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:09/September-2022 Impact Factor- 6.752 www.irjmets.com

NOVEL APPROACH OF MACHINE LEARNING ALGORITHMS IN CAR DATASET

Jency M. Shah*1, Dr. Ronak Panchal*2

*1Researcher, R.N.G.P.I.T, India.

*2Adjunct Professor, R.N.G.P.I.T, India.

ABSTRACT

In this work, we have provided various machine learning algorithms and using one of them we have predicted price of car in given dataset. Apart from this we have done Exploratory Data Analysis in which we have done customer segmentation along with various charts on relations between various fields. The given dataset was cleaned first during data pre-processing. The pie chart of Legal and Illegal cars indicates the number of cars whose emission norm is below BS 6 which is banned by the government. We have used Linear Regression for Price Prediction.

Keywords: Machine Learning Algorithms, Price Prediction, Exploratory Data Analysis.

I. INTRODUCTION

India is one of the leading countries in field of Automotive Industry. Automotive industry contributes around 8% of country's total export and in India's GDP. In today's world, everyone wants to own a car so, there is a need of automatic system which predicts the price of the car. This system will reduce the task and will make it easy for the buyers to search for a car under their budget. The price of a car is affected by various features like Car Model, its gas type, its emission norm, its type and many more. This project focuses on prediction of price using some of the features and doing Exploratory analysis of given dataset. It will help the individual and will give them better insights of this field. Also, this is an interesting topic for research and we hope to achieve significant growth.

Overview of Machine Learning Algorithms

Artificial Intelligence is a technology through which we can create intelligent machine which can work like a human brain. It can think like a human, behave like a human and can also make decisions. The top five languages used for AI are Python, Java, Lisp, R, Prolog. Machine Learning is a subset of AI and AI uses different machine learning algorithms to achieve intelligence. Machine Learning algorithms are divided into three categories:- supervised learning, unsupervised learning and reinforcement learning. The algorithms included are Linear Regression, Logistic Regression, Naïve Bayes, Support Vector Machine, K-Nearest Neighbours, Decision Tree, Random Forest and many more.

Table 1: Difference between Supervised and Unsupervised Learning

Supervised Learning	UnSupervised Learning
Supervised Learning can be used for 2 different types of problems i.e. regression and classification	It can be used for two different types of problems like clustering and association
Here, input and output data is provided.	Here, only input data is provided.
Here, the data is labelled.	Here, the data is not labelled.
Here, results are more accurate.	Here, the results are less accurate.
Here, feedback is also accepted.	Here, no feedback is taken.

Supervised Learning

Supervised Learning as the name suggest includes a supervisor and is a type of learning in which machine needs an external supervision to learn from data. Here we teach the machines to use label data to produce correct outcome. It is divided into two categories named Classification and Regression. The complexity of supervised learning is less than unsupervised learning. Also, it is more accurate than unsupervised learning. Classification and Regression both are used for prediction in machine learning. But the basic difference is Regression algorithm uses continuous values for prediction whereas Classification algorithms uses discrete

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:09/September-2022 Impact Factor- 6.752 www.irjmets.com

values for the same. The task of classification algorithms is to link the input variable with discrete output variable whereas regression will link input variable with continuous output variable. Classification algorithms are used to solve the problems like speech recognition, identification of cancer cells etc. whereas regression algorithms are used to solve problems like price prediction, weather prediction etc.

Linear Regression

Linear Regression is used to find a linear relationship between dependent and independent variables and for predictive analysis. Linear Regression has many practical uses due to its simplicity and various properties. The equation of Linear Regression is Y = A + B*X

Where

Y is the dependent variable

A is the intercept

B is the coefficient of regression

X is the independent variable.

Once we find the best A and B we will get the best fit line. So when we are finally using our model for prediction, it will predict the value of Y for the input value of X.

Logistic Regression

Logistic Regression is one of the most popular machine learning algorithms used to predict categorical dependent values using independent variables. So, the outcome must be categorical or discrete values. It can be either 0 or 1, yes or no, True or False etc. The graph of logistic regression is an 'S' shaped curve known as sigmoid curve. It predicts two maximum values as 0 or 1.

Logistic Regression comes under Classification when a decision threshold is brought into picture. Logistic Regression is also divided into three categories named Binomial, Multinomial and Ordinal.

Random Forest

Random Forest is an inseparable part of Machine Learning. It can be both Classification and Regression. It is a process of combining classifiers to solve a complex problem and to improve the model. Instead of relying on one decision tree, random forest takes the decision from every tree and based on that it makes the predictions. The greater number of trees lead to higher accuracy. Random forest is used widely because it takes less training time as compared to others, it predicts with high accuracy and it can maintain its accuracy even for large proportion of data.

II. RELATED WORKS

Several related works have been done previously to predict car prices using different approaches. Several machine learning algorithms like Decision tree, naïve bayes, K- nearest, multiple regression have been used to predict price. We will try to develop statistical model that can predict the price of car. To accurately anticipate the price of the car, many different approaches have been used. In [4] [5] [6] [7] and [3] all of the solutions took into distinct attributes while training the model.

[2]To anticipate the price of vehicle , Noor and Jan used multiple Linear Regression. Only few variables are included in data, which were used to create the model and the R-square score was 98% and the outcome was outstanding.

[1]Listiani, 2009 used Support vector machines to evaluate car prices and results showed that it is far more accurate in large dataset than multiple linear regression. Also, it includes samples up to 178 attributes which is very large than our study so, Linear regression would be better in our case.

[3] Kupier, 2008 collected data from general motor of cars that are produced where he used variable selection method in his model and it reduced the complexity of the model.

III. WORKFLOW

The project deals with Indian cars. After the dataset was collected the dataset was pre-processed and we removed null values in some cases we filled null values with mean or mode. We removed unwanted columns and prepared a clean car dataset on which we did exploratory data analysis and obtained various graphs on relations between various fields.

We obtained a heatmap, bar graph, pie charts, box plots etc. using python library named seaborn and matplotlib.

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:09/September-2022 Impact Factor- 6.752 www.irjmets.com

Data understanding and exploring is very important step of model building. As it gives insights of data and what fields are interconnected and can be used to produce results. It helps to correlate between various fields. Also, it helps us to understand the fields which are going to be used in price prediction.

Afterwards, when data is cleaned and transferred to a clean dataset, we can use that dataset to produce our results. After the study of models, we decided to go with the linear regression algorithm to built our model and we were successful in making one model. After our model is built we tested it and deployed it. Below figure gives the proper workflow in which we tried to work and our model was built.

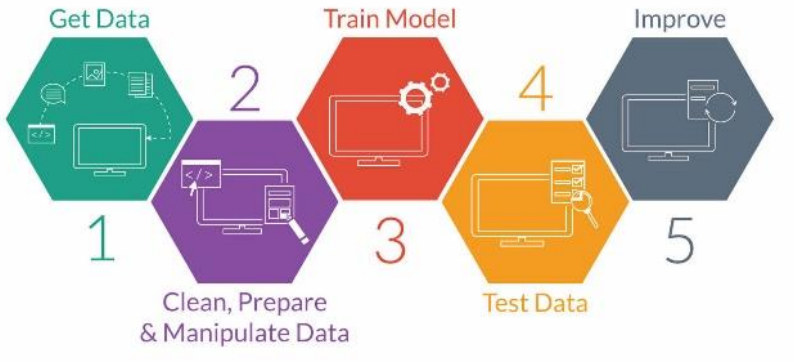

Figure 1: WorkFlow

IV. EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis is very essential step in building any model. To start working on any data we first need to understand the data and EDA is the step to learn and understand about the data. It includes steps like Data preparation, Data Pre-processing, Data description and Data analysis.

Data Preparation

This step includes collection of dataset and understanding the data. It is the essential step before data modelling. The main task is to clean the data and to ensure the quality and correctness of data. We would make changes to data to make it more useful and to remove unwanted data so that it would not create any further problems during model building.

Data Pre-Processing

Analysis is must for any data to process further. It gives us insights of the data so that we can use the information and can produce various useful results. Data analysis is performed before building the model. This type of analysis is important because it helps us to understand the data better so can find the correlations between various fields and can plot them via graphs and can even use the relations in our model building.

Dataset Description

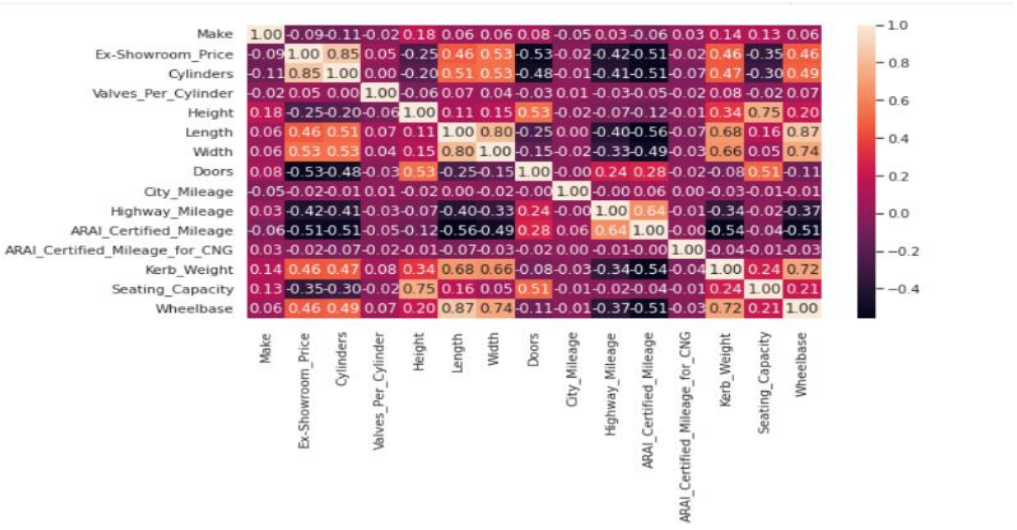

Figure 2: HeatMap

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:09/September-2022 Impact Factor- 6.752 www.irjmets.com

Our car dataset contains fields like Company, Variant, Model, Showroom price, Displacement, body type, fuel-type, cylinders, doors, mileage, horsepower etc. columns. The heatmap shows that Kerb-weight is positively correlated and city-mileage and highway mileage are negatively co-related.

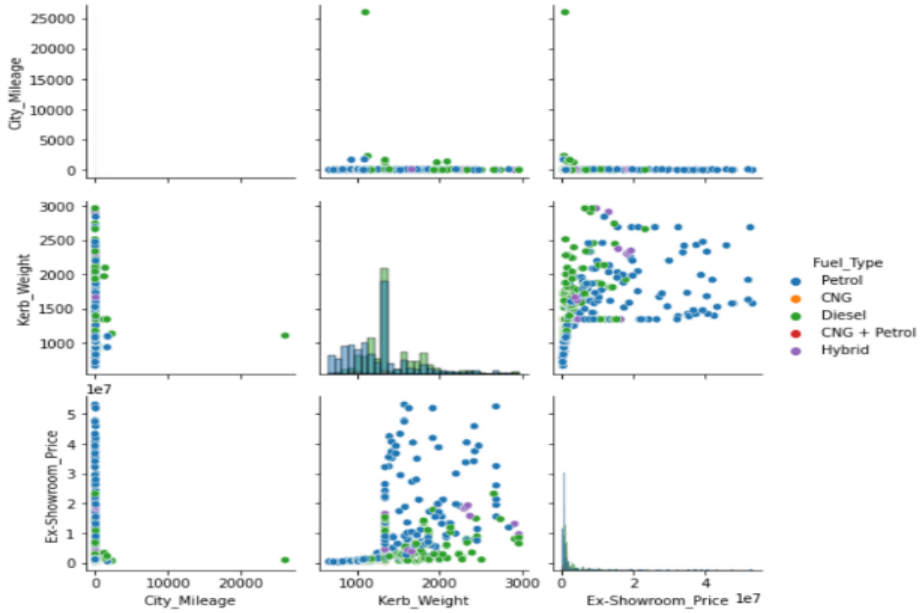

Figure 3: Mix-Graph

The mix graph shows that Vehicle mileage decreases as kerb-weight increases and the kerb-weight is positively corelated with price. It is plotted between three parameters named Kerb Weight, Showroom price and City Mileage.

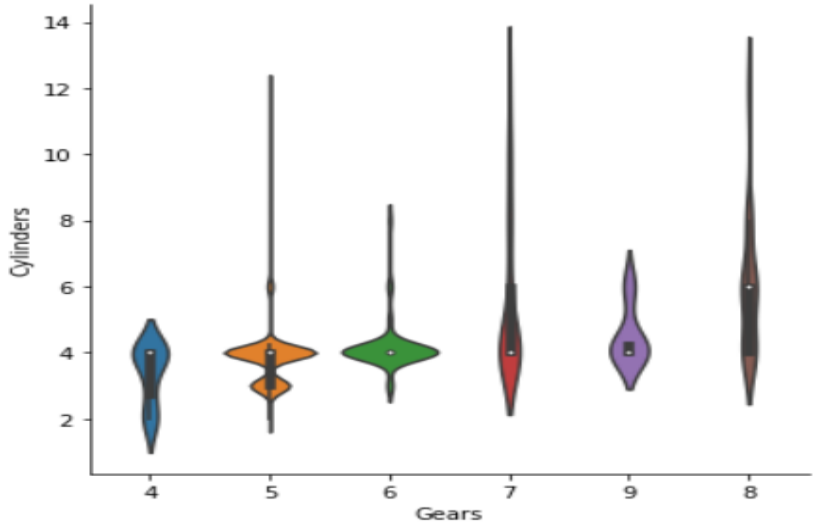

Figure 4: Violin Graph

This Violin plot is plotted between quantities named Cylinders and Gears. It shows that more number of cylinders will lead to more number of gears.

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:09/September-2022 Impact Factor- 6.752 www.irjmets.com

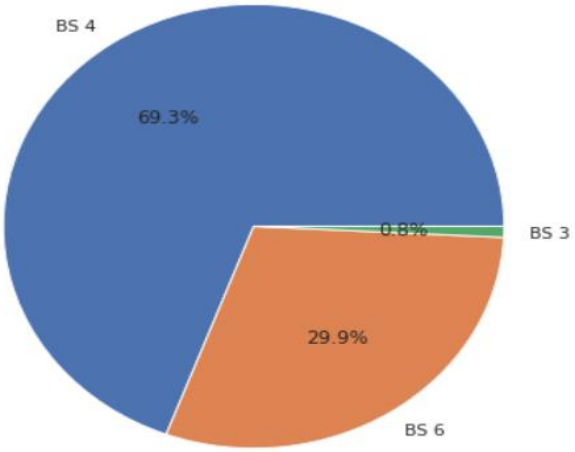

Figure 5: Pie-Chart

This is the pie-chart of the Emission Norms, According to above value counts there are 784 vehicles with BS 4 emission and 9 BS 3 emission. So, these vehicles will be considered illegal as the government has banned the vehicles below BS6 emission norms and hence, it's time for them to launch a new variant with BS 6 emission norm.

V. EXPERIMENTATION

Data exploration is done on the dataset and we have got insights of the data and have understood various columns and relation between them. We have understood various features used for price prediction. We have collected the data then have worked on the data and have cleaned it. After cleaning we did the Exploratory analysis on the data and found some important insights which can be used for model building. And build the model using linear regression algorithm.

Model Building

After cleaning and doing exploratory data analysis on the dataset we started building our model. The Linear regression model was imported using SKlearn library of python language and fitted the pipeline. After that we divided the data into training and testing and firstly we will fit the pipeline of training data. After that we will predict the Y and then we will fit the testing pipeline. We found the accuracy of our model using r2 score and it turned out to be 98% accurate.

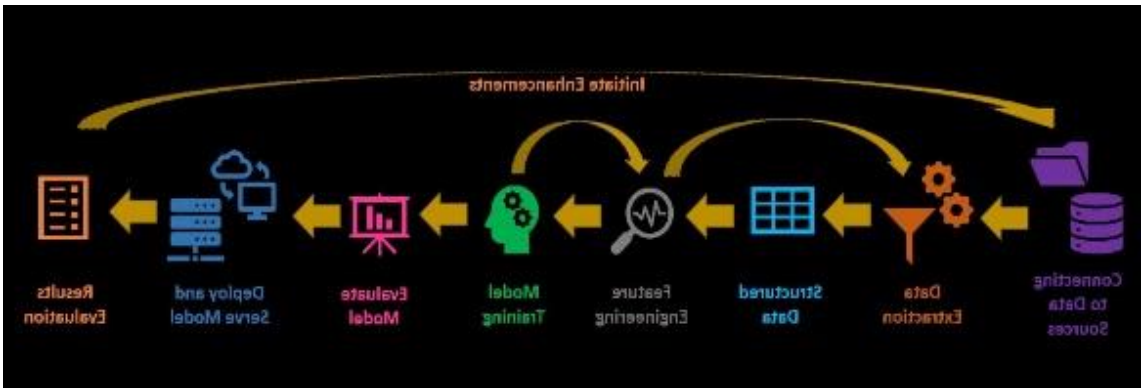

Figure 6: Model Building

Hence that was the technique we used to build the model and the above figure shows the methodology we used to make this model.

After model building the deployment stage comes. We made a website and attached our regression model to that website. We deployed our website on an app named Streamlit. Deployment is very important step of model building. After deploying the model it can be used by everyone as it will get deployed on server. The model can be used in real world only by deploying it. We can also make updates in our website by adding features to it so that people can use it and can get benefit of it.

Hence, that was all about model building and deployment.

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:09/September-2022 Impact Factor- 6.752 www.irjmets.com

VI. CONCLUSION

Using Data analysis and machine learning algorithm this project proposed a scalable model for price prediction in India. An efficient machine learning algorithm can be made only by training, testing and evaluating various algorithms. Each experiment was performed using Google colab environment as it took less training time in google colab.

In future, more data will be collected and we can train it using different algorithms. We can also convert our project into mobile based application with more features which can be used by a large number of users.

Hence, we would be having a real time processing program.

ACKNOWLEDGEMENTS

The dataset and the python Google Colab file are uploaded here:
https://github.com/Jencyshah1211/Microsoft-engage-2022

VII. REFERENCES

[1] Listiani, M. (2009). Support vector regression analysis for price prediction in a car leasing application. Master thesis. Hamburg: Hamburg University of technology.

[2] Noor, K., & Jan,S. (2017). Vehicle Price prediction System using machine Learning techniques. International Journal of Computer Applications, 27-31.

[3] Kupier, S. (2008). Introduction to multiple regression: how much is your car worth? Journal of statistics education.

[4] Comparative analysis of used car price evaluation models, Tongji University, Shanghai 200000, China.

[5] Nitis Monburinon, "Predictions of prices for used car by Regression models", 5th International Conference on Business and Industrial Research, Bangkok, Thailand,2018.

[6] Nabarun Pal, "A methodology for predicting used cars prices using Random Forest", Future of Information and Communications Conference,2018.

[7] Jaideep A Muley, "Prediction of used cars Prices using SAS EM", Oklahoma State University.