CAR RESALE VALUE PREDICTION

NALAIYA THIRAN PROJECT BASED LEARNING

On

PROFESSIONAL READINESS FOR INNOVATION,EMPLOYABILITY AND ENTREPRENEURSHIP

Submitted By

RANJITHKUMAR M	**710719104078**
RANJITHKUMAR S	**710719104079**
SRIMUTHUKRISHNA I	**710719104093**
SUDHARSAN G	**710719104099**

in partial fulfillment for the award of the degree of

Bachelor of ENGINEERING

IN

COMPUTER SCIENCE AND ENGINEERING

DR.NGP INSTITUTE OF TECHNOLOGY

(An Autonomous Institution)

COIMBATORE - 641048

ABSTRACT

The Car Resale value prediction which implements ,that the price of a new car in the industry is 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 prices of new cars and the financial incapability of thecustomers to buy them, Used Car sales are on a global increase. Therefore, there is an urgent need for a Used Car Price Prediction system which effectively determines the worthiness of the car using a variety of features.

Existing System includes a process where a seller decides a price randomly and buyer has no idea about the car and it's value in the present day scenario. In fact, selleralso has no idea about the car's existing value or the price he should be selling the car at.

To overcome this problem we have developed a model which will be highly effective. Regression Algorithms are used because they provide us with continuous value as an output and not a categorized value. Because of which it will be possible to predict the actual price a car rather than the price range of a car. User Interface has alsobeen developed which acquires input from any user and displays the Price of a car according to user's inputs.

TABLEOFCONTENTS

CHAPTER1

1. INTRODUCTION:

Almost everyone wants their own car these days, but because of factorslike affordability or economic conditions, many prefer to opt for pre-owned cars.Accurately predictingusedcarpricesrequiresexpertknowledgeduetothenature of their dependence on a variety of factors and features. Used car pricesare not constant in the market, both buyers and sellers need an intelligent systemthat will allow them to predict the correct price efficiently. In this intelligentsystem, the most difficult problem is the collection of the dataset which containsall important elements like the manufacturing year of the car, its gas type, itscondition, miles driven, horsepower, doors, number of times a car has beenpainted, customer reviews, the weight of the car, etc. It is necessary to pre-process and transform collected data in the proper format prior to feeding itdirectly to the data mining model. As a first step, the dataset was statisticallyanalysed and plotted. Missing, duplicated, and null values were identified anddealt with. Features were chosen and extracted using correlation matrices. Tobuild an efficient model, the most correlated features were retained, and otherswerediscarded.Thispredictionproblemcanbeconsideredaregressionproblemsin ceitbelongstothesupervisedlearning domain.

1.1 PROBLEMSTATEMENT:

Itiseasyforanycompanytopricetheirnewcarsbasedonthemanufacturing and marketing cost it involves. But when it comes to a used car itis quite difficult to define a price because it involves it is influenced by variousparameters like car brand, manufactured year and etc. The goal of our project isto predict the best price for a pre-owned car in the Indian market based on thepreviousdatarelatedtosoldcarsusingmachinelearning.

1.2 PROJECTGOAL:

Cars are more than just a utility for many. We all have different tasteswhen it comesto owning a car or at least when thinking of owning one. Somefit in our budget and some luxury brands are heavy on our pockets. But thatshould not stop us from owning it, at least used ones. The goal of this project topredict the costs of used cars to enable the buyers to make informed purchaseusing the data collected from various sources and distributed across variouslocationsinIndia.

1.3 MACHINELEARNING:

The goal of machine learning (ML) is to help a computer learn without beingexplicitlyinstructedtodosobymeansofmathematicalmodelsofdata.Artificial intelligence (AI) is a subset of machine learning. Data is analysedusing algorithms to identify patterns, which are then used to create predictivemodels. Like humans, machine learning becomes more accurate with more dataandexperience.Withmachinelearning,youcanadapttosituationswheredatais constantly changing, the nature of the request or task is shifting, or coding asolutionisn'tfeasible.

1.4 GRADIENTBOOSTINGREGRESSOR:

Gradient boosting is one of the most popular machine learning algorithms fortabulardatasets.Itispowerfulenoughtofindanynonlinearrelationshipbetweenyo ur model target and features and has great usability that can deal with missingvalues, outliers, and high cardinality categorical values on your features withoutanyspecialtreatment.

CHAPTER2

2. LITERATURESURVEY:

2.1 CARRESALEPREDICTIONSYSTEM

Author:DhwaniNimbark,AkshatPatel,SejalThakkar-2021

Used car resale market in India was marked at 24.2 billion US dollars in 2019.Due to the huge requirement of used cars and lack of experts who can determinethe correct valuation, there is an utmostneedof bridging this gap betweensellers and buyers. This project focuses on building a system that can accuratelypredict a resale value of the car based on minimal features like kms driven, yearof purchase etc. without manual or human interference and hence it remainsunbiased.

2.2 VEHICLE RESALE PRICE PREDICTION USING MACHINELEARNING

Author:B.Lavanya,Sk.Reshma,N.Nikitha,M.Namitha

The production of vehicles has been consistently expanding in the previousdecade, with more than 70 million traveler's vehicles being delivered in the year2016. This has brought about the trade-in vehicle market, which all alone hasbecome a roaring industry. The new approach of online gateways has workedwith the requirement for both the client and the merchant to be better educatedabout the patterns and examples that decide the worth of a pre-owned vehicle onthe lookout. Utilizing Machine Learning Algorithms like Linear Regression,MultipleRegression.wewillattempttofosterafactualmodelwhichwillac tually want to anticipate the cost of a pre-owned vehicle, in light of pastshopper information and a given arrangement of highlights. We will likewise becontrastingtheforecastprecisionofthesemodelstodecidethe idealone.

CHAPTER3

3. MODULEDESCRIPTION:

3.1 COLLECTDATASET:

MachineLearninghasbecomeatoolusedinalmosteverytaskthatrequires estimation. So we need to build a model to estimate the price of usedcars. The model should take car-related parameters and output a selling price.On sprint-1 the selling price of a used car depends on certain features datasetsare collectedfrom different open sources like kaggle.com, data.gov, UCImachinelearning repository, the dataset which contains a set of featuresthrough which the resale price of the car can be identified is to be collectedas

- price
- vehicleType
- yearOf Registration
- gearbox
- model
- kilometer
- monthOfRegistration
- fuelType
- brand
- notRepairedDamage

ML is a data hunger technology, it depends heavily on data, without data,it is impossible. It is the most crucial aspect that makes algorithm trainingpossible.Collects Data, Import necessary packages, Pre-process images, andpasses on to Network Model and Saves Model Weights. The libraries can beimported,

Pre-ProcessTheData:

Pre-processingthedatasetthatincludes:

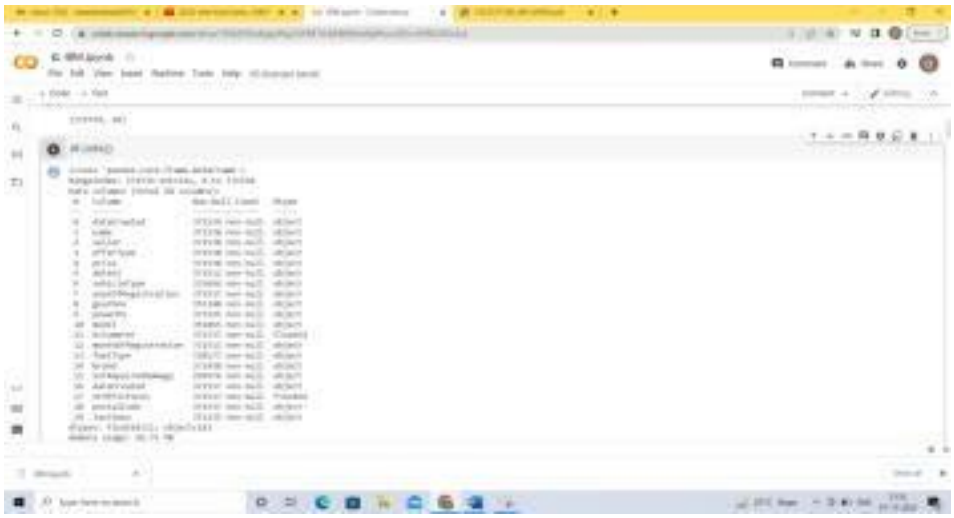
- Handlingthenullvalues.
- Handlingthecategoricalvaluesifany.

- Normalizethedataifrequired.
- Identifythedependentandindependentvariables.

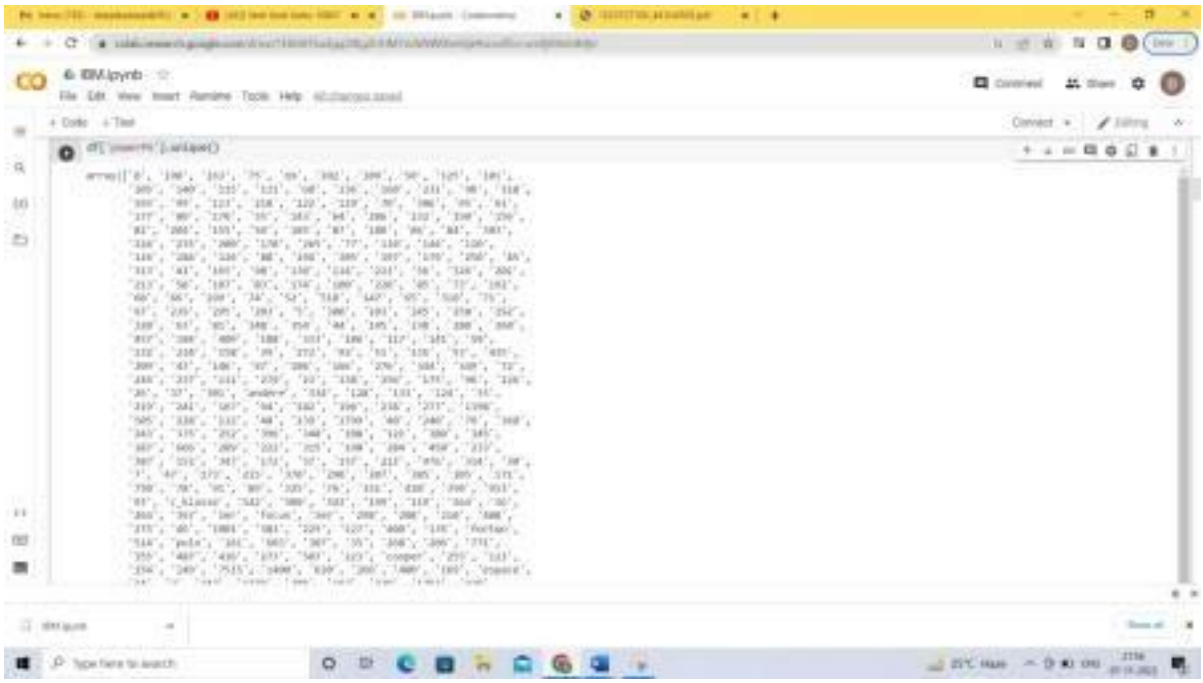
Datacleaningandwranglingmethodsareappliedonthe*usedcars*datafile.Before making data cleaning, some explorations and data visualizationswere applied on data set. This gave some idea and guide about how to dealwith missing values and extreme values. After data cleaning, data explorationwasappliedagaininordertounderstandcleanedversionofthedata.

```
df =
pd.read_csv("/content/drive/MyDrive/ColabNoteb
ooks/autos.csv")df.head()
```

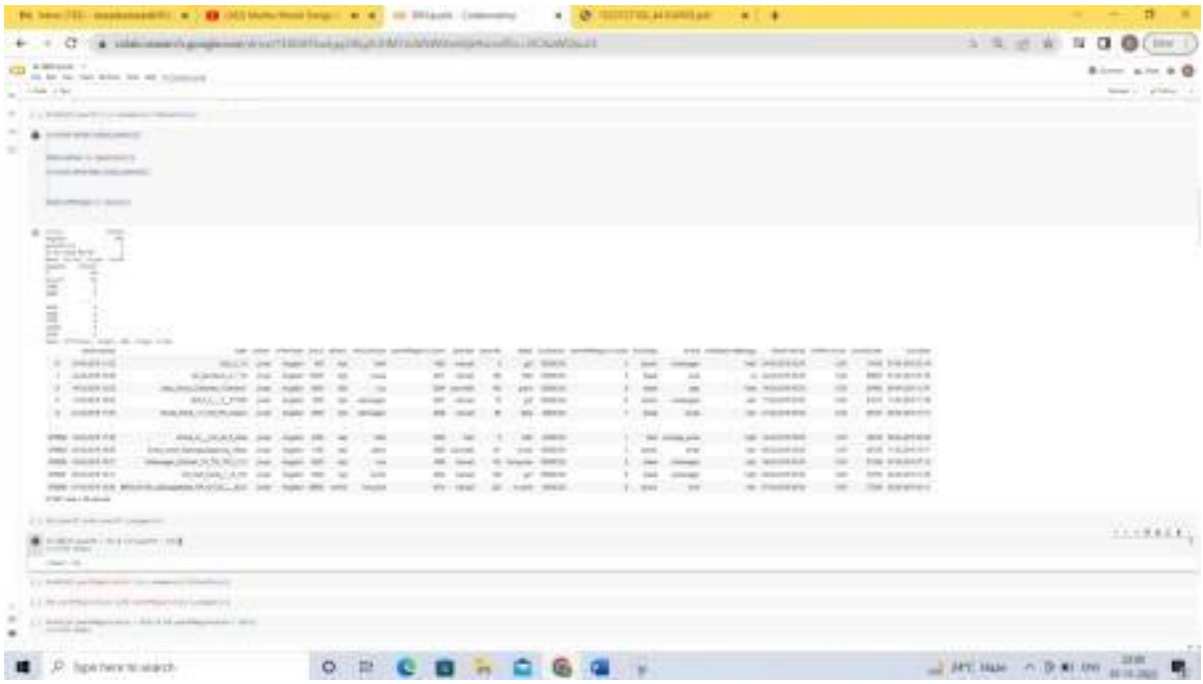
```
print(df.shape)
 (371539, 20)

df.info()
```



```
df['powerPS'].unique()
```



```
df=df[df['powerPS'].str.isnumeric().fillna(False)]print(df.seller.value_c
ounts())
df[df.seller !=
'gewerblich']print(df.offerType.value_counts()
)df[df.offerType !='Gesuch']
```



```
df['powerPS']=df['powerPS'].astype(int)df=df[
(df.powerPS >50) & (df.powerPS
<900)]print(df.shape)
df=df[df['yearOfRegistration'].str.isnumeric().fillna(False)]df
['yearOfRegistration']=df['yearOfRegistration'].astype(int)df=d
f[(df.yearOfRegistration >1950) & (df.yearOfRegistration
<2017)]print(df.shape)
```

```
df.drop(['name', 'abtest', 'dateCrawled', 'nrOfPictures',
'lastSeen','postalCode','dateCreated'],axis='columns',inplace=True)df
.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 308923 entries, 1 to 371538
Data columns (total 13 columns):
 #   Column               Non-Null Count   Dtype
---  ------               --------------   -----
 0   seller               308923 non-null  object
 1   offerType            308923 non-null  object
 2   price                308923 non-null  object
 3   vehicleType          297510 non-null  object
 4   yearOfRegistration   308923 non-null  int64
 5   gearbox              303629 non-null  object
 6   powerPS              308923 non-null  int64
 7   model                297134 non-null  object
 8   kilometer            308923 non-null  float64
 9   monthOfRegistration  308923 non-null  object
 10  fuelType             293046 non-null  object
 11  brand                308923 non-null  object
 12  notRepairedDamage    265507 non-null  object
dtypes: float64(1), int64(2), object(10)
memory usage: 33.0+ MB
```

```
new_df=df.copy()
new_df = new_df.drop_duplicates(['price',
'vehicleType','yearOfRegistration',
'gearbox', 'powerPS', 'model', 'kilometer',
'monthOfRegistration','fuelType',
'notRepairedDamage'])
new_df.gearbox.replace(('manuell', 'automatik'),
('manual','automatic'), inplace=True)
new_df.fuelType.replace(('benzin', 'andere', 'elektro'),
('petrol','others','electric'),inplace=True)new_df.notRepairedDamag
e.replace(('ja', 'nein'),('Yes', 'No'),inplace=True)
        new_df.vehicleType.replace(('kleinwagen', 'cabrio',
    'kombi','andere'), ('small car','convertible', 'combination',
                        'others'),inplace=True)
new_df['price'].unique()
```

```
new_df['price'].unique()
```

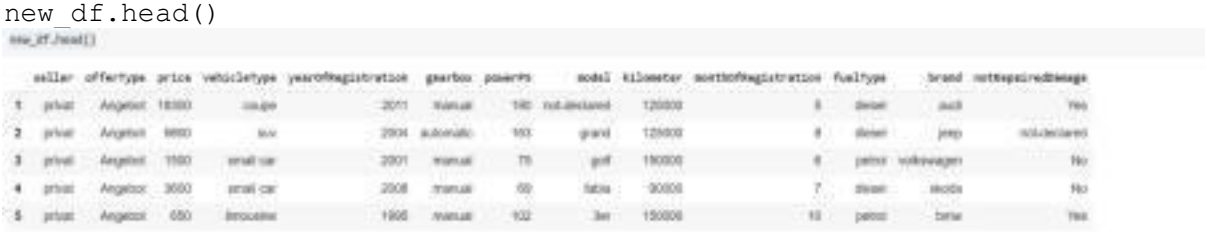
```
array(['18300', '9800', '1500', ..., '18429', '24895', '10985'],
      dtype=object)
```

```
new_df['price']=new_df['price'].astype(int)
```

```
new_df = new_df[(new_df.price >= 100) & (new_df.price
<=150000)] new_df['fuelType'].fillna (value='not-
declared',inplace=True) new_df['gearbox'].fillna (value='not-
declared',inplace=True)

new_df['notRepairedDamage'].fillna (value='not-
declared',inplace=True)

new_df[ 'vehicleType'].fillna (value='not-
declared',inplace=True) new_df['model'].fillna (value='not-
declared',inplace=True)new_df['kilometer']=new_df['kilomete
r'].astype(int)new_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 278363 entries, 1 to 371538
Data columns (total 13 columns):
 #   Column                Non-Null Count   Dtype
---  ------                --------------   -----
 0   seller                278363 non-null  object
 1   offerType             278363 non-null  object
 2   price                 278363 non-null  int64
 3   vehicleType           278363 non-null  object
 4   yearOfRegistration    278363 non-null  int64
 5   gearbox               278363 non-null  object
 6   powerPS               278363 non-null  int64
 7   model                 278363 non-null  object
 8   kilometer             278363 non-null  int64
 9   monthOfRegistration   278363 non-null  object
 10  fuelType              278363 non-null  object
 11  brand                 278363 non-null  object
 12  notRepairedDamage     278363 non-null  object
dtypes: int64(4), object(9)
memory usage: 29.7+ MB
```

```
new_df.head()
```

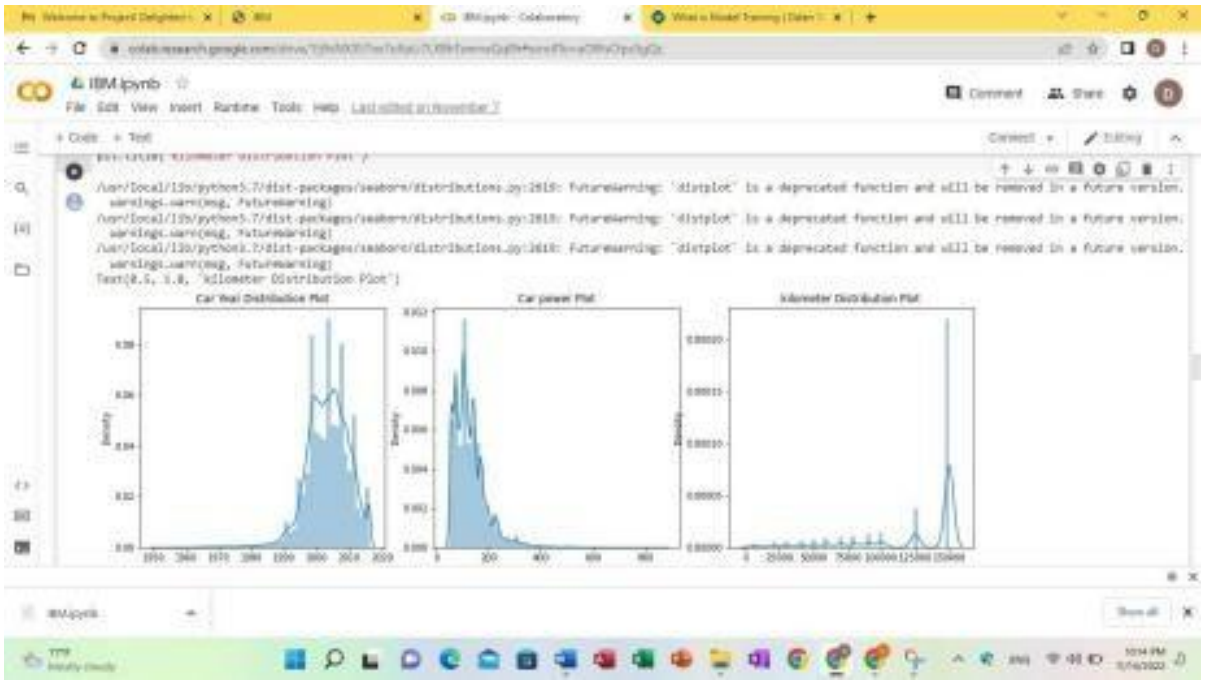


3.2 TRAININGANDTESTINGPHASE:

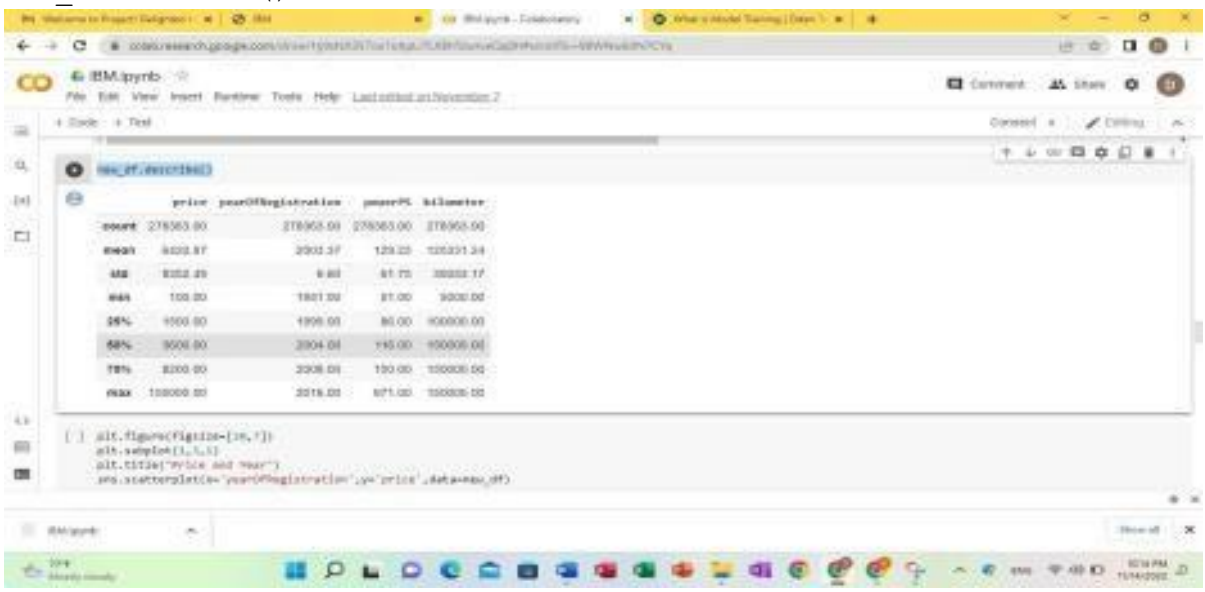
A training model is a dataset that is used to train an algorithm. It consists of thesample output data and the corresponding sets of input data that have aninfluence on the output. The training model is used to run the input data throughthe algorithm to correlate the processed output against the sample output. Theresult from this correlation is used to modify the model. This iterative process iscalled “model fitting”. The accuracy of the training dataset or the validationdataset is critical for the precision of the model. Model training is the process offeeding an algorithm with data to help identify and learn good values for allattributesinvolved.

```
import seaborn as
snsfrom matplotlib
import *importsys
from
pylabimport*plt.figure(figsize=[
11,5])sns.distplot(new_df['price
```

```
plt.figure(figsize=[17,5])plt.su
bplot(1,3,1)
sns.distplot(new_df['yearOfRegistration'])plt.
title('CarYearDistributionPlot')

plt.subplot(1,3,2)sns.distplot(n
ew_df['powerPS'])plt.title('Carp
owerPlot')
```




```
new_df.describe()
```

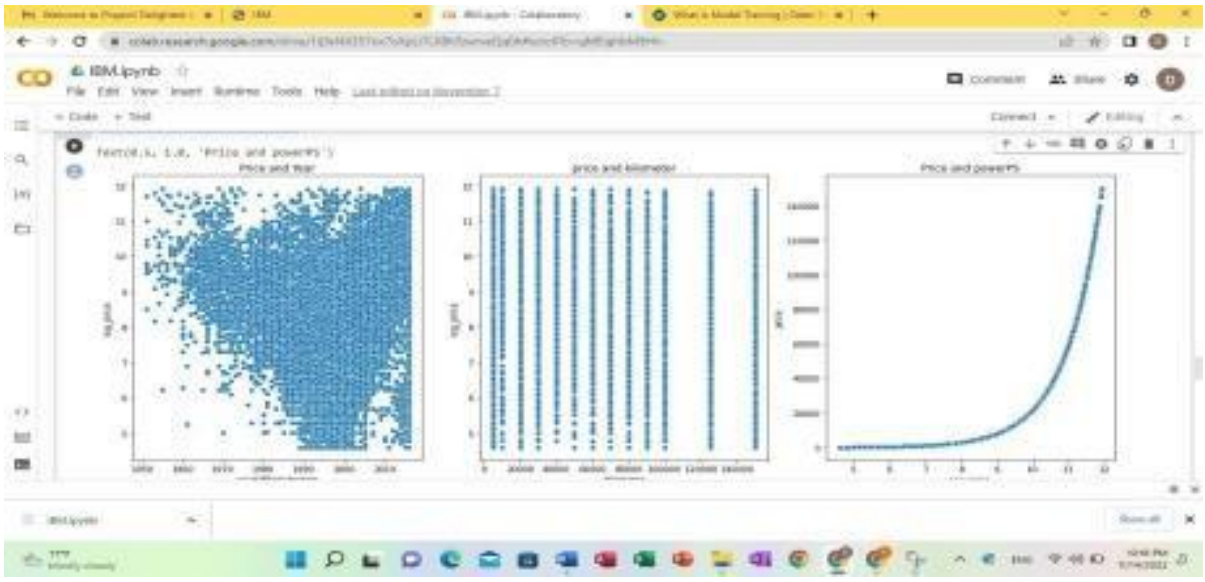



```
plt.figure(figsize=[20,7])p
lt.subplot(1,3,1)plt.title(
"PriceandYear")
sns.scatterplot(x='yearOfRegistration',y='price',data=new_df)

plt.subplot(1,3,2)plt.title("pri
ceandkilometer")
sns.scatterplot(x='price',y='kilometer',data=new_df)

plt.subplot(1,3,3)sns.scatterplot(y='price',x='powerPS',data=ne
w_df)plt.title("PriceandpowerPS")
```


```
log_price =
np.log(new_df['price'])new_df['log_
price'] = log_pricenew_df.head()
```

	seller	offerType	price	vehicleType	yearOfRegistration	gearbox	powerPS	model	kilometer	monthOfRegistration	fuelType	brand	notRepairedDamage
1	privat	Angebot	18300	coupe	2011	manual	190	not-declared	125000	5	diesel	audi	Yes
2	privat	Angebot	9800	suv	2004	automatic	163	grand	125000	8	diesel	jeep	not-declared
3	privat	Angebot	1500	small car	2001	manual	75	golf	150000	6	petrol	volkswagen	No
4	privat	Angebot	3600	small car	2008	manual	69	fabia	90000	7	diesel	skoda	No
5	privat	Angebot	650	limousine	1995	manual	102	3er	150000	10	petrol	bmw	Yes

```
plt.figure(figsize=[20,7])p
lt.subplot(1,3,1)plt.title(
"PriceandYear")
sns.scatterplot(x='yearOfRegistration',y='log_price',data=new_df)

plt.subplot(1,3,2)plt.title("pri
ceandkilometer")
sns.scatterplot(x='kilometer',y='log_price',data=new_df)

plt.subplot(1,3,3)sns.scatterplot(y='price',x='log_price',data=
new_df)plt.title("PriceandpowerPS")
```



```
new_df=new_df.drop(['price'],axis=1)new_df['monthOfRegistration']=new_df['m
onthOfRegistration'].astype(int)labels=['gearbox','notRepairedDamage','mod
el',
'brand','fuelType','vehicleType']
mapper={}
foriinlabels:
  mapper[i]
  =LabelEncoder()mapper[i].fit(new_
  df[i])tr=mapper[i].transform(new_
  df[i])
  np.save(str('classes'+i+'.npy'),
  mapper[i].classes_)print(i,":",mapper[i])
    new_df.loc[:,i+'_labels']=pd.Series(tr,index=new_df.index)


labeled
=new_df[
['log_price','yearOfRegistration','powerPS','kilometer','mo
nthOfRegistration']
+ [x+"_labels" for x in
labels]]print(labeled.columns)
```

```
gearbox : LabelEncoder()
notRepairedDamage : LabelEncoder()
model : LabelEncoder()
brand : LabelEncoder()
fuelType : LabelEncoder()
vehicleType : LabelEncoder()
Index(['log_price', 'yearOfRegistration', 'powerPS', 'kilometer',
       'monthOfRegistration', 'gearbox_labels', 'notRepairedDamage_labels',
       'model_labels', 'brand_labels', 'fuelType_labels',
       'vehicleType_labels'],
      dtype='object')
```

```
plt.figure(figsize=[15,7])sns.heatmap(new_df.c
orr(),annot=True)
```

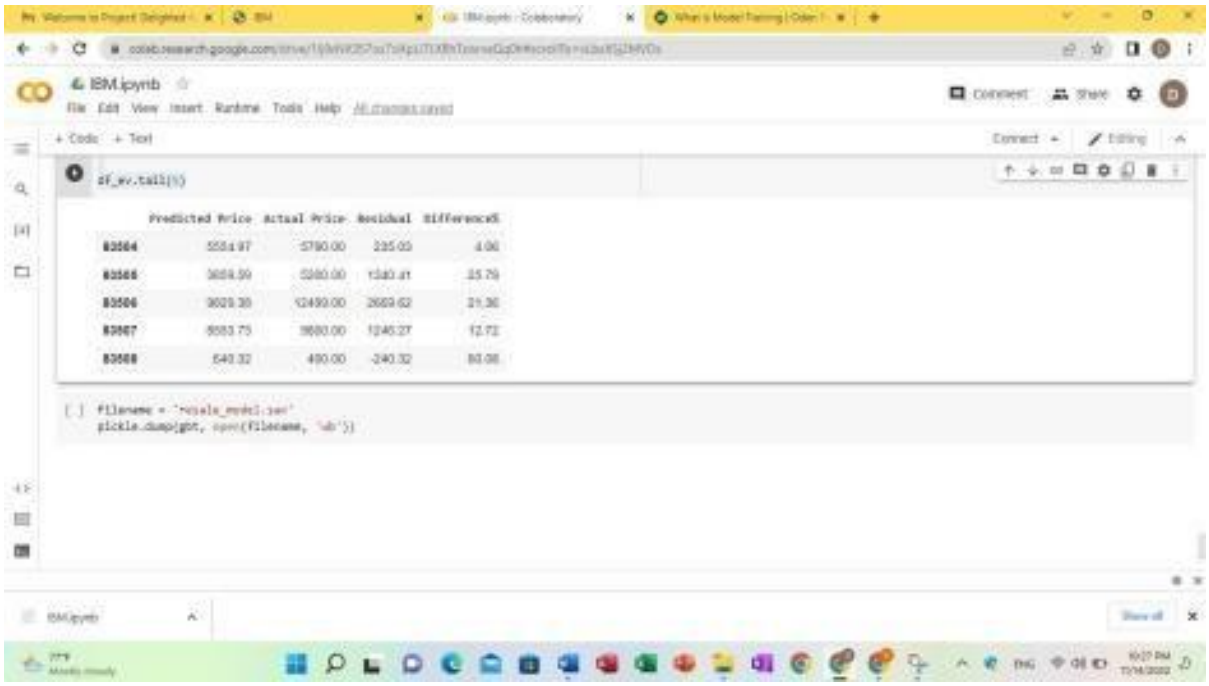

```
Y=labeled.iloc[:,0].values
X =
labeled.iloc[:,1:].valuesY=Y.
reshape(-1,1)
fromsklearn.model_selectionimporttrain_test_split,cross_val_s
coreX_train, X_test,
Y_train,Y_test=train_test_split(X,Y,test_size=0.3,random_stat
e=3)

from sklearn.ensemble import

RandomForestRegressorfrom

sklearn.metricsimportr2_score
regressor= RandomForestRegressor (n_estimators=1000,
max_depth=10,random_state=34)

regressor.fit(X_train, np.ravel (Y_train,
order='C'))y_pred=regressor.predict(X_test)
print(r2_score
(Y_test,y_pred))y_pred=regressor
.predict(X_test)print(r2_score(Y
_test,y_pred))
df_ev=pd.DataFrame(np.exp(y_pred),columns=['PredictedPrice'])

# We can also include the Actual price column in that data
frame(sowecanmanuallycomparethem)#Y_test=Y_test.reset_index(drop
=True)
df_ev['ActualPrice']=np.exp(Y_test)

# we can calculate the difference between the targets and
thepredictions
df_ev['Residual'] = df_ev['Actual Price'] -
df_ev['PredictedPrice']df_ev['Difference%']
```

```
=np.absolute(df_ev['Residual']/df_ev['ActualPrice']*100)

pd.set_option('display.float_format', lambda x: '%.2f' %
x)df_ev.sort_values(by=['Difference%'])
```

```
df_ev.tail(5)
```

	Predicted Price	Actual Price	Residual	Difference%
83504	4946.32	5790.00	843.68	14.57
83505	4177.92	5200.00	1022.08	19.66
83506	11025.04	12499.00	1473.96	11.79
83507	7967.92	9800.00	1832.08	18.69
83508	564.48	400.00	-164.48	41.12

```
from sklearn.linear_model import LinearRegression lr
=LinearRegression()
lr.fit(X_train,Y_train)y_pred_l
r=lr.predict(X_test)
r_squared =
r2_score(Y_test,y_pred_lr)print("R_squ
ared :",r_squared)

fromsklearn.ensembleimportGradientBoostingRegressorgbt
=
GradientBoostingRegressor()gbt.f
it(X_train,Y_train)y_pred_gbt=gb
t.predict(X_test)
r_squared =
r2_score(Y_test,y_pred_gbt)print("R_squ
ared :",r_squared)

df_ev = pd.DataFrame(np.exp(y_pred_gbt),
columns=['PredictedPrice'])df_ev['Actual Price'] =
np.exp(Y_test)df_ev['Residual'] = df_ev['Actual Price'] -
df_ev['PredictedPrice']df_ev['Difference%']
=np.absolute(df_ev['Residual']/df_ev['Actual
Price']*100)pd.set_option('display.float_format', lambda x:
'%.2f' % x)df_ev.sort_values(by=['Difference%'])

df_ev.tail(5)
```




```
filename =
'resale_model.sav'pickle.dump(gbt,open
(filename,'wb'))
```

CHAPTER4

4. SYSTEMSPECIFICATION

4.1 HARDWARESPECIFICATION

- Processors:Intel®Core™i5processor4300Mat2.60GHzor 2.59GHz (1socket,2cores,2threadspercore),8GB ofRAM
- Diskspace:320GB
- Operatingsystems:Windows®10,macOS*,andLinux*

4.2 SOFTWARESPECIFICATION

- Python3.7.4(64-bit)or(32-bit)
- HTML,CSS,javascript
- Flask1.1.1
- JupyterNotebook
- Windows1064–bit

4.3 SOFTWAREDESCRIPTION:

FLASK:

Flask is an open source web framework which offers us with the tools, libraryresources needed to create a web application. Flask is a microweb frameworkwritten in Python. It is classified as a microframework because it does notrequire particular tools or libraries. It has no database abstraction layer, formvalidation,oranyothercomponentswherepre-existingthird-partylibrariesprovide common functions. However, Flask supports extensions that can addapplication features as if they were implemented in Flask itself. Extensions existfor object relational mappers, form validation, upload handling, various openauthenticationtechnologiesandseveralcommon frameworkrelatedtools.

PANDAS:

Data analysis and related manipulation of tabular data in Data frames are themajor uses of Pandas. Data can be imported into Pandas from a variety offiletypes, including Microsoft Excel, JSON, Parquet, SQL database tables, andcomma-separated values. Data wrangling, data cleaning, and other operationslike merging, restructuring, and choosing are all possible with Pandas. Many ofthe R programming language's established functionality for working with dataframes were brought into Python with the introduction of pandas. The NumPylibrary, which is focused on effectively working with arrays rather than thecharacteristics of working with Data frames, is the foundation upon which thePandalibraryisconstructed.

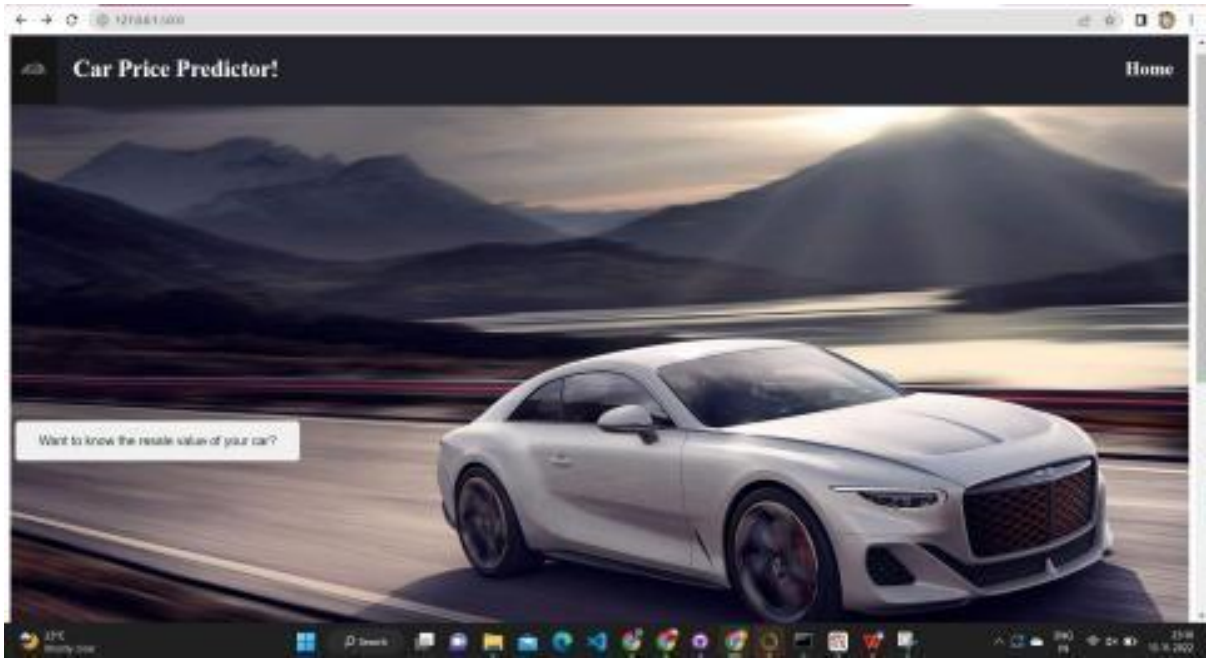
NumPy:

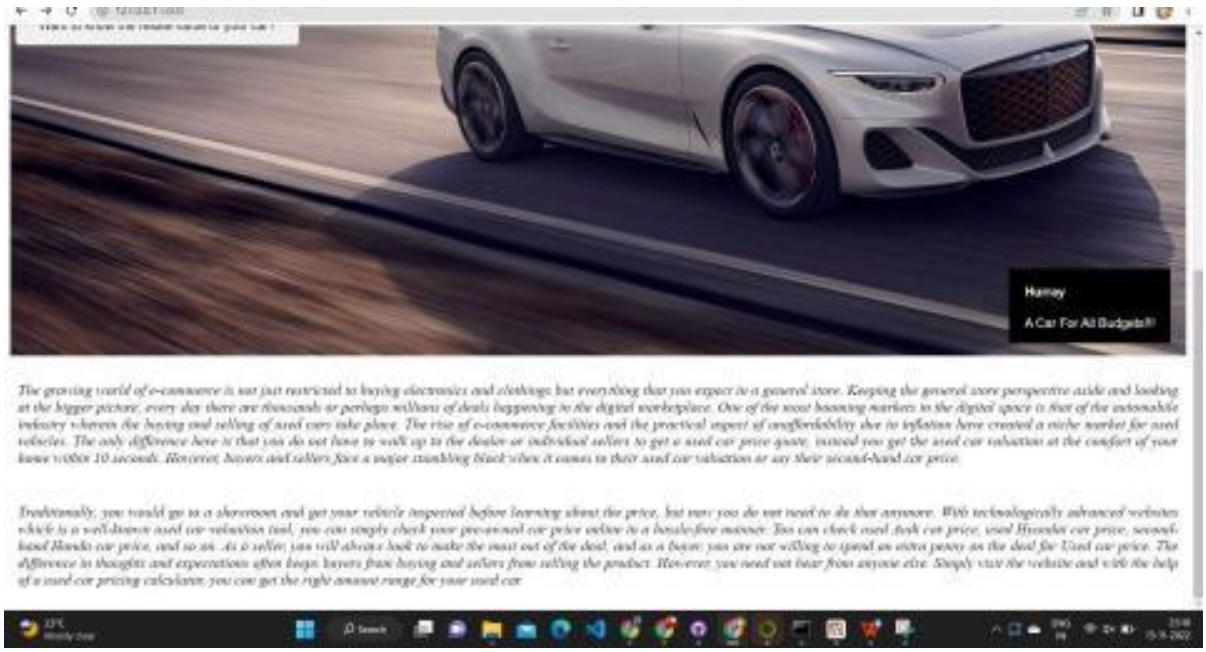
Acollectionofthemultidimensionalmatrixthatfacilitatescomplexmathematical operations. NumPy can be used to execute operations on arraysthat are related to mathematics, such as algebraic, statistical, and trigonometricpatterns. The image is transformed into a matrix. The Convolutional NeuralNetwork is utilized to understand and analyse the image in its matrix form. Theimage's annotationsthen adoptedaNumPyarraystyle.Finally,thedatasetcontains the precise labels for each image. On this, SciPy was also developed. Itprovides more noteworthy execution that utilizes NumPy arrays and is requiredforvariouslogicalandengineeringtasks.

CHAPTER5

5. OUTPUT:

Homepage

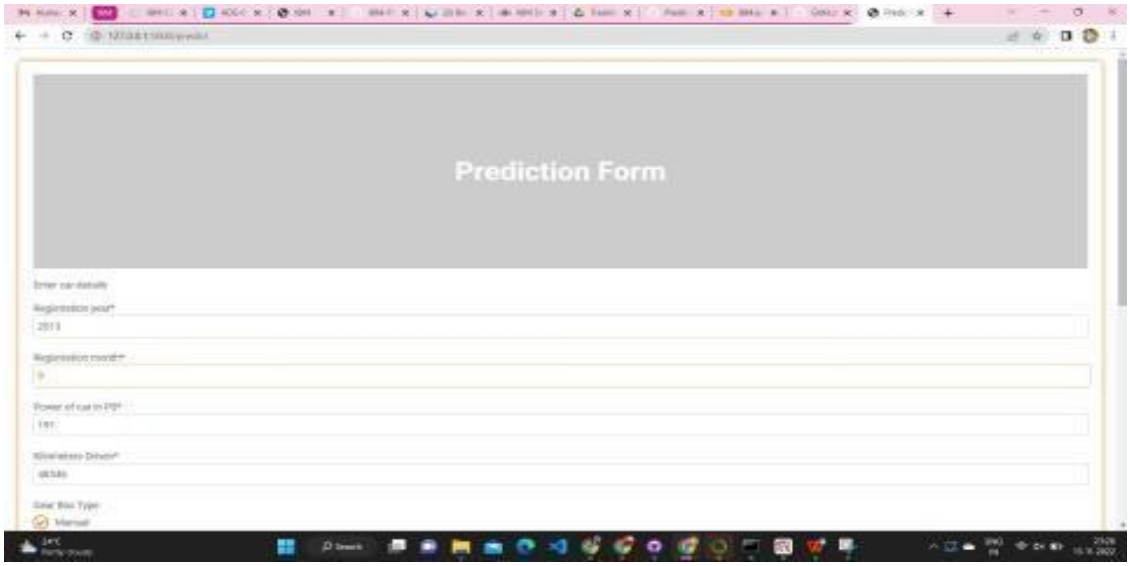


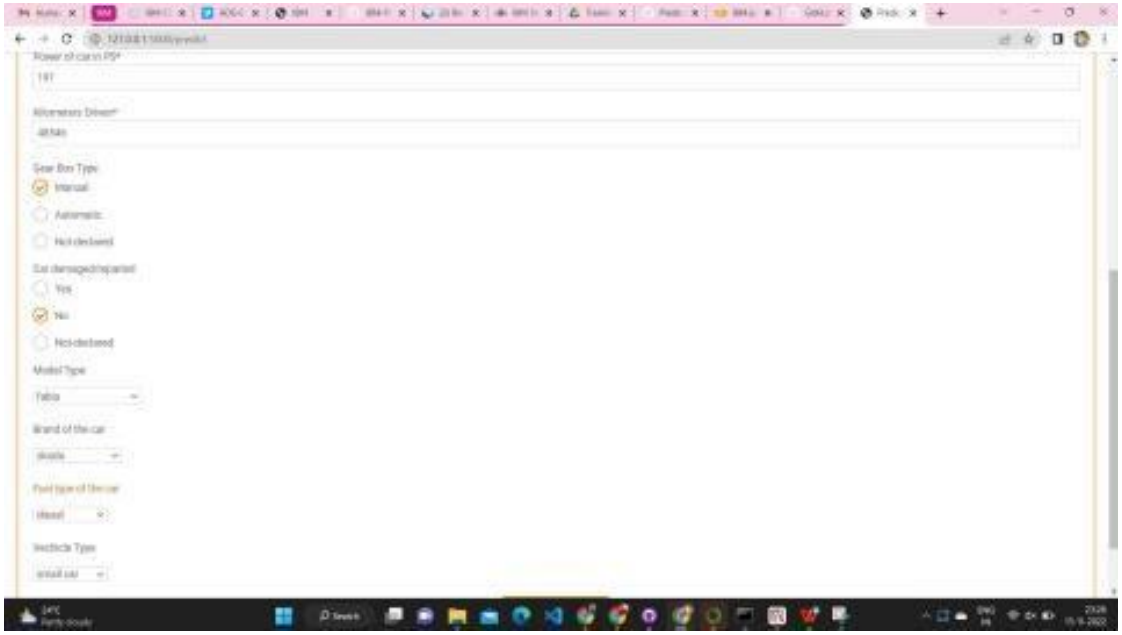


The growing world of e-commerce is not just restricted to buying electronics and clothings but everything that you expect in a general store. Keeping the general store perspective aside and looking at the bigger picture, every day there are thousands or perhaps millions of deals happening in the digital marketplace. One of the most booming markets in the digital space is that of the automobile industry wherein the buying and selling of used cars take place. The rise of e-commerce facilities and the practical aspect of unaffordability due to inflation have created a niche market for used vehicles. The only difference here is that you do not have to walk up to the dealer or individual sellers to get a used car price quote, instead you get the used car valuation at the comfort of your home within 10 seconds. However, buyers and sellers face a major stumbling block when it comes to their used car valuation or say their second-hand car price.

Traditionally, you would go to a showroom and get your vehicle inspected before learning about the price, but now you do not need to do that anymore. With technologically advanced websites which is a web-based used car valuation tool, you can simply check your pre-owned car price online in a hassle-free manner. You can check used Audi car price, used Hyundai car price, second-hand Honda car price, and so on. As a seller, you will always look to make the most out of the deal, and as a buyer, you are not willing to spend an extra penny on the deal for Used car price. The difference in thoughts and expectations often keeps buyers from buying and sellers from selling the product. However, you need not bear from anyone else. Simply visit the website and with the help of a used car pricing calculator, you can get the right amount range for your used car.

Predictionformpage:





CHAPTER6

6. CONCLUSION:

Determining whether the listed price of a used car is a challenging task, due tothe many factors that drive a used vehicle's price on the market. The focus ofthis project is developing machine learning models that can accurately predictthepriceofausedcarbasedonitsfeatures,inordertomakeinformedpurchases.
By performing ML models, we aim to get a better result or less errorwith max accuracy to predict the value of the used car. Initially, data cleaning isperformed to remove the null values and outliers from the dataset then MLmodels are implemented to predict the price of cars. Next, with the help of datavisualization features were explored deeply. The relation between the features isexamined.Fromthereport,itcanbesaidthatgradientregressionregressoristhebestmodelforthepredictionforusedcarprices.