

**DR. BR AMBEDKAR NATIONAL INSTITUTE OF TECHNOLOGY
, JALANDHAR (144011), PUNJAB**



Machine Learning (CSPC-204)

Mini-Project

Submitted by:

NAROTAM SINGH (20103097)

NAYAN ABHISHEK (20103098)

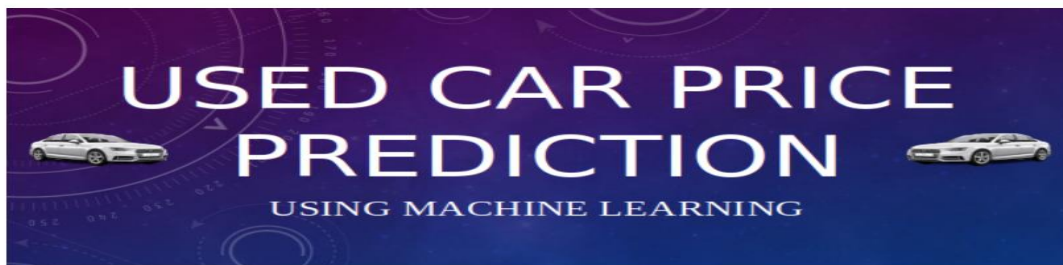
PRASANT (20103108)

Branch- CSE (B - G2)

B.Tech – 2 nd Year

Submitted to: Dr Jagdeep Kaur

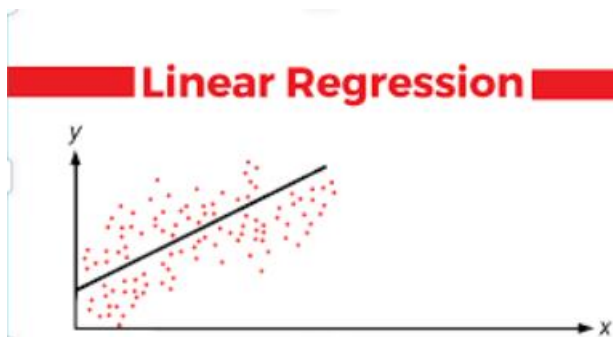
Introduction



Predict the price of an old car using some set of features given in dataset.
We Trained our model on two algorithms Linear Regression and XGBoost.

Algorithm Description

Linear Regression:-



Regression is used to study the relationship between two variables.

Linear Regression is used to study the relationship between two variables.

It assumes that there exists a linear relationship between a dependent variable and independent variable.

It tries to find out the best linear relationship that describes the data you have.

Single Variable Linear Regression

For simple linear regression, the form of the model is-

$Y_i = mX_i + b$

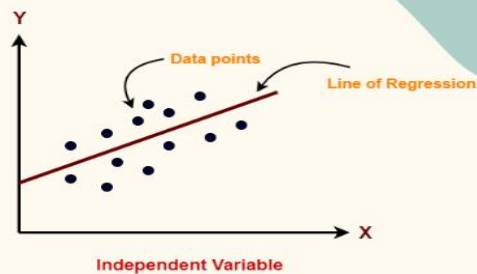
x: input training data

y: labels to data (supervised learning)

$$\text{minimize } \frac{1}{n} \sum_{i=1}^n (\text{pred}_i - y_i)^2$$

$$m = \frac{n \sum xy - \sum x \sum y}{n \sum x^2 - (\sum x)^2}$$

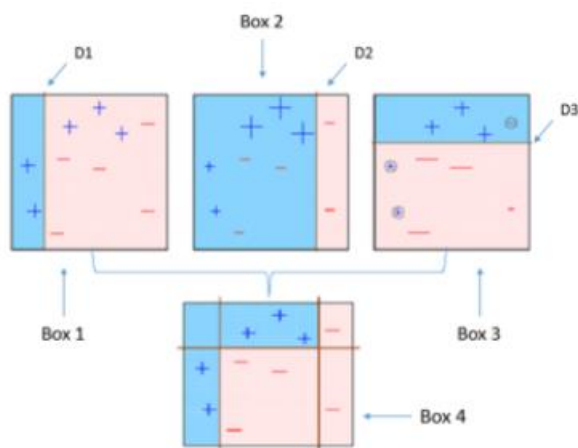
$$b = \frac{\sum y - m \sum x}{n}$$



XGBoost:-

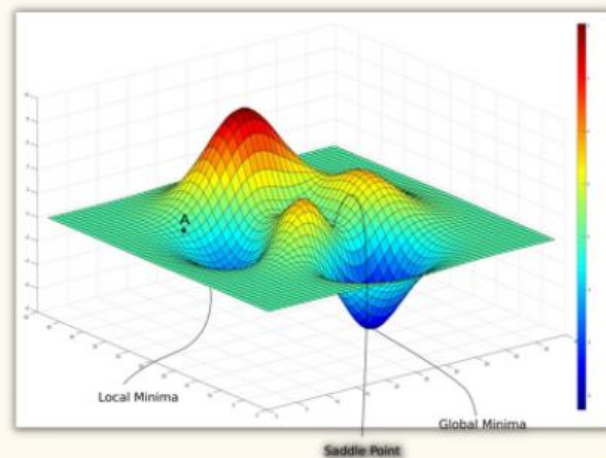
Stands for:

- eXtreme Gradient Boosting.
- XGBoost is a powerful iterative learning algorithm based on gradient boosting.
- Robust and highly versatile, with custom objective loss function compatibility.



Why use Xgboost?

- All of the advantages of gradient boosting, plus more.
- Utilizes CPU Parallel Processing by default.
- Two main reasons for use:
 1. Low Runtime
 2. High Model Performance



How does XGBoost work?

- Tree-Based Boosting algorithm.
- 4 Critical Parameters for Tuning:
 1. η : ETA or “Learning Rate”
 2. `max_depth`: Controls the “height” of the tree via splits.
 3. γ : Minimum required loss for the model to justify a split.
 4. λ : L2 (Ridge) regularization on variable weights.

Data Cleaning

One Hot Encoding for Categorical Variable

What is Categorical Data?

Categorical data are variables that contain label values rather than numeric values. The number of possible values is often limited to a fixed set.

Many machine learning algorithms including linear regression cannot operate on label data directly. They require all input variables and output variables to be numeric.

Using this encoding and allowing the model to assume a natural ordering between categories may result in poor performance or unexpected results (predictions halfway between categories).

For example:

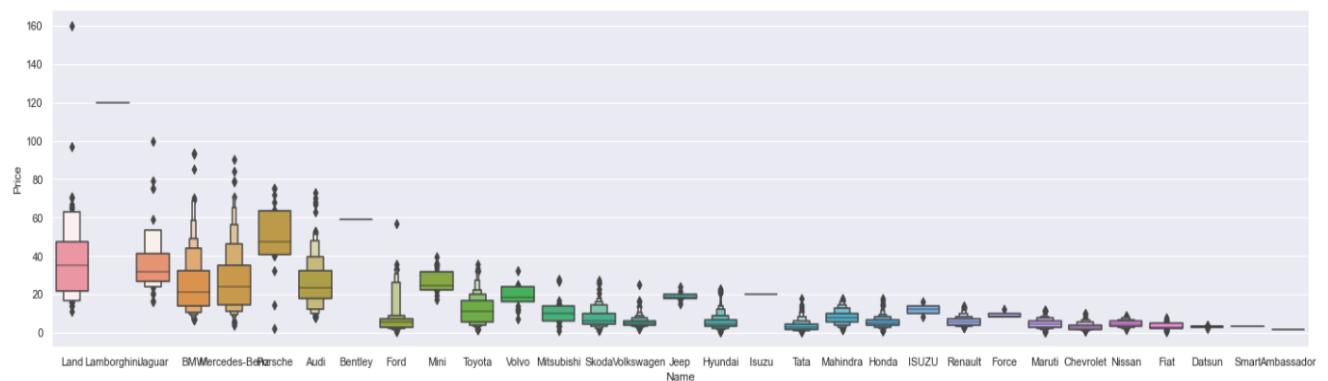
	red,	green,	blue
1	red,	green,	blue
2	1,	0,	0
3	0,	1,	0
4	0,	0,	1

The binary variables are often called “dummy variables” in other fields, such as statistics.

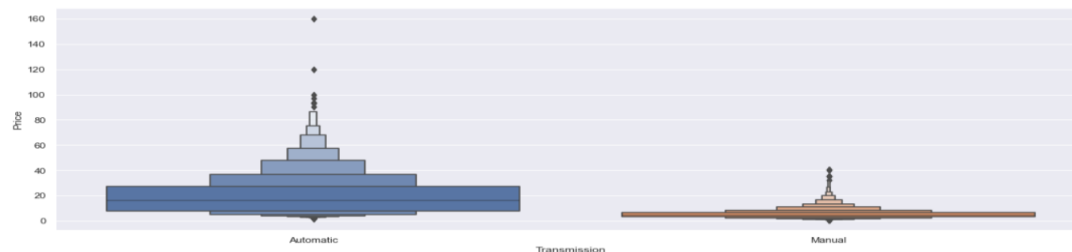
DataSet Description:-

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6019 entries, 0 to 6018
Data columns (total 14 columns):
#   Column                Non-Null Count  Dtype  
---  -
0   Unnamed: 0             6019 non-null   int64  
1   Name                   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 
dtypes: float64(2), int64(3), object(9)
memory usage: 658.5+ KB
```

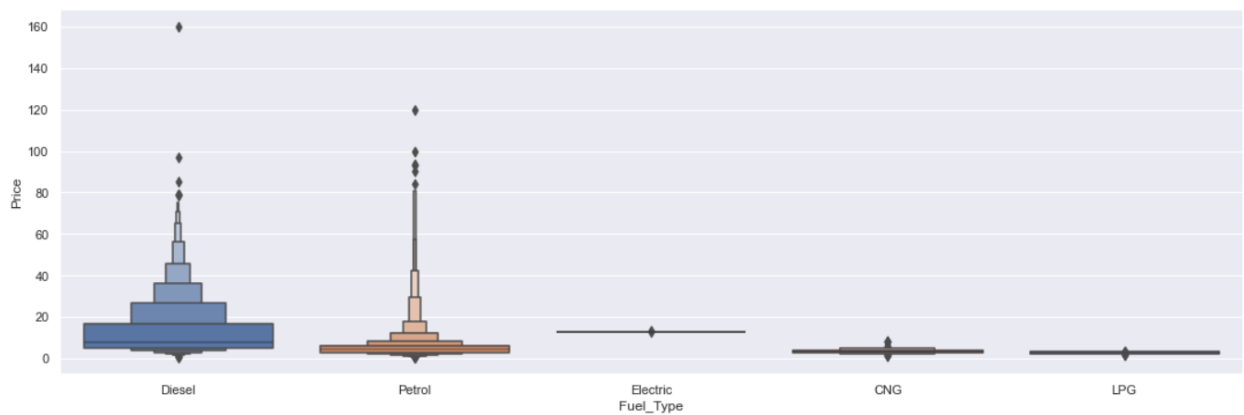
Price Vs Company Scatter Chart:-



Price Vs Transmission Scatter Chart:-



Price Vs Fuel Type Scatter Chart:-



Model Result:-

Linear Regression:-

Accuracy: - 77.35%

XGBoost :-

Accuracy:- 99.55%

Our Team

- ❖ Narotam Singh (20103097) – Data Preprocessing
- ❖ Nayan Abhishek (20103098) – Data Visualization
- ❖ Prasant (20103108) – Model Deployment