CAR PRICE PREDICTION

Submitted by:

Akash chaudhary

Intenship 17

ACKNOWLEDGMENT

The project car price prediction is a problem with different prices of cars on the behalf of the different features of the car. The data frame was collected by myself from the cartrade.com using selenium and beautiful soup. My mentor Mr. keshav bansal help me a lot to address the problems and solutions.

INTRODUCTION

• Business Problem Framing

Transportation and private cars having the biggest market in the world everyone wants their own car either new or old. It's the biggest problem to find the exact price of the used car. Whenever a new company want to comes in the existing market or some individual wants to buy a used car they also want to know the exact price of the used car and also tring to buy it in low costs. The used car price could be vary on many factors like location, transmission, number of owners, kilometres the car have runned, fuel types, cars brand, model and variant of the car, manufacturing year etc. so because of these many factors the prices changes everytime. So the consumer or company stuck and confused to took the idea of the prices of used car but by using machine learning algorithms this problem could be solved.

• Conceptual Background of the Domain Problem

With the covid 19 impact in the market, we have seen lot of changes in the car market. Now some cars are in demand hence making them costly and some are not in demand hence cheaper. One of our clients works with small traders, who sell used cars. With the change in market due to covid 19 impact, our client is facing problems with their previous car price valuation machine learning models. So, they are looking for new machine learning models from new data. We have to make car price valuation model. This project contains two phases.

Data Collection Phase

You have to scrape at least 5000 used cars data. You can scrape more data as well, it's up to you. more the data better the model

In this section You need to scrape the data of used cars from websites (Olx, cardekho, Cars24 etc.) You need web scraping for this. You have to fetch data for different locations. The number of
columns for data doesn't have limit, it's up to you and your creativity. Generally, these columns are Brand, model, variant, manufacturing year, driven kilometers, fuel, number of owners, location and at last target variable Price of the car. This data is to give you a hint about important variables in used car model. You can make changes to it, you can add or you can remove some columns, it completely depends on the website from which you are fetching the data.
Try to include all types of cars in your data for example- SUV, Sedans, Coupe, minivan, Hatchback.
After collecting the data, you need to build a machine learning model. Before model building do all data pre-processing steps. Try different models with different hyper parameters and select the best model.

- **Review of Literature**

Own car is one of the necessary need of each and every person around the globe and therefore used and new cars market is one of the markets which is one of the major contributors in the world's economy. It is a very large market and there are various companies working in the domain. Data science comes as a very important tool to solve problems in the domain to help the companies increase their overall revenue, profits, improving their marketing strategies and focusing on changing trends in car sales and purchases. Predictive modelling, Market mix modelling, recommendation systems are some of the machine learning techniques used for achieving the business goals for used cars companies. Our problem is related to one such used car company.

- Motivation for the Problem Undertaken

As of time is changing digitalisation is also increasing everyone finds comfort in all the thing and now the cars are become the best comfort zone for everyone because maybe everyone wants to have their own car either new or old.

Everyone switching their car after a period of time that's why cars market is a growing business so it make my analytical skills strong and the market can demand this type of analysis. It also enhance my personal skills. To make this type of project and being the perfectionist in such type of work could help the companies and car agents and concumer to getting the idea of the price of used car.

Analytical Problem Framing

- **Mathematical/ Analytical Modeling of the Problem**

The mathematical problems are included are this was the huge data so getting null values is not a big deal. So there is some null values we need to tackle with them because of huge data there are many rows i.e. 5350 rows first of all we have gather the data using web scrapping selenium and beautiful soup for collecting data we need to understand which information we will need to scrape than understand which will work for us to make prediction model for prices. After these we need to make the dataframe and check the data is balanced or not. Checking for outliers adjust them and some other works are accured at the time of making the model and collecting the data through web using web driver. We used random forest regressor to predict the prices and putting them into the test data to predict the prices of the used car.

- Data Sources and their formats

The data is gathered by web scraping from the cartrade.com website. The project was assigned to me by mr.sajid chaudhary.

The data was gathered by me from cartrade.com for making the machine learning model to predict the prices for entering into the new market of used cars. The data contains 5024 rows and 9 columns. The data includes both in integer and string forms.

```
df
```

t[3]:

	Unnamed: 0	car_brand	car_name	manufacturing_year	location	fuel_type	kilometers_driven	number_of_owners	price
0	0	Maruti Suzuki»	Celerio»	2016	Kolkata	Petrol	45386	First	3.29 Lakh
1	1	Honda»	Mobilio»	2015	Pune	Petrol	42492	First	5.75 Lakh
2	2	Maruti Suzuki»	Swift»	2009	Hyderabad	Petrol	69298	First	3.11 Lakh
3	3	Honda»	City»	2016	Delhi	Petrol	40138	Second	5.75 Lakh
4	4	Ford»	Ecosport»	2013	Mangalore	Diesel	21871	Second	6.25 Lakh
...	...	...	...	...	...	...	...	...	...
5019	5019	Mercedes-Benz»	S-Class»	2016	Delhi	Petrol	5800	First	73.5 Lakh
5020	5020	Kia»	Carnival»	2020	Lucknow	Diesel	12000	First	27 Lakh
5021	5021	Skoda»	Superb»	2015	Delhi	Petrol	57000	First	11.75 Lakh
5022	5022	Honda»	Mobilio»	2015	Dehradun	Petrol	74000	First	4.9 Lakh
5023	5023	Maruti Suzuki»	Ciaz»	2015	Bangalore	Petrol	74541	First	6.99 Lakh

5024 rows × 9 columns

- **Data Preprocessing Done**

First of all we got data the data in excel form so we import this into our model. we need to clean up the data although its almost cleaned because we have gathered it by ourselves so we just added only those columns which we need the most so first we need to clean the data in the data there are 5024 rows and 9 columns inclusive of nan values and an unnamed column which comes by default. So we decide to drop the unnamed numeric column than fill rest of null values with different mean and mode technique. To clean the data. Than we correlate the targeted variable with some important independent variables.

- **Data Inputs- Logic- Output Relationships**

As data was gathered by ourselves we used selenium and web driver to get the data. We have used a single code to collect the data from all the 230 pages and make the dataframe. So these codes were the inputs. And output going to be came after the making the prediction model by using the machine learning algorithms. So we have split the data into train and test dataset.we need to train the model on the behalf of gathered data and put them in the test data.

So we have done all the data cleaning steps to the test data too. After getting the best model put that model on test.

```
# Activating the chrome browser
driver = webdriver.Chrome("chromedriver.exe")
time.sleep(3)

# Opening the homepage of carwale
url = "https://www.cartrade.com/buy-used-cars/#sc=-1&so=-1&pn=1"
driver.get(url)
time.sleep(3)
```

```
car_url=[]
#scrapping the required details
start=0
end=100
for page in range(start,end):#for loop for scrapping 4 page
    url =driver.find_elements_by_xpath("//h2[@class='h2heading truncate']//a")
    for i in url:
        car_url.append(i.get_attribute("href"))
    nxt_button=driver.find_elements_by_xpath("//li[@class='next']//a")
    try:
        driver.get(nxt_button[1].get_attribute('href'))#getting the link from the list for next page
    except:
        driver.get(nxt_button[0].get_attribute('href'))
```

- **State the set of assumptions (if any) related to the problem under consideration**

Presumptions are includes the data was only for private used cars because there was not any commercial car like which are used in ola uber or any taxi service.

- **Hardware and Software Requirements and Tools Used**

We have used jupyter notebook to make the model and imported some python libraries these are includes numpy, pandas, maths, stats, seaborn, matplotlib, sklearn etc.

Model/s Development and Evaluation

- ## Identification of possible problem-solving approaches (methods)

It was a regression problem that's why we have used linear regression. We have used standard scaler to scale the data and also balancing of the data. We have used random forest regressor because it gives as the best r2 score and its cross validation score was also good. And the difference between accuracy and cross validation score is very less. So we took this and also done the hyper parameter tunning to tune the data.

- ## Testing of Identified Approaches (Algorithms)

Algorithms which are used for training and testing are:

Linear regression

Ada boost regressor

Decision tree regressor

Random forest regressor.

- ## Run and Evaluate selected models

There are the snapshots of the models that we used:

1. Linear regressor

```
#getting the MAE,MSE,RMSE for linear model
lr=LinearRegression()
lr.fit(x_train,y_train)
pred=lr.predict(x_test)

print('Mean absolute error:',mean_absolute_error(y_test,pred))
print('Mean squared error:',mean_squared_error(y_test,pred))
print('Root Mean squared error:',np.sqrt(mean_squared_error(y_test,pred)))
```

```
Mean absolute error: 1.1336694648168211e-09
Mean squared error: 1.933593527954184e-18
Root Mean squared error: 1.3905371364886966e-09
```

```
#getting the r2 score for our model
print('r2 score is:',r2_score(y_test,pred))
```

```
r2 score is: 1.0
```

2. Random forest regressor.

random forest regressor

```
: #getting best r2 score using Random Forest Regressor
from sklearn.ensemble import RandomForestRegressor

rfr=RandomForestRegressor()
rfr.fit(x_train,y_train)
print('r2_score is',rfr.score(x_train,y_train))
pred=rfr.predict(x_test)
```

```
r2_score is 0.9999832963029573
```

3. Decision tree regressor

Decision Tree Regressor

```
64]: #getting best r2 score using Decision Tree Regressor
from sklearn.tree import DecisionTreeRegressor

dtr=DecisionTreeRegressor()
dtr.fit(x_train,y_train)
print('r2_score is',dtr.score(x_train,y_train))
pred=dtr.predict(x_test)
```

```
r2_score is 1.0
```

4. Ada boost regressor

ada boost regressor

```
: #getting best r2 score using ada boost regressor
from sklearn.ensemble import AdaBoostRegressor

ada=AdaBoostRegressor()
ada.fit(x_train,y_train)
print('r2_score is',ada.score(x_train,y_train))
pred=ada.predict(x_test)
```

```
r2_score is 0.9904569211864629
```

So we have chose the decision tree regressor because of its best fit r2 score and cross validation score.

- **Key Metrics for success in solving problem under consideration**

The key metrics which are used are:

R2_score

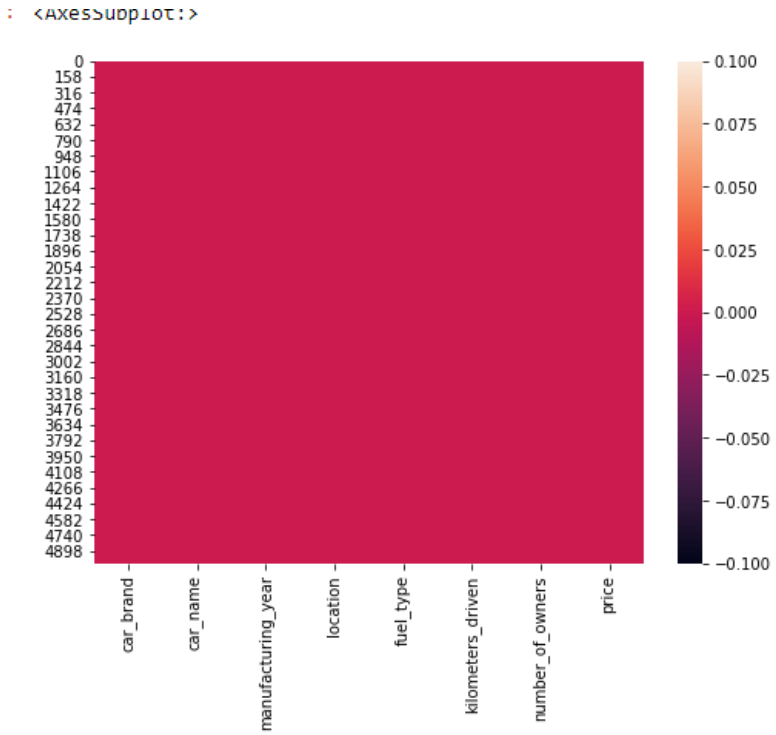
Cross validation score

And we check the cross val score. Because the difference between the fit score and cross validation score indicates models accuracy.

- **Visualizations and Interpretation of the Results**

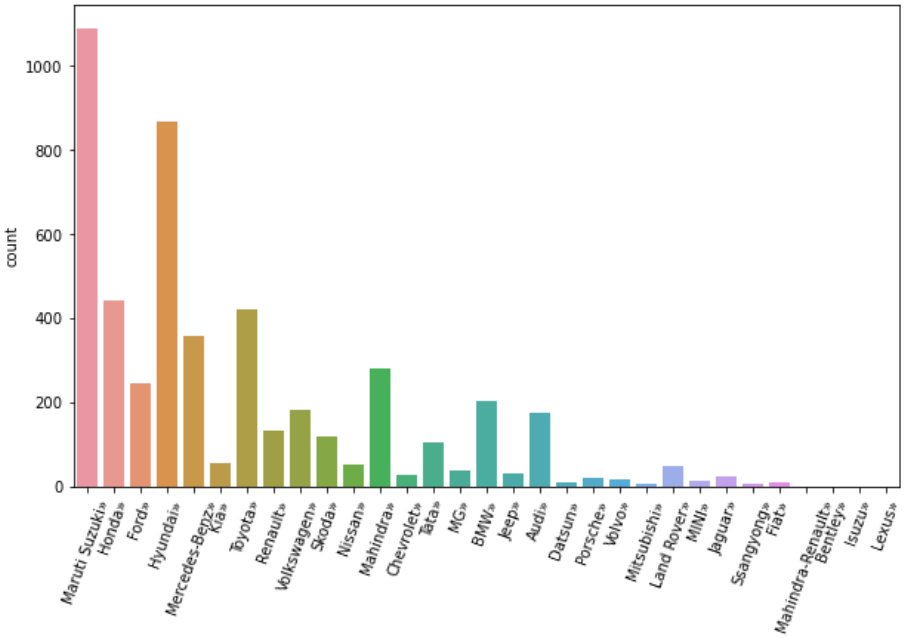
The visualizations are includes:

1. Heat map for null values



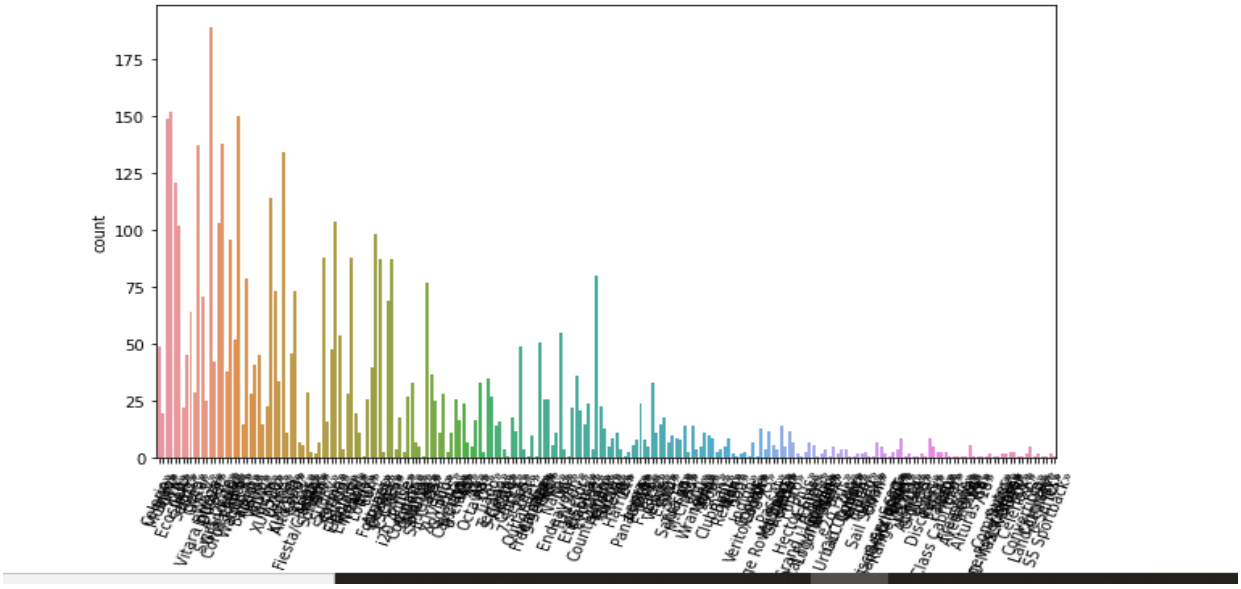
the dataset is free of null values.

2. Countplot for car brands.



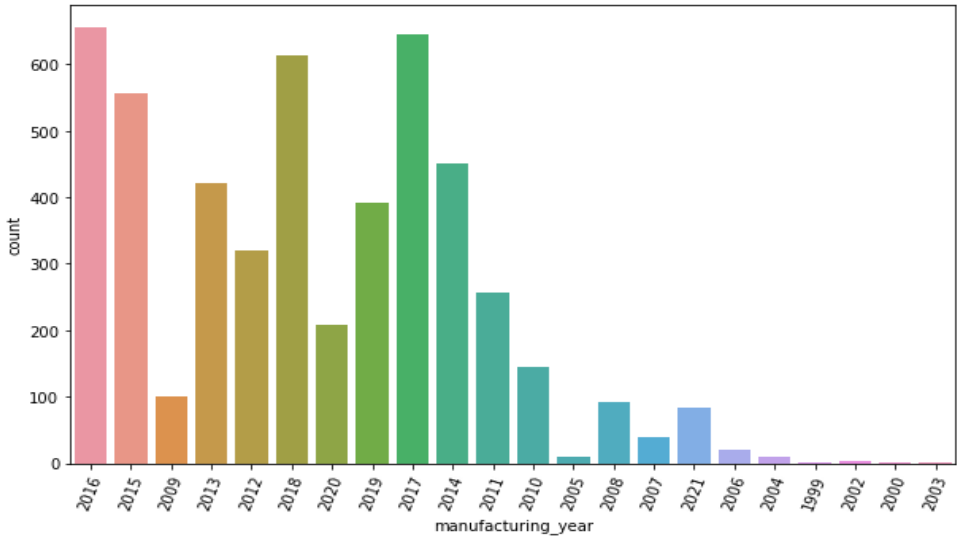
we can clearly see the market is demanding the maruti, hyundai car. mostly these brands are on sale and honda,toyota, mahindra have average contribution. Lexus , ishuzu, fiat are very rarely comes on sale.

3. Countplot based on car names.



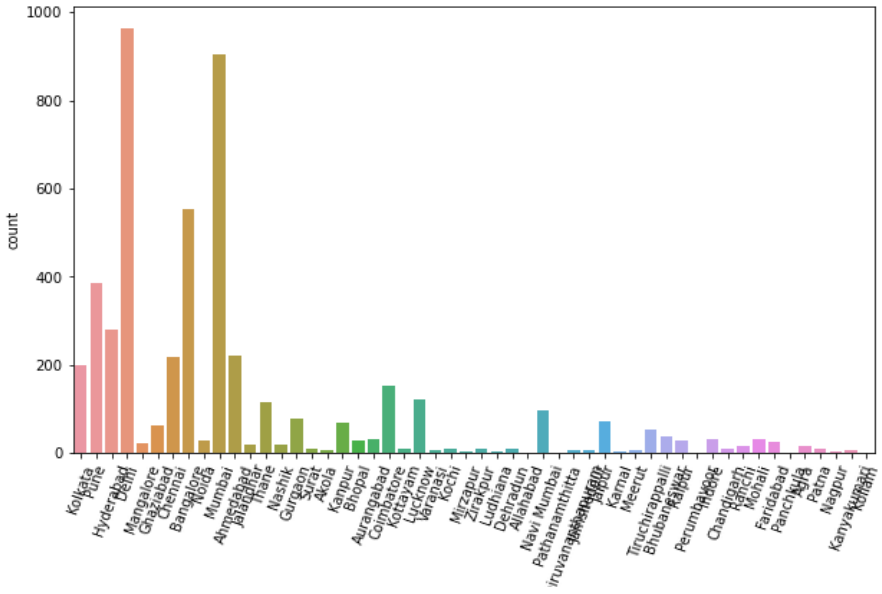
we can see there are maximum innova, city, wagonr, shift, grand, xuv etc are on sale. some models like baleno, i20, vitara brizza, have average numbers of sale. and also some types of car available who have only one car available for sale.

4. Countplot on the behalf of manufacturing year.



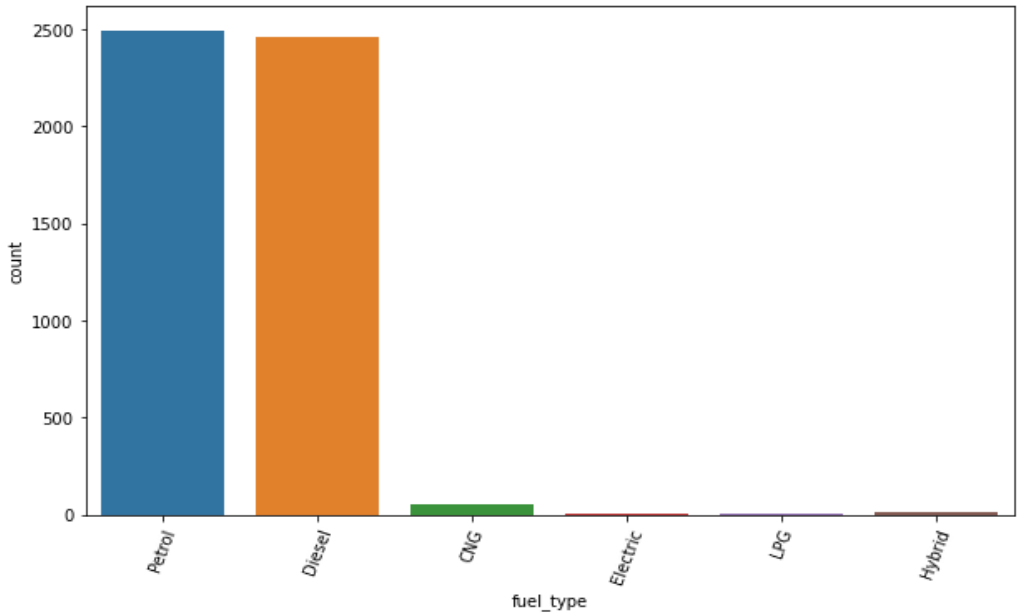
most of the cars are 4-6 years old those manufacturing years are between 2013 to 2018. very few cars of 1999 to 2004 are on sales maybe the reason is govt ban the 20 years old cars.

5. Countplot on the behalf of cities.



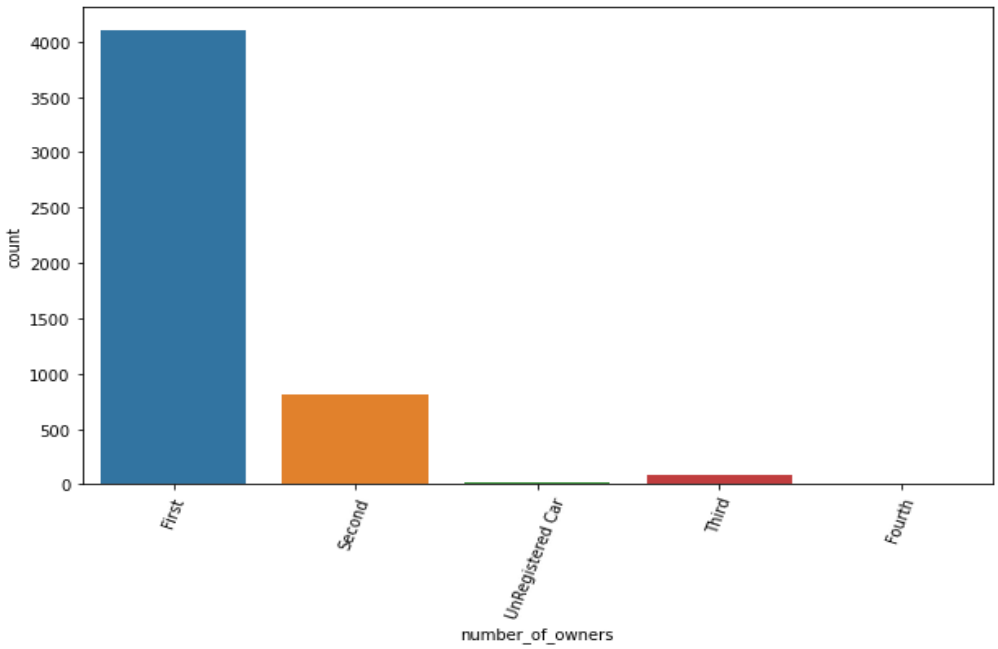
we can see most of the cars are from delhi, mumbai, banglore which comes for sale. than pune, ahemdabad, hyedrabad and chennai having the average nnumbers of car for sale than some cities have very minium numbers of car which are on sales maybe there are people oftenly uses car these cities are allahabad, kollam, karnal etc.

6. Countplot on the behalf of fuel types.



maximum petrol and diesel cars of the dataset are on sales. maybe in india people choose petrol and diesel car varients. after diesel and petrol cng cars are comes on sale. very few electric, LPG and hybrid cars are used in india.

7. Countplot on the behalf of number of owners.



most of cars in the dataset have only one owner right now. and some car like 15% cars have 2 owners before it comes for sale than third and fourth.

CONCLUSION

- ## Key Findings and Conclusions of the Study

The cars market is the biggest market in the world whether used or new and estimating the pricing or the used car is the toughest thing so we have used the linear regression because it's a regression problem. And used the different machine leaning algorithms to predict the prices of the used car on the bases of the previous data and the factors which effect the price of the used cars like manufacturing year, number of owners, fuel type, car brand , cars model, variant, transmission etc.

The conclusion is many factor as I mentioned earlier could effect the prices of the used in different ways. but the market demand of used cars are mostly depends upon the manufacturing year and the consumers demands the model between 2015 to 2018 manufactured cars and they also have more demand and there prices are also little high as compared to old model cars like which are made before 2015. Second factor is fuel type petrol car are more demanding as compared to diesel perhaps because of the permite. Number of owners also impacts on the price of car as first owner cars prices are little high as compared to 2^{nd} and 3^{rd}. brands are also the biggest factor because market and our dataframe have more Hyundai and maruti car so in the comparission of others the price of these cars are also high because of demand. As we all the the price and the demand are always equivalent if demand of a product increases then the price of the same product also increase.

- ## Learning Outcomes of the Study in respect of Data Science

Because the data is huge the number of rows and columns are also high so these make my understanding of data increased. I have tried many things to clean up the data than find which variables are used for prediction and make the model powerful I have correlate many useful independent variables with the targeted variable so that I can understand the data more. After visualisation I have correlate the data with target in numbers to understand which independent variables gives how much impact on the

targeted variable. So after all of these I used standard scaler to scale the data after scalling the data train the model than used the different machine learning models. And get the best score in random forest regressor than doing hyper parameter tuning etc. these steps improve my machine learning and model building skill. And these model surely will help the consumers, companies and used car agents to estimate the price of a particular car.

- **Limitations of this work and Scope for Future Work**

Because we get the good r2 score and we have choose the best one model. We can assume the estimated price are almost correct. But we all know cars prices are very dynamic it could change with many other factors which we didn't scraped and took in our dataframe like variant of the car as new cars base model is comparably cheaper than the top model and transmission too. So we can took the estimated prices by using these model but also need to see the market trend.