e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
Volume:03/Issue:03/March-2021 Impact Factor- 5.354 www.irjmets.com

OLD CAR PRICE PREDICTION WITH MACHINE LEARNING

Prashant Gajera [*1], Akshay Gondaliya[*2], Jenish Kavathiya[*3]

[*1]Student, Department of Computer Science & Engineering, Parul Institute of Technology, Vadodara, Gujarat, India.

[*2]Student, Department of Computer Science & Engineering, Parul Institute of Technology, Vadodara, Gujarat, India.

[*3]Student, Department of Computer Science & Engineering, Parul Institute of Technology, Vadodara, Gujarat, India.

ABSTRACT

The world is growing day by day and also expectations of every people are also growing up. Out of all the expectation one of them is to buy a car. But all are not able to buy always a new car, so they will buy used one. But new person don't know about the market price for his or her dream car for old one. That is where we have need a platform which helps new people for car price prediction. In this paper we are coming up with that platform which is made using machine learning technology. Using supervised machine learning algorithms such as linear-regression, KNN, Random Forest, XG boost and Decision tree, let's try to build a statistical model which will be able to predict the price of a used car. For that previous consumer data and a given set of features will helps us. And we will also be comparing the prediction accuracy of these models to determine the optimal one.

Keywords: Analysis, Research, Machine Learning, Random Forest, XG boost, Decision Tree, Linear Regression

I. INTRODUCTION

Concluding whether a used car is worth the posted price when you see listings online can be difficult. Several factors, including mileage, make, model, year, km driven etc. can affect the actual worth of a car. From the perspective of a seller, it is also a difficulty to price a used car appropriately. Based on existing data, the aim is to use machine learning algorithms to develop models for predicting used car prices. Features are, Kilometers traveled – We know that the number of kilometers traveled by a vehicle has a huge role to play while putting the vehicle up for sale. The more the vehicle has traveled, the older it is, horse power – It is the power output of the vehicle. More output yields better value out of a vehicle, Age – It is the year when the vehicle was registered with the Road Transport Authority. The newer the vehicle is; the better value it will yield. By every passing year, the value will depreciate, Fuel Type – There were four types of fuel types present in the dataset that we had. Diesel, Electrical, Gasoline and Unknown (for other types of fuel), Model, Gear Type – There were two gear types present in the dataset that we had, Automatic and manual.

II. METHODOLOGY

In this section, we discuss various algorithms and the required dataset that were implemented to build this module. A dataset containing 92386 records will be used for training the model. Attributes such as kilometers traveled, year of registration, fuel type, car model, fiscal power, car brand and gear type determine the worth of an automobile. Since this is a regression problem, we have implemented five algorithms – K Nearest Neighbors (KNN) Regressor, Random Forest Regressor, Linear Regression, XG Boost Regressor and Decision Tree Regressor.

KNN_Regressor

We have used KNN with neighbor range 1 to 100 and plotted a graph mean squared error vs number of neighbors and we found that with the neighbors 6 we can get good result or accuracy.

Random Forest Regressor

To account for the large number of features in the dataset and compare a bagging technique with the Gradient Boost method.

Linear Regression

Quick to train and test as a baseline algorithm.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
Volume:03/Issue:03/March-2021 Impact Factor- 5.354 www.irjmets.com

XG Boost Regressor

To improve performance compared to standard Gradient Boosting using regularization, second order gradients and added support for parallel compute.

Decision Tree Regressor

Decision tree is used to build regression models and the structure is in the form of a tree. It breaks down a dataset into smaller and smaller subsets based on the information gain value for the each individual features while at the same time an associated decision tree is incrementally developed. Finally the result is a tree with decision nodes and leaf nodes.

III. MODELING AND ANALYSIS

We inspect the frequency of the different categorical variables to get a better understanding of their distribution and potentially drop some variables that may act as outliers.

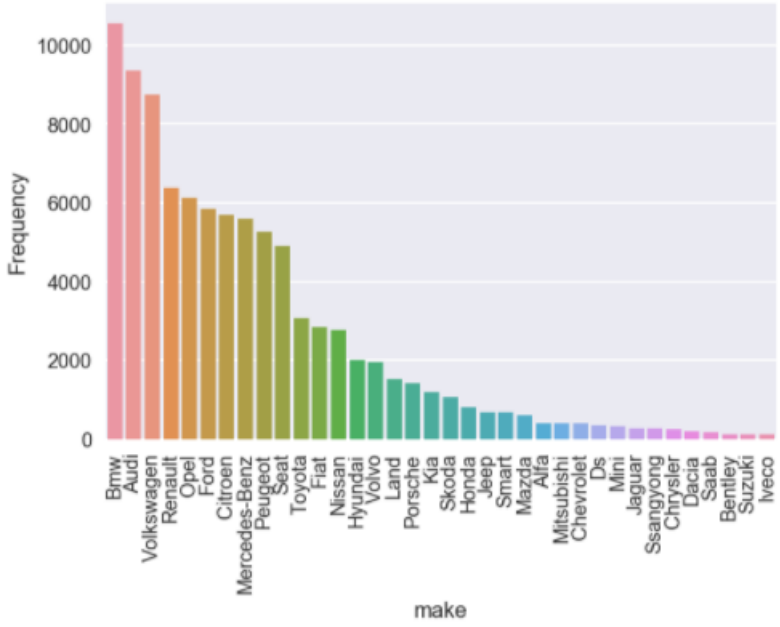

Figure 1: frequency plot of car make

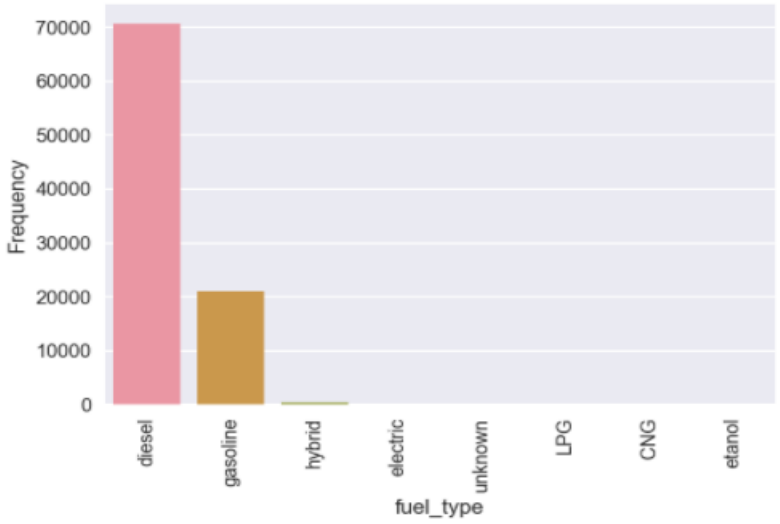

Figure 2: frequency plot of fuel type

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
Volume:03/Issue:03/March-2021 Impact Factor- 5.354 www.irjmets.com

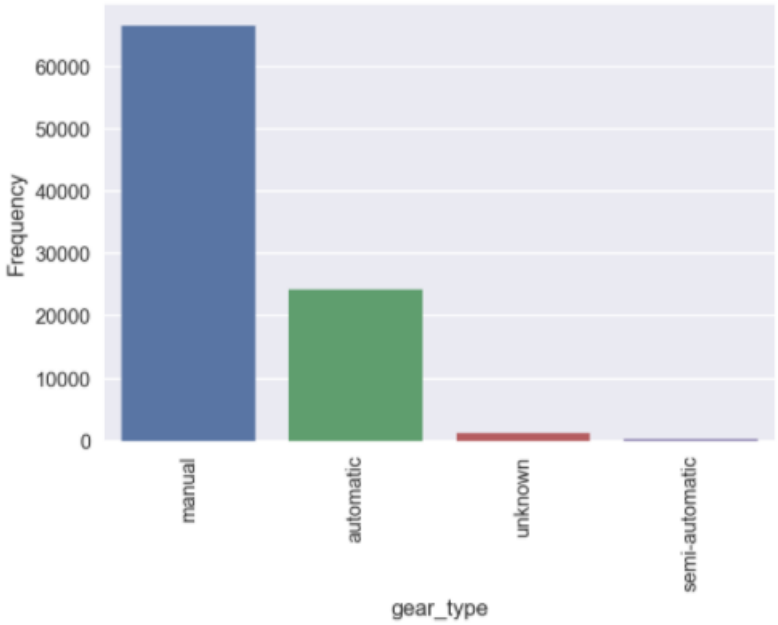

Figure 3: frequency plot of gear type

Based off the above frequency plots, I notice some categories may act "outliers" due to very low counts. We decided to drop two fuel type categories, ethanol & CNG. (Low frequency, very imbalanced)I decided to keep all of the categories within gear type. Lastly, I clean the outcome variable column 'price', from any outliers such as cars priced over 400,000 euros or less than 0 euros.

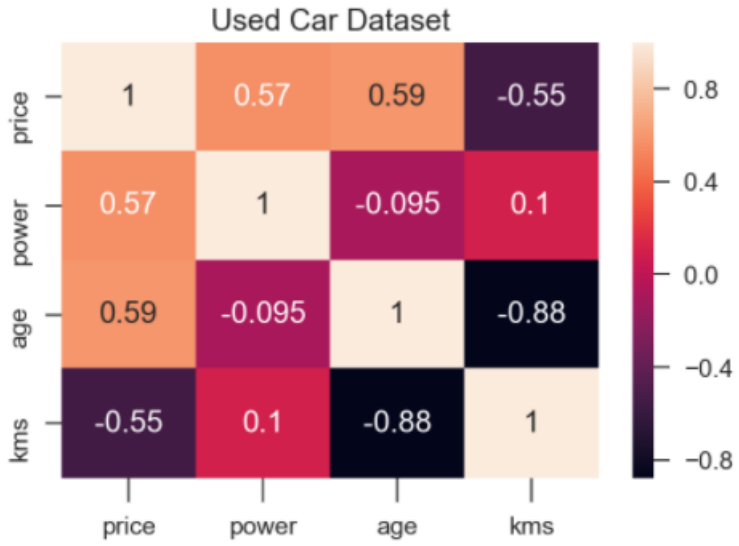

Figure 4: heat map

Using heat map assessing the correlation coefficients of the numeric variables. None of them are highly correlated with each other, so we cannot drop any feature out of these four.

IV. RESULTS AND DISCUSSION

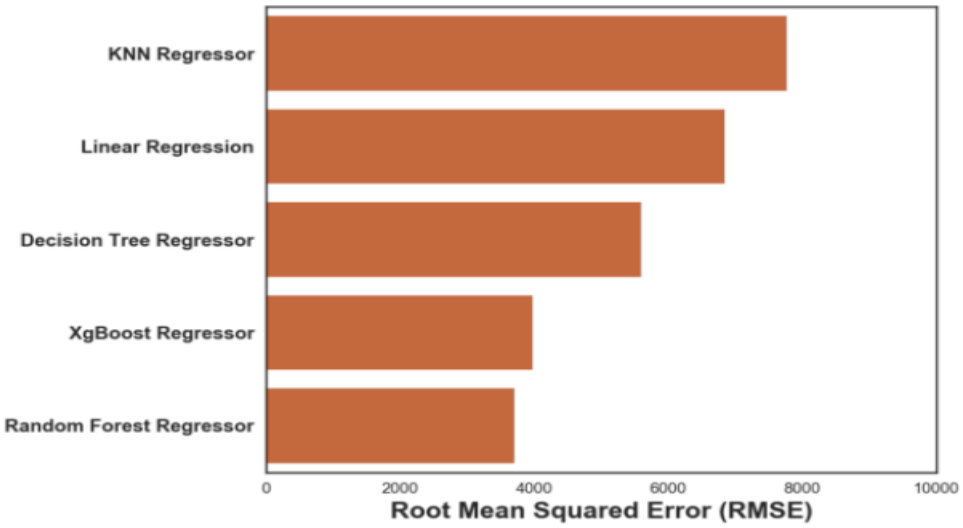

Figure 5: Root Mean Squared Error comparison

Figure 5 is showing that Random Forest Regressor has given lowest Root mean squared error on test data set as compared to all other algorithms. It simply means that Random forest has performed well.

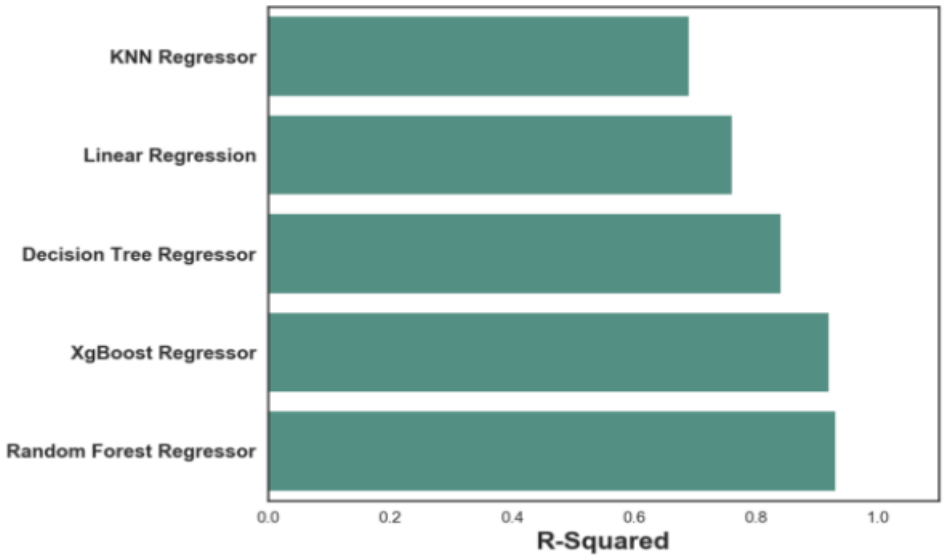

Figure 6: R-Squared Value Comparison

Figure 6 is used to analyze the R-Squared value of different-different algorithms. Here we can see that random forest and XG boost both has almost same and higher R2 Score as compare to other three algorithms. But Random forest has is the algorithm which has highest R2 score. So that we can say the variance between best fitted line and mean line of the data set is lowest compared to other algorithms based on the R2 Score formula.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
Volume:03/Issue:03/March-2021 Impact Factor- 5.354 www.irjmets.com

$$R^2 = 1 - \frac{\text{sum squared regression (SSR)}}{\text{total sum of squares (SST)}},$$
$$= 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}.$$

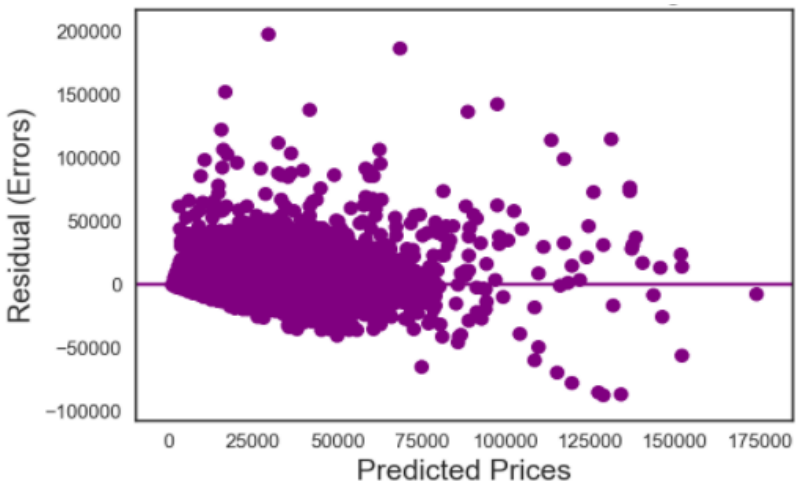

Figure 7: Residuals vs Fitted Plot KNN Regressor

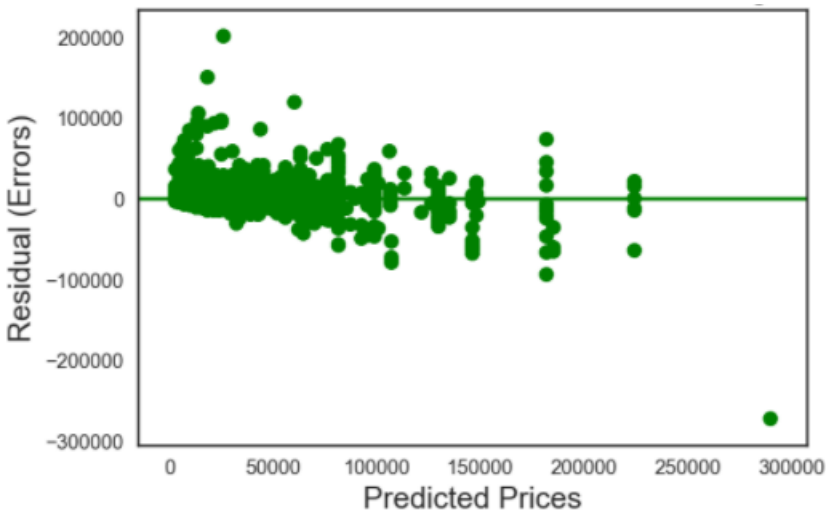

Figure 8: Residuals vs Fitted Plot Decision Tree Regressor

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
Volume:03/Issue:03/March-2021 Impact Factor- 5.354 www.irjmets.com

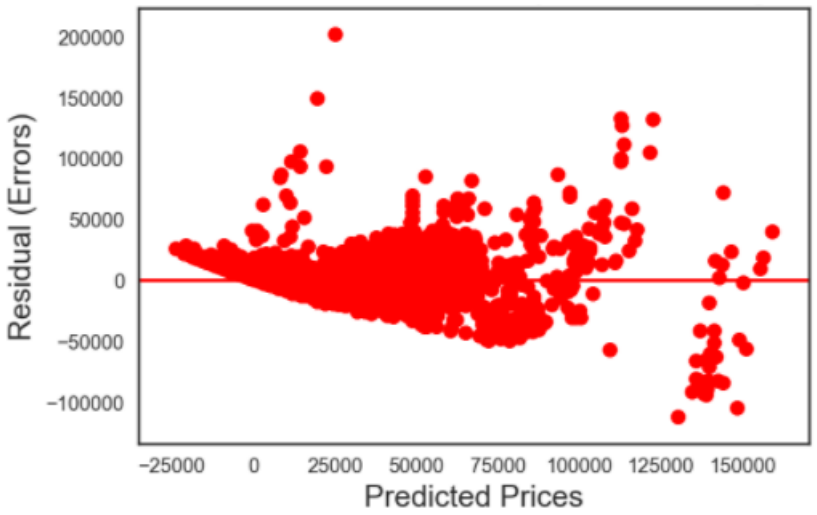

Figure 9: Residuals vs Fitted Plot Linear Regression

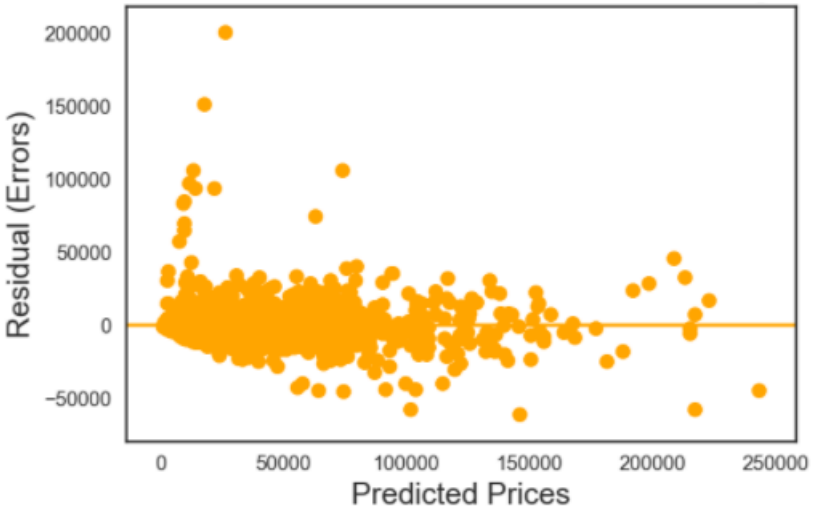

Figure 10: Residuals vs Fitted Plot Random Forest Regressor

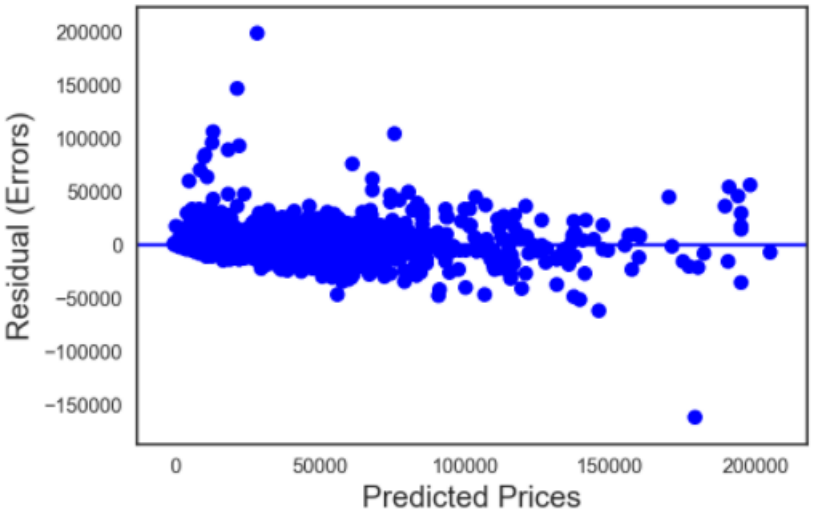

Figure 11: Residuals vs Fitted Plot XG Boost Regressor

Above Figures from 7 to 11 represents a comparison graph between Residual value vs predicted value, the dotted points from 0th line represent the deviation from actual against predicted for all five algorithms.

Table-1: Comparison of RMSE & Test data accuracy of all 5 models

SN.	Model Name	Root Mean Squared Error(RMSE)	Test Accuracy
1	Random Forest Regressor	3702.34	93.11 %.
2	KNN-Regressor	7771.91	69.66 %
3	Decision Tree Regressor	5590.43	84.30 %
4	XG Boost Regressor	3980.77	92.04 %
5	Linear Regression	6846.23	76.46 %

Table 1 contains the results of all five algorithms in a numeric format so than we can estimate correctly. We can see Random forest performed best with highest test accuracy 93.11 % and lowest RMSE value 3702.34.

V. CONCLUSION

We have used total five supervised machine learning models, and their root mean squared error are, KNN Regressor: 7771.09, Linear Regression: 6846.23, XG Boost: 3980.77, Random Forest: 3702.34 and Decision Tree Regressor: 5590.43. Out of all Random Forest has lowest RMSE, and performed well with highest R-squared value: 0.93. The limitation of this research is less number of records of old car. in future if we get more data than we can retrain our models and might get more accurate and stable model. This study used different models in order to predict used car prices. However, there was a relatively small dataset for making a strong inference because number of observations was only 92386. Gathering more data can yield more robust predictions. Secondly, there could be more features that can be good predictors. For example, here are some variables that might improve the model: number of doors, color, mechanical and cosmetic reconditioning time, used-to-new ratio and appraisal-to-trade ratio.

VI. REFERENCES

[1] Ashish Chandak, Prajwal Ganorkar, Shyam Sharma, Ayushi Bagmar, Soumya Tiwari, Car Price Prediction Using Machine Learning, International Journal of Computer Sciences and Engineering, Volume 7, Issue 5, May 2019.

[2] Durgesh k. Shrivstava, Lekha Bhambhu, “Data Classification Using Support Vector Machine”, Journal of Theoretical and Applied Information Technology, Sep. 2009.

[3] Pattabiraman Venkatasubbu, Mukkesh Ganesh, “Used Cars Price Prediction using Supervised Learning Techniques”, International Journal of Engineering and Advanced Technology (IJEAT), Vol. 9, Issue 1S3, Dec. 2019.

[4] Vrushali Y Kulkarni, Pradeep K Sinha, “Effective Learning and Classification using Random Forest Algorithm,” International Journal of Engineering and Innovative Technology (IJEIT), Vol. 3, Issue. 11, May 2014.