A Study on used Cars Price Prediction using Regression Model with Reference to Cartrade.Com

Y. Sudheer[1], Dr. P. Viswanath[2]

[1]Student, [2]Assistant Professor,
[1,2]School of Management Studies, JNTUA, Anantapur, Andhra Pradesh, India

ABSTRACT

Predicting the true value of used cars requires helps customers to know best price to buy in online used cars market. The aim of this project is developing machine learning models to predict used cars price accurately based on its features, such as car model, fuel type, number of owners, kilometers driven, etc. The results shows that Random Forest model and K-Means clustering with linear regression yield the best results, using python. Out of the three models, Random Forest model predicted price with more accuracy.

***KEYWORDS**: Kmeans, Linear Regression, Random forest regression*

***How to cite this paper**:* Y. Sudheer | Dr. P. Viswanath "A Study on used Cars Price Prediction using Regression Model with Reference to Cartrade.Com" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-6 | Issue-6, October 2022, pp.782-789, URL: www.ijtsrd.com/papers/ijtsrd51956.pdf

IJTSRD51956

INTRODUCTION

"**India Used Car Market By Vehicle Type, By Sector, By Sales Channel, By Fuel Type, Competition Forecast & Opportunities, 2012 – 2022**", India used car market is projected to reach over $ 66 billion by 2022, on the back of growing population and rising urbanization in the country. Increasing focus of automakers towards setting up used car networks in different parts of the country and growing inclination of consumers towards used cars owing to their affordability and improved after sales services are some of the other major factors expected to boost demand for used cars in India in the coming years.

Moreover, market growth is anticipated to be driven by rising penetration of online platforms such as cartrade.com, OLX, Quikr, etc., that enable used car dealers to boost their reach to a larger audience.

Some of the major players operating in India used car market are Maruti True Value, Mahindra First Choice Wheels, Hyundai H Promise, Das Welt Auto, Ford Assured, Toyota U Trust, Honda Auto Terrace, BMW Premium Selection, Audi Approved Plus, Mercedes-Benz Certified, etc.

DEFINITION:

Price prediction:

Price prediction uses an algorithm to analyze a product or service based on its characteristics, demand, and current market trends. Then the software sets a price at a level it predicts will both attract customers and maximize sales.

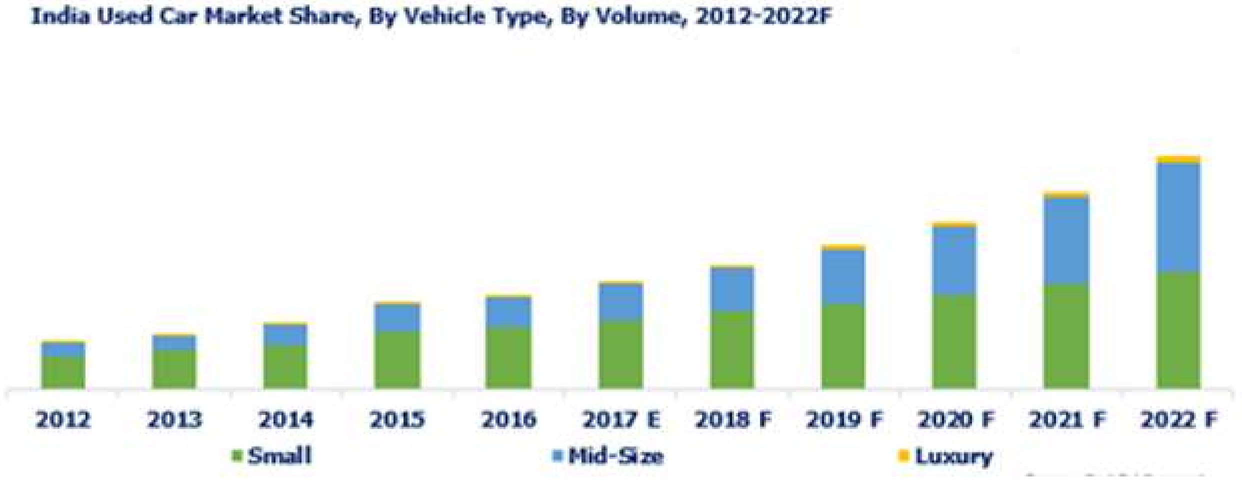
In some circles, the practice is called price forecasting or predictive pricing. And some people are a little sceptical of it. That’s because more experienced operators often feel they have a solid understanding of their industry prices.

Growing Demand for Luxury Used Cars to Play Key Role in the Market

The Indian pre-owned car market is growing due to a steady increase in the demand for luxury cars. The sales of used luxury cars observed a 20% growth.

➢ **Market, by Vehicle Type:**

- Small
- Mid-Size
- Luxury



Figer-1

Until a few years ago, it was not easy to purchase a luxury car due to its high cost. However, this trend is gradually changing as consumers can now buy pre-owned luxury vehicles. The market is becoming more organized with easy access to financing options, annual maintenance contracts, and lower entry prices. Also, the average age of used luxury vehicles coming into the market is between 2 and 3 years compared to 5-6 years for a mid-size or small-scale vehicle, making them a better option in some cases.

As per auto dealers, the demand for pre-owned luxury cars has been rising at ~35-40% on a Y-o-Y basis, as owners of luxury cars usually sell off their vehicles after a year or two and upgrade to better models.

Market Overview:

The Indian used car market was valued at USD 32.14 billion in 2021, and it is expected to reach USD 74.70 billion in 2027, registering a CAGR of 15.1% during the forecast period (2022-2027).

The COVID-19 pandemic had a minimal impact on the industry. With the increased number of people preferring individual mobility and more finance options available in the used car market, the market is set to grow considerably.

Reduced cash inflow due to the pandemic has forced buyers to look for alternatives other than new cars, and the used car industry has high growth potential in these terms. As the sales and production of new vehicles have been hindered due to the pandemic, the used car market is gaining traction among buyers.

However, the used car market evolved in the country with the growth of the organized and semi-organized sales sectors. The pre-owned car market recorded sales of 4.4 million units in FY2020 compared to only 2.8 million units of new passenger vehicles in the same year.

REVIEWE OF LITERATURE:

Monburinon, et al., 2018 Gathered data from a German e-commerce site that totalled to 304,133 rows and 11 attributes to predict the prices of used car using different techniques and measured their results using Mean Absolute Error (MEA) to compare their results. Same training dataset and testing dataset was given to each model. Highest results achieved was by using gradient boosted regression tree with a MAE of 0.28, and MEA of 0.35 and 0.55 for mean absolute error and multiple linear regression respectively. Authors suggested adjusting the parameters in future works to yield better results, as well as using one hot encoding instead of label encoding for more realistic data interpretations on categorical data.

Gongqi, Yansong, & Qiang, 2011 proposed using Artificial Neural Network (ANN) through a combined method of BP neural network and nonlinear curve fit and have achieved accurate value prediction with a feasible model.

Gegic, Isakovic, Keco, Masetic, & Kevric, 2019 from the International Burch University in Sarajevo, used three 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.

K. Samruddhi & Kumar, 2020 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 accuracy results are achieved. The model was also cross validated with 5 and 10 folds by using K fold method.

On the whole, Review of literature reveals that though Studies are made used cars price prediction using different models, there are very few studies made on analysis on used cars price prediction using regression models focusing Cartrade.com. For then unpredicted economic conditions viz., financial reason demonetization and covid 19 etc. may have a major influence on prices Therefore, there is a gap in analysing cars price prediction using regression model under this scenario.

NEED OF THE STUDY:

Customers now a days using used cars by comparing different features at best price with the help of technology cartrade.com. predicting used cars price got significance due to competing in availability of cars in used cars market. At this context the study under taken to predict selling price based on machine learning with different features.

SCOPE OF THE STUDY:

The study covers to analyse the price of used cars by collecting 19 years of past data with selected features like type of fuel, auto transmission, engine, power(bhp)and millage, etc.

OBJECTIVES OF THE STUDY:

- To predict the price of used cars.
- To compare cars price based on type of fuel.

RESEARCH METHODOLOGY:

To predict the used cars price secondary data has been used for 19 years. price predicted by using forecasting models like KMenas-cluster, linear regression, random forest regression models. The data is collected from the Kaggle website (www.Kaggle.com) relating to all used car companies and brands.

LIMITATIONS:

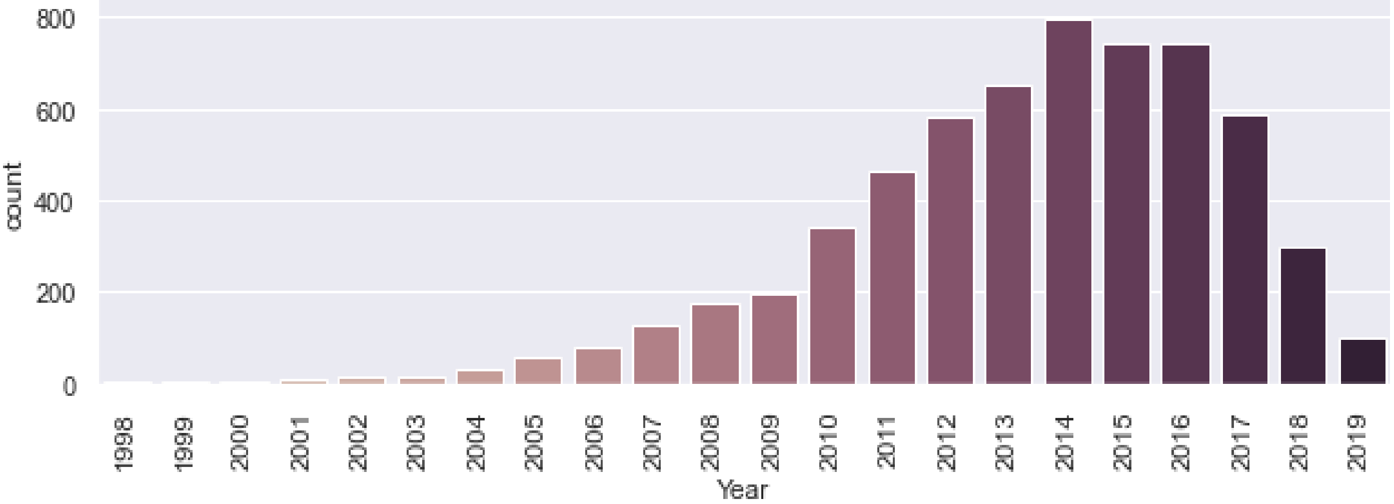
- The study is limited to 19 years of used cars.
- The study is confined to used car brands' price prediction.
- The study is limited to 45 days.

DATA ANALYSIS AND INTERPRETATION:

- To analyse the comparative price artificial intelligence algorithms used with the help of python program. The result are presented below figers 1,2,3,4,5,6.

- The present study used cars price has been predicted using independent variable features like type of fuel, auto transmission, engine, power(bhp)and millage, year of manufacturing, kilometors driven etc.
- To predict price their models like KMeans-cluster, linear regression and random forest regression models has been used.

Data analysis and interpretation, analysis of year attribute:

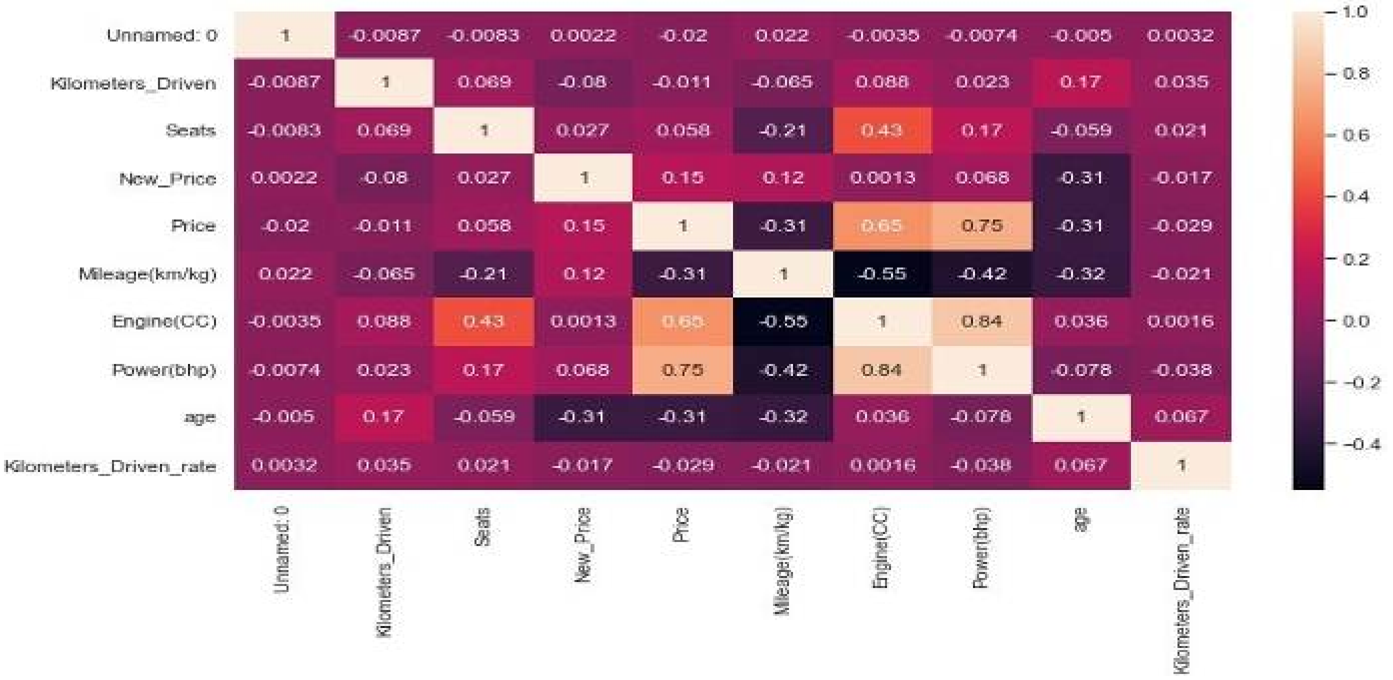


Figer-2

INTERPRETATION:

The analysis shows that count all years of used cars almost 6020 used cars from dataset and first highest year of 800 used cars count in 2014 year and the least of used cars in 1998.

CORRELATION ANALYSIS OF DATA

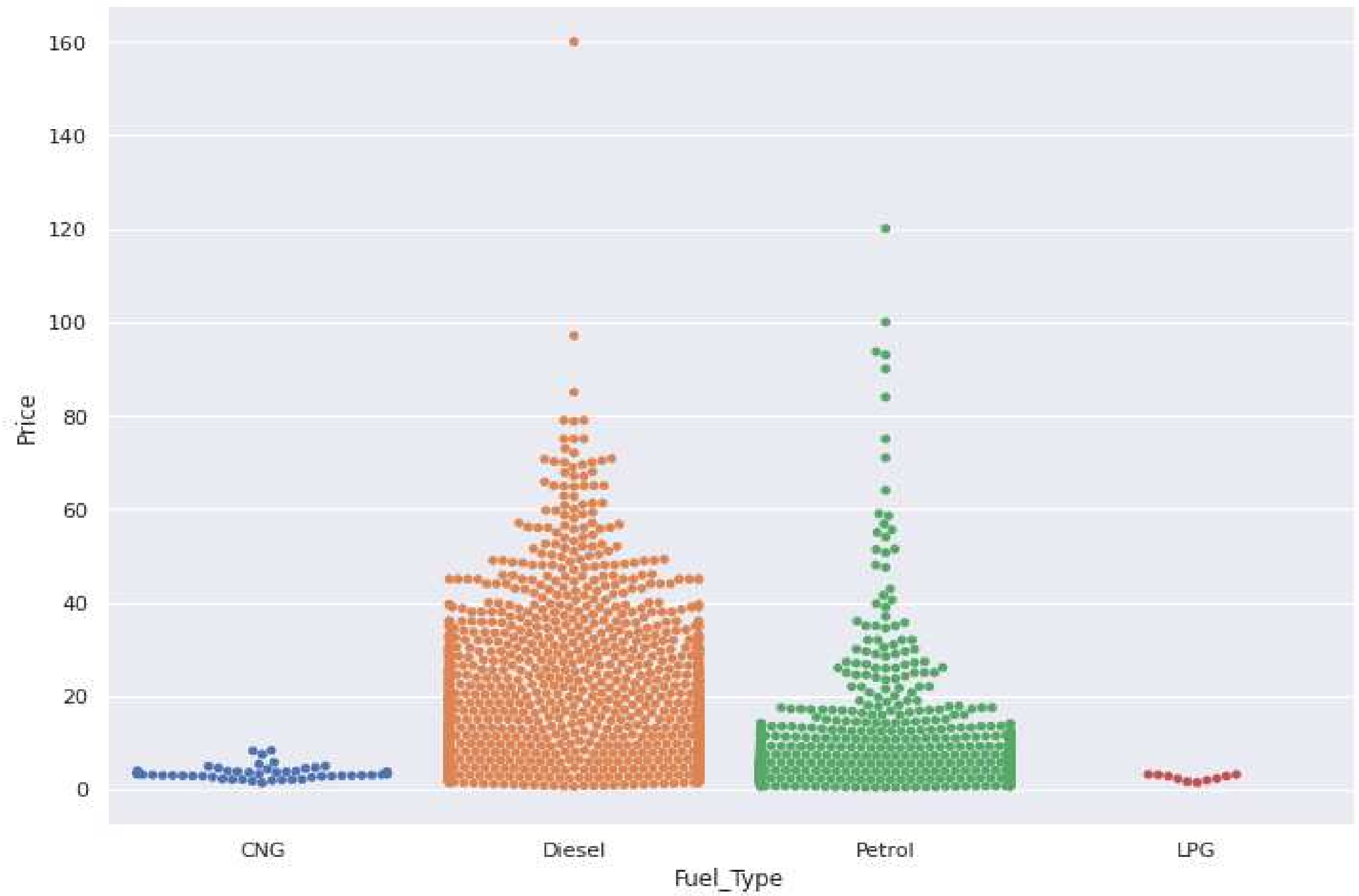


Figer-3

INTERPRETATION:

- The analysis visual shows that explain to the correlation of used cars given features.
- Analysis shows that correlation taken -0.1 to 1.0.
- Heat map of Correlation Features for Final Dataset: The correlation features of a dataset define the closeness of two variables to have a linear relationship with each other. Features having high correlation would be more linearly dependent and also have same impact on the dependent variable. In case two variables have a high correlation, we can always drop one of them. The following is the heat map of correlation where the darker color resembles a high correlation and light color represents low correlation. the many variables show a positive correlation.

ANALYSIS OF ENGINE FUEL TYPE CARS PRICE OF USED CARS:



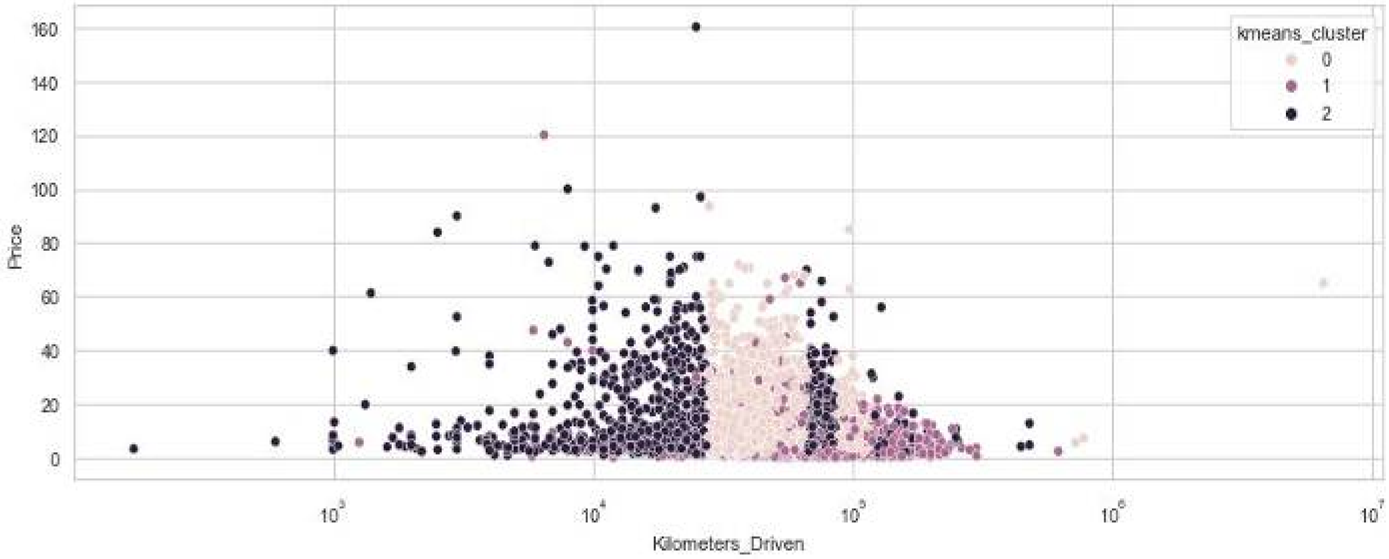
Figer-4

INTERPRETATION:

- The above analysis shows to explain engine fuel cars' prices of used cars.
- The Compared to used cars fuel type and price then diesel used cars very high prices and LPG fuel type and CNG fuel type engine of used cars is very low prices. And petrol used cars medium prices.

MODEL: 1

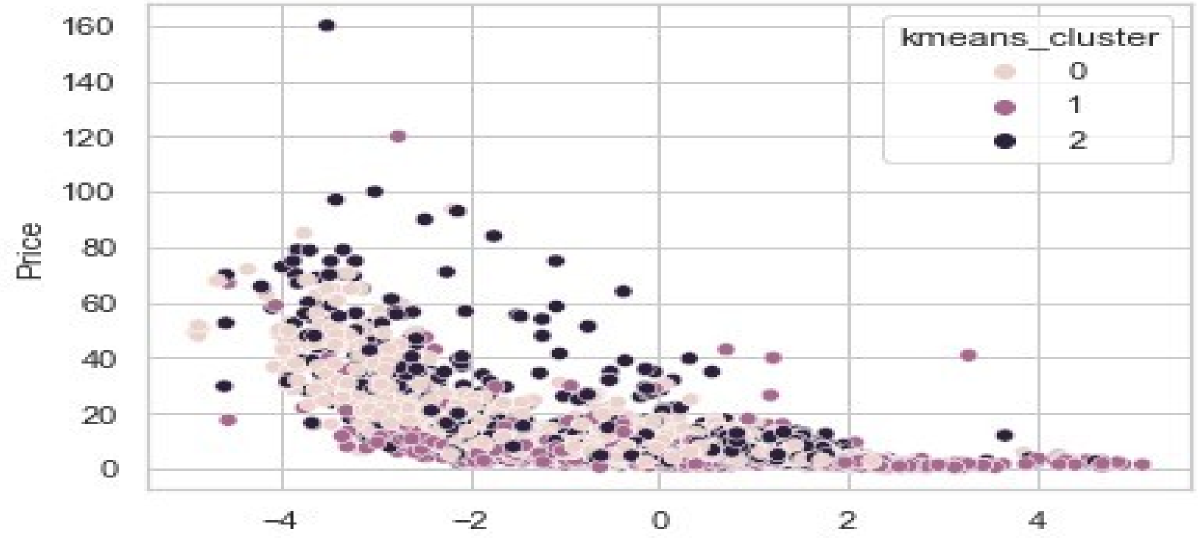
KMEANS_CLUSTER MODEL:



Figer-5

INTERPRETATION:

- The analysis shows that used cars price in Y axis and kilometers_Driven in X axis.
- The analysis shows that divided into three clusters is 0,1,2 and this clusters is lite wilate color to dark wilate color.



Figer-6

INTERPRETATION:

The analysis shows that divided into three clusters 0,1,2 based on the price segmentation of used cars. And the kmeans_cluster divided on the three colors.

MODEL:2

LINEAR REGRESSION:

```
print("R2 score on Traing set: %.2f " % linear_reg.score(X2_train,y2_train))
print("R2 score on Testing set: %.2f" % linear_reg.score(X2_test,y2_test))
print('Mean squared error: %.2f'% mean_squared_error(y2_test,y2_pred))
```

OUTPUT:

```
R2 score on Traing set: 0.92
R2 score on Testing set: 0.90
Mean squared error: 0.09
```

INTERPRETATION:

The linear regression algorithm was the best performer with r2 score on testing set of 0.90 and r2 score on traing set of 0.92 and mean squared error of 0.09.which simply signified the fact that it generated the most accurate predictions.

MODEL:3

RANDOM FOREST REGRESSION:

```
print("R2 score on Traing set: %.2f"% rf_reg.score(X2_train,y2_train))
print("R2 score on Testing set: %.2f"% rf_reg.score(X2_test,y2_test))
print('Mean squared error: %.2f'% mean_squared_error(y2_test,y2_pred))
```

OUTPUT:

```
R2 score on Traing set: 0.99
R2 score on Testing set: 0.92
 Mean squared error: 0.08
```

INTERPRETATION:

The Random forest regression algorithm was the best performer with highest r2 score on testing set of 0.92 and r2 score on traing set of 0.99 and mean squared error of 0.08.which simply signified the fact that it generated the most accurate predictions.

RESULT:

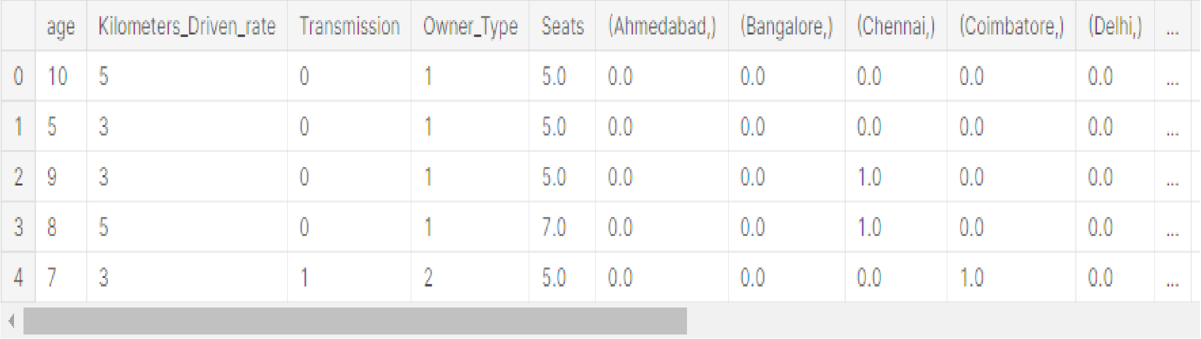
Table: 1

Algorithm	r2 Traing	r2 Testing	MSE
Random Forest regression	0.99	0.92	0.08
Linear regression	0.92	0.90	0.09

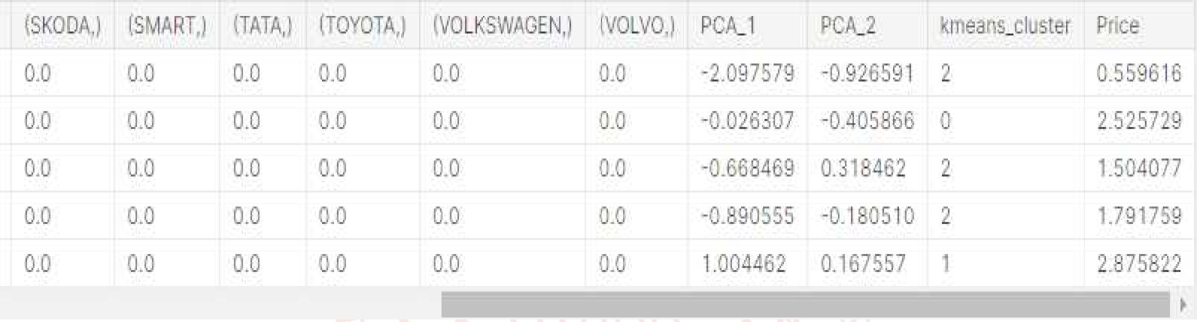
From the r_2 scores comparison of all regression algorithms, the random forest Algorithm has the best r_2 score of 0.92 which simply means that the Random forest Algorithm has given the most accurate predictions in comparison to the other algorithms.

FINAL OUTPUT FOR USED CARS PREDICTED PRICES:

Table 2:



	age	Kilometers_Driven_rate	Transmission	Owner_Type	Seats	(Ahmedabad,)	(Bangalore,)	(Chennai,)	(Coimbatore,)	(Delhi,)	...
0	10	5	0	1	5.0	0.0	0.0	0.0	0.0	0.0	...
1	5	3	0	1	5.0	0.0	0.0	0.0	0.0	0.0	...
2	9	3	0	1	5.0	0.0	0.0	1.0	0.0	0.0	...
3	8	5	0	1	7.0	0.0	0.0	1.0	0.0	0.0	...
4	7	3	1	2	5.0	0.0	0.0	0.0	1.0	0.0	...



(SKODA,)	(SMART,)	(TATA,)	(TOYOTA,)	(VOLKSWAGEN,)	(VOLVO,)	PCA_1	PCA_2	kmeans_cluster	Price
0.0	0.0	0.0	0.0	0.0	0.0	-2.097579	-0.926591	2	0.559616
0.0	0.0	0.0	0.0	0.0	0.0	-0.026307	-0.405866	0	2.525729
0.0	0.0	0.0	0.0	0.0	0.0	-0.668469	0.318462	2	1.504077
0.0	0.0	0.0	0.0	0.0	0.0	-0.890555	-0.180510	2	1.791759
0.0	0.0	0.0	0.0	0.0	0.0	1.004462	0.167557	1	2.875822

INTERPRETATION:

- The analysis table shows that used cars price prediction using regression models (machine learning models are random forest regression, linear regression and kmeans_ cluster) it is most of the accurate results for used cars price.
- The analysis shows that predicted of used cars price based on the kmeans- cluster and age of used car analysis to better understanding the prices.
- 10 years of used cars kmeans _cluster is 2 then predicted price of 0.559616.
- 3 years of used cars kmeans _cluster is 1 then predicted price of 2.875822.
- 5 years of used cars kmeans _cluster is 0 then predicted price of 2.525729.
- 9 years of used cars kmeans _cluster is 2 then predicted price of 1.504077.
- 8 years of used cars kmeans _cluster is 2 then predicted price of 1.791759.

FINDINGS:

- The count all years of used cars almost 6020 used cars from dataset and first highest year of 800 used cars count in 2014 year and the least of used cars in 1998.
- Analysis ofo correlation taken -0.1 to 1.0. Heatmap of Correlation Features for Final Dataset: The correlation features of a dataset define the closeness of two variables to have a linear relationship with each other.
- The Compared to used cars fuel type and price then diesel used cars very high prices and LPG fuel type and CNG fuel type engine of used cars is very low prices. And petrol used cars medium prices.
- The analysis divided into three clusters is 0,1,2 and this clusters is white wilate color to dark wilate color.
- From the r_2 scores comparison of all regression algorithms, the random forest Algorithm has the best r_2 score of 0.92 which simply means that the Random forest Algorithm has given the most accurate predictions in comparison to the other algorithms.
- The analysis shows that predicted of used cars price based on the kmeans- cluster and age of used car analysis to better understanding the prices.

CONCLUSION:

Predicting prices of a used car is a challenging task because of a high number of features and parameters that should be considered to generate accurate results. The first and foremost step is data gathering and preprocessing data. Then a model was defined and created for implementing algorithms and generating results. After applying various regression algorithms on the model, it could be concluded that random forest regression algorithm was the best performer with highest r2 score of 0.92 which simply signified the fact that it generated the most accurate predictions. Apart from a best r2 score, random forest also had the least Mean Squared Error and Root Mean Squared Values that shows that the errors in predictions were least among all and therefore the results generated are highly accurate.

SUGGESTIONS:

- It is trite knowledge that the value of used cars depends on several factors. The most important ones are usually the age of the car, its make (and model), the origin of the car (the original country of the manufacturer), its mileage (the number of kilometres it has run) and its horsepower.
- The five most important car specifications for predicting car price are **Age, Kilometers, Auto transmission, Fuel type and Automatic aircondition**. Compare the prediction errors of the training and validation sets by examining their RMS error and by plotting the box plots.

References:

[1] Sameerchand Pudaruth, Computer Science and Engineering Department, University of Mauritius, Reduit, MAURITIUS. Predicting the Price of Used Cars using Machine Learning Techniques. International Journal of Information & Computation Technology, 2014.

[2] Saamiyah Peerun, Nushrah Henna Chummun and Sameerchand Pudaruth, University of Mauritius, Reduit, Mauritius. Predicting the Price of Second-hand Cars using Artificial Neural Networks. Proceedings of the Second International Conference on Data Mining, Internet Computing, and Big Data, Reduit, Mauritius 2015.

[3] Ashish Chandak, Prajwal Ganorkar, Shyam Sharma, Ayushi Bagmar, Soumya Tiwari, Information Technology, Shri Ramdeobaba College of Engineering, Rashtrasant Tukadoji Maharaj Nagpur University, Nagpur. Car Price Prediction Using Machine Learning. India International Journal of Computer Sciences and Engineering, May 2019.

[4] Laveena D'Costa, Ashoka Wilson D'Souza, Abhijith K, Deepthi Maria Varghese. Predicting True Value of Used Car using Multiple Linear Regression Model. International Journal of Recent Technology and Engineering (IJRTE). January 2020.

[5] S. E. Viswapriya, Durbaka Sai Sandeep Sharma, Gandavarapu Sathya Kiran. Vehicle Price Prediction using SVM Techniques. International Journal of Innovative Technology and Exploring Engineering (IJITEE), June 2020.

[6] Enis Gegic, Becir Isakovic, Dino Keco, Zerina Masetic, Jasmin Kevric, International Burch University, Sarajevo, Bosnia and Herzegovina. Car Price Prediction using Machine Learning Techniques. TEM Journal, February 2019.

[7] https://www.researchgate.net/publication/318667714_Car_resale_price_forecasting_The_impact_of_regression_method_private_information_and_heterogeneity_on_forecast_accuracy/link/5a29c20f0f7e9b63e5352f8c/download

[8] https://www.ijrte.org/wpcontent/uploads/papers/v8i5s/E10100285S20.pdf

[9] https://www.irjet.net/archives/V8/i4/IRJET-V8I4278.pdf

[10] https://www.researchgate.net/publication/356756110_Price_Prediction_for_Pre-Owned_Cars_Using_Ensemble_Machine_Learning_Techniques

[11] **Source:**https://www.mordorintelligence.com/industry-reports/india-used-car-market

[12] **Source:**https://www.techsciresearch.com/report/india-used-car-market-by-vehicle-type-small-mid-size-luxury-by-sector-organized-vs-semi-organized-unorganized-by-sales-channel-dealership-broker-vs-c2c-by-fuel-type-petrol-others-competition-forecast-opportunities/1239.html