**[Prepared Solution FIT for Car Resale Value Prediction.](https://akhilanandkspa.medium.com/?source=post_page-----3dc412d24aa0--------------------------------)**

[Anand](https://akhilanandkspa.medium.com/?source=post_page-----3dc412d24aa0--------------------------------)

Dec 28, 2020

·

10 min read

·

Listen

**End-to-End Project on Used Car Price Prediction**



[Source](https://www.google.com/search?q=neeraj+kumar+pic+with+renault+duster&tbm=isch&ved=2ahUKEwjgmZS09OjtAhUIIisKHczuDiIQ2-cCegQIABAA&oq=neeraj+kumar+pic+with+renault+duster&gs_lcp=CgNpbWcQAzoECCMQJ1DtkwFYoLIBYKyzAWgAcAB4AIAB_gGIAc8WkgEGMC4xNy4ymAEAoAEBqgELZ3dzLXdpei1pbWfAAQE&sclient=img&ei=9bzlX6DLDojErAHM3buQAg&bih=625&biw=1366#imgrc=e12zkXTXgPw95M)

**Overview**

This dataset consists information about used car listed on cardekho.com. It has 9 columns each columns consists information about specific features like **Car\_Name** gives information about car company .which **Year**the brand new car has been purchased.**selling\_price**the price at which car is being sold this will be target label for further prediction of price.**km\_driven**number of kilometre car has been driven.**fuel**this feature the fuel type of car (CNG , petrol,diesel etc).**seller\_type** tells whether the seller is individual or a dealer. **transmission** gives information about the whether the car is automatic and manual.**owner** number of previous owner of the car.**Present\_price** what is the current showroom price of the car.

**Step 1 : Setting a virtual environment**

This should be initial step when you are building an end to end project.We need new virtual environment because each project required different set and different version of libraries so by making an individual environment for a specific project we can feed all the essential library to that environment.follow these step to do so….

conda create -n carfare python=3.6  
#some essential package for the environment will be installed #automatically then you will get option  
[y/n] ---> y #click y

After that we will activate the enviorment by using

>> activate carfare  
>> jupyter notebook #run jupyter notebook on newly created env

It might be possible that we will get issue about absence of jupyter notebook for that we have to do one more step

>> pip install jupyter notebook # installing jupyter notebook on env  
>> jupyter notebook

our enviorment is created , now we will do our complete project on this enviorment.

**Step 2 :- Acquiring data set and importing all the essential library**

I have taken data set from [**here**](https://www.kaggle.com/nehalbirla/vehicle-dataset-from-cardekho?select=car+data.csv) & The data set is in csv format.Now i will import all the essential library that will be needed for this project.

import pandas as pd  
import numpy as np  
import matplotlib.pyplot as plt  
import seaborn as sns  
from sklearn.preprocessing import StandardScaler  
from sklearn.model\_selection import train\_test\_split

**Step 3 :- Data Pre Processing**

df=pd.read\_csv("car data.csv")  
df.head() #Printing first five rows

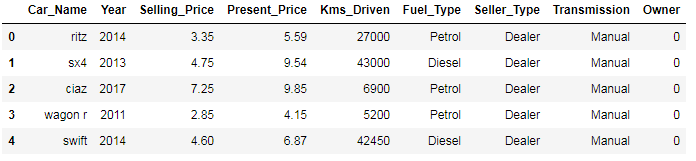


Figure 1

len(df["Car\_Name"].unique())[out]>> 98

Here we have cars of 98 different companies and name of companies won’t affect car’s price ,price depends upon how many year it’s been used ,fuel type etc,.so i will drop the column **car\_name** from original dataframe.

df.shape #number of rows and columns present in a dataset[out] >> (301, 9)--------------------------------------------------------------------  
df.columns #printing index of all the columns[out]>>Index(['Car\_Name', 'Year', 'Selling\_Price', 'Present\_Price', 'Kms\_Driven','Fuel\_Type', 'Seller\_Type', 'Transmission', 'Owner'],  
 dtype='object')-------------------------------------------------------------------  
#dropping the car\_name column  
df.drop("Car\_Name",axis=1,inplace=True)--------------------------------------------------------------------  
df.isnull().values.any() #is there any null value present [out]>> False # there is no null value present in dataset

Now we will check the data type of each column if the data type is numerical then we have no such issue but if datatype is categorical then we need to convert all those categorical features into numerical values.

df.dtypes

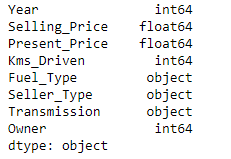


Figure 2

If we will observe above output we can say that there are some features which are having object data type now in my next step i will create a **cat\_df** and will store all the categorical feature into **cat\_df**.

list1=[] #storing all the features having categorical datatype  
for i in df.columns :  
 if df[df[i]=="object"]:  
 list1.append(i)[out]>> ['Fuel\_Type', 'Seller\_Type', 'Transmission']

Now we will make the categorical dataframe with all the features having categorical variables, and will drop all the categorical features from original dataframe.

cat\_df=df[list1]  
cat\_df.head() #top five rows of cat\_df

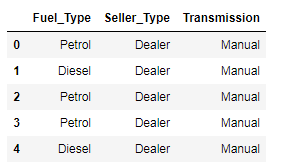


Figure 3

df.head()

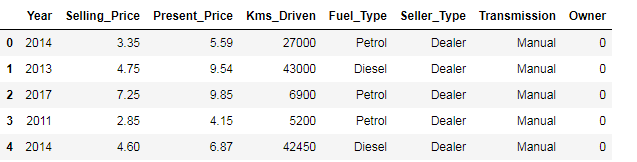


Figure 4

Year represent the year in which car have been purchased so how we can estimate the number of year car has been used ?

number of year car has been used = current year — previous year

So as we are in 2020 and car is of 2014 then number of years it’s been used will be :- number of year car has been used=2020–2014=6 to print number of year car has been used we need to add a column which represent current year.

df["Current\_Year"]=2020

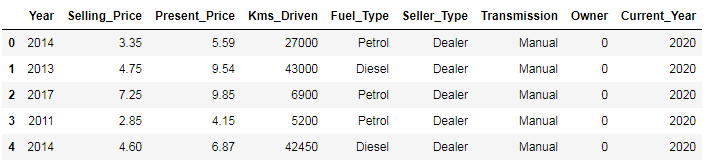


Figure 5

We have successfully added the **current\_ year**now we will add **number\_of\_years** column and drop **Year**and **Current \_year**column.

df["No\_of\_years"]=df["Current\_Year"]-df["Year"]  
df=df.drop(["Current\_Year","Year"],axis=1)  
df.head()

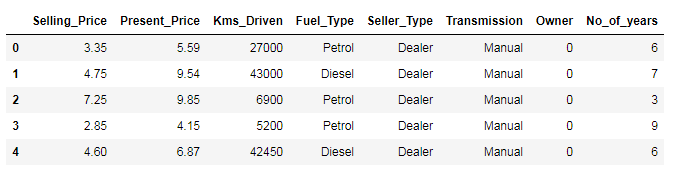


Figure 5

Getting statistical description about data

df.describe()

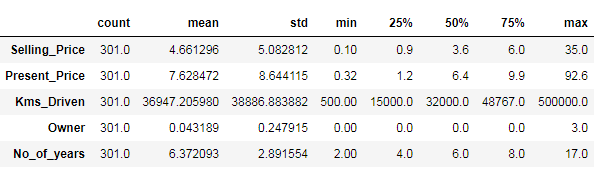


Figure 6

On an average car has been driven 36947 kilometres and max distance the car has been traveled is 5,00,000 kilometres. The car with highest ex-showroom selling price present in data set is 92.6 lakh. Maximum number of years car has been used and then come for sell is 17 years.maximum number of owner that has used a single car is 3.Maximum selling price for used car is 35 lakh rupees. This is how we make conclusion with statistical description of dataset.

**Step 4 :- Data Visualization**

This is most important step of data science life cycle, here we understand the behavior of data and try to make certain meaningful insight out of it. let’s understand it by doing…

sns.set\_style("darkgrid")  
sns.FacetGrid(df,hue="No\_of\_years",height=6).map(plt.scatter,"Present\_Price","Selling\_Price").add\_legend()  
plt.show()

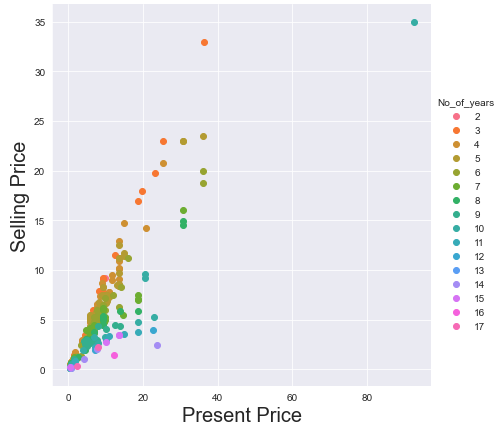


Figure 7

More number of Years you will use your car lesser the amount you will get.

sns.set\_style("darkgrid")  
sns.FacetGrid(df,hue="Present\_Price",height=6).map(plt.scatter,"Kms\_Driven","Selling\_Price")  
plt.xlabel("Present Price",fontsize=20)  
plt.ylabel("Selling Price",fontsize=20)  
plt.show()

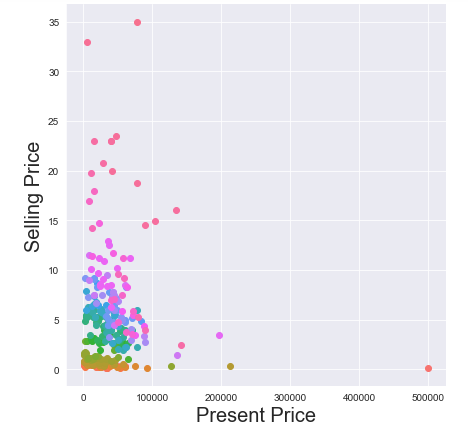


Figure 8

lesser the car would be driven higher will be the cost as we see the graph at max distance i e:- 500000 kilometres the car’s cost is near to Zero or we can say nobody is willing to pay any amount to those cars.

**plotting pair plot**

We cannot visualize multi dimensional scatter plot hence by using pair plot we can visualize each and every dimension of (Dimension with numerical variable )multidimensional data precisely.

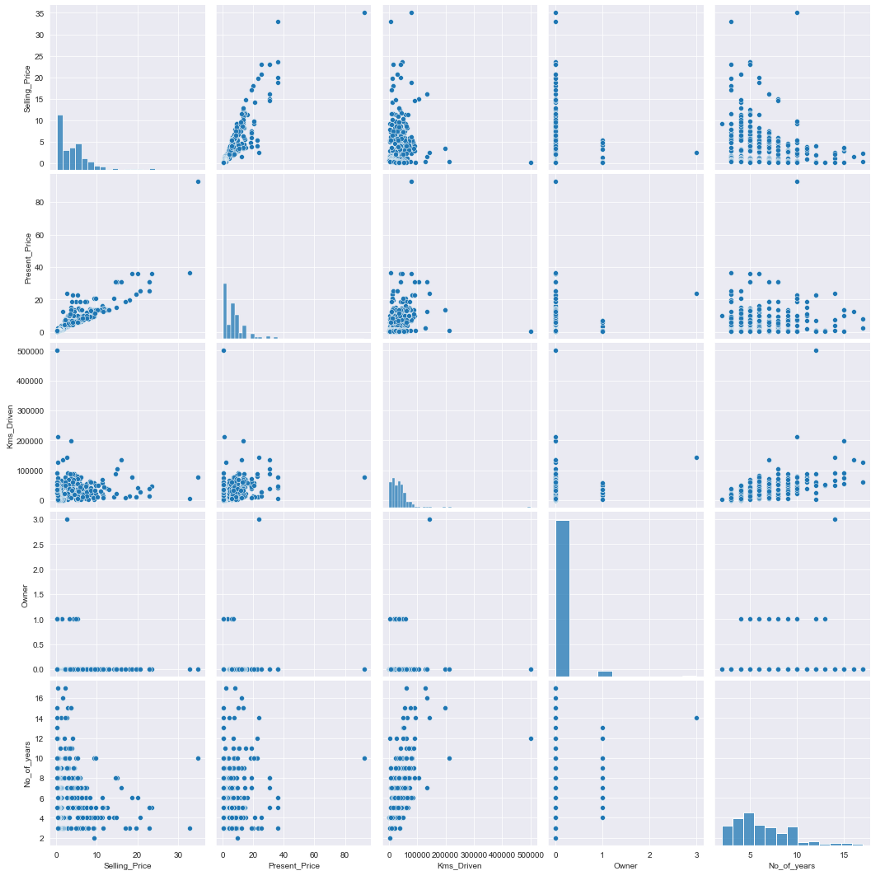


Figure 9

As we see there are very less overlapping in dataset is seen so we cannot use knn ,linear regression,svm and because of the dynamic nature of dataset we even cannot use decision tree so we will go with random forest and xgboost.

Uni variate analysis :- when analysis involve single variable most predominantly it is used to find the pattern in dataset.

sns.set\_style("darkgrid")  
sns.FacetGrid(df,height=6).map(sns.histplot,"Selling\_Price")  
plt.xlabel()  
plt.show()



Figure 10

Most number of car has been sold within a price range of 1–10 lakh and for a price range of 25 -35 lakh there are negligible amount of customer

sns.set\_style("darkgrid")  
sns.FacetGrid(df,height=6).map(sns.histplot,"Kms\_Driven")  
plt.xlabel("Distance Travel",Fontsize=20)  
plt.ylabel("Demand",Fontsize=20)  
plt.show()

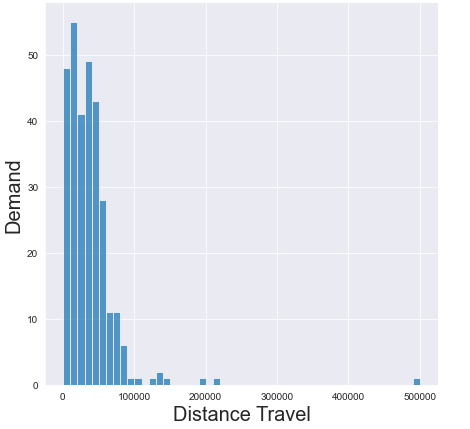


Figure 11

Demand for those car that has been traveled less distance are in more demand especially if car has traveled distance within a range of 0–5000 kilometre people are more attracted towards them.

**C.D.F Plot**

It defines how many percentage of variable has value less than and equal to corresponding x-axis. lets’ take above example how many percentage of vehicle has selling price less than 15 lakh then we can find this kind of answer by using C.D.F.

df\_Selling\_Price=df.loc[:,"Selling\_Price"]  
count,bin\_edges=np.histogram(df\_Selling\_Price,bins=10,density=True) #density=True gives normalized form od bin\_edges and count  
print(count)  
print(bin\_edges)  
PDF=count/sum(count)  
CDF=np.cumsum(PDF) #cdf is sum of all pdf values  
plt.figure(figsize=(8,6))  
plt.plot(bin\_edges[1:],PDF,label="PDF")  
plt.plot(bin\_edges[1:],CDF,label="CDF")  
plt.yticks(np.linspace(0,1,20))  
plt.legend(loc="lower left")  
plt.show()

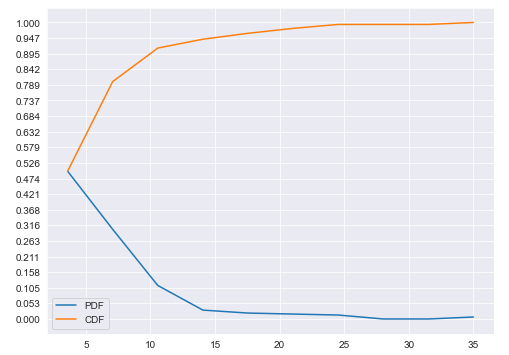


Figure 12

As we can see **94.7%**of cars that are on cardekho has price ≤15 lakh . So one thing is clear that if we want to purchase used car with a price range of 20–25 we won’t prefer to go to cardekho.com because there we won’t get so many options.

multivariate Analysis :- When we are analyzing two and more variables.

sns.set\_style("darkgrid")  
sns.jointplot("Present\_Price","Selling\_Price",data=df,kind="kde")  
plt.show()

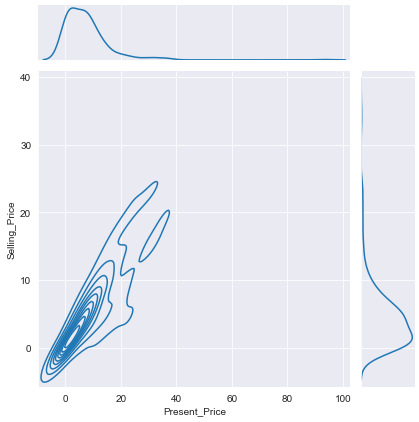


Figure 13

If we see above graph we can understand that for those vehicle whose original price lie within range of 0–20 lakh they are getting approximately 50% of their money when they sell their car after using certain period of time.