Resale Car Price Prediction

Dr. Vijay Gaikwad[1], Sanjyot Kotgire[2], Shravan Raina[3], Rutuja Shinde[4], Saksham Sharma[5], Aditya Sabde[6]

[1]Ph. D. (E&TC), Dean Quality Assurance, Vishwakarma Institute of Technology, Pune
[2, 3,4,5,6] Department of Computer Science, Vishwakarma Institute of Technology, Pune

[1] vijay.gaikwad@vit.edu, [2] sanjyot.kotgire20@vit.edu, [3] shravan.raina20@vit.edu, [4] rutuja.shinde20@vit.edu,
[5] saksham.sharma20@vit.edu, [6] aditya.sabde20@vit.edu

***Abstract*—**

The Pre-owned cars or the used cars have increased markets across the globe right now. There are several factors and facts according to which the price of an original or a pre-owned car can be stated. These factors play a major role for second hand car price predictions in this paper. The dataset gives the factors and actual prices about various AUDI car models. Examining the algorithms, the used classifiers in this paper are simple linear regression, decision trees and random forest and comparing these classifiers with # main factors (Adj.r.sqr, RMSE, MAE). The use of machine learning techniques can help forecast predicted prices of second-hand cars. This paper introduces a system that has been used to predict the fair value of any pre-owned vehicle.

Keywords— *Pre-owned AUDI models, Regression analysis, Random Forest, Decision Tree, caret*

I. INTRODUCTION

According to Indian Blue Book's most current survey of the used car market in India, 4 million used cars were acquired and sold between 2018 and 2019. The market for used cars has benefited buyers and sellers alike. The demand for used cars is growing due to the increasing need for private vehicles around the world, which presents opportunities for both buyers and sellers. The market for used cars has expanded even though the market for brand-new vehicles has decreased.

The vehicle price prediction is actually very crucial and important for seller and buyerMore and more automobile purchasers are looking for alternatives to outright purchasing new cars due to the rise in interest for used cars and up to 8% decline in demand for new cars in 2013. People like leasing their vehicles, which is a formal agreement between the buyer and seller. The seller category includes insurance companies, third parties, commercial organizations, and direct sellers. In a lease, the buyers pay for the item over a certain length of time with monthly payments.

The majority of people choose to acquire old automobiles since they are less expensive and they can resell them after a few years of use for a profit. However, a number of factors, like the age of the cars and their current condition, affect the price of a used car. In general, used car prices on the market remain constant. Consequently, a car price model is required to aid in trade. The cost of a used car is determined by a number of criteria, including the kind of gasoline, color, model, mileage, transmission, engine, and number of seats. The market price of second hand cars will continue to fluctuate.

Let us consider a situation where a person has to sell his old car and that person does not know the actual price of the car. Without knowing the price, how can a car be sold? If the person already knows the price of the car by giving some information related to the car's performance it would be easy for them to know the price of the car. The system we are building will help to predict the price of the old cars.

II. RELATED WORK

The paper [1] proposed a model that uses the K closest neighbor approach to predict the pricing of second hand cars, which is suitable for small data sets. They compiled and evaluated a dataset of used cars. They used various ratios of test and training data to assess the model's accuracy after it had been trained on the data. The K- Fold approach, which is straightforward to understand was used to cross-validate the same model in order to assess the model's performance.

In [2], they used supervised learning methods like Naive Bayes, decision trees, K-nearest Neighbor, and random forest algorithms to investigate various aspects of heart disease. The patient database for Cleveland heart disease provided a pre-existing dataset that was used in the study. It makes use of a previously collected dataset from the Cleveland heart disease patient database. The collection contained 303 instances with 76 attributes. In spite of their significance in demonstrating the effectiveness of different algorithms, only 14 out of the 76 attributes were examined. The purpose of this study report is to forecast who will get heart disease.

In [3] firstly they have done the simplifying of the dataset e.g assuming a numeric value for the data (Expensive as 3, Affordable as 2, Normal as 1, etc.). Then they apply cosine similarity on the customer's reviews and if the value of >= 0.5 then the review is positive otherwise it is negative. In the next step, they have applied the four supervised learning algorithms i.e. SVM, KNN, Random forest and Naive Bayes on their dataset have selected one algorithm which gives best results for the dataset.

To determine the values of variables that matched the training dataset the best, [4] employed Lasso regression. The Lasso Regression variables chosen are then employed in the Multiple Regression. The pricing is then modelled to the chosen set of variables using a regression tree. Before trimming, the tree had 344 leaf nodes; after pruning, there were only 152. The price of the test data, which consisted of 241 records, was then predicted using the trained models, and the error rates for lasso regression, multiple regression, and regression tree are provided.

In paper [5], statistical models are built using machine learning technologies. Employing supervised machine learning methods including KNN, Random Forest, Decision Trees, Linear Regression, and XG Boost. 92386 entries will be used in the dataset used to train the model. These five methods can be used to determine an automobile's value based on factors such kilometers driven, year of registration, fuel type, car model, financial strength, car brand, and gear type.

In order to create a price model for used automobiles, [6] compares gradient boosted regression, random forest regression, and multiple linear regression trees. The qualities (variables) that were severely out of balance and unable to predict prices were eliminated. After choosing the appropriate range, 19% of the data was removed because the average price was 17,295.14 and the standard deviation was 3,587,954. The data was divided for training and testing into 0.67 and 0.33, respectively.

In [7] the algorithm technique used goes as selecting random samples, building a decision tree for every subset, nextly collecting the output for each of the decision trees then averaging out the predicted values and finally the average value is the final predicted value. Mean Absolute Error, Mean Squared Error, and Root Mean Squared Error are metrics that are taken into account while evaluating the algorithm's robustness. Both regression techniques are assessed using these three measures.

The model in [8] is based on a number of methods. In that case, Random Forest, Support Vector Machine, and Artificial Neural Network (ANN) (RF). For prediction, ANN took into account a number of factors, including brand, expected automobile life, and kilometers driven. Compared to other linear models, the non-linear model had a higher degree of accuracy in predicting automobile prices. Combining these algorithms yields 92% accuracy compared to individual methods' 52% accuracy.

In [9], they introduced a model that uses the K closest neighbor approach, which is straightforward and suited for small data sets, to predict the pricing of secondhand automobiles. Here, we have gathered and examined a dataset of used autos. The model was trained on the data, and we looked at the model's accuracy using various ratios of the training and test sets. Using the simple to comprehend and apply K- Fold approach, the same model is cross-validated to evaluate the model's performance. This research suggests using modal analysis, which has an accuracy of 85% whereas linear regression has a 71% accuracy. With a K-value of 4, the

RMSE rate was 4.01, the MAE rate was 2.01, and the RMSE rate was 4.01. Using the K Fold Method, the suggested model is additionally verified with 5 and 10 folds.

III. METHODOLOGY

A. Data Processing

The KAGGLE datasets are one of the most commonly used dataset repo which are mainly used for implementing machine learning algorithms and Appling classifier models. The data must be reliable and well-structured in order to use machine learning algorithms. Data preprocessing must be done with a dataset before applying prediction algorithms. Data preprocessing involves omitting the null values, discarding the outliers, converting values to one of a kind, removing irrelevant data, syntax errors, and standardization and converting small values to null values. This dataset contains two categories of information: (a.) numerical and (b.) categorical.

The dataset variables which are character type must be converted to the factor or the numeric variables so as to apply prediction algorithms. Dummy data frames must be created to check whether any null values are present in variables.

```
> summary(data)
    model          year         transmission     mileage         fuelType
 A3     :1922   Min.   :1997   Automatic:2654   Min.   :     1   Diesel:5557
 Q3     :1390   1st Qu.:2016   Manual   :4350   1st Qu.:  6000   Hybrid:  28
 A4     :1380   Median :2017   Semi-Auto:3590   Median : 19095   Petrol:5009
 A1     :1346   Mean   :2017                    Mean   : 24892
 A5     : 879   3rd Qu.:2019                    3rd Qu.: 36538
 Q5     : 872   Max.   :2020                    Max.   :323000
(Other):2805
      tax              mpg          engineSize       price
 Min.   :  0.0   Min.   : 18.90   Min.   :1.000   Min.   :  1490
 1st Qu.:125.0   1st Qu.: 40.90   1st Qu.:1.500   1st Qu.: 15000
 Median :145.0   Median : 49.60   Median :2.000   Median : 20000
 Mean   :125.7   Mean   : 50.83   Mean   :1.939   Mean   : 22848
 3rd Qu.:145.0   3rd Qu.: 58.90   3rd Qu.:2.000   3rd Qu.: 27988
 Max.   :580.0   Max.   :188.30   Max.   :6.300   Max.   :145000
```

Fig.1 Data Summary

By summarizing the data, it gives basic information about the used dataset in the paper. The elements in variables which have less significance and those which have lesser values are removed from the dataset before applying any algorithm. In this case few models were removed from the dataset after summarizing. The values only where mileage was between range 1000 and 70000 were kept, omitting the other values. The cars which tend to have a higher mileage value are mostly considered as very old and in poor conditions, which are un-sellabe at second hand prices.

The given dataset had a total of 10668 values, considering no null values were present. After omitting the values which were less significant and had lesser impact on predictions the values present in the dataset were 9646.

SR.NO	VARIABLE	TYPE
1	Model	fct
2	Year	int
3	Transmission	fct
4	Mileage	int
5	FuelType	fct
6	Tax	int
7	Mpg	dbl
8	EngineSize	dbl
9	Price	int

Fig.2 Glimpse of dataset

These used functions give the variable type along with total rows and total number of variables or columns present in the dataset. Fig 2 gives all variables used in the dataset, the significance of each variable is considered at prediction time. Variables which are categorical are converted to numeric type for ease of calculations using levels function.

B. Data splitting

After data cleaning and pre-processing the raw data gets converted to processed data. The dataset should be further divided into: (a.) training and (b.) testing data. The training data applies the algorithms and predicts for the testing data. The training dataset contains 70% (6430 values) of random variables from the dataset. We consider random variables using the set.seed function which does not discriminate between training and testing data values. The caret function helps to split the data easily using createdatapartition function. The remaining values are considered into the testing data after removing the outliers. Removing outliers reviews the data which is present between 1st and 3rd quartile range and discards other values.

C. Machine learning algorithms

Machine learning explores multiple approaches and adapts to different settings over time. Experience is gained through training using a set of data known as training data. Machine learning algorithms predict or classify data without making explicit programs after training. In speculative mathematical algorithms, attributors with a coefficient of high correlation generally, have a greater influence on predictive variability. A mathematical statistic that illustrates the relationship between variables is the correlation coefficient. The two qualities' correlation coefficient always runs from 1 (the best relationship) to -1 (the worst relationship), with 0 denoting no relationship at all. We plot the ggcorr matrix to determine the correlation between the variables. Our prediction model will be fitted using training data, and its performance will be assessed using testing data.

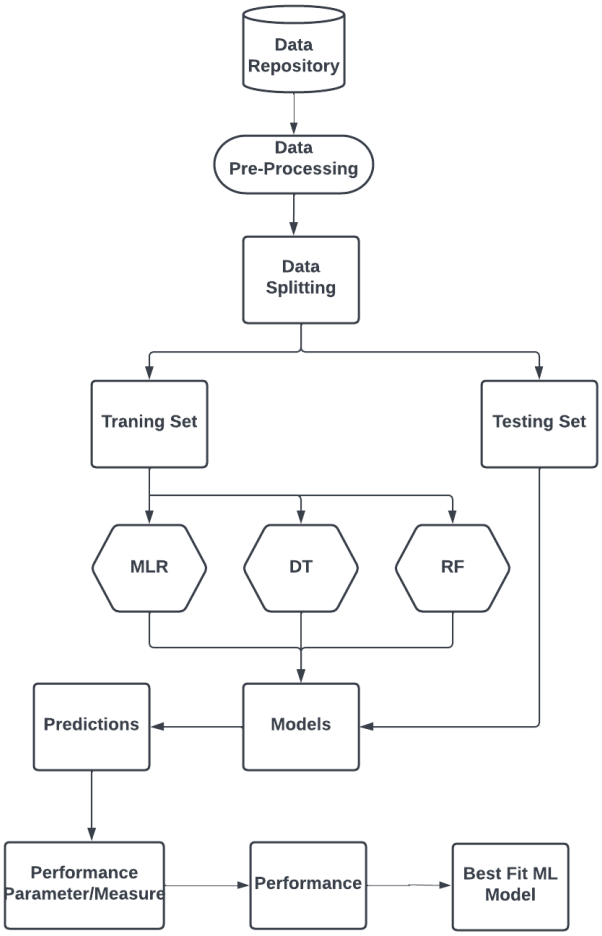

Fig.3 Data flow diagram

Fig 3 shows relation between all variables except the factor variables. The price is related most with engineSize,tax and year. Here most related variables are colored in orange scale. Three supervised techniques are used to predict the prices for second hand car models, these classifiers are considered according to the survey.

Fig.4 Correlation between variables

1) ***Regression:*** An expansion of linear regression into connections between more than two variables is known as multiple regression. While there is just one predictor and one response variable in a basic linear relationship, there are numerous predictor variables and response variables in multiple regression. In this study, the following equation is used to define multiple linear regression:

$$y = a + b^1x^1 + b^2x^2 + \ldots\ldots + b^nx^n$$

he y= response variable (prediction variable),

a, b^1, b^2,..., b^n, are coefficients, x^1,x^2,.... x^n, are predictor variables,

n= number of variables

2) ***Decision tree:*** A graph that expresses selection and its results as a tree is a decision tree. The graph's edges stand in for the rules of decision, while the notes on the graph represent the event, choice, or selection. To plot the trees on the console in this paper, libraries are used. The function accepts input from both dependent and independent variables. The tree splits into 2 trees at each point. The function of rpart prediction gives as:

$$rpart(formula, data)$$

Formula takes dependent variable which is to be predicted on the given data.

3) ***Random Forest (RF):*** Random Forest is a learning technique that creates many decision trees during training and outputs a class of individual trees. It may be used for regression as well as classification.
It filters a small number of feature columns out of all feature columns during bootstrapping. After training, predictions of unknown inputs may be expressed as:

$$f' = \frac{1}{B}\sum_{b=1}^{B} f_b(x')$$

Where, B= Optimal number of trees

Working:
1) It randomly selects ‘m’ features from the dataset
2) Finding the best fit for each split in the tree.
3) Node splitting using the best split.
4) Repeat (1-3) and finally build a forest of DT’s.

IV. RESULTS

Comprehensive simulations were run after the implementation of different machine learning methods. For each ML model 3 factors were compared to get the best fit classifier for the given dataset for predictions:

1) Adjusted R Square

Adjusted r-square is a revised form of r-square for which the value rises when new variables improve model performance and falls when new predictors do not boost efficiency as predicted. In regression analysis, the R-square test is used to measure the goodness of fit. The term "goodness of fit" refers to how well a regression model fits the data points. The closer the r-square number is to 1, the better the model. However, the difficulty is that the value of r-square grows as additional variables (qualities) are included in the model, regardless of whether the recently introduced attributes have a measurable impact on the model. Also, unless there are a lot of variables, this can result in overfitting of the model.

2) Root Mean Square Error(RMSE)

The standard deviation of the residuals is known as the root mean square error (RMSE) (prediction errors). The distance between the data points and the regression line is represented by residuals, and the RMSE is a measurement of how widely spaced out these residuals are. To put it another way, it shows how closely the data falls along the line of best fit. Root mean square error is frequently employed in climatology, forecasting, and regression analysis to examine the effectiveness of the experiment..

3) Mean Absolute Error(MAE)

The MAE statistic evaluates the total amount of forecasting errors without considering the direction of the errors. The accuracy of continuous variables is evaluated. In other words, the MAE is the average of the absolute differences between the forecast and the corresponding observation over the validation sample. Since the MAE is a linear score, it is implied that all small differences are equally weighted in the average.

A Matrix table is created for each classifier method for each test case to check the performance measure. For each of the classifier methods, a data frame is created to check the actual and the predicted price for the given second hand car models dataset.

a) ***Regression***: For regression lm function is used for predicting car prices. 3 regression models are built and checked for the best fit regression model for this paper. The 1st model checks with all the other independent variables and the last prediction is done on the basis where most relevant dependent variables are.

	actual	predicted1	predicted2	predicted
5	17300	20242.784	24681.128	21546.864
6	13900	14471.185	12320.372	12408.605
10	12000	14429.862	13663.654	13783.091
13	17000	17071.704	17442.062	18199.864
17	15700	20028.636	18757.539	19827.535

Fig.5 Comparing values with regression

In fig5. Column 1 given actual values with other columns give the prediction values according to the models used in regression. The adj.r.squared value for the regression model is **0.865.**

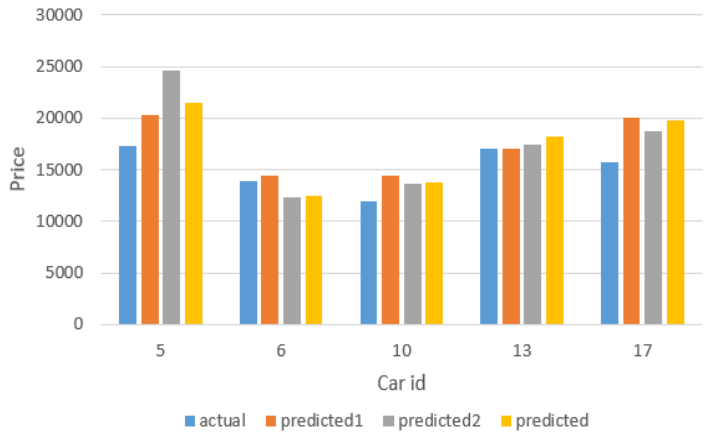

Fig.6 Regression comparing graph

In fig6. all the models used to predict prices are shown graphically. First bar is for the actual prices and last one for the best fit model for regression classifier.

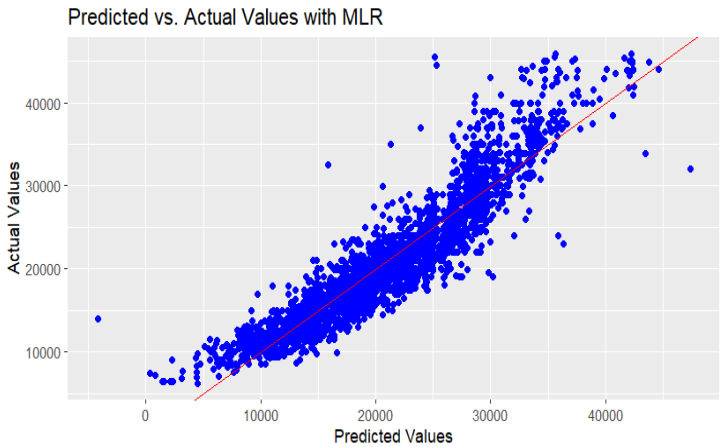

Fig.6 MLR plot for predicted values

Fig6. shows a detailed plot about the actual values and the predicted values with the regression model for the test data. A regression line is plotted to predict the values easily.

b) ***Decision Tree****:* Rpart and Rpart.plot function are used for predicting the prices. The decision tree is plotted using a function(where each node splits into 2 trees). Fig 6 shows the actual and the predicted values for a given dataset. The parament variables used to predict price in this model are:

(year+model+engineSize+mpg+mileage)

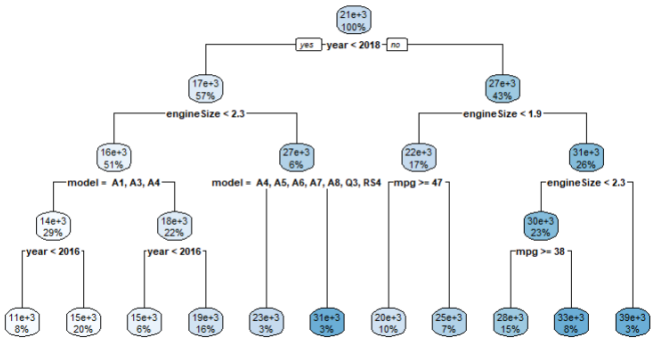

Fig.7 Decision Tree

Fig 7. describes the split at each node and finally gives the selected prices for each model

	actual	**predicted**
5	17300	19903.85
6	13900	15136.39
10	12000	15136.39
13	17000	19317.36
17	15700	19317.36

Fig.8 Comparing values with decision tree

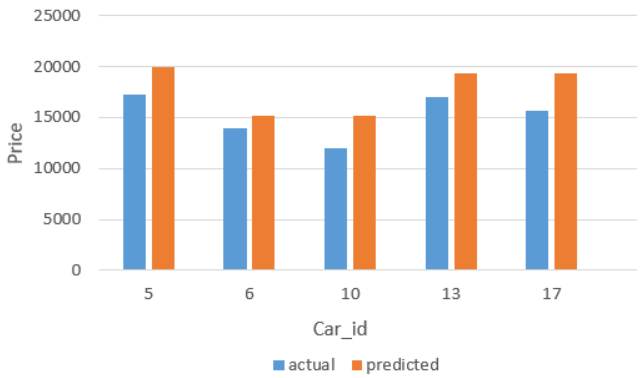

Fig.9 DT comparing graph

Fig 8,9. shows the prices actual vs predicted in tabular and graphics formats. This suggested that the predicted values for the head section of the test data are comparatively greater than the actual given car prices for those car ids. The adj.r.squared value for the Decision Tree model is **0.703.**

c) Random Forest: Basically, random forest in combination of multiple decision trees where the best fit nodes are split into 2 are each iteration.

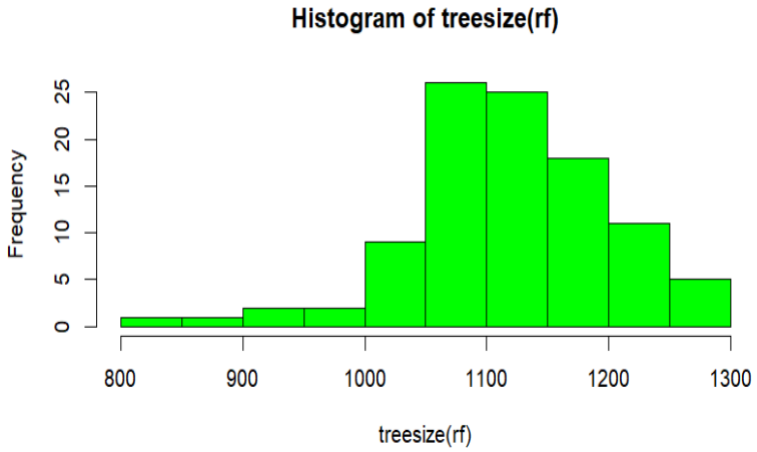

Fig.10 Tree size

Various functions of the random forest help us to get the variables which have the highest purity. The nodes which have highest purity can be used to predict the car prices efficiently with greater accuracy.

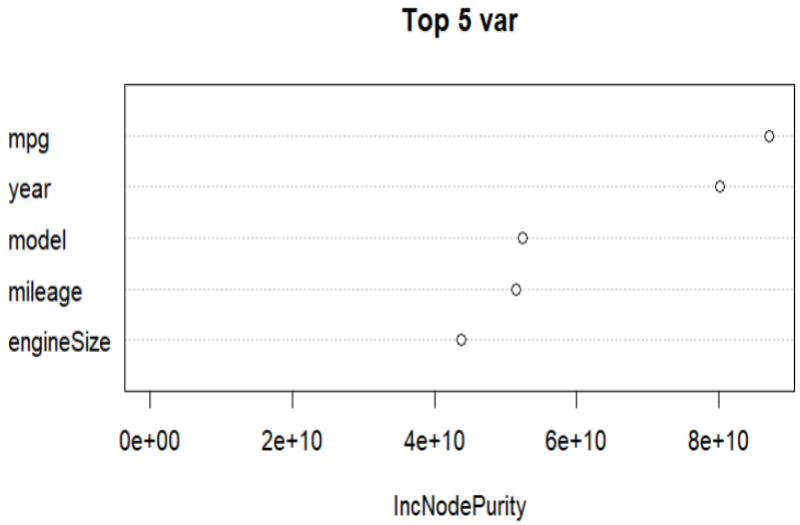

Fig.11 Top 5 variables in RF

Fig11. gives the top 5 variables chosen which are used for prediction in the next iteration of Random Forest. Here mpg has the highest node purity node which is in decreasing order till engine Size variable.

	actual	predicted
5	17300	18112.59
6	13900	12940.35
10	12000	12596.18
13	17000	16991.55
17	15700	17966.68

Fig.12 Comparing values with random forest

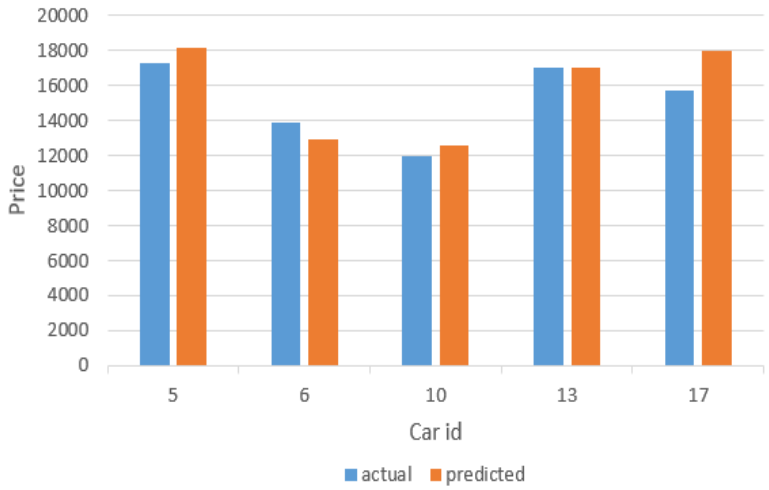

Fig.13 RF comparing graph

Above figs give the predicted prices for the random forest classifier. The predicted values were closer to the real values compared to other classifier outputs. The adj.r.squared value for the Decision Tree model is **0.932.** The best among the used classifiers in this paper.

V. CONCLUSION

Price prediction of second hand cars can be a challenging task as there are many numbers of attributes on which actual price prediction is to be done. In this we have 9 such attributes. We have used three algorithms that are regression, decision tree and random forest in our project.

ALGORITHM	R SQUARE	RMSE	MAE
Multiple Linear Regression	0.8643	2969.34	2177.49
Decision Tree	0.7035	3933.58	2989.08
Random Forest	0.9328	1896.25	1387.68

Fig.14 Comparing all algorithms

This is a comparison between all the required performance parameters along with their measures for all the classifiers used in the paper. According to the values of Adj R square, the best fit model or the best fit classifier for the given values of the dataset, Random Forest suits the best as it has the highest values of R square. This suggests that using Random Forest classifiers the predicted values are closer to the real values. Lesser the RMSE better is the performance of the classifier. Random forests hold the lowest value hence the error occurred for prediction are lesser in this algorithm. MAE also needs to be lesser while predicting. Random forest qualifies for all the 3 performance parameters. Hence Random Forest is the best fit classifier with given top 5 variables for this dataset. According to the performance paraments the algorithms used are ordered as :

1) Random Forest
2) Regression
3) Decision Tree

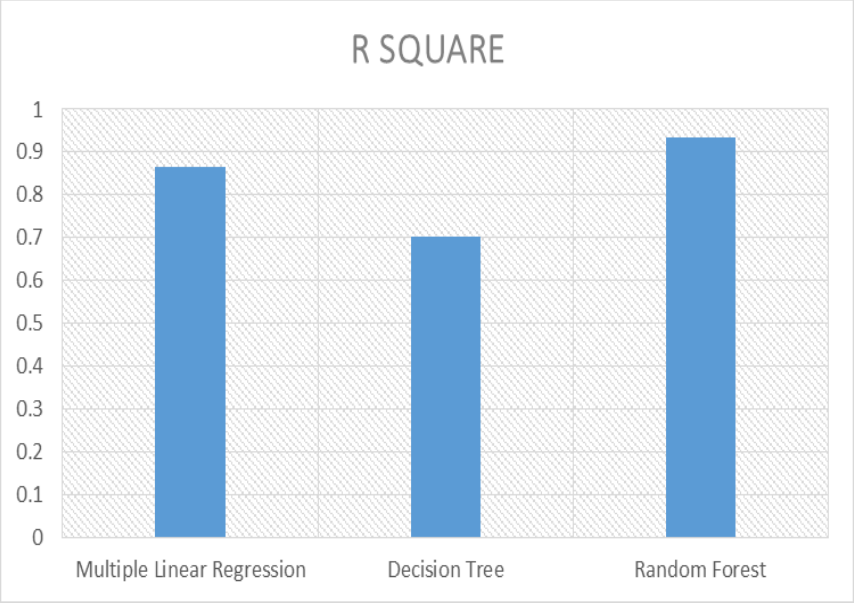

Fig.15 MLR comparison

As the R-Square values are much lesser compared to other parameters, Fig12. given a detailed graph for the same.

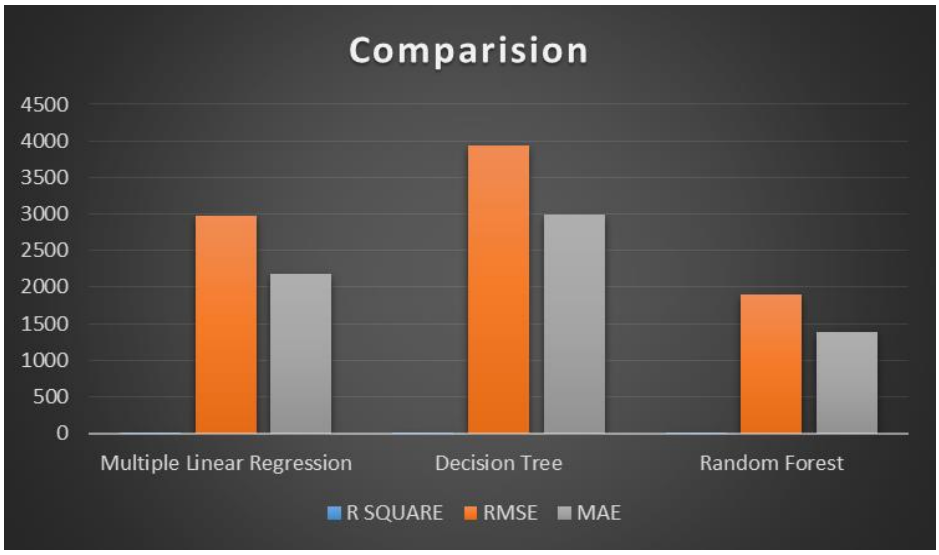

Fig.16 Comparison table

With the help of these three algorithms we have predicted the prices of the cars (regression Fig.5, decision tree Fig.6 and random forest Fig.8). R-Square, RMSE and MAE values are also evaluated Fig.11.A comparison is also made between these three values of each algorithm Fig.12 where all the three values of multiple linear regression are smaller than random forest values and these random forest values are smaller than values of decision tree.

VI. FUTURE SCOPE

However, due to the dataset's restricted number of observations, there was a very tiny dataset for drawing a firm conclusion. IN future for accurate price prediction the attributes can be increased or a big dataset can be taken. As we know gathering more dataset leads to better accuracy of prediction. We can apply some more algorithms such as KNN , Naive Bayes for more accurate prediction.

REFERENCES

[1] Kumar, K. Samruddhi1 Dr R. Ashok. "Used Car Price Prediction using K-Nearest Neighbor Based Model."

[2] Shah, D., Patel, S. & Bharti, S.K. Heart Disease Prediction using Machine Learning Techniques. *SN COMPUT. SCI.* 1, 345 (2020). https://doi.org/10.1007/s42979-020-00365-y

[3] A. Das Mou, P. K. Saha, S. A. Nisher and A. Saha, "A Comprehensive Study of Machine Learning algorithms for Predicting Car Purchase Based on Customers Demands," 2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD), 2021, pp. 180-184, doi: 10.1109/ICICT4SD50815.2021.9396868.

[4] Ganesh, Mukkesh & Venkatasubbu, Pattabiraman. (2019). Used Cars Price Prediction using Supervised Learning Techniques. International Journal of Engineering and Advanced Technology. 9. 216-223. 10.35940/ijeat.A1042.1291S319.

[5] Gajera, Prashant, Akshay Gondaliya, and Jenish Kavathiya. "Old Car Price Prediction With Machine Learning." *Int. Res. J. Mod. Eng. Technol. Sci* 3 (2021): 284-290.

[6] Monburinon, Nitis; Chertchom, Prajak; Kaewkiriya, Thongchai; Rungpheung, Suwat; Buya, Sabir; Boonpou, Pitchayakit (2018). [IEEE 2018 5th International Conference on Business and Industrial Research (ICBIR) - Bangkok, Thailand (2018.5.17-2018.5.18)] 2018 5th International Conference on Business and Industrial Research (ICBIR) - Prediction of prices for used car by using regression models. , (), 115–119. doi:10.1109/ICBIR.2018.8391177

[7] Gegic, Enis & Isakovic, Becir & Kečo, Dino & Mašetić, Zerina & Kevric, Jasmin. (2019). Car price prediction using machine learning techniques. TEM Journal. 8. 113-118. 10.18421/TEM81-16.

[8] C. R. Madhuri, G. Anuradha and M. V. Pujitha, "House Price Prediction Using Regression Techniques: A Comparative Study," 2019 International Conference on Smart Structures and Systems (ICSSS), 2019, pp. 1-5, doi: 10.1109/ICSSS.2019.8882834.

[9] Pingale, Kedar, Sushant Surwase, Vaibhav Kulkarni, Saurabh Sarage, and Abhijeet Karve. "Disease prediction using machine learning." *International Research Journal of Engineering and Technology (IRJET)* 6 (2019): 831-833.

[10] Venkatasubbu, Pattabiraman, and Mukkesh Ganesh. "Used Cars Price Prediction using Supervised Learning Techniques." *Int. J. Eng. Adv. Technol.(IJEAT)* 9, no. 1S3 (2019).

[11] Mammadov, Huseyn. "Car Price Prediction in the USA by using Linear Regression." *International Journal of Economic Behavior (IJEB)* 11, no. 1 (2021): 99-108.

[12] Longani, Chetna, Sai Prasad Potharaju, and Sandhya Deore. "Price Prediction for Pre-Owned Cars Using Ensemble Machine Learning Techniques." In *Recent Trends in Intensive Computing*, pp. 178-187. IOS Press, 2021.

[13] Pudaruth, Sameerchand. "Predicting the price of used cars using machine learning techniques." *Int. J. Inf. Comput. Technol* 4, no. 7 (2014): 753-764.

[14] Fathalla, Ahmed, Ahmad Salah, Kenli Li, Keqin Li, and Piccialli Francesco. "Deep end-to-end learning for price prediction of second-hand items." *Knowledge and Information Systems* 62, no. 12 (2020): 4541-4568.

[15] Noor, Kanwal, and Sadaqat Jan. "Vehicle price prediction system using machine learning techniques." *International Journal of Computer Applications* 167, no. 9 (2017): 27-31.

[16] Khan, Zamar. "Used Car Price Evaluation using three Different Variants of Linear Regression." *International Journal of Computational and Innovative Sciences* 1, no. 1 (2022).

[17] Pudaruth, Sameerchand. "Predicting the price of used cars using machine learning techniques." Int. J. Inf. Comput. Technol 4, no. 7 (2014): 753-764.