BUSINESS REPORT

Predictive Modelling

Contents

List of Tables

List of Figures

Problem Statement 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.

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.

Data Description

Column Name	Description	Data Type
lread	Reads (transfers per second) between system memory and user memory	Integer
lwrite	writes (transfers per second) between system memory and user memory	Integer
scall	Number of system calls of all types per second	Integer
sread	Number of system read calls per second	Integer
swrite	Number of system write calls per second	Integer
fork	Number of system fork calls per second.	Integer
exec	Number of system exec calls per second.	Integer
rchar	Number of characters transferred per second by system read calls	Integer
wchar	Number of characters transferred per second by system write calls	Integer
pgout	Number of page-out requests per second	Integer
ppgout	Number of pages, paged out per second	Integer
pgfree	Number of pages per second placed on the free list.	Integer
pgscan	Number of pages checked if they can be freed per second	Integer
atch	Number of page attaches (satisfying a page fault by reclaiming a page in memory) per second	Integer
pgin	Number of page-in requests per second	Integer
ppgin	Number of pages paged in per second	Integer
pflt	Number of page faults caused by protection errors (copy	Integer
vflt	Number of page faults caused by address translation.	Integer
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	Integer
freemem	Number of memory pages available to user processes	Integer
freeswap	Number of disk blocks available for page swapping.	Integer

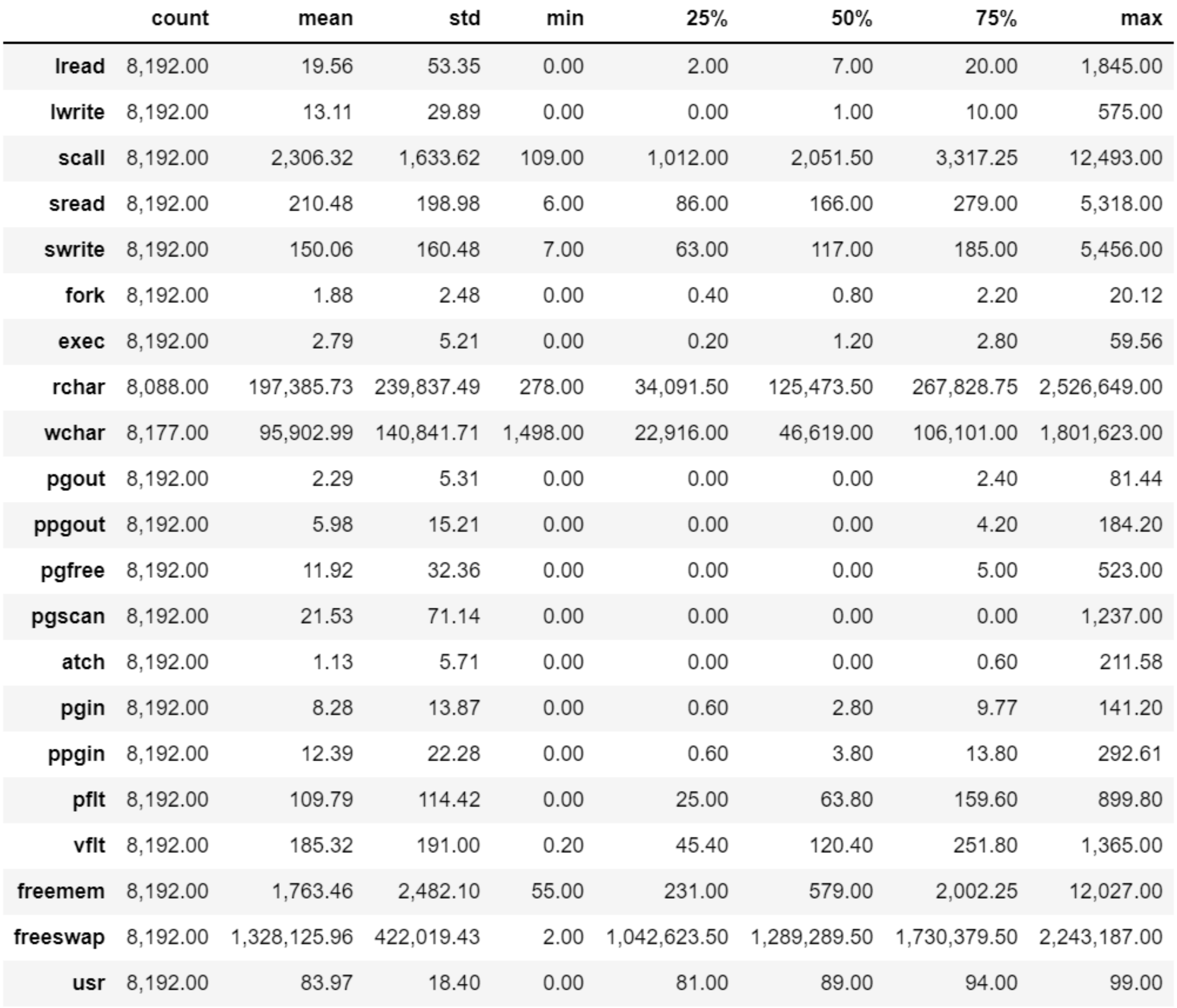
Table 1: Data Description - Dataset 1

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.

EDA

The data is imported, and the following are the observations:

- The data has 8192 rows and 22 columns. There is 1 object type data types and rest are float 7 int data types.



	count	mean	std	min	25%	50%	75%	max
lread	8,192.00	19.56	53.35	0.00	2.00	7.00	20.00	1,845.00
lwrite	8,192.00	13.11	29.89	0.00	0.00	1.00	10.00	575.00
scall	8,192.00	2,306.32	1,633.62	109.00	1,012.00	2,051.50	3,317.25	12,493.00
sread	8,192.00	210.48	198.98	6.00	86.00	166.00	279.00	5,318.00
swrite	8,192.00	150.06	160.48	7.00	63.00	117.00	185.00	5,456.00
fork	8,192.00	1.88	2.48	0.00	0.40	0.80	2.20	20.12
exec	8,192.00	2.79	5.21	0.00	0.20	1.20	2.80	59.56
rchar	8,088.00	197,385.73	239,837.49	278.00	34,091.50	125,473.50	267,828.75	2,526,649.00
wchar	8,177.00	95,902.99	140,841.71	1,498.00	22,916.00	46,619.00	106,101.00	1,801,623.00
pgout	8,192.00	2.29	5.31	0.00	0.00	0.00	2.40	81.44
ppgout	8,192.00	5.98	15.21	0.00	0.00	0.00	4.20	184.20
pgfree	8,192.00	11.92	32.36	0.00	0.00	0.00	5.00	523.00
pgscan	8,192.00	21.53	71.14	0.00	0.00	0.00	0.00	1,237.00
atch	8,192.00	1.13	5.71	0.00	0.00	0.00	0.60	211.58
pgin	8,192.00	8.28	13.87	0.00	0.60	2.80	9.77	141.20
ppgin	8,192.00	12.39	22.28	0.00	0.60	3.80	13.80	292.61
pflt	8,192.00	109.79	114.42	0.00	25.00	63.80	159.60	899.80
vflt	8,192.00	185.32	191.00	0.20	45.40	120.40	251.80	1,365.00
freemem	8,192.00	1,763.46	2,482.10	55.00	231.00	579.00	2,002.25	12,027.00
freeswap	8,192.00	1,328,125.96	422,019.43	2.00	1,042,623.50	1,289,289.50	1,730,379.50	2,243,187.00
usr	8,192.00	83.97	18.40	0.00	81.00	89.00	94.00	99.00

Table 2: Data Summary

- Majority of the times the process run queue size was Not CPU Bound.
- On an average 83.9% of times the cpus run in user mode.
- There are no duplicate rows present in the data.
- There are few missing values in variables ‘rchar’ & ‘wchar’. These will be treated later.
- A new feature can be calculated which is the ‘System Read-Write rate’ by the features Number of system read and write calls per second. So these features can be replaced by the newly created one.

Univariate Analysis

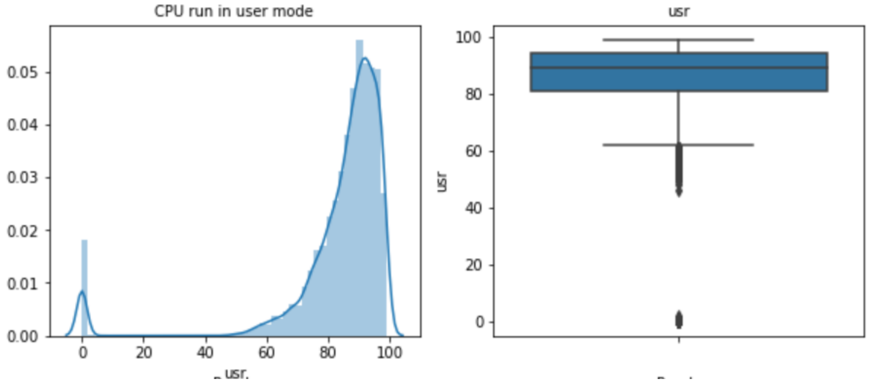


Figure 1: Univariate Analysis (Usr)

- The CPU runs in user mode 80 to 99% of the times or it stays idle.

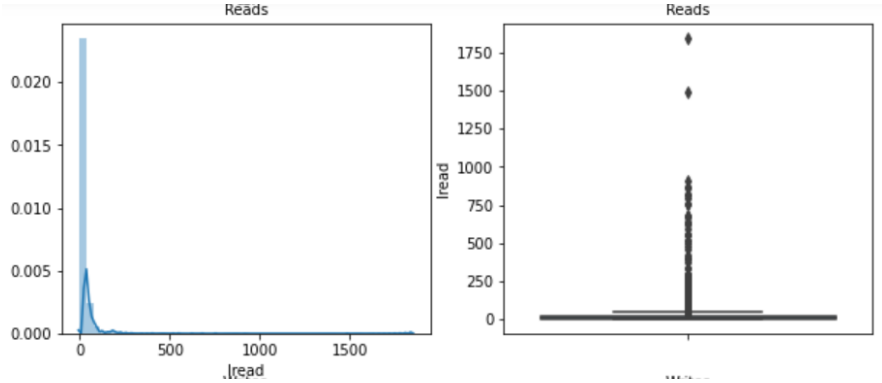


Figure 2: Univariate Analysis (lread)

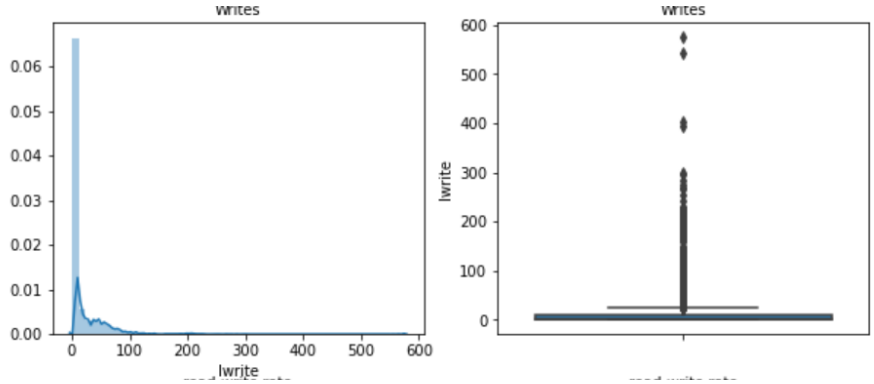


Figure 3: Univariate Analysis (lwrite)

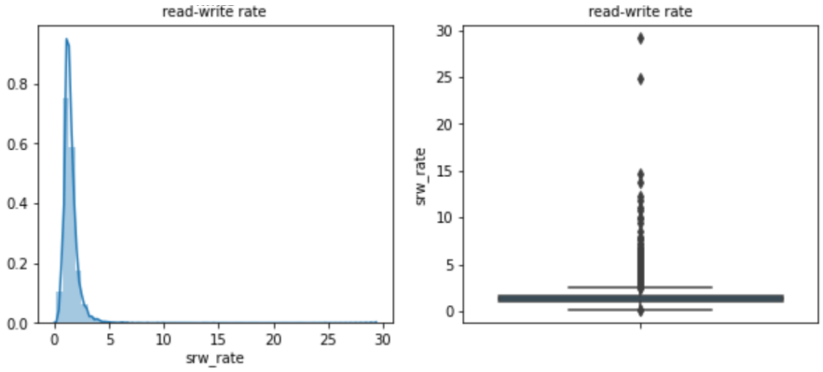


Figure 4: Univariate Analysis (srw rate)

- The transfers per seconds for read and write is pretty fast as majority of the transfers are happening quickly.
- The System read-write rate is also quick and majority of the transactions happen to be under 5%.
- It seems that there are not many activities that are happening.

Bivariate Analysis

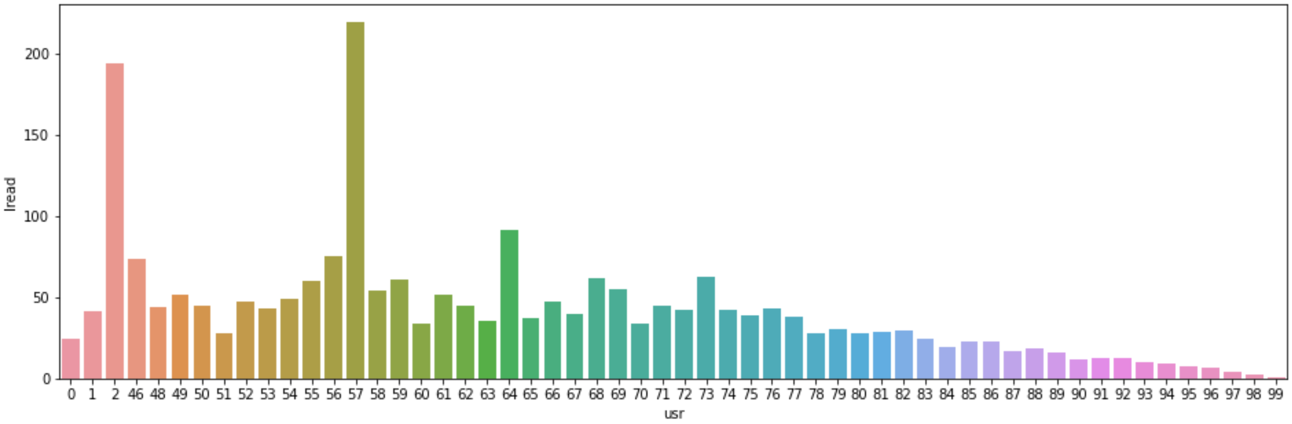


Figure 5: Bivariate analysis (usr vs lread)

- Two unusual spikes can be seen when comparing the number of reads per second with the CPUs running in user mode.

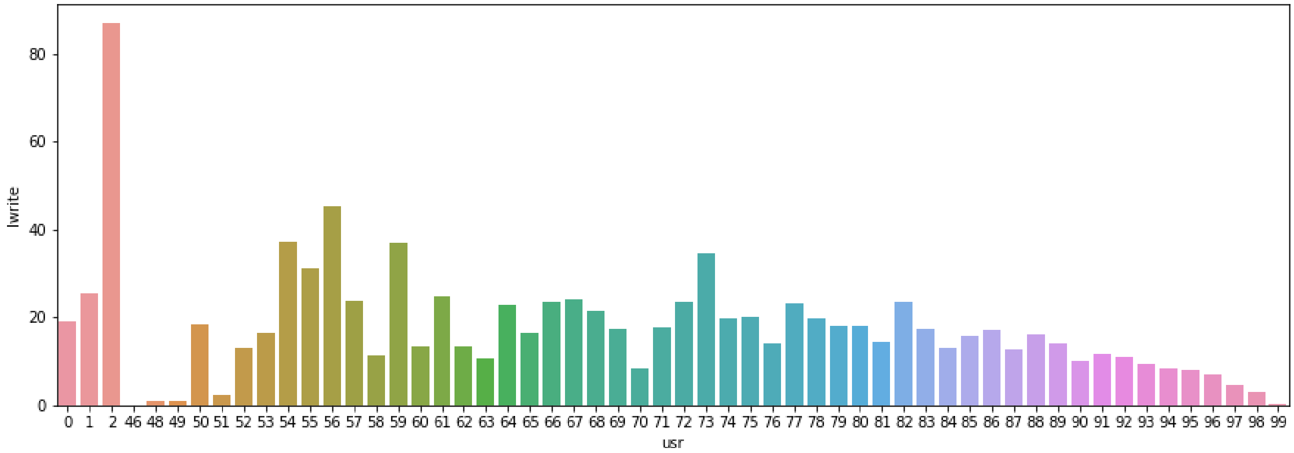


Figure 6: Bivariate analysis (usr vs lwrite)

- Similarly for the write, when the number or writes is high, only 2% of the CPU runs in user mode.
- This indicates that when the read/write is high, the majority of the CPU does not run in user mode.

Multivariate analysis

- From the pairplot shown below, we can see correlation between a few variables:

  - Linear correlation can be seen between 'vflt', 'pflt' & 'fork'. If the fork calls increase, number of page faults also tend to increase.

- Similarly, Number of page out requests per second is also highly correlated to the number of pages, paged out per second.

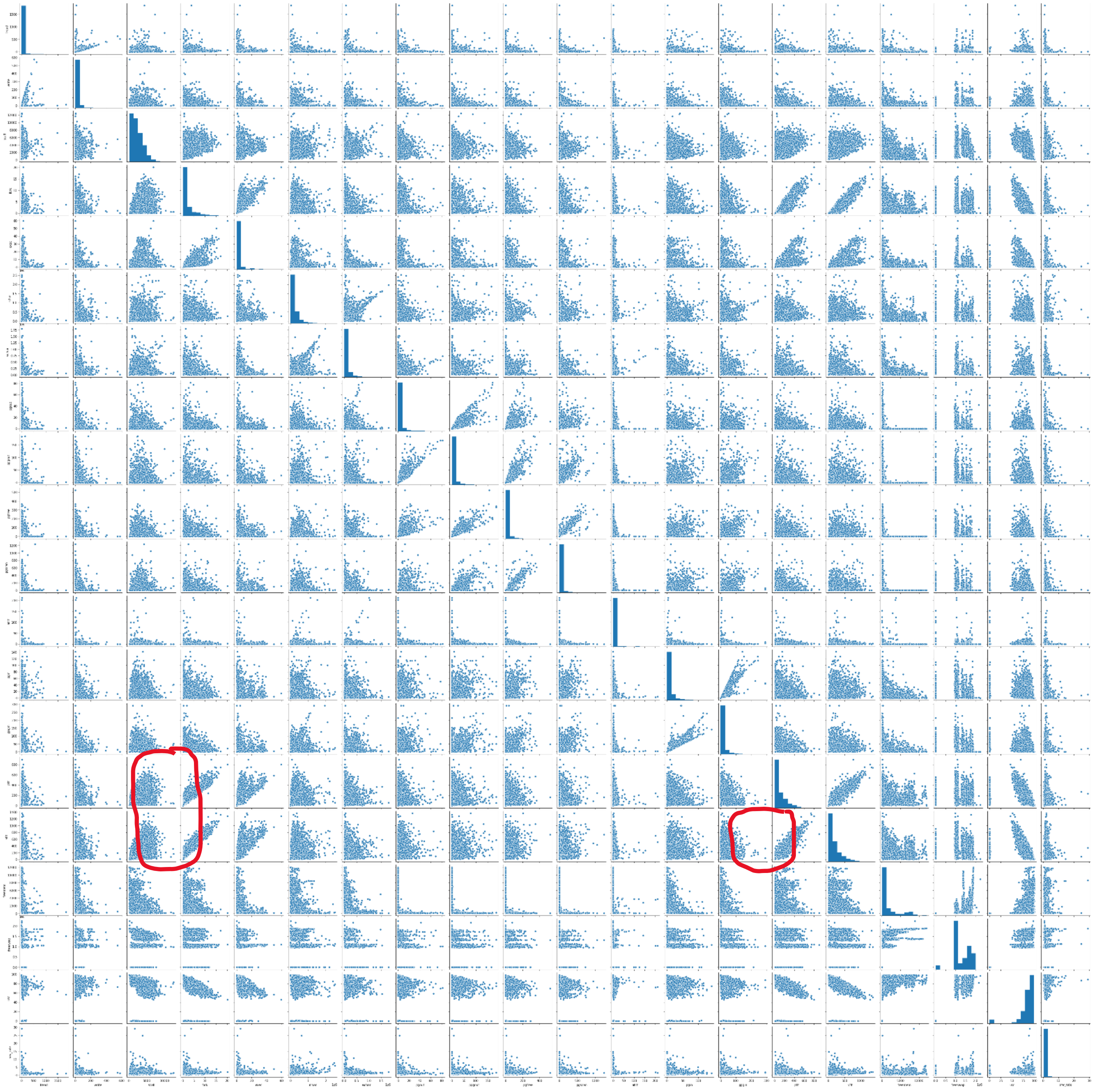


Figure 7: Pairplot

- Similar correlations can be observed from the heatmap
- Majority of the times ‘pgscan’ i.e the number of pages checked if they can be freed per second is 0.

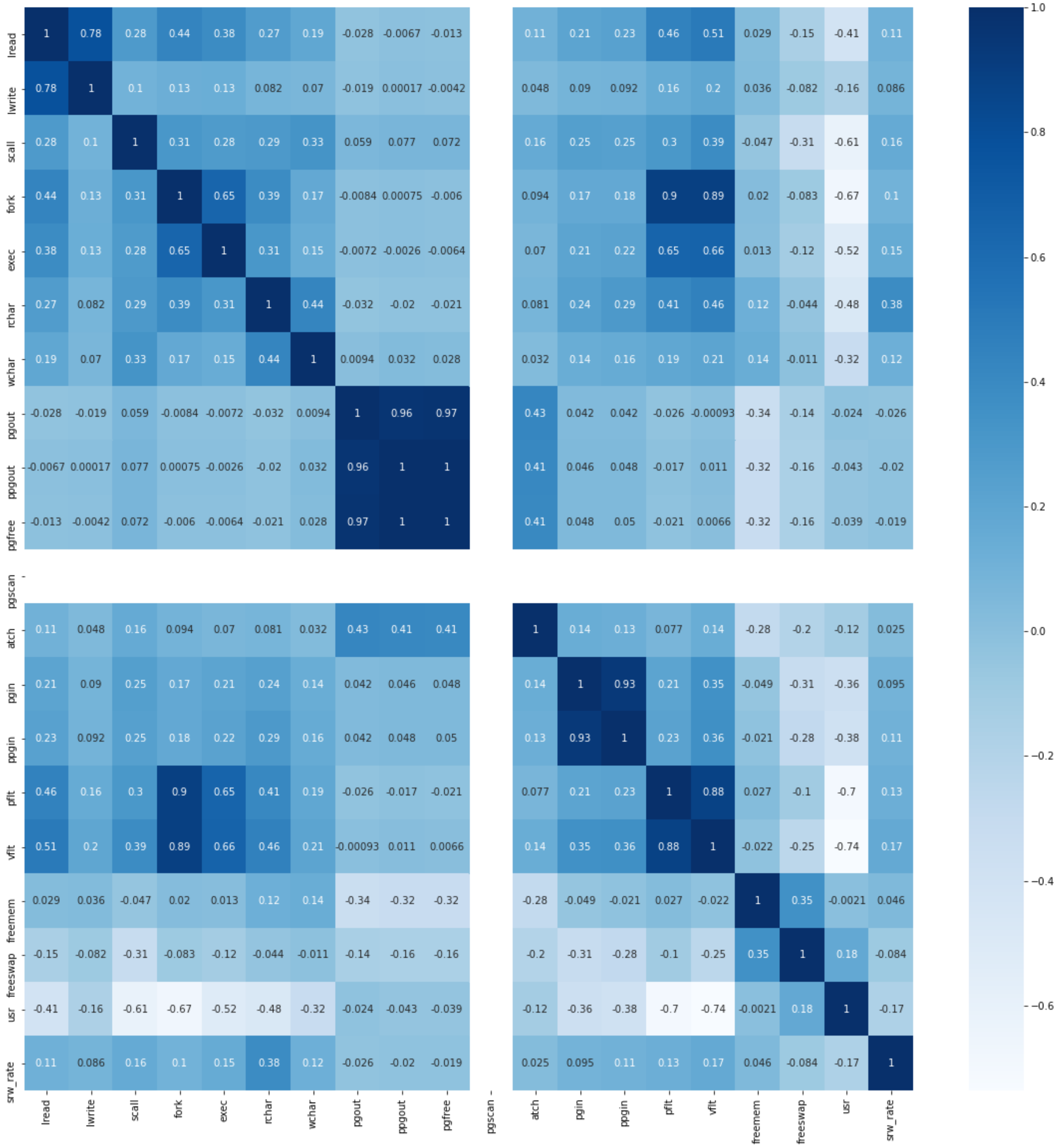


Figure 8: Correlation Heatmap

- As observed before, number of page out requests per second is also highly correlated to the number of pages, paged out per second variable.
- Similarly, both the page fault variables – pflt & vflt are highly correlated with the fork variable.
- We can try to drop these variables from the model and check the performance.
- We will also drop pgscan variable as it is 0, before building the models.

Outlier Detection & Treatment

- The red dots boxplots show that there is presence of outliers in all the variables.
- Majority of the variables are highly skewed as well.
- All the outliers are **treated** by adjusting them to the lower and upper bound values calculated by the IQR value.

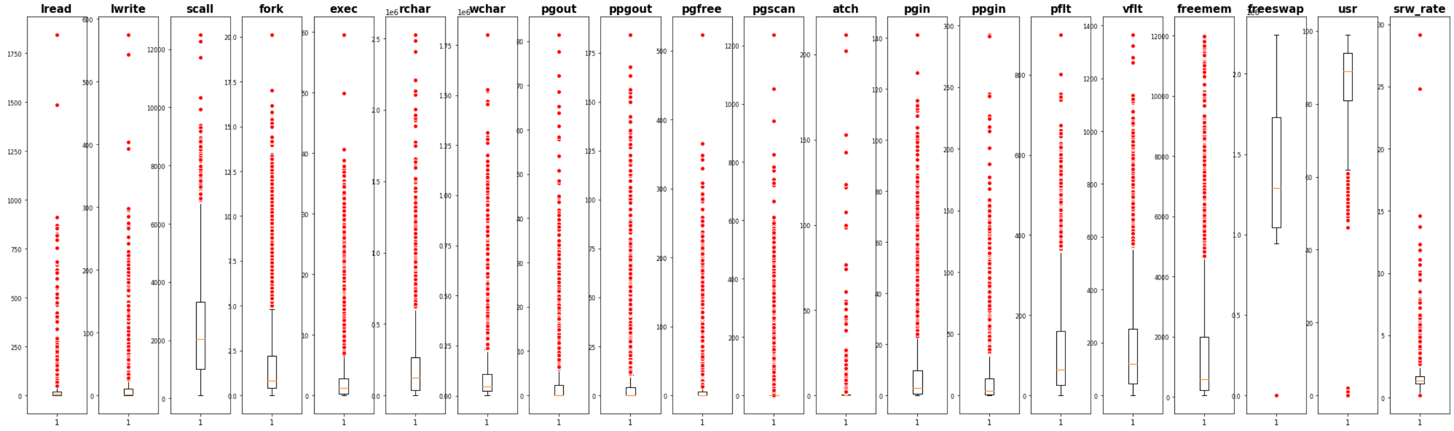


Figure 9: Boxplot for outlier detection

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.

The data is imported, and the following are the observations:

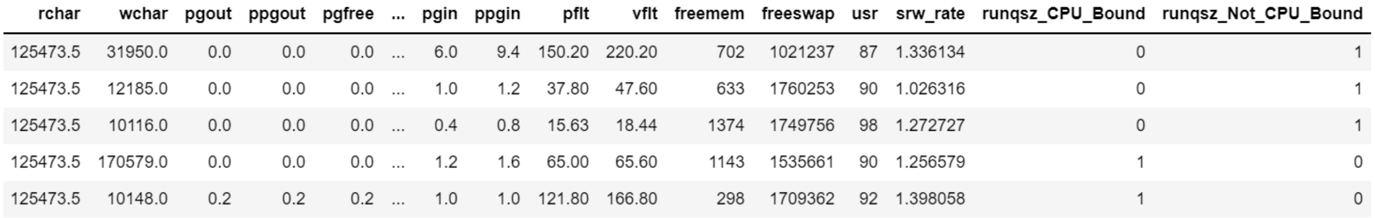
- As observed before, there were null values present in the variables 'rchar' & 'wchar'.
- Since there are only a few values missing, these null values are replaced with the median value.
- It has also been observed that there are 0s present for many dimensions. These are all valid values as a it is related to the activities in the computer system.
- No ordinal variables are available in the data hence an option to combine the sub-ordinal variables is not there.
- Instead, as described above; a new feature is generated i.e. 'srw_rate' (System Read-Write rate) which will be useful in model building and reducing multi-collinearity in the data.
- New features - number of page rate & page requests rate have also been created with the variables pgin, pgout, ppgin & ppgout.

- However these new features are not giving any significant output as majority of the values are 0 or inf.

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.

One Hot encoding is done on the only 'Object' types variable i.e 'runqsz'.

A new column is created, with 1 indicating that variable as True and 0 as False and this is how the extended variable's data looks.



rchar	wchar	pgout	ppgout	pgfree	...	pgin	ppgin	pflt	vflt	freemem	freeswap	usr	srw_rate	runqsz_CPU_Bound	runqsz_Not_CPU_Bound
125473.5	31950.0	0.0	0.0	0.0	...	6.0	9.4	150.20	220.20	702	1021237	87	1.336134	0	1
125473.5	12185.0	0.0	0.0	0.0	...	1.0	1.2	37.80	47.60	633	1760253	90	1.026316	0	1
125473.5	10116.0	0.0	0.0	0.0	...	0.4	0.8	15.63	18.44	1374	1749756	98	1.272727	0	1
125473.5	170579.0	0.0	0.0	0.0	...	1.2	1.6	65.00	65.60	1143	1535661	90	1.256579	1	0
125473.5	10148.0	0.2	0.2	0.2	...	1.0	1.0	121.80	166.80	298	1709362	92	1.398058	1	0

Table 3: Encoded data

Train – Test split & Model Building

The data set is split into training and testing data in the ratio of 70:30.

The Linear Regression model is built and fitted into the Training dataset.

The coefficients of all the variables are calculated, and it clearly shows that features like 'runqsz_CPU_Bound','pgout' will directly impact the value of the target variable if all the other variables are 0.

Similarly, is the case for the variables with negative coefficients.

```
The coefficient for lread is 0.03817618209248563
The coefficient for lwrite is -0.057922437287160726
The coefficient for scall is -0.0015063430043030404
The coefficient for fork is -0.4614891114963575
The coefficient for exec is -0.005118212759306604
The coefficient for rchar is -5.026192047489768e-06
The coefficient for wchar is -2.904685746643218e-06
The coefficient for pgout is 0.32735781004518383
The coefficient for ppgout is -0.24684549271472117
The coefficient for pgfree is -0.02942222087076294
The coefficient for pgscan is 3.608224830031759e-16
The coefficient for atch is 0.3356025437366802
The coefficient for pgin is 0.03680316328903658
The coefficient for ppgin is -0.12302670402364305
The coefficient for pflt is -0.02241295261454668
The coefficient for vflt is -0.010443658418113646
The coefficient for freemem is 0.00013337885817554004
The coefficient for freeswap is -1.3445222133921218e-06
The coefficient for srw_rate is 0.22499316904760963
The coefficient for runqsz_CPU_Bound is 0.37779909857991506
The coefficient for runqsz_Not_CPU_Bound is -0.3777990985799155
```

Model Performance

Model 1 – Sklearn method

To check the model's performance, we calculate the Rsquare values or the Coefficient of Determinants for both Train and test data.

Rsquare for Train data: 0.722

RMSE for Train data: 2.47

This is a good value. This shows that almost 72% of the variance of the training dataset was captured by the model.

Now evaluating the Rsquare and RMSE for test data.

Rsquare for Test data: 0.708

RMSE for Test data: 2.54

This is also a good value. This shows that almost 70% of the variance of the testing dataset was captured by the model.

The model seems to be neither overfitting nor under-fitting, therefore this is a good model to go with.

However, let's see if there is any improvement with the statsmodel approach.

Model 1- Statsmodel method

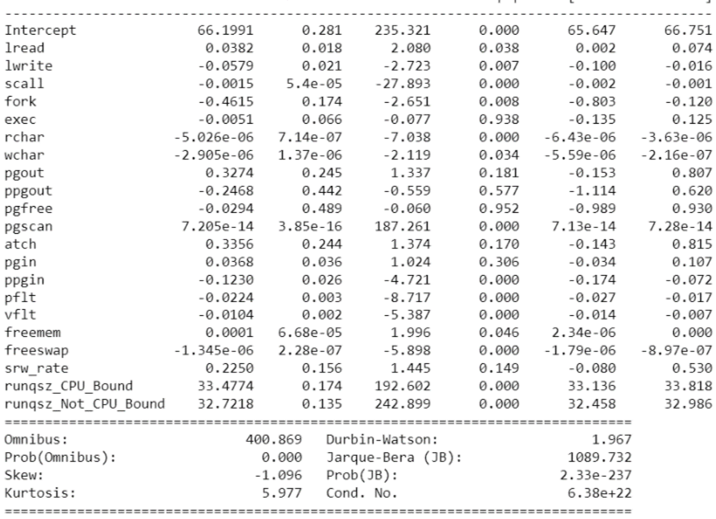
If we build the model using stats model and OLS method, we see that the Adjusted Rsquare is equal to the Rsqaure value which is 0.72.

This shows that there is no statistical fluke or sampling error present.

Looking at the p-values of the predictors, we see that variables like 'exec', 'pgout', 'ppgout' etc. have a p-value greater than 0.05. This shows that there is no relation between this variable and the target variable, hence these are not useful in prediction.

Rsquare for Train data: 0.72

OLS Output:



```
------------------------------------------------------------------------------------------
Intercept              66.1991      0.281    235.321      0.000      65.647      66.751
lread                   0.0382      0.018      2.080      0.038       0.002       0.074
lwrite                 -0.0579      0.021     -2.723      0.007      -0.100      -0.016
scall                  -0.0015    5.4e-05    -27.893      0.000      -0.002      -0.001
fork                   -0.4615      0.174     -2.651      0.008      -0.803      -0.120
exec                   -0.0051      0.066     -0.077      0.938      -0.135       0.125
rchar               -5.026e-06   7.14e-07     -7.038      0.000   -6.43e-06   -3.63e-06
wchar               -2.905e-06   1.37e-06     -2.119      0.034   -5.59e-06   -2.16e-07
pgout                   0.3274      0.245      1.337      0.181      -0.153       0.807
ppgout                 -0.2468      0.442     -0.559      0.577      -1.114       0.620
pgfree                 -0.0294      0.489     -0.060      0.952      -0.989       0.930
pgscan               7.205e-14   3.85e-16    187.261      0.000    7.13e-14    7.28e-14
atch                    0.3356      0.244      1.374      0.170      -0.143       0.815
pgin                    0.0368      0.036      1.024      0.306      -0.034       0.107
ppgin                  -0.1230      0.026     -4.721      0.000      -0.174      -0.072
pflt                   -0.0224      0.003     -8.717      0.000      -0.027      -0.017
vflt                   -0.0104      0.002     -5.387      0.000      -0.014      -0.007
freemem                 0.0001   6.68e-05      1.996      0.046    2.34e-06       0.000
freeswap            -1.345e-06   2.28e-07     -5.898      0.000   -1.79e-06   -8.97e-07
srw_rate                0.2250      0.156      1.445      0.149      -0.080       0.530
runqsz_CPU_Bound       33.4774      0.174    192.602      0.000      33.136      33.818
runqsz_Not_CPU_Bound   32.7218      0.135    242.899      0.000      32.458      32.986
==============================================================================
Omnibus:                      400.869   Durbin-Watson:                   1.967
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1089.732
Skew:                          -1.096   Prob(JB):                    2.33e-237
Kurtosis:                       5.977   Cond. No.                     6.38e+22
==============================================================================
```

Figure 10: Statsmodel Model 2

Model 2

Building another model using Statsmodel without the Page variables:

"pgin","ppgin","pgout","ppgout","pgscan"

Also using the newly created srw_rate feature.

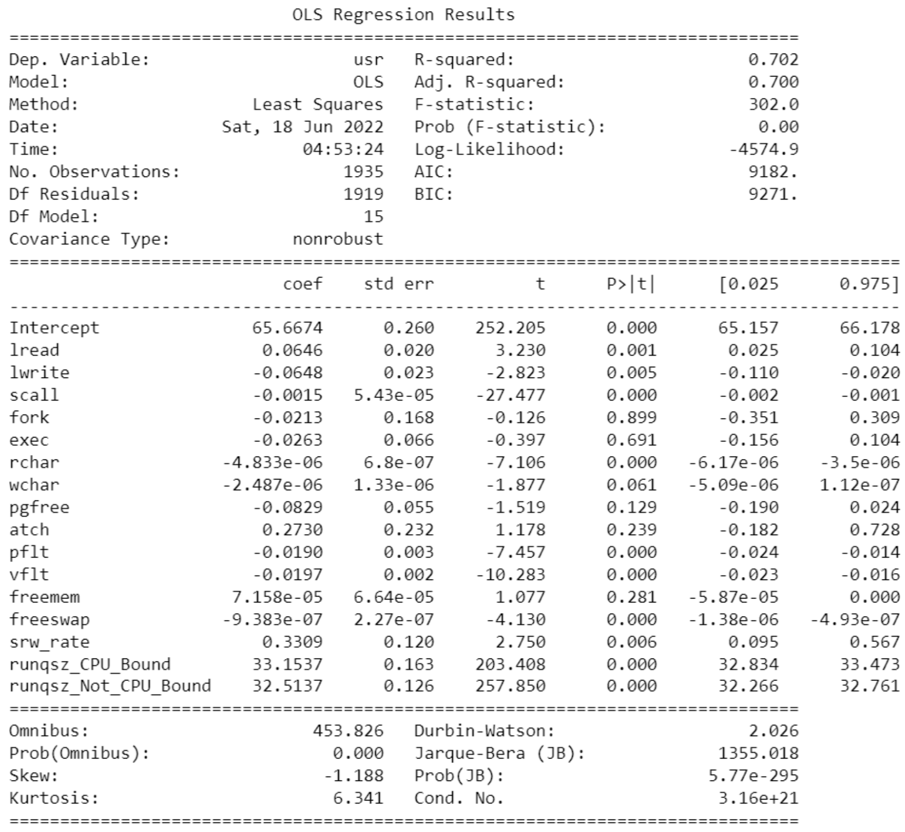
Rsquare for Train data: 0.70

RMSE for Train data: 2.57

Rsquare for Test data: 0.68

RMSE for Test data: 2.62

OLS Output:



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    usr   R-squared:                       0.702
Model:                            OLS   Adj. R-squared:                  0.700
Method:                 Least Squares   F-statistic:                     302.0
Date:                Sat, 18 Jun 2022   Prob (F-statistic):               0.00
Time:                        04:53:24   Log-Likelihood:                -4574.9
No. Observations:                1935   AIC:                             9182.
Df Residuals:                    1919   BIC:                             9271.
Df Model:                          15
Covariance Type:            nonrobust
==========================================================================================
                           coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------
Intercept               65.6674      0.260    252.205      0.000      65.157      66.178
lread                    0.0646      0.020      3.230      0.001       0.025       0.104
lwrite                  -0.0648      0.023     -2.823      0.005      -0.110      -0.020
scall                   -0.0015   5.43e-05    -27.477      0.000      -0.002      -0.001
fork                    -0.0213      0.168     -0.126      0.899      -0.351       0.309
exec                    -0.0263      0.066     -0.397      0.691      -0.156       0.104
rchar                -4.833e-06    6.8e-07     -7.106      0.000   -6.17e-06    -3.5e-06
wchar                -2.487e-06   1.33e-06     -1.877      0.061   -5.09e-06    1.12e-07
pgfree                  -0.0829      0.055     -1.519      0.129      -0.190       0.024
atch                     0.2730      0.232      1.178      0.239      -0.182       0.728
pflt                    -0.0190      0.003     -7.457      0.000      -0.024      -0.014
vflt                    -0.0197      0.002    -10.283      0.000      -0.023      -0.016
freemem               7.158e-05   6.64e-05      1.077      0.281   -5.87e-05       0.000
freeswap             -9.383e-07   2.27e-07     -4.130      0.000   -1.38e-06   -4.93e-07
srw_rate                 0.3309      0.120      2.750      0.006       0.095       0.567
runqsz_CPU_Bound        33.1537      0.163    203.408      0.000      32.834      33.473
runqsz_Not_CPU_Bound    32.5137      0.126    257.850      0.000      32.266      32.761
==============================================================================
Omnibus:                      453.826   Durbin-Watson:                   2.026
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1355.018
Skew:                          -1.188   Prob(JB):                    5.77e-295
Kurtosis:                       6.341   Cond. No.                     3.16e+21
==============================================================================
```

Figure 11: Statsmodel model 2

There seems not much of a difference in the Rsqaure values for the 2 models above. We can try to remove more non-significant variables, and build more models.

Model 3

Building another model using Statsmodel without the variables: "fork","exec","atch","pgfree","freemem" which have very high p-value.

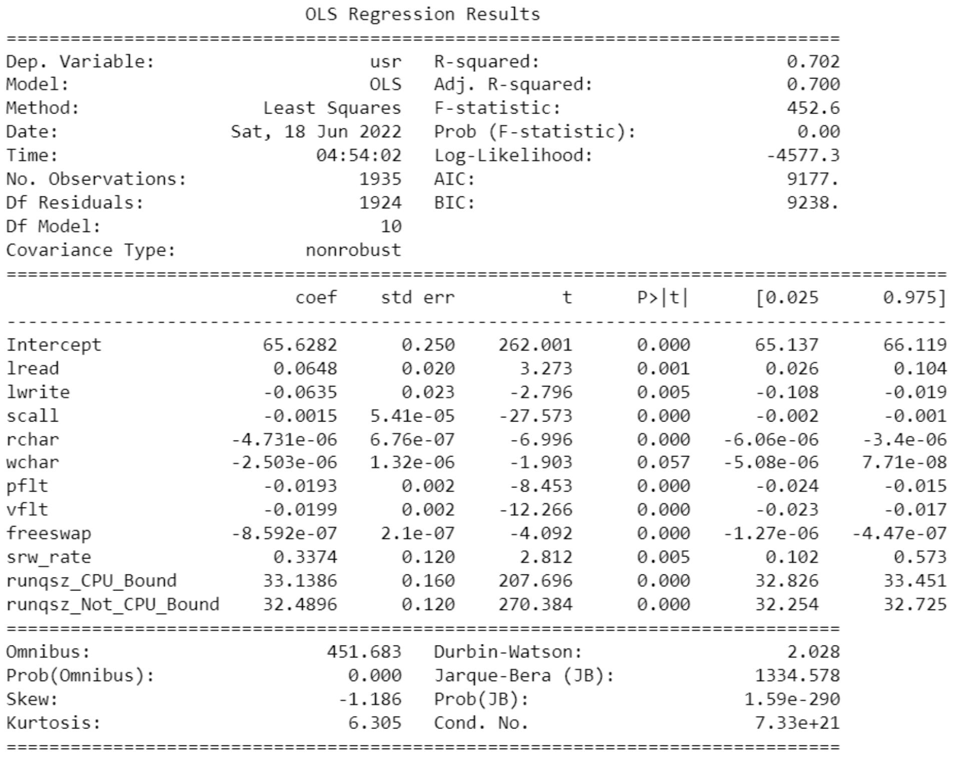
Rsquare for Train data: 0.70

RMSE for Train data: 2.57

Rsquare for Test data: 0.68

RMSE for Test data: 2.63

OLS Output:



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    usr   R-squared:                       0.702
Model:                            OLS   Adj. R-squared:                  0.700
Method:                 Least Squares   F-statistic:                     452.6
Date:                Sat, 18 Jun 2022   Prob (F-statistic):               0.00
Time:                        04:54:02   Log-Likelihood:                -4577.3
No. Observations:                1935   AIC:                             9177.
Df Residuals:                    1924   BIC:                             9238.
Df Model:                          10
Covariance Type:            nonrobust
========================================================================================
                          coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------
Intercept              65.6282      0.250    262.001      0.000      65.137      66.119
lread                   0.0648      0.020      3.273      0.001       0.026       0.104
lwrite                 -0.0635      0.023     -2.796      0.005      -0.108      -0.019
scall                  -0.0015   5.41e-05    -27.573      0.000      -0.002      -0.001
rchar               -4.731e-06   6.76e-07     -6.996      0.000   -6.06e-06    -3.4e-06
wchar               -2.503e-06   1.32e-06     -1.903      0.057   -5.08e-06    7.71e-08
pflt                   -0.0193      0.002     -8.453      0.000      -0.024      -0.015
vflt                   -0.0199      0.002    -12.266      0.000      -0.023      -0.017
freeswap            -8.592e-07    2.1e-07     -4.092      0.000   -1.27e-06   -4.47e-07
srw_rate                0.3374      0.120      2.812      0.005       0.102       0.573
runqsz_CPU_Bound       33.1386      0.160    207.696      0.000      32.826      33.451
runqsz_Not_CPU_Bound   32.4896      0.120    270.384      0.000      32.254      32.725
==============================================================================
Omnibus:                      451.683   Durbin-Watson:                   2.028
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1334.578
Skew:                          -1.186   Prob(JB):                    1.59e-290
Kurtosis:                       6.305   Cond. No.                     7.33e+21
==============================================================================
```

Figure 12: Statsmodel model 3

There seems to be no major improvement in the R Square and RMSE values after removing the not so significant variables.

It would be better to go with the SKlearn model for prediction and Statsmodel model for interpretation and understand which variables are playing a major role in the model.

<u>NOTE:</u> The VIF method can also be used for identifying important variables and eliminating the ones that are not significant and have high multicollinearity.

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

Business Insights

The following are the observations for the predictions made by the model:

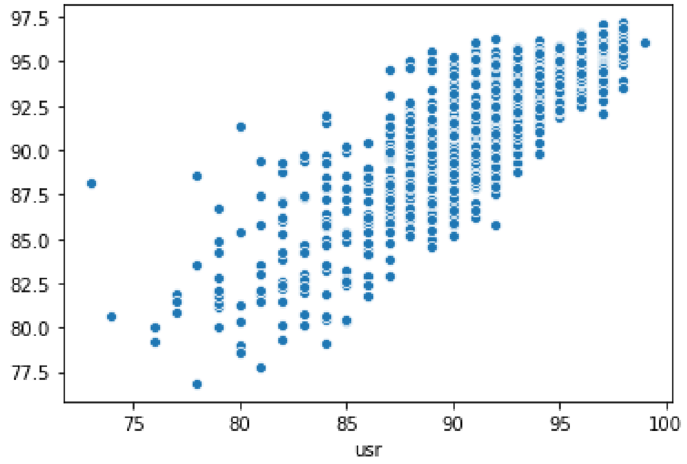


Figure 13: Scatterplot- Actual vs predicted

- Since this is regression model, we have plotted the predicted y values vs the actual y values for the test dataset. This is the plot obtained.
- From the plot, it is visible that the actual and the predicted values are close enough, except for a few. This shows that the model performed good as per the data.
- We get the following Linear Regression equation from the final model:

Usr = (66.2) * Intercept + (0.04) * lread + (-0.06) * lwrite + (-0.0) * scall + (-0.46) * fork + (-0.01) * exec + (-0.0) * rchar + (-0 .0) * wchar + (0.33) * pgout + (-0.25) * ppgout + (-0.03) * pgfree + (0.0) * pgscan + (0.34) * atch + (0.04) * pgin + (-0.12) * ppgin + (-0.02) * pflt + (-0.01) * vflt + (0.0) * freemem + (-0.0) * freeswap + (0.22) * srw_rate + (33.48) * runqsz_CPU_Bou nd + (32.72) * runqsz_Not_CPU_Bound

- We see that CPU run in user mode is highly influenced by the Process run queue size
- If the CPU bound queue size is increased by 1 unit, the % of time the CPU will run in user mode will increase by 33.5 times, keeping all other features constant.
- Similarly, if the non-CPU bound queue size is increased by 1 unit, the % of time the CPU will run in user mode will increase by 32.7 times, keeping all other features constant.
- Together the Process run queue size variable affects the % of time the CPU will run in user mode by a value of approx. 132 times, including the Intercept.
- All the other features are not impacting the CPU runtime too significantly.

Problem Statement 2: Classification

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.

Data Description

Column Name	Description	Data Type
Wife_age	Wife's age	numerical
Wife_ education	1=uneducated, 2, 3, 4=tertiary	Categorical
Husband_education	1=uneducated, 2, 3, 4=tertiary	Categorical
No_of_children_born		numerical
Wife_religion	Non-Scientology, Scientology	Binary
Wife_Working	Yes, No	Binary
Husband_Occupation	1, 2, 3, 4(random)	Categorical
Standard_of_living_index	1=very low, 2, 3, 4=high	Categorical
Media_exposure	Good, Not good	Binary
Contraceptive_method_used		Target

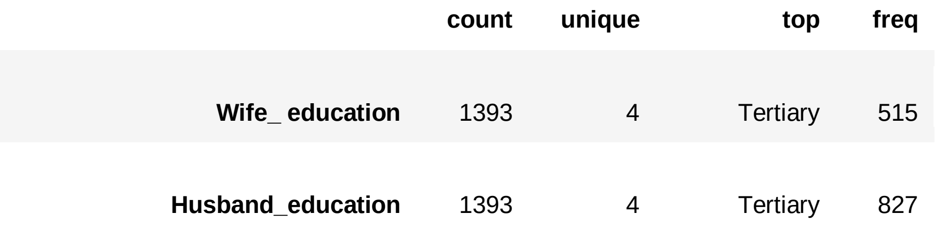
Table 4: Data description: Dataset 2

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.

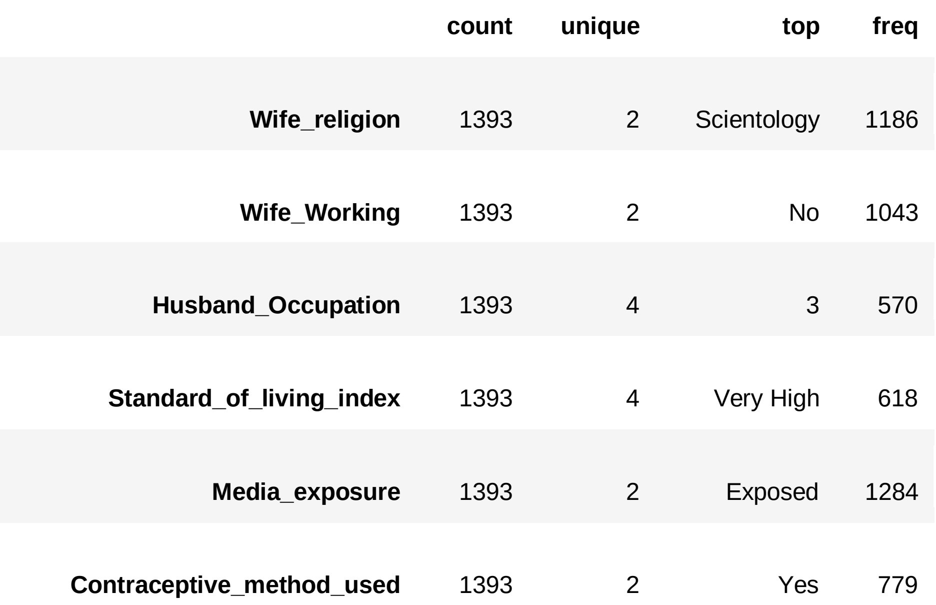
EDA

The data is imported, and the following are the observations:

- The data has 1473 rows and 10 variables. There are 5 variables that have object data types and rest are int data types.
- There are a few missing values in the dataset in the variables 'wife_age' and 'No_of_children_born'. These are replaced by the median values to remove the null entries.
- There are 80 duplicate rows which can be dropped from the dataset. The number for rows is 1393 now.
- The variable 'Husband_Occupation' has been also changed to Object data type as it is a categorical variable.



	count	unique	top	freq
Wife_ education	1393	4	Tertiary	515
Husband_education	1393	4	Tertiary	827



	count	unique	top	freq
Wife_religion	1393	2	Scientology	1186
Wife_Working	1393	2	No	1043
Husband_Occupation	1393	4	3	570
Standard_of_living_index	1393	4	Very High	618
Media_exposure	1393	2	Exposed	1284
Contraceptive_method_used	1393	2	Yes	779

Table 5: Data description

- From the 5-point summary of the object type variables, we can see that Tertiary is the most frequent education level of both Husband and Wife.
- Scientology is the most frequent religion that is followed by the women and majority of them are not working.
- Majority of the Husbands Occupation is of level 3.
- The Standard of living index is very high amongst the people and majority of them are exposed to media.
- This means that the people might be from a city or an urban area.
- Majority of the women have used a contraceptive method.

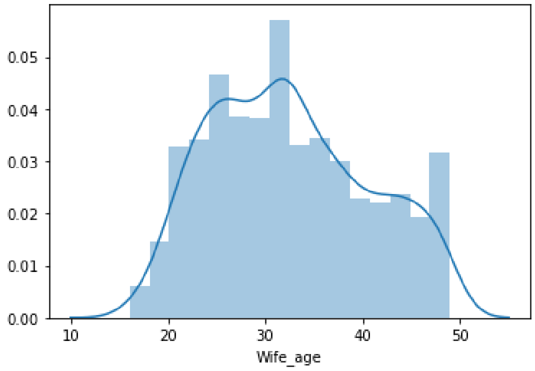


Figure 14: Histogram (wife age)

- The age of the wives ranges from 18 to 49 years where most of them are in their 30’s and mid 20’s early 50’s.
- Majority of the people had 1 or 2 children but a few have more than 15 children as well.

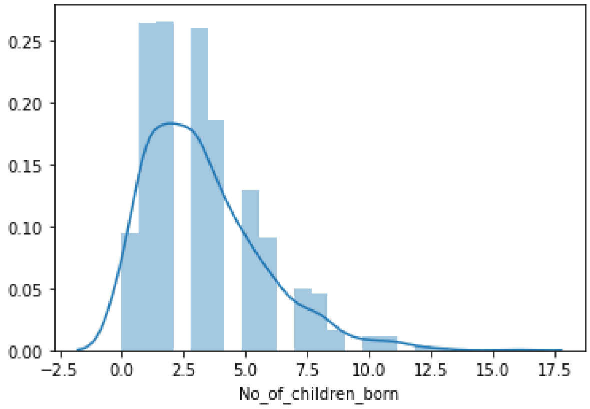


Figure 15: Histogram (No. of children born)

Univariate Analysis

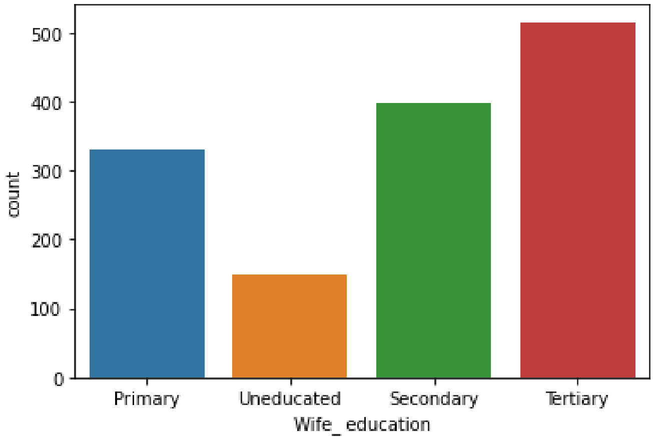


Figure 16: Countplot (Wife education)

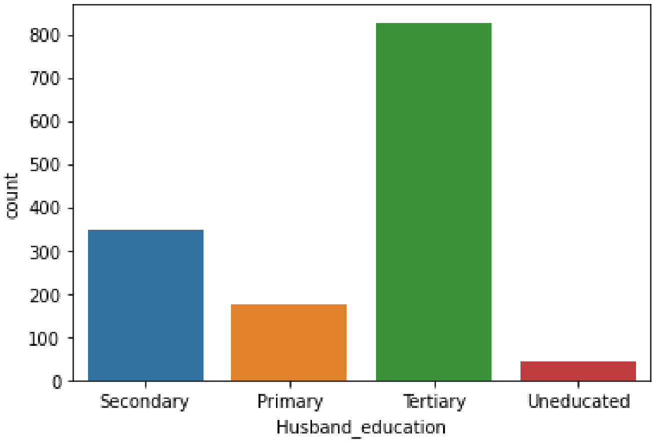


Figure 17: Husband education

- As mentioned, Tertiary is the most frequent education level of both Husband and Wife.
- Fewer Husbands are uneducated as compared to the wives.

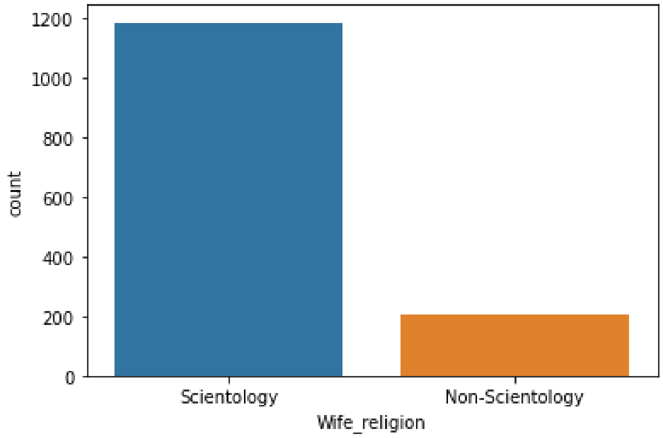


Figure 18: Countplot Wife religion

- Also, Scientology is followed the most.

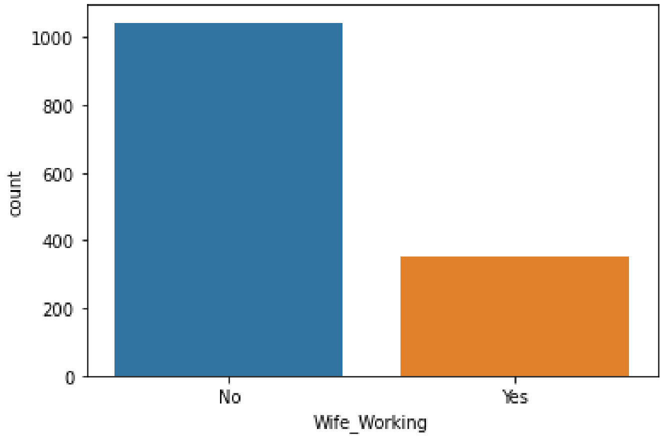


Figure 19: Countplot (wife working)

- Majority of the wives are not working.

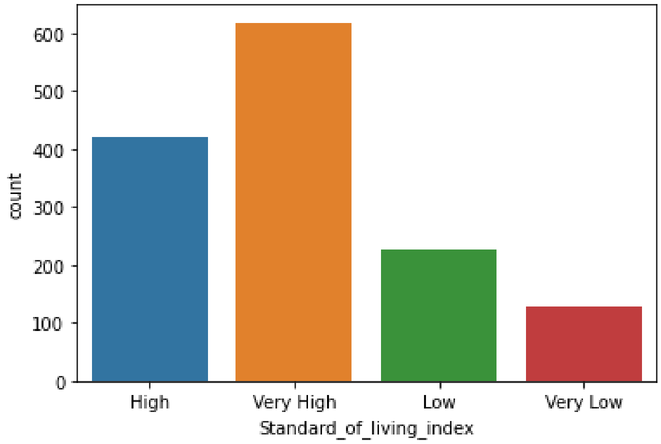


Figure 20: Countplot (Standard of living index)

- Major portion of the people are from the areas where the standard of living is Very High and High.
- In total around 350 people are from the areas with Low and Very low standard of living index.

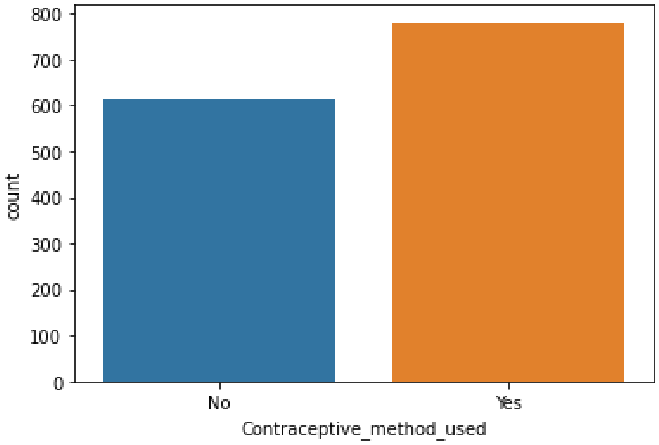


Figure 21: Countplot (Contraceptive method used)

- We already know that the majority of the women have used a contraceptive method, however there is a good proportion as well who have not used any.

Bivariate Analysis

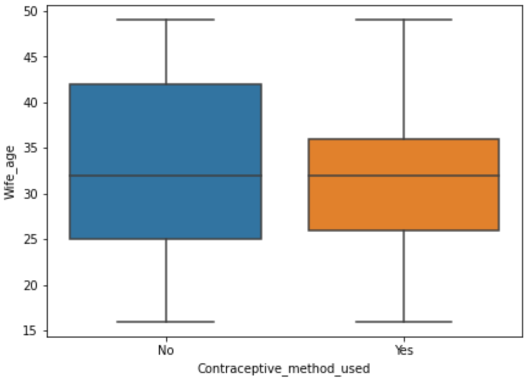


Figure 22: Contraceptives used vs Wife age

- Looks like females at an age of 25 to 35 have used contraceptive methods, with some extreme values.
- Many females from 25 to 43 have not used any contraceptive methods as well.

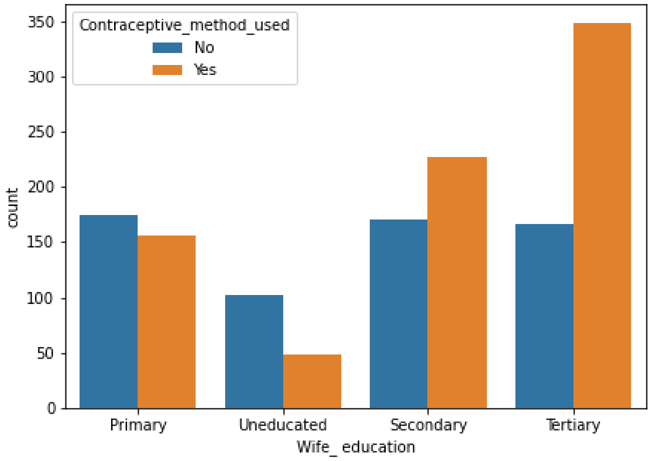


Figure 23: Contraceptives used vs wife education

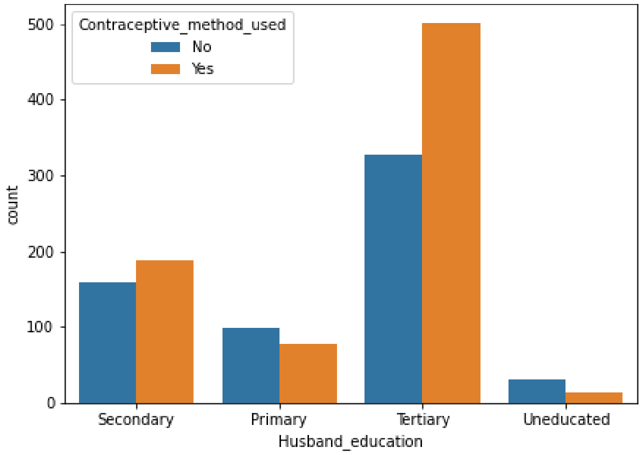


Figure 24: Contraceptives used vs Husband education

- Females who have completed their secondary and Tertiary education have used contraceptive methods more as compared to the others.
- Whereas, Females who are not educated or only completed Primary education tend not to use any contraceptive methods.
- Similar finding can be seen based on the Husband’s education level.

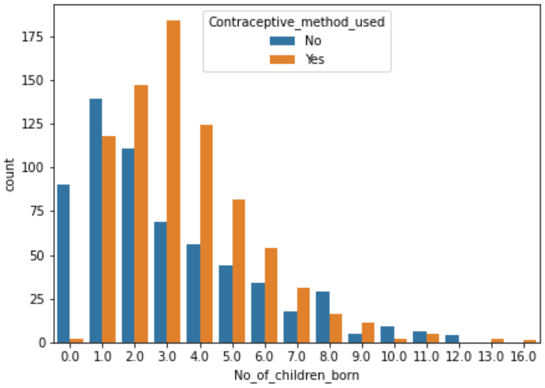


Figure 25: Contraceptives used vs No.of children

- It is interesting to see that the majority of women are using contraceptives after 3 childern.
- Majority Women who have only 1 child are not taking any contraceptives. This indicates that they have intentions to have more children.
- However, a very few women also take contraceptives even though they have no children.

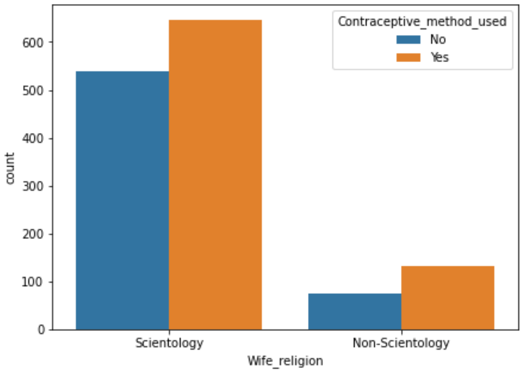


Figure 26: Contraceptives used vs Wife religion

- Religion does not seem to affect the use of Contraceptives.

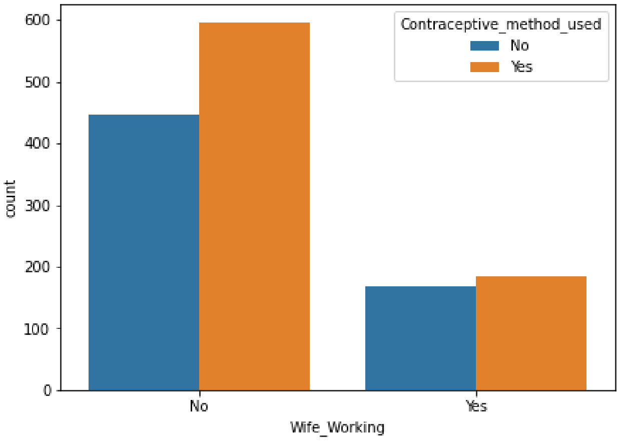


Figure 27: Contraceptives used vs Wife working

- The proportion of non-working women taking contraceptives as more as compared to the women who are working.

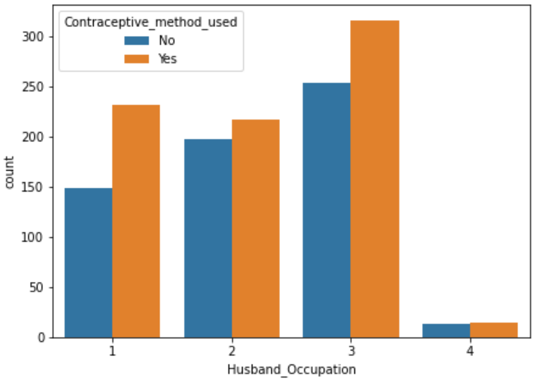


Figure 28: Contraceptive used vs Husband occupation

- Since we do not have clear definitions of the Husbands Occupation levels, assuming level 1 to be the lowest and 4 being the highest.
- The proportion of females using contraceptives as more for occupation level 1,2 & 3 as compared to 4.

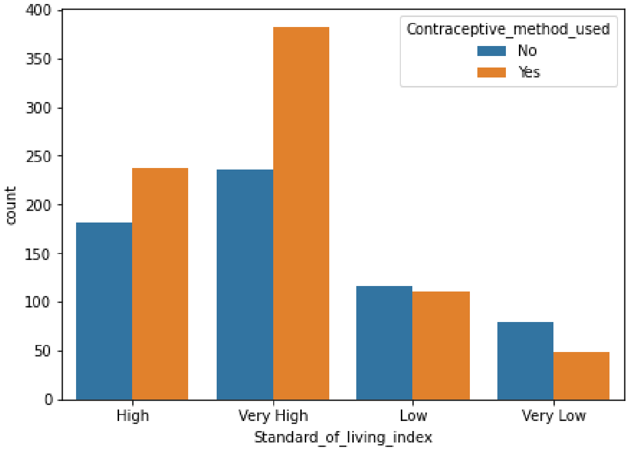


Figure 29: Contraceptives used vs Standard of living index

- As seen before, women from area with high and very high standard of living use contraceptive methods

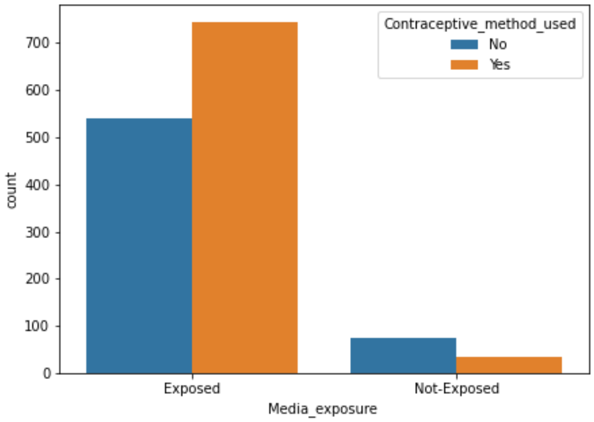


Figure 30: Contraceptives used vs Media Exposure

- Women with the exposure to media use contraceptives more as compared to the others.

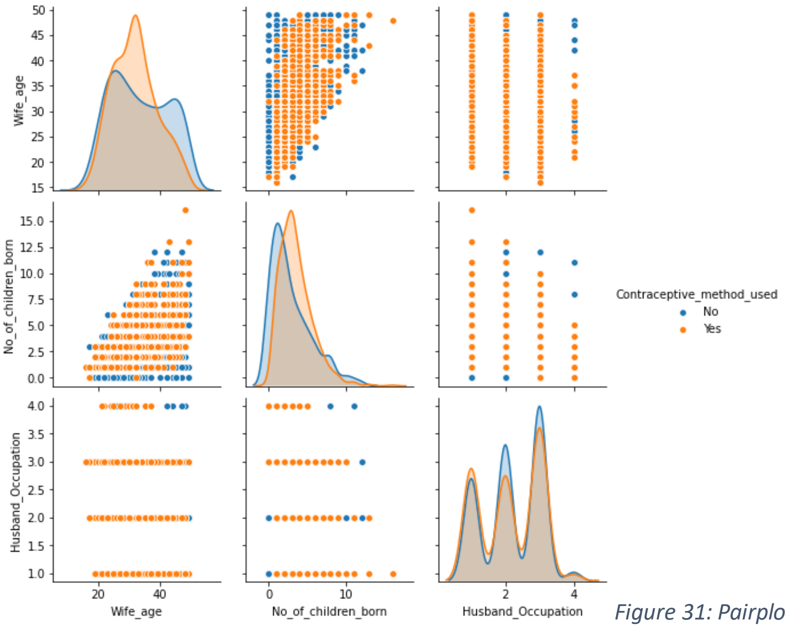


Figure 31: Pairplot

- The pairplot does not indicate any major trend/correlation between the variables.
- Some of the variables available in the pairplot, do not have the classes well separated. They will not be considered as good predictors.

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 LDA (linear discriminant analysis) and CART.

Encoding

- Since the data has string & categorical type variables, these variables must be encoded so that the Machine Learning model understands the data.
- In the target variable, "No" is replaced by 0 and "Yes" is replaced by 1 first.

- Similarly, ordinal numbers are given to the values in variables Wife_ education, Husband_education & Standard_of_living_index.
- After this dummy encoding us used to encode the data for the rest of the columns.
- The dataset looks like this.

	Wife_age	Wife_education	Husband_education	No_of_children_born	Standard_of_living_index	Wife_religion_Scientology	Wife_Working_Yes	Husband_Occupation_2	Husband_Occupation_3	Husband_Occupation_4	Media_exposure_Not-Exposed
0	24.0	2	3	3.0	3	1	0	1	0	0	0
1	45.0	1	3	10.0	4	1	0	0	1	0	0
2	43.0	2	3	7.0	4	1	0	0	1	0	0

Table 6: Encoded data

Train-Test split

- To build the Machine Learning models, we split the entire data set into a ratio of 70:30 into Training dataset and Testing dataset.
- Since there are 1 and 0 values in the dependent variable, we need to ensure that an equal number of 1 and 0 are split into both Training and Testing datasets.
- This will ensure a balance in the data and will not cause biasness while Training or Testing the model. Therefore, a function stratify=target is used while splitting.

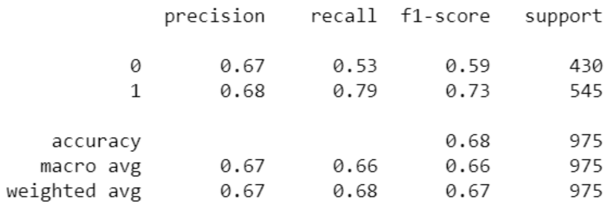
Model Building

Logistic Regression Model

After the data preprocessing Logistic Regression model is applied to the Train and Test datasets with default hyper-parameters and solver considered as to be ‘newton-cg’.

Performance metrics

- Classification report – Train Data



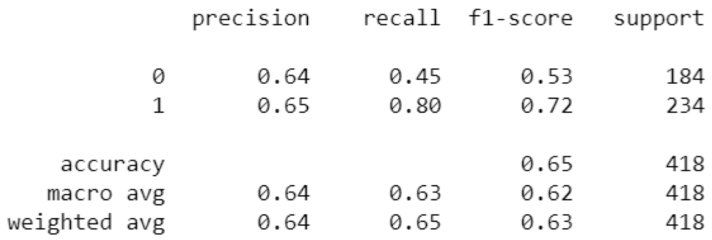
```
              precision    recall  f1-score   support

           0       0.67      0.53      0.59       430
           1       0.68      0.79      0.73       545

    accuracy                           0.68       975
   macro avg       0.67      0.66      0.66       975
weighted avg       0.67      0.68      0.67       975
```

Table 7: Classification report - Logistic regression model 1 - Train

- Classification report – Test Data



```
              precision    recall  f1-score   support

           0       0.64      0.45      0.53       184
           1       0.65      0.80      0.72       234

    accuracy                           0.65       418
   macro avg       0.64      0.63      0.62       418
weighted avg       0.64      0.65      0.63       418
```

Table 8: Classification report - Logistic Regression model 1 - Test

- AUC and ROC Curve – Train Data
  - AUC: 0.722

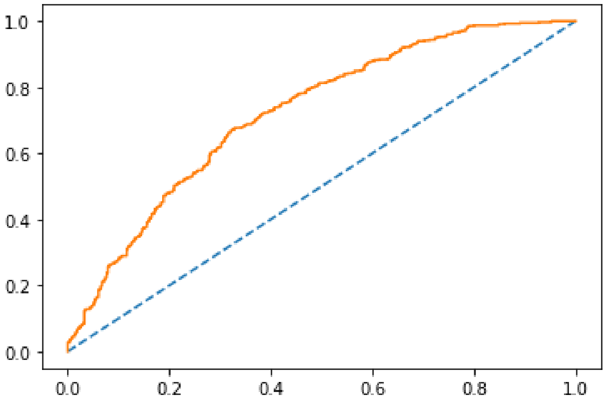


Figure 32: ROC Curve - Logistic regression model - Train

- AUC and ROC Curve – Test Data
  - AUC: 0.665

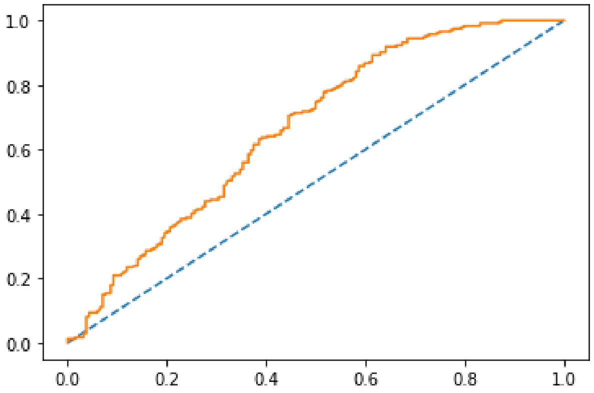


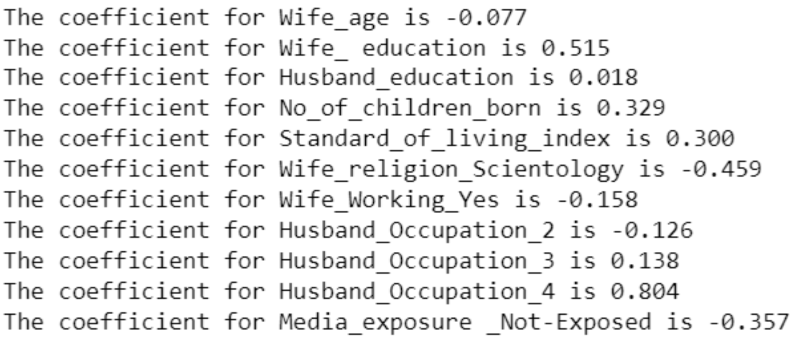
Figure 33: ROC Curve - Logistic regression model - Test

Inference

From the Accuracy and Recall values, the model seems to be performing fine. However, from the AUC values & ROC curve for Test data shows that it is not covering a large area as compared to the train data.

Therefore, there is a need to optimize this model.

Feature importance:



```
The coefficient for Wife_age is -0.077
The coefficient for Wife_ education is 0.515
The coefficient for Husband_education is 0.018
The coefficient for No_of_children_born is 0.329
The coefficient for Standard_of_living_index is 0.300
The coefficient for Wife_religion_Scientology is -0.459
The coefficient for Wife_Working_Yes is -0.158
The coefficient for Husband_Occupation_2 is -0.126
The coefficient for Husband_Occupation_3 is 0.138
The coefficient for Husband_Occupation_4 is 0.804
The coefficient for Media_exposure _Not-Exposed is -0.357
```

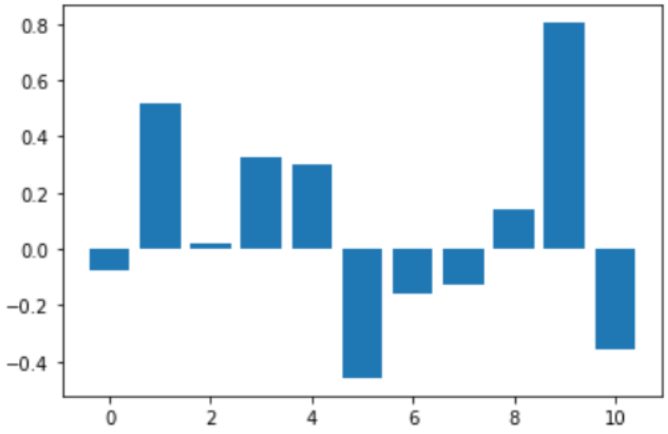


Figure 34: Important features

- The model also shows that features like Wife education, Husband Occupation 4 and highly important as they are directly correlated to the dependent variable.
- The variables Wife_religion_Scientology and Media_exposure_Not-Exposed are also very important features.

Optimized Logistic Regression Model

To optimize the Logistic Regression model, the best parameters are found using Grid Search Cross Validation technique.

These are the best parameters obtained:

'penalty': 'l1'

'solver': 'saga'

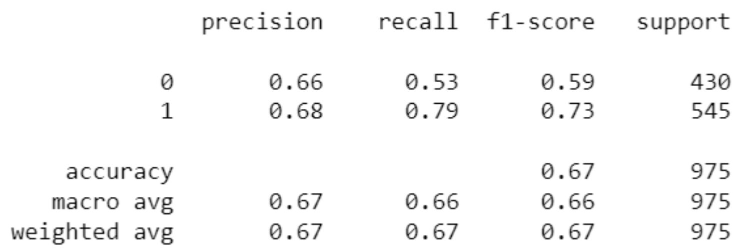
'tol': 0.0001

Another Logistic Regression model is built with these best parameters

The model evaluation score is calculated, along with the confusion matrix. The AUC-ROC curve is also plotted for both the versions of the model to check their performance. This will be described later.

Performance metrics

- Classification report – Train Data



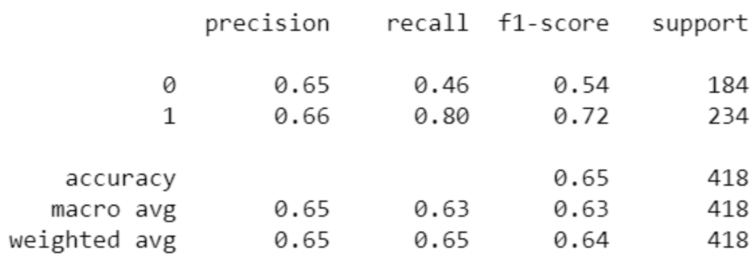
```
              precision    recall  f1-score   support

           0       0.66      0.53      0.59       430
           1       0.68      0.79      0.73       545

    accuracy                           0.67       975
   macro avg       0.67      0.66      0.66       975
weighted avg       0.67      0.67      0.67       975
```

Table 9: Classification report - Optimized Logistic Regression model – Train

- Classification report – Test Data



```
              precision    recall  f1-score   support

           0       0.65      0.46      0.54       184
           1       0.66      0.80      0.72       234

    accuracy                           0.65       418
   macro avg       0.65      0.63      0.63       418
weighted avg       0.65      0.65      0.64       418
```

Table 10: Classification report - Optimized Logistic Regression model – Test

- AUC and ROC Curve – Train Data
    - AUC: 0.722

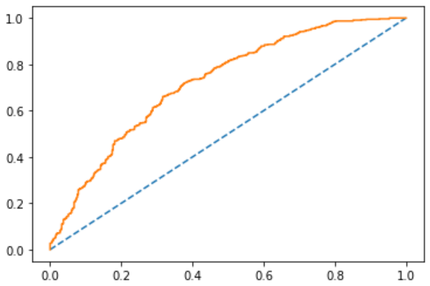


Figure 35: ROC Curve - Optimized Logistic Regression model – Train

- AUC and ROC Curve – Test Data
    - AUC: 0.722

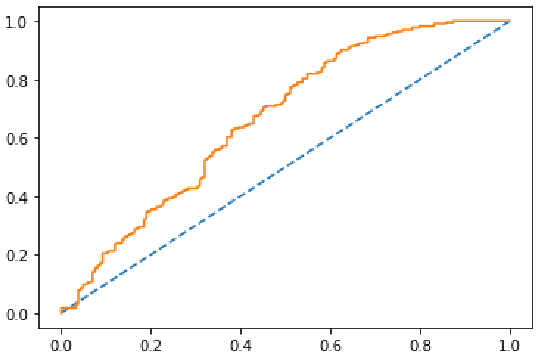


Figure 36: ROC Curve - Optimized Logistic Regression model – Test

Inference

The Accuracy, Recall and Precision seems to be the same as per the previous model, however there are slight variation in the AUC score.

There does not seem to be much of an improvement in the figures, therefore let us try to build an LDA model to get better performance.

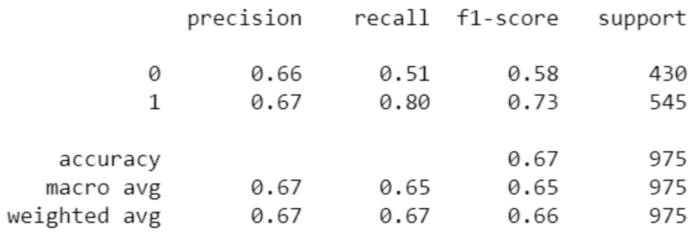
LDA Model

The LDA model is also built with default parameters. The default cut-off value of 0.5 is considered for prediction.

This model is also further evaluated with Accuracy score, along with the confusion matrix. The AUC-ROC curve is plotted for both the Train and Test data.

Performance metrics

- Classification report – Train Data



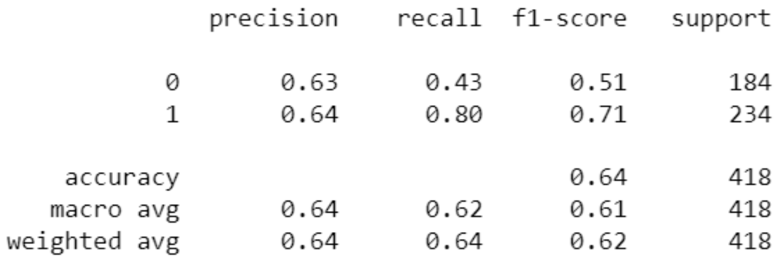
```
              precision    recall  f1-score   support

           0       0.66      0.51      0.58       430
           1       0.67      0.80      0.73       545

    accuracy                           0.67       975
   macro avg       0.67      0.65      0.65       975
weighted avg       0.67      0.67      0.66       975
```

Table 11: Classification report - LDA model - Train

- Classification report – Test Data



```
              precision    recall  f1-score   support

           0       0.63      0.43      0.51       184
           1       0.64      0.80      0.71       234

    accuracy                           0.64       418
   macro avg       0.64      0.62      0.61       418
weighted avg       0.64      0.64      0.62       418
```

Table 12: Classification report - LDA model – Test

- AUC and ROC Curve – Train Data
  - AUC: 0.722

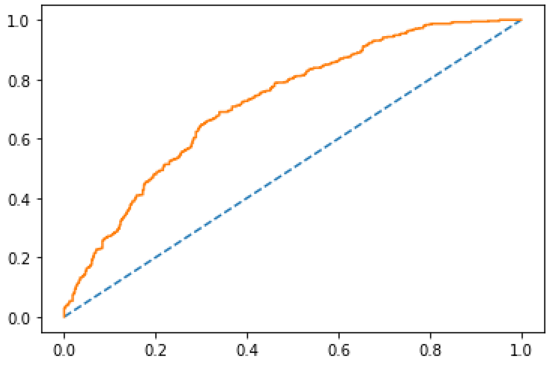


Figure 37: ROC Curve - LDA model – Train

- AUC and ROC Curve – Test Data
  - AUC: 0.662

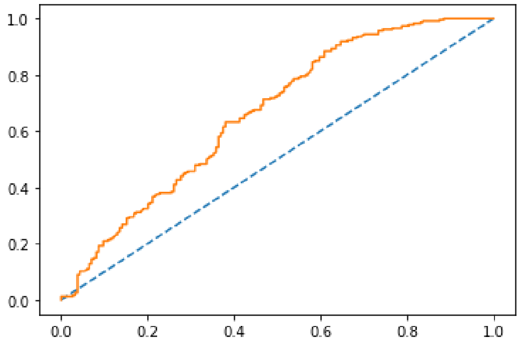


Figure 38: ROC Curve - LDA model - Test

<u>Inference</u>

The LDA model looks a bit better than the Logistic Regression models in terms of the Recall value for Train and test data. However, the Accuracy for the test data has taken a hit. The AUC and ROC curves also do not show a significant difference compared to the other models built.

CART Model

A CART model is also built using the following parameters:

criterion = 'gini',

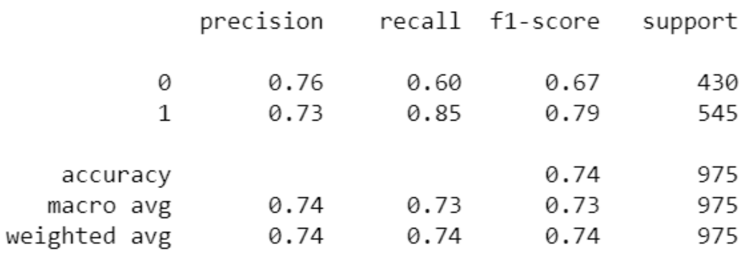
max_depth = 7,

min_samples_leaf=20,

min_samples_split=60

This model is also further evaluated with Accuracy score, along with the confusion matrix. The AUC-ROC curve is plotted for both the Train and Test data.

- Classification report – Train Data



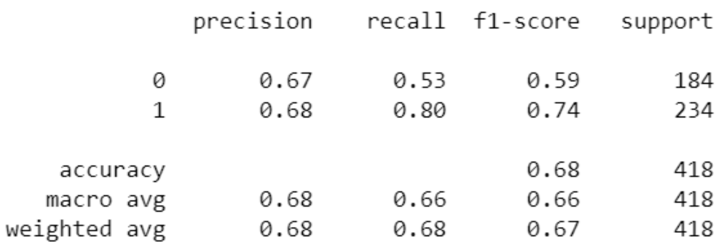
```
              precision    recall  f1-score   support

           0       0.76      0.60      0.67       430
           1       0.73      0.85      0.79       545

    accuracy                           0.74       975
   macro avg       0.74      0.73      0.73       975
weighted avg       0.74      0.74      0.74       975
```

Table 13: Classification report - CART model - Train

- Classification report – Test Data



```
              precision    recall  f1-score   support

           0       0.67      0.53      0.59       184
           1       0.68      0.80      0.74       234

    accuracy                           0.68       418
   macro avg       0.68      0.66      0.66       418
weighted avg       0.68      0.68      0.67       418
```

Table 14: Classification report - CART model - Test

- AUC and ROC Curve – Train Data
  - AUC: 0.821

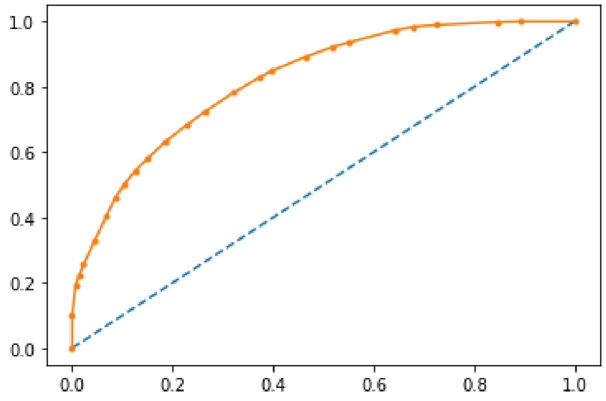


Figure 39: ROC Curve - CART model – Train

- AUC and ROC Curve – Test Data
  - AUC: 0.721

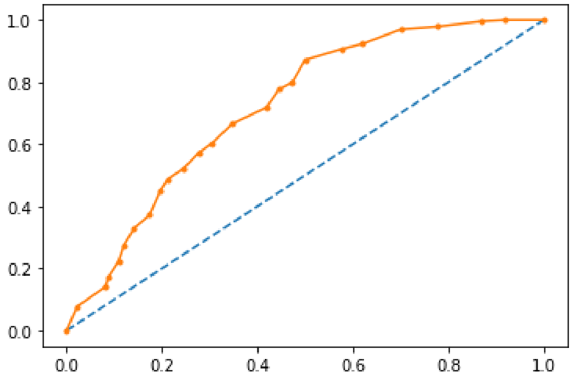
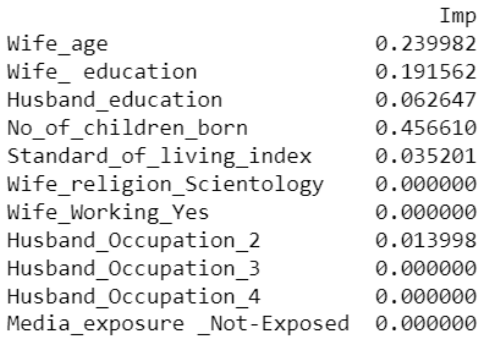


Figure 40: ROC Curve - CART model – Test

Inference

The CART model from all the other models seems to be performing the best in terms of Accuracy, Recall and Precision values.

Feature Importance:



```
                                  Imp
Wife_age                     0.239982
Wife_ education              0.191562
Husband_education            0.062647
No_of_children_born          0.456610
Standard_of_living_index     0.035201
Wife_religion_Scientology    0.000000
Wife_Working_Yes             0.000000
Husband_Occupation_2         0.013998
Husband_Occupation_3         0.000000
Husband_Occupation_4         0.000000
Media_exposure _Not-Exposed  0.000000
```

Table 15: Important features from CART model

The CART model also gives the most important features according to which the split in the Decision Tree was made.

Wife_age, Wife_ education & No_of_children_born are the important features. These are not the same as the Logistic Regression model suggested.

2.3 Performance 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.

Model Evaluation and Performance

To check performance of Predictions of every model built on Train and Test datasets, Accuracy score is calculated.

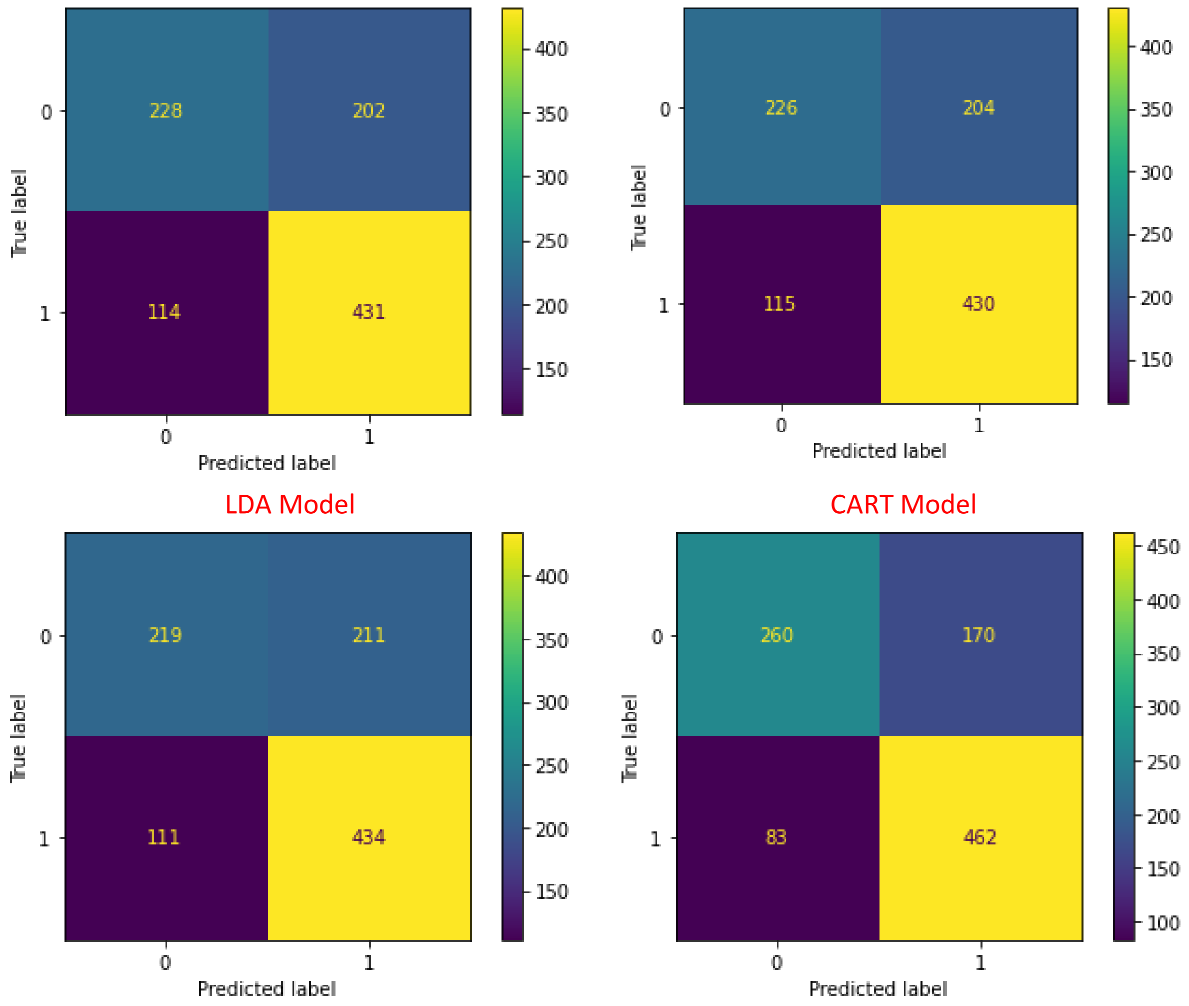
A Confusion Matrix, ROC curve and AUC-ROC score has been devised as well.

We have considered the ‘Contraceptive method used’ i.e both 0, 1 as the interest classes. Therefore, we will also look at the Accuracy scores of all models.

Comparing Confusion matrix of all models (Train data)

Logistic Regression

Tuned Logistic Regression Model



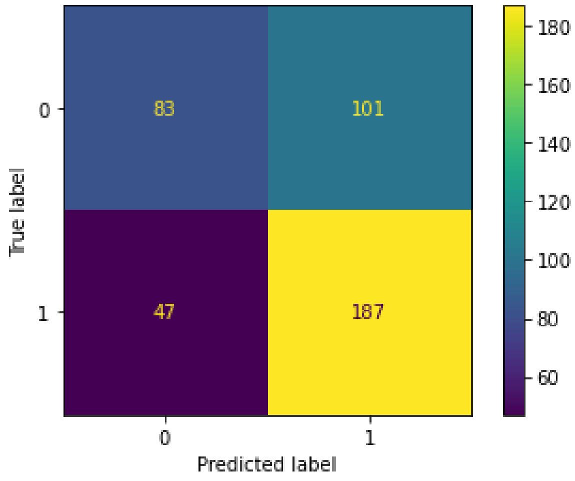
LDA Model

CART Model

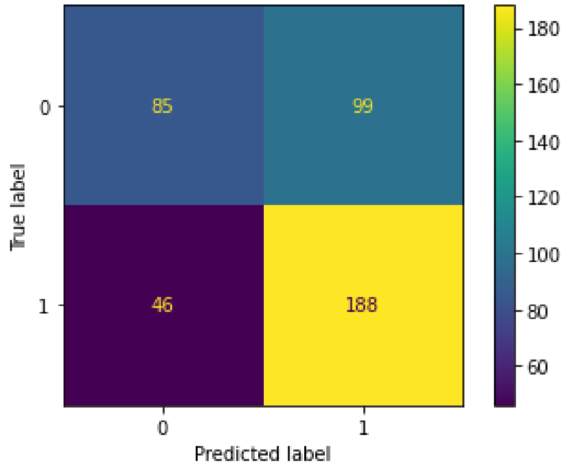
Figure 41: Confusion matrices of all models (Train data)

Comparing Confusion matrix of all models (Test data)

Logistic Regression



Tuned Logistic Regression Model



LDA Model

CART Model

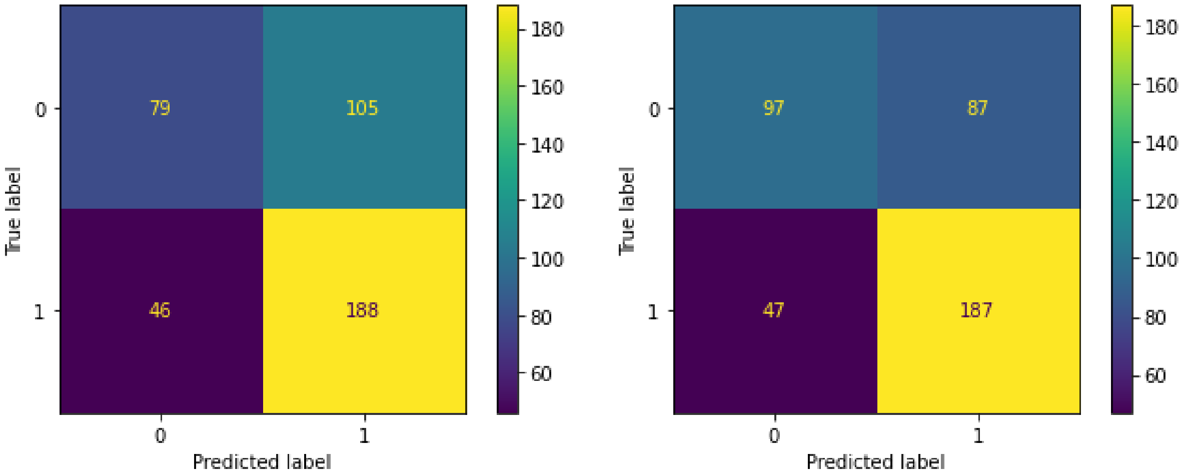


Figure 42: Confusion matrices of all models (Test data)

Model Name	Accuracy		Recall		Precision		AUC	
	Train	Test	Train	Test	Train	Test	Train	Test
Logistic Regression Model	0.68	0.65	0.79	0.80	0.68	0.65	0.72	0.66
Tuned Logistic Regression Model	0.67	0.65	0.79	0.80	0.68	0.66	0.72	0.72
LDA Model	0.67	0.64	0.80	0.80	0.67	0.64	0.72	0.66
CART model	0.74	0.68	0.85	0.8	0.73	0.68	0.82	0.72

Table 16: Different model parameters

From all the inferences above, we see that mostly all the models have similar performance.

The Accuracy score for all the models are above 65% for both test and train data.

Best model selection:

With this, it is also very clear that the CART model has performed above all the rest of the models. With an **Accuracy** value of 68%, it is predicting the highest percentage of both our classes of interest.

If we still look at the **Recall** value, the CART model is able to identify 80% of the true positives correctly. The LDA model also gives a similar Recall value, however the Accuracy of the CART model is slightly higher therefore it would be better to consider the CART model for doing the prediction.

Similarly, we see that the Area Under the Curve (AUC) captured is 82% for train data and 72% for the test data. It is not the best however; it still supersedes all the other models.

Therefore, it is safe to say that this model can be used for making predictions on any unseen data that is fed to the model.

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

Business Insights & Recommendations

From the important features from Logistic Regression Model & CART Model

- As per the Logistic Regression model, the wife's education, no. of children born is very important in deciding whether the women will use contraceptive methods or not.
- The CART model also indicates that the wife's education, no. of children born are important. Therefore, these features are highly important.
- Both the models also indicated that the Husband's education is also important, and in real life that makes sense. This feature can influence the wife's decision to use contraceptive methods.

Recommendations

- Women from area with high and very high standard of living are more likely to use contraceptive methods.
- Women between the age of 25 to 35 years are more likely to use contraceptives which have a good education level.
- The education level of the husband also plays a major role in contributing to the fact that the wife will use contraceptive methods or not.
- It would be helpful to get the viewpoint of the women who do not have any children and are still using contraceptives.
- The exposure to media also plays a key role.
- Republic of Indonesia Ministry of Health can reach out to women who do not use contraceptive and can educate them about its usage, affects etc.
- Wives who have 8, 10, 11 & 12 do not use contraceptives. It would be interesting to see if why this situation is there.