FINTECH RISK MANAGEMENT
www.fintech-ho2020.eu

REGTECH WORKSHOP I

Credit scoring model development: importance of dataset and dataset analysis

Andrea Sorrentino, Head of Fintech

modefinance

European Commission

WHAT IS CREDIT SCORING

A **credit score**, is an **opinion on the creditworthiness of a company**, finalized to understand the likelihood the company will meet its financial commitments within a given time horizon (usually one year).
Credit score is expressed via an established ranking system.

VS

A **Credit Rating**, as defined in *Article 3(1)(a) of the CRA Regulation*, include quantitative analysis and sufficient qualitative analysis, according to the rating methodology established by the credit rating agency.
A measure of creditworthiness derived from summarising and expressing data based only on a pre-set statistical system or model, without additional substantial qualitative rating-specific analytical input from a rating analyst should not be considered as a credit rating.

A credit score cannot be considered a credit rating

how a scoring model is developed

A credit score can derive from a completely automated model (algorithm).

In the filed of creditworthiness analyses, several kind of different algorithms have been developed (Heuristic models, Machine Learning, Causal models) as previously mentioned.

As seen previously, creditworthiness analysis, since Altman's pioneer work, has become a science, **and so did the approach.**

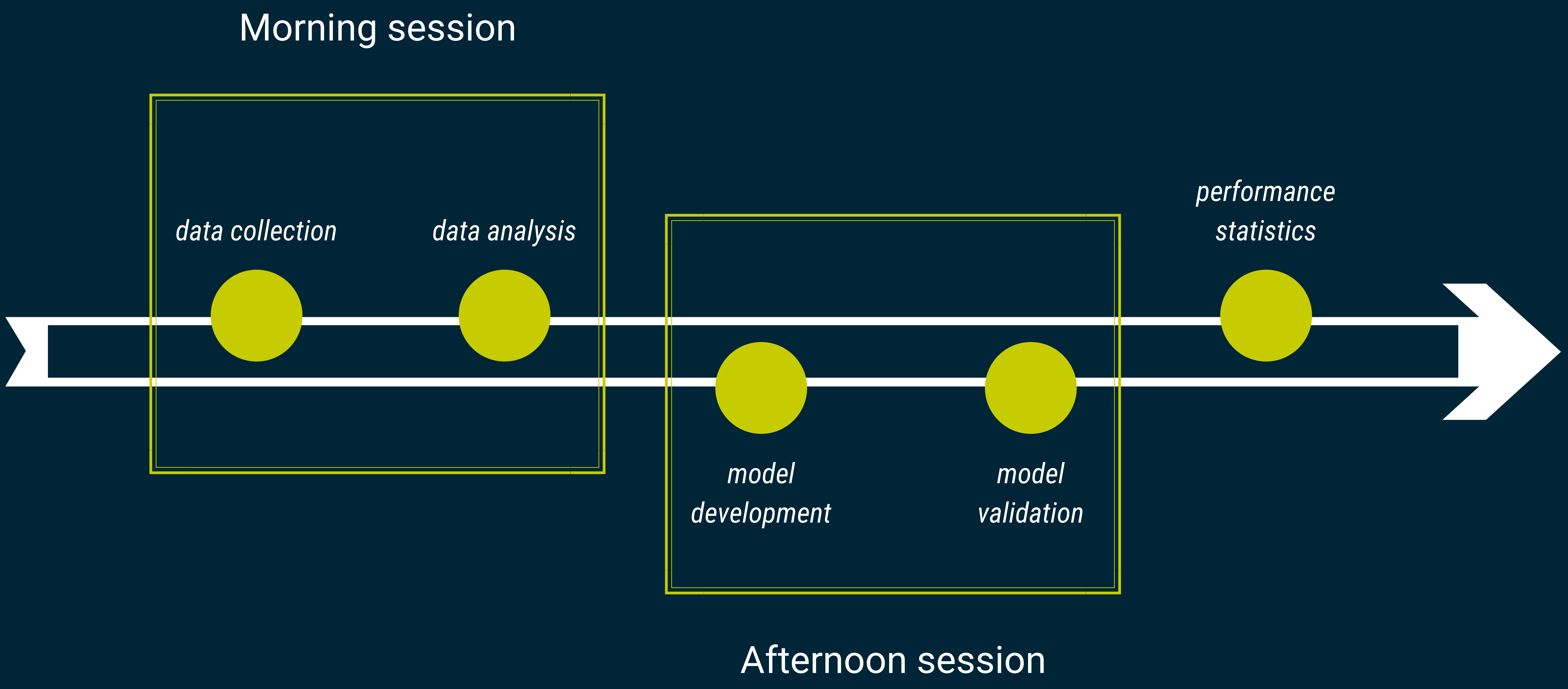
Nevertheless, **data science is a necessary but not sufficient condition.**

Financial knowledge of what numbers are telling us is mandatory.

$$\text{(I)} \qquad Z = .012X_1 + .014X_2 + .033X_3 + .006X_4 + .999X_5$$

where 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 equity/Book value of total debt
X_5 = Sales/Total assets
Z = Overall Index

how a scoring model is developed



data collection

Data collection is the development phase in which we’re creating a data set.

When we refer to a credit scoring model, the dataset will consist essentially of:

- **Registry information** about the counterparties (eg: country, industry classification, consolidation code of the accounts, etc.);
- Financials (**balance sheet and income statement**: issue of data availability and accounting differences among different countries);
- Sub sets of **Active and Defaulted companies**.

data collection: the dataset created

For the purpose of this event, we've collected data and created a dataset of companies with the following characteristics:

- Country: **Italy**;
- Size: **SMEs**;
- Type of entities: **Industrial companies**;
- Number of selected **active** companies: **13,413***;
- Number of selected **defaulted** companies: **1,632**.

*data collected to 2017

data collection: the dataset created

For these SMEs we've computed all financial ratios needed to perform a fundamentals' analysis:

ID	FORMULA	ID	FORMULA
RATIO001	(Total assets - Shareholders Funds)/Shareholders Funds	RATIO019	Interest paid/(Profit before taxes + Interest paid)
RATIO002	(Long term debt + Loans)/S. Funds	RATIO027	EBITDA/interest paid
RATIO003	Total assets/Total liabilties	RATIO029	EBITDA/Operating revenues
RATIO004	Current assets/Current liabilties	RATIO030	EBITDA/Sales
RATIO005	(Current assets - Current assets: stocks)/Current liabilties	RATIO036	Constraint EBIT
RATIO006	(Shareholders Funds + Non current liabilities)/Fixed assets	RATIO037	Constraint PL before tax
RATIO008	EBIT/interest paid	RATIO039	Constraint Financial PL
RATIO011	(Profit (loss) before tax + Interest paid)/Total assets	RATIO040	Constraint P/L for period th EUR
RATIO012	P/L after tax/Shareholders Funds	DPO	Trade Payables/Operating revenues
RATIO013	GROSS PROFIT/Operating revenues	DSO	Trade Receivables/Operating revenues
RATIO017	Operating revenues/Total assets	DIO	Inventories/Operating revenues
RATIO018	Sales/Total assets	NACE	Industry classification on NACE

focus: LEVERAGE

In creditworthiness analysis, **leverage and financial leverage represent fundamental metrics**.

Leverage *is defined as: Total liabilities/Shareholders funds*

It is a measure of the **level of indebtedness of a company**, respect its own sources.

The lower the better, BUT negative means disaster!

Financial leverage *is defined as: Financial liabilities/Shareholders funds*

It represents the **amount of bank debts** (i.e. high seniority) respect a company own funds.

focus: LEVERAGE

	2017-12-31 Chamber of commerce CONSOLIDATED	2016-12-31 Chamber of commerce CONSOLIDATED	2015-12-31 Chamber of commerce CONSOLIDATED
Turnover (€)	184,919,000	188,274,000	180,554,000
modefinance score	CCC	B	B
Probability of default			
1 year	25.33 %	5.08 %	4.70 %
Confidence	100.00 %	100.00 %	100.00 %
Solvency ratios			
Leverage ratio	17.51	11.05	11.92
Financial leverage	10.00	6.23	6.94
Total asset/Total liabilities	1.06	1.09	1.08
Liquidity ratios			
Current ratio	0.78	0.87	0.96
Quick ratio	0.49	0.58	0.66
Cash cycle ratio	40.00	38.00	53.00
Profitability ratios			
Return on investement ROI	0.35 %	8.76 %	7.63 %
Return on equity ROE	-55.52 %	18.99 %	30.53 %
Asset turnover	1.06	1.05	1.01
Interest Coverage ratios			
EBIT interest coverage ratio	1.03	2.83	2.63

focus: CASH CYCLE

Cash cycle represents the time (expressed in days) needed by a company to convert its inventory and trade receivables into cash, and the time used to repay the trade debtors.

The constituents of the cash cycle are:

DIO: *Inventories/Op. Revenues;*
DSO: *Trade receivables/Op. Revenues;*
DPO: *Trade payables/Op. Revenues.*

Cash Cycle: DIO + DSO - DPO

focus: CASH CYCLE

BALANCE SHEET (th €)	31/12/2017	31/12/2016	31/12/2015
Accounting practice	Local GAAP	Local GAAP	Local GAAP
Exchange rate USD - EUR	0.83382	0.94868	0.91853
Number of months	12	12	12
Total assets	**23,893,423**	**21,500,887**	**7,410,618**
Fixed assets	18,414,792	15,562,363	4,855,271
Intangible fixed assets	351,654	356,840	0
Tangible fixed assets	17,086,320	14,265,177	4,771,505
Other fixed assets	976,818	940,346	83,766
Current assets	5,478,631	5,938,524	2,555,347
Stocks	1,887,382	1,961,346	1,173,728
Debtors	429,735	473,525	155,199
Other current assets	3,161,514	3,503,654	1,226,420
Cash & cash equivalent	2,808,234	3,219,066	1,099,392
Shareholders funds	**3,533,097**	**4,508,977**	**995,411**
Capital	141	153	120
Other shareholders funds	3,532,956	4,508,825	995,291
Total liabilities	**20,360,326**	**16,991,910**	**6,415,207**
Non current liabilities	13,961,032	11,463,965	3,833,196
Long term debt	7,853,241	5,671,460	1,899,860
Other non-current liabilities	6,107,791	5,792,504	1,933,335
Current liabilities	6,399,294	5,527,946	2,582,011
Loans	747,561	1,091,118	576,768
Creditors	1,993,038	1,764,863	841,506
Other current liabilities	3,658,695	2,671,965	1,163,737
Total shareh. funds & liab.	**23,893,423**	**21,500,887**	**7,410,618**

NET DEBT (th €)	31/12/2017	31/12/2016	31/12/2015
Short term debts	747,561	1,091,118	576,768
Long term debt	7,853,241	5,671,460	1,899,860
Cash & cash equivalent	2,808,234	3,219,066	1,099,392
Net debt	**5,792,568**	**3,543,513**	**1,377,236**

	31/12/2017	31/12/2016	31/12/2015
Working capital	**324,079**	**670,008**	**487,421**
Net Current Assets	**-920,662**	**410,579**	**-26,664**

	31/12/2017	31/12/2016	31/12/2015
Days Sales Of Inventory (DIO)	70	108	115
Days Sales Outstanding (DSO)	16	26	15
Days Payable Outstanding (DPO)	74	97	83
Cash Conversion Cycle (DIO + DSO - DPO)	**12**	**37**	**47**

focus: CASH CYCLE

BALANCE SHEET (th €)	31/12/2017	31/12/2016	31/12/2015
Accounting practice	Local GAAP	Local GAAP	Local GAAP
Exchange rate EUR - EUR	1	1	1
Number of months	12	12	12
Total assets	**559,071**	**567,436**	**577,692**
Fixed assets	338,734	347,448	362,414
Intangible fixed assets	8,113	8,304	12,373
Tangible fixed assets	329,876	338,399	349,691
Other fixed assets	746	745	350
Current assets	220,336	219,988	215,278
Stocks	105,675	106,602	105,771
Debtors	35,732	34,940	35,307
Other current assets	78,929	78,446	74,199
Cash & cash equivalent	51,447	52,154	37,748
Shareholders funds	**127,134**	**128,772**	**131,806**
Capital	20,250	20,250	20,250
Other shareholders funds	106,884	108,522	111,556
Total liabilities	**431,937**	**438,664**	**445,886**
Non current liabilities	100,132	100,237	125,034
Long term debt	71,417	71,518	96,260
Other non-current liabilities	28,715	28,719	28,774
Current liabilities	331,805	338,427	320,852
Loans	73,953	75,184	79,811
Creditors	225,710	224,229	210,442
Other current liabilities	32,142	39,015	30,599
Total shareh. funds & liab.	**559,071**	**567,436**	**577,692**

NET DEBT (th €)	31/12/2017	31/12/2016	31/12/2015
Short term debts	73,953	75,184	79,811
Long term debt	71,417	71,518	96,260
Cash & cash equivalent	51,447	52,154	37,748
Net debt	**93,923**	**94,548**	**138,323**

	31/12/2017	31/12/2016	31/12/2015
Working capital	**-84,303**	**-82,687**	**-69,364**
Net Current Assets	**-111,469**	**-118,440**	**-105,575**

	31/12/2017	31/12/2016	31/12/2015
Days Sales Of Inventory (DIO)	35	35	34
Days Sales Outstanding (DSO)	12	12	11
Days Payable Outstanding (DPO)	75	74	68
Cash Conversion Cycle (DIO + DSO - DPO)	**-28**	**-27**	**-23**

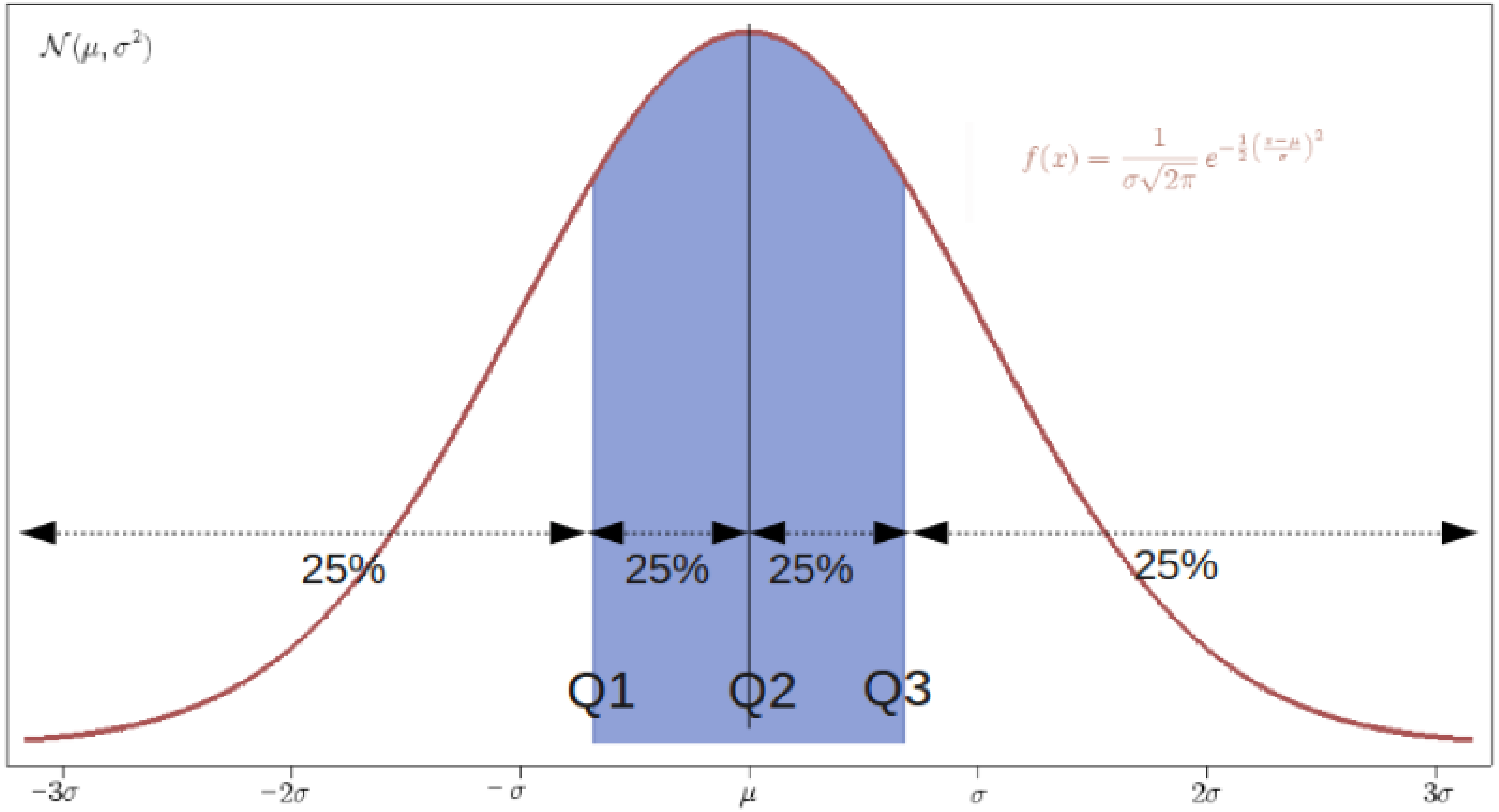
HANDS ON: DATA ANALYSIS

In this morning session we'll go through an algorithm that was developed so to carry on the following conceptual steps:

- Data comprehension: descriptive statistics;
- Correlation analysis;
- Discriminating capacity analysis.

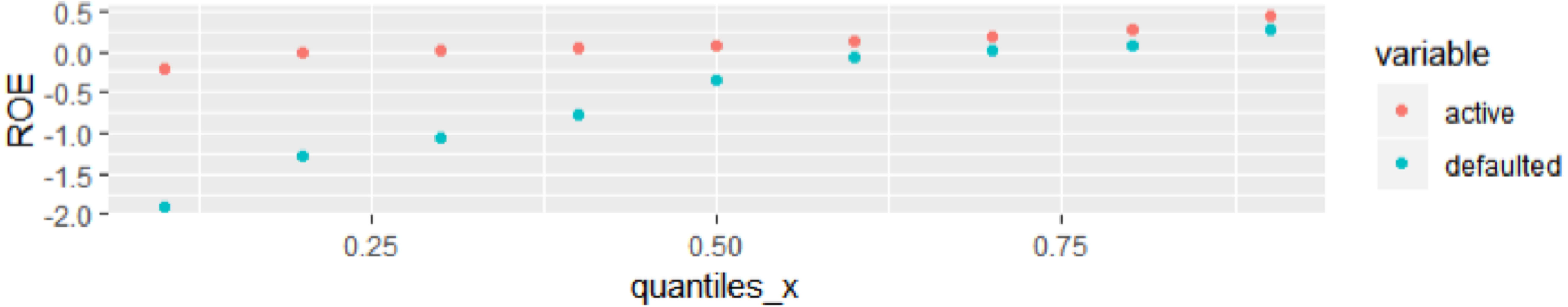
In a few minutes we'll run a code that will display some nice statistics. Here are the basics.

Percentile: represent the values that break the frequency distribution into parts, of preset frequencies or percentages. Of particular interest are the quartiles, which correspond to the values which divide the distribution into four equal parts*.

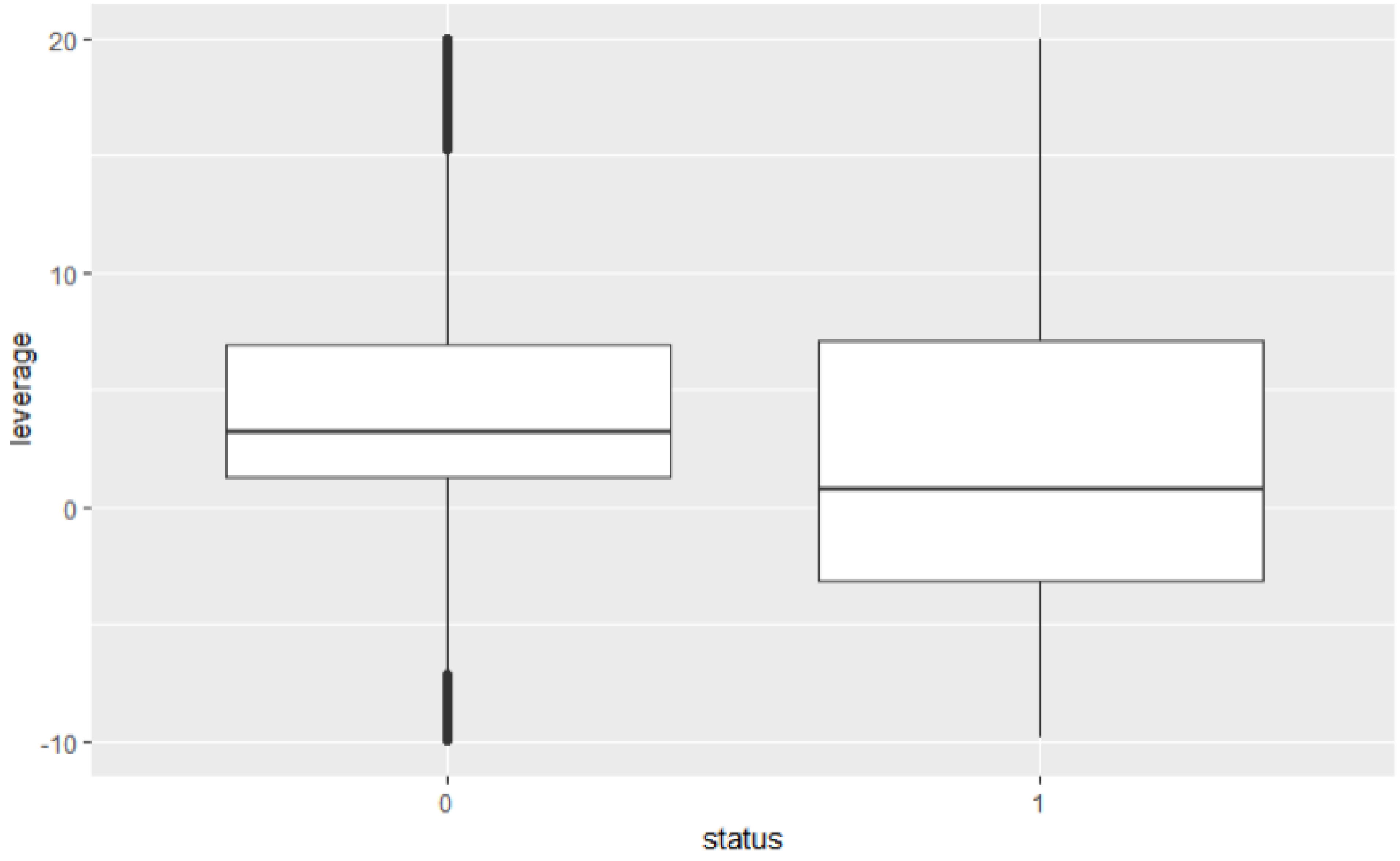


Applied Data Mining for Business and Industry, Second Edition Paolo Giudici and Silvia Figini © 2009 John Wiley & Sons, Ltd.

Percentile distribution example



BoxPlots: Box plots are very useful data visualization tools for depicting many different summary statistics and especially for graphically comparing multiple data sets. On each box, the central mark is the median, the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme data points not considered outliers, and outliers are plotted individually.



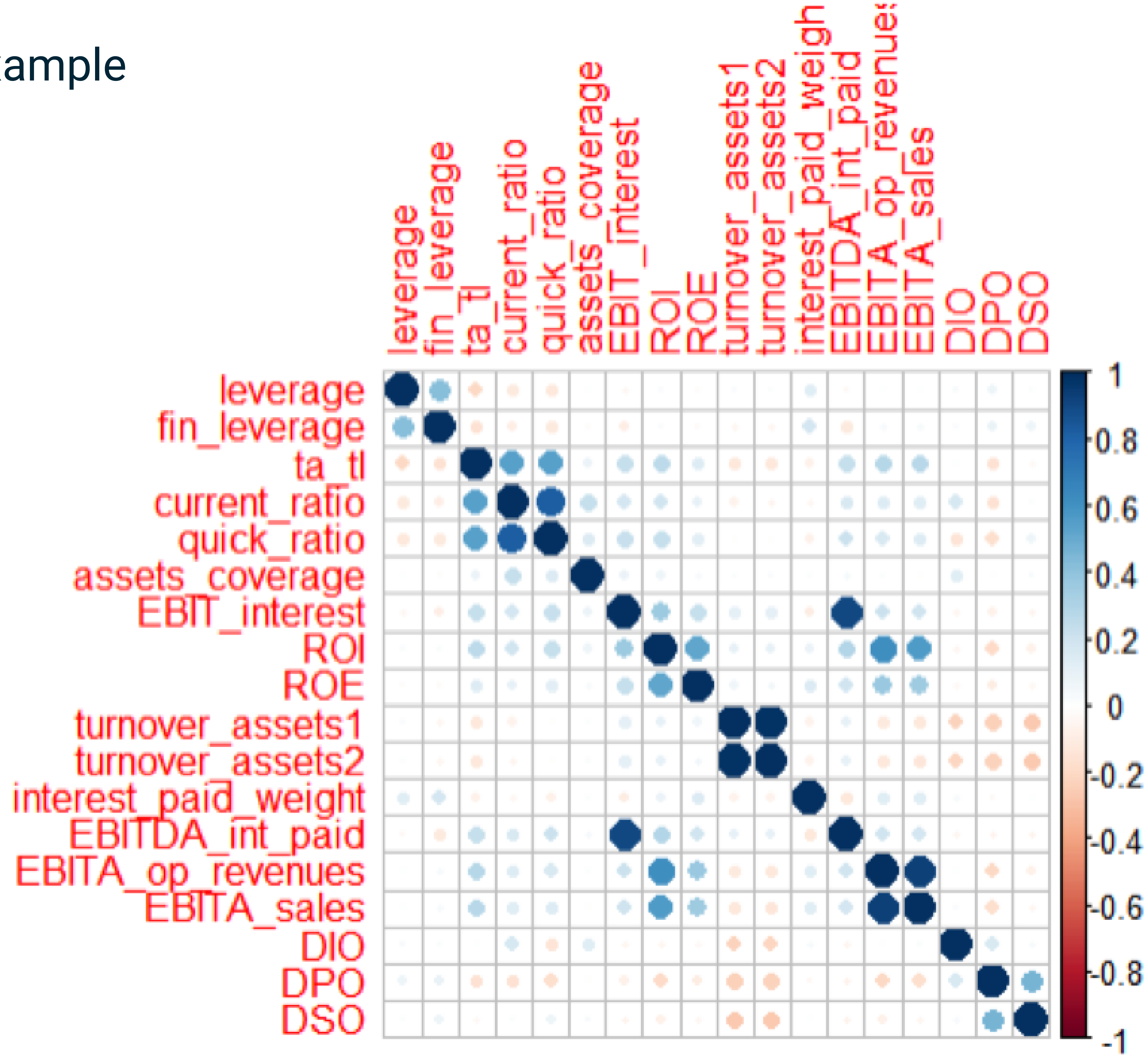
Correlation: The correlation coefficient is a measure of the direction and strength of the linear dependence between two continuous variables.

It is calculated using the covariance (*A*,*B*) and the standard deviation of both the *A* and *B* variables.

$$\rho(A, B) = \frac{\mathrm{cov}(A, B)}{\sigma_A \sigma_B}$$

data analysis: mathematical basics

Correlation matrix example



In the afternoon session we'll go through some credit scoring algorithms, namely:

- Logistic regression;
- Tree models;
- Random forests.

Before going hands on within those algorithms, we'll go through some theory underneath each of them.

afternoon session

model creation

There are several factors/shrewdness to take into account when developing/applying a scoring model, the following are probably the most important:

1) Divide the dataset in two sub samples: one for training/development of the model, the other one for validation;
2) Model’s choice: depending on the needs and the available data, different models have different pros and cons;
3) -Out of sample- validation of the model: there are several techniques to validate a model, the most common applied in credit scoring are represented by CAP/ROC curves;
4) Perform some statistics on the output, focus on “extreme” outcomes: they should not derive from “factual errors” and have to be somehow justifiable.

dataset division

- It’s very important that the model is not back-tested on the same data it has been developed (**Out of sample validation vs In sample validation**).
- Usually it results to be fair enough to split the dataset into **70-30** subsets, first for development, second for validation.
- Out of sample validation will confirm (or not) the final quality of the model, spotting potential overfitting.

model's choice

We've seen there are different "kind" of scoring models: heuristic, Machine learning, causal models, etc.

Depending on need and dataset, one model can represent a better choice respect another one.

For instance, if we apply a simple logistic regression model, we may want to factor qualitative elements, such as Country or Sector ex-ante; meaning developing sub-model for different sectors and/or countries.

If we develop a random forest, we've got to be careful with data availability and how the network reacts to missing data.

validation

Any model needs a thorough validation aimed to understand the model performance.

The key elements of a scoring model are:
- **Discriminating power of the model**: verify that the model discriminates "healthy" companies from bankrupt companies;
- **Bankruptcy dynamics**: for bankrupt companies, the assigned score deteriorates, approaching the default date.

Apart from the said validation, which is carried on “massively”, developers have to analyze carefully also “extreme” output, eventually spotting exceptions that lead to unwanted results, or to verify that extreme results are always “justifiable”.

An answer like *“the model said so”*, is **never acceptable** by the end user of the model in case of anomalies.

This last task is harder to carry on when dealing with Machine Learning models.

LET'S SEE THIS !

THANK YOU

FINTECH RISK MANAGEMENT
www.fintech-ho2020.eu

Andrea Sorrentino, Head of Fintech

modefinance

European Commission