Used Vehicle Price Prediction System Using Multiple Linear Regression

Madhuri M. Swami[1], Komal B.Valivadekar[2], Pradnya B. Patil[3], Srushti S. Ingle[4], Mr. S.T.Powar[5] [Guide]
[1,2,3,4,5] D.Y.Patil Collage of Engg and Technology,Kasaba Bawada

Abstract- *The multiple linear regression technique is used for vehicle price prediction system. Using multiple linear regression, multiple independent variables but one and only one dependent variable that's actual and predicted values are compared to find precision of results. Paper proposes a system where price is dependent variable which is to be predicted, and price is derived from factors like vehicle's model, make, city, price, fuel type, version, color, mileage and power PS.*

Keywords- Multiple linear regression, used vehicle price, Preprocessing

I. INTRODUCTION

Sale of the second hand vehicles is increasing due to economic conditions. More vehicle buyers are finding alternatives of buying new vehicles outright. These lease installments are dependent upon the estimated price of the vehicle and thus sellers are interested to know about fair estimated price of their vehicles. It is found that finding fair estimated price of a used vehicles are important as well as challenging. So, there is a need of accurate price prediction mechanism for the used vehicles. Prediction techniques of machine learning can be helpful in this regard.

Prediction technique of machine learning is helpful in which earlier data was trained. Factors were consider in vehicle price prediction are make, model, year, engine cylinders, transmission types, number of doors, highway mpg, and city mpg. The mileage of the car was considered as the major factor in the price prediction. On the analysis of above factors price was predicted. Predicted price can be used by a buyer to know the accurate market value of vehicle. The factors must be eliminated whose accurate information is not available because these factors affect on accuracy. The factor price is also a major factor must be present in the dataset. Multiple linear regression approach was used since it creates new value based on existing value. Here we use the data of vehicles which contains price details so that result could be verified.

The data associated with the investigation was very large because there were thousands of used vehicles and each vehicle's data comprises of values of many features. A distributed system was used for both data gathering and analysis. The developed system will provide the user with an interface that allows navigation to assess the fairness of prices compared to the predicted ones.

II. RELATED WORK

Most developing countries adopt the lease culture instead of buying a new vehicle due to affordability. Hence the used vehicles sales are exponentially increasing trade. Vehicles sellers sometimes take advantage of manual scenario by listing unrealistic prices. Therefore, arises a need for a model that can assign a price for a vehicle by evaluating its features.

The work is extremely hard to make sure for getting best prices. Using advanced price indexing tools, to bring the right prices to client so that client know exactly what price client should buy a vehicles.

III. METHODOLOGY

The focus of this research is to be building such a model which has the capabilities of dealing with high complexity and gives accurate results irrespective of the magnitude of data set. The input data is gathered from kaggle in a month or two. In the beginning, 1000 records of used vehicles were recorded. The collected data included variable values for price, engine capacity, color, mileage in kilometer, power steering, alloy rims, transmission, type of engine, version, make and model year. Once the data collection was over, then preprocessed data used multiple linear regression technique for price prediction. In this research, statistical software Minitab was used in which input the data and analyze the results via multiple linear regression application. Initially, all attributes were considered, but later applied the variable selection techniques on preprocessed input data and found the most significant variables and skipped all other insignificant variables.

Table 1. Sample Data Collection

Sr. No	Seller	Vehicle Type	Year Of Reg.	Brand	Price
1	Private	Suv	1993	Volkswagen	208500
2	Private	Coupe	2011	Audi	181500
3	Private	Suv	2004	Jeep	223500
4	Private	Kleinwage	2001	Volkswagen	140000
5	Private	Kleinwage	2008	Skoda	250000
6	Private	Cabrio	1995	BMW	143000
7	Private	Limousine	2004	Peugeot	307000

Above dataset was collected for each vehicles and factors are as follows Vehicle Type, Seller, Year of Reg, model, fuel type, Brand and price. Dataset collected from various web resources. Collected dataset was not in well structure. On the sample dataset cleaning, sampling, formatting was done and removed the unwanted columns and all null entries and converts it into structured format.

Implementation of MLR

Every machine learning process is the application of a chosen algorithm to a problem. Multiple linear regressions are similar to simple linear regression but instead of using one variable to predict the outcome of another variable, MLR uses two or more variables to do so.

Multiple linear regression algorithm that attempts to model the relationship between X independent (explanatory) variables and a Y independent (response) variables which analyzed the contents by fitting a linear equation to observed data. It Vehicle price is considered as the dependent variable while other attributes as the independent variables. Let X be the input and Y be the output, the linear regression correlation can be expressed as:

$$Y=\beta 0+\beta 1X1+\beta 2X2\ldots\ldots\ldots\beta nXn$$

In the above equation, X1, X2…Xn represent attributes in final dataset as multiple inputs.

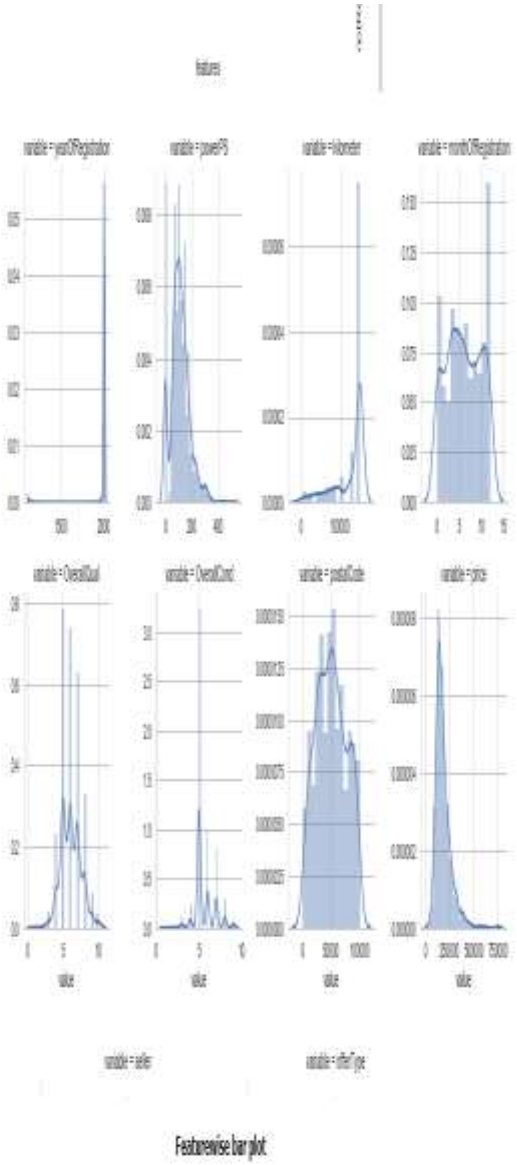

Fig1. Feature wise bar plot

The above bar graph shows individual plotting of variables. There were eight variables selected for price prediction and accuracy testing. In machine learning maximum accuracy is about 87%. If it is not found nearby 87% accuracy then add some more features to increase accuracy of price.

IV. CONCLUSION

The data set used in this paper can be very valuable in conducting similar research using different prediction techniques. The prices of vehicles can be predicted using dataset on same or different prediction software as well. The data obtained under the research facilitated in prediction of prices of used vehicles through multiple linear regression method. Many assumptions were made on the basis of the data set. The future price prediction of used vehicles with the help of same data set will comprise of using MLR algorithm.

REFERENCES

[1] IEEE Transactions on Control Systems Technology, Vol. 26, No. 1, Jan 2018

[2] 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

[3] Pudaruth,S. 2014. “Predicting the Price of Used Cars Using Machine Learning Techniques”, International Journal of information & Computation Technology, 4(7), p.753-764.

[4] Kuiper, S. 2008. “Introduction to Multiple Regression: How Much Is Your Car worth?”, Journal of Statistics Education, 16(3).

[5] International Journal of Computer Applications (0975 – 8887) Volume 167 – No.9, June 2017

[6] Limsombunchai, V. 2004. House price prediction: Hedonic price model vs. artificial neural network. In New Zealand Agricultural and Resource Economics Society Conference, New Zealand, pp. 25-26.