Jan 2021 | PG-DSBA-Online

FINANCE AND RISK ANALYTICS PROJECT REPORT

SUBMITTED BY
DEV TRIPATHI

Contents

List of Figures

Problem

Credit Default Dataset Analysis

Information about dataset

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 Net worth of the company in the following year (2016)** is provided which can be used to drive the labeled field.

Data Dictionary for Credit Default Dataset

The dataset contains total 67 features.

Explanation of data fields available in Data Dictionary, 'Credit Default Data Dictionary.xlsx'

EDA

Variable Information-

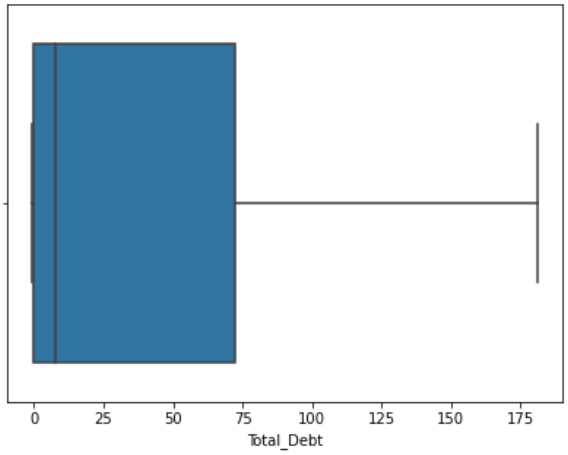
The dataset contains 3586 entries and 67 features including our target variable**'Networth_Next_Year'**. Datatype was integer and float for all the variables excel variable “Co_Name” which is an object type variable. Hence, no irregularities were found in data types

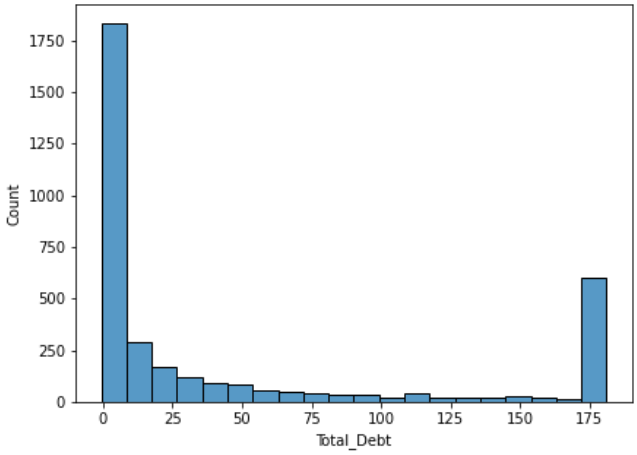
Null count was 1 for 12 variables in the dataset. Hence, it was better to simply drop the data point if it is same for each of those variables. After dropping that data point (whose Co_Name was **GM Breweries**) only two variables had missing values, namely **‘Inventory_Vel_Days’ (103)** and **‘Book_Value_Adj_Unit_Curr’ (4).**

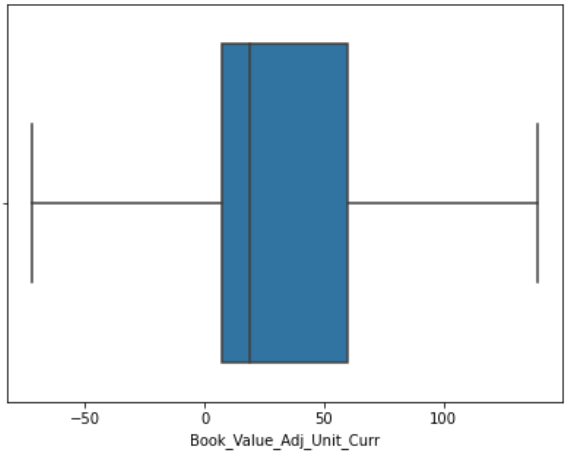
For getting interpretable plots, the missing value treatment and outliers treatment was done before proceeding to Univariate and Bivariate Analysis.

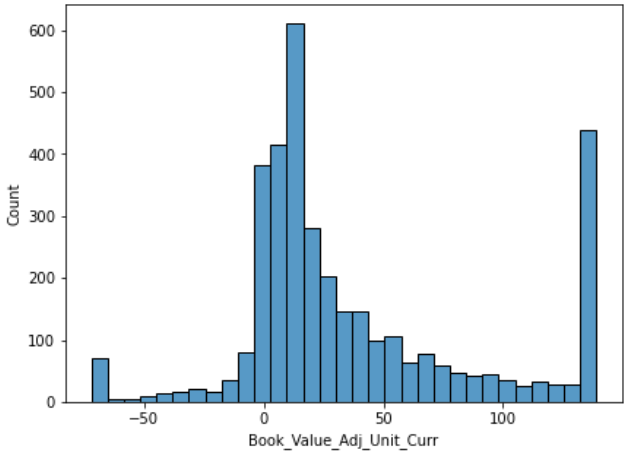
Now, for univariate and bivariate analysis we chose to include only those variables which were found out to be statistically significant later in the modelling. These variables are, namely **‘Equity_Paid_Up', 'Total_Debt', 'Book_Value_Adj_Unit_Curr', 'Market_Capitalisation', 'CEPS_annualised_Unit_Curr', 'Curr_Ratio_Latest', ‘BITM_perc_Latest' ,'Value_of_Output_to_Gross_Block'.**

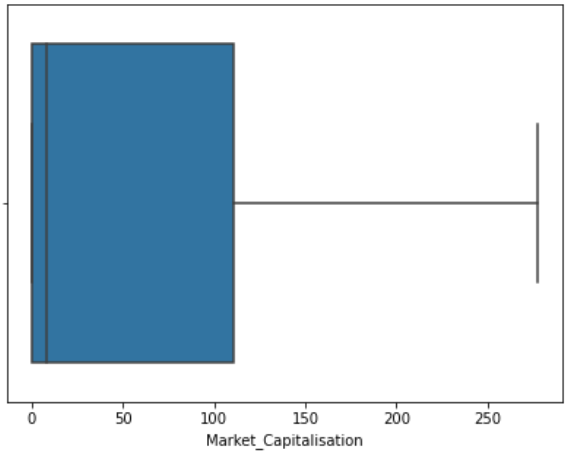
Univariate Analysis-

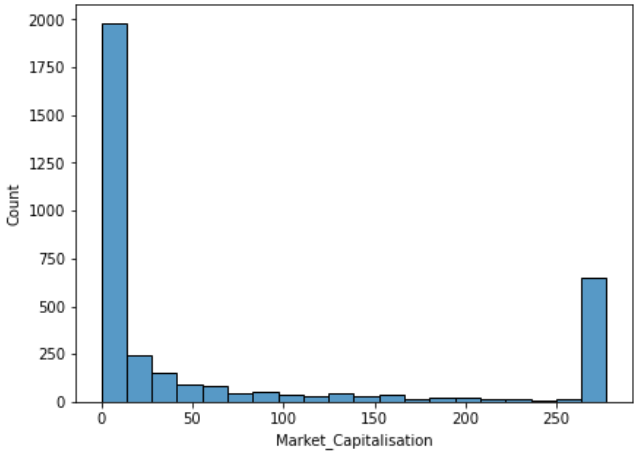


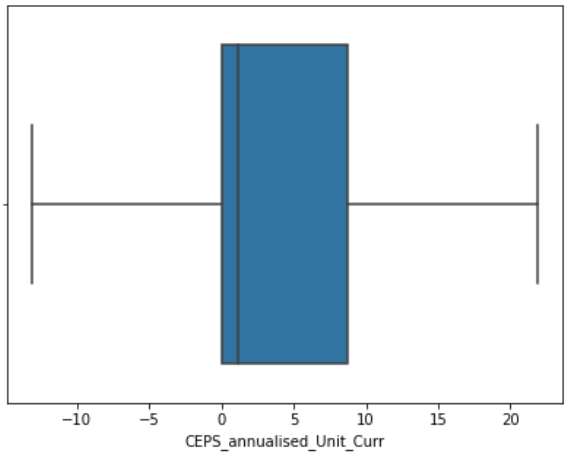


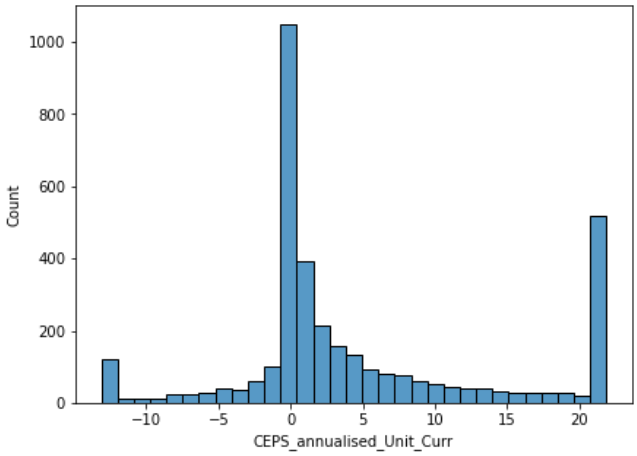


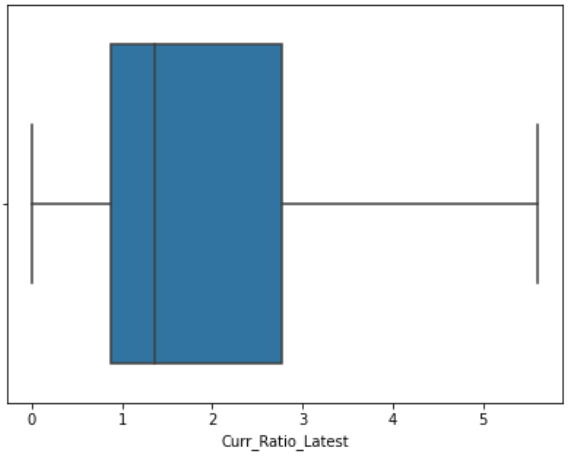


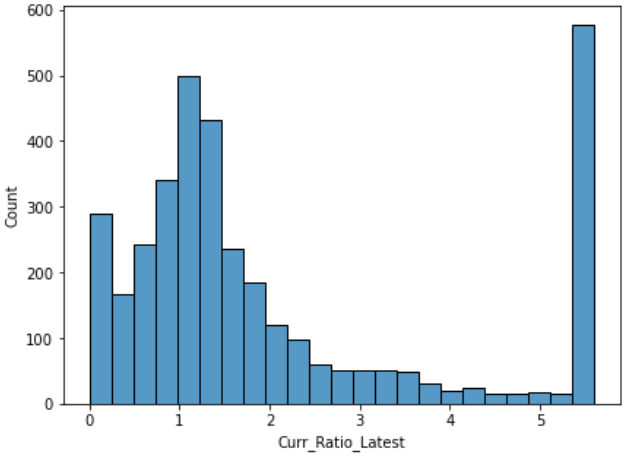

0
50
100
150
200
250
Market_Capitalisation


Count
2000
1750
1500
1250
1000
750
500
250
0
0
50
100
150
200
250
Market_Capitalisation


−10
−5
0
5
10
15
20
CEPS_annualised_Unit_Curr


Count
1000
800
600
400
200
0
−10
−5
0
5
10
15
20
CEPS_annualised_Unit_Curr


0
1
2
3
4
5
Curr_Ratio_Latest


Count
600
500
400
300
200
100
0
0
1
2
3
4
5
Curr_Ratio_Latest

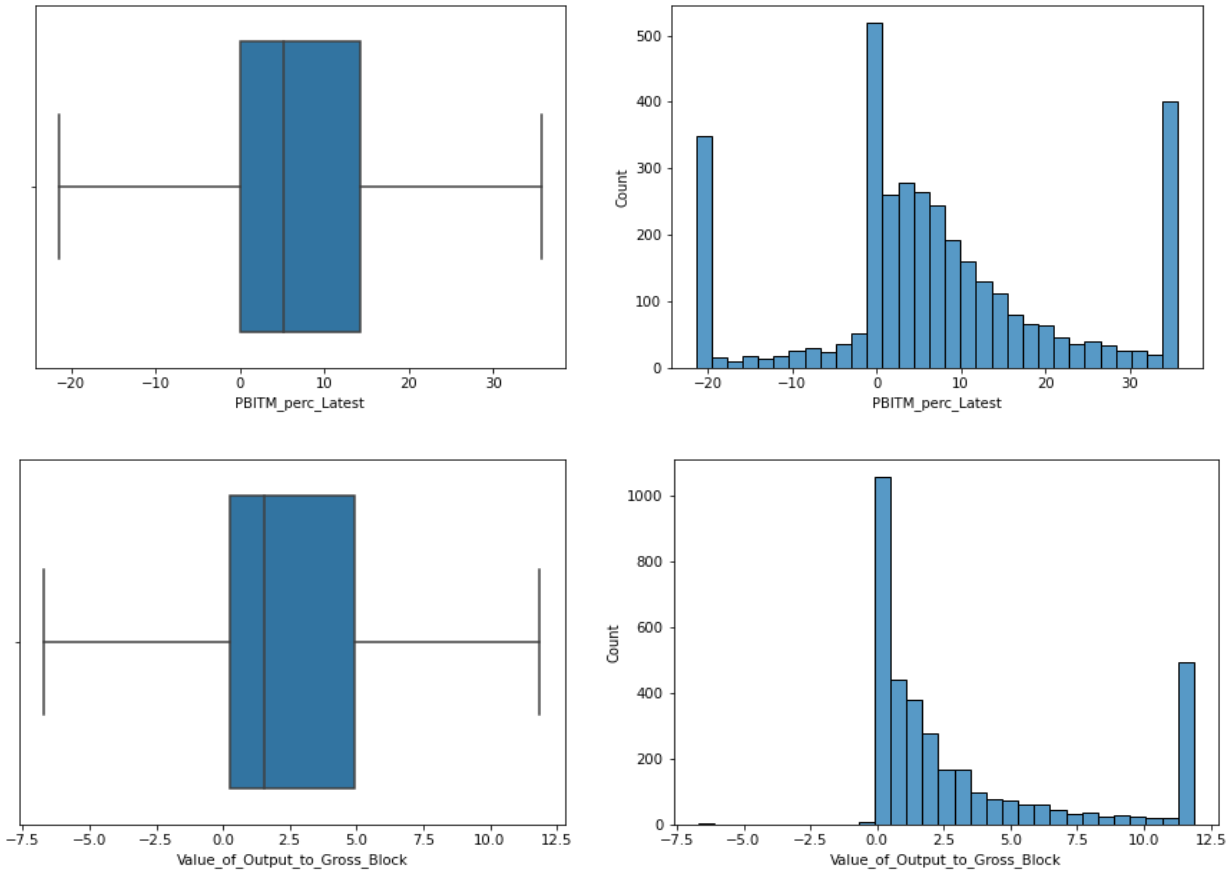


Figure 1: Univariate Curves for Significant features in the dataset

	Equity_Paid_Up	Total_Debt	Book_Value_Adj_Unit_Curr	CEPS_annualised_Unit_Curr	Curr_Ratio_Latest	Debtors_Ratio_Latest	PBITM_perc_Latest
count	3585.000000	3585.000000	3585.000000	3585.000000	3585.000000	3585.000000	3585.000000
mean	13.996070	47.475955	38.096106	4.805310	2.084286	5.990901	7.214545
std	14.004971	68.289989	50.058057	8.972073	1.806562	6.626860	15.284085
min	0.000000	-0.720000	-72.200000	-13.125000	0.000000	0.000000	-21.435000
25%	3.750000	0.030000	7.060000	0.000000	0.880000	0.420000	0.000000
50%	8.290000	7.480000	18.920000	1.140000	1.360000	3.820000	5.230000
75%	19.520000	72.430000	59.900000	8.750000	2.770000	8.520000	14.290000
max	43.175000	181.030000	139.160000	21.875000	5.605000	20.670000	35.725000

Figure 2: Statistical description of the 'significant' variables

Now, conclusions drawn from this analysis can be briefly explained in the following points:

1. All the distributions are right skewed.
2. The value of **'Total_Debt'** is between 0 to 72 for 50% of the observations.

3. The left whisker of **'Book_Value_Adj_Unit_Cur'** is at -72 with 50% of observations lying between 7 to 59.
4. **'Equity_Paid_Up'** is having 50% of the values lying between 3.75 to 19.52 and maximum value as 43.175.
5. **'CEPS_annualised_Unit_Curr'** is having 50% of the values lying between 0 to 8.75 with minimum as -13.125 and max as 21.875.
6. '**Curr_Ratio_Latest'** is having 50% of the values lying between 0 to 2.77 with min as 0 and max as 5.6m.
7. '**PBITM_perc_Latest'** which a percentage value, is having 50% of the values lying between 0 to 14 with minimum -21.4 as and maximum as 35.7.
8. **'Debtors_Ratio_Latest'** is having 50% of the values lying between 0.42 to 8.52 with minimum -0 and maximum 20.

Bi-Variate Analysis-

Now for bivariate analysis, we have created heatmap that simply shows a correlation for all the pairs of "significant" features present in the dataset.

Figure 3: Correlation plot for Credit Default Dataset

Now, one can observe that the ‘Equit_Paid_Up’ , ’Total_Debt’ and ‘Book_Value_Adj_Unit_Corr’ ,’CEPS_annaulisaed_Unit_Corr’ features have a moderate or high correlation between them (Which could be problematic for the logistic regression model). The ‘Networth_Next_Year’ can be observed to have moderate or high values with each of the fore mentioned variables. These correlations can be observed from the joint plots also.

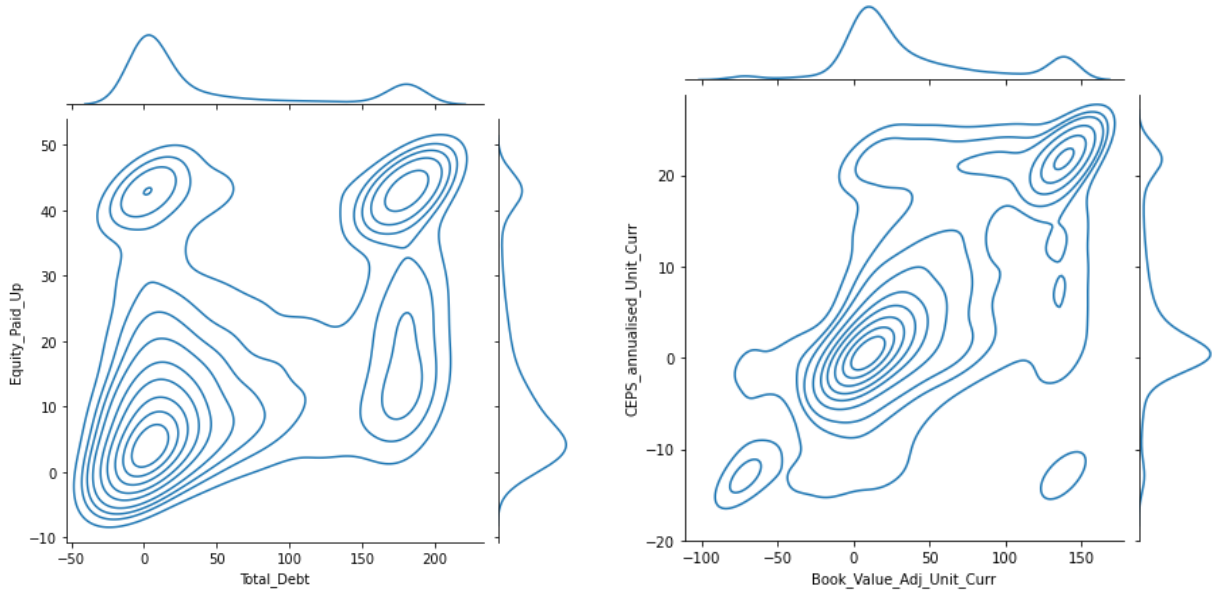


Figure 4: Joint Plots

Feature Selection

Feature Selection was performed as number of features present in the dataset are 67 which is quite high considering the interpretability of model. Hence, feature selection becomes a necessity in this scenario. Now, for feature selection various methods are available. Here, in this analysis we have used two methods:

1. Feature Selection based on VIF (Variation Inflation Factor)
2. Recursive Feature Elimination method (RFE)

Variance inflation factor 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.

$$VIF_i = \frac{1}{1 - R_i^2}$$

In general, if value of VIF for a variable is more than 5 then the variable should be removed as it is inducing high amount of multicollinearity in the dataset. Hence, using this thumb rule we were able to reduce the number of features to 33. But 33 is itself very large number. Hence, RFE was used for this purpose. But before applying RFE, **scaling of the data** and **train & test spilt [67:33]** was performed to fit the data RFE instance. Now, after some random iterations it was found that only 7 features are statistically significant for the logistic regression model. Hence, number of features to be selected in RFE instance was 7.

Logit model

Now, the Logistic regression model was fitted to train data, and predictions were made for train and test data. For logistic regression, statsmodel library was used which generates easily interpretable results, summary of which is given below:

Logit Regression Results

Dep. Variable:	y_train	**No. Observations:**	2401
Model:	Logit	**Df Residuals:**	2393
Method:	MLE	**Df Model:**	7
Date:	Sun, 09 Jan 2022	**Pseudo R-squ.:**	0.6299
Time:	15:40:13	**Log-Likelihood:**	-292.84
converged:	True	**LL-Null:**	-791.24
Covariance Type:	nonrobust	**LLR p-value:**	5.960e-211

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-7.1659	0.450	-15.916	0.000	-8.048	-6.283
Total_Debt	0.6512	0.161	4.045	0.000	0.336	0.967
Book_Value_Adj_Unit_Curr	-5.9327	0.531	-11.162	0.000	-6.974	-4.891
Market_Capitalisation	-0.5382	0.224	-2.405	0.016	-0.977	-0.100
CEPS_annualised_Unit_Curr	-0.9474	0.263	-3.607	0.000	-1.462	-0.433
Curr_Ratio_Latest	-0.9633	0.164	-5.890	0.000	-1.284	-0.643
PBITM_perc_Latest	-0.5722	0.127	-4.494	0.000	-0.822	-0.323
Value_of_Output_to_Gross_Block	-0.4008	0.159	-2.517	0.012	-0.713	-0.089

Figure 5: Summary of Logistic Regression Model

P>|t| is one of the most important statistics in the summary. It uses the t statistic to produce the p-value, a measurement of how likely a coefficient is measured through our model by chance. As one can clearly observe, from 'P>|z|' column that for all the variables values are less than 0.05 which is our significance level. Hence, we can say that all the variables are statistically significant for prediction of "Defaults" which is our target variable.

Performance Metrics

	precision	recall	f1-score	support
0	0.96	0.99	0.98	2156
1	0.92	0.67	0.78	245
accuracy			0.96	2401
macro avg	0.94	0.83	0.88	2401
weighted avg	0.96	0.96	0.96	2401

	precision	recall	f1-score	support
0	0.96	0.99	0.97	1042
1	0.88	0.72	0.79	142
accuracy			0.95	1184
macro avg	0.92	0.85	0.88	1184
weighted avg	0.95	0.95	0.95	1184

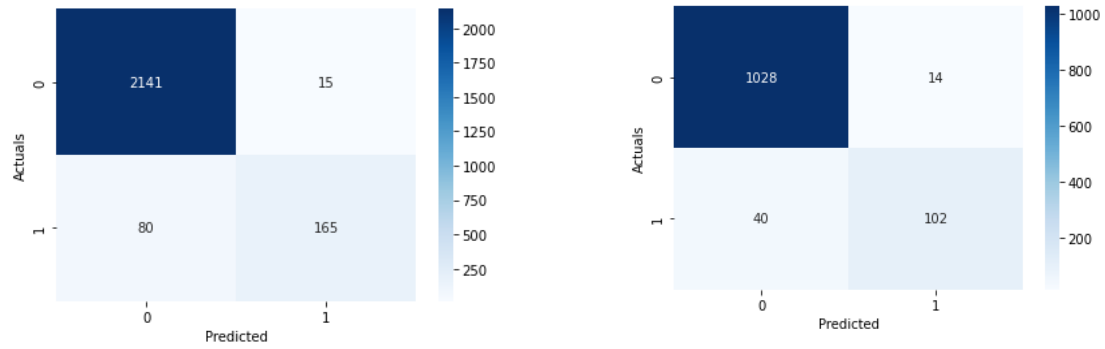


Figure 6: Performance metrics using Logit model for train set (left) and test set (Right)

Now, if we look at the performance metrics of models, these come out to be very similar. Though the accuracy of the model is good in both train and test set. But since we are the target variable is probability of 'Default', the recall value will matter the most. In our case, Recall is a more important metric as opposed to Precision given that we're more concerned with false negatives (our model predicting that someone is not going to default but they do) than false positives (our model predicting that someone is going to default but they don't).

Tuning of Logit Model

For both, train and test set performance are still not good considering the recall metric. This may be because we are working with highly imbalanced dataset as the number of non-defaulters (2156) is much higher than the people who are going to default (245).

```
[[   0 2156]
 [   1  245]]
```

Hence, for further improving performance we tried oversampling technique to take care of the class imbalance and again pick the best features using RFE in this case the feature which turned out to be statistically significant are:

```
'Equity_Paid_Up',
'Total_Debt',
'Book_Value_Adj_Unit_Curr',
'CEPS_annualised_Unit_Curr',
'Curr_Ratio_Latest',
'Debtors_Ratio_Latest',
'PBITM_perc_Latest'
```

Using these as independent variables, logit model was built again, summary of which is given below:

Logit Regression Results

Dep. Variable:	y_train	**No. Observations:**	2401
Model:	Logit	**Df Residuals:**	2393
Method:	MLE	**Df Model:**	7
Date:	Sun, 09 Jan 2022	**Pseudo R-squ.:**	0.6292
Time:	16:11:41	**Log-Likelihood:**	-293.43
converged:	True	**LL-Null:**	-791.24
Covariance Type:	nonrobust	**LLR p-value:**	1.068e-210

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-7.3451	0.462	-15.911	0.000	-8.250	-6.440
Equity_Paid_Up	-0.4405	0.159	-2.764	0.006	-0.753	-0.128
Total_Debt	0.8092	0.180	4.498	0.000	0.457	1.162
Book_Value_Adj_Unit_Curr	-6.3442	0.550	-11.530	0.000	-7.423	-5.266
CEPS_annualised_Unit_Curr	-1.0719	0.262	-4.091	0.000	-1.585	-0.558
Curr_Ratio_Latest	-1.0444	0.165	-6.334	0.000	-1.367	-0.721
Debtors_Ratio_Latest	-0.3091	0.132	-2.345	0.019	-0.567	-0.051
PBITM_perc_Latest	-0.6864	0.132	-5.216	0.000	-0.944	-0.428

Figure 7: Logistic Regression Model (built after class-balancing) Summary

Again, since p-value is less than 0.05 for all the independent variables, we can say that all the variables included are statistically significant.

Now, we can recheck if the performance the model has improved for recall metric or not.

	precision	recall	f1-score	support
0	0.77	0.99	0.87	2156
1	0.99	0.70	0.82	2156
accuracy			0.85	4312
macro avg	0.88	0.85	0.84	4312
weighted avg	0.88	0.85	0.84	4312

	precision	recall	f1-score	support
0	0.96	0.99	0.97	1042
1	0.88	0.70	0.78	142
accuracy			0.95	1184
macro avg	0.92	0.84	0.88	1184
weighted avg	0.95	0.95	0.95	1184

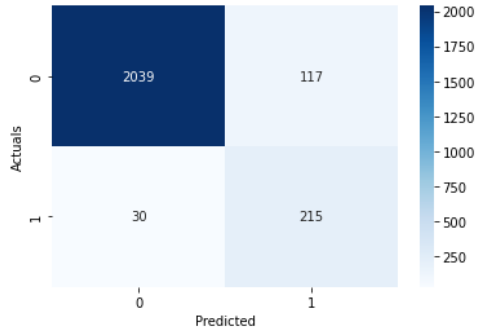
Figure 8: Classification Report After class balancing using SMOTE for train (above) and test (below) set

We can observe that the accuracy for training set has decreased but for test set it is coming out to be fine. Recall score which is coming to be 70 for both train and test set, has improved slightly if we compare with earlier ones.

Further, we tried changing the threshold for decision making. The roc curve can be used to find the optimum value of threshold. In this analysis, optimal index of threshold was found by finding maximum value of (True Positive Rate – False Positive Rate). This threshold came out to be 0.1953. Based on this threshold evaluated performance metrics are given below:

	precision	recall	f1-score	support
0	0.99	0.95	0.97	2156
1	0.65	0.88	0.75	245
accuracy			0.94	2401
macro avg	0.82	0.91	0.86	2401
weighted avg	0.95	0.94	0.94	2401

	precision	recall	f1-score	support
0	0.99	0.92	0.95	1042
1	0.60	0.92	0.73	142
accuracy			0.92	1184
macro avg	0.79	0.92	0.84	1184
weighted avg	0.94	0.92	0.92	1184



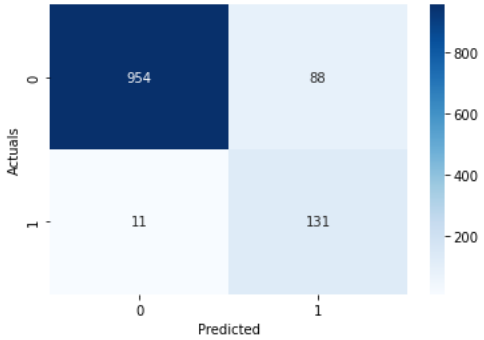


Figure 9: Performance metrics after using SMOTE and tuning for threshold value

Now, we can observe that the performance of the model has improved significantly considering the recall metrics. For train set the recall value is coming out to 88 percent and for test set it is coming out to be 92 percent. The accuracy value is 94 percent for train set and 92 percent for test set. Since, for the metrics the values are coming out to be pretty close, we can that the model is not overfitting.

Inferences

Now, we can look at the feature importance values calculated by the model:

	1
Intercept	-6.44
Equity_Paid_Up	-0.13
Total_Debt	1.16
Book_Value_Adj_Unit_Curr	-5.27
CEPS_annualised_Unit_Curr	-0.56
Curr_Ratio_Latest	-0.72
Debtors_Ratio_Latest	-0.05
PBITM_perc_Latest	-0.43

Figure 10: Coefficient for the variable present in the dataset for Logit model(left) and LDA (right)

Final model can in mathematical terms can be expressed as:

$$\boldsymbol{Logit(y)} = 1.16 * \boldsymbol{'Total\_Debt'} - 0.05 * \boldsymbol{'Debtors\_Ratio\_Latest'} - 0.13 * \boldsymbol{'Equit\_Paid\_Up'} - 0.72 * \boldsymbol{'Curr\_Ratio\_latest'} - 0.43 * \boldsymbol{'PBITM\_perc\_Latest'} - 5.27 * \boldsymbol{'Book\_Value\_Adj\_Unit\_Curr'} - 0.56 * \boldsymbol{'CEPS\_annualised\_Unit\_Curr'} - 6.44$$

Notice that the values for 'Debtors_Ratio_Latest' are almost zero which means that it is a weak predictor for the target variables. On the other hand, the 'Total_Debt' feature is more important for class 1 and 'Book_Value_Adj_Unit_Curr' is important for class 0 in this model.

Now, the following conclusion can be drawn from this analysis:

- By using only 7 independent variables (above mentioned), we are able to get >90% accuracy along with >90% recall value.
- Using SMOTE, was found to beneficial for this analysis.
- For this analysis, we have simply capped the outliers with the corresponding upper or lower whisker value. Other techniques can be also be used or a domain expert can also help in this regard.
- Although, VIF is general criteria for feature selection in regression analysis, but in this analysis RFE was able to perform satisfactorily.
- For further improving the performance other models such as Ensemble methods can be tried.