International Journal of Computational and Experimental Science and ENgineering (IJCESEN)

Vol. 9-No.1 (2023) pp. 11-16
http://dergipark.org.tr/en/pub/ijcesen

ISSN: 2149-9144

Research Article

Using Linear Regression For Used Car Price Prediction

Sümeyra MUTİ[1], Kazım YILDIZ[2,*]

[1]Marmara University, Institute of Pure and Applied Sciences, Department of Computer Engineering, Istanbul, Turkiye
Email: sumeyramuti@marun.edu.tr **ORCID:** 0000-0001-6489-0258

[2] Marmara University, Technology Faculty, Department of Computer Engineering, 34722, Istanbul, Turkiye
* **Corresponding Author Email** : kazim.yildiz@marmara.edu.tr **ORCID:** 0000-0001-6999-1410

Article Info:

DOI: 10.22399/ijcesen.1070505
Received : 09 February 2022
Accepted : 26 February 2023

Keywords

Machine Learning
Car Price Prediction
Linear Regression

Abstract:

Recently, there have been studies on the use of machine learning algorithms for price prediction in many different areas such as stock market, rent a house and used car sales. Studies give information about which algorithm is more successful in price prediction using different machine learning methods. The most commonly used method for price prediction is the linear regression model. In this study, the effectiveness of the linear regression model was examined for used car price prediction. The linear regression model was applied to the data set that includes the features and price information of vehicles in Turkey as the year 2020. As a result, when we selected 1/3 of the data set as the test data, it was observed that the R2 score for the prediction success of model was 73%. To improve the effectiveness of the results the dataset could be extend or preprocessing part be detailed.

1. Introduction

Selling process of a vehicle, the most important issue is to determine the most affordable price without giving a price below the value of the car. If a value is given above the market price, the probability of selling the car will either decrease or the selling time may be longer. In addition, no one, who wants to buy a used car, does not want to own a car by paying more than it is worth. The price of a used vehicle may vary depending on attributes such as the vehicle's model, year, number of kilometers, gear type, damage record, color, and additional comfort features. Using this data on a used car sale site, it is possible to calculate how much a used vehicle can be sold according to its features, using machine learning algorithms.

Machine learning methods are algorithms that process the given data to perform a specified task without being fully programmed, learn with this data and improve itself as a result of learning [1]. There are studies on the use of machine learning algorithms in many different areas such as the detection of COVID-19 cases [2], crop yield analysis [3], visiting time prediction [4], predicting motor vehicle theft [5]. In this study, we will apply the linear regression (LR) method as a machine learning method for used car price prediction. LR is both a statistical method and a machine learning method. It is aimed to create models to identify the relationship between input and output variables [6].

There are many studies using various machine learning techniques in price estimation applications. Home price prediction [7-10], stock price prediction [11-14], stock market prediction [15], bitcoin price prediction [16], price estimate for vacation rentals [17], car price prediction [18-19] are some of them. In this study the main focus is on used vehicle price estimation. Comparison of LR, Artificial Neural Networks (ANN) and Support Vector Machine (SVM) algorithms for price estimation of used truck models is one of these studies [20]. In another study, the success of the LR method for used car price estimation has been examined [21].

In a study for the estimation of house prices, different machine learning algorithms has been applied by analyzing the housing data of 5359 townhouses in Virginia collected by the Multiple Listing Service (MLS) of Metropolitan Regional Information Systems (MRIS). Data collected

through WEKA then classified various machine learning algorithms such as C4.5, RIPPER Naïve Bayesian and AdaBoost. It has been observed that the RIPPER model gives more successful results compared to others[7].

In another home price estimation study, Support Vector Regression (SVR) and LR methods were compared and it was found that LR had lower error rates [22]. The success of RF and LR methods in home price estimation was compared. It was observed that the training data being more than the test data increased the estimation success of the system and it was concluded that the LR method, which includes 80% data training and 20% data testing, gives lower Absolute Error and RMSE values [23]. A model was developed which predicts prices by using both textual and visual data of houses in housing price estimation. In order to determine the luxury level of the house, the KNN, SVM and Convolutional Neural Network (CNN) methods applied on the visual data were compared and it was observed that the CNN method gave the lowest Median Error Rate. It has been observed that the created model is more successful than Zillow's Zestimate method [24]. In another application stock price estimation was made, the decision tree model, multiple regression and random forest algorithms applied to five different stocks tried to predict the closing prices of the stocks the next day and the experimental results of these algorithms are successful [12].

In a study in which LR and SVM algorithms used to predict stock prices, it has been that LR give better results than SVM [15].

There is a study examining the effect of various machine learning algorithms on price estimation for Airbnb, which offers vacation rental services. In this study, LR, tree-based models, Decision Support Machines and ANN methods were applied to get the best results in terms of Mean Square Line, Mean Absolute Error (MAE) and R2 score. Among the methods tested, SVR gave the best result with an R2 score of 69% [17].

In a study conducted for vehicle price estimation, some machine learning algorithms are used and compared for an application that finds the best price that a truck company can give when buying used vehicles from customers. For this study, the previous offers of the company were used as the data set. LR, ANN and SVM were used for price estimation and the most successful results were obtained by applying 90% data splitting method with the SVM algorithm [20].

In another application, LR method was applied for used car price estimation. The data of 5041 used cars, in which 23 features that affect the price (such as brand and model) were selected, were used and the explanatory rate of these features was found to be 89.1%. The data set was divided into two, half of which was used as training data and the other half as test data, and when the results were compared, the prediction success rate was found to be 81.15%. More successful results can be obtained with more training data [21]. In another application for car price estimation, price prediction success was examined by applying various regression techniques on a data set containing information such as fuel type, mileage and model of vehicles collected from an e-commerce site called Avito. As a result, it was revealed that the Gradient Boosting Regression (GBR) method had the highest R2 score and the least MAE value [25].

In this paper, Section 2 describes the methodology and material, the research results was given in section 3. Finally conclusion and future work details were given.

2. Material and Methods

2.1. Dataset

As shown in Table 1, the dataset contains information about 15 attributes of used car in Turkey in 2020, including car brand, date of announcement, vehicle type group (such as Clio), vehicle type (such as 520i Standard), model year, fuel type, gear type, CCM, horsepower, color, body type (such as sedan), from whom it is sold (from the gallery) etc.), condition (used or new), mileage, price [26]. Since the date of the announcement and the information from whom it was sold have no effect on the price or are negligible, they will be removed from the data set. Horse power column has 61% “unknown” data so will not be included in this study. Status information, on the other hand, will be used distinctively since we will only make price estimations on used vehicles, and only data with second-hand status will be used in the data set. Also CCM data will be ignored.

In this study, the LR algorithm was examined for used car price estimation. Classification and LR processes will be performed with Python [27] in the JupyterLab environment.

2.2. Data Preprocessing

Data preprocessing is used to improve the performance of machine learning methods, especially when it comes to classification. The dataset may initially be noisy and inconsistent. With cleaning activities, such as removing outliers and correcting noisy values, the data set is made suitable for machine learning applications [28].

Table 1. *Sample of the data set used for car price prediction*

Date Annou	Brand	Vehicle Type Group	Vehicle Type	Model Year	Fuel Type	Gear Type	CCM	Horse Power	Color	Body Type	From Who	State	Km	Price
27.05. 2020	Jaguar	XF	2.0 D Prestige Plus	2017	Diesel	Autom atic	1801-2000 cc	176-200 HP	Navy Blue	Hatachback 5 Door	Galery	Used	26100	634500
16.06. 2020	Acura	CL	-	2015	Diesel	Semi-automa tic	1301-1600cc	101-125 HP	Blue	Sedan	Owner	Used	127000	151500
14.06. 2020	Acura	CL	2.2	1194	Benzine/ LPG	Manuel	1301-1600cc	101-125 HP	Turquoise	Sedan	Owner	Used	175000	19750
11.06. 2020	Acura	CL	-	2013	Diesel	Manuel	1301-1600cc	76-100 HP	Brown	Sedan	Owner	Used	325	52000
11.06. 2020	Acura	CL	2.2	2010	Diesel	automa tic	1801-2000cc	151-175 HP	White	Sedan	Owner	used	207000	148750

Incorrectly added data should be removed from the data set, as it will reduce the probability of the system guessing correctly. Similarly, removing data that has no effect on the estimation from the data set will increase the success of the algorithm. Outliers, which are far from other data in the data set, are also among the factors that reduce the success of the algorithm. It is possible to reduce these negative effects with data preprocessing.

We removed the information that was not useful to us, such as "Unknown", "-" from the data set. Then, we detected the outlier data of “price” attribute and removed from the dataset. We found that the price difference was large in two vehicles of the “Jaguar” brand with the same features, and we removed the low price from the dataset. We observed that the accuracy of the prediction decrease when the “price” value rises above 200000, so we also removed the data from the dataset with the “price” value above 200000. Since some mileage information is too low for a used vehicle, we removed lines below 1000 km from the dataset.

We converted all our data into numerical values in order to use machine learning algorithms. After these processes, the numbers of data features in the dataset are shown in Figure 3 and the relationship of the features with each other is shown in Figure 1.

3. Results and Discussions

We partition our dataset as 1/3 test data and train it with LR model. The relationship between the estimated price values and the actual price values in Figure 2 shows that the LR model is suitable for price prediction. The blue solid line shows the regression line and the red dots around it show the intersection of the estimated values and the real values. The closer the red dots are to the blue line, the more accurate the algorithm is making predictions.

As a result, we found that the R2 score [29] which gives the relationship between the real and estimated values of the LR model we applied to our data set was 0.73. While our model success was around 0.62 at the beginning, we observed that our model success reached this value with the data preprocessing process.

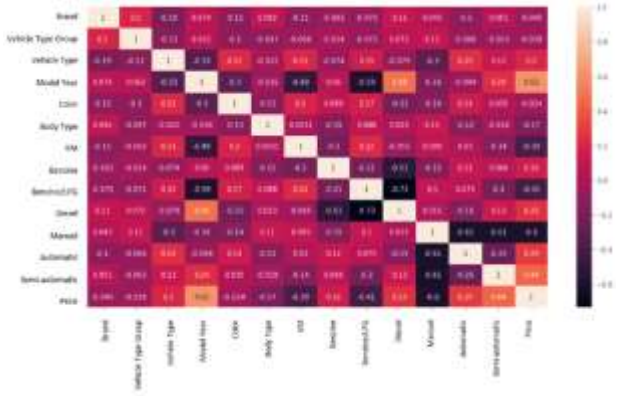

Figure 1. *Correlation of attributes*

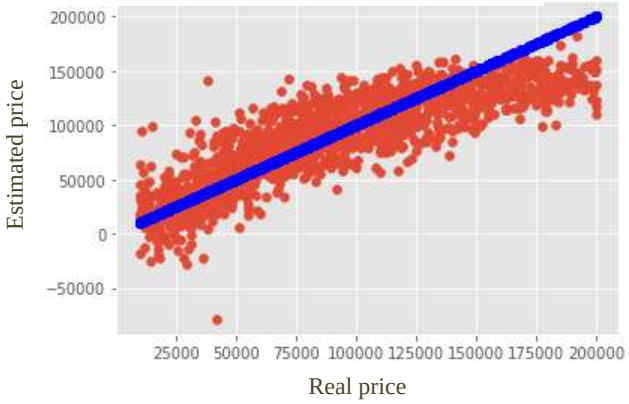

Figure 2. *Correlation between estimated and real price values*

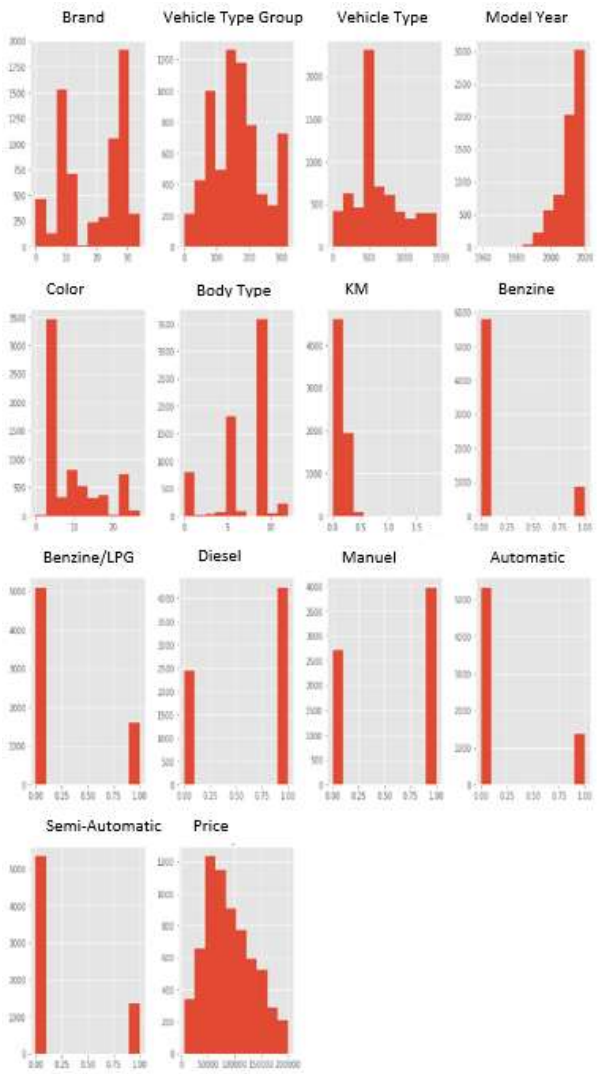

Figure 3. *Amount of attributes in the dataset*

Table 2 shows the comparison of the estimated price values of our model with the actual price values. Accordingly, we see that the difference between model estimates and actual values is quite close in some areas, and this difference increases in some areas. For example, the estimated price of the vehicle whose actual price is 56750 in line 6 is found 60003, and a successful estimate was obtained with a difference of approximately 3000 liras. On the other hand, there is a difference of approximately 40000 TL between the estimate and the actual value in the 2nd line. A more detailed examination of the data preprocessing section or the use of a more suitable data set can increase the success of the model.

4. Conclusions

LR method is a suitable model for price estimation, as features such as brand, km, year, gear type of used vehicles directly affects their price. We observed that the success of the model increased when we removed the non-logical values in the data set.

Table 2. *Real price vs estimated price values*

	Real value	Prediction
0	4800	31320.27
1	145000	105211.80
2	49250	64534.34
3	118000	133167.16
4	103500	108646.10
5	56750	60063.20
6	104000	113513.86
7	32900	38927.42
8	142950	135422.77
9	60000	67108.96
10	81650	91284.90
11	122500	125503.39
12	93000	100478.09
13	54900	70027.75
14	93500	120581.50
15	52950	83542.02
16	140500	131251.74
17	65500	109028.45
18	103500	124890.43
19	155000	109971.27

While our R^2 score was around 0.62 in the first stage, it increased to 0.73 when we removed the data that reduced the predictive ability of the model in the data preprocessing section. As seen in Figure 2, the relationship between estimated and actual price values shows that LR is suitable for used car price estimation. More appropriate datasets can be used to obtain better results.

In order to examine the success of LR with more up-to-date data, a study can be done by pulling the appropriate data from used car buying sites.

Author Statements:

- **Ethical approval:** The conducted research is not related to either human or animal use.
- **Conflict of interest:** The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper
- **Acknowledgement:** The authors declare that they have nobody or no-company to acknowledge.
- **Author contributions:** The authors declare that they have equal right on this paper.

- **Funding information:** The authors declare that there is no funding to be acknowledged.
- **Data availability statement:** The data that support the findings of this study are available n request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

References

[1] I. E. Naqa and M. J. Murphy. (2015). What Is Machine Learning ?. *Machine Learning in Radiation Oncology* 3;11 DOI 10.1007/978-3-319-18305-3_1.

[2] N. S. Özen, S. Saraç and M. Koyuncu, (2021). COVID-19 Vakalarının Makine Öğrenmesi Algoritmaları ile Tahmini: Amerika Birleşik Devletleri Örneği. *European Journal of Science and Technology (EJOSAT)*. 22;134-139 DOI: 10.31590/ejosat.855113.

[3] F. F. Haque, A. Abdelgawad, V. P. Yanambaka and K. Yelamarthi. (2020). Crop Yield Analysis Using Machine Learning Algorithms. *IEEE 6th World Forum on Internet of Things (WF-IoT)*. pp. 1-2, DOI: 10.1109/WF-IoT48130.2020.9221459.

[4] I. Hapsari, I. Surjandari and K. (2018). Visiting Time Prediction Using Machine Learning Regression Algorithm. *6th International Conference on Information and Communication Technology (ICoICT)*. pp. 495-500, DOI: 10.1109/ICoICT.2018.8528810.

[5] N. Nafi'iyah and K. F. Mauladi. (2021). Linear Regression Analysis and SVR in Predicting Motor Vehicle Theft. *International Seminar on Application for Technology of Information and Communication (iSemantic)*, pp. 54-58 DOI: 10.1109/ISEMANTIC52711.2021.9573225.

[6] Ms Kavita and P. Mathur. (2020). Crop Yield Estimation in India Using Machine Learning. *5th International Conference on Computing Communication and Automation (ICCCA)*, pp. 220-224. DOI: 10.1109/ICCCA49541.2020.9250915.

[7] J. K. Bae and B. Park. (2015). Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data. *Expert systems with applications*. 42(6);2928-2934, DOI: 10.1016/j.eswa.2014.11.040.

[8] A. Varma, A. Sarma, S. Doshi and R. Nair. (2018). House Price Prediction Using Machine Learning and Neural Networks. *Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*, pp. 1936-1939, DOI: 10.1109/ICICCT.2018.8473231.

[9] I. Imran, U. Zaman, M. Waqar, A. Zaman. (2021) Using Machine Learning Algorithms for Housing Price Prediction: The Case of Islamabad Housing Data. *Soft Computing and Machine Intelligence*, 1(1);11-23.

[10] B. Jia (2021). Computer mathematical statistics applied in the housing price investigation through machine learning and linear regression model. *International Conference on Data Science and Computer Application (ICDSCA)*, pp. 769-772, DOI: 10.1109/ICDSCA53499.2021.9650136.

[11] C. K.-S. Leung, R. K. MacKinnon and Y. Wang. (2014). A machine learning approach for stock price prediction. *IDEAS '14: Proceedings of the 18th International Database Engineering & Applications Symposium*, pp. 274-277, DOI:10.1145/2628194.2628211.

[12] Z. D. Akşehir and E. Kılıç, (2019). Makine Öğrenmesi Teknikleri ile Banka Hisse Senetlerinin Fiyat Tahmini. *Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi*, 12(2);30.

[13] M. Nikou, G. Mansourfar, and J. Bagherzadeh (2019). Stock price prediction using DEEP learning algorithm and its comparison with machine learning algorithms. *Intelligent Systems in Accounting, Finance and Management*, 26(4)164-174, DOI: 10.1002/isaf.1459.

[14] W. Lu, W. Ge, R. Li and L. Yang. (2021). A Comparative Study on the Individual Stock Price Prediction with the Application of Neural Network Models. *International Conference on Computer Engineering and Artificial Intelligence (ICCEAI)*, pp. 235-238, DOI: 10.1109/ICCEAI52939.2021.00046,

[15] B. Panwar and P. J. Gaurav Dhuriya. (2021). Stock Market Prediction Using Linear Regression and SVM. *International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, pp. 629-631, DOI: 10.1109/ICACITE51222.2021.9404733.

[16] V. Siddhi, S. Valecha and M. Shreya. (2018). Bitcoin price prediction using machine learning. *20th International Conference on Advanced Communication Technology (ICACT)*, pp. 144-147, DOI: 10.23919/ICACT.2018.8323676.

[17] P. R. Kalehbasti, L. Nikolenko and H. Rezaei. (2021). Airbnb Price Prediction Using Machine Learning and Sentiment Analysis. *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, pp. 173-184, DOI: 10.1007/978-3-030-84060-0_11.

[18] E. Gegic, B. Isakovic, D. Keco, Z. Masetic and J. Kevric, (2019). Car Price Prediction using Machine Learning Techniques. *TEM Journal*, 8(1)113, DOI: 10.18421/TEM81-16.

[19] S. Selvaratnam, B. Yogarajah, T. Jeyamugan and N. Ratnarajah, (2021). Feature selection in automobile price prediction: An integrated approach. *International Research Conference on Smart Computing and Systems Engineering (SCSE)*, 4;106-112, DOI: 10.1109/SCSE53661.2021.9568288.

[20] Namlı, E., Gül, E., and Ünlü, R., (2019). Fiyat Tahminlemesinde Makine Öğrenmesi Teknikleri ve Doğrusal Regresyon. *Konya Mühendislik Bilimleri Dergisi*, 7;806-821, DOI: 10.36306/konjes.654952.

[21] Çelik, Ö and Osmanoğlu, U.Ö, (2019). Prediction of The Prices of Second-Hand Cars. *European Journal of Science and Technology (EJOSAT)*, 16;77-83, DOI: 10.31590/ejosat.542884.

[22] Amaresh, V., Singh, R. R., Kamal, R., & Kulkarni, A. (2022). Linear Regression Models based Housing

Price Forecasting. *2022 International Conference on Industry 4.0 Technology (I4Tech)* (pp. 1-5). IEEE.

[23] Septianingrum, S. A., Dzikri, M. A., Soeleman, M. A., Pujiono, P., & Muslih, M. (2022). Performance Analysis of Multiple Linear Regression and Random Forest for an Estimate of the Price of a House. *International Seminar on Application for Technology of Information and Communication (iSemantic)* (pp. 415-418). IEEE.

[24] Nouriani, A., & Lemke, L. (2022). Vision-based housing price estimation using interior, exterior & satellite images. *Intelligent Systems with Applications*, 14;200081.

[25] M. Hankar, M. Birjali and A. Beni-Hssane (2022). Used Car Price Prediction using Machine Learning: A Case Study. *11th International Symposium on Signal, Image, Video and Communications (ISIVC)*, El Jadida, Morocco, pp. 1-4, doi: 10.1109/ISIVC54825.2022.9800719.

[26] KAGGLE (10.10.2021), https://www.kaggle.com/alpertemel/turkey-car-market-2020.

[27] G. V. Rossum, Python Development Team (2020). Python Tutorial Release 3.8.1 *The Python Software Foundation.*

[28] L. Moreira, C. Dantas, L. Oliveira, J. Soares and E. Ogasawara, (2018). On Evaluating Data Preprocessing Methods for Machine Learning Models for Flight Delays. *International Joint Conference on Neural Networks (IJCNN)*, pp. 1-8, DOI: 10.1109/IJCNN.2018.8489294.

[29] D. Chicco, M. J. Warrens and G. Jurman, (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *Peer J Computer Science*, 7; DOI: 10.7717/peerj-cs.623.