**PREDICTIVE**

**MODELLING**

**Project Report**

Submitted by,

Sindhu R Udupa.

BATCH: PGPDSBA.O. NOV22.B

**Contents**

|  |  |  |
| --- | --- | --- |
| **Sl. No.** | **Details** | **Page #** |
|  | **Part -1: Linear Regression** |  |
| 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. | 6 |
| 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. | 17 |
| 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. | 19 |
| 1.4 | Inference: Basis on these predictions, what are the business insights and recommendations. | 30 |
|  | **Part – 2: Logistic Regression and LDA** |  |
| 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. | 31 |
| 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. | 38 |
| 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. | 38 |
| 2.4 | Inference: Basis on these predictions, what are the insights and recommendations. | 45 |

**List of Figures**

|  |  |  |
| --- | --- | --- |
| **Sl. No** | **Figure Details** | **Page #** |
| 1 | Fig. 1: Boxplots of all continuous variables | 9 |
| 2 | Fig. 2: Histograms of all continuous variables. | 11 |
| 3 | Fig. 3: Count plot of the categorical variable. | 12 |
| 4 | Fig. 4: Scatter plots of the continuous variables v/s the target variable. | 14 |
| 5 | Fig. 5: Scatterplots of all variables v/s the target variable. | 15 |
| 6 | Fig. 6: Heatmap. | 16 |
| 7 | Fig. 7: Boxplots after outlier treatment. | 18 |
| 8 | Fig. 8: Actual values v/s fitted values | 29 |
| 9 | Fig. 9: Residuals v/s fitted values. | 29 |
| 10 | Fig. 10: Boxplots of numerical variables. | 34 |
| 11 | Fig. 11: Histograms of numerical variables. | 34 |
| 12 | Fig. 12: Count plots of categorical variables. | 35 |
| 13 | Fig. 13: Histograms depicting the poor predictors. | 36 |
| 14 | Fig. 14: Box plots of the numerical variable. | 36 |
| 15 | Fig. 15: Bivariate analysis using countplots. | 37 |
| 16 | Fig. 16: Confusion Matrix – Logistic Regression. | 40 |
| 17 | Fig. 17: ROC Curve – Logistic Regression. | 40 |
| 18 | Fig. 18: Confusion Matrix – LDA | 42 |
| 19 | Fig. 19: ROC Curve – LDA | 42 |
| 20 | Fig. 20: Confusion Matrix – CART. | 44 |
| 21 | Fig. 21: ROC Curve –CART. | 44 |
| 22 | Fig. 22: Decision tree. | 45 |

**List of Tables**

|  |  |  |
| --- | --- | --- |
| **Sl. No** | **Table Details** | **Page #** |
| 1 | Table 1: First five records of the dataset. | 6 |
| 2 | Table 2: Last five records of the dataset. | 6 |
| 3 | Table 3: Information of the dataset. | 7 |
| 4 | Table 4: Statistical description of the data. | 8 |
| 5 | Table 5: Zero values present in different columns. | 17 |
| 6 | Table 6: Regression results of model 1. | 20 |
| 7 | Table 7: VIF values of variables in model 1. | 21 |
| 8 | Table 8: Regression results of model 2. | 22 |
| 9 | Table 9: VIF values of variables in model 2. | 23 |
| 10 | Table 10: Regression results of model 3. | 24 |
| 11 | Table 11: VIF values of variables in model 3. | 25 |
| 12 | Table 12: Regression results of model 3. | 26 |
| 13 | Table 13: VIF values of variables in model 4. | 27 |
| 14 | Table 14: Comparison of all models. | 27 |
| 15 | Table 15: Regression results of Test data on model 4. | 28 |
| 16 | Table 16: First five records of the dataset. | 31 |
| 17 | Table 17: Last five records of the dataset. | 31 |
| 18 | Table 18: Information of the dataset | 32 |
| 19 | Table 19: Statistical description of the dataset. | 32 |
| 20 | Table 20: Different categories in the categorical variables | 33 |
| 21 | Table 21: Report on Train data – Logistic Regression | 39 |
| 22. | Table 22: Report on Test data – Logistic Regression | 39 |
| 23. | Table 23: Report on Train data – LDA | 41 |
| 24. | Table 24: Report on Test data – LDA | 41 |
| 25. | Table 25: Report on Train data – CART | 43 |
| 26. | Table 26: Report on Test data – CART | 43 |

**Part -1**

**Linear Regression**

* 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 data has been read. We observe the following.

1. The data has 8192 rows and 22 columns.
2. Out of the 22 columns, 13 are float, 8 are integer data type and one object type.
3. There is only one categorical column – runqsz. The different values it takes are CPU Bound and Not CPU Bound. All the remaining columns are numerical in nature.
4. There are **104** missing values in the column **rchar** and **15** missing values in **wchar**.
5. When we look at the statistical description of the data, the mean and median of most of the columns are far apart, which indicates the possible outliers.
6. There are no duplicates present in the data.
7. usr Portion of time (%) that cpus run in user mode) is out target variable. We will build a model to predict this dependent variable.

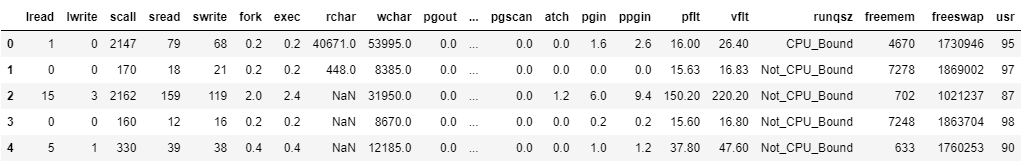


Table. 1: First five records of the data set.

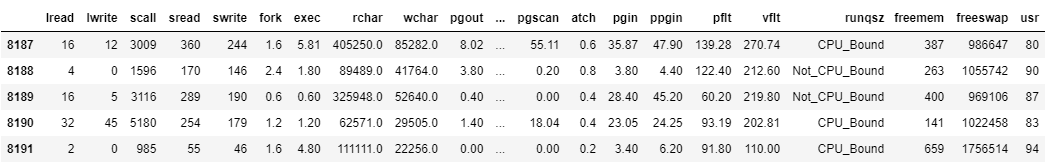


Table. 2: Last five records of the data set.

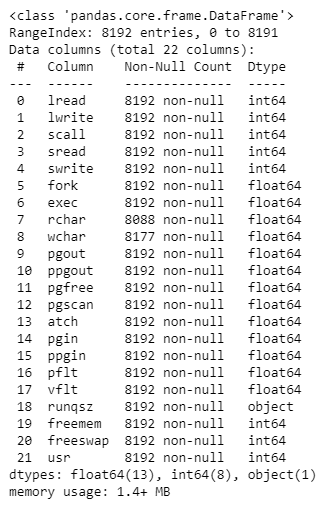
****

Table. 3: Information of the data set.

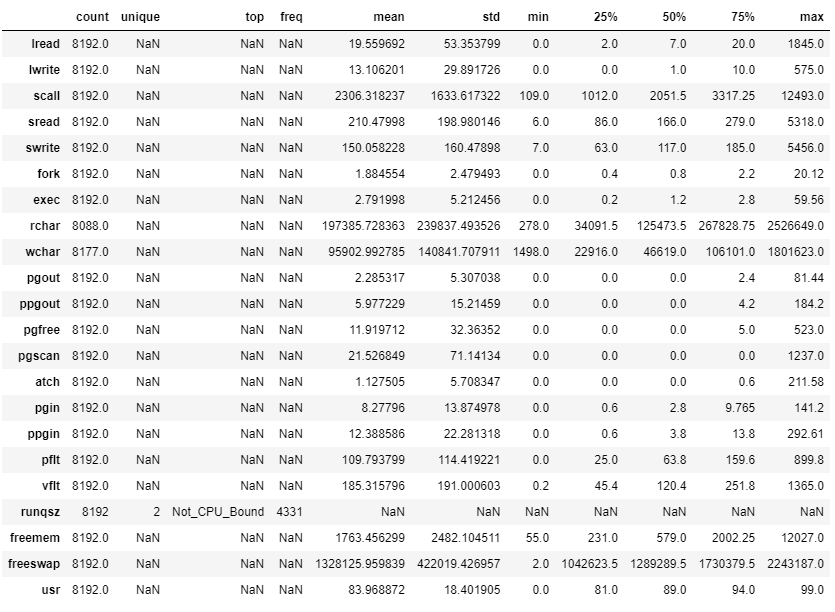
****

Table. 4: Statistical Description of the data set.

**Univariate / Bivariate and multivariate Analysis:**

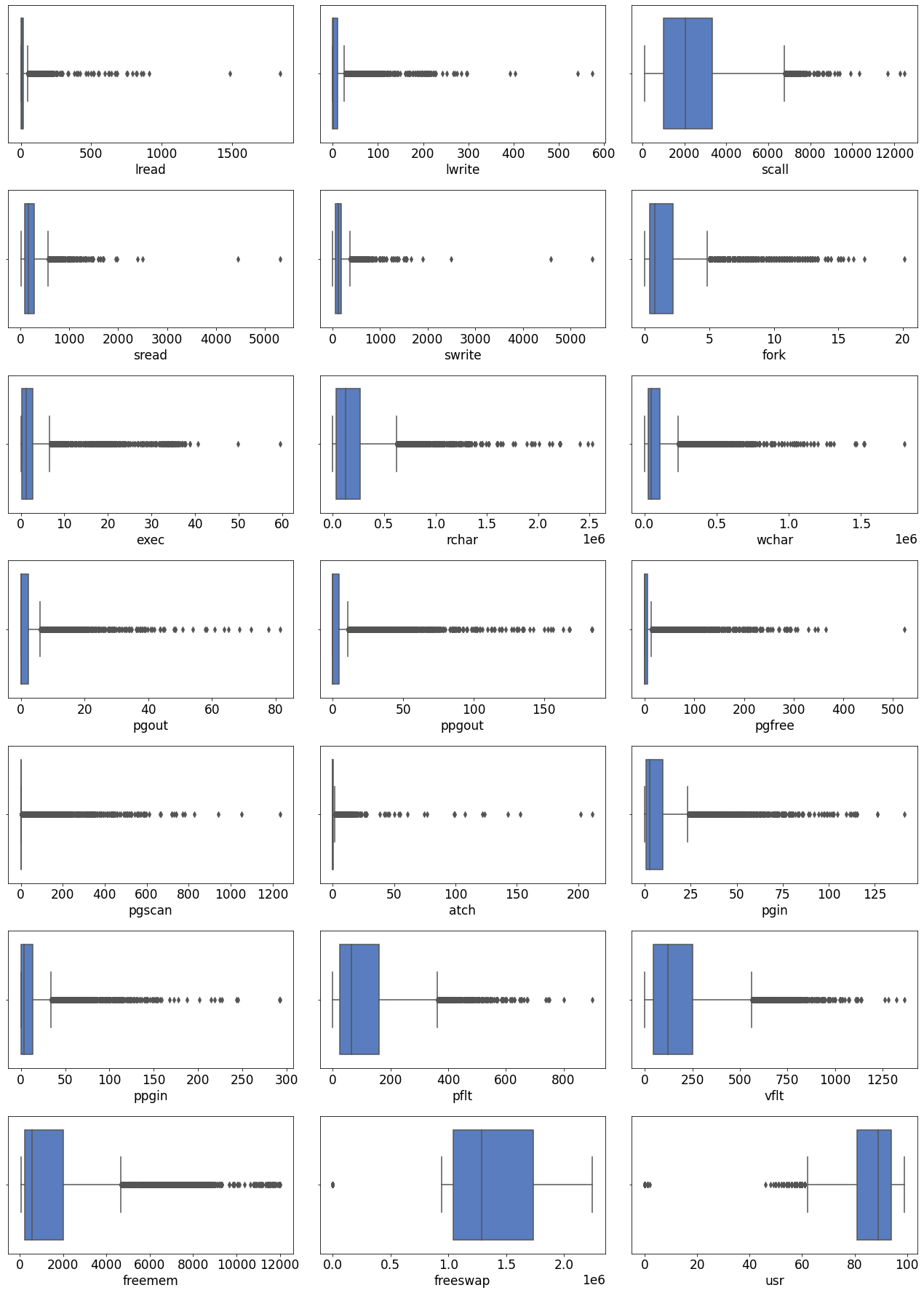
****

Fig. 1: Boxplots of all continuous variables

From the boxplots above, we can see that

1. There are outliers present in all the columns.
2. For the columns lread, lwrite, sread, swrite, fork, exec, pgout, ppgout,pgfree,pgscan,atch, pgin and ppgin, the median is very close to zero. This tells us that there are a lot of zeros present in these columns.
3. Outliers on the lowers side are present in only two columns – freeswap and usr.
4. Since linear regression algorithm is sensitive to outliers, we will treat these in the later stage to get better results from the model.

Now, we will plot histograms of these continuous variables to get an idea of their distribution.

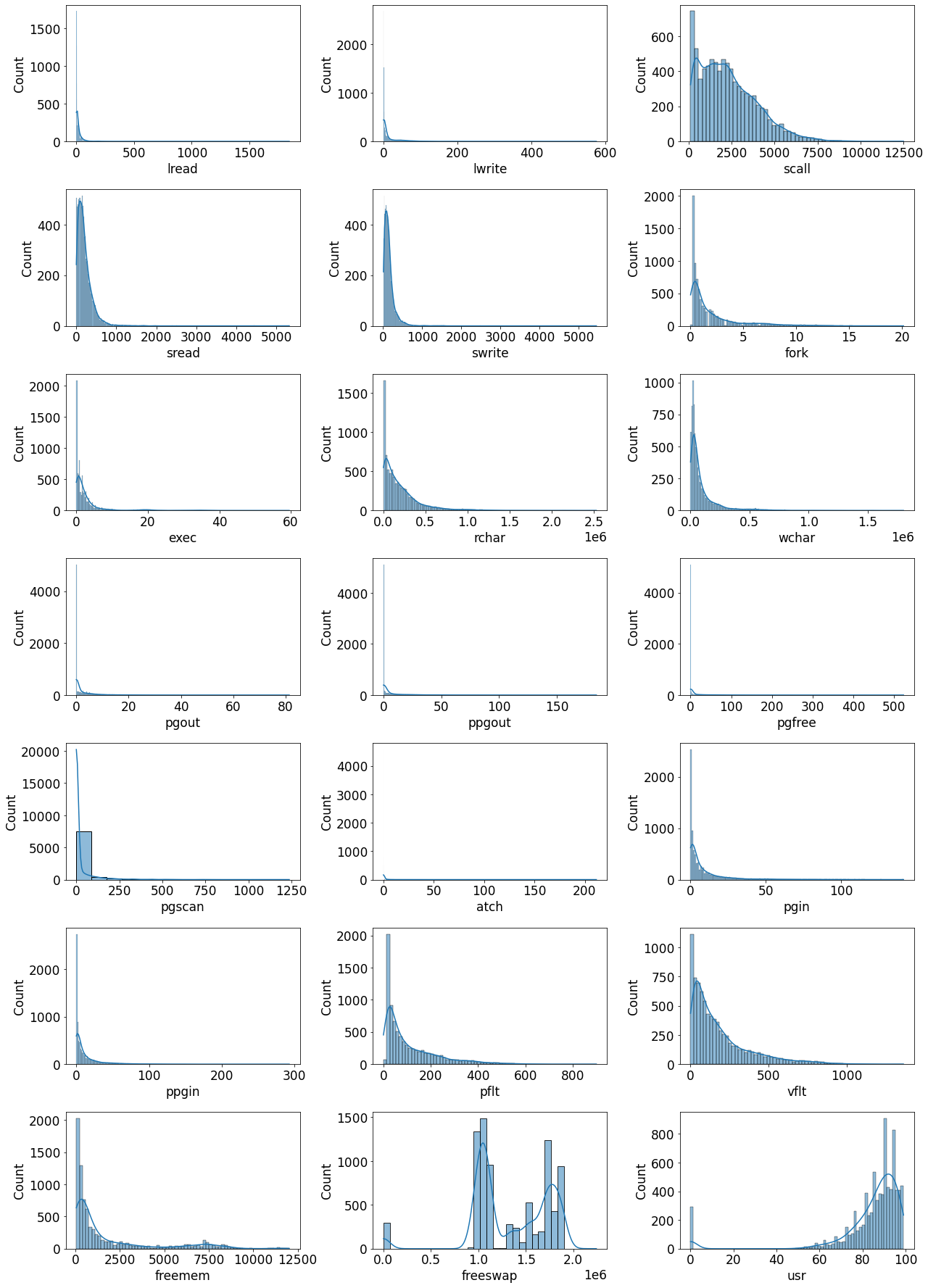
****

Fig. 2: Histograms of all continuous variables.

We observe the following:

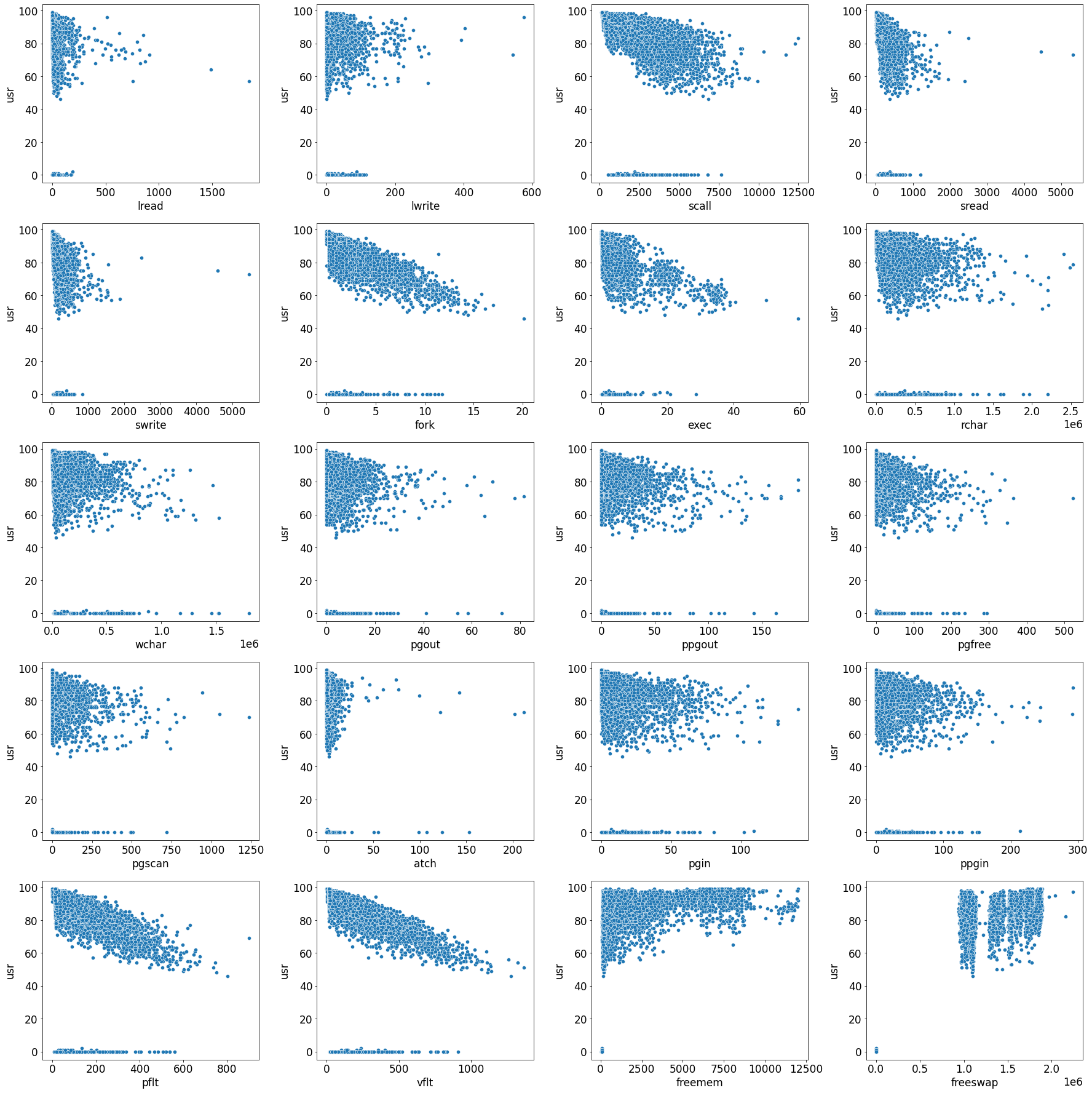
1. We can barely see a thin straight line at 0 in the columns lread, lwrite, pgout, ppgout, pgfree, atch, pgin, ppgin. This indicates that zero values are dominating these columns.
2. The remaining columns scall, sread, swrite, fork, exec, rchar, wchar, pgin, ppgin, pflt, vflt and freemem are all right skewed.
3. The column freeswap is bimodal. We can observe a tiny bump near 0, which is because of the zero values.
4. The column usr is left skewed. Here also we observe a tiny bump near 0, which is again because of the zeros present in the column.

Now, we shall move on to the categorical variable.



Fig. 3: Count plot of the categorical variable.

There are 52.86% of the data which is not CPU bound. The remaining 47.14% of the data is CPU bound.

****

Scatter plots of the continuous variables v/s the target variable.

From the above scatter plot, a linear relation between scall v/s usr, fork v/s usr, exec v/s usr, pflt v/s usr and vflt v/s usr is very evident. Even though the scatter plot of other variables v/s usr appears clouded, it is too soon to pass any judgement. We shall investigate further and come to a conclusion.

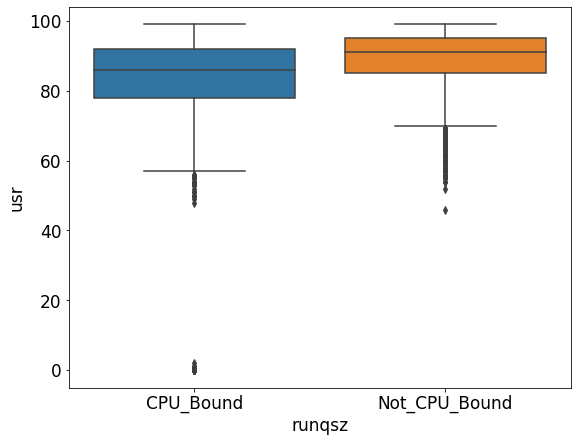


Fig. 4: Boxplot of categorical variable v/s the target variable.

We see that the amount of time the CPU runs in user mode is more when the run queue size is not CPU bound. When the run queue size is not CPU bound we may observe that the amount of time CPU runs in user mode has never become zero.

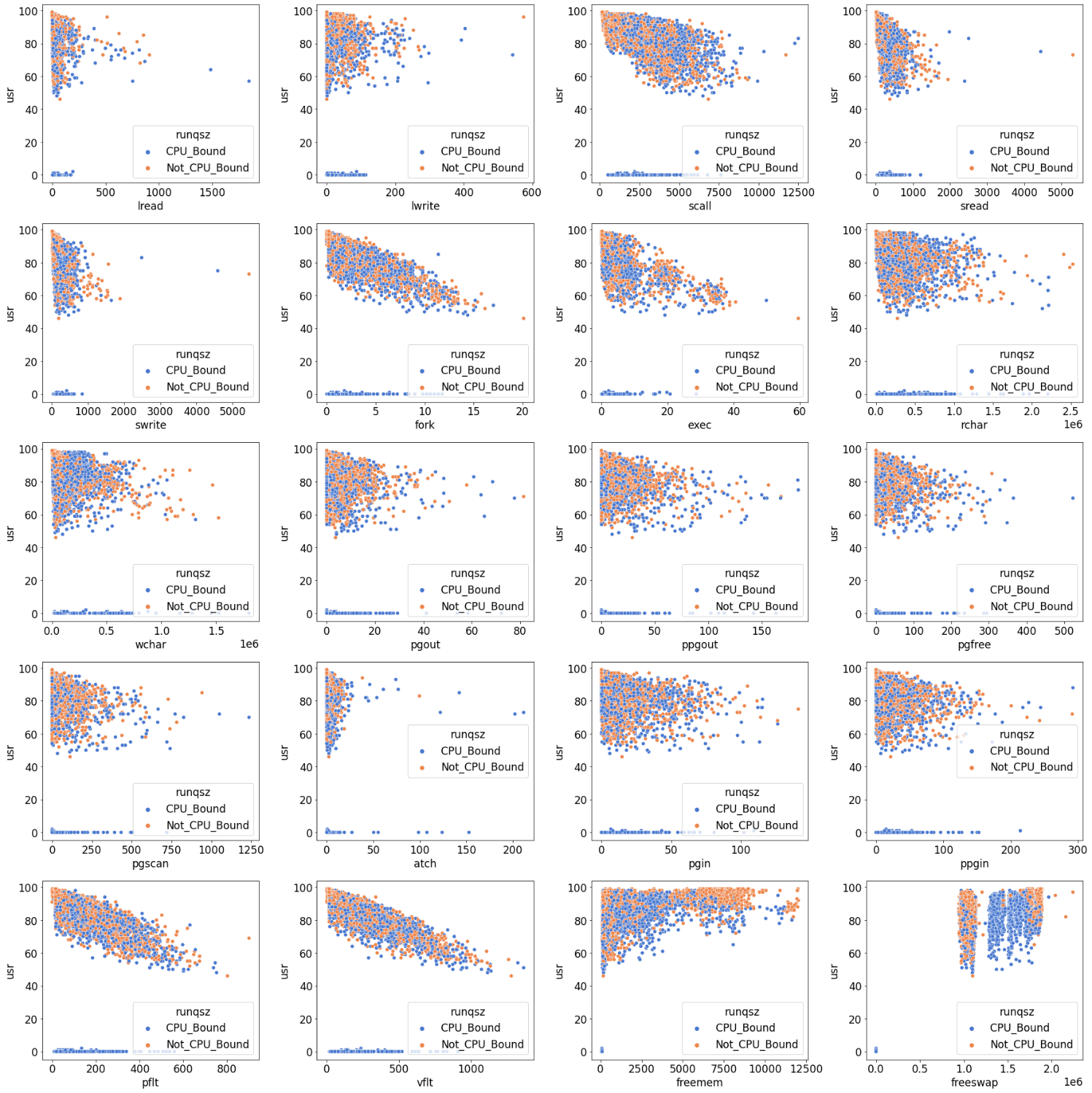
****

Fig. 5: Scatterplots of all variables v/s the target variable.

We do not see any clear bifurcation happening from the categorical column. We shall see the heatmap to understand the correlation among the variables.

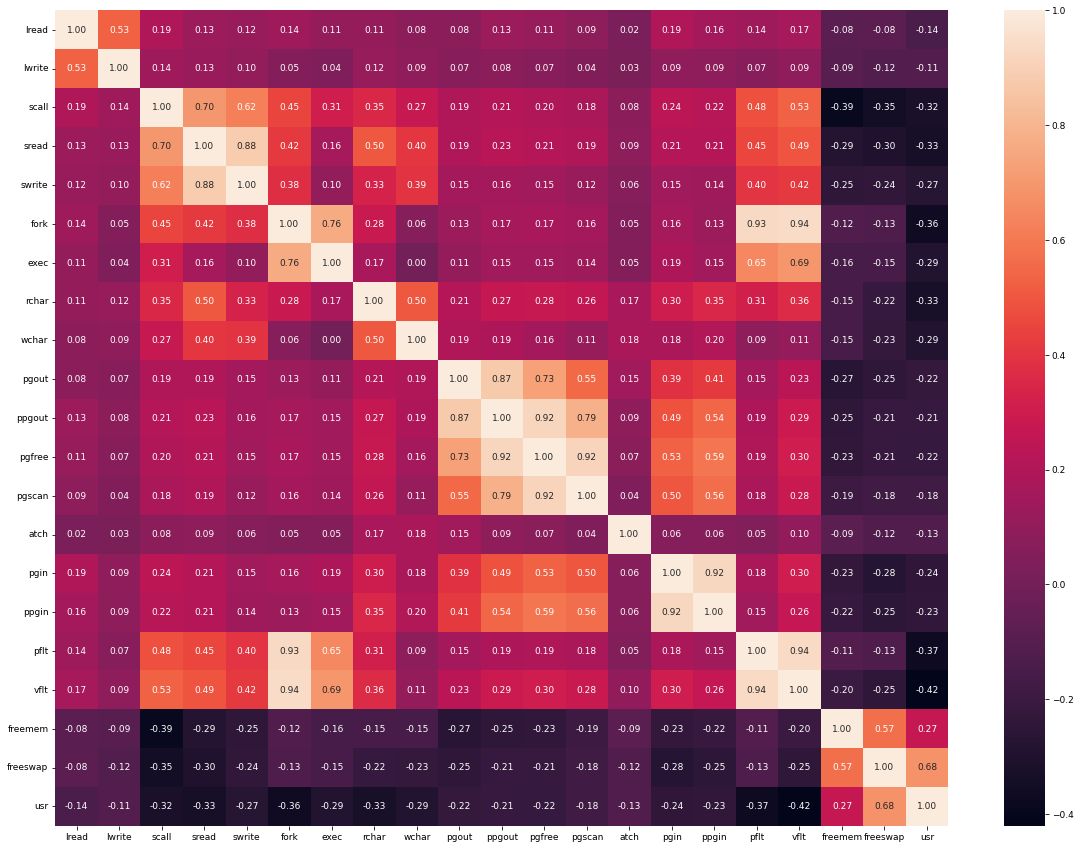


Fig. 6: Heatmap.

We clearly see very strong correlation(more than 0.85) among the following variables:

1. swrite and sread
2. ppgout and pgout
3. pgfree and ppgout
4. pgscan and pgfree
5. ppgin and pgin
6. pflt and fork
7. vflt and fork
8. vflt and pflt.
   1. **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.**

**Null values:**

There were null values present in the columns rchar and wchar. Since there are outliers in both the columns, we will impute the null values using the median of the respective columns.

**Zero values:**

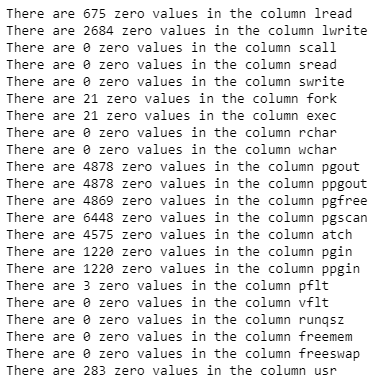


Table 5: Zero values present in different columns.

As we can see there are zero values are present in lread, lwrite, fork, exec, pgout, ppgout, pgfree, pgscan, atch, pgin, ppgin, pflt, usr.

Pgscan has the highest number of zero values (78.7%)

We think these values are not bad data or anomalies since when the system is not in use, these variables could take the value of 0. Hence, we are not going to treat the zeros in the data.

**Possibility of creating new variables:**

We don’t see any scope to create any new variables as of now.

**Duplicates:**

As seen earlier, there are no duplicates in the data.

**Outliers:**

As seen in the univariate analysis, there are outliers present in every column. Since the linear regression algorithm is sensitive to outliers, we will treat them using the boxplot method.

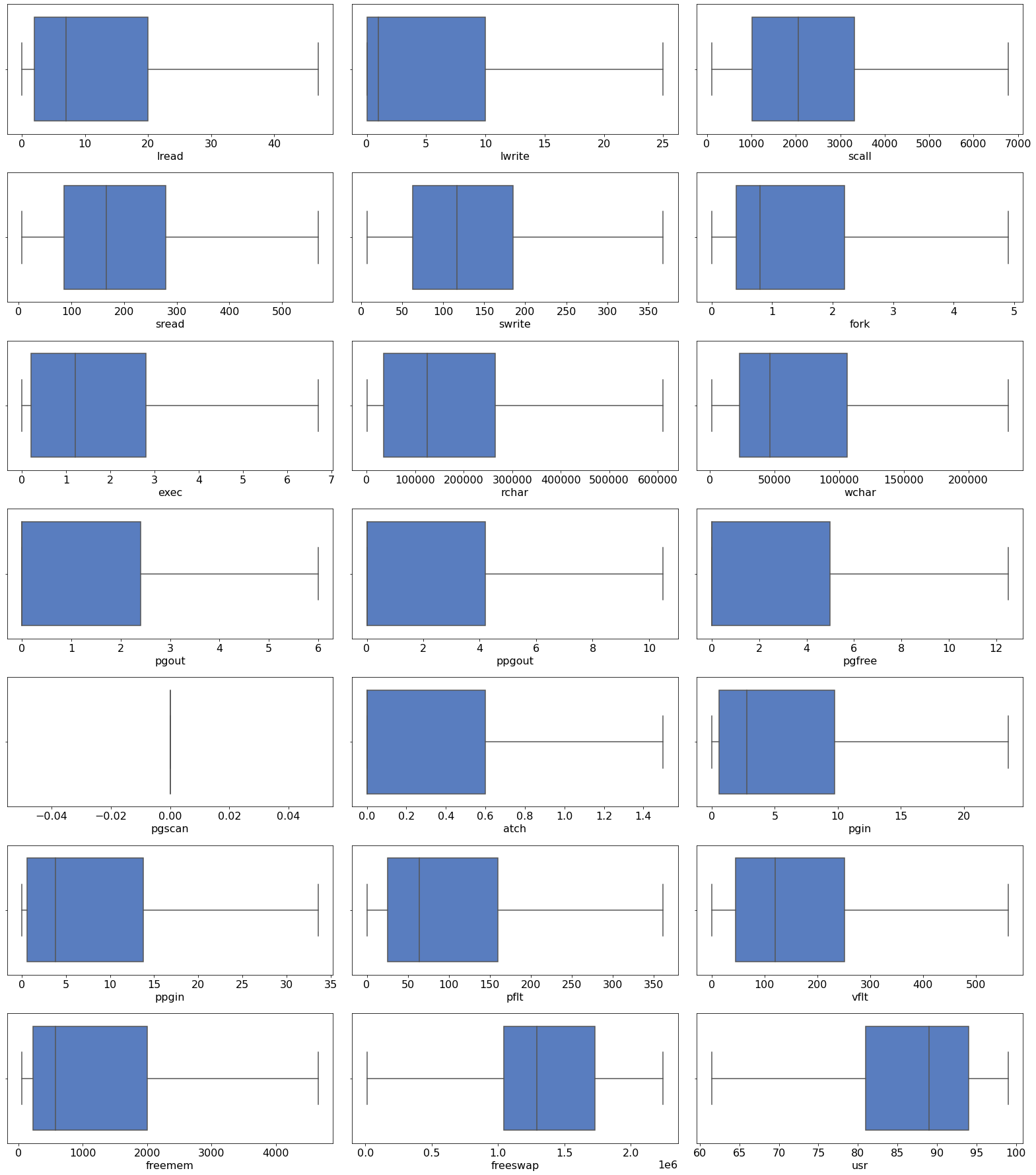


Fig. 7: Boxplots after outlier treatment.

From the above figure we see that all the values in the **pgscan** have been pushed to 0 after the outlier treatment. Hence, **we will drop this variable and proceed further.**

* 1. **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.**

**Encoding the data:**

The column runqsz is the only categorical variable. Since this is a nominal variable, we will perform one hot encoding using the get dummies of the pandas package.

**Splitting the data into test and train:**

The independent variables and the dependent variables are separated. These are then split into training and testing data in the ratio 70:30.

The training data contains 5734 records and the test data contains 2458 records.

**Linear Regression:**

We use both the libraries scikit learn and statsmodel to build models.

**Model 1:**

We have used stats model to build this model. All the independent variables are taken into account to build this model.

The summary of the model is given below.

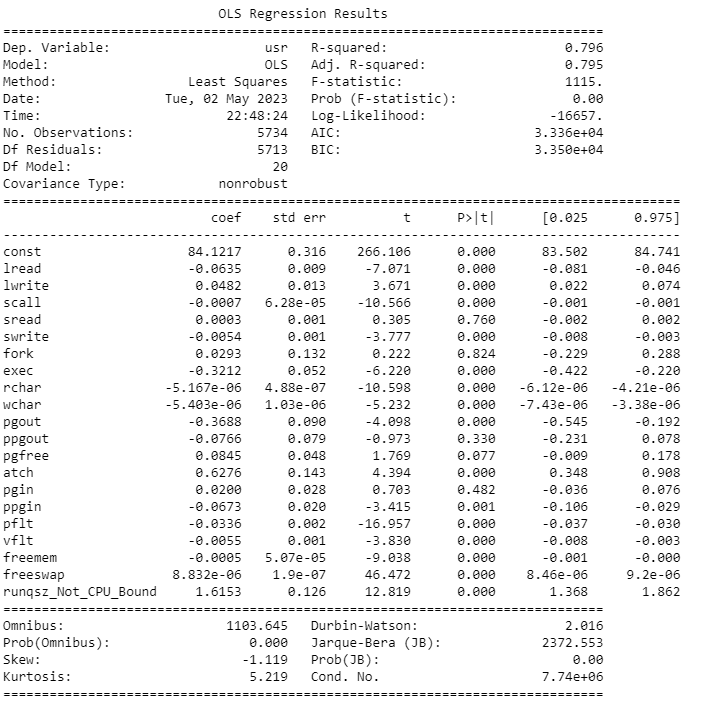


Table 6: Regression results of model 1.

We had seen multi collinearity among various variables in the heatmap already. Here, we will check VIF scores to identify the multicollinearity and we address this issue by removing those variables.

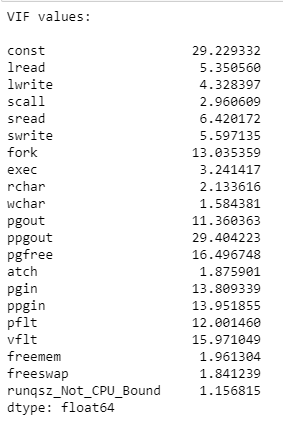


Table 7: VIF values of variables in model 1.

Here, we are considering VIF values to be high if it is more than 5.

We are dropping one variable at a time and checking for the value of R-square and adjusted R-square. If these scores drop considerably, then the particular variable that was dropped in the model is considered an important variable.

To check the drop in R-square and adjusted R-square, we are building the model in **scikit learn** library.

Making use of VIF scores and the heatmap, we are going to remove the variables sread, pgout, ppgout and build our next model.

**Model 2:** The summary of this model is given below. We see that even after dropping the above three variables, the value of R-square and adjusted R-square are still comparable.

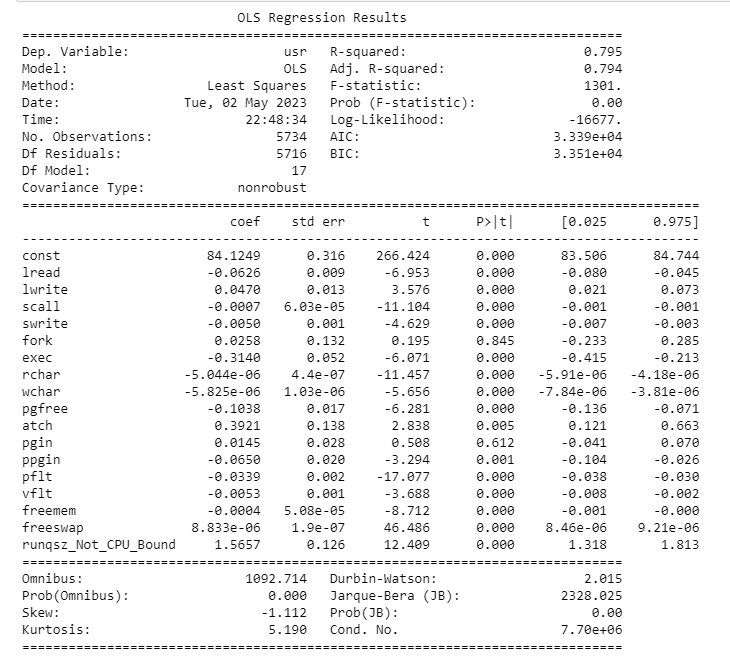
****

Table 8: Regression results of model 2.

Now we will check again for the VIF scores of the variables in this model.

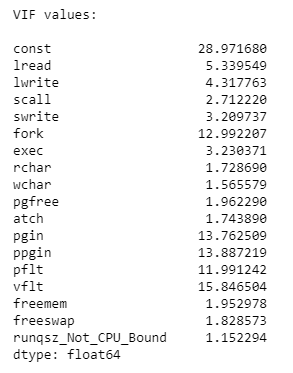


Table 9: VIF values of variables in model 2.

We still see that multicollinearity is present in the model. We again follow the above-mentioned procedure to identify important variables. At the end of this procedure we come to the conclusion to drop the variables fork, vflt,ppgin in the next model.

**Model 3:**

The summary is given below.

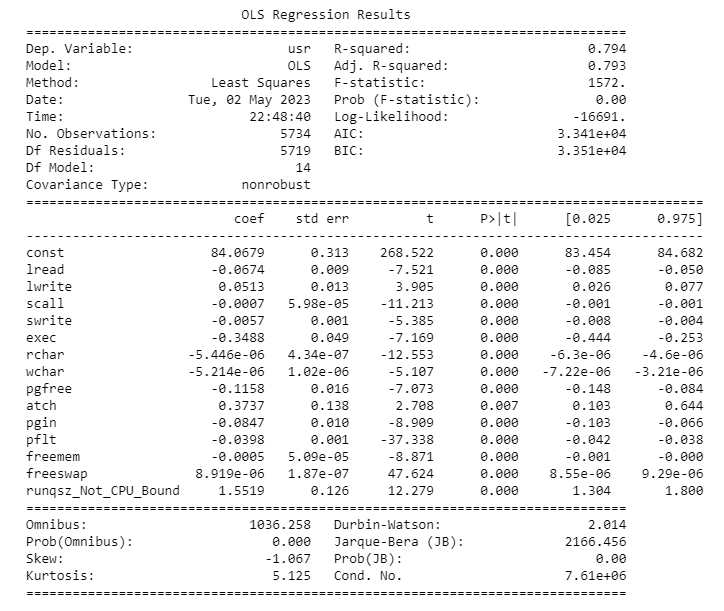


Table 10: Regression results of model 3.

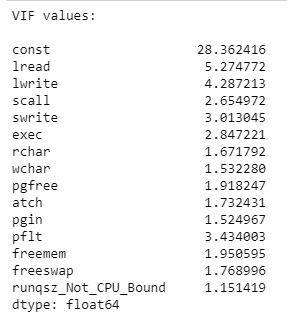


Table 11: VIF values of variables in model 3.

We see there is only one variable, lread, whose VIF value is greater than 5. We shall remove this variable in our next model.

**Model 4:**

This is our final model. The summary is given below.

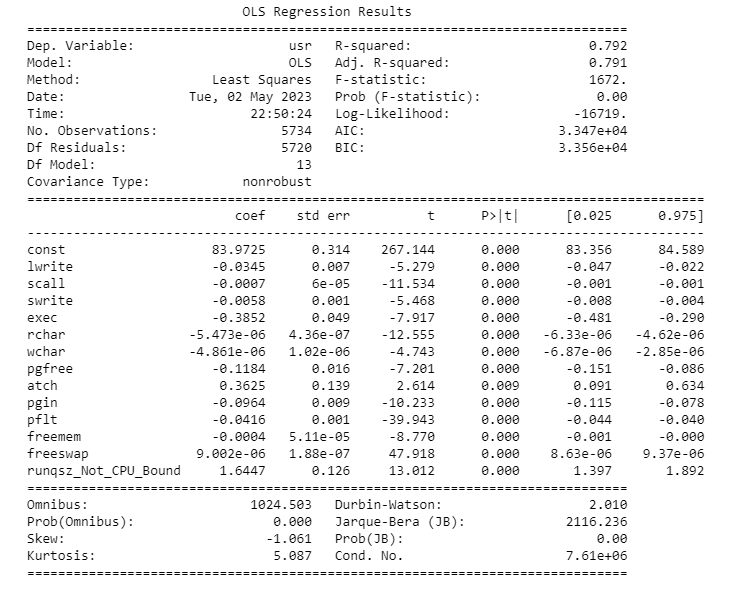


Table 12: Regression results of model 3.

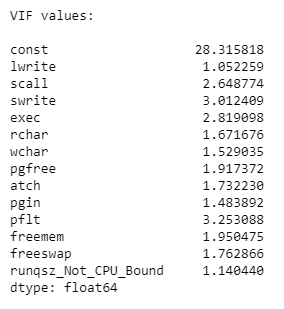
****

Table 13: VIF values of variables in model 4.

The VIF values in this model are all less than 5. This indicates that there is no multi collinearity present in our model. Even though we see the R-square value dropping by 0.002, we take a call to eliminate the variable.

Now that our model has no multi collinearity, we can interpret the p-values.

If the p-values are greater than 0.05, the variable is not significant. If it is less than 0.05, the variable is significant. In our model we see that p-values for all the variables are 0, which tells us that all the variables that we have used to build the model are significant.

The below table will summarize all the models we built so far. The values mentioned are calculated on train data.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **R-square** | **Adj. R-square** | **RMSE** | **Multicollinearity** |
| Model 1 | 0.796 | 0.795 | 4.419 | Present |
| Model 2 | 0.795 | 0.794 | 4.435 | Present |
| Model 3 | 0.794 | 0.793 | 4.445 | Present |
| Model 4 | 0.792 | 0.791 | 4.467 | Absent |

Table 14: Comparison of all models.

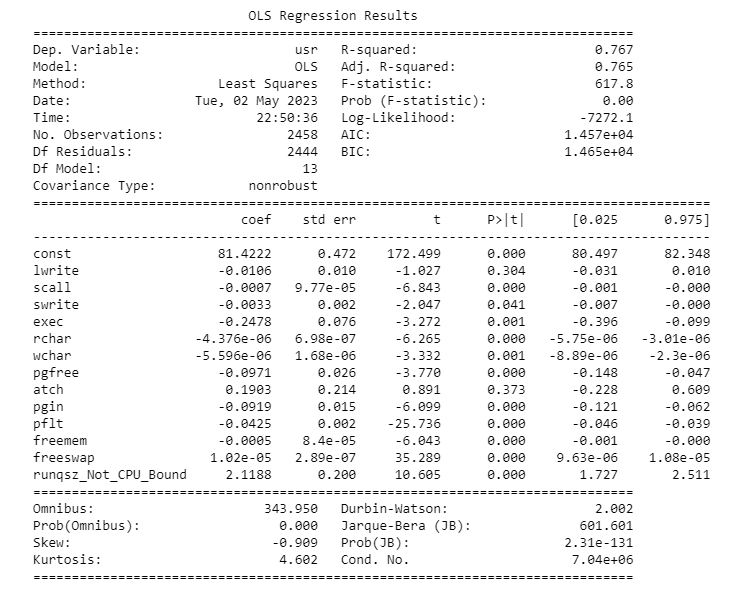
****

Table 15: Regression results of Test data on model 4.

The R-square for the test data is 0.767. A model is said to be overfitting when the regression results on the test data vary by 10% of the values of the train data. So, even though the R-square value of the test data is slightly less than that of train data, we cannot say this model is overfitting.

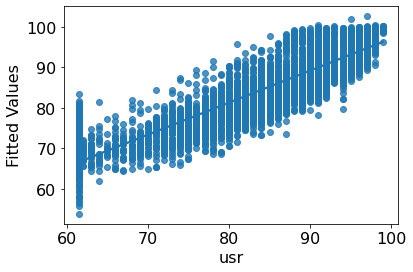


Fig. 8: Actual values v/s fitted values

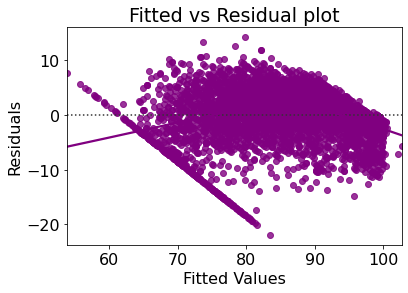
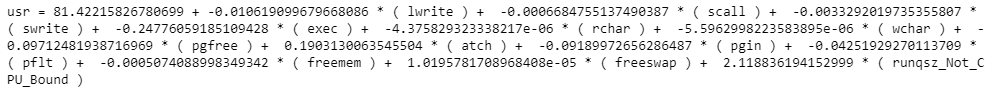
****

Fig. 9: Residuals v/s fitted values.

The residuals are not following any pattern; hence we can say the model is linear. Hence, we conclude that Model 4 is the best model with no multicollinearity and acceptable R-square values.

**Final Equation:**

****

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

We have concluded that model 4 is the best model to predict the amount of time the CPU runs in user mode. There are many features which are highly correlated to each other.

We see negative correlation between the user time and fork, pflt and vflt. If one of them decreases the other one increases. From our final model we see that exec variables carries the maximum weight. Atch variables is also important.

Recommendations:

The company needs to hire a subject matter expert who can explain these variables.

If values of exec variables are increased, the cpu time that runs in user mode will decrease.

On the other hand, if the values of atch are increased, the user time will increase.

**Part -2**

**Logistic Regression / LDA / CART**

**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 data has been read.

1. Data has 1473 records and 10 columns.
2. Out of 10 columns, two are float, one is integer and remaining seven are object data types.
3. The data has no unique identifier.
4. Out of the 7 categorical data, there are 4 ordinal datatypes; Wife education, Husband education, standard of living index and media exposure (Since it is given in the data dictionary that the media exposure is either good or not good, we consider this to be ordinal.)
5. There are null values in the columns Wife\_age and No\_of\_children\_born.
6. There are 80 duplicates present in the data.

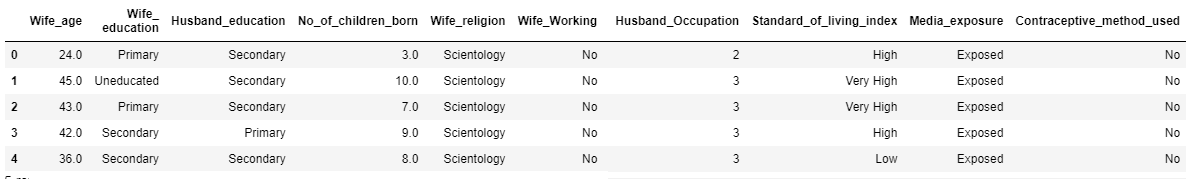


Table 16: First five records of the dataset.

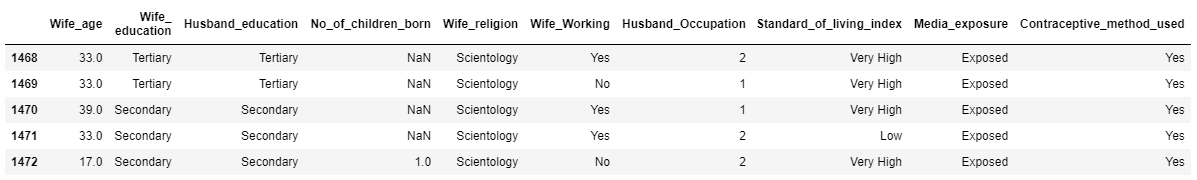


Table 17: Last five records of the dataset.

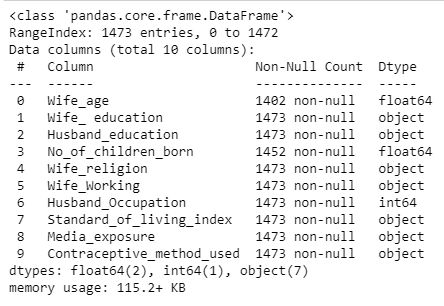
****

Table 18: Information of the dataset

.

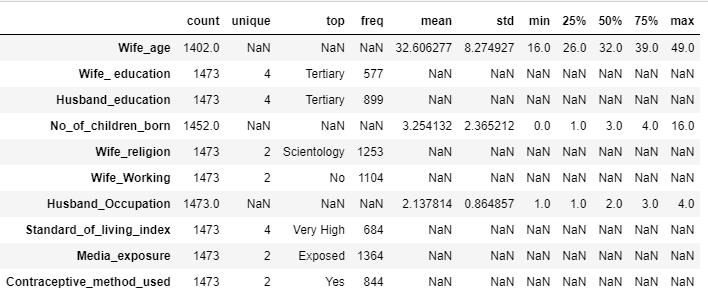


Table 19: Statistical description of the dataset.

Different categories in the categorical columns are given below:

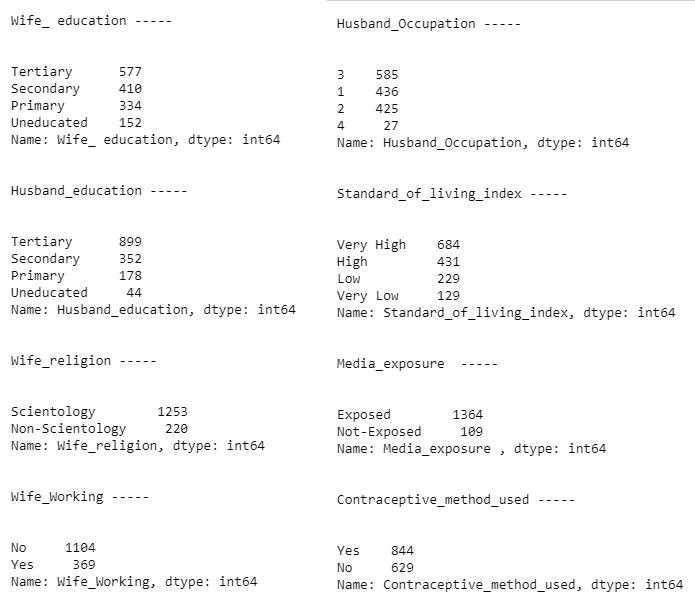


Table 20: Different categories in the categorical variables.

**Null values:**

There are 71 null values in the column Wife age and 21 null values in the column No of children born. We are treating the missing values with the median of the same column.

**Duplicates:**

There are 80 duplicate rows present in the data. The data has no unique identifier. And since the columns are generic in nature, there might be really 80 people who are of same age, whose educations are same, etc., in the whole country of Indonesia. Considerng this fact, we will not treat our data for duplicate values.

**Outliers:**

There are outliers present in the number of children born. Since these outliers are not anomalies but real data, we take a call to keep them without any treatment and proceed further.

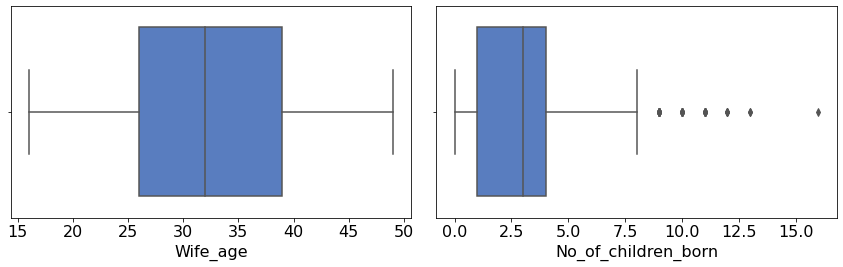


Fig. 10: Boxplots of numerical variables.

**Univariate / Bivariate and multivariate Analysis:**

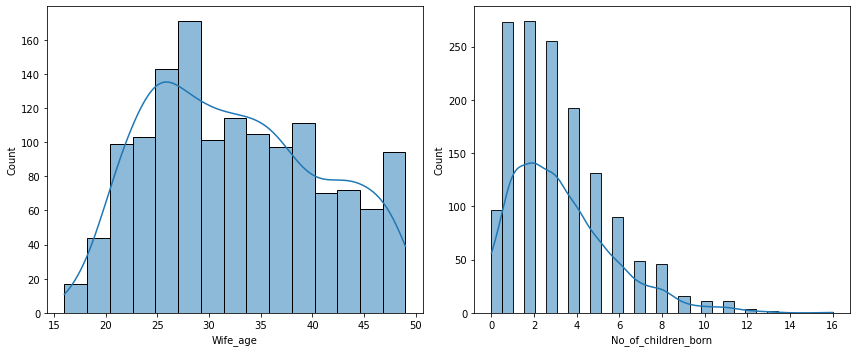
****

Fig. 11: Histograms of numerical variables.

The distributions are skewed. The reported age of the wife is from 15 to 50. Most number of women have had 1 or 2 pregnancies reported. Less number of people have more than 3 pregnancies.

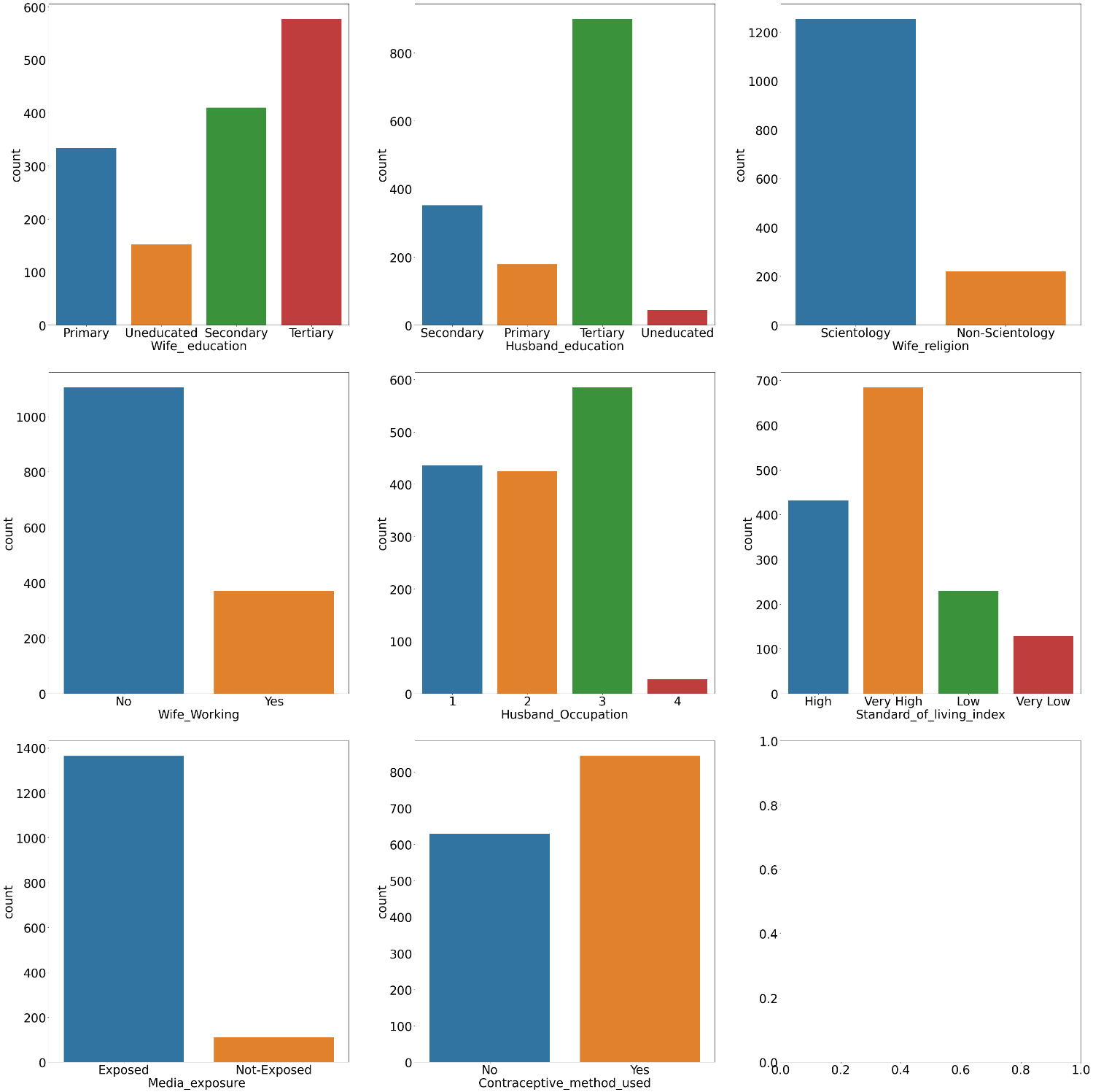


Fig. 12: Count plots of categorical variables.

As we see, the number of people who are using the contraceptive methods are higher. The data contains more number of people who have a very high standard of living index.

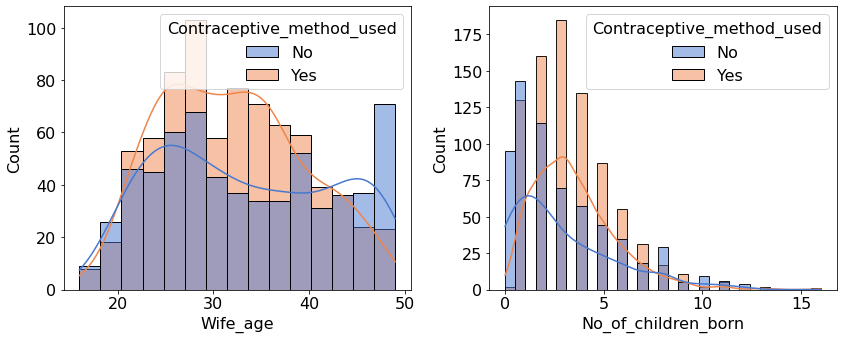
****

Fig. 13: Histograms depicting the poor predictors.

As we can see from the kernel density functions on the histograms, they are overlapping on each other, hence both the numerical variables are going to be poor predictors of the target variable.

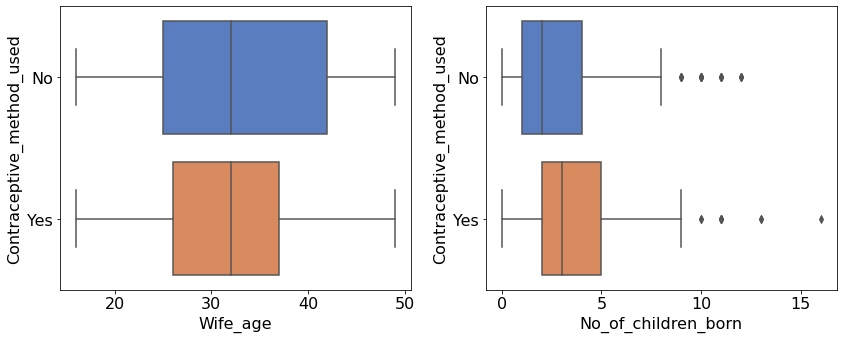
****

Fig. 14: Box plots of the numerical variable.

The average age of wives with and without contraceptives are same. From the second boxplot, we may say people who have more children are more likely to use contraceptives.

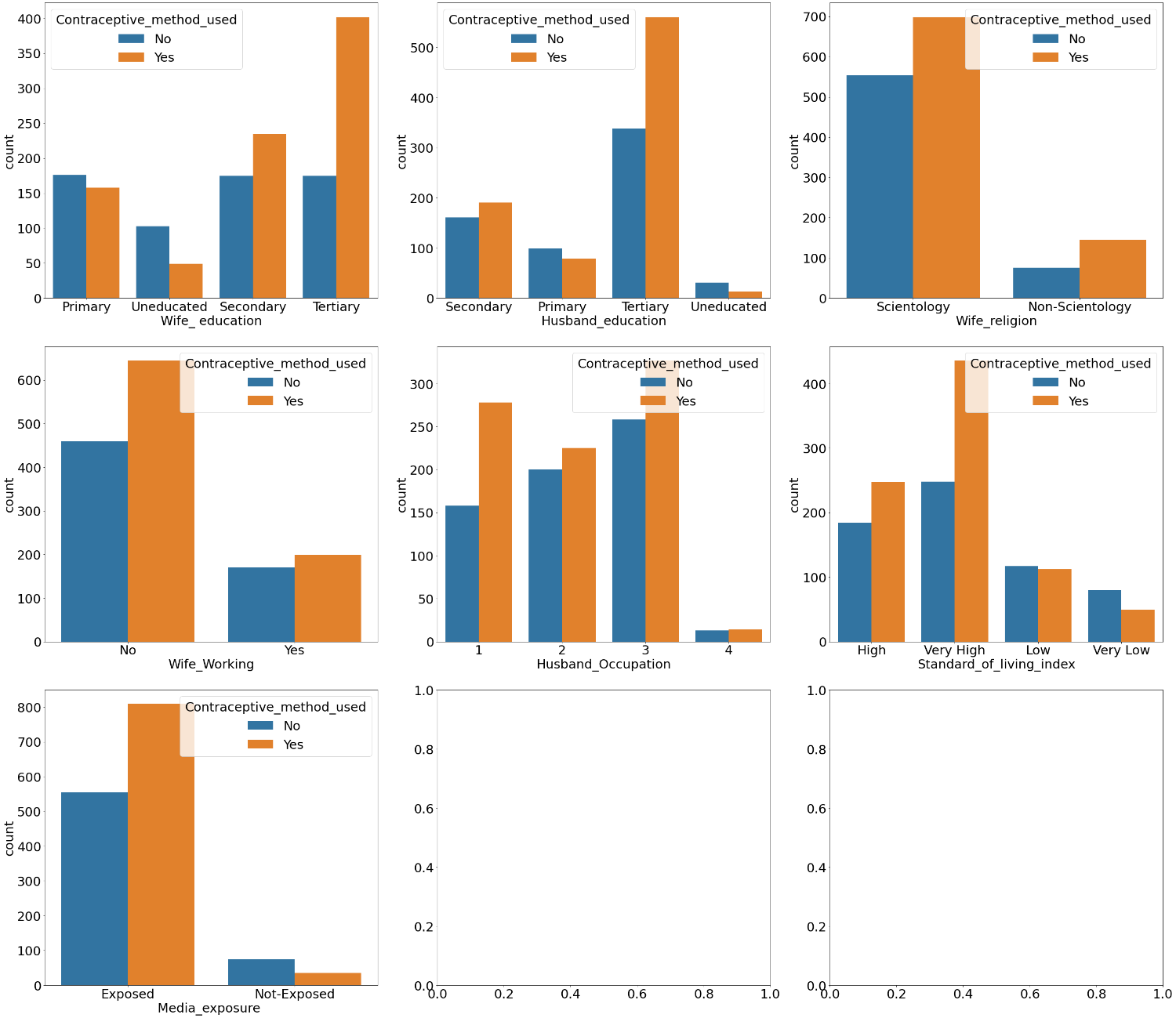
****

Fig. 15: Bivariate analysis using countplots.

People who have a very high and high standard of living are using more contraceptives. Also ladies who had higher education are aware of contraception. People who are not exposed to media are less likely to use contraceptives.

**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.**

As mentioned, we are not scaling the data. We are encoding the data which has string values depending on whether the column is an ordinal or nominal categorical column.

**Ordinal column: Wife’s Education**

Values encoded: Tertiary = 4, Secondary = 3, Primary = 2, Uneducated = 1.

**Ordinal column: Husband’s Education**

Values encoded: Tertiary = 4, Secondary = 3, Primary = 2, Uneducated = 1.

**Ordinal column: Standard of living index**

Values encoded: Very high = 4, High = 3, Low = 2, Very Low = 1.

**Ordinal column: Media Exposure** (Since it is given in the data dictionary that the values are good or not good, we are considering this column to be ordinal and encoding it accordingly.)

Values encoded: Exposed = 1, Not exposed =0.

**Target column: Contraceptive method used**

Values encoded: Yes = 1, No = 0.

**Nominal categorical columns:**

Wife’s religion and wife working are one hot encoded using the dummy variables. Since these two columns are binary in nature, the one hot encoding will not add more dimensions to the data.

The independent and dependent variables are separated and split into train and test sets in the ratio 70:30. The train set contains 1031 records and the test set contains 442 records.

Logistic regression, Linear Discriminant Analysis and CART models are applied on the data.

The performance report are as follows.

**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.**

**Logistic Regression**

**Accuracy:**

Training data: 0.6876818622696411

Testing data: 0.6538461538461539

**Classification Report:**

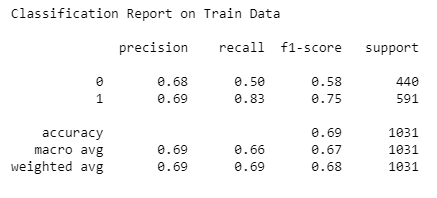
****

Table 21: Report on Train data – Logistic Regression

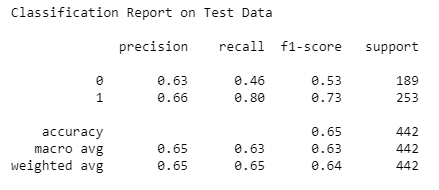


Table 22: Report on Test data – Logistic Regression

The recall and precision for class 1 is pretty good in both train and test data. Even though the model accuracy is less for test data, it is not overfitting.

**Confusion Matrix:**

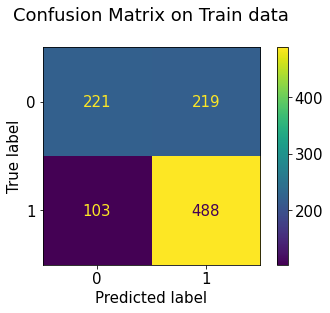
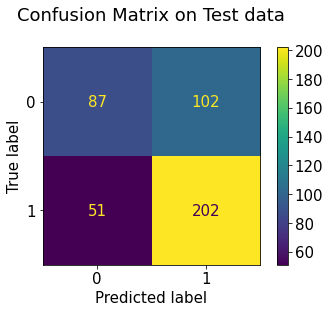
** **

Fig. 16: Confusion Matrix – Logistic Regression.

**Area under the curve:**

Training Data: 0.714

Testing Data: 0.714

**ROC Curve:**

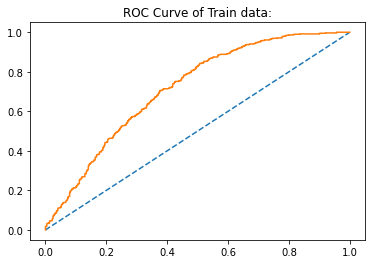
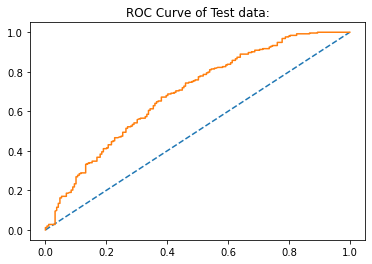
** **

Fig. 17: ROC Curve – Logistic Regression.

In an ideal case, we would expect the ROC curve to pass through the point (0,1). But in real cases, it’s highly unlikely. This curve covers 71.4% of the total area under the curve. We mays say this model is doing a pretty good job.

**Linear Discriminant Analysis**

**Accuracy:**

Training data: 0.6867119301648884

Testing data: 0.6561085972850679

**Classification Report:**

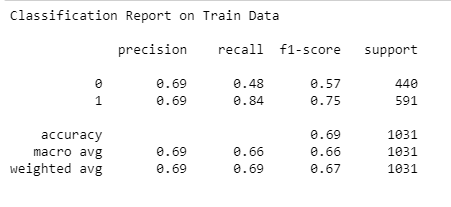
****

Table 23: Report on Train data – LDA

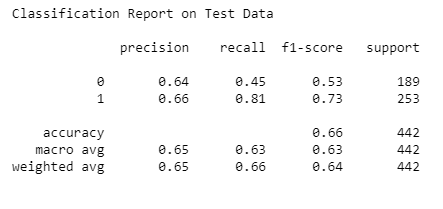


Table 24: Report on Test data – LDA

The accuracy scores, precision and recall are similar to that of the logistic regression model. We will check the confusion matrix, ROC curve to see how the model is performing.

**Confusion Matrix:**

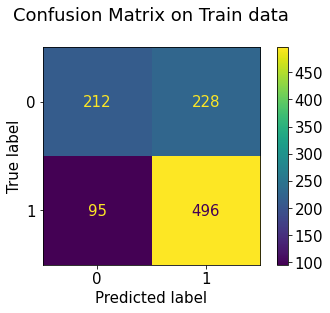
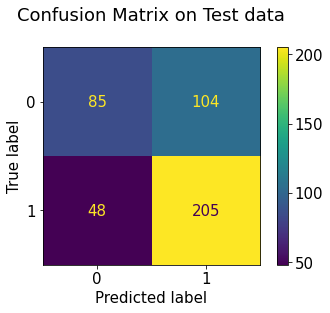
 

Fig. 18: Confusion Matrix – LDA

**Area under the curve:**

Training Data: 0.714

Testing Data: 0.714

**ROC Curve:**

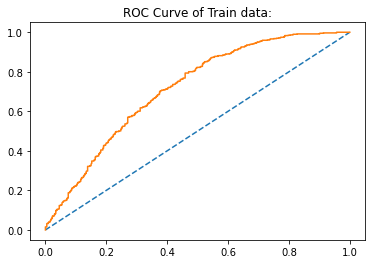
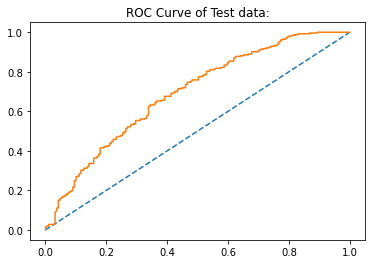
** **

Fig. 19: ROC Curve – LDA

The ROC curve from the LDA model also covers 71.4% of the area under the curve. From the confusion matrix and the ROC and AUC, we can say that LDA is performing as good as Logistic regression model.

**Classification and Regression Tree**

**Accuracy:**

Training data: 0.7507274490785645

Testing data: 0.6719457013574661

**Classification Report:**

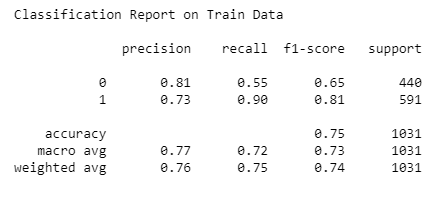
****

Table 25: Report on Train data – CART

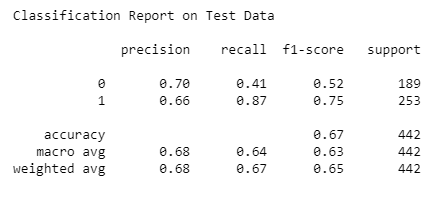


Table 26: Report on Test data – CART

We see that accuracy of the model in train data is 75.2%, but in the test data, the accuracy has dropped to 67.1%, which is a 8.1% huge drop.

**Confusion Matrix:**

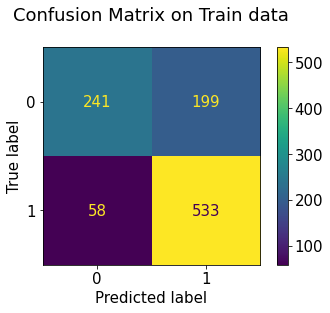
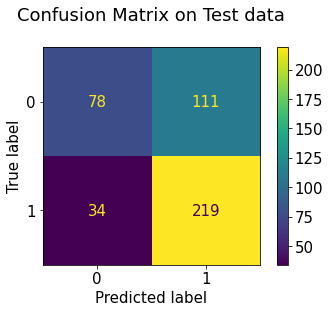
** **

Fig. 20: Confusion Matrix – CART.

**Area under the curve:**

Training Data: 0.816

Testing Data: 0.816

**ROC Curve:**

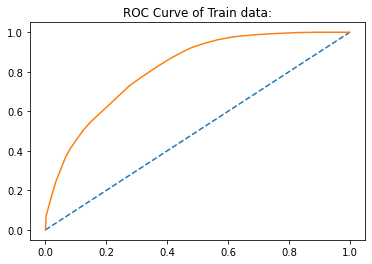
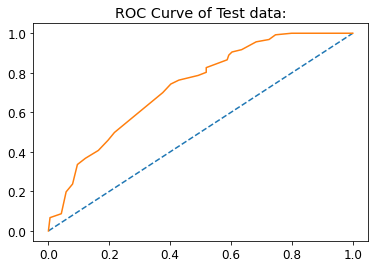
** **

Fig. 21: ROC Curve –CART.

The area under the curve for both the train and test data are 81.6%. This is a better number compared to the previous models.

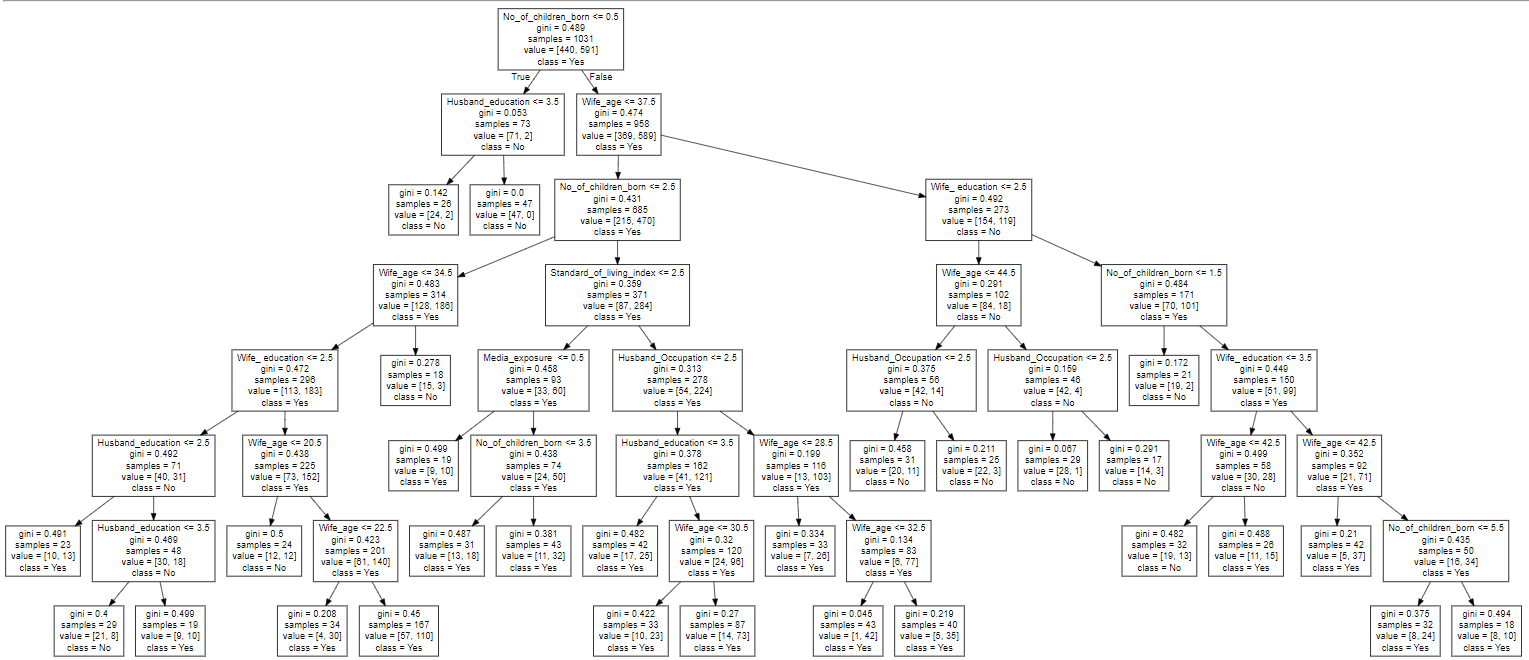


Fig. 22: Decision tree.

On comparing the three models we built, logistic regression model and the LDA models looks to be doing a good job. When we look at the accuracy scores of the model, we get an impression that CART model is overfitting. Also, since the data is not highly imbalanced (57% class 1 and 43% class 0 in the target variable), accuracy scores are trust worthy. Hence, Logistic regression and LDA models are doing a better job compared to CART.

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

Our aim was to predict if a woman is using the contraceptives or not, using the independent variables, such as the women’s age, education, standard of living, husband’s education, etc. From the EDA, we see that women who have more education are more likely to have been using a contraceptive and the uneducated people women are not very aware of this. Most of the women who have very high standard of living are more likely to have a contraception. People who have been not exposed to any media are likely to be unaware of this.

Before we built the model, we encoded the data on the basis of ordinal and nominal categories. On comparing the models, we suggest logistic regression and LDA models are performing better compared to the CART model.

Recommendations:

The government may conduct programs to create awareness in those women who already have had a specific number of children. Mostly uneducated women and men are to be considered since the use of contraceptives is very less in this category.