Executive Summary

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.

Introduction

The purpose of this project is to explore the data which contains information from the financial statements 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 labelled field.

Data description

Data consists of 3586 rows and 67 columns.

Info of dataset(After removing special characters from variables name)

Co_Code	3586 non-null	int64
Co_Name	3586 non-null	object
NetworthNextYear	3586 non-null	float64
EquityPaidUp	3586 non-null	float64
Networth	3586 non-null	float64
CapitalEmployed	3586 non-null	float64
TotalDebt	3586 non-null	float64
GrossBlock	3586 non-null	float64
NetWorkingCapital	3586 non-null	float64
CurrentAssets	3586 non-null	float64
CurrentLiabilitiesandProvisions	3586 non-null	float64
TotalAssetsLiabilities	3586 non-null	float64
GrossSales	3586 non-null	float64
NetSales	3586 non-null	float64
OtherIncome	3586 non-null	float64
ValueOfOutput	3586 non-null	float64
CostofProduction	3586 non-null	float64
SellingCost	3586 non-null	float64
PBIDT	3586 non-null	float64
PBDT	3586 non-null	float64

PBIT	3586 non-null	float64
PBT	3586 non-null	float64
PAT	3586 non-null	float64
AdjustedPAT	3586 non-null	float64
CP	3586 non-null	float64
Revenueearningsinforex	3586 non-null	float64
Revenueexpensesinforex	3586 non-null	float64
Capitalexpensesinforex	3586 non-null	float64
BookValueUnitCurr	3586 non-null	float64
BookValueAdjUnitCurr	3582 non-null	float64
MarketCapitalisation	3586 non-null	float64
CEPSannualisedUnitCurr	3586 non-null	float64
CashFlowFromOperatingActivities	3586 non-null	float64
CashFlowFromInvestingActivities	3586 non-null	float64
CashFlowFromFinancingActivities	3586 non-null	float64
ROGNetWorth	3586 non-null	float64
ROGCapitalEmployed	3586 non-null	float64
ROGGrossBlock	3586 non-null	float64
ROGGrossSales	3586 non-null	float64
ROGNetSales	3586 non-null	float64
ROGCostofProduction	3586 non-null	float64
ROGTotalAssets	3586 non-null	float64
ROGPBIDT	3586 non-null	float64
ROGPBDT	3586 non-null	float64
ROGPBIT	3586 non-null	float64
ROGPBT	3586 non-null	float64
ROGPAT	3586 non-null	float64
ROGCP	3586 non-null	float64
ROGRevenueearningsinforex	3586 non-null	float64
ROGRevenueexpensesinforex	3586 non-null	float64
ROGMarketCapitalisation	3586 non-null	float64
CurrentRatioLatest	3585 non-null	float64
FixedAssetsRatioLatest	3585 non-null	float64
InventoryRatioLatest	3585 non-null	float64
DebtorsRatioLatest	3585 non-null	float64
TotalAssetTurnoverRatioLatest	3585 non-null	float64
InterestCoverRatioLatest	3585 non-null	float64
PBIDTMLatest	3585 non-null	float64
PBITMLatest	3585 non-null	float64
PBDTMLatest	3585 non-null	float64
CPMLatest	3585 non-null	float64
APATMLatest	3585 non-null	float64
DebtorsVelocityDays	3586 non-null	int64
CreditorsVelocityDays	3586 non-null	int64
InventoryVelocityDays	3483 non-null	float64
ValueofOutputTotalAssets	3586 non-null	float64
ValueofOutputGrossBlock	3586 non-null	float64

-Variables are float, integer and object types
- There are null values in some columns.

Question 1- Outlier treatment

Mostly all columns have outliers. Outliers are treated by capping extreme values within upper range and lower range of interquartile range.

Question 2- Missing value treatment

As missing values of some variables are less than 1%, so dropping all missing values

Question 3 - Transform target value into 0 and 1

A new column is constructed by name of default which contains value 1 when net worth next year is negative and value 0 when net worth next year is positive.

Question 4 - Univariate and bivariate analysis

- Univariate analysis
  - IT is done by using a distribution plot.
  -Whole data is not normally distributed
  - Many variables like net with next year,Equity paid up, net worth, capital employed, total dept , gross block have bimodal distribution
  - ROG Net worth,ROG Capital employed,ROG Gross block, ROG gross sales are normally distributed and have 3 modal distribution
  -Fixed Assets ratio, Inventory ratio,Debtors Ratio, Total Asset Turnover ratio are positively skewed
  -PBIDTM, PBITM, PBDTM,CPM are negatively skewed

- Bivariate analysis
- IT is done by heat map to find correlation between different variables.
- Net worth and net worth next year are showing positive correlation with each other.
- Capital employed and total asset liabilities are showing positive correlation with each other.
- Value of output and Gross sales are showing positive correlation with each other.
- Value of output and Net sales are showing positive correlation with each other.

Question 5 - Train test split

- First data is split into x containing independent variables and y containing dependent variables.
- Train test split function is imported from sklearn.
- Data is split into train and test sets in 67:33 ratio.
- X and y are later concatenated back because the stats model does not take data separately.

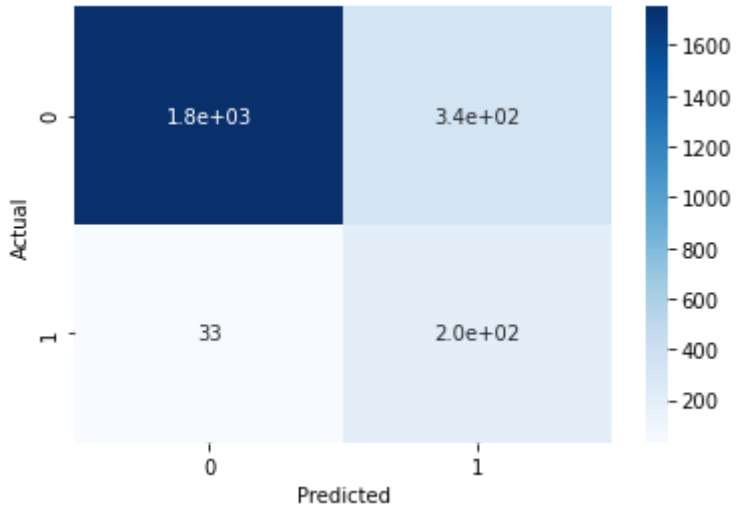
Question 6 - Building Logistic regression model

- Logistic regression model is made by using stats model
- Step1 - Correlation between many variables are treated variance inflation factor(VIF)
- - All variables with VIF more than 5 are removed.

- Step2 - All variables with P>0.05 in ols summary because these variables are not playing an important role in making predictions.

- Step 3 - Finding optimal threshold by using ROC curve.
  -Optimal threshold = 0.19

- Step 4 - Model is validated on train set and then confusion matrix and classification report are created
  Confusion matrix of train set-



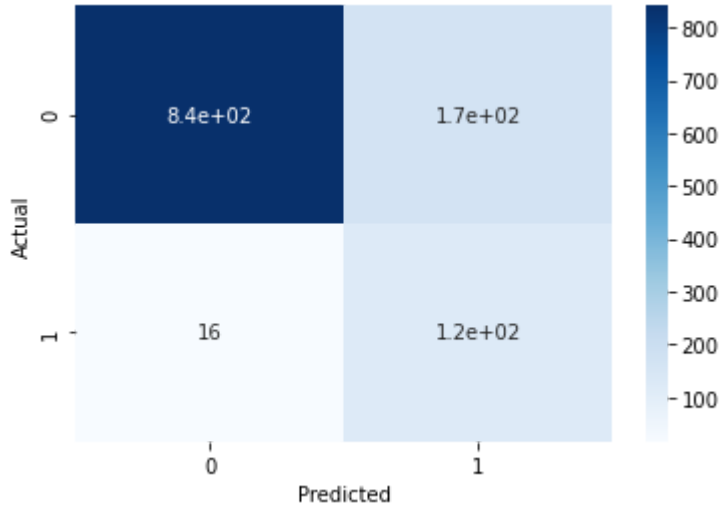
Classification report of train set-

	precision	recall	f1-score	support
0	0.981	0.837	0.903	2092
1	0.375	0.861	0.522	238

Here recall is important because we have to correctly find out true positives.
Recall= 86%
Precision= 37%

Question 7 - Validate the model on test set and state the performance matrix

Confusion matrix of test set-



Classification report of test set-

	precision	recall	f1-score	support
0	0.981	0.833	0.901	1009
1	0.423	0.885	0.572	139

Recall = 88%
Precision = 42%