|  |  |
| --- | --- |
| A picture of a winding road and trees  Finance and risk analytics  Project | **Created by-**  **SHUBHASREE SARKAR** |

**INDEX**

|  |  |  |
| --- | --- | --- |
| **Sr No** | **Topic: Credit Default Data** | **Page No** |
| **1** | **Problem Statement** | **3** |
| **2** | **Data Dictionary** | **4-5** |
| **3** | **1.1. Outlier Treatment** | **6-8** |
| **4** | **1.2 Missing Value Treatment** | **9** |
| **5** | **1.3 Transform Target variable into 0 and 1** | **10** |
| **6** | **1.4 Univariate (4 marks) & Bivariate ( 6marks) analysis with proper interpretation. (** | **11-15** |
| **7** | **1.5 Train Test Split** | **16** |
| **8** | **1.6 Build Logistic Regression Model (using statsmodel library) on most important variables on Train Dataset and choose the optimum cutoff. Also showcase your model building approach** | **17-21** |
| **9** | **1.7 Validate the Model on Test Dataset and state the performance matrices. Also state interpretation from the model** | **22-25** |
| **10** | **Recommendations** | **25** |
| **11** | **Reference** | **25** |

**Topic: Credit Default Data**

|  |  |
| --- | --- |
| **Description** | **Page no.** |
| **Boxplots with outliers before treatment- Figure 1** | **6-7** |
| **Boxplots with outliers after treatment- Figure 2** | **8** |
| **Default vs Non-Default- Figure 3** | **10** |
| **Check the shape of the dataset- Figure -4** | **12** |
| **Distribution plot – Figure-5** | **13** |
| **Correlation plot – Figure-6** | **15** |
| **Confusion matrix-Train dataset 1– Figure- 7** | **24** |
| **Confusion matrix-Test dataset 1 - Figure - 8** | **24** |
| **Confusion matrix-Train dataset 2 – Figure-9** | **25** |
| **Confusion matrix-Test dataset 2 – Figure- 10** | **25** |

**List of Figures**

**List of Tables**

|  |  |
| --- | --- |
| **Description** | **Page no.** |
| **Missing values- Table 1** | **9** |
| **Check for Head/Sample of the dataset – Table 2** | **11** |
| **Check for info – Table 3** | **11** |
| **Check for Summary stats- Table** *4* | **12** |
| **Statsmodel output – 1** | **19** |
| **Statsmodel output- 2** | **20** |
| **VIF values – Table-7** | **21** |
| **Classification report – Train dataset-1** | **24** |
| **Classification report – Test dataset-1** | **24** |
| **Classification report – Train dataset-2** | **25** |
| **Classification report – Test dataset-2** | **25** |

**Topic: Credit Default Data**

**Problem Statement**

Businesses or companies can fall prey to default if they are not able to keep up their debt obligations. Defaults will lead to a lower credit rating for the company which in turn reduces its chances of getting credit in the future and may have to pay higher interests on existing debts as well as any new obligations. From an investor's point of view, he would want to invest in a company if it is capable of handling its financial obligations, can grow quickly, and is able to manage the growth scale.

A balance sheet is a financial statement of a company that provides a snapshot of what a company owns, owes, and the amount invested by the shareholders. Thus, it is an important tool that helps evaluate the performance of a business.

Data that is available includes information from the financial statement of the companies for the previous year (2015). Also, information about the Networth of the company in the following year (2016) is provided which can be used to drive the labeled field.

**Data Dictionary**

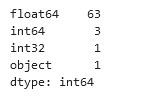


**1.1. Outlier Treatment**

**Outliers are basically the extreme values in the dataset.** Outliers increase the variability in your data, which decreases statistical power.

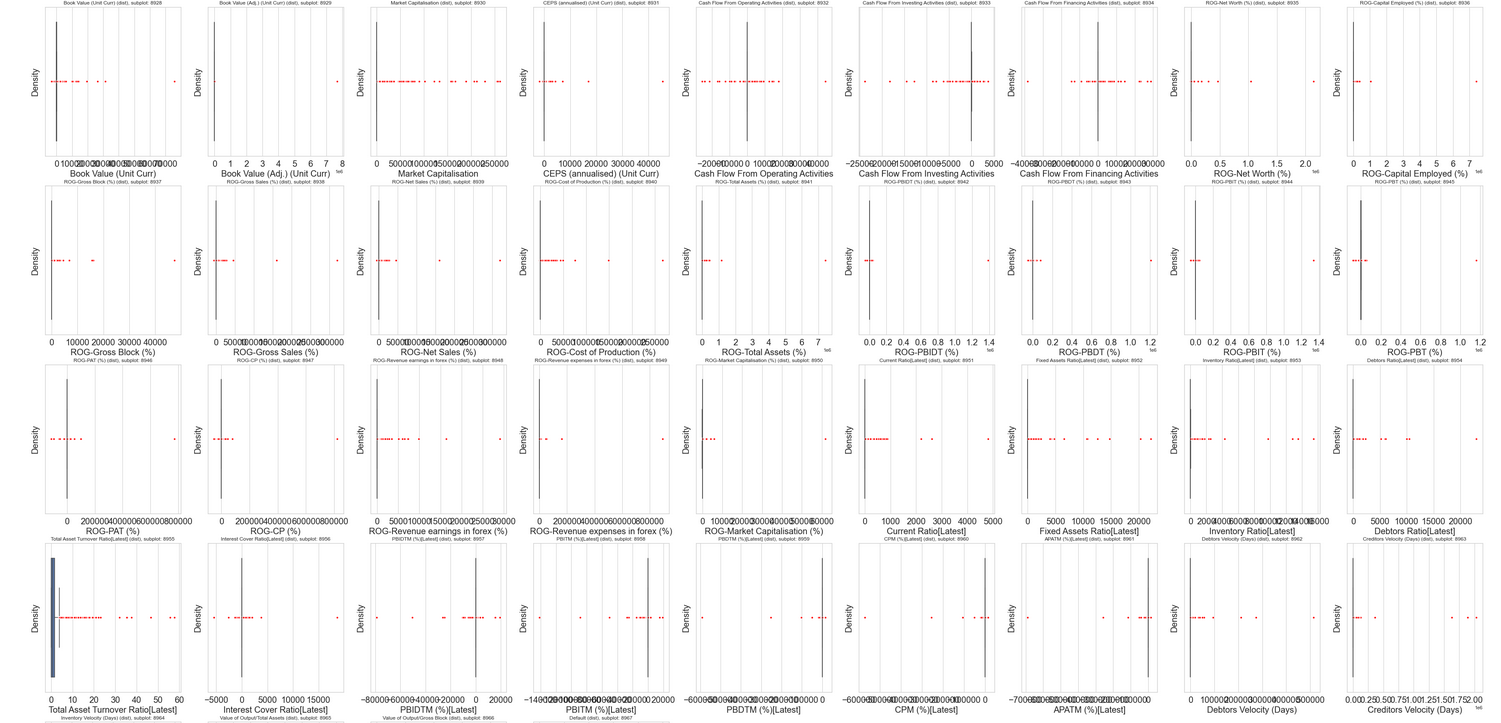
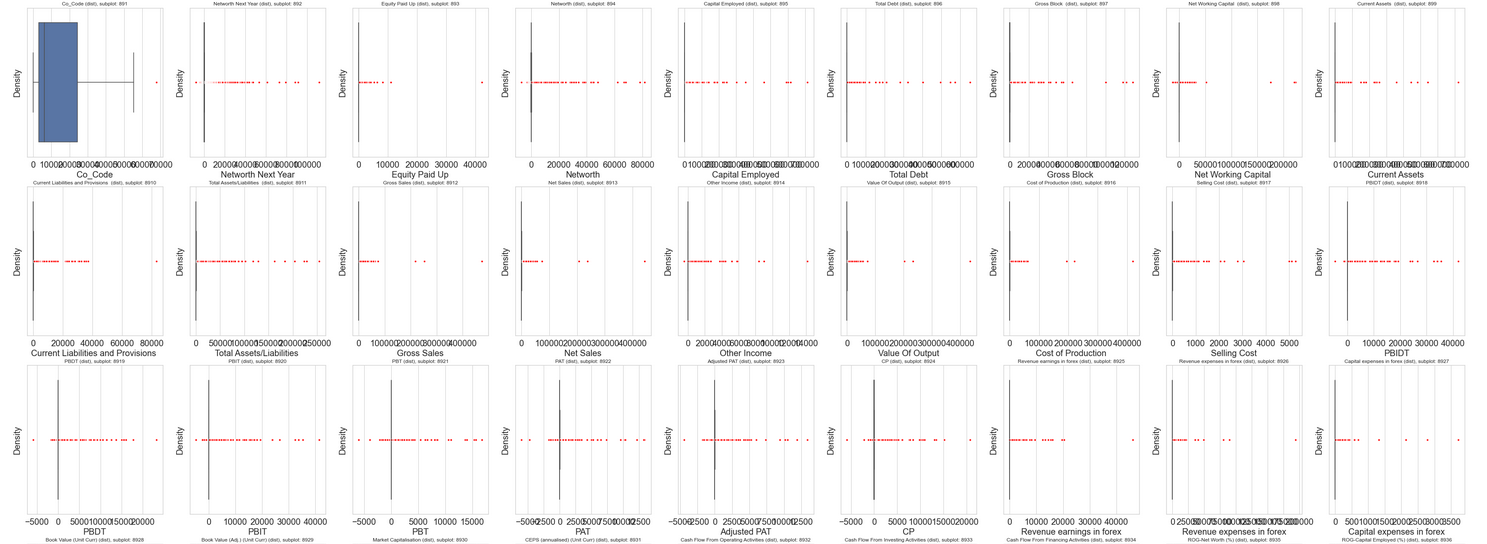
**As the dataset has huge no. of outliers in most of the continuous variables, we created a user defined function to remove the outliers from the data.**

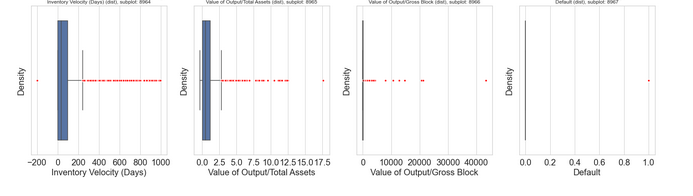
**First we need to check different data types in this dataset-**



We have dropped the object datatype (Co\_Names) from this dataset and went ahead with the overview regarding the outliers present. **The boxplot before outlier removal is show below. For the analysis after outlier treatment, we have dropped insignificant variables- Co\_code and Networth\_Next\_Year from the dataset.**

**Boxplots with outliers before treatment- Figure 1**





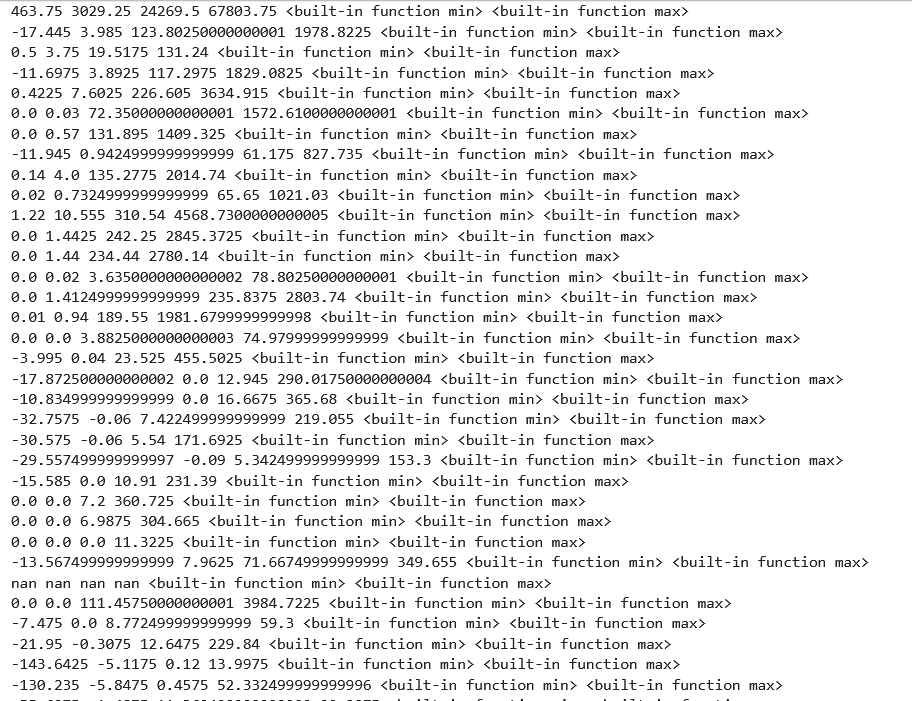
**The outliers are basically treated by using the Interquartile range for each of the numerical variables**

Values greater than Upper quartile range are being capped with 75% of quartile value

Values lesser than Lower quartile range are being capped with 25% of quartile value

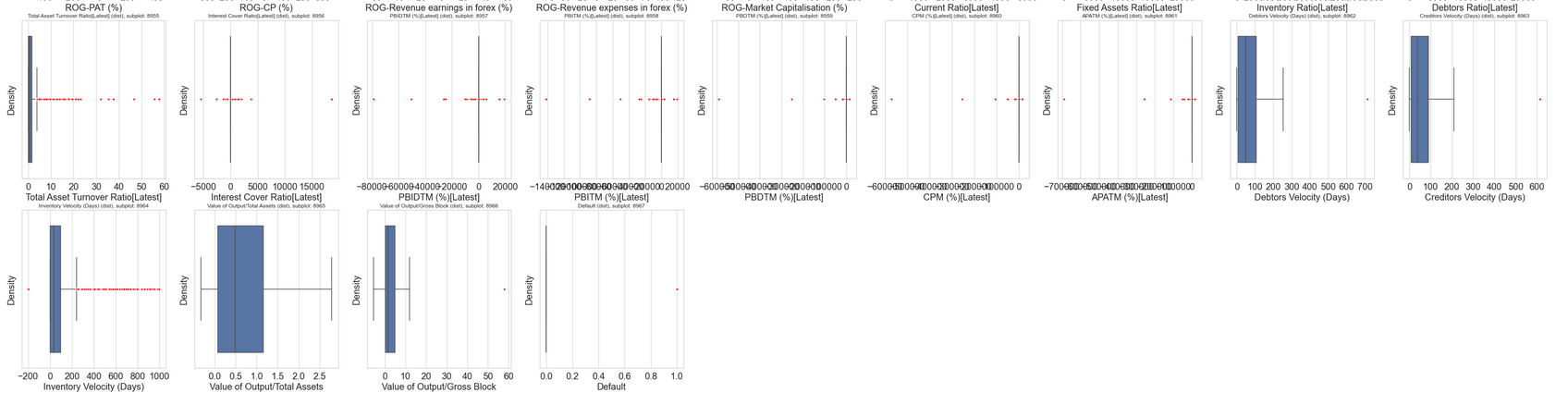
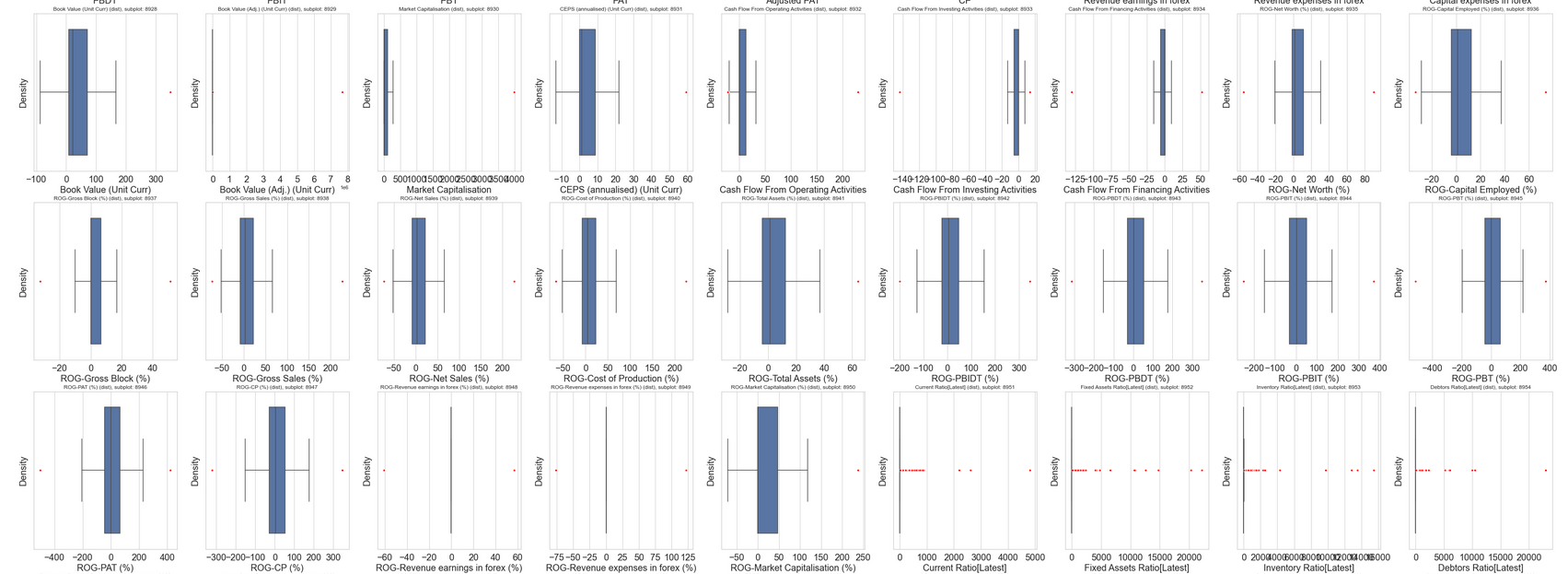
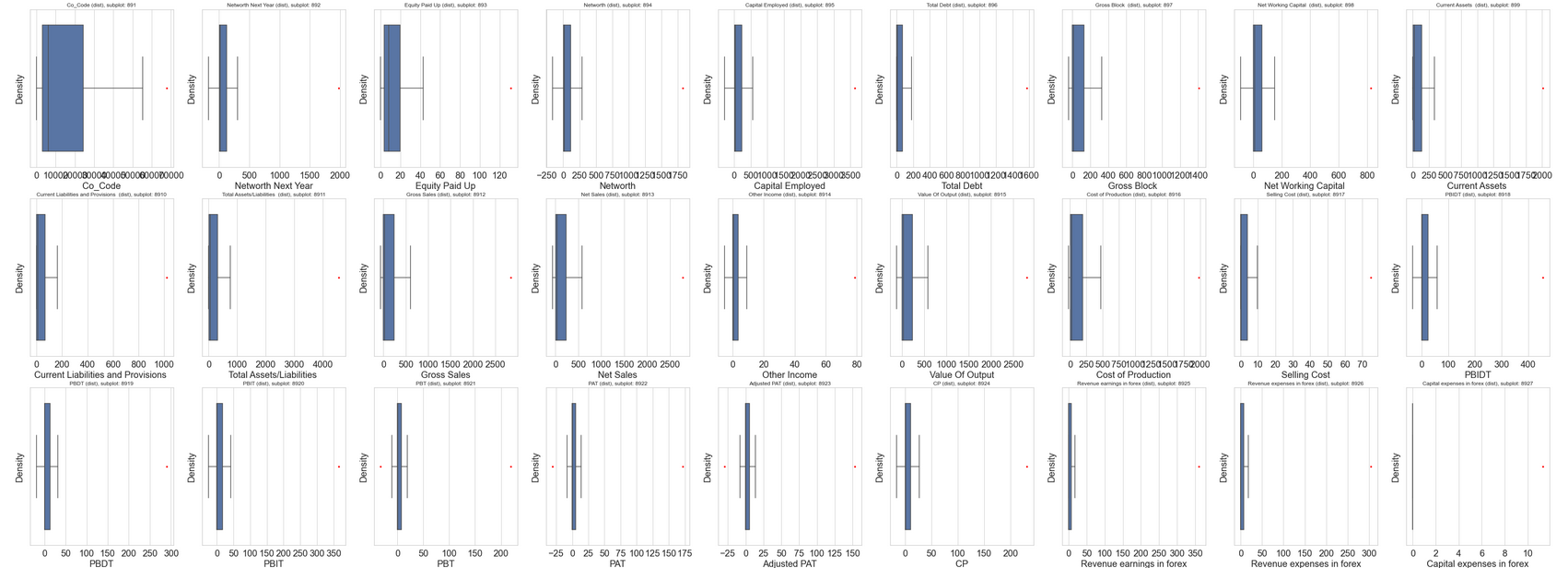
The outliers thus are being replaced with Upper Quartile values or Lower using the user defined functions

**The output of the outlier treatment is shown below-**



**The boxplot after outlier removal is shown below-**

**Boxplots with outliers after treatment- Figure 2**



**1.2 Missing Value Treatment**

The missing values or ‘NaN’ or ‘NA’ values are in general to be cleaned during the data cleaning process.

They are basically –

1. Removed or dropped
2. Replaced with mode function (Categorical variables)
3. Replaced with Median function (Continuous variables)
4. Replace with other processes

In this dataset, missing value is present. The same has been shown below-

**Missing values- Table** **1**

**The most no. of missing values are being found within Inventory Velocity (Days).**

All the variables in this case being continuous in nature, hence we will replace the missing ones with the **median values** for each one of them.

**1.3 Transform Target variable into 0 and 1**

In this business problem, there is no target variable defined – but since the objective here is basically to build a model for investor to decode or determine which company to invest in – the variable Networth\_Next\_Year could be used to transform into target variable.

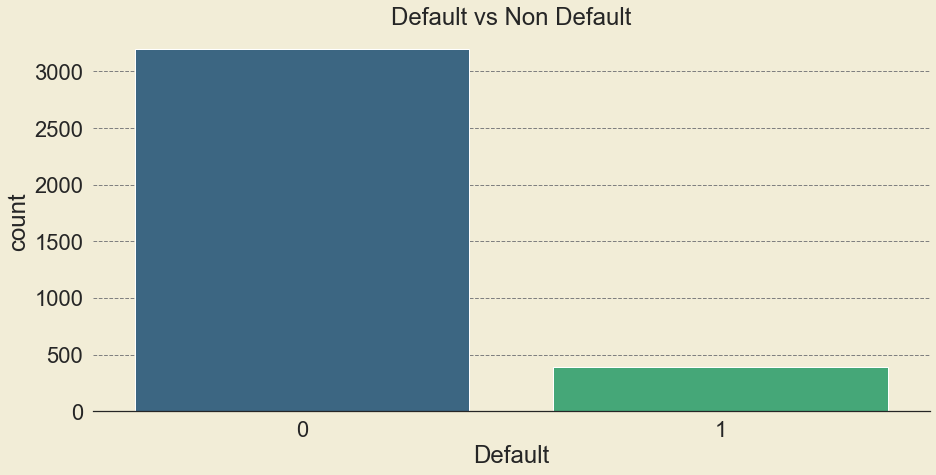
In this case, the target variable is created based on the following criteria-

Dependent variable - We need to create a default variable that should take the value of 1 when net worth next year is negative & 0 when net worth next year is positive.

Thus it means, if the company’s Networth Next Year is positive and greater than 0 – then the company would continue to return good investment for investor and thus could be transformed as 0 –NON-DEFAULT

And, if the company’s Networth Next Year is negative and lesser than or equal to 0 – then the company would not like to return good investment for investor and thus could be transformed as 1 – DEFAULT

**Default vs Non-Default- Figure 3**



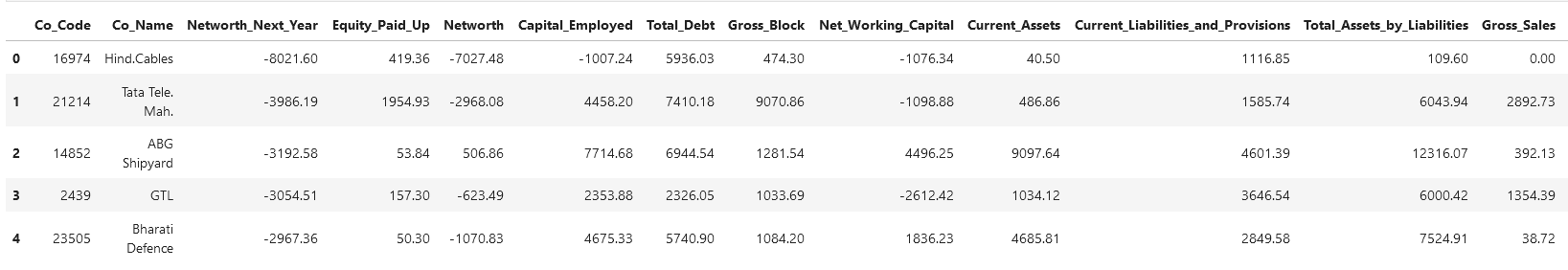
From the above visualization, we can say that the data has been imbalanced. The overall distribution is being stated below-

**Non Default(0) = 89.20%**

**Default (1) = 10.79%**

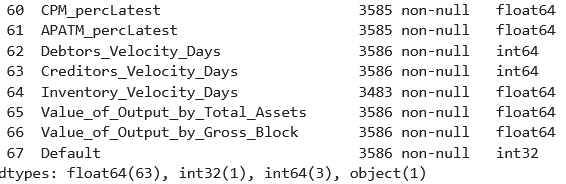
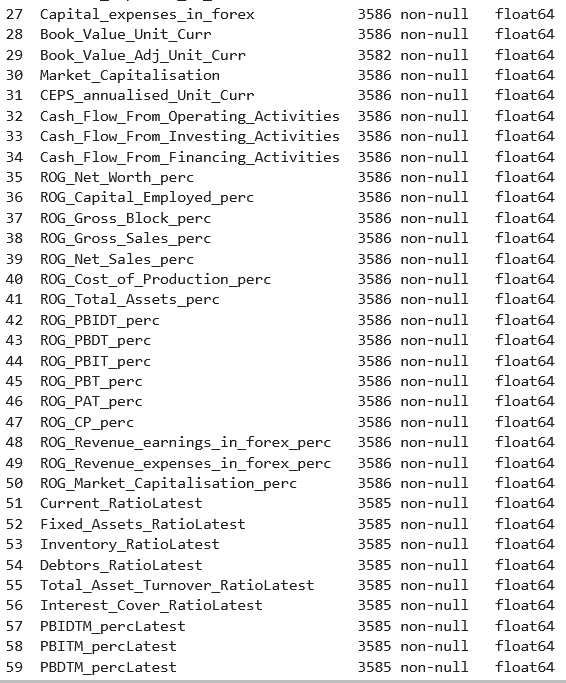
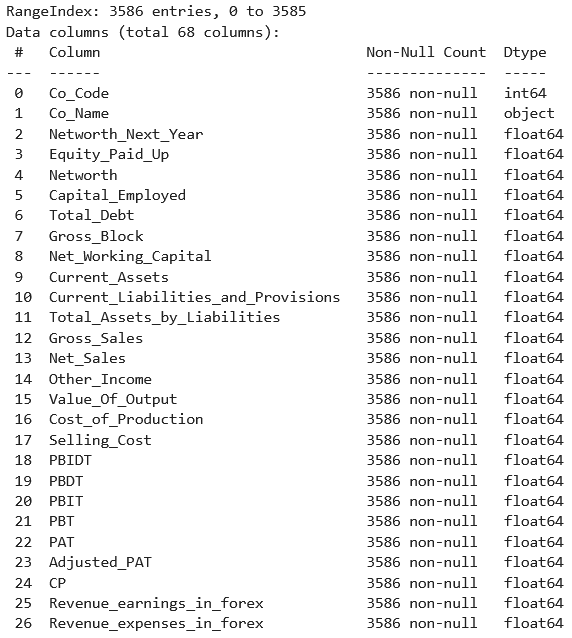
**1.4 Univariate (4 marks) & Bivariate ( 6marks) analysis with proper interpretation. (You may choose to include only those variables which were significant in the model building)**

**Check for Head/Sample of the dataset – Table 2**



The above figure shows the output of the head function, showing top 5 records with few of the variables out of 68 in total

**Check for info – Table 3**



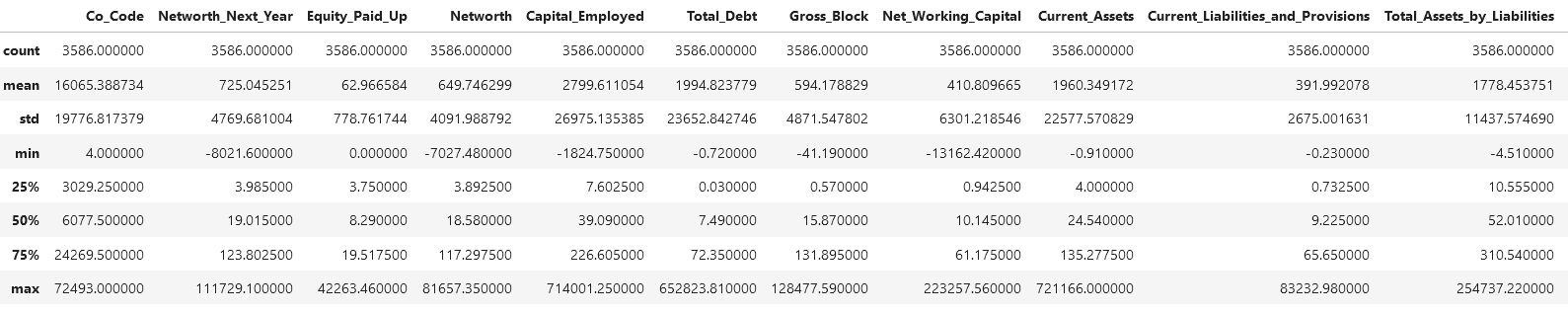
The above figure shows the different datatypes using the info function. To have a better count of the datatypes, we have the output of the value counts function as follows.



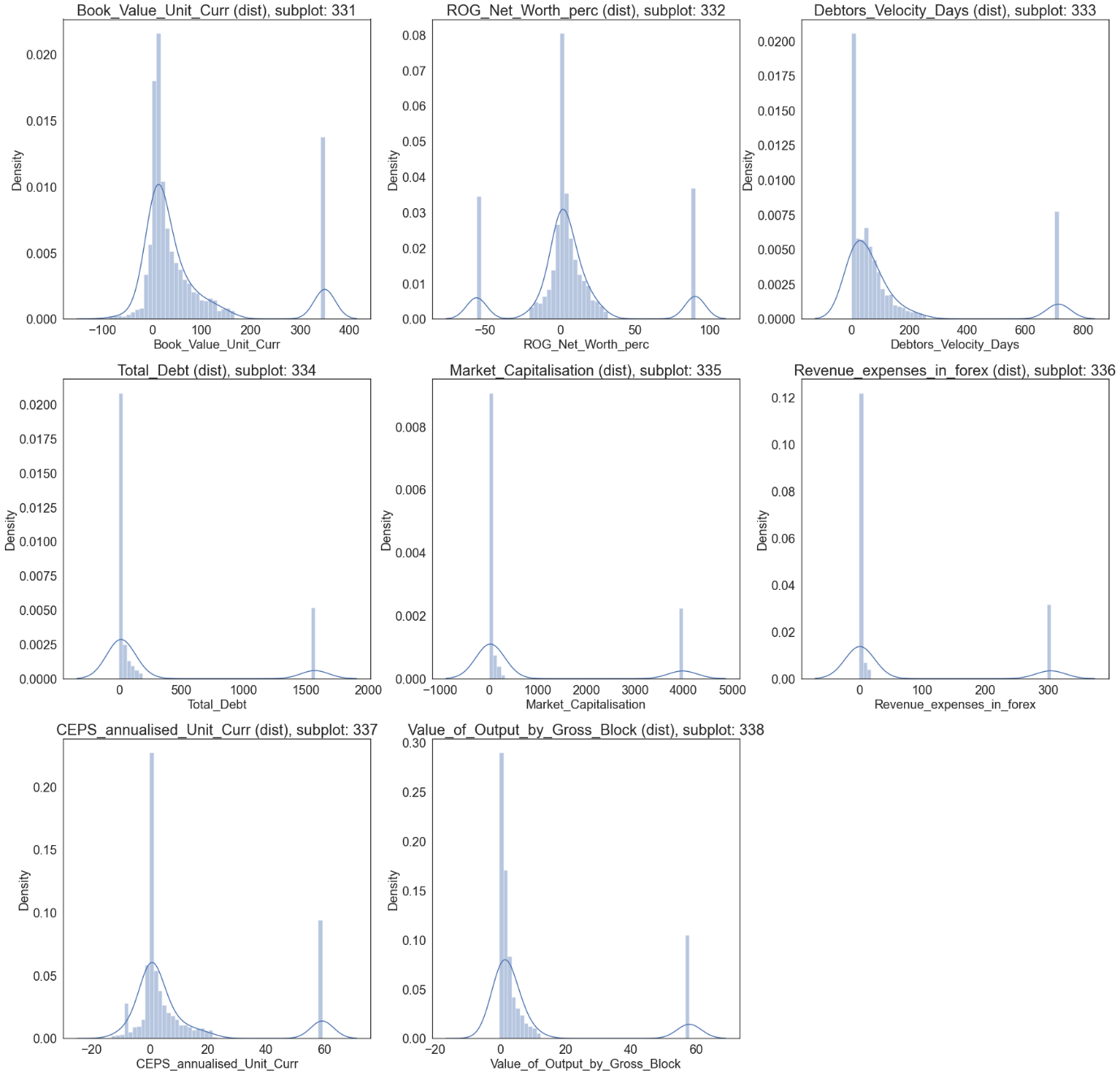
**Check the shape of the dataset- Figure -4**



The dimensions of the whole dataset is shown in the above figure

**Check for Summary stats- Table 4**

**Distribution plot – Figure-5**

****

**After plotting the histogram plots for all the numerical variables, the following can be concluded-**

* Multi modal distribution and bimodal distribution can be seen in few of the variables.
* Book\_Value\_Unit\_Curr - The distribution here seems to have Bimodal appearance. The peak in the first mode, range(-100->100), has the highest frequency, while the peak in the second mode, range(300-> 400) has the second highest frequency
* ROG\_Net\_Worth\_perc - The distribution here seems to have Trimodal appearance. The highest peak range is approximately lying between the modes with range between -10-> 50 units.
* Debtors\_Velocity\_Days - The distribution here seems to have Bimodal appearance. The highest peak range is approximately lying between the mode with range between 0-> 200 units.
* Total\_Debt - The distribution here seems to have Bimodal appearance. The highest peak range is approximately lying between the mode with range between 0-> 500 units.
* Market\_Capitalisation - The distribution here seems to have Bimodal appearance. The highest peak range is approximately lying between the mode with range between 0-> 500 units.
* Revenue\_expenses\_in\_forex - - The distribution here seems to have Bimodal appearance. The highest peak range is approximately lying between the mode with range between 0-> 50 units.
* CEPS\_annualised\_Unit\_Curr - The distribution here seems to have Bimodal appearance. The highest peak range is approximately lying between the mode with range between -20-> 20 units.
* Value\_of\_Output\_by\_Gross\_Block- The distribution here seems to have Bimodal appearance. The highest peak range is approximately lying between the mode with range between 0-> 20 units.

**Correlation plot – Figure-6**

Heatmap or Corrleation plot is basically being used to evaluate the relationship between the different numeric variables within a dataset

****

**From the above Correlation plots using Heatmap, the following facts can be concluded-**

* Presence of correlation can be observed between the features like-

Book\_Value\_Unit\_Curr and CEPS\_annualised\_Unit\_Curr, while that between Debtors\_Velocity\_Days and Total\_Debt it is significantly low.

* In case of multicollinearity, it is not very significant between most of the independent variables.
* Default has moderate to low correlation with that of the independent features like Book\_Value\_Unit\_Curr and ROG\_Net\_Worth\_perc.
* In case of Total\_Debt, it has moderate correlation with Market\_Capitalisation and Revenue\_expenses\_in\_forex

**1.5 Train Test Split**

The Target variable in this case is the new variable created based on Networth\_Next\_year and is named Default .

The independent variables such as **Co\_Code , Co\_Name, Networth Next Year** are not contributing much to this model, which is why they are being dropped before the model development.

The remaining dataset is thus split into train and test data into 67: 33 ratio with random state being 42 and stratified on default to make sure both train and test data have similar proportion of defaulters and non-defaulters. This is being done as the dataset is imbalanced with more of non-defaulters as compared to defaulters.

Model Building is to be done on Train Dataset and Model Validation is to be done on Test Dataset.

**Train- dataset**



**Test- dataset**



**1.6 Build Logistic Regression Model (using statsmodel library) on most important variables on Train Dataset and choose the optimum cutoff. Also showcase your model building approach**

**LOGISTIC REGRESSION**

Logistic regression is a classification algorithm used to assign observations to a discrete set of classes. Some of the examples of classification problems are Email spam or not spam, Online transactions Fraud or not Fraud

Logistic regression transforms its output using the logistic sigmoid function to return a probability value.

Logistic Regression is thus a Machine Learning algorithm which is used for the classification problems, it is a predictive analysis algorithm and is based on the concept of probability.

## **What are the types of logistic regression**

* Binary (eg. Tumor Malignant or Benign)
* Multi-linear functions fails -Class (eg. Cats, dogs or Sheep's)

*#Steps involved*

Before the model development, we have to work on feature selection and using VIF or Variation Inflation Factor helps to eliminate the insignificant features

Using the train dataset, we have created Logistic model and then further testing the same on the test dataset.

For the Sigmoid formation, we have imported the LogisticRegression module from the **statsmodel library**

With the help of afore mentioned package, Logistic regression model is created, in order to fit the training data into this model.

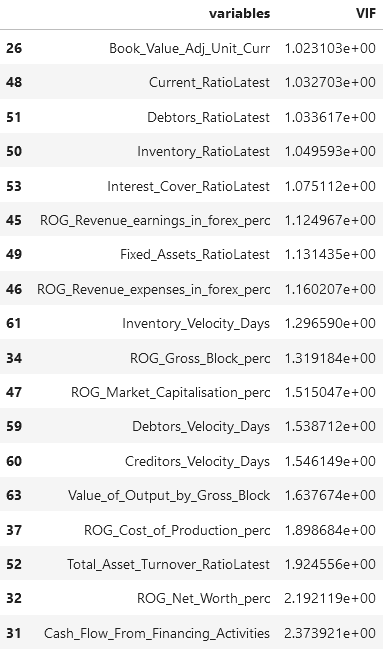
Once the model is being created after eliminating the variables on the basis of VIF and P-value, we have applied the model building formula in order to predict the respective Target variable on the train and test dataset

**Variation Inflation Factor**

###### Variance inflation factor (vif) is a measure of the amount of multicollinearity in a set of multiple regression variables.

###### Mathematically, the vif for a regression model variable is equal to the ratio of the overall model variance to the variance of a model that includes only that single independent variable.

###### This ratio is calculated for each independent variable. A high vif indicates that the associated independent variable is highly collinear with the other variables in the model.



**Here, we see that the value of VIF is high for many variables. Hence, we have considered those variables with VIF less than 5 (very high correlation) to build our model**

**The output from statsmodel where we have taken only those variables with VIF < 5**

**Statsmodel output - 1**

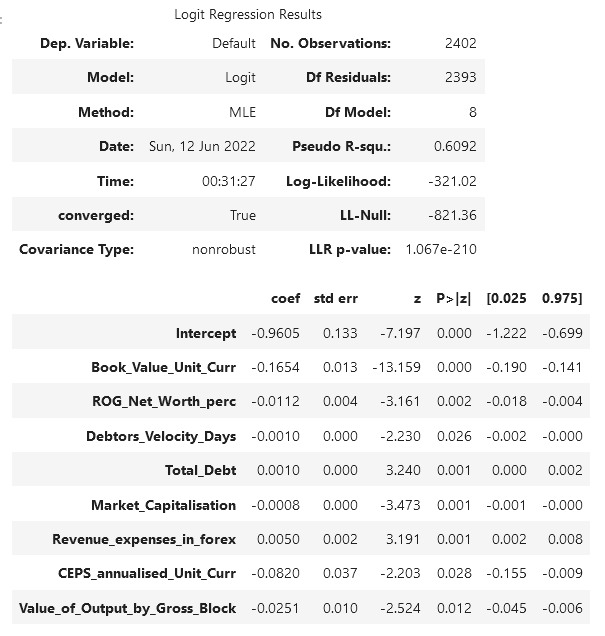




In the above case, the **Adjusted pseudo R-square seems to be lower than Pseudo R-square value which means there are insignificant variables present in the model. Hence, we tried to remove variables whose p value is greater than 0.05 and rebuild our model.**

**The output from statsmodel where we have taken only those variables with Pvalue <0.05**

**Statsmodel output- 2**



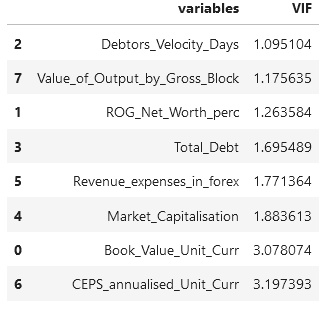


We see that adjusted R sq is now close to Rsq, thus suggesting lesser insignificant variables in the model

We also noticed that current model has no insignificant variables and can be used for prediction purposes.

The VIF of the final no. of variables in the model are-

**VIF values – Table-7**

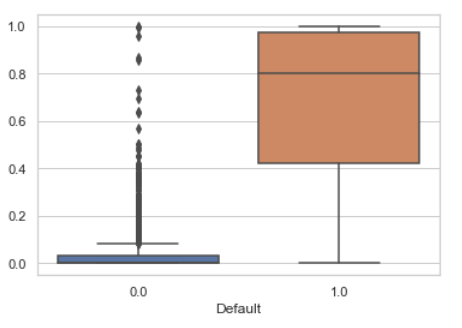


**The optimum cut off of 0.07 is being considered first and later being compared with a cut off of 0.08. The details is being discussed in the model validation based on test output.**

**1.7 Validate the Model on Test Dataset and state the performance matrices. Also state interpretation from the model**

The model has been validated on the basis of various parameters taken into consideration.

First on the basis of the predicted values based on the train dataset, we have tried to use a boxplot to differentiate between the Default and Non-Default values .



**So, based on the boxplot ,we are required to decide on one such value of a cut-off which will give us the most reasonable descriptive power of the model. At first, we took a cut-off of 0.07 and checked.**

The parameters being considered are stated below

***ACCURACY***

It is a part of metrices derived from confusion matrix which is basically a **NxN** matrix, where **N** is the **number of classes to be predicted**

It is the **proportion** of the total number of **predictions** that were **correct**.

It is easily suited for **binary** as well as a **multiclass** **classification** **problem** which are **well** **balanced** and **not skewed** or **No class imbalance**.



***PRECISION***

Similar to the Accuracy, it is a metric derived from confusion matrix

**Positive Predictive Value or Precision** is also defined as the proportion of positive cases that were correctly identified.

In other words, it determines the proportion of **predicted Positives**which is truly Positive

Precision is a valid choice of evaluation metric when we want to be very sure of our prediction.



***RECALL***

Similar to the Precision, it is a metric derived from confusion matrix.

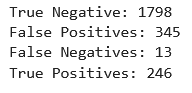
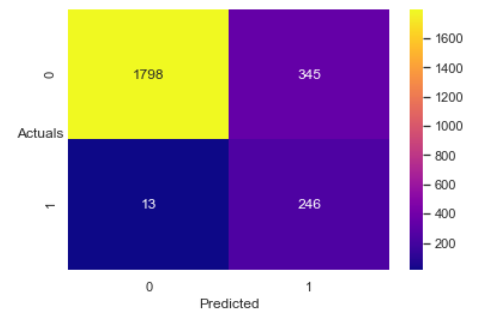
**Sensitivity or Recall** is also defined as the proportion of actual positive cases which are correctly identified.

In other words, it determines the proportion of **actual Positives**is correctly classified

Recall is a valid choice of evaluation metric when we want to capture as many positives as possible.



**Confusion matrix-Train dataset 1– Figure- 7**



Overall confusion matrix shows that True positives and negatives are being correctly determined in majority, but it has high no. of false positives as well

**Classification report – Train dataset-1**

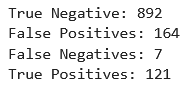
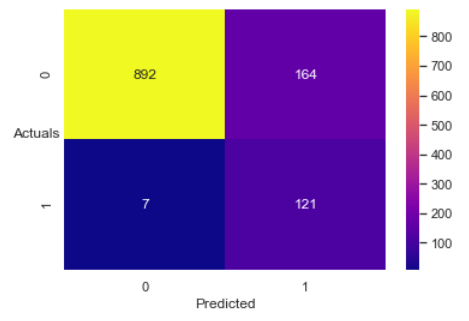


As observed above, accuracy of the model i.e. %overall correct predictions is 85.1%

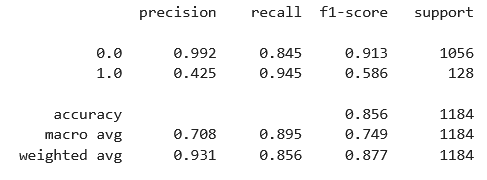
Sensitivity of the model is 95% i.e. 95% of those defaulted were correctly identified as defaulters by the model

**Confusion matrix-Test dataset 1 - Figure - 8**

Overall confusion matrix shows that True positives and negatives are being correctly determined in majority, but it has high no. of false positives as well



**Classification report – Test dataset-1**

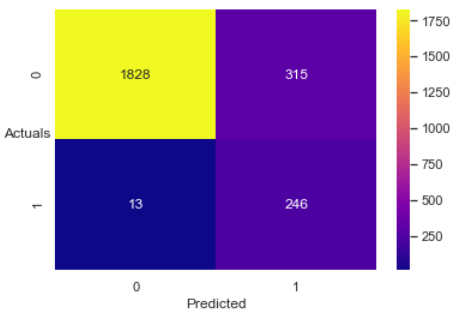


As observed above, accuracy of the model i.e. %overall correct predictions is 85.6%

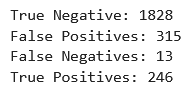
Sensitivity of the model is 94.5% i.e. 94.5% of those defaulted were correctly identified as defaulters by the model

**Now we took a cut-off of 0.08 and checked.**

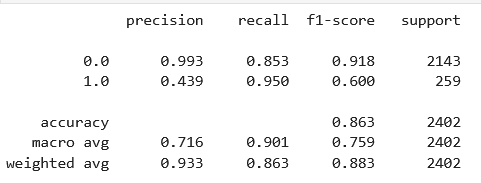
**Confusion matrix-Train dataset 2 – Figure-9**



Overall confusion matrix shows that True positives and negatives are being correctly determined in majority with an increase in no., but it has high no. of false positives as well



**Classification report – Train dataset-2**

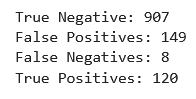
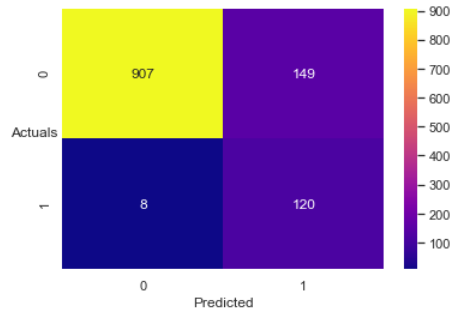


As observed above, accuracy of the model i.e. %overall correct predictions **has increased from 85.1% to 86.3%**

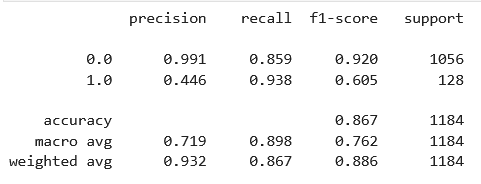
Sensitivity of the model is 95% i.e those defaulted **has been same as before ie., 95%**

**Confusion matrix-Test dataset 2 – Figure-10**

Overall confusion matrix shows that True positives and negatives are being correctly determined in majority with an increase in no., but it has high no. of false positives as well



**Classification report – Test dataset-2**



As observed above, accuracy of the model i.e. %overall correct predictions **has increased from 85.6% to 86.7%**

Sensitivity of the model is decreased to 93.8% i.e those defaulted **has been lower than before ie.,** 93.8%

**We may choose cutoff of 0.08 as it gave higher model sensitivity & overall accuracy of the model in test dataset**

**Recommendations**

In Financial Industry or Healthcare industry, it is to be taken into consideration that the value of False negatives is low, which means a test result which wrongly indicates that a particular condition or attribute is absent. So, in the model building process, the false negative values was quite low compared to its true counterpart.

To increase the accuracy and to determine the classes in a better way, it is recommended to clarify the values at the data entry level and check accordingly as there can be mismatch in data which can lead to the increase in the no. of false positives and negatives.

**References**

* *https://adataanalyst.com/data-analysis-resources/visualise-categorical-variables-in-python/#:~:text=Visualise%20Categorical%20Variables%20in%20Python%20using%20Bivariate%20Analysis,a%20pre%2Ddefined%20significance%20level.*
* *https://www.investopedia.com/terms/v/variance-inflation-factor.asp#:~:text=Variance%20inflation%20factor%20(VIF)%20is,only%20that%20single%20independent%20variable.*

**\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\***