

Used Car Price Prediction Using Random Forest Algorithm

**Prof. Dipti A. Gaikwad¹, Pratik S. Suwarnakar², Yash R. Mahajan³,
Amita U. Petkar⁴, Shreyasi G. Theurkar⁵**

¹Assistant Professor, Information Technology, JSPM's Jayawantrao Sawant College of Engineering, Pune, India

^{2,3,4,5}Student, Information Technology, JSPM's Jayawantrao Sawant College of Engineering, Pune, India

Abstract

The price of a new car in the industry is fixed by the manufacturer with some additional costs incurred by the Government in the form of taxes. So, customers buying a new car can be assured of the money they invest to be worthy. But, due to the increased prices of new cars and the financial incapability of the customers to buy them, used car sales are on a global increase. Therefore, there is an urgent need for a used car price prediction system which effectively determines the worthiness of the car based on multiple aspects, including vehicle mileage, year of manufacturing, fuel consumption, transmission, road tax, fuel type, and engine size. We have developed a model which will be highly effective. This model can benefit sellers, buyers, and car manufacturers in the used cars market. Upon completion, it can output a relatively accurate price prediction based on the information that user's input. Various regression methods were applied in the research to achieve the highest accuracy. Because of which it will be possible to predict the actual price a car rather than the price range of a car. User Interface has also been developed which acquires input from any user and displays the Price of a car according to user's inputs. To evaluate the performance of each regression, R-square was calculated.

Keywords: Used car price prediction, Regression, Linear Regression, Lasso Regression, SVM, Random Forest, and Machine Learning.

1. Introduction

The used car market is a burgeoning industry with a significant market value that has almost doubled in recent years. To estimate the market worth of a used car, there are numerous internet resources and other tools available. These tools have made it simpler for both buyers and sellers to gain a better knowledge of the elements that go into determining a used car's market value. Any automobile's price can be predicted using machine learning algorithms based on a variety of variables.

The data set will contain details on a range of vehicles. For each car, details about the technical components of the vehicle, such as the engine type, fuel type, miles per gallon, and so forth, will be provided.

Since different websites use different methods to calculate the retail price of used cars, there is no comprehensive mechanism for doing so. Using statistical models, it is possible to forecast pricing

without having to enter all the information into the desired website. This study's main goal is to examine the precision of several forecasting algorithms for determining the suggested retail price of used cars.

Machine learning can be used to automate operations, enhance processes, forecast results, and make judgements based on prior experiences. Additionally, machine learning can be utilized to develop robust algorithms that can handle massive amounts of data. It enables software programmes to predict outcomes more accurately without having to be expressly designed to do so. In order to forecast new output values, machine learning algorithms use historical data as input.

As a result, we provide a machine learning-based methodology for estimating used automobile costs based on their specifications. The effectiveness of different machine learning algorithms, including Linear Regression, Lasso Regression, Support Vector Machine, and Random Forest, will be compared, and the best one will be chosen. We will figure out the cost of the car based on a number of factors. Because regression

algorithms provide us a continuous number rather than a categorized value as an output, it is possible to estimate a car's exact price rather than just its price range. Then, to analyze our findings, a user interface that accepts input from any user and displays the price of a car in accordance with user inputs has also been constructed. This methodology can help consumers who are looking to buy a second-hand car make better informed decisions. Customers can now look for all automobiles without any physical efforts, anytime and from any location.

2. Motivation Behind Project Topic

Deciding whether a used car is worth the posted price when you see listings online can be difficult. Several factors, including mileage, make, model, year, etc. can influence the actual worth of a car. From the perspective of a seller, it is also a dilemma 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.

3. Aim and Objective(S) of the Work

Project aim:

The aim of this project is to predict the car price as to develop an efficient and effective model which predicts the price of a used car according to user's inputs.

Project objectives:

- To predict the car price as per the data set (previous consumer data like engine capacity, distance travelled, year of manufacture, etc.
- To achieve good accuracy.
- To develop a User Interface (UI) which is user-friendly and takes input from the user and predicts the price.

4. Literature Survey

According to author Doan Van Thai et al [1], they used data inference, meaning extraction approaches, and rules for qualitative data in this research. The major goal of the current research is to investigate various automotive data types with the aim of developing an automated method to forecast car prices. They compared and built models using random forest, XGBoost, and LightGBM with r2 values utilising Kaggle and Vietnamese Datasets.

According to author Praful Rane, Deep Pandya and Dhawal Kotak [2], in this paper Regression Algorithms like Lasso, Linear, Ridge Regression are used because they provide us with continuous value as an output and not a categorized value. As a result, it will be feasible to forecast the exact cost of an automobile rather than its price range. A user interface that accepts input from any user and displays the price of a car based on their inputs has also been developed.

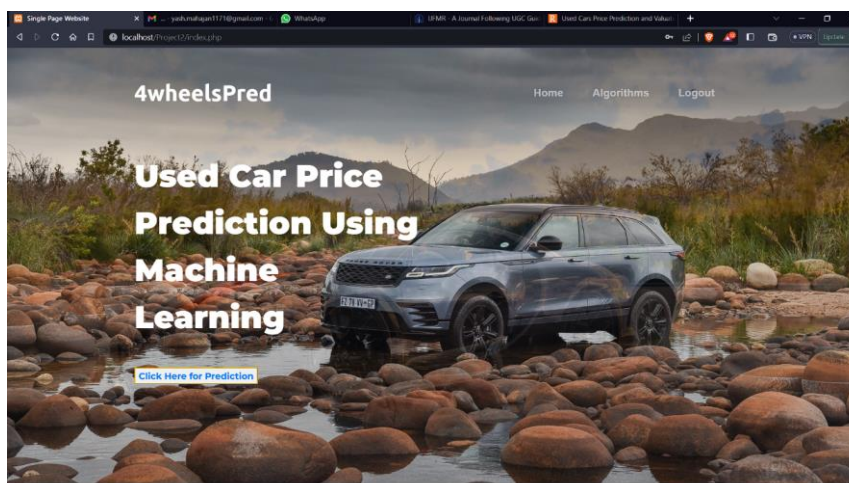
According to author Anamika Das Mou et al [3], to improve accuracy for a car purchase, they suggested some well-known algorithms in this paper, including SVM, Naive Bayes, and KNN. These algorithms were used on their dataset, which consists of 50 data. With a prediction accuracy of 86.7%, Support Vector Machine (SVM) produces the best result of the group. Additionally, they compare the precision, recall, and F1 score for all data samples using various algorithms in this study.

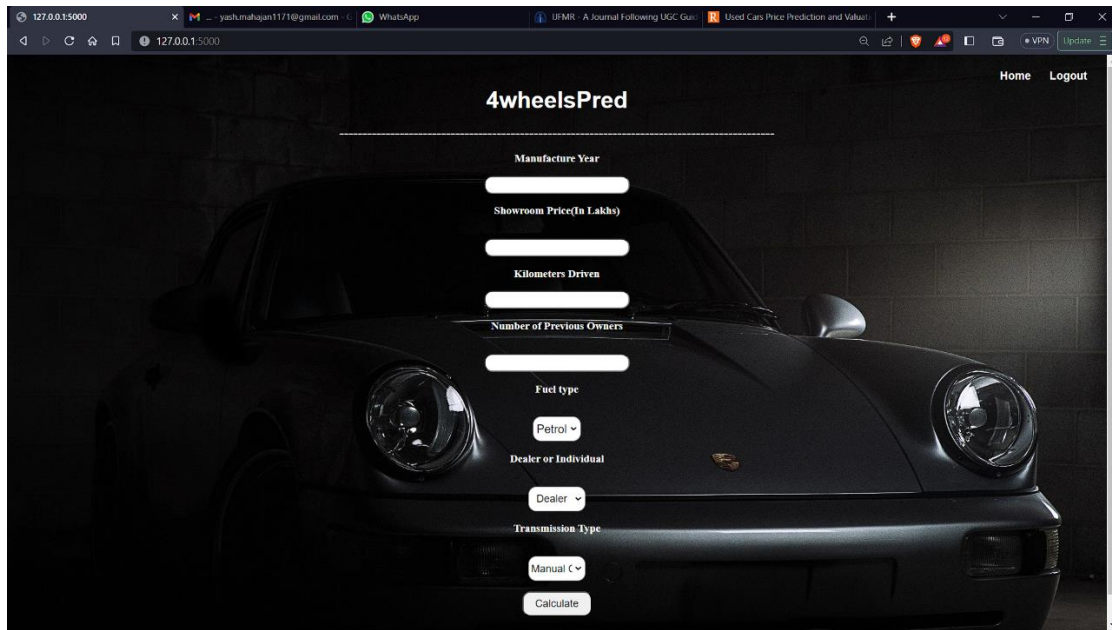
According to author S. E. Vishwapriya, Durbaka Sai Sandeep Sharma and Gandavarapu Sathya kiran [4], This research uses three machine learning techniques—Artificial Neural Network, Support Vector Machine, and Random Forest—to construct an accurate model to forecast the price worth of second-hand cars. These methods were used to many data points. This data set was obtained via a web portal, which was also utilized to forecast prices. The information needs to be gathered using a PHP web scraper. To acquire the best result from the supplied data set, numerous machine learning methods with different results had been compared. The last prediction model was added to a Java programme.

According to author Laveena D’Costa et al [5], in this paper they are applying machine learning algorithms to determine the true value of cars when selling them to the dealers. They 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.

5. Proposed System

In this, we will apply suitable/best fit algorithm to predict car price. The most necessity ingredient for prediction is brand and model, period usage of vehicle, mileage of vehicle. Different features like year of manufacture, fuel type, transmission type, dealer type, etc. will also influence the vehicle price. Here we will see algorithms like Linear Regression, Lasso Regression, Support Vector Machine (SVM), Random Forest and then choose best fit algorithm. Model is built according to best fit algorithm and then deployed





Algorithms

1. Linear Regression-

In statistics, linear regression is a linear approach for modelling the relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables). Simple linear regression is used when there is only one explanatory variable, and multiple linear regression is used when there are numerous explanatory variables. This term is distinct from multivariate linear regression, where multiple correlated dependent variables are predicted, rather than a single scalar variable. Regression coefficients as observed in the dataset are provided via linear regression.

Steps:

1. Collection of the data
2. Plotting of the data
3. Choosing the regression model
4. Determination of the regression equation
5. Calculation of slope and intercept
6. Evaluation of the model
7. Predictions Making
8. Validating the model

Equation:

The simple equation for linear regression with one independent variable is:

$$y = \beta_0 + \beta_1 * x$$

where:

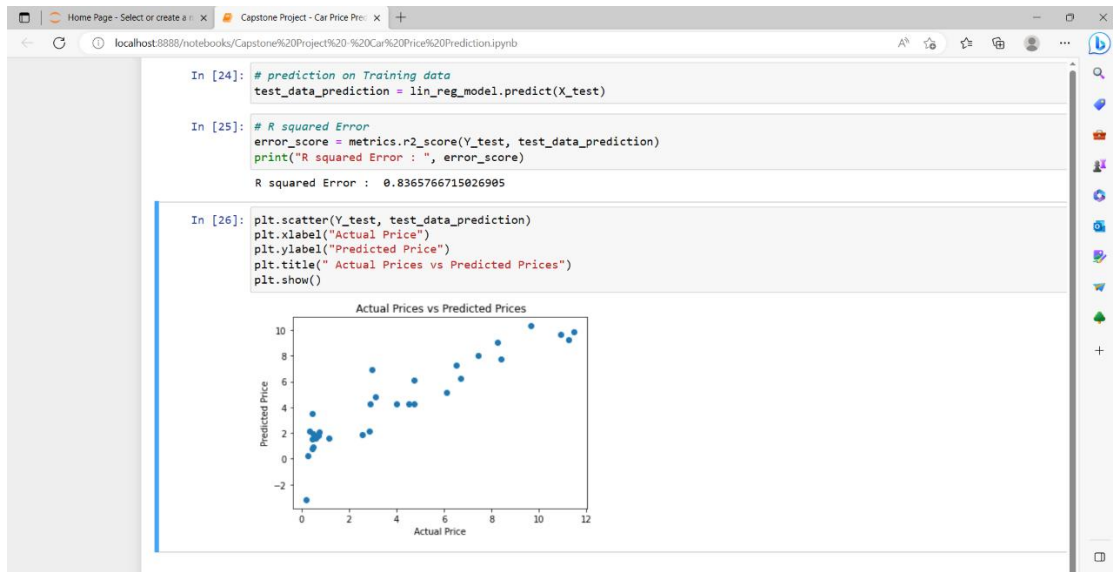
y is the response or dependent variable

x is the predictor or independent variable

β_0 is the intercept or constant term

β_1 is the coefficient or slope of x

R2 Score: **0.8365**



2. Lasso Regression-

Least Absolute Shrinkage and Selection Operator is referred to as "LASSO." It is a statistical algorithm for selecting features and regularizing data models. The lasso regression analysis method in statistics and machine learning combines variable selection and regularization to improve the predictability and understandability of the produced statistical model. Like linear regression, lasso regression employs the "shrinkage" strategy, which involves reducing the coefficients of determination until they are zero. To prevent overfitting and improve their performance on other datasets, you can reduce or regularize these coefficients using the lasso regression method.

Steps:

1. Collection of the data
2. Preprocessing of the data
3. Splitting of the data
4. Choosing the regression model
5. Choosing the regularization parameter
6. Fitting of the model
7. Evaluation of the model
8. Tuning of the regularization parameter
9. Validating the model

Equation:

The simple equation for Lasso regression with one independent variable is:

$$y = \beta_0 + \beta_1 * x$$

subject to the constraint that the sum of the absolute values of the coefficients is less than or equal to a constant value (i.e., the L1-norm of the coefficients):

$$|\beta_1| + |\beta_2| + \dots + |\beta_n| \leq t$$

where:

y is the response or dependent variable

x is the predictor or independent variable

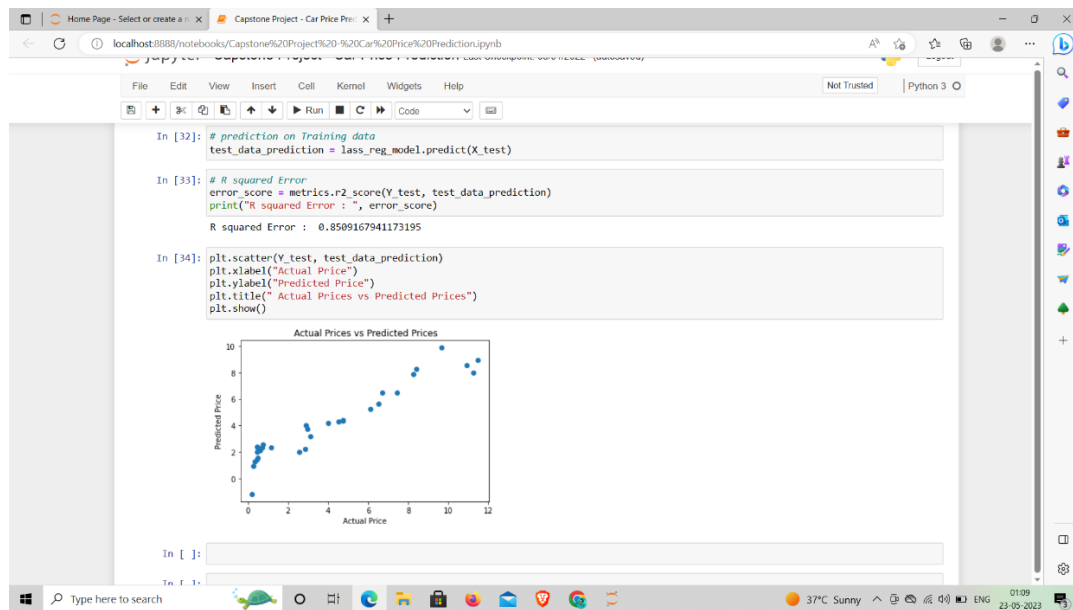
β_0 is the intercept or constant term

β_1 is the coefficient or slope of x

β_2, \dots, β_n are the coefficients of other predictor variables (if present)

t is a constant value that controls the amount of regularization applied to the model

R2 Score: **0.8509**



3. Support Vector Machine (SVM)-

SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future(hyperplane). SVM selects the extreme vectors and points that aid in the creation of the hyperplane. Support vectors, which are used to represent these extreme instances, form the basis for the SVM method. The challenges of utilising linear functions in the high-dimensional feature space can be overcome using SVM.

Steps:

1. Collection of the data
2. Preprocessing of the data
3. Splitting of the data
4. Choosing the kernel function
5. Choosing the regularization parameter
6. Fitting of the model
7. Evaluation of the model
8. Tuning of the hyperparameter
9. Validating the model

Equation:

The simple equation for a SVM model for binary classification with two predictor variables is:

$$y = \beta_0 + \beta_1x_1 + \beta_2x_2$$

where:

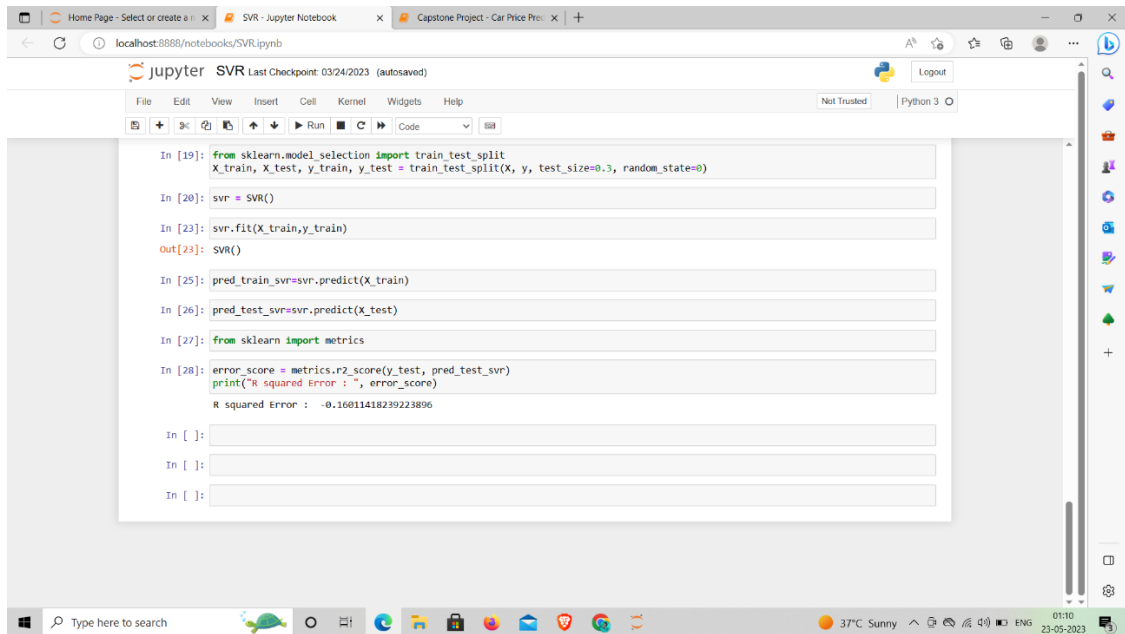
y is the predicted class label (+1 or -1)

x1 and x2 are the predictor variables (features)

β_0 is the intercept or bias term

β_1 and β_2 are the coefficients or weights of x1 and x2, respectively

R2 Score: **-0.1601**



```
In [19]: from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=0)

In [20]: svr = SVR()

In [23]: svr.fit(X_train, y_train)
Out[23]: SVR()

In [25]: pred_train_svr=svr.predict(X_train)

In [26]: pred_test_svr=svr.predict(X_test)

In [27]: from sklearn import metrics

In [28]: error_score = metrics.r2_score(y_test, pred_test_svr)
print("R squared Error : ", error_score)
R squared Error : -0.16011418239223896

In [ ]:

In [ ]:

In [ ]:
```

4. Random Forest Regression-

Random Forest, as the name implies, is a classifier that uses a number of decision trees on different subsets of the provided dataset and averages them to increase the dataset's predictive accuracy. Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. It is called a Random Forest because we use Random subsets of data and features and we end up building a Forest of decision trees (many trees). As we employ several subsets of data in each model to produce predictions, Random Forest is another well-known illustration of a bagging strategy.

Steps:

1. Collection of the data
2. Preprocessing of the data
3. Splitting of the data
4. Choosing the number of trees
5. Choosing the maximum depth of trees
6. Choosing the number of features to consider

7. Fitting of the model
8. Evaluation of the model
9. Tuning of the hyperparameter
10. Validating the model

Equation:

The equation for Random Forest regression is a bit more complex than for some of the other models, as it involves multiple decision trees working together. However, we can write the basic equation as follows:

$$y = \sum_{j=1}^J f_j(x)$$

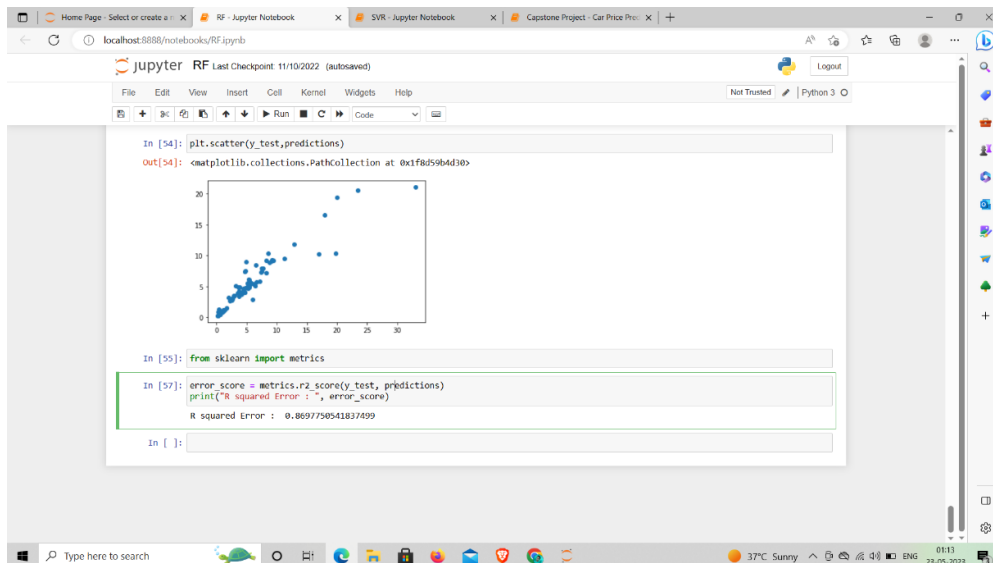
where:

y is the predicted value of the response variable

J is the total number of trees in the Random Forest model

$f_j(x)$ is the predicted value of the j^{th} decision tree for the given set of predictor variables (features) x

R2 Score: **0.8697**



System Requirements

Hardware requirements

- Operating system- Windows 7,10,11
- Processor- dual core 2.4 GHz (i5 or i7 series Intel processor or equivalent AMD)
- RAM-4GB

Software requirements

- Python
- Jupyter Notebook
- Pycharm / Spider
- Chrome
- Flask

• Xampp

6. Acknowledgement

It gives us great pleasure in presenting the preliminary project report on ‘Used Car Price Prediction Using Random Forest Algorithm. We would like to take this opportunity to thank our internal guide **Prof. D. A. Gaikwad** for giving us all the help and guidance we needed. We are grateful to them for their kind support. Their valuable suggestions were very helpful.

We are also grateful to **Prof. V.V. Kalunge**, Head of the Information Technology Department, JSPM’s Jaywantrao Sawant College of Engineering for his indispensable support, suggestions. In the end our special thanks to **Prof. V.V. Kalunge** for providing various resources such as laboratory with all needed software platforms, continuous Internet connection, for our Project.

7. Conclusion

The increased prices of new cars and the financial incapability of the customers to buy them, Used Car sales are on a global increase. Therefore, there is an urgent need for a Used Car Price Prediction system which effectively determines the worthiness of the car using a variety of features. The proposed system will help to determine the accurate price of used car price prediction. This paper compares 4 different algorithms of Machine Learning: Linear Regression, Lasso Regression, SVM and Random Forest. As, the R2 score value for Random Forest is greater than the other three algorithms so we build the model using Random Forest which predicts value for used cars.

8. References

1. Doan Van Thai, Luong Ngoc Son, Pham Vu Tien, Nguyen Nhat Anh, Nguyen Thi Ngoc Anh, “Prediction car prices using quantify qualitative data and knowledge-based system”, IEEE – 2020.
2. Praful Rane, Deep Pandya, Dhawal Kotak, “Used Car Price Prediction”, International Research Journal of Engineering and Technology, Apr 2021.
3. Anamika Das Mou, Protap Kumar Saha, Sumiya Akter Nisher, Anirban Saha, “A Comprehensive Study of Machine Learning algorithms for Predicting Car Purchase Based on Customers Demands”, IEEE –2021.
4. 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) ISSN: 2278-3075, Volume-9 Issue-8, June 2020.
5. 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), ISSN: 2277-3878, Volume-8, Issue-5S, January 2020.