iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** VII **Month of publication:** July 2022

DOI: https://doi.org/10.22214/ijraset.2022.45469

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VII July 2022- Available at www.ijraset.com

Predicting the Price of Pre-Owned Cars Using Machine Learning and Data Science

G. Kalpana[1], Dr. A. Kanaka Durga[2], T. Anoop Reddy[3], Dr. G. Karuna[4]
[1]*Asst. Professor, Department of Computer Science And Engineering, Vidya Jyothi Institute of Technology, Hyderabad, India*
[2]*Professor, Department of Computer Science And Engineering, Stanley College of Engineering & Technology for Women Hyderanbad, India*
[3]*Student, Koneru Lakshmiah University, Hyderanbad, India*
[4]*Professor, Department of Computer Science And Engineering, Gokaraju Rangaraju Institute Of Engineering & Technology, Hyderanbad, India*

Abstract: ***Storm Motors Is An E-Commerce Company Who Act As Mediators Between Parties Interested In Selling And Buying Pre-Owned Cars. They Have Recorded Data About The Seller And Car Details, Registration Details, Web Advertisement Details, Make And Model Information And Price. The Company Wishes To Develop An Algorithm To Predict The Price Of Pre-Owned Cars Based On Various Attributes Associated With The Car To Make A Sale Quickly, If The Price Is Reasonable And Satisfies Both The Seller And Buyer, By Comparing The Price Of Various Car Models Based On Car Features To Improve Their Business. In This Paper, We Have Conducted A Comparative Study Using Machine Learning Algorithms Like Linear Regression And Random Forest Algorithms Which Is Implemented With Jupyter Note Book. The Study Shows That Linear Regression Algorithm Performance Is More Than Random Forest Algorithm. We Have Also Experimented With Auto Ai Experimentation In Ibm Cloud Watson Studio, Which Automatically Builds The Best Predictive Model By Comparing With Other Algorithm, With Accurate Measures. In This Auto Ai Experiment We Have Found That Linear Regression Is Performing Better Than Ridge Algorithm And Random Forest. The Main Objective Of This Paper Is To Find The Best Predictive Model For Predicting Pre-Owned Car Price.***

Keywords: Regression, dataset, machine learning, prediction

I. INTRODUCTION

From the private car under assessment conditions.(2) The data we acquired were from the used car trade market The automotive industry is a cornerstone of the national economy, and forecasting automobile sales correctly is of great importance. [1] The pre-owned automobile market is an ever-increasing industry, almost doubling its market value in recent years. The advent of online portals such as CarDheko, Quikr, Carwale, Cars24 and many others has made the customer's and many others' needs simpler[2]. It's common in many developed countries to rent a car instead of buying it outright.

A lease is a binding contract between a buyer and a seller (or a third party – usually a bank, insurance company or other financial institutions) where the buyer has to pay fixed instalments to the seller / financing company for a predefined number of months / years. The buyer has the option to purchase the car at its residual value , i.e. its expected resale value, after the lease period is over. Sellers / financers are therefore of commercial interest in being able to predict the salvage value (residual value) of cars with precision[3].

Every year the car industry has become more and more dynamic and has expanded globally. Therefore, in this competitive car market, an exact price must be set for both customers and manufacturers. Customers and manufacturers are confused about the purchase or sale price for the car. Consequently, on the Internet , customers and manufacturers are trying to seek advice from auto-dealers, car magazines or the website. This information, however, takes a long time and could confuse the customers on the market.[4]

Some modeling hypotheses have been set as follows : (1) the pre-owned car we mentioned here was private car only, not including the car used as a commercial car like taxi or as a chauffeur-driven car in government or company, which is different in Shanghai which may be a little different from the other place.[5]

Predicting vehicle prices is considered a challenging issue, as there are many different factors that affect the price of vehicles. Besides the characteristics of vehicles such as brands, manufacturers, models, engines, fuel, etc., there are also many external factors that affect the price of automobiles such as taxes or distance traveled[4].

From previous studies it can be seen that different factors were chosen by the authors as input variables for forecasting car prices. These characteristics are diverse, and they consist of many qualitative variables. Quantifying the qualitative data is therefore a crucial step in pre-processing data before it is placed into the model for predicting vehicle prices. It is one of the big contributions of the paper, too.[4]

II. MACHINE LEARNING ALGORITHMS

Machine learning algorithms, based on a certain set of features, can be used to predict a car's retail value. Different websites have different algorithms to generate the retail price of the used cars, and therefore there is no unified price algorithm. By training statistical models to predict prices, a rough price estimate can be easily obtained without actually entering the details on the desired website.

The main objective of this paper is to use different prediction models to predict the retail price of a used car and compare their levels of accuracy.[2]

A. Random Forest

Random forest is mainly used for classification, but we used it as a model of regression by turning the issue into an equivalent issue of regression. Their trees (weak-learner) are trained on small parts of the dataset individually and help to learn highly unpredictable patterns by growing very deeply. This solves overfitting problem by combining the predictions of individual trees with a view to raising the variance and maintaining consistency[6][4].

B. Linear Regression

Regression is a supervised-learning approach. Continuous variables can be modelled and predicted. In Regression we have the labeled datasets and the value of the output variable is determined by the values of the input variable-so this is the supervised learning approach. The simplest form of regression is linear regression, where attempts are made to fit a straight line (straight hyperplane) into the dataset, and when the relationship between the data set variables is linear. Linear regression has the advantage of being easy to understand and regularization also makes it easy to avoid over fitting. We can also use SGD to update the linear templates with new data. Linear Regression is a good fit if the covariate-response variable relationship is known to be linear. It shifts from statistical modeling to analyzing and preprocessing the data. Linear Regression is useful for thinking about the method of data analysis. However, for most practical applications this is not a suitable approach because it oversimplifies real world problems[7].

C. Ridge Logistic Regression

(Hoerl and Kennard, 1970; Cessie and Houwelingen, 1992; Schaefer et al., 1984), A maximization of the likelihood function is obtained by applying a penalized parameter to all coefficients except intercept. The ordinary logistic regression with binary response is given by the probability of success of the answer.[8]

III. DATA SET DESCRIPTION

The car data set used in this research were collected from website. This dataset consists of 50,002 car observations and the 19 attributes of pre-owned car are from an e-commerce site as shown in Table I and II. These datasets may contain a significant number of pre-owned cars information with several presumably requiring some tweaking and engineering. For example, duplicated observations can affect model performance and must be removed in advance[10]. For this action the study used python programming language. [9] [11]

A descriptive statistics of categorical variables is shown in table I. Technically, attributes such as dateCrawled, lastSeen, postal-code, and dateCreated have no effect whatsoever on price prediction, so they can be removed to improve model performance.[12] Since their values are highly unbalanced, attributes such as seller, offerType, abtest, and nrOfPicture were also removed with the data preparation process by inspecting more detail on dataset. Finally, the name was removed as well, because it contains too many unique values.[9]

B. Comparative analysis on price prediction This work incorporates many machine learning algorithms available in the machine learning library Scikit-learn[13]. Each model is trained using the same training data and tested with the same test data. The result was then compared in the next section, and described. The regression-based method has been proven reliable in predicting a continuous variable in supervised machine learning[14]

Table1:Descriptive Statistic of Categorical Variables

Attributes	Count	Unique	Top	Freq.
dataCrawled	371,528	280500	2016-03-24	7
name	371,528	233531	Ford_Flesta	657
seller	371,528	2	pivat	371525
offerType	371,528	2	Angebot	371516
abtest	371,528	2	test	192585
vehicleType	333,659	8	limousine	95894
gearbox	351,319	2	manuell	274214
model	351,044	251	golf	30070
fuelType	338,142	7	benzin	223857
brand	371,528	40	volkswagen	79640
notReparedDamage	299,468	2	nein	263182
dateCreated	371,528	114	2016-04-03	14450
lastSeen	371,528	182806	2016-04-07	17

Attributes	Mean	Std.	Min	Max
price	17,295.14187	3.59E+06	0	2.15E+09
Year Of Registration	2004.577997	9.29E+01	1000	1.00E+04
powerPS	115.549477	1.92E+02	0	2.00E+04
kilometer	125618.6882	4.01E+04	5000	1.50E+05
MonthOf Registration	5.734445	3.71E+00	0	1.20E+01
Nr Of Pictures	0	0.00E+00	0	0.00E+00
postalCode	50820.66764	2.58E+04	1067	1.00E+05

Table2.Descriptive Statistic Numerical Variables

1) In predictive statistics and machine learning, attributes with a high coefficient of correlation frequently, but not always, have greater effect on

2) variable prediction [15] As its name suggests, the correlation coefficient is a statistical measure describing the relation between variables. The correlation coefficient of two attributes is always between 1 (Positive Relation) and í1 (Negative Relationship), while 0 does not imply any correlation

$$r = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2\sum(y_i-\bar{y})^2}} \qquad \text{.....(1)}$$

A. The Qualitative and Quantitative Impact

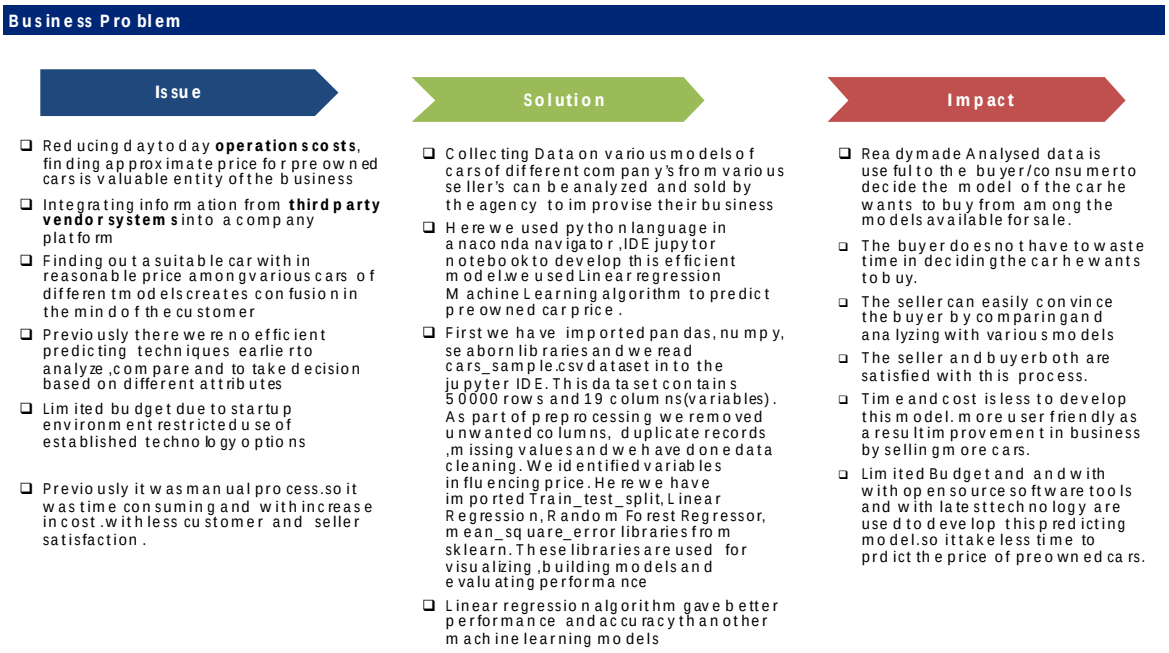


Fig1: The qualitative and quantitative impact

B. Architecture

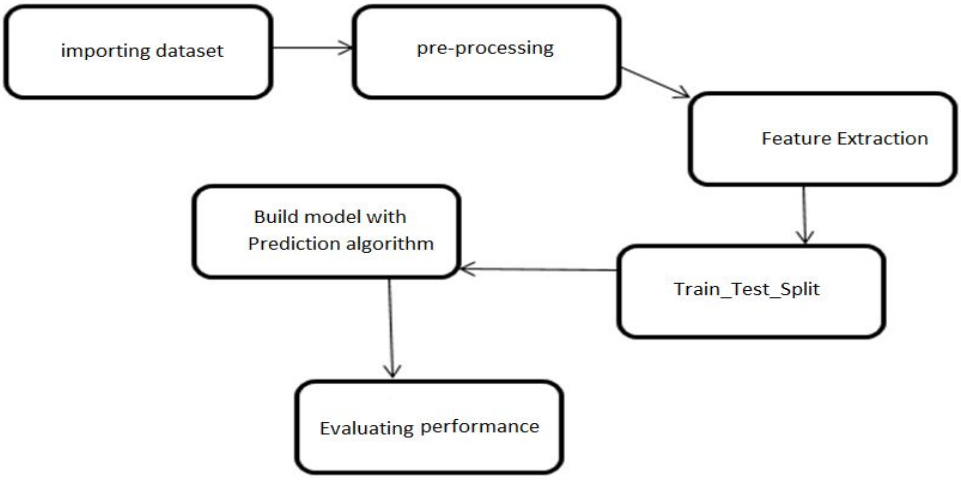


Fig2.Architecture of the system

C. Factors Influencing

Here are some key factors that influencing Design

1) *Data Set:* To analyze large data set with dimension 50000X19 design is effective
2) *Resources:* Availability of plenty of open source software's in the market
3) *Technology:* Good, popular, latest and easy language & technology support
4) *Libraries:* Plenty of predefined libraries are available to import, preprocess, train, test, analyze and to construct the model
5) *Visualization:* Tools are available to plot various types of graphs such as line graph, bar graph, Box plot plot, scatter plot etc.. for better understanding the insights of each graph.
6) *Easy to Learn:* It is easy to use, lean, develop analyze the complex data
7) *Performance Evaluation:* Availability of different performance metrics to compute and compare each model's performance and accuracy

IV. METHODOLOGY

To give the solution for the above problem statement here we have used Machine Learning Algorithms like Linear Regression and Random Forest. Here we have used python language to develop source code. we have csv dataset in to the jupyter IDE. This data set contains 50000 rows and 19 columns(variables). As part of preprocessing we removed unwanted columns, duplicate records ,missing values and we have done data cleaning. The variables can be grouped into different buckets based on the imported pandas, numpy, seaborn libraries and we read cars_sample. information. We identified variables influencing price and looked for relationship among variables for that here we have used correlation, boxplot, scatterplot. We also identified outliers for that we have used box plots, histograms etc. We filtered data based on logical checks for that we have used variables price, year of registration, power and then reduced number of data. Here we have imported Train_test_split, Linear Regression, Random Forest Regressor, mean_square_error libraries from sklearn. These libraries are used for visualizing ,building models and evaluating performance. With these models one can predict the selling price of various car models and compare the price of car based on various features of the car to satisfy the buyer. The output models, metrics we can see in the console of jupyter.

Regression is a statistical approach to finding the relationship between variables. In machine learning regression models are used to predict a continuous value. Here we used linear regression based on supervised learning.We have also compared with random forest regressor algorithm.

Pricing issue Predicting car prices is a problem of regression analysis in which the price of the car is a dependent variable and the characteristics of the vehicle (brand, car model, year of registration, type of gearbox, type of fuel, ...) are independent variables.

The input is denoted by X = {X1,X2, ... ,XN}and denote the output by Y. The regression model represents the relation of dependence between Y and X[4].

$$Y=f(X;\theta) \quad \text{..... (2)}$$

A. Solution Conceptualization

Import necessary libraries
Identify if data is clean
Look for missing values
Identify variables influencing price and look for possible relationships between variables
Data Visualization using box plot,scatter plot,bar graphsetc
Split the dataset into train and test dataset
Build a model with reduced number of variables to predict pre-owned car price.
Evaluate the performance of regressor model

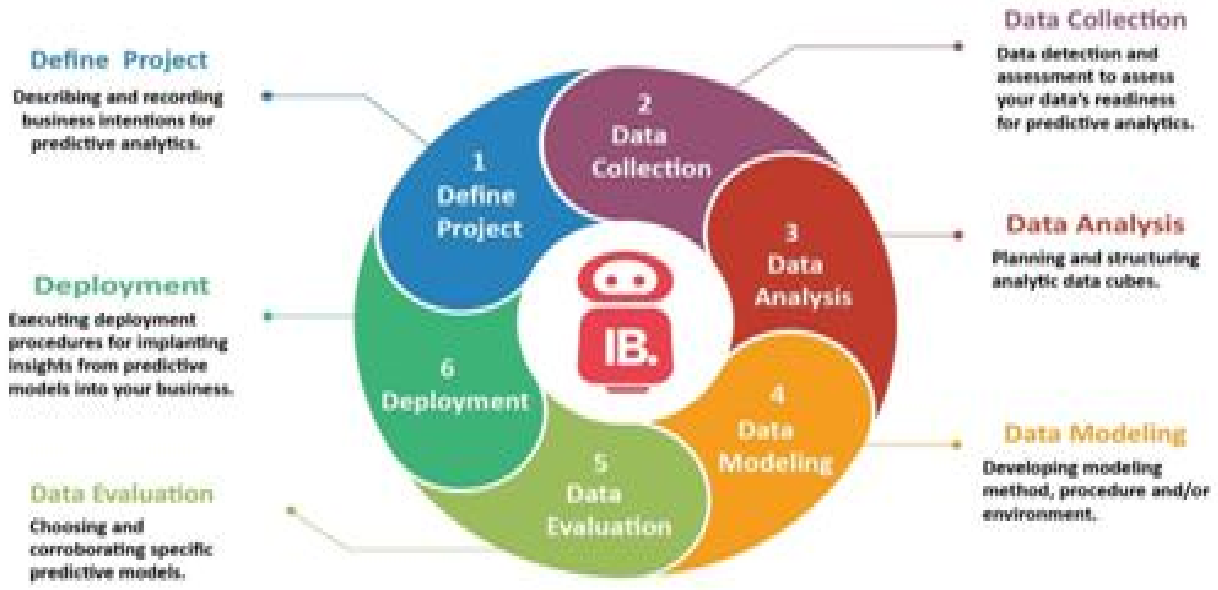


Fig3. steps to predictive Analysis

V. RESULT ANALYSIS

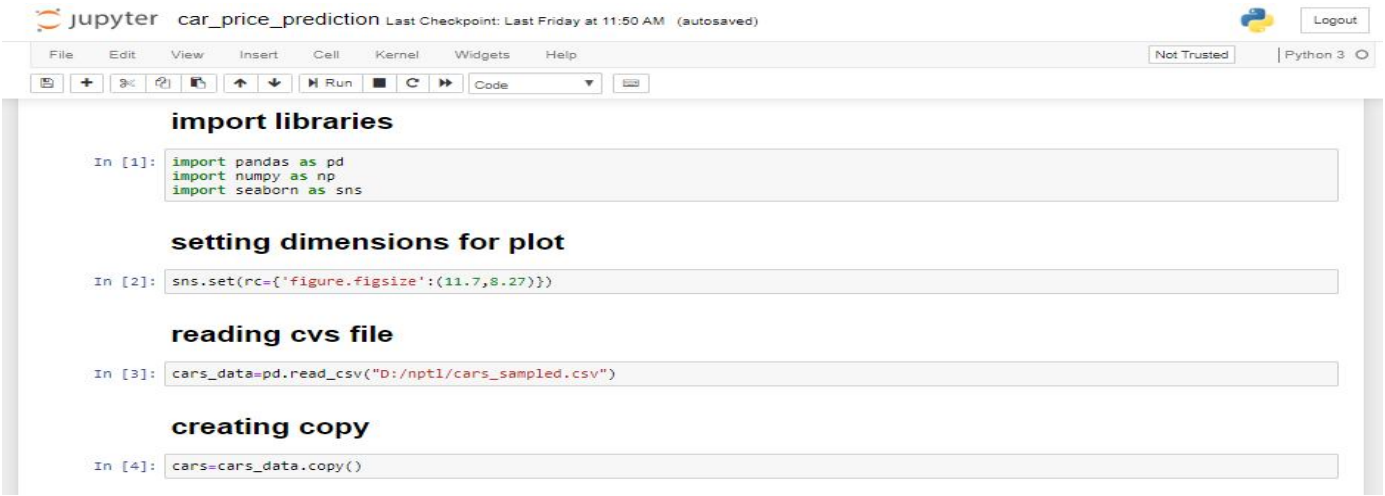


Fig4: Implementation with jupyter notebook

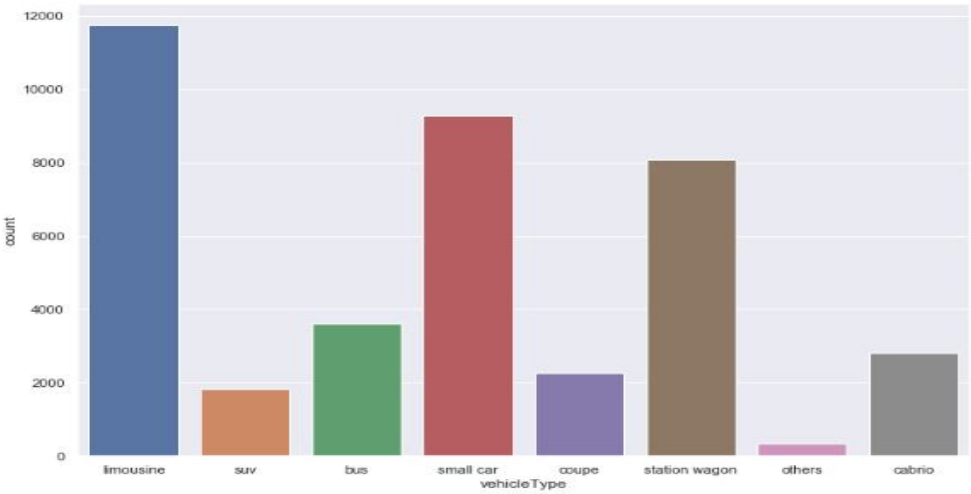


Fig5: Bar plot for Vehicle type VS Count

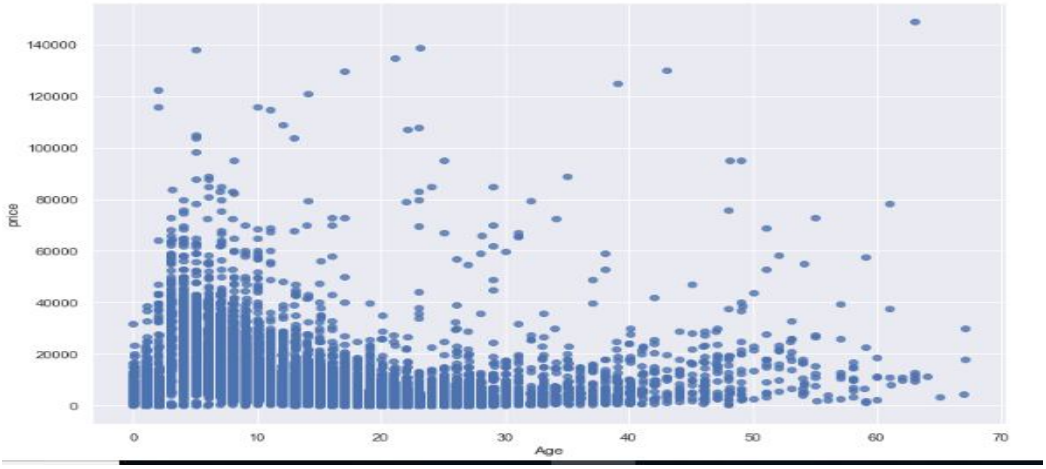


Fig6: Scatter Plot For Age Vs Price.

From the above scatter plot we can understand that cars priced higher are newer with increase in age price decreases

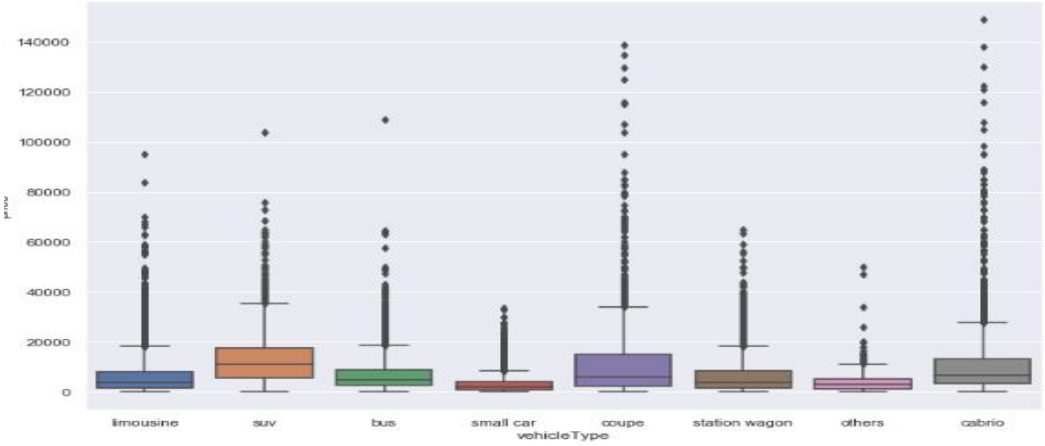


Fig7: Box plot for vehicle type vs price

From the above figure we can say that various vehicle types affect the price.

Categorical such as gearbox, notRepairedDamage, model, brand, fuelType, and vehicleType are not suitable for regression based on machine learning algorithm. Thus, label encoding algorithm was implemented to help normalize these attributes. Label encoding is just a simple approach for handling categorical variables which convert each value in an attribute.

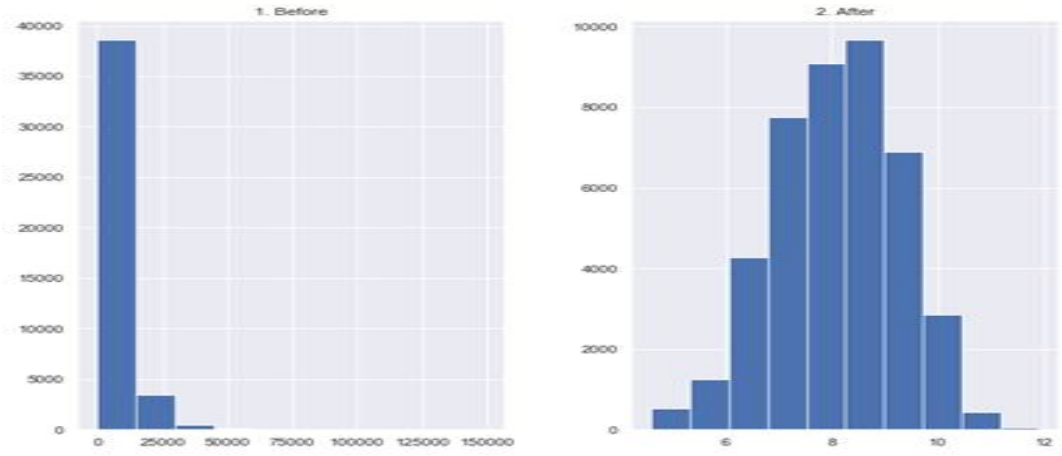


Fig 8. a) A right-skewed distribution of price before log transformation (price Vs kilometers)
b)A bell curve distribution of price after log transformation

```
metrics for models built from data where missing values were omited
 R square value for train from linear regression= 0.7808936978183916
 R square value for test from linear regression= 0.7658615091649237
 R square value for train from Random Forest= 0.9282494705146291
 R square value for test from Random Forest= 0.8504818147750623
base RMSE of model built from data where missing values were omitted 1.1274483657478247
RMSE value for test from linear regression =0.5455481266513847
RMSE value for test from Random Forest =0.4368736289370223
```

Fig9: output metrics Linear Regression Vs Random Forest Regressor

From the above figure we can say that RMSE value for Linear regression is 0.54 which is greater than the RMSE value of Random forest ie. 0.43 .And we have also seen with imputed data the RMSE value of Linear Regression is 0.64 .we received above Result while experimenting with Jupyter notebook.

Fig10: Building model with Auto AI Experiment in IBM cloud Watson Studio

From the above figure 10 shows the process of constructing the best performing model using Auto AI experimentation in series of steps. It first Read Dataset, Split Dataset Read training data, Pre-processing, finally Model selection is done based on hyper parameter optimization and feature engineering.

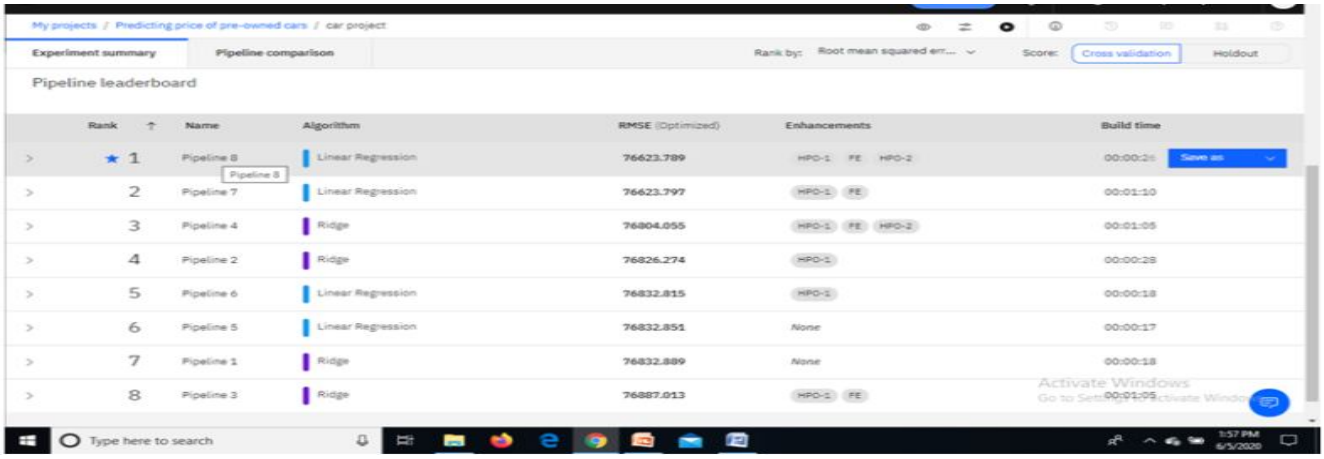


Fig 11: Linear regression Algorithm gives rank1 performance it is showing in pipeline 8.

From the above figure 11 it shows the Linear regression cross validation score is 76623.709 which gives better performance than the Ridge algorithm ie the RMSE cross validation score for ridge algorithm is 76304.055. we received above ranks while experimenting with Auto AI experimentation with Watson studio in IBM cloud

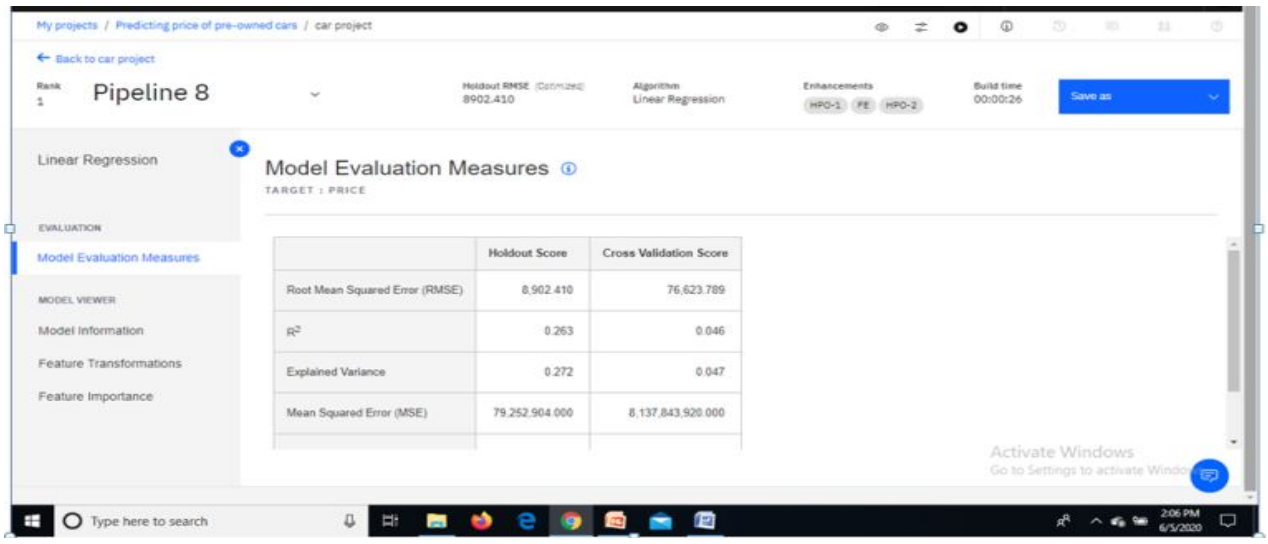


Fig12: Linear Regression Model Evaluation Measures

From the above figure12 shows the model accurate values of Linear regression for different measures while experimenting with Auto AI Experimentation with Watson studio in IBM cloud

A. Performance Measures

R-squared (R2) is a statistical measure representing the ratio of the variance of dependent variable that is explained in a regression model by independent variables. The R-squared formula is:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - Y_i)^2}$$

----------(3) where ˆ Yi is the value prediction.[4]

1) Mean Square Error (MSE) :

MSE is like the MAE, but the only difference is that it squares the difference between actual and predicted output values rather than using the absolute value. In the following equation the difference can be noted

$$MSE = \frac{1}{n}\sum(Y - \hat{Y})$$

----------(4)

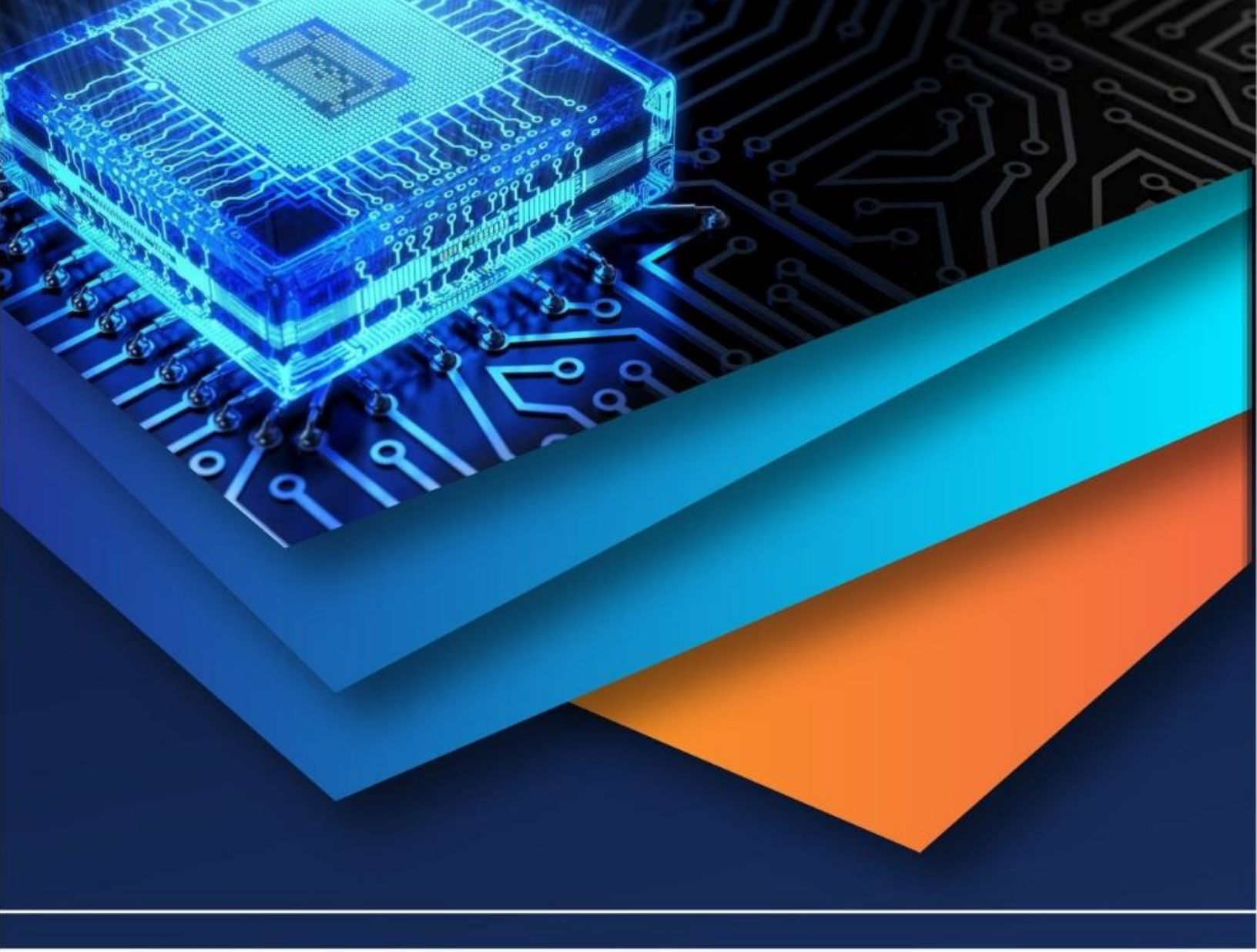
Where Y is the Actual out put value, And Y^ = Predicted Output Values

VI. CONCLUSION

This project is more helpful for all e-commerce companies who act as mediators for selling and buying pre-owned cars. The customer can easily be convinced in taking a decision to buy a pre-owned car out of various car models with various features. The seller can easily convince the buyer by comparing and analyzing various models.The seller and buyer both are satisfied with this process. This model reduces time and cost and is also more user friendly as a result of which there is improvement in business by selling more cars. Here we are also conducting a comparative study on performance of regression based on supervised machine learning models. Each model is trained using data of used car market collected from e-commerce website. As a result, Linear regression gives the best performance with Root mean square error (RMSE) =8902.410 . Followed by ridge, random forest regression algorithms respectively. We can also extend this project by considering more attributes like Resale history, Lic ,Accidents history,image etc to the data set for getting clear and ccurate analysis.

REFERENCES

[1] Yuan Qingyu, Liu Ying, Peng Geng, Lv Benfu" A Prediction Study On The Car Sales Based On Web Search Data ", 978-1-4244-8694-6/11/$26.00 ©2011 Ieee
[2] Pattabiraman Venkatasubbu, Mukkesh Ganesh "Used Cars Price Prediction using Supervised Learning Techniques" International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-9 Issue-1S3, December 2019
[3] Sameerchand Pudaruth1" Predicting The Price Of Used Cars Using Machine Learning Techniques" International Journal Of Information & Computation Technology. Issn 0974-2239 Volume 4, Number 7 (2014), Pp. 753-764
[4] Doan Van Thai, Luong Ngoc Son, Pham Vu Tien, Nguyen Nhat Anh, Nguyen Thi Ngoc Anh On" Prediction Car Prices Using Quantify Qualitative Data And Knowledge-Based System", 978-1-7281-3003-3/19/$31.00 C 2019 Ieee
[5] Shen Gongqi, Wang Yansong, Zhu Qiang" A New Model For Residual Value Prediction Of The Used Car Based On Bp Neural Network And Nonlinear Curve Fit" 011 Third International Conference On Measuring Technology And Mechatronics Automation, 978-0-7695-4296-6/11 $26.00 © 2011 Ieee
[6] N. Pal, P. Arora, D. Sundararaman, P. Kohli, And S. Sumanth Palakurthy, "How Much Is My Car Worth? A Methodology For Predicting Used Cars Prices Using Random Forest," Arxiv E-Prints, P. Arxiv:1711.06970, Nov 2017.
[7] Susmita Ray Department Of Computer Science & Technology "A Quick Review Of Machine Learning Algorithms "2019 International Conference On Machine Learning, Big Data, Cloud And Parallel Computing (Com-It-Con), India, 14th -16th Feb 2019
[8] Jose Manuel Pereira[a]*, Mario Bastoa , Amelia Ferreira Da Silva" The Logistic Lasso And Ridge Regression In Predicting Corporate Failure" Sciencedirect, 3rd Global Conference On Business, Economics, Management And Tourism, 26-28 November 2015, Rome, Italy
[9] Nitis Monburinon, Prajak Chertchom,Thongchai Kaewkiriya, Suwat Rungpheung, Sabir Buya, Pitchayakit Boonpou Title" Prediction Of Prices For Used Car By Using Regression Models"
[10] G.Rossum, "Python Reference Manual," Amsterdam, The Netherlands, The Netherlands, Tech. Rep., 1995.
[11] A. K. Elmagarmid, P. G. Ipeirotis, And V. S. Verykios, "Duplicate Record Detection: A Survey," Ieee Transactions On Knowledge And Data Engineering, Vol. 19, No. 1, Pp. 1–16, Jan 2007.
[12] G.Chandrashekar And F. Sahin, "A Survey On Featureselection Methods," Computers & Electrical Engineering, Vol. 40, No. 1, Pp. 16–28, 2014. [Online]. Available:
[13] J.Morgan, "Classification And Regression Tree Analy-Sis," Bu.Edu, No. 1, P. 16, 2014. [Online]. Available: Http://Www.Bu.Edu/Sph/Files/2014/05/Morgancart.Pdf
[14] N. Kanwal And J. Sadaqat, "Vehicle Price Prediction System Using Machine Learning Techniques," International Jounal Of Computer Ap-Plications, Vol. 167, No. 9, Pp. 27–31, 2017.
[15] R.Taylor, "Interpretation Of The Correlation Coefficient: A Basic Review," Journal Of Diagnostic Medical Sonography, Vol. 6, No. 1, Pp. 35–39, 1990.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY