# Used car price prediction

# Machine Learning Algorithms for Used Car Price Prediction

# A

# Project

# For the award of

# Diploma in Advanced computing (PG-DBDA)

# In

# CDAC

****

**Submitted By: Guided By-**

Prajakta Waykos [220310125014] Mr. Pappu Kapgate

Balaji Kamble [2220310125014]

Mansi Patil [2203101225030]

Submitted to

**ACTS-CDAC NEW DELHI**

**SESSION MARCH 2022**

**ACTS-CDAC NEW DELHI**

****

**CERTIFICATE**

This is to certify that Prajakta Waykos, Balaji Kamble and Mansi Patil of PG-DBDA have completed their project entitled “**USED CAR PRICE PREDICTION**” during the academic session March 2022 under the guidance and supervision of Mr. Paapu Kapgate

We approve the project for the submission for the partial fulfilment of the requirement for the award of diploma of DBDA in ACTS-CDAC.

**Mr. PAPPU KAPGATE Mr. Ankit Khurana** Project Co-ordinator Co-Ordinator PG-DAC Acts-CDAC New Delhi Acts-CDAC New Delhi

# LIST OF CONTENTS

[1 ABSTRACT 2](file:///D:\PROJ\3.+List+of+Tables.docx#_Toc499487479)

[2 INTRODUCTION 4](file:///D:\PROJ\3.+List+of+Tables.docx#_Toc499487480)

[3 EXECUTIVE SUMMARY](file:///D:\PROJ\3.+List+of+Tables.docx#_Toc499487481) 5

[4 PROBLEM STATEMENT 7](file:///D:\PROJ\3.+List+of+Tables.docx#_Toc499487482)

[5 DATASET ANALYSIS 11](file:///D:\PROJ\3.+List+of+Tables.docx#_Toc499487483)

[6 DATA PREPROCESSING 12](file:///D:\PROJ\3.+List+of+Tables.docx#_Toc499487484)

[7 EXPLORATORY DATA ANALSIS 19](file:///D:\PROJ\3.+List+of+Tables.docx#_Toc499487485)

[8 METODOLOGY 20](file:///D:\PROJ\3.+List+of+Tables.docx#_Toc499487486)

9 PROPOSED SYSTEM 24

10 ASSUMPTIONS FOR ALGORTHM 26

11 FUTURE WORK 27

12 CONCLUSIONS 28

# LIST OF FIGURES

[1 Price Box plot 2](file:///D:\PROJ\3.+List+of+Tables.docx#_Toc499487479) [2 2 Odometer box plot 16](file:///D:\PROJ\3.+List+of+Tables.docx#_Toc499487480)

[3 Count brand](file:///D:\PROJ\3.+List+of+Tables.docx#_Toc499487481)  18

[4 Count type of vehicle 18](file:///D:\PROJ\3.+List+of+Tables.docx#_Toc499487482)

[5 Condition 19](file:///D:\PROJ\3.+List+of+Tables.docx#_Toc499487483)

[6 Transmission Type 19](file:///D:\PROJ\3.+List+of+Tables.docx#_Toc499487484)

[7 Drive 19](file:///D:\PROJ\3.+List+of+Tables.docx#_Toc499487485)

[8 Fuel VS No of Cars 20](file:///D:\PROJ\3.+List+of+Tables.docx#_Toc499487486)

9 No of Cars VS Manufacturer 21

10 Linear Regression 23

11 Ridge Regression 25

12 Random Forest 26

13 XGBoost 27

ACKNOWLEDGEMENT

First and foremost, we praise and thank almighty God for showering his perennial blessing on us and for giving us the courage all through the execution of our work.

At the outset, we are indebted to our Mentor Mr. PAPPU KAPGATE for his time, valuable inputs and guidance. His experience, support and structured thought process guided us to be on the right track towards completion of this project.

We are extremely gifted and fortunate to have Ms. Ankit Khurana sir as our Academic counsellor. Her in-depth knowledge coupled with her passion in delivering the subjects in a lucid manner has helped us a lot. We are thankful to her for her guidance towards entire coursework.

We also thank all the course faculty of the DSE program for providing us a strong foundation in various concepts of analytics & machine learning.

Last but not the least, we would like to sincerely thank our respective families for giving us the necessary support, space and time to complete this project.

We certify that the work done by us for conceptualizing and completing this project is original and authentic.

Date: 25th sept, 2022

Place: New Delhi

1. Abstract

As the automobile industry is booming, the demand for new cars is increasing. With the demand for new cars, used cars have also gained tremendous value. Estimating the price of these used cars has become crucial by which we can provide the actual value of the car to the buyers and also to the sellers for transparency.

2. INTRODUCTION

The used car market in 2019 is valued at $ 828.24 billion and is expected to be worth $ 1.355 billion by 2027. During the 2013-2018 evaluation period, India's used car market experienced double-digit CAGR revenue growth. The market was in the early stages of expansion due to higher vehicle replacement rates, faster time to market for new cars, growth in the middle class, higher average ticket sizes, and lower import taxes on new cars. Older cars are preferred in India because new cars are expensive and popular among people in the socio-economic group below the middle class. Also, renting a car is becoming a widespread practice, and rather than purchasing a car outright, you can do so by making fixed monthly payments for a predetermined number of months. You have the option to purchase the vehicle at the residual value, or the anticipated selling price, at the end of the lease. To make it easier for people to buy a car, we have developed a predictive model that predicts the price of used cars as demand increases.

3. EXECUTIVE SUMMARY

3.1.1 Background and need for study:

The 17th century industrial revolution started from the British colony. Automobile industry started developing and increasing demand. As more manufacturing competition increased according to demand. But as technology gets developed, old transportation gets avoided by the consumers.

Used car prices are based on Brand, Manufacturer and Transmission type. This process done by professionals who understand the condition and market demand of cars. Condition of cars depends on the engine, body of cars, average and condition of cars. But due to the increase in price of new cars and the inability of customers to buy new cars due to lack of funds, used car sales are on a global increase. There is a need for a used car price prediction system to effectively determine the worthiness of the car using a variety of features.

3.1.2 Scope and objectives of the study:

The main aim of this project is to determine the price of used cars using professional evaluation. Predict the price of a used car using various machine Learning Models. This helps the customer to make decisions based on different input or factors such as Price, brand, Condition of vehicles.

The objectives of the research are outlined as follows:

1. To investigate the modern approach to determine the used car price.
2. To predict human capitalism with high accuracy using machine learning models.
3. To evaluate performance of various supervised models.
4. To improve automobile economy.
5. To determine the most favorable technique to estimate the parameters for learning models.

3.1.4 Approach and Methodology:

The data is scraped from used car listing on Ebay-Kleinanzeigen (German), retrieved from Kaggle. After processing the dataset and cleaning the inconsistencies, the numerical and categorical features used in the purchasing intention prediction model is generated. Various Regression algorithms are used to predict used car price based on set of independent variables like brand, model, kilometre, powers, Fuel Type, Vehicle Type, etc. The predictive models are also used to identify the variables that strongly influence the price using variable importance and probabilistic approaches. The models are evaluated using relevant model performance measures to arrive at the most robust models for prediction.

3.1.5 Key Learnings:

The data obtained from Kaggle convey important information about the features of a used car, and how they influence its market price.

3.1.6 Recommendations & actionable insights:

The high-level recommendations for the project are developed by predicting used car prices. These are then linked to the model findings to recommend actionable insights.

4. PROBLEM STATEMENT

The challenging part for used car dealers is to predict the price of used cars with accuracy. The prices of new cars in the industry are fixed by the manufacturer with some additional costs incurred by the Government in the form of taxes. So, customers buying a new car can be assured of the money they invest to be worthy. But due to the increased price of new cars and the incapability of customers to buy new cars due to the lack of funds, used cars sales are on a global increase. There is a need for a used car price prediction system to effectively determine the worthiness of the car using a variety of features. Even though there are websites that offers this service, their prediction method may not be the best. Besides, different models and systems may contribute on predicting power for a used car’s actual market value. It is important to know their actual market value while both buying and selling

4.1.1. Need of Study

The increase in e-commerce usage over the past few years has created potential in the used car market, enabling the used car dealers to reach out to a huge customer base. To be able to predict used cars market value can help both buyers and dealers (sellers in the organized used car market).

4.1.2. Target Business:

Online User-Car Dealers They are one of the biggest target groups that can be interested in results of this study. If used car dealers better understand what makes a car desirable, what the important features are for a used car, then they may consider this knowledge and market products with better price. E-commerce is being preferred over other businesses as the proportion of revenue investment, which is apparently better than other options and requires less investment.

4.1.3. Target Customers

Buyers who would like to purchase used cars via online portal, wherein, it’s a big corner to pay too much or sell less then it’s market value

4.1.4. Scope of Study

The organized sector is dominating the German used car market due to higher network chains with a record of satisfactory relationship with their customers. This gives the organized dealerships an edge over the unorganized dealers in terms of enhanced quality of documentation process, certified inspection and many other factors. Additionally, the multi-brand dealerships recorded a larger market share in the organized used car market as multi-brand used car dealers have various brands and models available with them and the consumers have the choice of comparing and then purchasing the used cars. Furthermore, the organized used car dealers have a larger geographical presence in the country. To predict the price of used cars, the business would require pointers about various features pertaining to the used car. The study will aid in price prediction based on the trained model on the used car dataset and give insights on how the price varies depending on the features.

4.1.5 The objectives of the study are:

I. Perform data cleaning and visualization, in order to reach an elementary understanding of each car feature and its influence on the market price.

II. Build and evaluate models using machine learning algorithms for price prediction in order to provide a real-time used car evaluation service.

4.1.6. Complexity Involved

The dataset has some complexity which needs to be resolved in order to get better results of the predicted price. The complexity in the dataset includes: • The dataset has many outliers and missing values that need to be treated. • High number of categories for features like model was difficult to handle as data would be spread over a large area, and so encoding techniques were done.

5. DATA ANALYSIS

For this project, we are using the dataset on used car sales from all over the United States, available on Kaggle. The features available in this dataset are Mileage, VIN, Make, Model, Year, State and City.

5.1 Data Preparation

In the following section, we are going to experiment with the dataset by data cleaning, feature selection, feature extraction, and feature engineering

5.1.1. Data Cleaning

Data cleaning process of dropping null values, or fill null values using mean, median. In the project we have drop the null values from the dataset.

5.1.2 Outlier Treatment

Outliers are data points that are far from other data points. In other words, they're unusual values in a dataset. Outliers are problematic for many statistical analyses because they can cause tests to either miss significant findings or distort real results. Let’s get a sense of how we handled outliers in our dataset.

5.1.3. Missing Value Imputation

A common problem when dealing with real-world data is missing values. These can arise for many reasons and have to be either filled in or removed before we train a machine learning model. First, let’s get a sense of how many missing values are in each column.

5.1.4 Feature Selection

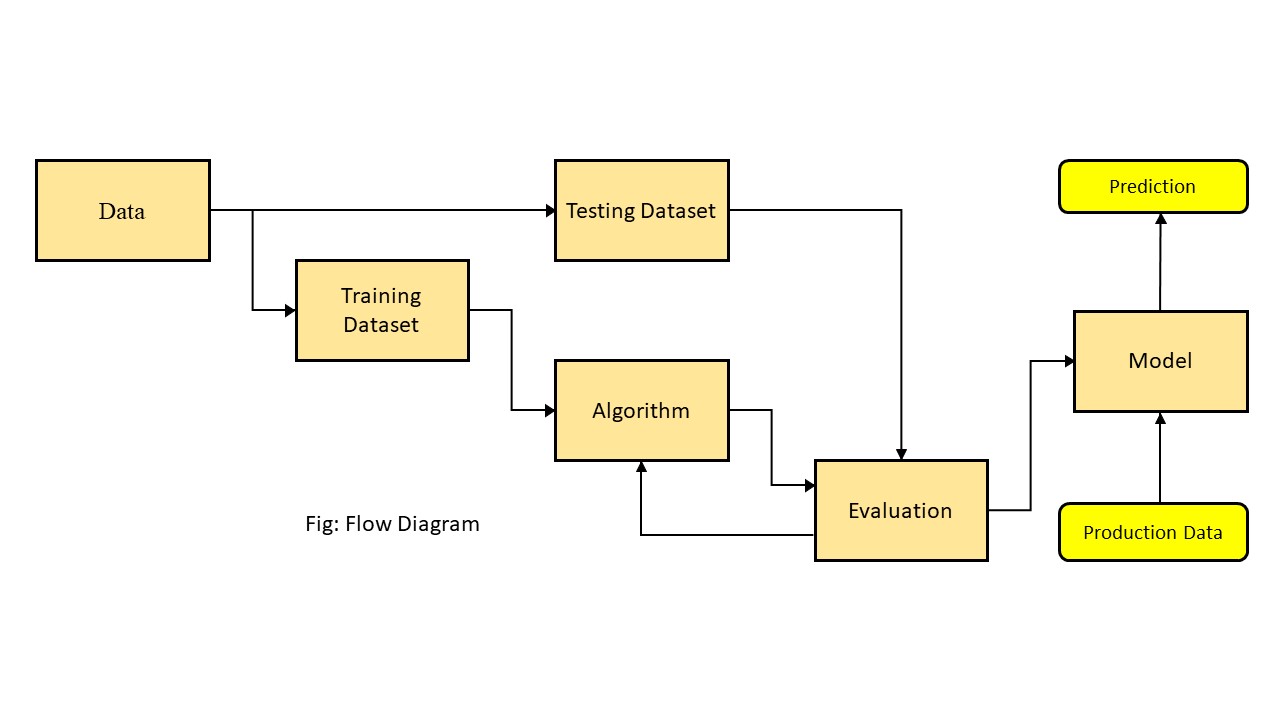
A major chunk of our feature selection was done based on the statistical analysis of the features. All of the features passed the significance test except for abtest for which we got a p-value greater than alpha (0.05).

5.1.5 Feature Extraction

feature was extracted from the Code feature. It gave an additional support to price prediction depending on in which state the vehicle is available

5.1.6 Feature Engineering

There were many categorical features in our dataset, so finding the best encoding technique for these variables was certainly a challenging task. We decided to divide the categorical features into high and low cardinality features.



6. PRE-PROCESSING

In order to get a better understanding of the data, we plotted a histogram of the data. We noticed that the dataset had many outliers, primarily due to large price sensitivity of used cars. Typically, models that are the latest year and have low mileage sell for a premium, however, there were many data points that did not conform to this. This is because accident history and condition can have a significant effect on the car’s price. Since we did not have access to vehicle history and condition, we pruned our dataset to three standard deviations around the mean in order to remove outliers. We converted the Make, Model and State into one-hot vectors. Since we had unique cities in the dataset, we replaced the string representing the city with a Boolean which was set if the population of the city was above a certain threshold i.e., a major city.

Certain features such as URL, image URL are dropped during training as these were unique to each vehicle, thus adding no value to training process. Apart from this, while implementing certain Machine Learning frameworks, we realised that certain categorical features had common values which causes problems while using frameworks such as XGBoost which require unique feature names (across all features). For e.g.: having the string “URL” in both Make and Model features is not allowed. These common feature names were hence filtered out and renamed to work with these frameworks.

6.1.1 Analysing Linearity in Dataset

To analyse the degree to which our features are linearly related to price, we plotted the Price against Mileage and Year for a particular Make and Model. There seemed to be a fair degree of linearity for these two features.

7. EXPLORATORY DATA ANALYSIS

In the following section, we are going to experiment with visualization, in order to reach an elementary understanding of each car feature and its influence on the used car market price.

The purpose of exploratory data analysis is two-fold:

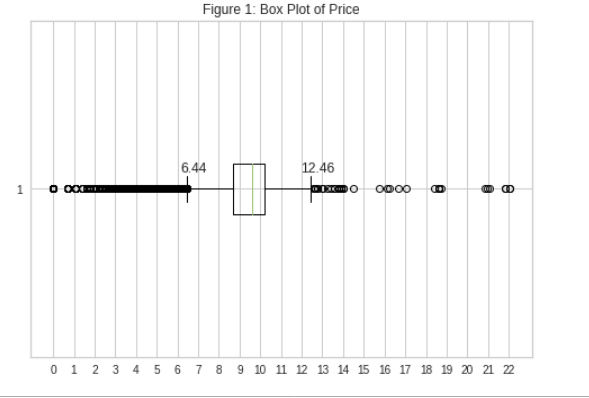
• To understand the data in terms of price across various independent variables/features.

• Get insights on various features.

7.1. Univariate Analysis

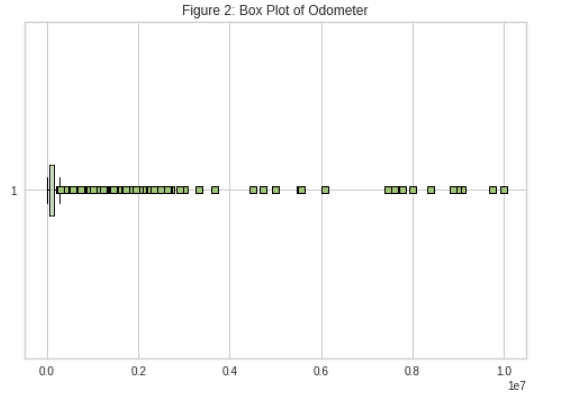
7.1.1. Price:

Box plot representation used for selling price data analysis and visualization. Price in dataset are integer variables



7.1.2. Odometer:

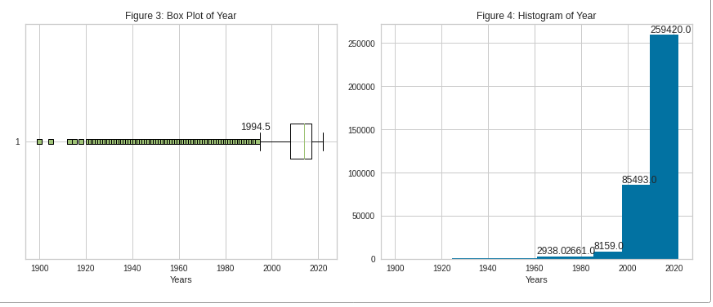
Odometer in dataset shows vehicles total run and Box plot graphical representation used float values of odometer are given in dataset.

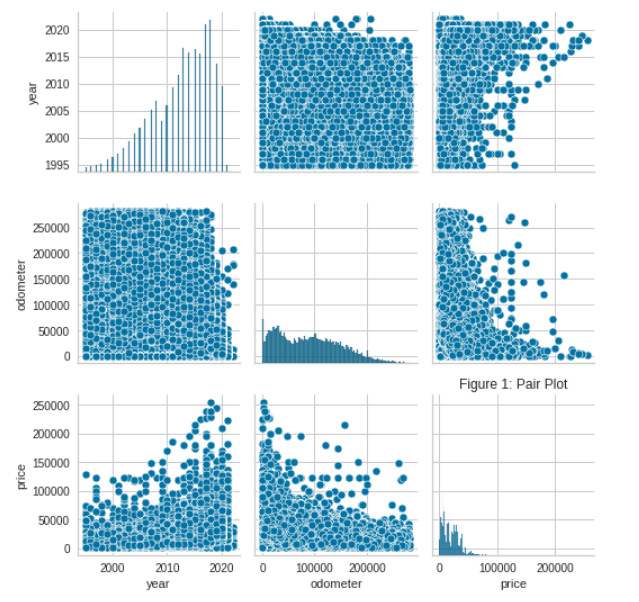


7.1.3 Age:

Year of vehicle manufacture and total age factor effect on selling price of vehicle.

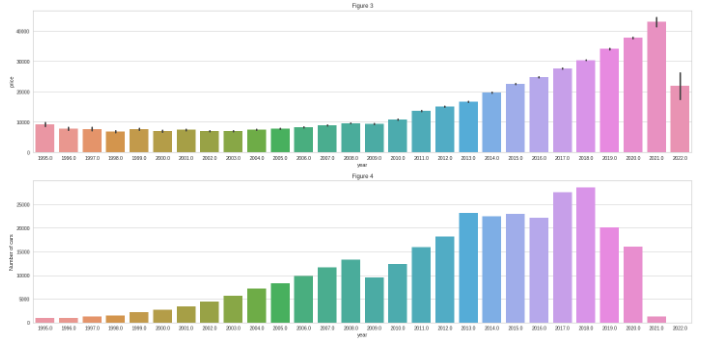
Box plot and histogram used for graphical representation of age of vehicle.





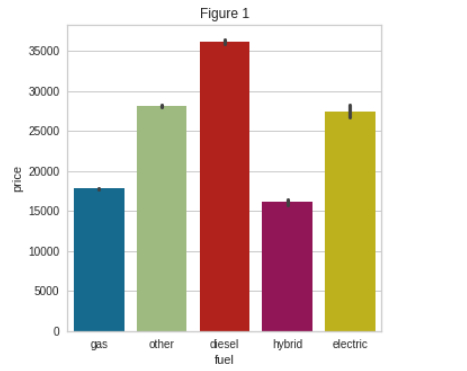
7.1.4 Count of brand:

Total count of manufactures shown on bar plot with respective to year.



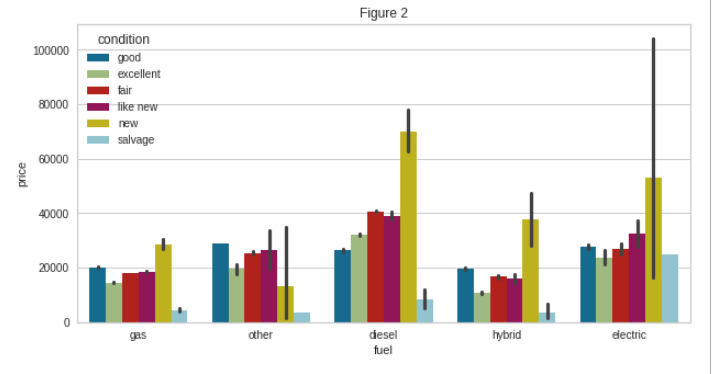
7.1.5 Count of vehicle Type:

Dataset contain different type of vehicles based on fuel systems are gas, diesel hybrid, electric. Bar plot used to fuel type data visualization.



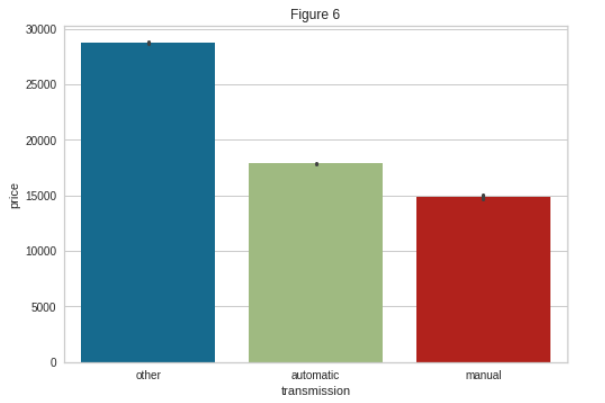
7.1.6 CONDITION:

Depend on different type of fuel type vehicle condition of vehicle data visualisation using Bar plot.



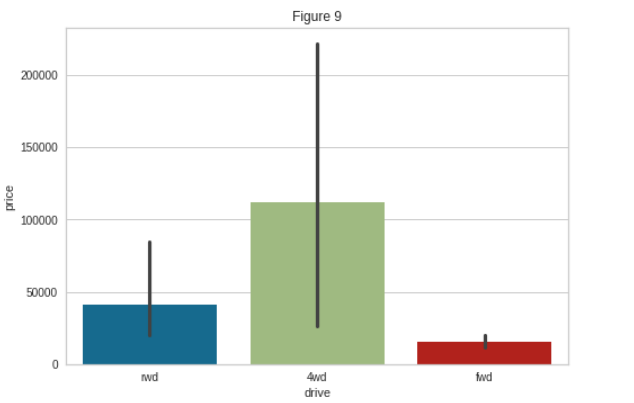
7.1.7 TRANSMISSION TYPE:

Dataset shows transmission type manual, automatic which directly affect on price of vehicle. Data visualisation shown using bar plot.



7.1.8 DRIVE:

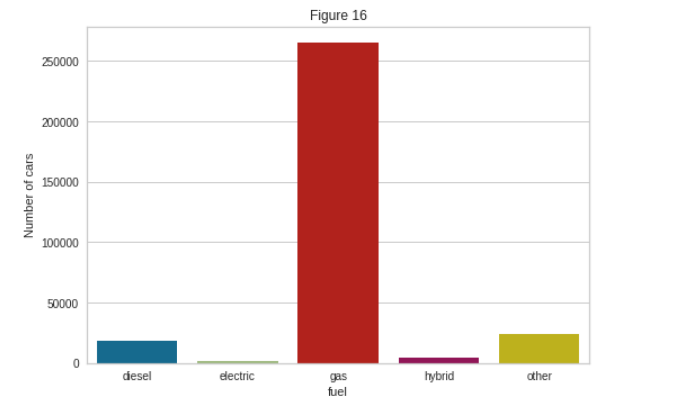
Dataset shows rwd,4wd, fwd. different type of drive which affect on price of vehicle shown using Bar plot.



7.2.1. Bivariate Analysis

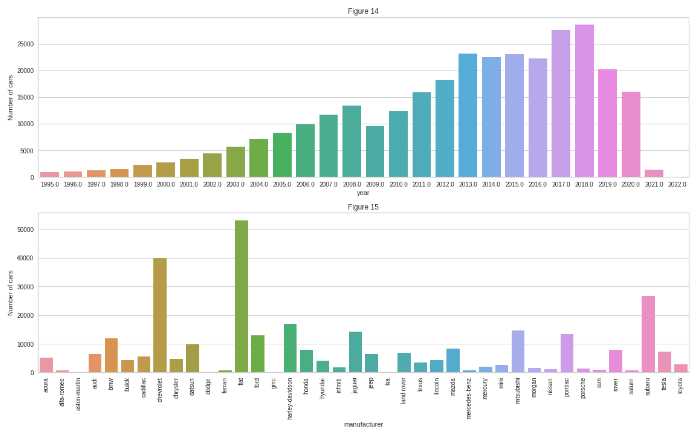
7.2.2 FUEL VS NO. OF CARS:

Following Bar plot shows relationship between fuel type and number of cars data visualisation using Bar plot. Gas fuel type vehicles maximum used to see according to visualisation. Electric, hybrid, other type of car is less used compare to the gas type vehicles.



7.2.3 NO. OF CARS VS MANUFACTURER:

Number of cars and the manufacturer relationship sown using bar plot. Number of cars and manufacture year relationship shown using bar plot.



8. METHODOLGY

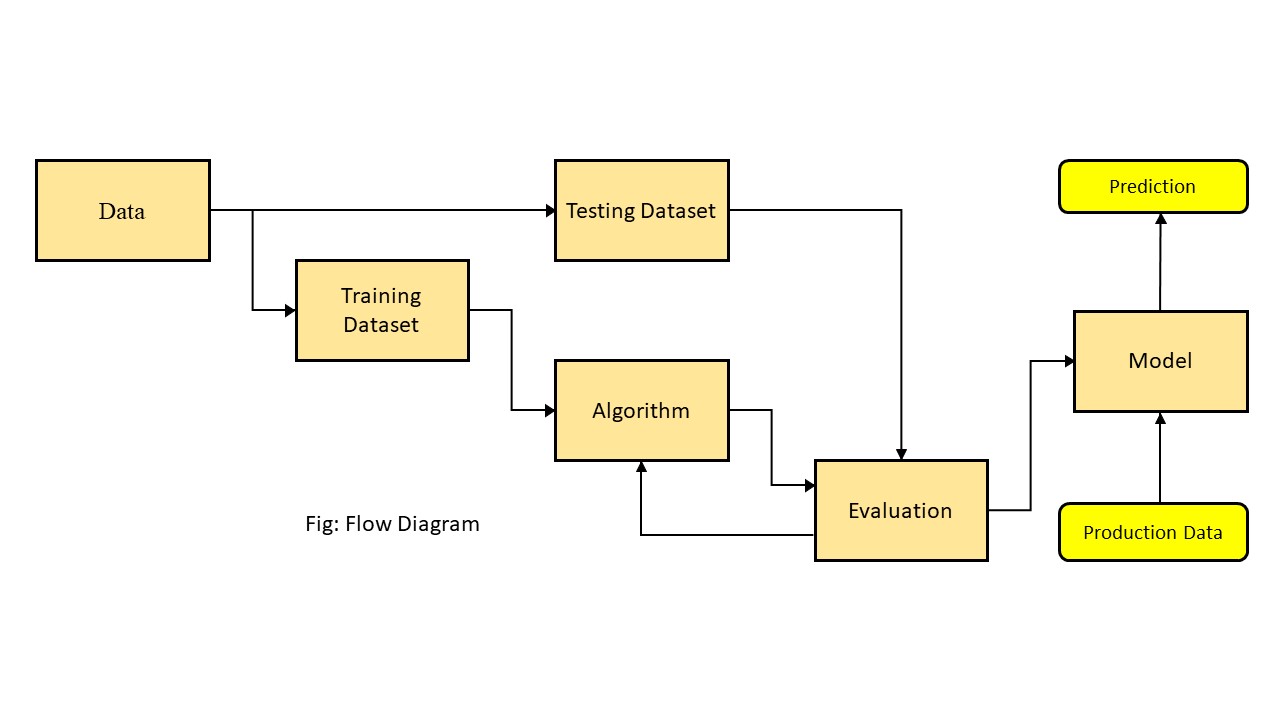
There are two primary phases in the system:

8.1.1 Training phase:

The system is trained by using the data in the data stand fits a model based on the algorithm chosen accordingly.

8.1.2 Testing phase:

The system is provided with the inputs and is tested for its working. The accuracy is checked. And therefore, the data that is used to train the model or test it, has to be appropriate. The system is designed to detect and predict price of used car and hence appropriate algorithms must be used to do the two different tasks. Before the algorithms are selected for further use, different algorithms were compared for its accuracy. The well-suited one for the task was chosen.



8.1.3 Objective

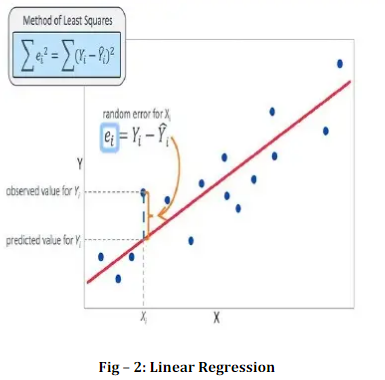
* To develop an efficient and effective model which predicts the price of a used car according to user’s inputs.
* To achieve good accuracy.
* To develop a User Interface which is user-friendly and takes input from the user and predicts the price

9. PROPOSED SYSTEM

As shown in the above figure, the process starts by collecting the dataset. The next step is to do Data Pre-processing which includes Data cleaning, Data reduction, Data Transformation. Then, using various machine learning algorithms we will predict the price. The algorithms involve Linear Regression, Ridge Regression and Lasso Regression. The best model the “LASSO” stands for Least Absolute Shrinkage and Selection Operator. Lasso regression is a regularization technique. It is used over regression methods for a more accurate prediction. This model uses shrinkage. Shrinkage is where data values are shrunk towards a central point as the mean. The lasso procedure encourages simple, sparse models (i.e., models with fewer parameters). This particular type of regression is well-suited for models showing high levels of multicollinearity or when you want to automate certain parts of model selection, like variable selection/parameter elimination which predicts the most accurate price is selected. After selection of the best model the predicted price is displayed to the user according to user’s inputs. User can give input through website to for used car price prediction to machine learning model

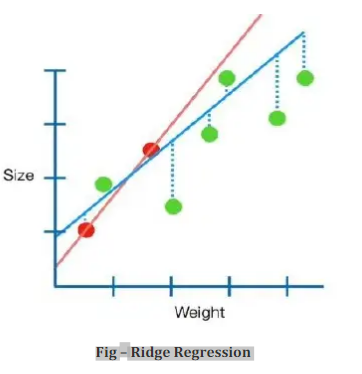
9.1 Linear Regression

Linear Regression attempt to model the relationship between two variables by fitting a linear equation to observed data. The other is considered to be dependent variable. For Example: A modeler might want to relate weights of individuals to their heights using a linear regression mode



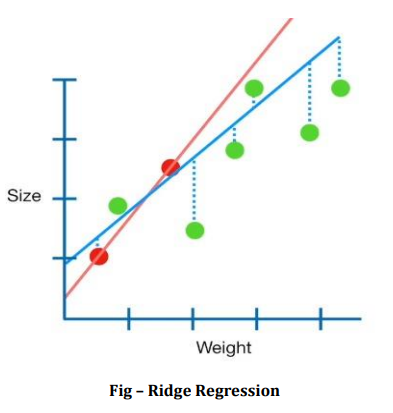
9.2 Ridge Regression

A Ridge regressor is basically a regularized version of Linear Regressor.The regularized term has the parameter ‘alpha’ which controls the regularization of the model i.e., helps in reducing the variance of the estimates.



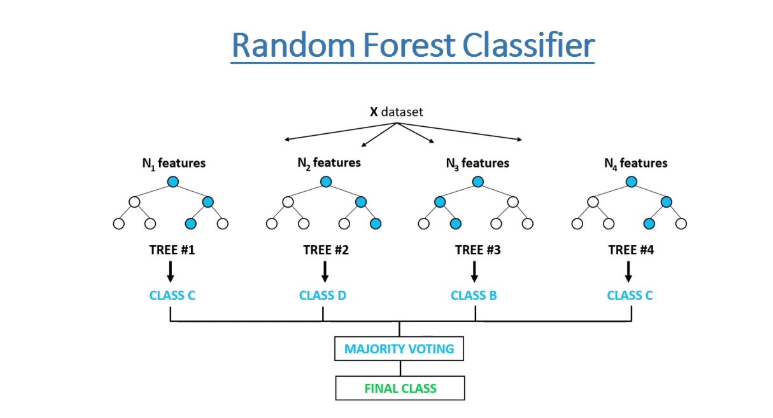
9.3 Lasso Regression

The “LASSO” stands for Least Absolute Shrinkage and Selection Operator. Lasso regression is a regularization technique. It is used over regression methods for a more accurate prediction. This model uses shrinkage. Shrinkage is where data values are shrunk towards a central point as the mean. The lasso procedure encourages simple, sparse models (i.e. models with fewer parameters). This particular type of regression is well-suited for models showing high levels of multicollinearity or when you want to automate certain parts of model selection, like variable selection/parameter elimination



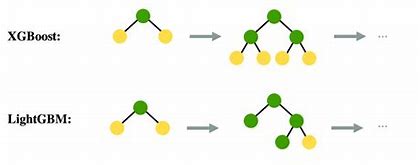
9.4. Random Forest

Random Forest is an ensemble learning based regression model. It uses a model called decision tree, specifically as the name suggests, multiple decision trees to generate the ensemble model which collectively produces a prediction. The benefit of this model is that the trees are produced in parallel and are relatively uncorrelated, thus producing good results as each tree is not prone to individual errors of other trees. This uncorrelated behaviour is partly ensured by the use of Bootstrap Aggregation or bagging providing the randomness required to produce robust and uncorrelated trees. This model was hence chosen to account for the large number of features in the dataset and compare a bagging technique with the following gradient boosting methods.



9.5. XGBoost

Extreme Gradient Boosting or XGBoost [4] is one of the most popular machine learning models in current times. XGBoost is quite similar at the core to the original gradient boosting algorithm but features many additive features that significantly improve its performance such as built-in support for regularization, parallel processing as well as giving additional hyperparameters to tune such as tree pruning, sub sampling and number of decision trees. A maximum depth of used and the algorithm was run on all cores in parallel.



10. Assumptions of algorithms

10.1.1. No Auto Correlation

For Durbin-Watson test we got a value of 2.004 which lies in the acceptance range of 1.5 - 2.5, hence, no autocorrelation. Moreover, the ACF plot also clearly confirms the same, that is no auto-correlation between the residuals.

10.1.2. Normality of residuals

The Jarque Bera test resulted in a p-value of 0.0 along with a test statistic value of 279629.60 which is greater than the t-critical value of 5.99. Moreover, our residuals deviated from normality towards the extreme which we can clearly see from the Q-Q - plot below. So, we rejected the null hypothesis and concluded that residuals are not normal.

10.1.3 Homoscedasticity

If the variance of the residuals is symmetrically distributed across the regression line, then the data is said to be homoscedastic. The Goldfeld-Quandt test gives a p-value of 0.3132 which is higher than 0.05. Moreover, we can visually see that Homoscedasticity is present.

10.1.4 No Multicollinearity

The variance inflation factor (VIF) quantifies the extent of correlation between one predictor and the other predictors in a model. Values of more than 4 or 5 are sometimes regarded as being moderate to high, with values of 10 or more being regarded as very high.

10.1.5 Model Performance Measures Used for Evaluating Models

The various models built, must be evaluated based on certain model performance measures to identify the most robust models. The choice of the right model performance measures is highly critical hence, Root Mean Squared Error (RMSE) was also considered, in addition to R-squared.

10.1.6 Root Mean Squared Errors (RMSE)

The RMSE is the square root of the variance of the residuals. It indicates the absolute fit of the model to the data–how close the observed data points are to the model’s predicted values. As the square root of a variance, RMSE can be interpreted as the standard deviation of the unexplained variance, and has the useful property of being in the same units as the response variable. Lower values of RMSE indicate better fit. RMSE is a good measure of how accurately the model predicts the response, and it is the most important criterion for fit if the main purpose of the model is prediction.

10.1.7 R-Squared

R-squared is the coefficient of determination; it ranges from zero to one, with zero indicating that the proposed model does not improve prediction over the mean model, and one indicating perfect prediction. Improvement in the regression model results in proportional increases in R-squared

Machine learning algorithms are classified as two distinct groups: parametric and non-parametric. Herein, parametric Ness is related to a pair of model complexity and the number of rows in the train set. We can classify algorithms as non-parametric when model becomes more complex if number of samples in the training set increases. Vice versa, a model would be parametric if model becomes stable when number of examples in the training set increases.

11. Future Work

For better performance, we plan to judiciously design deep learning network structures, use adaptive learning rates and train on clusters of data rather than the whole dataset. To correct for overfitting in Random Forest, different selections of features and number of trees will be tested to check for change in performance. This package can be used in internet environment that will make handling of different customers easier. Future expandability and interconnectivity are the features, which are considered for scope in future.

12. CONCLUSION

* By performing different ML models, we aim to get a better result or less error with max accuracy. Our purpose was to predict the price of the used cars having 25 predictors and 509577 data entries.
* Initially, data cleaning is performed to remove the null values and outliers from the dataset then ML models are implemented to predict the price of cars.
* Next, with the help of data visualization features were explored deeply. The relation between the features is examined.
* From the below table, it can be concluded that XGBoost is the best model for the prediction for used car prices. XGBoost as a regression model gave the best MSLE and RMSLE values.