Proceedings of the International Conference on Inventive Computation Technologies (ICICT 2023)
IEEE Xplore Part Number: CFP23F70-ART; ISBN: 979-8-3503-9849-6

Car Price Prediction: An Application of Machine Learning

Snehit Shaprapawad
2nd Year student Post Graduate Diploma in Management (PGDM) (Artificial Intelligence and Machine Learning Specialization)
Ashoka School Of Business
Hyderabad, India
ssnehitreddy@gmail.com

Premkumar Borugadda
Assistant Professor
Ashoka School Of Business
Hyderabad, India
premkumar.jones@gmail.com

Nirmala Koshika
2nd Year student Post Graduate Diploma in Management (PGDM) (Artificial Intelligence and Machine Learning Specialization)
Ashoka School Of Business
Hyderabad, India
nirmalakoshika@gmail.com

***Abstract*— In order to determine the worthiness of a car based on a variety of factors using machine learning models. In this study, the challenge is to prevent the model from becoming overfit and to generalize it. A combination of regularization techniques as well as hyperparameter tuning techniques was employed to overcome this challenge. Develop linear regression, lasso regression, ridge regression, elastic net regression, random forest, decision tree and Support Vector Machine models with hyperparameters. The objective of this article is to build a generalized model that can predict the price of used cars based on some factors, such as the car's mileage, the year it was made, the road tax, the type of fuel it uses, the size of its engine etc. Optimal model can help sellers, buyers, and car manufacturers. A relatively accurate prediction of price can be made based on information provided by users. Among the seven models, the support vector regressor is the optimal model based on the evaluation metrics such as R Squared (R^2) of 95.27 %, Mean Absolute Error (MAE) of 0.142, Mean Squared Error (MSE) of 0.047, and Root Mean Squared Error (RMSE) of 0.218 at 90% of the train data and 10% of the validation data.**

Keywords— Car price Prediction, Machine Learning, Mean Absolute Error, Mean Squared Error, Root Mean Squared Error, Train Data, Validation Data.

I. INTRODUCTION

In recent years, the used automobile market has grown to become one of the fastest growing industries in the world. This is because it has a market value that has nearly doubled over the past few years as a result. Compared to a few years ago, online websites have become quite popular, as have other tools that are similar to them. This has made it quite easy for buyers and sellers to get an insight into the different factors that matter when determining a used car's market value. The price of any automobile can be predicted using Machine Learning (ML) algorithms that are based on a set of factors [1].

A variety of information about a variety of automobiles is colleted in the proposed data set. For each of the vehicles, there will be technical information regarding the engine type, the fuel type, the kilometers perlite, and more. Information about these vehicles describes its performance. Although the retail price of a used car can vary depending on the website that customer is visiting, it is generally possible to get a price forecast without entering all of the details that you wish to see in the report. This is because each website has its own method of calculating the value of a car [2].

Using seven different prediction models to estimate the value of a used car at retail, this study primarily aims to compare the accuracy of each prediction model in order to estimate the value of a used car.

There are a variety of vehicles in the market, ranging from cars, sedans, coupes, support cars, station wagons, hatchbacks, convertibles, sport-utility vehicles (SUVs), minivans, trucks, and so on. There are different types of vehicles on this market with different features and uses, and as a result the prices change based on that.

Many people have expressed an interest in purchasing a used vehicle in the past. This is since customers were either trying to sell or purchase a used car. In this process, need to be careful not to overpay for buying or selling it for less than its market value. For that a good forecasting model may be available to customers. As a result, consumers are provided with a more accurate forecast Therefore, the developed model in this research useful to online services [3].

II. LITERATURE REVIEW

Table 1: Comparison of Machine Learning Algorithms on Car Price Dataset

Reference	Approach	Objective	Algorithms Used	R^2 score (%)	Merits	Demerits
[4]	Using regression techniques for building the best model	Find the best predicting model for estimating the used car price	SVM ANN	90 85.71	Analysis and results are explained in details	Comparison between proposed work and existing work is not mentioned
[5]	Linear regression with the OLS method used	To forecast the used car's price	Lasso Linear Regression	91	Methodology explained clearly in detail	Accuracy of each individual model is not mentioned
[6]	Supervised ML Techniques Used	ML-based prediction of used car	Lasso MLR decision Tree	3.58 (Error Rate) 3.46 3.51	Used statistical tests like ANOVA, P-	Only error rate is mentioned instead

2023 International Conference on Inventive Computation Technologies (ICICT) | 979-8-3503-9849-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICICT57646.2023.10134142

		retail prices			Value	of evaluation metrics
[7]	Using Supervised ML	Developing a model to estimate the used car price	Linear Regression Ridge Regression Lasso Regression	Not mentioned	Methodology explained clearly with flowchart	Results not mentioned clearly
[8]	Machine learning techniques like linear and ensemble models are used.	To make the ML model for predicting car price	Random forest KNN Decision Tree XG Boost Linear regression	93.11 69.66 84.30 92.04 76.46	Analysis of the study explained in detail	Evaluation metrics of each model is not mentioned

III. METHODOLOGY

Figure 1 illustrates the ML method's structure. The whole procedure for its actual implementation can be split into five stages. The five stages are data collection, data pre-processing, train a ML model, evaluate the model and predicting the target variable.

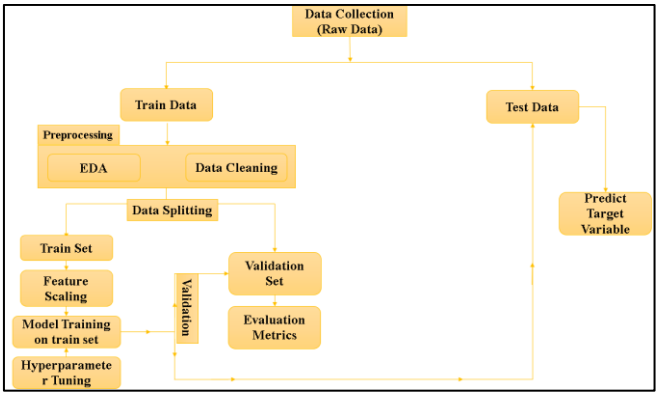

Figure 1. Framework of Machine Learning Models

3.1 Data Collection

Data is collected from Kaggle [20] information of dataset is given in the following figure 2, which includes raw data of variety characteristics.

```
#   Column             Non-Null Count  Dtype
---  ------             --------------  -----
0   Name               6019 non-null   object
1   Location           6019 non-null   object
2   Year               6019 non-null   int64
3   Kilometers_Driven  6019 non-null   int64
4   Fuel_Type          6019 non-null   object
5   Transmission       6019 non-null   object
6   Owner_Type         6019 non-null   object
7   Mileage            6017 non-null   object
8   Engine             5983 non-null   object
9   Power              5983 non-null   object
10  Seats              5977 non-null   float64
11  New_Price          824 non-null    object
12  Price              6019 non-null   float64
```

Figure 2. Information of Dataset

3.2 Data Preprocessing

Preprocessing the data is a very important part of ML, as it will allow the most accurate possible results to be obtained. At this stage, data pre-processing is divided into two subdivisions that are referred to as Exploratory Data Analysis (EDA) and feature scaling. EDA is a pre-processing technique to help gain insights and gain a better understanding of the data which help to eliminate irregularities and unnecessary values that may arise from it. Provides you with assistance when it comes to preparing dataset for analysis. As a result, dataset can be predicted more accurately with a machine learning model and gives the more accurate results. EDA consist of various subdivisions like converting categorical to numerical, data cleaning and feature scaling. Categorical feature is converted to numerical form using a nominal encoder [9] at the beginning. Then, those numerical values are converted into binary numbers by a binary encoder [10] at the end. After this process has been completed, the binary values have been divided into different columns. When there are many categories, binary encoding is a very effective way to encode them. Data cleaning is the process that involves removing corrupt data, incorrectly formatted data, null values, and duplicate values from a dataset. The purpose of feature scaling is to distribute the data points closer together by transforming the data into a specific scale, or in a range with a specific size, such as 0 to 1. This practice is required when the data has data points that are far apart in any condition. Scaling is the technique to move them closer together.

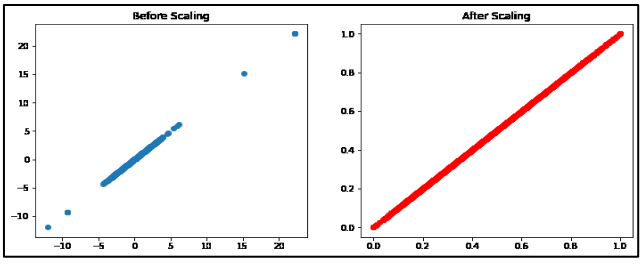

Figure 3. Comparison of Data Before and After applying scaling

In order to scale the features, the MinMax scalar is used, which allows the features to be reduced to a range of 0 to 1 and thereby decrease the complexity of the training model. Figure 3 shows that the distribution of the data before and after the scaling is significantly different. This is also known

as the Minmax scalar method. The Minmax Scaler divides the result by the range after dividing the feature's minimal value from it [11].

3.3 Training ML Model

This data set is divided into training and validation data at 90%-10% respectively. Fit the model on 90% of training dataset and then checked against a validation data set of 10% to assess how well the trained model is performing. As the name implies, ML is primarily intended to reduce errors associated with a model.

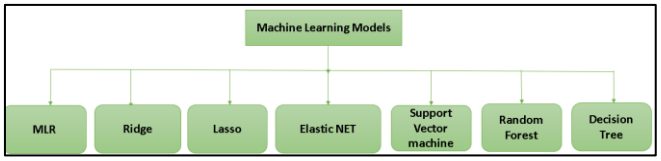

Figure 4. Machine Learning Regression Models

Figure 4 illustrates a visual representation of all the machine learning algorithms are applied during training the models.

A. Multiple Linear Regression (MLR)

Based on the fact that multiple linear regression is a regression model that uses straight lines to estimate the relationship between two or more independent variables and a quantitative dependent variable [12]. The linear equation for regression is shown in below. There is a linear relationship between the variables.

$$y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_n x_n \quad \cdots\cdots\cdots \quad (1)$$

Where, x = exploratory variable

y = dependent variable, b = slope, b_0 = intercept

B. Ridge Regression

The ridge regression technique is a regularization method which is also known as L2 norm. It is used to estimate the coefficients of multiple regression models when the variables are independent are highly correlated. Equation 2 shows the formula of ridge regression [13].

$$\sum_{i=1}^{M} (y_i - \hat{y_i})^2 = \sum_{i=1}^{M} (y_i - \sum_{i=1}^{M} w_j \times x_{ij})^2 \cdots\cdots\cdots \quad (2)$$

Where,

C. Lasso Regression

The lasso regression technique is a regularization method which is also known as L1 norm. It is used over regression methods to make predictions that are more accurate since shrinkage is the process of shrinking the values of the data towards a central point as the mean [14]. Equation 3 shows the formula of lasso regression.

$$\sum_{i=0}^{n} (y_i - \sum_{j} x_{ij})^2 + \alpha \sum_{j=1}^{p} |\beta_j| \quad \cdots\cdots\cdots \quad (3)$$

D. Elastic-Net Regression (E-NET)

There are two parameters in elastic net regression, alpha and L1_ratio, and elastic net regression is a technique that performs well on a large dataset. Equation 4 shows the formula of elastic net regression.[15]

$$\sum_{i=1}^{r} (y_i - \sum_{i=0}^{c} m_j \cdot x_{ij})^2 + \alpha \sum_{j=0}^{c} |m_j| + \alpha \sum_{j=0}^{c} m_j^2 \quad \cdots\cdots\cdots \quad (4)$$

E. Support Vector Regressor (SVR)

Using support vector machine, one can predict discrete values in a supervised manner. This is the same principle that is used in Support vector regressor, as well as in Support Vector Machines (SVMs). Finding the ideal best fit line is the basic aspect of SVR, which considers the hyperplane with the most points to be the best fit line [16].

F. Random Forest Regressor

By combining several categorization decision trees upon a sample of data, the random forest regressor can enhance prediction performance and control overfitting [17].

G. Decision Tree Regressor (DTR)

In a decision tree, each node is an example of a feature, and every branch represents a decision rule [18].

3.4 Evaluation of Models

The model evaluation process is a process by which different evaluation metrics are used in order to examine a machine learning model's performance. The evaluation metrics are R^2 score, MAE, MSE, RMSE [19].

A. R Squared (R^2)

R^2 is a metric which evaluates how effectively a regression model fits the data. If R-square is close to 1,

the model will fit better. Therefore, if this is 100%, the two variables are perfectly correlated, or there are no variances at all. (Total variance explained by model) / total variance. Equation 5 shows the formula of R^2 score.

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \quad \text{..... (5)}$$

B. Mean Absolute Error (MAE)

As a machine learning model is being developed, there are many metrics for evaluating the quality of the model. An example of a measurement method is the MAE, which is one of the metrics that measures the average magnitude of errors. Equation 6 shows the formula of MAE.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \quad \cdots\cdots\cdots \quad (6)$$

C. Mean Squared Error (MSE)

An MSE is an accurate measure of the degree to which regression problems are modeled using the following equation 7.

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad \cdots\cdots\cdots \quad (7)$$

The MSE is determined by averaging throughout the whole dataset the square of the original values minus the square of the predicted values.

D. Root Mean Squared Error (RMSE)

To assess how well a regression model is performing, one can use the RMSE metric which is comparable to the standard deviation (SD) for an ideal measurement model. As a result of the SD, we can estimate the deviation from the sample mean x. Equation 6 shows the formula of RMSE.

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (\hat{y}_i - y_i)^2} \quad \cdots\cdots\cdots \quad (8)$$

3.5 Predicting the Target Variable

When the training process has been completed, the model will be able to predict the target variable on a test data set. It is always the case that test datasets do not have a target variable.

IV. EXPETIMENTAL SETUP AND HYPERPARAMETER TUNING

The present experiment was executed on Jupyter notebook support to implement ML algorithms in Python 3.10.6 The hardware configuration is given in Table 3.

Table 3. System Configuration

Sr.no.	Hardware and Software	Characteristics
1	Memory (RAM)	16 GB
2	Processor	11th Gen Intel(R) Core (TM) i5-1135G7 @ 2.40GHz 2.42 GHz
3	Graphics (GPU)	Intel iris 2GB
4	Operating system	Windows 11 64-bit operating system
5	Integrated Development Environment (IDE)	Jupyter notebook

Hyperparameters are values that are determined during model training that were optimized to improve the model results. The hyperparameters of ML models are shown in table 4.

To tune the hyperparameters, there are certain steps that need to be taken, including defining the search space, selecting the search technique (GridsearchCV, RandomsearchCV), choosing the performance metrics, assessing the model, and researching and refining.

Table 4. Hyperparameters of Machine Learning Models

Sr. No.	Models	Hyperparameters	Optimal Values
1	Ridge Regression	Learning_rate	0.01
2	Lasso Regression	Learning_rate	0.001
3	Elastic Net Regression	Learning_rate	0.0001
4	Decision Tree Regressor	criterion	Friedman_mse
		Splitter	Random
		Min_sample_split	600
5	Support Vector Regressor	Degree	1
		Gamma	Auto
		Kernel	Rbf
6	Random Forest Regressor	Criterion	Absolute_error
		Max_depth	11
		n_estimators	150

The above table 4 shows the optimal hyperparameters for the regression algorithms used in this research. Applied the GridsearchCV on the hyperparameters with a CV is 10 to all models. Among the seven models, support vector regressor given the best accuracy.

V. ANALYSIS OF RESULTS

The purpose of this section is to summarize the results of an extensive experiment in which several machine learning models have been developed, including Lasso, Ridge,

Elastic Net, Random Forest, Decision Tree, and Support Vector Regression. The following table 5 shows the performance measurements of ML models. From this table, it is observed that the support vector model provides better results compared to other models.

Table 5. Analysis of Results

Models	Train & Test (%)	R^2 score	MAE	MSE	RMSE
MLR	90-10	86.64	0.26	0.134	0.366
Lasso	90-10	86.64	0.265	0.134	0.366
Ridge	90-10	86.64	0.265	0.134	0.366
E-NET	90-10	86.65	0.26	0.134	0.366
Decision Tree	90-10	90.6	0.217	0.093	0.306
Support Vector Regressor	**90-10**	**95.27**	**0.142**	**0.047**	**0.218**
Random Forest	90-10	94.04	0.164	0.059	0.244
MLR	80-20	87.16	0.261	0.134	0.366
Lasso	80-20	87.16	0.261	0.134	0.366
Ridge	80-20	87.16	0.261	0.134	0.366
E-NET	80-20	87.17	0.261	0.134	0.366
Decision Tree	80-20	88.76	0.217	0.093	0.306
Support Vector Regressor	80-20	95.06	0.151	0.051	0.226
Random Forest	80-20	93.34	0.175	0.069	0.263
MLR	70-30	87.25	0.261	0.134	0.366
Lasso	70-30	87.25	0.261	0.130	0.360
Ridge	70-30	87.25	0.261	0.130	0.360
E-NET	70-30	87.25	0.26	0.130	0.360
Decision Tree	70-30	86.88	0.242	0.133	0.365
Support Vector Regressor	70-30	94.80	0.153	0.053	0.230
Random Forest	70-30	92.84	0.182	0.073	0.270

An analysis of R^2 scores for distinct ML models visual representation in the following figure 5. Among the seven ML models, the Support vector regressor has the maximum R^2 score is 95.27% contrast to other ML models with 90% training data set and 10% validation data set.

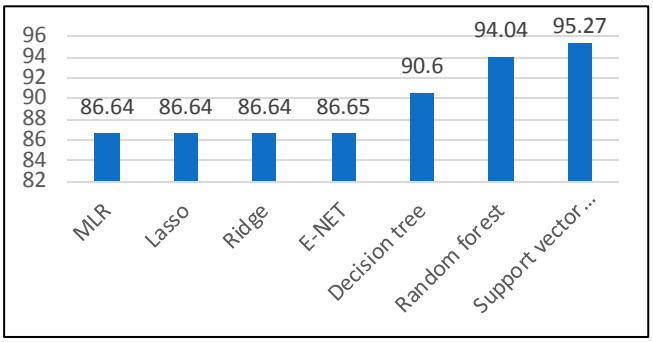

Figure 5. Comparison of R^2 score

Figure 6 compares the MAE for various machine-learning models. The support vector regressor model has the lowest MAE of 0.16 out of all seven ML models with 90% training data set and 10% validation data set.

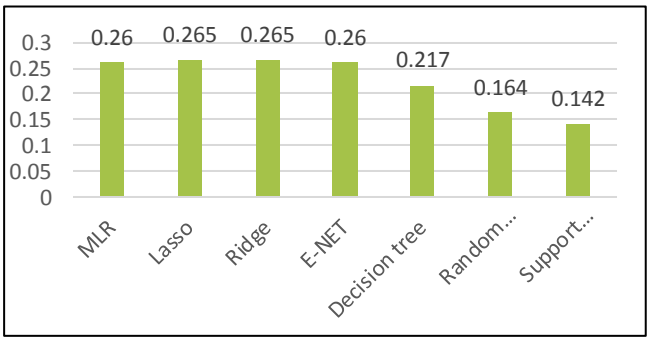

Figure 6. Comparison of MAE

Figure 7 compares the MSC for various machine-learning models. The support vector regressor model has the lowest MSC of 0.059 out of all seven ML models with 90% training data set and 10% validation data set.

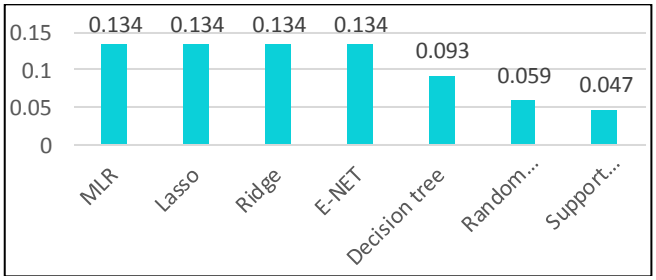

Figure 7 Comparison of MSE

Figure 8 compares the RMSC for various machine-learning models. The support vector regressor model has the lowest RMSC of 0.244 out of all seven ML models with 90% training data set and 10% validation data set.

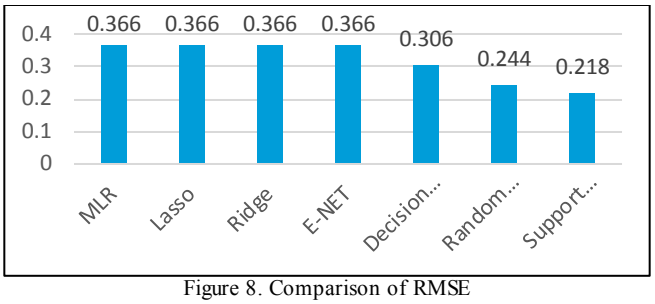

Figure 8. Comparison of RMSE

Table 6. Performance Comparison of Proposed work with Previous Models

Sr.no	Methods	R^2 score
1	Kshitij Kumbar (Deep neural network)	85%
2	Asghar, M., Mehmood, K., Yasin, S., & Khan (OLS regression with VIF method)	90%
3	Gajera, P., Gondaliya, A., & Kavathiya, (Random forest regressor)	93.11%
4	T. Veda Reddy (Gradient boosting with hyperparameter tuning)	94%
5	**Proposed work (Support vector regressor with hyperparameter tuning)**	**95.27%**

The suggested model's efficiency was assessed regarding that of the ML models developed previously by M. Asghar, Prashant Gajera, Kshitij Kumbar, T Veda Reddy.
As a result of this proposed study, techniques such as hyperparameter tuning, model selection, and regularization techniques were used to enhance the performance of the model.

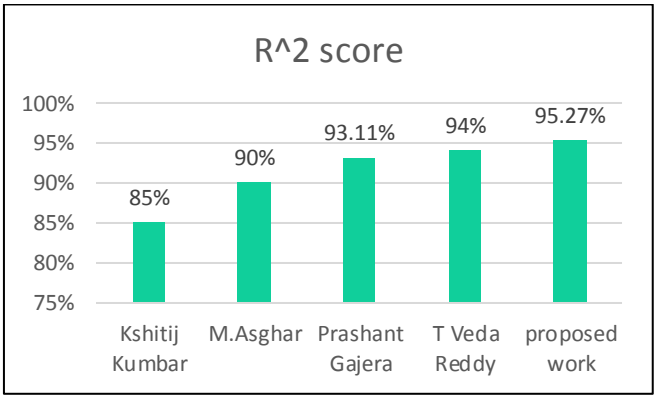

Figure 9. Contrasting the Validation Comparing the suggested model's R^2 score to earlier models

Figure 9 compares the suggested model with the original model on a collection of data related toused car price.
The current model's validation accuracy was enhanced and evaluated to earlier models like kshitij Kumbar (85%), M.Asghar (90%), Prashant Gajera(93.11%), T Veda Reddy (95.27%)

VI. CONCLUSION

Due to the numerous aspects that must be taken into account for an effective automobile forecast in India, there is an increasing desire for individual transportation in the present market, making it challenging for car estimates to be correct.Consequently, there is a booming used-car market in India, with buyers being able to choose from broad selection, simple access to finance, and usage of practical digital sales methods. A brief overview of each of the seven machine learning models are used in this research has been presented. The support vector regressor achieved the highest R^2 score of 95.27 percent using 90 percent of training dataset and 10 percent of the validation dataset, respectively. Furthermore, all ML algorithms are also evaluated for their performance Each of the ML algorithms was compared with the actual target values.

REFERENCES

[1] Chandak, A., Ganorkar, P., Sharma, S., Bagmar, A., & Tiwari, S. (2019). Car Price Prediction Using Machine Learning. International Journal of Computer Sciences andEngineering, 7(5), 444-450.
[2] Mammadov, H. (2021). Car Price Prediction in the USA by using Liner Regression. International Journal of Economic Behavior (IJEB), 11(1), 99-108.
[3] Samruddhi, K., & Kumar, R. A. (2020). Used Car Price Prediction using K-Nearest Neighbor Based Model. Int. J. Innov. Res. Appl. Sci. Eng. (IJIRASE), 4, 629-632.
[4] Gegic, E., Isakovic, B., Keco, D., Masetic, Z., & Kevric, J. (2019). Car price prediction using machine learning techniques. TEM Journal, 8(1), 113.
[5] Asghar, M., Mehmood, K., Yasin, S., & Khan, Z. M. (2021). Used Cars Price Prediction using Machine Learning with Optimal Features. Pakistan Journal of Engineering and Technology, 4(2), 113-119.
[6] Venkatasubbu, P., & Ganesh, M. (2019). Used Cars Price Prediction using Supervised Learning Techniques. Int. J. Eng. Adv. Technol. (IJEAT), 9(1S3).
[7] Praful Ranel, Deep Pandya2, Dhawal Kotak3 "Used car price prediction "International Research Journal of Engineering and Technology (IRJET). (2021)
[8] Gajera, P., Gondaliya, A., & Kavathiya, J. (2021). Old Car Price Prediction with Machine Learning. Int. Res. J. Mod. Eng. Technol. Sci, 3, 284-290.
[9] Potdar, K., Pardawala, T. S., & Pai, C. D. (2017). A comparative study of categorical variable encoding techniques for neural network classifiers. International journal of computer applications, 175(4), 7-9.
[10] Lu, C., Hu, X., Yang, H., & Gong, Q. (2013). All-optical logic binary encoder based on asymmetric plasmonic nanogrooves. Applied Physics Letters, 103(12), 121107.
[11] Shaheen, H., Agarwal, S., & Ranjan, P. (2020). MinMaxScaler binary PSO for feature selection. In First International Conference on Sustainable Technologies for Computational Intelligence: Proceedings of ICTSCI 2019 (pp. 705-716). Springer Singapore.
[12] Mrs. T Veda Reddy, Y. Praneeth, Y. Sai Kiran, G. Sai Pavan Car Price Prediction using Machine Learning
[13] Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: applications to nonorthogonal problems. Technometrics, 12(1), 69-82.
[14] Ranstam, J., & Cook, J. A. (2018). LASSO regression. Journal of British Surgery, 105(10), 1348-1348.
[15] Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the royal statistical society: series B (statistical methodology), 67(2), 301-320.
[16] Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. Statistics and computing, 14, 199-222.
[17] Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M., & Chica-Rivas, M. J. O. G. R. (2015). Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. Ore Geology Reviews, 71, 804-818.
[18] Tso, G. K., & Yau, K. K. (2007). Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks. Energy, 32(9), 1761-1768.

[19] Premkumar.B, R. Lakshmi, Bichitrananda Behera. (2020). Performance Analysis and Evaluation of Machine Learning Algorithms in Rainfall Prediction. International Journal of Advanced Science and Technology, 29(05), 5727 - 5741.
[20] https://www.kaggle.com/datasets/avikasliwal/used-cars-price-prediction?select=traindata.csv