e-ISSN: 2582-5208

International Research Journal of Modernization in Engineering Technology and Science

(Peer-Reviewed, Open Access, Fully Refereed International Journal)

Volume:04/Issue:02/February-2022 Impact Factor- 6.752 www.irjmets.com

CAR PRICE PREDICTION USING MACHINE LEARNING TECHNIQUES

Abishek R*1

*1Student, Big Data Analytics, SRM University, Chennai, Tamil Nadu, India.

ABSTRACT

As a result of incredible technological advancements and research of new technical expertise and huge economical growth of our country, people started to buy cars more than other vehicles. Therefore, there arises an enormous demand for cars, as the demand increases for new cars the used car market also booms alike. But the used car market is highly manipulated by few numbers of people who govern the rates of the used cars and also online selling websites designate the values for the used cars. This paper tries to study and investigate the trends in used car prices and predicts the price of used cars with the help of supervised machine learning algorithms. And to suggest which machine learning algorithm performs well among the selected methods for predicting the cars price. There has been related work done with machine learning algorithms like linear regression, multiple regression, random forest and so on. We wanted to study which algorithm predicts the car price more reliably and accurately. So that this solution will be helpful for first time used car buyers and also for sellers for determining the selling cost of the car.

Keywords: Supervised Machine Learning Algorithms, Linear Regression, Multiple Regression, Random Forest.

I. INTRODUCTION

In India the automobile market is a biggest business for international and Indian automobile companies. As the boom and demand for automobiles increase there is also a big market opening for used cars. The used car market is being manipulated and controlled by some of the online advertisement websites like olx and quickr, but customers who want to buy a used car is easily being manipulated and cheated to a higher price which the car isn't worth buying for. I would like to propose a solution for this problem by using the help of artificial intelligence and machine learning by using some supervised learning machine learning techniques and algorithms to predict the used car prices based on some parameters[2]. And I want to investigate and compare the accuracy which different algorithms produce on testing and predicting with the used car data. During 2019-20 the entire automobile production in India was 26,353,293, But in 2020-21 the automobile production in India was 22,652,108. We can see that there is a huge decline in automobile industry, people are preferring more on used and second-hand vehicles than new vehicles. Therefore the system of used cars must be standardized and a clear pricing system needs to be implemented. This paper suggests few machine techniques which can be used to predict the prices of used cars with historical used car prices data and considering a mean value from the list of prices for a specific car and assigning it as the predicted price for the given features and parameters[3]. There has been many related work done regarding this topic and field but only very few or one or two authors have done for Indian dataset, Thus I wanted to find a solution for this problem and find prediction method to give the prices for used cars in a correct method. The data for this car price prediction experiment is taken from various sources like Kaggle, web-scrapping and open source data websites which provide free data. A car price prediction has been a high-interest research area, as it requires noticeable effort and knowledge of the field expert. Considerable number of distinct attributes are examined for the reliable and accurate prediction. As the demand for cars increase the demand for second hand and used cars also increases so due to this high demand, we need to build a AI solution for solving this demand in a customer friendly way. The customers are getting cheated and tricked for a higher price for a less worth used car if the customer wants to buy it from a dealer who sells used cars. The dealer tries to sell a damaged or repaired car for high price to customers who don't know much about buying cars and stuff. The customer who doesn't know about the technical specifications and other prices of spare parts and how to deduct the price will easily be cheated with high price. I wanted to solve this type of problem where the customer has to know exact price the car is worth for. Using machine learning it is possible to predict the correct and worthy price for a given used car based on previous data from various sellers and buyers. This can be done by training the model using used cars dataset which has several features and parameters such as year of manufacturing, model year, number of cylinders, number of kms/miles driven, diesel or petrol, automatic or manual or other type of transmission, the gearing system of the cars, the number of owners of the car etc., like this there are many features from which the cars price can be predicted. And also we can add if there is any damage or is it flood affected or accidental damaged car these factors can also be considered for predicting the correct and exact price of the car.

II. METHODOLOGY

I have selected the required used car prices dataset with needed features and parameters from Kaggle. Kaggle is an open source Machine learning and data science platform which offers data and notebooks for data scientists and data analysts. The required data is cleaned and pre-processed used machine learning techniques before applying any algorithm for predicting the price. Then after pre-processing and cleaning the data first we need apply train test split to keep the data into two parts for training and validation using train and test data respectively. Then we must apply a simple linear regression model and predict the output and test its test and train accuracy with the help of roc auc score then again, we need to train and test it with multiple linear regression model and validate its accuracies. Then we need to use clustering methods and logistic regression methods and knn methods for predicting the output of car price. Also, we can use random forests and decision tree algorithms. At last, we need to compare all the accuracies of all the machine learning algorithms and choose the best algorithms for the prediction.

III. MODELING AND ANALYSIS

Between the following examples, we do some linear regression.

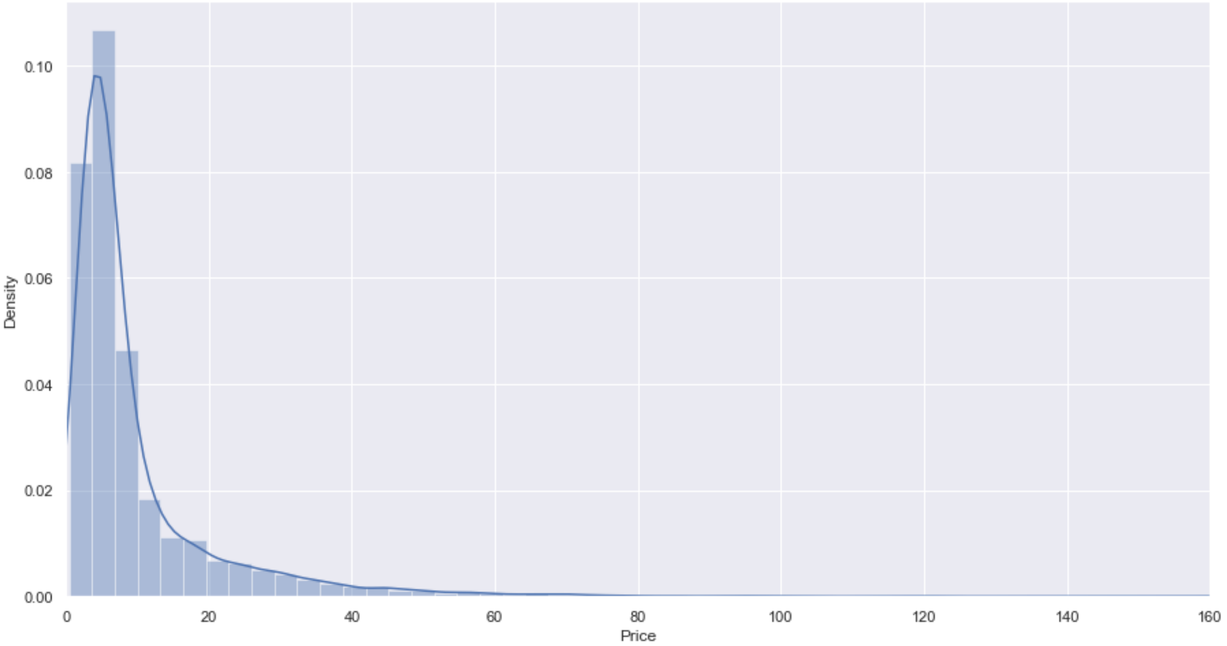


Figure 1: Comparisons of price and density

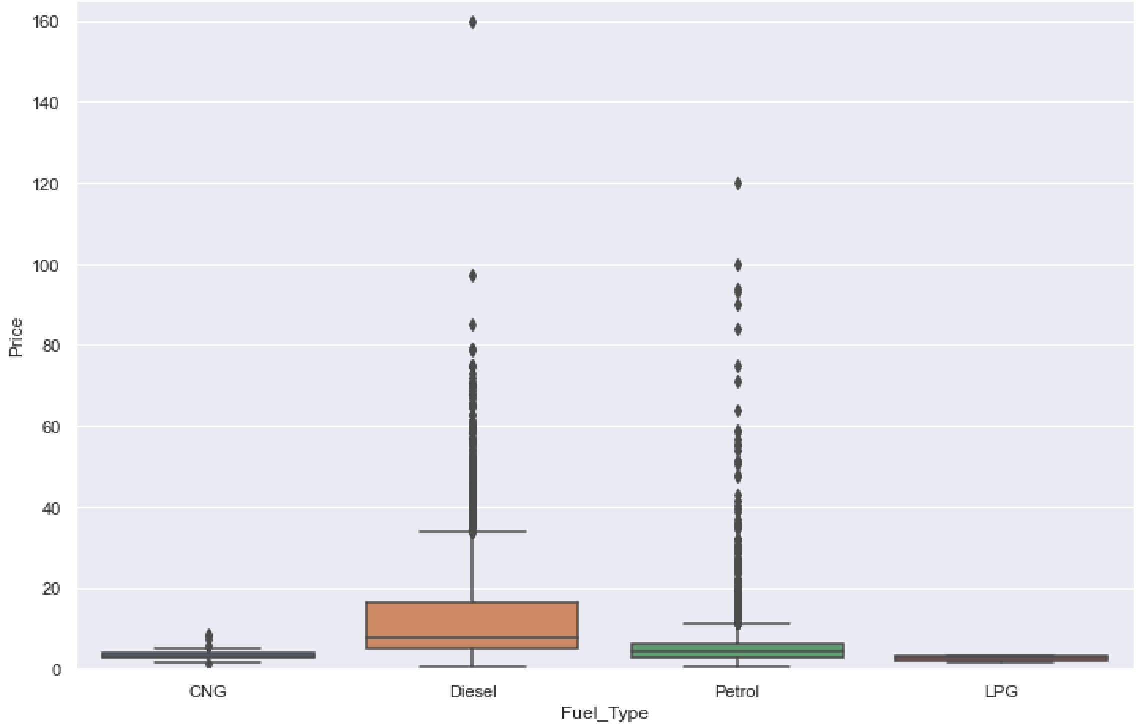


Figure 2: Box plot between various fuel types of cars.

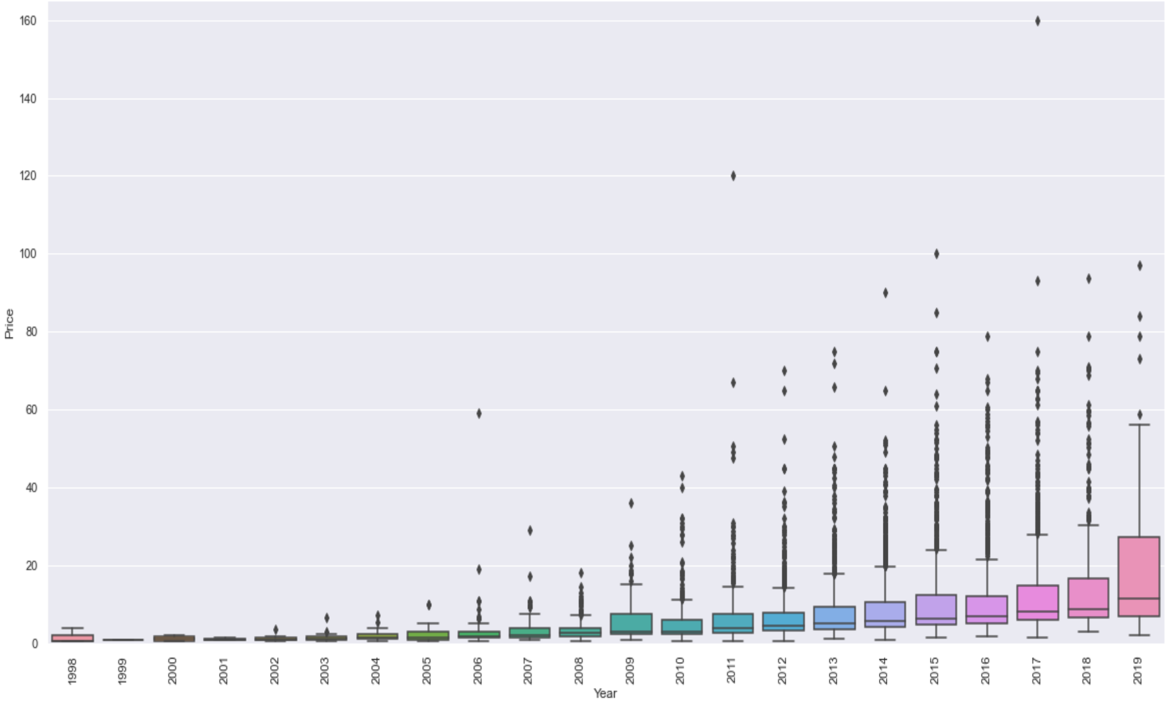
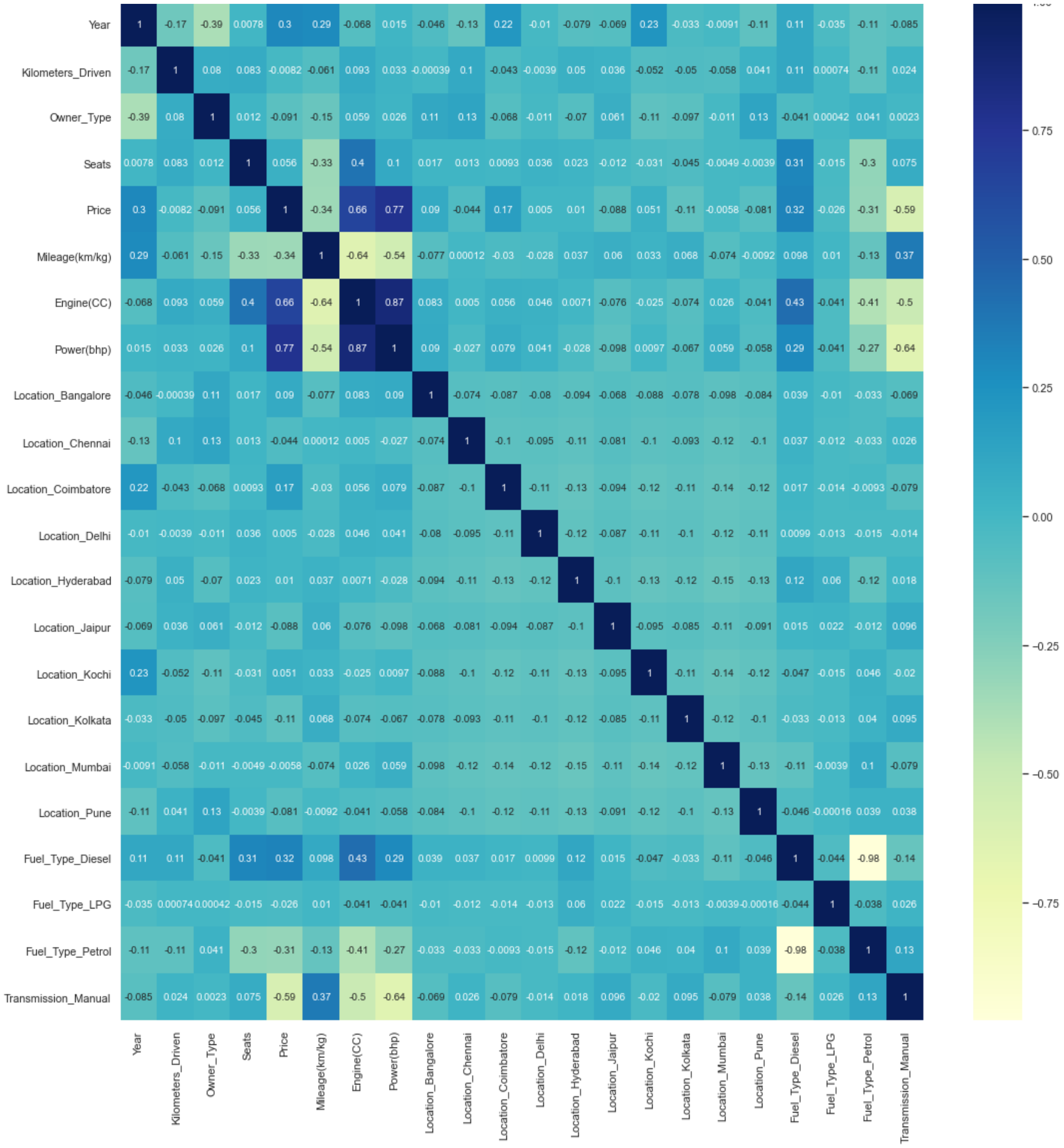


Figure 3: Box plot between various years of cars.

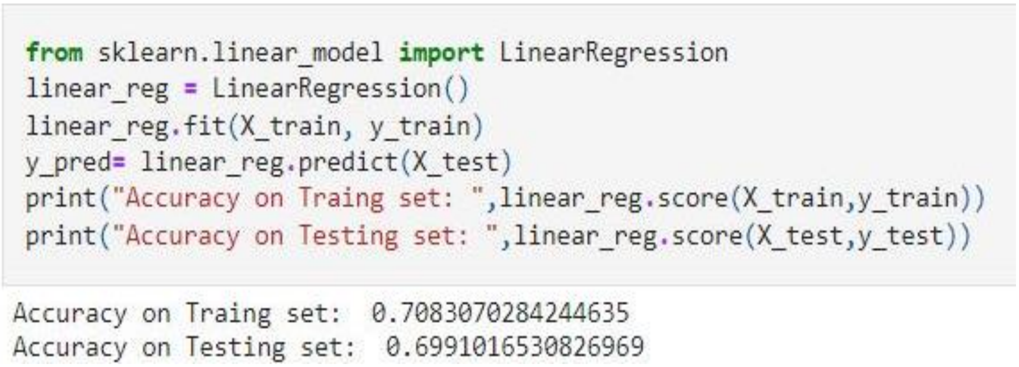
IV. RESULTS AND DISCUSSION

From the pair plot we can understand that the most affecting features are price, engine, power, mileage, fuel type, transmission types. So we need to build the model based on those features. This project mainly aims to develop a solution for car price prediction for customers and also sellers who sell cars online and as well as offline, The used car prices are predicted using kms, transmission, owner type, mileage, cc of the engine and various features. These prices are accurate and are trust worthy for the customers, further this car be developed with hyperparamter tuning and by using neural networks and other deep learning algorithms.



	Year	Kilometers_Driven	Owner_Type	Seats	Price	Mileage(km/kg)	Engine(CC)	Power(bhp)	Location_Bangalore	Location_Chennai	Location_Coimbatore	Location_Delhi	Location_Hyderabad	Location_Jaipur	Location_Kochi	Location_Kolkata	Location_Mumbai	Location_Pune	Fuel_Type_Diesel	Fuel_Type_LPG	Fuel_Type_Petrol	Transmission_Manual
Year	1	-0.17	-0.39	0.0078	0.3	0.29	-0.068	0.015	-0.046	-0.13	0.22	-0.01	-0.079	-0.069	0.23	-0.033	-0.0091	-0.11	0.11	-0.035	-0.11	-0.085
Kilometers_Driven	-0.17	1	0.08	0.083	-0.0082	-0.061	0.093	0.033	-0.00039	0.1	-0.043	-0.0039	0.05	0.036	-0.052	-0.05	-0.058	0.041	0.11	0.00074	-0.11	0.024
Owner_Type	-0.39	0.08	1	0.012	-0.091	-0.15	0.059	0.026	0.11	0.13	-0.068	-0.011	-0.07	0.061	-0.11	-0.097	-0.011	0.13	-0.041	0.00042	0.041	0.0023
Seats	0.0078	0.083	0.012	1	0.056	-0.33	0.4	0.1	0.017	0.013	0.0093	0.036	0.023	-0.012	-0.031	-0.045	-0.0049	-0.0039	0.31	-0.015	-0.3	0.075
Price	0.3	-0.0082	-0.091	0.056	1	-0.34	0.66	0.77	0.09	-0.044	0.17	0.005	0.01	-0.088	0.051	-0.11	-0.0058	-0.081	0.32	-0.026	-0.31	-0.59
Mileage(km/kg)	0.29	-0.061	-0.15	-0.33	-0.34	1	-0.64	-0.54	-0.077	0.00012	-0.03	-0.028	0.037	0.06	0.033	0.068	-0.074	-0.0092	0.098	0.01	-0.13	0.37
Engine(CC)	-0.068	0.093	0.059	0.4	0.66	-0.64	1	0.87	0.083	0.005	0.056	0.046	0.0071	-0.076	-0.025	-0.074	0.026	-0.041	0.43	-0.041	-0.41	-0.5
Power(bhp)	0.015	0.033	0.026	0.1	0.77	-0.54	0.87	1	0.09	-0.027	0.079	0.041	-0.028	-0.098	0.0097	-0.067	0.059	-0.058	0.29	-0.041	-0.27	-0.64
Location_Bangalore	-0.046	-0.00039	0.11	0.017	0.09	-0.077	0.083	0.09	1	-0.074	-0.087	-0.08	-0.094	-0.068	-0.088	-0.078	-0.098	-0.084	0.039	-0.01	-0.033	-0.069
Location_Chennai	-0.13	0.1	0.13	0.013	-0.044	0.00012	0.005	-0.027	-0.074	1	-0.1	-0.095	-0.11	-0.081	-0.1	-0.093	-0.12	-0.1	0.037	-0.012	-0.033	0.026
Location_Coimbatore	0.22	-0.043	-0.068	0.0093	0.17	-0.03	0.056	0.079	-0.087	-0.1	1	-0.11	-0.13	-0.094	-0.12	-0.11	-0.14	-0.12	0.017	-0.014	-0.0093	-0.079
Location_Delhi	-0.01	-0.0039	-0.011	0.036	0.005	-0.028	0.046	0.041	-0.08	-0.095	-0.11	1	-0.12	-0.087	-0.11	-0.1	-0.12	-0.11	0.0099	-0.013	-0.015	-0.014
Location_Hyderabad	-0.079	0.05	-0.07	0.023	0.01	0.037	0.0071	-0.028	-0.094	-0.11	-0.13	-0.12	1	-0.1	-0.13	-0.12	-0.15	-0.13	0.12	0.06	-0.12	0.018
Location_Jaipur	-0.069	0.036	0.061	-0.012	-0.088	0.06	-0.076	-0.098	-0.068	-0.081	-0.094	-0.087	-0.1	1	-0.095	-0.085	-0.11	-0.091	0.015	0.022	-0.012	0.096
Location_Kochi	0.23	-0.052	-0.11	-0.031	0.051	0.033	-0.025	0.0097	-0.088	-0.1	-0.12	-0.11	-0.13	-0.095	1	-0.11	-0.14	-0.12	-0.047	-0.015	0.046	-0.02
Location_Kolkata	-0.033	-0.05	-0.097	-0.045	-0.11	0.068	-0.074	-0.067	-0.078	-0.093	-0.11	-0.1	-0.12	-0.085	-0.11	1	-0.12	-0.1	-0.033	-0.013	0.04	0.095
Location_Mumbai	-0.0091	-0.058	-0.011	-0.0049	-0.0058	-0.074	0.026	0.059	-0.098	-0.12	-0.14	-0.12	-0.15	-0.11	-0.14	-0.12	1	-0.13	-0.11	-0.0039	0.1	-0.079
Location_Pune	-0.11	0.041	0.13	-0.0039	-0.081	-0.0092	-0.041	-0.058	-0.084	-0.1	-0.12	-0.11	-0.13	-0.091	-0.12	-0.1	-0.13	1	-0.046	-0.00016	0.039	0.038
Fuel_Type_Diesel	0.11	0.11	-0.041	0.31	0.32	0.098	0.43	0.29	0.039	0.037	0.017	0.0099	0.12	0.015	-0.047	-0.033	-0.11	-0.046	1	-0.044	-0.98	-0.14
Fuel_Type_LPG	-0.035	0.00074	0.00042	-0.015	-0.026	0.01	-0.041	-0.041	-0.01	-0.012	-0.014	-0.013	0.06	0.022	-0.015	-0.013	-0.0039	-0.00016	-0.044	1	-0.038	0.026
Fuel_Type_Petrol	-0.11	-0.11	0.041	-0.3	-0.31	-0.13	-0.41	-0.27	-0.033	-0.033	-0.0093	-0.015	-0.12	-0.012	0.046	0.04	0.1	0.039	-0.98	-0.038	1	0.13
Transmission_Manual	-0.085	0.024	0.0023	0.075	-0.59	0.37	-0.5	-0.64	-0.069	0.026	-0.079	-0.014	0.018	0.096	-0.02	0.095	-0.079	0.038	-0.14	0.026	0.13	1

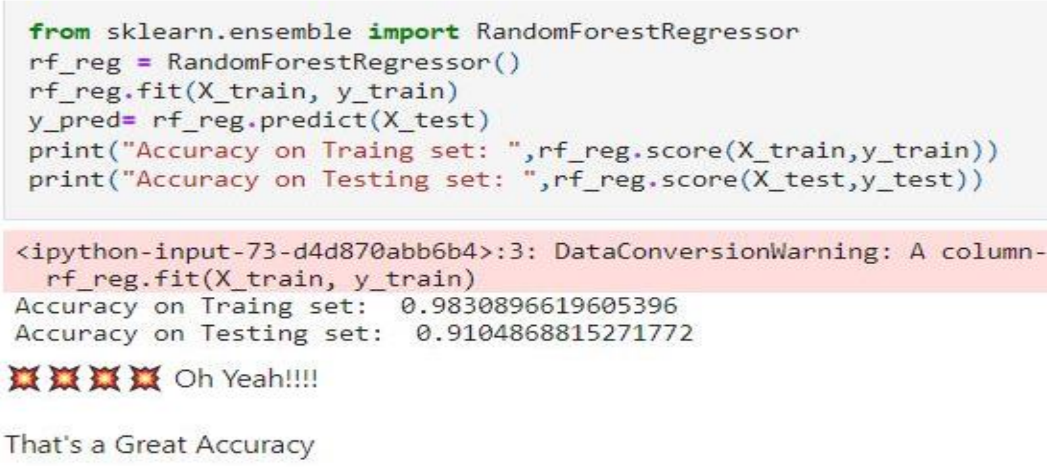
Figure 4: Pair plot between all the features.



```
from sklearn.linear_model import LinearRegression
linear_reg = LinearRegression()
linear_reg.fit(X_train, y_train)
y_pred= linear_reg.predict(X_test)
print("Accuracy on Traing set: ",linear_reg.score(X_train,y_train))
print("Accuracy on Testing set: ",linear_reg.score(X_test,y_test))
```

```
Accuracy on Traing set:  0.7083070284244635
Accuracy on Testing set:  0.6991016530826969
```

Figure 5: Accuracy of linear regression model



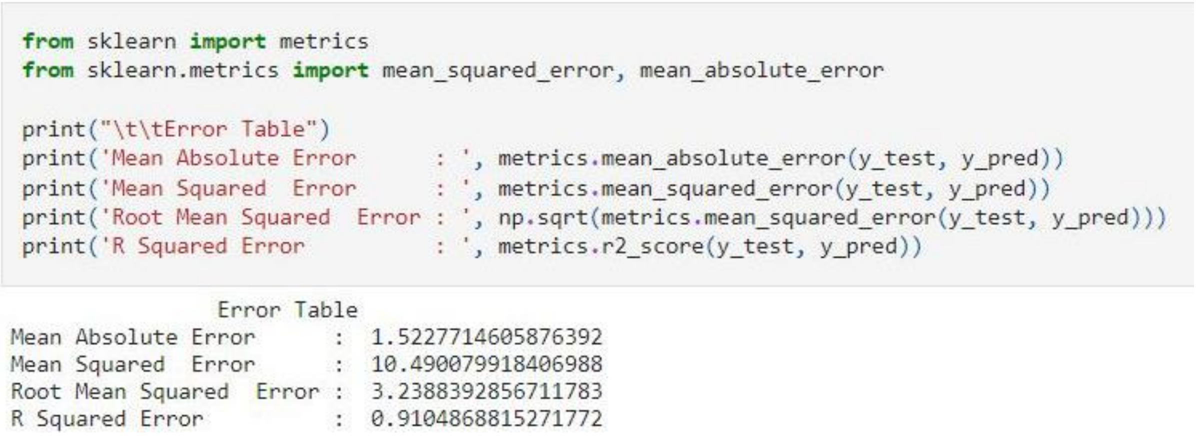
```
from sklearn.ensemble import RandomForestRegressor
rf_reg = RandomForestRegressor()
rf_reg.fit(X_train, y_train)
y_pred= rf_reg.predict(X_test)
print("Accuracy on Traing set: ",rf_reg.score(X_train,y_train))
print("Accuracy on Testing set: ",rf_reg.score(X_test,y_test))
```

```
<ipython-input-73-d4d870abb6b4>:3: DataConversionWarning: A column-
  rf_reg.fit(X_train, y_train)
Accuracy on Traing set:  0.9830896619605396
Accuracy on Testing set:  0.9104868815271772
```

💥💥💥💥 Oh Yeah!!!!

That's a Great Accuracy

Figure 6: Accuracy of random forest



```
from sklearn import metrics
from sklearn.metrics import mean_squared_error, mean_absolute_error

print("\t\tError Table")
print('Mean Absolute Error      : ', metrics.mean_absolute_error(y_test, y_pred))
print('Mean Squared  Error      : ', metrics.mean_squared_error(y_test, y_pred))
print('Root Mean Squared  Error : ', np.sqrt(metrics.mean_squared_error(y_test, y_pred)))
print('R Squared Error          : ', metrics.r2_score(y_test, y_pred))
```

```
                Error Table
Mean Absolute Error       :  1.5227714605876392
Mean Squared  Error       :  10.490079918406988
Root Mean Squared  Error  :  3.2388392856711783
R Squared Error           :  0.9104868815271772
```

Figure 7: Accuracy and errors of random forest model

When comparing the accuracies and error scores of linear regression model and random forest we can infer than the random forest model performs far better with high accuracy score of 91%. Therefore it is better to use random forest model for car price prediction project.

V. CONCLUSION

I have selected the required used car prices dataset with needed features and parameters from Kaggle. Kaggle is an opensource Machine learning and data science platform which offers data and notebooks for data scientists and data analysts. The required data is cleaned and pre-processed used machine learning techniques before applying any algorithm for predicting the price. Then after pre-processing and cleaning the data first we need apply train test split to keep the data into two parts for training and validation using train and test data respectively. Then we must apply a simple linear regression model and predict the output and test its test and train accuracy with the help of roc auc score then again, we need to train and test it with multiple linear regression model and validate its accuracies. Then we need to use clustering methods and logistic regression methods and knn methods for predicting the output of car price. Also, we can use random forests and decision tree algorithms. At last, we need to compare all the accuracies of all the machine learning algorithms and choose the best algorithms for the prediction.

VI. REFERENCES

[1] NATIONAL TRANSPORT AUTHORITY. 2014. Available from: http://nta.gov.mu/English/Statistics/Pages/Archive.aspx [Accessed 15 January 2014]

[2] MOTORS MEGA. 2014. Available from: http://motors.mega.mu/news/2013/12/17/auto-market8-decrease-sales-newcars/ [Accessed 17 January 2014].

[3] LISTIANI, M., 2009. Support Vector Regression Analysis for Price Prediction in a Car Leasing

Application. Thesis (MSc). Hamburg University of Technology.

[4] Oprea, C, Making the decision on buying second-hand car market using data mining techniques (Special, 2010), pp.17-26.

[5] C Ozgur, Z Hughes, G Rogers and S Parveen, Multiple Linear Regression Applications Automobile Pricing (International Journal of Mathematics and Statistics Invention, 2016), pp.01-10

[6] Lessmann, Stefan, M. Listiani, and S. Voß, Decision support in car leasing: A forecasting model for residual value estimation (2010).

[7] G.Chandrashekar and F. Sahin, "A survey on feature selection methods," Computers Electrical Engineering, vol. 40, no. 1, pp. 16–28, 2014. [Online]. Available:

http://www.sciencedirect.com/science/article/pii/S0045790613003066

[8] M.C.Newman,"Regression analysis of log-transformed data: Statistical bias and its correction," Environmental Toxicology and Chemistry, vol. 12, no. 6, pp. 1129–1133, 1993. [Online]. Available: http://dx.doi.org/10.1002/etc.5620120618

[9] R.Taylor, "Interpretation of the Correlation Coefficient: A Basic Review," Journal of Diagnostic Medical Sonography, vol. 6, no. 1, pp. 35–39, 1990

[10] Sameerchand Pudaruth, "Predicting the Price of Used Cars using Machine Learning Techniques";(IJICT 2014)

[11] Enis gegic, Becir Isakovic, Dino Keco, Zerina Masetic, Jasmin Kevric, "Car Price Prediction Using Machine Learning"; (TEM Journal 2019)

[12] Ning sun, Hongxi Bai, Yuxia Geng, Huizhu Shi, "Price Evaluation Model In Second Hand Car System Based On BP Neural Network Theory"; (Hohai University Changzhou, China)

[13] Nitis Monburinon, Suwat Rungpheung, Sabir Buya, Pitchayakit Boonpou, "Prediction of Prices for Used Car by using Regression Models" (ICBIR 2018).