ANNUAL SALES-FORECASTING OF PASSENGER CAR USING RANDOM FOREST MODEL: MACROECONOMICS PERSPECTIVE OF BANGLADESH

AL-AMIN

MANZURUL ALAM

AMIN MAHMUD

JOY DASH

A THESIS SUBMITTED FOR

THE DEGREE OF BACHELOR OF SCIENCE

Dept. Computer Science and Engineering,

University of Information Technology and Sciences.

2021

SUPERVISOR'S APPROVAL

This thesis paper titled **"ANNUAL SALES-FORECASTING OF PASSENGER CAR USING RANDOM FOREST MODEL: MACROECONOMICS PERSPECTIVE OF BANGLADESH"**, submitted by the group as mentioned below has been accepted as satisfactory in partial fulfillment of the requirements for the degree B.Sc. in Computer Science and Engineering in 2021.

RISHAD ISLAM
Assistant Lecturer of CSE Dept, UITS
Military Institute of Science and
Technology

DECLARATION

This is to certify that the work presented in this thesis paper, titled, "ANNUAL SALES-FORECASTING OF PASSENGER CAR USING RANDOM FOREST MODEL: MACROECONOMICS PERSPECTIVE OF BANGLADESH", is the outcome of the investigation and research carried out by the following students under the supervision of RISHAD ISLAM, Assistant Lecturer of CSE Dept, UITS, Military Institute of Science and Technology .

It is also declared that neither this thesis paper nor any part thereof has been submitted anywhere else for the award of any degree, diploma or other qualifications.

AL-AMIN
Roll: 17251003
25 August 2021

MANZURUL ALAM
Roll: 17251004
25 August 2021

AMIN MAHMUD
Roll: 17251005
25 August 2021

JOY DASH
Roll: 17251018
25 August 2021

ABSTRACT

According to World Bank, Bangladesh has made remarkable progress in poverty reduction with sustainable economic growth. It has been one of the fastest growing and now 37th largest economy all over the world. Remarkable economic growth of Bangladesh for last decade has greatly influenced on automotive industry. As a result Bangladesh is now 3rd largest assembler in South Asia. Previous study of automotive industry says that economy has immense impact on automotive business. Evaluating the entire economy of a country is not easy. But Macroeconomics is a term of Economics which can evaluate the entire state of an economy by some indicators. Automotive industry of Bangladesh assembles and imports sort of cars included heavy weight and light weight passenger cars. The proposed model is for light weight passenger cars (contains not more than 9 or 10 passengers) selecting various macroeconomic variables. Model takes input as 'UnemploymentRate', 'CPI', 'PopD', 'PerCapGNI', 'PerCapGDP', 'AnnulGDP', 'AnnualGNI' and outputs number of passenger car will be sold in future. But this model will be also able to estimate heavy weight passenger cars as well as cargo. All data are taken from 1971 to 2019. As a result some macroeconomic indicators and car sales records are missing from the very beginning of Bangladesh. Due to covid-pandemic previous and recent years of data are dropped.

ACKNOWLEDGEMENT

We are thankful to Almighty Allah for his blessings for the successful completion of our thesis. Our heartiest gratitude, profound indebtedness and deep respect go to our supervisor, RISHAD ISLAM, Assistant Lecturer of CSE Dept, UITS, Military Institute of Science and Technology , for his constant supervision, affectionate guidance and great encouragement and motivation. His keen interest on the topic and valuable advice throughout the study was of great help in completing thesis. We are especially grateful to the Department of Computer Science and Engineering (CSE) of University Of Information Technology and Sciences (UITS) for providing their all out support during the thesis work.

Finally, we would like to thank our families and our course mates for their appreciable assistance, patience and suggestions during the course of our thesis.

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

List of Algorithms

CHAPTER 1
INTRODUCTION

Machine learning (ML) is the study of computer algorithms that improves automatically through experience and by the use of data. It is seen as a part of artificial intelligence. Machine learning algorithms build a model based on sample data, known as ”training data”, in order to make predictions or decisions without being explicitly programmed to do so.
Machine learning algorithms are used in a wide variety of applications, such as in medicine, email filtering, speech recognition, and computer vision, where it is difficult or unfeasible to develop conventional algorithms to perform the needed tasks. [1]

1.1 Macroeconomics

Macroeconomics is a part of Economics which discusses about the entire scenario of an economy. Basically, it concentrates on broad issues like growth of production, the number of unemployment people, the inflationary increase in prices, government deficits, and levels of exports and imports. [2]

Economists were always concerned for the topic like unemployment, price, growth and trade from the very beginning of 1940s. So it can say that that the concept of Macroeconomics is not much old. Study of Macroeconomics was specified in 20th and 21th century. Adam Smith and John Stuart Mill brought up the issues that is now considered as the domain of Macroeconomics. However, John Maynard Keynes illustrated the reasons of ”Great Depression” in his book The General Theory of Employment, Interest, and Money in 1936. He depicted why the goods remain unsold too. Since then macroeconomics got modern form. [3]

Macroeconomics has broad area to research. But two fields consider as the most. The first research area discuss about long term economic growth-factor or increases the national income (NI). Another includes the reasons and outcomes of short term fluctuation in national income and employment which is called business cycle.
Economic growth demonstrates the aggregate production increasing and which factors are behind the development, progress and rising of standard living. Growth is generally modelled as mathematical function of physical capital, human capital, labor force and technology.

Business cycle impacts on long term economic growth. Fluctuation of major macroeconomic variable like employment and national output go up or down expansion or recession can be confirm from business cycle. [4]

1.1.1 Significant of Macroeconomics

- As Macroeconomics demonstrates the entire state of an economy, it helps to understand how functions macroeconomic variable and determines the level of national income and employment according to demand and supply.
- Macroeconomics suggests the way to obtain economic goal of a country as well as individuals since it concentrates on economic growth which concludes the large scale of GDP (Gross Domestic Product) growth, inflation rate, unemployment etc.
- Economy is a backbone of a country. So Every country wants sustainable economy to achieve prosperity form all aspects. Macroeconomics helps to understand the behavior of vital economic term in large scale to achieve sustainable economy.
- Macroeconomics brings up consistency of an economy. As it approaches fluctuation of business activities and price level and indicates the way for stability.
- It suggests the policy makers to control inflation and deflation.
- It elucidates payment balance with its factor at the same time finds out the causes of deficit in payment balance and helps to take steps.
- It solves economic difficulties like poverty, unemployment, inflation, deflation etc. in macro level.
- Proper knowledge of functions and behavior in macro scale of an economy helps to formulate economic policies as correct as possible.
- Microeconomics is complex as it contains smaller scale. So it is much complex to apply on the difficulties of whole economy. Ultimately Macroeconomics has given the flexibility to perceive health of an economy. [5]

1.1.2 Macroeconomics in Automotive Industry

- China, USA, Japan and Germany are the top four automotive country in the world. Macroeconomics has great influence on these automotive industry. Real GDP, car production, gasoline price influence positively on car selling whereas per capita GDP , inflation and exchange rate work as opposite. [6]

- Many microeconomic issues such as interest rate, employment, oil prices, cost of living, consumer confidence etc affect indirectly macroeconomic variables. This is comparatively sluggish way macroeconomics impacts on automotive industry. [7]
- Malaysia, Singapore, Indonesia, Philippine are also significantly influenced by macroeconomic indicators. GDP, inflation, unemployment rate and loan rate have long term effect on automobile sales. However, several country has got different dominant factor. [8]

1.2 Objectives

To estimate passenger car selling in every year in Bangladesh automotive industry how many passenger cars should assemble for the country. Our main objectives are given below:

- How many passenger cars will be sold in every year.
- How macroeconomics factors impact on automotive industry and finding which influences the most.

1.3 Motivation

As growing economy of Bangladesh has been influencing life style, people are more concerned about transportation to save and utilize time. Moving one place to another place within short time contributes business or individuals financially. So people are involving more with transport as well as automotive industries.

On accordance with demand of people automotive industries are also assembling to meet up consumers demand. But the fastest growing economy often troubles automotive industries to estimate the number of production of light weight passenger car.

We have taken some vital macroeconomics indicators such as annual GDP, per captita GDP, GDP growth, per capita GDP growth, annual GNI, per capita GNI, per capita GNI growth, inflation, unemployment rate and foreign exchange for our data set. Various methods are applied to explore data but regression method performs the best to predict.

1.4 Our Contribution

We have suggested indigenous automotive industries to use prediction to estimate the production of light weight passenger cars with the fastest changing financial capability of people. To measure the estimate we build a model that uses regression algorithm to give the expected output after analysing data set and gives the 0.87 percent accuracy. Hope this accuracy will help to make proper decision producing the approximate number of passenger cars of the industries at the same time company can reduce the loss.

1.5 Report Layout

In this research paper chapter-2 discusses literature review where all the previous research related to the topic are included.In Chapter-3, it contains research methodology and Chapter-4 discusses all the algorithm we applied and sample of input and output are explained.

CHAPTER 2
LITERATURE REVIEW

Machine Learning (ML) is now the most rapidly growing field. What machine learning actually does it builds a computer model recognizing the pattern underlying on data set. At the same time it improves the model from training experience. Since the computing model is created and learned from the statistical data, machine learning is seen as the intersection of computer science and statistics. ML is now using many aspects of life for decision making such as healthcare, manufacturing, education, financial modeling, policing, marketing and so on. [9]

2.1 Macroeconomics in Predicting Car Sales

With the progress of economy, automobile industry is extending worldwide. Enterprises are also changing their strategy to cope up with largely growing of this sector. Enterprises are using machine learning to know the prediction of car sales according to evaluate the state of economy in every year as well as for long term economical effect. These prediction can be reliable source in every aspects of car business.

Automobile industry is now world's no.1 consumer market with the rapid growing of economy. Increasing the number of car production has an effect on auto finance. The effect is now almost 40% whereas five year back it was 13%. Car sales depend on various factors such as steel car production, rubber tire car production, monetary supply of consumer goods, consumer price index etc. [10]

A research on Indonesian automobile market which has taken five macroeconomic variables such as exchange rate, Gross Domestic Product (GDP), growth of GDP, inflation and interest rate shows that GDP and GDP growth have a significant impact on car and motorcycles sales. 30 years of data is taken and SPSS (Statistical Package for the Social Sciences) model with regression method was used. Empirical result shows that the exchange rate of USD to IDR, inflation, and interest rate do not impact on automobile sales in Indonesia. [11]

Monetary policy appointed by the Brazilian government from 1994 to 2014, represented the performance of automotive industry. The findings emerged from regression and correlation state the relation between monetary policy and automotive sector. Performance of the

automotive sector demonstrated by production and export level is associated to the variables of monetary policy. Reserve requirements and GDP difference impacted production in a balanced way. When these variables are growing, performance of the sector is extended and vice versa. On the contrary, some variables drive the performance in reverse order. This is also confirmed that tight monetary policy decreases and expansionary monetary policy increases the performance of automotive sector. [12]

A study from co-integration and causality test relates inflation and automotive industry of South Africa. Empirical data results that inflation and trending of new vehicle sales are co-integrated as equivalency exists in a long run. Based on Granger-Causility test, inflation has 5% significant level to new vehicles sales with undirectional causal effect. [13]

A research concentrated on five Asian countries: Indonesia, Thailand, Malaysia, Singapore, and Vietnam examines behaviour of macroeconomic variables and auto sales. Statistical data of macroeconomic variables has brought out that inflation, foreign exchange rate and interest rate have significantly impact on car sales of these five Asian countries. Growth of GDP per capita does not influence auto sales. So undoubtedly these macroeconomic variables steer the competitiveness among five Asian countries. [14]

2.2 Details of Macroeconomics variables

As Macroeconomics is a branch of Economics and discusses behavior and state of an overall economy on large scale, it contains some variable to measure. These variables are also said Macroeconomic indicators such as GDP, GNI, inflation, PPP, Unemployment Rate, Price Indices etc. [15]

We have taken some vital indicators to measure the state of Bangladeshi economy. Details of all variables are given below:

Variable	Description	Source
Gross Domestic Product (GDP)	GDP is the ultimate value of goods and services produced within a geographic area for a specific time period. GDP = Real GDP - Taxes + Subsidies. (output method) [16]	macro-trends.net
GDP Per Capita	GDP per capita is a metric that crumbles economic output of per person. [17] GDP per capita = Total GDP / Total Population.	macro-trends.net
GDP Growth Rate	Growth rate measures the change of growth year-over-year which indicates how fast or slow growing the economy. [18] GDP Growth/Economic Growth = (Final - Initial) / Initial.	macro-trends.net
Grows National Income (GNI)	GNI previously known as GNP (Gross National Product), sums up total domestic and foreign output gained by total residents. GNI = GDP + Factor Income of living foreign resident - Domestic Income by non-resident. [19]	macro-trends.net
GNI Per Capita	GNI per capita is final income of each person of country. Per Capita GNI = Total GNI / Total Population. [20]	macro-trends.net
GNI Growth Rate	Likewise GDP growth rate.	macro-trends.net
Purchasing Power Parity (PPP)	PPP is a metric that compares value of currency of different countries through 'Basket of Goods' approach. Exchange Rate (form currency 1 to 2) = P1 / P2. P1 = Cost of good X in currency 1 P2 = Cost of good X in currency 2. [21]	data. world-bank.org
Foreign Exchange Reserve	Foreign Exchange Reserves are assets denominated in a foreign currency that are held by central bank. [22]	data. world-bank.org
Inflation Rate	Inflation is the overall increase price with in fixed economy. Inflation Rate = (Current CPI - Prior CPI) / Prior CPI. [23]	macro-trends.net
Unemployment Rate	Unemployment refers to the share of the labor force that is without work but available for and seeking employment. Ratio is expressed as percentage. U-3 = (Unemployed / Labor force) x 100. [24]	macro-trends.net
Consumer Price Index (CPI)	CPI is a metric that measures average change in prices over time that consumers pay for a basket of goods and services. CPI = Cost of Market Basket in Given Year / Cost of Market Basket in Base Year. [25]	data. world-bank.org

And a Non-macroeconomic variable, Number of Passenger Car (NumCar) has taken from theglobaleconomy.com.

CHAPTER 3
RESEARCH METHODOLOGY

In this chapter a framework has been suggested to predict sales of passenger cars using macroeconomic variables. This proposed framework has been established thorough some procedures:

- Main Goal
- Environment Setup
- Data Analysis
- Statistical Description
- Data Visualisation
- Data Pre-processing
- Feature Selection
- Model Design
- Final Model
- I/O Sample And Result

3.1 Proposed Methodology

There are several steps to go through to arrange methodology. Each step was to bring transparency to contribute as much as avoiding error in the entire research. A part of this methodology is mentioned above as part as framework. The first step remarks aim of the pursuing. After that we have provided environment set up to explore data. Before pre-processing we have gone through data analysis, statistical description and data visualisation respectively. Likewise final model has come up having feature selection, algorithm selection, test-train, parameter tuning before.

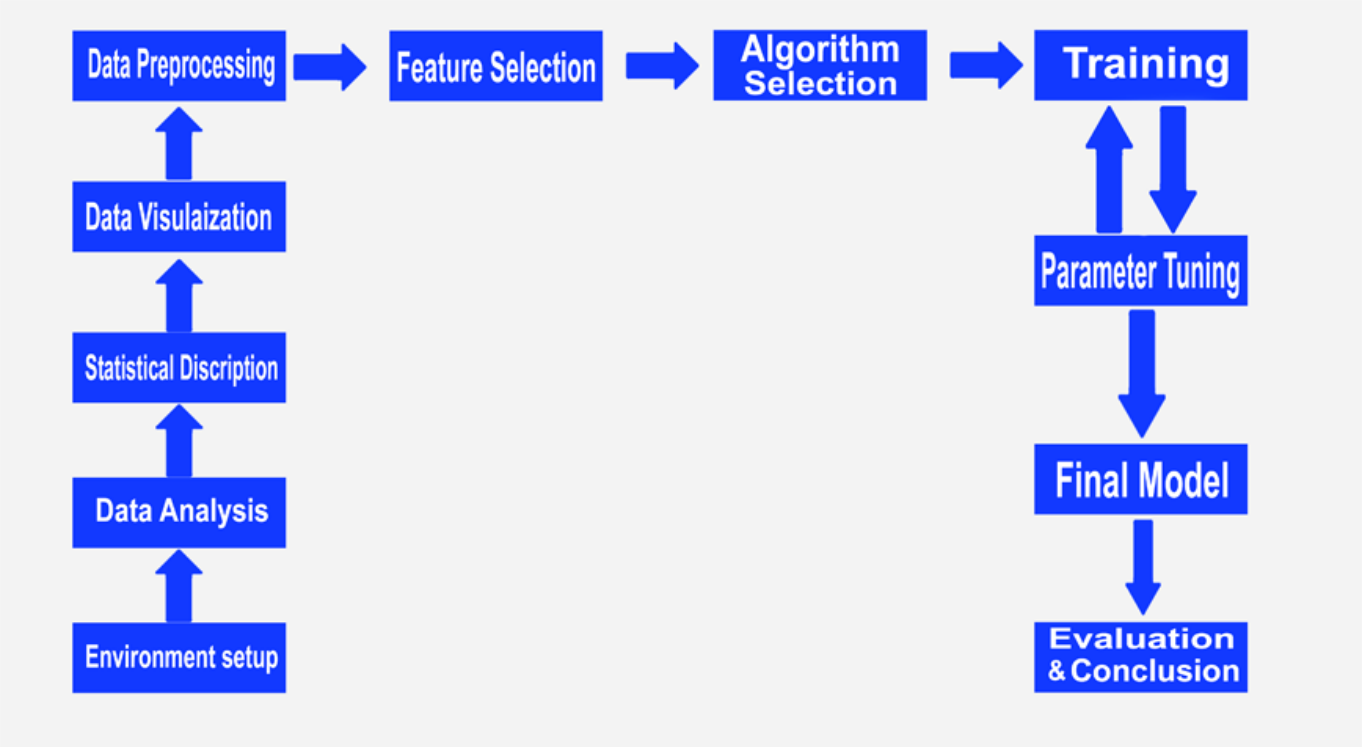

Figure 3.1: The Framework of Research Methodology

3.2 Main Goal

This research intends to estimate of car sales assessing macroeconomic variables based on ML. ML refers to deploy a model examined by many steps before attaining goal. The most relevant technique is to employ in considering better performance of model. Various applied techniques are given below:

- Gradient Boosting Regressor.
- Decision Tree Regressor.
- Random Forest Regressor.
- Ridge Regression.
- Lasso Regression.

Following some procedures the most scoring technique will be picked up to build model.

3.3 Environment Setup

Firstly we have imported libraries like pandas numpy, seaborn, matplotlib and matrix, Then we read the data from a pandas Data frame. The data set contains 49 rows and 16 columns.

	Year	NumCar	AnnulGDP	GDPgrowth	PerCapGDP	PerCapGDPgrowth	AnnualGNI	GNIgrowth	PerCapGNI	PerCapGNIgrowth	InflationRate	ForExcReserve	UnemploymentRate	CPI	ppp	PopD
count	49.00000	49.000000	49.000000	49.000000	49.000000	49.000000	49.000000	49.000000	49.000000	49.000000	49.000000	49.000000	49.000000	49.000000	49.000000	49.000000
mean	1995.00000	9358.163265	68.605714	5.177347	505.387755	6.656939	70.672449	5.381429	522.040816	8.204082	7.450000	19.673224	3.037143	60.552388	1626.846306	774.297959
std	14.28869	6672.107261	73.566442	2.496340	429.124245	15.683964	74.587608	2.491582	433.925779	10.540111	2.518985	19.464000	0.944555	45.235524	1115.203918	211.656195
min	1971.00000	2500.000000	6.290000	-4.090000	94.000000	-49.130000	8.400000	-4.230000	120.000000	-15.000000	2.010000	1.077000	2.200000	24.280000	850.505000	443.890000
25%	1983.00000	6500.000000	19.450000	3.890000	234.000000	0.520000	20.470000	4.610000	230.000000	3.130000	5.670000	2.625000	2.200000	24.280000	850.505000	583.500000
50%	1995.00000	6500.000000	37.940000	5.180000	329.000000	7.150000	39.240000	5.260000	340.000000	7.220000	7.530000	14.319000	2.460000	42.530000	1021.431000	780.130000
75%	2007.00000	6500.000000	79.600000	6.520000	558.000000	11.580000	86.320000	6.650000	610.000000	10.950000	9.870000	32.126000	4.060000	80.550000	2065.155000	966.340000
max	2019.00000	29300.000000	302.570000	13.970000	1856.000000	52.490000	316.240000	9.800000	1940.000000	33.330000	11.400000	87.625000	5.000000	170.160000	4954.761000	1104.420000

Figure 3.2: Datset from Data frame

We start off analyzing data, then select the features to build a machine learning model and predict.

3.4 Data Analysis

In statistics, exploratory data analysis is the process of summarizing the main characteristics of a dataset to understand what the data can tell us beyond the formal modeling or hypothesis testing task. We start off getting an overview of the whole dataset. We want to know how many categorical and numerical variables there are and the proportion of missing data. So we plot a heatmap of the dataframe and visualize columns type and missing data.

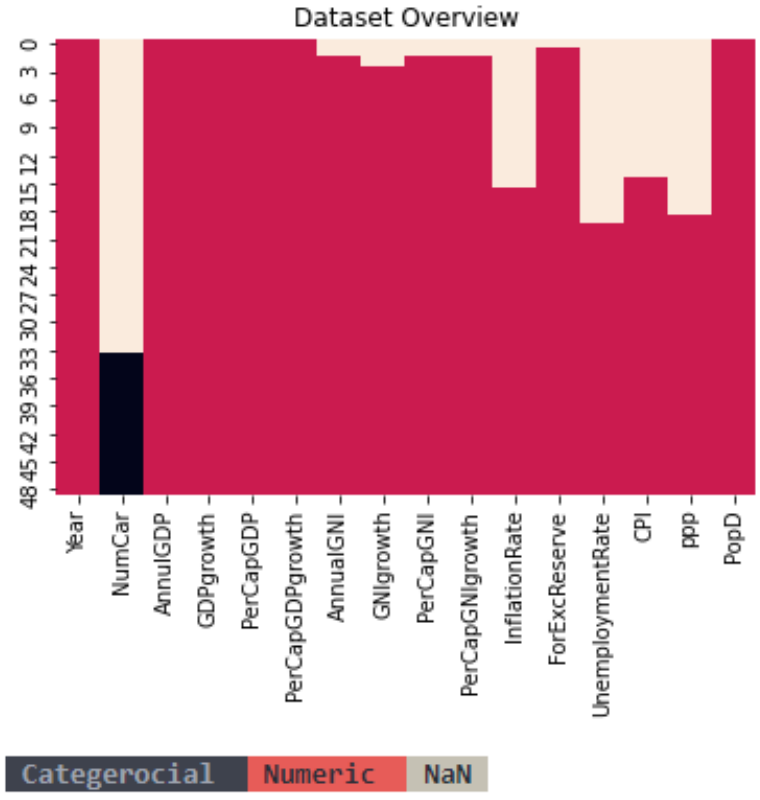

Figure 3.3: Dataset Overview

There are 49 rows and 16 columns and each row of the table represents an observation. We have

found all the data are numerical and some of data have missing value. List of these are given below:

Variable	Missing Value
NumCar	34
AnnualGNI	2
GNIgrowth	3
PerCapGNI	2
PerCapGNIgrowth	2
InflationRate	16
ForExcReserve	1
UnemploymentRat	20
CPI	15
PPP	19

We have used Backward filling method to handle missing values. From Pandas, dataframe.bfill() is to call to fill the missing values in the dataset. It fills with the NaN (Not a Number) values that are present in the pandas dataframe.

	Year	NumCar	AnnulGDP	GDPgrowth	PerCapGDP	PerCapGDPgrowth	AnnualGNI	GNIgrowth	PerCapGNI	PerCapGNIgrowth	InflationRate	ForExcReserve
0	1971	6500.0	8.75	5.48	134	-4.61	8.40	9.32	120.0	33.33	9.87	27.047
1	1972	6500.0	6.29	13.97	94	-29.33	8.40	9.32	120.0	33.33	9.87	27.047
2	1973	6500.0	8.09	3.33	120	26.68	8.40	9.32	120.0	33.33	9.87	14.319
3	1974	6500.0	12.51	9.59	182	52.24	10.95	9.32	160.0	33.33	9.87	13.820
4	1975	6500.0	19.45	-4.09	278	52.49	14.73	-4.23	210.0	31.25	9.87	14.826
5	1976	6500.0	10.12	5.66	141	-49.13	14.61	5.58	200.0	-4.76	9.87	28.891
6	1977	6500.0	9.65	2.67	131	-6.96	12.75	2.32	170.0	-15.00	9.87	24.149
7	1978	6500.0	13.28	7.07	176	34.00	12.46	7.20	170.0	0.00	9.87	32.126
8	1979	6500.0	15.57	4.80	201	14.05	15.00	4.96	190.0	11.76	9.87	41.365
9	1980	6500.0	18.14	0.82	228	13.44	18.48	0.50	230.0	21.05	9.87	33.118
10	1981	6500.0	20.25	7.23	248	8.74	21.50	9.80	260.0	13.04	9.87	15.968

Figure 3.4: After Backward Filling Method

3.5 Statistical Decription

Pandas dataframes additionally have techniques for summarizing the numeric values in the dataframe. For example, utilize the method. To perform summary statistics on all numeric columns in a pandas dataframe, can be used describe(): newdf.describe()
Count, mean, minimum, and maximum values are only a few examples.
The.describe() function returns a beautifully structured dataframe. Because the column called precip is the only column in the example dataset having numeric values, the result of.describe() only includes that column.

	Year	NumCar	AnnulGDP	GDPgrowth	PerCapGDP	PerCapGDPgrowth	AnnualGNI	GNIgrowth	PerCapGNI	PerCapGNIgrowth	InflationRate	ForExcReserve	UnemploymentRate	CPI	ppp	PopD
count	49.00000	49.000000	49.000000	49.000000	49.000000	49.000000	49.000000	49.000000	49.000000	49.000000	49.000000	49.000000	49.000000	49.000000	49.000000	49.000000
mean	1995.00000	9358.163265	68.605714	5.177347	505.387755	6.656939	70.672449	5.381429	522.040816	8.204082	7.450000	19.673224	3.037143	60.552388	1626.846306	774.297959
std	14.28869	6672.107261	73.566442	2.496340	429.124245	15.683964	74.587608	2.491582	433.925779	10.540111	2.518985	19.464000	0.944555	45.235524	1115.203918	211.656195
min	1971.00000	2500.000000	6.290000	-4.090000	94.000000	-49.130000	8.400000	-4.230000	120.000000	-15.000000	2.010000	1.077000	2.200000	24.280000	850.505000	443.890000
25%	1983.00000	6500.000000	19.450000	3.890000	234.000000	0.520000	20.470000	4.610000	230.000000	3.130000	5.670000	2.625000	2.200000	24.280000	850.505000	583.500000
50%	1995.00000	6500.000000	37.940000	5.180000	329.000000	7.150000	39.240000	5.260000	340.000000	7.220000	7.530000	14.319000	2.460000	42.530000	1021.431000	780.130000
75%	2007.00000	6500.000000	79.600000	6.520000	558.000000	11.580000	86.320000	6.650000	610.000000	10.950000	9.870000	32.126000	4.060000	80.550000	2065.155000	966.340000
max	2019.00000	29300.000000	302.570000	13.970000	1856.000000	52.490000	316.240000	9.800000	1940.000000	33.330000	11.400000	87.625000	5.000000	170.160000	4954.761000	1104.420000

Figure 3.5: Statistical View of Dataset

Here found some parameter like count, mean, std, min, 25%,50%, 75%, max Optional percentiles list of numbers The percentiles that should be included in the final product. All of the numbers should be between 0 and 1. [.25,.5,.75] is the default, which returns the 25th, 50th, and 75th percentiles.
newdf.count: Count number of non-NA/null observations.
newdf.max: Maximum of the values in the object.
newdf.min: Minimum of the values in the object.
newdf.mean: Mean of the values.
newdf.std: Standard deviation of the observations.
The result's index will provide count, mean, standard deviation, min, max, as well as lower, 50, and upper percentiles for numeric data. The lower percentile is set to 25 by default, while the upper percentile is set to 75. The median and the 50th percentile are the same.
The result's index will comprise count, unique, top, and freq . The most common value is the top. The freq is the frequency of the most common value. The first and last items are also included in the timestamps.
If more than one object value has the greatest count, the count and top results will be chosen at random from those with the highest count.The default for mixed data types provided via a DataFrame is to return o.

3.6 Data Visualisation

The process of presenting and converting data and information in a visual environment, generally using a graph, chart, bar, or other visual assistance, is known as data visualization. Images are often used in visualization to explain the links between different kinds of data. Information visualization, information graphics, and statistical graphics are all terms used to describe data visualization. It's a phase in the ML(Machine Learning) process that states that once all of the data has been collected, processed, and modeled, the information must be visualized so that users may make conclusions from it.

3.6.1 Why do we need data visualization?

You can have all the greatest, most valuable data analysts and other data researchers in the world, but if the customers and users can't comprehend it, it's meaningless. As a result, the facts must be presented in easy-to-understand ways for the ordinary layperson. That is why data visualization exists. "A picture is worth a thousand words," as the adage goes. Data visualization aids in the creation of that image, which leads to a better comprehension. We're going to talk about data visualization today. We'll look at its description, distinct types, relevance, and how it's used in various businesses and sectors, as well as different data visualization approaches.
Dataset includes 16 rows and 49 columns of data. We have visualized all of rows individually and jointly too. Firstly, we are going to explain individual data visualization.
For single row visualization, seaborn distplot has been taken. We can see (figure 3.4) that the count of different tip value present in the dataset and infer that most of the tips are between 1 and 9. There are also negative values that is shown left side of the figure.

```
In [17]: sns.distplot(newdf['AnnulGDP'])
Out[17]: <matplotlib.axes._subplots.AxesSubplot at 0xad50e8>
```

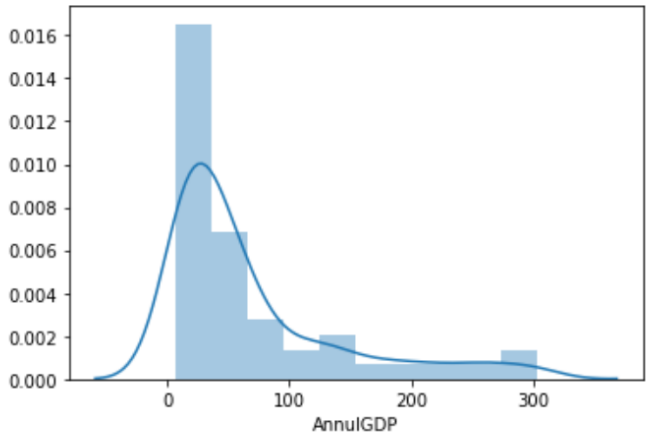

Figure 3.6: Annual GDP

```
sns.distplot(df['GDPgrowth'],kde=True)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x20e9acbe4c8>
```

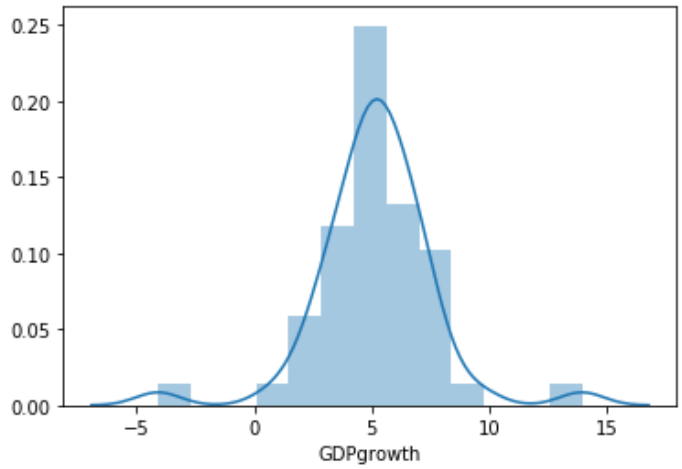

Figure 3.7: GDPgrowth

```
: sns.distplot(df['PerCapGDP'],kde=True)
```

```
: <matplotlib.axes._subplots.AxesSubplot at 0x20e9b2cd588>
```

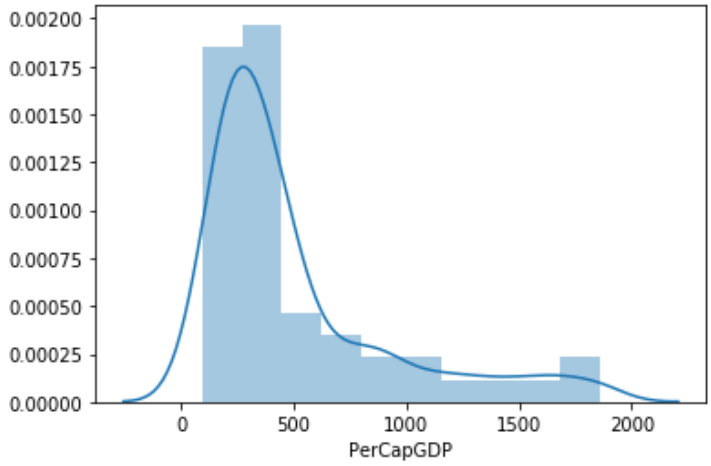

Figure 3.8: Per Capita GDP

As figure 3.5 has no negative value, bar has been started from right side of zero. This plot has used KDE (Kernel Density Estimation) to show a line of average value. As all variables are numerical. Visualization of these is almost the similar kind of figure.

```
In [21]: sns.distplot(newdf['AnnualGNI'])
Out[21]: <matplotlib.axes._subplots.AxesSubplot at 0x13a6bc8>
```

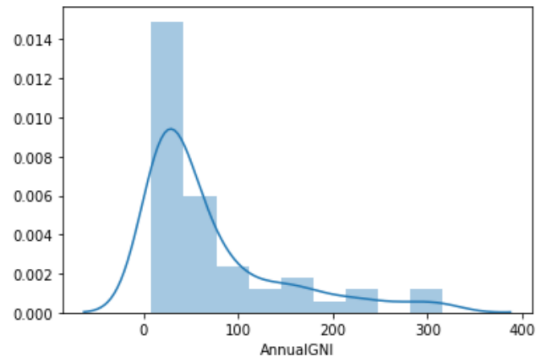

Figure 3.9: Annual GNI

```
In [24]: sns.distplot(newdf['PerCapGNIgrowth'])
Out[24]: <matplotlib.axes._subplots.AxesSubplot at 0x142c340>
```

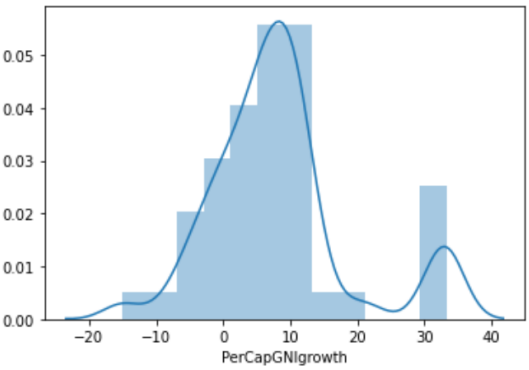

Figure 3.10: Annual GNI

```
In [22]: sns.distplot(newdf['GNIgrowth'])
Out[22]: <matplotlib.axes._subplots.AxesSubplot at 0x13e3988>
```

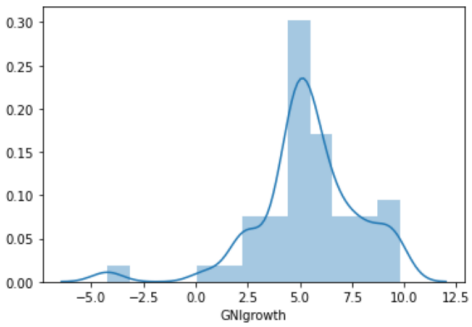

Figure 3.11: GNI Growth

```
In [27]: sns.distplot(newdf['UnemploymentRate'])
Out[27]: <matplotlib.axes._subplots.AxesSubplot at 0x9ae2268>
```

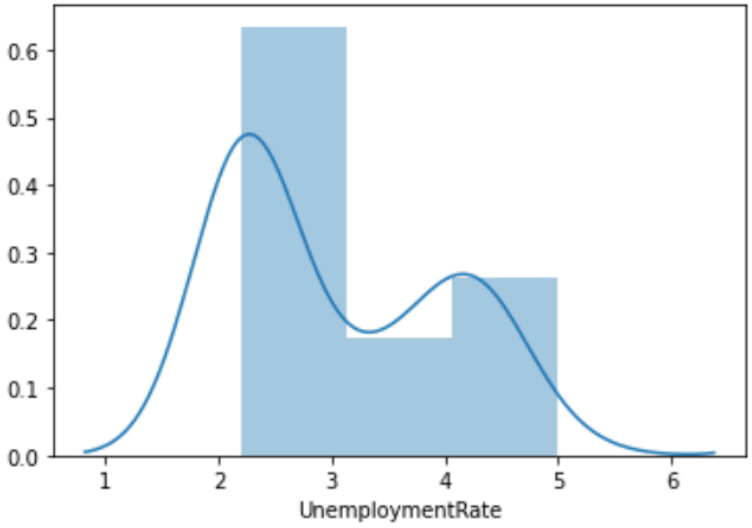

Figure 3.12: Unemployment Rate

```
In [28]: sns.distplot(newdf['CPI'])
Out[28]: <matplotlib.axes._subplots.AxesSubplot at 0x14cfb68>
```

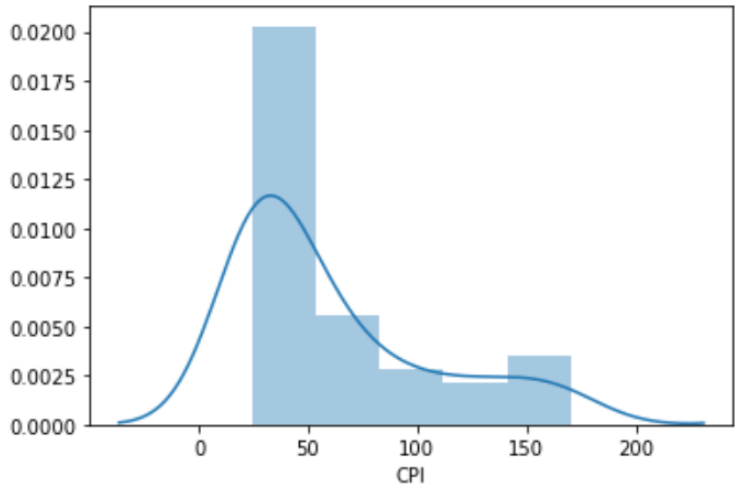

Figure 3.13: CPI

```
In [30]: sns.distplot(newdf['PopD'])
```

```
Out[30]: <matplotlib.axes._subplots.AxesSubplot at 0x1547fb8>
```

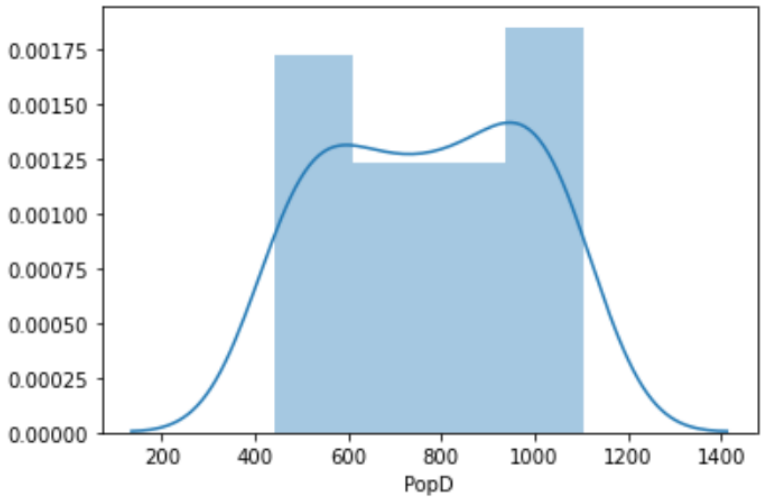

Figure 3.14: Population Density

```
In [31]: sns.distplot(newdf['NumCar'])
         import warnings
         warnings.filterwarnings("ignore")
```

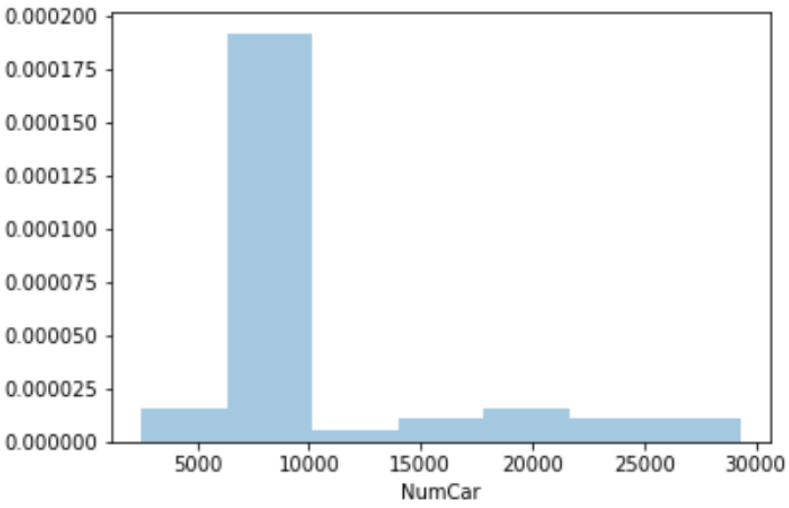

Figure 3.15: Number of Car Sold in a Year

Joint plot shows the change of variable for every year. Two samples are shown below:

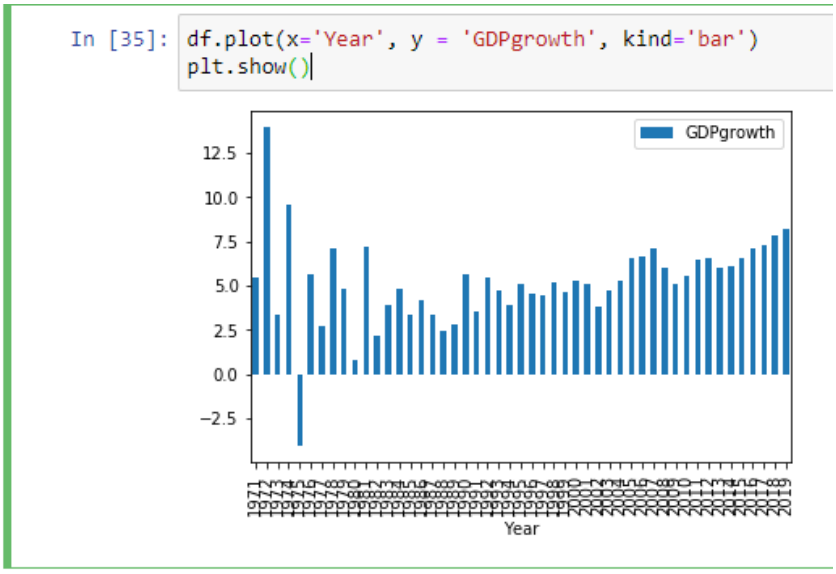

Figure 3.16: GDPgrowth

Changes of variable with positive value in every year, indicates upper from zero and bar gets increased according to the size of changed. On contrast to negative change bar starts from down to zero and gets increased based on size of value.

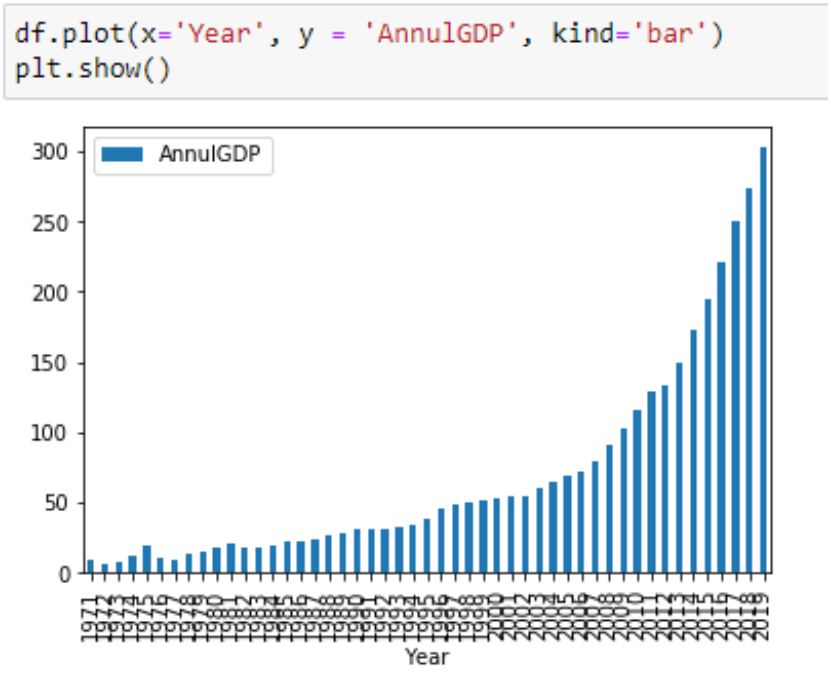

Figure 3.17: AnnualGDP

3.7 Data Preprocessing

Data preprocessing is the phase of preparing raw data to make it suitable for a machine learning model. Single row represents each observation.The dataset is partitioned into two sets. 33% data of dataset has been taken as test-set and the rest of data use as train-set. Afterwards StandardScaler is needed to scale initializing the data. It assists to normalize the data within a particular range and speed up the calculations in an algorithm as well as scale the features.

3.8 Feature Selection

Feature selection is the process of selecting a subset of relevant variables to build the machine learning model. It makes the model easier to interpret and reduces over-fitting. Correlation Matrix is bought about mapping correlated variables.

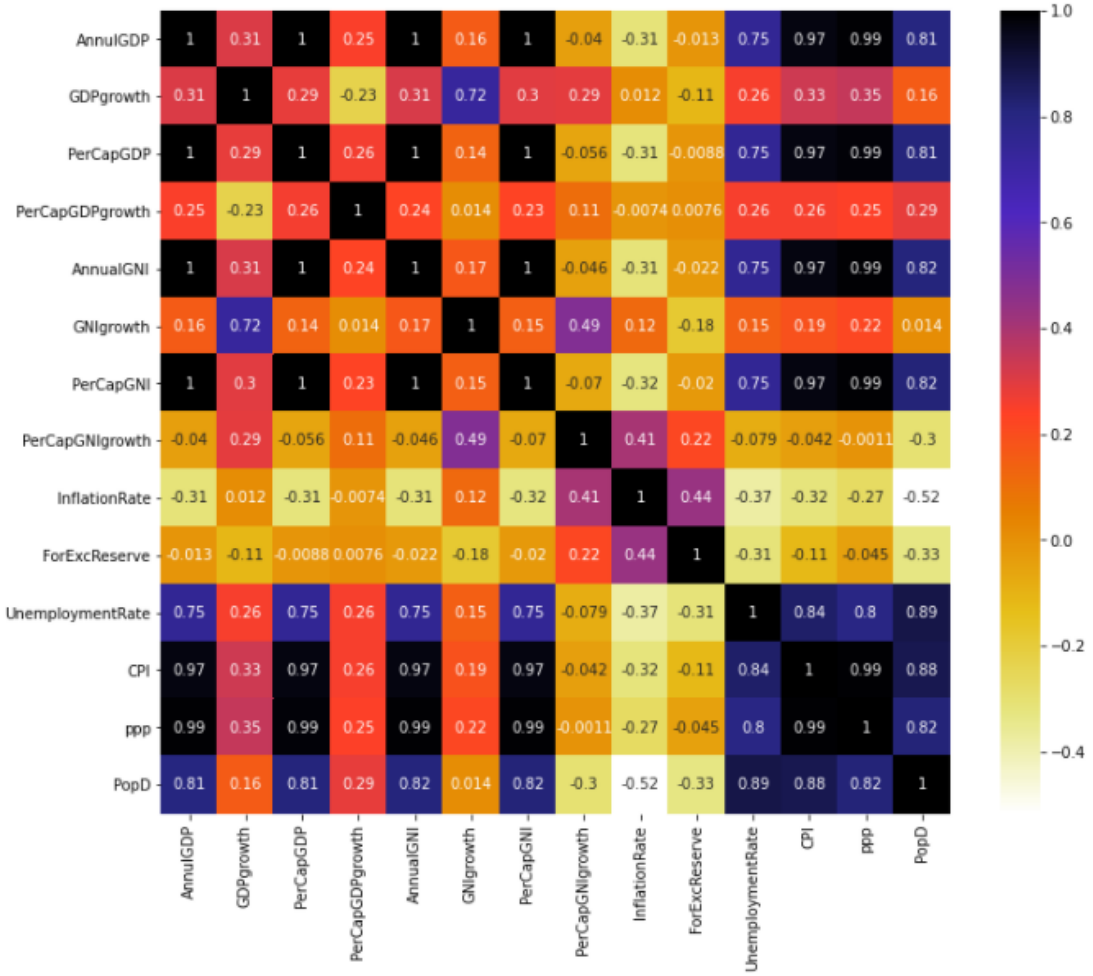

Figure 3.18: x-train Correlation heatmap

In train-set we get 8 correlation feature which are noted below: 'AnnualGNI', 'CPI', 'GNIgrowth', 'PerCapGDP', 'PerCapGNI', 'PopD', 'UnemploymentRate', 'PPP'.
We have found that 'CPI' and 'PPP' are highly correlated.

In test-set we find 12 correlation feature which are – 'AnnualGNI', 'CPI', 'ForExcReserve', 'GDPgrowth', 'GNIgrowth', 'InflationRate', 'PerCapGDP', 'PerCapGNI', 'PerCapGNIgrowth', 'PopD', 'UnemploymentRate', 'PPP'.
Again 'CPI' and 'PPP' are found highly correlated.

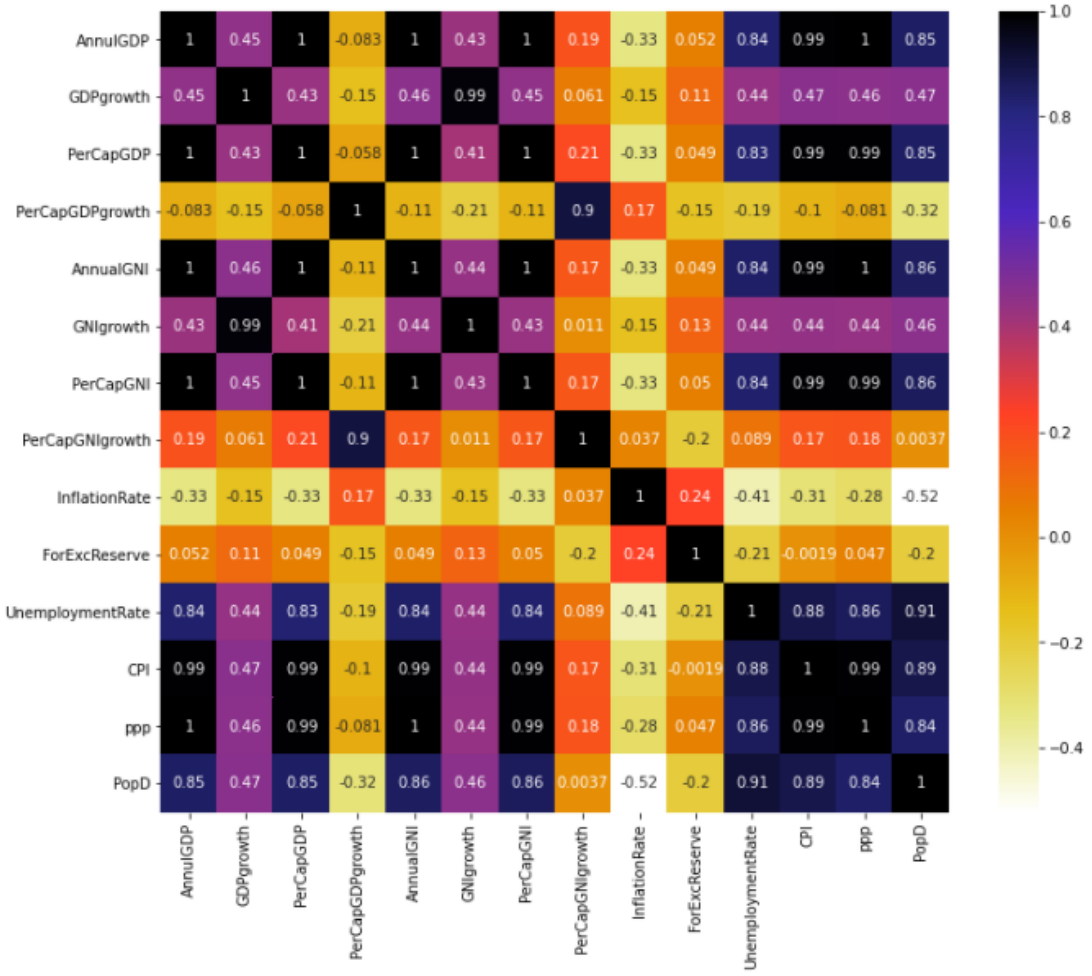

Figure 3.19: x-test Correlation heatmap

For feature importance we have used f-regression and select k-best-algorithm and evaluates f-score and p-value for the features.
P-value: The P-value is the probability of obtaining the estimated value of the parameter if the actual parameter value is zero. The smaller the value of P-value, the more significant the parameter and the less likely that the actual parameter value is zero. For instance, AnnulGDP's computed P-value is 0.01 then this indicates that there is only a 0.01 (1%) chance that the actual value of the parameter can be zero and GDPgrowth's computed P-value is 0.29 then this indicates that there is only a 0.29 (29%) chance that the actual value of the parameter can be zero.

f-score: The f-test of the overall significance is a specific form of the f-test. It compares a model with no predictors to the model that we specify. We'll find the f-test for overall significance in the analysis of variance.

	Input_Features	F_Score	P_Value
10	UnemploymentRate	19.602703	0.00
11	CPI	17.851643	0.00
13	PopD	15.810377	0.00
12	ppp	12.676902	0.00
6	PerCapGNI	8.468756	0.01
2	PerCapGDP	8.458976	0.01
4	AnnualGNI	8.403868	0.01
0	AnnulGDP	8.331556	0.01
1	GDPgrowth	1.127313	0.29
3	PerCapGDPgrowth	0.564966	0.46

Figure 3.20: Feature Importance

As we see previously 'CPI' and 'PPP' are highly correlated we have dropped PPP(12.676902) because its f-score is less than CPI(17.851643).

3.9 Model Design

We have called grid searchcv.GridSearchCV is a funcition that comes in scikit-learn's (or SK-learn) mode selection package which indicates hyper-parameter tuning in order to determine the optimal values for a given model. The performance of a model significantly depends on the value of hyper-parameter and try all possible values to know the optimal values. We have passed predefined values for hyper-parameter to the GridSearchCV function. We've done this by defining a dictionary in which we mentioned a particular hyper-parameter along with the values it can take.

We will compare the regression R squared with the grid search cv's one using max 10, min 5 time cross-validation, a procedure that consists in splitting the data 10 times into train and validation sets and for each split, the model is trained and tested. It's used to check how well the model is able to get trained by some data and predict unseen data. We found:

Selective Algorithm	Average R-squared Value
Gradient Boosting Regressor	0.67
Decision Tree Regressor	0.73
Random Forest Regressor	0.75
Ridge Regression	0.49
Lasso Regression	0.65

The Random Forest Regressor model presents better performances (average R squared of 0.75). So We must use it to get better prediction accuracy.

Algorithm Description

3.9.1 Gradient Boosting Regressor

Gradient boosting regressor:Gradient boosting is a machine learning technique for regression problems. which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. it is a type of machine learning boosting. The key idea is to set the target outcomes for this model in order to minimize the error and find good accurecy. [26] We use four types of parameter in this regression analysis which are 'learning rate','max depth', 'n estimators' and 'sub sample'. Learning rate determines the impact of each tree on the final outcome and controls the magnitude of this change in the estimates. Sub-sample is the fraction of observations to be selected for each tree. Selection is done by random sampling. n-estimators is The number of sequential trees to be modeled and tuned using CV for a particular learning rate. So, The best estimator across all searched parameters(learning rate=0.04, max depth=8, n estimators = 1000, sub-sample = 0.2) and The best score across all searched parameters: 0.87856940913959 or 87% and the accuracy is 63%. [27]

3.9.2 Decision Tree Regressor

Decision Tree generates regression or classification model according to follow tree data structure. Basically, it divides a dataset into smaller and smaller subset while a decision tree grows up at the same time. Decision nodes and leaf nodes results final output.

A decision node has two or more branch nodes. Each of these consumes value for attribute test. Leaf node represents a decision based on numerical target. Top most decision node called Root node gives the best prediction. [28]

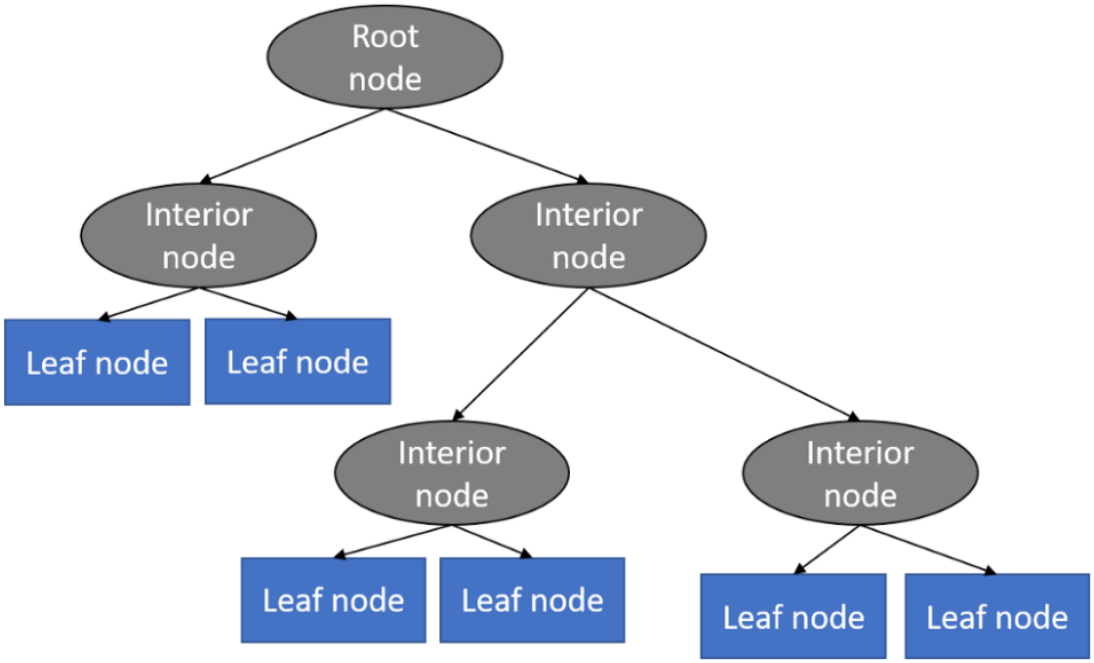

Figure 3.21: Decision Tree

3.9.3 Random Forest Regressor

Random forest is a flexible, easy to use machine learning algorithm that produces, even without hyper-parameter tuning, a great result most of the time. It is also one of the most used algorithms, because of its simplicity and diversity (it can be used for both classification and regression tasks). In this post we'll learn how the random forest algorithm works, how it differs from other algorithms and how to use it.

Random forest is a supervised learning algorithm. The "forest" it builds, is an ensemble of decision trees, usually trained with the "bagging" method. The general idea of the bagging method is that a combination of learning models increases the overall result. Random forest adds additional randomness to the model, while growing the trees. Instead of searching for the most important feature while splitting a node, it searches for the best feature among a random subset of features. This results in a wide diversity that generally results in a better model.

Therefore, in random forest, only a random subset of the features is taken into consideration by the algorithm for splitting a node. You can even make trees more random by additionally using random thresholds for each feature rather than searching for the best possible thresholds (like a normal decision tree does). [29]

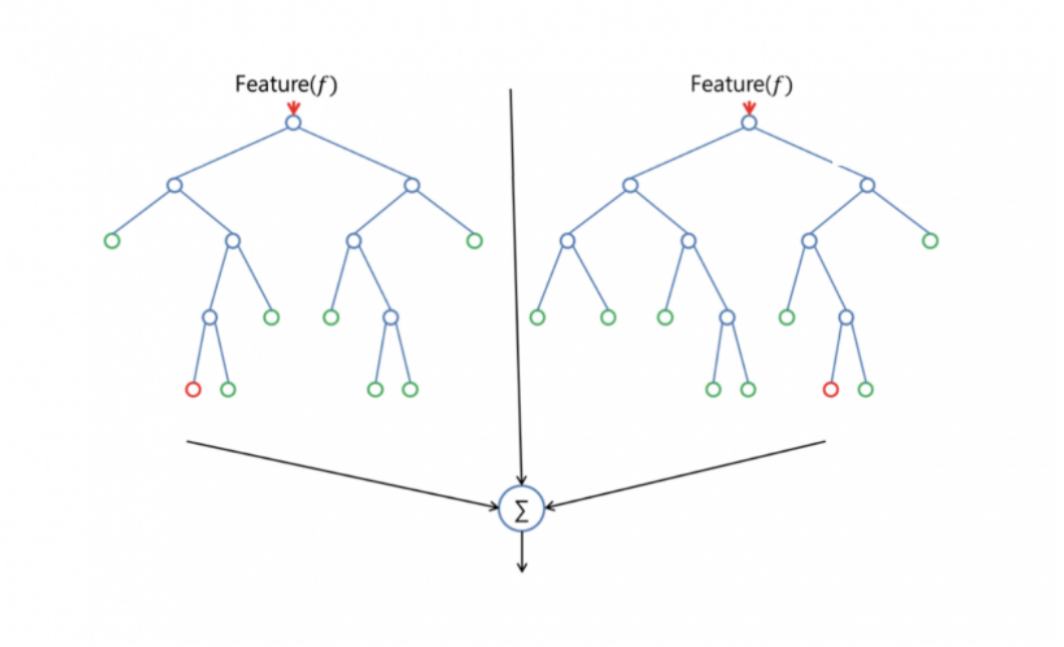

Figure 3.22: Random Forest Tree

3.9.4 Ridge Regression

Ridge Regression is a form of regularized linear regression with an L2 penalty that is widely used. This causes the coefficients for input variables that don't contribute significantly to the prediction job to decrease. Cost Function for Ridge Regression: [30]

$$\text{Cost} = \sum_{i=1}^{N} \left\{ y_i - \sum_{j=0}^{M} \beta_j x_{ij} \right\}^2 + \lambda \sum_{j=0}^{M} w_j^2$$

The punishment term is lambda. The ridge function's alpha argument denotes the value supplied here. We can adjust the penalty term by altering the values of alpha. The greater the alpha value, the greater the penalty, and therefore the size of the coefficients is lowered. In this regression study, we employ four different types of parameters: 'alpha' , 'copy-X', 'fit intercept', 'solver' 'max-iter'. The best score across all searched params: 0.8761857971298411 or 87% and the accuracy is 48%

3.9.5 Lasso Regression

Lasso Regression is a popular type of regularized linear regression that includes an L1 penalty. This has the effect of shrinking the coefficients for those input variables that do not contribute much to the prediction task. This penalty allows some coefficient values to go to the value of zero, allowing input variables to be effectively removed from the model, providing a type of automatic feature selection.
Lasso regression penalizes less important features of your dataset and makes their respective coefficients zero, thereby eliminating them. Thus it provides you with the benefit of feature selection and simple model creation. So, if the dataset has high dimensionality and high correlation, lasso regression can be used. [31]

3.10 Final Model

In this section we have tried to predict a value based on dominant variables which are 'UnemploymentRate', 'CPI', 'PopD', 'PerCapGNI', 'PerCapGDP', 'AnnulGDP', 'AnnualGNI'. We make a list of independent (feature importance) values and called this variable X and Put the dependent(Number of Cars) values in a variable called y. We use ensemble methods from the sklearn module to create Random Forest Regressor object. This object has a method called .fit() that takes the independent and dependent values as parameters and fills the regression object with data. After that model is ready to predict.

3.11 I/O Sample And Result

The methodology has five different ML algorithm and select the best one which has better performance to estimate the number of light weight passenger car sells in Bangladesh. We found Random Forest Regressor is the best algorithm for the model to predict.

INPUT: 'UnemploymentRate', 'CPI', 'PopD', 'PerCapGNI', 'PerCapGDP', 'AnnulGDP', 'AnnualGNI'

OUTPUT: Number of Passenger Car

IMF (International Monetary Fund), World Bank, The Economist (Magazine) and various organization project these macroeconomic indicators. Taking these values as input the mode outputs light weight passenger car for 2022 is 19405.5 units and for 2023 is 18017.5 units. Snapshot are given below:

```
predictedNumCar = RFR.predict([[5.40, 318.38, 1137.76, 2542.2, 1310.00, 340.00, 13462.00]])
```

```
print(predictedNumCar)
```

```
[19405.5]
```

Figure 3.23: Sales Prediction (2022)

```
predictedNumCar = RFR.predict([[4.70, 334.30, 1147.92, 2330.84, 2542.2, 430.00, 14270.00]])
```

```
print(predictedNumCar)
```

```
[18017.5]
```

Figure 3.24: Sales Prediction (2023)

CHAPTER 4
DISCUSSION AND CONCLUSION

Due to the fastest growing economy for last decade automotive sales were increasing. Fluctuation of rising economy of Bangladesh often troubles automotive sales. Since automotive industry of Bangladesh assembles car from Japan, China, Germany and other countries, local industry often assembles less or more from the real number of sold cars. This results lack of car in the market or car gets more available. For this reason, old model cars are subjected to sell in the market with loss as at the same time buyer doesn't want to purchase old model car. This research will lessen the difference of assembling car from the real number and reduce the loss of local automotive industries as the models estimates the passenger cars. Local industry can get the idea about the number of car they should assemble from abroad.

4.1 Limitations

There are some limitations of the research.

- We have tried to collect all data from 1971 to 2019. But we didn't get all the data. So the model has been trained by less data with some missing values.
- This model predicts car only economical aspect. In reality car sale dependents on various factors.
- This model has skipped natural calamities as well as pandemic. For 2020 and 2021 model won't output properly.

4.2 Future Scope

- This research has considered a small area like Bangladesh and estimated cars from only economical point of view.The model will be applicable to large area by adding demographic factors, Social and cultural aspects, e-WM (Electronic Word of Mouth) and political aspects.
- Advance algorithm like ANN (Artificial Neural Network) and GA (Genetic Algorithm) can use in the model taking categorical values.

REFERENCES

[1] "Machine learning." Last accessed on June 29, 2021, at 04:31:00AM. [Online]. Available:https://en.wikipedia.org/wiki/Machine$_{l}earning$.

[2] P. S. U. STEVEN A. GREENLAW, UNIVERSITY OF MARY WASHINGTON DAVID SHAPIRO, *Principal Of Macroeconomics 2e*. Timothy Taylor, 2011.

[3] "History of macroeconomics." Last accessed on July 01, 2021, at 02:33:00PM. [Online]. Available:https://www.investopedia.com/terms/m/macroeconomics.asp.

[4] "Areas of macroeconomics research." Last accessed on July 01, 2021, at 10:49:00PM. [Online]. Available:https://www.investopedia.com/terms/m/macroeconomics.asp.

[5] HELNA, "The importance of macroeconomics," Nov 12, 2014. Last accessed on July 02, 2021, at 01:35:00AM. [Online]. Available:https://owlcation.com/stem/Meaning-and-importance-of-Macroeconomics.

[6] F. Pehlivanoğlu and R. Riyanti, "Macroeconomic effect on the automobile sales in top four automobile production countries," *Kocaeli Üniversitesi Sosyal Bilimler Dergisi*, pp. 139 – 161, 2018.

[7] L. WOFFORD, "How do both microeconomic and macroeconomic factors influence the global automobile industry?." Last accessed on July 03, 2021, at 09:28:00PM. [Online]. Available:https://www.enotes.com/homework-help/how-both-microeconomic-macroeconomic-factors-1344842.

[8] A. A. R. Fidlizan Muhammad1, Mohd Yahya Mohd Hussin1, "Automobile sales and macroeconomic variables: A pooled mean group analysis for asean countries," *IOSR Journal of Business and Management (IOSRJBM)*, vol. 02, no. 2, pp. 15–21, 2012.

[9] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science*, vol. 349, no. 6245, pp. 255–260, 2015.

[10] F. Zhang, J. Yang, Y. Guo, and H. Gu, "Multi-source heterogeneous and xboost vehicle sales forecasting model," in *International Conference on Machine Learning and Big Data Analytics for IoT Security and Privacy*, pp. 340–347, Springer, 2020.

[11] S. Johan, "Macroeconomic determinants of automobile sales in indonesia: an empirical study in 1986-2016," *Binus Business Review*, vol. 10, no. 3, pp. 159–166, 2019.

[12] T. M. Bach, A. M. Machado, C. Kudlawicz-Franco, T. S. Martins, and C. P. Da Veiga, "Monetary policy and the automotive retail performance in brazil," *Journal of Business and Retail Management Research*, vol. 11, no. 2, 2017.

[13] R. Chifurira, I. Mudhombo, M. Chikobvu, and D. Dubihlela, "The impact of inflation on the automobile sales in south africa," *Mediterranean Journal of Social Sciences*, vol. 5, no. 7, pp. 200–200, 2014.

[14] S. Johan, "Macroeconomic determinants of auto sales in asean: An empirical study in five major asean countries," *JAS (Journal of ASEAN Studies)*, vol. 8, no. 2, pp. 95–110, 2020.

[15] "Macroeconomics." Last accessed on August 19, 2021, at 05:06:00AM. [Online]. Available:https://en.wikipedia.org/wiki/Macroeconomics.

[16] "Definition of 'gross domestic product'." Last accessed on August 20, 2021, at 05:34:00AM. [Online]. Available:https://economictimes.indiatimes.com/definition/gross-domestic-product.

[17] T. BROCK, "Per capita gdp," Dec 25, 2020. Last accessed on August 21, 2021, at 03:34:00PM. [Online]. Available:https://www.investopedia.com/terms/p/per-capita-gdp.as.

[18] "Economic growth rate," Dec 25, 2020. Last accessed on August 21, 2021, at 11:47:00PM. [Online]. Available:https://www.investopedia.com/terms/e/economicgrowthrate.asp.

[19] "Gross national income." Last accessed on August 21, 2021, at 12:17:00AM. [Online]. Available:https://en.wikipedia.org/wiki/Gross$_n ational_i ncome.$

[20] "Methodology." Last accessed on August 21, 2021, at 03:20:00PM. [Online]. Available:https://en.wikipedia.org/wiki/List$_o f_c ountries_b y_G NI_{(nominal)}{}_p er_c apita.$

[21] T. I. TEAM, "What is purchasing power parity (ppp)?." Last accessed on August 21, 2021, at 09:27:00PM. [Online]. Available:https://www.investopedia.com/updates/purchasing-power-parity-ppp/.

[22] M. HARGRAVE, "What are foreign exchange reserves?." Last accessed on August 21, 2021, at 10:11:00PM. [Online]. Available:https://www.investopedia.com/terms/f/foreign-exchange-reserves.asp.

[23] the MasterClass staff, "What is inflation rate? learn how to calculate the inflation rate." Last accessed on August 21, 2021, at 10:41:00PM. [Online]. Available:https://www.masterclass.com/articles/what-is-inflation-ratewhat-is-the-inflation-rate.

[24] T. I. TEAM, "Unemployment rate." Last accessed on August 21, 2021, at 11:08:00PM. [Online]. Available:https://www.investopedia.com/terms/u/unemploymentrate.asp.

[25] J. FERNANDO, "Consumer price index (cpi)." Last accessed on August 21, 2021, at 11:20:00PM. [Online]. Available:https://www.investopedia.com/terms/c/consumerpriceindex.asp.

[26] "Gradient boosting." Last accessed on August 31, 2021, at 01:35:00PM. [Online]. Available:https://en.m.wikipedia.org/wiki/Gradient$_b oosting.$

[27] A. JAIN, “Complete machine learning guide to parameter tuning in gradient boosting (gbm) in python,” Feb21,2016. Last accessed on August 31, 2021, at 01:43:00PM. [Online]. Available:https://www.analyticsvidhya.com/blog/2016/02/complete-guide-parameter-tuning-gradient-boosting-gbm-python/.

[28] G. M. K, “Machine learning basics: Decision tree regression,” Jul15, 2020. Last accessed on August 31, 2021, at 4:47:00AM. [Online]. Available:https://towardsdatascience.com/machine-learning-basics-decision-tree-regression-1d73ea003fda.

[29] N. Donges, “A complete guide to the random forest algorithm,” Jul15, 2020. Last accessed on August 31, 2021, at 9:47:00AM. [Online]. Available:https://https://builtin.com/data-science/random-forest-algorithm.

[30] A. BILOGUR, “Ridge regression cost function.” Last accessed on August 31, 2021, at 5:47:00AM. [Online]. Available:https://www.kaggle.com/residentmario/ridge-regression-cost-function.

[31] J. Brownlee, “How to develop lasso regression models in python,” October 12,2020. Last accessed on August 31, 2021, at 6:25:00AM. [Online]. Available:https://https://machinelearningmastery.com/lasso-regression-with-python/.