Detecting Financial Danger Zones with Machine Learning

Research Consortium for Systemic Risk Meeting

Jointly Sponsored with the

Systemic Risk Tomography (SYRTO) Project

December 15, 2014 - MIT Media Lab

Marika Vezzoli
University of Brescia
Main Objective

What’s our main target → stratify the “Risk Temperature” of:

- Country-risk
- Bank(and other Financial)-risk
- Corporate-risk

How do we do → Data Mining techniques:

- Regression Trees (RT) & Heatmap in each RT risk cluster (final node)
- Random Forests (RF)
- Tree on ensemble predictions (FRT – Final Regression Tree)

Where in SYRTO → EWS Implementation (WP 7)

- Risk Thresholds
- Warning Signals

Marika Vezzoli
University of Brescia
Regression Trees (RT)

- Regression Trees are non parametric methods that partition the predictor space $X$ into **homogeneous subsets with respect to the dependent variable $Y$**.
- They explain non-linear patterns between dependent variable and covariates.
- The main advantages are:
  - They identify the most important variables and corresponding split points thereby finding **risky/non risky** final zone (and their paths).
Regression Trees (RT)

- Regression Trees are non-parametric methods that partition the predictor space $X$ into homogeneous subsets with respect to the dependent variable $Y$.
- They explain non-linear patterns between dependent variable and covariates.
- The main advantages are:
  - They identify the most important variables and corresponding split points thereby finding risky/non risky final zone (and their paths).

Marika Vezzoli
University of Brescia
Regression Trees (RT)

- Regression Trees are non parametric methods that partition the predictor space $X$ into homogeneous subsets with respect to the dependent variable $Y$.
- They explain non-linear patterns between dependent variable and covariates.
- The main advantages are:
  - They identify the most important variables and corresponding split points thereby finding risky/non risky final zone (and their paths).

Marika Vezzoli
University of Brescia
Regression Trees (RT)

- Regression Trees are non-parametric methods that partition the predictor space $X$ into homogeneous subsets with respect to the dependent variable $Y$.
- They explain non-linear patterns between dependent variable and covariates.
- The main advantages are:
  - They identify the most important variables and corresponding split points thereby finding \textit{risky/non risky} final zone (and their paths).

Marika Vezzoli
University of Brescia
Regression Trees (RT)

- Regression Trees are non parametric methods that partition the predictor space $X$ into homogeneous subsets with respect to the dependent variable $Y$.
- They explain non-linear patterns between dependent variable and covariates.
- The main advantages are:
  - They identify the most important variables and corresponding split points thereby finding risky/non risky final zone (and their paths).

Marika Vezzoli
University of Brescia
Regression Trees (RT)

- Regression Trees are non parametric methods that partition the predictor space $X$ into **homogeneous subsets with respect to the dependent variable $Y$**.
- They explain non-linear patterns between dependent variable and covariates.
- The main advantages are:
  - They identify the most important variables and corresponding split points thereby finding **risky/non risky** final zone (and their paths).
Regression Trees (RT)

- Regression Trees are non parametric methods that partition the predictor space $X$ into homogeneous subsets with respect to the dependent variable $Y$.
- They explain non-linear patterns between dependent variable and covariates.
- The main advantages are:
  - They identify the most important variables and corresponding split points thereby finding risky/non risky final zone (and their paths).

Marika Vezzoli
University of Brescia
Regression Trees (RT)

- Regression Trees are non-parametric methods that partition the predictor space $X$ into \textit{homogeneous subsets with respect to the dependent variable} $Y$.
- They explain non-linear patterns between the dependent variable and covariates.
- The main advantages are:
  - They identify the most important variables and corresponding split points thereby finding \textit{risky/non risky} final zone (and their paths).

Marika Vezzoli
University of Brescia
Regression Trees (RT)

- Regression Trees are non parametric methods that partition the predictor space $X$ into **homogeneous subsets with respect to the dependent variable** $Y$.
- They explain non-linear patterns between dependent variable and covariates.
- The main advantages are:
  - They identify the most important variables and corresponding split points thereby finding **risky/non risky** final zone (and their paths).
Regression Trees (RT)

- Regression Trees are non-parametric methods that partition the predictor space $X$ into homogeneous subsets with respect to the dependent variable $Y$.
- They explain non-linear patterns between dependent variable and covariates.
- The main advantages are:
  - They identify the most important variables and corresponding split points thereby finding risky/non-risky final zone (and their paths).

Marika Vezzoli
University of Brescia
Regression Trees (RT)

- Regression Trees are non parametric methods that partition the predictor space \( X \) into **homogeneous subsets with respect to the dependent variable** \( Y \).
- They explain non-linear patterns between dependent variable and covariates.
- The main advantages are:
  - They identify the most important variables and corresponding split points thereby finding **risky/non risky** final zone (and their paths).

Marika Vezzoli
University of Brescia
Regression Trees (RT)

- Regression Trees are non parametric methods that partition the predictor space $X$ into homogeneous subsets with respect to the dependent variable $Y$.
- They explain non-linear patterns between dependent variable and covariates.
- The main advantages are:
  - They identify the most important variables and corresponding split points thereby finding risky/non risky final zone (and their paths).

Marika Vezzoli
University of Brescia
Regression Trees (RT)

- Regression Trees are non parametric methods that partition the predictor space $X$ into homogeneous subsets with respect to the dependent variable $Y$.
- They explain non-linear patterns between dependent variable and covariates.
- The main advantages are:
  - They identify the most important variables and corresponding split points thereby finding **risky/non risky** final zone (and their paths).

![Diagram of Regression Trees](image)
Regression Trees (RT)

- Regression Trees are non parametric methods that partition the predictor space $X$ into homogeneous subsets with respect to the dependent variable $Y$.
- They explain non-linear patterns between dependent variable and covariates.
- The main advantages are:
  - They identify the most important variables and corresponding split points thereby finding risky/non risky final zone (and their paths).
Heatmap

In order to understand what happens inside each terminal node, we use a graphical representation.

For each region, we visualize the values of all covariates by means of colors: from blue (low values) to red (high values).

In this way we have an idea of how variables are “expressed” in each node.

Marika Vezzoli
University of Brescia
Heatmap

$X_1 \leq \alpha_1$

$X_2 \leq \alpha_2$

$T_2$

$T_3$

$X_1 \leq \alpha_2$

$X_2 \leq \alpha_4$

$T_4$

$T_5$
Heatmap

$X_1 \leq \alpha_1$

$X_2 \leq \alpha_2$

$X_1 \leq \alpha_2$

Heatmap on node 3

$X_2 \leq \alpha_3$

$X_1 \leq \alpha_3$

$X_2 \leq \alpha_4$
Heatmap

$X_1 \leq \alpha_1$

$X_2 \leq \alpha_2$

$T_1$

$T_2$

$X_1 \leq \alpha_2$

$T_5$

Heatmap on node 4

Color Key

Value

Marika Vezzoli
University of Brescia
Heatmap

$X_1 \leq \alpha_1$

$X_2 \leq \alpha_2$

$\overline{T}_1$

$\overline{T}_2$

$\overline{T}_3$

$\overline{T}_4$

Heatmap on node 5

Color Key

- Value 1

Marika Vezzoli
University of Brescia
Ensemble learning techniques (P&C techniques) have been introduced to increase the accuracy of the results:

- Combining multiple versions of unstable classifiers increases the accuracy of the predictors.
Ensemble methods

Ensemble learning techniques (P&C techniques) have been introduced to increase the accuracy of the results:

✓ Combining multiple versions of unstable classifiers increases the accuracy of the predictors
Ensemble learning techniques (P&C techniques) have been introduced to increase the accuracy of the results:

- Combining multiple versions of unstable classifiers increases the accuracy of the predictors
Ensemble learning techniques (P&C techniques) have been introduced to increase the accuracy of the results:

- Combining multiple versions of unstable classifiers increases the accuracy of the predictors
Ensemble learning techniques (P&C techniques) have been introduced to increase the accuracy of the results:

- **Combining multiple versions** of unstable classifiers **increases** the accuracy of the predictors.
Ensemble learning techniques (P&C techniques) have been introduced to increase the accuracy of the results:

- Combining multiple versions of unstable classifiers increases the accuracy of the predictors.
Random Forests

- RF grow a non pruned tree on a training set which is a different bootstrap sample drawn from the data.

- An important issue of RF is about the use of Out-Of-Bag (OOB) predictions, where for each observation $z_i = (x_i; y_i)$ the algorithm computes the predictions by averaging only those trees grown using a training set not containing $z_i$.

- For improving the accuracy, the injected randomness has to maximize the differences between the trees. For this reason, in each tree node a subset of predictors is randomly chosen.

- RF provide an accuracy level that is in line with Boosting algorithm with better performance in terms of computational burden (Breiman, 2001).
RF increase the accuracy of the predictions but loose the interpretability of a single tree.

A possible simple solution is the FRT (Savona, Vezzoli 2013)

- The results of the RF are combined with RT. More precisely, we fit a RT using the RF predictions in place of the original dependent variable $Y$

- The substitution of $y$ with $\hat{y}$ mitigates the effects of the noisy data on the estimation process that affect both the predictions and the dependent variables itself.
Detecting Danger Zones in the Euro Svgn CDS

Joint Project by: V. Arakelian, P. Dellaportas, R. Savona, M. Vezzoli

✓ **What we inspect:** 5 yrs svgn CDS daily quotes of 7 Euro countries
  - GIIPS (Greece, Ireland, Italy, Portugal, Spain)
  - Core (France and Germany)
✓ **Period:** January 2008 – October 2013
✓ **Covariates:**
  - *Contagion-based measures:* non parametric daily correlations through Kendall’s Tau & Copulas (Arakelian&Dellaportas, 2010 - Contagion tests via copula threshold models, Quantitative Finance);
  - *Country-specific fundamentals:*
    1. inflation;
    2. export/GDP;
    3. GDP growth;
    4. industrial production;
    5. unemployment rate;
    6. debt/GDP.
The «Financial Greek Tragedy»

Marika Vezzoli
University of Brescia
Results – Regression Tree

- medcorr_GiIPS $\geq 0.3167$
- unempl. rate $< 0.1175$
- Debt/GDP $< 1.196$
- inflation $< 0.0315$
- Debt/GDP $< 0.9365$
- medcorr_other $< 0.4872$
- medcorr_GiIPS $\geq 0.4924$
- vs_svgnUS $< 0.08706$
- vs_EU Fin $\geq 0.3546$

Safe Zone

Marika Vezzoli
University of Brescia
Results – Regression Tree

- medcorr_GIIPS $\geq 0.3167$
- unempl. rate $< 0.1175$
- Debt/GDP $< 1.196$
- inflation $< 0.0315$
- Debt/GDP $< 0.9365$
- Debt/GDP $< 0.6645$
- medcorr_other $< 0.4872$
- medcorr_GIIPS $\geq 0.4924$
- vs_svgnUS $< 0.08706$
- vs_EU Fin $\geq 0.3546$
- medcorr_GIIPS $\geq 0.4924$
- vs_svgnUS $< 0.08706$
- 717
- 1217
- 575
- 842
- 717
- 1217

Risky Zones

Marika Vezzoli
University of Brescia
Results – Regression Tree

Extremely Risky Zones

Median correlation GIIPS \geq 0.3167

Unemployment rate < 0.1175

Debt/GDP < 1.196

Inflation < 0.0315

Debt/GDP < 0.9365

Median correlation other < 0.4872

Median correlation GIIPS \geq 0.4924

vs. svgnUS < 0.08706

vs. EU Fin \geq 0.3546

575 842

717 1217

Marika Vezzoli
University of Brescia
Results – Regression Tree

- $\text{medcorr}_{\text{GIIPS}} \geq 0.3167$
- $\text{unempl. rate} < 0.1175$
- $\text{Debt/GDP} < 1.196$
- $\text{inflation} < 0.0315$
- $\text{Debt/GDP} < 0.9365$
- $\text{Debt/GDP} < 0.6645$
- $\text{medcorr}_{\text{other}} < 0.4872$
- $\text{medcorr}_{\text{GIIPS}} \geq 0.4924$
- $\text{vs}_\text{EU Fin} \geq 0.3546$
- $\text{vs}_\text{svgnUS} < 0.08706$

Marika Vezzoli
University of Brescia
Results – Variable Importance

Marika Vezzoli
University of Brescia
This project has received funding from the European Union’s Seventh Framework Programme for research, technological development and demonstration under grant agreement no° 320270

www.syrtoproject.eu

This document reflects only the author’s views. The European Union is not liable for any use that may be made of the information contained therein.