Companies Bankruptcy in Poland

BY
Sheetal Golecha
Shraddha Avasthy

Executive Summary:

After 2008 Financial crisis the whole world is on alert and getting prepared for any future financial crsis that may occur. Hence the study to identify what are the factors that will lead to the finanacial crsis are very important. Currently we have a dataset of a set of 5 polish companies which contain ratios of the financial companies. This data was collected for the year 2007 to 2013.

But before we do a headstart to analysis this data, lets understand why do the companies face bankruptcy.Reasons that a company face bankruptcy

- Insufficient Demand: if the demand for the goods or services reduces which could be due to various reason then the loss incurred by the company increases.
- Competition: If the product you are selling is also sold by the other companies then the competition between the firm increases. For example: its senseless to open a gym in an area where there are 10 other gyms within a kilometer
- Failure to control cost: even if the company is able to generate a huge amount of revenue, it might still not have huge profit if it is having huge cost incurred.
- Market Decline: Shift in market preference will also lead to decline of the purchase of your product.

The above are the few causes which leads to closure of the small business but most of it will not be applicable for a huge firm, Because even if a multi million dollar firm invests in a new venture, the new venture even if its in loss won't make a huge difference to firm, as it is its side business. And if he or she still wants it, they can prefer to continue the firm.
Few of the above points are taken "https://smallbusiness.chron.com/causes-business-bankruptcy-49407.html"

Business Problem:

From the ratios of the dataset given to us , identify the factors that can cause bankruptcy . create a model which will help us to predict the bank rupcy of a company.

Introduction:

What is Bankruptcy: Bankruptcy is a legal term for when a person or business cannot repay their outstanding debts. The bankruptcy process begins with a petition filed by the debtor, which is most common, or on behalf of creditors, which is less common. All of the debtor's assets are measured and evaluated, and the assets may be used to repay a portion of outstanding debt.

BREAKING DOWN Bankruptcy

Bankruptcy offers an individual or business a chance to start fresh by forgiving debts that simply cannot be paid, while offering creditors a chance to obtain some measure of repayment based on the individual's or business's assets available for liquidation. In theory, the ability to file for bankruptcy can benefit an overall economy by giving persons and businesses a second chance to gain access to consumer credit and by providing creditors with a measure of debt repayment. Upon the successful completion of bankruptcy proceedings, the debtor is relieved of the debt obligations incurred prior to filing for bankruptcy.

The above lines are taken from "https://www.investopedia.com/terms/b/bankruptcy.asp"

DataSet:

We have 5 different years dataset starting from year 2007 to 2013. Each year dataset has about 65 columns.

Year	Rows	Bankrupcy	Running
1	7027	271	6756
2	10173	400	9773
3	10503	495	10008
4	9792	515	9277
5	5910	410	5500

Table 1: details of data

About 206 cells in the dataset have ? or basically NA data in them, I converted all the question mark to NA and ran "centralImputation" to remove the NA data and manually imputed the data.

column Name	Frequency
V37	18984
V48	9501
V7	9368
V14	9368
V18	9368
V1	9364
V35	9319
V3	9281
V57	9254
V11	9228
V25	9179
V2	8879
V10	8878
V51	8681
V38	8618
V22	8366

V36	7239
V29	5951
V21	5854
V6	5727
V59	5416

Above are the few column where the NA cells are more than 5000K lines.

df=centralImputation(df)

Column Names:
net profit _ total assets , total liabilities _ total assets , working capital _ total assets , current assets _ short-term liabilities , cash _ short-term securities _ receivables - short-term liabilities _ operating expenses - depreciation * 365, retained earnings _ total assets , EBIT _ total assets , book value of equity _ total liabilities , sales _ total assets , equity _ total assets , gross profit _ extraordinary items _ financial expenses _ total assets , gross profit _ short-term liabilities , gross profit _ depreciation _ sales , gross profit _ interest _ total assets , total liabilities * 365 _ gross profit _ depreciation, gross profit _ depreciation _ total liabilities , total assets _ total liabilities , gross profit _ total assets , gross profit _ sales , inventory * 365 _ sales, sales n _ sales n-1, profit on operating activities _ total assets , net profit _ sales , gross profit in 3 years _ total assets, equity - share capital _ total assets , net profit _ depreciation _ total liabilities , profit on operating activities _ financial expenses , working capital _ fixed assets , logarithm of total assets , total liabilities - cash _ sales , gross profit _ interest _ sales , current liabilities * 365 _ cost of products sold, operating expenses _ short-term liabilities , operating expenses _ total liabilities , profit on sales _ total assets , total sales _ total assets , current assets - inventories _ long-term liabilities , constant capital _ total assets , profit on sales _ sales , current assets - inventory - receivables _ short-term liabilities , total liabilities _ profit on operating activities _ depreciation * 12_365, profit on operating activities _ sales , rotation receivables _ inventory turnover in days , receivables * 365 _ sales, net profit _ inventory , current assets - inventory _ short-term liabilities , inventory * 365 _ cost of products sold, EBITDA profit on operating activities - depreciation _ total assets , EBITDA profit on operating activities - depreciation _ sales , current assets _ total liabilities , short-term liabilities _ total assets , short-term liabilities * 365 _ cost of products sold, equity _ fixed assets , constant capital _ fixed assets , working capital , sales - cost of products sold _ sales , current assets - inventory - short-term liabilities _ sales - gross profit - depreciation , total costs _total sales , long-term liabilities _ equity , sales _ inventory , sales _ receivables , short-term liabilities *365 _ sales, sales _ short-term liabilities , sales _ fixed assets, class

As we can see most of the column have same numerator there will definitely be some collinearity in the dataset, lets have a look.

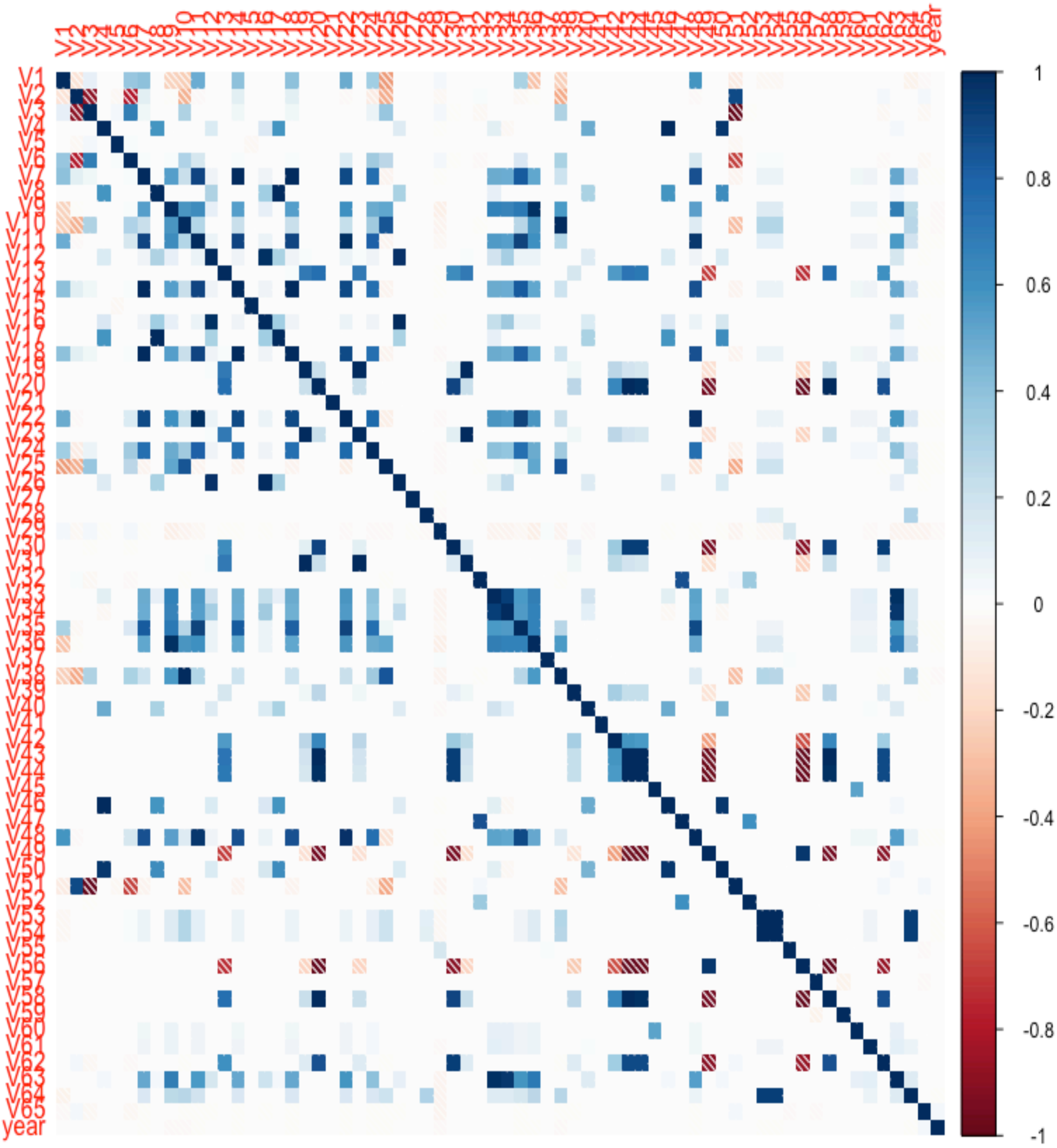
Correlation :

Lets have look at correlation of the dataset with column to be predicted just to idenfy what is happening:

```
V1          V2           V3            V4            V5           V6           V7
V8
[1,] -0.02562381 0.03111519 -0.03129992 -0.001648409 -0.001329903 -0.03075562 -
0.01525365 -0.002684099
               V9          V10         V11          V12           V13          V14
V15          V16
[1,] -0.002850635 -0.01181572 -0.0153483 -0.01608436 -0.001408471 -0.01525398
0.005469821 -0.01380733
             V17          V18           V19           V20           V21          V22
V23
[1,] -0.00271774 -0.01535474 -0.001220889 -0.001024785 -0.002367038 -0.01403495 -
0.001163174
             V24          V25          V26           V27           V28        V29
V30
[1,] -0.01524488 -0.01429602 -0.01355566 -0.006746806 -0.004269772 -0.050733
0.0002277069
               V31         V32          V33           V34          V35           V36
V37
[1,] -0.0009653139 0.01156119 0.002640294 0.0004174358 -0.01679755 -0.0005166604 -
0.002445497
             V38          V39          V40          V41         V42           V43
V44
[1,] -0.01173615 -0.01979423 0.000312245 -0.00123667 0.00169188 -0.001302782 -
0.001393211
               V45           V46           V47          V48           V49           V50
V51
[1,] -0.0009613342 -0.001887197 -0.001859912 -0.01520351 0.0004377862 -0.001896495
0.03105972
              V52          V53          V54         V55           V56         V57
V58          V59
[1,] -0.001787644 0.004604274 0.004626553 -0.0221001 0.0009494614 -0.0218023 -
0.001219028 -0.001326727
              V60           V61            V62            V63          V64 V65
year
[1,] -0.002150847 -0.0007005754 -0.0001183522 -0.0008509868 0.003145637   1
0.04348803
```

There is no such huge correlation in the dataset with the class.

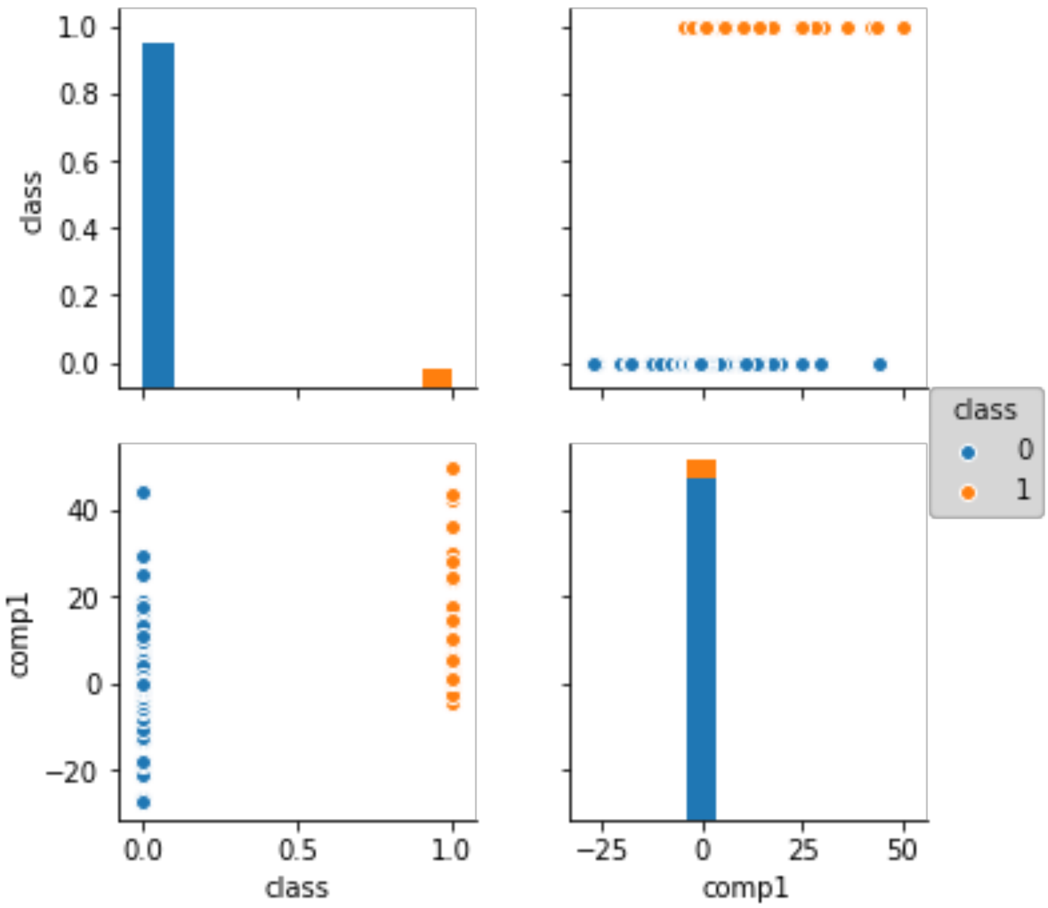
Lets have a look at multi collinearity



As we can seet there are lot of columns where we can see correlation, I created a subset of the dataset and removed all column where there could be potential correlation.
Based on the above correlation I have created the subset of the data and taken only columns which are helpful in predicting and has no correlation with other clumns.

Df$V65=as.factor(df$V65)
Subset_df<-
df[c("V1","V2","V4","V5","V7","V8","V9","V10","V12","V13","V15","V17","V21","V27","V28","V29","V30","V32","V33","V37","V39","V40","V41","V42","V47","V52","V53","V54","V55","V57","V59","V60","V61","year","V65")]

When I applied the LDA on the dataset, could generate the one component. When we looked at the graph on it, its difficult to find a clear demarcation to differentiate it.



Lets look at the proportion of the data overall. Between bankruptcy and currently running company.

prop.table(table(df$V65))
0 1
0.95182583 0.04817417

Model:

The model I have worked with are below:

1. Logistic Regression:
2. Poison Distribution
3. Negative Binomial Distribution
4. Zero Inflation
5. Hurdle
6. Naïve Bayer's Algorithm

7. LDA
8. Decision tree
9. KNN Algorithm
10. SVM
11. Random Forest
12. Ensemble Algorithm

For all the algorithm I have used the same formula

"V65 ~ V1+V2+V4+V5+V7+V8+V9+V10+V12+V13+V15+V17+V21+V27+V28+V29+V30+V32+V33+V37+V39+V40+V41 +V42+V47+V52+V53+V54+V55+V57+V59+V60+V61"

Below is the Accuracy for each run

Logistic Regression:

The first run was basic logistic regression, the test and results are given below.

Call:
glm(formula = V65 ~ ., data = df)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-0.74224	-0.05559	-0.04853	-0.03666	1.02217

Training DataSet

	0	1
0	30981	1547
1	9	13

Accuracy : 0.9522
95% CI : (0.9498, 0.9545)

Sensitivity : 0.999710
Specificity : 0.008333

Test DataSet Accuracy:

Confusion Matrix and Statistics

	0	1
0	10317	527
1	7	4

Accuracy : 0.9508
95% CI : (0.9466, 0.9548)

Sensitivity : 0.999322
Specificity : 0.007533

When we look at the accuracy of the above , it clearly shows that due to data imbalance, there are lot of false positive.

Poison distribution:

The next we applied poison distribution, to see it will take care of excessive 0.

It was giving the same problem.

Train Dataset Accuracy:
Confusion Matrix and Statistics

	0	1
0	30988	1553
1	2	7

Accuracy : 0.9522
95% CI : (0.9499, 0.9545)
Sensitivity : 0.999935
Specificity : 0.004487

Test Dataset Accuracy:
Confusion Matrix and Statistics

	0	1
0	10321	530
1	3	1

Accuracy : 0.9509
95% CI : (0.9467, 0.9549)
Sensitivity : 0.999709
Specificity : 0.001883

The False positive improved by three values form logistic regression but is still. Not good enough.

Negative Binomial:

When the 0's are a lot we must use the Negative Binomial:

Train dataset Accuracy:

Confusion Matrix and Statistics

	0	1
0	30988	1553
1	2	7

Accuracy : 0.9522
95% CI : (0.9499, 0.9545)
Sensitivity : 0.999935
Specificity : 0.004487

Test dataset Accuracy:

Confusion Matrix and Statistics

	0	1
0	10321	530
1	3	1

Accuracy : 0.9509
95% CI : (0.9467, 0.9549)
Sensitivity : 0.999709
Specificity : 0.001883

The result is same as the poisson no difference at all.

Zero Inflation:

To handle the 0's lets try the statistics model "zero hurdle"

Train Dataset Accuracy:

Confusion Matrix and Statistics

	0	1
0	30940	1552
1	50	8

Accuracy : 0.9508
95% CI : (0.9484, 0.9531)
Sensitivity : 0.998387
Specificity : 0.005128

Test dataset Accuracy:
Confusion Matrix and Statistics

	0	1
0	10299	530
1	25	1

Accuracy : 0.9489
95% CI : (0.9446, 0.9529)
Sensitivity : 0.997578
Specificity : 0.001883

The accuracy reduces in the zero inflation model when compared to previous negative binomial model.

Hurdle:

Train Dataset Accuracy:
Confusion Matrix and Statistics

	0	1
0	30897	1545
1	81	14

Accuracy : 0.95
95% CI : (0.9476, 0.9524)
Sensitivity : 0.99739
Specificity : 0.00898

Test Dataset Accuracy:
Confusion Matrix and Statistics

	0	1
0	10283	528
1	36	2

Accuracy : 0.948
95% CI : (0.9437, 0.9521)
Sensitivity : 0.996511
Specificity : 0.003774

This gives way better result when compared to rest.

Naïve Bayer's Model:

This is the last statical model, we are trying the accuracy has decreased a huge amount. Below is the confusion matrix for

Train Dataset Accuracy:
Confusion Matrix and Statistics

	0	1
0	744	42
1	30246	1518

Accuracy : 0.0695
95% CI : (0.0668, 0.0723)
Sensitivity : 0.02401
Specificity : 0.97308

Test Dataset accuracy:

Confusion Matrix and Statistics

	0	1
0	235	19
1	10089	512

Accuracy : 0.0688
95% CI : (0.0641, 0.0737)
Sensitivity : 0.02276
Specificity : 0.96422

LDA:

Lets look at the accuracy of training dataset.

Confusion Matrix and Statistics

	0	1
0	30982	1552
1	8	8

Accuracy : 0.9521
95% CI : (0.9497, 0.9544)
Sensitivity : 0.999742
Specificity : 0.005128

Test Daataset Accuracy:
Confusion Matrix and Statistics

	0	1
0	10320	528
1	4	3

Accuracy : 0.951
95% CI : (0.9468, 0.955)
Sensitivity : 0.99961
Specificity : 0.00565

Decision tree:

Lets look at the accuracy of training dataset.

Train Dataset Accuracy:

Confusion Matrix and Statistics

	0	1
0	30950	1125
1	40	435

Accuracy : 0.9642
95% CI : (0.9621, 0.9662)
Sensitivity : 0.9987
Specificity : 0.2788

Test Dataset Accuracy:

Confusion Matrix and Statistics

	0	1
0	10304	379
1	20	152

Accuracy : 0.9632
95% CI : (0.9595, 0.9667
Sensitivity : 0.9981
Specificity : 0.2863

The accuracy is much better when compared to all the other model, but true negatives is higher in this model.

KNN:

We have run KNN for K values for all numbers between 1 to 21 with only odd values all give same result.
Below is th result from run.

K	true_positive	true_negative	false_positive	false_negative
1	9848	476	493	38
2	9877	447	482	49
3	10191	133	520	11
4	10196	128	520	11
5	10280	44	526	5
6	10278	46	526	5
7	10304	20	530	1
8	10301	23	530	1
9	10313	11	531	0
10	10309	15	531	0
11	10315	9	531	0
12	10316	8	531	0
13	10318	6	531	0
14	10322	2	531	0
15	10322	2	531	0
16	10321	3	531	0
17	10324	0	531	0
18	10324	0	531	0
19	10324	0	531	0
20	10324	0	531	0
21	10324	0	531	

SVM:

The svm confusion matrix

Train Dataset Accuracy

Confusion Matrix and Statistics

Confusion Matrix and Statistics

	0	1
0	29535	1497
1	1455	63

Accuracy : 0.9093
95% CI : (0.9061, 0.9124)

Sensitivity : 0.95305
Specificity : 0.04038

Test Dataset Accuraacy:

Confusion Matrix and Statistics

	0	1
0	9852	506
1	472	25

Accuracy : 0.9099
95% CI : (0.9044, 0.9152)
Sensitivity : 0.95428
Specificity : 0.04708

Random Forest:

Train Dataset Accuraacy
Confusion Matrix and Statistics

	0	1
0	30990	2
1	0	1558

Accuracy : 0.9999
95% CI : (0.9998, 1)
Sensitivity : 1.0000
Specificity : 0.9987

Test Dataset Accuracy:
Confusion Matrix and Statistics

	0	1
0	10283	446
1	41	85

Accuracy : 0.9551
95% CI : (0.9511, 0.959)

Sensitivity : 0.9960
Specificity : 0.1601
The accuracy has dropped down tremendously in the test dataset, so we cant choose it.

Ensemble:

Created various ensemble model with combination of poisson,negative binomial,lda,decision tree, random forest and svm. The accuracy dint increase.

Ensemble Model 1(RandomForest, LDA and Poisson):

Train Dataset Accurcay:

Confusion Matrix and Statistics

	0	1
0	30988	1552
1	2	8

Accuracy : 0.9523
95% CI : (0.9499, 0.9545)
Sensitivity : 0.999935
Specificity : 0.005128

Test Dataset Accuracy:
Confusion Matrix and Statistics

	0	1
0	10322	530
1	2	1

Accuracy : 0.951
95% CI : (0.9468, 0.955)
Sensitivity : 0.999806
Specificity : 0.001883

Ensmeble Model 2(RandomForest, LDA,decision tree and Poisson)
Train Dataset Accurcay:

Confusion Matrix and Statistics

	0	1
0	30988	1530
1	2	30

Accuracy : 0.9529
95% CI : (0.9506, 0.9552)
Sensitivity : 0.99994
Specificity : 0.01923

TEST DataSet Accuracy:
Confusion Matrix and Statistics

	0	1
0	10323	530
1	1	1

Accuracy : 0.9511
95% CI : (0.9469, 0.9551)
Sensitivity : 0.999903
Specificity : 0.001883

ROSE:

Finally when none of the normal model worked I tried to implented ROSE. Creates a sample of synthetic data by enlarging the features space of minority and majority class examples. Operationally, the new examples are drawn from a conditional kernel density estimate of the two classes,

When we applied the ROSE algorithm, we used only 0.05 probablity on the bankruptcy class, we dint change the proportion and ran the algorithm.

Logit on ROSE Model:

Train Dataset Accuracy:

Confusion Matrix and Statistics

	0	1
0	30944	1596
1	0	10

Accuracy : 0.951
95% CI : (0.9486, 0.9533)

Sensitivity : 1.000000
Specificity : 0.006227

Test DataSet Accuracy:

Confusion Matrix and Statistics

	0	1
0	10329	524
1	0	2

Accuracy : 0.9517
95% CI : (0.9475, 0.9557)
Sensitivity : 1.000000
Specificity : 0.003802

Naïve Bayer's on ROSE Dataset:

Train Dataset Accuracy:
Confusion Matrix and Statistics

	0	1
0	30927	34
1	17	1572

Accuracy : 0.9984
95% CI : (0.9979, 0.9988)
Sensitivity : 0.9995
Specificity : 0.9788

Test Dataset Accuracy:
Confusion Matrix and Statistics

	0	1
0	10328	10
1	1	516

Accuracy : 0.999
95% CI : (0.9982, 0.9995)

Sensitivity : 0.9999
Specificity : 0.9810

LDA:

Train Dataset Accuracy:
Confusion Matrix and Statistics

	0	1
0	30938	1594
1	6	12

Accuracy : 0.9508
95% CI : (0.9484, 0.9532)
Sensitivity : 0.999806
Specificity : 0.007472

Test Dataset Accuracy:
Confusion Matrix and Statistics

	0	1
0	10326	524
1	3	2

Accuracy : 0.9515
95% CI : (0.9472, 0.9554)
Sensitivity : 0.999710
Specificity : 0.003802

Decsion Tree:

Train Dataset Accuracy:

Confusion Matrix and Statistics

	0	1
0	30935	36
1	9	1570

Accuracy : 0.9986
95% CI : (0.9982, 0.999)
Sensitivity : 0.9997
Specificity : 0.9776

Test Dataset Accuracy:

Confusion Matrix and Statistics

	0	1
0	10313	20
1	16	506

Accuracy : 0.9967
95% CI : (0.9954, 0.9977).
Sensitivity : 0.9985
Specificity : 0.9620

Random Forest:

Train data set accuracy:
Confusion Matrix and Statistics

	0	1
0	30944	0
1	0	1606

Accuracy : 1
95% CI : (0.9999, 1)

Sensitivity : 1.0000
Specificity : 1.0000

Test Dataset Accuracy:

Confusion Matrix and Statistics

	0	1
0	10329	2
1	0	524

Accuracy : 0.9998
95% CI : (0.9993, 1)

Sensitivity : 1.0000
Specificity : 0.9962

Conclusion:

Columna name	Dummy Name
net profit _ total assets	V1
total liabilities _ total assets	V2
current assets _ short-term liabilities	V4
cash _ short-term securities _ receivables - short-term liabilities _ operating expenses - depreciation * 365	V5
EBIT _ total assets	V7
book value of equity _ total liabilities	V8
sales _ total assets	V9
equity _ total assets	V10
gross profit _ short-term liabilities	V12
gross profit _ depreciation _ sales	V13
total liabilities * 365 _ gross profit _ depreciation	V15
total assets _ total liabilities	V17
sales n _ sales n-1	V21
profit on operating activities _ financial expenses	V27
working capital _ fixed assets	V28
logarithm of total assets	V29
total liabilities - cash _ sales	V30
current liabilities * 365 _ cost of products sold	V32
operating expenses _ short-term liabilities	V33
current assets - inventories _ long-term liabilities	V37
profit on sales _ sales	V39
current assets - inventory - receivables _ short-term liabilities	V40
total liabilities _ profit on operating activities _ depreciation * 12_365	V41
profit on operating activities _ sales	V42
inventory * 365 _ cost of products sold	V47
short-term liabilities * 365 _ cost of products sold	V52
equity _ fixed assets	V53

constant capital _ fixed assets	V54
working capital	V55
current assets - inventory - short-term liabilities _ sales - gross profit - depreciation	V57
long-term liabilities _ equity	V59
sales _ inventory	V60
sales _ receivables	V61

In Conclusion above are the columns when applied kernel density function will help you to identify the Bankruptcy status of the company. The Model which are created with random forest, Decision tree and Naïve Bayer's . We will personally choose Decision tree because the model generated is visible with the below plot, so we know how a particular company is selected for bankruptcy.