PREDICTIVE MODELING PROJECT REPORT 2023

Sangeeth A

PGP-DSBA Online

July - 2022

CONTENTS

LIST OF FIGURES

LIST OF TABLES

PROBLEM 1 – SUMMARY

The comp-activ databases is a collection of a computer systems activity measures . The data was collected from a Sun Sparcstation 20/712 with 128 Mbytes of memory running in a multi-user university department. Users would typically be doing a large variety of tasks ranging from accessing the internet, editing files or running very cpu-bound programs.

As you are a budding data scientist you thought to find out a linear equation to build a model to predict 'usr'(Portion of time (%) that cpus run in user mode) and to find out how each attribute affects the system to be in 'usr' mode using a list of system attributes.

INTRODUCTION

The purpose of this exercise is to conduct linear regression on the given dataset and develop a linear equation to predict 'usr'(Portion of time (%) that cpus run in user mode) and to find out how each attribute affects the system to be in 'usr' mode using a list of system attributes.

DATA DESCRIPTION

System measures used:

1. lread - Reads (transfers per second) between system memory and user memory
2. lwrite - writes (transfers per second) between system memory and user memory
3. scall - Number of system calls of all types per second.
4. sread - Number of system read calls per second.
5. swrite - Number of system write calls per second.
6. fork - Number of system fork calls per second.
7. exec - Number of system exec calls per second.
8. rchar - Number of characters transferred per second by system read calls.
9. wchar - Number of characters transfreed per second by system write calls.
10. pgout - Number of page out requests per second.
11. ppgout - Number of pages, paged out per second.
12. pgfree - Number of pages per second placed on the free list.
13. pgscan - Number of pages checked if they can be freed per second.
14. atch - Number of page attaches (satisfying a page fault by reclaiming a page in memory) per second.
15. pgin - Number of page-in requests per second.

16. ppgin - Number of pages paged in per second.
17. pflt - Number of page faults caused by protection errors (copy-on-writes).
18. vflt - Number of page faults caused by address translation.
19. runqsz - Process run queue size (The number of kernel threads in memory that are waiting for a CPU to run.
    Typically, this value should be less than 2. Consistently higher values mean that the system might be CPU-bound.).
20. freemem - Number of memory pages available to user processes.
21. freeswap - Number of disk blocks available for page swapping.
22. usr - Portion of time (%) that cpus run in user mode

SAMPLE OF THE DATASET

	lread	lwrite	scall	sread	swrite	fork	exec	rchar	wchar	pgout	ppgout	pgfree	pgscan	atch	pgin	ppgin	pflt	vflt	runqsz
0	1	0	2147	79	68	0.2	0.2	40671.0	53995.0	0.0	0.0	0.0	0.0	0.0	1.6	2.6	16.00	26.40	CPU_Bound
1	0	0	170	18	21	0.2	0.2	448.0	8385.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	15.63	16.83	Not_CPU_Bound
2	15	3	2162	159	119	2.0	2.4	NaN	31950.0	0.0	0.0	0.0	0.0	1.2	6.0	9.4	150.20	220.20	Not_CPU_Bound
3	0	0	160	12	16	0.2	0.2	NaN	8670.0	0.0	0.0	0.0	0.0	0.0	0.2	0.2	15.60	16.80	Not_CPU_Bound
4	5	1	330	39	38	0.4	0.4	NaN	12185.0	0.0	0.0	0.0	0.0	0.0	1.0	1.2	37.80	47.60	Not_CPU_Bound

freemem	freeswap	usr
4670	1730946	95
7278	1869002	97
702	1021237	87
7248	1863704	98
633	1760253	90

Table 1 – Sample dataset

The dataset contains 22 features, our dependent variable is “usr”, we have to create linear regression equation based on other 21 independent features.

EXPLORATORY DATA ANALYSIS

The dataset contains 8192 entries and 22 columns, there are null value present in the dataset.

NO.	Column	Non – Null content	Data Type
1	lread	8192	Int64
2	lwrite	8192	Int64
3	scall	8192	Int64
4	sread	8192	Int64
5	swrite	8192	Int64
6	fork	8192	Float64

7	exec	8192	Float64
8	rchar	8088	Float64
9	wchar	8177	Float64
10	pgout	8192	Float64
11	ppgout	8192	Float64
12	pgfree	8192	Float64
13	pgscan	8192	Float64
14	atch	8192	Float64
15	pgin	8192	Float64
16	ppgin	8192	Float64
17	pflt	8192	Float64
18	vflt	8192	Float64
19	runqsz	8192	Object
20	freemem	8192	Int64
21	freeswap	8192	Int64
22	usr	8192	Int64

Table 2 – Exploratory Data Analysis

DESCRIPTIVE DATA ANALYSIS

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
lread	8192.0	NaN	NaN	NaN	19.559692	53.353799	0.0	2.0	7.0	20.0	1845.0
lwrite	8192.0	NaN	NaN	NaN	13.106201	29.891726	0.0	0.0	1.0	10.0	575.0
scall	8192.0	NaN	NaN	NaN	2306.318237	1633.617322	109.0	1012.0	2051.5	3317.25	12493.0
sread	8192.0	NaN	NaN	NaN	210.47998	198.980146	6.0	86.0	166.0	279.0	5318.0
swrite	8192.0	NaN	NaN	NaN	150.058228	160.47898	7.0	63.0	117.0	185.0	5456.0
fork	8192.0	NaN	NaN	NaN	1.884554	2.479493	0.0	0.4	0.8	2.2	20.12
exec	8192.0	NaN	NaN	NaN	2.791998	5.212456	0.0	0.2	1.2	2.8	59.56
rchar	8088.0	NaN	NaN	NaN	197385.728363	239837.493526	278.0	34091.5	125473.5	267828.75	2526649.0
wchar	8177.0	NaN	NaN	NaN	95902.992785	140841.707911	1498.0	22916.0	46619.0	106101.0	1801623.0
pgout	8192.0	NaN	NaN	NaN	2.285317	5.307038	0.0	0.0	0.0	2.4	81.44
ppgout	8192.0	NaN	NaN	NaN	5.977229	15.21459	0.0	0.0	0.0	4.2	184.2
pgfree	8192.0	NaN	NaN	NaN	11.919712	32.36352	0.0	0.0	0.0	5.0	523.0
pgscan	8192.0	NaN	NaN	NaN	21.526849	71.14134	0.0	0.0	0.0	0.0	1237.0
atch	8192.0	NaN	NaN	NaN	1.127505	5.708347	0.0	0.0	0.0	0.6	211.58
pgin	8192.0	NaN	NaN	NaN	8.27796	13.874978	0.0	0.6	2.8	9.765	141.2
ppgin	8192.0	NaN	NaN	NaN	12.388586	22.281318	0.0	0.6	3.8	13.8	292.61
pflt	8192.0	NaN	NaN	NaN	109.793799	114.419221	0.0	25.0	63.8	159.6	899.8
vflt	8192.0	NaN	NaN	NaN	185.315796	191.000603	0.2	45.4	120.4	251.8	1365.0
runqsz	8192	2	Not_CPU_Bound	4331	NaN	NaN	NaN	NaN	NaN	NaN	NaN
freemem	8192.0	NaN	NaN	NaN	1763.456299	2482.104511	55.0	231.0	579.0	2002.25	12027.0
freeswap	8192.0	NaN	NaN	NaN	1328125.959839	422019.426957	2.0	1042623.5	1289289.5	1730379.5	2243187.0
usr	8192.0	NaN	NaN	NaN	83.968872	18.401905	0.0	81.0	89.0	94.0	99.0

Table 3 – Descriptive Data Analysis

1. Most of the time runqsz is in not_CPU_Bound mode
2. Total 8192 entries are there.
3. Only runqsz has object value among all the columns.

Problem 1 – LINEAR REGRESSION

The comp-activ databases is a collection of a computer systems activity measures . The data was collected from a Sun Sparcstation 20/712 with 128 Mbytes of memory running in a multi-user university department. Users would typically be doing a large variety of tasks ranging from accessing the internet, editing files or running very cpu-bound programs.

We are expected to find out a linear equation to build a model to predict 'usr'(Portion of time (%) that cpus run in user mode) and to find out how each attribute affects the system to be in 'usr' mode using a list of system attributes.

1.1 Read the data and do exploratory data analysis. Describe the data briefly. (Check the Data types, shape, EDA, 5 point summary). Perform Univariate, Bivariate Analysis, Multivariate Analysis.

The problem statement is to read the data and do the initial necessary steps and conduct exploratory data analysis on the dataset provided.

	lread	lwrite	scall	sread	swrite	fork	exec	rchar	wchar	pgout	ppgout	pgfree	pgscan	atch	pgin	ppgin	pflt	vflt	runqs
0	1	0	2147	79	68	0.2	0.2	40671.0	53995.0	0.0	0.0	0.0	0.0	0.0	1.6	2.6	16.00	26.40	CPU_Boun
1	0	0	170	18	21	0.2	0.2	448.0	8385.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	15.63	16.83	Not_CPU_Boun
2	15	3	2162	159	119	2.0	2.4	NaN	31950.0	0.0	0.0	0.0	0.0	1.2	6.0	9.4	150.20	220.20	Not_CPU_Boun
3	0	0	160	12	16	0.2	0.2	NaN	8670.0	0.0	0.0	0.0	0.0	0.0	0.2	0.2	15.60	16.80	Not_CPU_Boun
4	5	1	330	39	38	0.4	0.4	NaN	12185.0	0.0	0.0	0.0	0.0	0.0	1.0	1.2	37.80	47.60	Not_CPU_Boun

freemem	freeswap	usr
4670	1730946	95
7278	1869002	97
702	1021237	87
7248	1863704	98
633	1760253	90

Table 1.1.1 – Sample dataset

The above shown is the head of the dataset, the dataset contains twenty two columns and they are,

System measures used:

1. lread - Reads (transfers per second) between system memory and user memory
2. lwrite - writes (transfers per second) between system memory and user memory
3. scall - Number of system calls of all types per second.
4. sread - Number of system read calls per second.

5. swrite - Number of system write calls per second.
6. fork - Number of system fork calls per second.
7. exec - Number of system exec calls per second.
8. rchar - Number of characters transferred per second by system read calls.
9. wchar - Number of characters transfreed per second by system write calls.
10. pgout - Number of page out requests per second.
11. ppgout - Number of pages, paged out per second.
12. pgfree - Number of pages per second placed on the free list.
13. pgscan - Number of pages checked if they can be freed per second.
14. atch - Number of page attaches (satisfying a page fault by reclaiming a page in memory) per second.
15. pgin - Number of page-in requests per second.
16. ppgin - Number of pages paged in per second.
17. pflt - Number of page faults caused by protection errors (copy-on-writes).
18. vflt - Number of page faults caused by address translation.
19. runqsz - Process run queue size (The number of kernel threads in memory that are waiting for a CPU to run.
    Typically, this value should be less than 2. Consistently higher values mean that the system might be CPU-bound.).
20. freemem - Number of memory pages available to user processes.
21. freeswap - Number of disk blocks available for page swapping.
22. usr - Portion of time (%) that cpus run in user mode

All the variables in the dataset is numerical except runqsz. Usr column is the dependent variable and based on other 21 independent variables we have to built a linear equation to find out how each attribute affects the system to be in 'usr' mode

Now let us check the data types of the different features present in the dataset.

NO.	Column	Non – Null content	Data Type
1	lread	8192	Int64
2	lwrite	8192	Int64
3	scall	8192	Int64
4	sread	8192	Int64
5	swrite	8192	Int64
6	fork	8192	Float64
7	exec	8192	Float64
8	rchar	8088	Float64
9	wchar	8177	Float64

10	pgout	8192	Float64
11	ppgout	8192	Float64
12	pgfree	8192	Float64
13	pgscan	8192	Float64
14	atch	8192	Float64
15	pgin	8192	Float64
16	ppgin	8192	Float64
17	pflt	8192	Float64
18	vflt	8192	Float64
19	runqsz	8192	Object
20	freemem	8192	Int64
21	freeswap	8192	Int64
22	usr	8192	Int64

Table 1.1.2 – Data info

There are total 8192 entries and 22 columns in the dataset.

Now we can check for the null values present in the dataset,

NO.	Column	Null Value Present
1	lread	0
2	lwrite	0
3	scall	0
4	sread	0
5	swrite	0
6	fork	0
7	exec	0
8	rchar	104
9	wchar	15
10	pgout	0
11	ppgout	0
12	pgfree	0
13	pgscan	0
14	atch	0
15	pgin	0
16	ppgin	0
17	pflt	0
18	vflt	0
19	runqsz	0
20	freemem	0

21	freeswap	0
22	usr	0

Table 1.1.3 – Null value check

There is no null value present in rchar and wchar columns.

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
lread	8192.0	NaN	NaN	NaN	19.559692	53.353799	0.0	2.0	7.0	20.0	1845.0
lwrite	8192.0	NaN	NaN	NaN	13.106201	29.891726	0.0	0.0	1.0	10.0	575.0
scall	8192.0	NaN	NaN	NaN	2306.318237	1633.617322	109.0	1012.0	2051.5	3317.25	12493.0
sread	8192.0	NaN	NaN	NaN	210.47998	198.980146	6.0	86.0	166.0	279.0	5318.0
swrite	8192.0	NaN	NaN	NaN	150.058228	160.47898	7.0	63.0	117.0	185.0	5456.0
fork	8192.0	NaN	NaN	NaN	1.884554	2.479493	0.0	0.4	0.8	2.2	20.12
exec	8192.0	NaN	NaN	NaN	2.791998	5.212456	0.0	0.2	1.2	2.8	59.56
rchar	8088.0	NaN	NaN	NaN	197385.728363	239837.493526	278.0	34091.5	125473.5	267828.75	2526649.0
wchar	8177.0	NaN	NaN	NaN	95902.992785	140841.707911	1498.0	22916.0	46619.0	106101.0	1801623.0
pgout	8192.0	NaN	NaN	NaN	2.285317	5.307038	0.0	0.0	0.0	2.4	81.44
ppgout	8192.0	NaN	NaN	NaN	5.977229	15.21459	0.0	0.0	0.0	4.2	184.2
pgfree	8192.0	NaN	NaN	NaN	11.919712	32.36352	0.0	0.0	0.0	5.0	523.0
pgscan	8192.0	NaN	NaN	NaN	21.526849	71.14134	0.0	0.0	0.0	0.0	1237.0
atch	8192.0	NaN	NaN	NaN	1.127505	5.708347	0.0	0.0	0.0	0.6	211.58
pgin	8192.0	NaN	NaN	NaN	8.27796	13.874978	0.0	0.6	2.8	9.765	141.2
ppgin	8192.0	NaN	NaN	NaN	12.388586	22.281318	0.0	0.6	3.8	13.8	292.61
pflt	8192.0	NaN	NaN	NaN	109.793799	114.419221	0.0	25.0	63.8	159.6	899.8
vflt	8192.0	NaN	NaN	NaN	185.315796	191.000603	0.2	45.4	120.4	251.8	1365.0
runqsz	8192	2	Not_CPU_Bound	4331	NaN	NaN	NaN	NaN	NaN	NaN	NaN
freemem	8192.0	NaN	NaN	NaN	1763.456299	2482.104511	55.0	231.0	579.0	2002.25	12027.0
freeswap	8192.0	NaN	NaN	NaN	1328125.959839	422019.426957	2.0	1042623.5	1289289.5	1730379.5	2243187.0
usr	8192.0	NaN	NaN	NaN	83.968872	18.401905	0.0	81.0	89.0	94.0	99.0

Table 1.1.4 – Summary statistics

The above shown table shows the summary statistics of the dataset, runqsz is the only categorical column present in the dataset. Most of the time the system in not CPU bound mode.

EXPLORATORY DATA ANALYSIS

Now let us do the exploratory data analysis of the given data.

UNIVARIATE ANALYSIS

Fig 1.1.1 –Univariate analysis (HIst Plot)

The above shown is the hist plot of the different features of the dataset. The observations from the hist plot is given below.

1. Most of the time system run in not CPU bound mode.
2. Zero value is present in most of the features.
3. In the case of atch feature almost all the values are very low.
4. The portion of time CPU run in usr mode is more in between 80 and 100.
5. Number of pages out request and number of pages out is very low.

Now let us check the boxplot of the numerical features present in the dataset.

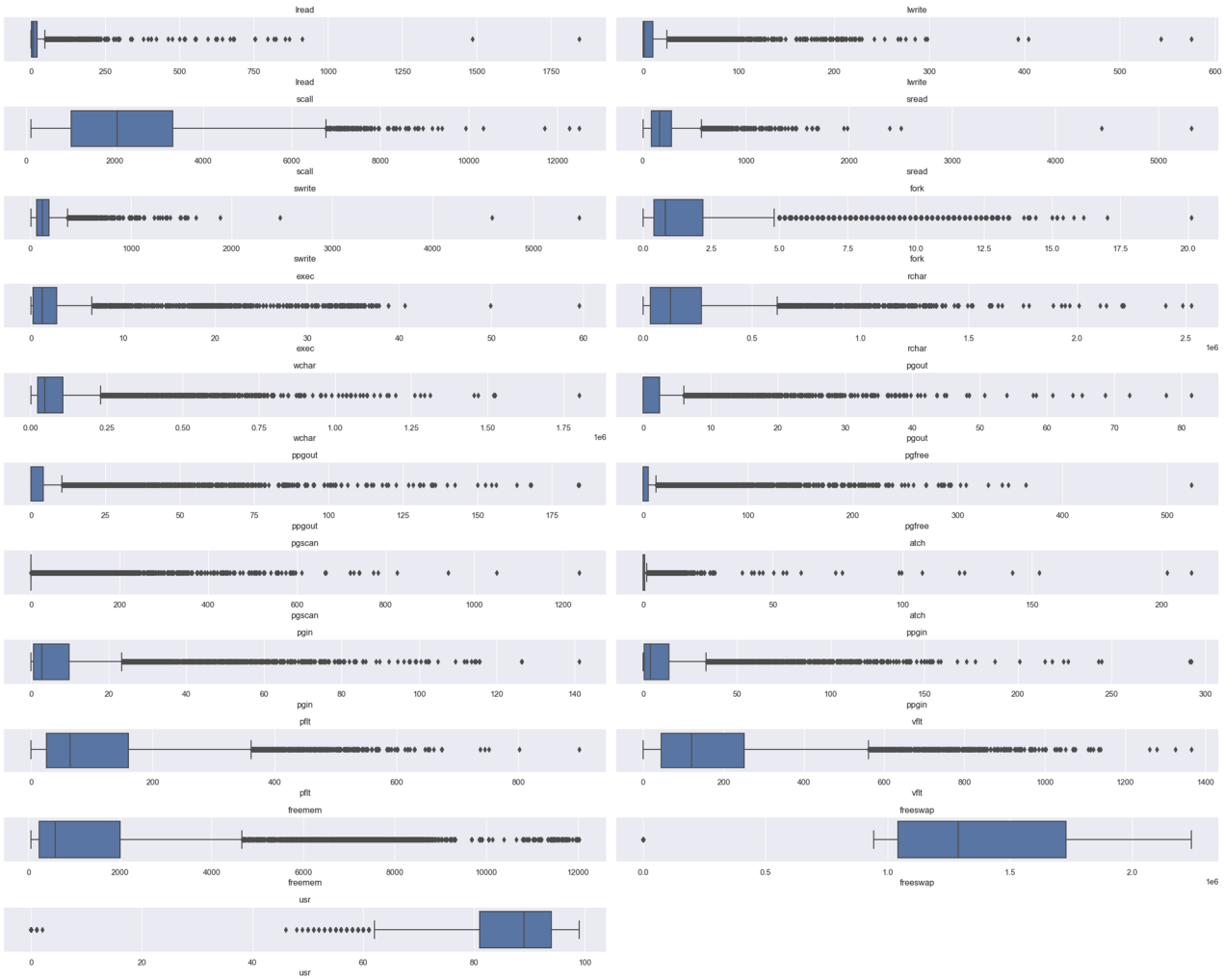


Fig 1.1.2 –Univariate analysis (Box Plot)

The above shown is the box plot of different features in the dataset. The observations from the dataset is below,

1. All the features in the dataset contains outliers.
2. From the boxplot it is clear that almost all the columns in the dataset is right skewed.

We can check treating about outliers in the next session.

BIVARIATE ANALYSIS

Now let us do the bivariate analysis of the data.

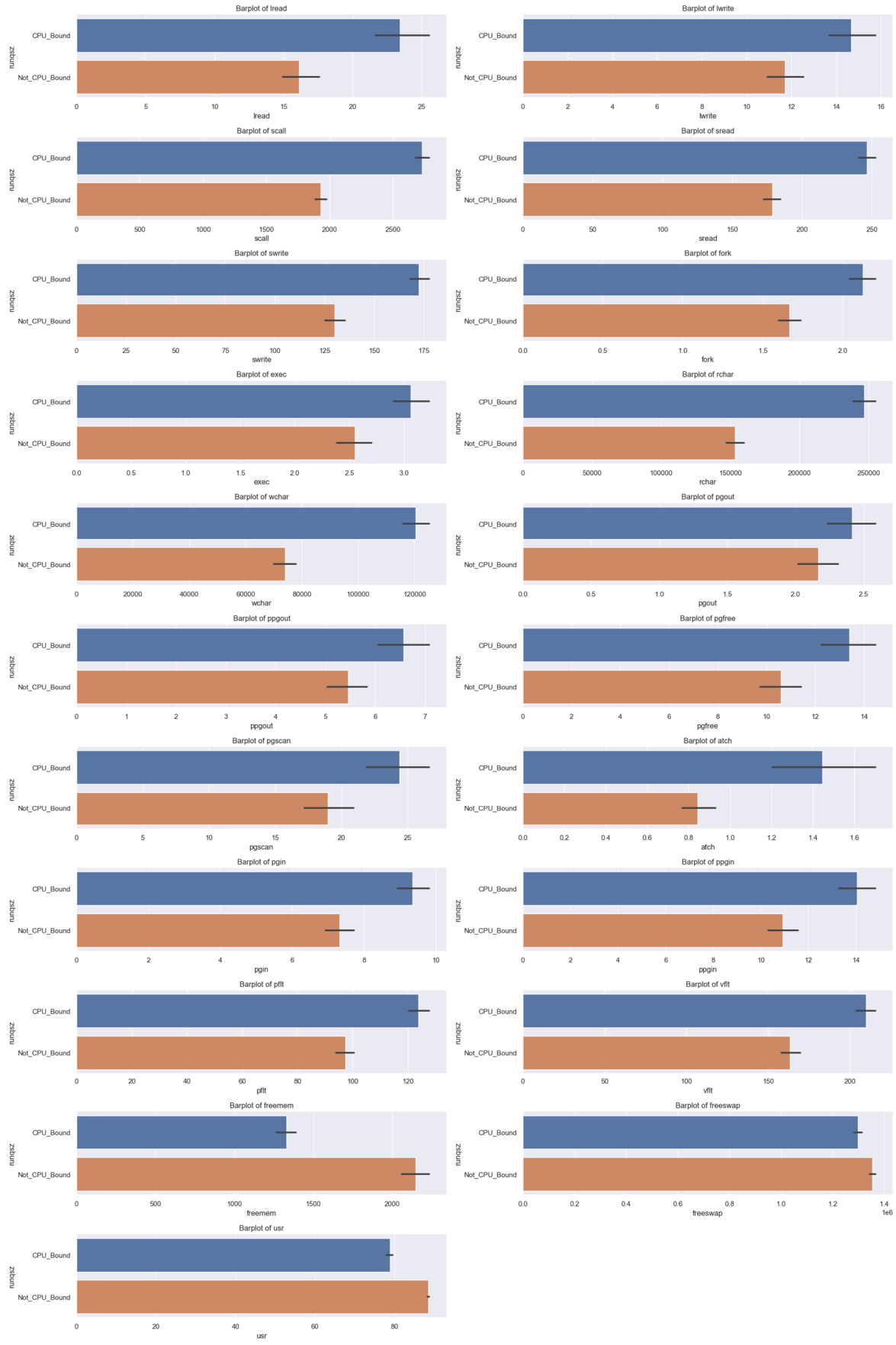


Fig 1.1.3 –Bivariate analysis (runqsz v/s others)

While considering other features with runqsz from the above plot it I clear that most of the time the system in CPU bound mode. Only in the case of usr, freemem and freeswap the system run most of the time in not CPU bound mode.

MULTIVARIATE ANALYSIS

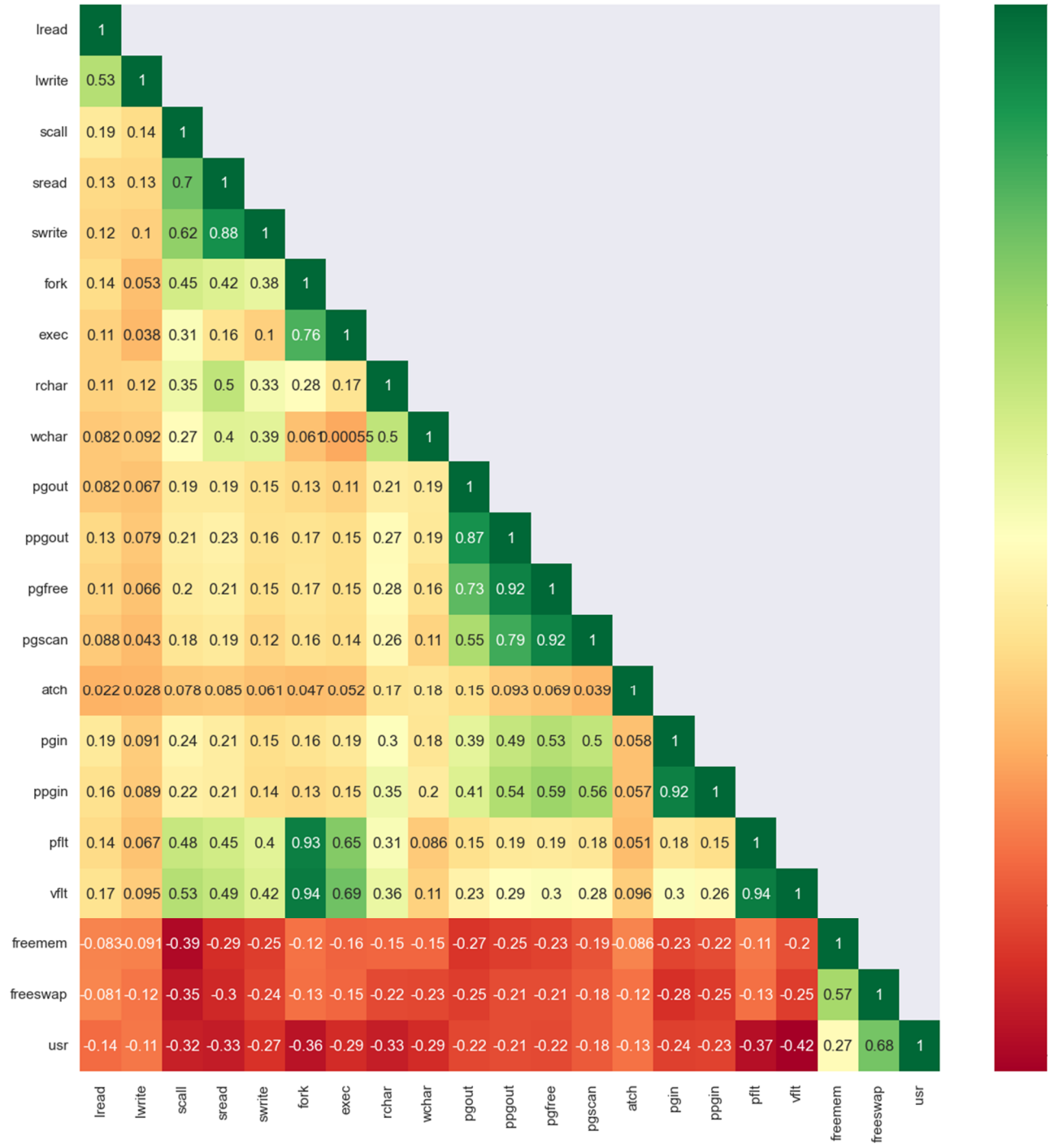


Fig 1.1.4 –Multivariate analysis (Heatmap)

The figure represents the correlation heatmap between different features of the dataset. From the dataset it is clear that there is high correlation between pgfree and ppgout, Number of page faults caused by protection errors and Number of page faults caused by address translation.

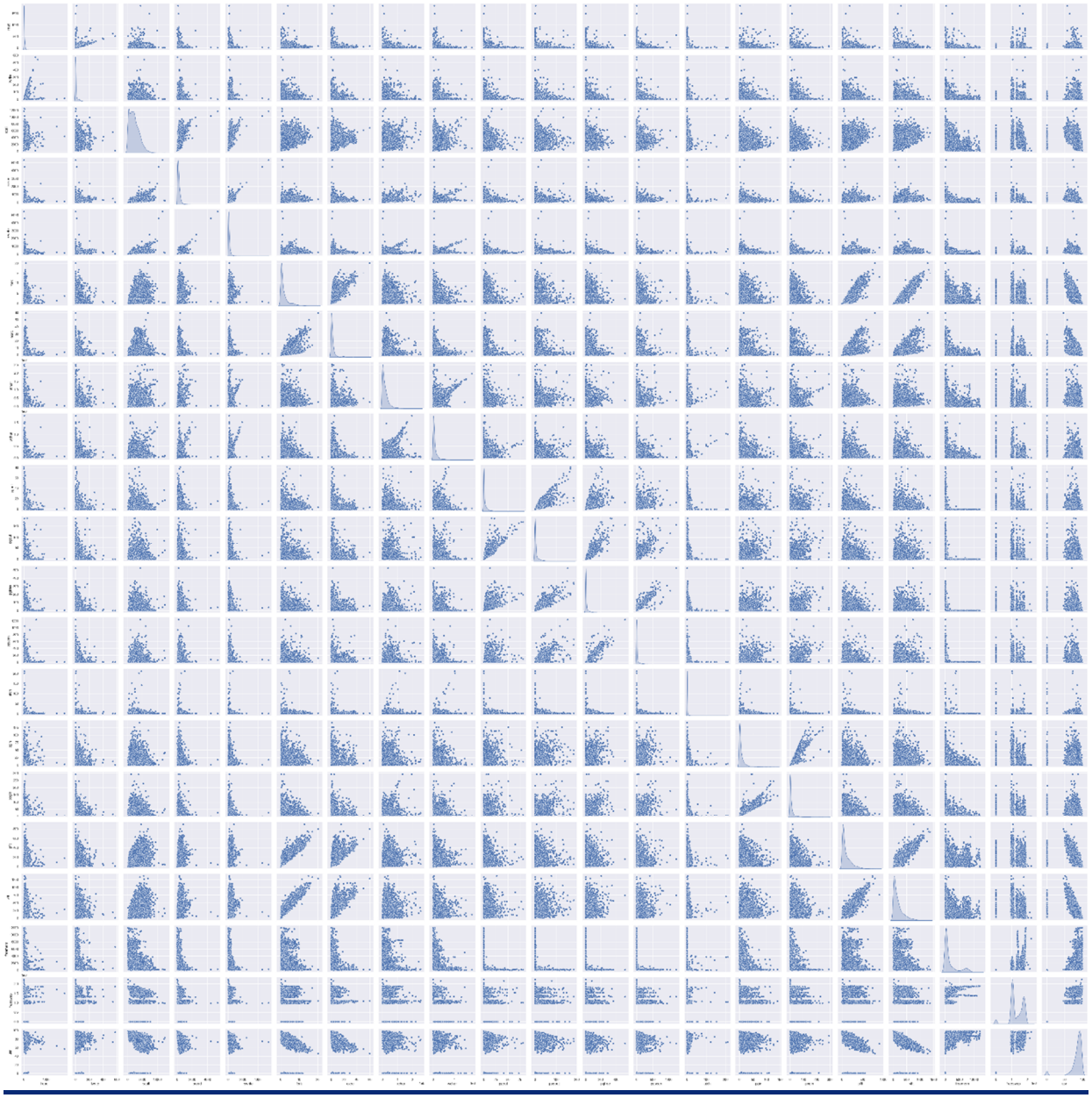


Fig 1.1.5 –Multivariate analysis (Pairplot)

Pairplot is similar to heatmap, it shows the correlation between the features of the dataset.

Next let us check the five point summary of the dataset,

```
Column Name :  lread
Minimum Value :  0
25% =  2.0
50% or Median=  7.0
75% =  20.0
Maximum Value =  1845
IQR =  18.0
```

```
Column Name :  swrite
Minimum Value :  7
25% =  63.0
50% or Median=  117.0
75% =  185.0
Maximum Value =  5456
IQR =  122.0
```

```
Column Name :  lwrite
Minimum Value :  0
25% =  0.0
50% or Median=  1.0
75% =  10.0
Maximum Value =  575
IQR =  10.0
```

```
Column Name :  fork
Minimum Value :  0.0
25% =  0.4
50% or Median=  0.8
75% =  2.2
Maximum Value =  20.12
IQR =  1.8000000000000003
```

```
Column Name :  scall
Minimum Value :  109
25% =  1012.0
50% or Median=  2051.5
75% =  3317.25
Maximum Value =  12493
IQR =  2305.25
```

```
Column Name :  exec
Minimum Value :  0.0
25% =  0.2
50% or Median=  1.2
75% =  2.8
Maximum Value =  59.56
IQR =  2.5999999999999996
```

```
Column Name :  sread
Minimum Value :  6
25% =  86.0
50% or Median=  166.0
75% =  279.0
Maximum Value =  5318
IQR =  193.0
```

```
Column Name :  rchar
Minimum Value :  278.0
25% =  34091.5
50% or Median=  125473.5
75% =  267828.75
Maximum Value =  2526649.0
IQR =  nan
```

```
Column Name :  wchar
Minimum Value :  1498.0
25% =  22916.0
50% or Median=  46619.0
75% =  106101.0
Maximum Value =  1801623.0
IQR =  nan
```

```
Column Name :  pgscan
Minimum Value :  0.0
25% =  0.0
50% or Median=  0.0
75% =  0.0
Maximum Value =  1237.0
IQR =  0.0
```

```
Column Name :  pgout
Minimum Value :  0.0
25% =  0.0
50% or Median=  0.0
75% =  2.4
Maximum Value =  81.44
IQR =  2.4
```

```
Column Name :  atch
Minimum Value :  0.0
25% =  0.0
50% or Median=  0.0
75% =  0.6
Maximum Value =  211.58
IQR =  0.6
```

```
Column Name :  ppgout
Minimum Value :  0.0
25% =  0.0
50% or Median=  0.0
75% =  4.2
Maximum Value =  184.2
IQR =  4.2
```

```
Column Name :  pgin
Minimum Value :  0.0
25% =  0.6
50% or Median=  2.8
75% =  9.765
Maximum Value =  141.2
IQR =  9.165000000000001
```

```
Column Name :  pgfree
Minimum Value :  0.0
25% =  0.0
50% or Median=  0.0
75% =  5.0
Maximum Value =  523.0
IQR =  5.0
```

```
Column Name :  ppgin
Minimum Value :  0.0
25% =  0.6
50% or Median=  3.8
75% =  13.8
Maximum Value =  292.61
IQR =  13.200000000000001
```

```
Column Name :  pflt
Minimum Value :  0.0
25% =  25.0
50% or Median=  63.8
75% =  159.6
Maximum Value =  899.8
IQR =  134.6
---------------------------------
Column Name :  vflt
Minimum Value :  0.2
25% =  45.4
50% or Median=  120.4
75% =  251.8
Maximum Value =  1365.0
IQR =  206.4
---------------------------------
Column Name :  freemem
Minimum Value :  55
25% =  231.0
50% or Median=  579.0
75% =  2002.25
Maximum Value =  12027
IQR =  1771.25
---------------------------------
Column Name :  freeswap
Minimum Value :  2
25% =  1042623.5
50% or Median=  1289289.5
75% =  1730379.5
Maximum Value =  2243187
IQR =  687756.0
---------------------------------
Column Name :  usr
Minimum Value :  0
25% =  81.0
50% or Median=  89.0
75% =  94.0
Maximum Value =  99
IQR =  13.0
---------------------------------
```

Table 1.1.5 – Five point Summary.

The above table shows the five point summary of the dataset, which is similar to the descriptive statistics of the data.

1.2 Impute null values if present, also check for the values which are equal to zero. Do they have any meaning or do we need to change them or drop them? Check for the possibility of creating new features if required. Also check for outliers and duplicates if there.

NO.	Column	Null Value Present
1	lread	0
2	lwrite	0
3	scall	0
4	sread	0
5	swrite	0
6	fork	0
7	exec	0
8	rchar	104
9	wchar	15
10	pgout	0
11	ppgout	0
12	pgfree	0
13	pgscan	0
14	atch	0
15	pgin	0
16	ppgin	0
17	pflt	0
18	vflt	0
19	runqsz	0
20	freemem	0
21	freeswap	0
22	usr	0

Table 1.2.1 –Null Values

From the table it is clear that null values present in columns rchar and wchar, we can replace the null values using the mean of the observations of the respective columns.

	lread	lwrite	scall	sread	swrite	fork	exec	rchar	wchar	pgout	ppgout	pgfree	pgscan	atch	pgin	ppgin	pflt	vflt	runqsz
0	1	0	2147	79	68	0.2	0.2	40671.000000	53995.0	0.0	0.0	0.0	0.0	0.0	1.6	2.6	16.00	26.40	CPU_Bound
1	0	0	170	18	21	0.2	0.2	448.000000	8385.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	15.63	16.83	Not_CPU_Bound
2	15	3	2162	159	119	2.0	2.4	197385.728363	31950.0	0.0	0.0	0.0	0.0	1.2	6.0	9.4	150.20	220.20	Not_CPU_Bound
3	0	0	160	12	16	0.2	0.2	197385.728363	8670.0	0.0	0.0	0.0	0.0	0.0	0.2	0.2	15.60	16.80	Not_CPU_Bound
4	5	1	330	39	38	0.4	0.4	197385.728363	12185.0	0.0	0.0	0.0	0.0	0.0	1.0	1.2	37.80	47.60	Not_CPU_Bound

freemem	freeswap	usr
4670	1730946	95
7278	1869002	97
702	1021237	87
7248	1863704	98
633	1760253	90

Table 1.2.2 –Dataset after imputing null values.

Next let us check the total zero values present in each column,

```
Count of zeros in column  lread  is :  675
Count of zeros in column  lwrite  is :  2684
Count of zeros in column  scall  is :  0
Count of zeros in column  sread  is :  0
Count of zeros in column  swrite  is :  0
Count of zeros in column  fork  is :  21
Count of zeros in column  exec  is :  21
Count of zeros in column  rchar  is :  0
Count of zeros in column  wchar  is :  0
Count of zeros in column  pgout  is :  4878
Count of zeros in column  ppgout  is :  4878
Count of zeros in column  pgfree  is :  4869
Count of zeros in column  pgscan  is :  6448
Count of zeros in column  atch  is :  4575
Count of zeros in column  pgin  is :  1220
Count of zeros in column  ppgin  is :  1220
Count of zeros in column  pflt  is :  3
Count of zeros in column  vflt  is :  0
Count of zeros in column  runqsz  is :  0
Count of zeros in column  freemem  is :  0
Count of zeros in column  freeswap  is :  0
Count of zeros in column  usr  is :  283
```

Table 1.2.3 –Number of zeros in columns.

Most of the columns contains zero values, the column pgscan contains most number of zeros, ie, 6448. Since we don't have clear domain expert advice on the dataset and the linear regression equation is sensitive to manipulation of the dataset, we decide not to drop the columns contains most number of zeros.

Dataset contains only one object type dataset and remaining all other columns are continuous variables, so creating new columns is not require here, since we don't have clear information on what basis we will split the continuous variable and create new variable and in the case of object variable during pd.getdummies conversion new columns will create automatically.

There are no duplicate values present in the dataset.

Next let us check the case of outliers,

	% OUTLIERS
pgscan	21.29
pgfree	18.98
ppgout	16.05
lwrite	15.93
atch	14.76
freemem	14.47
pgout	12.06
fork	11.51
ppgin	10.02
pgin	9.63
lread	9.19
exec	8.67
swrite	6.04
vflt	5.91
usr	5.25
pflt	4.82
sread	4.15
freeswap	3.59
scall	1.32
wchar	0.00
rchar	0.00

Table 1.2.4 –Outlier Proportion.

The table shows the outlier proportion of the dataset, most number of outlier present in the column, pgscan. Since outlier treatment is mentioned in the problem and linear regression equation is sensitive to outliers we decide not to treat the outliers.

1.3 Encode the data (having string values) for Modelling. Split the data into train and test (70:30). Apply Linear regression using scikit learn. Perform checks for significant variables using appropriate method from statsmodel. Create multiple models and check the performance of Predictions on Train and Test sets using Rsquare, RMSE & Adj Rsquare. Compare these models and select the best one with appropriate reasoning.

Using pd.get_dummies we can encode the given data, the only categorical variable present in the dataset is runqsz, and only two type of variable present in the column, therefore even after encoding the data total number of column remains same.

Here in the dataset the column usr is the dependent variable and all other columns are independent variables. So we can split the data to dependent and independent variable, and we can perform train-test split on the given dataset. We can split the dataset in the ratio 70:30.

	const	lread	lwrite	scall	sread	swrite	fork	exec	rchar	wchar	pgout	ppgout	pgfree	pgscan	atch	pgin	ppgin	pflt	vflt	freemem
694	1.0	1	1	1345	223	192	0.6	0.6	198703.0	293578.0	0.60	6.20	23.40	56.4	2.60	3.80	7.40	28.20	56.60	121
5535	1.0	1	1	1429	87	67	0.2	0.2	7163.0	24842.0	0.00	0.00	0.00	0.0	0.00	1.60	1.60	15.77	30.74	1476
4244	1.0	49	71	3273	225	180	0.6	0.4	83246.0	53705.0	5.39	7.19	7.19	0.0	2.79	3.99	4.59	59.88	74.05	82
2472	1.0	13	8	4349	300	191	2.8	3.0	96009.0	70467.0	0.00	0.00	0.00	0.0	0.00	2.80	3.20	129.00	236.80	772
7052	1.0	17	23	225	13	13	0.4	1.6	17132.0	12514.0	0.00	0.00	0.00	0.0	0.00	0.00	0.00	19.80	23.80	4179

freeswap	runqsz_Not_CPU_Bound
1375446	0
1021541	1
18	0
993909	0
1821682	1

Table 1.3.1 –Sample train Dataset.

	const	lread	lwrite	scall	sread	swrite	fork	exec	rchar	wchar	pgout	ppgout	pgfree	pgscan	atch	pgin	ppgin	pflt	vflt	freemem
3894	1.0	27	39	1252	53	118	0.2	0.2	26592.0	54394.0	0.0	0.0	0.0	0.0	0.0	0.4	0.6	19.44	20.04	7762
4276	1.0	1	0	996	85	55	0.4	0.4	16667.0	36431.0	0.0	0.0	0.0	0.0	0.0	1.0	1.4	35.53	52.10	2979
3414	1.0	9	7	1530	247	135	0.4	0.4	14513.0	61905.0	13.8	19.2	30.4	24.2	10.4	14.8	18.4	26.80	186.20	89
4165	1.0	32	4	3243	182	140	5.2	5.6	337517.0	94832.0	0.8	1.0	1.0	0.0	1.4	4.6	7.0	250.60	420.20	1300
7385	1.0	16	3	5017	259	249	2.8	1.4	73537.0	237547.0	0.0	0.0	0.0	0.0	0.0	5.6	5.8	142.80	276.20	2114

freeswap	runqsz_Not_CPU_Bound
1875466	1
1010114	1
11	0
1535309	0
988600	0

Table 1.3.2 –Sample test Dataset.

Next let us create the base model using all the columns in the dataset, OLS regression result of the base model is shown below,

OLS Regression Results

Dep. Variable:	usr	R-squared:	0.643
Model:	OLS	Adj. R-squared:	0.642
Method:	Least Squares	F-statistic:	489.5
Date:	Sat, 11 Feb 2023	Prob (F-statistic):	0.00
Time:	15:27:18	Log-Likelihood:	-21788.
No. Observations:	5734	AIC:	4.362e+04
Df Residuals:	5712	BIC:	4.377e+04
Df Model:	21		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	44.6423	0.746	59.830	0.000	43.180	46.105
lread	-0.0199	0.003	-6.219	0.000	-0.026	-0.014
lwrite	0.0048	0.006	0.804	0.422	-0.007	0.017
scall	0.0010	0.000	7.452	0.000	0.001	0.001
sread	-0.0004	0.002	-0.231	0.817	-0.004	0.003
swrite	-0.0021	0.002	-1.046	0.296	-0.006	0.002
fork	-1.7207	0.244	-7.046	0.000	-2.199	-1.242
exec	-0.0899	0.048	-1.885	0.060	-0.183	0.004
rchar	-4.128e-06	8.29e-07	-4.978	0.000	-5.75e-06	-2.5e-06
wchar	-1.156e-05	1.28e-06	-9.059	0.000	-1.41e-05	-9.06e-06
pgout	-0.1743	0.064	-2.724	0.006	-0.300	-0.049
ppgout	0.0990	0.037	2.701	0.007	0.027	0.171
pgfree	-0.0702	0.020	-3.502	0.000	-0.110	-0.031
pgscan	0.0086	0.006	1.359	0.174	-0.004	0.021
atch	-0.0784	0.027	-2.942	0.003	-0.131	-0.026
pgin	0.0918	0.029	3.120	0.002	0.034	0.149
ppgin	-0.0596	0.019	-3.137	0.002	-0.097	-0.022
pflt	-0.0415	0.004	-9.698	0.000	-0.050	-0.033
vflt	0.0223	0.003	6.666	0.000	0.016	0.029
freemem	-0.0016	7.53e-05	-21.476	0.000	-0.002	-0.001
freeswap	3.219e-05	4.54e-07	70.987	0.000	3.13e-05	3.31e-05
runqsz_Not_CPU_Bound	7.7882	0.303	25.682	0.000	7.194	8.383

Omnibus:	1506.830	Durbin-Watson:	2.057
Prob(Omnibus):	0.000	Jarque-Bera (JB):	4764.836
Skew:	-1.333	Prob(JB):	0.00
Kurtosis:	6.583	Cond. No.	7.48e+06

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 7.48e+06. This might indicate that there are strong multicollinearity or other numerical problems.

Table 1.3.3 –OLS Regression result of the base model.

We got the R-squared and Adj. R-squared of the base model as 0.643 and 0.642 respectively. Let us check the VIF(Variation Inflation Factor) of the features to check the multi collinearity among the features.

VIF values:

const	27.196572
lread	1.472619
lwrite	1.405963
scall	2.414389
sread	6.835434
swrite	5.318842
fork	18.210258
exec	3.059757
rchar	1.972604
wchar	1.552236
pgout	5.776002
ppgout	15.907035
pgfree	20.437679
pgscan	9.237578
atch	1.087534
pgin	8.075029
ppgin	8.670693
pflt	11.834021
vflt	20.231053
freemem	1.677431
freeswap	1.761133
runqsz_Not_CPU_Bound	1.119086

dtype: float64

Table 1.3.4 –VIF of the base model.

By using the thumb rule we can remove the feature which have higher multi collinearity, that is more than five and we can build multiple model based on that. In the first step let us remove the column pgfree which have high VIF value compared to other columns,

OLS Regression Results

Dep. Variable:	usr	R-squared:	0.642
Model:	OLS	Adj. R-squared:	0.641
Method:	Least Squares	F-statistic:	512.4
Date:	Sat, 11 Feb 2023	Prob (F-statistic):	0.00
Time:	15:27:18	Log-Likelihood:	-21794.
No. Observations:	5734	AIC:	4.363e+04
Df Residuals:	5713	BIC:	4.377e+04
Df Model:	20		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	44.7133	0.747	59.889	0.000	43.250	46.177
lread	-0.0197	0.003	-6.145	0.000	-0.026	-0.013
lwrite	0.0047	0.006	0.785	0.433	-0.007	0.017
scall	0.0010	0.000	7.512	0.000	0.001	0.001
sread	-0.0001	0.002	-0.076	0.940	-0.004	0.003
swrite	-0.0023	0.002	-1.149	0.251	-0.006	0.002
fork	-1.6538	0.244	-6.786	0.000	-2.132	-1.176
exec	-0.0853	0.048	-1.787	0.074	-0.179	0.008
rchar	-4.183e-06	8.3e-07	-5.040	0.000	-5.81e-06	-2.56e-06
wchar	-1.156e-05	1.28e-06	-9.057	0.000	-1.41e-05	-9.06e-06
pgout	-0.1746	0.064	-2.726	0.006	-0.300	-0.049
ppgout	0.0270	0.030	0.888	0.374	-0.033	0.086
pgscan	-0.0081	0.004	-1.945	0.052	-0.016	6.41e-05
atch	-0.0775	0.027	-2.907	0.004	-0.130	-0.025
pgin	0.0861	0.029	2.930	0.003	0.028	0.144
ppgin	-0.0580	0.019	-3.054	0.002	-0.095	-0.021
pflt	-0.0411	0.004	-9.599	0.000	-0.050	-0.033
vflt	0.0209	0.003	6.284	0.000	0.014	0.027
freemem	-0.0016	7.53e-05	-21.349	0.000	-0.002	-0.001
freeswap	3.214e-05	4.54e-07	70.841	0.000	3.13e-05	3.3e-05
runqsz_Not_CPU_Bound	7.7851	0.304	25.646	0.000	7.190	8.380

Omnibus:	1510.983	Durbin-Watson:	2.055
Prob(Omnibus):	0.000	Jarque-Bera (JB):	4794.588
Skew:	-1.335	Prob(JB):	0.00
Kurtosis:	6.597	Cond. No.	7.48e+06

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 7.48e+06. This might indicate that there are
strong multicollinearity or other numerical problems.

Table 1.3.5 –OLS Regression result of the model 2.

In model 2 the value of R-square and the ADJ.R-square value decreased by 0.001. Now let us check the VIF values of the model 2 features,

VIF values:

const	27.176502
lread	1.472051
lwrite	1.405924
scall	2.413532
sread	6.821993
swrite	5.314195
fork	18.098997
exec	3.057417
rchar	1.971897
wchar	1.552230
pgout	5.775992
ppgout	10.897442
pgscan	3.993835
atch	1.087438
pgin	8.050707
ppgin	8.666044
pflt	11.825640
vflt	19.937702
freemem	1.675517
freeswap	1.759249
runqsz_Not_CPU_Bound	1.119076

dtype: float64

Table 1.3.6 –VIF of the model 2.

Here vflt feature have high VIF value, so in the next step we can remove the feature vflt and we can build the next model.

OLS Regression Results

Dep. Variable:	usr	R-squared:	0.640
Model:	OLS	Adj. R-squared:	0.638
Method:	Least Squares	F-statistic:	533.7
Date:	Sat, 11 Feb 2023	Prob (F-statistic):	0.00
Time:	15:27:19	Log-Likelihood:	-21814.
No. Observations:	5734	AIC:	4.367e+04
Df Residuals:	5714	BIC:	4.380e+04
Df Model:	19		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	45.7508	0.731	62.624	0.000	44.319	47.183
lread	-0.0209	0.003	-6.507	0.000	-0.027	-0.015
lwrite	0.0078	0.006	1.292	0.196	-0.004	0.020
scall	0.0011	0.000	8.401	0.000	0.001	0.001
sread	0.0008	0.002	0.412	0.681	-0.003	0.004
swrite	-0.0033	0.002	-1.659	0.097	-0.007	0.001
fork	-0.7984	0.203	-3.937	0.000	-1.196	-0.401
exec	-0.1034	0.048	-2.164	0.030	-0.197	-0.010
rchar	-3.719e-06	8.29e-07	-4.484	0.000	-5.34e-06	-2.09e-06
wchar	-1.234e-05	1.28e-06	-9.678	0.000	-1.48e-05	-9.84e-06
pgout	-0.1952	0.064	-3.041	0.002	-0.321	-0.069
ppgout	0.0414	0.030	1.363	0.173	-0.018	0.101
pgscan	-0.0057	0.004	-1.363	0.173	-0.014	0.002
atch	-0.0542	0.026	-2.045	0.041	-0.106	-0.002
pgin	0.1312	0.029	4.587	0.000	0.075	0.187
ppgin	-0.0736	0.019	-3.892	0.000	-0.111	-0.037
pflt	-0.0276	0.004	-7.427	0.000	-0.035	-0.020
freemem	-0.0016	7.55e-05	-21.285	0.000	-0.002	-0.001
freeswap	3.151e-05	4.44e-07	70.968	0.000	3.06e-05	3.24e-05
runqsz_Not_CPU_Bound	7.7845	0.305	25.558	0.000	7.187	8.382

Omnibus:	1576.240	Durbin-Watson:	2.058
Prob(Omnibus):	0.000	Jarque-Bera (JB):	5168.158
Skew:	-1.382	Prob(JB):	0.00
Kurtosis:	6.741	Cond. No.	7.29e+06

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 7.29e+06. This might indicate that there are
strong multicollinearity or other numerical problems.

Table 1.3.7 –OLS Regression result of the model 3.

The R-square and Adj. R-square value of model 3 is 0.640 and 0.638 respectively. Now let us check the VIF values of model 3 features.

VIF values:

const	25.847445
lread	1.466877
lwrite	1.396756
scall	2.370295
sread	6.780950
swrite	5.279092
fork	12.451700
exec	3.046202
rchar	1.956283
wchar	1.537704
pgout	5.760871
ppgout	10.834977
pgscan	3.959399
atch	1.066379
pgin	7.571121
ppgin	8.519188
pflt	8.850111
freemem	1.675514
freeswap	1.673843
runqsz_Not_CPU_Bound	1.119076

dtype: float64

Table 1.3.8 –VIF of the model 3.

The feature fork has higher VIF value in model 3, so in the next step we can remove this feature.

OLS Regression Results

Dep. Variable:	usr	R-squared:	0.639
Model:	OLS	Adj. R-squared:	0.637
Method:	Least Squares	F-statistic:	561.1
Date:	Sat, 11 Feb 2023	Prob (F-statistic):	0.00
Time:	15:27:19	Log-Likelihood:	-21821.
No. Observations:	5734	AIC:	4.368e+04
Df Residuals:	5715	BIC:	4.381e+04
Df Model:	18		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	46.0253	0.728	63.208	0.000	44.598	47.453
lread	-0.0217	0.003	-6.753	0.000	-0.028	-0.015
lwrite	0.0088	0.006	1.463	0.143	-0.003	0.021
scall	0.0012	0.000	8.732	0.000	0.001	0.001
sread	0.0010	0.002	0.531	0.596	-0.003	0.005
swrite	-0.0046	0.002	-2.316	0.021	-0.008	-0.001
exec	-0.2191	0.038	-5.801	0.000	-0.293	-0.145
rchar	-3.838e-06	8.3e-07	-4.624	0.000	-5.46e-06	-2.21e-06
wchar	-1.201e-05	1.27e-06	-9.425	0.000	-1.45e-05	-9.51e-06
pgout	-0.1958	0.064	-3.047	0.002	-0.322	-0.070
ppgout	0.0428	0.030	1.409	0.159	-0.017	0.102
pgscan	-0.0060	0.004	-1.444	0.149	-0.014	0.002
atch	-0.0541	0.027	-2.041	0.041	-0.106	-0.002
pgin	0.1310	0.029	4.574	0.000	0.075	0.187
ppgin	-0.0711	0.019	-3.759	0.000	-0.108	-0.034
pflt	-0.0402	0.002	-20.963	0.000	-0.044	-0.036
freemem	-0.0016	7.56e-05	-21.259	0.000	-0.002	-0.001
freeswap	3.147e-05	4.44e-07	70.801	0.000	3.06e-05	3.23e-05
runqsz_Not_CPU_Bound	7.7570	0.305	25.443	0.000	7.159	8.355

Omnibus:	1565.042	Durbin-Watson:	2.058
Prob(Omnibus):	0.000	Jarque-Bera (JB):	5112.776
Skew:	-1.373	Prob(JB):	0.00
Kurtosis:	6.723	Cond. No.	7.25e+06

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 7.25e+06. This might indicate that there are
strong multicollinearity or other numerical problems.

Table 1.3.9 –OLS Regression result of the model 4.

The R-square and Adj. R-square value of model 3 is 0.639 and 0.637 respectively. Now let us check the VIF values of model 4 features.

VIF values:

const	25.612032
lread	1.461328
lwrite	1.394094
scall	2.355212
sread	6.774680
swrite	5.144057
exec	1.896413
rchar	1.953691
wchar	1.530906
pgout	5.760838
ppgout	10.833405
pgscan	3.957671
atch	1.066379
pgin	7.571096
ppgin	8.509881
pflt	2.344289
freemem	1.675514
freeswap	1.672780
runqsz_Not_CPU_Bound	1.118488

dtype: float64

Table 1.3.10 –VIF of the model 4.

In model 4 ppgout has higher VIF value, so in the next step we can remove this feature.

OLS Regression Results

Dep. Variable:	usr	R-squared:	0.638
Model:	OLS	Adj. R-squared:	0.637
Method:	Least Squares	F-statistic:	593.8
Date:	Sat, 11 Feb 2023	Prob (F-statistic):	0.00
Time:	15:27:20	Log-Likelihood:	-21822.
No. Observations:	5734	AIC:	4.368e+04
Df Residuals:	5716	BIC:	4.380e+04
Df Model:	17		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	45.9162	0.724	63.412	0.000	44.497	47.336
lread	-0.0211	0.003	-6.624	0.000	-0.027	-0.015
lwrite	0.0088	0.006	1.449	0.147	-0.003	0.021
scall	0.0012	0.000	8.682	0.000	0.001	0.001
sread	0.0012	0.002	0.661	0.509	-0.002	0.005
swrite	-0.0048	0.002	-2.403	0.016	-0.009	-0.001
exec	-0.2166	0.038	-5.742	0.000	-0.291	-0.143
rchar	-3.84e-06	8.3e-07	-4.626	0.000	-5.47e-06	-2.21e-06
wchar	-1.192e-05	1.27e-06	-9.365	0.000	-1.44e-05	-9.42e-06
pgout	-0.1189	0.034	-3.506	0.000	-0.185	-0.052
pgscan	-0.0019	0.003	-0.635	0.525	-0.008	0.004
atch	-0.0563	0.026	-2.127	0.033	-0.108	-0.004
pgin	0.1291	0.029	4.511	0.000	0.073	0.185
ppgin	-0.0690	0.019	-3.661	0.000	-0.106	-0.032
pflt	-0.0402	0.002	-21.000	0.000	-0.044	-0.036
freemem	-0.0016	7.56e-05	-21.255	0.000	-0.002	-0.001
freeswap	3.152e-05	4.43e-07	71.173	0.000	3.07e-05	3.24e-05
runqsz_Not_CPU_Bound	7.7570	0.305	25.441	0.000	7.159	8.355

Omnibus:	1562.992	Durbin-Watson:	2.058
Prob(Omnibus):	0.000	Jarque-Bera (JB):	5098.384
Skew:	-1.372	Prob(JB):	0.00
Kurtosis:	6.717	Cond. No.	7.21e+06

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 7.21e+06. This might indicate that there are
strong multicollinearity or other numerical problems.

Table 1.3.11 –OLS Regression result of the model 5.

The R-square and Adj. R-square value of model 3 is 0.638 and 0.637 respectively. Now let us check the VIF values of model 5 features.

VIF values:

const	25.322217
lread	1.435481
lwrite	1.393953
scall	2.351604
sread	6.719377
swrite	5.126053
exec	1.892475
rchar	1.953685
wchar	1.527007
pgout	1.602247
pgscan	1.992399
atch	1.062712
pgin	7.553432
ppgin	8.459080
pflt	2.342984
freemem	1.675511
freeswap	1.660663
runqsz_Not_CPU_Bound	1.118488

dtype: float64

Table 1.3.12 –VIF of the model 5.

In model 5 ppgin has higher VIF value, so in the next step we can remove this feature.

OLS Regression Results

Dep. Variable:	usr	R-squared:	0.638
Model:	OLS	Adj. R-squared:	0.637
Method:	Least Squares	F-statistic:	628.8
Date:	Sat, 11 Feb 2023	Prob (F-statistic):	0.00
Time:	15:27:20	Log-Likelihood:	-21829.
No. Observations:	5734	AIC:	4.369e+04
Df Residuals:	5717	BIC:	4.381e+04
Df Model:	16		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	46.0322	0.724	63.565	0.000	44.613	47.452
lread	-0.0206	0.003	-6.477	0.000	-0.027	-0.014
lwrite	0.0082	0.006	1.358	0.175	-0.004	0.020
scall	0.0012	0.000	8.736	0.000	0.001	0.001
sread	0.0012	0.002	0.669	0.504	-0.002	0.005
swrite	-0.0048	0.002	-2.411	0.016	-0.009	-0.001
exec	-0.2135	0.038	-5.654	0.000	-0.288	-0.139
rchar	-4.303e-06	8.21e-07	-5.240	0.000	-5.91e-06	-2.69e-06
wchar	-1.205e-05	1.27e-06	-9.465	0.000	-1.45e-05	-9.56e-06
pgout	-0.1207	0.034	-3.556	0.000	-0.187	-0.054
pgscan	-0.0052	0.003	-1.832	0.067	-0.011	0.000
atch	-0.0540	0.027	-2.039	0.042	-0.106	-0.002
pgin	0.0358	0.013	2.746	0.006	0.010	0.061
pflt	-0.0398	0.002	-20.798	0.000	-0.044	-0.036
freemem	-0.0016	7.57e-05	-21.239	0.000	-0.002	-0.001
freeswap	3.146e-05	4.43e-07	71.011	0.000	3.06e-05	3.23e-05
runqsz_Not_CPU_Bound	7.7210	0.305	25.308	0.000	7.123	8.319

Omnibus:	1570.138	Durbin-Watson:	2.054
Prob(Omnibus):	0.000	Jarque-Bera (JB):	5141.349
Skew:	-1.377	Prob(JB):	0.00
Kurtosis:	6.733	Cond. No.	7.21e+06

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 7.21e+06. This might indicate that there are
strong multicollinearity or other numerical problems.

Table 1.3.13 –OLS Regression result of the model 6.

The R-square and Adj. R-square value of model 3 is 0.638 and 0.637 respectively. Now let us check the VIF values of model 6 features.

VIF values:

const	25.273707
lread	1.433215
lwrite	1.393106
scall	2.350925
sread	6.719340
swrite	5.126009
exec	1.891486
rchar	1.908248
wchar	1.525727
pgout	1.601914
pgscan	1.808570
atch	1.062116
pgin	1.565863
pflt	2.334796
freemem	1.675504
freeswap	1.658519
runqsz_Not_CPU_Bound	1.117322

dtype: float64

Table 1.3.14 –VIF of the model 6.

In model 6 sread has higher VIF value, so in the next step we can remove this feature

OLS Regression Results

Dep. Variable:	usr	R-squared:	0.638
Model:	OLS	Adj. R-squared:	0.637
Method:	Least Squares	F-statistic:	670.7
Date:	Sat, 11 Feb 2023	Prob (F-statistic):	0.00
Time:	15:27:21	Log-Likelihood:	-21829.
No. Observations:	5734	AIC:	4.369e+04
Df Residuals:	5718	BIC:	4.380e+04
Df Model:	15		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	46.0506	0.724	63.639	0.000	44.632	47.469
lread	-0.0206	0.003	-6.494	0.000	-0.027	-0.014
lwrite	0.0084	0.006	1.383	0.167	-0.003	0.020
scall	0.0012	0.000	9.363	0.000	0.001	0.001
swrite	-0.0037	0.001	-3.112	0.002	-0.006	-0.001
exec	-0.2150	0.038	-5.704	0.000	-0.289	-0.141
rchar	-4.082e-06	7.52e-07	-5.431	0.000	-5.56e-06	-2.61e-06
wchar	-1.211e-05	1.27e-06	-9.534	0.000	-1.46e-05	-9.62e-06
pgout	-0.1205	0.034	-3.549	0.000	-0.187	-0.054
pgscan	-0.0051	0.003	-1.796	0.073	-0.011	0.000
atch	-0.0544	0.027	-2.053	0.040	-0.106	-0.002
pgin	0.0355	0.013	2.725	0.006	0.010	0.061
pflt	-0.0397	0.002	-20.799	0.000	-0.043	-0.036
freemem	-0.0016	7.57e-05	-21.233	0.000	-0.002	-0.001
freeswap	3.144e-05	4.42e-07	71.128	0.000	3.06e-05	3.23e-05
runqsz_Not_CPU_Bound	7.7214	0.305	25.311	0.000	7.123	8.319

Omnibus:	1571.706	Durbin-Watson:	2.054
Prob(Omnibus):	0.000	Jarque-Bera (JB):	5152.052
Skew:	-1.378	Prob(JB):	0.00
Kurtosis:	6.738	Cond. No.	7.20e+06

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 7.2e+06. This might indicate that there are
strong multicollinearity or other numerical problems.

Table 1.3.15 –OLS Regression result of the model 7.

The R-square and Adj. R-square value of model 3 is 0.638 and 0.637 respectively. Now let us check the VIF values of model 7 features.

VIF values:

const	25.237235
lread	1.432468
lwrite	1.391224
scall	2.141703
swrite	1.866842
exec	1.884698
rchar	1.599294
wchar	1.518556
pgout	1.601760
pgscan	1.802358
atch	1.061655
pgin	1.563999
pflt	2.324804
freemem	1.675171
freeswap	1.651131
runqsz_Not_CPU_Bound	1.117317

dtype: float64

Table 1.3.16 –VIF of the model 7.

In model 7 all the feature have less than 5 VIF value, so we can stop dropping tihe columns based on the VIF values, next let us check the p-value of the feature of the model 7, and we can remove the columns having p-value greater than 0.05.

In model 7 the feature lwrite has highest p-value 0.167, greater than 0.05, so we can drop this column in the next step.

OLS Regression Results

Dep. Variable:	usr	R-squared:	0.637
Model:	OLS	Adj. R-squared:	0.637
Method:	Least Squares	F-statistic:	718.4
Date:	Sat, 11 Feb 2023	Prob (F-statistic):	0.00
Time:	15:27:21	Log-Likelihood:	-21830.
No. Observations:	5734	AIC:	4.369e+04
Df Residuals:	5719	BIC:	4.379e+04
Df Model:	14		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	46.1714	0.718	64.271	0.000	44.763	47.580
lread	-0.0184	0.003	-6.714	0.000	-0.024	-0.013
scall	0.0012	0.000	9.416	0.000	0.001	0.001
swrite	-0.0037	0.001	-3.126	0.002	-0.006	-0.001
exec	-0.2161	0.038	-5.734	0.000	-0.290	-0.142
rchar	-4.044e-06	7.51e-07	-5.382	0.000	-5.52e-06	-2.57e-06
wchar	-1.208e-05	1.27e-06	-9.512	0.000	-1.46e-05	-9.59e-06
pgout	-0.1206	0.034	-3.554	0.000	-0.187	-0.054
pgscan	-0.0050	0.003	-1.772	0.076	-0.011	0.001
atch	-0.0547	0.027	-2.063	0.039	-0.107	-0.003
pgin	0.0345	0.013	2.650	0.008	0.009	0.060
pflt	-0.0398	0.002	-20.818	0.000	-0.044	-0.036
freemem	-0.0016	7.57e-05	-21.233	0.000	-0.002	-0.001
freeswap	3.14e-05	4.41e-07	71.219	0.000	3.05e-05	3.23e-05
runqsz_Not_CPU_Bound	7.7153	0.305	25.291	0.000	7.117	8.313

Omnibus:	1573.184	Durbin-Watson:	2.055
Prob(Omnibus):	0.000	Jarque-Bera (JB):	5166.229
Skew:	-1.378	Prob(JB):	0.00
Kurtosis:	6.745	Cond. No.	7.15e+06

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 7.15e+06. This might indicate that there are
strong multicollinearity or other numerical problems.

Table 1.3.17 –OLS Regression result of the model 8.

The R-square and Adj. R-square value of model 3 is 0.637 and 0.637 respectively. Now let us check the VIF values of model 8 features.

VIF values:

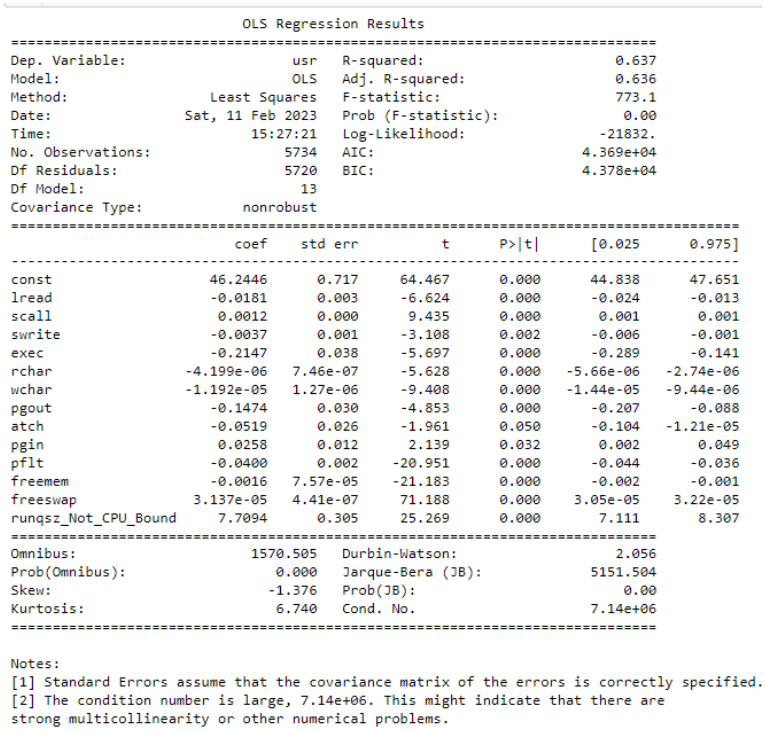
const	24.869698
lread	1.067352
scall	2.139125
swrite	1.866665
exec	1.883868
rchar	1.597080
wchar	1.518171
pgout	1.601744
pgscan	1.801825
atch	1.061597
pgin	1.558864
pflt	2.324407
freemem	1.675168
freeswap	1.641836
runqsz_Not_CPU_Bound	1.117081

dtype: float64

Table 1.3.18 –VIF of the model 8.

VIF values of the model 8 are acceptable, so again we can check the p-value of the features of the model 8.

The p-value of the feature pgscan is greater than 0.05, so we can drop this feature,



OLS Regression Results

Dep. Variable:	usr	R-squared:	0.637
Model:	OLS	Adj. R-squared:	0.636
Method:	Least Squares	F-statistic:	773.1
Date:	Sat, 11 Feb 2023	Prob (F-statistic):	0.00
Time:	15:27:21	Log-Likelihood:	-21832.
No. Observations:	5734	AIC:	4.369e+04
Df Residuals:	5720	BIC:	4.378e+04
Df Model:	13		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	46.2446	0.717	64.467	0.000	44.838	47.651
lread	-0.0181	0.003	-6.624	0.000	-0.024	-0.013
scall	0.0012	0.000	9.435	0.000	0.001	0.001
swrite	-0.0037	0.001	-3.108	0.002	-0.006	-0.001
exec	-0.2147	0.038	-5.697	0.000	-0.289	-0.141
rchar	-4.199e-06	7.46e-07	-5.628	0.000	-5.66e-06	-2.74e-06
wchar	-1.192e-05	1.27e-06	-9.408	0.000	-1.44e-05	-9.44e-06
pgout	-0.1474	0.030	-4.853	0.000	-0.207	-0.088
atch	-0.0519	0.026	-1.961	0.050	-0.104	-1.21e-05
pgin	0.0258	0.012	2.139	0.032	0.002	0.049
pflt	-0.0400	0.002	-20.951	0.000	-0.044	-0.036
freemem	-0.0016	7.57e-05	-21.183	0.000	-0.002	-0.001
freeswap	3.137e-05	4.41e-07	71.188	0.000	3.05e-05	3.22e-05
runqsz_Not_CPU_Bound	7.7094	0.305	25.269	0.000	7.111	8.307

Omnibus:	1570.505	Durbin-Watson:	2.056
Prob(Omnibus):	0.000	Jarque-Bera (JB):	5151.504
Skew:	-1.376	Prob(JB):	0.00
Kurtosis:	6.740	Cond. No.	7.14e+06

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 7.14e+06. This might indicate that there are
strong multicollinearity or other numerical problems.

Table 1.3.19 –OLS Regression result of the model 9.

The R-square and Adj. R-square value of model 3 is 0.637 and 0.636 respectively. Now let us check the VIF values of model 9 features,

VIF values:

const	24.787361
lread	1.063997
scall	2.138833
swrite	1.866489
exec	1.883050
rchar	1.575163
wchar	1.510403
pgout	1.282852
atch	1.057840
pgin	1.337064
pflt	2.316645
freemem	1.673393
freeswap	1.639830
runqsz_Not_CPU_Bound	1.116947

dtype: float64

Table 1.3.20 –VIF of the model 9.

VIF and p-value model 9 are acceptable, so we can conclude this one as the final model.

Next we can drop these same features in test data and build OLS model of the test data.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    usr   R-squared:                       0.632
Model:                            OLS   Adj. R-squared:                  0.630
Method:                 Least Squares   F-statistic:                     322.6
Date:                Sat, 11 Feb 2023   Prob (F-statistic):               0.00
Time:                        15:27:23   Log-Likelihood:                -9509.1
No. Observations:                2458   AIC:                         1.905e+04
Df Residuals:                    2444   BIC:                         1.913e+04
Df Model:                          13
Covariance Type:            nonrobust
===========================================================================================
                              coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------
const                      40.6346      1.100     36.938      0.000      38.477      42.792
lread                      -0.0156      0.005     -3.287      0.001      -0.025      -0.006
scall                       0.0014      0.000      6.326      0.000       0.001       0.002
swrite                     -0.0025      0.002     -1.045      0.296      -0.007       0.002
exec                       -0.1255      0.064     -1.973      0.049      -0.250      -0.001
rchar                  -1.802e-06   1.31e-06     -1.377      0.169   -4.37e-06    7.64e-07
wchar                  -8.92e-06    2.14e-06     -4.162      0.000   -1.31e-05   -4.72e-06
pgout                      -0.1619      0.051     -3.178      0.002      -0.262      -0.062
atch                        0.0556      0.041      1.360      0.174      -0.025       0.136
pgin                        0.0096      0.019      0.497      0.619      -0.028       0.048
pflt                       -0.0416      0.003    -12.814      0.000      -0.048      -0.035
freemem                    -0.0018      0.000    -14.742      0.000      -0.002      -0.002
freeswap                3.47e-05    6.89e-07     50.358      0.000    3.33e-05    3.6e-05
runqsz_Not_CPU_Bound       8.3687      0.493     16.984      0.000       7.402       9.335
==============================================================================
Omnibus:                      458.433   Durbin-Watson:                   2.012
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              967.154
Skew:                          -1.082   Prob(JB):                     9.67e-211
Kurtosis:                       5.181   Cond. No.                     6.62e+06
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 6.62e+06. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Table 1.3.21 –OLS Regression result of the test model.

The R-square and Adj. R-square value of test data is 0.632 and 0.630 respectively.

Similarly using scikit learn also create multiple model, delete those variables which we deleted in stats model, and create base model and best model. The observed RMSE, R-square, Adj. R-square values of train and test data is shown in the below table.

NO.	Stats Model	Base model(Train data)	Base model (Test Data)
1	R square	0.643	0.641
2	Adj. R-square	0.642	0.637
3	RMSE	10.84	11.65

Table 1.3.22 –Stats model output for base model

NO.	Stats Model	Best model(Train data)	Best model (Test Data)
1	R square	0.637	0.632
2	Adj. R-square	0.636	0.630
3	RMSE	10.87	11.703

Table 1.3.23 –Stats model output for best model

NO.	Scikit learn	Base model(Train data)	Base model (Test Data)
1	R sqare	0.642	0.631
2	Adj. R-sqaure	0.641	0.628
3	RMSE	10.813	11.597

Table 1.3.24 –Scikit Learn model output for base model

NO.	Scikit learn	Best model(Train data)	Best model (Test Data)
1	R sqare	0.637	0.624
2	Adj. R-sqaure	0.641	0.628
3	RMSE	10.81	11.62

Table 1.3.25 –Scikit Learn model output for best model

1.4 Inference: Basis on these predictions, what are the business insights and recommendations

usr = 46.24461982001648 + -0.01814701122031362 * (lread) + 0.00121007490600
47823 * (scall) + -0.0037166984239540776 * (swrite) + -0.21469537953195517 *
(exec) + -4.199441528139076e-06 * (rchar) + -1.192104720989713e-05 * (wcha
r) + -0.14743212620982396 * (pgout) + -0.05188517076557647 * (atch) + 0.025
782090458607406 * (pgin) + -0.03995910766066664 * (pflt) + -0.001603467440
2948211 * (freemem) + 3.1371163153816664e-05 * (freeswap) + 7.7093819052
35096 * (runqsz_Not_CPU_Bound)

The above shown is the final linear regression equation developed after the complet ion of all the process, The business insights and recommendation based on the equa tion are,

1. The factor lread will affect negatively user mode, lread is the Reads (transfers per second) between system memory and user memory. So lread needs to de crease to run the CPU in the user mode.
2. Increase in the number of system call will increase the time of user mode of t he CPU.
3. Number of characters transferred per second by system read calls need to dec rease for running CPU in user mode.
4. Number of disk blocks available for page swapping will increase the user mod e time of the CPU.
5. The number of kernel threads in memory that are waiting for a CPU to run ne ed to decrease which will increase the user mode time of CPU.

These are the basic insights we obtained from the above data, since we are not domain experts in this area, it is difficult to provide a clear cut information.

PROBLEM 2 – SUMMARY

You are a statistician at the Republic of Indonesia Ministry of Health and you are provided with a data of 1473 females collected from a Contraceptive Prevalence Survey. The samples are married women who were either not pregnant or do not know if they were at the time of the survey.

The problem is to predict do/don't they use a contraceptive method of choice based on their demographic and socio-economic characteristics.

INTRODUCTION

The purpose of this exercise is to build logistic regression model to predict do/don't they use a contraceptive method of choice based on their demographic and socio-economic characteristics.

DATA DESCRIPTION

1. Wife's age (numerical)
2. Wife's education (categorical) 1=uneducated, 2, 3, 4=tertiary.
3. Husband's education (categorical.) 1=uneducated, 2, 3, 4=tertiary.
4. Number of children ever born (numerical).
5. Wife's religion (binary) Non-Scientology, Scientology.
6. Wife's now working? (binary) Yes, No.
7. Husband's occupation (categorical) 1, 2, 3, 4(random).
8. Standard-of-living index (categorical) 1=verlow, 2, 3, 4=high.
9. Media exposure (binary) Good, Not good.
10. Contraceptive method used (class attribute) No,Yes.

SAMPLE OF THE DATASET

	Wife_age	Wife_education	Husband_education	No_of_children_born	Wife_religion	Wife_Working	Husband_Occupation	Standard_of_living_index	Media_exposure	Contraceptive_method_used
0	24.0	Primary	Secondary	3.0	Scientology	No	2	High	Exposed	No
1	45.0	Uneducated	Secondary	10.0	Scientology	No	3	Very High	Exposed	No
2	43.0	Primary	Secondary	7.0	Scientology	No	3	Very High	Exposed	No
3	42.0	Secondary	Primary	9.0	Scientology	No	3	High	Exposed	No
4	36.0	Secondary	Secondary	8.0	Scientology	No	3	Low	Exposed	No

Table 4 – Sample dataset

The dataset contains ten features.

EXPLORATORY DATA ANALYSIS

NO.	Column	Non – Null content	Data Type
1	Wife_age	1402	float64
2	Wife_ education	1473	object
3	Husband_education	1473	object
4	No_of_children_born	1452	float64
5	Wife_religion	1473	object
6	Wife_Working	1473	object
7	Husband_Occupation	1473	int64
8	Standard_of_living_index	1473	object
9	Media_exposure	1473	object
10	Contraceptive_method_used	1473	object

Table 5 – Exploratory data analysis

There are total 1473 entries and 10 columns present in the dataset, and there are null content present in the dataset. Husband

DESCRIPTIVE DATA ANALYSIS

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Wife_age	1402.0	NaN	NaN	NaN	32.606277	8.274927	16.0	26.0	32.0	39.0	49.0
Wife_ education	1473	4	Tertiary	577	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Husband_education	1473	4	Tertiary	899	NaN	NaN	NaN	NaN	NaN	NaN	NaN
No_of_children_born	1452.0	NaN	NaN	NaN	3.254132	2.365212	0.0	1.0	3.0	4.0	16.0
Wife_religion	1473	2	Scientology	1253	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Wife_Working	1473	2	No	1104	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Husband_Occupation	1473.0	4.0	3.0	585.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Standard_of_living_index	1473	4	Very High	684	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Media_exposure	1473	2	Exposed	1364	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Contraceptive_method_used	1473	2	Yes	844	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Table 6 – Descriptive data analysis

1. The data contains 1473 entries
2. The age of the wife varies from 16 to 49 and the mean age is 32.6.
3. Most of the people from data use contraceptive methods.
4. Most of the people exposed to media.
5. Number of children born varies from zero to 16.

Problem 2 – LOGRISTIC REGRESSION

You are a statistician at the Republic of Indonesia Ministry of Health and you are provided with a data of 1473 females collected from a Contraceptive Prevalence Survey. The samples are married women who were either not pregnant or do not know if they were at the time of the survey.

The problem is to predict do/don't they use a contraceptive method of choice based on their demographic and socio-economic characteristics.

2.1 Data Ingestion: Read the dataset. Do the descriptive statistics and do null value condition check, check for duplicates and outliers and write an inference on it. Perform Univariate and Bivariate Analysis and Multivariate Analysis.

The problem statement is to read the data and do the initial necessary steps and conduct exploratory data analysis on the dataset provided.

	Wife_age	Wife_education	Husband_education	No_of_children_born	Wife_religion	Wife_Working	Husband_Occupation	Standard_of_living_index	Media_exposure	Contraceptive_method_used
0	24.0	Primary	Secondary	3.0	Scientology	No	2	High	Exposed	No
1	45.0	Uneducated	Secondary	10.0	Scientology	No	3	Very High	Exposed	No
2	43.0	Primary	Secondary	7.0	Scientology	No	3	Very High	Exposed	No
3	42.0	Secondary	Primary	9.0	Scientology	No	3	High	Exposed	No
4	36.0	Secondary	Secondary	8.0	Scientology	No	3	Low	Exposed	No

Table 2.1.1 – Sample Dataset.

1. Wife's age (numerical)
2. Wife's education (categorical) 1=uneducated, 2, 3, 4=tertiary
3. Husband's education (categorical) 1=uneducated, 2, 3, 4=tertiary
4. Number of children ever born (numerical)
5. Wife's religion (binary) Non-Scientology, Scientology
6. Wife's now working? (binary) Yes, No
7. Husband's occupation (categorical) 1, 2, 3, 4(random)
8. Standard-of-living index (categorical) 1=verlow, 2, 3, 4=high
9. Media exposure (binary) Good, Not good
10. Contraceptive method used (class attribute) No,Yes

NO.	Column	Non – Null content	Data Type
1	Wife_age	1402	float64
2	Wife_ education	1473	object
3	Husband_education	1473	object
4	No_of_children_born	1452	float64
5	Wife_religion	1473	object
6	Wife_Working	1473	object
7	Husband_Occupation	1473	int64
8	Standard_of_living_index	1473	object
9	Media_exposure	1473	object
10	Contraceptive_method_used	1473	object

Table 2.1.2 – Data info.

There are total 1473 entries in the data, and out of that one feature is in int64 type, that one need to convert to object type, 2 features are in float 64, and 7 features are in object type data type.

Now let us check the null values in the dataset.

NO.	Column	Null content
1	Wife_age	71
2	Wife_ education	0
3	Husband_education	0
4	No_of_children_born	21
5	Wife_religion	0
6	Wife_Working	0
7	Husband_Occupation	0
8	Standard_of_living_index	0
9	Media_exposure	0
10	Contraceptive_method_used	0

Table 2.1.3 – Null values.

There are null value present in the dataset. Wife age column has 71 null values and no.of children born has 21 null values.

Next we can check the summary statistics of the given dataset.

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Wife_age	1402.0	NaN	NaN	NaN	32.606277	8.274927	16.0	26.0	32.0	39.0	49.0
Wife_ education	1473	4	Tertiary	577	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Husband_education	1473	4	Tertiary	899	NaN	NaN	NaN	NaN	NaN	NaN	NaN
No_of_children_born	1452.0	NaN	NaN	NaN	3.254132	2.365212	0.0	1.0	3.0	4.0	16.0
Wife_religion	1473	2	Scientology	1253	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Wife_Working	1473	2	No	1104	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Husband_Occupation	1473.0	4.0	3.0	585.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Standard_of_living_index	1473	4	Very High	684	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Media_exposure	1473	2	Exposed	1364	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Contraceptive_method_used	1473	2	Yes	844	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Table 2.1.4. – Summary statistics.

1. The data contains 1473 entries
2. The age of wife varies from 16 to 49 and the mean age of wife is 32.
3. Most of the people from dataset use contraceptive methods.
4. Most of the people exposed to media.
5. Number of children born varies from zero to 16.

Next let us check the duplicate values in the dataset,

	Wife_age	Wife_ education	Husband_education	No_of_children_born	Wife_religion	Wife_Working	Husband_Occupation	Standard_of_living_index
79	38.0	Tertiary	Tertiary	1.0	Scientology	Yes	1	Very High
167	26.0	Tertiary	Tertiary	1.0	Scientology	No	1	Very High
224	47.0	Tertiary	Tertiary	4.0	Scientology	No	1	Very High
270	30.0	Tertiary	Tertiary	2.0	Scientology	No	1	Very High
299	26.0	Tertiary	Tertiary	1.0	Scientology	No	1	Very High
394	29.0	Tertiary	Tertiary	0.0	Scientology	Yes	2	Very High
414	20.0	Primary	Secondary	3.0	Scientology	No	3	Very High
462	36.0	Tertiary	Tertiary	3.0	Scientology	No	1	Very High
492	37.0	Tertiary	Tertiary	3.0	Scientology	No	1	Very High
528	29.0	Tertiary	Tertiary	2.0	Scientology	Yes	1	High

Table 2.1.5. – Duplicate rows.

We found about 80 duplicate rows in the dataset, currently we decide not to drop the rows since the logistic regression equation is sensitive to data manipulation and we don't have clear information about how to deal with duplicate values.

Next we can check the outlier proportion of the given data,

We have only two numerical columns in the dataset and the outlier proportion of the data is shown below,

	% OUTLIERS
No_of_children_born	3.05
Wife_age	0.00

Table 2.1.6. – Outlier proportion.

The wife column has no outliers and the number of children born column has a outlier of 3.05%.

EXPLORATORY DATA ANALYSIS

Now let us do the exploratory data analysis of the given data.

UNIVARIATE ANALYSIS

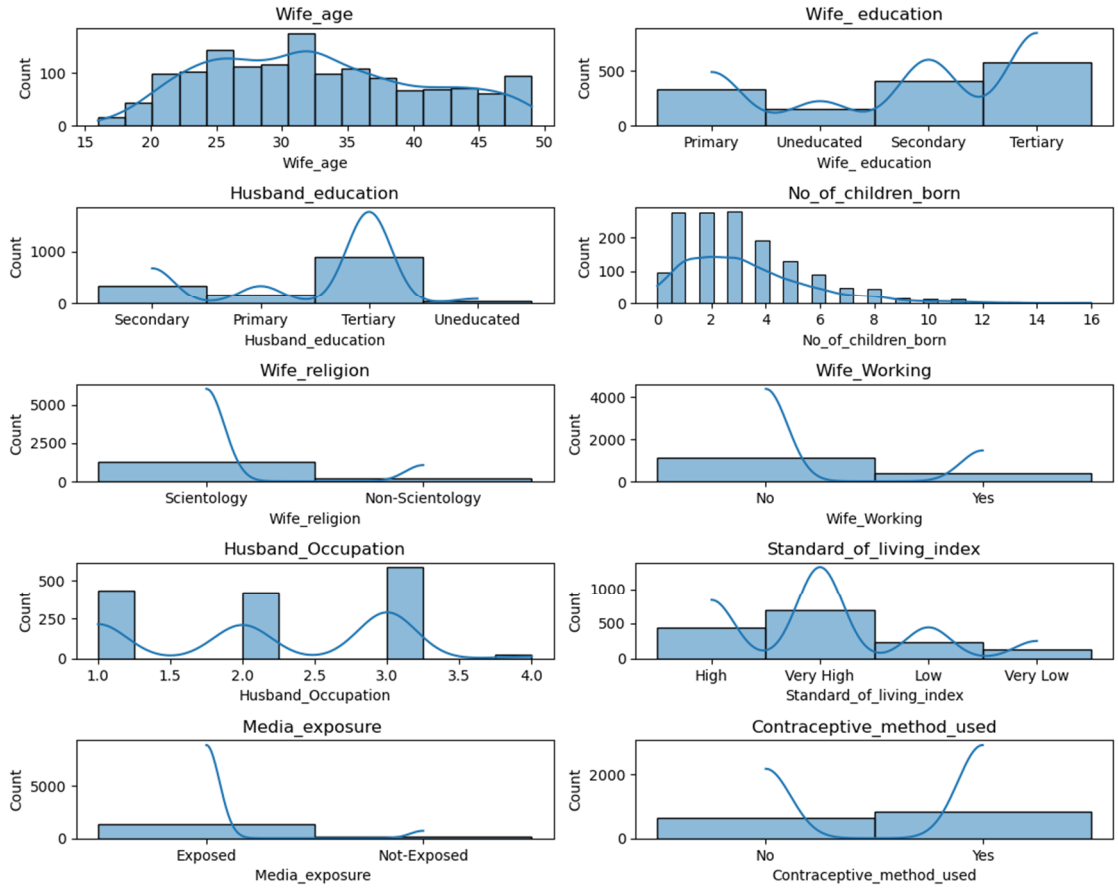


Fig 2.1.1 –Univariate analysis (HIst Plot)

The above shown is the hist plot of the different features of the dataset. The observations from the hist plot is given below.

1. Maximum age of wife lies between 30 and 35
2. Uneducated people count is very low in wife and husband.
3. Most of the people have number of children between 1 and 4
4. Most of the wife religion is scientology.
5. Most of the wives in the given data is not working.
6. Standard of living is high or very high for most of the people.
7. Most of the people is exposed to media.
8. Most of the people use contraceptive method.

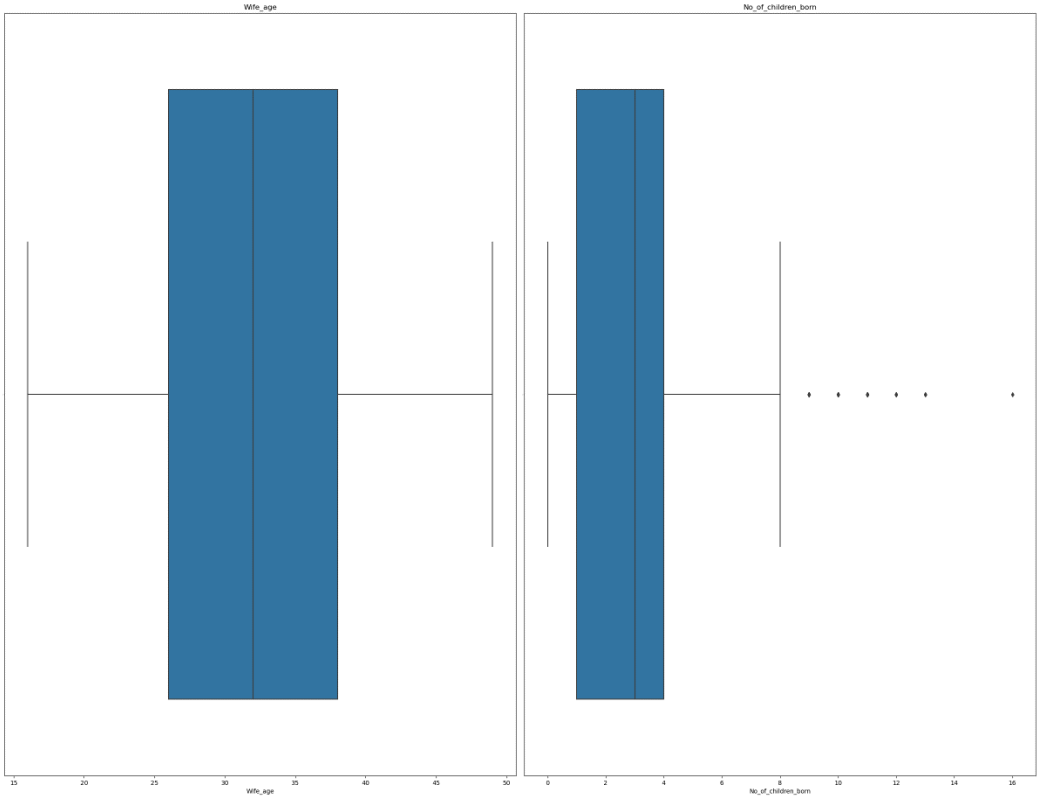


Fig 2.1.2 –Univariate analysis (Box Plot)

Now let us check the box plot of the numerical variable,

1. Wife age data is seems to be uniformly distributed.
2. There are no outliers in wife age data.
3. Number of children born is left skewed, and there are outliers in the data.

We can treat the outlier in this case, and we can cap it with upper limit. After outlier treatment box plot is shown below.

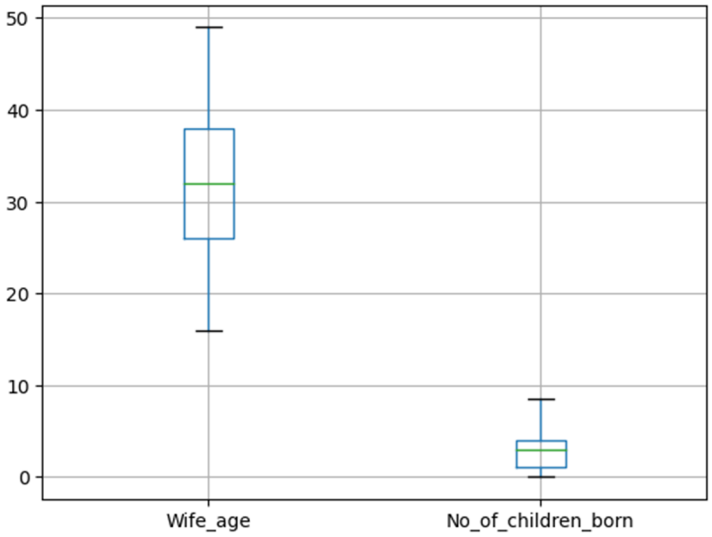


Fig 2.1.3 –Univariate analysis (Box Plot after outlier treatment)

Now there are no outlier present in the dataset.

BIVARIATE ANALYSIS

Now let us check the bivariate analysis of the categorical variables.

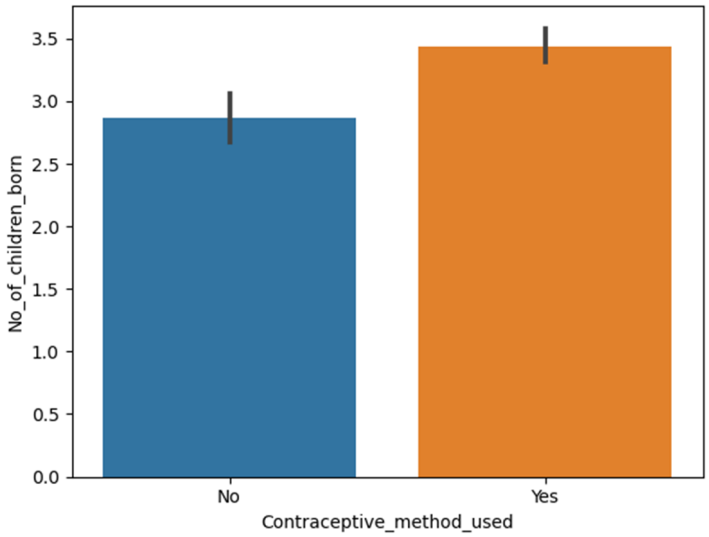


Fig 2.1.4 –Bivariate analysis (Contraceptive method used – No.of children)

The above plot shows between the contraceptive method used and number of children born, the funny part is the number of children born is more for the people who use contraceptive method.

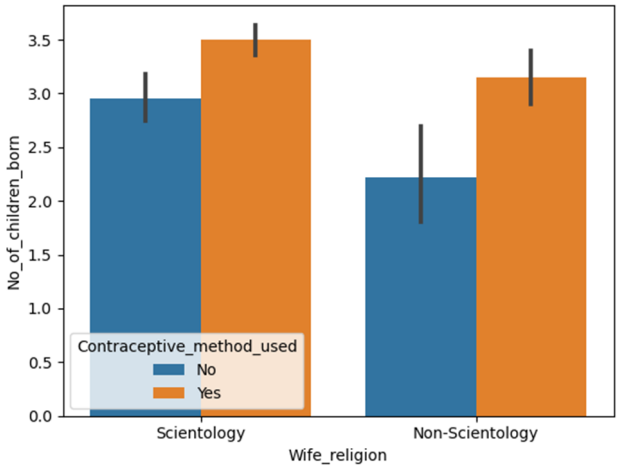


Fig 2.1.5 –Bivariate analysis (religion – No.of children)

The above plot shows the relation between the religion and the number of children born, for scientology people number of children born is more compared to other. Most of the scientology people use contraceptive method.

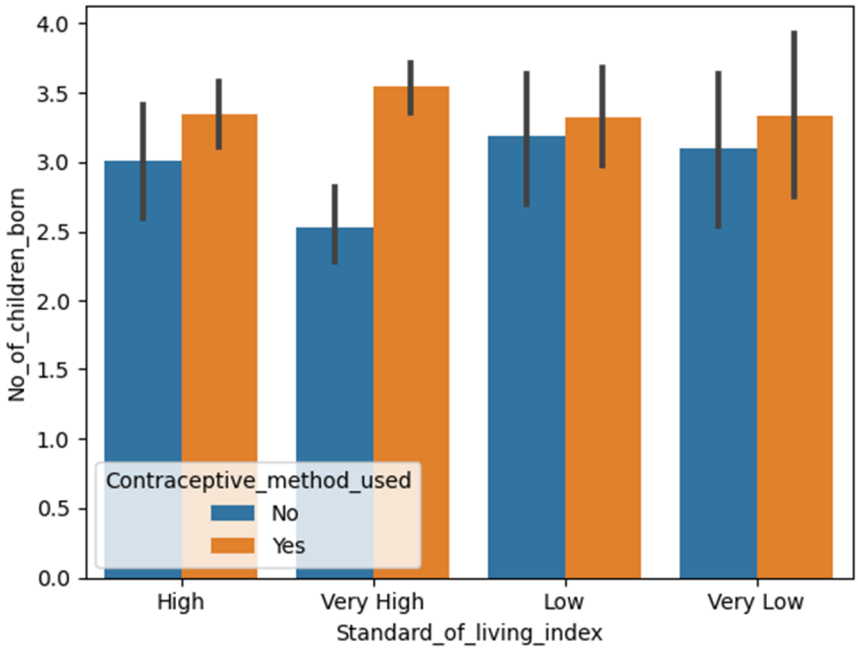


Fig 2.1.6 –Bivariate analysis (Standard of living index– No.of children)

Number of people using contraceptive method is more in all living index. Number of children born is more for the people living in high living index.

MULTIVARIATE ANALYSIS

Now let us check the pair plot and heat map of multivariate analysis.

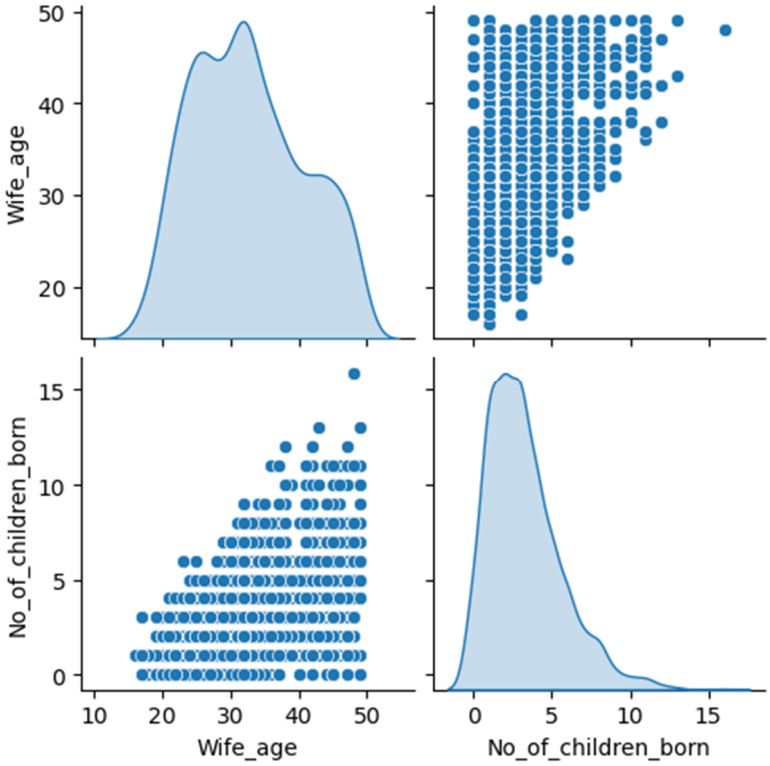


Fig 2.1.7 –Multivariate analysis (Pairplot)

The above figure shows the pairplot of the numerical features of the dataset.

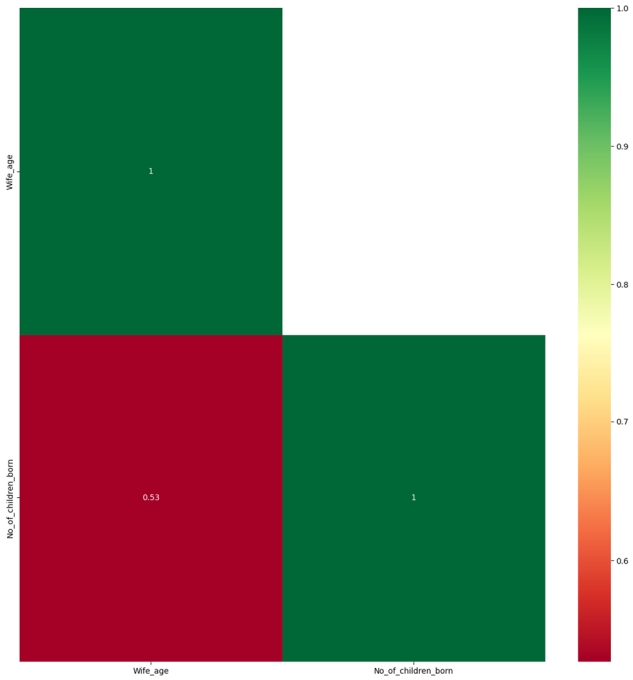


Fig 2.1.8 –Multivariate analysis (Heatmap)

2.2 Do not scale the data. Encode the data (having string values) for Modelling. Data Split: Split the data into train and test (70:30). Apply Logistic Regression and build multiple models with different predictors.

First we can transform the column contraceptive method used using label encoder, then encode the data using pd.get_dummies and split the data using train test split function.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1473 entries, 0 to 1472
Data columns (total 18 columns):
 #   Column                                Non-Null Count  Dtype
---  ------                                --------------  -----
 0   Wife_age                              1473 non-null   float64
 1   No_of_children_born                   1473 non-null   float64
 2   Contraceptive_method_used             1473 non-null   int32
 3   Wife_education_Secondary              1473 non-null   uint8
 4   Wife_education_Tertiary               1473 non-null   uint8
 5   Wife_education_Uneducated             1473 non-null   uint8
 6   Husband_education_Secondary           1473 non-null   uint8
 7   Husband_education_Tertiary            1473 non-null   uint8
 8   Husband_education_Uneducated          1473 non-null   uint8
 9   Wife_religion_Scientology             1473 non-null   uint8
 10  Wife_Working_Yes                      1473 non-null   uint8
 11  Husband_Occupation_2                  1473 non-null   uint8
 12  Husband_Occupation_3                  1473 non-null   uint8
 13  Husband_Occupation_4                  1473 non-null   uint8
 14  Standard_of_living_index_Low          1473 non-null   uint8
 15  Standard_of_living_index_Very High    1473 non-null   uint8
 16  Standard_of_living_index_Very Low     1473 non-null   uint8
 17  Media_exposure_Not-Exposed            1473 non-null   uint8
dtypes: float64(2), int32(1), uint8(15)
memory usage: 50.5 KB
```

Table 2.2.1. – Data info after encoding.

Next let us create the base model using all the features.

Logit Regression Results

Dep. Variable:	Contraceptive_method_used	No. Observations:	1473
Model:	Logit	Df Residuals:	1455
Method:	MLE	Df Model:	17
Date:	Sun, 12 Feb 2023	Pseudo R-squ.:	0.1275
Time:	11:37:18	Log-Likelihood:	-877.07
converged:	True	LL-Null:	-1005.3
Covariance Type:	nonrobust	LLR p-value:	1.039e-44

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	1.7217	0.436	3.947	0.000	0.867	2.577
Wife_age	-0.0863	0.010	-8.713	0.000	-0.106	-0.067
No_of_children_born	0.3831	0.037	10.413	0.000	0.311	0.455
Wife_education_Secondary	0.5035	0.167	3.012	0.003	0.176	0.831
Wife_education_Tertiary	1.2544	0.193	6.507	0.000	0.877	1.632
Wife_education_Uneducated	-0.3322	0.241	-1.377	0.168	-0.805	0.141
Husband_education_Secondary	0.1459	0.213	0.684	0.494	-0.272	0.564
Husband_education_Tertiary	-0.0295	0.222	-0.133	0.894	-0.464	0.405
Husband_education_Uneducated	-0.3129	0.399	-0.784	0.433	-1.095	0.469
Wife_religion_Scientology	-0.4391	0.176	-2.501	0.012	-0.783	-0.095
Wife_Working_Yes	-0.1318	0.134	-0.981	0.326	-0.395	0.131
Husband_Occupation_2	-0.1837	0.168	-1.093	0.274	-0.513	0.146
Husband_Occupation_3	0.0746	0.166	0.450	0.652	-0.250	0.399
Husband_Occupation_4	0.5920	0.455	1.301	0.193	-0.300	1.483
Standard_of_living_index_Low	-0.1735	0.183	-0.950	0.342	-0.532	0.185
Standard_of_living_index_Very_High	0.2376	0.143	1.666	0.096	-0.042	0.517
Standard_of_living_index_Very_Low	-0.5573	0.230	-2.427	0.015	-1.007	-0.107
Media_exposure_Not_Exposed	-0.5361	0.255	-2.104	0.035	-1.035	-0.037

Table 2.2.2. – Logit regression table of base model

The pseudo R-square value is 0.1275 for the base model. Let us check the VIF value of the feature of the base model.

```
Wife_age  VIF =  1.66
No_of_children_born  VIF =  1.54
Wife_education_Secondary  VIF =  1.77
Wife_education_Tertiary  VIF =  2.62
Wife_education_Uneducated  VIF =  1.58
Husband_education_Secondary  VIF =  2.52
Husband_education_Tertiary  VIF =  3.59
Husband_education_Uneducated  VIF =  1.25
Wife_religion_Scientology  VIF =  1.15
Wife_Working_Yes  VIF =  1.04
Husband_Occupation_2  VIF =  1.72
Husband_Occupation_3  VIF =  1.93
Husband_Occupation_4  VIF =  1.12
Standard_of_living_index_Low  VIF =  1.36
Standard_of_living_index_Very_High  VIF =  1.51
Standard_of_living_index_Very_Low  VIF =  1.29
Media_exposure_Not_Exposed  VIF =  1.27
```

Table 2.2.3. – VIF table of base model

Using the thumb rule let us drop the columns whose VIF value is more than 2, Let us drop the column husband education tertiary and create the second model

Logit Regression Results

Dep. Variable:	Contraceptive_method_used	No. Observations:	1473
Model:	Logit	Df Residuals:	1456
Method:	MLE	Df Model:	16
Date:	Sun, 12 Feb 2023	Pseudo R-squ.:	0.1275
Time:	11:37:19	Log-Likelihood:	-877.08
converged:	True	LL-Null:	-1005.3
Covariance Type:	nonrobust	LLR p-value:	2.567e-45

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	1.6988	0.401	4.239	0.000	0.913	2.484
Wife_age	-0.0863	0.010	-8.717	0.000	-0.106	-0.067
No_of_children_born	0.3835	0.037	10.456	0.000	0.312	0.455
Wife_education_Secondary	0.4977	0.161	3.083	0.002	0.181	0.814
Wife_education_Tertiary	1.2460	0.182	6.843	0.000	0.889	1.603
Wife_education_Uneducated	-0.3286	0.240	-1.371	0.170	-0.798	0.141
Husband_education_Secondary	0.1663	0.148	1.121	0.262	-0.124	0.457
Husband_education_Uneducated	-0.2978	0.382	-0.779	0.436	-1.047	0.452
Wife_religion_Scientology	-0.4382	0.175	-2.498	0.012	-0.782	-0.094
Wife_Working_Yes	-0.1311	0.134	-0.977	0.329	-0.394	0.132
Husband_Occupation_2	-0.1805	0.166	-1.085	0.278	-0.507	0.145
Husband_Occupation_3	0.0772	0.164	0.470	0.638	-0.245	0.399
Husband_Occupation_4	0.5946	0.454	1.309	0.191	-0.296	1.485
Standard_of_living_index_Low	-0.1711	0.182	-0.941	0.347	-0.527	0.185
Standard_of_living_index_Very_High	0.2374	0.143	1.665	0.096	-0.042	0.517
Standard_of_living_index_Very_Low	-0.5541	0.228	-2.427	0.015	-1.002	-0.107
Media_exposure_Not_Exposed	-0.5335	0.254	-2.100	0.036	-1.031	-0.036

Table 2.2.4. – Logit regression table of model 2

The pseudo R-square value is 0.1275 for the base model. Let us check the VIF value of the feature of the model 2.

```
Wife_age  VIF =  1.66
No_of_children_born  VIF =  1.53
Wife_education_Secondary  VIF =  1.66
Wife_education_Tertiary  VIF =  2.32
Wife_education_Uneducated  VIF =  1.55
Husband_education_Secondary  VIF =  1.22
Husband_education_Uneducated  VIF =  1.14
Wife_religion_Scientology  VIF =  1.14
Wife_Working_Yes  VIF =  1.04
Husband_Occupation_2  VIF =  1.69
Husband_Occupation_3  VIF =  1.9
Husband_Occupation_4  VIF =  1.11
Standard_of_living_index_Low  VIF =  1.35
Standard_of_living_index_Very_High  VIF =  1.51
Standard_of_living_index_Very_Low  VIF =  1.28
Media_exposure_Not_Exposed  VIF =  1.26
```

Table 2.2.5. – VIF table of model 2

Next let us drop the column Husband_education_Tertiary,

Logit Regression Results

Dep. Variable:	Contraceptive_method_used	No. Observations:	1473
Model:	Logit	Df Residuals:	1457
Method:	MLE	Df Model:	15
Date:	Sun, 12 Feb 2023	Pseudo R-squ.:	0.1034
Time:	11:37:19	Log-Likelihood:	-901.33
converged:	True	LL-Null:	-1005.3
Covariance Type:	nonrobust	LLR p-value:	5.365e-36

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	2.7146	0.369	7.351	0.000	1.991	3.438
Wife_age	-0.0800	0.010	-8.337	0.000	-0.099	-0.061
No_of_children_born	0.3432	0.035	9.729	0.000	0.274	0.412
Wife_education_Secondary	-0.1401	0.133	-1.057	0.291	-0.400	0.120
Wife_education_Uneducated	-0.8054	0.227	-3.550	0.000	-1.250	-0.361
Husband_education_Secondary	-0.0747	0.142	-0.525	0.599	-0.354	0.204
Husband_education_Uneducated	-0.4850	0.378	-1.283	0.200	-1.226	0.256
Wife_religion_Scientology	-0.5787	0.171	-3.391	0.001	-0.913	-0.244
Wife_Working_Yes	-0.0844	0.132	-0.641	0.521	-0.342	0.174
Husband_Occupation_2	-0.4378	0.159	-2.758	0.006	-0.749	-0.127
Husband_Occupation_3	-0.2371	0.154	-1.537	0.124	-0.539	0.065
Husband_Occupation_4	0.0591	0.444	0.133	0.894	-0.811	0.929
Standard_of_living_index_Low	-0.2396	0.179	-1.342	0.180	-0.590	0.110
Standard_of_living_index_Very_High	0.3006	0.139	2.155	0.031	0.027	0.574
Standard_of_living_index_Very_Low	-0.7154	0.224	-3.189	0.001	-1.155	-0.276
Media_exposure_Not_Exposed	-0.6519	0.253	-2.576	0.010	-1.148	-0.156

Table 2.2.6. – Logit regression table of model 3

The pseudo R-square value is 0.1034 for the base model. Let us check the VIF value of the feature of the model 3.

```
Wife_age  VIF =  1.66
No_of_children_born  VIF =  1.5
Wife_education_Secondary  VIF =  1.13
Wife_education_Uneducated  VIF =  1.41
Husband_education_Secondary  VIF =  1.16
Husband_education_Uneducated  VIF =  1.14
Wife_religion_Scientology  VIF =  1.12
Wife_Working_Yes  VIF =  1.04
Husband_Occupation_2  VIF =  1.6
Husband_Occupation_3  VIF =  1.75
Husband_Occupation_4  VIF =  1.09
Standard_of_living_index_Low  VIF =  1.35
Standard_of_living_index_Very_High  VIF =  1.5
Standard_of_living_index_Very_Low  VIF =  1.26
Media_exposure_Not_Exposed  VIF =  1.26
```

Table 2.2.7. – VIF table of model 3

All the VIF values are less than 2, next let us check the p-value, Husband_occupation_4 has high p-value, so let us drop that column in the next step,

Logit Regression Results

Dep. Variable:	Contraceptive_method_used	No. Observations:	1473
Model:	Logit	Df Residuals:	1458
Method:	MLE	Df Model:	14
Date:	Sun, 12 Feb 2023	Pseudo R-squ.:	0.1034
Time:	11:37:20	Log-Likelihood:	-901.34
converged:	True	LL-Null:	-1005.3
Covariance Type:	nonrobust	LLR p-value:	1.372e-36

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	2.7206	0.367	7.420	0.000	2.002	3.439
Wife_age	-0.0801	0.010	-8.360	0.000	-0.099	-0.061
No_of_children_born	0.3431	0.035	9.729	0.000	0.274	0.412
Wife_education_Secondary	-0.1399	0.133	-1.055	0.291	-0.400	0.120
Wife_education_Uneducated	-0.8020	0.226	-3.557	0.000	-1.244	-0.360
Husband_education_Secondary	-0.0739	0.142	-0.520	0.603	-0.353	0.205
Husband_education_Uneducated	-0.4824	0.378	-1.277	0.201	-1.223	0.258
Wife_religion_Scientology	-0.5789	0.171	-3.391	0.001	-0.913	-0.244
Wife_Working_Yes	-0.0837	0.131	-0.636	0.525	-0.341	0.174
Husband_Occupation_2	-0.4420	0.156	-2.841	0.005	-0.747	-0.137
Husband_Occupation_3	-0.2416	0.150	-1.606	0.108	-0.536	0.053
Standard_of_living_index_Low	-0.2394	0.179	-1.341	0.180	-0.589	0.111
Standard_of_living_index_Very_High	0.3007	0.139	2.156	0.031	0.027	0.574
Standard_of_living_index_Very_Low	-0.7140	0.224	-3.186	0.001	-1.153	-0.275
Media_exposure_Not_Exposed	-0.6499	0.253	-2.573	0.010	-1.145	-0.155

Table 2.2.8. – Logit regression table of model 4

The pseudo R-square value is 0.1034 for the base model. Let us check the VIF value of the feature of the model 4.

```
Wife_age  VIF =  1.65
No_of_children_born  VIF =  1.5
Wife_education_Secondary  VIF =  1.13
Wife_education_Uneducated  VIF =  1.39
Husband_education_Secondary  VIF =  1.15
Husband_education_Uneducated  VIF =  1.13
Wife_religion_Scientology  VIF =  1.12
Wife_Working_Yes  VIF =  1.03
Husband_Occupation_2  VIF =  1.53
Husband_Occupation_3  VIF =  1.66
Standard_of_living_index_Low  VIF =  1.34
Standard_of_living_index_Very_High  VIF =  1.5
Standard_of_living_index_Very_Low  VIF =  1.26
Media_exposure_Not_Exposed  VIF =  1.26
```

Table 2.2.9. – VIF table of model 4

VIF values seems to be okay, so let us check the p-value, wife education secondary has higher p-value so let us drop the column in the next step,

Logit Regression Results

Dep. Variable:	Contraceptive_method_used	No. Observations:	1473
Model:	Logit	Df Residuals:	1459
Method:	MLE	Df Model:	13
Date:	Sun, 12 Feb 2023	Pseudo R-squ.:	0.1032
Time:	11:37:21	Log-Likelihood:	-901.48
converged:	True	LL-Null:	-1005.3
Covariance Type:	nonrobust	LLR p-value:	3.807e-37

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	2.7265	0.366	7.440	0.000	2.008	3.445
Wife_age	-0.0801	0.010	-8.363	0.000	-0.099	-0.061
No_of_children_born	0.3421	0.035	9.717	0.000	0.273	0.411
Wife_education_Secondary	-0.1506	0.131	-1.149	0.250	-0.407	0.106
Wife_education_Uneducated	-0.8077	0.225	-3.585	0.000	-1.249	-0.366
Husband_education_Uneducated	-0.4565	0.374	-1.219	0.223	-1.190	0.277
Wife_religion_Scientology	-0.5859	0.170	-3.444	0.001	-0.919	-0.252
Wife_Working_Yes	-0.0861	0.131	-0.655	0.512	-0.344	0.171
Husband_Occupation_2	-0.4559	0.153	-2.975	0.003	-0.756	-0.156
Husband_Occupation_3	-0.2567	0.148	-1.739	0.082	-0.546	0.033
Standard_of_living_index_Low	-0.2459	0.178	-1.381	0.167	-0.595	0.103
Standard_of_living_index_Very_High	0.3030	0.139	2.174	0.030	0.030	0.576
Standard_of_living_index_Very_Low	-0.7214	0.224	-3.226	0.001	-1.160	-0.283
Media_exposure_Not_Exposed	-0.6428	0.252	-2.550	0.011	-1.137	-0.149

Table 2.2.10. – Logit regression table of model 5

The pseudo R-square value is 0.1032 for the base model. Let us check the VIF value of the feature of the model 5.

```
Wife_age  VIF =  1.65
No_of_children_born  VIF =  1.5
Wife_education_Secondary  VIF =  1.1
Wife_education_Uneducated  VIF =  1.39
Husband_education_Uneducated  VIF =  1.11
Wife_religion_Scientology  VIF =  1.11
Wife_Working_Yes  VIF =  1.03
Husband_Occupation_2  VIF =  1.49
Husband_Occupation_3  VIF =  1.6
Standard_of_living_index_Low  VIF =  1.34
Standard_of_living_index_Very_High  VIF =  1.5
Standard_of_living_index_Very_Low  VIF =  1.26
Media_exposure_Not_Exposed  VIF =  1.25
```

Table 2.2.11. – VIF table of model 5

VIF values seems to be okay, so let us check the p-value, wife working yes has higher p-value so let us drop the column in the next step,

Logit Regression Results

Dep. Variable:	Contraceptive_method_used	No. Observations:	1473
Model:	Logit	Df Residuals:	1460
Method:	MLE	Df Model:	12
Date:	Sun, 12 Feb 2023	Pseudo R-squ.:	0.1030
Time:	11:37:22	Log-Likelihood:	-901.69
converged:	True	LL-Null:	-1005.3
Covariance Type:	nonrobust	LLR p-value:	1.094e-37

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	2.7117	0.366	7.414	0.000	1.995	3.429
Wife_age	-0.0806	0.010	-8.432	0.000	-0.099	-0.062
No_of_children_born	0.3449	0.035	9.861	0.000	0.276	0.413
Wife_education_Secondary	-0.1479	0.131	-1.130	0.259	-0.405	0.109
Wife_education_Uneducated	-0.8048	0.225	-3.573	0.000	-1.246	-0.363
Husband_education_Uneducated	-0.4637	0.374	-1.239	0.215	-1.198	0.270
Wife_religion_Scientology	-0.5840	0.170	-3.433	0.001	-0.917	-0.251
Husband_Occupation_2	-0.4593	0.153	-2.999	0.003	-0.760	-0.159
Husband_Occupation_3	-0.2571	0.148	-1.742	0.081	-0.546	0.032
Standard_of_living_index_Low	-0.2479	0.178	-1.393	0.164	-0.597	0.101
Standard_of_living_index_Very_High	0.2989	0.139	2.147	0.032	0.026	0.572
Standard_of_living_index_Very_Low	-0.7153	0.223	-3.200	0.001	-1.153	-0.277
Media_exposure_Not_Exposed	-0.6489	0.252	-2.574	0.010	-1.143	-0.155

Table 2.2.12. – Logit regression table of model 6

The pseudo R-square value is 0.1030 for the base model. Let us check the VIF value of the feature of the model 6.

```
Wife_age  VIF =  1.64
No_of_children_born  VIF =  1.47
Wife_education_Secondary  VIF =  1.1
Wife_education_Uneducated  VIF =  1.39
Husband_education_Uneducated  VIF =  1.11
Wife_religion_Scientology  VIF =  1.11
Husband_Occupation_2  VIF =  1.49
Husband_Occupation_3  VIF =  1.6
Standard_of_living_index_Low  VIF =  1.34
Standard_of_living_index_Very_High  VIF =  1.5
Standard_of_living_index_Very_Low  VIF =  1.26
Media_exposure_Not_Exposed  VIF =  1.25
```

Table 2.2.13. – VIF table of model 6

VIF values seems to be okay, so let us check the p-value, wife education secondary has higher p-value so let us drop the column in the next step,

Logit Regression Results

Dep. Variable:	Contraceptive_method_used	No. Observations:	1473
Model:	Logit	Df Residuals:	1461
Method:	MLE	Df Model:	11
Date:	Sun, 12 Feb 2023	Pseudo R-squ.:	0.1024
Time:	11:37:23	Log-Likelihood:	-902.33
converged:	True	LL-Null:	-1005.3
Covariance Type:	nonrobust	LLR p-value:	4.511e-38

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	2.6731	0.364	7.345	0.000	1.960	3.386
Wife_age	-0.0802	0.010	-8.404	0.000	-0.099	-0.062
No_of_children_born	0.3435	0.035	9.832	0.000	0.275	0.412
Wife_education_Uneducated	-0.7527	0.220	-3.416	0.001	-1.185	-0.321
Husband_education_Uneducated	-0.4541	0.374	-1.214	0.225	-1.187	0.279
Wife_religion_Scientology	-0.5910	0.170	-3.479	0.001	-0.924	-0.258
Husband_Occupation_2	-0.4801	0.152	-3.158	0.002	-0.778	-0.182
Husband_Occupation_3	-0.2779	0.146	-1.899	0.058	-0.565	0.009
Standard_of_living_index_Low	-0.2533	0.178	-1.424	0.154	-0.602	0.095
Standard_of_living_index_Very_High	0.3115	0.139	2.246	0.025	0.040	0.583
Standard_of_living_index_Very_Low	-0.7114	0.223	-3.186	0.001	-1.149	-0.274
Media_exposure_Not_Exposed	-0.6458	0.252	-2.565	0.010	-1.139	-0.152

Table 2.2.14. – Logit regression table of model 7

The pseudo R-square value is 0.1024 for the base model. Let us check the VIF value of the feature of the model 7.

```
Wife_age  VIF =  1.64
No_of_children_born  VIF =  1.47
Wife_education_Uneducated  VIF =  1.32
Husband_education_Uneducated  VIF =  1.11
Wife_religion_Scientology  VIF =  1.11
Husband_Occupation_2  VIF =  1.47
Husband_Occupation_3  VIF =  1.58
Standard_of_living_index_Low  VIF =  1.34
Standard_of_living_index_Very_High  VIF =  1.49
Standard_of_living_index_Very_Low  VIF =  1.26
Media_exposure_Not_Exposed  VIF =  1.25
```

Table 2.2.15. – VIF table of model 7

VIF values seems to be okay, so let us check the p-value, husband education uneducated has higher p-value so let us drop the column in the next step,

Logit Regression Results

Dep. Variable:	Contraceptive_method_used	No. Observations:	1473
Model:	Logit	Df Residuals:	1462
Method:	MLE	Df Model:	10
Date:	Sun, 12 Feb 2023	Pseudo R-squ.:	0.1016
Time:	11:37:24	Log-Likelihood:	-903.08
converged:	True	LL-Null:	-1005.3
Covariance Type:	nonrobust	LLR p-value:	1.999e-38

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	2.6977	0.363	7.424	0.000	1.985	3.410
Wife_age	-0.0809	0.010	-8.483	0.000	-0.100	-0.062
No_of_children_born	0.3430	0.035	9.829	0.000	0.275	0.411
Wife_education_Uneducated	-0.8040	0.216	-3.715	0.000	-1.228	-0.380
Wife_religion_Scientology	-0.5964	0.170	-3.512	0.000	-0.929	-0.264
Husband_Occupation_2	-0.4899	0.152	-3.229	0.001	-0.787	-0.193
Husband_Occupation_3	-0.2839	0.146	-1.942	0.052	-0.570	0.003
Standard_of_living_index_Low	-0.2577	0.178	-1.451	0.147	-0.606	0.090
Standard_of_living_index_Very_High	0.3158	0.139	2.278	0.023	0.044	0.588
Standard_of_living_index_Very_Low	-0.7275	0.223	-3.265	0.001	-1.164	-0.291
Media_exposure_Not_Exposed	-0.6579	0.251	-2.619	0.009	-1.150	-0.166

Table 2.2.16. – Logit regression table of model 8

The pseudo R-square value is 0.1016 for the base model. Let us check the VIF value of the feature of the model 8.

```
Wife_age  VIF =  1.63
No_of_children_born  VIF =  1.47
Wife_education_Uneducated  VIF =  1.27
Wife_religion_Scientology  VIF =  1.11
Husband_Occupation_2  VIF =  1.46
Husband_Occupation_3  VIF =  1.57
Standard_of_living_index_Low  VIF =  1.34
Standard_of_living_index_Very_High  VIF =  1.48
Standard_of_living_index_Very_Low  VIF =  1.25
Media_exposure_Not_Exposed  VIF =  1.25
```

Table 2.2.17. – VIF table of model 8

VIF values seems to be okay, so let us check the p-value, standard of living index low, has higher p-value so let us drop the column in the next step,

Logit Regression Results

Dep. Variable:	Contraceptive_method_used	No. Observations:	1473
Model:	Logit	Df Residuals:	1463
Method:	MLE	Df Model:	9
Date:	Sun, 12 Feb 2023	Pseudo R-squ.:	0.1006
Time:	11:37:24	Log-Likelihood:	-904.14
converged:	True	LL-Null:	-1005.3
Covariance Type:	nonrobust	LLR p-value:	1.121e-38

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	2.6160	0.358	7.298	0.000	1.913	3.318
Wife_age	-0.0804	0.010	-8.437	0.000	-0.099	-0.062
No_of_children_born	0.3427	0.035	9.823	0.000	0.274	0.411
Wife_education_Uneducated	-0.8073	0.216	-3.735	0.000	-1.231	-0.384
Wife_religion_Scientology	-0.6041	0.170	-3.557	0.000	-0.937	-0.271
Husband_Occupation_2	-0.5037	0.151	-3.328	0.001	-0.800	-0.207
Husband_Occupation_3	-0.2994	0.146	-2.055	0.040	-0.585	-0.014
Standard_of_living_index_Very_High	0.3973	0.126	3.141	0.002	0.149	0.645
Standard_of_living_index_Very_Low	-0.6277	0.212	-2.964	0.003	-1.043	-0.213
Media_exposure_Not_Exposed	-0.7119	0.248	-2.866	0.004	-1.199	-0.225

Table 2.2.18. – Logit regression table of model 9

The pseudo R-square value is 0.1006 for the base model. Let us check the VIF value of the feature of the model 9.

```
Wife_age  VIF =  1.63
No_of_children_born  VIF =  1.47
Wife_education_Uneducated  VIF =  1.27
Wife_religion_Scientology  VIF =  1.11
Husband_Occupation_2  VIF =  1.46
Husband_Occupation_3  VIF =  1.57
Standard_of_living_index_Very_High  VIF =  1.24
Standard_of_living_index_Very_Low  VIF =  1.13
Media_exposure_Not_Exposed  VIF =  1.23
```

Table 2.2.19. – VIF table of model 9

All the p-values are now less than 0.05 and VIF is less than 2, so we can consider the model 9 as the final model.

2.3Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC_AUC score for each model Final Model: Compare Both the models and write inference which model is best/optimized.

Now let us check the confusion matrix of the train and test data of the base model and the best model,

Train Data(Base model)

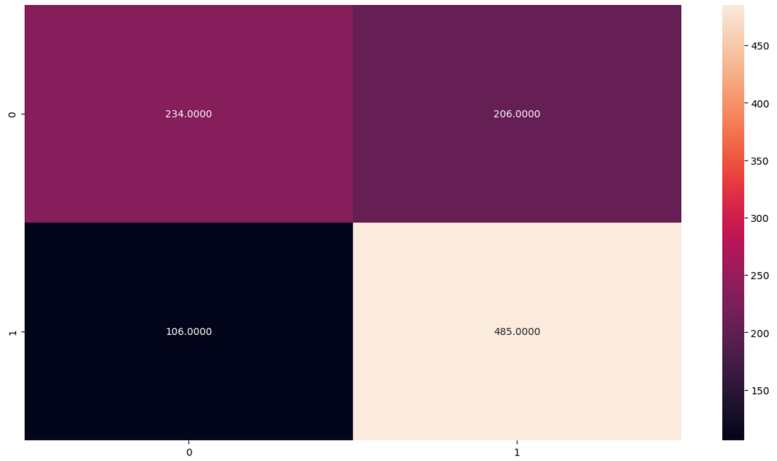


Fig 2.3.1 –Confusion Matrix – Train data – Base model

$$\text{Accuracy} = \frac{(TP+TN)}{(TP+TN+FP+FN)}$$

TP : True Positives i.e. positive classes that are correctly predicted as positive.

FP: False Positives i.e negative classes that are falsely predicted as positive.

TN: True Negatives i.e. negative classes that are correctly predicted as negative.

FN: False Negatives i.e positive classes that are falsely predicted as negative.

$$\text{Accuracy} = \frac{(485+234)}{(485+234+206+106)}$$

$$= \frac{(719)}{(1031)} = 0.697$$

Test Data(Base model)

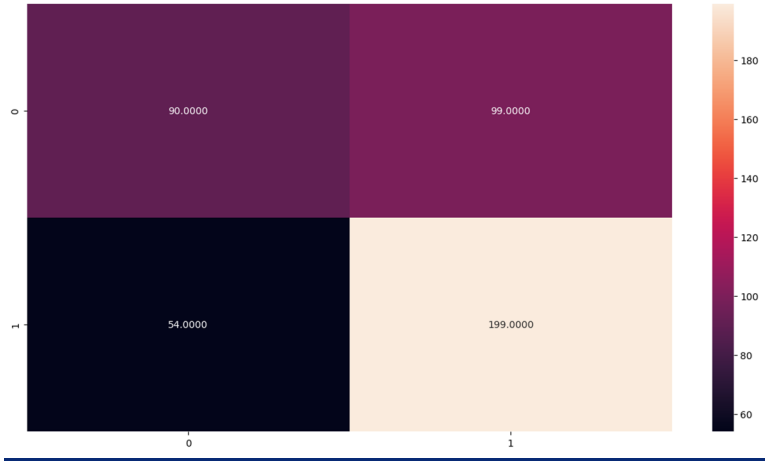


Fig 2.3.2 –Confusion Matrix – Test data – Base model

$$\text{Accuracy} = \frac{(TP+TN)}{(TP+TN+FP+FN)}$$

$$\text{Accuracy} = \frac{(199+90)}{(199+90+99+54)}$$

$$= \frac{(289)}{(442)} = 0.654$$

Train Data(Best model)

Let us first create the confusion matrix of the train data

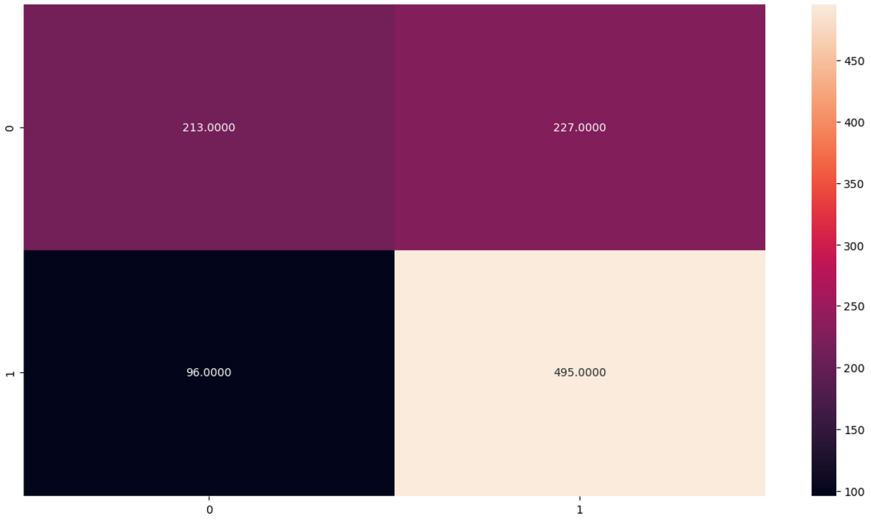


Fig 2.3.3 –Confusion Matrix – Train data – Best model

$$\text{Accuracy} = \frac{(TP+TN)}{(TP+TN+FP+FN)}$$

$$\text{Accuracy} = \frac{(495+213)}{(495+213+96+227)}$$

$$= \frac{(708)}{(1031)} = 0.6867$$

Test Data(Best model)

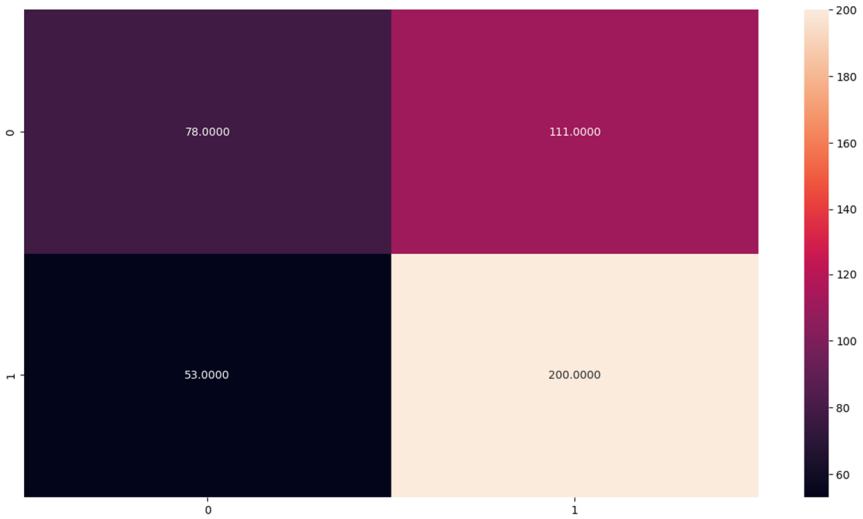


Fig 2.3.4 –Confusion Matrix – Test data – Best model

$$\text{Accuracy} = \frac{(TP+TN)}{(TP+TN+FP+FN)}$$

$$\text{Accuracy} = \frac{(200+78)}{(200+78+111+53)}$$

$$= \frac{(278)}{(442)} = 0.6289$$

In both model, the accuracy value of the train data and is more than the accuracy value of the test data. This indicates the over fitting of the model. Since only 60% accuracy showing in both model we should consider increase the amount of data sample size.

Let us consider the above confusion matrix test data,

NO.	Predicted value	
Actual	0(Predicted negative)	1(Predicted positive)
0(Actual negative)	90	99

1(Actual positive)	54	199

Table 2.3.1 –Prediction Test data base model

NO.	Predicted value	
Actual	0(Predicted negative)	1(Predicted positive)
0(Actual negative)	78	111
1(Actual positive)	53	200

Table 2.3.2 –Prediction Test data best model

In both model, the value of true positive is more compared to the value of true negative. Let us calculate the sensitivity of the model to determine how apt the model is to detecting events in the positive class.

Sensitivity of Base Model

$$\text{Sensitivity} = \frac{(TP)}{(TP+FN)}$$

$$\text{Sensitivity} = \frac{(199)}{(199+54)}$$

$$= \frac{(199)}{(253)} = 0.786$$

Sensitivity of Best Model

$$\text{Sensitivity} = \frac{(TP)}{(TP+FN)}$$

$$\text{Sensitivity} = \frac{(200)}{(200+53)}$$

$$= \frac{(200)}{(253)} = 0.791$$

The value of sensitivity is high, so we can conclude that model has good accuracy predicting true positive values.

ROC – AUC CURVE OF BASE MODEL

Now let us check the ROC – AUC curve of testing and training data of Base model

Train Data

We use the probability value of positive outcomes for the construction of ROC curve as well as the AUC score, We use ROC – AUC curve to compare the model. The probability value of train dataset is shown below,

By using the positive outcomes from the probability of the predicted training dataset we will calculate the area under the curve (AUC) value of the train dataset, The AUC value we got for training dataset is,

AUC for Training data = 0.731

Next let us construct the ROC (Receiver Operating Characteristics) curve for the training dataset of the base model, we will use roc_curve parameter from sklearn.metrics for the construction of the ROC curve, we will use dependant train data and probability of the predicted training data as the parameters for the ROC curve. The constructed ROC curve is shown below,

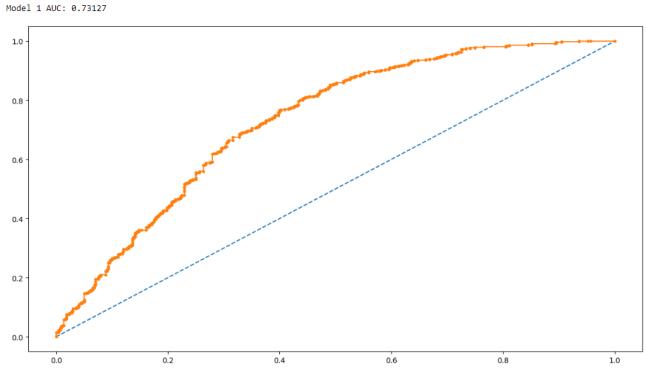


Fig 2.3.5 –ROC Curve – Train data – Base model

Test Data

For calculating the AUC score and constructing the ROC curve for the test data of the base model, we will use the same procedure we did for the train data,

By using the positive predicted values from the predicated probability test data set we will calculate the value of AUC score,

AUC for Testing data = 0.686

Next using dependant test data and probability of the predicted test data, we will create the ROC curve for the test data and it is shown below,

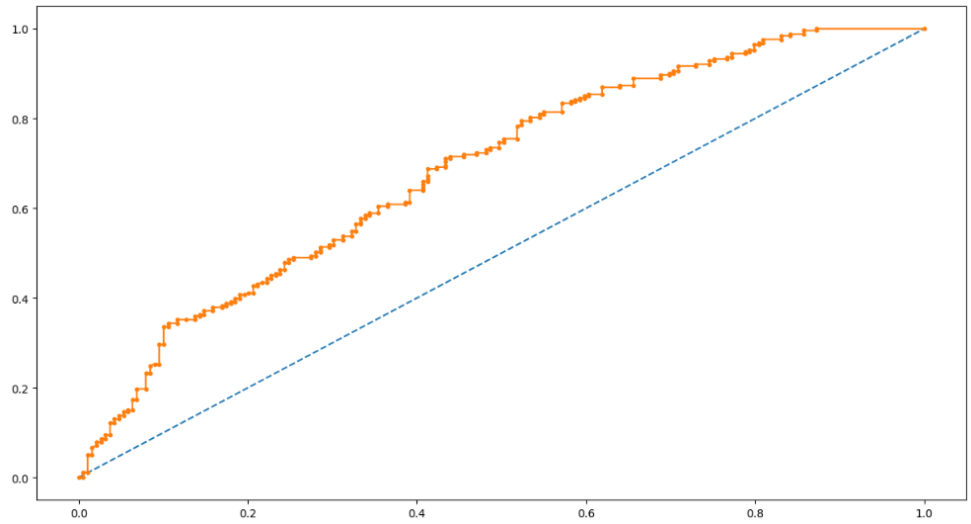


Fig 2.3.6.–ROC Curve – Test data – Base model

We got the AUC score of the test and train data in the range of 0.6 to 0.7 which is acceptable, as expected the AUC score of the testing data is less than the AUC score of the training data. The same is visible in the case of ROC curve too.

ROC – AUC CURVE OF BEST MODEL

Next let us calculate the AUC score and construct the ROC curve of best model for test as well as train data,

Train Data

As mentioned in the case of base model, similar steps will follow in the case of best model too,

By using the positive predicted values from the predicated probability test data set we will calculate the value of AUC score,

AUC for Training data = 0.7138

Next using dependant train data and probability of the predicted train data, we will create the ROC curve for the test data and it is shown below,

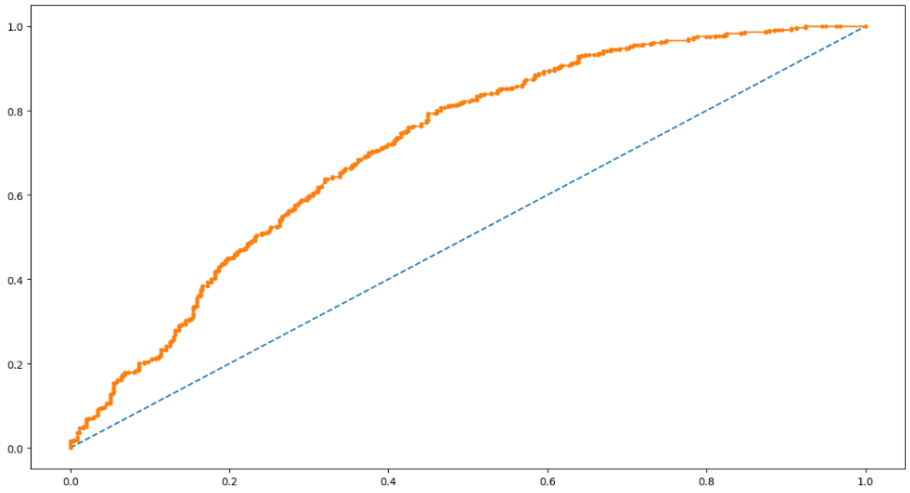


Fig 2.3.7 –ROC Curve – Train data – Best model

Test Data

By using the positive predicted values from the predicated probability test data set we will calculate the value of AUC score,

AUC for Test data = 0.644

Next using dependant test data and probability of the predicted test data, we will create the ROC curve for the test data and it is shown below,

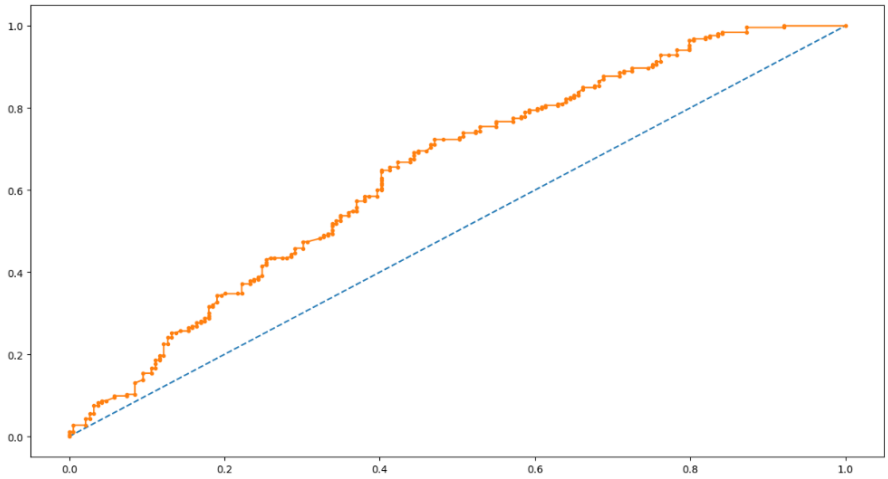


Fig 2.3.8 –ROC Curve – Test data – Best model

As similar to base model the AUC score of the best model also is high for the train data than the test data. We got the AUC score in the range 0.6 to 0.7 considered good. In the case of ROC curve also same scenario is visible, area under the curve is more for the train data than the test data.

CLASSIFICATION REPORT FOR BASE MODEL

A classification report is a performance evaluation metric in machine learning. It is used to show the precision, recall, F1 Score, and support of your trained classification model. Let understand each parameters,

Precision - Precision is defined as the ratio of true positives to the sum of true and false positives.

$$\text{Precision} = \frac{(TP)}{(TP+FP)}$$

Recall - Recall is defined as the ratio of true positives to the sum of true positives and false negatives.

$$\text{Recall/Sensitivity} = \frac{(TP)}{(TP+FN)}$$

TP : True Positives i.e. positive classes that are correctly predicted as positive.

FP: False Positives i.e negative classes that are falsely predicted as positive.

TN: True Negatives i.e. negative classes that are correctly predicted as negative.

FN: False Negatives i.e positive classes that are falsely predicted as negative.

F1-score - The F1 is the weighted harmonic mean of precision and recall. The closer the value of the F1 score is to 1.0, the better the expected performance of the model is.

Support - Support is the number of actual occurrences of the class in the dataset. It doesn't vary between models, it just diagnoses the performance evaluation process.

Let us check the classification report of train and test data of CART model,

Train data

The classification report for train data of the base model is shown below,

Model 1

	precision	recall	f1-score	support
0	0.69	0.53	0.60	440
1	0.70	0.82	0.76	591
accuracy			0.70	1031
macro avg	0.70	0.68	0.68	1031
weighted avg	0.70	0.70	0.69	1031

Table 2.3.3 –Classification Report – Train Data – Base Model

Test data

The classification report for test data of the base model is shown below,

Model 1

	precision	recall	f1-score	support
0	0.62	0.48	0.54	189
1	0.67	0.79	0.72	253
accuracy			0.65	442
macro avg	0.65	0.63	0.63	442
weighted avg	0.65	0.65	0.64	442

Table 2.3.4 –Classification Report – Test Data – Base Model

CLASSIFICATION REPORT FOR BEST MODEL

Similar to base model the classification report of train and test data best model is shown below,

Train data

The classification report for train data of the best model is shown below,

Model 2

	precision	recall	f1-score	support
0	0.69	0.48	0.57	440
1	0.69	0.84	0.75	591
accuracy			0.69	1031
macro avg	0.69	0.66	0.66	1031
weighted avg	0.69	0.69	0.67	1031

Table 2.3.5 –Classification Report –Train Data – Best Model

Test data

The classification report for test data of the best model is shown below,

Model 2

	precision	recall	f1-score	support
0	0.60	0.41	0.49	189
1	0.64	0.79	0.71	253
accuracy			0.63	442
macro avg	0.62	0.60	0.60	442
weighted avg	0.62	0.63	0.61	442

Table 2.3.6 –Classification Report –Test Data – Best Model

Now let us compare all the performance metrics values of train data and test data of base model and best model, consider below table,

	BASE TRAIN	BASE TEST	BEST TRAIN	BEST TEST
ACCURACY	0.70	0.65	0.69	0.63
AUC	0.731	0.686	0.713	0.644
PRECISION	0.70	0.67	0.69	0.64
RECALL	0.82	0.79	0.84	0.79
F1-SCORE	0.76	0.72	0.75	0.72

Table 2.3.7 –Performance metrics comparison

Let us compare the ROC curve of test and train data of CART and RF as well,

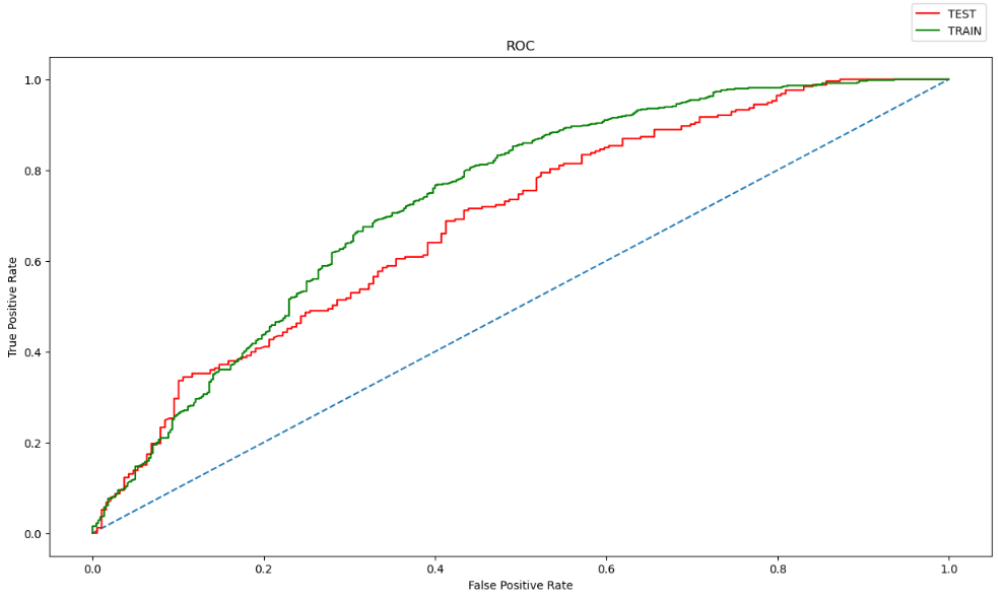


Fig 2.3.9 –Test-Train – ROC Curve – Base model

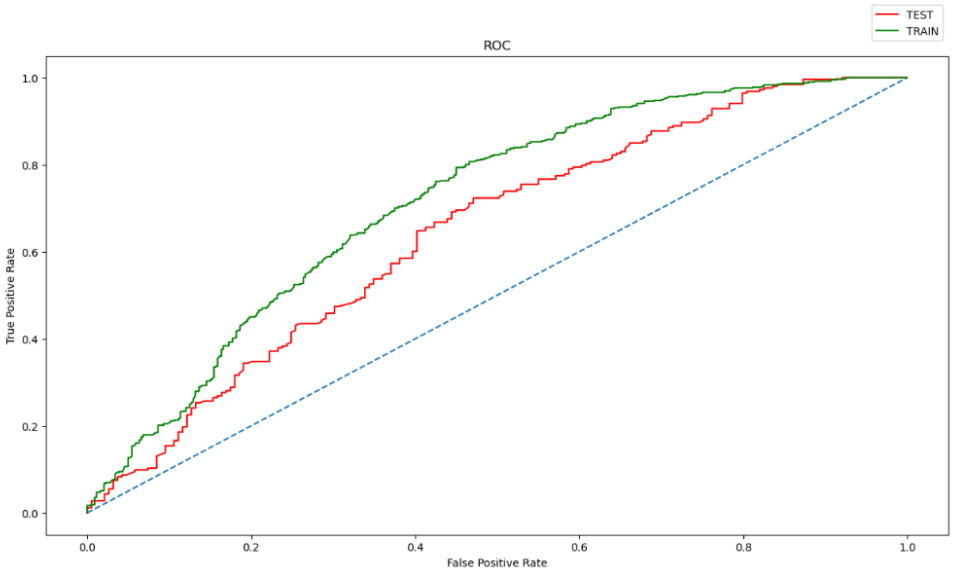


Fig 2.3.10 –Test-Train – ROC Curve – Best model

2.4 Inference: Basis on these predictions, what are the insights and recommendations.

#P(Y=1 or Contraceptive method used yes)= 1/(1+e** -(2.6160 - 0.0804* Wife_age+ 0.3427* No_of_children_born+ -0.8073 * Wife_education_Uneducated+ -0.6041* Wife_religion_Scientology+ -0.5037* Husband_Occupation_2 + -0.2994* Husband_Occupation_3 + 0.3973* Standard_of_living_index_Very_High+ -0.6277* Standard_of_living_index_Very_Low+ -0.7119* Media_exposure_Not_Exposed))

The above shown equation is the final output of the logistic regression. Since as per the performance matrix evaluation the probability to predict true positive is high for this model, so the model accuracy is good. Unlike linear regression the coefficients of logistic regression are difficult to interpret. Other insights and recommendations are,

1. Wife age will negatively affect the use of contraceptive method.
2. The number of children born increase the probability of use contraceptive method.
3. People who have a very high living index increase the probability of use of contraceptive method and vice versa.
4. Exposure to media is also play a key role, exposure to media increase the use of contraceptive method.
5. So in conclusion the factors mainly affecting the use of contraceptive methods are media exposure, age of the wife and education.

THE END

CLICK HERE TO GO TO CONTENTS