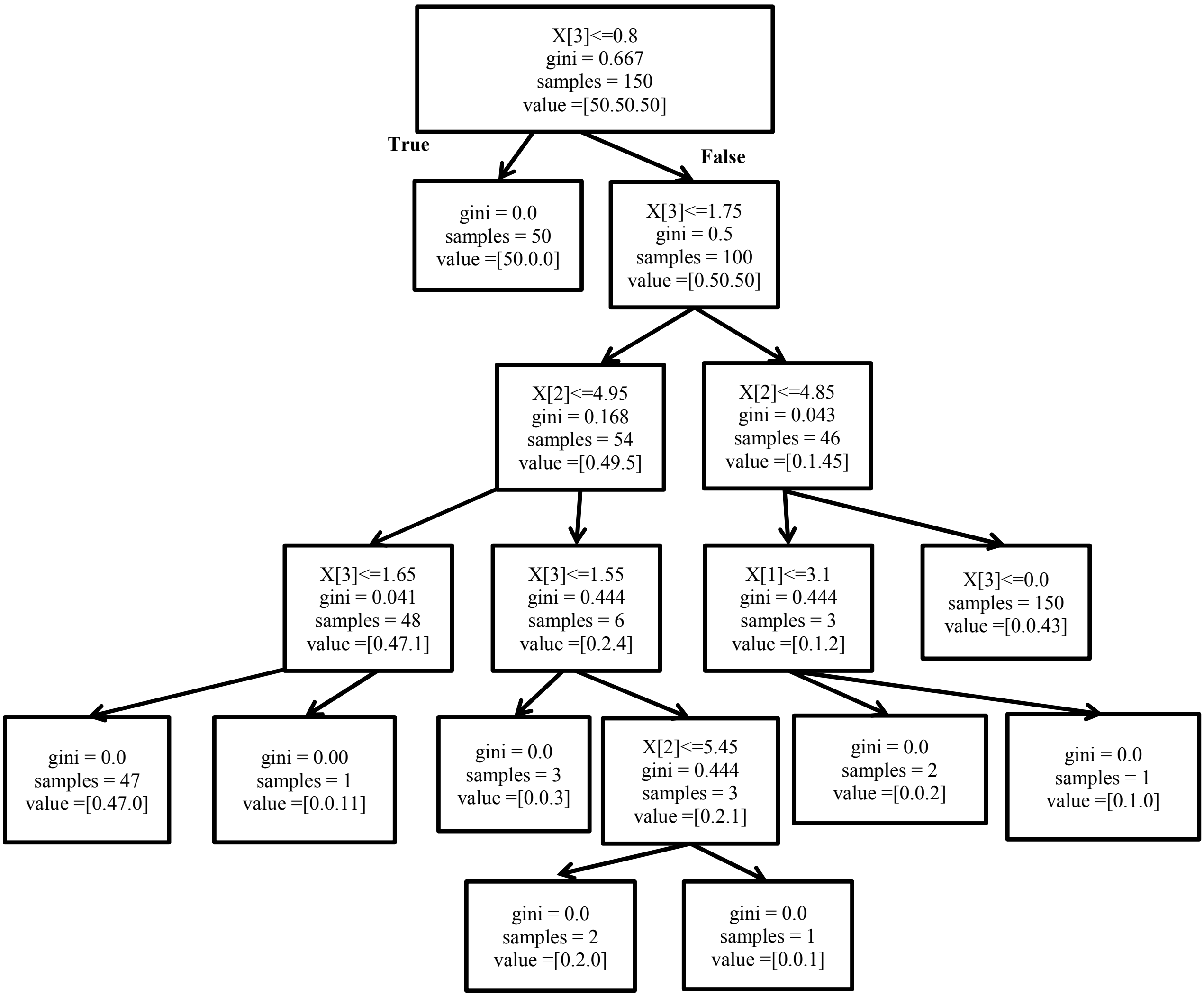

Fig. 5. An example visualization of regression tree using iris data

VI. Results

The following results are evaluating using testing data as input to multiple linear regression, random forest regression, and gradient boosted regression trees. The results are then compared using mean absolute error as a criterion. Table III shows mean absolute error (5) of multiple linear regression, random forest regression, and gradient boosted regression trees, in order. Gradient boosted regression yield the best performance with only MSE =0.28. Random forest regression is in second place with MSE = 0.35. Multiple linear regression has relatively large MSE of 0.55 when compared with the other.

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i| \quad (5)$$

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

IV. Conclusion

In this research, authors conducted a comparative study on regression based model performance. Data used in this research are scraped from German e-commerce site and then data preparation processed by using python programming language. As a result, final data have 304,133 rows and 11 attributes. We tested data by using multiple linear regression, random forest regression, and gradient boosted regression trees on that particular dataset. Each model was evaluated by using the same testing data. The results are then compared by using mean absolute error as a criterion. With gradient boosted regression trees gave the highest performance with only MAE =0.28. Followed by random forest regression with 0.35 errors, and multiple linear regression with 0.55 errors. Thus, we concluded that gradient boosted regression trees is recommended to develop the price evaluation model. The future work can be developed from this research by fine tuning each model parameter. More appropriate data engineering can be utilize to create the better training data. As mentioned in Section III, one hot encoding can be used as an alternative to label encoding for more realistic data interpretation on categorical data. The models can also be implemented in real world application, however, more polishing is required.

References

[1] S. Pudaruth, “Predicting the Price of Used Cars using Machine Learning Techniques,” International Journal of Information & Computation Technology, vol. 4, no. 7, pp. 753–764, 2014.

[2] 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.

[3] 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.

[4] 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.

[5] G.Rossum, “Python Reference Manual,” Amsterdam, The Netherlands, The Netherlands, Tech. Rep., 1995.

[6] 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.

[7] 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

[8] 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

[9] R.Taylor, “Interpretation of the Correlation Coefficient: A Basic Review,” Journal of Diagnostic Medical Sonography, vol. 6, no. 1, pp. 35–39, 1990.

[10] 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.

[11] 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