iJRASET

International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 Issue: IV Month of publication: April 2022

DOI: https://doi.org/10.22214/ijraset.2022.41300

Used Car Price Prediction using Different Machine Learning Algorithms

Prof. Pallavi Bharambe[1], Bhargav Bagul[2], Shreyas Dandekar[3], Prerna Ingle[4]
[1]Assistant Professor, [2, 3, 4] Student Department of Computer Engineering, Shivajirao S. Jondhale College of Engineering, Mumbai University

Abstract: A car price prediction has been a high-interest research area, as it needed recognizable effort and knowledge of the field expert. This paper mainly focuses on working of three different kind regression algorithms which are used to predict price of a used car. In this project, We have Considered number of distinct attributes which are examined for the reliable and accurate prediction. To build a model for predicting the price of used cars we have used three different kinds of machine learning techniques which comes under supervised machine learning type of algorithm which are linear regression, lasso regression and ridge regression respectively. we have used Python libraries to design GUI for our project and some other machine learning related libraries like Numpy, Pandas, Sklearn etc. we have calculated and compared the accuracies of three machine learning algorithms. The accuracies for linear regression, lasso and Ridge regression were 83.65%, 87.09% and 84.00% respectively. The final main price is predicted according to lasso regression as it gives highest accuracy amongst three different algorithms.

Keywords: car price prediction, machine learning, Regression techniques, linear regression, lasso regression, ridge regression

I. INTRODUCTION

Car price prediction is anyhow interesting and popular problem. Accurate car price prediction involves expert knowledge, because price usually depends on many unique features and factors. Generally, most important ones are brand name and model, years, KMs driven and mileage. The fuel type used in the car as well as fuel consumption per mile highly affected price of a car due to often changes in the price of a fuel. Distinct features like exterior color, door number, type of transmission, dimensions, safety, air condition, interior, whether it has navigation or not will also results in the car price. In this paper, we applied dissimilar and unique methodologies in order to achieve higher precision of car price like Car Model Name, Year, Selling, Price, Present Price, KMs Driven, Fuel Type, Seller Type, Transmission Type, Owner Type. In this paper, we applied different methodologies and ML techniques and models in order to accomplish higher precision of the used car price prediction.

This paper is organized in the following manner: Section II having related work in the field of price prediction of used cars. In section III, the Existing system model is explained. Section IV explains the requirements and model of proposed system along with various ML algorithms to verify their respective performances to predict the price of the used cars. Finally, in section V, a conclusion of our work is given.

II. LITERATURE REVIEW

While making this project we examined many of research papers, these papers disclose the newest research during this field, a summary of some of the papers we referred to are mentioned below:

Enis Gegic, Becir Isakovic, Dino Keco, Zerina Masetic, Jasmin Kevric[1] During this paper , authors have first considered number of distinct attributes are examined for the reliable and accurate prediction. They built a model for predicting the price value of used cars, using three machine learning techniques (Artificial Neural Network, Support Vector Machine and Random Forest).The authors have compared Respective performances of various algorithms to search out one that best that most closely fits the available data set. The ultimate prediction model was integrated into Java application. Furthermore, the model was evaluated using test data and also the accuracy of 87.38% was obtained.

Pattabiraman Venkatasubbu, Mukkesh Ganesh[2] In this research, the authors attempt to construct a statistical model that would estimate the price of a used car based on previous customer data and a collection of attributes using Algorithms such as Lasso, Multiple regression and Regression Trees. The authors have also analysed the forecast accuracy of different models in order to calculate the car's price using an algorithm that is more accurate.

.

Sameerchand Pudaruth[3] The Author has researched the application of supervised machine learning techniques to predict the price of used cars in Mauritius. The predictions are based on historical data collected from daily newspapers. Different techniques like multiple linear regression analysis, k-nearest neighbours, naïve bayes and decision trees are accustomed to make the predictions. The predictions are then evaluated and compared so as tob search out those which offer the most effective performances. A seemingly easy problem clothed to be indeed very difficult to resolve with high accuracy. All the four methods which are used provided comparable performance.

K.Samruddhi, Dr. R.Ashok Kumar[4] During ths paper Authors proposed a supervised machine learning model using KNN (K Nearest Neighbor) regression algorithm to research the price value of used cars. Authors have trained the model with data of used cars which is collected from the Kaggle website. Through this experiment, the information was examined with different trained and test ratios. As a result, the accuracy of the proposed model is around 85% and is fitted because the optimized model.

Nabarun Pal, Dhanasekar Sundararaman, Priya Arora, Puneet Kohli, Sai Sumanth Palakurthy [5] During this paper, Authors have used supervised learning method namely Random Forest to predict the costs of used cars. The model has been chosen after careful exploratory data analysis to work out the impact of every feature on price. A Random Forest with 500 Decision Trees were created to train the data. From experimental results, the training accuracy was discovered to be 95.82%, and therefore the testing accuracy was 83.63%. The model can predict the price value of cars accurately by choosing the fore most correlated features.

III. EXISTING SYSTEM

The existing system includes a basic car price prediction using the inputted data and displays its output on simple static website and there is no use of different kinds of regression algorithms. There is only use of simple linear regression algorithm in the project. The existing gives less accuracy in result. The below mentioned flowchart includes the behavior of the existing system.

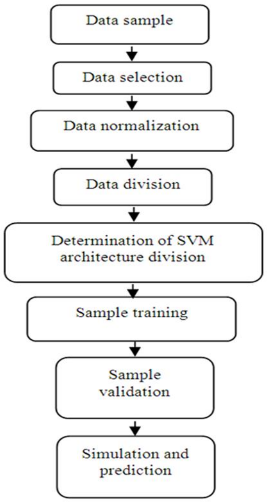


Fig.1 Existing System Architecture

IV. PROPOSED SYSTEM

In the proposed System, we have tried to solve the issues in existing system. We have developed a GUI. Also we have used three machine learning algorithms which are linear, lasso and ridge regression respectively in our system to predict the output more correctly with higher accuracy.

The Requirement analysis and model implementation of proposed system is briefly explained below.

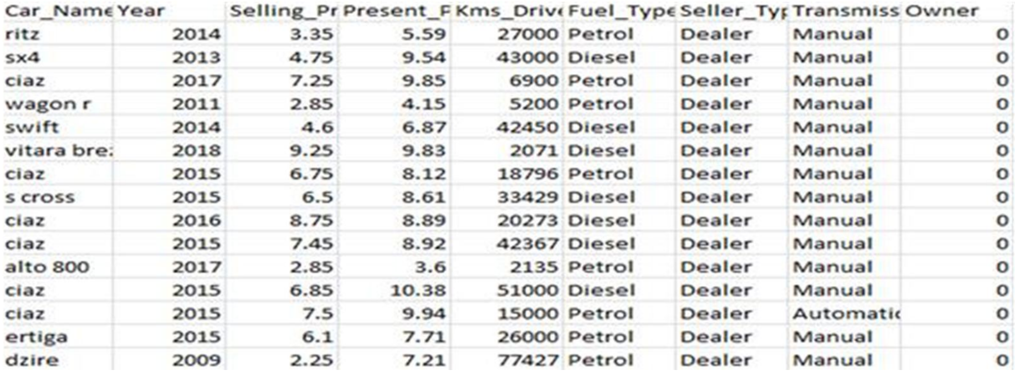
A. Requirement Analysis

The Data required is collected from a kaggle.com. The following attributes were captured for each car: Car Name, Year of car purchase, Present Price, Kilometers Drive, Fuel Type (0:Petrol, 1:Diesel, 2:CNG), Seller Type(0:Dealer, 1:Individual),Transmission Type(0:Manual, 1:Automatic), Owner previous cars (yes/no) etc.

We have used python Tkinter Library for developing GUI (Graphical user Interface) for our system and libraries like pandas, numpy, sklearn for data extraction and machine learning purposes.

Since manual data collection is time consuming task, especially when there are numerous records to process, we have taken the car dataset from kaggle.com. Data is Manually collected from kaggle.com.

In the training phase 75% of the total data is used and for testing purpose of machine learning model 25% of the total data is used.



Car_Name	Year	Selling_Pr	Present_F	Kms_Drive	Fuel_Type	Seller_Typ	Transmiss	Owner
ritz	2014	3.35	5.59	27000	Petrol	Dealer	Manual	0
sx4	2013	4.75	9.54	43000	Diesel	Dealer	Manual	0
ciaz	2017	7.25	9.85	6900	Petrol	Dealer	Manual	0
wagon r	2011	2.85	4.15	5200	Petrol	Dealer	Manual	0
swift	2014	4.6	6.87	42450	Diesel	Dealer	Manual	0
vitara bre:	2018	9.25	9.83	2071	Diesel	Dealer	Manual	0
ciaz	2015	6.75	8.12	18796	Petrol	Dealer	Manual	0
s cross	2015	6.5	8.61	33429	Diesel	Dealer	Manual	0
ciaz	2016	8.75	8.89	20273	Diesel	Dealer	Manual	0
ciaz	2015	7.45	8.92	42367	Diesel	Dealer	Manual	0
alto 800	2017	2.85	3.6	2135	Petrol	Dealer	Manual	0
ciaz	2015	6.85	10.38	51000	Diesel	Dealer	Manual	0
ciaz	2015	7.5	9.94	15000	Petrol	Dealer	Automati	0
ertiga	2015	6.1	7.71	26000	Petrol	Dealer	Manual	0
dzire	2009	2.25	7.21	77427	Petrol	Dealer	Manual	0

Fig.2. Processed data set sample in CSV format

The collected raw data set contains 302 samples. Since data is collected using kaggle.com, there are many samples that have only sufficient number of attributes. The data is present in the CSV file. The Pandas is the python machine learning library which is used to use the data in csv file to fit it in the machine learning model for training and testing purpose.

The Total Eight attributes related to used car were used for predicting selling price of the car in this project. The attributes like Car_name, Year of car Purchase, Present_price (Present price of car), Kms_Driven (Total distance driven by the car in kilometers). Fuel Type(Fuel Type of car), Seller type (whether the car is sell by dealer or by the individual owner),Transmission Type (whether the car is Manual or automatic), Owner previous number of cars etc. were used fitting them into the three machine learning models for training and testing purpose of the car. Then selling price of the car was predicted using the algorithm which achieves higher accuracy out of three algorithms used (i.e Linear regression, Lasso regression, Ridge regression) based on the data which is given as input by the user .

.

B. Model Implementation and Evaluation

Approach for car price prediction proposed in this paper is composed of several steps, shown in Fig. 3

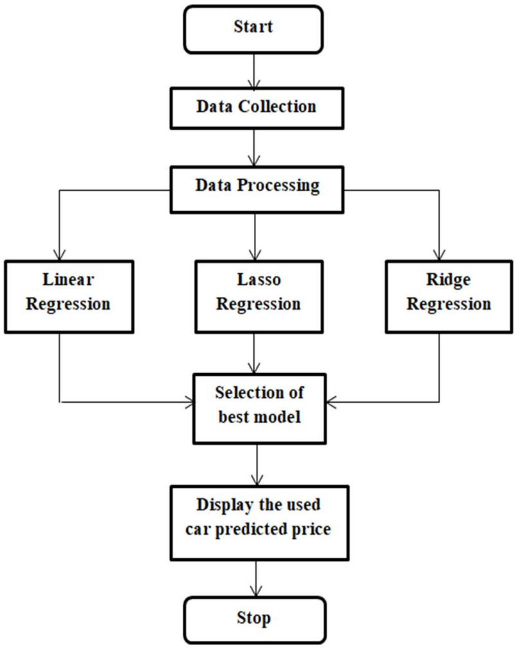


Fig.3. Proposed System Architecture

The Above Diagram Shows Flowchart or processing steps in our project. So at very first step data is collected for project. For our project we have taken the data related to car attributes in the csv form from kaggle.com which a data science learning website. We have considered 302 samples of car data in our research. Then in next step data processing is done. We have python pandas module to extract the data in csv from stored in an Excel sheet and used it further in our project for building the required machine learning models.

Then by using the data in excel sheet training and testing of models is done. We Have used regression machine learning algorithms for prediction purpose in our project. Three algorithms used in our project are Linear regression, lasso regression and ridge regression respectively. Using SVM (Support Vector Machine) classifier the data is divided into two parts i.e. 75% of data is used for training purpose and then 25% of data is used for testing purpose.

Then when user enters the car details on interface and inputs it to the system. Further system process on that input using the machine learning models and in the final step output is produced and displayed based on the algorithm which achieves higher accuracy.

In our project, lasso regression algorithm achieved the highest out of three algorithms. Hence we have displayed the final output based on lasso regression algorithm.

We have used three algorithms in our project which are Linear Regression, Lasso Regression, Ridge Regression respectively. The detail explanation of Algorithms which are used in this project is given below:-

1) *Linear Regression:* Linear regression is one in all simplest and preferred Machine Learning algorithms. It's a statistical method that's used for predictive analysis. Simple Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (x) variables, therefore called as linear regression. Since linear regression shows the linear relationship, which suggests it finds how the value of the dependent variable is changing with respect to the value of the independent variable.

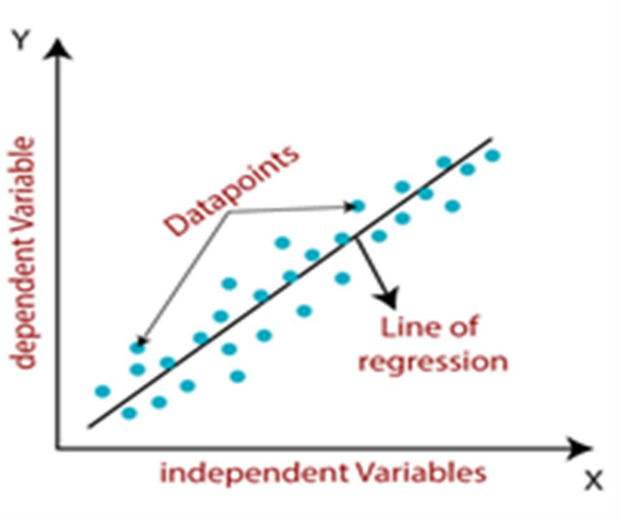


Fig.4. Graphical Representation of Linear Regression

The Formula for Simple Linear Regression is,

y= a0+a1x+ ε

Here,
Y= Dependent Variable (Target Variable)
X= Independent Variable (predictor Variable)
a0= intercept of the line (Gives an additional degree of freedom)
a1 = Linear regression coefficient (scale factor to each input value).
ε = random error
The values for x and y variables are training datasets for Linear Regression model representation.

2) *Lasso Regression:* Lasso regression is another regularization technique to minimize the complexity of the model. It stands for Least Absolute and Selection Operator. It's almost like the Ridge Regression except that the penalty term contains only the absolute weights rather than square of weights. Since it takes absolute values, hence, it can shrink the slope to 0. whereas Ridge Regression can only shrink it just about to 0.It is called as L1 Regularization.

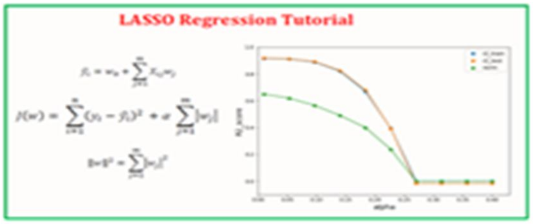


Fig.5. Graphical Representation of Lasso Regression

The equation for the cost function of Lasso regression will be:

$$\sum_{i=1}^{M}(y_i - y'_i)^2 = \sum_{i=1}^{M}\left(y_i - \sum_{j=0}^{n}\beta_j * x_{ij}\right)^2 + \lambda\sum_{j=0}^{n}|\beta_j|^{\square}$$

Fig.6. Mathematical Formula of Lasso Regression

Where,

X_{ij} = Features of Y or Independent Variable
Y_i = Dependent Variable
βi = Weights or Magnitude shows importance of a feature
λ = minimize the cross-validation prediction error rate.

The weight or Magnitude of features is selected based on importance of feature. selection of more features creates complexity and reduces prediction accuracy. Hence some of the features in this technique are completely neglected for model evaluation.

In our project present_price(present of the car) and Kms_Driven (Distance driven in kilometers) are two most important features hence there magnitude or importance factor is more than other attributes. Lasso regression achieves highest prediction accuracy except other regression in our project

Hence, the Lasso regression can help us to reduce the overfitting in the model as well as the feature selection.

3) *Ridge Regression:* Ridge regression is one of the types of linear regression in which a small amount of bias is introduced in order to get better long-term predictions. Ridge regression is a regularization technique, which is used to reduce the complexity of the model. It is also called as L2 regularization. In this technique, the cost function is altered by adding the penalty term to it. The amount of bias added to the model is called Ridge Regression penalty. We can calculate it by multiplying with the lambda to the squared weight of each individual label.

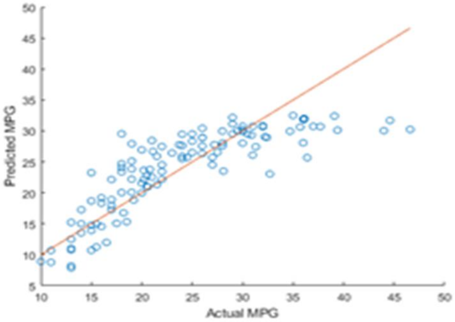


Fig.7. Graphical Representation of Ridge Regression

The equation for the cost function in ridge regression will be:

$$\sum_{i=1}^{M}(y_i - y'_i)^2 = \sum_{i=1}^{M}\left(y_i - \sum_{j=0}^{n}\beta_j * x_{ij}\right)^2 + \lambda\sum_{j=0}^{n}\beta_j{}^2$$

Fig.8. Mathematical Formula of Ridge Regression

X_{ij} = Features of Y or Independent Variable
Y_i = Dependent Variable
βi^2 = Weights or Magnitude shows importance of a feature λ λ = minimize the cross-validation prediction error rate

V. CONCLUSION

Car price prediction will be a challenging task because of the high number of attributes that should be considered for the accurate prediction. The most important step within the prediction process is collection and preprocessing of the information. During the research, Car data collected from kaggle.com is converted into CSV form and used for building the machine learning algorithms.

Three algorithms which are Linear, Lasso and Ridge Regression were utilized in this project. The data was divided into two parts training and testing purpose by SVM classifier (Support Vector Machine).i.e. 75% of data was used for training purpose and 25% of data was used for testing purpose of the machine learning. The accuracies of the three machine learning models were checked and compared with one another. The Final result was predicted consistent with the algorithm which achieves higher accuracy. The main drawback of this project was less number of records that have been utilized. As future work, we expect to collect more information and to utilize further advanced developed methods like Random Forest, ANN (Artificial Neural Network), CNN (Convolutional Neural Network) with a better user computer user interface experience.

REFERENCES

[1] Enis Gegic, Becir Isakovic, Dino Keco, Zerina Masetic, Jasmin Kevric. “Car Price Prediction using Machine Learning Techniques” TEM Journal Volume 8, Issue 1, Pages 113-118, ISSN 2217-8309, DOI: 10.18421/TEM81-16, February 2019

[2] Pattabiraman Venkatasubbu, Mukkesh Ganesh “Used Cars Price Prediction using Supervised Learning Techniques” IJEAT ISSN: 2249 – 8958, Volume-9 Issue-1S3, December 2019

[3] Sameerchand Pudaruth “Predicting the Price of Used Cars using Machine Learning Techniques” January 2014

[4] K.Samruddhi, Dr. R.Ashok Kumar “Used Car Price Prediction using K-Nearest Neighbor Based Model” IJIRASE Volume 4, Issue 3, DOI: 10.29027/IJIRASE.v4.i3.2020.686-689, September 2020 September 2020

[5] Nabarun Pal, Dhanasekar Sundararaman, Priya Arora, Puneet Kohli, Sai Sumanth Palakurthy “How much is my car worth?” A methodology for predicting used cars prices using Random Forest” FICC 2018