Revanth Janapriyan

 PGP-DSBA ONLINE

10/06/2023

PREDICTIVE MODEL

Contents

[**PROBLEM 1: LINEAR REGRESSION** 2](#_Toc137483299)

[1.1) Read the data and do exploratory data analysis. Describe the data briefly. (Check the null values, data types, shape, EDA). Perform Univariate and Bivariate Analysis. 2](#_Toc137483300)

[**EXECUTIVE SUMMARY** 2](#_Toc137483301)

[**INTRODUCTION** 2](#_Toc137483302)

[**DATA DESCRIPTION** 2](#_Toc137483303)

[**DATA SAMPLE** 3](#_Toc137483304)

[**EXPLORATORY DATA ANALYSIS** 3](#_Toc137483305)

[1.2) Impute null values if present? Do you think scaling is necessary in this case? 12](#_Toc137483306)

[1.3) Encode the data (having string values) for Modelling. Data Split: Split the data into test and train (30:70). Apply Linear regression. Performance Metrics: Check the performance of Predictions on Train and Test sets using R-square, RMSE. 13](#_Toc137483307)

[1.4) Inference: Based on these predictions, what are the business insights and recommendations. (6 marks) 13](#_Toc137483308)

[**PROBLEM 2: Logistic Regression and Linear Discriminant Analysis** 14](#_Toc137483309)

[2.1) Data Ingestion: Read the dataset. Do the descriptive statistics and do null value condition check, write an inference on it. Perform Univariate and Bivariate Analysis. Do exploratory data analysis. 14](#_Toc137483310)

[**EXECUTIVE SUMMARY** 14](#_Toc137483311)

[**INTRODUCTION** 14](#_Toc137483312)

[**DATA DESCRIPTION** 15](#_Toc137483313)

[**DATA SAMPLE** 16](#_Toc137483314)

[**EXPLORATORY DATA ANALYSIS** 16](#_Toc137483315)

[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). 25](#_Toc137483316)

[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. Compare both the models and write inferences, which model is best/optimized. 25](#_Toc137483317)

[**2.4)** Inference: Based on these predictions, what are the insights and recommendations. 29](#_Toc137483318)

# **PROBLEM 1: LINEAR REGRESSION**

### 1.1) Read the data and do exploratory data analysis. Describe the data briefly. (Check the null values, data types, shape, EDA). Perform Univariate and Bivariate Analysis.

## **EXECUTIVE SUMMARY**

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.

## **INTRODUCTION**

In this first problem, various steps have been taken in order to predict the sales of 759 firms. Data exploration have been carried out on different variables including capital, patents, R & D stocks, Employment, Stock market value of the firm, etc using data exploration techniques such as univariate analysis, bivariate analysis, graphical representations. Based the data exploration many usual insights have been obtained and also steps are taken to eliminate outliers and missing values. Thus, made sure a clean data is ready for the Linear Regression model. Based on the predicts and insights the Investment firm can plan their future investments across the firms accordingly.

## **DATA DESCRIPTION**

The Description of the variables provided in the dataset given below.

1. sales: Sales (in millions of dollars).

2. capital: Net stock of property, plant, and equipment.

3. patents: Granted patents.

4. randd: R&D stock (in millions of dollars).

5. employment: Employment (in 1000s).

6. sp500: Membership of firms in the S&P 500 index. S&P is a stock market index that measures the stock performance of 500 large companies listed on stock exchanges in the United States

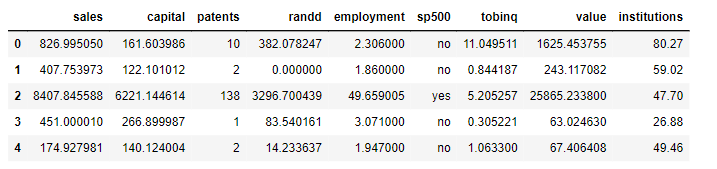
7. tobinq: Tobin's q (also known as q ratio and Kaldor's v) is the ratio between a physical asset's market value and its replacement value.

8. value: Stock market value.

9. institutions: Proportion of stock owned by institutions.

## DATA SAMPLE

The table below shows the first few rows of the given dataset.

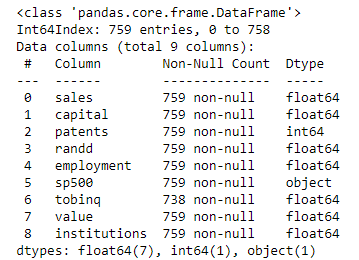


**Table 1.1**

There are 9 variables in the given dataset.

## **EXPLORATORY DATA ANALYSIS**

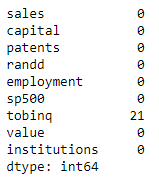
Let’s see the different types of variables and number of datapoints each has, data types each of variable and missing values.



**Table 1.2**

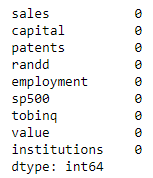
The table 1.2 shows that there are 7 float type features, one integer and one object data type each having 759 entries.

Let’s check for the missing values in the dataset. Upon checking up there are 21 null values in the tobinq variables and it’s shown in table 1.3.



**Table 1.3**

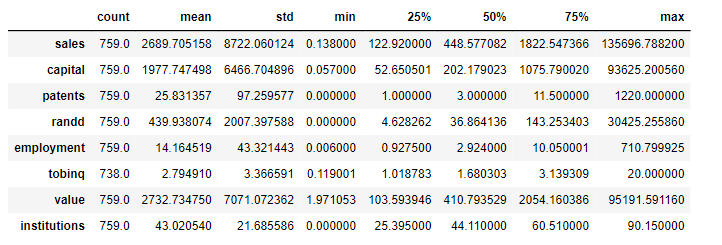
The null values are treated and the table 1.4 shows the output.



**Table 1.4**

Descriptive Analysis

The table below shows the descriptive stats of the given dataset.



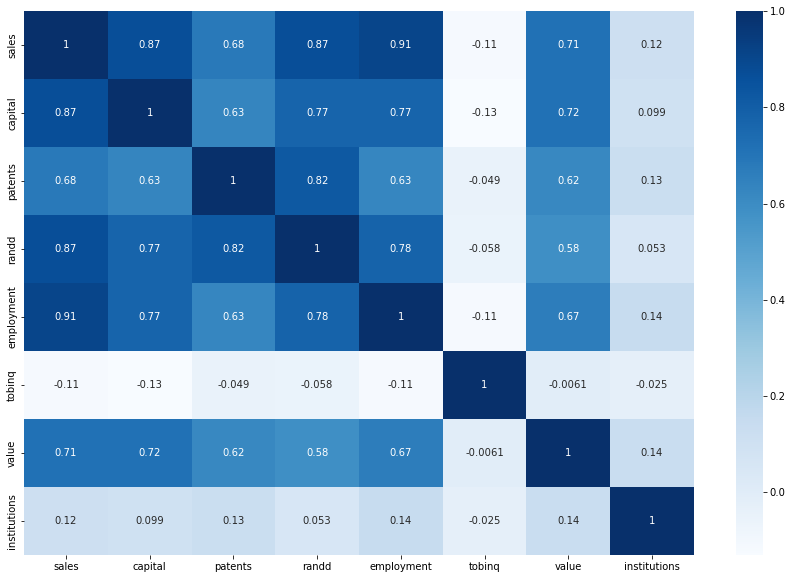
**Table 1.5**

With the help of the descriptive statistics, we can see various measure such as mean, Q1, Q2, Q3, min, max and standard deviation of each feature. Below are few of the insights from above table:

* The average sales of the firms are $ 2689 million.
* The highest capital is $ 93635 million.
* The least stock market value is $1.97 million and the highest is $95191.5 million.

Visualizing each variable:

Below figure, show the correlation across variables in the form of Heatmap.



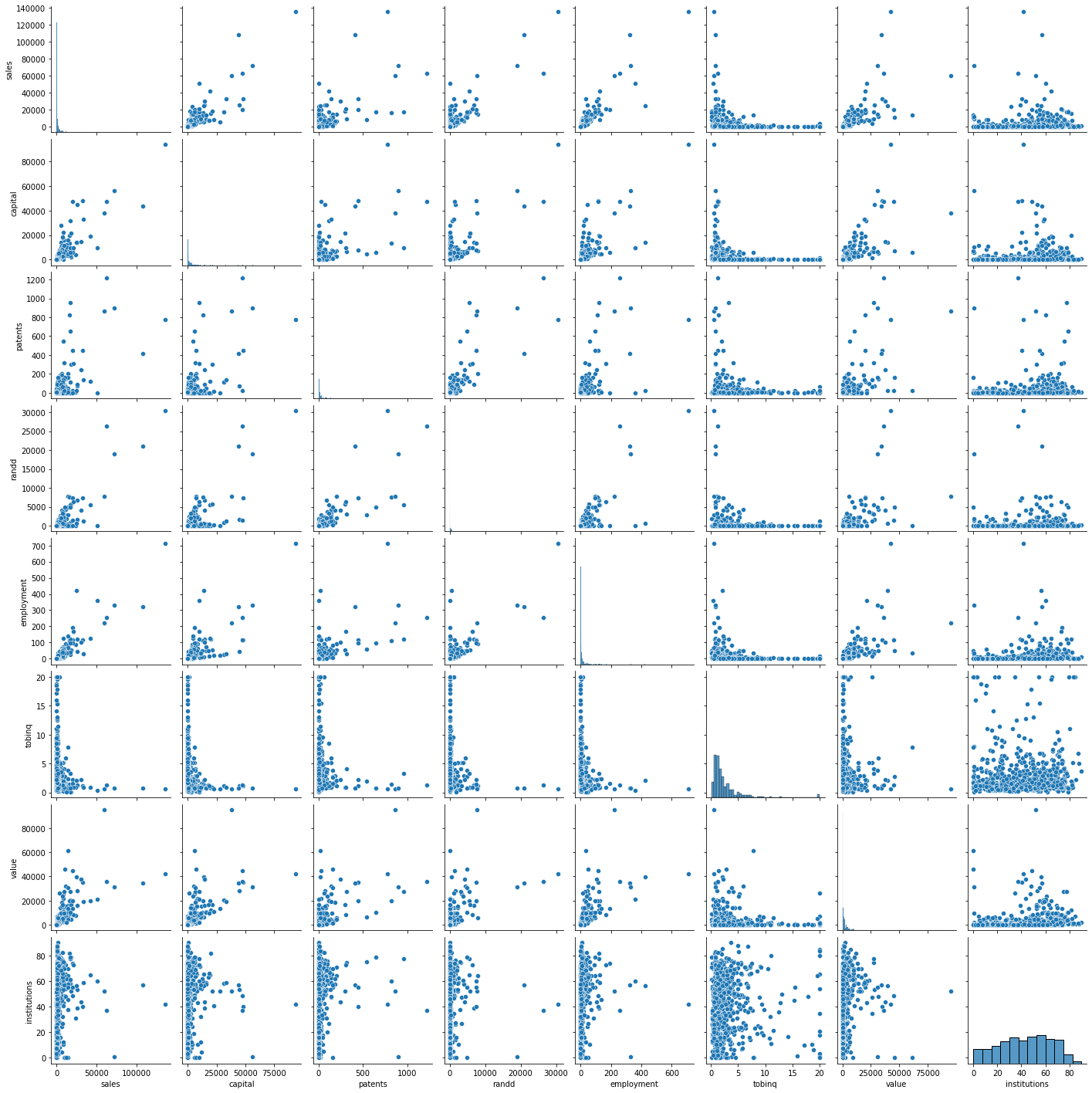
**Fig 1.1**

The above heatmap shows the following:

* Highest correlation is between sales and the employment variables and the second highest correlation is between sales and randd and also between sales and capital.
* Lowest correlation is between sales and tobniq.

Pair plot:

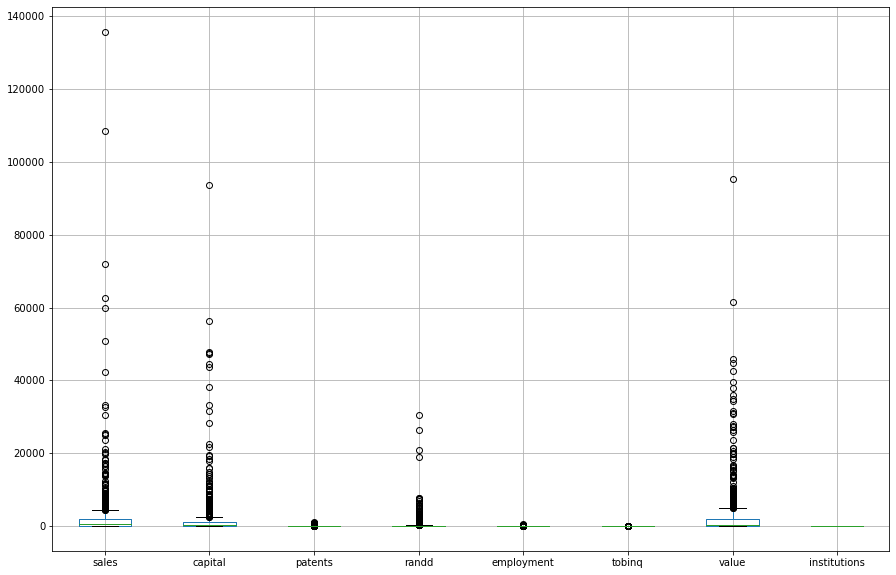
The fig 1.2 shows the pair plot of the given dataset.



**Fig 1.2**

Outliers:

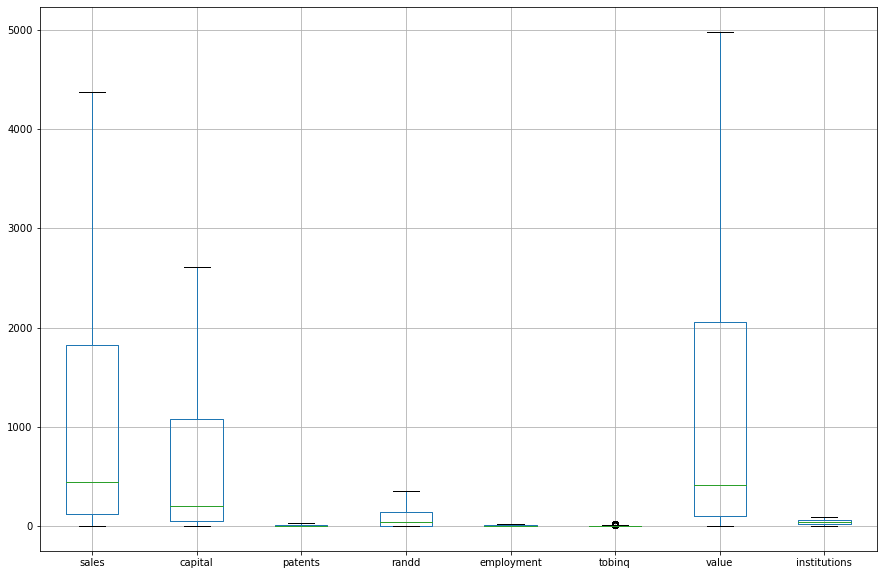
The outliers present in the dataset have been checked and it is treated. Fig 1.3 shows the boxplot of the numeric variables present along with its outliers.



**Fig 1.3**

It can be clearly seen that the varibles have lots of outliers. Hence, necessary action have been taken such that the outliers present are eliminated.

The boxplot of the variables after the outliers have been treated is shown in the fig 1.4.



**Fig 1.4**

Visualizing the variables (univariate & bivariate):

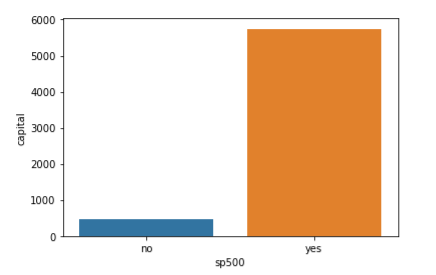
Let’s visualize the each variables and see what insights we can get from them.

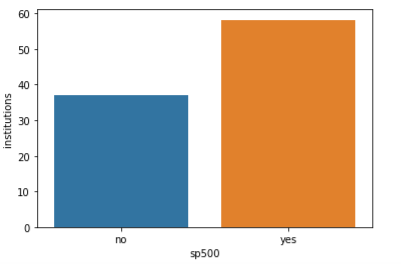
****

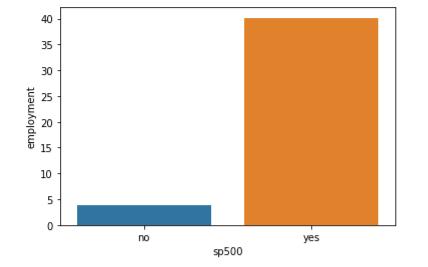
**Fig 1.5**

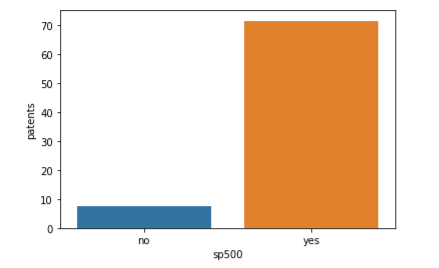
The fig 1.5 shows the counts of the membership of firms in the S&P 500 index. It can be clearly seen that most of the firms of are not a member of S & P 500 index.

Bivariate:

Let’s visualize the variables using bivariate analysis.



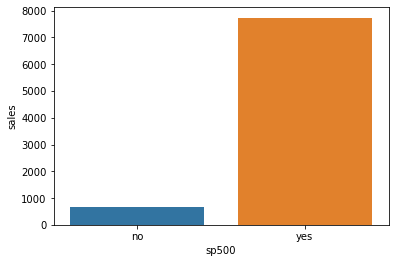
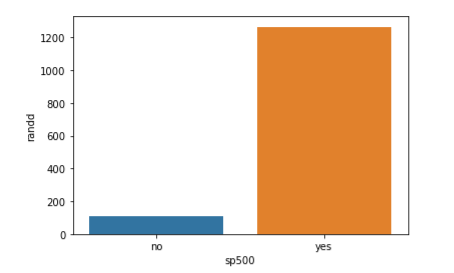


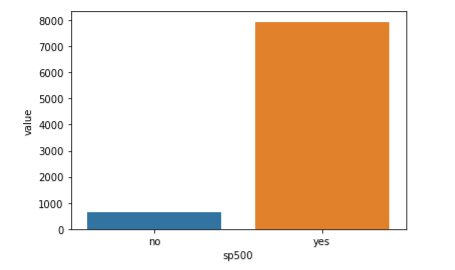
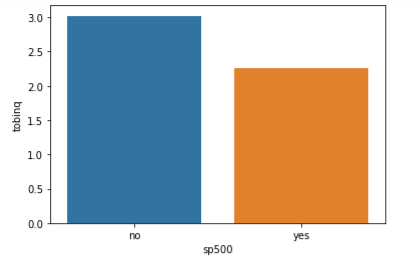


**Fig 1.6.1**

The figure 1.6.1 shows the bivariate analysis of the variable’s capital, Institutions, employment and patents, each against sp500.

The patents, capital and employment in sp500 firms are relatively higher than that of those non sp500 firms.

****

****

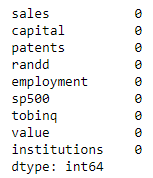
**Fig 1.6.2**

The above figure 1.6.2 also shows the bivariate representations of the variables present.

In the fig 1.6.2, bivariate analysis has been carried out between sp500 and randd, sales, tobniq and value individually. We can clearly see that sales are way too high for sp500 firms than non sp500 firms, whereas tobinq is higher in non sp500 firms than the firms which are members of S & P 500 Index.

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

The null values in the datasets have been treated and no variables is left with null values. (Refer below table).



**Table 1.6**

Feature scaling is a method used to standardize the range of independent variables or features of data. In data processing, it is also known as data normalization or standardization. Feature scaling is generally performed during the data pre-processing stage, before training models using machine learning algorithms.

Upon looking up at the different variables, it is evident that the values across each variable is measured in different scales. So, it is a necessary step to scale the dataset such that the accuracy of the model while predicting does not get affected in a bad way.

### 1.3) Encode the data (having string values) for Modelling. Data Split: Split the data into test and train (30:70). Apply Linear regression. Performance Metrics: Check the performance of Predictions on Train and Test sets using R-square, RMSE.

All the actions required has been taken including encoding string data types and splitting the data into training and test data.

A Linear Regression model has been built and both training and testing dataset has been used to predict and evaluate the performance of the model.

The model performed well on both training and testing dataset with 100% accuracy.

The Rmse value of the model on test dataset is quite high while the R-Squared values remained good.

### 1.4) Inference: Based on these predictions, what are the business insights and recommendations. (6 marks)

Based on the prediction the Linear Regression model works well with current problem statement.

Firms which are members of S&P 500 performed well in all the aspects. I will be a good move for the investment firm to have only S&P 500 member firms so that the sales can be well promised.

# **PROBLEM 2: Logistic Regression and Linear Discriminant Analysis**

### 2.1) Data Ingestion: Read the dataset. Do the descriptive statistics and do null value condition check, write an inference on it. Perform Univariate and Bivariate Analysis. Do exploratory data analysis.

## **EXECUTIVE SUMMARY**

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.

## **INTRODUCTION**

In this problem survival rate of the persons coming across car crash has been predicted using Logistic Regression and Linear Discriminant Analysis method. Univariate analysis, Bivariate analysis, descriptive statistics are few of the data analysis and data exploratory techniques has been used. The dataset provided contains various features such as airbag availability, seatbelt usage, etc which were sufficient enough in training the models and making the models predict with higher accuracy. Both Logistic Regression model and Linear Discriminant analysis model has been put into predicting the survival rate and metrics such as confusion matrix, classification report, ROC curve and ROC\_AUC score are used here to evaluate the predictions made by each model. Thus, it helped in choosing the best model for this problem scenario.

## **DATA DESCRIPTION**

The following description of each feature in the dataset is provided below:

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, not deploy and unavailable.

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

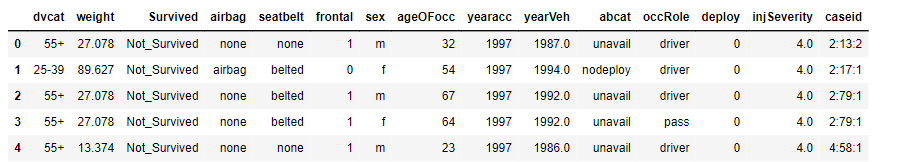
13. **deploy**: a numeric vector: 0 if an airbag was unavailable or did not deploy; 1 if one or more bags deployed.

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.

## DATA SAMPLE

The sample of the dataset is given below:

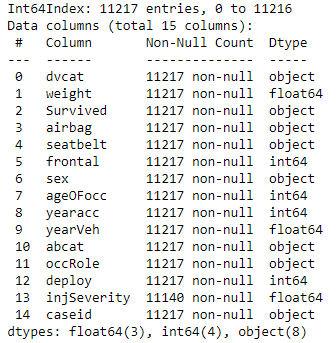


**Table 2.1**

There are 15 features present in the given dataset. There are 11217 rows, 15 columns in the given dataset.

## **EXPLORATORY DATA ANALYSIS**

Let’s see the different types of variables and number of datapoints each has, data types each of variable and missing values.

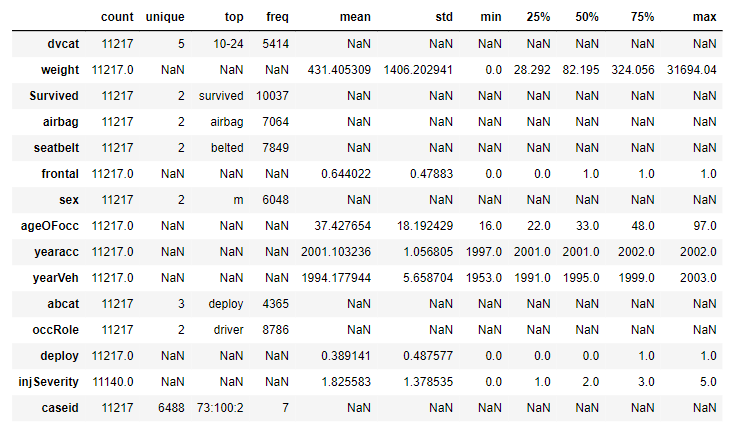


**Table 2.2**

There are 7 string variables, 3 float variables and 4 integer variables.

Descriptive Analysis

Below table shows descriptive statistics of the given dataset.



**Table 2.3**

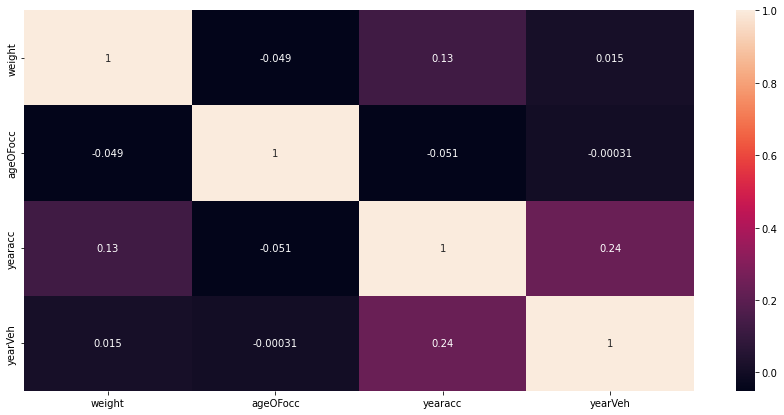
With the help of the descriptive statistics, we can see various measure such as mean, Q1, Q2, Q3, min, max and standard deviation of each feature. Below are few of the insights from above table:

* Average weight is 431.4 kg.
* Severity of Injury is between 1-2.

Visualization:

Now let’s visualize the features in the dataset.

Heatmap:

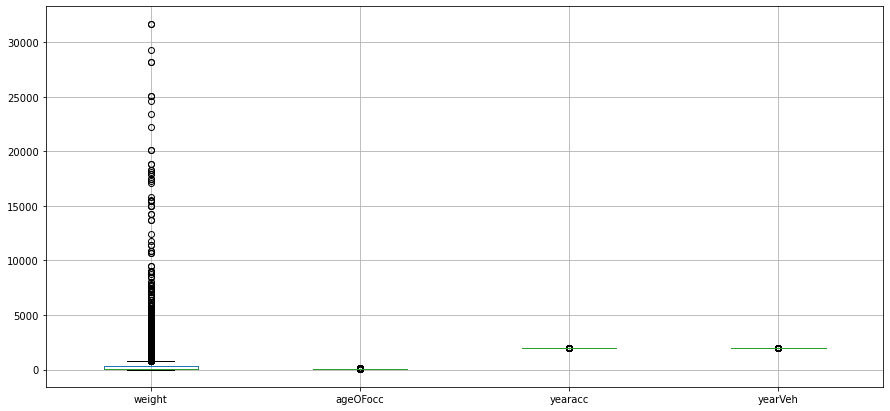


**Fig 2.1**

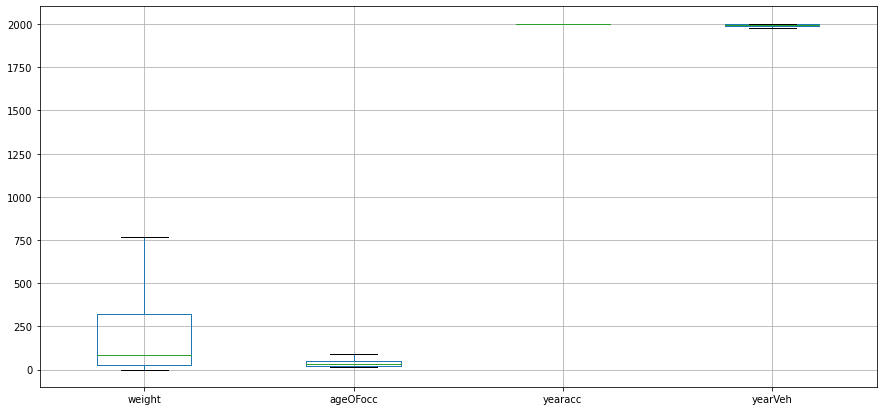
The above heatmap shows no high correlation among the variables in the dataset.

Outliers:

The features in the dataset have been visualized using boxplot to check for the presence of outliers.



**Fig 2.2**

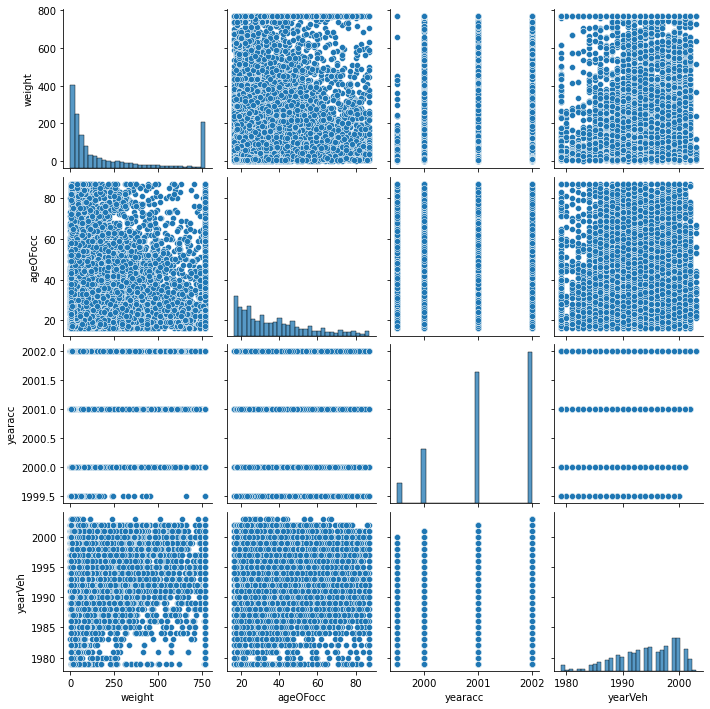
****

**Fig 2.3**

The figures 2.2 and 2.3 shows the presence of outliers in the features and the featured with no outliers after it is treated respectively.

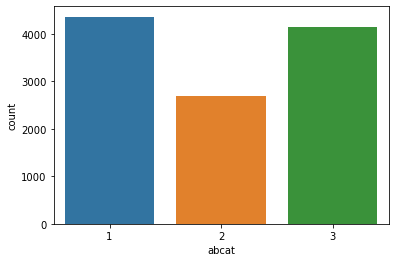
Pair plot:

The figure below shows the pair plot of the features present in the dataset. It reconfirms that there is no correlation among the variables.



**Fig 2.4**

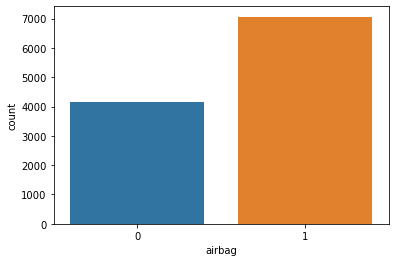
Now let’s see the univariate analysis of each variable present in the dataset.



**Fig 2.5**

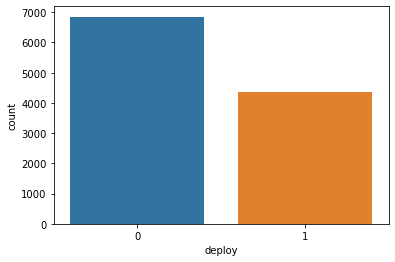
The above figure shows the three categories of whether one or more passengers used airbags. Here 1 represents deployed.

The below shows the overall count of airbags deployed and non delployed.



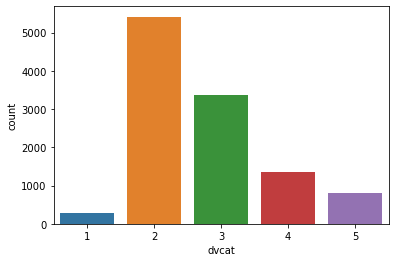
**Fig 2.6**

The figure below shows the count of airbags deployed and not deployed.



**Fig 2.7**

The fig 2.8 show the different speed levels at which the vehicle it driven. It has been classified into 5 levels and counts of each are shown in the graph below.



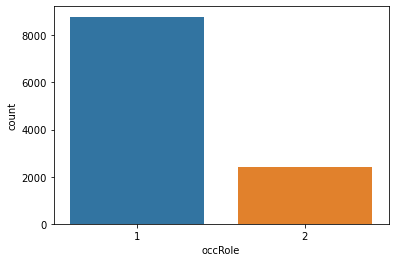
**Fig 2.8**

The count of frontal damage and non-frontal damage is shown in the fig 2.9 below.

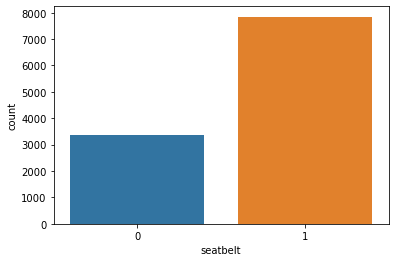


**Fig 2.9**

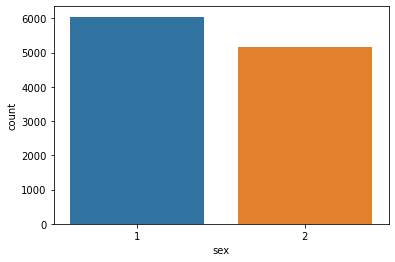
The figure below counts of occRole variable.



**Fig 2.10**



**Fig 2.11**



**Fig 2.12**

The above two figures 2.11 and 2.12 show the counts of seatbelt worn and sexuality of the passengers and drivers respectively.

### 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).

Once the data analysis and data exploration are done the string values are encoded.

The encoded data is then split in into training and testing dataset in 70:30 ratio. The split dataset is then used to train and test on both the logistic regression model and the linear discriminant analysis model.

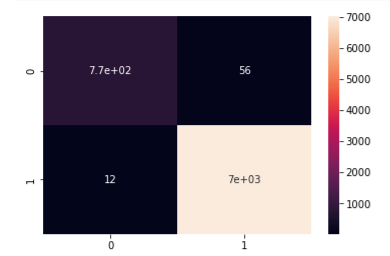
### 2.3) Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC\_AUC score for each model. Compare both the models and write inferences, which model is best/optimized.

The performance of the model is evaluated using the metrics such as Accuracy score, confusion matrix, ROC\_AUC score and ROC curve on both Logistic Regression model and Linear Discriminant Analysis model.

**Accuracy**

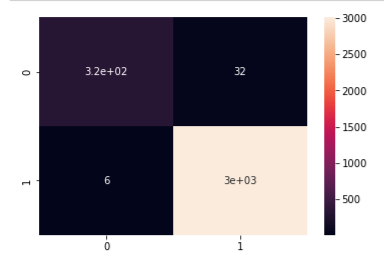
* The Accuracy score of the Logistic Regression model on training and testing dataset are 99.1% and 98.8% respectively.
* The Accuracy score of the Linear Discriminant Analysis model on training and testing dataset are 99% and 98.8% respectively.

**Confusion Matrix**



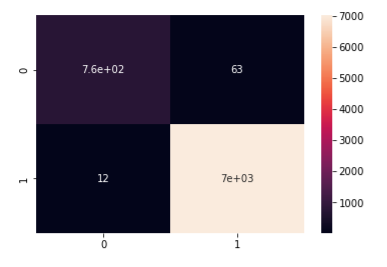
**Fig 2.13**

The figure 2.13 shows the confusion matrix of Logistic regression model on training dataset. The type I error is higher than the type II error here.



**Fig 2.14**

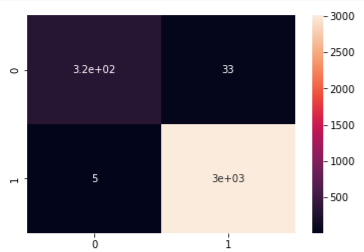
The figure above shows the confusion matrix of logistic regression model on the test dataset. Clearly, we can see the reduction in type I and type II error when compared to the performance of model on training dataset.



**Fig 2.15**

The figure 2.15 shows the confusion matrix of Linear Discriminant Analysis model on train dataset. Here, the type I and type II error are higher than that of logistic regression model on same training dataset.

The figure below shows the confusion matrix of Linear Discriminant Analysis model on the test dataset. Both the Linear Discriminant analysis model and logistic regression model have performed so similarly in terms of confusion matrix.

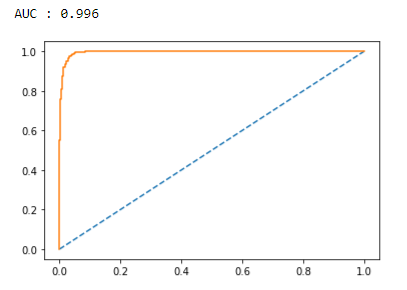


**Fig 2.16**

**ROC and AUC curve**

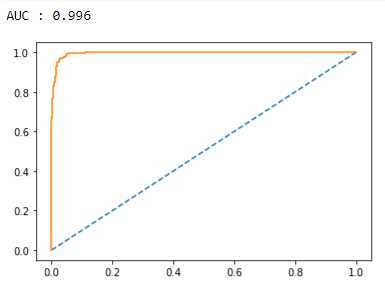
Now let’s look into the auc score and roc\_auc curve of logistic regression model and linear discriminant analysis model on both training and test dataset.

Below figure provides us with the auc score and roc\_auc curve of logistic regression model on training dataset.



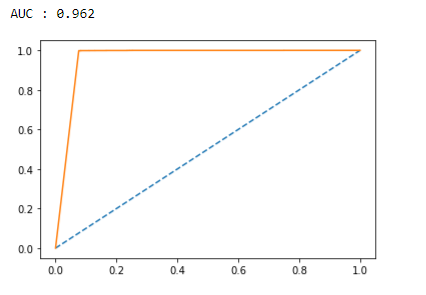
**Fig 2.17**

The figure 2.18 shows the auc score and roc\_auc score of the logistic regression model on the test dataset. The performance of the logistic model on test dataset is same as it did on training dataset.



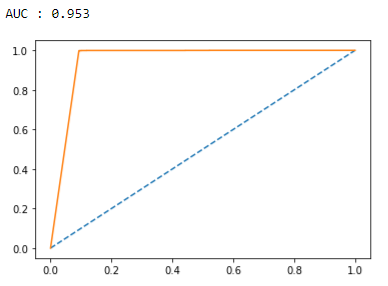
**Fig 2.18**

Figure 2.19 shows the shows the auc score and roc\_auc score of the Linear Discriminant analysis model on the train dataset. The auc score of the model is closer to that of the logistic regression model on the training dataset.

****

**Fig 2.19**

The auc score of the Linear Discriminant analysis model on the test dataset has dropped and it is also lesser than that of auc score of the Logistic regression model on training dataset.

****

**Fig 2.20**

### **2.4)** Inference: Based on these predictions, what are the insights and recommendations.

From the results of data analysis, data exploration and model prediction Logistic Regression model is the best model for the given scenario.

* Frontal damage count is the most, which indicates that the driver is careless and not very active while driving. So necessary training or sleep alert can be usual if implemented.
* Most of the cars didn’t deploy or have airbags. Airbags in cars should be made mandatory to save more lives.
* Just like Airbags, seatbelts must wear mandatorily and severe punishments and fee should be given if seatbelt is not used by both driver and co passenger.
* Survival rate is quite good when compared to death. But still life is life so necessary deep dive is needed in data collection to find the exact reason.