B.M.S COLLEGE OF ENGINEERING

(An Autonomous College under VTU, Belagavi)

Bull Temple Road, Bangalore - 560 019

A Project Report-2021-22

On

"VEHICLE PRICE PREDICTION"

Submitted as a part of Alternate Assessment for the Elective course

MACHINE LEARNING

Offered by

ELECTRONICS AND COMMUNICATIONS ENGINEERING

In association with

NOKIA NETWORKS, BANGALORE

Submitted By

NAME:	USN:
1. Bhuvan.D.B	1BM19EC030
2. Bhargav	1BM19EC028
3. Amogha.A.Acharya	1BM19EC013
4. Anil.R.Devarakki	1BM19EC016
5. Atul.Ram	1BM19EC023
6. Thejas J	1BM19EC191

FIC:	Dr.Suma M N	Dr.Geetishree Mishra
Designation	Professor	Assistant Professor

Table of Contents

ABSTRACT

Problem Statement

Nowadays numerous cars are being launched daily, which also means that most of these vehicles will enter into the resale market. In order to effectively gauge the resale value of these vehicles, both as a seller and a buyer, it is necessary to have a good predictive mechanism to predict the price of a car we intend to buy or sell.

The Price of the car can be predicted by analyzing several data pertaining to the Features present in the car and the performance it has to offer.

Our Model aims to utilize these features and help predicting the corresponding price of the vehicles and thereby helps people in effectively analyzing their requirements and selecting the best vehicles based on their needs and budget.

Problem Solution

Here we have tried to solve the problem mentioned above through the concept of machine learning, especially through Linear Regression.

We will first train the dataset of car prices using a linear regression model and then that model will be used to predict the prices.

From multiple regression models we will choose the one which has minimum mean square error and then this model will be used for prediction.

INTRODUCTION

In recent years, the number of vehicles produced has increased considerably, which in turn has increased competition in the market. Due to advancement in technology many features are added to the vehicles to make our life comfortable. Due to these Features the prices of the vehicles are highly variable and the difficulty to get the right car with right features is increasing for everyone. Especially for the people who want to buy second hand, third hand cars.

Recent improvement in technologies such as Automated driving ,Electric vehicles ,and many more have brought a transition in automobile sector and hence vehicular sales has been increased considerably .In order to predict the prices of cars we have made a model which will intake features as input and gives the approximate price of the car as the output. The multiple features which are taken into consideration are vehicle mileage, year of manufacturing, Fuel consumption, transmission, Fuel type, Engine Power. The model Benefits sellers and buyers.

The model building process includes Machine learning. Before the actual start of model-building, this project visualized the data to understand the dataset better. The dataset was divided and modified to fit the regression, thus ensuring the performance of the regression. To evaluate the performance of each regression, R-square and mean absolute error was calculated.

Literature Survey

[1] V.C.Sanap, Mohammed Munawwar Rangila, Sufiyaan Rahi, Samiksha Badgujar, Yashodhan Gupta. “Car Price Prediction using Linear Regression Technique of Machine Learning”. International Journal of Innovative Research in Science, Engineering and Technology. Volume 11, Issue 4, April 2022.

[2] Laveena D’Costa, Ashoka Wilson D’Souza, Abhijith K, Deepthi Maria Varghese. “Predicting True Value of Used Car using Multiple Linear Regression Model”. International Journal of Recent Technology and Engineering. ISSN: 2277-3878, Volume-8, Issue-5S, January 2020.

[3] Kanwal Noor, Sadaqat Jan. “Vehicle Price Prediction System using Machine Learning Techniques”. International Journal of Computer Applications. Volume 167 – No.9, June 2017.

[4] Praful Rane, Deep Pandya, Dhawal Kotak. “USED CAR PRICE PREDICTION”. International Research Journal of Engineering and Technology. Volume: 08 Issue: 04 | Apr 2021.

[5] Chuyang Jin. “Price Prediction of Used Cars Using Machine Learning”. IEEE International Conference on Emergency Science and Information Technology. IEEE 22-24 November 2021.

[6] Enis Gegic, Becir Isakovic, Dino Keco, Zerina Masetic, Jasmin Kevric. “Car Price Prediction using Machine Learning Techniques” .TEM Journal. Volume 8, Issue 1, Pages 113-118, ISSN 2217-8309, February 2019.

[7] K.Samruddhi ¸ Dr. R.Ashok Kumar. “Used Car Price Prediction using K-Nearest Neighbor Based Model”. International Journal of Innovative Research in Applied Sciences and Engineering.Volume 4, Issue 3, September 2020.

METHODOLOGY

Prediction dataset

	name	year	selling_price	km_driven	fuel	seller_type	transmission	owner	mileage	engine	max_powe
0	Maruti Swift Dzire VDI	2014	450000	145500	Diesel	Individual	Manual	First Owner	23.4 kmpl	1248 CC	74 bh
1	Skoda Rapid 1.5 TDI Ambition	2014	370000	120000	Diesel	Individual	Manual	Second Owner	21.14 kmpl	1498 CC	103.52 bh
2	Honda City 2017-2020 EXi	2006	158000	140000	Petrol	Individual	Manual	Third Owner	17.7 kmpl	1497 CC	78 bh
3	Hyundai i20 Sportz Diesel	2010	225000	127000	Diesel	Individual	Manual	First Owner	23.0 kmpl	1396 CC	90 bh
4	Maruti Swift VXI BSIII	2007	130000	120000	Petrol	Individual	Manual	First Owner	16.1 kmpl	1298 CC	88.2 bh
5	Hyundai Xcent 1.2 VTVT E Plus	2017	440000	45000	Petrol	Individual	Manual	First Owner	20.14 kmpl	1197 CC	81.86 bh
6	Maruti Wagon R LXI DUO BSIII	2007	96000	175000	LPG	Individual	Manual	First Owner	17.3 km/kg	1061 CC	57.5 bh
7	Maruti 800 DX BSII	2001	45000	5000	Petrol	Individual	Manual	Second Owner	16.1 kmpl	796 CC	37 bh
8	Toyota Etios VXD	2011	350000	90000	Diesel	Individual	Manual	First Owner	23.59 kmpl	1364 CC	67.1 bh
9	Ford Figo Diesel Celebration Edition	2013	200000	169000	Diesel	Individual	Manual	First Owner	20.0 kmpl	1399 CC	68.1 bh

This Dataset by “CarDheko.com” has been taken from the Kaggle database. For the purpose of trying to predict the resale value of vehicles in India.

Dataset Description:

Name - Company and model name of the vehicle.

Year - Year of Manufacture.

Selling Price - Recent sale value of the vehicle.

Km_driven - Total Kms traveled

Fuel_type - Fuel intake of vehicle

Seller_type - Either owner/individual or broker.

Transmission - Either manual or automatic.

Owner - First/Second/Third/fourth and above.

Mileage - Mileage of that vehicle

Engine - Denotes the capacity of the engine.

Max-power - Brake horse power of the vehicle.

Torque - Torque value of engine.

Seats - Number of seats in a vehicle.

Phase 1

Data Exploration and Cleaning

Name – Thejas J

USN - 1BM19EC191

The dataset consists of 13 columns and 8128 rows.

As we can observe, most of the columns in the dataset are of type “object”, which is unexpected since most of the columns such as torque, mileage, engine, max power are considered to numerical fields.

This shows the presence of string contamination in the numerical data fields.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8128 entries, 0 to 8127
Data columns (total 13 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   name           8128 non-null   object
 1   year           8128 non-null   int64
 2   selling_price  8128 non-null   int64
 3   km_driven      8128 non-null   int64
 4   fuel           8128 non-null   object
 5   seller_type    8128 non-null   object
 6   transmission   8128 non-null   object
 7   owner          8128 non-null   object
 8   mileage        7907 non-null   object
 9   engine         7907 non-null   object
 10  max_power      7913 non-null   object
 11  torque         7906 non-null   object
 12  seats          7907 non-null   float64
dtypes: float64(1), int64(3), object(9)
memory usage: 825.6+ KB
```

	name	year	selling_price	km_driven	fuel	seller_type	transmission	owner	mileage	engine	max_power	torque	seats
0	Maruti Swift Dzire VDI	2014	450000	145500	Diesel	Individual	Manual	First Owner	23.4 kmpl	1248 CC	74 bhp	190Nm@ 2000rpm	5.0
1	Skoda Rapid 1.5 TDI Ambition	2014	370000	120000	Diesel	Individual	Manual	Second Owner	21.14 kmpl	1498 CC	103.52 bhp	250Nm@ 1500-2500rpm	5.0
2	Honda City 2017-2020 EXi	2006	158000	140000	Petrol	Individual	Manual	Third Owner	17.7 kmpl	1497 CC	78 bhp	12.7@ 2,700(kgm@ rpm)	5.0
3	Hyundai i20 Sportz Diesel	2010	225000	127000	Diesel	Individual	Manual	First Owner	23.0 kmpl	1396 CC	90 bhp	22.4 kgm at 1750-2750rpm	5.0
4	Maruti Swift VXI BSIII	2007	130000	120000	Petrol	Individual	Manual	First Owner	16.1 kmpl	1298 CC	88.2 bhp	11.5@ 4,500(kgm@ rpm)	5.0
5	Hyundai Xcent 1.2 VTVT E Plus	2017	440000	45000	Petrol	Individual	Manual	First Owner	20.14 kmpl	1197 CC	81.86 bhp	113.75nm@ 4000rpm	5.0
6	Maruti Wagon R LXI DUO BSIII	2007	96000	175000	LPG	Individual	Manual	First Owner	17.3 km/kg	1061 CC	57.5 bhp	7.8@ 4,500(kgm@ rpm)	5.0
7	Maruti 800 DX BSII	2001	45000	5000	Petrol	Individual	Manual	Second Owner	16.1 kmpl	796 CC	37 bhp	59Nm@ 2500rpm	4.0
8	Toyota Etios VXD	2011	350000	90000	Diesel	Individual	Manual	First Owner	23.59 kmpl	1364 CC	67.1 bhp	170Nm@ 1800-2400rpm	5.0
9	Ford Figo Diesel Celebration Edition	2013	200000	169000	Diesel	Individual	Manual	First Owner	20.0 kmpl	1399 CC	68.1 bhp	160Nm@ 2000rpm	5.0

Our primary step is to separate out the company name from the name of the vehicle. This will help in analyzing the company market share, in the resale market and could be a means of categorizing the vehicles based on their company of make.

The name of the company is separated out as the first string in the vehicle name

And a separate column is created for the same, erstwhile separating the vehicle names too.

```
company_list = []

for i in df.name:
    company_list.append(i.split()[0])

df.insert(0, "company", company_list, True)
```

```
for i in range(0,len(df.name)):
    df.at[i, 'name'] = df.name[i].split(" ",1)[1]
```

Before

	name
8118	Hyundai i20 Magna
8119	Maruti Wagon R LXI Optional
8120	Hyundai Santro Xing GLS
8121	Maruti Wagon R VXI BS IV with ABS
8122	Hyundai i20 Magna 1.4 CRDi
8123	Hyundai i20 Magna
8124	Hyundai Verna CRDi SX
8125	Maruti Swift Dzire ZDi
8126	Tata Indigo CR4
8127	Tata Indigo CR4

After

	company	name
0	Maruti	Swift Dzire VDI
1	Skoda	Rapid 1.5 TDI Ambition
2	Honda	City 2017-2020 EXi
3	Hyundai	i20 Sportz Diesel
4	Maruti	Swift VXI BSIII
...	...	...
8123	Hyundai	i20 Magna
8124	Hyundai	Verna CRDi SX
8125	Maruti	Swift Dzire ZDi
8126	Tata	Indigo CR4
8127	Tata	Indigo CR4

The null values in the dataset are identified.

Since the number of null values account for a miniscule amount compared to the size of the data set, these null values can be safely dropped without any adverse effects.

```
df.isnull().sum()
```

```
company           0
name              0
year              0
selling_price     0
km_driven         0
fuel              0
seller_type       0
transmission      0
owner             0
mileage         221
engine          221
max_power       215
torque          222
seats           221
dtype: int64
```

```
df = df.dropna()
df.reset_index(drop = True, inplace = True)
```

Some of the values in the “mileage”

Column consisted of ‘0’ instead of ‘nan’, and also contaminated with decimal values.

Since the mileage of a vehicle cannot be a decimal value, the irregularities were identified and removed.

```
for i in range(0,len(df.name)):
    try:
        if float(df.mileage[i]) < 1:
            df = df.drop(i)
    except Exception as e:
            print(i,df.mileage[i],e)
df.reset_index(drop = True, inplace = True)
```

Name – Bhargav Natekar

USN - 1BM19EC028

The string contaminations are cleaned and removed from the numeric fields

```
for i in range(0,len(df.name)):
    try:
        df.at[i, 'engine'] = df.engine[i].split(" ",1)[0]
        df.at[i, 'mileage'] = df.mileage[i].split(" ",1)[0]
        df.at[i, 'owner'] = df.owner[i].split(" ",1)[0]
        df.at[i, 'max_power'] = df.max_power[i].split(" ",1)[0]
    except:
        pass
```

Once the contamination strings are removed, the data can safely be typecast into its corresponding numerical types.

```
df['engine'] = df['engine'].apply(np.int64)
df['seats'] = df['seats'].apply(np.int64)
df['transmission'] = df['transmission'].apply(np.int64)
df['mileage'] = df['mileage'].apply(np.float64)
df['max_power'] = df['max_power'].apply(np.float64)
df['torque'] = df['torque'].apply(np.int64)
```

Before

owner	mileage	engine	max_power	torque
First Owner	23.4 kmpl	1248 CC	74 bhp	190Nm@ 2000rpm
Second Owner	21.14 kmpl	1498 CC	103.52 bhp	250Nm@ 1500-2500rpm
Third Owner	17.7 kmpl	1497 CC	78 bhp	12.7@ 2,700(kgm@ rpm)
First Owner	23.0 kmpl	1396 CC	90 bhp	22.4 kgm at 1750-2750rpm
First Owner	16.1 kmpl	1298 CC	88.2 bhp	11.5@ 4,500(kgm@ rpm)
...	...	...	...	...
First Owner	18.5 kmpl	1197 CC	82.85 bhp	113.7Nm@ 4000rpm
Fourth & Above Owner	16.8 kmpl	1493 CC	110 bhp	24@ 1,900-2,750(kgm@ rpm)
First Owner	19.3 kmpl	1248 CC	73.9 bhp	190Nm@ 2000rpm
First Owner	23.57 kmpl	1396 CC	70 bhp	140Nm@ 1800-3000rpm
First Owner	23.57 kmpl	1396 CC	70 bhp	140Nm@ 1800-3000rpm

After

owner	mileage	engine	max_power	torque
First	23.4	1248	74	190.0
Second	21.14	1498	103.52	250.0
Third	17.7	1497	78	124.57176
First	23.0	1396	90	219.71712
First	16.1	1298	88.2	112.8012
...	...	...	...	...
First	18.5	1197	82.85	113.7
Fourth	16.8	1493	110	235.4112
First	19.3	1248	73.9	190.0
First	23.57	1396	70	140.0
First	23.57	1396	70	140.0

The torque field consists of values/measurements in both Nm and Kgm

This difference in measurement standards, leads to numerical irregularities in their values.

Hence these values must be all brought to one standard of measurement.

```
y = []
for i in range(0, len(df['torque'])):
    str = df.torque[i].split()[0]
    t1 = re.findall('\d*\.?\d+', str)[0]
    if 'Nm' in str:
      df.at[i,'torque'] = float(t1)
    else:
      df.at[i,'torque'] = float(t1) * 9.8088
      if(int(df.at[i,'torque'] > 1000)):
        df.at[i,'torque'] = df.at[i,'torque'] / 9.8088
```

Here the values in Kgm were identified using regular expressions and were converted to Nm, as Nm is the adopted standard of measuring the torque produced.

Before

torque
190Nm@ 2000rpm
250Nm@ 1500-2500rpm
12.7@ 2,700(kgm@ rpm)
22.4 kgm at 1750-2750rpm
11.5@ 4,500(kgm@ rpm)
...
113.7Nm@ 4000rpm
24@ 1,900-2,750(kgm@ rpm)
190Nm@ 2000rpm
140Nm@ 1800-3000rpm
140Nm@ 1800-3000rpm

After

torque
190.0
250.0
124.57176
219.71712
112.8012
...
113.7
235.4112
190.0
140.0
140.0

Data Visualization

After the cleaning process is complete, we are left with a clean dataset, which can be used for statistical analysis and visualization

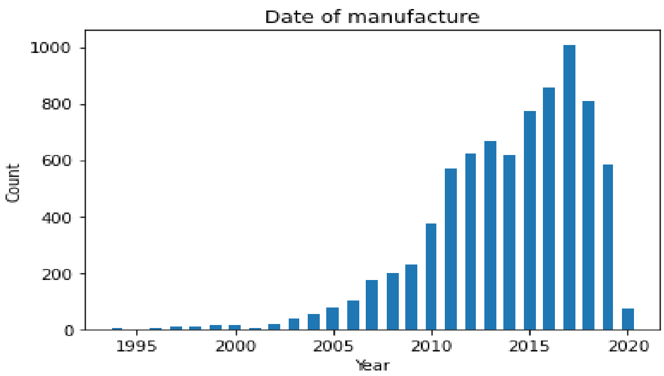
Name – Atul Ram

USN - 1BM19EC023

Distribution parameters

Year of Make

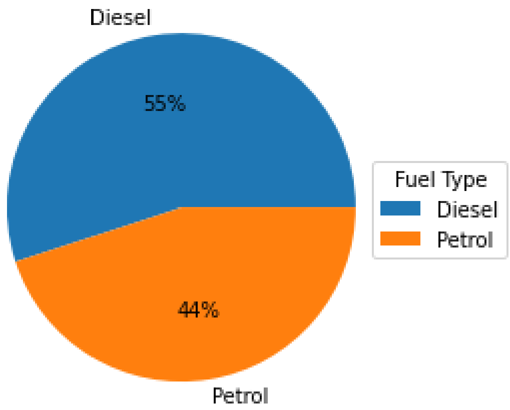
The graph shows the number of vehicles manufactured in a particular year, a sharp spike can be seen in the year 2010 and onwards, this suggests that people tend to lean towards buying newer cars, specifically not older than 10 years.



Fuel Type Percentage

Contrary to popular belief, it is seen that diesel vehicles are being resold more and have a higher market share in the resale domain, from further investigation it was found that most of these resale vehicles belong to “taxi” category, rather than regular house vehicles and SUV’s

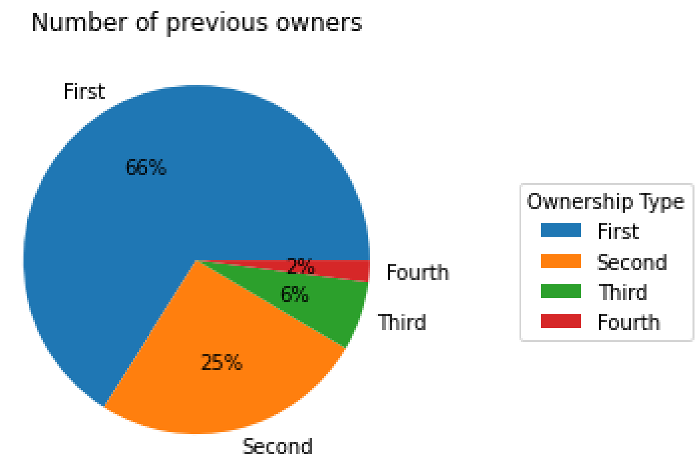
The percentage of vehicles running on CNG and LPG are less than 1%, hence have been omitted here.



Number of previous owners

Maximum number of vehicles being resold are the ones being resold for the first time.

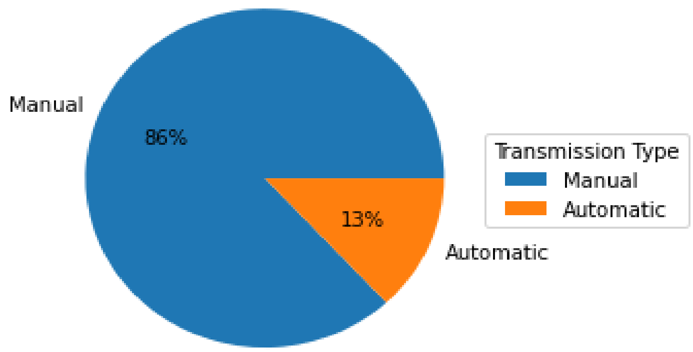
This leads us to believe that most of the owners of these vehicles tend to keep and use these vehicles till the end of their lifespans.



Transmission Type

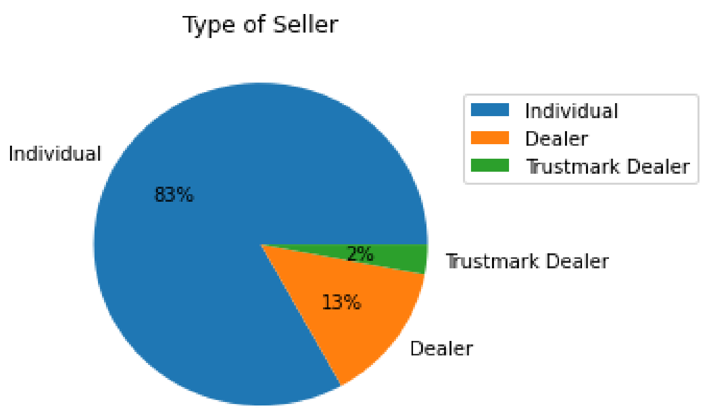
As the general trend would suggest, the number of manual vehicles drastically out scale the number of vehicles with automatic transmission.

Since they usually cost less, provide better mileage and have a lesser maintenance cost.



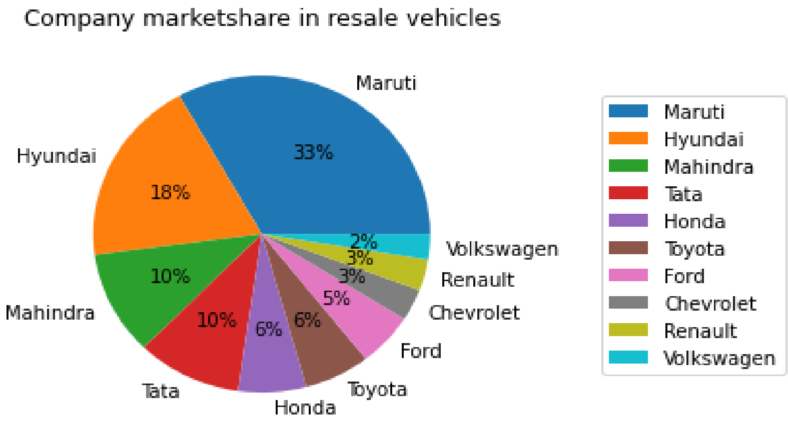
Seller Type

Types of Seller also has impact on resale as buying vehicles from owner's is preferred rather than buying from dealers. Buying vehicles from individual owner's generally provides a greater sense of security and trust.



Company Market Share

Companies such as Maruti, Hyundai, Mahindra, Tata alone accounts for more than 70% of market share. This shows that Company name (brand value) has huge impact on resale vehicles. Since these companies are better trusted their demand is always high in market.



Name – Amogha A Acharya

USN - 1BM19EC013

Relation between mean selling price and some of the features

Figure indicates the mean selling price of vehicle with respect to the number of times it has been resold.

It is clearly observed that a vehicle being resold for the second time is worth only half the price of the same vehicle being resold for the first time.

The reduction in value is lesser when a vehicle is resold more than 2 times.

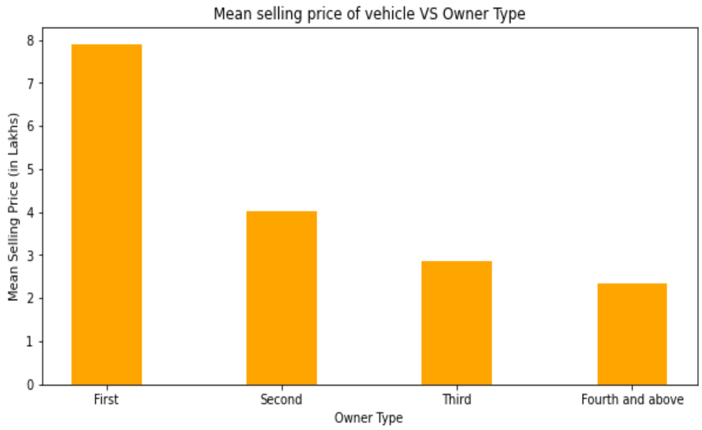
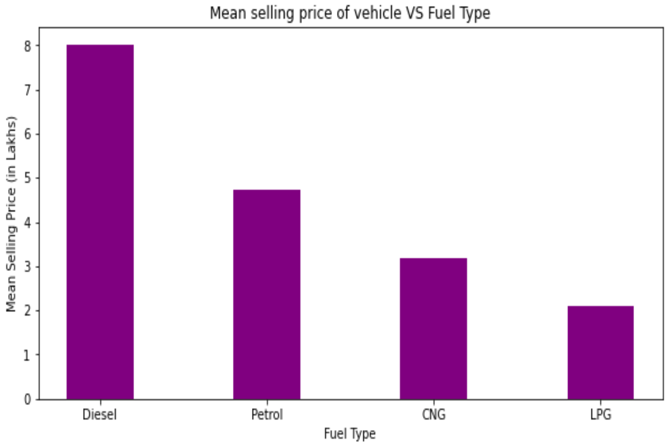


Figure shows mean selling price of the vehicle with respect to its fuel type

From the graph it is inferred that diesel vehicles cost nearly double the price of petrol vehicles, although there are other factors contributing to this difference in prices. The difference is still substantial.

CNG and LPG vehicles being the cheapest fuel types, they also have almost similar prices.



Descriptive analysis

Name - Anil R Devarakki

USN - 1BM19EC016

	year	selling_price	km_driven	transmission	mileage	engine	max_power	torque	seats
count	7889.000000	7.889000e+03	7.889000e+03	7889.000000	7889.000000	7889.000000	7889.000000	7889.000000	7889.000000
mean	2013.987831	6.496753e+05	6.919859e+04	0.131195	19.461709	1458.378628	91.588665	177.876157	5.418050
std	3.863460	8.134766e+05	5.682769e+04	0.337635	3.938527	503.299977	35.731275	92.962807	0.958526
min	1994.000000	2.999900e+04	1.000000e+00	0.000000	9.000000	624.000000	32.800000	47.000000	4.000000
25%	2012.000000	2.700000e+05	3.500000e+04	0.000000	16.780000	1197.000000	68.050000	111.000000	5.000000
50%	2015.000000	4.500000e+05	6.000000e+04	0.000000	19.330000	1248.000000	82.000000	170.000000	5.000000
75%	2017.000000	6.900000e+05	9.550000e+04	0.000000	22.320000	1582.000000	102.000000	209.000000	5.000000
max	2020.000000	1.000000e+07	2.360457e+06	1.000000	42.000000	3604.000000	400.000000	941.000000	14.000000

This Table describes the features of the dataset with the help of mean, min and max values, standard deviation and quartiles.

Mean year of manufacture is 2013, which says that most vehicles manufactured are around 2013.

Mean selling price is 6.5 lakhs, and average at about 5 seats, which suggests that most of the resold cars are mid segment vehicles, being driven an average of 69,000 kms.

The Mean mileage is 19.5 kmpl, once again proving the emphasis on fuel utilisation and mileage in Indian vehicles.

From this data it evident that all of these columns contain skewed data, with a mixture of both left and right skewed features. This may prove worry some during prediction leading to unexpected or subpar results.

Bivariate Analysis

Correlation Matrix

	year	selling_price	km_driven	transmission	mileage	engine	max_power	torque	seats
year	1.00	0.61	-0.40	0.15	0.40	-0.08	0.10	0.06	-0.00
selling_price	0.61	1.00	-0.18	0.26	0.01	0.40	0.59	0.51	0.26
km_driven	-0.40	-0.18	1.00	-0.13	-0.22	0.31	0.11	0.22	0.23
transmission	0.15	0.26	-0.13	1.00	-0.09	0.07	0.25	0.08	-0.06
mileage	0.40	0.01	-0.22	-0.09	1.00	-0.58	-0.38	-0.18	-0.48
engine	-0.08	0.40	0.31	0.07	-0.58	1.00	0.65	0.68	0.69
max_power	0.10	0.59	0.11	0.25	-0.38	0.65	1.00	0.74	0.30
torque	0.06	0.51	0.22	0.08	-0.18	0.68	0.74	1.00	0.41
seats	-0.00	0.26	0.23	-0.06	-0.48	0.69	0.30	0.41	1.00

- Year of manufacture, Engine capacity, Max_power, Torque are highly correlated with Selling Price. So these parameters are included for linear regression analysis.

- Transmission type, Number of seats, mileage are poorly correlated with selling price hence they are neglected and are not considered for training.

- km_driven is negatively correlated but poorly correlated hence can't be used to train model.

Features Considered for Model Development

Features such as year of make, max power of the vehicle and torque, were found to be the features with the best correlation with the selling price.

Hence these 3 features were considered for model development.

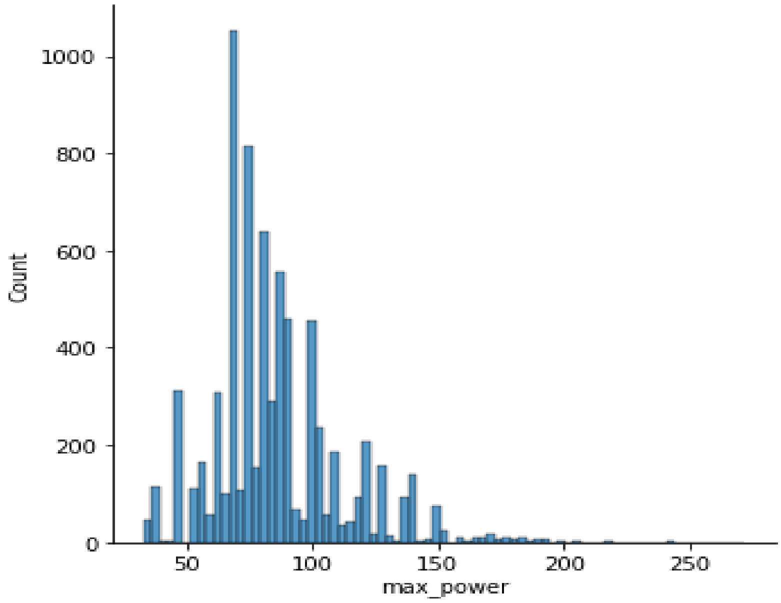
	year	selling_price	max_power	torque
year	1.00	0.61	0.10	0.06
selling_price	0.61	1.00	0.59	0.51
max_power	0.10	0.59	1.00	0.74
torque	0.06	0.51	0.74	1.00

Distribution Parameters

Name – Bhuvan DB

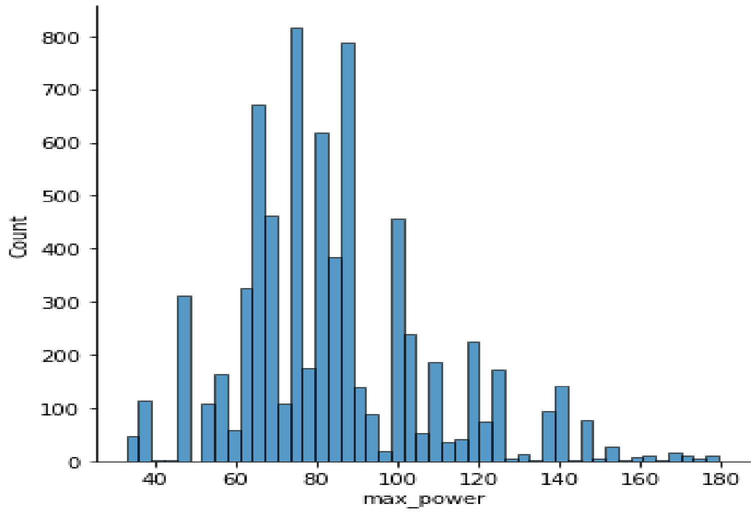
USN - 1BM19EC030

Distribution of max_power before removing outliers.



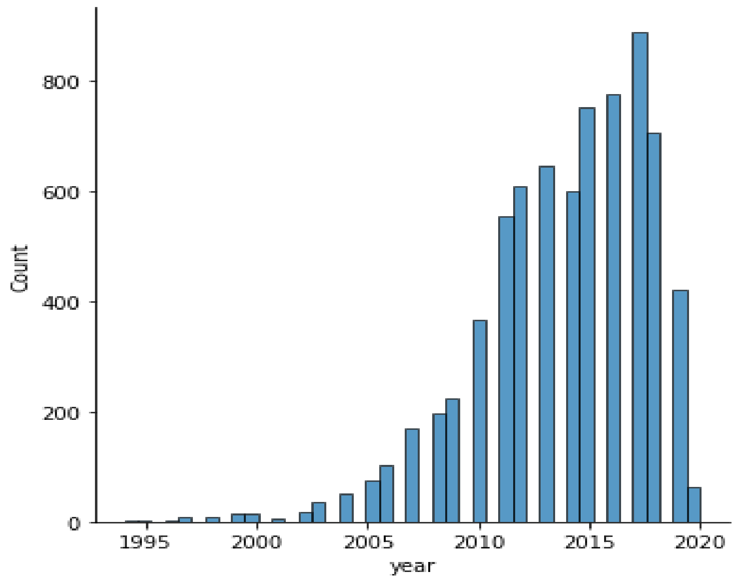
Here there are many outliers present in this distribution of max_power and skew value is found to be 1.16 which is pretty high.so in order to reduce skew remove vehicles having max_power>180 .after doing this the skew is reduced to 0.82 which is in the preferred range. Vehicles with power greater than 180 bhp usually are costlier, and second hand costlier vehicles are not in great demand in market.

Distribution of max_power after removing outliers.



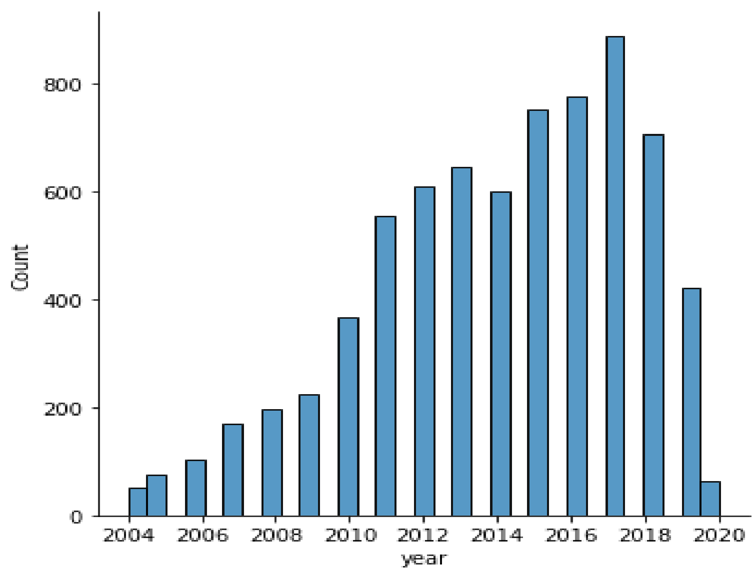
Year of manufacture

- Distribution of Year of manufacture before removing outliers.



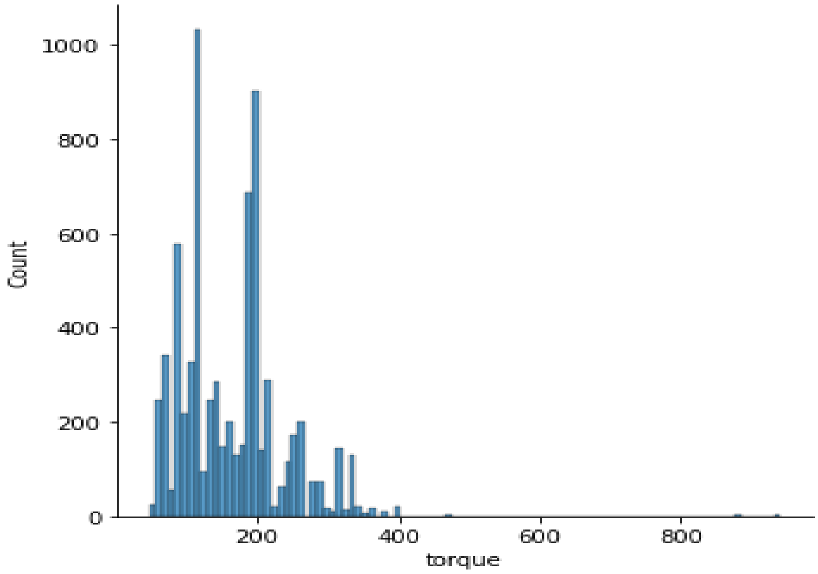
There are many outliers, which will effect in prediction by adding weights. This may cause increased errors. Skew value is found to be -0.97 which is not in preferred range**.** Normal Vehicles lifespan will be of 15 years and after that it will have lot of issues, Also vehicles manufactured before 2004 have less demand in market. Hence considering vehicles which are manufactured after 2004. Skew value is reduced to -0.57.

- Distribution of Year of manufacture after removing outliers.



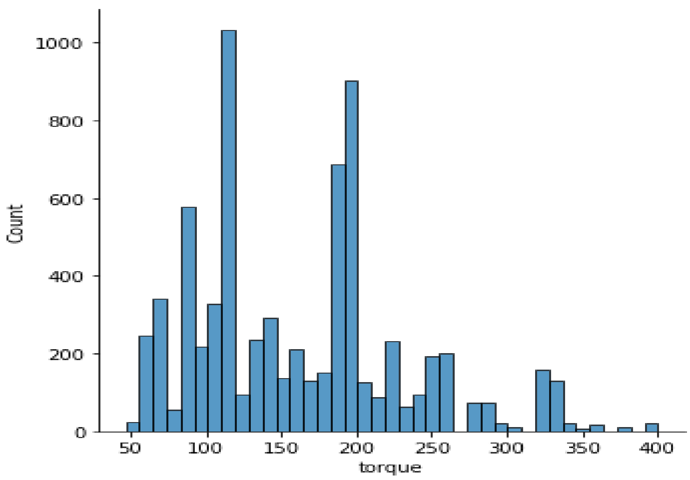
Torque

- Distribution of Torque before removing skew

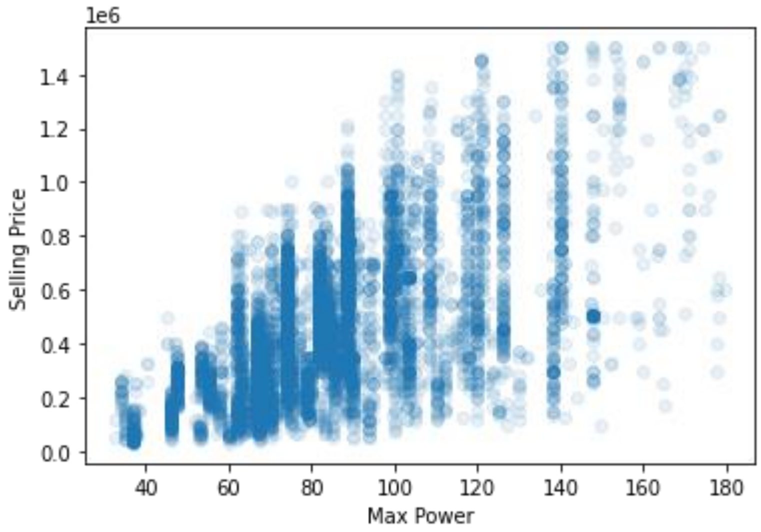


Torque distribution also has many outliers. Its skew value is found to be 1.57 which is high. To reduce this skew value remove vehicles having torque values higher than 400, skew gets reduced to 0.67 which is preferred. Vehicles having higher torque are usually sports car and they are not resold, hence removing those will help in betterment of model.

- Distribution of Torque after removing skew



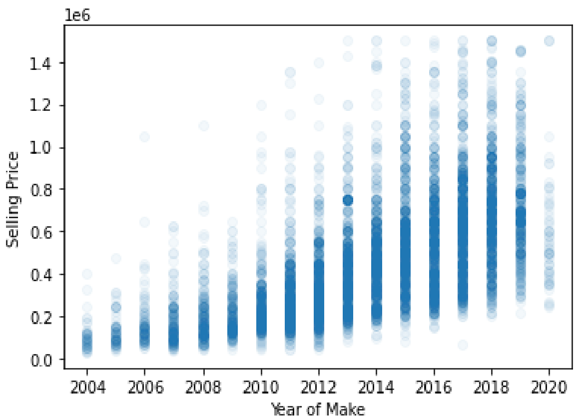
Prediction variable(Selling Price) and it's interaction with other parameters-

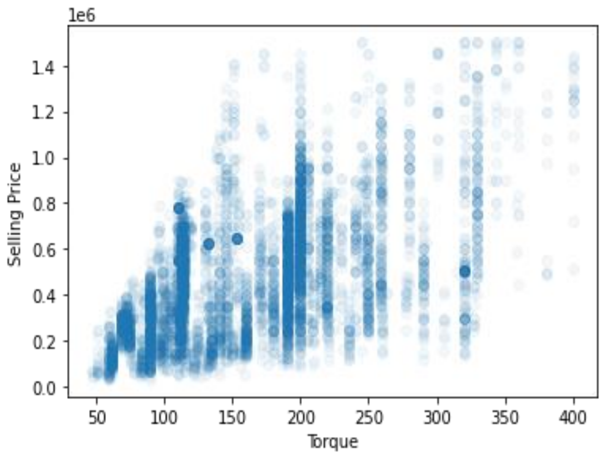


- Following plots shows that the selling price is very much dependent on max_power, Torque, Year of make.

- Vehicles manufactured in recent times are costlier, since they are longer lifespans.

- Vehicles having High Torque and power also have High Selling Price, but their demand is not much in the market as most don't prefer second hand cars having higher price.





Phase 2

Algorithm used in Model Building

Linear Regression: Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables they are considering, and the number of independent variables getting used.

Here we are using Multiple Linear Regression with 3 variables

A regression model involving multiple variables can be

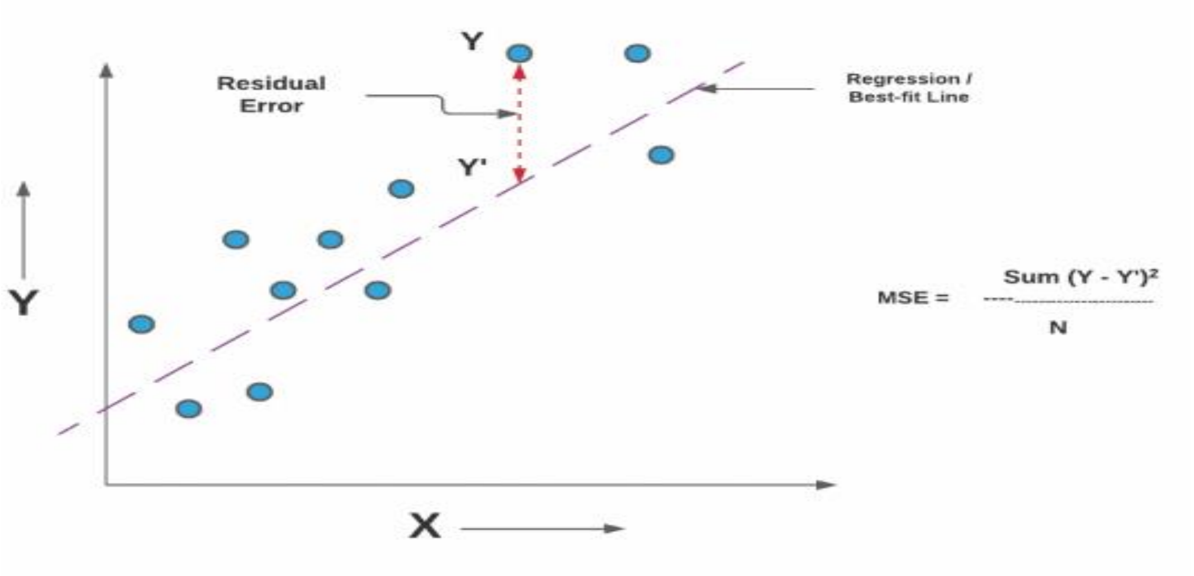
Represented as:

• $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \dots \dots \beta_n x_n$

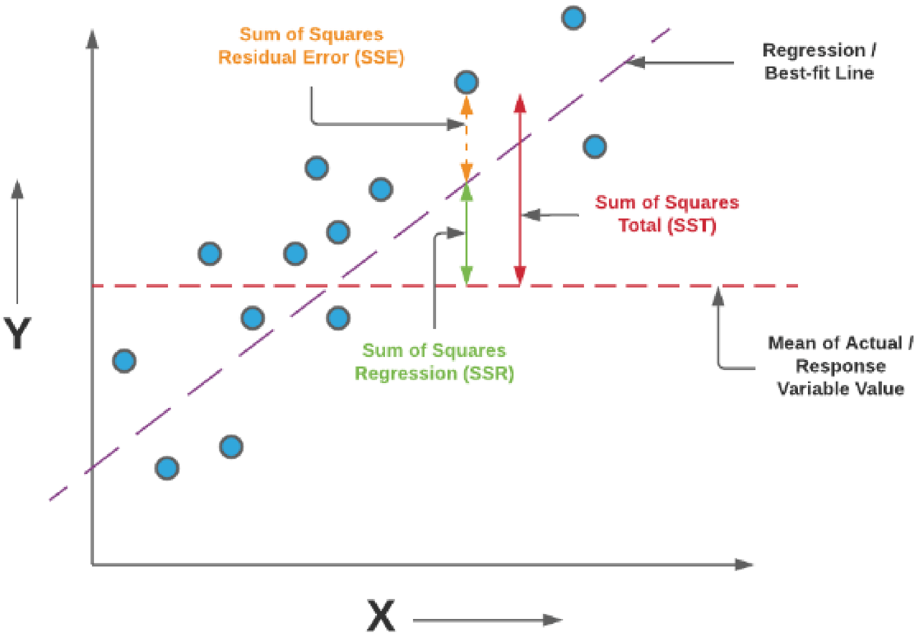
X1,x2,x3 as Year of manufacture, Torque and Max power.

Errors

Mean Squared Error - The average squared difference between the estimated values and the actual value.



R Squared Error - R-squared represents the fraction of variance of the actual value of the response variable captured by the regression model rather than the MSE which captures the residual error.



R-Squared = 1 – (SSE/SST)

Mean absolute error - It is a measure of errors between paired observations expressing the same phenomenon.

$$\mathrm{ME} = \frac{\sum_{i=1}^{n} y_i - x_i}{n}.$$

Here we have created 3 models for comparison and we have picked the best.

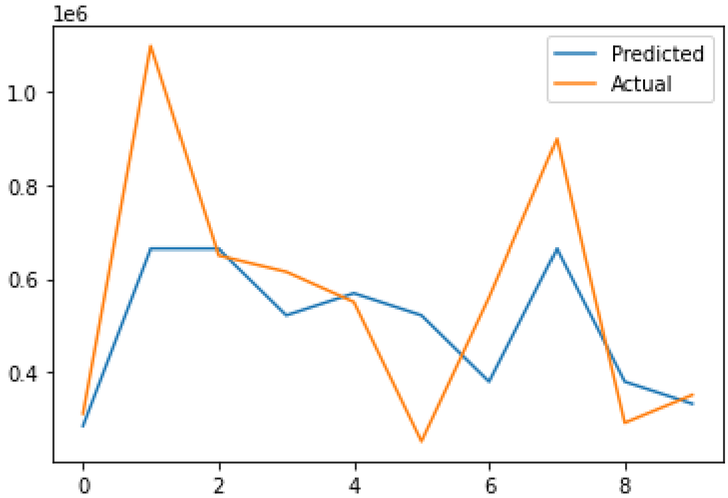
Models

Model 1

Linear regression using single variable

Independent Variable - Year of manufacturing
dependent Variable - Selling Price

Plot of actual vs predicted values after training



Mean absolute error = 154069.843

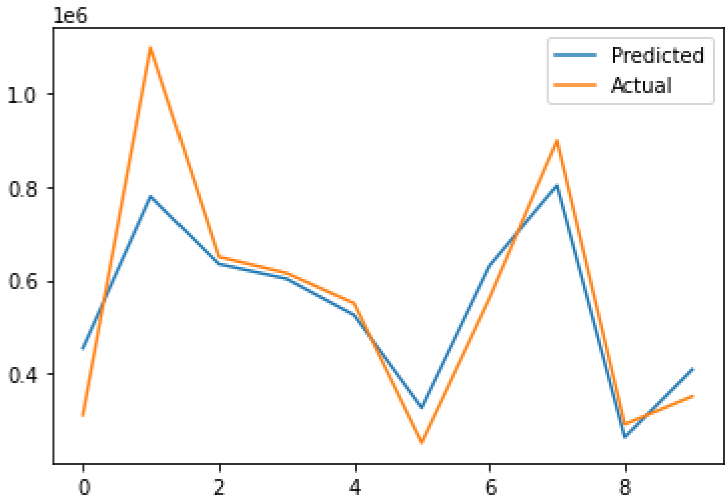
R Squared error = -0.31
Mean squared error = 0.4249

Model 2

Multiple Linear Regression (using the 3 most highly correlated features)

Independent Variable - Year of manufacturing, Max_power, Torque.
Dependent Variable - Selling Price

Plot of actual vs predicted values after training



Mean absolute error = 111531.211

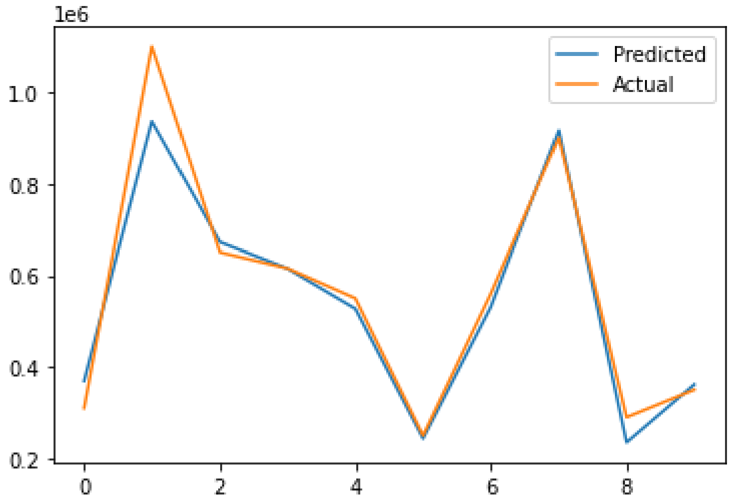
R squared error = 0.60

Mean squared error = 0.217

Model 3

Multiple Regression (nonlinear)

Polynomial scaled features are applied to get optimal model for prediction and this is the best model. Degree of polynomial is taken as '2'. Polynomial Regression.

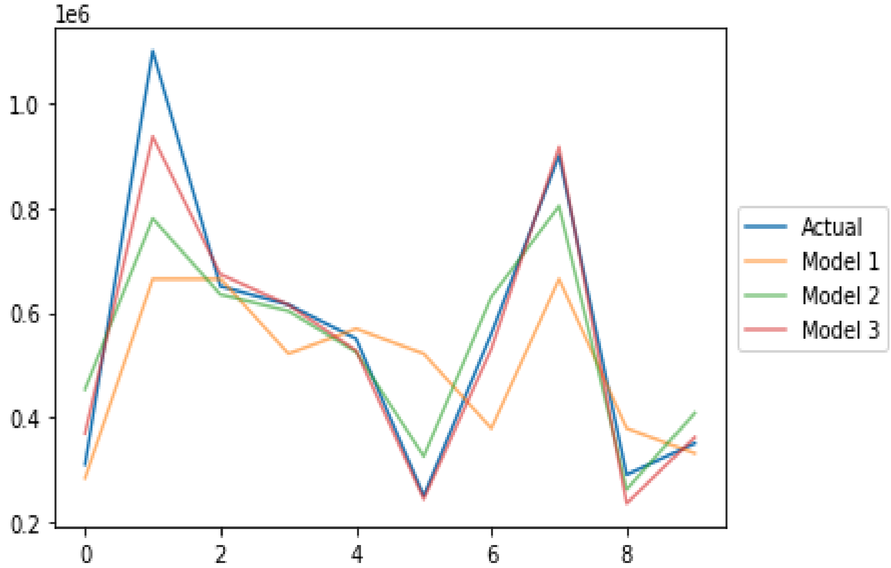


Mean absolute error = 88393.026

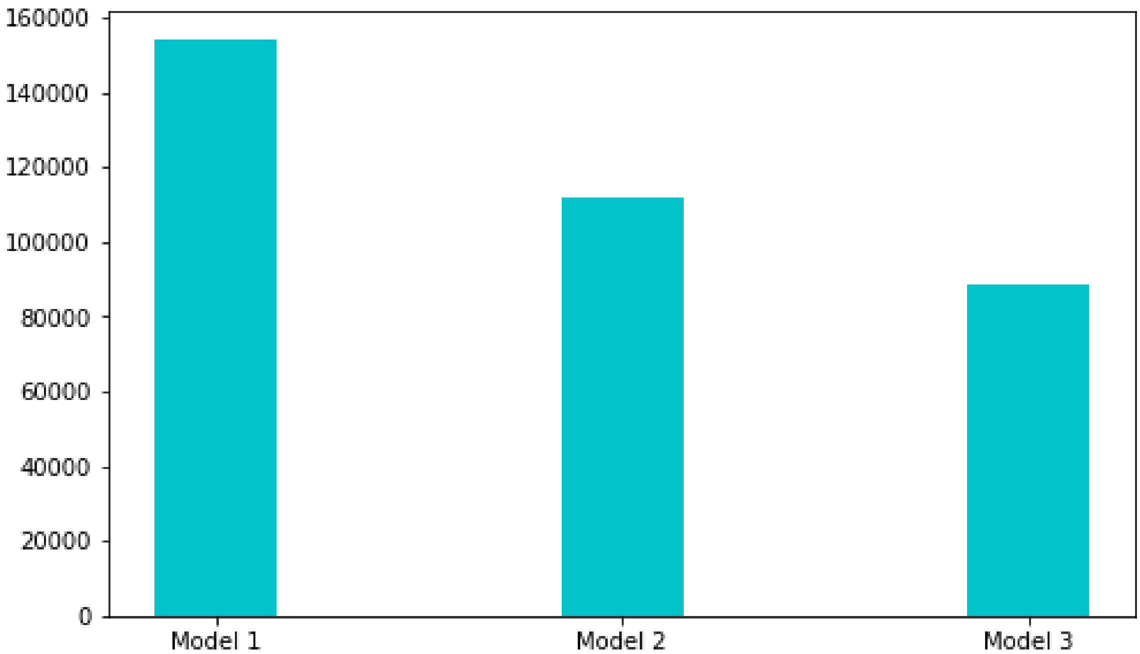
R squared error = 0.72

Mean squared error = 0.164

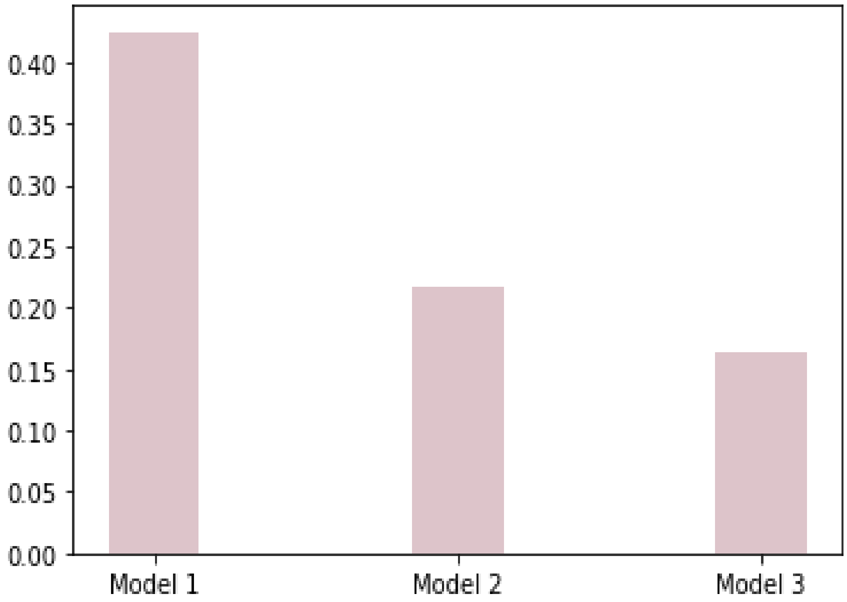
Comparing all 3 models



Variation of Mean Absolute error vs particular models.



Variation of mean squared error vs individual models



From above all models we can say that polynomial featured model fits best and hence it is used for predicting the outputs.

Results and Discussions

We have successfully predicted the Price of Vehicles based on the concept of Linear Regression. This model will be helpful in predicting the price.

In future there is a scope of increasing the variables which are highly correlated and thus will help in increasing the model accuracy, reducing errors.

Conclusion

Due to various features of the car and various models it’s difficult to predict the vehicle prices of used vehicles. In this model we have successfully been able to predict the prices of vehicles with good accuracy based on features such as Vehicle’s year of manufacture, Maximum Power and Torque since these are highly correlated compared to other features. By observing the scatter plots of various features we concluded that the best fit curve is not linear, its Polynomial We got the best fit model by considering a polynomial of degree 2. The final model selected which is the polynomial regression model has the least mean absolute error, mean squared error and it fits best.

References

[1] Pudaruth,S. 2014. “Predicting the Price of Used Cars Using Machine Learning Techniques”, International Journal of information & Computation Technology,4(7), p.753-764.

[2] Kuiper, S. 2008. “Introduction to Multiple Regression: How Much Is Your Car Worth?”, Journal of Statistics Education, 16(3).

[3] Listiani M. 2009. Support Vector Regression Analysis for Price Prediction in a Car Leasing Application. Master Thesis. Hamburg University of Technology.

[4] Limsombunchai, V. 2004. House price prediction: Hedonic price model vs. artificial neural network. In New Zealand Agricultural and Resource Economics Society Conference, New Zealand, pp. 25-26.

[5] Bourassa, S.C., Cantoni, E. and Hoesli, M. 2007. “Spatial dependence, housing submarkets, and house price prediction”, The Journal of Real Estate Finance and Economics, 35(2), p.143-160.