Forecasting Firm Failure via Random Forests and Logit

Mila Andreani, Lorenzo Bellomo, Andrea Pugnana, Laura State, Jack Tacchi

April 2021

Table of Contents

Research Question

Do Machine Learning algorithms provide better results than traditional statistical methods [1, 2] for **firm failure prediction**? To answer this question, we are going to present an unsupervised approach based on several **clustering algorithms** and a supervised approach exploiting two models, **Logit** and **Random Forests**.

Data pulling

AIDA covers over 2 million firms starting from 2001.
Around 110k firms are labeled as failed, which account for 5 % of the total firms.
We considered only firms failed after 2011, restricting the sample of failed firms to $\sim$21,000.
We randomly sampled other firms in order to attain a total number of 70,000 firms.

Dataset Variables

- The dataset contained several variables related to demographics of the firms and balance sheets data.
- For each of the balance sheet variables we had the lagged version of the same variable, up to 5 years in the past.

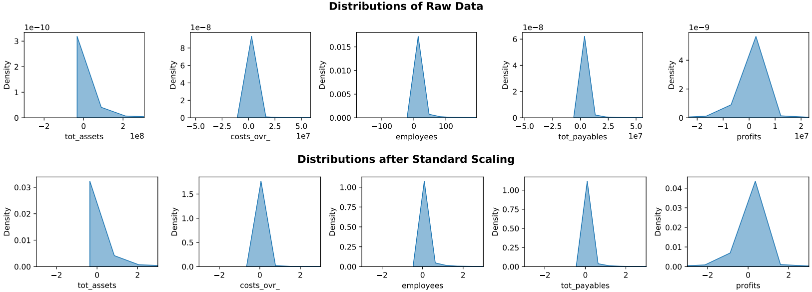
Variables in the dataset:

- company name, ateco code, legal form, ID
- location (longitude/lattitude)
- account closing date (last entry)
- number of employees
- total payables + costs ovr + retained earnings
- profits + revenues + total assets
- shareholder funds, working capital
- fail/no fail

Final Dataset

The sample was reduced further due to missing values.
The final sample contains 37,328 firms.
Roughly 35 % of these are failed firms.

Preprocessing



Standard scaler for the individual variables, i.e. subtract mean and divide by standard deviation, for each column individually.

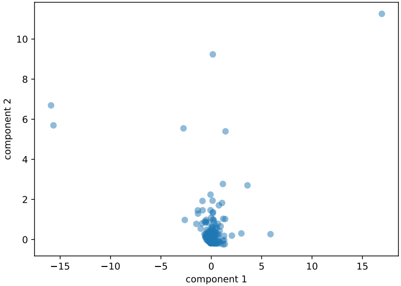
$$\frac{x_{ij} - \overline{x_i}}{\sigma_i} \quad \text{with} \quad i \quad \text{indexing columns,} \quad j \quad \text{indexing rows}$$

Clustering

- We take the data from the latest available time point (at time t) and subsample. We use only balance sheet information.
- Clustering follows after a PCA that extracts the two components that explain most variance in the data (approx. 0.6/0.3 as ratio of explained variance).

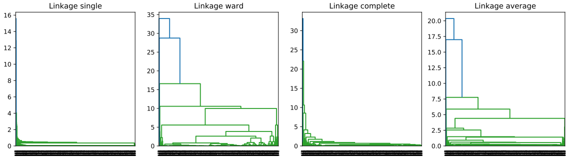
Clustering Algorithms

1. Hierarchical Clustering
2. K-Means Clustering
3. DBSCAN (density-based spatial clustering of applications with noise)
4. Mean-Shift Clustering

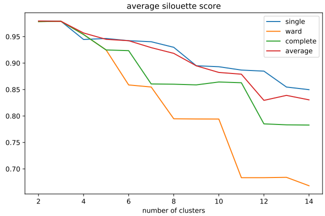


Hierarchical Clustering

Dendrogram for different linkage functions

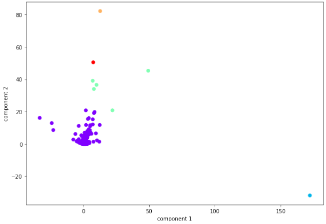


Average silhouette score for different linkage functions, and over an increasing number of clusters

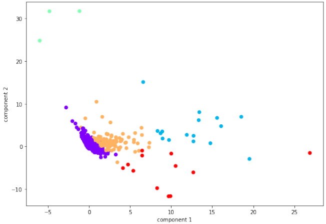


KMeans - Density Imbalance

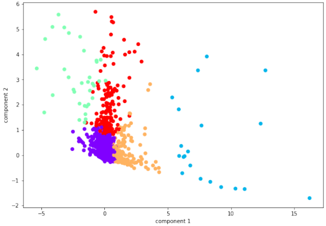
First application - k = 5



Second application - k = 5



Third application - k = 5



Cluster sizes at the end

9541

20

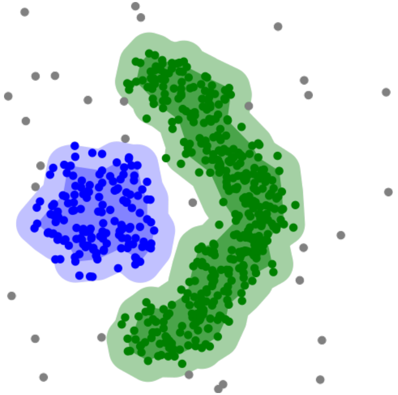
34

158

201

Clustering - DBSCAN Concept

- *Density* based algorithm [3]
- Three types of point:
  - *Core* Points
  - *Border* Points
  - *Noise* Points
- *Driven* by two parameters
  - ϵ: radius
  - **min_points**: requirement for cluster creation
- *Robust* to non linearly-separable clusters
- *Robust* to noise
- *Bad* results when data has various densities



WARNING: not our results :)

Clustering - DBSCAN Algorithm

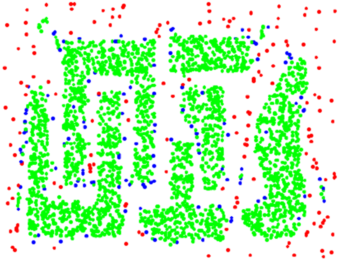


Figure: Green core, Blue border, Red noise

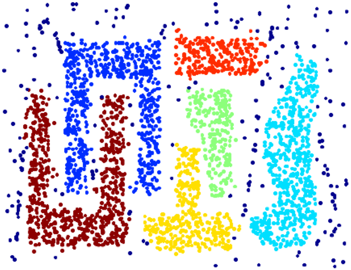


Figure: Cluster labels spread

Clustering - DBSCAN Parameter Estimation

How to estimate parameters?

- Try several **min_points**
- *Plot* the sorted distances to all the various *min_points* parameters choice
- *Pick the elbow of the curve*

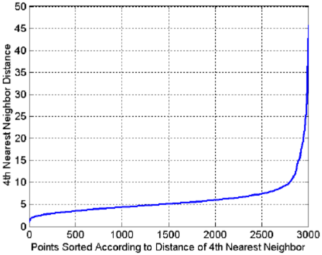


Figure: A good k-th NN plot

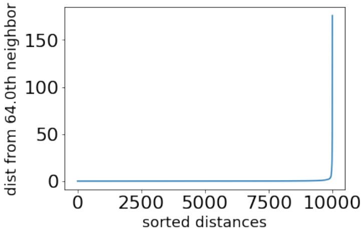
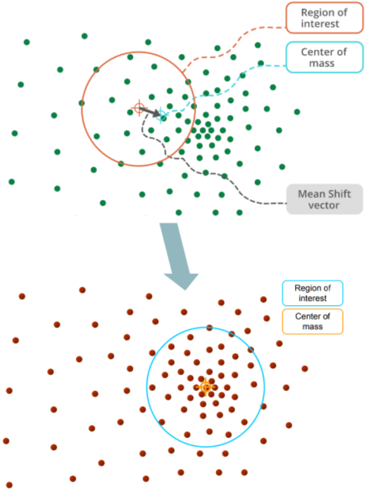


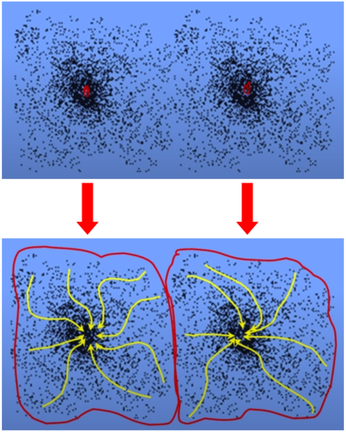
Figure: The harsh reality

Mean-Shift



- Another density-based clustering algorithm [4]
- For each point p:
  - Take the region of interest surrounding p
  - Calculate its centre of mass
  - “Shift” the region of interest to the new centre of mass
  - Repeat until there is no change (mode found)

Mean-Shift

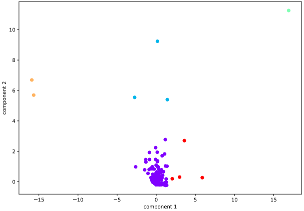


- After performing Mean-Shift for all data points:
  - Group all points within each *attraction basin* into a cluster

Clustering - Results

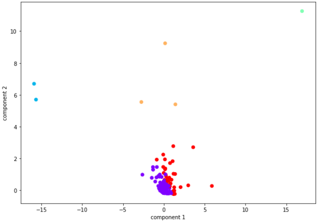
1) Hierarchical Clustering

linkage ward, # of clusters = 5



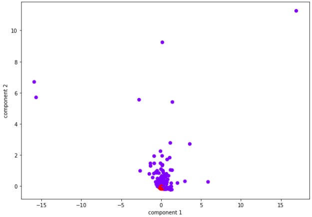
2) K-Means Clustering

of clusters = 5



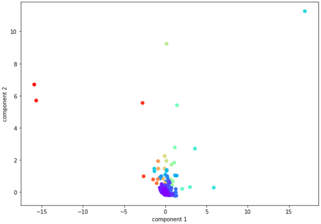
3) DBSCAN

$\epsilon = 0.9$ and **mp** $= 64$



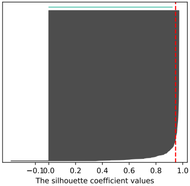
4) Mean-Shift Clustering

bandwidth $= 0.25$



Clustering - Results cont.

- General trend of silhouette curves: very high average silhouette score for small numbers of clusters, lower silhouette score with a larger numbers of clusters
- We find similar results for all four clustering algorithms and all tried combinations of parameter settings
- Repeatedly found **one very dense and big cluster**, even if we "zoom in" it is not possible to get a meaningful sub-partition



Example silhouette plot from hierarchical clustering, # of clusters = 5

Methodology

Our sample comprises $37,328$ observations from 2010 to 2020.
The 35% of which are failed firms.
The response variables are modeled as:

$$Y_i = f(\mathbf{X}_i, \mathbf{Z}_i)$$

where $Y_i \in [0, 1]$.

Predictor variables

By considering Altman (1968), $\mathbf{X}_i$ is composed of:

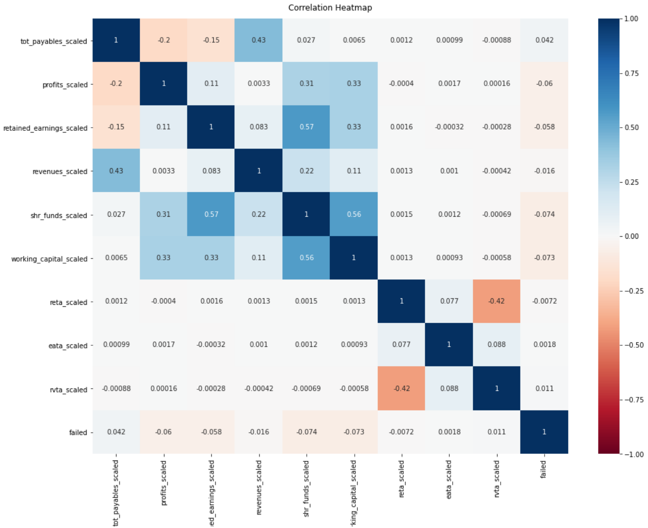
- X_1: Working Capital/Total Assets
- X_2: Retained Earnings/Total Assets
- X_3: Earnings before Interest and Taxes/Total Assets
- X_4: Market Value of Equity/ Total Debt
- X_5: Sales/Total Assets

whereas $\mathbf{Z}_i$ consists in balance sheet data.

Choice of Variables

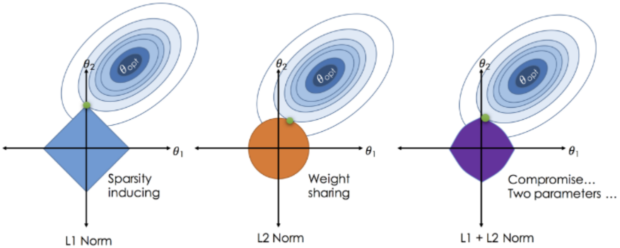
- Starting with 62 variables
- Removal of lagged variables, 17 variables remaining (including 4 ratios)
- Removal of variables correlated above a threshold of 0.65, 9 variables remaining (including 3 ratios)

Final Correlations



Logistic Regression (Logit)

- Logit
  - Classification algorithm
- Lasso
  - Introduces L1 Normalisation
- Ridge Regression
  - Introduces L2 Normalisation
- Elastic Net
  - Introduces L1 and L2 Normalisation



Random Forests

Ensemble algorithm aimed at mitigating the **bias-variance** trade-off of **decision trees**

A simple example

Let's say I have to weigh my cat, but I have no scales

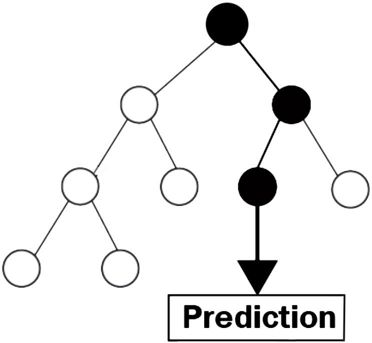
A simple example

I could ask my friends here their guess!

In Galton's seminal paper (1907)[6] something pretty similar happened.

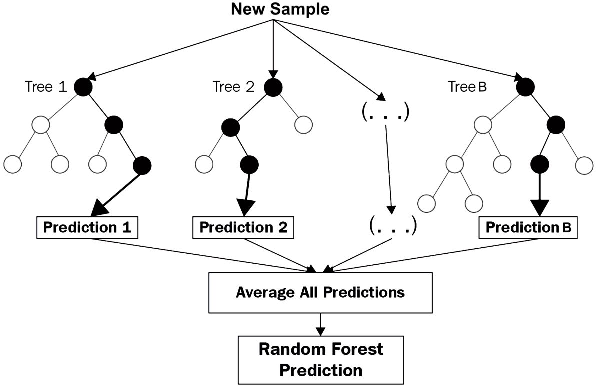
- Aggregating several opinions can drive to pretty accurate estimates;
- This is the so-called nowadays as **Wisdom of the Crowds**;
- Random Forests [5] are built on this principle.

Decision Trees



Random Forest

The prediction of the **Random Forest** is computed as the **average** prediction of all the individual decision trees.



Random Forest

Some details of the implementation:

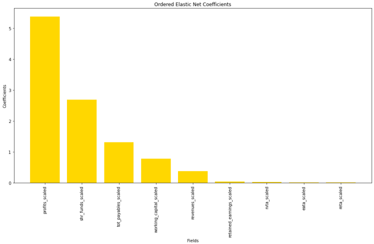
- Bootstrap the training set;
- Randomly extract features you want to use;
- Train a decision tree over this bootstrapped training set with limited features;
- Repeat *n* times;
- Aggregate trees prediction.

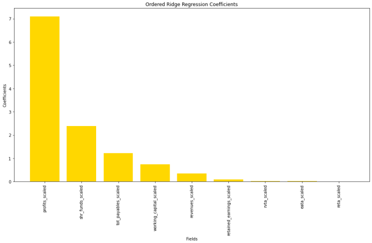
Final results

Model	AUC	Accuracy
Random Forest	0.78	0.81
Logit	0.70	0.78
Lasso	0.71	0.78
Ridge	0.71	0.78
Elastic Net	0.71	0.78

Notes: Lasso and Elastic Net didn't eliminate any variables.

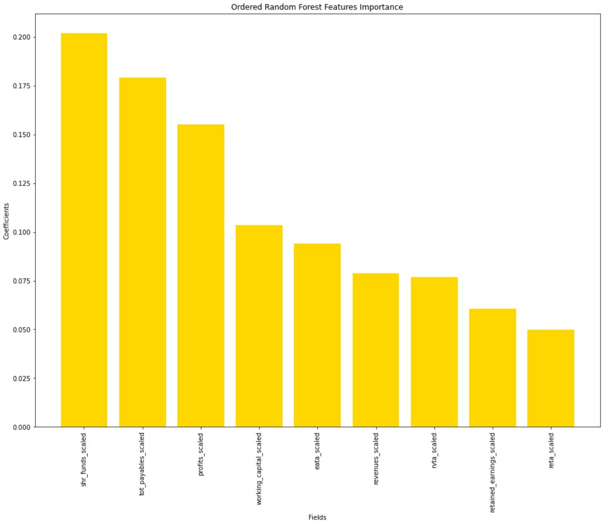
Coefficients





Notes: Coefficients are in the same order of importance for all 4 Logit models

Variable Importance



Random Forest specifications

Model	AUC	Accuracy
All features	0.81	0.83
Top10	0.78	0.81
Top5	0.77	0.80
Ratios+Categories	0.79	0.82
Logit Specification	0.78	0.81
Ratios	0.78	0.81

Conclusions

Machine Learning algorithms allow to extend statistical models by uncovering complex patterns among variables.

References I

- Edward I. Altman.
  Financial ratios, discriminant analysis and the prediction of corporate bankruptcy.
  The Journal of Finance, 23(4):589–609, sep 1968.

- Edward I Altman.
  Predicting financial distress of companies: revisiting the z-score and zeta® models.
  In *Handbook of research methods and applications in empirical finance*. Edward Elgar Publishing, 2013.

- Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu.
  A density-based algorithm for discovering clusters in large spatial databases with noise.
  pages 226–231. AAAI Press, 1996.

References II

- K. Fukunaga and L. Hostetler.
  The estimation of the gradient of a density function, with applications in pattern recognition.
  pages 32–40. IEEE Transactions on information theory, 1975.
- Leo Breiman.
  Random forests.
  Machine learning, 45(1):5–32, 2001.
- Francis Galton.
  Vox populi.
  Nature, 75(1949):450–451, mar 1907.