# Bankruptcy prediction modeling in real world conditions:

## A contrast between Boosting algorithm and Logistic regression

## CIDE LILLE – November 2019


humanOrg
INSTITUT DE RECHERCHE
EN DEVELOPPEMENT HUMAIN
ET DES ORGANISATIONS DE L'UMONS

UMONS
Université de Mons

# Plan

I. Introduction

II. Literature review

III. Methodology

- III.1. Data
- III.2. Variables
- III.1. Models

IV. Results

V. Discussion and conclusions

# I. Introduction

- **Data** and their characteristics are the **most crucial elements** of any prediction model (Anderson, 2007).

- **if** the model is run on an **imbalanced dataset**, it **optimizes the overall prediction accuracy but** it does not take the disproportion between the number of failed and non-failed firms into account (Lopez et al., 2013). In turn, this **results in a poor classification rate for the minority class** (Wilson & Sharda, 1994). Most precisely, **type I error** (misclassifying a bankrupt firm as non-bankrupt) tends to be high in models using imbalanced datasets.

→ most bankruptcy prediction models use datasets that **do not represent the real-world conditions.**

They use paired samples of firms that contain the same number of failed and non-failed firms (Daily & Dalton, 1996; Ciampi et al., 2015) although bankruptcy is rarely observed in the real-world.

- **Solutions** as the sequential **boosting technique** and **resampling** may help to solve this issue.
- Few studies focus on imbalanced datasets in the bankruptcy prediction field (Kim et al., 2015; Zhou, 2013; Séverin & Veganzones, 2018).

# I. Introduction

## ➢ Aim of the paper :

Compare the accuracy of **different prediction models** based on information from firms' balance sheets and income statements.

## ➢ Original data :

A dataset of 2,266 Belgian firms including 153 bankrupt firms and 2,113 non-failed firms.

## ➢ Methods:

First, **logit** modelization on the original dataset.
Second, **boosting** (Schapire, 1990)
Third, **resampling** methods aiming to create a balanced distribution.

# II. Literature review

## ➢ Models

- ➢ Beaver's (1966) **discriminant analysis** on a single financial ratio.
- ➢ Altman (1968), Olhson (1980) and Zmijewski (1984) developed **statistical methods**.
- ➢ In the 90's, to **artificial intelligence** methods such as neural networks (Odom & Sharda, 1990).
- ➢ Recently, **ensemble methods**, as boosting (du Jardin et al., 2017), have been used for corporate failure prediction.

# II. Literature review

## ➢ Data

- ➢ Financial information represents the main element in bankruptcy prediction.
- ➢ Most bankruptcy modelization use balanced samples including the same proportion of failed and non-failed firms. This 'paired sample' (generally by size and/or industry) technique prevents the model from neglecting failed firms class prediction accuracy. Nevertheless, in this case, sample-selection bias may occur (Zmijewski, 1984).

Models build upon balanced samples outperformed the ones built upon imbalanced ones, especially for failed firms (Wilson & Sharda, 1994; McKee & Greenstein, 2000).

- ➢ **Solutions to improve the accuracy** of models (and especially the classification rate of failed firms) built upon imbalanced datasets.
  - ➢ **Resampling** the data ;
  - ➢ Assigning different **weights** (penalties) to observations depending on their misclassification instances.

# II. Literature review

## ➢ Methods

### ➢ Resampling

Resampling the dataset **to make their distributions balanced**.
This data manipulation prevents the model from neglecting failed firms class prediction accuracy.

Two categories methods exist: **under or oversampling**.

- **Under-sampling** consists in **removing** observations from the majority class while **oversampling** duplicates or **creates synthetic observations** to increase the number of cases of the minority class.
  - Under-sampling techniques allow to reduce the time spent to train the models but suffer from the loss of information because observations have been deleted (Seiffert et al., 2008). In the case of bankruptcy prediction, real word conditions results in an enormous the loss of healthy firms.
  - In contrast, the use of oversampling methods does not imply any loss of information but requires more time to train the models (Japkowicz & Stephen, 2002; Seiffert et al., 2008) and can lead to over-fitting (Drummond et al., 2003).

# II. Literature review

## ➢ Methods

### ➢ Resampling

**Zhou (2013) and Kim and Ahn (2015)** used sampling techniques on originally imbalanced datasets; their results report an improved accuracy following the **resampling**.

**Séverin and Veganzones (2018)** report that **SMOTE** (Synthetic Minority Oversampling Technique) **outperforms** other sampling techniques.

# II. Literature review

## ➢ Methods

### ➢ Cost-sensitive classification methods

Cost-sensitive classification methods consist in **assigning penalties to misclassified instances.**

Cost-sensitive classification methods may be highly sensitive to samples and thus generate unstable classifiers (Kim et al, 2015).

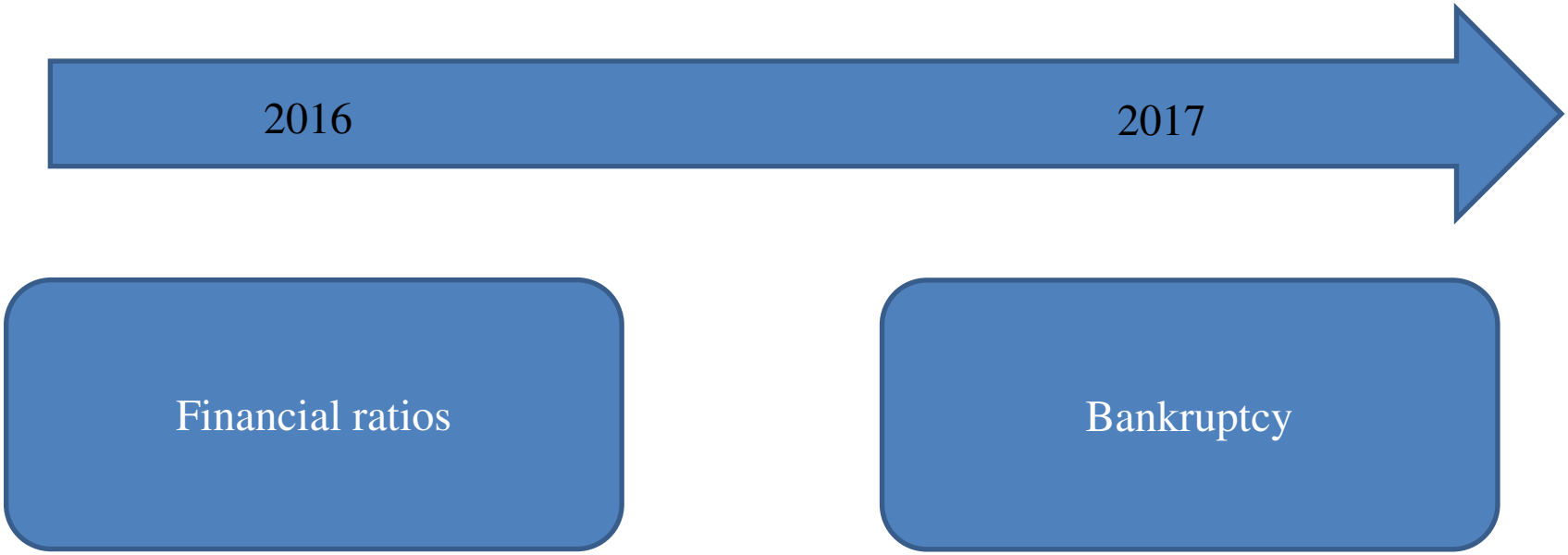
**Boosting technique (Schapire, 1990) sequentially builds models in which higher weight (penalty) is assigned to incorrectly classified observations.**
Boosting provides more learning opportunities for minority class samples and therefore represents an appropriate technique to solve data imbalance problem.

In the field of bankruptcy prediction, as per **du Jardin et al. (2017),** on the whole, **boosting** leads to more accurate models than single models.

# III. Methodology

## ➢ Data

- ➢ Bureau Van Dijk Bel-First database → Belgian firms
- ➢ Since the purpose of this paper is to predict bankruptcy 1 year in advance, financial ratios of both type of companies were calculated from year 2016.

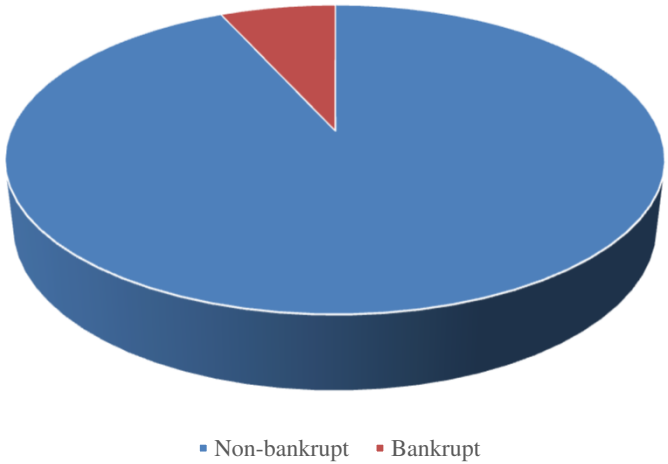


# III. Methodology

## ➢ Data

- We identify 153 firms that went bankrupt in the year 2017 and 2,114 non-bankrupt companies in the same year.

| Classification | Companies | Participation |
|---|---|---|
| Non-bankrupt | 2,113 | 93.2% |
| Bankrupt | 153 | 6.8% |
| Total | 2,266 | 100.0% |



# III. Methodology

## ➢ Variables

*Dependent variable:*

**Bankruptcy** takes the value of one when a company is bankrupt and the value of zero otherwise.

*Independent variables:*

Liquidity, profitability and debt ratios were considered in the study to predict bankruptcy (Ben, 2017).

| Category | Variable | Calculation |
|---|---|---|
| **Liquidity** | Free cash flow | Net cash from operating activities+Capex |
| **Liquidity** | Current ratio | Current assets/Current liabilities |
| **Profitability** | Ebitda | EBIT+Depreciation+Amortization |
| **Profitability** | ROA | Net profit/Assets |
| **Profitability** | ROE | Net profit/Equity |
| **Profitability** | Net added value | Operating income - Purchases - Services and other goods |
| **Debt** | Debt concentration | Current liabilities/Total liabilities |
| **Debt** | Debt level | Total liabilities/Total Assets |
| **Debt** | Financial Leverage | Financial liabilities/Equity |

# III. Methodology

## ➢ Models

### ➢ Logistic regression

$$P_i = \frac{e^z}{1 + e^z}$$

Where $P_i$ represents the likelihood of a specific firm enters in bankruptcy and z represents the independent variables that were mentioned before.

### ➢ Boosting

Boosting technique (Schapire, 1990) sequentially builds models in which higher weight (penalty) is assigned to incorrectly classified observations.

AdaBoost (Freund & Schapire, 1997) assigns the same weight $1/N$ to a set of data. The algorithm generates several iterations $m = 1,2,3, \ldots, M$. **In each iteration the weight of all observations are modified in accordance to their classification accuracy.** At round $m$, the weights are decreased for observations that were classified properly and the weights are increased for those that were misclassified.

# III. Methodology

## ➢ Models

### ➢ SMOTE (Synthetic Minority Oversampling TEchnique)

SMOTE is an oversampling technique proposed by Chawla, Bowyer, Hall, & Kegelmeyer (2002) in order to **create synthetic observations that create a balanced dataset.**

The algorithm generates a new sample considering specific observations with K (nearest neighbours) similar minority class.

# IV. Résults

## ➢ Descriptives

| Variable | Mean Non-bankrupts | Sd Non-bankrupts | Mean Bankrupts | Sd Bankrupts | Test of equal means |
|---|---|---|---|---|---|
| FCF | 569.30 | 6291.09 | 87.4 | 404.08 | 3.2** |
| Current ratio | 2.59 | 5.06 | 5.42 | 12.35 | -3.48*** |
| Ebitda | 674.52 | 7611.96 | 78.64 | 252.39 | 3.51*** |
| Net added value | 1532.39 | 14390.42 | 252.99 | 506.57 | 3.12** |
| ROE | 0.19 | 0.58 | 0.09 | 0.47 | 2.58* |
| ROA | 0.05 | 0.11 | 0.02 | 0.17 | 1.66· |
| Financial leverage | 1.38 | 3.54 | 1.24 | 4.09 | 3.54*** |
| Debt level | 0.59 | 0.24 | 0.55 | 0.28 | 1.11 |
| Debt concentration | 0.51 | 0.36 | 0.7 | 0.32 | -7.17*** |

# IV. Results

## ➢Logit and boosting on original database

**Table : Confusion matrix**

| | Boosting algorithm | | Logistic regression | |
|---|---|---|---|---|
| Clasificación | Bankrupt | Non-bankrupt | Bankrupt | Non-bankrupt |
| Bankrupt | 73.3% | 12.3% | 0.0% | 0.0% |
| Non-bankrupt | 26.7% | 87.7% | 100.0% | 100.0% |
| Total | 100.0% | 100.0% | 100.0% | 100.0% |
| Error type I | 26.7% | | 100% | |
| Error type II | 12.3% | | 0% | |
| Global accuracy rate | 86.7% | | 93.2% | |

**Boosting algorithm**: 73.3% of bankrupt and 87.7% of Non-bankrupt companies were classified correctly.

**Logistic regression**: bankrupt companies were predicted incorrectly while all non-bankrupt firms were classified properly.

→ According to Liang et al. (2016) error type I is more critical because it implies the loss of the credit granted and not only an opportunity cost as type 2 error.

→ **error type I si still high → resampling.**

# IV. Results

## ➢Resampling

We do **test the models on different proportions** (using smote) to reduce the type 1 error and evaluate its sensibility. .

Five groups (Kim, Kang, & Bae, 2015) were created according to different balance rates :

- 1:1;
- 1:3;
- 1:5;
- 1:10;
- 1:20.

# IV. Results : Confusion matrixes for different imbalanced proportions

**Boosting algorithm**

| Clasification | Initial sample | | 1:1 | | 1:3 | | 1:5 | | 1:10 | | 1:20 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Bankrupt | Non-bankrupt | Bankrupt | Non-bankrupt | Bankrupt | Non-bankrupt | Bankrupt | Non-bankrupt | Bankrupt | Non-bankrupt | Bankrupt | Non-bankrupt |
| **Bankrupt** | 73.3% | 12.3% | 95.3% | 6.5% | 91.8% | 9.2% | 80.0% | 5.2% | 70.0% | 0.7% | 48.4% | 1.8% |
| **Non-bankrupt** | 26.7% | 87.7% | 4.7% | 93.5% | 8.2% | 90.8% | 20.0% | 94.8% | 30.0% | 99.3% | 51.6% | 98.2% |
| **Total** | **100.0%** | **100.0%** | **100.0%** | **100.0%** | **100.0%** | **100.0%** | **100.0%** | **100.0%** | **100.0%** | **100.0%** | **100.0%** | **100.0%** |
| **Error type I** | 26.7% | | 4.7% | | 8.2% | | 20.0% | | 30.0% | | 51.6% | |
| **Error type II** | 12.3% | | 6.5% | | 9.2% | | 5.2% | | 0.7% | | 1.8% | |
| **Global Accuracy** | 86.7% | | 94.4% | | 91.0% | | 92.3% | | 96.7% | | 95.8% | |

**Logistic Regression**

| Clasification | Initial sample | | 1:1 | | 1:3 | | 1:5 | | 1:10 | | 1:20 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Bankrupt | Non-bankrupt | Bankrupt | Non-bankrupt | Bankrupt | Non-bankrupt | Bankrupt | Non-bankrupt | Bankrupt | Non-bankrupt | Bankrupt | Non-bankrupt |
| **Bankrupt** | 0.0% | 0.0% | 85.0% | 25.2% | 26.2% | 6.5% | 17.8% | 0.4% | 10.0% | 1.3% | 0.0% | 0.0% |
| **Non-bankrupt** | 100.0% | 100.0% | 15.0% | 74.8% | 73.8% | 93.5% | 82.2% | 99.6% | 90.0% | 98.7% | 100.0% | 100.0% |
| **Total** | **100.0%** | **100.0%** | **100.0%** | **100.0%** | **100.0%** | **100.0%** | **100.0%** | **100.0%** | **100.0%** | **100.0%** | **100.0%** | **100.0%** |
| **Error Type I** | 100.0% | | 15.0% | | 73.8% | | 82.2% | | 90.0% | | 100.0% | |
| **Error Type II** | 0.0% | | 25.2% | | 6.5% | | 0.4% | | 1.3% | | 0.0% | |
| **Global Accuracy** | 93.2% | | 79.9% | | 76.7% | | 85.9% | | 90.6% | | 95.2% | |

# IV. Discussion and conclusion

- Prediction **accuracy** of both models **decreases as the asymmetry is greater**.

- **Boosting algorithm has better prediction results** for bankrupt firms. Error type I increases while the imbalance in the dataset is greater. In the case of boosting algorithm, error type I is 4.7% when the dataset is symmetric. However, when the dataset is 1:20 this error is 51.6%.

- Using **logistic regression** error type I is 15% when the dataset is symmetric and 100% when the dataset is 1:20. In this context, there is no way to classify correctly a bankrupt company using logistic regression when the dataset is imbalanced at 1:20.

- Through **boosting** algorithm is possible to **reduce the probability error type I and II in comparison to logistic regression**.

- Ultimately, the **best model** is the one using the **boosting** algorithm on a **balanced sample** created through the SMOTE oversampling technique.

# IV. Discussion and conclusion

- Our results are **in line** with Zhou (2013) and Kim and Ahn (2015) and Sévérin and Veganzones (2018) report an improved accuracy of the models following the **resampling**.

- We are also in line **with Kim (2015) and du Jardin et al. (2017)** reporting that **boosting** technique is suitable to bankruptcy prediction modeling in real world conditions.

- Limitations:

  - we rely on a database of firms from only one country, **Belgium**, which has specific characteristics that may influence the results.
  - Study based on financial variables only although it has been proved that the **inclusion of non-financial variables** into models can improve their accuracy (Ciampi, 2015, Tobback et al., 2017).

# III. Methodology

## ➢ Variables

*Independent variables:*

One of the challenges in bankruptcy studies is variance stability since the financial information present different distributions, outliers and asymmetry (Jones, Johnstone, & Wilson, 2017). These characteristics of the data affect bankruptcy prediction. Data transformation proposed by Yeo and Johnson (2000) was applied in order to overcome these issues.