ISSN 2063-5346

PREDICTION OF THE PRE-OWNED CAR SELLING VALUE DETERMINATION USING XGBOOST REGRESSION TECHNIQUE

Sahana h[1] Mrs Ashlesha Pandhare [2]

Article History: Received: 10.05.2023 Revised: 29.05.2023 Accepted: 09.06.2023

Abstract

Due to factors including high prices, limited supply, financial inability, and other factors, newly built cars are unable to reach buyers despite the significant expansion in car usage. As a result, the used automobile industry is growing rapidly all over the world, but it is still in its infancy in India and is largely dominated by the unorganised sector. This creates the potential for deception when purchasing a used car. Thus, a highly accurate model that can estimate the cost of a used car without favouring either the customer or the merchandiser is needed. This model develops an XGBoost Regression model based on supervised learning that can learn from the input automobile dataset. This project offers a low error workable model for estimating secondhand car prices. For dependable and accurate forecasts, a large number of unique attributes are considered. The findings achieved are in line with theoretical predictions, and they outperform models that rely on straightforward linear models. The XGBoost Regression algorithm is used to build a model, and for comparison, other machine learning algorithms such as Linear Regression, Random Forest Regression, Lasso Regression, Ridge Regression, Decision Tree Regression, ElasticNet Regression, Adaboost Regression, and Gradient Boosting Regressions are also built. The automobile dataset is used to test these techniques. According to experimental findings, the XGBoost Regression model has a training accuracy of 99.86%, a test accuracy of 94.45%, and a root mean square error of 0.05237.among all the other methods, has produced the least inaccuracy. The work reported here has significant ramifications for future research on the XGBoost Regression model for Used Car Price Prediction, and it may one day contribute to a 100% accurate solution to the fraud problem.

Keywords—Keras, Used car price prediction, Regression, Random Forest, Machine Learning, Ridge, LASSO, Linear regression.

[1]Research Scholar, Dept. of Computer Science and Engineering, Lords Institute of Engineering & Technology, Hyderabad, Telangana

[2]Associate Professor, Dept. of Computer Science and Engineering, Lords Institute of Engineering & Technology, Hyderabad, Telangana

DOI:10.48047/ecb/2023.12.9.84

I. INTRODUCTION

Predicting the price of used cars in both an important and interesting problem. According to data obtained from the National Transport Authority [1], the number of cars registered between 2003 and 2013 has witnessed a spectacular increase of 234%. From 68, 524 cars registered in 2003, this number has now reached 160, 701. With difficult economic conditions, it is likely that sales of second-hand imported (reconditioned) cars and used cars will increase. It is reported in [2] that the sales of new cars has registered a decrease of 8% in 2013. In many developed countries, it is common to lease a car rather than buying it outright. A lease is a binding contract between a buyer and a seller (or a third party – usually a bank, insurance firm or other financial institutions) in which the buyer must pay fixed instalments for a pre-defined number of months/years to the seller/financer. After the lease period is over, the buyer has the possibility to buy the car at its residual value, i.e. its expected resale value. Thus, it is of commercial interest to seller/financers to be able to predict the salvage value (residual value) of cars with accuracy. If the residual value is under-estimated by the seller/financer at the beginning, the instalments will be higher for the clients who will certainly then opt for another seller/financer. If the residual value is over-estimated, the instalments will be lower for the clients but then the seller/financer may have much difficulty at selling these high-priced used cars at this over-estimated residual value. Thus, we can see that estimating the price of used cars is of very high commercial importance as well. Manufacturers' from Germany made a loss of 1 billion Euros in their USA market because of mis-calculating the residual value of leased cars [3]. Most individuals in Mauritius who buy new cars are also very apprehensive about the resale value of their cars after certain number of years when they will possibly sell it in the used cars market. Predicting the resale value of a car is not a simple task. It is trite knowledge that the value of used cars depends on a number of factors. The most important ones are usually the age of the car, its make (and model), the origin of the car (the original country of the manufacturer), its mileage (the number of kilometers it has run) and its horsepower. Due to rising fuel prices, fuel economy is also of prime importance. Unfortunately, in practice, most people do not know exactly how much fuel their car consumes for each km driven. Other factors such as the type of fuel it uses, the interior style, the braking system, acceleration, the volume of its cylinders (measured in cc), safety index, its size, number of doors, paint colour, weight of the car, consumer reviews, prestigious awards won by the car manufacturer, its physical state, whether it is a sports car, whether it has cruise control, whether it is automatic or manual transmission, whether it belonged to an individual or a company and other options such as air conditioner, sound system, power steering, cosmic wheels, GPS navigator all may influence the price as well.

I. RESEARCH BACKGROUND

Problem Statement:

With the extensive growth in usage of cars, the newly produced cars are unable to reach the customers for various reasons like high prices, less availability, financial incapability, and so on. Hence the used car market is escalated across the globe but in India, the used car market is in a very nascent stage and mostly dominated by the unorganized sector. This gives chance for fraud while buying a used car.

Aim of the project

The main of the project is to predict the price of used cars based on its features. We have used the used cars market dataset and fed it to the proposed model that is built using advanced machine learning learning techniques

Scope of the project

The scope of the project is limited to the compute the accuracy of the proposed model and predict the price of used cars. The admin of the system trains the proposed model with training data and tests the model with test data.

Technical approach

Below is the technical approach to address the problem:

1. Identification of dataset
2. Explorative Data Analysis
3. Cleaning the dataset and data pre-processing
4. Feeding the dataset to multiple regression algorithms and finding the best algorithm that suits the scenario
5. Training the final classifier and creating a model for the final classifier
6. Testing the final classifier and saving the results.

II. SYSTEM ANALYSIS

Data Analysis:

Range Index: 6019 entries, 0 to 6018
Data columns (total 14 columns):

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

Project target: Predict Used car price based on car specifications

1. Data cleaning
2. Feature engineering
3. Getting more insights
4. Data pre-processing
5. Modeling
6. Evaluation

Data cleaning

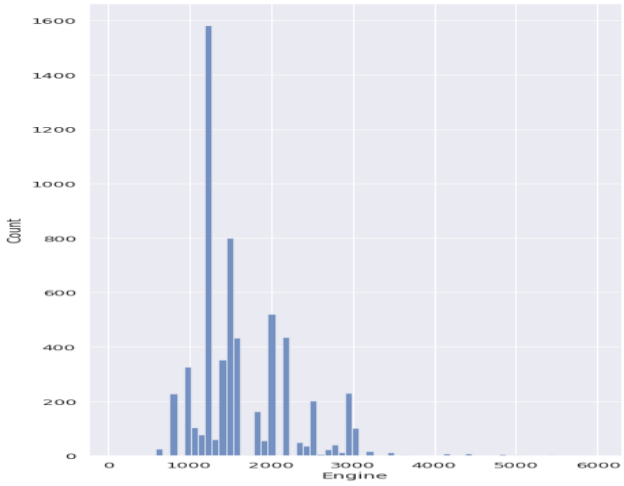
1.1 Investigation
Notes
Numerical values mixed with text in (Mileage, Engine, Power) columns.
in Milage column we have 2 units (kmpl & km/kg)
Most of New_Price Column is null values (so, we have 2 solution.)
remove it or scrap some data to fill it.
Null values values in other columns
Seats columns have some values with Zero !!
duplicated rows founded
Power column have values "null bhp"
1.2 Working with data issues
Dropping duplicated columns
Drop "New_Price" column because most of them is null
Another solution is to Scrap New price
Dropping Null values (KMPL) is refered to Kilometers Per Litre
(km/kg) is refered to kilometers Per kilogram
1 liter = 1 kilogram

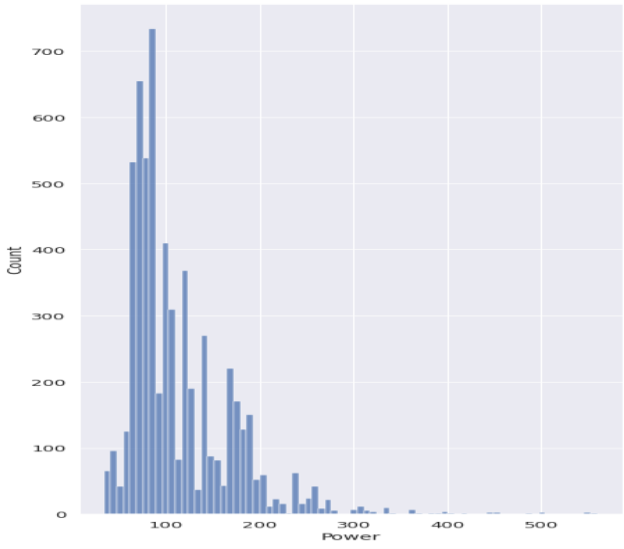
2. Feature Engineering

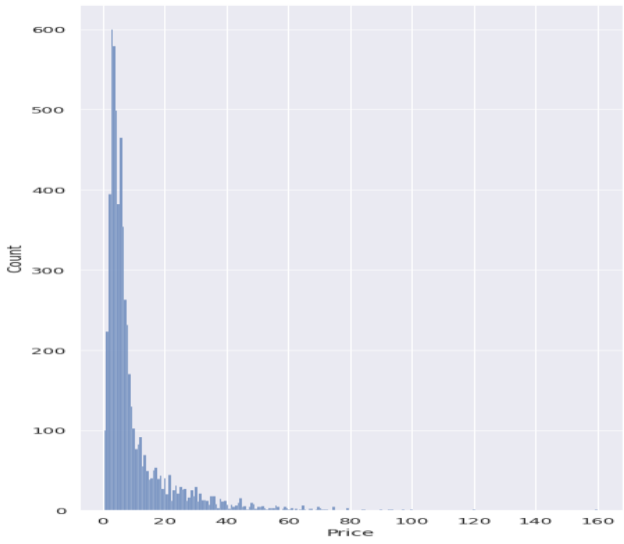
Percentage of uniques 31 %
"Name" feature has no affect that's because it has so many unique values
So let's make it useful and impactful

	Maruti Wagon R LXI CNG
	Hyundai Creta 1.6 CRDi SX Option
	Honda Jazz V
	Maruti Ertiga VDI
	Audi A4 New 2.0 TDI Multitronic
	...
6014	Maruti Swift VDI
6015	Hyundai Xcent 1.1 CRDi S
6016	Mahindra Xylo D4 BSIV
6017	Maruti Wagon R VXI
6018	Chevrolet Beat Diesel

Name: Name, Length: 5912, dtype: string
We can notice that the first word of the name is (Brand), so let's get it
A huge difference here, From this columns we can make a big affect.
Another observation that first two word can express wich car we want.
So, let's change name column with just first 2 words.

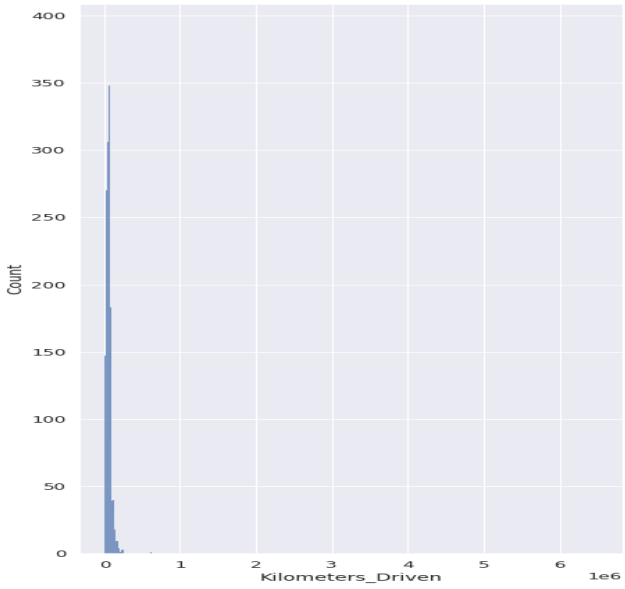


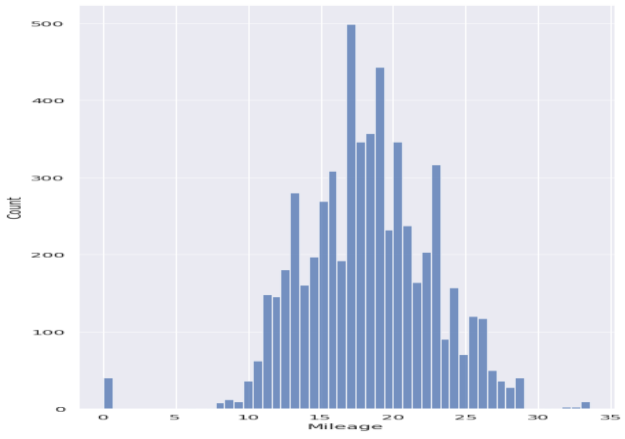




3. Data understanding

3.1 Univariate analysis





4. Data pre-processing

4.1 Data transformation
4.1.1 Catogerical transformation
4.1 Data splitting

5. Modeling

5.1 Linear regression (OLS method

5.2 Linear regression (Lasso method)
5.3 Linear regression (Ridge method)
5.4 Linear regression (ElasticNetCV method)

Proposed System:

When buying and selling cars, it can be a challenge to assign the correct price. Artificial neural networks, a branch of artificial intelligence, are frequently used for such calculations. In this study we designed two different artificial neural networks for car price forecasting and tested them using data from a car sales website. For data, a software was developed using the C# programming language and the MSSQL Server database management system, also HTMLAgilityPack library was used to read the data on the website. A procedure was written with TSQL language to digitize the data. As seen, although the number of layers and neurons increases, when the amount of data used for training decreases, the accuracy also decreases. In this study, a 91.38% success rate in car price estimation was achieved. In order to achieve more accurate results, a much larger dataset is required. Since the website blocks more than a certain number of requests for security reasons, more data could not be retrieved. Still, the predictions parallel those of a person. In the next study, it is planned to investigate the effect of the related site on the price increase by comparing the prices to be received from the people who buy and sell cars with the predictions of the developed program and what the sellers add to the website.

Advantages:

- Able to predict the prediction of used car prices accurately.

Minimize loss and maximize profits

III. ALGORITHMS AND TECHNIQUES

Root Mean Square Error (RMSE) is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are; RMSE is a measure of how spread out these residuals are. In other words, it tells you how concentrated the data is around the line of best fit. Root mean square error is commonly used in climatology, forecasting, and regression analysis to verify experimental results.

The formula is:

$$RMSE = \sqrt{\overline{(f-o)^2}}$$

Where:

- f = forecasts (expected values or unknown results),
- o = observed values (known results).

The bar above the squared differences is the mean (similar to $\bar{x}$). The same formula can be written with the following, slightly different, notation (Barnston, 1992):

$$\mathrm{RMSE}_{fo} = [\sum_{i=1}^{N} (z_{f_i} - z_{o_i})^2 / N]^{1/2}$$

- Σ = summation (“add up”)
- $(z_{fi} – Z_{oi})^2$ = differences, squared
- N = sample size.

You can use whichever formula you feel most comfortable with, as they both do the same thing. **If you don’t like formulas, you can find the RMSE by:**

1. Squaring the residuals.
2. Finding the average of the residuals.
3. Taking the square root of the result.

Linear Regression:

R-Squared Train 87.81 %

R-Squared Test 88.4 %

RMSE: 0.11027731416662538

Lasso Regression:

R-Squared Train 94.96 %

R-Squared Test 94.43 %

RMSE: 0.05299341980335

RidgeCV Regression:

R-Squared Train 95.31 %

R-Squered Test 94.16 %

RMSE: 0.05551218083455348

ElasticNet Regression:

R-Squared Train 94.98 %

R-Squared Test 94.43 %

RMSE: 0.05295312143740462

Decision Tree Regressor:

R-Squared Train 100.0 %

R-Squared Test 87.48 %

RMSE: 0.11908324518899932

Random Forest Regressor:

R-Squared Train 99.01 %

R-Squared Test 94.14 %

RMSE: 0.055704716159959776

Adaboost Regressor:

R-Squared Train 84.15 %

R-Squared Test 82.92 %

RMSE: 0.16243898099137

Gradient Boosting Regressor:

R-Squared Train 95.78 %

R-Squared Test 94.16 %

RMSE: 0.05550681960388575

Model created successfully using XGBoost Regression

R-Squared Train 99.86 %

R-Squared Test 94.49 %

RMSE: 0.05237308442027828

Creating Model:

This project helps to understand the emotional well being of the users through their posts and detect and self destructive intentions from their tweets. Timely support can curb suicides and save lives. In this work, we consequently removed casual inactive subjects from tweeter which communicating self-destructive ideations. We first abstractly assessed the dormant subjects and afterward contrasted them with chance components. Proposed models are also used to predict the urgency of the posts.

This notebook includes the following:

Project target: Predict Used car price based on car specifications

1. Data cleaning
2. Feature engineering
3. Getting more insights
4. Data pre-processing
5. Modeling
6. Evaluation

Most significant features for prediction

IV. PROJECT IMPLIMENTATION

Below is the proposed modular implementation of the project. It consists of modules:

Admin Module:

1. Login
2. Upload used cars market dataset that was downloaded from Kaggle
3. Exploratory Data Analysis
4. Data Pre-processing
    a. Transforming Categorical features using one hot encoding
    b. Normalizing numeric inputs using MinMax Scalar

5. Feeding the dataset to multiple regression algorithms
    a. Random Forest
    b. Lasso
    c. Ridge
    d. Linear regression
    e. XGBoost
    f. ElasticNet
6. Creation of model using Feed forward neural network

V. IMPLEMENTATION AND RESULT ANALYSIS

Admin Login:

This is the login page for the admin module. The admin need to login into the system with his credentials in order to perform operations like uploading the dataset, Training the dataset, Exploratory data Analysis of the dataset, Feeding the dataset to different Machine learning Algorithms to find the Algorithm that can meet the best accuracy and Create a model that can be hosted on the Flask Application to be used by the users.

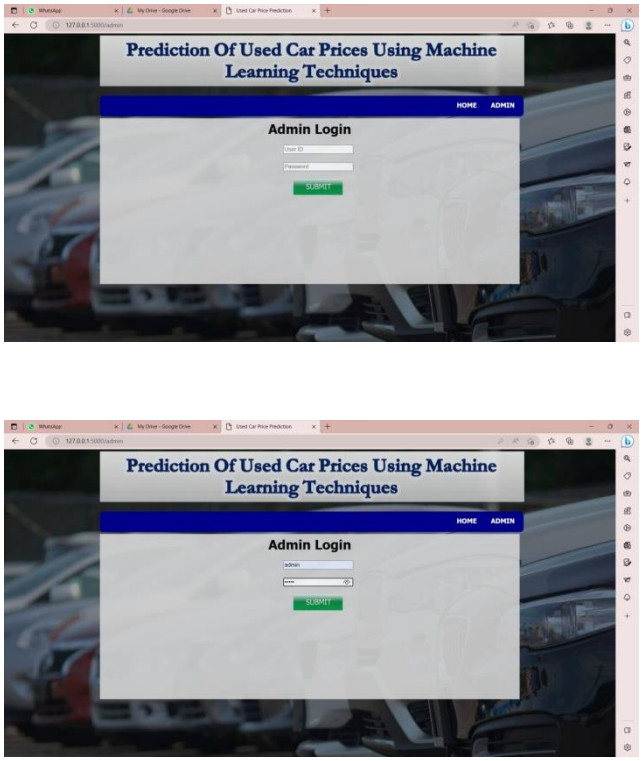


Fig: Admin Login

Upload Dataset:

On this page, the administrator of the system can upload datasets that are used for training the machine learning models. The admin has to select the file by clicking on the Choose file button and click on the upload button to upload the file to the server. Once the upload is complete, a success message would be displayed that the file is successfully uploaded. For this project we are using train-data.csv as a dataset.

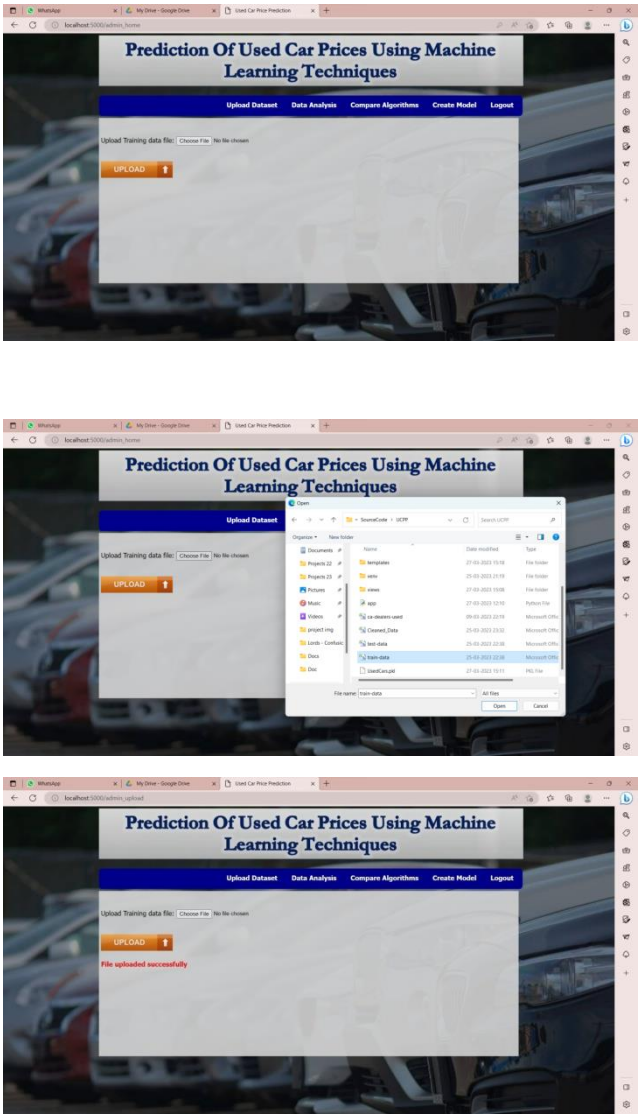


Fig: Upload Dataset & File Uploaded Successfully.

Data Analysis:

Exploratory Data Analysis is performed on the dataset in order to clean the dataset for any missing data, identify patterns, identify the

relationships of various parameters of the outputs with the help of graphs, statistics etc.

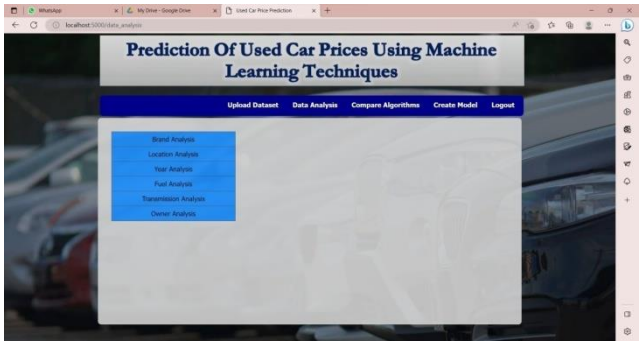


Fig: Data Analysis

Brand Analysis:

The below graph shows the Brand Analysis over data present in the dataset.

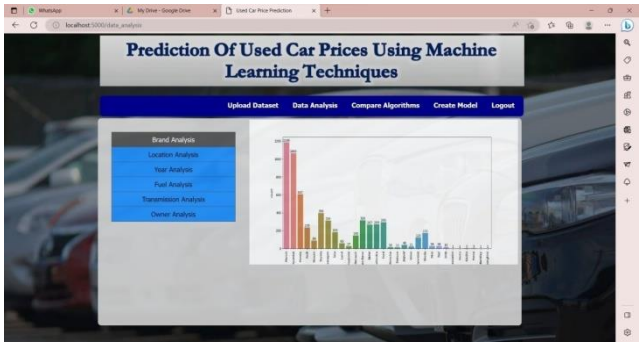


Fig: Brand Analysis

Location Analysis:

The below graph shows the Location Analysis over data present in the dataset.

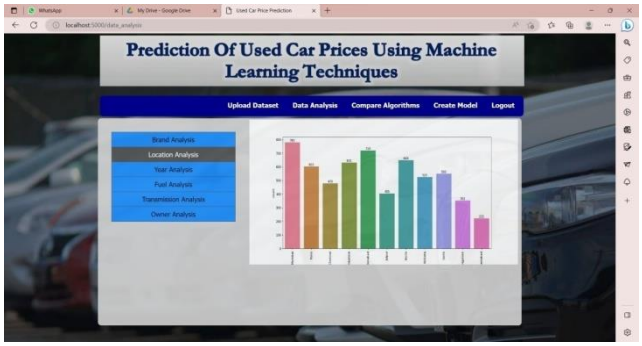


Fig: Location Analysis

Year Analysis:

The below graph shows the Year Analysis over data present in the dataset.

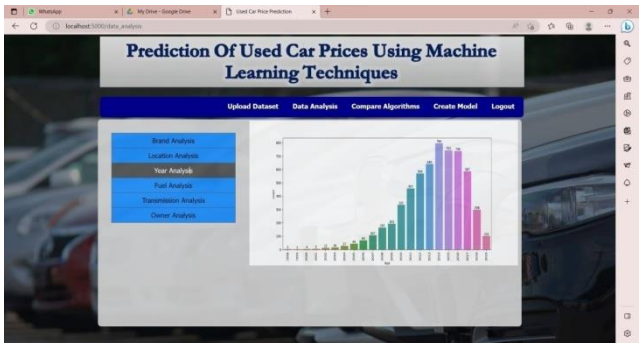


Fig: Year Analysis

Fuel Analysis:

The below graph shows the Fuel Analysis over data present in the dataset.

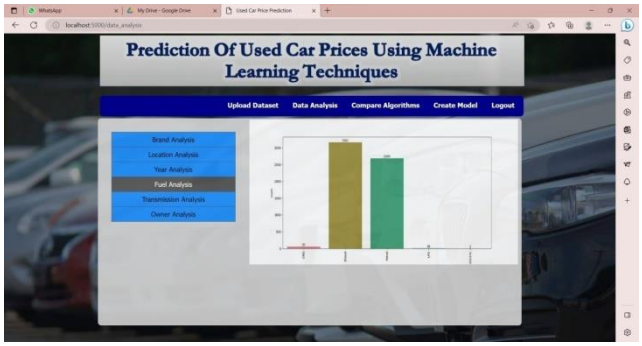


Fig: Fuel Analysis

Transmission Analysis:

The below graph shows the Transmission Analysis over data present in the dataset.

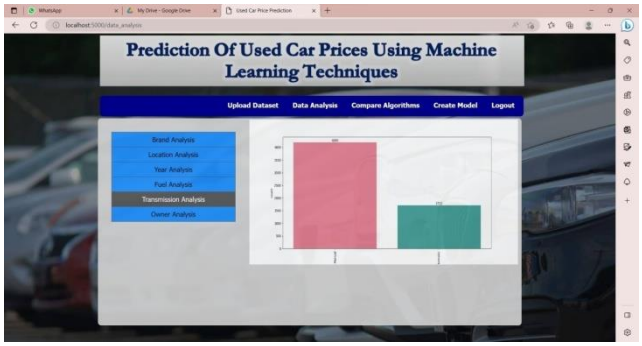


Fig Transmission Analysis

Owner Analysis:

The below graph shows the Owner Analysis over data present in the dataset.

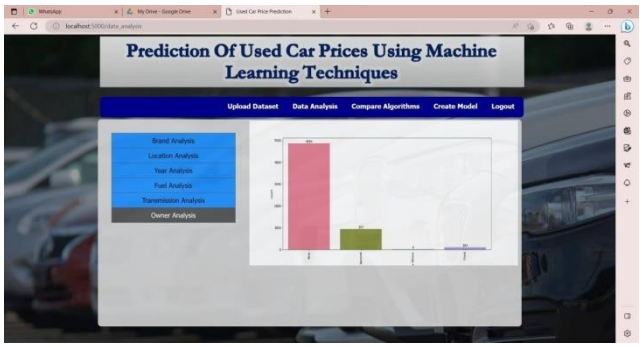
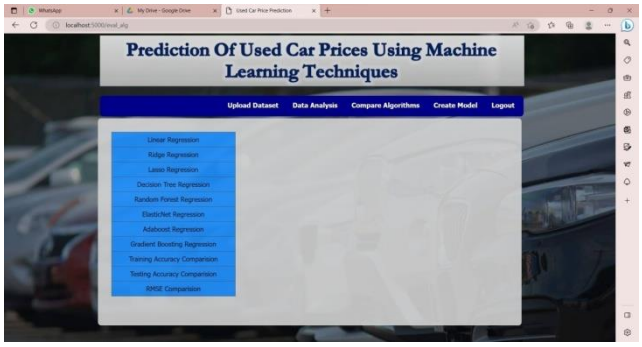


Fig: Owner Analysis

Compare Algorithms:

On this page, the admin can feed the dataset to various Algorithms to train them and get the test accuracy for each algorithm. When the dataset is feed to various algorithms to evaluate the situation with some parameters like Train Accuracy, Test Accuracy, Root Mean Square Error …



Linear Regression Algorithm:

When the dataset is feed to Logistic Regression algorithm we observe that the Training Accuracy of the Model is 87.81% and Test Accuracy of the Model is 88.4% and Root Mean Square Error is 0.11027.

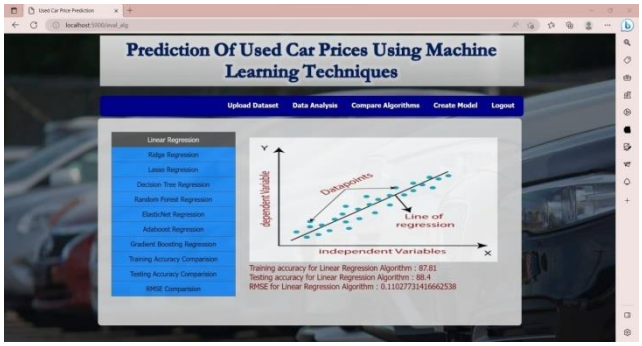


Fig: Logistic Regression

Ridge Regression:

When the dataset is feed to Ridge Regression algorithm we observe that the Training Accuracy of the Model is 95.31% and Test Accuracy of the Model is 94.16% and Root Mean Square Error is 0.0555.

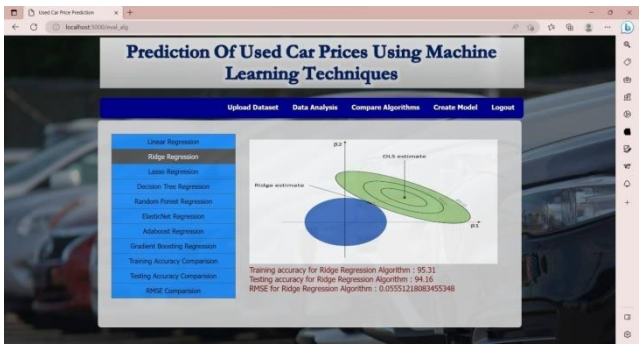


Fig: Ridge Regression

Lasso Regression:

When the dataset is feed to Lasso Regression algorithm we observe that the Training Accuracy of the Model is 94.96% and Test Accuracy of the Model is 94.43% and Root Mean Square Error is 0.05299.

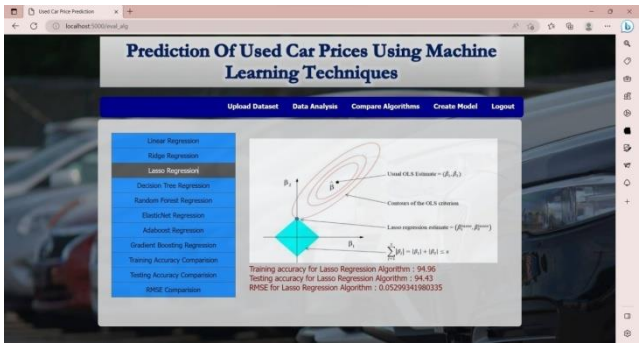


Fig: Lasso Regression

Decision Tree Regression:

When the dataset is feed to Decision Tree Regression algorithm we observe that the Training Accuracy of the Model is 100% and Test Accuracy of the Model is 87.48% and Root Mean Square Error is 0.11908.

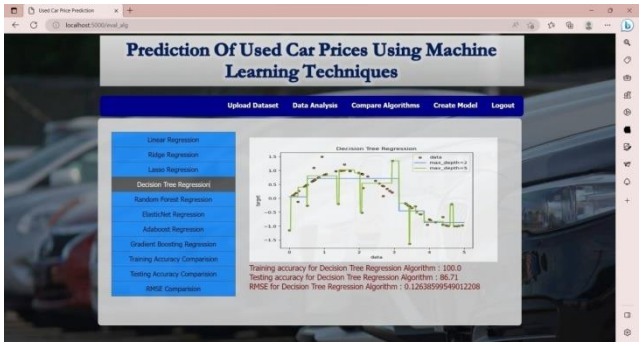


Fig: Decision Tree Regression

Random Forest Regression:

When the dataset is feed to Random Forest Regression algorithm we observe that the Training Accuracy of the Model is 99.01% and Test Accuracy of the Model is 94.14% and Root Mean Square Error is 0.0557.

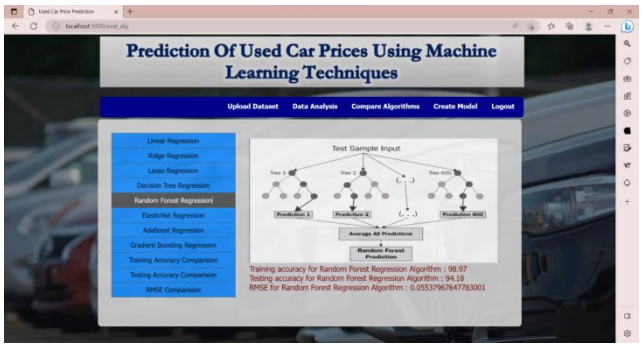


Fig: Random Forest Regression

ElasticNet Regression:

When the dataset is feed to ElasticNet Regression algorithm we observe that the Training Accuracy of the Model is 94.98% and Test Accuracy of the Model is 94.43% and Root Mean Square Error is 0.0529.

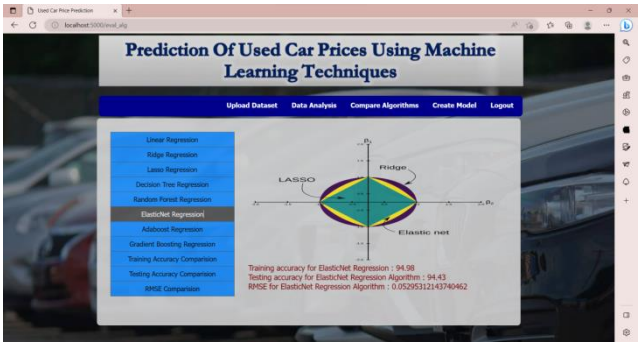


Fig: ElasticNet Regression

Adaboost Regression:

When the dataset is feed to Adaboost Regression algorithm we observe that the Training Accuracy of the Model is 84.15% and Test Accuracy of the Model is 82.92% and Root Mean Square Error is 0.1624.

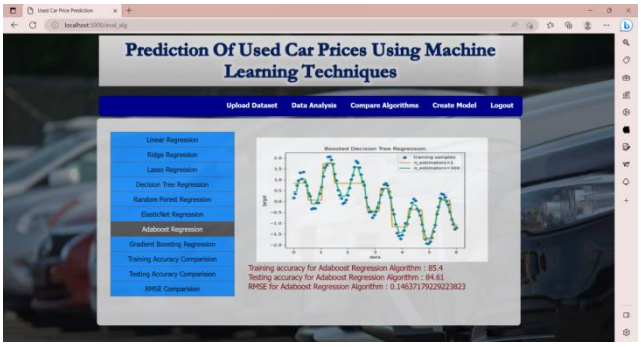


Fig: Adaboost Regression

Gradient Boosting Regression:

When the dataset is feed to Gradient Boosting Regression algorithm we observe that the Training Accuracy of the Model is 95.78% and Test Accuracy of the Model 94.16% and Root Mean Square Error is 0.0555.

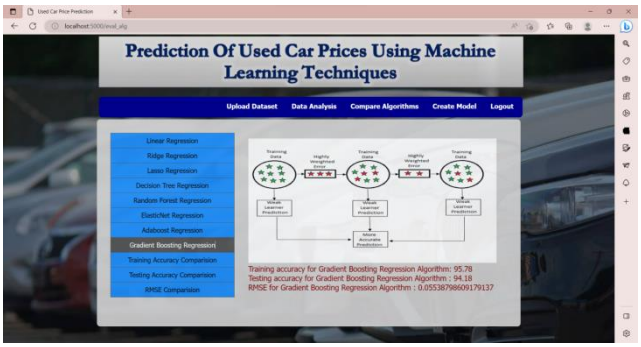


Fig: Gradient Boosting Regression

Training Accuracy Comparison:

When the dataset is feed to Comparison algorithms we observe that the Training Accuracies of all.

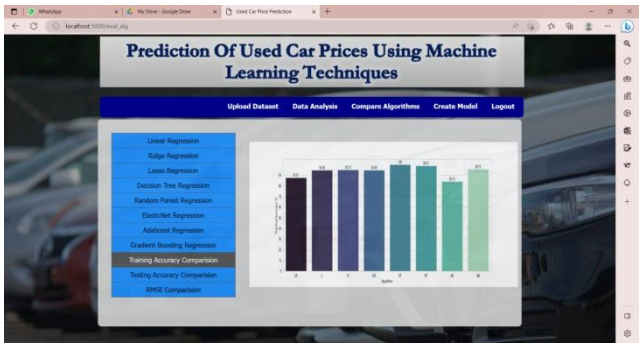


Fig: Training Accuracy Comparison

Test Accuracy Comparison:

When the dataset is feed to Comparison algorithms we observe that the Test Accuracies of all.

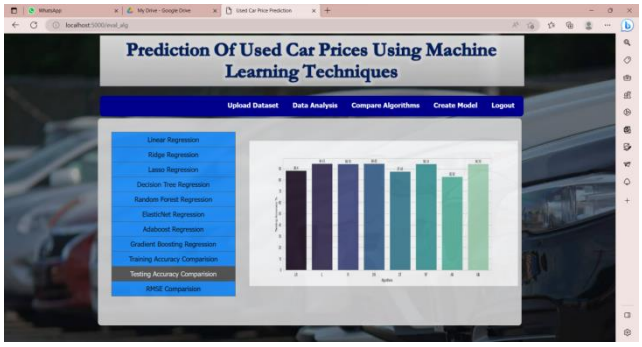


Fig: Test Accuracy Comparison

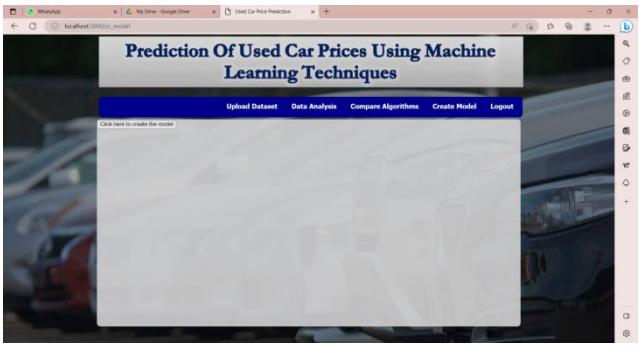
RMSE Comparison:

When the dataset is feed to Comparison algorithms we observe that the Root Mean Square Error of all.

Fig: RMSE Comparison

Create Model:

This screen shows the Training Accuracy of the Model is 99.86% and Test Accuracy of the Model is 94.45% and Root Mean Square Error is 0.05237.



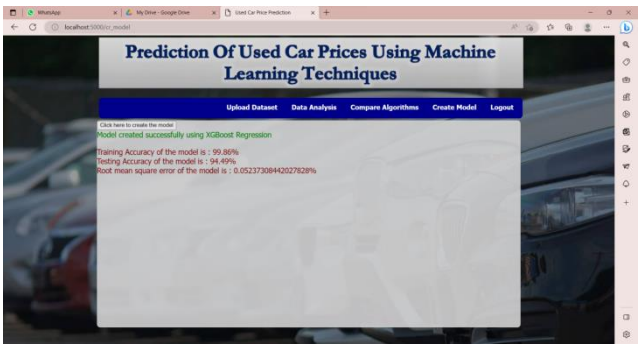


Fig: Create Model

VI. CONCLUSION

Predicting used car prices is a difficult task due to the large number of features and parameters that must be examined in order to get reliable findings. The first and most important phase is data collection and preprocessing. A Supervised learning-based XGBoost Regression model are developed which can learn from the car dataset provided to it. This project presents a working model for used car price prediction with a low error value. A considerable number of distinct attributes are examined for reliable and accurate predictions. The results obtained agree with theoretical predictions and have shown improvement over models which use simple linear models. The XGBoost Regression algorithm is used to build a model, and for comparison, other machine learning algorithms such as Linear Regression, Random Forest Regression, Lasso Regression, Ridge Regression, Decision Tree Regression, ElasticNet Regression, Adaboost Regression, and Gradient Boosting Regressions are also built. The automobile dataset is used to test these techniques. According to experimental findings, the XGBoost Regression model, which has the lowest error among all other methods, has a training accuracy of the model of 99.86%, a test accuracy of the model of 94.45%, and a root mean square error of 0.05237. The work reported here has significant ramifications for future research on the XGBoost Regression model for Used Car Price Prediction, and it may one day contribute to a 100% accurate solution to the fraud problem.

Future scope: The created machine learning model can be deployed as an open source, ready-to-use price prediction model and exported as a "Python class" so that it can be quickly integrated with other websites. By adopting deep learning network topologies, adjustable learning rates, and training on data clusters rather than the complete dataset, the model can be significantly improved using neural networks.

References

[1] Vehicle Price Prediction using SVM Techniques S.E.Viswapriya Durbaka Sai Sandeep Sharma Gandavarapu Sathya kiran International Journal of Innovative Technology and Exploring Engineering (IJITEE), vol. 9, no. 8, June 2020, ISSN 2278–3075.

[2] Pandey Abhishek, Rastogi Vanshika and Singh Sanika, Car's Selling Price Prediction using Random Forest Machine Learning Algorithm, March 2020.

[3] Ganesh Mukkesh and Venkatasubbu Pattabiraman, "Used Cars Price Prediction using Supervised Learning Techniques", International Journal of Engineering and Advanced Technology, vol. 9, pp. 216-223, 2019.

[4] Pubushed in International Journal of Trend inScientificResearch and Development (ijtsrd), vol. 5, no. 4, 2021, ISSN 2456-6470.

[5] [online] Available: https://www.kaggle.com!datasets.
Show in Context

[6] N. Monburinon, P. Chertchom, T. Kaewkiriya, S. Rungpheung, S. Buya and P. Boonpou, "Prediction of prices for used car by using regression models", 2018 5th International Conference on Business and Industrial Research (ICBIR), pp. 115-119, 2018.

[7] N. Sun, H. Bai, Y. Geng and H. Shi, "Price evaluation model in second-hand car system based on BP neural network theory", 2017 18th IEEE/ACIS International Conference on Software Engineering Artificial Intelligence Networking and Parallel/Distributed Computing (SNPD), pp. 431-436, 2017.

[8] C. V. Narayana, C. L. Likhitha, S. Bademiya and K. Kusumanjali, "Machine Learning Techniques To Predict The Price Of Used Cars: Predictive Analytics in Retail Business", 2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC), pp. 1680-1687, 2021.