International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

# Used Car Price Prediction Using One Hot Encoding

**Rakesh Sambyal[1], Rachit Kaushik[2], Tushar Tyagi[3], Rudransh Sharma[4]**

[1]Assistant Professor MIET, Department of Information Technology Meerut Institute of Engineering and Technology, Meerut, U.P.
[2,3,4]Student Department of Information Technology Meerut Institute of Engineering and Technology, Meerut, U.P.

**Abstract**
Guessing the value of old cars is a topic of intense interest since it calls for distinctive effort from a subject-matter specialist. The manufacturer in the industry determines the cost of a new automobile, plus any additional taxes that the government must pay. Customers who purchase a new automobile are therefore certain that their financial commitment will be beneficial. However, old car sales are rising internationally as a outcome of new car price increases and consumers' financial inability to purchase them. Almost the previous ten years, the number of automobiles manufactured has constantly increased; in 2022, there will be over 80 million passenger cars produced. With the use of machine learning techniques like Extra Trees Regressor, Random Forest Regressor, and Regression Trees. we'll try to expend the model that predicts the cost of a old vehicle using past customer data and a specified set of characteristics. Regression algorithms are employed because they give clients continuous results as opposed to categorical final results. As a result, it will be feasible to forecast the exact cost of an automobile rather than just its price range. The user interface, which requests input from any user and shows a car's cost in response to that input, was likewise built using React js. The main aim of this exploration is to produce machine learning models that can directly predict an old car's cost based on its parameters so that customer or user may make best decisions.

**Keywords**: Old Car Price Prediction, One Hot Encoding, Random Forest Regression, Randomized Search CV.

## 1. INTRODUCTION:

It can be challenging to tell if the advertised price for a used or old car is accurate due to the many factors that influence a used car's cost on the market. In order to make informed purchases, this research focuses on creating machine learning models that can properly guess an old car's cost on the basis its qualities. We use a variety of learning techniques on the huge data that includes the sale prices of various manufacturers and framework. The effectiveness of many machine learning techniques, such as Extra Trees Regressor, Random forest Regressor, and Regression Trees, will be compared, and the best one will be chosen. The used vehicle pricing has been guess based on a variety of variables. We utilize regression technique because, unlike classified values, the results they provide are continuous values, making it feasible to anticipate the real price. Users may submit information and see the pricing of a car based on their inputs using user interfaces built using ReactJS.

The cost of manufacturing, shipping, and GST on motor vehicles are some of the factors that affect how much new cars cost. A car's value continues to decline by 20percent20 percent every time after that. The majority of people choose to acquire used cars instead of going to the market to get new one. Prediction is nothing more than an estimate based on compliances. It makes arriving possible for the scenario with the use of visible wonders. Our monitoring are depend on the prior data, that are so abundant which drawing conclusion only from a cursory glance at the data is laborious. Homemade data explaination is a featureless endeavor. Different machine learning techniques has to be utilized to reduce this embarrassment and give time-ahead predictions.

**The following are the parameters used:**

1. Name of Car: The Car name with its brand.
2. Year: Year in which car has been purchased.
3. Selling Price: The price in which car has been sold.
4. Fuel: Type of fuel which is used in car like petrol, diesel, CNG, electric.
5. KM driven: It shows how much car has been driven.
6. Seller type: It basically shows that the car has been sold by individual owner or by the dealer.
7. Transmission: It show car type like Automatic or Manual.
8. Owner: It shows that how many times car has been sold like First Owner, Second Owner, etc.
9. Cylinder: How many cylinders present in the car.
10. Horsepower: It basically shows the power of car in Cubic centimetres.

## 2. Literature Survey:

The study is titled Predicting the cost of an Old Cars Using Machine Learning Algorithm. The implementation of machine learning algorithms to forecast the cost of secondhand vehicles in India is examined in this research. The forecasts are supported by information gathered from daily publications. Numerous algorithms, such as multivariate linear regression, KNN & decision trees, were working to create the predictions. Utilizing machine learning techniques, predicting car prices is the second paper. For a dependable and accurate forecast, a number of unique qualities are analyzed. They used three machine learning techniques (Convolutional Neural Network (CNN), Support Vector Machine (SVM), & Random Forest) to make a machine learning model for guessing the cost of old vehicles.

The last study presents a used automobile price evaluation model from BP neural networks. Here, a large data-based pricing assessment methodology is put forth. This model uses the optimized BP neural network method to get pricing data for each type of car by utilizing widely disseminated vehicle data as well as a broad number of cars transaction data. The main aim of this research is to develop a model for assessing used vehicle pricing in order to identify the cost that best matches the automobile..

Therefore, this model was chosen to differentiate a adjure policy with the following ramp improving methods and account for the dataset's huge number of features. When determining a pre-owned vehicle's price, a variety of factors are taken into account, including the moment of purchase, the mileage of the car, its exchange value, its availability, and the car's machine and seating capacities.

Other elements like if the dealer is the vehicles owner or a dealer, whether the owner is the first or second owner of a vehicle, whether the gear is manual or automatic, and whether gasoline or diesel is used as energy are also taken into account when creating the model.

The goal of this study, according to authors Enids Gegic et al., is to use a web scraping approach to extract data from an online source. To anticipate car pricing simply, these are differentiated with the use of various

machine-learning algorithms. They separated the prices into various, previously determined price categories.

The authors claim that Pattabiraman concentrates more on the interaction between the vendor and the buyer. Additional features, such as mentioned cost, mileage, year, car's type like suv, cylinder, doors, cruise, and the sound are needed to estimate the price of four-wheelers.

## 3. Methodology

**:** For each system, data collecting is the first and most important phase. We created a system for vintage cars using data from Kaggle that was obtained from vintage automobiles. Using the Beautiful Soup (BS4) package, a total of 4341 records were deleted. 2454 records are kept from the data after the null records in the data that were brought in are eliminated. We used both traditional and modern approaches, such as ensemble learning strategies.

**The system is divided into two main phases:**

1. Training phase: A model (line/curve) based on the Decision tree technique is fitted once the system has been trained using data from the data set.
2. Training phase: The input values for the system have been given to it, and its correctness has been tested. It is examined for correctness. Selected methods have been employed to complete the two distinct jobs since the technology aims to recognize and foresee the cost of an outdated car. Many algorithms were examined for accuracy before the ones that would be used later were chosen.

**Following the Machine learning algorithm which has been used to train and test the models:**

1. Linear Regression: As a result of its ease of use and proportional quick training & testing period, linear regression was selected as the first model. A value of the parameters can be guessed with the use of linear regression analysis based on another variable. Since the findings clearly demonstrated minimal variance, regularisation was not applied.
2. One Hot Encoding: One hot encoding is the crucial process of translating the variables in categorical data to machine and deep learning algorithms, which enhances a model's classification and prediction accuracy. Figures or statistics can be transformed with the use one hot encoding that means of kniwing the best prediction and using the figures for an algorithm. Using one-hot, new columns have been created according their category and for each category they have some binary value which are 0 or 1.

| | selling_price | km_driven | no_year | fuel_Diesel | fuel_Electric | fuel_LPG | fuel_Petrol | seller_type_Individual | seller_type_Trustmark Dealer | transmission_Manual | owner_Fourth & Above Owner | owner_Second Owner |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 60000 | 70000 | 15 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 |
| 1 | 135000 | 50000 | 15 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 |
| 2 | 600000 | 100000 | 10 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 |
| 3 | 250000 | 46000 | 5 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 |
| 4 | 450000 | 141000 | 8 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 1 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 4335 | 409999 | 80000 | 8 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 1 |
| 4336 | 409999 | 80000 | 8 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 1 |
| 4337 | 110000 | 83000 | 13 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 1 |
| 4338 | 865000 | 90000 | 6 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 |
| 4339 | 225000 | 40000 | 6 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 |

4340 rows × 14 columns

**Figure 1: One Hot Encoding**

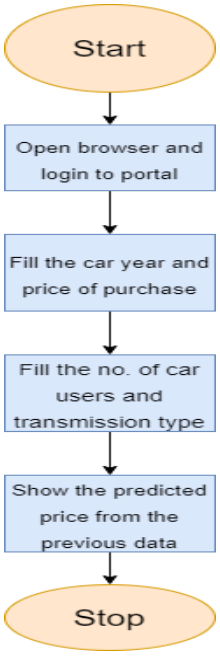
3. Random Forest: A group learning-based regression model is Random Forest. The best way to create the group model, which accordingly submit the forecast, it engages the decision tree model, precisely as the name implies, several decision trees.

Every tree is built with the use of dissimilar sample of rows in random forest technique, and a dissimilar sample of attributes is select for break each node. Each tree gives the different forecast on its own. The sole result is then turn out by calculating the forecasts.

**Objective**

- To develop a model that, depending on input from the user, predicts the price of a used automobile with accuracy.
- To attain best accuracy.
- To help user to predict the best price for its vehicle.
- To help user or buyer to buy the vehicle according to their need and according to their budget.
- Using ReactJS, create an intuitive Web Interface that solicits information from the buyer and guess the price.

## 3. Proposed System:




**Proposed System Flowchart**

1. The method begins with gathering the dataset, as seen in the above graphic. The following stage consists of using data filtering, which encompasses data cleansing, data lightning, and data modification, to clean or filter the data. As a result, we will employ machine learning techniques to estimate the price. Extra Trees Regressor and Random Forest Regressor are used in the algorithms. The best model that has produced the most accurate price predictions will be chosen. After choosing the best model, the user is provided the estimated price based on their inputs. Through a website, users may provide input for a machine-learning model that predicts the price of used cars.
2. Random Forest Regressor: Averaging is used by a random forest meta-estimator to enhance predicted

accuracy and reduce overfitting. This is done by fitting different categorization decision trees to different dataset subsamples. More accurate than other algorithms is the random forest. When a significant fraction of the data is missing, it nevertheless retains accuracy thanks to an efficient estimation algorithm.

## 4. Results

The final result of the old car price prediction will be the exact or nearby prediction of the price of the old car. In which, it basically shows three values of the car on the basis of parameters on which the model has predicted the price of the car which will be the best price, average price, and worst price of the old car.
The results of the car cost prediction are based on the accuracy of the model by using the various machine learning algorithms and the parameters of the data sample on which the model has been trained and tested and produce the final output.

**Heat Map**:

It is the corelation matrix between the data set and the finalize data set or parameters which is used to see the relation or symmetry in the parameters and it helps to analyse the data.

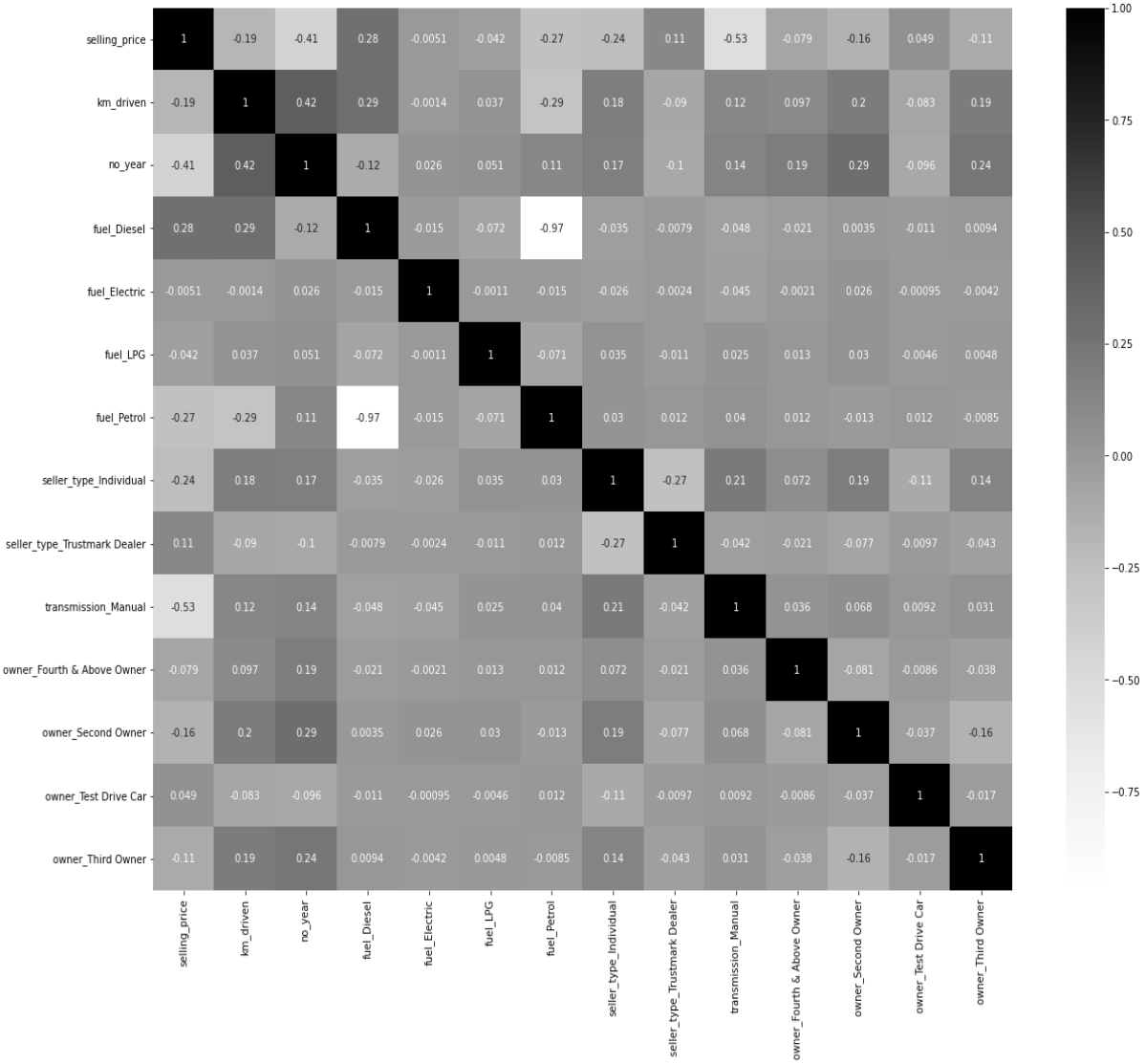



**Figure 2: Heat Map**

**Histogram of parameter:**

These are the graphs in which the data of each and every parameter has been shown. Each of these factors has a specific value, and they are used to estimate the price of an old automobile. For each parameter like for fuel, there are different histograms in which each kind of fuel has been shown fuel has different types petrol, diesel, CNG, and electric, and for seller type, there are two types of the seller which are Individual or Trustmark Dealer.

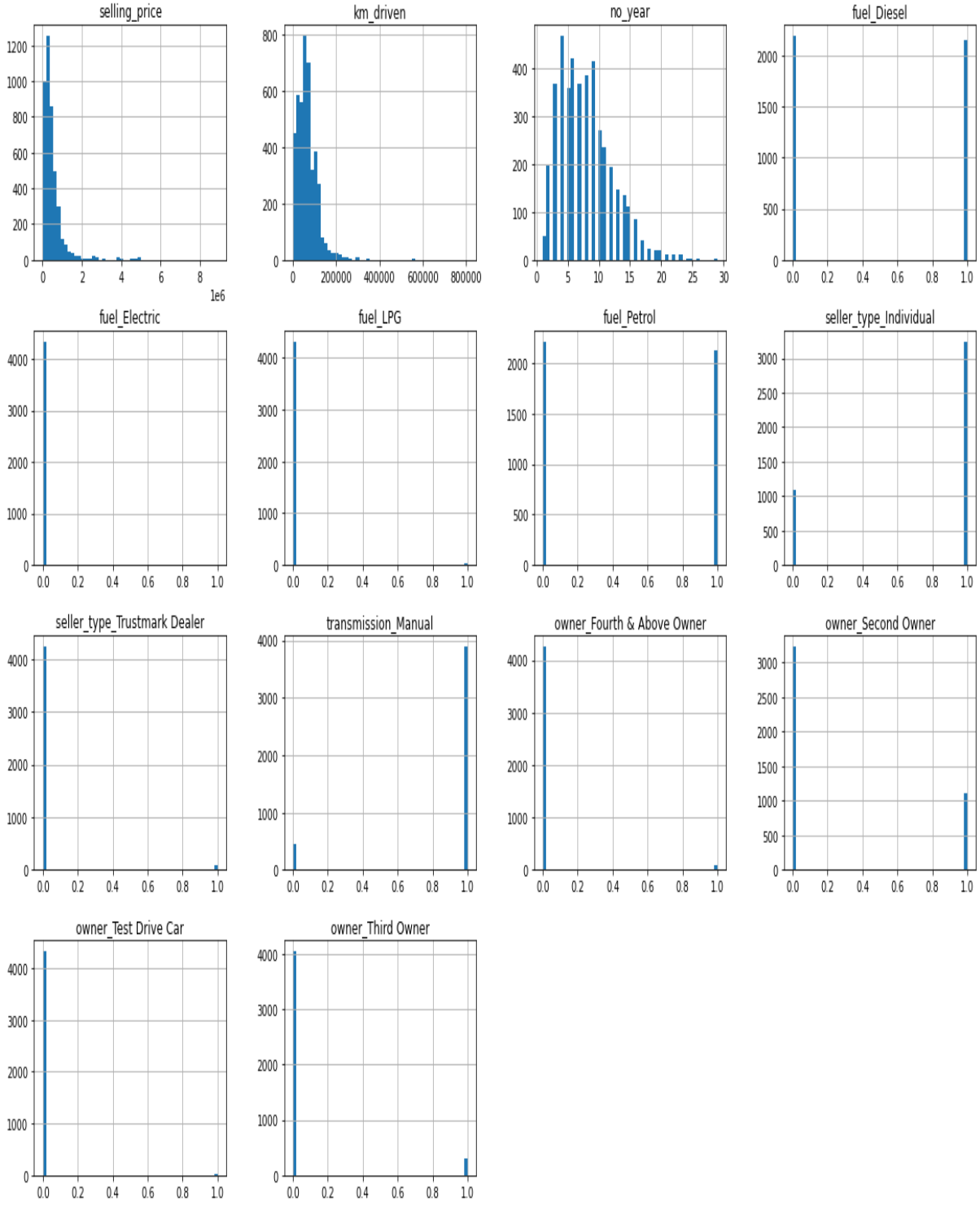



**Figure 3: Histogram of Parameters**

**Feature Importance Graph:**

These variables are used to forecast the price of secondhand cars and these parameters are the most important feature of car. Without there features or parameters the accuracy of the price prediction of car has been decrease.

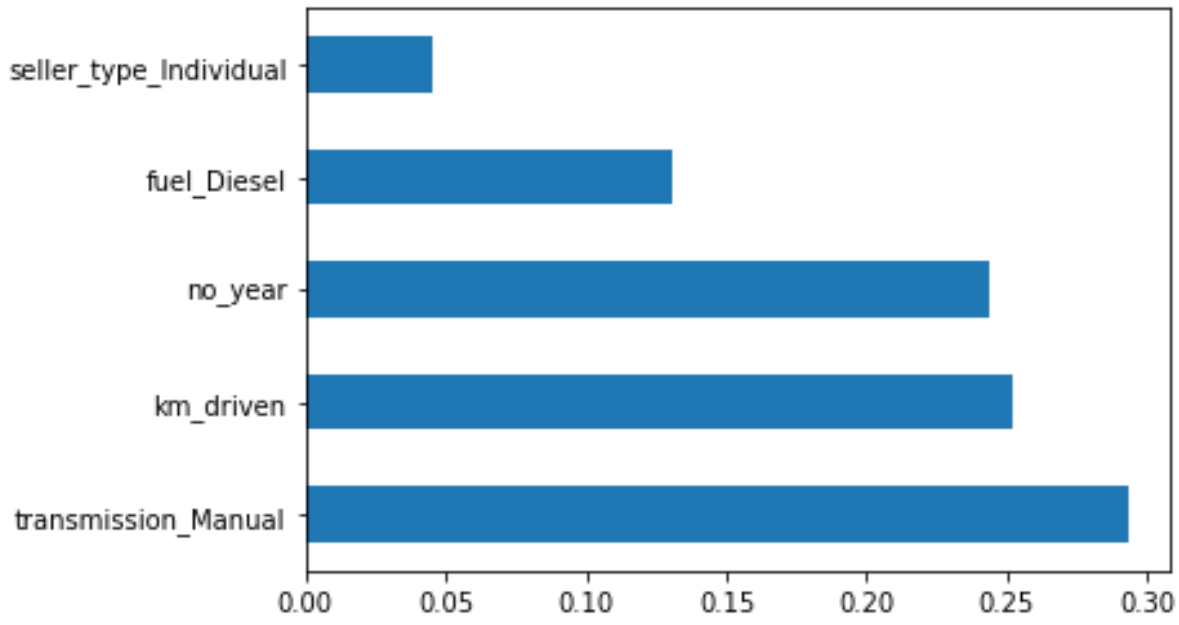



**Figure 4: Feature Importance Graph**

**Heatmap for Important Feature:**

**It is the corelation matrix of importance feature which are shown below:**

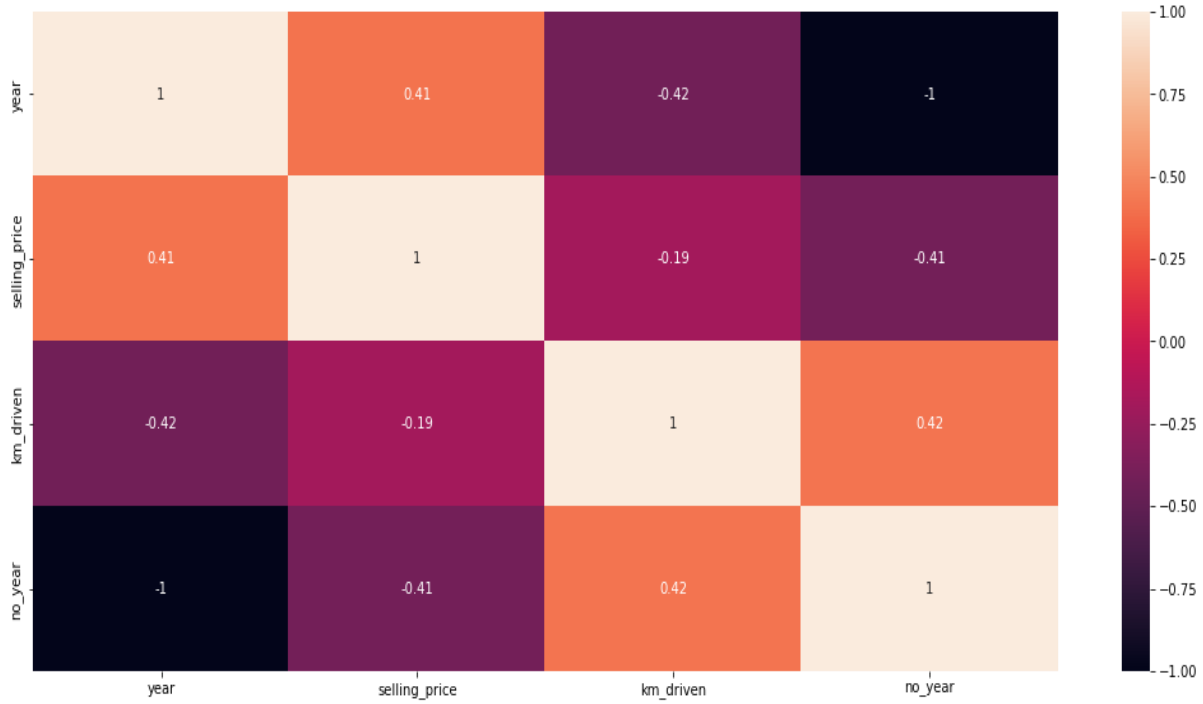



**Figure 5: HeatMap of Feature Importance Graph**

**Distpolt of Price Prediction:**

The distplot of price prediction are highly appear in between -1 to 1 which increase the accuracy of the model.


**Figure 6: Distplot of Pridiction**

**Accuracy Between testing and prediction:**

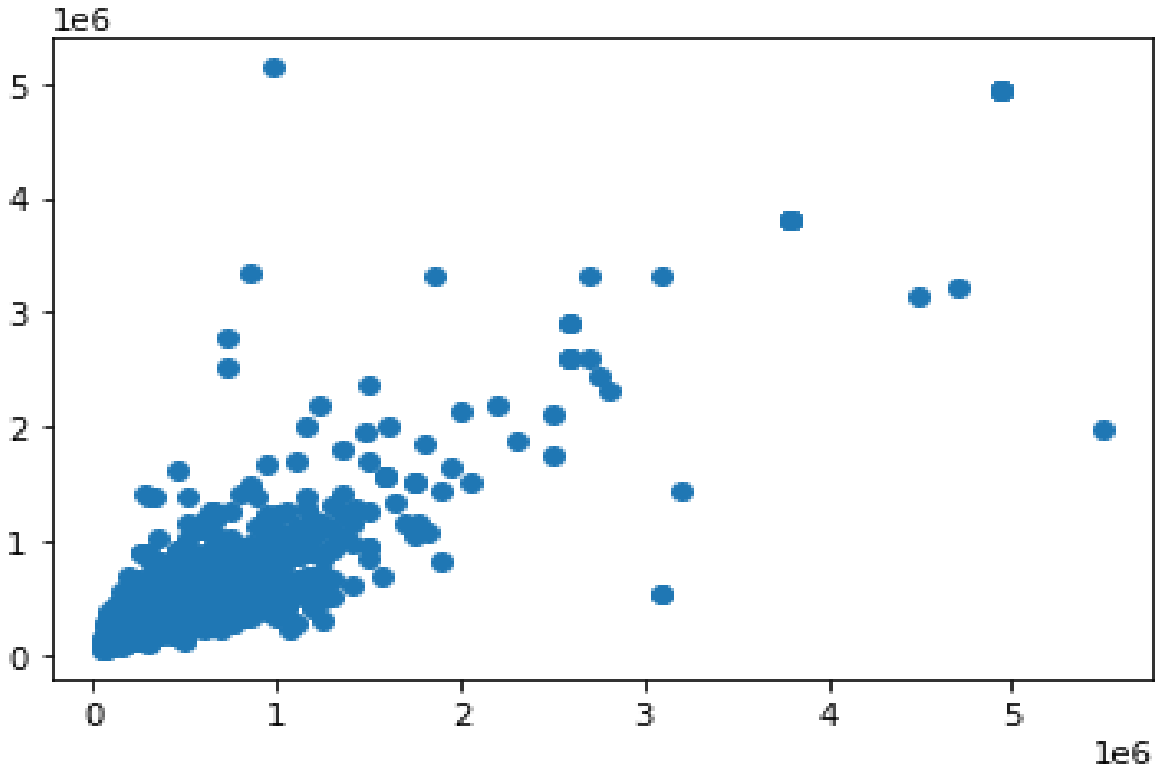



**Figure 7: Accuracy between Test and prediction using Random Forest Regressor.**

## 5. Conclusion and Future Scope

It may one day be able to integrate our machine learning model with several web pages that offer real-time data for cost prediction. Additionally, we could include a tonne more automobile price-related characteristics, which would boost the machine learning model's accuracy. As user interfaces for engaging with the user, we may create an iOS or an Android application. In order to carefully build deep learning network topologies, variable learning rates and training on data clusters rather than the entire dataset can both increase performance.

Due to rising new automobile prices and consumers' limited financial resources, sales of used cars are rising globally. For the purpose of predicting the price of old cars, a system that properly evaluates the worth of the vehicle using a variety of factors is therefore critically needed. The suggested approach will make it possible to anticipate the price of used cars with greater accuracy. Instead of training on the entire dataset, the notion of carefully designing machine learning network topologies, using adaptive learning rates, and improving performance is lacking. In order to adjust for overfitting in Random Forest, various feature selections and tree counts will be examined to look for performance improvements.

## References

1. Enis gegic, Becir Isakovic, Dino Keco, Zerina Masetic, Jasmin Kevric, “Car Price Prediction Using Machine Learning Technique”, (TEM Journal 2019)
2. Nitis Monburinon, Prajak Chertchom, Thongchai Kaewkiriya, Suwat Rungpheung, Sabir Buya, Pitchayakit Boonpou, “Prediction of Prices for Old Car by using Regression Models” (ICBIR 2018)
3. Used cars database. (n.d.) Retrieved from: https://www.kaggle.com/orgesleka/used-carsdatabase. [accessed: June 04, 2018].
4. 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.
5. Sameerchand Pudaruth, “Predicting the Price of Used Cars using Machine Learning Techniques”;(IJICT 2014)
6. Praful Rane1, Deep Pandya2, Dhawal Kotak3 “Used car price prediction “International Research Journal of Engineering and Technology (IRJET). (2021)
7. Abhishek pandey1, Vanshika Rastogi2, Shanika Singh “Car selling price prediction using random forest Machine learning Algorithm, MAY 2021.
8. Dr. M. J. Garbade – “Clearing the Confusion:AI vs Machine Learnings Deep LearningDifferences” Available: https://towardsdatascience.com/clearing-the-confusion-ai-vs-machine- learning-vs deep-learning differences-fce69b21d5eb
9. Frost,J. 2013. Multiple Regression Analysis: Use Adjusted R-squared and Predicted R-squared to Include the Correct Number of Variables. Available online from:http://blog.minitab.com/blog/adventures-instatistics/multiple-regession-analysis-use-adjusted-rsquared-and-predicted-r-squared-to-include-the-correctnumber-of-variables (Last accessed: 29-11-2016).
10. 3.2.4.3.1. sklearn.ensemble.RandomForestClassifier — scikit-learn 0.19.2 documentation. (n.d.). Retrieved from: http://scikitlearn.org/stable/modules/generated/sklearn.ensemble .RandomForestClassifier.html [accessed: August 30, 2018].
11. Aizerman, M. A. (1964). Theoretical foundations of the potential function method in pattern

recognition learning. Automation and remote control, 25, 821- 837.
12. 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.
13. **Kaggle Datasets** - For real-world used car datasets.
14. **Scikit-learn Documentation** - For more on RandomForestRegressor, GridSearchCV, etc.