Journal of Analysis and Computation (JAC)
(An International Peer Reviewed Journal), www.ijaconline.com, ISSN 0973-2861
Volume XVII, Issue I, Jan-June 2023

CAR PRICE PREDICTION IN MACHINE LEARNING USING PYTHON

Divya Katkar[1], Renuka Vanamala[1], Siddhanti Pampattiwar[1], Pinky Gangwani[2]

[1]Student, Department of Computer Engineering, Cummins College of Engineering for Women, Nagpur, India
[2]Assistant Professor, Department of Computer Engineering, Cummins College of Engineering for Women, Nagpur, India

ABSTRACT

This work addresses the problem of car price prediction in Machine Learning; this work is an effort that tries to study and investigate the trends in used car prices and predicts the price of used cars with the help of supervised machine learning algorithms. And to suggest which machine learning algorithm performs well among the selected methods for predicting the cars price. We wanted to study which algorithm predicts the car price more reliably and accurately, So that this solution will be helpful for first time used car buyers and also for sellers for determining the selling cost of the car. For this research work and to predict the prices we have considered different machine learning regression models which are Linear Regression, Lasso Regression and Random Forest Regression. The research objective of this work is to predict used cars prices using machine learning techniques, by collecting data from websites like Kaggle, and analyzing the different aspects and factors that lead to the actual used car price valuation and To enable consumers to know the actual worth of their car or desired car, by simply providing the program with a set of attributes from the desired car to predict the car price. While buying a car it is very important to know its worth, so to make this work easy we may also use other more advanced regression models such as XGBoost Regression and so on for more better results and also we may add large historical data of car price which can help to improve accuracy of the machine learning model.

Keywords - Car Price Prediction, Machine Learning, Linear Regression, Lasso Regression, Random Forest Regression

[1] INTRODUCTION

Predicting the price for new vehicles is more interesting and needed problem by many users. The data set has been collected from an online website, Kaggle. We have chosen to work with linear regression, Lasso regression and Random Forest models since we feel these are the basic models that can predict approximately and also to compare the result and explore which model gives the better R squared value. We imported various dependencies or can be called as

python libraries since the project is performed using python. The technologies or libraries used in this work are Pandas, Matplotlib, NumPy, Sklearn, Seaborn and the considered machine learning algorithms.

After the dataset collection and importing libraries, we analysed the dataset by some data visualization techniques. We checked if there were any missing values or null values and also the number of attributes and tuples. We visualized the data in the form of graphs using matplotlib library.

Data Cleaning: Cleaning data with a data cleaning library like NumPy, pandas for Datasets, and NumPy for undesired data. Pre-Processing Data: In order to use the Machine Learning models, we must convert these categorical variables to numerical variables. The Sklearn module Label Encoder was used to tackle this problem.

Data for Training and Testing: In this process, 10% of the data was split for testing purposes and 90% of the data was used for training. That should help the management to understand how exactly the prices vary with the independent variables. They can accordingly manipulate the design of the cars, the business strategy etc. to meet certain price levels. Training all the regression models by importing them and finding all the R squared errors was next step of this work. After obtaining all the R squared errors we examined which error was close to 1 as the model with which the error value was closed to 1 is the one which is good fit for the prediction and fits data good.

As per the results Random Forest Regression outperformed well. This work should help the management to understand how exactly the prices vary with the independent variables. They can accordingly manipulate the design of the cars, the business strategy etc. to meet certain price level.

[2] LITERATURE REVIEW

Several studies and related works have been done previously to predict used car prices around the world using different methodologies and approaches, with varying results of accuracy from 50% to 90%.

Researcher [1] proposed to predict used car prices in Mauritius, where he applied different machine learning techniques to achieve his results like decision tree, K-nearest neighbours, Multiple Regression and Naïve Bayes algorithms to predict the used cars prices, based on historical data gathered from the newspaper. Achieved results ranged from accuracy of 60-70 percent, the author suggested using more sophisticated models and algorithms to make the evaluation, with the main weakness off the decision tree and naïve Bayes that it is required to discretize the price and classify it which accrue to more inaccuracies. Moreover, he suggested a larger set of data of data to train the models hence the data gathered was not sufficient.

Researchers [2] were able to achieve high level of accuracy using Multiple linear regression models topredict the price of cars collected from used cars website in Pakistan called Pak Wheels that totalled to 1699 records after pre-processing, and where able to achieve accuracy of 98%, this was done after reducing the total amount of attributes using variable selection technique to include significant attributes only and to reduce the complexity of the model.

Researchers [3] used a supervised learning methodknown as Random Forest. Kaggle's dataset was used as a basis for predicting used car prices. In order to determine the price impact of each feature, careful exploratory data analysis was performed. 500 DecisionTrees

were trained with Random Forests. It is most commonly used for classification, but they turned it into a regression model by transforming the problem into an equivalent regression problem. Using experimental results, it was found that training accuracy was 95.82%, and testing accuracy was 83.63%. By selecting the most correlated features, the model can accurately predict the car price. Hence, from all literature review it is concluded that used cars price prediction is an important topic whichis the area of many researchers nowadays.

So far, the best achieved accuracy is 83.63% on kaggle's datasetusingrandom forest technique. The researchers have tested multiple regressors and final model is regressionmodel using linear regression.

Researchers [4] usedthree different machine learning techniques to predict used car prices. Using data scrapped from a local Bosnian website for used cars totalled at 797 car samples after pre-processing, and proposed using these methods: Support Vector Machine, Random Forest and Artificial Neural network. Results have shown using only one machine learning algorithm achieved results less than 50%, whereas after combing the algorithms with pre calcification of prices using Random Forest, results with accuracies up to 87.38% was recorded.

Researchers [5] proposed using Supervised machine leaning model using K-Nearest Neighbour to predict used car prices from a data set obtained from Kaggle containing 14 different attributes,using this method accuracy reached up to 85% after different values of K as well as Changing the percent of training data to testing data, expectedly when increasing the percent of data that is tested better accuracyresults are achieved. The model was also cross validated with 5 and 10 folds by using K fold method.

[3] TECHNOLOGIES

There are a lot of libraries that we are importing. Using libraries make things easy as it is not needed to write the code of their functionality from scratch.

- ✓ **Pandas** is a python library used for working on big data sets. It has functions like analyzing, cleaning, exploring and manipulating data. It allows us to analyze big data and make conclusion based on statistical theories. It cleans messy data sets, and make them readable and relevant.
- ✓ **Matplotlib** is a visualization library in python for 2D plots arrays. One of the greatest benefits of visualization is that it allows us visual access to huge amounts of data in easily digestible visuals. Matplotlib consists of several plots like line, bar, scatter, histogram, etc.
- ✓ **Seaborn** is a library that uses matplotlib underneath to plot graphs. It provides a high-level interface that is used to draw informative statistical plots. Functions in Seaborn library expose a declarative, dataset-oriented API that makes it easy to translate questions about data into graphics that can answer them.
- ✓ **NumPy** is used for performing a wide variety of mathematical operations for arrays and matrices. It aims to provide an array object that is up to 50x faster than tradition python lists. It is a python library and is written partially in python, but most of the parts that require fast computation are written in C or C++.

- ✓ **Sklearn** is one of the most popular and useful libraries that are used for machine learning in python. It provides a list of efficient techniques and tools for machine learning and statistical modeling including classification, regression, clustering, and dimensionality reduction.

Used three regression models to predict car prices are Linear Regression, Lasso Regression models and Random forest Regression.

[4] METHODOLOGY

There are two primary phases in the system:
1. Training phase: The system is trained by using the data in the data set and fits a model (line/curve) based on the algorithm chosen accordingly.
2. Testing phase: the system is provided with the inputs and is tested for its working. The accuracy is checked. And therefore, the data that is used to train the model or test, it has to be appropriate.
The system is designed to detect and predict price of used car and hence appropriate algorithms must be used to do the two different tasks. Before the algorithms are selected for further use, different algorithms were compared for its accuracy. The well-suited one for the task was chosen.

Linear Regression:

Linear Regression attempt to model the relationship between two variables by fitting a linear equation to observed data. The other is considered to be dependent variable. For Example: A modeller might want to relate weights of individuals to their heights using a linear regression model Linear regression is useful for finding relationship between multiple continuous variables There are multiple independent variables and single independent variable.

Lasso Regression:

Lasso regression is a regularization technique. It is used over regression methods for a more accurate prediction. This model uses shrinkage. Shrinkage is where data values are shrunk towards a central point as the mean. The lasso procedure encourages simple, sparse models (i.e. models with fewer parameters). This particular type of regression is well-suited for models showing high levels of multi collinearity or when you want to automate certain parts of model selection, like variable selection/parameter elimination.

Random Forest Regression:

Random Forest Regression is a supervised learning algorithm that uses ensemble learning method for regression. Ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model. A Random Forest Regression model is powerful and accurate. It usually performs great on many problems, including features with non-linear relationships. Disadvantages, however, include the following: there is no interpretability, overfitting may easily occur, we must choose the number of trees to include in the model.

[5] RESULTS AND DISCUSSION

This section explains final output, which predicts car price:

1. Linear Regression :

Fig: 1 Linear Regression on training data

The above figure depicts the visual representation of car price after applying Linear Regression to Training data and comparing the actual price and predicted price of car.

Fig: 2 Linear Regression on testing data

The above figure depicts the visual representation of car price after applying Linear Regression to Test data and comparing the actual price and predicted price of car.

2. Lasso Regression :

Fig: 3 Lasso Regression on training data

The above figure depicts the visual representation of car price after applying Lasso Regression to Training data and comparing the actual price and predicted price of car.

Fig: 4 Lasso Regression on testing data

The above figure depicts the visual representation of car price after applying Lasso Regression to Test data and comparing the actual price and predicted price of car.

3. Random Forest Regression :

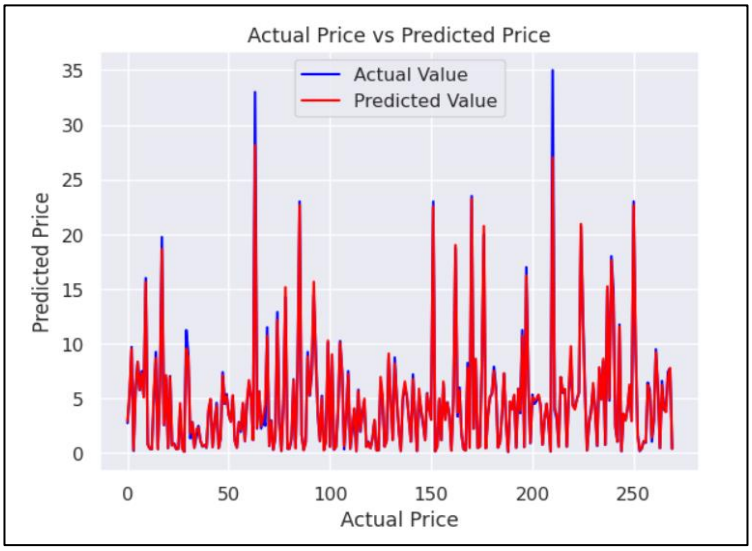

Fig: 5 Random Forest Regression in Training data

The above figure depicts the visual representation of car price after applying Random Forest Regression to Training data and comparing the actual price and predicted price of car.

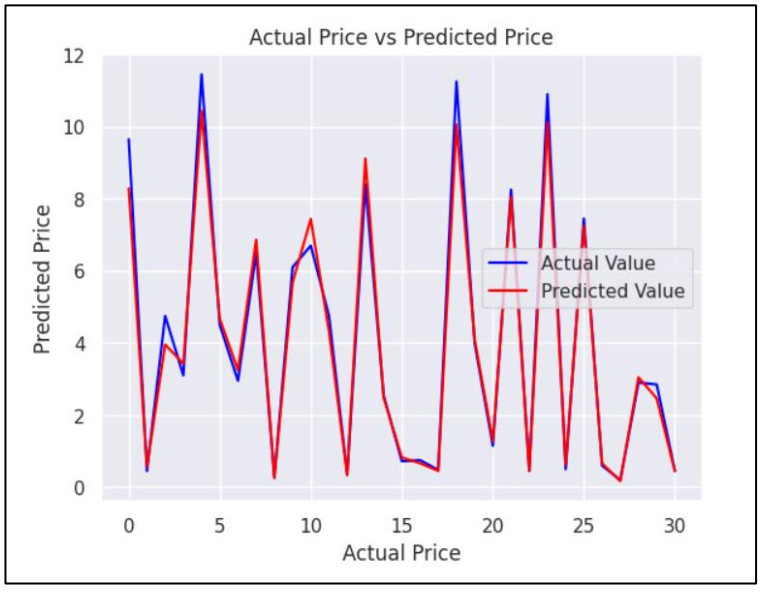

Fig: 6 Random Forest Regression on testing data

The above figure depicts the visual representation of car price after applying Random Forest Regression to Test data and comparing the actual price and predicted price of car.

The comparisons of all the experiments are shown below:

Table: 1 Results Comparison

SR	ALGORITHMS	R Squared error
1	Linear Regression	0.88
2	Lasso Regression	0.85
3	Random Forest Regression	0.98

Given the evaluation parameters the Random Forest Regression outperformed as it has the highest R squared error of the three different algorithms as 0.98 which is close to 1 and proves that it fits good, as well as the lower error in all three-evaluation parameter. Second in accuracy is the linear regression with 0.88 R squared error, even though it has a higher error parameter than linear regression. Least accurate was the Lasso regression with 0.85 R squared error thought it had a lower error value than Linear Regression.

[5] RESULTS AND DISCUSSIONS

The price prediction of second-hand items has not been widely addressed, which was the main motivation for this research, as various sellers generate the price of the vehicle mainly by the manufacturer brand. Only a few studies have addressed the price prediction of used products in a specific domain, specifically, the price prediction of second-hand cars. In this paper, the proposed approach uses exploratory data analysis along with features extracted from actual and historical attributes to predict the future behavior of the used-cars market. The prediction model uses supervised machine learning techniques and validation methods regarding statistical outputs.

To summarize,

- ✓ Data were collected from an online seller of used cars and important features were identified that reflect the price;
- ✓ Non-available values and entries were removed, and we discarded features not relevant for the prediction of the price;
- ✓ Supervised Machine Learning techniques applied in first data set and validation was compared with the price prediction outputs of the second data set regarding important features;
- ✓ The predicted model has the highest accuracy with linear regression where main features (price and model) are available.

This is performed by considering different types of vehicles, their usage condition, and prices. Furthermore, different techniques for numeric data pre-processing as well as text analysis for handling the unstructured data are considered. The competitive advantage of second-hand market trend prediction achieved by data mining and analysis includes the optimal price for the

vehicle observations, avoiding misclassification and risks along with improving the customer's awareness of the market, leading to accurate buying decisions.

REFERENCES

[1] Sameerchand Pudaruth, "Predicting the Price of Used Cars using Machine Learning Techniques", International Journal of Information & Computation Technology, 2014, 754-764.
[2] Noor, K., & Jan, S., Vehicle Price Prediction System using Machine Learning Techniques. International Journal of Computer Applications, 2017, 27-31.
[3] Nabarun Pal, "A methodology for predicting used cars prices using Random Forest", Future of Information and Communications Conference, 2018.
[4] Enis Gegic, Becir Isakovic, Dino Keco, Zerina Masetic, Jasmin Kevric, "Car Price Prediction using Machine Learning Techniques", TEM Journal-2019.
[5] K.Samruddhi, & Kumar, D. R., Used Car Price Prediction using K-Nearest Neighbor Based Model. International Journal of Innovative Research in Applied Sciences and Engineering (IJIRASE), 4(3), 2020, 686-689.
[6] Listiani, M. (2009). Support vector regression analysis for price prediction in a car leasing application (Doctoral dissertation, Master thesis, TU Hamburg-Harburg)
[7] Aizerman, M. A. (1964). Theoretical foundations of the potential function method in pattern recognition learning. Automation and remote control, 25, 821- 837.
[8] Ho, T. K. (1995, August). Random decision forests. In Document analysis and recognition, 1995, proceedings of the third international conference on (Vol. 1, pp. 278-282). IEEE.
[9] International Journal of Information & Computation Technology. ISSN 0974-2239 Volume 4, Number 7 (2014), pp. 753-764 © International Research Publications House http://www. irphouse.com.
[10] Enis Gegic, Becir Isakovic, Dino Keco, Zerina Masetic, Jasmin Kevric International Burch University, Sarajevo, Bosnia and Herzegovina Car Price Prediction Using Machine Learning , June 2021| IJIRT | Volume 8 Issue 1 | ISSN: 2349-6002.
[11] Ceriottia, M. (2019). Unsupervised machine learning in atomistic simulations, between predictions and understanding. 150-155.
[12] Q. Yuan, Y. Liu, G. Peng, and B. Lv, "A prediction study on the car sales based on web search data," in The International Conference on E-Business and E-Government (Index by EI), 2011, p.5.
[13] Yang, R.R.; Chen, S.; Chou, E. AI Blue Book: Vehicle Price Prediction Using Visual Features. *arXiv* **2018**, arXiv:1803.11227.
[14] Chen, C.; Hao, L.; Xu, C. Comparative analysis of used car price evaluation models. *AIP Conf. Proc.* **2017**, *1839*, 020165.
[15] Liu, E.; Li, J.; Zheng, A.; Liu, H.; Jiang, T. Research on the Prediction Model of the Used Car Price in View of the PSO-GRA-BP Neural Network. *Sustainability* **2022**, *14*, 8993.
[16] Siva, R.; Adimoolam, M. Linear Regression Algorithm Based Price Prediction of Car and Accuracy Comparison with Support Vector Machine Algorithm. *ECS Trans.* **2022**, *107*, 12953–12964.
[17] Monburinon, N.; Chertchom, P.; Kaewkiriya, T.; Rungpheung, S.; Buya, S.; Boonpou, P. Prediction of prices for used car by using regression models. In Proceedings of the 2018 5th International Conference on Business and Industrial Research (ICBIR), Bangkok, Thailand, 17–18 May 2018; pp. 115–119.
[18] Bharambe, P.P.; Bagul, B.; Dandekar, S.; Ingle, P. Used Car Price Prediction using Different Machine Learning Algorithms. *Int. J. Res. Appl. Sci. Eng. Technol.* **2022**, *10*, 773–778.
[19] Puteri, C.K.; Safitri, L.N. Analysis of linear regression on used car sales in Indonesia. *J. Phys. Conf. Ser.* **2020**, *1469*, 012143.
[20] Alex M. Goh and Xiaoyu L. Yann, (2021), "A Novel Sentiments Analysis Model Using Perceptron Classifier" Int. J. of Electronics Engineering and Applications, Vol. 9, No. 4, pp. 01-10, DOI 10.30696/IJEEA.IX.IV.2021.01-10.
[21] Dolly Daga, Haribrat Saikia, Sandipan Bhattacharjee and Bhaskar Saha, (2021), "A Conceptual Design Approach For Women Safety Through Better Communication Design" Int. J. of Electronics Engineering and Applications, Vol. 9, No. 3, pp. 01-11, DOI 10.30696/IJEEA.IX.III.2021.01-11
[22] Alex M. Goh and Xiaoyu L. Yann, (2021), "Food-image Classification Using Neural Network Model" Int. J. of Electronics Engineering and Applications, Vol. 9, No. 3, pp. 12-22, DOI 10.30696/IJEEA.IX.III.2021.12-22

[23] Jeevan Kumar, Rajesh Kumar Tiwari and Vijay Pandey, (2021), "Blood Sugar Detection Using Different Machine Learning Techniques" Int. J. of Electronics Engineering and Applications, Vol. 9, No. 3, pp. 23-33, DOI 10.30696/IJEEA.IX.III.2021.23-33

[24] Nisarg Gupta, Prachi Deshpande, Jefferson Diaz, Siddharth Jangam, and Archana Shirke, (2021), " F-alert: Early Fire Detection Using Machine Learning Techniques" Int. J. of Electronics Engineering and Applications, Vol. 9, No. 3, pp. 34-43, DOI 10.30696/IJEEA.IX.III.2021.34-43.

[25] Reeta Kumari, Dr. Ashish Kumar Sinha and Dr. Mahua Banerjee, (2021), "A Comparative Study Of Software Product Lines And Dynamic Software Product Lines" Int. J. of Electronics Engineering and Applications, Vol. 9, No. 2, pp. 01-10, DOI 10.30696/IJEEA.IX.I.2021.01-10

[26] MING AI and HAIQING LIU, (2021), "Privacy-preserving Of Electricity Data Based On Group Signature And Homomorphic Encryption" Int. J. of Electronics Engineering and Applications, Vol. 9, No. 2, pp. 11-20, DOI 10.30696/IJEEA.IX.I.2021.11-20

[27] Osman Goni, (2021), "Implementation of Local Area Network (lan) And Build A Secure Lan System For Atomic Energy Research Establishment (AERE)" Int. J. of Electronics Engineering and Applications, Vol. 9, No. 2, pp. 21-33, DOI 10.30696/IJEEA.IX.I.2021.21-33.

[28] XIAOYU YANG, (2021), "Power Grid Fault Prediction Method Based On Feature Selection And Classification Algorithm" Int. J. of Electronics Engineering and Applications, Vol. 9, No. 2, pp. 34-44, DOI 10.30696/IJEEA.IX.I.2021.34-44.

[29] Xiong LIU and Haiqing LIU, (2021), "Data Publication Based On Differential Privacy In V2G Network" Int. J. of Electronics Engineering and Applications, Vol. 9, No. 2, pp. 34-44, DOI 10.30696/IJEEA.IX.I.2021.45-53.

[30] Mandava Siva Sai Vighnesh, MD Shakir Alam and Vinitha.S, (2021), "Leaf Diseases Detection and Medication" Int. J. of Electronics Engineering and Applications, Vol. 9, No. 1, pp. 01-07, doi 10.30696/IJEEA.IX.I.2021.01-07

[31] Pradeep M, Ragul K and Varalakshmi K,(2021), "Voice and Gesture Based Home Automation System" Int. J. of Electronics Engineering and Applications, Vol. 9, No. 1, pp. 08-18, doi 10.30696/IJEEA.IX.I.2021.08-18

[32] Jagan K, Parthiban E Manikandan B,(2021), "Engrossment of Streaming Data with Agglomeration of Data in Ant Colony" Int. J. of Electronics Engineering and Applications, Vol. 9, No. 1, pp. 19-27, doi 10.30696/IJEEA.IX.I.2021.19-27

[33] M. Khadar, V. Ranjith, K Varalakshmi (2021), "Iot Integrated Forest Fire Detection and Prediction using NodeMCU" Int. J. of Electronics Engineering and Applications, Vol. 9, No. 1, pp. 28—35, doi 10.30696/IJEEA.IX.I.2021.28-35.

[34] Gayathri. M, Poorviga. A and Mr. Vasantha Raja S.S, (2021), "Prediction Of Breast Cancer Stages Using Machine Learning" Int. J. of Electronics Engineering and Applications, Vol. 7, No. 1, pp. 36-42, doi 10.30696/IJEEA.IX.I.2021.36-42