**PREDICTIVE MODELLING**

**PROJECT (extended)**

**LINEAR REGRESSION**

**LOGISTIC REGRESSION**

**LDA & CART**

BUSINESS

REPORT

**Prepared by: AMAYIKA**

**Great Learning.**

## Email- Kashyap.amayika009@gmail.com

January 1st, 2023

Table of Contents

[Problem1-Linear Regression   
You are a part of an investment firm and your work is to do research about these 759 firms. You are provided with the dataset containing the sales and other attributes of these 759 firms. Predict the sales of these firms on the bases of the details given in the dataset so as to help your company in investing consciously. Also, provide them with 5 attributes that are most important . 7](#_bookmark0)

[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.. 8](#_bookmark1)

[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? Do you think scaling is necessary in this case?. 22](#_bookmark2)

[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.. 27](#_bookmark3)

[1.4 Inference: Basis on these predictions, what are the business insights and recommendations.. 41](#_bookmark4)

[Problem -2 Logistic Regression, LDA, CART – You are hired by a government to do an anlysis of car crashes, you are provided with details of car crashes among which some people survived and some didn’t. you have to help the government in predicting whether a person will survive or not based on the information provided so as to provide insights that will help the government to make stronger laws and help the car manufacturers to ensure safety measures. Also, provide important features based on the models  43](#_bookmark18)

[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.. 43](#_bookmark21)

[2.2 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. 70](#_bookmark23)

[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.? 82](#_bookmark28)

[2.4 Inference: Basis on these predictions, what are the insights and recommendations. There should be proper business interpretation and actionable insights present.. 88](#_bookmark18)

List of Tables

[Table 1. Data set of Investment firm (1-5th rows) 7](#_bookmark7)

[Table 2. Data set of Investment firm (Last 5 rows) 7](#_bookmark8)

[Table 3. Information of data set 8](#_bookmark9)

[Table 4. Dataset after dropping the column Unnamed:0 – first and last 5 rows 9](#_bookmark10)

[Table 5. Information after dropping the column 9](#_bookmark13)

[Table 6. Description of the data set 10](#_bookmark15)

[Table 7. Missing or null values in the data set 10](#_bookmark16)

[Table 8. Numerical columns and their Frequencies 11](#_bookmark17)

[Table 9. Unique value Analysis 12](#_bookmark19)

[Table 10. Skewness table 15](#_bookmark20)

[Table 11. Correlation table for attributes 17](#_bookmark22)

[Table 12. Missing value percentage in the data set 22](#_bookmark25)

[Table 13. After missing data imputation 22](#_bookmark7)

[Table 14. Data set after missing values imputation 22](#_bookmark8)

[Table 15. Description of the data set after missing values imputation 23](#_bookmark9)

[Table 16. Unique values analysis after missing values treatment 24](#_bookmark10)

[Table 17. Numerical data set before scaling 25](#_bookmark13)

[Table 18. Numerical data set after scaling 26](#_bookmark15)

[Table 19. Data set after dummy variable encoding 27](#_bookmark16)

[Table 20. Information of data set after encoding 28](#_bookmark17)

[Table 21. Coefficients of variables of Linear Regression 29](#_bookmark19)

[Table 22. OLS regression Results on Model 1 – Train data 30](#_bookmark20)

[Table 23. OLS regression Results on Model 1 – Test data 32](#_bookmark22)

[Table 24. VIF values of variables 34](#_bookmark25)

[Table 25. OLS regression Results on Model 2 – Train data 35](#_bookmark7)

[Table 26. OLS regression Results on Model 3 – Train data 36](#_bookmark8)

[Table 27. OLS regression Results on Model 5 – Train data 38](#_bookmark9)

[Table 28. OLS regression Results on Model 6 – Train data 39](#_bookmark9)

[Table 29. Models’ comparison table 40](#_bookmark10)

[Table 30. Actual vs predicted probabilities table 41](#_bookmark10)

[Table 31. Model parameters and their importance 41](#_bookmark13)

[Table 32. Car crash data set top-5 rows 43](#_bookmark15)

[Table 33. Car crash data set last-5 rows 43](#_bookmark16)

[Table 34. Information of car crash data set 44](#_bookmark17)

[Table 35. Null values in the data set 45](#_bookmark19)

[Table 36. Description of the data set 45](#_bookmark16)

[Table 37. Numerical columns in data set 47](#_bookmark17)

[Table 38. Frequencies of numerical columns in data set 47](#_bookmark7)

[Table 39. Unique values analysis 48](#_bookmark8)

[Table 40. Categorical columns in the data set 48](#_bookmark9)

[Table 41. Skewness in the data set 60](#_bookmark9)

[Table 42. Correlation table 61](#_bookmark10)

[Table 43. Null value imputation table 68](#_bookmark10)

[Table 44. Description after null values imputation 68](#_bookmark13)

[Table 45. Data set after dummy variables encoding 70](#_bookmark15)

[Table 46. Predicted probabilities of Train data of LR 71](#_bookmark16)

[Table 47. Predicted probabilities of Test data of LR 72](#_bookmark17)

[Table 48. Coefficients of attributes in LR 72](#_bookmark19)

[Table 49. Classification Report on Training data for LR 73](#_bookmark16)

[Table 50. Classification Report on Testing data for LR 73](#_bookmark17)

[Table 51. Classification Report on Training data for LR best parameters 75](#_bookmark7)

[Table 52. Classification Report on Test data for LR best parameters 75](#_bookmark8)

[Table 53. Class label predictors of training data set for LDA 76](#_bookmark9)

[Table 54. Class label predictors of test data set for LDA 77](#_bookmark9)

[Table 55. Coefficients of variables for LDA model 77](#_bookmark10)

[Table 56. Classification Report on Training data for LDA 78](#_bookmark10)

[Table 57. Classification Report on Test data for LDA 78](#_bookmark13)

[Table 58. Classification Report on Test data for LDA on default cut-off 80](#_bookmark15)

[Table 59. Classification Report on Test data for LDA on customized cut-off 80](#_bookmark16)

[Table 60. Important features of CART model 82](#_bookmark17)

[Table 61. Model comparisons of LR, LDA and CART 87](#_bookmark19)

List of Equations and Formulas

[Equation 1. Min- Max scaler formula 26](#_bookmark8)

[Equation 2. MAE Formula 29](#_bookmark8)

[Equation 3. RMSE Formula 29](#_bookmark9)

[Equation 4. VIF formula 34](#_bookmark13)

[Equation 5. IQR formula 68](#_bookmark15)

List of Figures

[Figure 1. Null values percentage 11](#_bookmark9)

[Figure 2. Count plot for the attribute “patents” 13](#_bookmark10)

[Figure 3. Line plot of Sales vs Capital’ 13](#_bookmark13)

[Figure 4. Count plot of attribute sp500 14](#_bookmark15)

[Figure 5. Histograms of all the variables 14](#_bookmark16)

[Figure 6. Boxplots of variables 16](#_bookmark17)

[Figure 7. Pair plots of variables 17](#_bookmark19)

[Figure 8. Heat map of variables 18](#_bookmark20)

[Figure 9. Bar graphs of various variables 19](#_bookmark20)

[Figure 10. Scatter plots of various attributes and their analysis 19](#_bookmark16)

[Figure 11. Multivariate analysis of various attributes 21](#_bookmark17)

[Figure 12. Boxplots after outliers’ treatment 23](#_bookmark19)

[Figure 13. Boxplots before outliers’ treatment in one graph for comparing 24](#_bookmark20)

[Figure 14. Data distribution range before scaling – violin plots 25](#_bookmark22)

[Figure 15. Data distribution range after scaling – violin plots 27](#_bookmark17)

[Figure 16. Scatter plot Actual vs Predicted 33](#_bookmark19)

[Figure 17. Actual vs Fitted graph after multicollinearity treatment 41](#_bookmark20)

[Figure 18. Null values percentage of Car crash data set 46](#_bookmark22)

[Figure 19. Frequency distribution of dvcat 48](#_bookmark22)

[Figure 20. Frequency distribution of Survived 49](#_bookmark17)

[Figure 21. Frequency distribution of airbag 49](#_bookmark19)

[Figure 22. Frequency distribution of seatbelt 50](#_bookmark20)

[Figure 23. Frequency distribution of sex 50](#_bookmark22)

[Figure 24. Frequency distribution of abcat 51](#_bookmark17)

[Figure 25. Frequency distribution of occRole 51](#_bookmark19)

[Figure 26. Histogram of Weight 52](#_bookmark20)

[Figure 27. Boxplot of Weight 52](#_bookmark22)

[Figure 28. Histogram of frontal 53](#_bookmark22)

[Figure 29. Boxplot of frontal 53](#_bookmark25)

[Figure 30. Histogram of ageOFocc 54](#_bookmark22)

[Figure 31. Boxplot of ageOFocc 54](#_bookmark25)

[Figure 32. Histogram of yearacc 55](#_bookmark22)

[Figure 33. Boxplot of yearacc 55](#_bookmark25)

[Figure 34. Histogram of yearVeh 56](#_bookmark22)

[Figure 35. Boxplot of yearVeh 56](#_bookmark25)

[Figure 36. Histogram of deploy 56](#_bookmark25)

[Figure 37. Boxplot of deploy 57](#_bookmark22)

[Figure 38. Histogram of injSeverity 57](#_bookmark22)

[Figure 39. Boxplot of injSeverity 58](#_bookmark17)

[Figure 40. Histograms and KDE of all the variables 59](#_bookmark19)

[Figure 41. Boxplot of before outliers’ treatment of all variables 60](#_bookmark20)

[Figure 42. Pair plot of variables 62](#_bookmark22)

[Figure 43. Pie chart of airbags, Gender and Weight 63](#_bookmark17)

[Figure 44. Bar graph of caseid 63](#_bookmark19)

[Figure 45. Multivariate analysis of target variable 64](#_bookmark20)

[Figure 46. Facet grid of airbags, injSeverity and dvcat 64](#_bookmark22)

[Figure 47. Facet grid of seatbelt, injSeverity and dvcat 65](#_bookmark22)

[Figure 48. Facet grid of yearacc, injSeverity and dvcat 66](#_bookmark25)

[Figure 49. Point plot of dvcat and injSeverity 66](#_bookmark22)

[Figure 50. Point plot of dvcat and abcat 67](#_bookmark25)

[Figure 51. Heat map 67](#_bookmark22)

[Figure 52. Boxplots after outliers’ treatment 69](#_bookmark25)

[Figure 53. Boxplots after outliers’ treatment on one graph 69](#_bookmark22)

[Figure 54. AUC-ROC Curve of LR – Training data 73](#_bookmark25)

[Figure 55. AUC-ROC Curve of LR – Testing data 74](#_bookmark25)

[Figure 56. AUC-ROC Curve of LDA – Training data 78](#_bookmark22)

[Figure 57. AUC-ROC Curve of LDA – Testing data 79](#_bookmark25)

[Figure 58. Tree flow chart of CART model 81](#_bookmark25)

[Figure 59. Confusion matrix and AUC-ROC Curve of LR – Training data 82](#_bookmark25)

[Figure 60. Confusion matrix and AUC-ROC Curve of LR – Test data 83](#_bookmark22)

[Figure 61. Confusion matrix and AUC-ROC Curve of LDA – Training data 84](#_bookmark25)

[Figure 62. Confusion matrix and AUC-ROC Curve of LDA – Test data 85](#_bookmark22)

[Figure 63. Confusion matrix and AUC-ROC Curve of CART model – Training data 86](#_bookmark25)

[Figure 64. Confusion matrix and AUC-ROC Curve of CART model – Test data 87](#_bookmark25)

List of Dictionary

* sales: Sales (in millions of dollars).
* capital: Net stock of property, plant, and equipment.
* patents: Granted patents.
* randd: R&D stock (in millions of dollars).
* employment: Employment (in 1000s).
* 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
* 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.
* value: Stock market value.
* institutions: Proportion of stock owned by institutions.
* dvcat: factor with levels (estimated impact speeds) 1-9km/h, 10-24, 25-39, 40-54, 55+
* weight: Observation weights, albeit of uncertain accuracy, designed to account for varying sampling probabilities. (The inverse probability weighting estimator can be used to demonstrate causality when the researcher cannot conduct a controlled experiment but has observed data to model)
* Survived: factor with levels Survived or not survived
* airbag: a factor with levels none or airbag
* seatbelt: a factor with levels none or belted
* frontal: a numeric vector; 0 = non-frontal, 1=frontal impact
* sex: a factor with levels f: Female or m: Male
* ageOFocc: age of occupant in years
* yearacc: year of accident
* yearVeh: Year of model of vehicle; a numeric vector
* abcat: Did one or more (driver or passenger) airbag(s) deploy? This factor has levels deploy, no-deploy and unavailable
* occRole: a factor with levels driver or pass: passenger
* deploy: a numeric vector: 0 if an airbag was unavailable or did not deploy; 1 if one or more bags deployed.
* injSeverity: a numeric vector; 0: none, 1: possible injury, 2: no incapacity, 3: incapacity, 4: killed; 5: unknown, 6: prior death
* caseid: character, created by pasting together the populations sampling unit, the case number, and the vehicle number. Within each year, use this to uniquely identify the vehicle.

# PROBLEM – 1

# LINEAR REGRESSION

You are a part of an investment firm and your work is to do research about these 759 firms. You are provided with the dataset containing the sales and other attributes of these 759 firms. Predict the sales of these firms on the bases of the details given in the dataset so as to help your company in investing consciously. Also, provide them with 5 attributes that are most important.

Table 1 – Data set of investment firm (1st – 5 Rows)

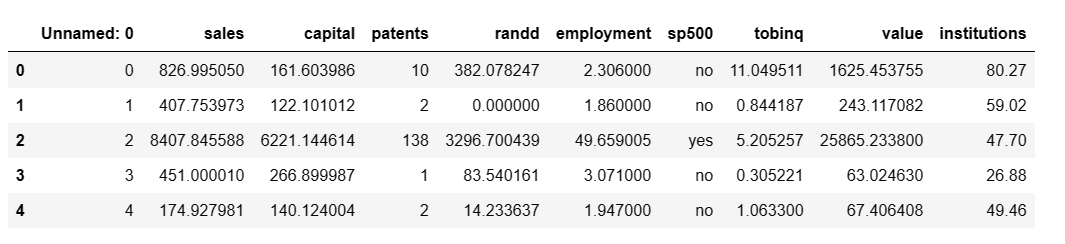
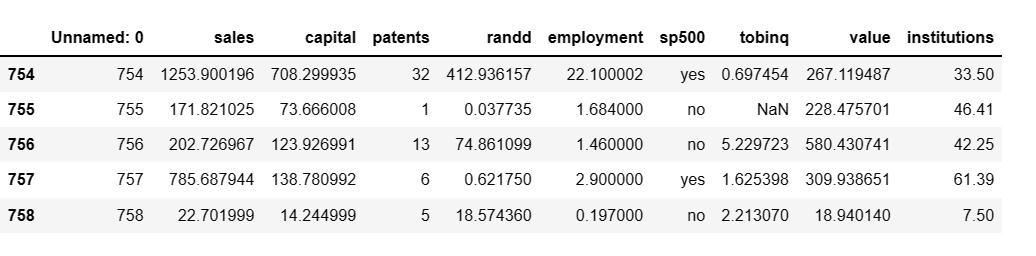


Table 2 – Data set of investment firm (Last – 5 Rows)



**Data Dictionary of the given Data set:**

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.

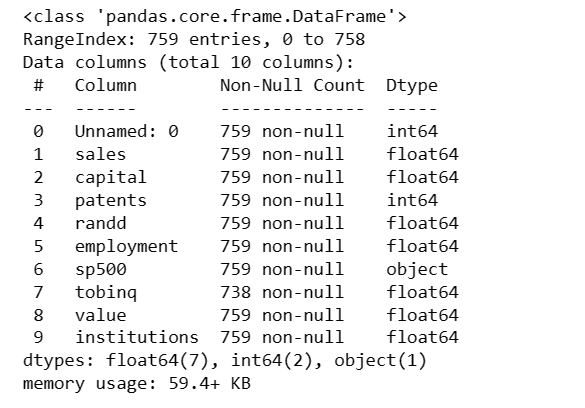
9. institutions: Proportion of stock owned by institutions.

**Part 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.**

**Exploratory Data Analysis**

So now we will start exploring the data first and we will get to know the structure and observations in the data set along with the information and description of the data. We will check the missing values and 5-point summary of the dataset and then impute the same if there are any missing values or outliers present.

Table 3 – Information of dataset



1. The data set contains 759 rows and 10 columns, whose descriptions are mentioned above.

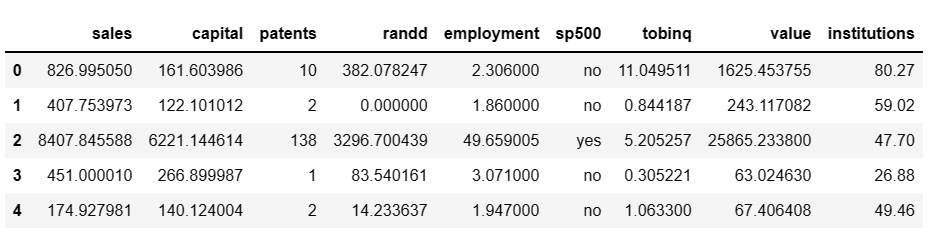
2. We can drop the column ‘Unnamed: 0’as it is just the index in the form of column not going to contribute in our data analysis and its exploration.

3. Here we can see that 9 out of 10 columns are numerical in nature with either Float or Int data type. ‘sp500’ which represents S&P is a stock market index that measures the stock performance of 500 large companies listed on stock exchanges in the United States is in the ‘object’ data type format. We will check the features and their value counts and then encode them using one hot encoding.

4. Dropping the column ‘Unnamed’ from the data set and then again checking the summary and description of the data set

5. The data set now contains only 9 columns after dropping one of the unnecessary columns.

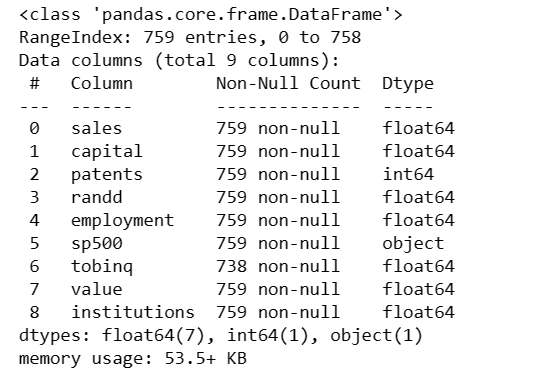
Table 4 – Dataset after dropping the column “unnamed:0”- top 5 and last 5 rows

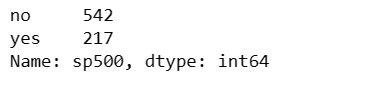




* Now we will again check the information after dropping the data set

Table 5 – Information after dropping the column “unnamed:0”

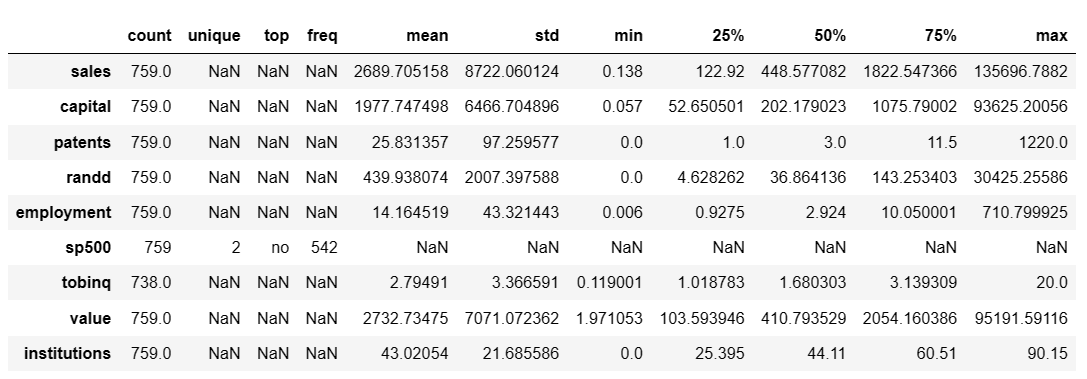




* Since this variable “sp500” is in binary form which means in the form of “YES” or “NO”, we can change it in integer datatype so that it can have values 0 or 1 for ‘Not a sp500 firm’ and ‘a sp500 firm’ respectively. So, we can create Dummy variables using One hot Encoding method afterwards before we proceed with the modelling part.

**Description Table for the above data set:**

Table 6 – Description of the data set

****

**Summary of dataset**

1. The mean and median are largely different for almost all the variables except for the feature “institutions”. This large gap between mean and median where Mean is higher in nature shows that data can be positively skewed having outliers, we will see that when we plot the boxplots later in the analysis part.

2. All variables are continuous variables except for “sp500” which is categorical in nature and ‘patents’ has discrete data.

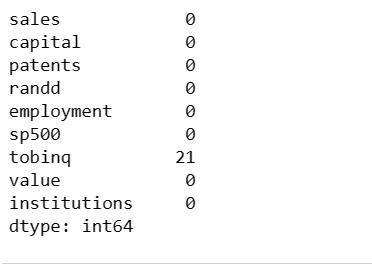
3. Almost all the variables have minimum amount values as zero or almost equals to zero, except for “value” and “tobinq”.

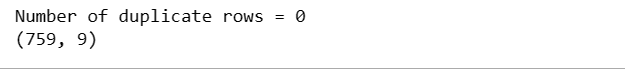
4. For most of the variables except ‘institutions’, the mean is far greater than median indicating positive skewness in the data

5. We can see that Standard deviation is also high for all numerical variables indicating large deflections in the values of the given variables.

6. The maximum values of almost all the variables are much higher than the 3rd Quartile value that is the 75th percentile value again indicating presence of large number of outliers in the data set.

Table 7 – Missing values in the data set





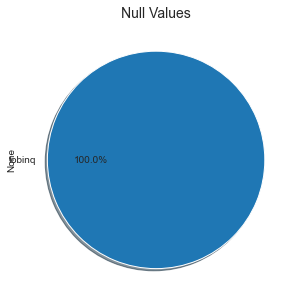
**Inference based on Null value Detection and Duplicate rows Detection**

1. The null values present in the data set is less than 2.76% of the total observations present. So, we will treat this data rather than dropping these observations.

2. Observed null values in only 1 field that is tobinq.

3. We can impute the null values with median value of the data set as the variables are continuous in nature.

4. Now checking the duplicate rows in the dataset is the next step and if we find any duplicate values, we will drop those columns else we will move forward with our further analysis. So, we used the corresponding duplicated function and it gives out that we have zero (0) duplicate rows in our dataset. There are no duplicate rows in the data set.

 fig -1 Null values percentage

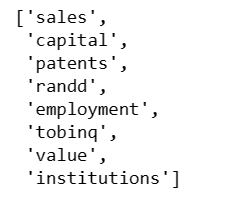
.

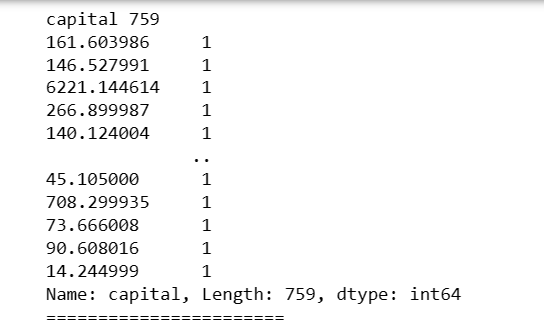
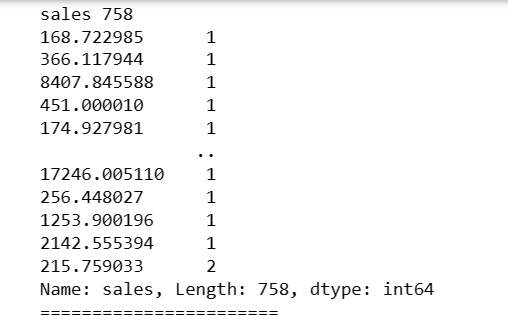
**UNIVARIATE ANALYSIS:**

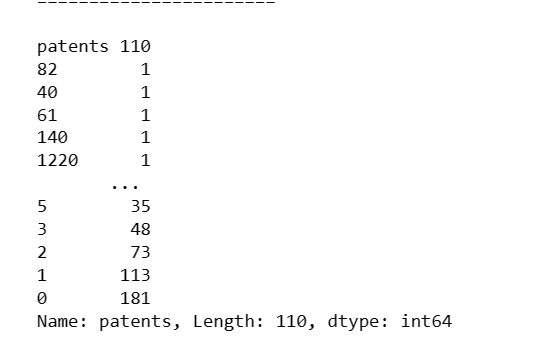
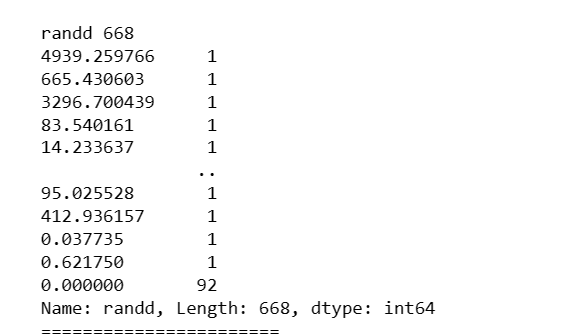
**Numerical feature levels and their frequencies:**

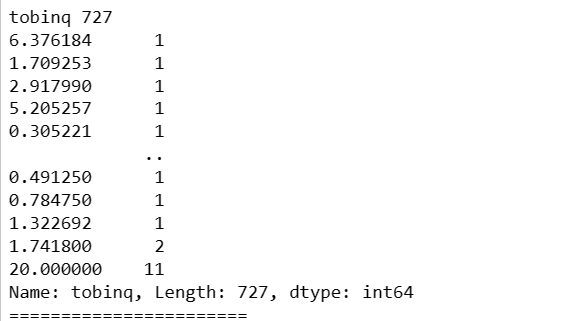
The following are the numerical columns present in the data set along with their frequencies and unique values in the form of tables below:

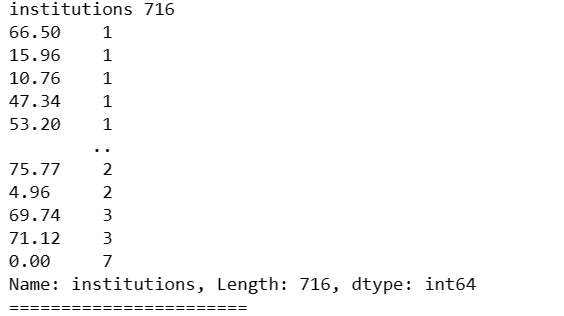
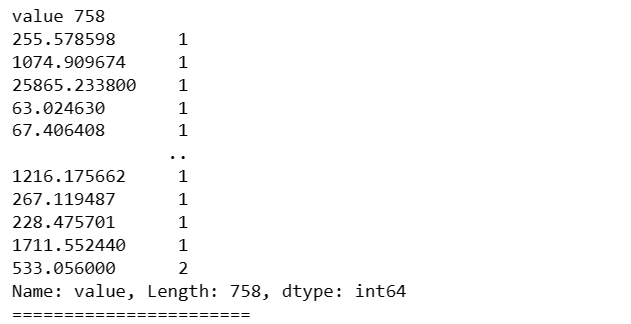
Table 8 – Numerical columns and their Frequencies



****

****

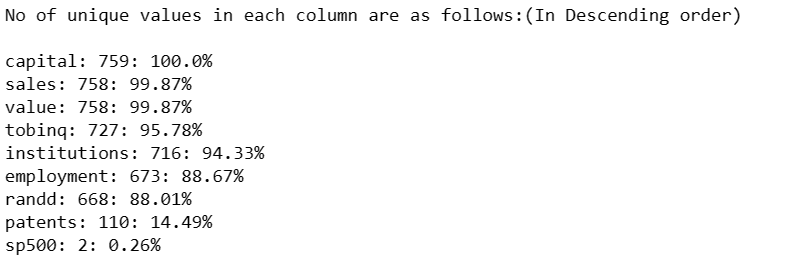
****

****

**Unique Values Analysis:**

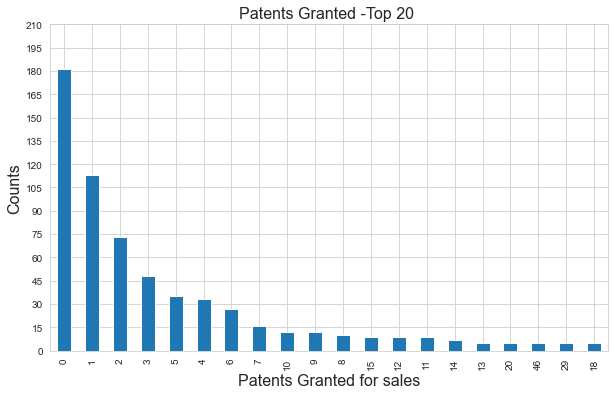
We will analyse the number of unique values in each of our attributes.

Table 9 – Unique value analysis

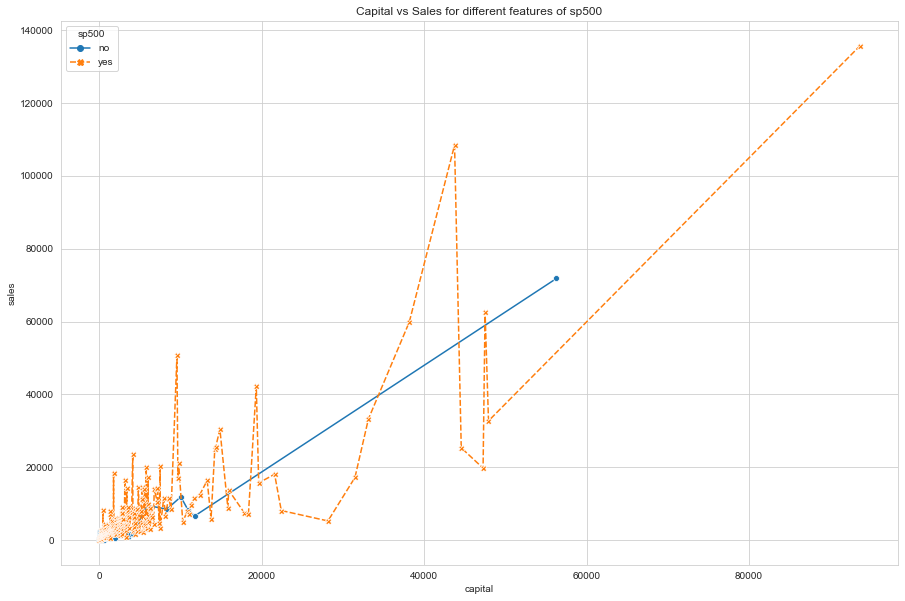


**Graphical representations of univariate analysis:**

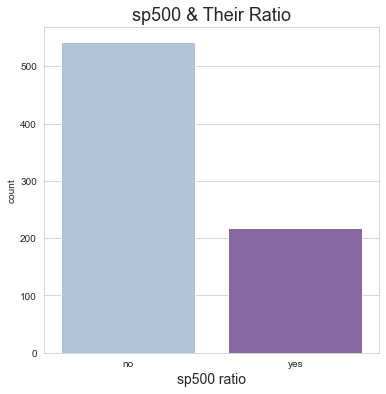
Patents Granted and their Distribution – Top 20

 fig -2 Count plot of Patents

Capital vs Sales for different features of sp500

 fig -3 Line plot of sales vs capital

* Maximum firms have no patents followed by 1 patent or 2 patents and it has gone upto 1220 also where one of the firms have this much patents as seen in the freqency if patents above.
* The line plot also shows that the companies who are registered in the list of top 500 firms have variable kind of sales even if the capital is increasing, showcasing the fact that there are other raesons as well for their increase or decrease in sales.
* But in case of the companies those are not registered they show a linear growing pattern of sales with increase of capital ratio.
* The below graph also shows the attribute “sp500” and its distribution ratio which shows that “no” is much more than “yes” in the ratio.

 fig -4 Count plot of sp500

**Histograms plots of various variables and their analysis:**

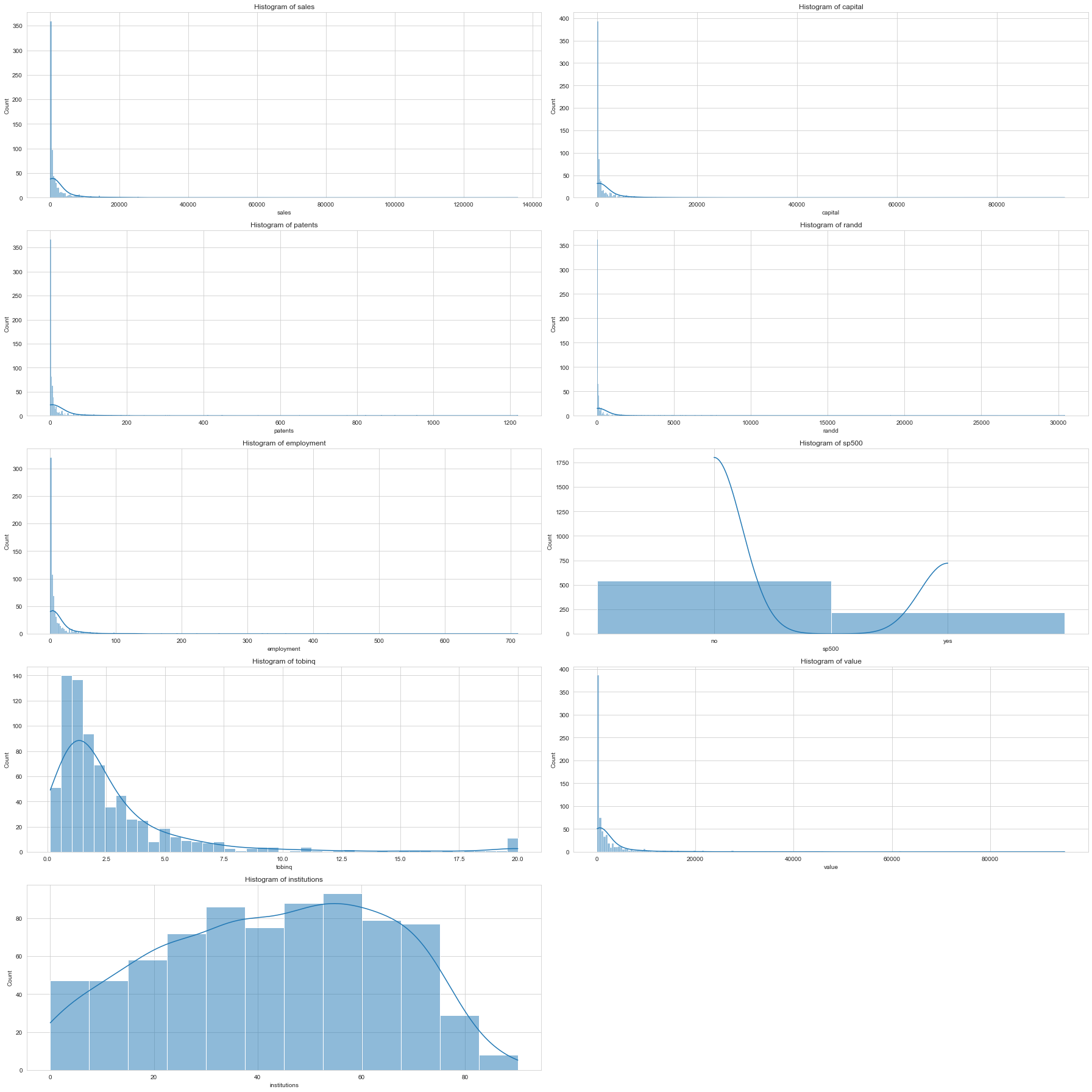


fig -5 Histograms of the variables

**Checking Skewness in the dataset:**

Table 10 – Skewness table

****

**Inferences:**

1. The data for almost all the columns shows that the data is positively skewed except for “Institutions” which is almost normally distributed. The highest Skew is present in the variable “randd” which is almost around 10.27

2. The data for “Institutions” is almost normally distributed with a single peak.

3. The data for target variable “Sales” shows that the data is highly positively skewed having the skewness value around 9.21 just slightly below than “randd”.

4. The data for “sp500” shows that the data has only two observations “Membership of firms in the S&P 500 index as Yes” and “Membership of firms in the S&P 500 index as No”. This is the only “object” data type present in the data set else all are numerical data types.

5. The Target variable in the data set is “sales” which is dependent on all the other features.

6. All the continuous variables in the data set have different range of values so we will have to make the scale similar and then only we will proceed with the modeling.

**Univariate analysis using boxplot:**

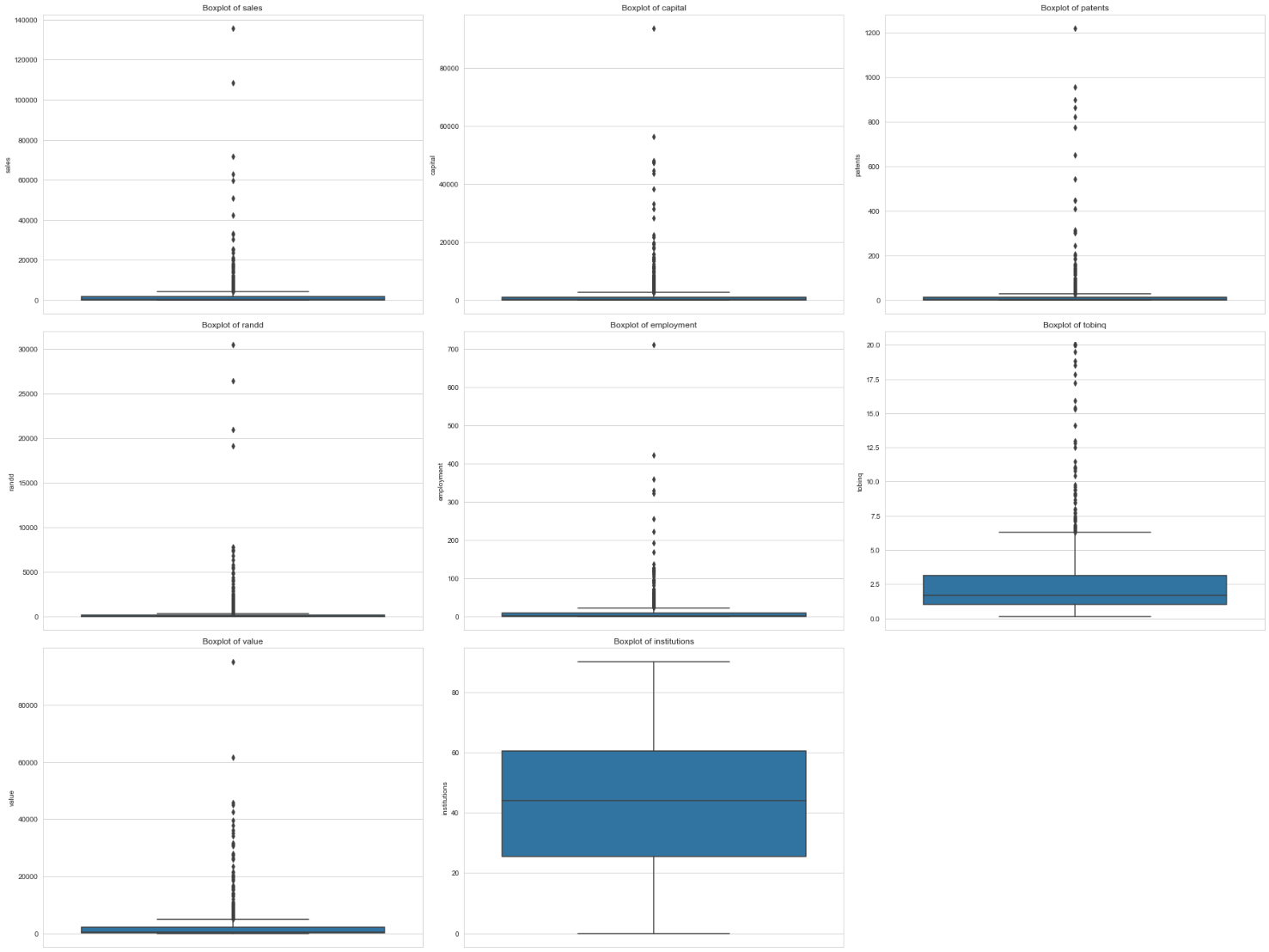
****

fig -6 Boxplots of the variables

**Observations:**

1. Almost all the continuous variables have outliers except for "Institutions" which has no outlier present.

2. We can treat these outliers using the IQR approach

3. The target variable "sales" has outliers more than that of the maximum value as we have already seen that it is a positively skewed curve.

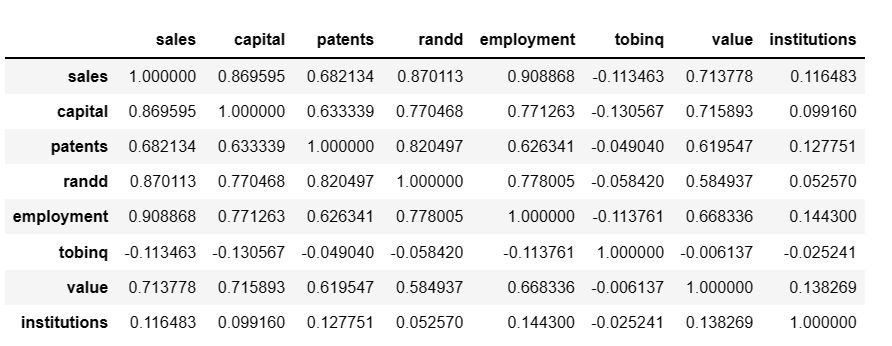
4. In this case, it may be necessary to scale the data as, different variables are in different scale or magnitude, even if they all are continuous in nature. For example- "patents" are whole numbers, “tobinq" is a ratio converted in decimal numbers and similarly all other variables are in having different range of variables.

5. For most of the variables except 'institutions’, the mean is far greater than median indicating positive skewness in the data and can be also observed from the boxplots.

**Bivariate and Multivariate Analysis:**

We can do the bivariate and multivariate analysis using count plot and bar plot of various variables with respect to each other and also with the target variable.

Table 11 – Correlation table

****

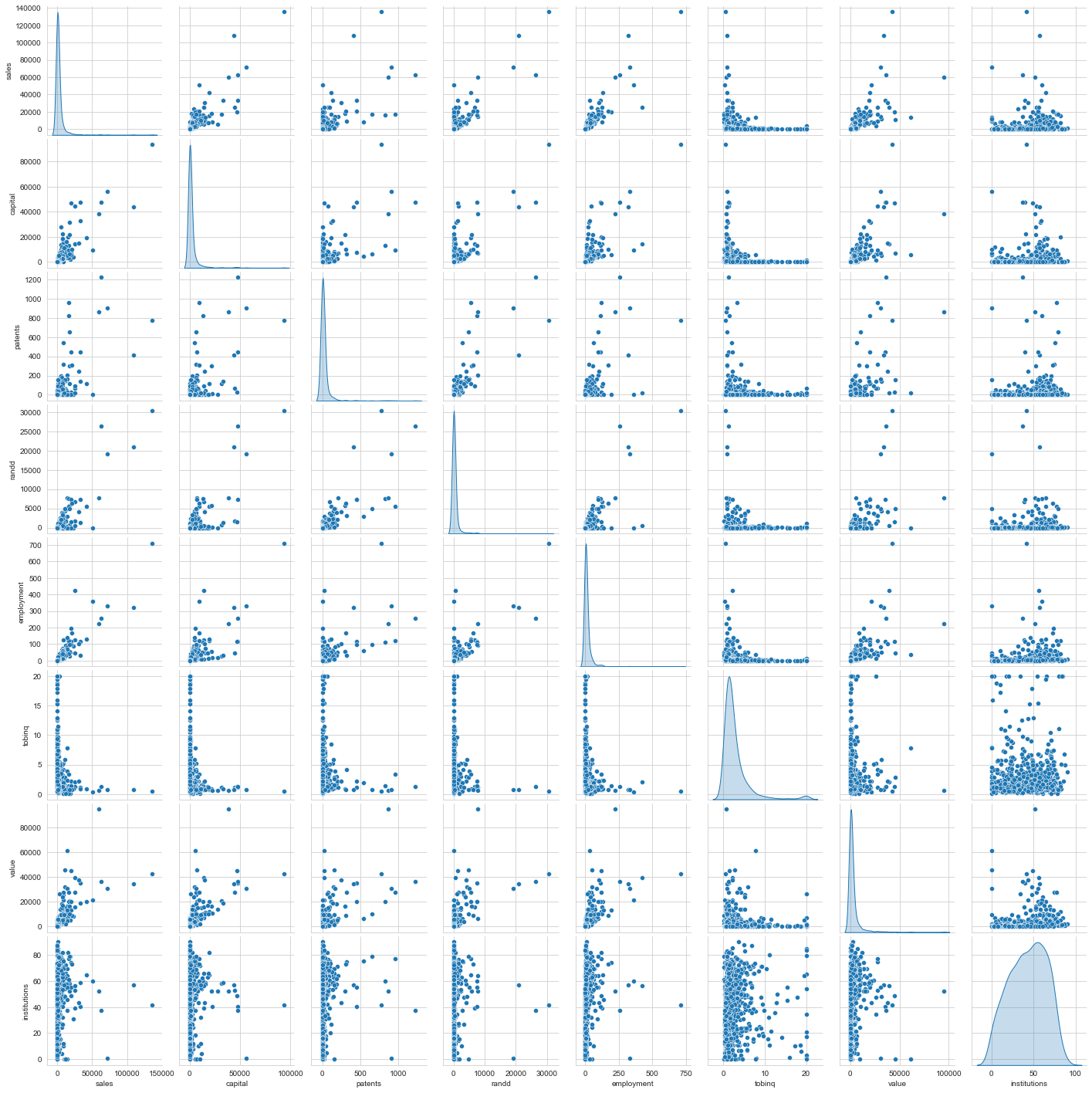
****

fig -7 Pair plots of the variables

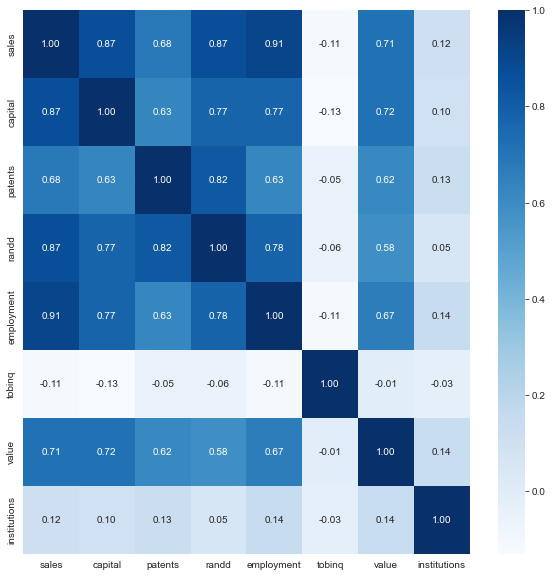
****

fig -8 Heat map of the variables

**Observations based on pair plot:**

Inference: As per the Heat Map, we can conclude that the following variables are highly correlated:

- employment and sales

- randd and sales

- randd and patents

- randd and capital

- randd and employment

- capital with employment

1. 'Sales’ is the target variable or dependent variable and rest are predictor variables or also known as independent variables.

2. Looking into the fields in the univariate analysis, we see there are outliers that needs to be treated.

3. Multivariate analysis indicates that there is strong positive correlation between the target variable "sales" and the predictor variables "randd" and "employment" and also some of the independent variables like "patents" with "randd" and "capital".

4. "Employment" and "Randd" has very high positive correlation with "Sales".

5. For bivariate analysis We can plot Bar plots and Count plots with target variables and these independent variables to see how they are behaving with the target variable.

**Bar graphs and scatter plots of distribution of various attributes:**

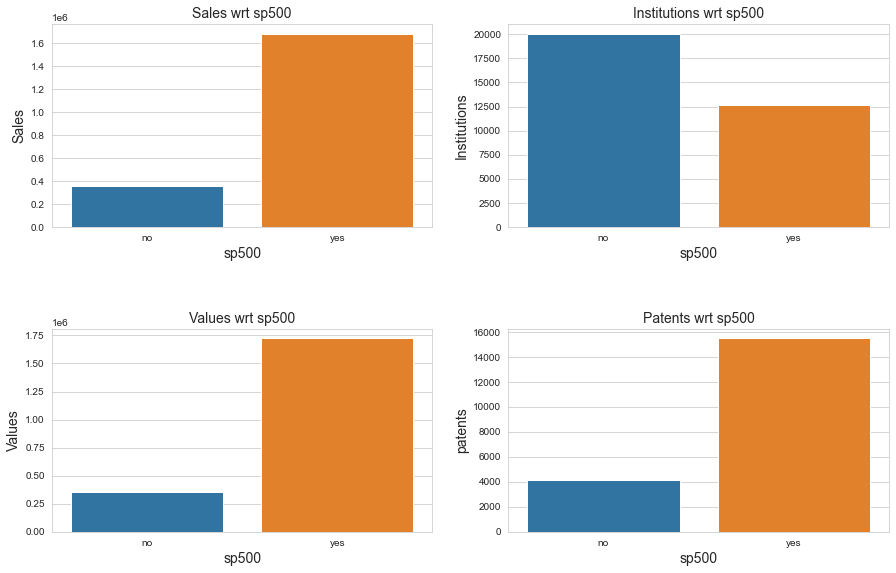
****

fig -9 Bar graphs of the few variables

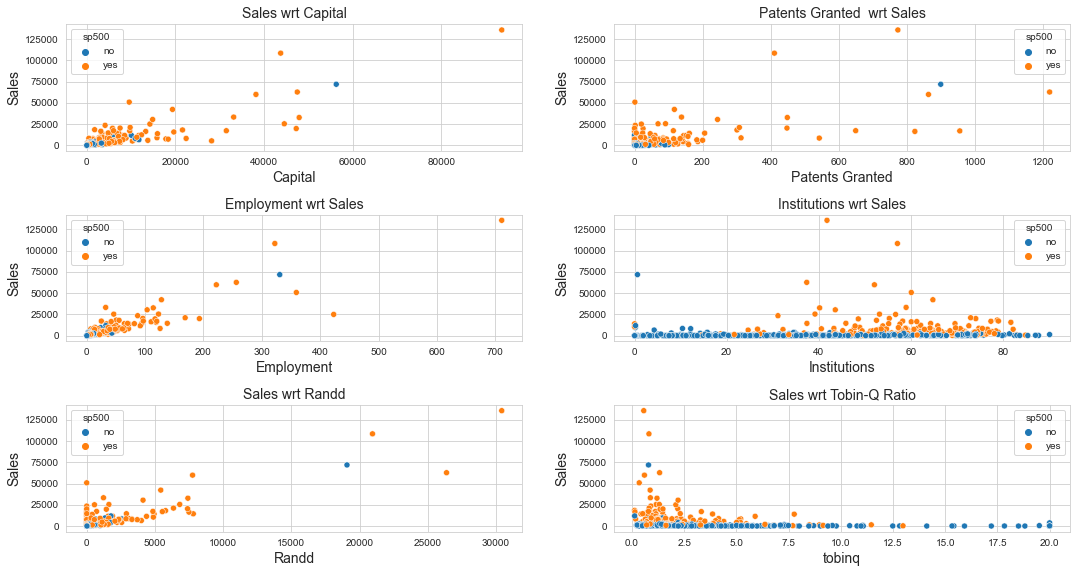
****

fig -10 Scatter plots of the various attributes

**Observations based on Bivariate & Multivariate Analysis:**

Inference: As per the Bivariate and Multivariate analysis these are the following observations:

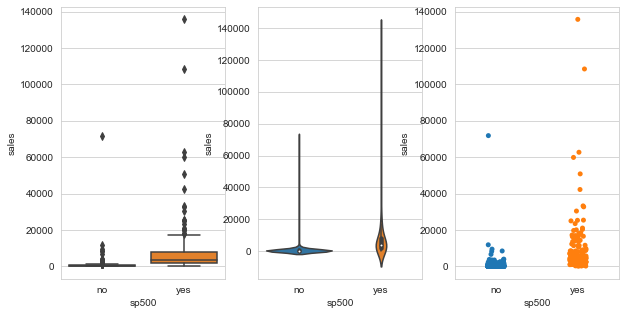
1. Based on the scatter plot we can see that Employment and sales are having positive correlation graph where the firms having the Membership in the S&P 500 index are much more in the ratio as compared to firms not having the membership.

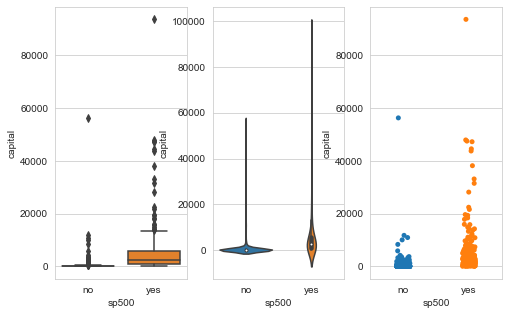
2. Institutions has no correlation as such with target variable sales but when we see the bar graph, we can see that the ratio of firms having no membership in S&P500 index is higher in this case as compared to all the other predictor variables like values, patents and employment.

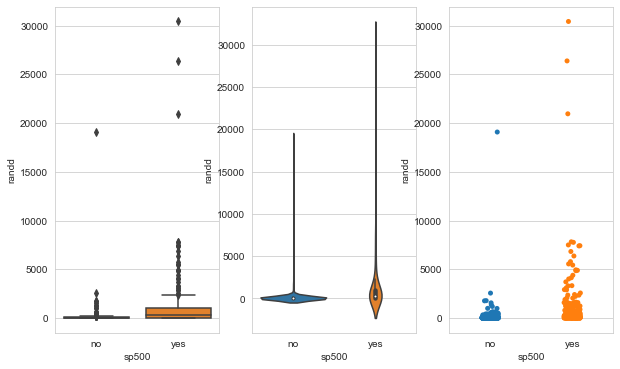
3. Sales shows a positive correlation with randd that is R&D stock but the firms having membership in S&P 500 index are much more than those firms who are not having the membership.

4. Patents has also similar effect on sales as that of capital and employment.

**Multivariate Analysis:**





****

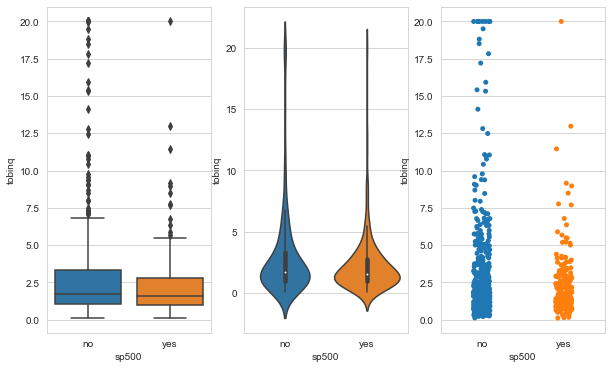
****

fig -11 Multivariate analysis of the various attributes

**Observations based on these triplet graphs including boxplots, strip plots and violin plots**

1. When it comes to sales the firms those are having the membership in S&P 500 index have much larger sales amount than those of the firms which are not having the membership. The maximum is almost near to 20000 in the former case.

2. In case of capital also these firms which are having the membership have much more capital which includes net stock of property, plants and equipment’s. This somehow explains their market growth, capital and investment ratio as well along with the large amount of sales. This further explains their entry in the list of S&P 500 index which actually includes the largest top 500 firms in US.

3. When it comes to randd which is R&D stock (in millions of dollars) the membership firms have invested an average of around 2500 million dollars and those firms which are not the members they have invested quite a low amount compared to the listed firms.

4. Now if we talk about Tobin's q which is the ratio between a physical asset's market value and its replacement value then the two types which are the firms having membership and the firms which are not having the membership are not having much of difference in this case as shown by the strip plot and violin plot.

**Part 1 (2) - Impute null values if present, also check for the values which are equal to zero. Do you think scaling is necessary in this case?**

Table 12 – Missing values % in the data set

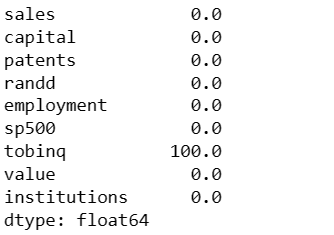
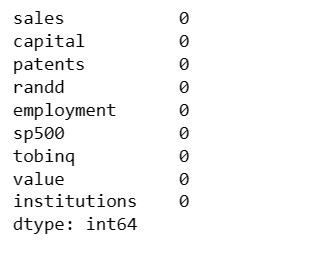


Table 13 – After missing values imputation



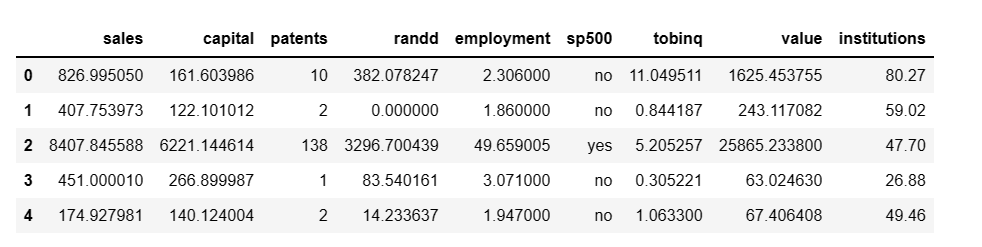
****

Table 14 – Data set After missing values imputation

**Summary after missing values imputation:**

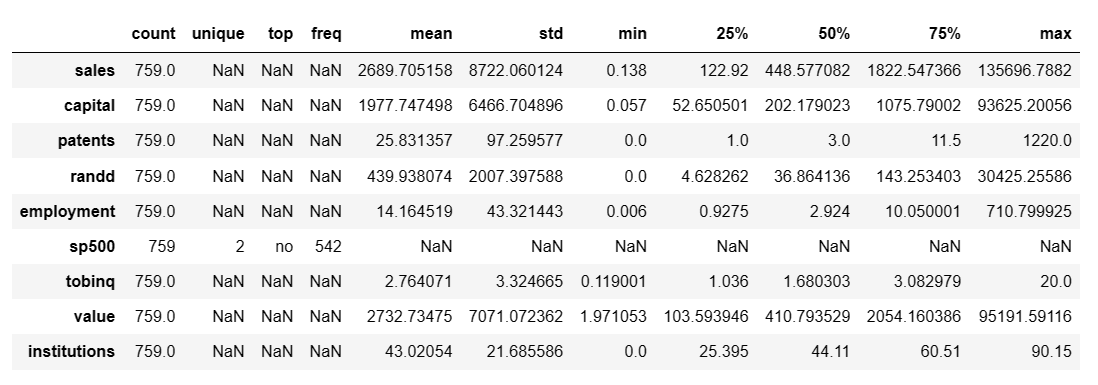
****

Table 15 – Description After missing values imputation

Number of duplicate rows = 0

**Treating the outliers:**

As we have already seen that outliers are present in the dataset, we need to treat those outliers as they can affect the modelling algorithm. So, we will treat the outliers using IQR method and then again check the 5-point summary for the continuous variables.

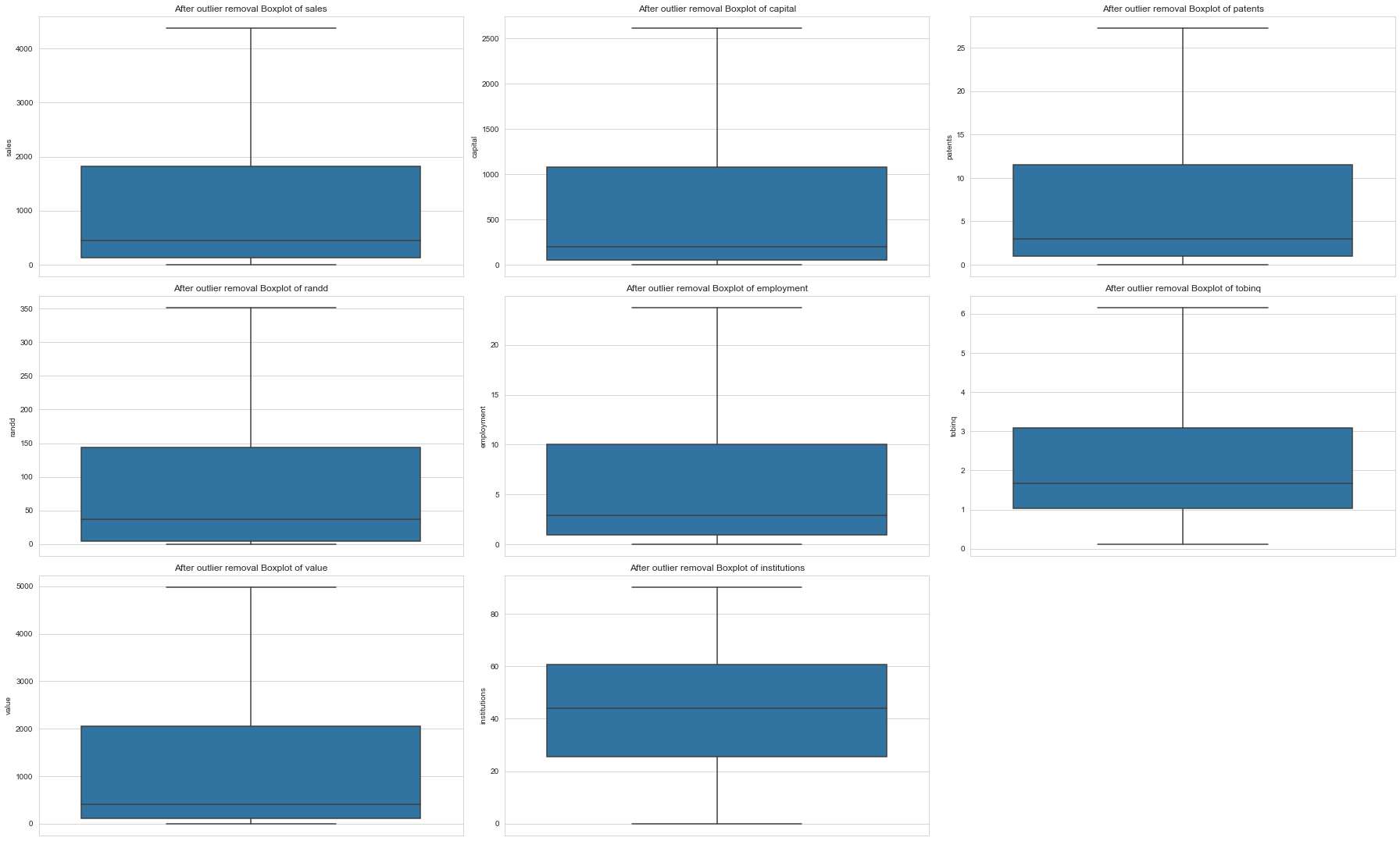


Fig 12 – Boxplots after outliers’ treatment

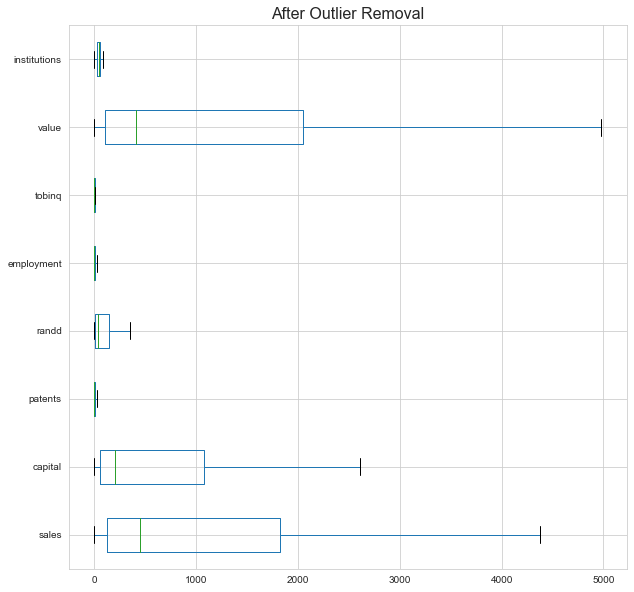
****

Fig 13 – Boxplots after outliers Treatment in One graph for comparing

For values which are equal to Zero (0)

* After treating the outliers, we can see that almost all the variables have their minimum values as zero (0).
* Removing the records with 0 values will not be a good idea in this case as it might have an impact on the model building

**Unique values in the variables after null value imputation and outlier treatment:**

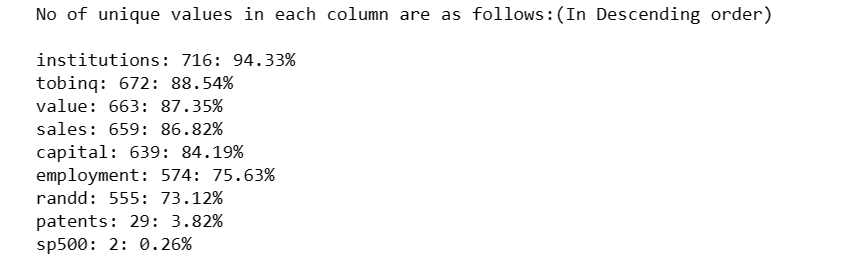
****

Table 16 – Unique values % After missing values imputation

**Scaling of the data set - Required or Not?**

As we have observed the summary of our data set, we can see that different attributes are in different scale or magnitude, For Example

* patents are whole numbers with lower range, whereas sales capital and values are decimal numbers having significantly larger range
* tobinq is a ratio converted in decimal numbers and similarly all other variables are in different scale

Hence scaling will be required because the linear regression model does distance calculation for training the model which is further utilised for testing. So, the scaled data which will have similar range will make the distance calculation easier because the comparison will be on the same scale.

**Before Scaling the data set:**

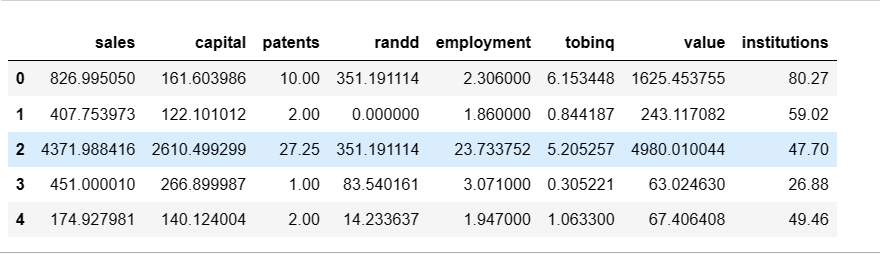
****

Table 17 – Numerical Data set before scaling

**Data Distribution range before scaling the data set:**

Here we are using violin plot to showcase the distribution of data set and the various ranges of all the numerical attributes. After scaling the data set, we will again see the violin plot and compare the two plots so that we can find out the difference in the distribution range.

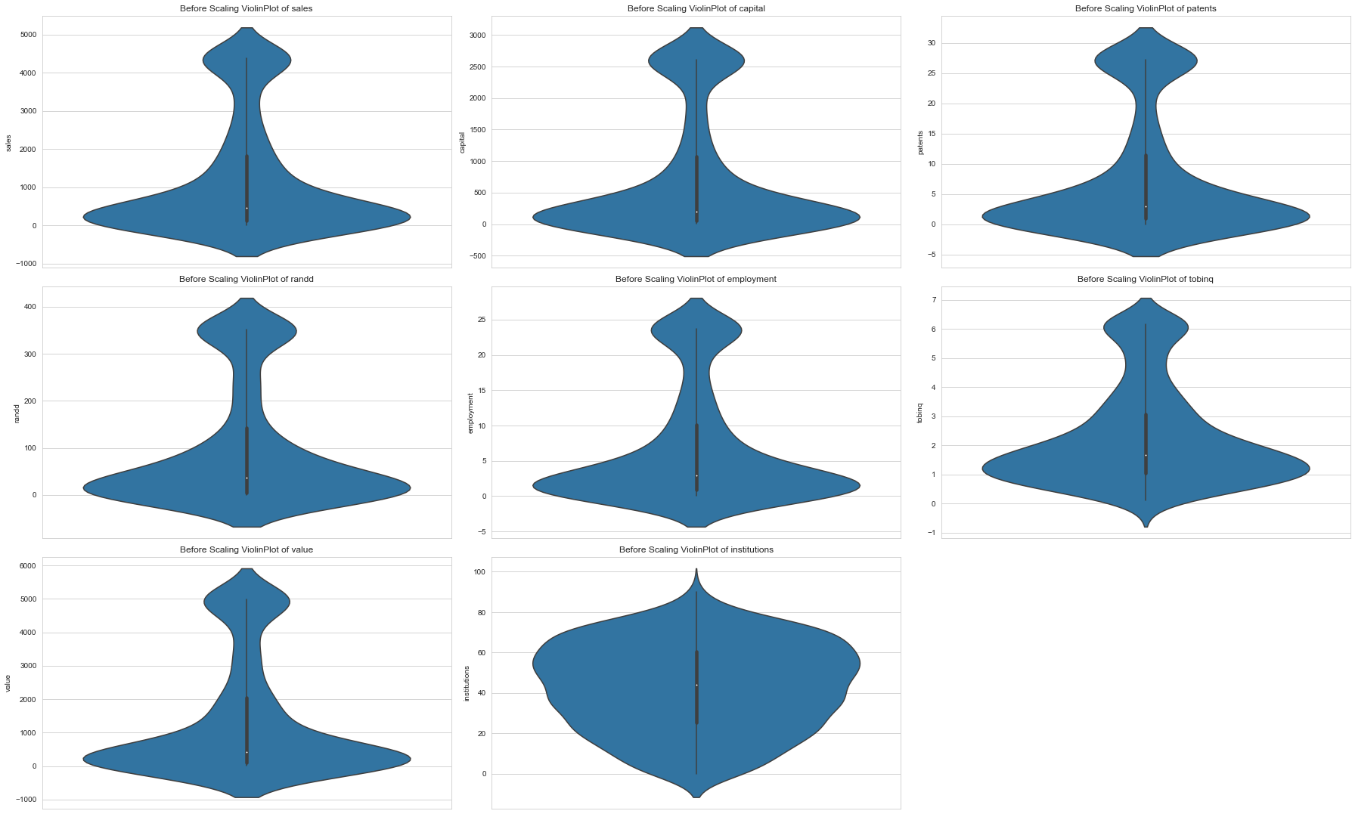


Fig 14 – Data distribution range before scaling- violin plots

**Observations:**

1. We can see that each attribute has different range of values stored in it so its difficult to compare the distances calculated.

2. We need to perform Feature Scaling when we are dealing with Gradient Descent Based algorithms (Linear and Logistic Regression, Neural Network) and Distance-based algorithms (KNN, K-means) as these are very sensitive to the range of the data points.

3. Now we will see what is happening after scaling the attributes.

4. We will first do it just for the Numerical data set named as "Firm\_df\_num" and we will analyse the transformation and then we will do it for the complete data set after Encoding the data det before splitting the data set.

**After scaling the data set:**

Min - Max Scaler

Transform features by scaling each feature to a given range.

This estimator scales and translates each feature individually such that it is in the given range on the training set, e.g., between zero and one.

The transformation is given by:

X std = (X - X.min(axis=0)) / (X.max(axis=0) - X.min(axis=0))

X scaled = X std \* (max - min) + min

where min, max = feature\_range. - Equation1

1. This transformation is often used as an alternative to zero mean, unit variance scaling.

2. We can use Min-Max scaler here in this case since we have a categorical variable in our data set which we need to encode first and then split the data into train and test model and then we can do scaling of both train and test data individually.

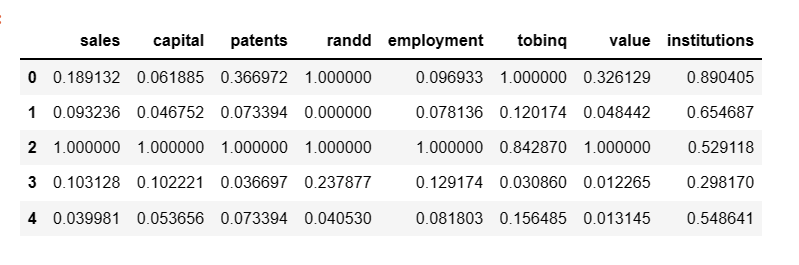


Table 18 – Numerical Data set after scaling

**Data Distribution Range after scaling the data set:**

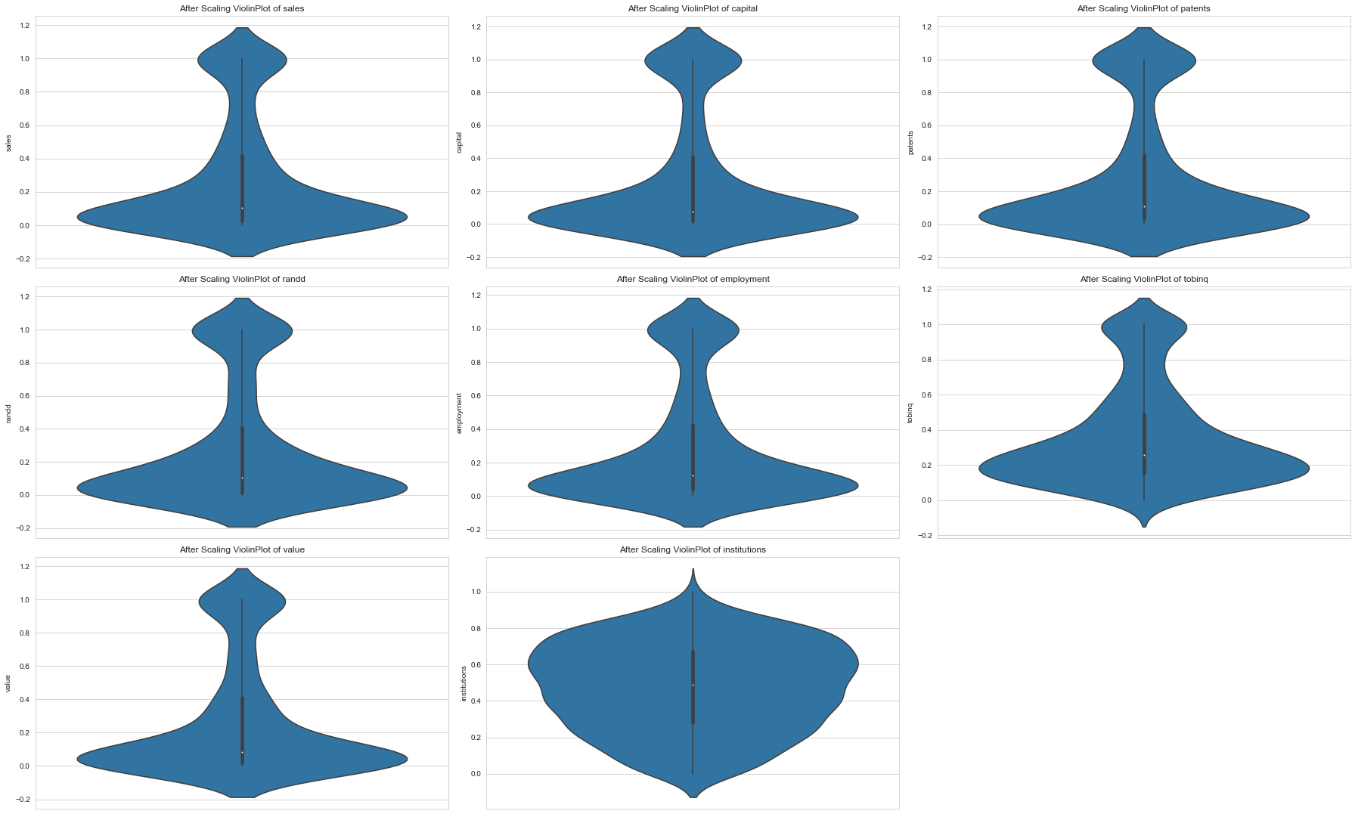


Fig 15 – Data distribution range after scaling- violin plots

**Scaling the data set does not affect the output of the model anyways, although it can affect the interpretability of the coefficients and their effect on the target variable. For that we can reverse the scaling and then interpret the beta coefficients**

**Part 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 stats model. 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:**

There are two types of categorical data:

- Ordinal: Order based like 'good', 'bad',' worst' like Clothing sizes

- Nominal: Without any order or ranks like city names, Genders, etc.

Here we will use "One Hot Encoding" because we don’t have the problem of high dimensionality in this case as it’s a Boolean type of categorical variable. So, we will Encode the attribute "sp500" which is in string format to integer format

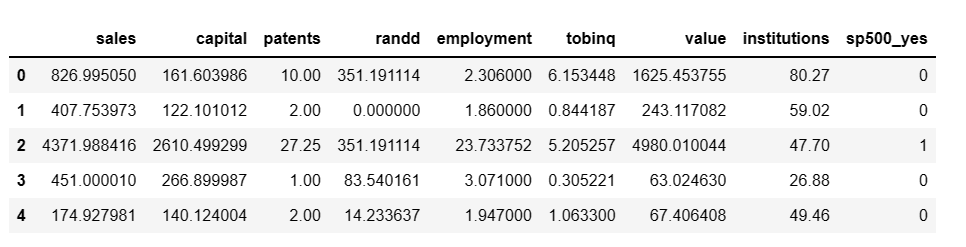


Table 19 – Data set After dummy variable encoding

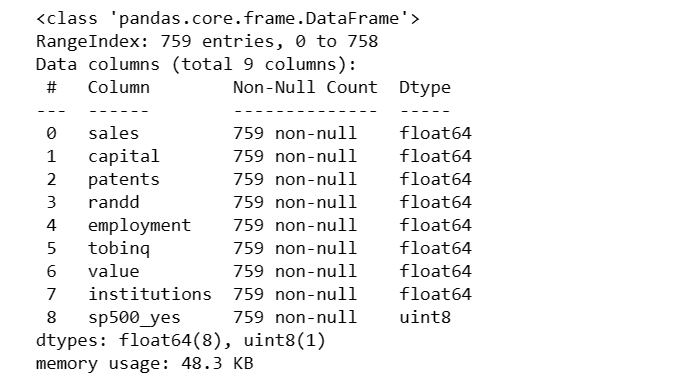


Table 20– Information After Encoding the data set

### After encoding the data set, we can observe that the attribute "sp500" has been converted into 8-bit unsigned integer data type which has values either 0 or 1 and the column is named as "sp500\_yes". Now we will scale the complete data set after Encoding again and then we will proceed with the splitting and modeling.

### Since we only scaled the numerical data set earlier to see how scaling helps in setting the range on a similar scale for each attribute. Now we will scale the completely Encoded data set on train and test data and then perform the modeling on this scaled data set.

### Preparation of the data: Checking the shape of Training and Test data sets respectively after scaling:

### Train – Test split & Model Building

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

### Now let’s split the data set in the ratio of 70:30 first where 70% is for Training data set on which the model will be built and 30% for the test data set on which the performance will be measured.

### If suppose the number of data points are very less on which our model is built then we have to change the ratio let's say 60:40 or 80:20 as it might give a better performing model.

### 3. We will check the shape of the data set on which we are going to build our model and this will help us in developing a better performing model.

### Shape of X train scaled data set = 531 rows and 8 columns

### Shape of X test scaled data set = 228 rows and 8 columns

### From the above findings we see that after splitting our data set in the ratio of 70:30 we have our Training data set with 531 rows and Test data set with 228 rows, which is quite good number for developing a model. So, we will go ahead with these data sets and will develop our model.

### Linear Regression Model –

Evaluation of Linear regression model-Evaluation helps to judge the performance of any machine learning model that would provide best results to our test data. Fundamentally three types of evaluation metrics are used to evaluate linear regression model. -R2 measure (discussed with least square method)-Mean Absolute Error (MAE)-Root Mean Square Error (RMSE) Mean Absolute Error (MAE)-Mean Absolute Error is the average of the difference between actual and predicted value of target variable.

𝑀𝐴𝐸=1𝑛∑|𝑦𝑖−𝑦ℎ𝑎𝑡𝑖|𝑛𝑖=1 – Equation 2

Root Mean Square Error (RMSE)-defined as:

𝑅𝑀𝑆𝐸=√1𝑛∑(𝑦𝑖−𝑦ℎ𝑎𝑡𝑖)2𝑛𝑖=1 – Equation 3

Pros and cons of Linear Regression: Pros-Linear regression models are very simple and easy to implement. These models are said to be most interpretable. Cons-Linear regression models are largely affected by the presence of outlier in training data.

1. We will first invoke the Linear Regression function and find the best fit model on training data.
2. Then we will explore the coefficients for each of the independent attributes.

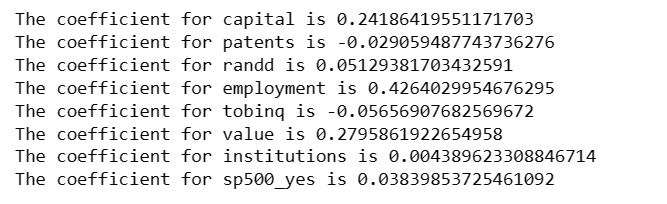


Table 21 – Coefficients of variables for Linear Regression

**Performance Metrics**

Performance measures are a way to evaluate and compare our business models and to decide which model works well for the business scenario. To evaluate our Linear Regression Model, we will take two measures namely, R square and RMSE which we will compute for both train and test datasets. R-squared (R2) is a statistical measure that represents the proportion of the variance for a dependent variable that is explained by an independent variable or variables in a regression model.

Root Mean Square Error (RMSE) is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are; RMSE is a measure of how spread out these residuals are. In other words, it tells you how concentrated the data is around the line of best fit.

**Following are the performance metrics of our data set on Model 1 using sklearn method:**

1. The intercept for our model is 0.01594262016671344
2. R square on training data = 0.9359702538559449
3. R square on testing data = 0.9240311293641785
4. RMSE on Training data = 0.09019914421833766
5. RMSE on Testing data = 0.08611411139061899

### Insights:

### R-squared is always between 0 and 100%:

### 0% indicates that the model explains none of the variability of the response data around its mean.

### 100% indicates that the model explains all the variability of the response data around its mean. In general, the higher the R-squared, the better the model fits data.

### 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: 0.935

### RMSE for Train data: 0.090

### This is a very good value when it comes to performance of the model. This shows that almost 93% 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: 0.924

### RMSE for Test data: 0.0861

### This is also quite a good value. This shows that almost 92% of the variance of the testing dataset was captured by the model.

### The model seems to be a good model as it is neither overfitting nor under-fitting, so we can go with this model. Although, we need to see if there is any improvement with the stats model approach.

### Linear Regression using stats models (OLS):

### Model 1 - Using all the independent variables

### Performance on Train data

Table 22 – OLS for regression model -1 on Training data

### 

### 

### Model 1

### Performance on Test data:

Table 23 – OLS for regression model -1 on Test data

### 

### 

### Model 1- Stats model method

### Interpretation of p-values (P > |t|)

### For each predictor variable there is a null hypothesis and alternate hypothesis.

### Null hypothesis: Predictor variable is not significant

### Alternate hypothesis: Predictor variable is significant

### (P > |t|) gives the p-value for each predictor variable to check the null hypothesis

### If the level of significance is set to 5% (0.05), the p-values greater than 0.05 would indicate that the corresponding predictor variables are not significant.

### However, due to the presence of multicollinearity in our data, the p-values will also change.

### We need to ensure that there is no multicollinearity in order to interpret the p-values.

### If we build the model using stats model and OLS method, we see that the Adjusted Rsquare is equal to the R-square value which is 0.93.5.

### This shows that this is not coincidental model and we can say that there are no sampling error present.

### Now, looking at the p-values of the predictors, we see that variables like ‘patents’ and 'institutions' have a p-value greater than 0.05. So, we can eventually say that there is no relation as such between these variables and the target variable, hence these attributes cannot be that not useful in prediction. However, we need to see other factors such as VIF and Multicollinearity effect before we conclude anything about the model.

### Insights:

### Null hypothesis (Ho) is true i.e., there is no relation between dependent and independent variable in the universe, where "sales" is dependent variable and others like "patents", "capital", "employment", "randd", "values" and "sp500" are independent variables.

### For example, all the variables are showing p value below the 0.05 (alpha), but some variables like patents and institutions are showing some p value greater than 0.05, if more sample is collected, we can predict better if these variables are good predictor of sales or not.

### For now, capital, randd, tobinq, employment and values are good predictor of sales. As we can see variable tobinq and patents have negative co-efficient, which means higher the value of "patents" and "tobinq", the sales will be lower.

### Interpretation of Coefficients

### - The coefficients tell us how one unit change in X can affect y.

### - The sign of the coefficient indicates if the relationship is positive or negative.

### - Multicollinearity occurs when predictor variables in a regression model are correlated. This correlation is a problem because predictor variables should be independent. If the collinearity between variables is high, we might not be able to trust the p-values to identify independent variables that are statistically significant.

### - When we have multicollinearity in the linear model, the coefficients that the model suggests are unreliable.

### 

Fig 16 – Scatter plot for actual vs predicted

### The Sklearn model:

### - The R-squared / Co-efficient of determinant (Training data) = 0.935

### - The R-squared / Co-efficient of determinant (Test data) = 0.924

### - RMSE/ Root mean squared error (Training data) = 0.090

### - RMSE/ Root mean squared error (Test data) = 0.860

### - As the training data & testing data score are almost inline, we can conclude this model is a Right-Fit Model.

### The Stats models:

### - The R-squared / Co-efficient of determinant and Adjusted R –squared are same = 0.935- VIF (Variation Inflation Factor): - There is high multi-collinearity between the independent variables.

- The coefficient of constant that is b0 is equals to 0.02

- When capital increases by 1 unit, sales increases by 0.24 units keeping all other predictors constant.

- When employment increases by 1 unit, sales increases by 0.43 units keeping all other predictors constant.

- When randd increases by 1 unit, sales increases by 0.05 units keeping all other predictors constant.

- When patents increase by 1unit, sales decreases by 0.03 units keeping all other predictors constant.

- Similarly, 1 unit increase of value, increase the sales by 0.28 units

### VIF

### Equation4

### Reviewing Linear Regression: Now we will check the VIF of the predictors

### 

### Table 24 – VIF values for variables

### Observations:

### 1. The VIF values indicate that the features "capital","employment","sp500" and "values" are a bit correlated with one or more independent features.

### 2. Multicollinearity affects only the specific independent variables that are correlated. Therefore, in this case we can also see the p values of these coefficients and can figure out that variables "capital" and "employment' can play important role and rather "Tobinq" may have some effect of multicollinearity.

### 3. To treat multicollinearity, we will have to drop one or more of the correlated features, and we will form different models and see which 5 features are performing the best for the model.

### 4. We will drop the variable that has the least impact on the adjusted R-squared of the model.

### MODEL DEVELOPMENT:

### Now we will drop few columns which are having p value higher than 5% or those who are having VIF higher to treat multicollinearity and then we will try to check the performance of various models after dropping the non-significant columns and then we will compare those models with model 1 which has all the features or attributes present in it.

### Model 2 - After dropping the variables "patents","institutions","sp500\_yes", "value", "tobinq" and "randd"

### Performance on Train data

### As we know that "capital" and "Employment" are highly significant when it comes to target variable "sales", and can be a strong predictor of "sales" so let’s make a model which includes only these two variables and see the performance of the model. Also, we can compare this model with model 1 which has all the variables.

* **R-squared: 0.919**
* **Adjusted R-squared: 0.918**

### 

Table 25 – OLS for regression model -2 on Train data

### Model 2 –

### Performance on Test data set

* **R-squared: 0.906**
* **Adjusted R-squared: 0.906**

**Observation:**

1. As compared to the first model where R2 was 91.9%, by just using capital and employment we have achieved R2 91.8%.
2. Thus, the second model with only two variables capital and employment explains almost 92% of response variable variation, which is just 1.5% less of full model (including all dependent variables).
3. Also, if we see the performance on the test data it gives the R-squared value as 90.6%, which is almost similar to the train data. This explains that our model is stable enough with these variables.

**Model 3 - After dropping variables 'patents', 'institutions', and 'tobinq' and including all the other variables**

Performance on Train data set:

* **R-squared: 0.934**
* **Adjusted R-squared: 0.933**

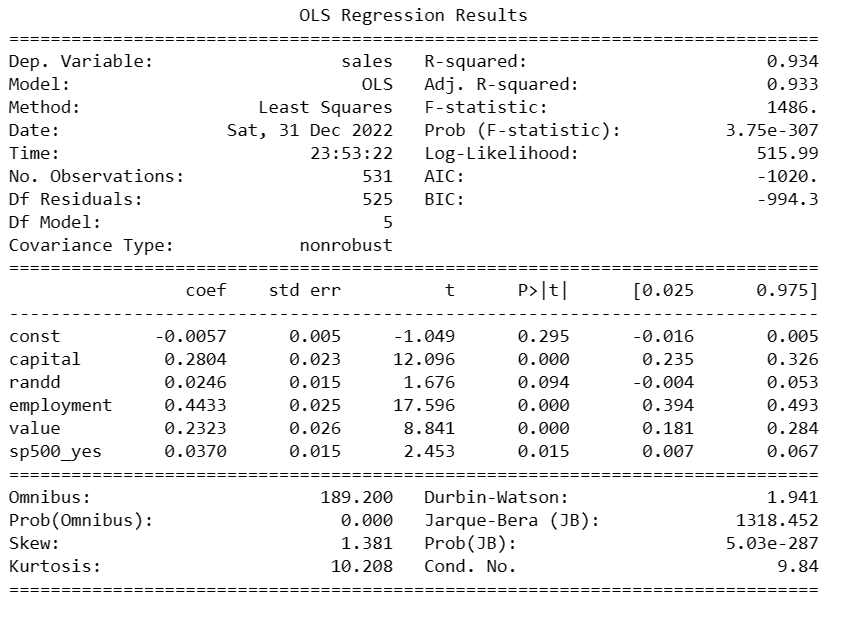
****

Table 26 – OLS for regression model -3 on Train data

**Model 3**

**Performance on Test data set:**

* R-squared: 0.93
* Adjusted R-squared: 0.928

**Observation:**

1. As compared to the first model where R2 was 93.6%, by just using capital and employment we have achieved R2 93.4%.
2. Thus, the third model with only 5 variables capital and employment explains almost 93.4% of response variable variation, which is just 0.4% less of full model (including all dependent variables).
3. The model performance for train data is similar as compared to our earlier models but at the same time, the model is performing very well for the test data or the unseen data as the R-squared value comes out to be 93%, Which means the model is reliable and its performance can be predictable when the model is deployed.

### Model 4 - After dropping variables 'patents', 'institutions', and 'value' and including all the other variables

### Performance on Training data set

* R-squared: 0.935
* Adjusted R-squared: 0.934

### Model 4

### Performance on Test data set:

* R-squared: 0.931
* Adjusted R-squared: 0.93

### Observation:

### When we compare the first model where R2 was 93.6%, by just using these 5 variables we have achieved R2 93.5%.

### Thus, the fourth model with only 5 variables capital, tobinq, value, sp500\_yes and employment explains almost 93.5% of response variable variation, which is just 0.1% less of full model (including all dependent variables).

### This shows that this model can be a really good model when we want a similar model as model 1.

### Also, the model performance for train data is similar as compared to our earlier models but at the same time, the model is performing very well for the test data or the unseen data with R-squared value equals to 93.1%.

### This shows that the model is reliable and its performance can be predictable when the model is deployed because both on train data and test data model is doing quite good and not deteriorating much.

### Model 5 - After dropping variables 'randd', 'sp500\_yes',"value","employment" and 'capital' and including all the other variables.

### Performance on Train data:

* R-squared: 0.4
* Adjusted R-squared: 0.396

Table 27 – OLS for regression model -5 on Train data

### 

### Model 5

### Performance on Test data set:

* R-squared: 0.366
* Adjusted R-squared: 0.357

**Observation:**

* Where we have dropped the 5 important attributes randd, employment, value, capital and sp500\_yes, which are 5 important variables with high multicollinearity
* Also, when we compare the first model where R2 was 93.6%, by just using these 5 variables we have reduced the performance of model to quite a low value of 40%.
* Thus, the fifth model with only 3 variables, institutions, tobinq, and patents which has low values of VIF explains 40% of response variable variation, which is almost 50% less than the full model (including all dependent variables).
* We can see that the model has become quite unstable post dropping the variables with high multicollinearity because 5 important variables are gone after that.
* The performance on Test data set has also deteriorated and became unstable with R-squared value 36.6%. So, this idea of removing attributes using VIF can not to be a good way to deal with multicollinearity.
* To get rid of the multicollinearity, we can also look at how we can apply PCA for dimensionality reduction

### Model 6 - After dropping variables 'patents', and 'institutions' and including all the other variables.

### Performance on Train data set:

* R-squared: 0.936
* Adjusted R-squared: 0.935

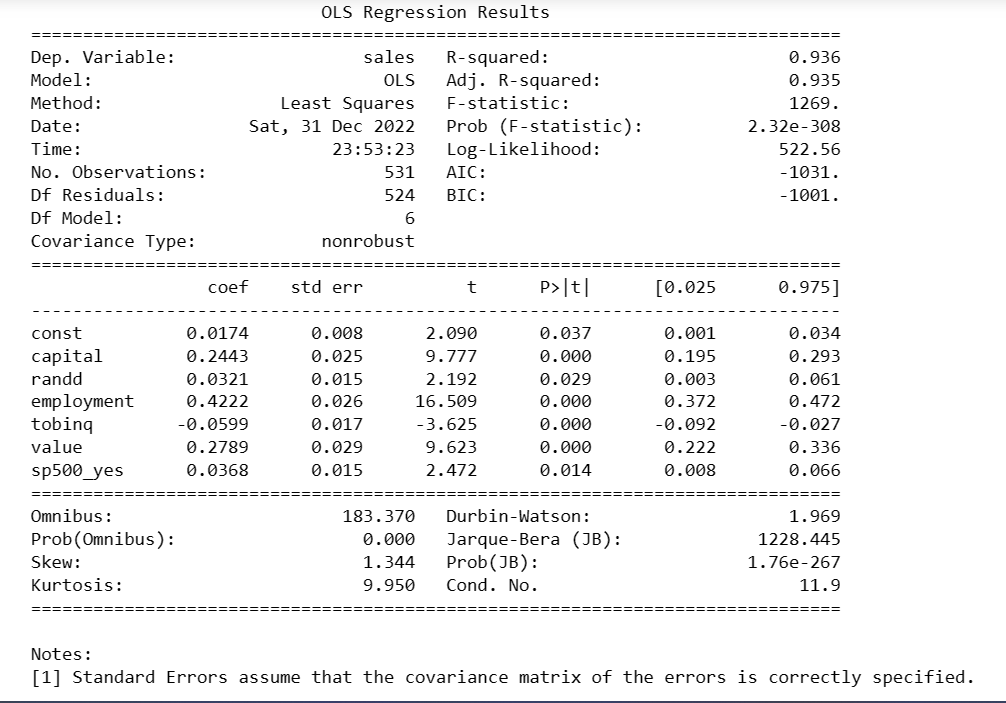


Table 28 – OLS for regression model -6 on Train data

**Model 6**

**Performance on Test data:**

* R-squared: 0.932
* Adjusted R-squared: 0.93

**Observation:**

* We have dropped the two attributes’ patents, and institutions we get a very stable model as the R square is equals
* to 93.6%
* So, when we compare the first model where R2 was 93.6%, by just using these 6 variables we have achieved R2 93.6% which is exactly equals to the first model.
* Thus, the sixth model with only 6 variables capital, tobinq, value, randd, sp500\_yes & employment explains all the 93.6% of response variable variation, which exactly equals to that of full model (including all dependent variables).
* This shows that this model can be a really good model when we want a similar model as model 1. Now, if we compare the model performance for train data and test data, the model is performing very well for the train data and the test data or the unseen data with R-squared value equals to 93.2% even with just 6 of the given attributes.
* This shows that the model can also be considered for prediction of the business model and its performance can be predictable when the model is deployed because both on train data and test data model is doing quite good and not deteriorating much.
* Now that we do not have multicollinearity in our data, the p-values of the coefficients have become reliable and we can remove the non-significant predictor variables.

### 

Table 29 – Model comparisons

### Model Selection

### Model for prediction If we compared all the models then Model 1 and Model 6 are doing better with respect to prediction. To select best among them it would be better to have more data for training, validating and testing. As of now, Model 6 looks to be more balanced. However, Model 3 not that is bad as it is giving similar result but will require further analysis with more data. Model for prescriptive analysis Model 6 with only 6 independent variable is most suitable. Model performance very close to full model. It is the simplest model with no transformation and with least variable. There is no multicollinearity, as the independent variable in the model are not correlated among each other as observed in EDA section.

### Linear regression Model 1 including all the variables:

**Before dropping the columns, this states our linear model which can be written as:**

### Sales = (0.02) \* const + (0.24) \* capital + (-0.03) \* patents + (0.05) \* randd + (0.43) \* employment + (-0.06) \* tobinq + (0.28) \* value + (0.0) \* institutions + (0.04) \* sp500\_yes.

**Insights:**

• When capital increases by 1 unit, sales increases by 0.24 units keeping all other predictors constant.

• When employment increases by 1 unit, sales increases by 0.43 units keeping all other predictors constant.

• When patents decrease by 1unit, sales increases by 0.05 units keeping all other predictors constant

**After removing the multi collinearity and dropping the columns these are what we are getting as coefficients:**

Model 6: After removing two attributes’ Patents and Institutions

**Sales = (0.02) \* const + (0.24) \* capital + (0.03) \* randd + (0.42) \* employment + (-0.06) \* tobinq + (0.28) \* value + (0.04) \* sp500\_yes**

**After solving the multicollinearity problem, we are getting the following parameters and model:**

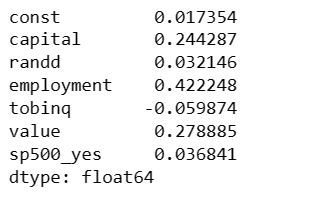
** **

Table 30– Parameters of fitted value

Table 31 – Model 6 Parameters and their importance

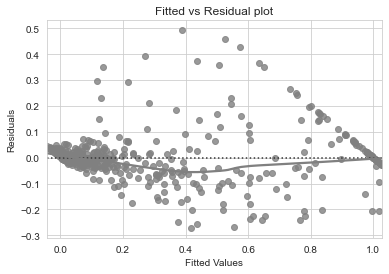
****

Fig 17 – Fitted vs Residual plot after multicollinearity treatment

**We observe that the pattern has slightly decreased and the data points seems to be randomly distributed and the prediction error should not be linked to the magnitude of the value predicted.**

### Making predictions on the train and test set:

### let's make predictions on the train and test set

* The MAE on the train data = 0.051401174063236015
* The MAE on the test data = 0.04891925827549611

**Part 1(4) – Inference: Basis on these predictions, what are the business insights and recommendations. There should be proper business interpretation and actionable insights present.**

**Inferences of the Test data performance of model**:

1. We can see that RMSE on the train and test sets are comparable. So, our model is not suffering from overfitting.
2. MAE indicates that our current model is able to predict usr within a mean error of 0.051 units on the test data and 0.048 on the train data. The range of MAE is 0 to infinity. The lower bound zero is achieved if all your model's predictions are exactly correct, else it will give positive value depending on how worse your model's predictions are.
3. Hence, we can conclude the model "ols\_res11" is good for prediction as well as inference purposes.

**Business Insights:**

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

Here we are developing this Regression model, we have plotted the predicted y values vs the actual y values for the test dataset. This is the plot obtained.

1. 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.
2. We get the following Linear Regression equation from the final model:
3. (0.02) \* const + (0.24) \* capital + (-0.03) \* patents + (0.05) \* randd + (0.43) \* employment + (-0.06) \* tobinq + (0.28) \* value + (0.0) \* institutions + (0.04) \* sp500\_yes
4. The variable "Capital" has strong positive impact (0.24 units) on the sales implying more the stock of property and equipment’s, more is the prosperity of the firm in terms of increased sales.
5. Number of patents and tobinq have slight negative impact on the sales. 0.03 units of increase in the number of patents actually decreases the sales by 1 unit and similar is the case with Tobin’s q ratio.
6. Employment is the variable which has the highest contribution on the sales performance of the business firm whereas randd has positive impact on the sales as it increases 1 unit with 0.05 times increase in the R&D stock which is not that significant though when a small amount is considered but can make a significant impact if the amount increases.
7. Both Stock market value and Proportion of stock owned by institutions enhances the sales of the business where the value plays quite a significant role.
8. With low beta coefficient the current the above attribute should be able to predict the sales performance significantly in the future.

**Cause of concern and Areas of Improvement:**

1. Patents are something which boosts the confidence of the organisation employees and helps in achieving royalties but in this case the business is unable to get any benefit from the number of patents it is registering, which means that the company is not able to get any attention from the market for its patented products on basis of its innovations and uniqueness in the market and service quality. This should be cause of concern for the firm and company should definitely look into this matter to change the negatives into positives.

2. 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. This again shows a negative impact on sales showcasing that the firm has not good marketing strategy team and not able to get good deals for the replacement of physical assets. It can be also assumed that rather the company is compromising on its physical asset's quality and company is not very focused about the Three R's which has Reduce Recycle and Reuse. This issue again needs to be focused on.

**Recommendations:**

**Short-term strategy:**

1) The firm needs to work on these three attributes (Employment rate, capital increase and value of stocks) to increase the sales and focus less on other parameters, which are not able to influence the sales like no of patents granted and replacement of physical assets.

**Long-term strategy:**

1) The company needs to focus on increasing the significance of few attributes on the sales like R&D stock in long run so that it gives a strong support to increase the sales

2) The firm needs to find other features also with which they can influence the market in a better way and increase their sales. The institutions and sp500\_yes may not look a differentiator in this data but if a company establishes a brand in the market and gets an entry in the top 500 club along with the increase in the stock value it can help the business in the long run.

3) Further market research needs to be done to see how competitors are doing in the similar market and what are their attributes to increase the sales.

# PROBLEM STATEMENT– 2

## **About Data**

# Logistic Regression, LDA and CART

You are hired by the Government to do an analysis of car crashes. You are provided details of car crashes, among which some people survived and some didn't. You have to help the government in predicting whether a person will survive or not on the basis of the information given in the data set so as to provide insights that will help the government to make stronger laws for car manufacturers to ensure safety measures. Also, find out the important factors on the basis of which you made your predictions.

### Part 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 Exploratory Data Analysis**

# Table-32 -Car crash data set – Top 5 rows

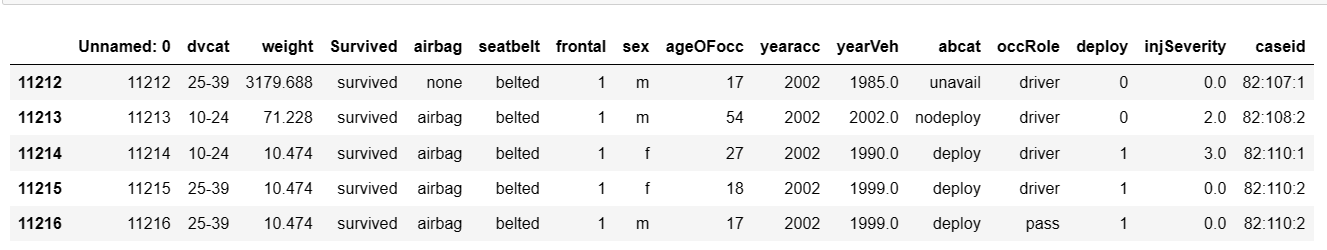


Table-33 -Car crash data set – last 5 rows

**Information and Null Values counts of the complete dataset**:

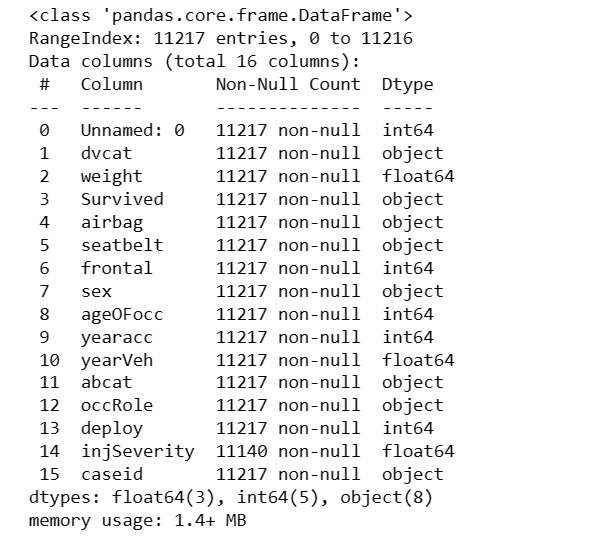
****

Table-34-Information of data set

**Inference:**

1. The data set contains 11217 rows and 16 columns, whose descriptions are mentioned above.
2. We can drop the column 'Unnamed: 0'as it is just the index in the form of column not going to contribute in our data analysis and its exploration.
3. Here we can see that 8 out of 16 columns are numerical in nature with 3 Float and 5 Int data type one of which is "Unnamed:0" and we are going to drop that before analysis.
4. In our data set we have 8 columns which are object data types including gender, seatbelt information and airbag information etc.

**Data Dictionary**

1. dvcat: factor with levels (estimated impact speeds) 1-9km/h, 10-24, 25-39, 40-54, 55+

2. weight: Observation weights, albeit of uncertain accuracy, designed to account for varying sampling probabilities. (The inverse probability weighting estimator can be used to demonstrate causality when the researcher cannot conduct a controlled experiment but has observed data to model)

3. Survived: factor with levels Survived or not survived

4. airbag: a factor with levels none or airbag

5. seatbelt: a factor with levels none or belted

6. frontal: a numeric vector; 0 = non-frontal, 1=frontal impact

7. sex: a factor with levels f: Female or m: Male

8. ageOFocc: age of occupant in years

9. yearacc: year of accident

10. yearVeh: Year of model of vehicle; a numeric vector

11. abcat: Did one or more (driver or passenger) airbag(s) deploy? This factor has levels deploy, no-deploy and unavailable

12. occRole: a factor with levels driver or pass: passenger

13. deploy: a numeric vector: 0 if an airbag was unavailable or did not deploy; 1 if one or more bags deployed.

14. injSeverity: a numeric vector; 0: none, 1: possible injury, 2: no incapacity, 3: incapacity, 4: killed; 5: unknown, 6: prior death

15. caseid: character, created by pasting together the populations sampling unit, the case number, and the vehicle number. Within each year, use this to uniquely identify the vehicle.

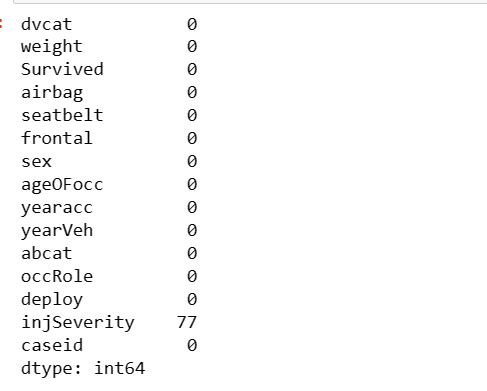
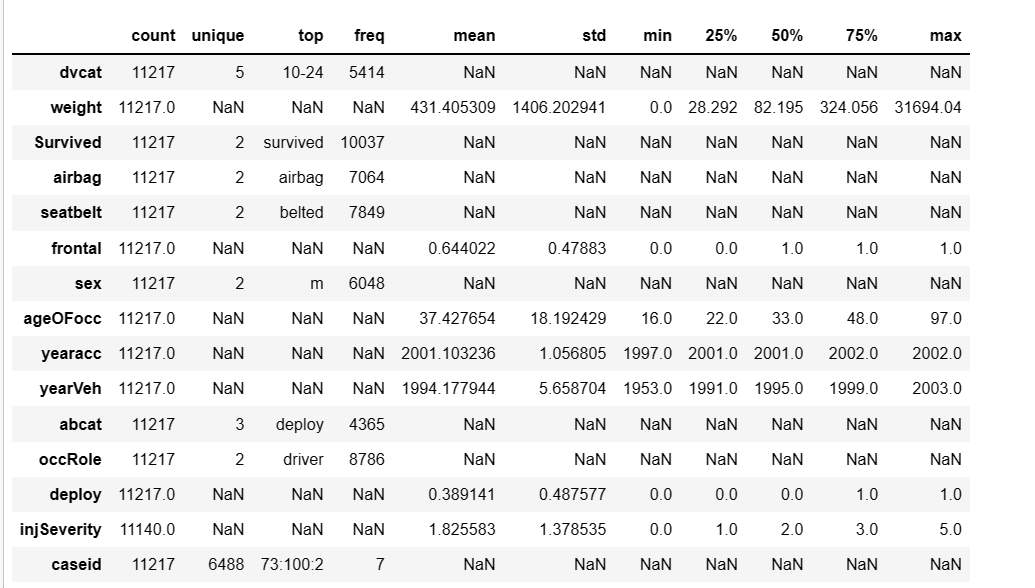


Table-35-Missing values in the data set

**Description of the complete dataset**:

Now let’s check the null and duplicate values of the given sample. If null values or duplicate data are present in the sample data set, we will treat them and if not, we will further proceed with the outlier’s detection and their treatment.

Table -36 Description of data set.



**Summary of dataset:**

1. The mean and median of weight shows a huge gap indicating a skew in the data or may be presence of outliers.
2. dvcat which represents the factor with levels (estimated impact speeds) shows that maximum frequency of speed range is between the speed group of 10-24km/h where the category has 5 unique types of groups.
3. ageOFocc which is age of occupant in years shows a mean value (37) greater than median (33) which is an indication of slight positive skew in the data. We will get more clarity on skewness once we plot density plots and find out skewness values.
4. The data indicates male gender is getting caught with accident more than Female gender.
5. abcat and deploy represents similar kind of data as both are talking about the deployment of airbags where deploy is Boolean in nature having values 0 and 1 and "deploy" has 3 categories including unavail, nodeploy and deploy
6. Frontal or non-frontal impact data is present in the column "frontal" having Boolean values 0 and 1, but the frequency is not very clear from the data description. We will further find out the unique values in the data set in the univariate analysis.

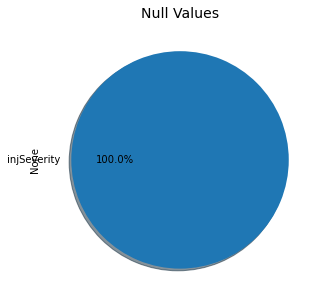


Fig 18 – Null values Percentage

**Observations:**

Inference based on Null value Detection and Duplicate rows Detection

* The null values present in the data set is less than 1% of the total observations present. So, we can either treat this data rather or can drop these observations depending upon the suitability. Here Since the Null values are present in the column "injSeverity", it can have an impact on the modelling and analysis part so we will rather impute the null values.
* Observed null values in only 1 field that is "injSeverity" and the number of null values is 77.
* We can impute the null values with mode of the data as the variable is categorical in nature although it’s a float data type.
* There are no duplicate rows in the data set.

### Univariate analysis:

### Let us define a function 'univariateAnalysis\_numeric' to display information as part of univariate analysis of numeric variables. The function will accept column name and number of bins as arguments.

### The function will display the statistical description of the numeric variable, histogram or density plot to view the distribution and the box plot to view 5-point summary and outliers if any.

Numerical feature levels frequencies



Table-37 - All numerical values in the dataset

### 

### 

Table-38 - All numerical columns & value counts in the dataset

### Unique Values Analysis:

### 

Table-39 - Unique value counts in the dataset

### Checking the Unique values in each categorical columns which has different factors present in it:

### 

Table-40 – Categorical columns in the dataset

### 

Fig 19 – Frequency distribution of dvcat

### 

Fig 20 – Frequency distribution of Survived

### 

Fig 21 – Frequency distribution of airbag

### 

Fig 22 – Frequency distribution of seatbelt

### 

Fig 23 – Frequency distribution of sex

### 

Fig 24 – Frequency distribution of abcat

### 

Fig 25 – Frequency distribution of occRole

### Inference based on Univariate Analysis of categorical columns:

### Estimated impact speeds showcase 5 levels 1-9km/h, 10-24, 25-39, 40-54, 55+ where 10-24km/h is largest occurring group which is equals to 5414.

### The frequency of Male drivers or passengers are a bit higher than female drivers in the given data set and in that too those who have survived the accident are much larger more than 10000, than those who have not survived the accident. This ratio is almost 1/10th.

### One of the reasons of survival can be frequency of airbags being higher than that of non-airbags and in that too belted drivers or passenger representing the larger number which is equal to 7849 out of total.

### The numbers of drivers engaging and getting the impact of accident are much larger than that to passengers.

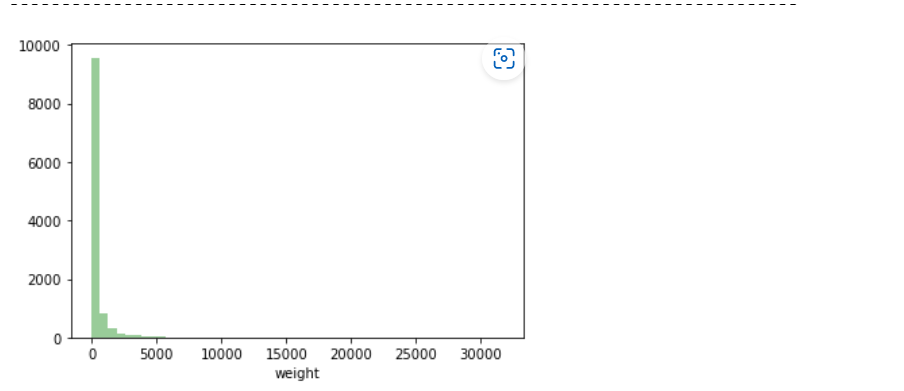
### Caseid representing character, created by pasting together the populations sampling unit, the case number, and the vehicle number. Within each year, use this to uniquely identify the vehicle hence it has almost 6400 unique samples.

### Almost 4000+ (~ 40%) of the vehicles did not have airbags and almost 3000+ (~ 30%) passengers or drivers in vehicles were not belted

### More than 60% of vehicles wither did not deploy the airbags or the airbags were not available

### Univariate analysis of Numerical columns:

### 

 Fig 26– Histogram of Weight

### Fig 27– Boxplot of Weight

### 

### Fig 28– Histogram of frontal

### Fig 29– Boxplot of frontal

### 

### Fig 30– Histogram of ageOFocc

### Fig 31– Boxplot of ageOFocc

### 

### Fig 32– Histogram of yearacc

### Fig 33– Boxplot of yearacc

### 

### Fig 34– Histogram of yearVeh

### Fig 35– Boxplot of yearVeh

### 

### Fig 36– Histogram of deploy

### Fig 37– Boxplot of deploy

### 

### Fig 38– Histogram of injSeverity

### Fig 39– Boxplot of injSeverity

### Inference based on Univariate Analysis of Numerical columns:

### The Observation weights is highly positively skewed column with minimum value as 0 and maximum more than 30000 having a large number of outliers.

### The Frontal column is actually categorical in nature having Boolean values that is either 0s or 1s, where 1 is representing the Frontal impact and its more in number than those of non-frontal impacts.

### Age of occupant in years is also a positively skewed curve with few numbers of outliers whereas year of accident is highest in case of year 2002 and is slightly increasing over the years and hence the plot is showing a negative skew.

### In case of column deploy 0 is representing if an airbag was unavailable or did not deploy and 1 is representing if one or more bags deployed and the data set has more 0s than 1s.

### Last but not the least the column "injSeverity" describes the after effect of accident from 0 to 6 where 3 is highest in number which is representing the "incapacity" after the accident.

### More than 7000 (~ 65%) crashes involved frontal impact

### In terms of injury severity 10% of the accidents were fatal

### Histograms plots of various variables and their analysis:

### 

### Fig 40 – Histograms and KDE of all variables

### Checking Skewness in the dataset

### 

Table-41 – Skewness in the car crash data set

### Outliers and their Treatment:

### 

Fig 41– Boxplots before outliers Treatment

### Inference based on Boxplots, histograms and skewness of data:

### 1. The data for almost all the columns shows that the data is positively skewed except for "frontal", "yearacc" and "yearVeh" which are negatively skewed. The highest Skew is present in the variable "weight" which is almost around 11.11

### 2. The data for "dvcat" is having few peaks where the group 10-24 represents the highest peak.

### 3. The newer vehicles have suffered more accidents and crashes as the density plot is negatively skewed.

### 4. The column "injSeverity" is showing a multimodal kind of curve where level 3 is representing the highest mode.

### 5. 4 of the numerical columns have outliers present in it named -"ageOFocc" representing age of occupant in years, “yearacc ‘which is year of accident, "yearVeh" and "weight".

### Bivariate and Multivariate analysis

### 

Table-42 – Correlation table of the variables

### 

Fig 42 – Pair plot of numerical variables

### From the Bivariate analysis and pair plot we can say that There does not seem to have multicollinearity in the data as the correlation is not that strong between various variables.

### Multivariate EDA of various columns:

### Sum of Weights based on Airbag Type and The Gender of Drivers using bar plots and pie charts and analysis based on that result

### 

Fig 43 – Pie chart of airbags, weights and Gender

### 

Fig 44 – Bar graph of caseid

### 

Fig 45 – Multivariate analysis of target variable

### 

Fig 46 – Facet grid for dvcat, injSeverity and airbag

### 

Fig 47 – Facet grid for dvcat, injSeverity and seatbelt

### Observations based on above Bivariate and Multivariate analysis:

### We can see that as the impact speed increases there are less chances of survival in case of a crash with as the highest not survived %age is for speed more than 55 km/h followed by 40-54 km/h and then with 25-39km/h. Although the maximum crashes has been for speed range 10-24km/h still the survival rate is high for that group reason being low speeding. Hence, following speed limit's protocol is must to avoid crashes.

### The chances of survival decreases in case of no availability of airbag or airbag not deployed

### Frontal impact is also an important factor however the survival %age is higher in the frontal impact than non-frontal impact if the airbags are deployed and speed limit is normal.

### So, we can say that, Frontal impact, Severity of the injury, speed impact level, seatbelt availability, airbag deployment are important features but again speed is being the major factor when it comes to surviving. We can see from the stacked graph that although the driver or passenger is belted, if speed is above 40km/h, it is leading to level 4 or level 5 "injSeverity" which is either killing or vanishing the identity.

### Being Belted is helping as the airbags are then coming in picture if its deployed even if the speed is higher its helping in surviving or less injury.

### 

Fig 48 – Facet grid for dvcat, injSeverity and yearacc

### 

Fig 49 – point plots for injSeverity and abcat

### 

Fig 50 – point plots for dvcat and abcat

### Inference based on the facet grid and point plots:

### Weight, Year of the car or age of the car is not impacting the survival greatly during a crash but Probability of survival with low speed at accident prone locations are high.

### Severity of crashes are minimized in those cases where the airbags are deployed and weight is high, but in case when weight is low deployment is not playing any major role.

### Crashes in those cases where the passengers have deployment of airbags and the seatbelts are on have higher survival rates.

### When speed is high and weight is also minimalistic then the airbags and their deployment are not very helpful in survival although they are reducing the severity of injury.

### 

### Fig 51 – Heat map of numerical variables

### Imputing Missing values

### Although the column "injSeverity" is Numerical in nature, it is actually representing categories of Severity in a crash, so to impute the null values we will use mode rather than median of the data.

### 

Table-43 – After missing value imputation

### 

Table-44 Description after null values Imputation

### Removing the outliers using IQR method:

### IQR = Q3- Q1

### upper= Q3+(1.5 \* IQR)

### lower= Q1-(1.5 \* IQR)- Equation 5

### 

Fig 52– Boxplots After outliers Treatment

### 

Fig 53– Boxplots After outliers Treatment in one graph

Now since we have removed the outliers, we will move forward with our data preparation and split the data set for training and testing purposes.

### Part 2.2 – 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.?

### Data is split into 70:30 (Logistic Regression): -

### The split data of both independent and dependent trained data (70%) is fit into the logistic regression model to predict target variable

### Logistic regression uses newton-cg (solver), iteration inputs to execute the model

### Before Encoding the data set: About dropping column "caseid"

### Since caseid is object data type character column, created by pasting together the populations sampling unit, the case number, and the vehicle number. There is no such use of this column for developing a model as it can be used for identification purpose as Within each year, we can use this to uniquely identify the vehicle. This has no greater impact on the model where target variable is "Survived".

Shape before encoding the data set **=** 11217 rows and 14 columns

**Encoding**

1. We are encoding and getting the dummy variables first and then we will start the logistic regression. We need to encode the data because
2. The data has string & categorical type variables; these variables must be encoded so that the Machine Learning model understands the data.
3. In the target variable, "Survived" is replaced by 0 and "Not\_Survived" is replaced by 1.
4. Similarly, ordinal numbers are given to the values in variables airbag, seatbelt, dvcat, sex, occRole & abcat.
5. After this dummy encoding is used to encode the data for the rest of the columns.
6. Shape after encoding the data set **=** 11217 rows and 18 columns

### 

### 

Table-45 data set after encoding

### Columns after encoding the data set:

### 

### Value counts of the target column after encoding:

### 

### Logistic Regression Model:

### We will create a model based on logistic regression and calculate the coefficients of the variables along with the predicted probabilities. Then based on that model we will interpret about our model weather the model is a good fit or overfit and underfit.

### So are building the data set for training and testing after splitting the data set and we have the following outputs as shape of each tarin and test data set:

1. **(X) logistic regression train data set shape = (7851 rows and 17 columns)**
2. **(X) logistic regression test data set shape = (3366 rows and 17 columns)**
3. **(y) logistic regression train data set shape for target variable = (7851 rows and 1 column)**
4. **(y) logistic regression test data set shape for target variable = (3366 rows and 1 column)**

### Observations:

### From the above findings we see that after splitting our data set in the ratio of 70:30 we have our Training data set with 7851 rows and Test data set with 3366 rows, which is quite good number for developing a model. So, we will go ahead with these data sets and will develop our model.

### To ensure an equal ratio of number of 1s and 0s are split into both Training and Testing datasets so that there is a balance in the data and our model do not give biased results while Training or Testing the model we are using a function stratify= y\_log that is target variable while splitting the data set.

### We can see that the ratio of 1s and 0s in train data is 8.504 and in test data set it is around 8.508 which is almost same in both the cases.

### Predicted probabilities on Training data set is as follows for the first five rows:

### 

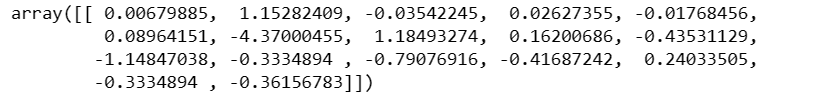
### Table-46-Predicted probabilities on Train data for LR

### Predicted probabilities on Testing data set is as follows for the first five rows:

### 

### Table-47-Predicted probabilities on Test data for LR

* Model intercept comes out to be = [-0.00355248]
* Model coefficients comes out to be =



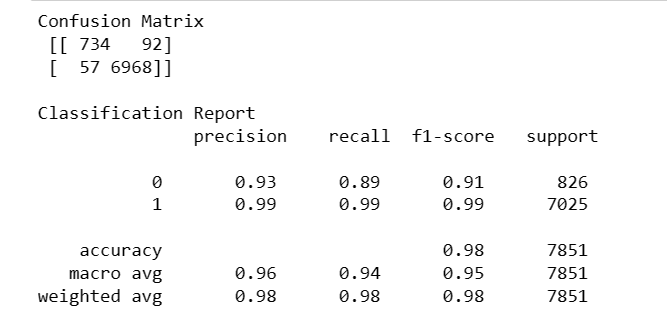
### Feature Importance and coefficients of various features:

### 

### Table- 48-Coefficients of various features in LR

### Evaluation of Logistic Regression on Train data:

Accuracy Score is 0.9811488982295249



### Table-49 -Classification report on Training data -LR

**AUC and ROC on Train data for Logistic Regression:**

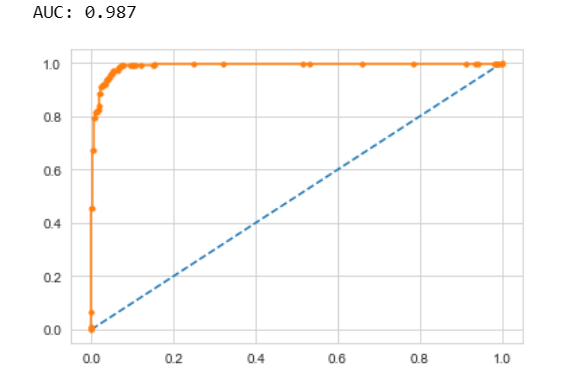
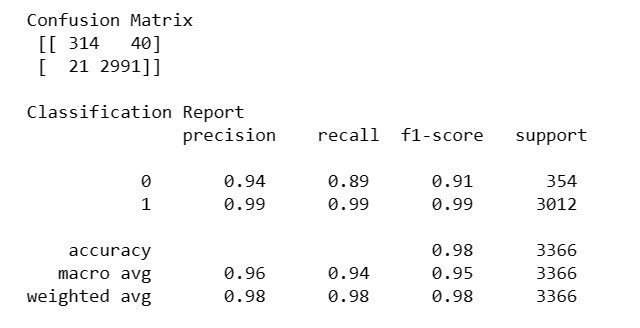
****

Fig 54 – AUC-ROC curve on training data for Logistic Regression

### Evaluation of Logistic Regression on Test data:

Accuracy Score is 0.9812834224598931



### Table- 50 -Classification report on Testing data -LR

**AUC and ROC for the Testing data for Logistic Regression model:**

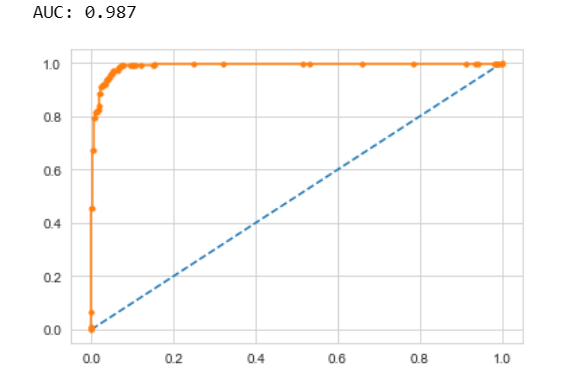
****

Fig 55 – AUC-ROC curve on test data for Logistic Regression

**Inference:**

From the Accuracy and Recall values, the model seems to be performing really good. However, also from the AUC values & ROC curve for Test data shows that it is covering almost all the values as compared to the train data**.**

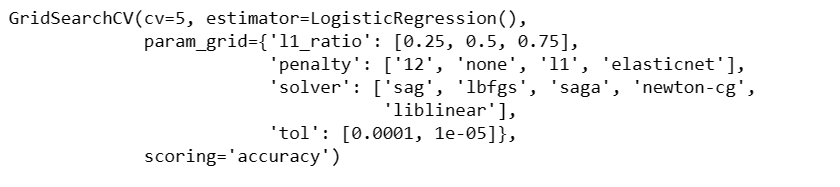
**Optimized Logistic Regression Model**

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

technique.

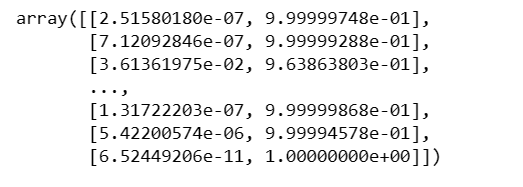
So, we are obtaining the best parameters using this technique which is given as:

Grid Search CV-



According to grid search best parameters:

* **'l1\_ratio': 0.25,**
* **'penalty': 'none'**
* **'solver': 'newton-cg'**
* **'tol': 0.0001**

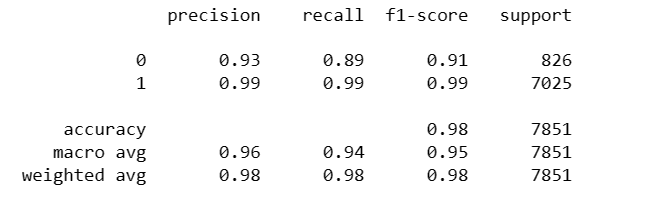


**Observations:**

* Although the model with normal features is showing quite good result on train and test data there is no need as such to build another model on these best parameters.
* But we can build Another Logistic Regression model with these best parameters and check its performance if it’s getting better in any way.
* 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.

**Performance metrics: On best parameters**

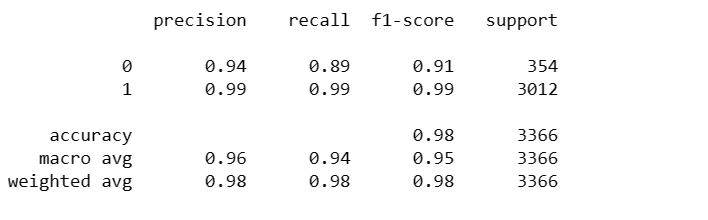
**Classification report – Train Data**



### Table-51-Classification report on Training data -LR best parameters

**Performance metrics: On best parameters**

**Classification report – Test Data**



### Table- 52 -Classification report on Test data -LR best parameters

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.

**Linear Discrimination Analysis**

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.

1. **(X) LDA train data set shape = (7851 rows and 17 columns)**
2. **(X) LDA test data set shape = (3366 rows and 17 columns)**

**Checking the class proportions of the train and test data set for LDA model:** LinearDiscriminantAnalysis ()

* Value counts of target variable on train data

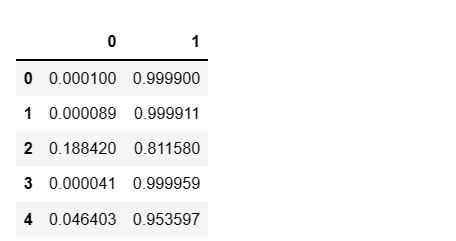


* Value counts of target variable on test data



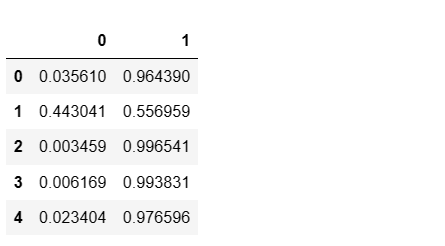
Class Label Predictions for LDA

* For training data set



### Table- 53 -Class label predictors Training data -LDA

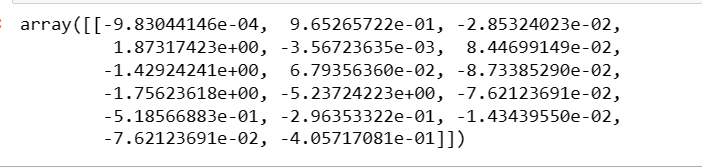
* For testing data set

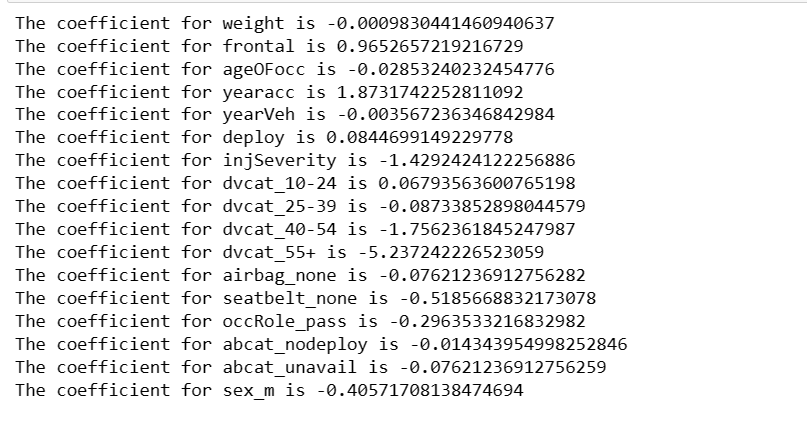


### Table- 54 -Class label predictors Test data -LDA

The class proportions almost come out to be the same for both train and test data set for LDA model. Now let’s calculate the intercept value and coefficients for the LDA model.

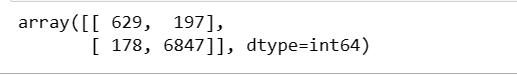
* Model intercept comes out to be = [-3731.9173267]
* Model coefficients comes out to be =

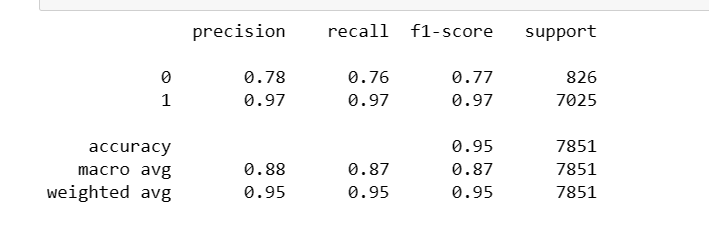




### Table- 55 -Coefficients of -LDA

**Classification Report on Train data and AUC-ROC curve:**

****

****

### Table-56-Classification report on Training data -LDA

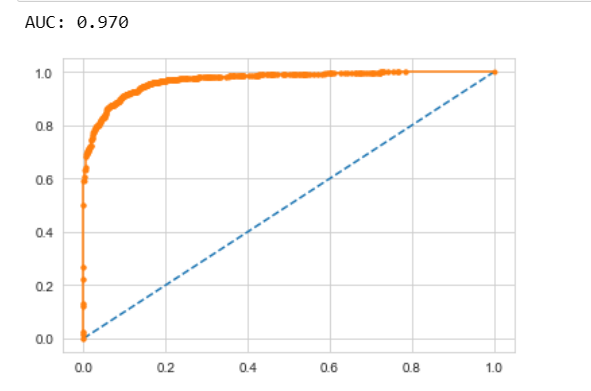
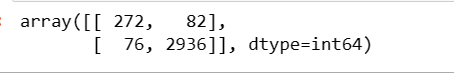
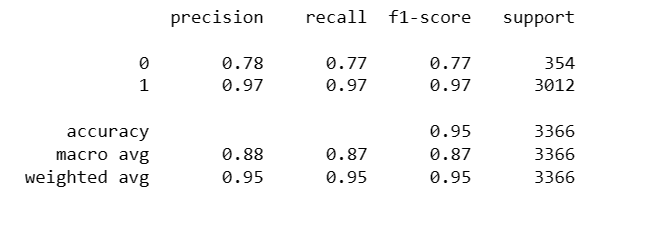


Fig 56 – AUC-ROC curve on training data for LDA

**Classification Report on Test data and AUC-ROC curve:**

****

****

### Table-57-Classification report on Testing data -LDA

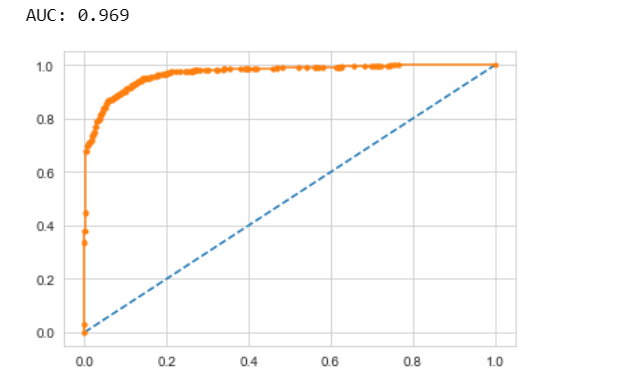
****

Fig 57 – AUC-ROC curve on test data for LDA

**LDA Model Developed:**

LDF= -3731.9173267+ X1\*(-0.0009830441460940637) + X2(0.9652657219216729) + X3(-0.02853240232454776) \n+X4(1.8731742252811092) +X5(-0.003567236346842984) + X6(0.0844699149229778) + X7(-1.4292424122256886) \n+X8(0.06793563600765198) + X9(-0.08733852898044579) +X10(-1.7562361845247987) \n+X11(-5.237242226523059) + X12( -0.07621236912756282) +X13(-0.5185668832173078) + X14(-0.2963533216832982) \n+X15(-0.014343954998252846) + X16(-0.0762123691275625) + X17(-0.40571708138474694)

**Inference:**

The LDA model is quite good in terms of accuracy but it does not look better than the Logistic Regression model in terms of the Recall and precision value for Train and Test data. Also, the Accuracy for the test data is lower in case of LDA model. The AUC and ROC curves also do not show a significant difference compared to the other models built.

**So, from the above equation the following things can be summarized as**

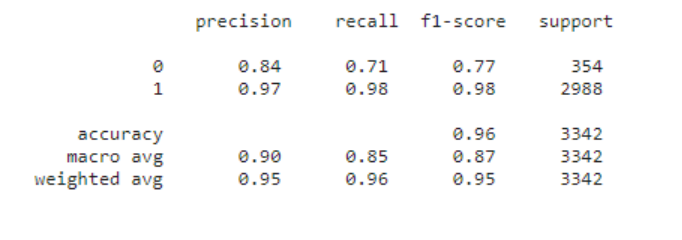
1. the coefficient of X11 predictor is largest in magnitude thus it helps in discriminating the target the best
2. the coefficient of X1 predictor is smallest in magnitude thus it helps in discriminating the target the least.
3. all the DS can be computed for each row using the above f(x) which will aid in classification

**Classification by Probability (with cut- off score 0.4 instead of default 0.5)**

* Cut off = 0.4
* Accuracy score = 0.9569
* F1 score = 0.9762

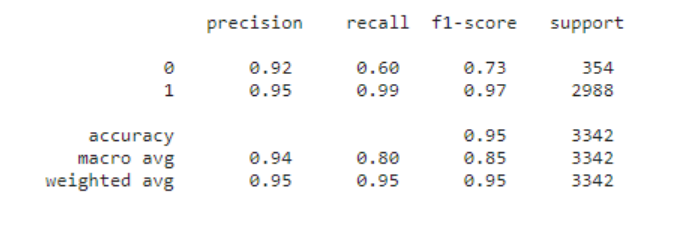
But 0.4 cut-off gives us the best 'recall'. Here, we will take the cut-off as 0.4 to get the optimum 'recall' score.

Classification report of default cut- off test data:



### Table-58-Classification report on Testing data -LDA default cut- off

Classification report of customised cut- off test data:



### Table-59-Classification report on Testing data -LDA customized cut- off

**CART Model:**

A CART model is also built using the following parameters:

criterion = 'Gini',

maximum depth = 7,

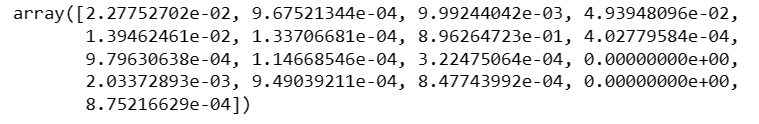
minimum samples leaf=20,

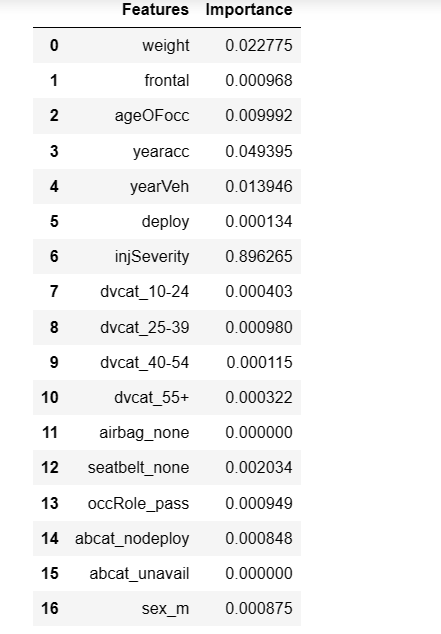
This model is also further evaluated with Accuracy score, along with the confusion matrix. The AUC-ROC curve is plotted for both the Train and Test data.

****

Fig 58 – Tree flow chart of the cart model

Model coefficients of CART model is given as:



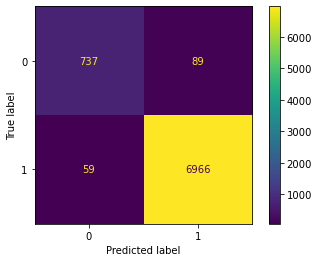
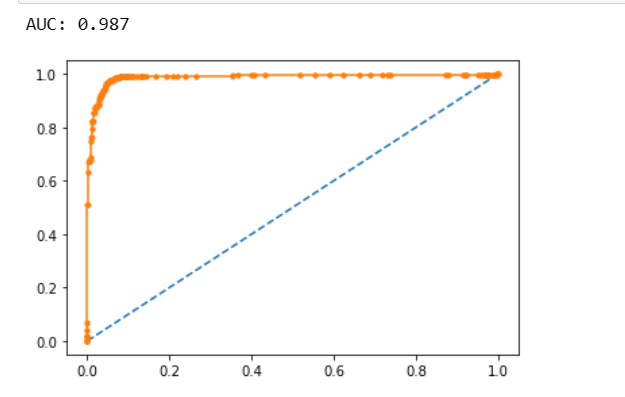


### Table- 60 – Important features in CART

**Part 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.?**

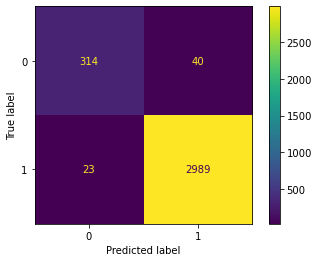
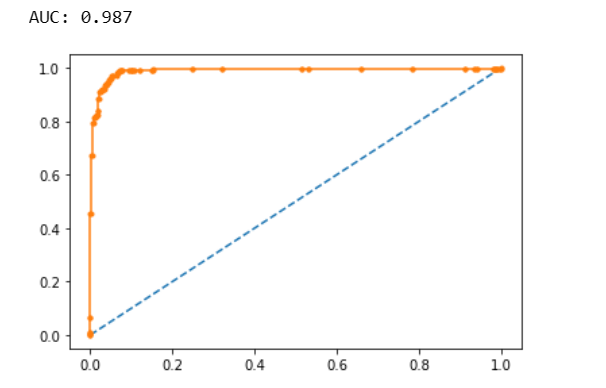
**Evaluation of Logistic regression model:**

**Confusion Matrix – Training set:**

**** ****

* Fig 59. Confusion matrix and AUC-ROC Curve- LR – Training
* True negative of logistic Regression model is – 737
* False Positive of Logistic Regression model is – 89
* False Negative of Logistic Regression model is - 59
* True positive of Logistic Regression model is – 6966
* **Model score = 0.98**
* **AUC – ROC curve for the training data set is given as follows where AUC = 0.987**

**LOGISTIC REGRESSION- Evaluation on Test data**

**** ****

* Fig 60. Confusion matrix and AUC-ROC Curve- LR – Test data
* True negative of logistic Regression model is – 314
* False Positive of Logistic Regression model is – 40
* False Negative of Logistic Regression model is - 23
* True positive of Logistic Regression model is – 2989
* **Model score = 0.98**
* **AUC – ROC curve for the training data set is given as follows where AUC = 0.987**

**Insights:**

1. Logistic Regression model seems to be a really good model specially on Accuracy, followed by precision and recall.

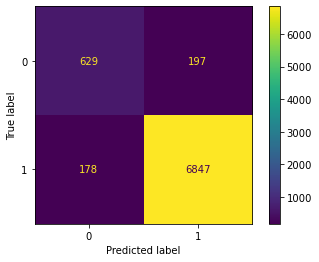
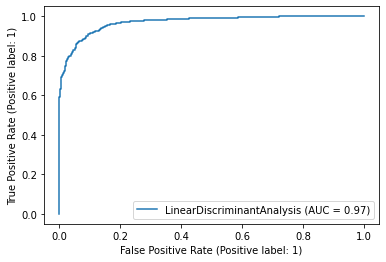
2. The AUC-ROC curve on Test data covers almost all the part as on Train data.

3. The most important features according to this model which plays the most important part in predicting the Survival of Drivers and passengers based on our model, comes out to be

* injSeverity
* dvcat (dvcat\_55+ and dvcat\_10-24)
* Frontal

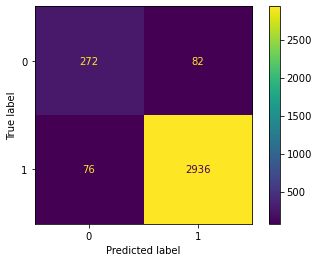
**Evaluation of LDA Model**

**LDA Model performance: On Train Data: Along with AUC-ROC curve**

** **

* Fig 61. Confusion matrix and AUC-ROC Curve- LDA – Train data
* True negative of logistic Regression model is – 629
* False Positive of Logistic Regression model is – 197
* False Negative of Logistic Regression model is - 178
* True positive of Logistic Regression model is – 6847
* **AUC – ROC curve for the training data set is given as follows where AUC = 0.97**

**LDA Model performance: On Test Data: Along with AUC-ROC curve**

** **

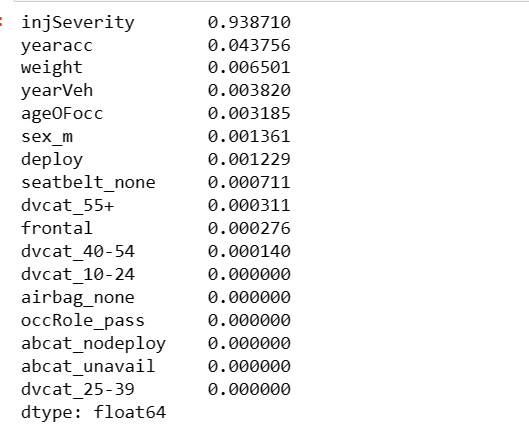
* Fig 62. Confusion matrix and AUC-ROC Curve- LDA – Test data
* True negative of logistic Regression model is – 272
* False Positive of Logistic Regression model is – 82
* False Negative of Logistic Regression model is - 76
* True positive of Logistic Regression model is – 2936
* **AUC – ROC curve for the training data set is given as follows where AUC = 0.97**

**Insights:**

* + Since we are dealing with a model which is about "Survived" and "Not\_Survived" predictions, both accuracy and recall becomes very important for our model and hence LDA model looks good with very little difference from Logistic Regression model.
  + However, Logistics Regression slightly scores higher with regards to Recall, hence Logistics Regression model is preferred. As all performance parameters are quite high, we can use this data and features for recommendations based on the model performance.
  + AUC-ROC curve also shows a little deviation in case of our LDA model both on Train and Test data sets.
  + But we have other options available as well which can give us even better recall values such as CART and decision tree model. So, our next agenda is to compare all these models and then choose based on the performance of each model.
  + In case of LDA model here, we are taking cut-off value as 0.5, if we want to get better accuracy and precision score, we can even change the cut-off values to 0.4 or 0.3 or 0.6 and see which are giving better results and then we can use that model for future reference.

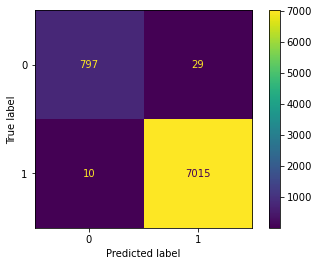
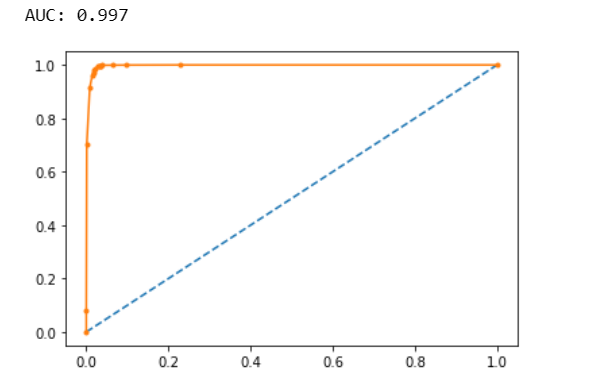
**Evaluation of Cart Model:**

* **Accuracy Score of Train data (DT Model) = 0.99503**
* **Accuracy Score of Test data (DT Model) = 0.9907**
* **Important features of the CART model are as follows:**

****

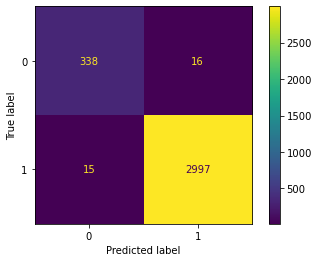
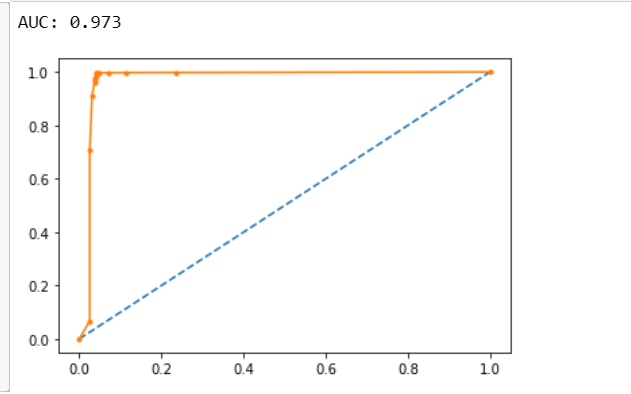
### Table- 60 – Important features in CART

**Evaluation of CART model on train data**

**** ****

* Fig 63. Confusion matrix and AUC-ROC Curve- CART – Train data
* True negative of logistic Regression model is – 797
* False Positive of Logistic Regression model is – 29
* False Negative of Logistic Regression model is - 10
* True positive of Logistic Regression model is – 7015
* **AUC – ROC curve for the training data set is given as follows where AUC = 0.997**

**Evaluation of CART model on test data:**

**** ****

* Fig 64. Confusion matrix and AUC-ROC Curve- CART – Test data
* True negative of logistic Regression model is – 338
* False Positive of Logistic Regression model is – 16
* False Negative of Logistic Regression model is - 15
* True positive of Logistic Regression model is – 2997
* **AUC – ROC curve for the training data set is given as follows where AUC = 0.973**

**Inference:**

1. From the above Evaluated models it seems The CART model is performing the better than LDA and Logistic Regression in terms of Accuracy, Recall, F1score and Precision values.

2. It also shows the important features in the model in terms of their value and "injSeverity" is the most important feature followed by "yearacc" and "weight".

3. We already saw in case of Logistic regression dvcat\_40-54 is -1.7562361845247987 and dvcat\_55+ is -5.237242226523059 which are the top 2 features for Survival followed by injSeverity.

**Model Comparison: On Testing Data and Training Data**

**Logistic regression Accuracy –**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Logistic Train** | **Logistic Test** | **LDA Train** | **LDA Test** | **CART Train** | **CART Test** |
| **Accuracy** | 0.98 | 0.98 | 0.95 | 0.94.9 | 0.9958 | 0.9907 |
| **Recall** | 0.99 | 0.99 | 0.97 | 0.97 | 0.99 | 0.98 |
| **F1 Score** | 0.99 | 0.99 | 0.97 | 0.97 | 0.99 | 0.99 |
| **Precision** | 0.99 | 0.99 | 0.97 | 0.97 | 0.98 | 0.99 |
| **AUC** | 0.987 | 0.98 | 0.97 | 0.969 | 0.99 | 0.975 |

### Table- 61 – Model comparisons

**Part 2.4 -Inference: Basis on these predictions, what are the insights and recommendations?**

**Business Insights & Recommendations**

From the important features from Logistic Regression Model, LDA model & CART Model

1. As per the Logistic Regression model, injSeverity, dvcat\_10-24, dvcat\_55+ and frontal are very important in deciding whether the driver or passenger will survive the crash or not.
2. In case of LDA model also we see that dvcat\_55, dvcat\_40-54, yearacc and injSeverity comes out to be the 4 top important factors when we are developing a model.
3. The CART model also indicates that the injSeverity, yearacc (0.049340) and Weight are important features.
4. Therefore, based on these 3 models we can say that

* dvcat\_55+, dvcat\_10-24
* injSeverity
* yearacc
* frontal
* Weight are top 5 highly important features.

1. All the models also indicated that the seatbelt and airbag is also important, and in real life that makes sense. This feature can actually help in surviving a major crash and can reduce the severity.
2. Although Weight, Year of the car or age of the car are significant features they are not impacting the survival greatly during a crash
3. Probability of survival with low speed at accident prone locations are high and the severity is also low after the survival.

**Recommendations for Government, Car manufacturers and Road safety authorities:**

- As we can see that speed is a major factor in crashes so Governments can implement stricter safety regulations for vehicles, such as mandatory safety features (e.g., airbags, stability control, etc.) and can set speed limits for various areas, minimum crash test ratings, and more stringent vehicle inspection requirements.

- Promote safer driving behaviours: Governments and car manufacturers can work together to promote safer driving behaviours, such as making it mandatory for all manufacturers to provide with airbags in the cars, and making it mandatory to have airbags for both front and back seats to ensure safety during frontal or non-frontal impact accidents, not driving while impaired by alcohol or drugs, not texting or using a phone while driving, and always wearing a seatbelt.

- Provide education and training: Governments and car manufacturers can provide education and training programs to help drivers improve their skills and knowledge of safe driving practices.

- Government should be stricter and should not allow cars after a particular year on the roads depending on the age of car and year of manufacture as these cars can be damaging and are not good for environment as well.

- Manufacturers should deploy warning signals in case of non-deployment of airbags or seatbelts.

- Manufacturers can deploy intelligent system to advise speed reduction at accident prone locations or crowded places. They can also Encourage the use of advanced driver assistance systems (ADAS): Car manufacturers can continue to develop and promote the use of advanced driver assistance systems (ADAS), such as lane departure warning systems, adaptive cruise control, and automatic emergency braking.

- Government can implement reward vs penalty clauses for manufacturers based on the record of safety measures deployed in the vehicles and the weight of cars should be in a particular limit for particular areas such as cities, hills etc so that heavy crashes can be avoided.

- Improve emergency response systems: Governments can improve emergency response systems, such as by investing in faster and more reliable ambulance and first responder services, and ensuring that hospitals have the necessary resources and equipment to treat crash victims immediately to reduce the level of severity after the crash.

- Provide education and training: Governments and car manufacturers can provide education and training programs to help drivers improve their skills and knowledge of safe driving practices.

**CONCLUSION:**

We can see that AUC curve of Both LDA, and Logistic regression shows similar results which is around 98% for test data and 99% in case of CART, hence we can say both LDA and Logistic regression can be a good model in this case but when it comes to Accuracy Logistic Regression and CART are performing better than LDA which has the lowest accuracy on Test data which is around 96% but we can say that for the given data set all the models are almost similar and quite stable.

*Reference – Great Learning lecture videos and Mentors*