BANKRUPTCY PREDICTION USING INTELLIGENT COMPUTATION TECHNIQUES

The financial crisis has brought about an increased interest in an area of research which responds to the need to predict bankruptcy. This project tried to identify non-traditional predictors of bankruptcy using intelligent mining techniques.

Authors:

Athulya Menon J
Dhananjay Neelakantan
Kiran Prasannadas
Shalini Sasidharan

Under the guidance of:

Dr. Yoganand Balagurunathan

Muma College of Business

University of South Florida

Contents

1. Introduction

Bankruptcy is a term given to an organization that does have the capability of repaying debt to creditors. It is often imposed by a legal court order which is initiated by the debtor themselves. In such cases, the bankrupt firm may liquidize their assets to cover the debt. There have been several studies in the past that have developed models for predicting bankruptcy. These models have used traditional industrial baseline predictors for the model.

2. Problem Statement & Objective

Since the 2008 recession, study of bankruptcy has gained momentum. Companies are more prudent about analyzing past data, trying to understand whether they are heading down the bankruptcy path. Often, the traditional parameters can fail to properly signal whether the organization is in financial distress.

The objective of this project is to identify non-traditional variables that can help predict bankruptcy of an organization. This is due to the fact that there have been instances where major organizations have had strong financial statements and yet declared bankruptcy. This analysis will help identify the next best set of factors that can analyze the financial standing of an organization.

3. Methodologies Used

3.1 Random Forest

A Random Forest Classifier is an ensemble algorithm, meaning that several individual classifiers are used to produce more accurate classification of the data than an individual model. This is an important technique mainly because it can be used for both classification and regression purposes. In addition, the RF algorithm has wide applicability, high predictive performance and low computational cost.

3.2 Decision Tree

Decision Tree is an algorithm with a flow-chart structure commonly used in Operations Research for decision analysis. The structure is made of nodes, branches, and leaves which are used for classifying data. It is an algorithm that makes use of conditional control statements.

4. Dataset

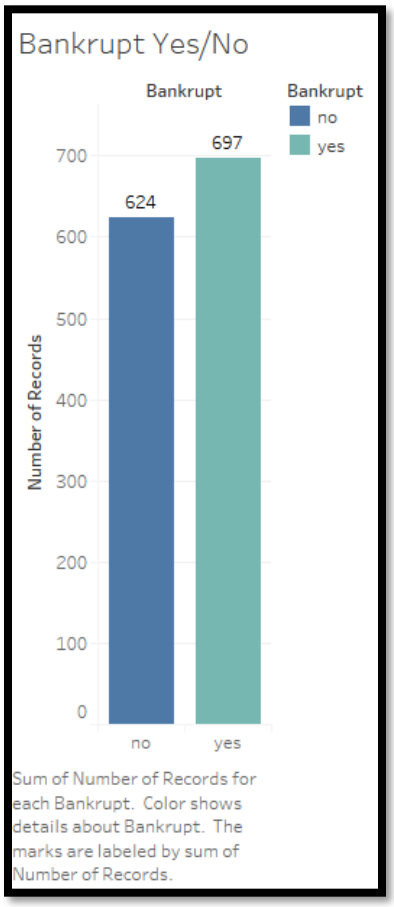
The dataset financial data of 100 companies that went bankrupt and that of 100 companies that did not go bankrupt. This time-period of the data is between 2001 and 2010. The original source of the dataset is the Compustat database. The dataset consists of 1321 observations and 17 variables. The different variables in the dataset are:

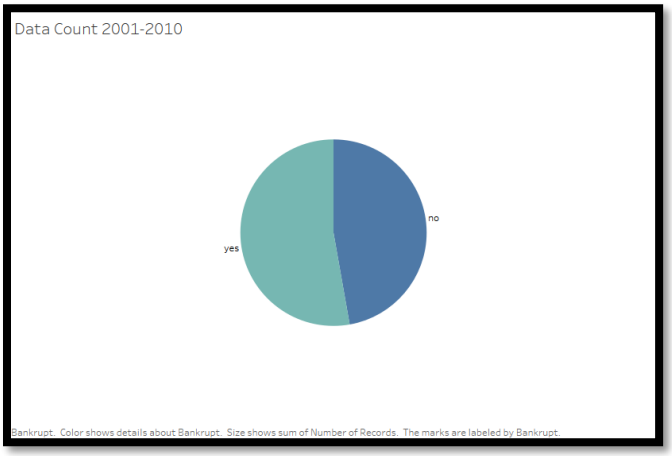
Sl No.	Variable name	Description
1	fyear	Data year — fiscal
2	at	Assets — total
3	bkvlps	Book value per share
4	invt	Inventories — total
5	Lt	Liabilities — total
6	rectr	Receivables — trade
7	cogs	Cost of goods sold

8	dvt	Dividends — total
9	ebit	Earnings before interest and taxes
10	gp	Gross profit (loss)
11	ni	Net income (loss)
12	oiadp	Operating income after depreciation
13	revt	Revenue — total
14	dvpsx_f	Dividends per share – ex-date – fiscal
15	mkvalt	Market value – total – fiscal
16	prch_f	Price high – annual – fiscal
17	bankruptcy	Bankruptcy (output variable)

5. Exploratory Data Analysis

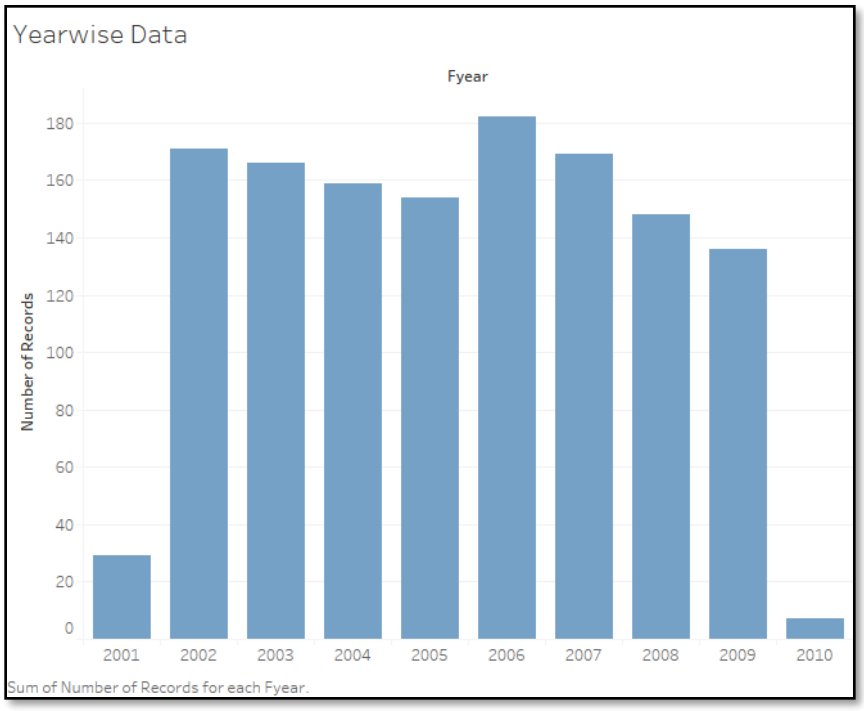
5.1 Count of observations belonging to Bankrupt vs Not Bankrupt





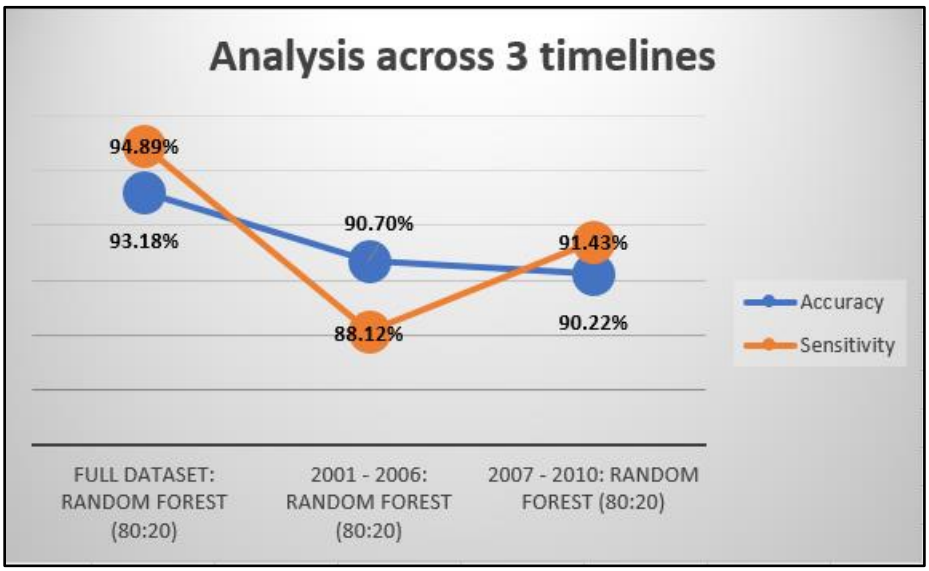
Over the period of study (2001 -2010) the data collected has almost equal number of data points belonging to either class (Bankruptcy = “Yes” & Bankruptcy = “No”). This is important as the count of data belonging to one class should not adversely impact the classification model.

5.2 Distribution of Data across the years



The dataset has fewer observations in the years 2001 and 2010. These were acceptable as this helped in keeping the total number of companies that went bankrupt and not bankrupt to be equal; also, the financial year factor didn't play a significant role in the model.

5.3 Analysis of Impact of Financial Crisis on the Data and Model



The data analysis was carried out using Weka, and variable selection using Excel and R. Of the models developed, Random Forest and J48 were the primary ones.

The data is split into two sets: from 2001 – 2006 (pre-recession) and 2007 – 2010 (during and post-recession). This analysis was carried out to understand whether the financial crisis would affect the dataset or not. On running the Random Forest model on the three datasets, we find that there is not much variation in the sensitivity values. In fact, using the entire dataset provided the best accuracy and sensitivity, showing that the recession does not really impact the financial standing of an organization.

6. Variable Selection

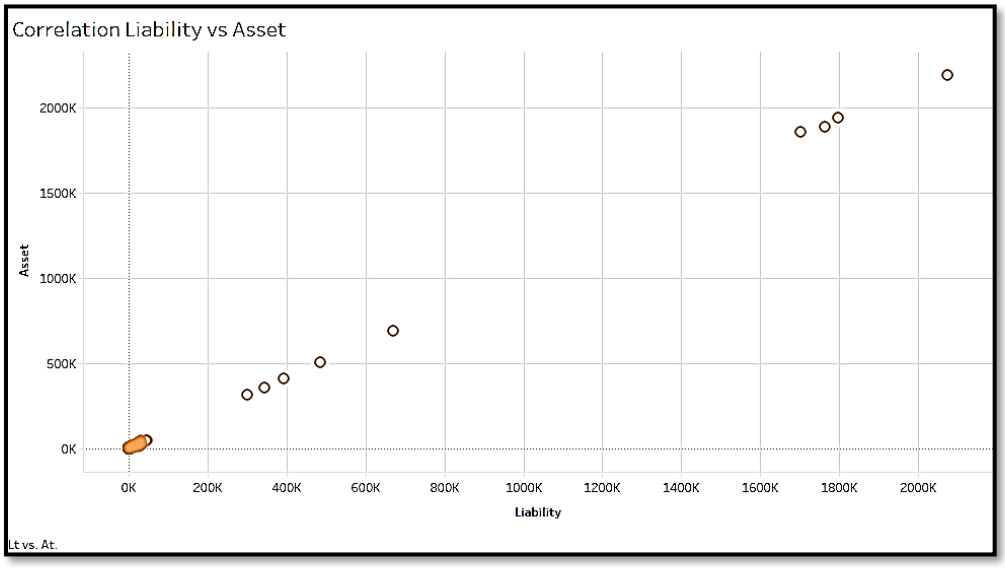
This dataset has 17 variables. The variables have been selected using

1. correlation, that is, highly correlated variables have been removed and the least correlated variables kept
2. step method, where the end result gave a set of variables

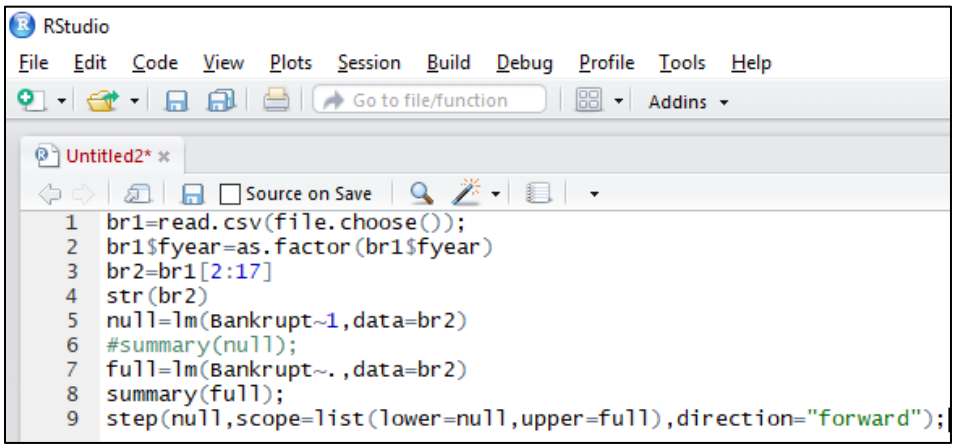
The final set of variables that have been selected are common between the two methods.

6.1 Correlation Matrix

	fyear	at	bkvlps	invt	Lt	rectr	cogs	dvt	ebit	gp	ni	oiadp	revt	dvpsx_f	mkvalt	prch_f
fyear	1															
at	0.05479	1														
bkvlps	0.053923	0.056983	1													
invt	0.039755	0.933406	0.085452	1												
Lt	0.054183	0.999777	0.051625	0.937348	1											
rectr	0.049311	0.990615	0.04663	0.90112	0.990351	1										
cogs	0.074239	0.888168	0.005571	0.760674	0.887824	0.87985	1									
dvt	0.054594	0.960466	0.041198	0.842228	0.959115	0.982119	0.874809	1								
ebit	0.027951	0.683987	0.098374	0.861585	0.69147	0.665697	0.489963	0.626042	1							
gp	0.044048	0.695333	0.090633	0.852821	0.702546	0.67239	0.56912	0.641262	0.979783	1						
ni	-0.01206	-0.00771	0.111078	0.127157	-0.00787	0.007543	-0.16076	0.032469	0.423582	0.384285	1					
oiadp	0.027951	0.683987	0.098374	0.861585	0.69147	0.665697	0.489963	0.626042	1	0.979783	0.423582	1				
revt	0.071271	0.916022	0.038217	0.881375	0.918485	0.90126	0.950903	0.88583	0.729732	0.795656	0.02621	0.729732	1			
dvpsx_f	-0.01357	0.01479	0.068296	0.015209	0.014701	0.015484	0.012465	0.034372	0.016263	0.015687	0.007737	0.016263	0.015087	1		
mkvalt	0.04692	0.520555	0.006778	0.391821	0.511528	0.476886	0.492341	0.485077	0.108231	0.12992	-0.02242	0.108231	0.411605	-0.00064	1	
prch_f	0.065961	0.116065	0.271553	0.166859	0.114243	0.091162	0.110492	0.093257	0.192128	0.220268	0.109825	0.192128	0.164305	0.2631	0.002289	1



6.2 R-code for Step Method




```
Call:
lm(formula = Bankrupt ~ prch_f + cogs + rectr + gp + ebit, data = br2)

Coefficients:
(Intercept)       prch_f         cogs        rectr           gp         ebit
  4.102e-01    3.006e-03    1.273e-05   -1.778e-06    2.201e-04   -2.129e-04
```

7. Data Modelling

7.1 Classification using Traditional Asset-Liability Ratio

To compare the performance of traditional predictors with the non-traditional ones, we developed a Random Forest Model (80:20 split) using Asset-to-Liability ratio as a predictor. The accuracy sensitivity thus obtained was 62.8% and accuracy was 59.85%.

```
Correctly Classified Instances         158               59.8485 %
Incorrectly Classified Instances       106               40.1515 %
Kappa statistic                          0.1949
Mean absolute error                      0.4045
Root mean squared error                  0.6332
Relative absolute error                 81.0918 %
Root relative squared error            126.6961 %
Total Number of Instances              264

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.628    0.433    0.610      0.628   0.619      0.195    0.595     0.577     yes
               0.567    0.372    0.585      0.567   0.576      0.195    0.595     0.537     no
Weighted Avg.  0.598    0.404    0.598      0.598   0.598      0.195    0.595     0.558

=== Confusion Matrix ===

  a  b   <-- classified as
 86 51 |  a = yes
 55 72 |  b = no
```

7.2 Classification Models

The model with five variables was run using several classifiers and the following data splits were taken:

- 10 Fold
- 80:20
- 70:30

To recap, the variables selected were:

- Receivables-trade
- Cost of goods sold
- Earnings before interest and taxes
- Gross Profit
- Price high

From the results obtained, the top models were selected and these constituted Random Forest and Decision Tree. These models have been primarily selected based on Sensitivity, Specificity and Accuracy.

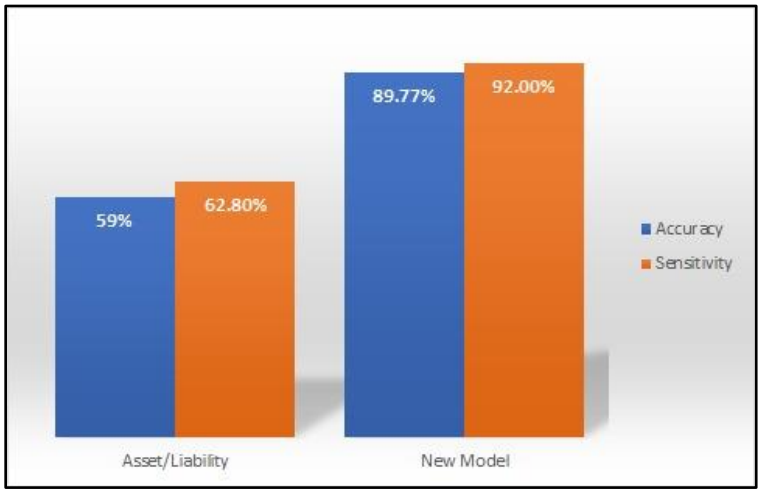
Sensitivity gives a measure of how good the model is in demonstrating that an organization is in financial distress, while specificity gives a measure of how good the model is in demonstrating that an organization is not in financial distress.

Models	True Negative	False Positive	False Negative	True Positive	Accuracy	Sensitivity	Specificity
Random Forest - 80:20	111	16	11	126	89.77%	91.97%	87.40%
Random Forest - 70:30	168	25	18	185	89.14%	91.13%	87.05%
Random Forest - 10 Fold	546	78	59	638	89.63%	91.54%	87.50%
J48 - 10 Fold	509	115	68	629	86.15%	90.24%	81.57%
J48 - 70:30	161	32	19	184	87.12%	**90.64%**	83.42%
J48 - 80:20	104	23	10	127	87.50%	**92.70%**	81.89%

```
=== Summary ===

Correctly Classified Instances         237               89.7727 %
Incorrectly Classified Instances        27               10.2273 %
Kappa statistic                          0.7949
Mean absolute error                      0.1023
Root mean squared error                  0.3198
Relative absolute error                 20.5007 %
Root relative squared error             63.9912 %
Total Number of Instances              264

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.920     0.126     0.887       0.920   0.903      0.795    0.897     0.858     yes
               0.874     0.080     0.910       0.874   0.892      0.795    0.897     0.856     no
Weighted Avg.  0.898     0.104     0.898       0.898   0.898      0.795    0.897     0.857

=== Confusion Matrix ===

   a   b   <-- classified as
 126  11 |   a = yes
  16 111 |   b = no
```

Random Forest 80:20 split gave an accuracy of 89.77%, sensitivity of 91.97% and specificity of 87.04%, while J48 with an 80:20 split gave an accuracy of 87.5%, sensitivity of 92.7% and specificity of 81.89%.

7.3 Comparison between model created by traditional asset to liability ratio and our new model



8. Inference

- The dataset with the given set of variables is not affected by the 2007 recession crisis, and hence we have used the entire dataset for further analysis
- The variables selected through correlation are the next best set of variables needed for a secondary analysis of the financial condition of an organization (after the common market used standards)
- Sensitivity is the most appropriate metric to gauge the performance of a model, as it gives a measure of how good the model is in demonstrating whether an organization is in financial distress or not
- We see that Random Forest with an 80:20 training-test separation is the best model after variable reduction, considering Sensitivity (91.97%) in combination with Specificity (87.4%) and Accuracy (89.77%)
- Although this is a 3% decrease from using all 17 variables, we can still use these secondary variables to reasonably analyze the financial situation of an organization

9. Conclusion

Traditional bankruptcy prediction methods almost always use standard industry baselines for the process. Our project aimed at identifying non-traditional prediction variables, with the intent to provide organizations a parallel method to analyze their financial standing. Based on the model parameters obtained, we have concluded that Random Forests are the best model for making these predictions in conjunction with the next best set of financial variables identified.

Task Allocation

Team Member	Tasks
Athulya MJ	Collect datasets, clean datasets, build model 1, visualize models, interim presentation
Dhananjay Neelakantan	Collect references, understand business, build model 2, evaluate models, final presentation
Kiran Prasannadas	Collect references, understand business, build model 3, evaluate models, interim presentation
Shalini Sasidharan	Collect datasets, clean datasets, study datasets, tune the model, final presentation

References

1. Comparative analysis of data mining methods for bankruptcy prediction - Article in Decision Support Systems · January 2012 – Prof David Olson, University of Nebraska, Lincoln
2. An improved boosting based on feature selection for corporate bankruptcy prediction - (Yang, Ma, & Wang, 2014)
3. Dynamics of Modeling in Data Mining: Interpretive Approach to Bankruptcy Prediction - (Sung, Chang, & Lee, 1999)
4. Ensemble Booster Trees with Synthetic Features Generation in Application to Bankruptcy Prediction - (Zieba, Tomczak, & Tomczak, 2016)