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

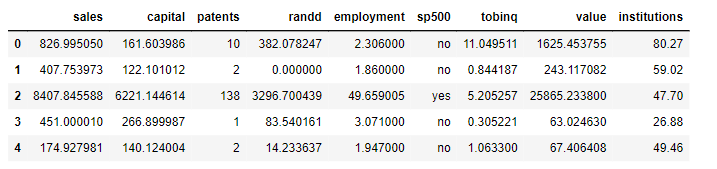
7. tobinq: Tobin's q (also known as q ratio and Kaldor's v) is the ratio between a physical asset's market value and its replacement value.

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.

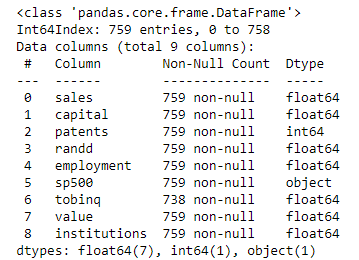


**Table 1.1**

There are 9 variables in the given dataset.

## **EXPLORATORY DATA ANALYSIS**

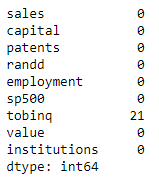
Let’s see the different types of variables and number of datapoints each has, data types each of variable and missing values.



**Table 1.2**

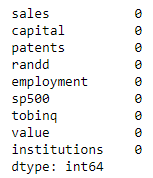
The table 1.2 shows that there are 7 float type features, one integer and one object data type each having 759 entries.

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**

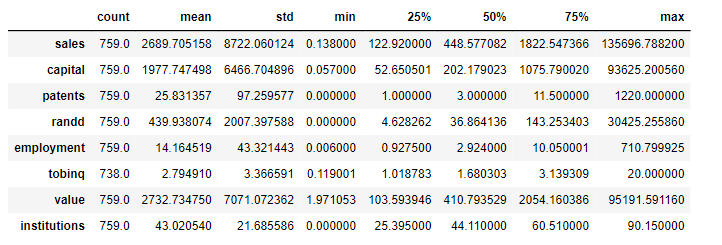
The null values are treated and the table 1.4 shows the output.



**Table 1.4**

Descriptive Analysis

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



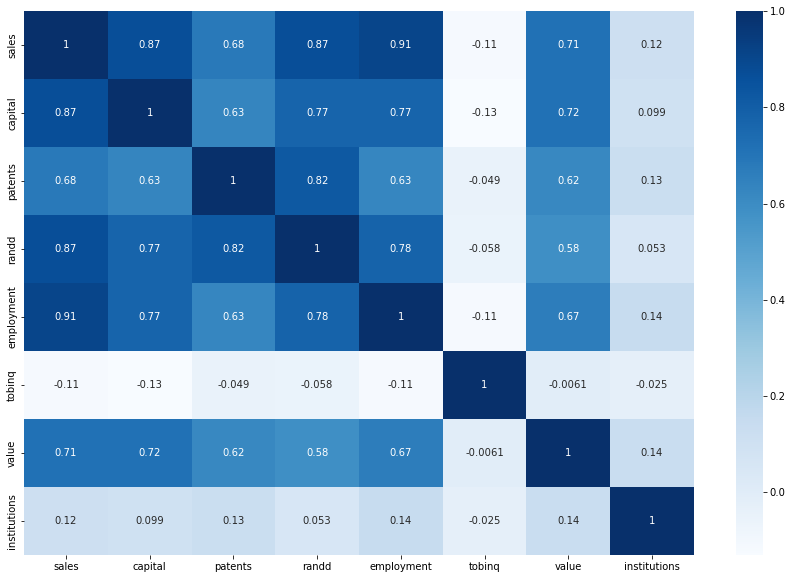
**Table 1.5**

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.



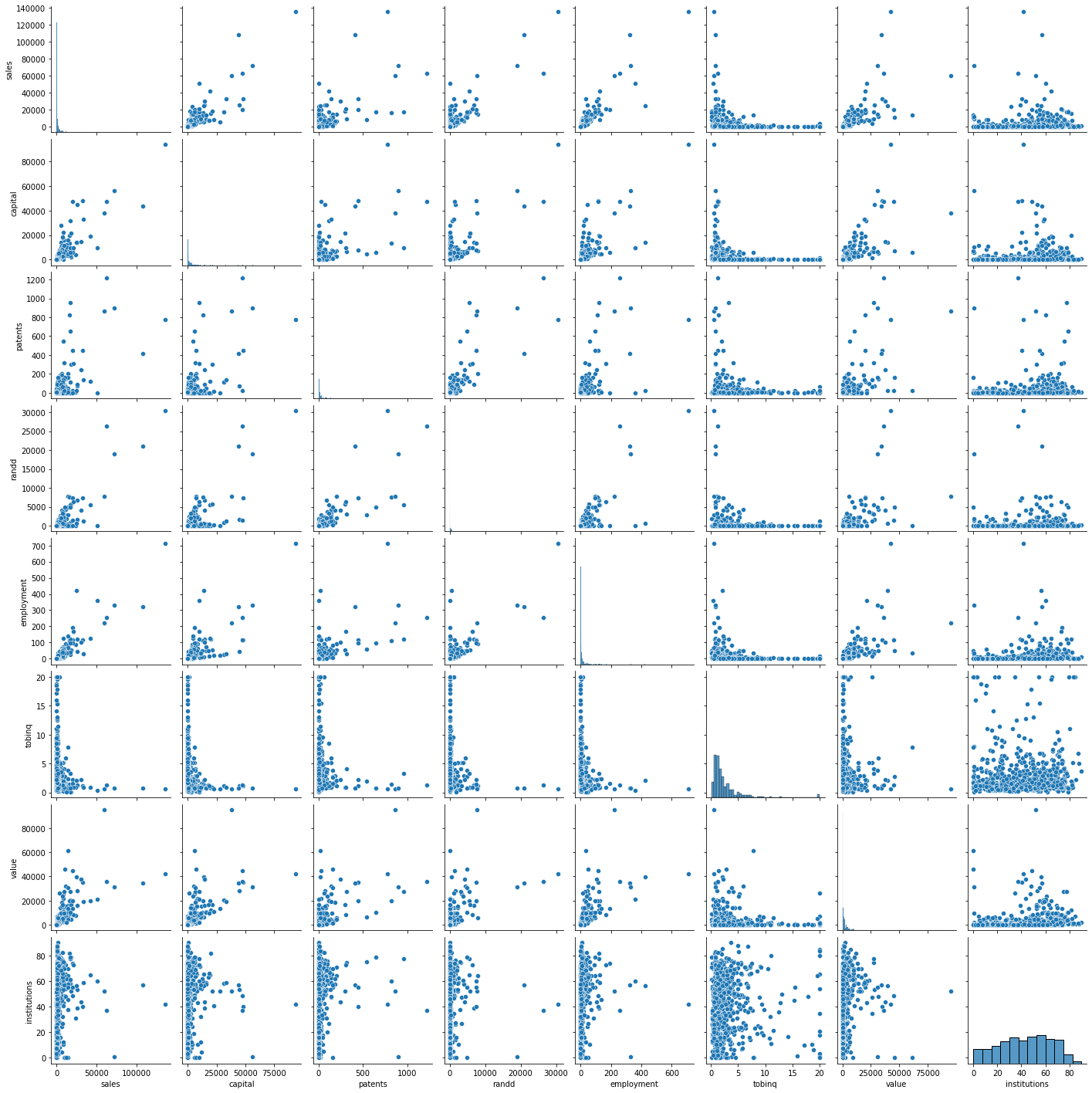
**Fig 1.1**

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:

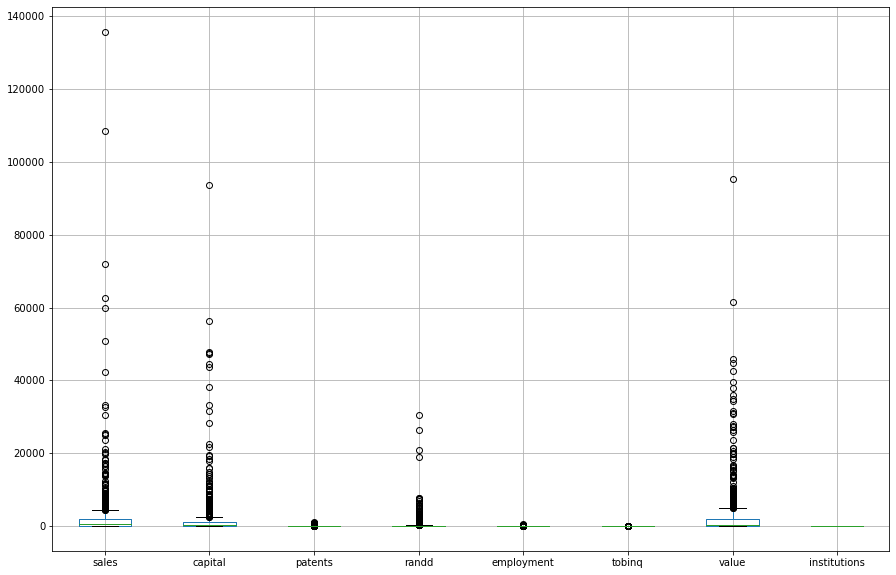
The fig 1.2 shows the pair plot of the given dataset.



**Fig 1.2**

Outliers:

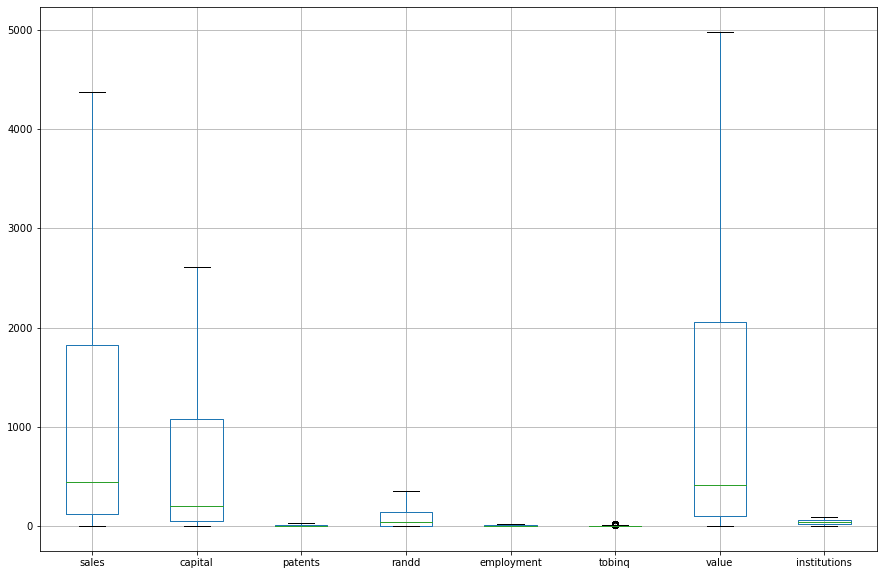
The outliers present in the dataset have been checked and it is treated. Fig 1.3 shows the boxplot of the numeric variables present along with its outliers.



**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):

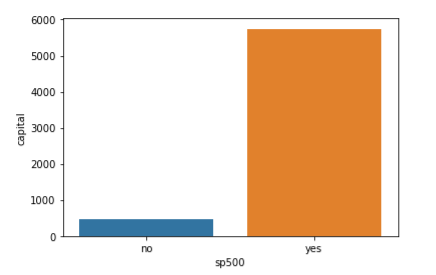
Let’s visualize the each variables and see what insights we can get from them.

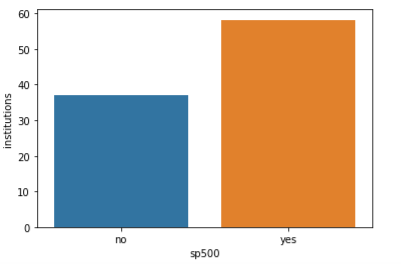
****

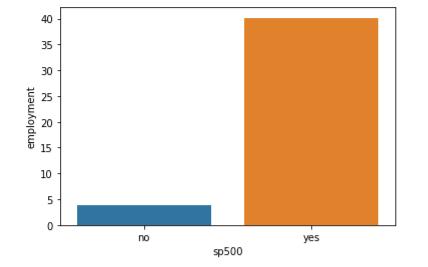
**Fig 1.5**

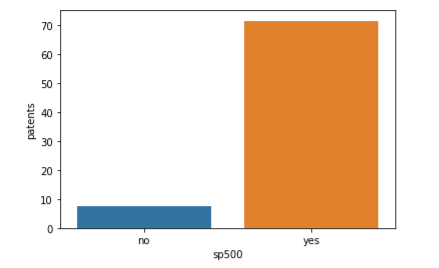
The fig 1.5 shows the counts of the membership of firms in the S&P 500 index. It can be clearly seen that most of the firms of are not a member of S & P 500 index.

Bivariate:

Let’s visualize the variables using bivariate analysis.



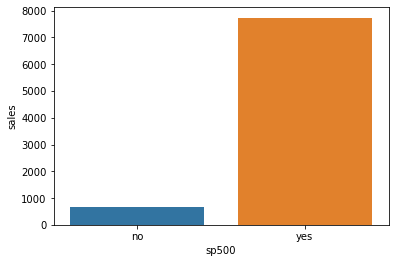
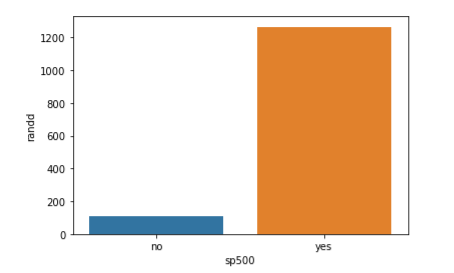


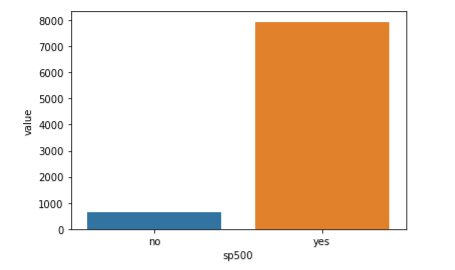
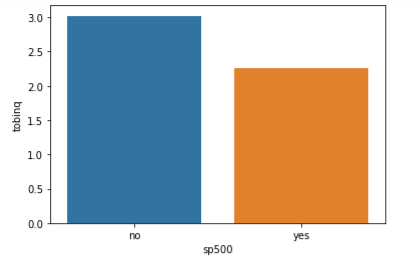


**Fig 1.6.1**

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.

****

****

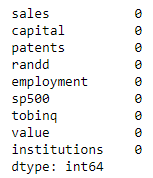
**Fig 1.6.2**

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.