RIDS
RES INSTITUTE FOR DATA SCIENCE

UNIVERSITÀ
DI PAVIA

Credit data science risk models for SMEs

Silvia Figini

University of Pavia

Credit Risk

The problem

Problem

> **Default prediction** for Small and Medium Enterprises.

> **Probability Of Default (PD)** is the probability of an enterprise go to default.

> The main aim of credit risk is to **estimate PD**.

Credit Risk

The problem

Problem & Research Results

> **Default prediction** for Small and Medium Enterprises on a real data set provided by Unicredit.

> **Figini, Bonelli and Giovannini (2017) Solvency prediction for small and medium enterprises in banking, to appear in Decision Support Systems**

State of the art

> **"Linear"** and **"Parametric"** models are the most used.

Aim

> Propose **suitable non-standard models**.

Project Overview

Literature Research

> What can possibly be done?

Literature on Multivariate Outlier Detection and non-standard prediction methods.

Model Comparison

> What really works?

Compare performance between standard and non-standard models on the provided dataset.

Proposal

> What is of interest?

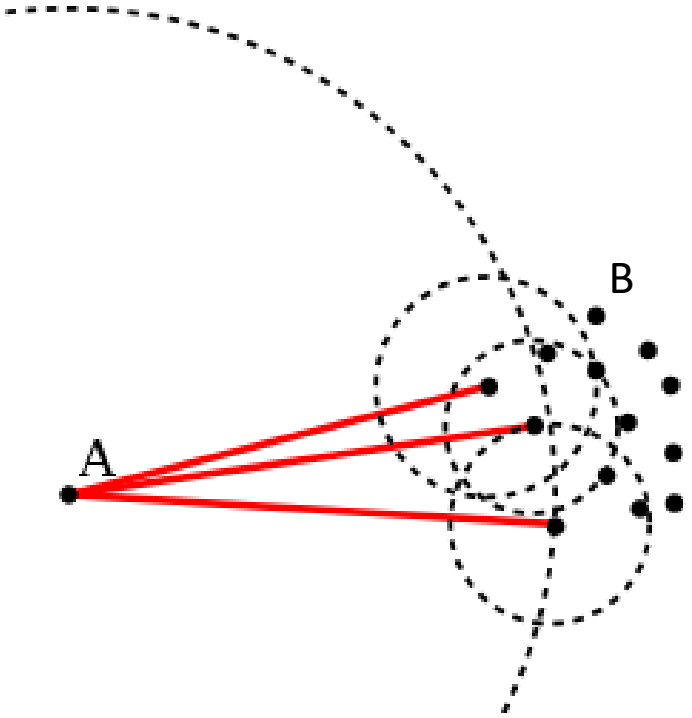
Introduction of models that, despite being non-standard, can appeal to the Risk Industry: many methods can be proposed, but only some can be used.

Local Outlier Factor

Multivariate Outlier Analysis

How likely A is an outlier? And B?

If we compare the local density of a point with the densities of its neighbors we find that A has a much lower density than any of its neighbors



Outlier detection

- Univariate outlier detection techniques are usually unsatisfactory when dealing with multivariate problems, because the relations among variables are not considered.
- The majority of multivariate techniques has strong assumptions about the data distribution, which are not respected when high dimensional real data are considered.

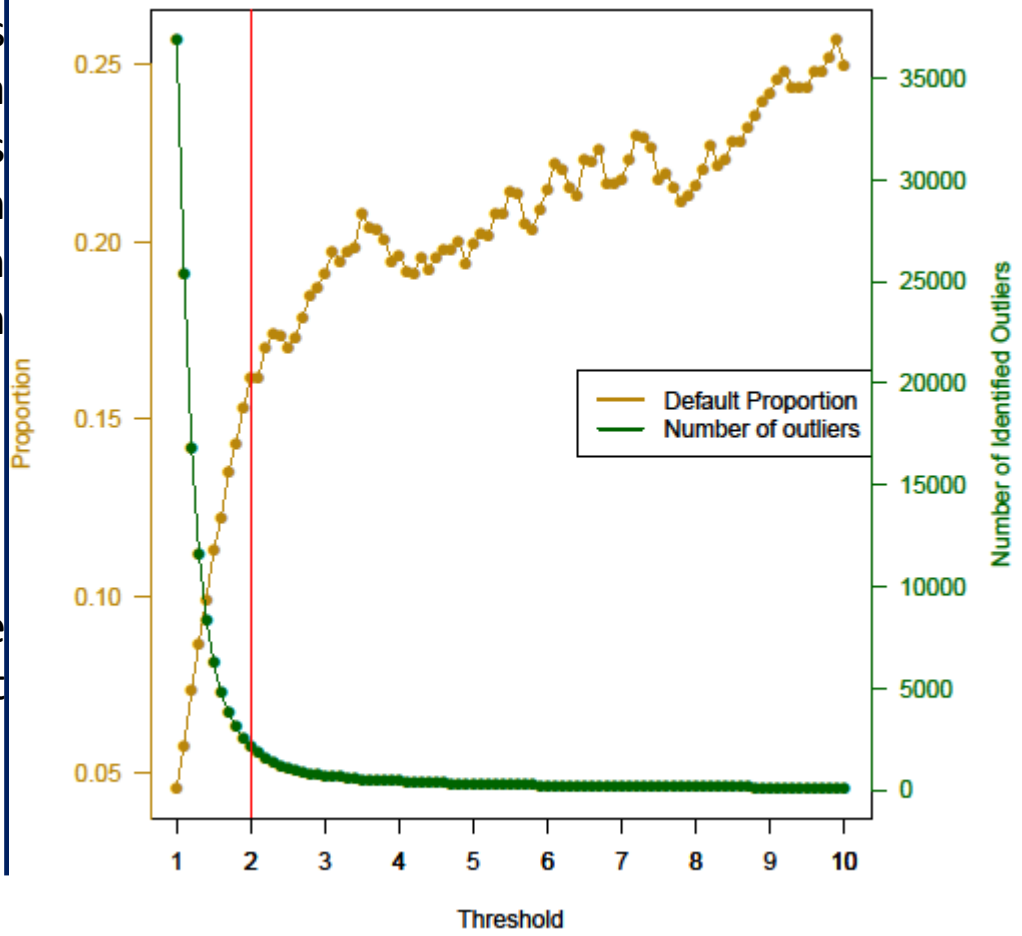
Local Outlier Factor

Results on the dataset

The Local Outlier Factor technique (LOF) is a multivariate technique which is consistent on high dimensional data without resorting to strong assumptions about the distribution. LOF provides an index which denotes how likely an observation can be considered as an outlier.

For each point we evaluate the LOF value

Points with high LOF are more likely to be outliers and it provides information about the “outlierness” of each point.



Local Outlier Factor

- This technique compares the local density of an object, dened by groups of its K-nearest neighbours, with the local densities of the closest neighbours themselves: if the local density of the object is too different from that of the neighbours' the observation can be considered an outlier.

- The LOF is a continuous variable that summarises a degree of "outlierness" for each observation. Of course, choosing a threshold c, the LOF index can be transformed in a binary variable which reports a value of 1 if the LOF > c and 0 otherwise.

Predictive models

Binary Generalised Extreme Value Model (BGEV)

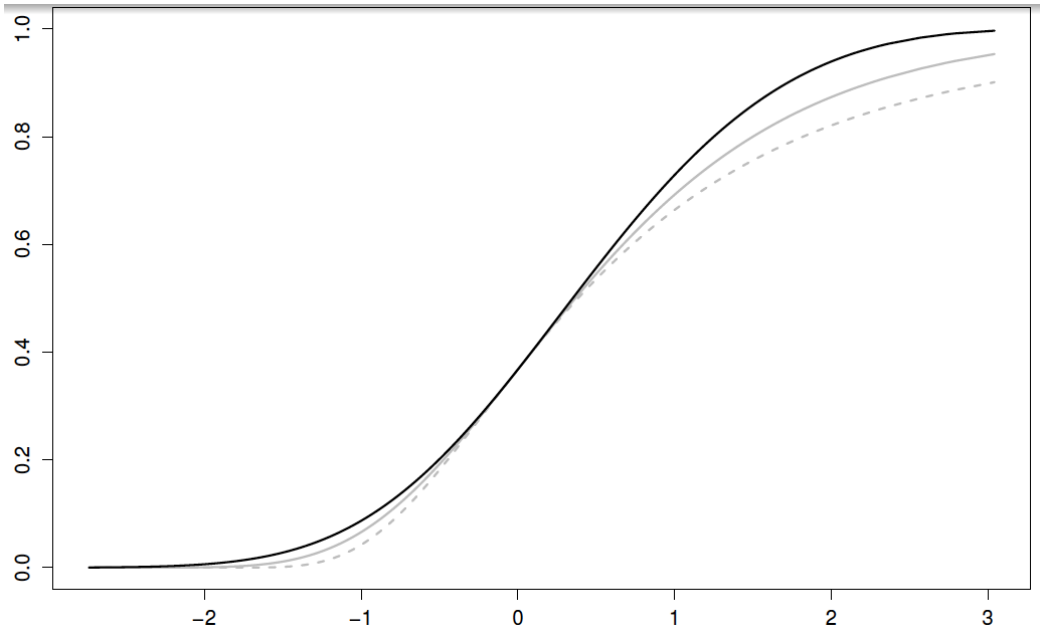
Parametric Linear Model

Link function: quantile function of the GEV distribution, where τ is a tail shape parameter

Model that **focus on predicting rare events** (Default)

Advantages: **no oversampling needed**

$$\frac{[-ln(PD)]^{-\tau} - 1}{\tau} = X\beta$$



Models under comparison

- Logistic regression (GLM)
- Linear Discriminant Analysis (LDA)
- Binary Generalised Extreme Value Model (BGEV)
- K-NN
- Classification Tree (CT)
- Random Forest (RF)
- Generalized Boosting (GBM)

Data description

- The real data set, provided by UniCredit bank, concerning credit risk of Italian SMEs has been analysed to predict the PD.

- The available sample refers to enterprises with annual revenue not exceeding 5 million euros in January 2015. Although information about the enterprises was collected throughout the years, the data available for the analysis consider only two years.

- Information about SMEs includes generic data (such as dimension, legal form, default status), financial ratios derived from the balance sheet, tendency and central credit register variables observed monthly.

- The target is a binary variable which represents the default status for each enterprise registered in 2015. The independent variables at hand are related to leverage liquidity, profitability, financial ratios, operations with bank, cash flow management, coverage, activity, size, including information about the number of employees, number of directors and number of subsidiaries.

Data description (2)

- The variables have been checked for linear dependence, and highly collinear ones have not been used in the analysis.
- After the descriptive data analysis we have deleted the variables with a percentage of missing values greater than 35% and we have removed highly correlated variables.
- The final data set is composed of 38036 rows and 43 variables with an a priori default probability of around 5%.

Data set

- **Filtered Data** This data set is derived from the original data set discarding all variables with too many missing values and all variables with high correlations.

- **Difference Data** This data set contains the same information of the ltered data set, with some transformations on the original variables. The transformed variables maintain the interpretability of the original one and give an idea of the variable trend.

- **LOF Data** This data set is obtained from the Difference data set. Two additional independent variables are added to identify if (dummy variable) and how much (LOF value) an enterprise is considered an outlier or not.

Logistic regression: results

	H	Gini	AUC	AUCH	KS	Spec.Sens95
Filtered Data	0.30	0.69	0.85	0.85	0.56	0.47
Differences Data	0.30	0.70	0.85	0.85	0.56	0.44
LOF Data	0.34	0.73	0.86	0.87	0.58	0.52

Table 1: Out of Sample performance measures for the Logistic Regression Model

Out of sample results

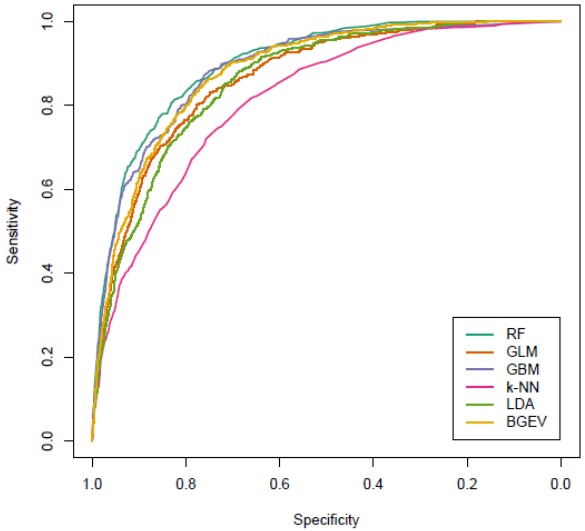


Figure 2: ROC curves for every method fitted on LOF Data

	H	Gini	AUC	AUCH	KS	Spec.Sens95
CT	0.23	0.54	0.77	0.77	0.52	0.58
k-NN	0.24	0.63	0.81	0.82	0.48	0.41
LDA	0.36	0.72	0.86	0.86	0.57	0.52
GLM	0.34	0.73	0.86	0.87	0.58	0.52
BGEV	0.41	0.76	0.88	0.89	0.62	0.56
GBM	0.41	0.78	0.89	0.89	0.63	0.59
RF	0.44	0.80	0.90	0.90	0.64	0.57

Table 2: Out of Sample (Test Set) Performance Measures for LOF Data

Greater AUC	Smaller AUC	p-value
BGEV	GLM	1.49e-11
GBM	GLM	4.39e-04
RF	GLM	1.06e-07
RF	GBM	1.11e-02
RF	BGEV	5.96e-04

Table 3: DeLong unilateral test with $H_0 : AUC_{greater} = AUC_{smaller}, H_1 : AUC_{greater} \geq AUC_{smaller}$ on LOF data set

Model Uncertainty in credit risk

- A «pool» approach to ensemble
  - which models to include in the pool of models
  - how to combine the model predictions

Figini, Savona and Vezzoli (2016) Corporate Default Prediction Model Averaging: A Normative Linear Pooling Approach, in Intelligent Systems in Accounting, Finance and Management

- Credit risk and model averaging

Figini et al. (2017) Credit risk assessment with Bayesian model averaging, in Communication in Statistics – Theory and Methods

Coherent assessment

- The Receiver Operating Characteristic (ROC) curve describes the performance of a classification or diagnostic rule (Lusted, 1971).

- Comparing ROC curves directly has never been easy, especially when those curves cross each other.

- Hence, summaries, such as the whole and the partial areas under the ROC curve, have been proposed (see, e.g. Hand, 2009), including H index.

ROC Curve or stochastic dominance indexes?

- In our opinion, the issue of model selection when ROC curves cross should be more adequately handled in the statistical literature.

- We have proposed a class of indices useful for model comparisons, when ROC curves show intersections.

- Referring to the literature on stochastic dominance and to the results therein obtained in case of crossing Lorenz curves, a novel method is provided for checking for unanimous rankings when the ROC curve dominance fails.

- Our method has the main advantage of establishing whether one distribution can be ranked superior to another according to the discriminative power, by looking at the entire distribution of the scores.

For more details see e.g. **Figini, Gigliarano and Muliere (2014), Making classifier performance comparison when ROC Curve intersect, Computational Statistics and Statistics and Data Analysis.**

ROC Curve for temporal models or polarization indexes for model comparison?

- We propose a new class of discrimination index for characterizing the predictive accuracy of survival models based on a restricted version of the Gini concentration index.
- The class of indexes proposed shows interesting mathematical properties.
- IDEA: Differences in the concentration of two survival distributions may suggest the presence of a differential covariates effect for some subgroups, thus providing evidence of high discriminative power.

See e.g. **Figini, Gigliarano and Muliere (2017), Polarization-based discrimination indexes for survival risk models, under review.**

Possible collaborations

- **International PhD program** in «Data science and Computational Statistics» (head Department of Mathematics, University of Pavia).

- Agreement for mobility in research for PhD students.

- Research Project in Data Sciences applied to insurance and finance (research project founded)

RIDS
RES INSTITUTE FOR DATA SCIENCE

UNIVERSITÀ
DI PAVIA

Thank you for your attention!

Silvia Figini

silvia.figini@unipv.it