Analysis of Used Car Price Prediction Using Machine Learning Algorithms

line 1: 2nd Given Name Surname
line 2: *dept. name of organization*
(of Affiliation)
line 3: *name of organization*
(of Affiliation)
line 4: City, Country
line 5: email address or ORCID

***Abstract*— As we can see, nowadays the automobile market is on boom everywhere. Due to the increase in demand, car prices are also increasing day by day so there is also a huge market for used cars. Sometimes, people avoid buying first-hand cars because according to their budget, they cannot get better products so most of the time they try to buy used cars. But sometimes, organizations of used cars started selling the products in huge amounts or providing good products as well because they saw us like this first hand but in actuality, it may be second hand. There may be defects in the engine or batteries as well. This paper suggests some machine learning techniques in which we will apply machine learning techniques on the dataset so that based on existing features we can predict like this is the actual price of used cars. Data is collected here.**

I. INTRODUCTION

Nowadays, the transportation industry is supporting the current GDP of the economy because when we use any new car, we have to give a price. On the road, car price is not much but according to the state, we have to give that much amount. Due to the cost line of the middle-class people, they are not able to buy new cars according to the time. So, in this case, they preferred to buy second-hand cars. Determining the exact selling price of the vehicle is a very difficult process we are getting the car at a proper rate according to the conditions whether the engine is good, the battery is good, or after checking all perspectives. Here we have to calculate price which is a continuous value that we prefer to apply regression techniques. To provide the actual rate of the used vehicles or cars, artificial intelligence techniques will be very helpful in predicting the car price with proper understanding. The main goal of this research analysis is to provide the actual price of the used vehicles to the people so that they will be helpful to buy or sell the cars to make them better in the automobile industry. One of the other goals we can also say implement multiple machine learning techniques and understand which model is better at providing better efficient results or accuracy. After data collection, initially, data is in raw format which is very necessary to be cleaned at any cost otherwise we can apply the techniques and then find out the better results. So, we applied the data cleaning and pre-processing techniques in which we removed the missing values and dropped unwanted columns. Selecting proper features is very important because some attributes are not that important as compared to other features. Label encoding is also implemented properly in which we convert the string data into integer value.

Data science methods and machine learning algorithms have become increasingly effective instruments for used car price prediction in recent years. Machine learning models can find patterns and relationships that help improve the accuracy of price predictions by utilizing large amounts of historical sales data in addition to pertinent features. Predictions can be continuously improved and refined over time thanks to these model's ability to adapt and learn from new data.

II. RELATED WORK

This recent study presents a research analysis of so many machine learning models to predict the prices of used cars. They use Python modules such as pandas, and matplotlib are employed for data pre-processing and model development. The authors discuss the impact of different feature sets on prediction accuracy and compare the performance of regression-based and tree-based models.

III. METHODOLOGY

A. Data Collection:

Used Car Price dataset is collected from the Kaggle to implement the machine learning algorithms. This folder contains the train and test dataset.

	Unnamed: 0	Name	Location	Year	Kilometers_Driven	Fuel_Type	Transmission	Owner_Type	Mileage	Engine	Power	Seats	New_Price	Price
0	0	Maruti Wagon R LXI CNG	Mumbai	2010	72000	CNG	Manual	First	26.6 km/kg	998 CC	58.16 bhp	5.0	NaN	1.75
1	1	Hyundai Creta 1.6 CRDI SX Option	Pune	2015	41000	Diesel	Manual	First	19.67 kmpl	1582 CC	126.2 bhp	5.0	NaN	12.50
2	2	Honda Jazz V	Chennai	2011	46000	Petrol	Manual	First	18.2 kmpl	1199 CC	88.7 bhp	5.0	8.61 Lakh	4.50
3	3	Maruti Ertiga VDI	Chennai	2012	87000	Diesel	Manual	First	20.77 kmpl	1248 CC	88.76 bhp	7.0	NaN	6.00
4	4	Audi A4 New 2.0 TDI Multitronic	Coimbatore	2013	40670	Diesel	Automatic	Second	15.2 kmpl	1968 CC	140.8 bhp	5.0	NaN	17.74

```
#Head Test Dataset
test_data=pd.read_csv("test-data.csv")
test_data.head()
```

	Unnamed: 0	Name	Location	Year	Kilometers_Driven	Fuel_Type	Transmission	Owner_Type	Mileage	Engine	Power	Seats	New_Price
0	0	Maruti Alto K10 LXI CNG	Delhi	2014	40929	CNG	Manual	First	32.26 km/kg	998 CC	58.2 bhp	4.0	NaN
1	1	Maruti Alto 800 2016-2019 LXI	Coimbatore	2013	54493	Petrol	Manual	Second	24.7 kmpl	796 CC	47.3 bhp	5.0	NaN
2	2	Toyota Innova Crysta Touring Sport 2.4 MT	Mumbai	2017	34000	Diesel	Manual	First	13.68 kmpl	2393 CC	147.8 bhp	7.0	25.27 Lakh
3	3	Toyota Etios Liva GD	Hyderabad	2012	139000	Diesel	Manual	First	23.59 kmpl	1364 CC	null bhp	5.0	NaN
4	4	Hyundai i20 Magna	Mumbai	2014	29000	Petrol	Manual	First	18.5 kmpl	1197 CC	82.85 bhp	5.0	NaN

Figure 1: Data Collection

We have checked the shape of the dataset so the training dataset contains 6019 rows and 14 columns while the test data contain 1234 rows and 13 columns.

```
[ ] #Dimension of train and test dataset
    print("Shape of Training Data:- ",train_data.shape)
    print("Shape of Test Data:- ",test_data.shape)

    Shape of Training Data:-  (6019, 14)
    Shape of Test Data:-  (1234, 13)
```

Figure 2: Dimension of Dataset

We have calculated the description of data which provides us the count, mean, standard deviation, minimum, maximum, 25%, 50%, and 75% values of numerical columns of the dataset.

```
[ ] #Data Description of train data
    train_data.describe()
```

	Unnamed: 0	Year	Kilometers_Driven	Seats	Price
count	6019.000000	6019.000000	6.019000e+03	5977.000000	6019.000000
mean	3009.000000	2013.358199	5.873838e+04	5.278735	9.479468
std	1737.679967	3.269742	9.126684e+04	0.808840	11.187917
min	0.000000	1998.000000	1.710000e+02	0.000000	0.440000
25%	1504.500000	2011.000000	3.400000e+04	5.000000	3.500000
50%	3009.000000	2014.000000	5.300000e+04	5.000000	5.640000
75%	4513.500000	2016.000000	7.300000e+04	5.000000	9.950000
max	6018.000000	2019.000000	6.500000e+06	10.000000	160.000000

```
[ ] #Data Description of test data
    test_data.describe()
```

	Unnamed: 0	Year	Kilometers_Driven	Seats
count	182.000000	182.000000	182.000000	182.000000
mean	626.269231	2015.939560	38771.681319	5.269231
std	352.312093	2.167483	26916.604504	0.792951
min	2.000000	2008.000000	1000.000000	2.000000
25%	333.500000	2015.000000	19814.000000	5.000000
50%	624.500000	2017.000000	31947.000000	5.000000

Figure 3: Data Description

B. *Data Cleaning:*

Data cleaning is a crucial step in the process of building a used car price prediction model. It involves identifying and correcting errors, handling missing values, and transforming the data into a format suitable for analysis.

We checked the missing values for train and test data using the isnull function and New_price almost contained a large number of missing values.

```
[ ] #check missing values of train data
    train_data.isnull().sum()

    Unnamed: 0           0
    Name                 0
    Location             0
    Year                 0
    Kilometers_Driven    0
    Fuel_Type            0
    Transmission         0
    Owner_Type           0
    Mileage              2
    Engine              36
    Power               36
    Seats               42
    New_Price         5195
    Price                0
    dtype: int64
```

```
[ ] #check missing values of test data
    test_data.isnull().sum()

    Unnamed: 0           0
    Name                 0
    Location             0
    Year                 0
    Kilometers_Driven    0
    Fuel_Type            0
    Transmission         0
    Owner_Type           0
    Mileage              0
    Engine              10
    Power               10
    Seats               11
    New_Price         1052
    dtype: int64
```

Figure 4: Check Missing Values

To remove the missing values, we dropped the missing rows using the drop function.

Some unwanted columns are also present in the dataset so, we have dropped the columns using the drop function. So, the dropped columns are name, location, transmission, and new price.

```
[ ] #Drop missing rows
    train_data_data = train_data.dropna(how='any')
    test_data = test_data.dropna(how='any')
    print("Shape of Training Data:- ",train_data.shape)
    print("Shape of Test Data:- ",test_data.shape)

    Shape of Training Data:-  (6019, 14)
    Shape of Test Data:-  (182, 13)
```

Figure 5: Drop Missing Rows

C. *Data Preprocessing:*

Data preprocessing is also one of the important steps to implement before applying machine learning techniques. Some columns are in object format so we have changed the data type into float and numeric. Some columns are also containing some spaces so we have used the rstrip function to remove the whitespaces.

```
[ ]  #Data Processing
     train_data1['Engine'] = train_data1['Engine'].str.
     train_data1['Power'] = train_data1['Power'].str.rs
     train_data1['Mileage'] = train_data1['Mileage'].st
```

```
[ ]  #Data Processing
     train_data1['Engine'] = train_data1['Engine'].asty
     train_data1['Power']=pd.to_numeric(train_data1['Po
```

Figure 6: Data Preprocessing

As we checked, there are some columns whose data type is a string, and machine learning never works on string datasets. So, we have applied the label encoding to convert the string data into an integer value. There are some missing values in the seat column also so, we implement the mean imputation to fill the missing values with the mean of the column.

```
[ ] #Label Encoding

from sklearn.preprocessing import LabelEncoder

train_data1[['Fuel_Type','Owner_Type','Mileage','Engine','Power']] = train_data1[['Fuel_Type','Owner_Type','Mileage','Engine','Power']].apply(LabelEncoder().fit_
test_data1[['Fuel_Type','Owner_Type','Mileage','Engine','Power']] = test_data1[['Fuel_Type','Owner_Type','Mileage','Engine','Power']].apply(LabelEncoder().fit_t
train_data1.head()
```

	Year	Kilometers_Driven	Fuel_Type	Owner_Type	Mileage	Engine	Power	Seats	Price
0	2010	72000	0	0	414	10	24	5.0	1.75
1	2015	41000	1	0	291	51	189	5.0	12.50
2	2011	46000	4	0	249	24	106	5.0	4.50
3	2012	87000	1	0	323	26	108	7.0	6.00
4	2013	40670	1	2	155	70	209	5.0	17.74

```
[ ] train_data1['Seats'].fillna(train_data1['Seats'].mean(), inplace=True)
```

Figure 7: Label Encoding

D. Data Visualization:

Data visualization plays a crucial role in understanding patterns, relationships, and trends within the dataset, which is essential for building accurate machine learning models for used car price prediction.

The first diagram is plotted the count plot of owner type which provides the information if the owner is first, second, third, or fourth and above so that they can easily identify the price based on how much vehicle is used.

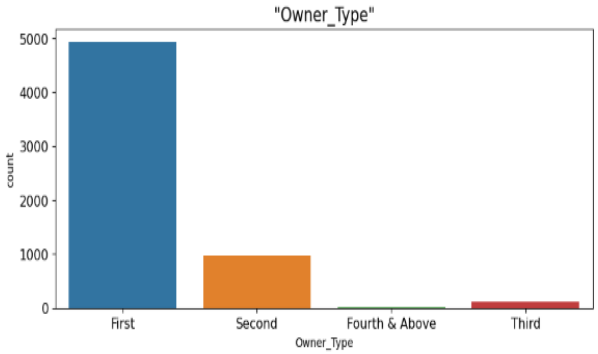

Figure 8: count Plot of Owner Type

The second diagram is a count plot of fuel type which provides us the type of vehicle it is means vehicle if diesel, CNG, petrol, LP or electric vehicle and according to the graph, most of the user cars are using diesel or petrol.

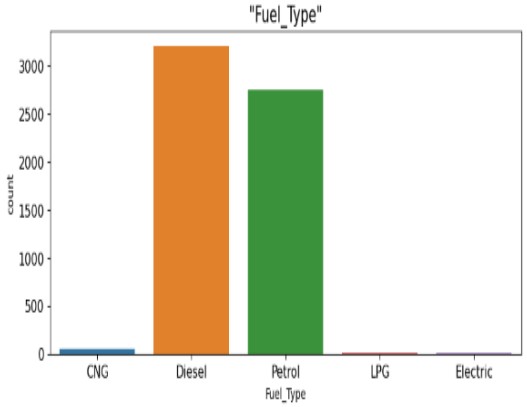

Figure 9: Count Plot of Fuel Type

The third diagram is a strip plot between owner type and price and we analyzed that for the first owner, there are higher prices of cars.

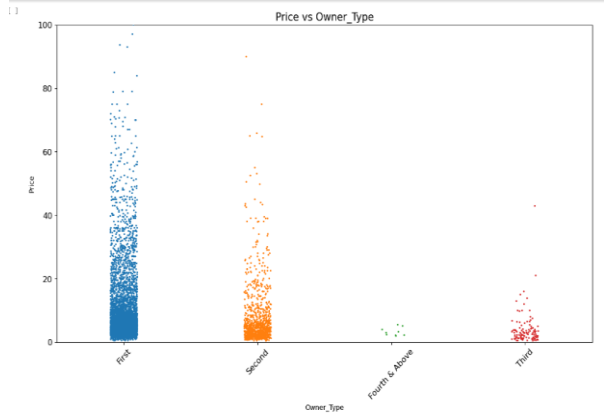

Figure 10: Strip plot b/w owner type and price

The fourth diagram is where we plotted the scatter plot between year and price. According to the analysis, between 2015 to 2020, there is the maximum price of used car vehicles.

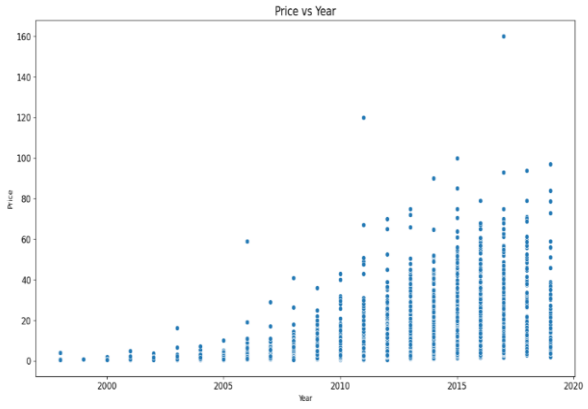

Figure 11: Scatter Plot between Year and Price

Fifth diagram is also a scatter plot between price and engine.

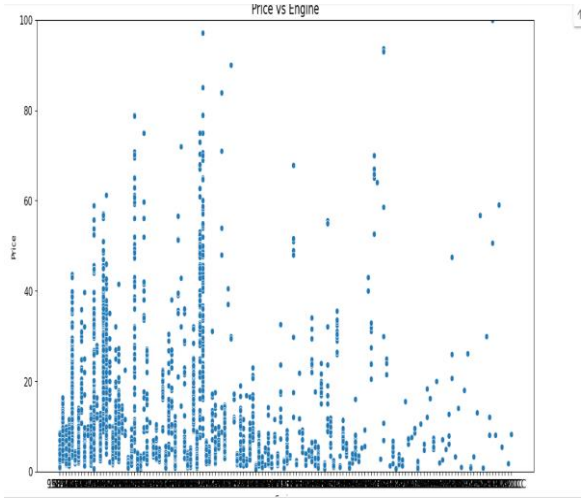

Figure 12: Scatter Plot b/w price and engine

The sixth diagram we plotted is a heatmap which represents the correlation matrix of the whole dataset with numerical values only.

```
[ ] #Plot heatmap
    plt.figure(figsize=(10, 6))
    sns.heatmap(train_data.corr(), annot=False, cmap='Blues')
    plt.show()
```

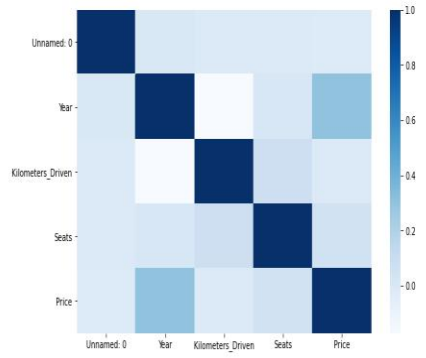

Figure 13: Heatmap

Figure 14 Predict values vs actual price

IV. Model Evaluation

After data cleaning and preprocessing, it is time to apply the machine learning algorithms. We took the dependent and independent variables to make the prediction. So, column price is the dependent variable while all other remaining variables and independent variables.

Now we split the dataset into training and test data. 70% of the data is for training purposes while the remaining 30% is for test purposes.

Data Splitting

```
[ ] #Data Splitting
    x = train_data1.drop(['Price'],axis=1)
    y = train_data1['Price']
```

```
[ ] from sklearn.model_selection import train_test_split
    x_train, x_test, y_train , y_test = train_test_split(x,y, test_size = 0.30, random_state = 42)
```

Figure 15: Data Splitting

We have applied the machine learning algorithms and calculated the accuracy score.

A. *Linear Regression:*

Linear regression is one of the regression models that find the linear relationship between the dependent and independent variables.

Linear Regression

```
[ ] #Linear Regression
    from sklearn.linear_model import LinearRegression
    regressor = LinearRegression()
    regressor.fit(x_train,y_train)
    y_pred = regressor.predict(x_test)
```

```
from sklearn.metrics import mean_squared_error, r2_score, mean_absolute_error
import numpy as np
mae = mean_absolute_error(y_test, y_pred)
mse = mean_squared_error(y_test, y_pred)
rmse = np.sqrt(mean_squared_error(y_test, y_pred))
r_squared = r2_score(y_test, y_pred)
print("Mean Absolute Error of Linear Regression:- ",mae)
print("Mean Squared Error of Linear Regression:- ",mse)
print("Root Mean Squared Error of Linear Regression:- ",rmse)
print("R Sq uared Error of Linear Regression:- ",r_squared)
```

```
Mean Absolute Error of Linear Regression:-  4.764177673015051
Mean Squared Error of Linear Regression:-  82.74580074585904
Root Mean Squared Error of Linear Regression:-  9.096432086065041
R Sq uared Error of Linear Regression:-  0.3272684702494352
```

Figure 16: Linear Regression

B. *Random Forest Regression:*

Random Forest is a supervised learning model that performs classification and regression data to handle the continuous and binary data.

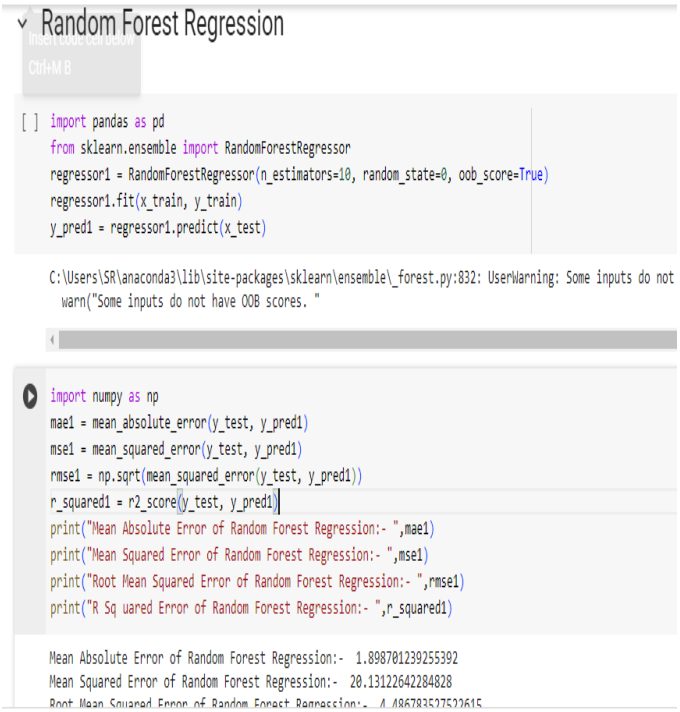

Figure 17: Random Forest Regression

C. Support Vector Regression:

The goal of support vector recognition (SVR) is to find the best hyperplane in a high-dimensional feature space to enable precise prediction. One of the key characteristics of SVR is the addition of a tolerance margin represented by the parameter ε (epsilon) which establishes the maximum amount that predicted values may deviate from actual values. This method seeks to ensure that the majority of data points fall within the tolerance while minimizing errors.

Figure 18: Support Vector Regression

	Model Name	Mean Absolute Error	Mean Squared Error	Root Mean Squared Error	R Suared Error
0	Linear Regression	4.764178	82.745081	9.096432	0.327268
1	Random Forest	1.898701	20.131226	4.486784	0.836330
2	Support Vector Machine	6.047488	134.738874	11.607708	-0.095450

Figure 19: Accuracy Score

V. CONCLUSION

In the end, we come to the point that it's now become very easy with the help of machine learning techniques to detect and predict the price of cars based on the few important parameters for which we used our codes. And we can say that based on the accuracy of the reports we can predict it. Data science and machine learning have become very useful and play a vital role in the company to reduce the workload and can easily check the conditions of used cars. This method works by merging the results of different models in Random Forest this merge is done with a group of decision trees. This technique assumes an association between the independent variables and the target variable. In the scenario of forecasting used car prices, linear regression serves as a tool to estimate the price of a pre-owned vehicle by considering its attributes such as mileage, age, brand, model, and other pertinent factors.

VI. FUTURE SCOPE

In the future, machine learning and artificial intelligence may combine the data of various websites and predict the real-time price as well by comparing the results of cars. We can also apply various deep learning techniques as well to improve the performance score of used car prices for better productivity by applying the best learning rates and training the clusters on the historical data as well.

REFERENCES

1. Chaganti, R., Rustam, F., De La Torre Díez, I., Mazón, J.L.V., Gupta, S., Vijarania, M. and Udbhav, M., 2023. A Machine Learning Approach for Predicting Price of Used Cars and Power Demand Forecasting to Conserve Non-renewable Energy Sources. In Renewable Energy Optimization, Planning and Control: Proceedings of ICRTE 2022 (pp. 301-310). Singapore: Springer Nature Singapore.
2. Poufinas, T., Gogas, P., Papadimitriou, T. and Zaganidis, E., 2023. Machine learning in forecasting motor insurance claims. Risks, 11(9), p.164.
3. Mazhar, T., Asif, R.N., Malik, M.A., Nadeem, M.A., Haq, I., Iqbal, M., Kamran, M. and Ashraf, S., 2023. Electric vehicle charging system in the smart grid using different machine learning methods. Sustainability, 15(3), p.2603.
4. Ahmed, S., Hossain, M.A., Ray, S.K., Bhuiyan, M.M.I. and Sabuj, S.R., 2023. A study on road accident prediction and contributing factors using explainable machine learning models: Analysis and performance. Transportation research interdisciplinary perspectives, 19, p.100814.

5. Li, Q., Cheng, R. and Ge, H., 2023. Short-term vehicle speed prediction based on BiLSTM-GRU model considering driver heterogeneity. Physica A: Statistical Mechanics and its Applications, 610, p.128410.
6. Wang, S., Zhuge, C., Shao, C., Wang, P., Yang, X. and Wang, S., 2023. Short-term electric vehicle charging demand prediction: A deep learning approach. Applied Energy, 340, p.121032.
7. Borup, D., Christensen, B.J., Mühlbach, N.S. and Nielsen, M.S., 2023. Targeting predictors in random forest regression. International Journal of Forecasting, 39(2), pp.841-868.
8. Naseri, H., Jahanbakhsh, H., Foomajd, A., Galustanian, N., Karimi, M.M. and D. Waygood, E.O., 2023. A newly developed hybrid method on pavement maintenance and rehabilitation optimization applying Whale Optimization Algorithm and random forest regression. International Journal of Pavement Engineering, 24(2), p.2147672.
9. Cai, W., Wen, X., Li, C., Shao, J. and Xu, J., 2023. Predicting the energy consumption in buildings using the optimized support vector regression model. Energy, 273, p.127188.
10. Nong, X., Lai, C., Chen, L., Shao, D., Zhang, C. and Liang, J., 2023. Prediction modelling framework comparative analysis of dissolved oxygen concentration variations using support vector regression coupled with multiple feature engineering and optimization methods: A case study in China. Ecological Indicators, 146, p.109845.