Business Report

Lavanya Sreeram

DSBA, Nov202 Batch

Business Report

Predictive Modeling



Contents

[List Of Figures: 2](#_Toc130867978)

[List of Tables: 2](#_Toc130867979)

[Data and Data Dictionary: 3](#_Toc130867980)

[**1** **Problem 1: Linear Regression** 4](#_Toc130867981)

[**2** **Problem 2: Logistic Regression, LDA, CART** 9](#_Toc130867982)

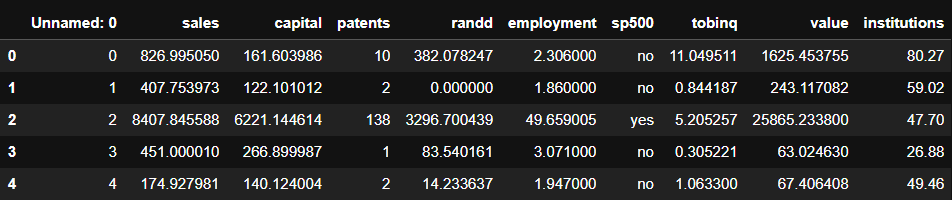
# List Of Figures:

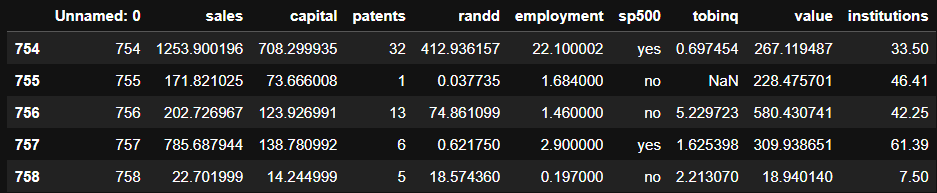
|  |  |  |  |
| --- | --- | --- | --- |
| **S No** | **Name** | **Description** | **Source** |
| 1 | Fig1 | Cover Page Picture | https://www.techrepublic.com/article/types-of-predictive-modeling/ |
| 2 | Fig2 | Linear Regression Data Dictionary | Provided Dataset. |
| 3 | Fig3 | Logistic Regression, LDA, CART Data Dictionary | Provided Dataset. |
| 4 | Fig1.1.1 | Linear Regression: Dataset Head | EDA |
| 5 | Fig1.1.2 | Linear Regression: Dataset Info | EDA |
| 6 | Fig1.1.3 | Linear Regression: Dataset Description | EDA |
|  | Fig1.1.4 | Linear Regression: pairplot | EDA |
| 7 | Fig1.2.1 | Boxplot of 'rchar' | Data Preparation |
| 8 | Fig1.2.2 | Boxplot of 'wchar' | Data Preparation |
| 9 | Fig1.2.3 | Linear regression: Outliers | Data Preparation |
| 10 | Fig1.3.1 | Linear Regression:Encoding 'runqsz' | Modeling |
| 11 | Fig1.3.2 | Linear Regression: Train Dataset Head | Modeling |
| 12 | Fig1.3.3 | Linear Regression: Test Dataset Head | Modeling |

# List of Tables:

# **Problem 1: Linear Regression**

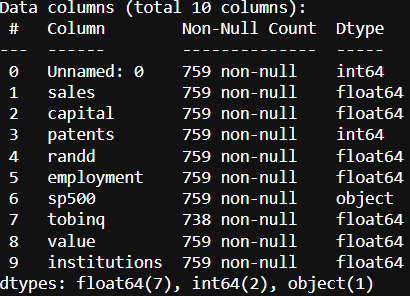
# Read the data and do exploratory data analysis. Describe the data briefly. (Check the null values, data types, shape, EDA). Perform Univariate and Bivariate Analysis. (8 marks). Answer: [Firm\_level\_data.csv](https://olympus.mygreatlearning.com/courses/87096/files/7893754/download?wrap=1" \t "_blank" \o "Firm_level_data.csv)[View in a new window](https://olympus.mygreatlearning.com/courses/87096/files/7893754/download?wrap=1) data set is provided. Data Dictionary for Firm\_level\_data is as below: 1. sales: Sales (in millions of dollars). 2. capital: Net stock of property, plant, and equipment. 3. patents: Granted patents. 4. randd: R&D stock (in millions of dollars). 5. employment: Employment (in 1000s). 6. sp500: Membership of firms in the S&P 500 index. S&P is a stock market index that measures the stock performance of 500 large companies listed on stock exchanges in the United States 7. tobinq: Tobin's q (also known as q ratio and Kaldor's v) is the ratio between a physical asset's market value and its replacement value. 8. value: Stock market value. 9. institutions: Proportion of stock owned by institutions

The Dataset has 759 rows, 10 columns. ‘sales’ is the Target Variable. The below figure shows the dataset first 5 observations (Fig1.1.1).  


The dataset sample observations from bottom are as shown below (Fig1.1.2).  


On first look, it can be inferred that the feature ‘Unnamed: 0’ has to be dropped. Data is not scaled. Drop ‘Unnamed:0’ feature. Now, the dataset top 5 observations are as shown below (Fig 1.1.3):  
A picture containing screenshot, text

Description automatically generated

Except ‘sp500’, all of the features in the data are numeric in nature ('int64' or 'float64' type). ‘sp500’ is object type, and as per data dictionary the feature is binary. The below figure shows the datatypes of all features of the dataset (Fig1.1.4).  


The statistical data descriptions is as shown in below figure.(Fig1.1.5)  
A picture containing text, screenshot

Description automatically generated

Univariate analysis on the dataset is as shown below using histograms (Fig 1.1.6):  
Fix- comment on skewness of data.

A graph of sales

Description automatically generated with low confidence

A picture containing text, screenshot, display, diagram

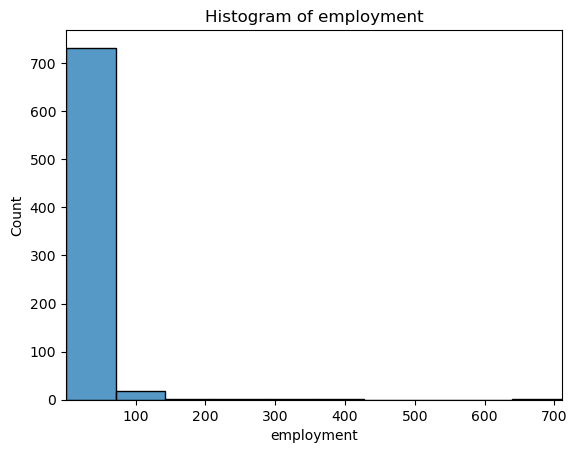
Description automatically generated

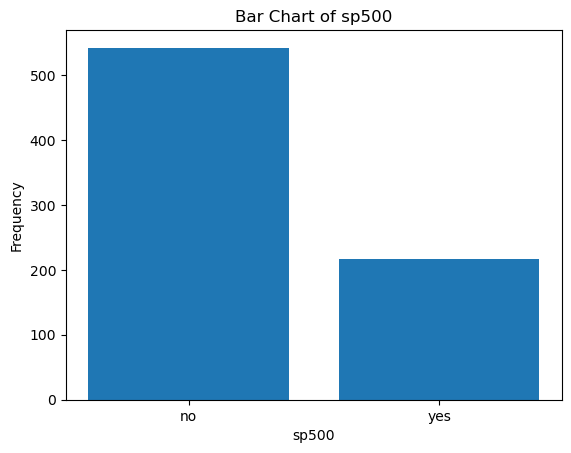
A picture containing text, screenshot, display, diagram

Description automatically generated

A picture containing text, screenshot, display, diagram

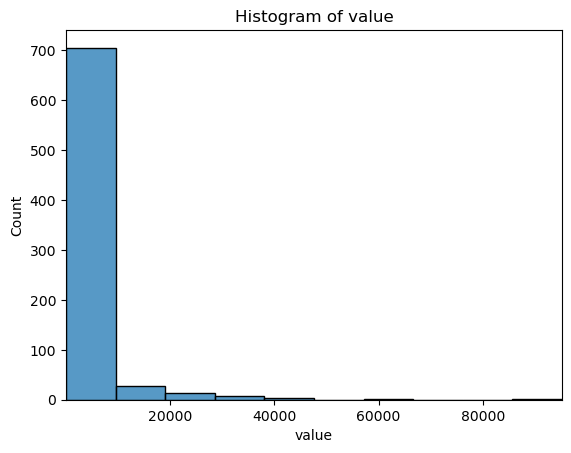
Description automatically generated

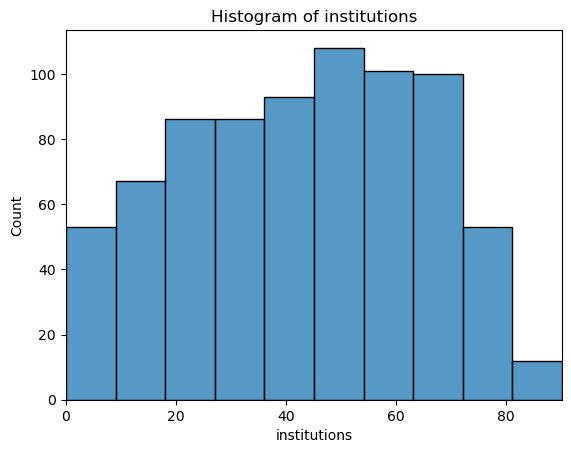




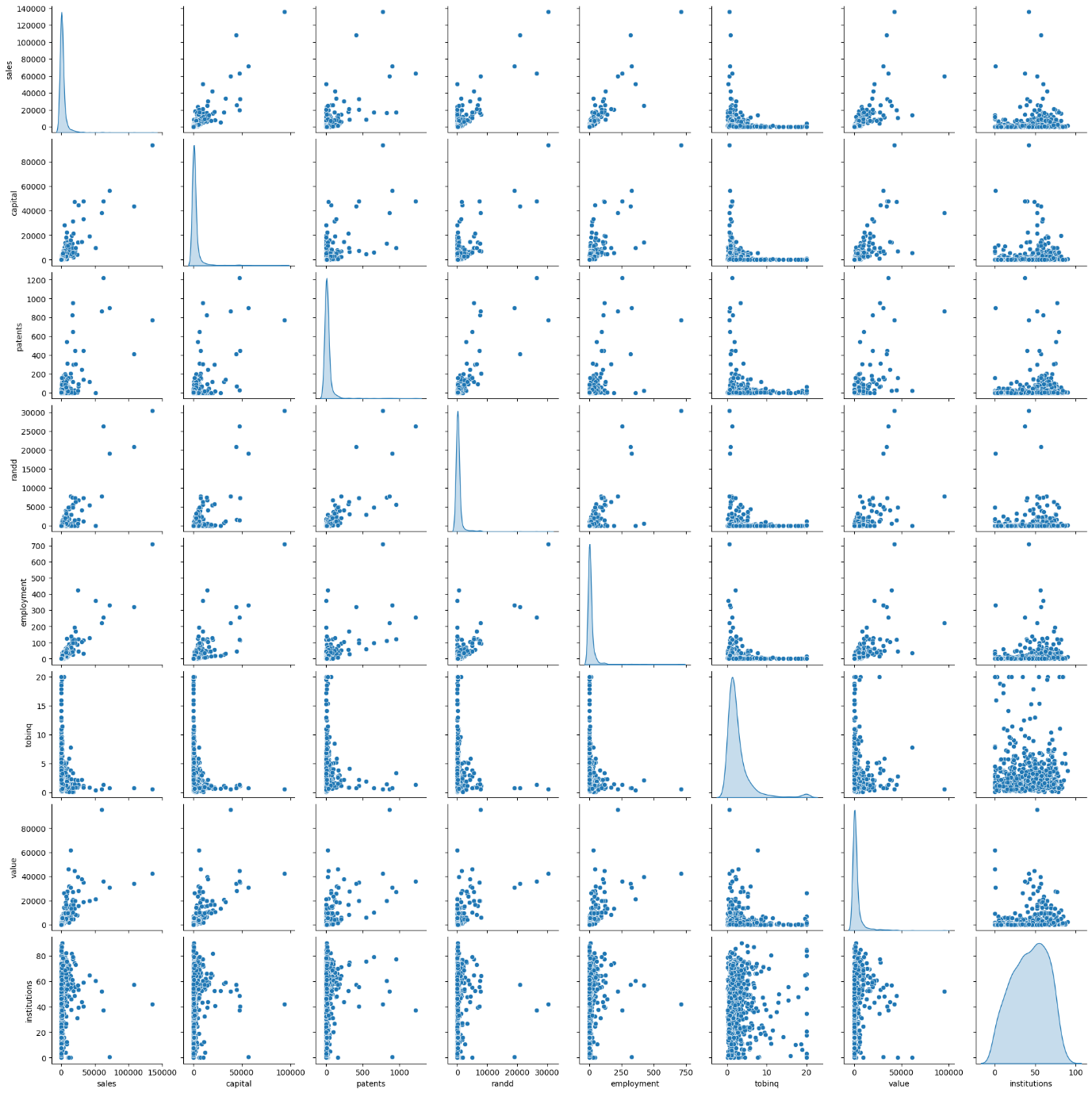
A picture containing text, screenshot, diagram, plot

Description automatically generated



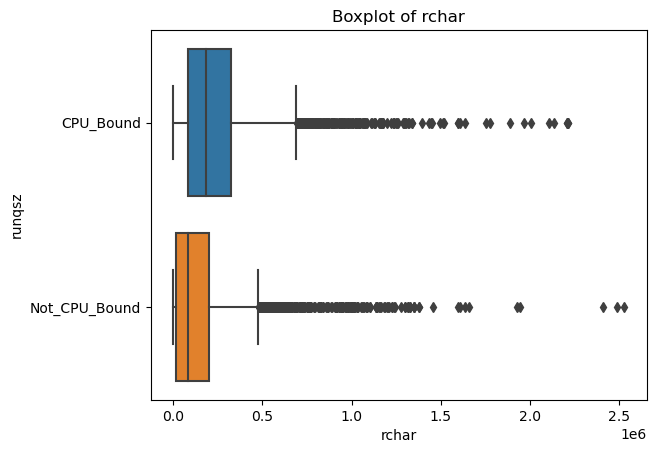


### Bivariate Analysis

Bivariate analysis is performed on the data and the pairplot is shown below (Fig 1.1.7):  
Fix- highlight co-relation features.  


1.2 Impute null values if present? Do you think scaling is necessary in this case? (8 marks)

Answer:

There are no duplicates in the data. There are 104 null values in ‘rchar’, 15 in ‘wchar’ respectively. The data in both these columns is skewed as seen in boxplots below (Fig1.2.1).  


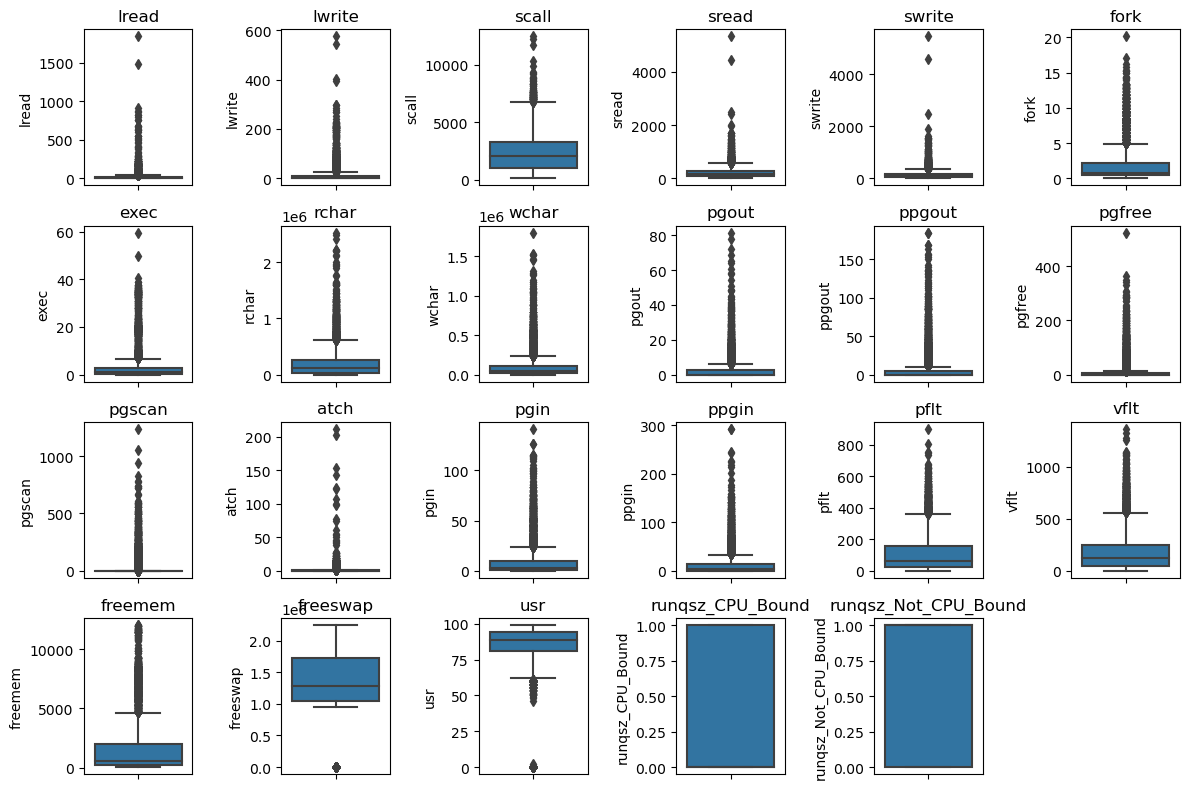
Fig(1.2.2)

Chart

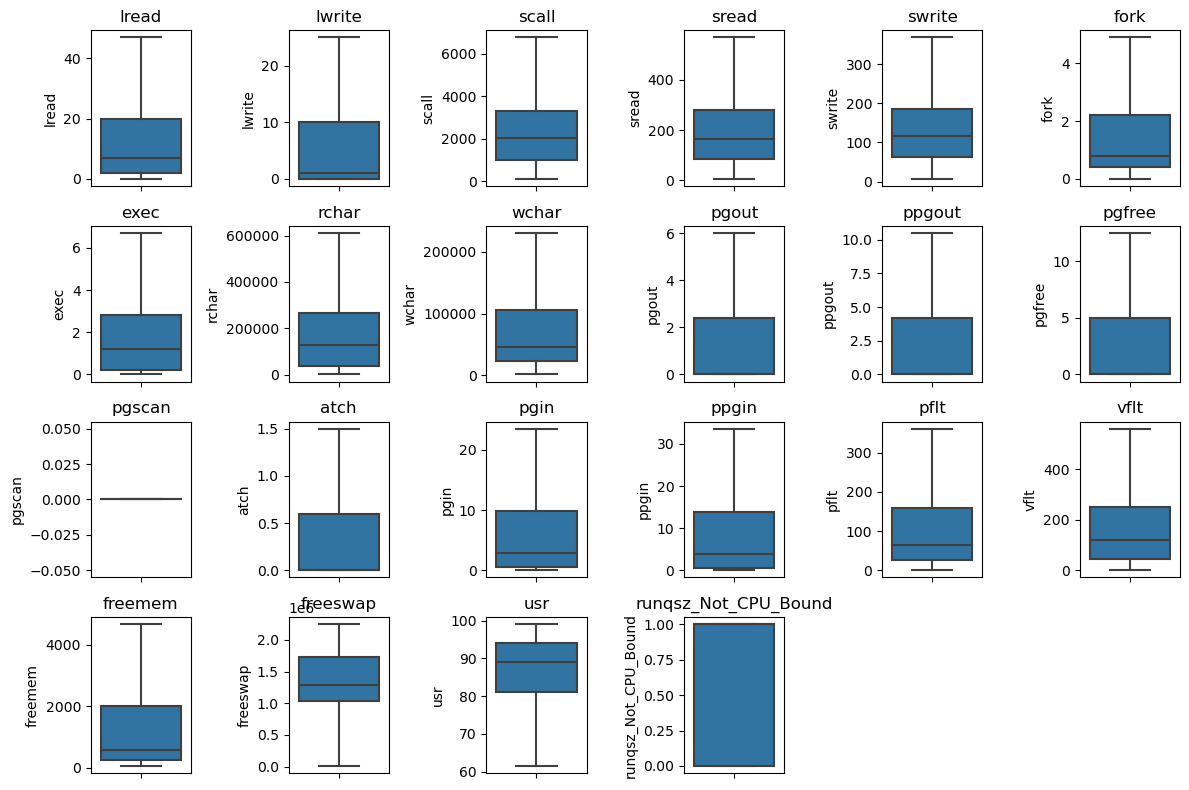
Description automatically generated

The missing values in ‘rchar’, ‘wchar’ columns are treated by median imputation. Median imputation is preferred when the distribution is skewed, as the median is less sensitive to outliers than the mean.

There are multiple outliers in many columns as shown in below figure (Fig1.2.3).



Treat outliers:

The outliers shall be treated via standard Q1 - 3 \* IQR or Q3 + 3 \* IQR method. There are outliers in the dataset as shown in below boxplot (Fig1.3.1)  
  


The below chart shows there is lots of multi-collinearity and co-relation between variables , before treating outliers.  
A picture containing text, screenshot, square, rectangle

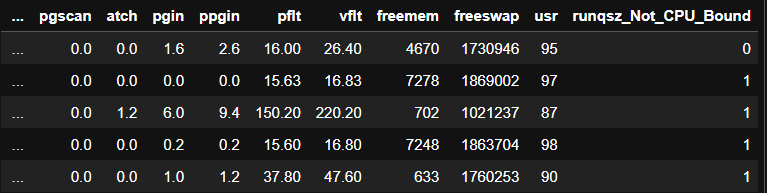
Description automatically generatedEven after treating outliers, below co-relation exists:  
A picture containing text, screenshot, rectangle, square

Description automatically generated

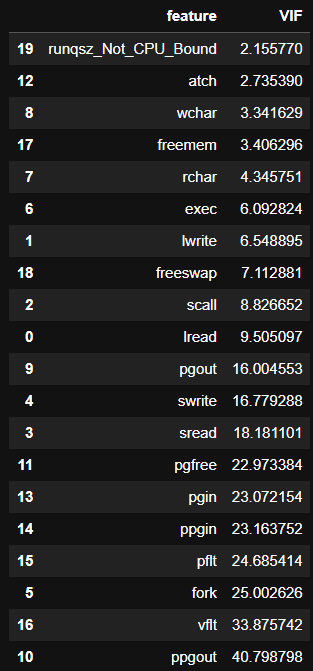
* 1. Encode the data (having string values) for Modelling. Split the data into train and test (70:30). Apply Linear regression using scikit learn. Perform checks for significant variables using appropriate method from statsmodel. Create multiple models and check the performance of Predictions on Train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

Answer:  
Encoding:

The ‘runqsz’ feature is the only categorical feature and contains binary values. After one-hot encoding, the dataset looks like below figure (Fig1.3.1).



VIF:

Using VIF, important features are extracted. The below figure shows the sorted eatures by Importance.  


The features with Imp above 10 are considered for linear regression model.

Split Data:  
The ‘usr’ column is target variable y. Rest of the data is X. The data X and y is split into train and test sets in a 70:30 ratio. The Train dataset top 5 rows are shown below (Fig1.3.2).

A picture containing calendar

Description automatically generated

The test dataset top 5 rows are shown below (Fig1.3.3).

A picture containing calendar

Description automatically generated

The training and test dataset sample for ols stats model is shown below:  




Below snapshots show the sample train and test data used for linear regression:





* 1. Inference: Basis on these predictions, what are the business insights and recommendations. Please explain and summarise the various steps performed in this project. There should be proper business interpretation and actionable insights present.

Answer:

Scikit-learn model RMSE: 4.652295704193024

OLS stats model RMSE: 4.652295704193209

The model co-efficient and intercept are given below for linear regression model.  
Coefficients: [[-6.34815062e-02 4.81612871e-02 -6.63828011e-04 3.08252103e-04

-5.42182230e-03 2.93127272e-02 -3.21166484e-01 -5.16684176e-06

-5.40287524e-06 -3.68819064e-01 -7.65976821e-02 8.44841447e-02

6.27574157e-01 1.99879077e-02 -6.73338398e-02 -3.36028294e-02

-5.46366880e-03 -4.58467188e-04 8.83184026e-06 1.61529785e+00]]

Intercept: [84.1217408]

----

Finally, After dropping the Features with VIF high, t**he Linear regression model gives below results for performance:**

Training set performance:

RMSE: 4.420

R-squared: 0.796

MSE 21.643855319252868

Test set performance:

RMSE: 4.652

R-squared: 0.768

MSE 19.532299277152052

These values are good.

# **Problem 2: Logistic Regression, LDA, CART**

* 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.

Answer:  
Data Description:

[**Contraceptive\_method\_dataset.xlsx**](https://olympus.mygreatlearning.com/courses/87096/files/7893734/download?verifier=HeuusfY52K3brnnpI6fXBmVuhzyAVT1AKTfTT9mN&wrap=1) **data set is provided. The given Data Dictionary is shown below.**Text

Description automatically generated

The data has 1473 rows, 10 columns. ‘Contraceptive\_method\_used’ is the target or feature of interest with ‘Yes’ having higher value count and hence the positive class.

Sample Data:

The Data provided has these top rows as shown below.  


The bottom five rows from the dataset are shown below.  
A screenshot of a computer screen

Description automatically generated

The data contains many numerical and categorical columns. Further exploratory data analysis is as below.

Exploratory Data Analysis:  
The datatype of each feature is shown below:  
Wife\_age float64

Wife\_ education object

Husband\_education object

No\_of\_children\_born float64

Wife\_religion object

Wife\_Working object

Husband\_Occupation int64

Standard\_of\_living\_index object

Media\_exposure object

Contraceptive\_method\_used object

Inference: Out of 10 columns 7 are type object, rest are either of type float or integer. Most of the existing *numerical* columns are of type float64, int64 already.

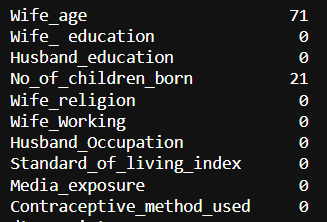
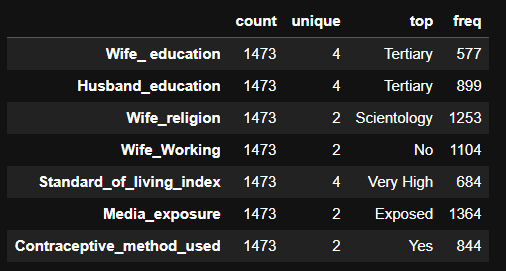
'Husband\_Occupation' data type is already encoded from categorical variable. Categorical columns like Education (both 'Wife\_ education', 'Husband\_education'), 'Wife\_religion', ‘Wife\_Working' , 'Standard\_of\_living\_index', 'Media\_exposure', 'Contraceptive\_method\_used' are object type.

'Media\_exposure ', 'Wife\_Working', 'Wife\_religion' columns will be converted to numerical by one-hot encoding process.

‘Contraceptive\_method\_used’, ‘Wife\_ education’, ‘Husband\_education’, ‘Standard\_of\_living\_index’ shall be converted to numerical by using the data description provided in data inputs.

Additionally, ‘'Wife\_ education' can be renamed as ‘Wife\_education', and ‘Media\_exposure ‘ as ‘Media\_exposure so as to remove the white space in the feature name.

Missing values:

Below information shows the missing values in the data.  
  
There are 71 missing values in ‘Wife\_age’ column, 21 missing values in ‘No\_of\_children\_born’ column.  
Data description is shown below:  


At first look, for most frequent/top observations for ‘Wife\_ education’ and ‘Husband\_education’ are tertiary. Standard of living is high, Contraceptive used is Yes. Need to perform regression to understand more insights from the dataset.

Below are the outliers before treating / encoding the data.  
Chart, box and whisker chart

Description automatically generated

Highest outliers are present in no of children. These outliers shall be ignored for now as these are very small number of outliers. Outliers in more features can be understood as the data is converted from categorical to numerical.

* 1. 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.

Answer:

The features that have categorical data needed to be converted to numerical, and also features are converted to int/float. Numerical conversion of data is performed using the data description provided. 'Wife\_Working', 'Wife\_religion' columns will be converted to numerical by one-hot encoding process.

‘Contraceptive\_method\_used’, ‘Wife\_ education’, ‘Husband\_education’, 'Media\_exposure ', ‘Standard\_of\_living\_index’ shall be converted to numerical by using the data description provided in data inputs. Additionally, the columns are now of numerical type. Data type of the dataset is shown below: Text

Description automatically generated

Sample dataset after encoding is shown below:

The pairplot is shown below:  
A picture containing text, diagram, parallel, screenshot

Description automatically generated

The dataset now has all numerical values.

The univariate and bivariate analysis shows the below charts:  
A picture containing text, screenshot, diagram, line

Description automatically generated

A picture containing text, screenshot, diagram, line

Description automatically generated

A picture containing diagram, plan, square, rectangle

Description automatically generated

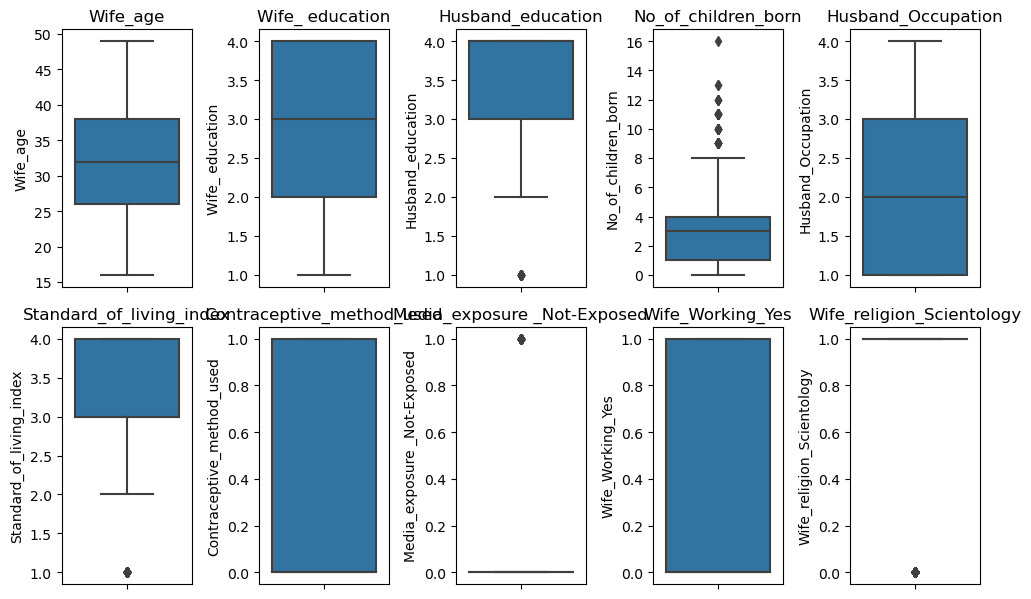
A picture containing diagram, plan, square, rectangle

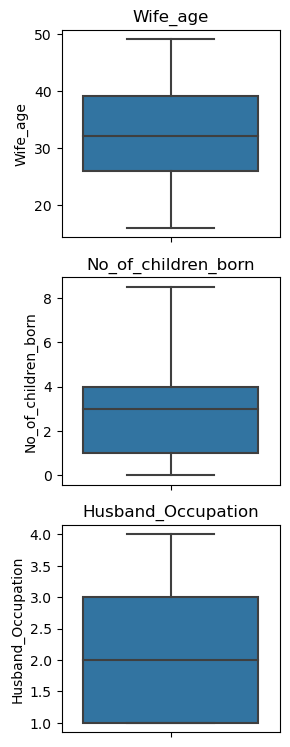
Description automatically generated

A picture containing text, diagram, screenshot, rectangle

Description automatically generated

After treating/encoding the features as numerical, there are outliers in no of children, standard of living. The outliers are shown below.



After treating outliers using the IQR+1.5 rule, we find the boxplot as below with no outliers:  


Using sklearn train test split method, the dataset is split into 70 30 for training and testing using random state 1 to replicate results.

The sample train dataset is shown below:   
A screenshot of a black screen

Description automatically generated with low confidenceThe sample test dataset is shown below:  
A screenshot of a computer

Description automatically generated with medium confidence

However..  
The dataset has lot of corelation & multi collinearity as shown in below heatmaps and VIF Importance.  
A picture containing text, screenshot, number, font

Description automatically generated

A picture containing text, screenshot, number, font

Description automatically generated

The VIF importance sorted shall look like:  

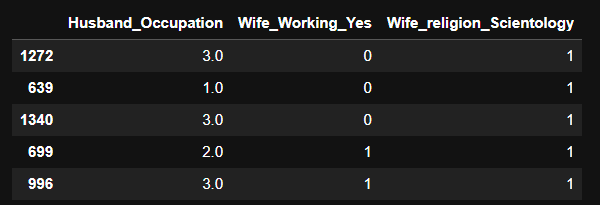

Drop the features with high vif and below corelation heat map is shown:  
Decision trees is as below:  
A picture containing text, diagram, screenshot, font

Description automatically generated

Logistic Regression, LDA, CART are applied on the dataset using training and performance is checked using the test data.

Hyper parameters used for the models are max\_iter=1000 in Logistic regression, n\_components=1 in LDA, criterion=’gini’ in CART.

Drop these features with high VIF:

'No\_of\_children\_born','Husband\_education','Wife\_age','Wife\_education','Standard\_of\_living\_index','Media\_exposure\_1'.  
Then train data will look like:  


Test data sample shall look like:  
A screenshot of a black screen

Description automatically generated with low confidences

* 1. 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.

When the model is run without any treatment of Feature Importances or multi -collinearity, below are the classification results.  
Classification Report of Train Data using Logistic Regression:

precision recall f1-score support

0.0 0.68 0.53 0.59 424

1.0 0.69 0.81 0.74 542

accuracy 0.69 966

macro avg 0.68 0.67 0.67 966

weighted avg 0.69 0.69 0.68 966

Classification Report of Test Data using Logistic Regression:

precision recall f1-score support

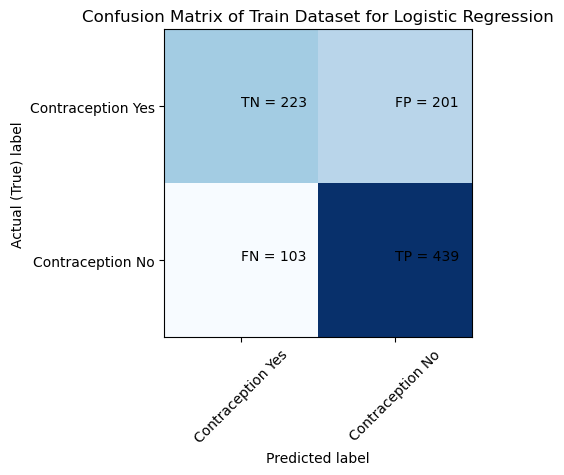
0.0 0.69 0.50 0.58 205

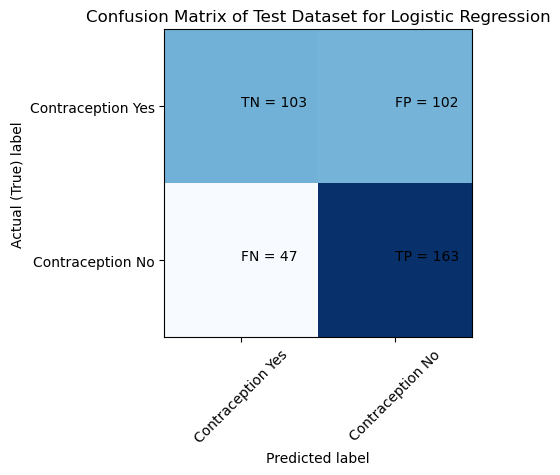
1.0 0.62 0.78 0.69 210

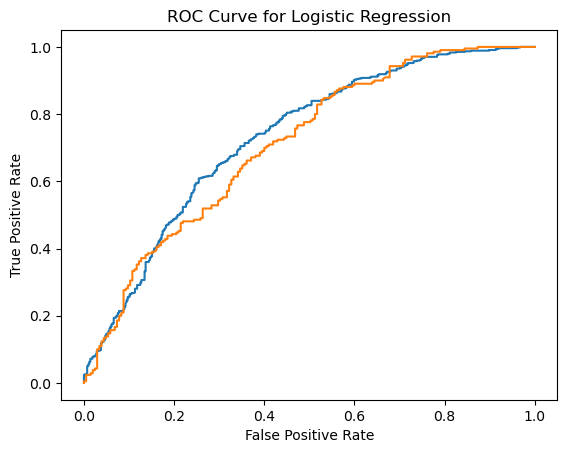
accuracy 0.64 415

macro avg 0.65 0.64 0.63 415

weighted avg 0.65 0.64 0.63 415







------

Classification Report of Train Data using LDA:

precision recall f1-score support

0.0 0.69 0.51 0.58 424

1.0 0.68 0.82 0.74 542

accuracy 0.68 966

macro avg 0.68 0.66 0.66 966

weighted avg 0.68 0.68 0.67 966

Classification Report of Test Data using LDA:

precision recall f1-score support

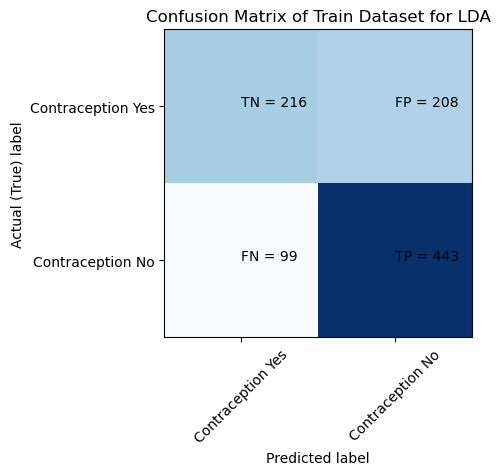
0.0 0.68 0.49 0.57 205

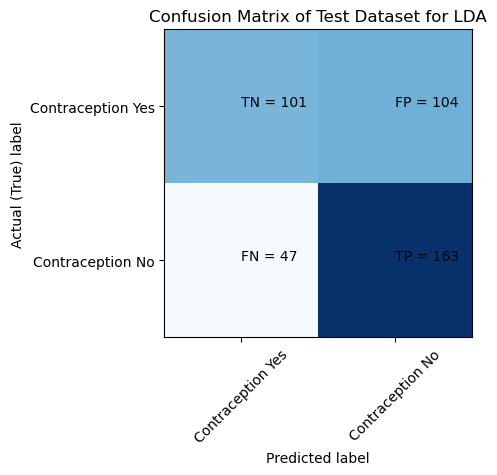
1.0 0.61 0.78 0.68 210

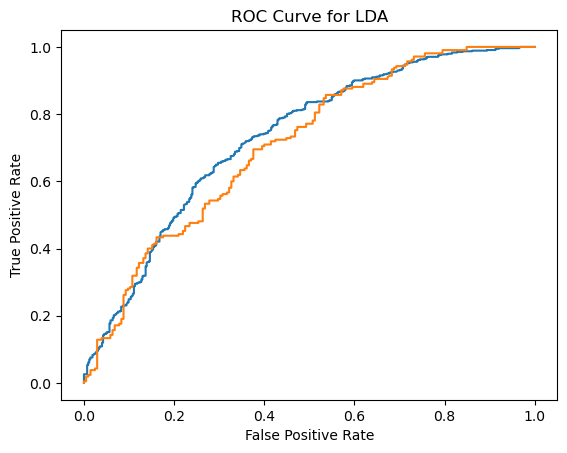
accuracy 0.64 415

macro avg 0.65 0.63 0.63 415

weighted avg 0.65 0.64 0.63 415







------

Classification Report of Train Data using CART:

precision recall f1-score support

0.0 0.97 1.00 0.98 424

1.0 1.00 0.97 0.99 542

accuracy 0.98 966

macro avg 0.98 0.98 0.98 966

weighted avg 0.98 0.98 0.98 966

Classification Report of Test Data using CART:

precision recall f1-score support

0.0 0.61 0.57 0.59 205

1.0 0.60 0.65 0.63 210

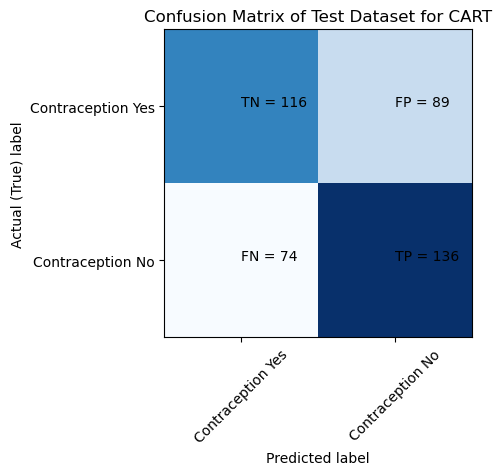
accuracy 0.61 415

macro avg 0.61 0.61 0.61 415

weighted avg 0.61 0.61 0.61 415

A picture containing text, screenshot, diagram, font

Description automatically generated



A picture containing text, line, screenshot, plot

Description automatically generated

------



Clearly Logistic regression is the best model here when VIF is not treated.

* 1. Inference: Basis on these predictions, what are the insights and recommendations. Please explain and summarise the various steps performed in this project. There should be proper business interpretation and actionable insights present.

Answer:

Untreated model gives slightly better performance results as shown below.  
Classification Report of Train Data using Logistic Regression:

precision recall f1-score support

0.0 0.58 0.30 0.39 424

1.0 0.60 0.83 0.70 542

accuracy 0.60 966

macro avg 0.59 0.56 0.54 966

weighted avg 0.59 0.60 0.56 966

Classification Report of Test Data using Logistic Regression:

precision recall f1-score support

0.0 0.59 0.31 0.40 205

1.0 0.54 0.79 0.64 210

accuracy 0.55 415

macro avg 0.56 0.55 0.52 415

weighted avg 0.56 0.55 0.52 415

A picture containing text, screenshot, font, diagram

Description automatically generated

A picture containing text, screenshot, font, diagram

Description automatically generated

A graph with blue and orange lines

Description automatically generated with low confidence

------

Classification Report of Train Data using LDA:

precision recall f1-score support

0.0 0.58 0.30 0.39 424

1.0 0.60 0.83 0.70 542

accuracy 0.60 966

macro avg 0.59 0.56 0.54 966

weighted avg 0.59 0.60 0.56 966

Classification Report of Test Data using LDA:

precision recall f1-score support

0.0 0.59 0.31 0.40 205

1.0 0.54 0.79 0.64 210

accuracy 0.55 415

macro avg 0.56 0.55 0.52 415

weighted avg 0.56 0.55 0.52 415

A picture containing text, screenshot, font, diagram

Description automatically generated

A picture containing text, screenshot, font, diagram

Description automatically generated

A picture containing text, line, plot, diagram

Description automatically generated

------

Classification Report of Train Data using CART:

precision recall f1-score support

0.0 0.57 0.37 0.45 424

1.0 0.61 0.78 0.69 542

accuracy 0.60 966

macro avg 0.59 0.57 0.57 966

weighted avg 0.59 0.60 0.58 966

Classification Report of Test Data using CART:

precision recall f1-score support

0.0 0.58 0.41 0.48 205

1.0 0.55 0.71 0.62 210

accuracy 0.56 415

macro avg 0.57 0.56 0.55 415

weighted avg 0.57 0.56 0.55 415

A picture containing text, screenshot, font, diagram

Description automatically generated

A picture containing text, screenshot, font, diagram

Description automatically generated

A picture containing text, line, plot, diagram

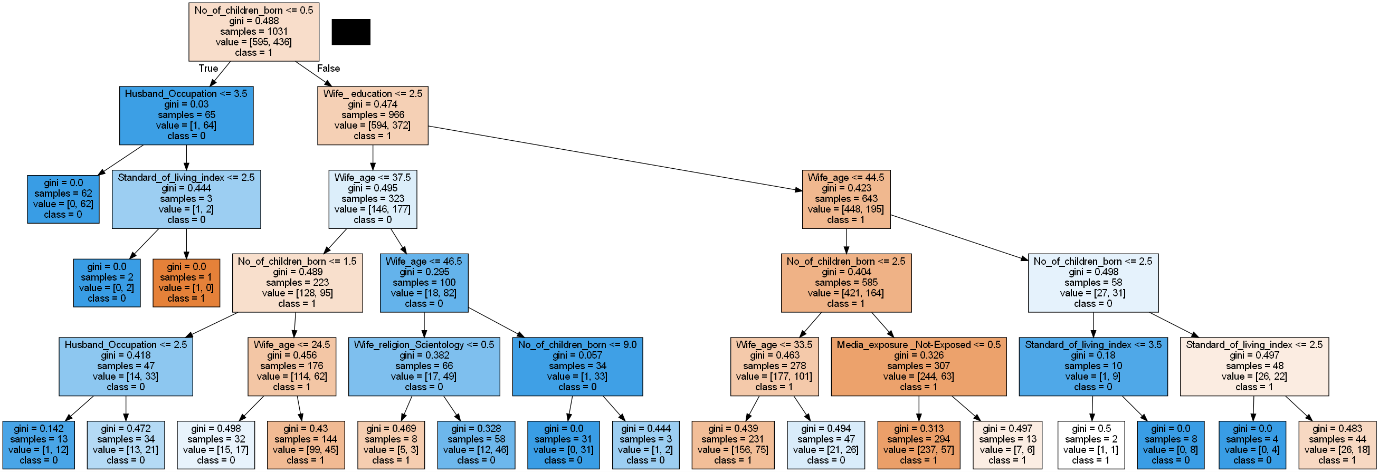
Description automatically generated

------

Regularized CART is slightly a better model by all metrics.

A screenshot of a computer

Description automatically generated with medium confidence

‘Husband\_education’, ‘Wife\_Working’ features have least importance In CART Tree where as ‘Wife\_ education’, ‘Wife\_age’, ‘No\_of\_children\_born’ have high importance in the regularized decision tree.  
The decision tree can be visualized as below (Fig):  


After treating features by VIF importance:

Decision trees is as below:  
A picture containing text, diagram, screenshot, font

Description automatically generated

1. Reflection Report:

Please reflect on all that you learnt and fill this reflection report. You have to copy the link and paste it on the URL bar of your respective browser. https://docs.google.com/forms/d/e/1FAIpQLScKuVyrmTTM7Pboh0IB4YIBUbJp2NrDZcsY4SCRn3ZUkwmLGg/viewform

<Completed>