Predicting True Value of Used Car using Multiple Linear Regression Model

Laveena D'Costa, Ashoka Wilson D'Souza, Abhijith K, Deepthi Maria Varghese

Abstract— Predicting the true value of used cars requires lot of analysis. This prediction takes into account variables such as car model, fuel type, number of owner and so on. In this paper we are applying machine learning algorithms to determine the true value of cars when selling them to the dealers. We have used multiple linear regression model by dividing the data into training and test. Vehicle price forecast is both a critical and significant job, particularly when the car is used and does not come directly from the factory.

Keywords: Multiple Linear Regression, True value.

I. INTRODUCTION

It is an exciting and much needed issue to tackle predicting the price of used cars. Several people are not capable of buying a new car because of the lack of funds. Sometimes, the customers are abused by the sellers. Therefore, there is a need for predicting the price of used cars. Whenever a car is sold to dealers, the price should be calculated. Prediction techniques of machine learning algorithms are used to predict the estimate price of used cars.

II. RELATED WORKS

Kanwal Noor and SadaqatJan[1] proposed a model in which they had used deductive approach of multiple linear regression because new values are created on the basis of existing values. The main goal is to identify the model which gives the more accuracy in price prediction that had been used in the research. For the implementation of the model, Noor and Jan have used data from pakwheels (www.pakwheels.com), which is a recognized online car reselling company in Pakistan. The variables such as mileage, car's model, engine capacity, version city, color, alloy rims, power steering and price were considered. The statistical software, Minitab is used to input and analyze the data using linear regression. Variable selection method is used to find the most significant variables. Since, some of the variables are categorical in nature, it is converted and coded as category codes (0, 1). Least square method is used for model estimation.

In the model proposed by Nabarun Pal et al. [2], factors such as age of the car, its make, the origin of the car (original country of the manufacturer), its mileage (number of kilometers has run) and its horsepower were considered as the attributes. Fuel price is also given great importance. Other variables are the fuel used, style, braking system, cylinder volume, acceleration, the number of doors, safety index, size, weight, height, paint color, consumer reviews, prestigious awards won by the car manufacturer.

They proposed a machine learning technique called Random Forest. Car price was considered was the dependent variable. As the first step, they collected the data about the used cars from kaggle ("Used car database", scraped from eBayKleinanzeigen, the German subsidiary of eBay) and pre-processed to remove the irrelevant features.After pre-processing, the most important features (price, kilometer, brand, vehicle type) are derived. Box plots, graph, histograms are used to represent the age of the vehicle, average price of the vehicle, top ten average prices by brand etc. Even though Random Forest is used for classification, Nabarun Pal et al. had tried the model with Linear Regression and Random Forest Regression. In the Random Forest Algorithm, a Grid search algorithm was used in order to find the optimum number of trees and the best accuracy was found when 500 decision trees are used.

NitisMonburinon et al. [3] used supervised machine learning models for the data collected from www.kaggle.com which uploaded by OrgesLeka under the public domain license. In order to identify the model with high accuracy, they conducted a comparative study on multiple linear regression, random forest regression and gradient boosted regression trees. Comparison was done using mean absolute error as the criterion. From the analysis, they suggest that gradient boosting trees are best for predicting the price of used cars.

III.DATA AND EMPIRICALANALYSIS

A. Data Collection

The data used for analysis is collected from Mandovi Motors Pvt Ltd, Mangalore. Sample size was 1870 which comprised of maruti and non-maruti vehicles. In order to proceed with our model and get a good prediction, we considered only maruti vehicles. Variables like manufacturer, car model, sub model, fuel type, emission, mileage ownership and year of registration are the independent variables where buying price is the continuous output. After preprocessing irrelevant variables like manufacturer and sub model is removed. Attributes that are considered for the analysis is shown in Table I.

Revised Manuscript Received on January 30, 2020.

* Correspondence Author

Laveena D'Costa*, department of Big Data Analytics, AIMIT, St. Aloysius College, Mangalore, India. Email: lavishalet@gmail.com

Ashoka Wilson D'Souza,department of Statistics, Mangalore University, Mangalore, India. Email:ashokdesouza@gmail.com

Abhijith K, department of Big Data Analytics, AIMIT, St. Aloysius College, Mangalore, India. Email:abhijithvijayan66@gmail.com

Deepthi Maria Varghese, department of Big Data Analytics, AIMIT, St. Aloysius College, Mangalore, India. Email:deepthimaria61@gmail.com

DOI: 10.35940/ijrte.E1010.0285S20

TABLE I. SAMPLE DATA

CAR MODEL	OMNI	ALTO	M800	RITZ
FUEL TYPE	PET	PET	PET	PET
EMISSION	EURO4	EURO3	EURO2	EURO4
MILEAGE	91450	48660	128116	42773
OWNERSHIP	1	1	3	2
YEAR OF REG	2013	2009	2001	2015
PRICE	130000	155000	30000	400000

B. Procedure and Data Analysis

The collected data was then processed. Categorical variables such as car model, fuel type and emission were converted to factors. Fig.1 and Fig.2 implies that actual buying price is right-skewed distribution. This means that dependent variable is not distributed normally. Whenever the data is not normally distributed, log or square root transformation can be used. Fig.3 and Fig.4 depicts that normal distribution of the price.

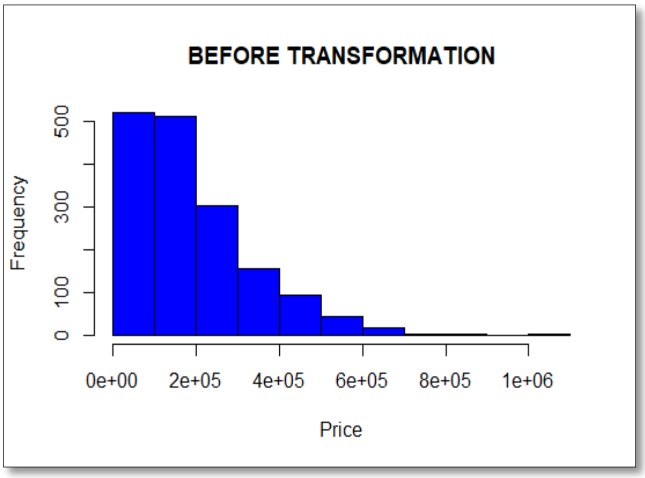


Fig.1:: Distribution of price before transformation

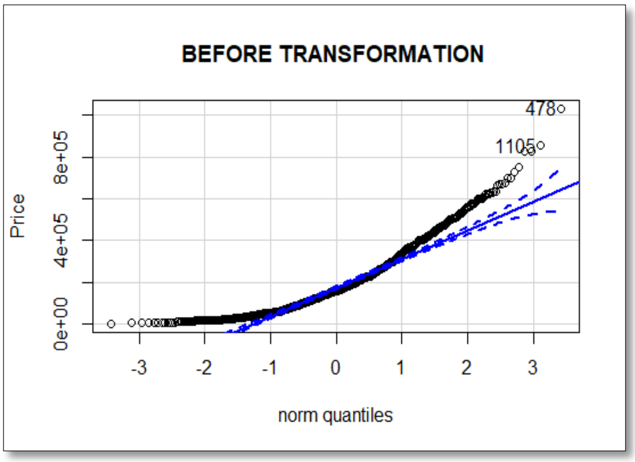


Fig. 2:QQ-Plot before transformation

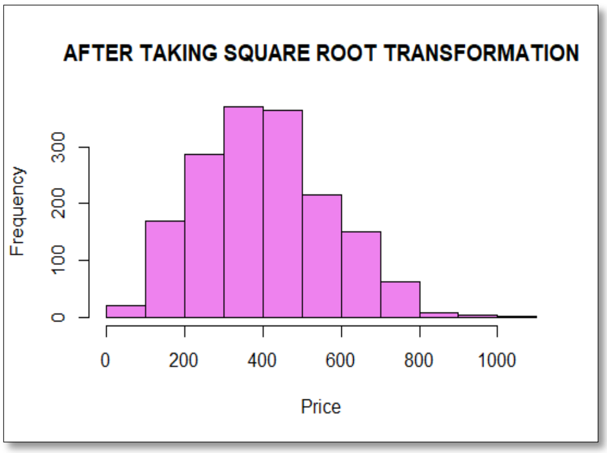


Fig.3: Histogram after taking square root transformation

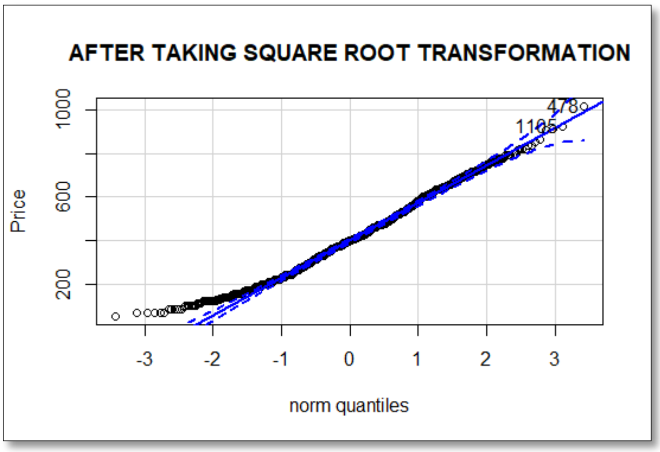


Fig.4: QQ-Plot after taking square root transformation

Fig.5 shows the distribution of mileage and the buying price. It implies that mileage is inversely proportional to price of cars. Correlation between mileage and price is calculated as -0.2823 and that of mileage and number of owners is -0.4588.The price of car that has more mileage will have less demand in the market. It also shows that one of the reason for high mileage depends on the year of manufacture or age of the car.

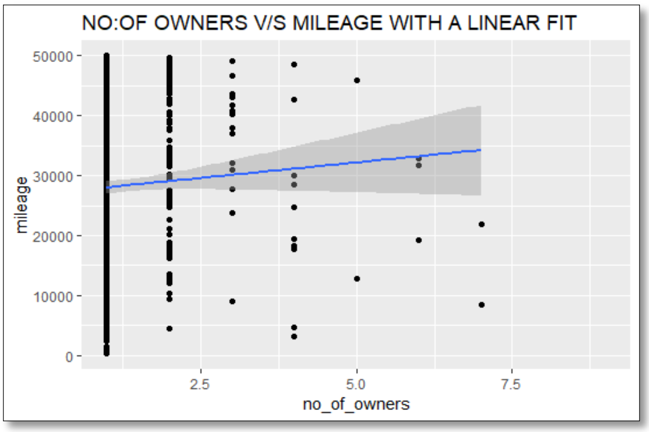


Fig.5: Scatter plot of distribution of ownership and mileage

Multiple Linear Regression

Multiple Linear regression is an improvised version of linear regression model where it helps in fitting relationship between more than two variables.

Retrieval Number: E10100285S20/2020©BEIESP
DOI: 10.35940/ijrte.E1010.0285S20

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Mathematical equation of multiple regression model is:

$$Y=\beta_0 + \beta_1X_1 + \beta_2X_2 + \ldots\ldots + \beta_nX_n$$

Here, Y is the dependent variable and X_1, X_2, ….X_n are the independent variable. β_0, β_1,β_2, …., β_n are the regression coefficients of the independent variables X_1,X_2,…,$X_{n..}$

Use of independent variables helps in determining the values of regression coefficients by fitting the model. This model is used to identify the dependent variable.

IV.RESULT

A. R – Square and Adjusted R –Square

The R-Square and the adjusted R-Square helps in fitting the best model. The model with high R-Square values implies the model that is well suited to the data. Table 1 shows the R-Square and adjusted R-Square value of the model.

Table 1: R-Square and Adjusted R-Square

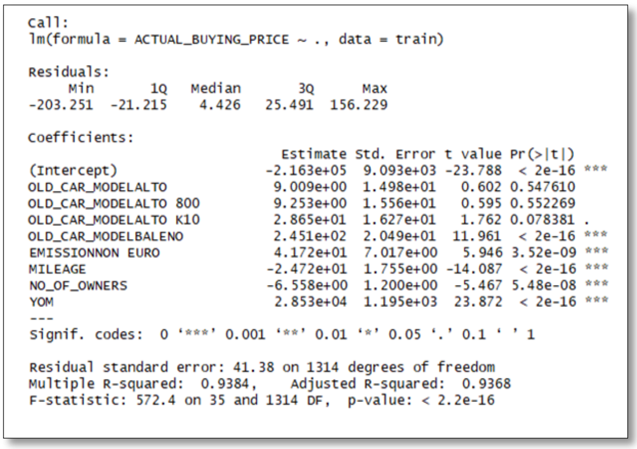
R - Square	98.34%
Adjusted R - Square	93.68%

From the table, it is clear that about 98.34% of the variation in price can be explained using the predictors. Remaining 2% depends on the attributes that are not considered in the model creation. Adjusted R-Square value helps in determining the accuracy of the model in advanced way. It means that R-Square value and the adjusted R-Square are the main factors to check the accuracy of the model.

B. Summary of the model

Summary of the model clearly explains about the remaining concepts such as Residual Standard Error and F-Statistic. The p-value was found to be less than 2.2e-16. And the F-Statistic was 572.4 . Summary of the fitted model is given in Table 2.

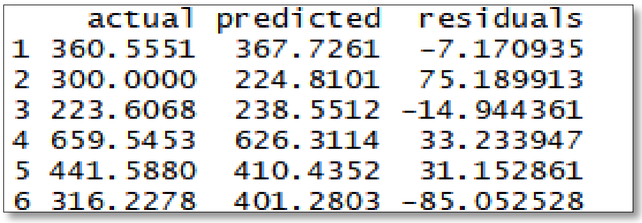
Table 2: Summary of model



```
Call:
lm(formula = ACTUAL_BUYING_PRICE ~ ., data = train)

Residuals:
     Min       1Q   Median       3Q      Max 
-203.251  -21.215    4.426   25.491  156.229 

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)    
(Intercept)                -2.163e+05  9.093e+03 -23.788  < 2e-16 ***
OLD_CAR_MODELALTO           9.009e+00  1.498e+01   0.602 0.547610    
OLD_CAR_MODELALTO 800       9.253e+00  1.556e+01   0.595 0.552269    
OLD_CAR_MODELALTO K10       2.865e+01  1.627e+01   1.762 0.078381 .  
OLD_CAR_MODELBALENO         2.451e+02  2.049e+01  11.961  < 2e-16 ***
EMISSIONNON EURO            4.172e+01  7.017e+00   5.946 3.52e-09 ***
MILEAGE                    -2.472e+01  1.755e+00 -14.087  < 2e-16 ***
NO_OF_OWNERS               -6.558e+00  1.200e+00  -5.467 5.48e-08 ***
YOM                         2.853e+04  1.195e+03  23.872  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 41.38 on 1314 degrees of freedom
Multiple R-squared:  0.9384,	Adjusted R-squared:  0.9368 
F-statistic: 572.4 on 35 and 1314 DF,  p-value: < 2.2e-16
```

The prediction may not be always accurate. There will be misclassifications and these are shown using residuals. Table 3 consists of actual and predicted columns obtained from the dependent variable and the residuals or the variation between actual and predicted price.

Table 3: Actual price, Predicted price and the Residuals



```
    actual predicted  residuals
1 360.5551  367.7261  -7.170935
2 300.0000  224.8101  75.189913
3 223.6068  238.5512 -14.944361
4 659.5453  626.3114  33.233947
5 441.5880  410.4352  31.152861
6 316.2278  401.2803 -85.052528
```

C. Assumptions of MLR model

There should be a linear relationship between the variables. In the model it is found that there exists a linear relationship among variables. It is shown in Fig 6.

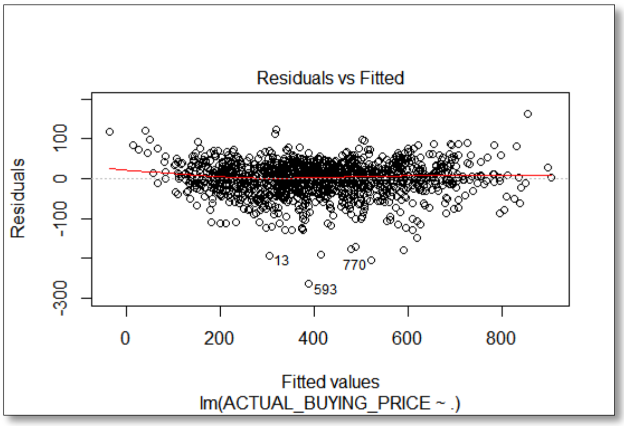


Fig 6: Residuals vs Fitted

The correlation matrix of the numeric variables helps identifying the significant variables in the data. It is an advanced form where we can diagnose the variables using the summarized information. Correlation matrix of numeric variables in the dataset is demonstrated in Fig.6.The asterisk (*) symbol in the figure denotes the significance of variables along with histograms.

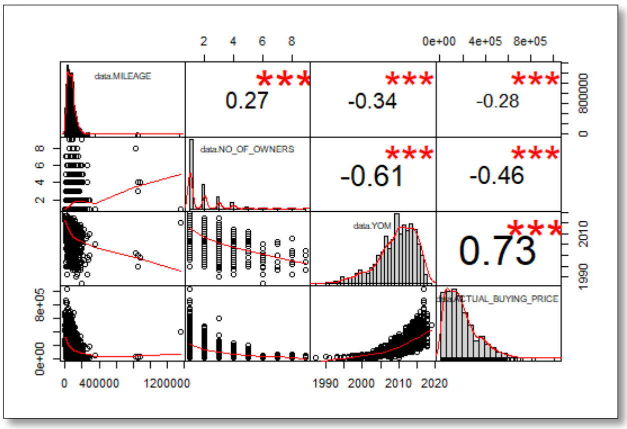
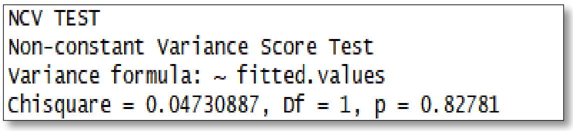


Fig 7: Correlation Chart of numeric variables

One of the assumption of multiple linear regression is that there should not be any heteroscedasticity among residuals. It can be diagnosed using Breush Pagan Test and NCV Test. In this study, we have used NCV Test. The output is shown in Table 4. Since the p-value is greater than 0.05, we conclude that variance of error terms are equal.

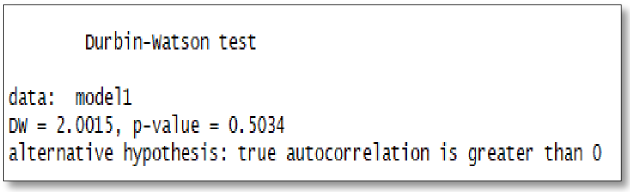
Table 4: NCV Test



```
NCV TEST
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 0.04730887, Df = 1, p = 0.82781
```

Autocorrelation refers to the correlation of one variable with itself. It can be detected using Durbin-Watson Test. Even though there is a minute variation in the value from 2 (no autocorrelation, if dw =2), we conclude that there is no autocorrelation in the model. Table 5 indicates the Durbin-Watson Test Value

Table 5: Durbin-Watson Test



```
        Durbin-Watson test

data:  model1
DW = 2.0015, p-value = 0.5034
alternative hypothesis: true autocorrelation is greater than 0
```

V.CONCLUSION

The model that is created using this data can be used to predict the true value of used cars. The accuracy of the model was found to be 89.33%. The model is created using multiple linear regression and it provides good accuracy. We The same data can be used to implement more algorithms and it may give accurate result.

ACKNOWLEDGMENT

We the authors would like to acknowledge our thanks to Mandovi Motors Pvt. Limited, Hampankatta Mangalore for providing us with data which has helped us to come up with this paper.

REFERENCES

1. Noor, Kanwal, and Sadaqat Jan. "Vehicle price prediction system using machine learning techniques." International Journal of Computer Applications 167.9 (2017): 27-31.
2. Pal, Nabarun, et al. "How Much Is My Car Worth? AMethodology for Predicting Used Cars' Prices UsingRandom Forest." Future of Information andCommunication Conference. Springer, Cham, 2018.
3. Monburinon, Nitis, et al. "Prediction of prices for used carby using regression models." 2018 5th InternationalConference on Business and Industrial Research (ICBIR), IEEE, 2018

AUTHORS PROFILE

Laveena D'Costa, has cleared her NET in Management. She holds Master's degree in Statistics and Business Administration from Mangalore University and Madras University. She also has a Post graduate diploma in HRM from KSOU. She has published 14 papers in international Journals. Her area of interest includes Machinelearning, Time series analysis and Data Analytics.

Ashoka Wilson Dsouza, is currentlypursuing his PhD (Statistics) from Mangalore University. His areas of interest are machine learning, limited dependent variables in Econometrics and time series analysis. He is also a guest faculty in the department of Big Data Analytics at AIMIT, St Aloysius College, Mangalore.

Abhijith K, student at AIMIT, St Aloysius College, Mangalore. He is pursuing his Master's in Big Data Analytics. He is interested in the area of machine learning.

Deepthi Maria Varghese, is a Student at AIMIT, St Aloysius College, Mangalore. She is pursuing her Master's in Big Data Analytics. She is interested in the area of machine learning.