Scoring

M2 D3S/EGR 2025-2026

Louis Olive

louis.olive@gmail.com / louis.olive@ut-capitole.fr

September 10, 2025

Outline

Outline

- Course logistics
- Corporate bankruptcies
- Scoring
- (Warm Up) Desbois case study

Course logistics

Course logistics

Logistics/Prerequisites

- Material available/uploaded at https://github.com/louis-olive/teaching_Scoring and partially on Moodle
- Email: louis.olive@gmail.com / louis.olive@ut-capitole.fr (copy both with [TSE - SCORING] as object)
- Personal laptop or university desktop / Running RStudio using Quarto or R Markdown for reports
- Basic knowledge of math, statistics / Good command of R (if not you'll have to learn it throughout the course)

Course logistics

Agenda

- Part 1 (9h): Methods

  Wed 10 Sep / Tue 16 Sep / Wed 24 Sep / Wed 1 Oct / Mon 6 Oct / Wed 8 Oct

- Part 2 (15h): Case study / project by group of students

  Mon 13 Oct / Wed 15 Oct / Mon 27 Oct / Wed 29 Oct / Mon 3 Nov / Wed 5 Nov / Wed 12 Nov / Mon 17 Nov / Wed 19 Nov / Wed 26 Nov

Throughout the course, the emphasis is put on numerical examples with a hands on / code from scratch approach.

Codes are given in R.

Course logistics

Grading

- Part 1 (8 Oct: 1.5 hour in-class exam / additional take home due 19 Oct):

  In-class data analysis on a real data set including data exploration and applications from the course, followed by a take-home assignment with additional tasks (report with text + R code, 40%).

- Part 2 (19 Oct group composition / 26 Nov: in-class presentation of your project progress, should be almost finished / final report due date: TBD but before Christmas):

  Case study / project per group of students (in-class presentation + report, 60%).

Course logistics

Agenda - Part 1 (9h): Methods

- Statistical learning goals and vocabulary are defined in the context of scoring and classification (10 Sep).

- Introduction and study of two workhorses:
  - Logistic Regression: model fit, inference, selection, penalization (16/24 Sep);

  - Decision Trees for Classification / (Gradient) Boosting (1/6 Oct).

Course logistics

Agenda - Part 2 (15h): Case study / project by group

- Credit risk is one of the various application of scoring.
- Case study using a real life data set, you will be given a set of tasks to perform in group:
  - Defining the problem / Building the data set [Getting/Cleansing/Enriching Data] / Exploration / Baseline model / Model evaluation pipeline
  - Implementing Logistic Regression and Tree based methods
  - “Expert” approach: Implementing ideas from academic literature on credit risk (involving in particular feature engineering)
  - “Off the shelf” approach: lasso for variable selection, gradient boosting
  - (Optionnal/Time permitting) more advanced topics: dealing with data imbalance, model calibration, interpretable ML (SHAP/PDP)
- A final presentation (group/one-to-one short meeting sketching the highlights your project, 26 Nov). The writing of a final report implementing/documenting your analysis (due date, TBD before Christmas).

Course logistics

Material

- **Logistic regression/Scoring:**

  Hosmer D.W., Lemeshow S. and Sturdivant R.X. (2013). Applied Logistic Regression, Wiley.

  Cornillon P.A., Matzner-Løber E., Rouvière L. (2019). Régression avec R, Springer (in French). I borrow a lot from this one.

- **Statistical Learning/Decision Trees/Boosting:**

  Hastie T., Tibshirani R. and Friedman. J. (2009). The Elements of Statistical Learning: Data Mining, Inference and Prediction, Springer. available here

  Lindholm, A. and Wahlström, N., Lindsten, F. and Schön, T. (2022). Machine Learning - A First Course for Engineers and Scientists, Cambridge University Press. latest draft available here

  Murphy, K. P. (2022). Probabilistic Machine Learning: An introduction, MIT Press. available here or here

Corporate bankruptcies

Corporate bankruptcies - US - 2024

Newspaper headline

US corporate bankruptcies hit 14-year high as interest rates take toll

Consumer companies such as Party City hit by persistent inflation and weak spending

Party City — once the largest supplier of balloons and other fun-time supplies in the US — filed for bankruptcy last month © David Paul Morris/Bloomberg

Will Schmitt in New York JANUARY 7 2025 137

US corporate bankruptcies have hit their highest level since the aftermath of the global financial crisis as elevated interest rates and weakened consumer demand punish struggling groups.

Source: Financial Times, January 7 2025

Corporate bankruptcies - US - 2024

Newspaper headline

Aux Etats-Unis, les faillites d'entreprises à leur plus haut depuis 2010

Le nombre de dépôts de bilan a augmenté de 8 % sur un an en 2024, à son plus haut niveau depuis la dernière crise financière mondiale. Les taux d'intérêt élevés et la faible demande mettent en difficulté de nombreux groupes.

Le vendeur d'articles de fête Party City a mis la clé sous la porte en 2024, fermant ses 700 magasins aux Etats-Unis. (Richard B. Levine/Newscom/SIPA)

Par **Les Echos**
Publié le 7 janv. 2025 à 07:55 | Mis à jour le 7 janv. 2025 à 15:26

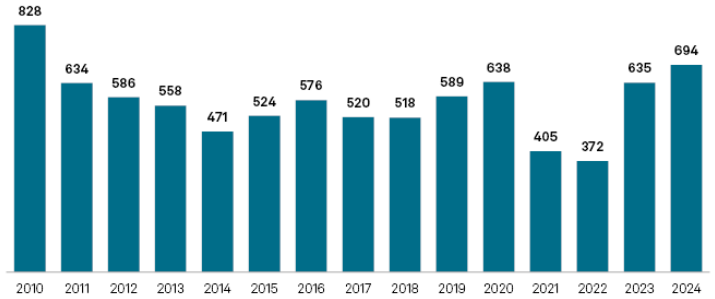
C'est un signal inquiétant pour l'économie américaine et mondiale, à quelques jours du retour de Donald Trump à la Maison-Blanche. Les Etats-Unis ont enregistré en 2024 un nombre de faillites d'entreprise à son plus haut niveau depuis 2010 et la crise financière qui avait alors durement affecté l'économie mondiale, selon des données de « S&P Global Market Intelligence », relayées par le « Financial Times ».

Source: Les Echos, January 7 2025

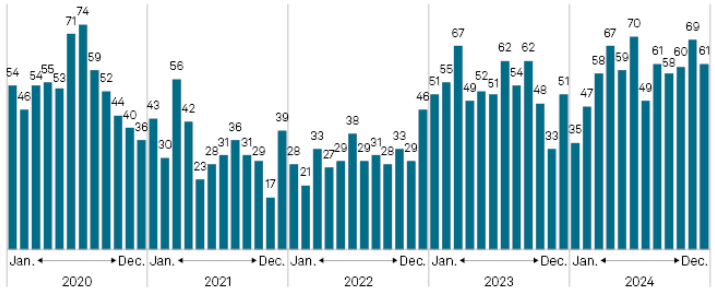
Corporate bankruptcies - US - 2024

Newspaper source

US bankruptcy filings by year*



US bankruptcy filings by month**



Data compiled Jan. 1, 2025.
* Includes US companies covered by S&P Global Market Intelligence that announced a bankruptcy between Jan. 1, 2010, and Dec. 31, 2024.
** Includes US companies covered by S&P Global Market Intelligence that announced a bankruptcy between Jan. 1, 2020, and Dec. 31, 2024.
S&P Global Market Intelligence's bankruptcy coverage is limited to public companies or private companies with public debt where either assets or liabilities at the time of the bankruptcy filing are greater than or equal to $2 million, or private companies where either assets or liabilities at the time of the bankruptcy filing are greater than or equal to $10 million. Involuntary bankruptcy filings are also included.
Source: S&P Global Market Intelligence.

Source: S&P Global Market, January 7 2025

Corporate bankruptcies - US - 2025 YTD

Update

63 US corporate bankruptcies in June set up 2025 for highest pace since 2010

By **Sean Longoria** and **Umer Khan**

The fast pace of monthly US corporate bankruptcies extended into June and put 2025 on track to be one of the busiest years for filings in more than a decade.

S&P Global Market Intelligence recorded 63 new bankruptcy filings from certain public and private companies in June, down from a revised count of 64 in May. The data includes companies with public debt and assets or liabilities of at least $2 million or private companies with assets or liabilities of at least $10 million at the time of filing.

Still, 371 bankruptcy filings have been recorded throughout 2025, the highest total for the first half of the year since 2010.

Corporate liquidity has largely worsened in 2025 as debt levels for many companies have risen and the US Federal Reserve is poised to hold benchmark interest rates at their current level through the summer. Consumer spending, meanwhile, is straining under the weight of a cooling job market, inflation still above monetary policymakers'

Source: S&P Global Market, July 8 2025

Corporate bankruptcies - US - 2025 YTD

Update

06 AUG, 2025

July US corporate bankruptcy filings hit highest monthly total in 5 years

By **Nick Lazzaro** and **Annie Sabater**

US corporate bankruptcies in July reached their highest monthly volume since 2020.

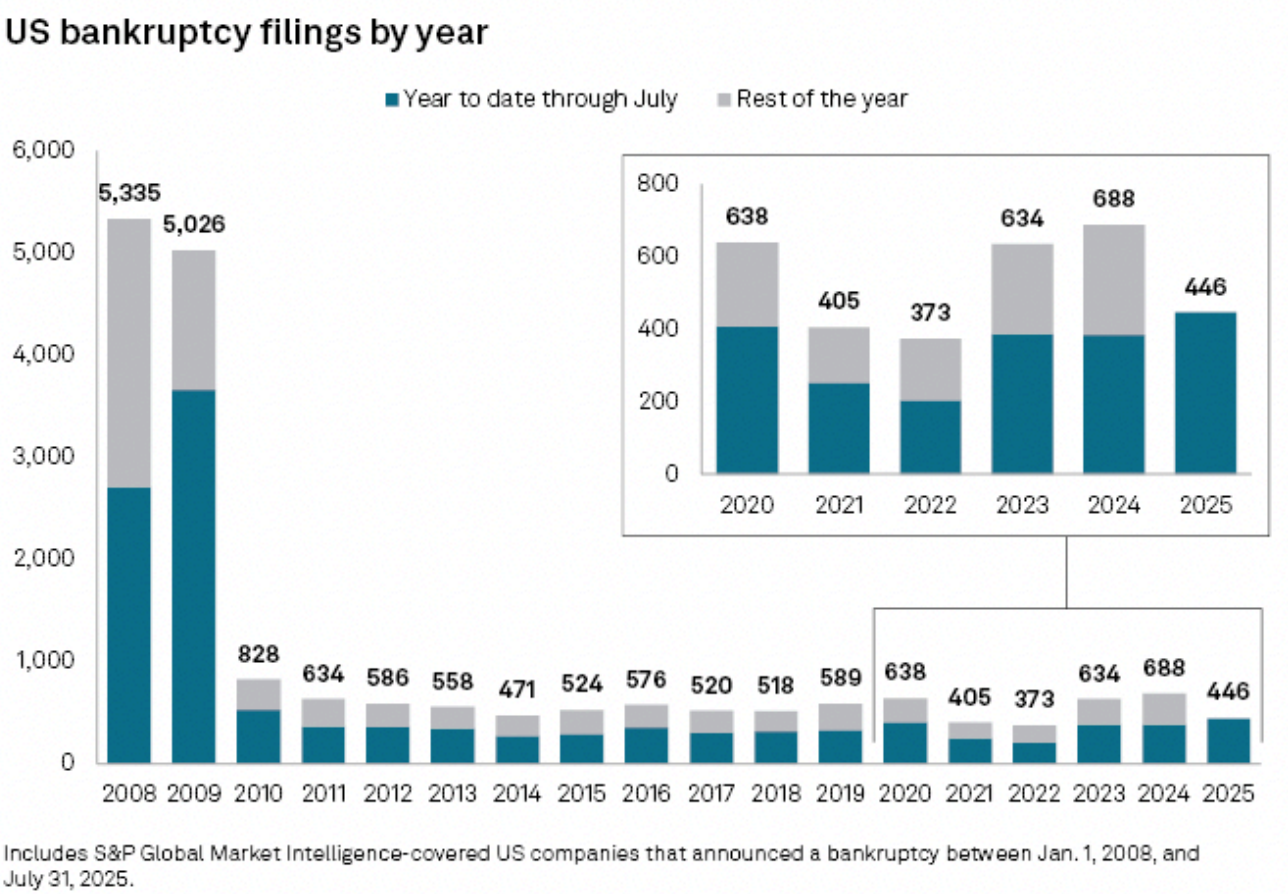
Large public and private company bankruptcy filings increased to 71 in July from a revised 66 in June and the highest single-month tally since July 2020, according to S&P Global Market Intelligence data. Year-to-date bankruptcy filings totaled 446 through the end of July, the most for this seven-month period since 2010. The data includes companies with public debt and assets or liabilities of at least $2 million or private companies with assets or liabilities of at least $10 million at the time of filing.

Companies are contending with elevated interest rates as uncertainty from US tariff policy pressures costs and supply chain resilience. The US Federal Reserve in July chose to hold its benchmark interest rate for the seventh-straight month at 4.25%–4.50%. Yet markets now project a nearly 90% probability that the central bank will issue a 25-basis-point rate cut in September, following downward revisions to job growth in an August labor report, according to CME FedWatch data Aug. 5.
Companies could benefit from the rate cut if such a move impacts US Treasury yields or market sentiment.

Source: S&P Global Market, August 6 2025

Corporate bankruptcies - US - 2025 YTD

Update



Includes S&P Global Market Intelligence-covered US companies that announced a bankruptcy between Jan. 1, 2008, and July 31, 2025.

Source: S&P Global Market, August 6 2025

Corporate bankruptcies - FR - 2024

Newspaper headline

Record de défaillances d'entreprises en France, pas d'éclaircie en vue

Quelque 66.420 entreprises françaises ont fait défaut en 2024, selon le décompte provisoire du groupe BPCE. Cela représente environ 260.000 emplois qui se sont retrouvés menacés. Ce record pourrait encore être dépassé cette année.

Fin novembre, la maison mère du Coq Sportif a annoncé que l'équipementier sportif a été placé en redressement judiciaire par le tribunal de commerce de Paris. (FRANCK FIFE/AFP)

Par **Nathalie Silbert**
Publié le 8 janv. 2025 à 10:39 | Mis à jour le 8 janv. 2025 à 15:59

« L'année 2024 est la pire année que l'on ait connue depuis 2010 en termes de défaillances d'entreprises », résume Alain Tourdjman, directeur des études économiques du groupe bancaire BPCE. Selon le décompte provisoire de l'établissement, 66.422 entreprises françaises ont fait défaut l'an dernier, un niveau qui n'avait pas été atteint même au lendemain de la crise financière de 2008.

Source: Les Echos, January 8 2025

Corporate bankruptcies - FR - 2024

Newspaper source

Les défaillances continuent de progresser au 4e trimestre 2024

… et atteignent un plus haut niveau depuis au moins 15 ans

⊙ Un tournant depuis la fin 2023 :

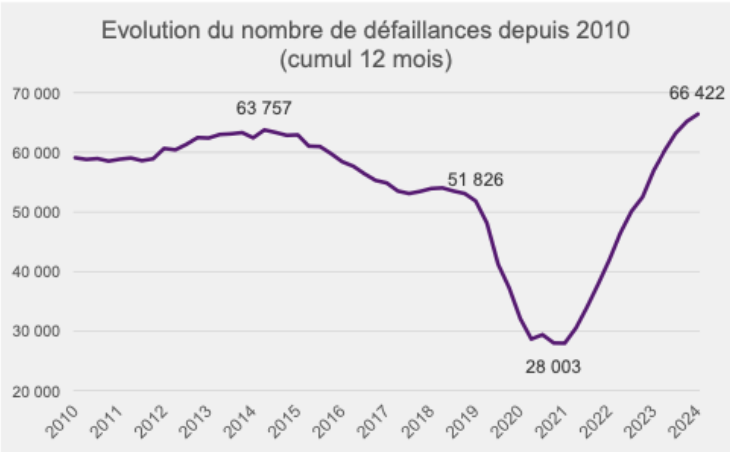
- Ralentissement économique (sans récession)
- Impact inflationniste sur les marges et l'activité
- Hausse des taux : financement plus difficile
- Remboursement des PGE
- Reprise des recouvrements Urssaf depuis un an

➔ **Un record depuis 2010 mais un nombre global à relativiser…**

➔ **avec un rattrapage limité des défaillances évitées de 2020 à 2022…**

➔ **Mais de nombreux points d'alerte !**

66 422 **défaillances d'entreprises en France en 2024 (+28% par rapport à 2019)**



Source : BPCE L'Observatoire

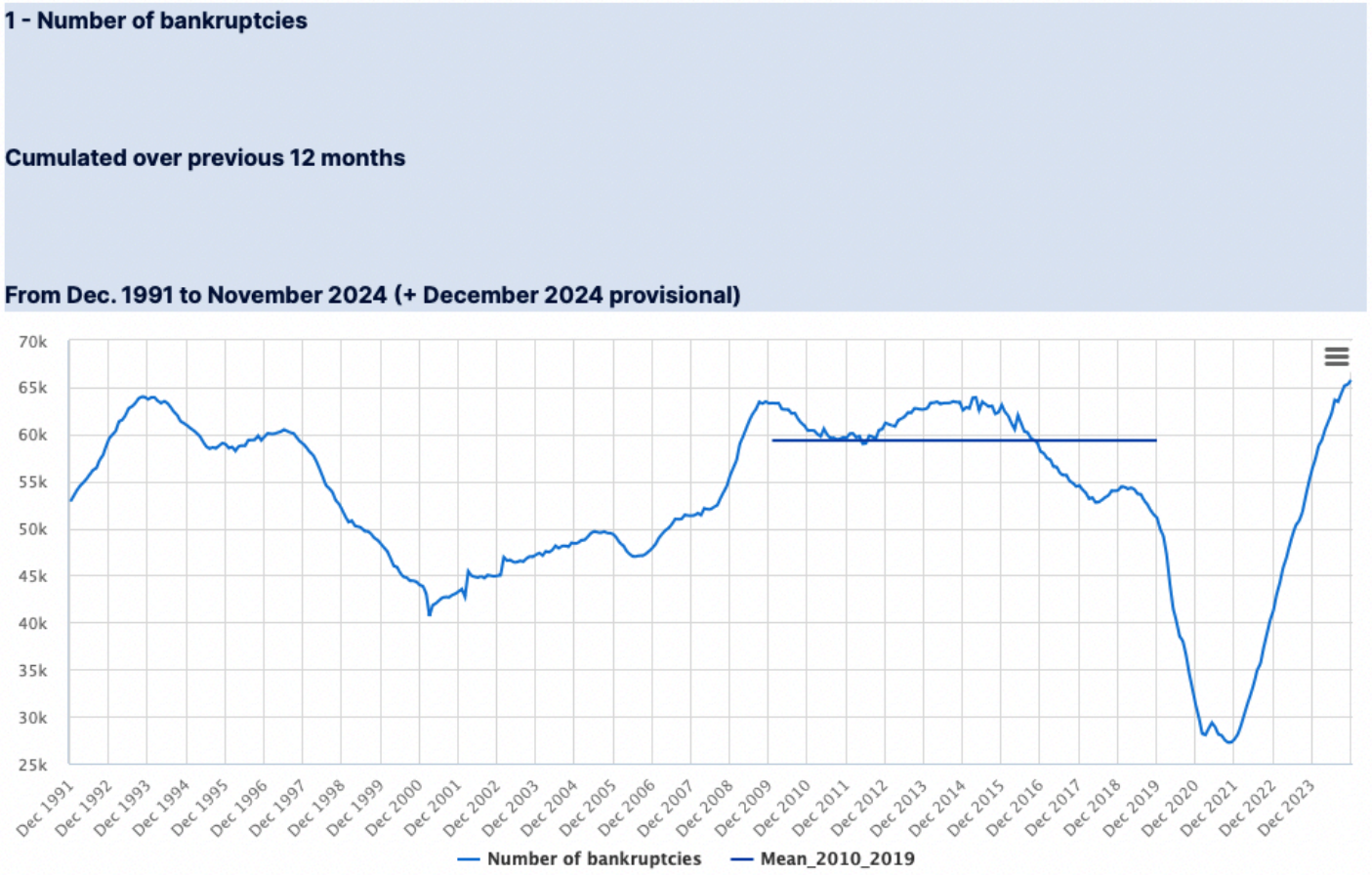
11 • BPCE L'OBSERVATOIRE – DÉFAILLANCES D'ENTREPRISES EN FRANCE : BILAN 2025 ET PERSPECTIVES

GROUPE BPCE

Source: BPCE L'Observatoire, January 8 2025

Corporate bankruptcies - FR - 2024

Alternative source



Source: Banque de France, January 17 2025

Corporate bankruptcies - FR - 2024

Alternative source

BANQUE DE FRANCE
EUROSYSTÈME

Contact us | Find us | Language: EN

Business failures - France - 2024-12

Published on the 17th of January 2025

Lastest publications

Business failures - France - 2025-07
Business failures - France - 2025-06
Business failures - France - 2025-05

All publications

statistics

At the end of December, the year-on-year rise in the number of corporate bankruptcies continues to slow down

At the end of December 2024, the cumulative number over twelve months of corporate bankruptcies (provisional data) stood at 65,764, compared with 65,298 at the end of November (see graph 1).

This increase is partly due to a catch-up effect, following the sharp slowdown in insolvencies during the covid period (2020-2021).

In December, however, the year-on-year rise in insolvencies was lower than in November (16.8% vs. 18.8% in November) across all sectors and most company sizes (see tables A, B and graph 2). The number of corporate bankruptcies among intermediate-sized enterprises (ISEs) and large companies has stabilized, but remains higher than its pre-pandemic average.

The public services and those of the Banque de France are mobilized to help companies in difficulty.

To find out more: data on business start-ups are reported by the French National Institute of Statistics and Economic Studies (INSEE): ***Business births on the INSEE's website***

Source: Banque de France, January 17 2025

Corporate bankruptcies - FR - 2024

Alternative source

A - Corporate bankruptcies by sector W

Bankruptcies in number of legal units, year on year change (%)

	Aggregate over previous 12 months (a) (raw data)							
	Mean 2010-2019	Nov. 24	Nov. 24/Nov. 23	Nov. 24/2010-2019	Dec. 23	Dec. 24 prov.	Dec. 24/Dec. 23	Dec. 24/2010-2019
Business sector								
Agriculture, forestry and fishing (AZ)	1,359	1,454	16.0%	7.0%	1,308	1,386	6.0%	2.0%
Industry (BE)	4,442	4,237	8.9%	-4.6%	3,934	4,249	8.0%	-4.3%
Construction (FZ)	14,684	14,433	25.8%	-1.7%	11,816	14,743	24.8%	0.4%
Wholesale and retail trade; repair of motor vehicles and motorcycles (G)	13,071	13,726	16.1%	5.0%	12,073	13,790	14.2%	5.5%
Transportation and storage (H)	1,901	2,953	34.2%	55.3%	2,287	2,989	30.7%	57.2%
Accommodation and food service activities (I)	7,374	8,503	11.0%	15.3%	7,819	8,557	9.4%	16.0%
Information and communication (JZ)	1,480	1,955	16.5%	32.1%	1,720	1,991	15.8%	34.5%
Financial and insurance activities (KZ)	1,150	1,683	31.2%	46.3%	1,330	1,688	26.9%	46.8%
Real estate activities (LZ)	1,984	2,586	39.3%	30.3%	1,959	2,570	31.2%	29.5%
Advisory & Business support activities (MN)	6,379	7,784	21.5%	22.0%	6,541	7,859	20.1%	23.2%
Education, human health and social work services, Arts, entertainment and recreation, Other service activities (P to S)	5,311	5,894	10.2%	11.0%	5,457	5,850	7.2%	10.1%
All firms (b)	**59,342**	**65,298**	**18.8%**	**10.0%**	**56,313**	**65,764**	**16.8%**	**10.8%**

Source: Banque de France - database: Fiben. Data available early January 2025: final for November 2024, provisional for December 2024.

Source: Banque de France, January 17 2025

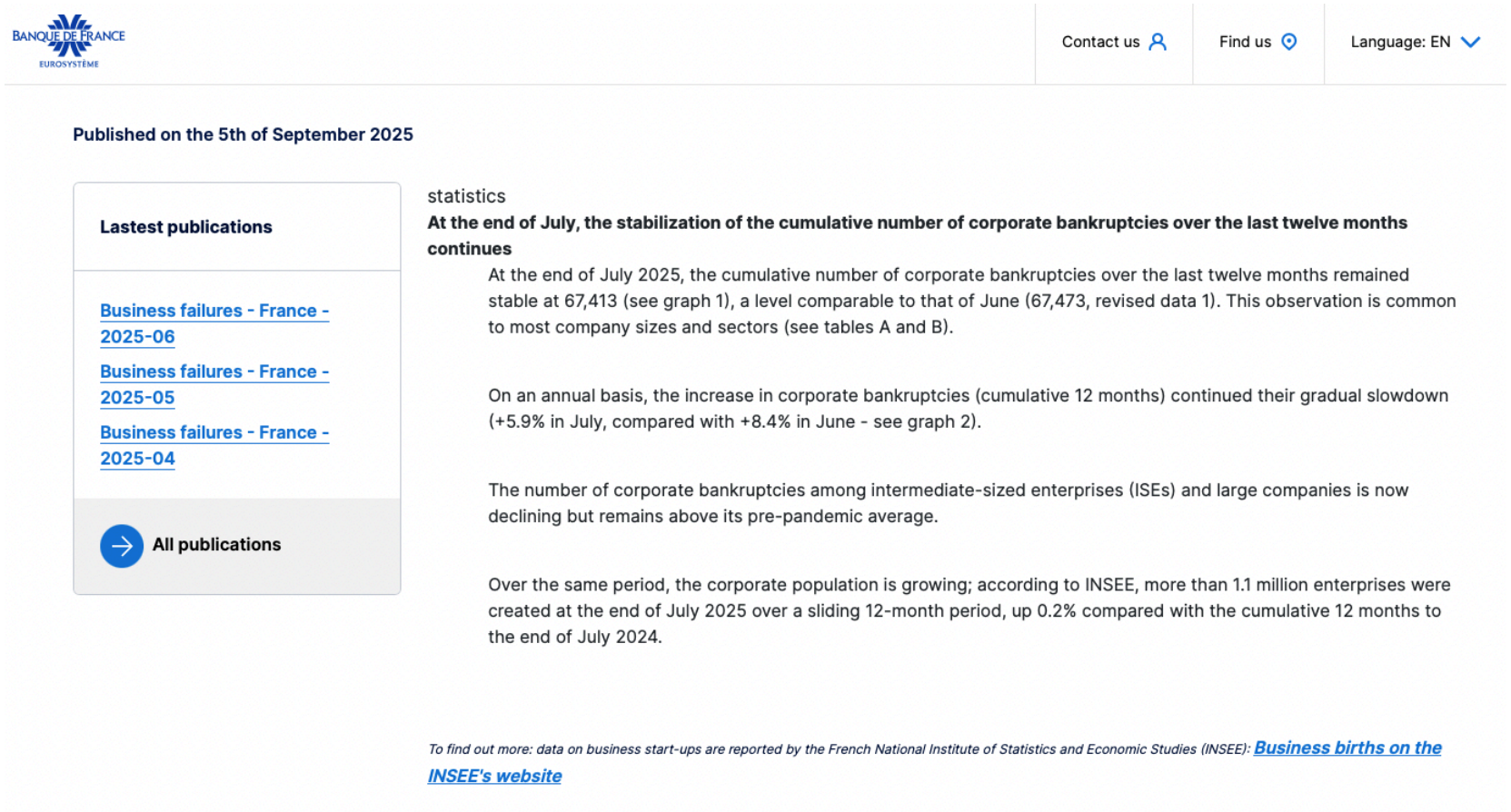
Corporate bankruptcies - FR - 2025 YTD

Alternative source

Source: Banque de France, September 5 2025

Corporate bankruptcies - FR - 2025 YTD

Alternative source



BANQUE DE FRANCE
EUROSYSTÈME

Contact us | Find us | Language: EN

Published on the 5th of September 2025

Lastest publications

Business failures - France - 2025-06

Business failures - France - 2025-05

Business failures - France - 2025-04

All publications

statistics

At the end of July, the stabilization of the cumulative number of corporate bankruptcies over the last twelve months continues

At the end of July 2025, the cumulative number of corporate bankruptcies over the last twelve months remained stable at 67,413 (see graph 1), a level comparable to that of June (67,473, revised data 1). This observation is common to most company sizes and sectors (see tables A and B).

On an annual basis, the increase in corporate bankruptcies (cumulative 12 months) continued their gradual slowdown (+5.9% in July, compared with +8.4% in June - see graph 2).

The number of corporate bankruptcies among intermediate-sized enterprises (ISEs) and large companies is now declining but remains above its pre-pandemic average.

Over the same period, the corporate population is growing; according to INSEE, more than 1.1 million enterprises were created at the end of July 2025 over a sliding 12-month period, up 0.2% compared with the cumulative 12 months to the end of July 2024.

To find out more: data on business start-ups are reported by the French National Institute of Statistics and Economic Studies (INSEE): ***Business births on the INSEE's website***

Source: Banque de France, September 5 2025

Corporate bankruptcies

Some definitions

FR: “défaillance”

Défaillance d'entreprise

Définition

Une unité légale est en situation de défaillance ou de dépôt de bilan à partir du moment où une procédure de redressement judiciaire est ouverte à son encontre.
Cette procédure intervient lorsqu'une unité légale est en état de cessation de paiement, c'est-à-dire qu'elle n'est plus en mesure de faire face à son passif exigible avec son actif disponible.

Remarque

Il ne faut pas confondre la notion de défaillance et la notion de cessation. La notion de cessation correspond à l'arrêt total de l'activité économique d'une entreprise. Toutes les défaillances ne donnent pas des cessations. Par exemple, un jugement d'ouverture de procédure de défaillance (dépôt de bilan d'une entreprise inscrite dans le cadre d'une procédure judiciaire) ne se résout pas forcement par une liquidation.
Toutes les cessations n'ont pas donné lieu à une défaillance. Par exemple, un entrepreneur individuel peut cesser son activité suite à un départ en retraite.

Enterprise bankruptcy

Definition

A business is in a situation of failure or filing for bankruptcy from the moment when a judicial settlement procedure is opened against it.
This procedure occurs when a legal unit has suspended payments, that is, when it is no longer capable of covering its current liabilities with its available assets.

Note

The notion of failure must not be confused with the notion of death of the enterprise. The notion of death corresponds to the total stoppage of the company's economic activity. Not all failures lead to a death. For instance, a judgement opening failure proceedings (filing for bankruptcy of a company as part of judicial proceedings) does not necessarily culminate in liquidation.
Not all deaths are linked with a failure. For instance, a sole proprietorship can stops his activity because of retirement.

US: Chapter 11 bankruptcy / Chapter 7 bankruptcy

Corporate bankruptcies

A complex phenomenon

> Indeed, corporate failure is a complex phenomenon for which the actual causal variables are difficult to access. The score functions therefore make use of symptoms such as descriptors of the company's situation before its failure. In other words, it is impossible to accurately define companies' failure processes, contrary to what occurs in other fields of application of discriminant analysis that are closer to physical science, such as shape recognition, where overlearning is easier to master and techniques such as neural networks are successfully applied.

Source: (Bardos, 2007)

Scoring

Scoring

Supervised learning - a reminder

This an informal and very rapid overview of Statistical Learning, where we introduce concepts and definitions.

For a rigorous introduction see for example the dedicated lesson from Sébastien Gadat - M1 course on Mathematical Statistics or the book by Francis Bach - Learning Theory from First Principles.

Scoring

Supervised learning - Inputs-Outputs

We follow here the terminology of Hastie et al. (2009).

The problem: we are given a data set where some variables, denoted as **inputs**[1], have some influence on one or more **output(s)**[2].

We will denote inputs by the symbol X having values in $\mathbf{X}$ (to explicit things we will consider $\mathbf{X}$ is $\mathbb{R}^p$).

If X is a vector, its components can be accessed by subscripts X_j, $j = 1, \ldots, p$.

Outputs will be denoted by Y having value in $\mathbf{Y}$.

1. In statistics, inputs are often called the predictors, factors or the independent variables. In machine learning the term features is also used.
2. The outputs are also called responses, labels or dependent variables.

Scoring

Supervised learning - Framework

Outputs are denoted by Y having value in $\mathbf{Y}$:

- When $\mathbf{Y}$ is $\mathbb{R}$, it is a **regression problem**;
- When $\mathbf{Y}$ is a discrete set, it is a **classification problem** (for example $\mathbf{Y}$ is $\{0, 1\}$ for binary classification). In this course we will restrict to the case of **binary classification**.

We need data to construct predictions or classifications. We suppose we have available a set of observed data:

$(x_i, y_i) \in \mathbf{X} \times \mathbf{Y}$, $i = 1, \cdots, n$, known as the **learning** or **training** set, with which to construct our prediction.

We assume $(x_i, y_i) \in \mathbf{X} \times \mathbf{Y}$ are generated i.i.d. from $\mathbf{P}_{\mathbf{X}\times\mathbf{Y}}$ where we denote $\mathbf{P}_{\mathbf{X}\times\mathbf{Y}}$, the data generating distribution (or underlying probability distribution or joint law) which is unknown.

The main goal of supervised learning will be to predict unobserved outputs y given unobserved inputs x using the learning set.

Such new or unobserved data is referred as the **testing** set.

Scoring

Supervised learning - Loss

To achieve that goal we try to find a “best” or optimal **classifier** (or prediction function in general):

$$f : \mathrm{X} \to \mathrm{Y}$$

We first must define some criterion to assess the classifier performance (what is optimal?).

For that purpose we consider a **loss** function $\ell : \mathrm{Y} \times \mathrm{Y} \to \mathbb{R}^+$:

$$\begin{cases} \ell(y, z) > 0 & \text{if } y \neq z \\ \ell(y, z) = 0 & \text{if } y = z \end{cases}$$

$\ell(y, f(x))$ is the loss or cost of predicting $f(x)$ while the true output is y.

In the context of binary classification a natural selection for loss is the 0-1 loss:

$$\ell(y, z) = \mathbf{1}_{y \neq z}$$

We will see later in the course that other losses are used.

Scoring

Supervised learning - Risk

We then define the expected **risk** of a prediction function f given the loss ℓ as:

$$\mathrm{R}(f) = \mathbb{E}[\ell(Y, f(X))]$$

In the context of binary classification $Y \in \{0, 1\}$ and 0-1 loss, we have:

$$\mathrm{R}(f) = \mathbb{P}[Y \neq f(X))]$$

Given a data set $(x_i, y_i) \in \mathbf{X} \times \mathbf{Y}$, we also define the empirical risk of a classifier f as:

$$\hat{\mathrm{R}}(f) = \frac{1}{n} \sum_i \ell(y_i, f(x_i))$$

Scoring

Supervised learning - Bayes classifier

We now define the **Bayes(ian) classifier**[1] $f^* : \mathrm{X} \to \mathrm{Y}$ as a function that achieves the minimal expected risk among all possible functions:

$$\operatorname*{argmin}_f \mathrm{R}(f)$$

The Bayes classifier depends on ℓ and $\mathbf{P}_{\mathrm{X} \times \mathrm{Y}}$ which is generally unknown (otherwise the job would be easy).

The Bayes classifier for the 0-1 loss is:

$$f^*(x) = \begin{cases} 1 & \text{if } \eta(x) \geq \frac{1}{2} \\ 0 & \text{otherwise} \end{cases}$$

where:

$$\eta(x) = \mathbb{P}[Y = 1 | X = x)] = \mathbb{E}[Y = 1 | X = x)]$$

$\eta(x)$ is called the regression function.

1. It is sometimes designed as oracle in the literature.

Scoring

Supervised learning - Theory vs practice

In practice $\mathbf{P}_{\mathbf{X}\times\mathbf{Y}}$ is unknown. What we have is the learning set. The Bayes classifier minimizes the expected risk but is not a function of the learning set. What to do? Conceptually two approaches are used[1]:

- the "probabilistic" approach: we estimate the regression function $\eta(x)$ and use/plug this estimate "inside" the Bayes classifier (Plugin); within this approach mainly two sub-approaches:
  - the "discriminative" approach: we directly model or make an assumption on $\eta(X) = \mathbf{P}_{\mathbf{Y}|\mathbf{X}}$ and estimate it; for example we can make the assumption that $\eta(x)$ is a function of a particular form.
  - the "generative" approach: we model the conditional distribution $\mathbf{P}_{\mathbf{X}|\mathbf{Y}}$ and use Bayes formula to get: $\eta(X) = \frac{\mathbb{P}[X|Y=1)]\mathbb{P}[Y=1)]}{\mathbb{P}[X]} = \frac{\mathbb{P}[X|Y=1)]\mathbb{P}[Y=1)]}{\mathbb{P}[X|Y=1)]\mathbb{P}[Y=1)]+\mathbb{P}[X|Y=0)]\mathbb{P}[Y=0)]}$
- the optimization or "machine learning" approach: finding f minimizing the empirical risk; optimization algorithm or heuristics are used to estimate f; the 0-1 loss is not convex so usually it is replaced by a convex surrogate loss or upper bound; usually re-sampling methods are used to compute empirical risk (f is trained on a training set and empirical risk evaluated on the testing set, f is chosen to minimize empirical risk).

1. See Sébastien Gadat - TSE - Maths of Deep and Machine Learning course here. Other very good references are Philippe Rigollet - MIT - 18.657: Mathematics of Machine Learning here or Erwan Le Pennec - Polytechnique - Introduction to Machine Learning (M2 MSV) slides here. We will see in another course that in the context of the Logistic Regression the two approaches are strongly linked: it can be seen as a probabilistic|discriminative approach or an optimization approach.

Scoring

Toy example - the Mixture data

To illustrate what we have seen we use the **Mixture** data set described in Hastie et al. (2009, pp. 12–16).

The data set is generated as follows (p. 16):

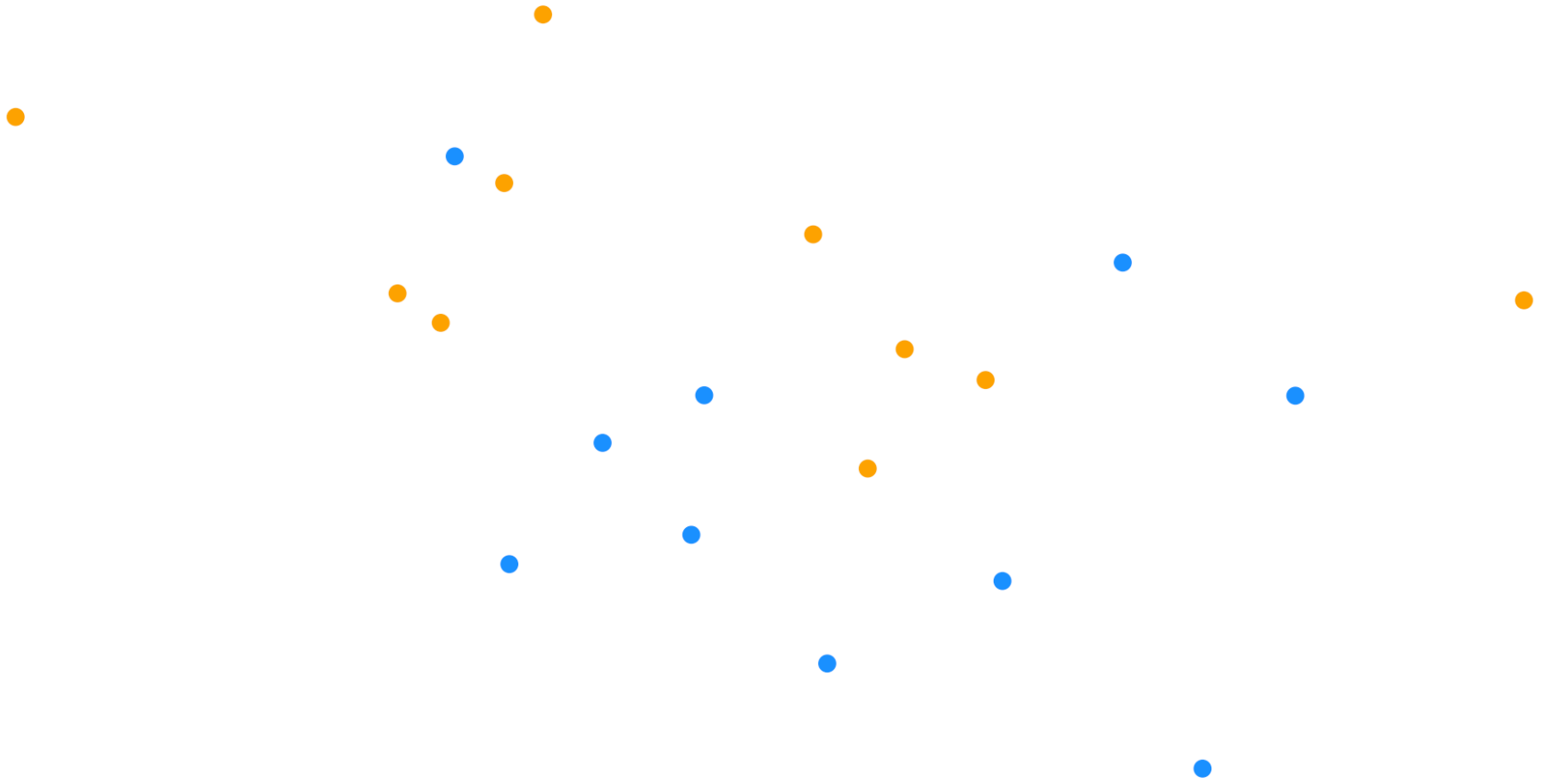
- First the authors generate 10 "means" m_k in $\mathbb{R}^2$ from a bivariate Gaussian distribution $\mathcal{N}((1,0),I)$ and label this class *BLUE*.
- Similarly, 10 more are drawn from $\mathcal{N}((0,1),I)$ and labeled *ORANGE*.
- Then for each class authors generate 100 observations as follows: for each observation, m_k is picked at random with probability $\frac{1}{10}$ and used to generate a $\mathcal{N}(m_k, I/5)$, thus leading to a mixture of Gaussian clusters for each class. They generate similarly an additional data set of 5k observations per class for testing purposes.

The task is to create a classifier that, based on the coordinates x_1 and x_2, determines whether the point is *ORANGE* or *BLUE*.

Scoring

Toy example - the Mixture data

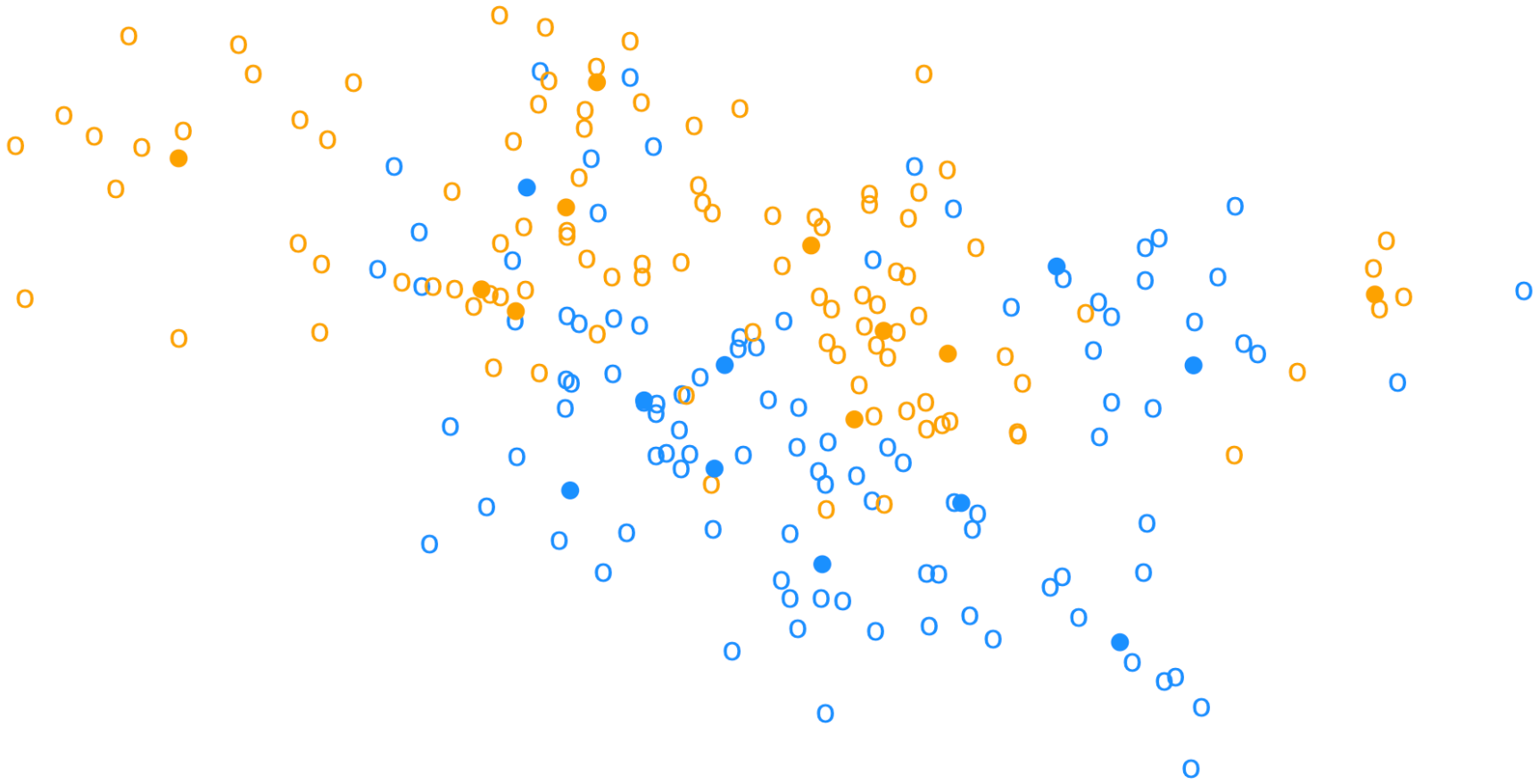
First we load the data set and plot the “means”:



Scoring

Toy example - the Mixture data

Then the training set together with the “means”:

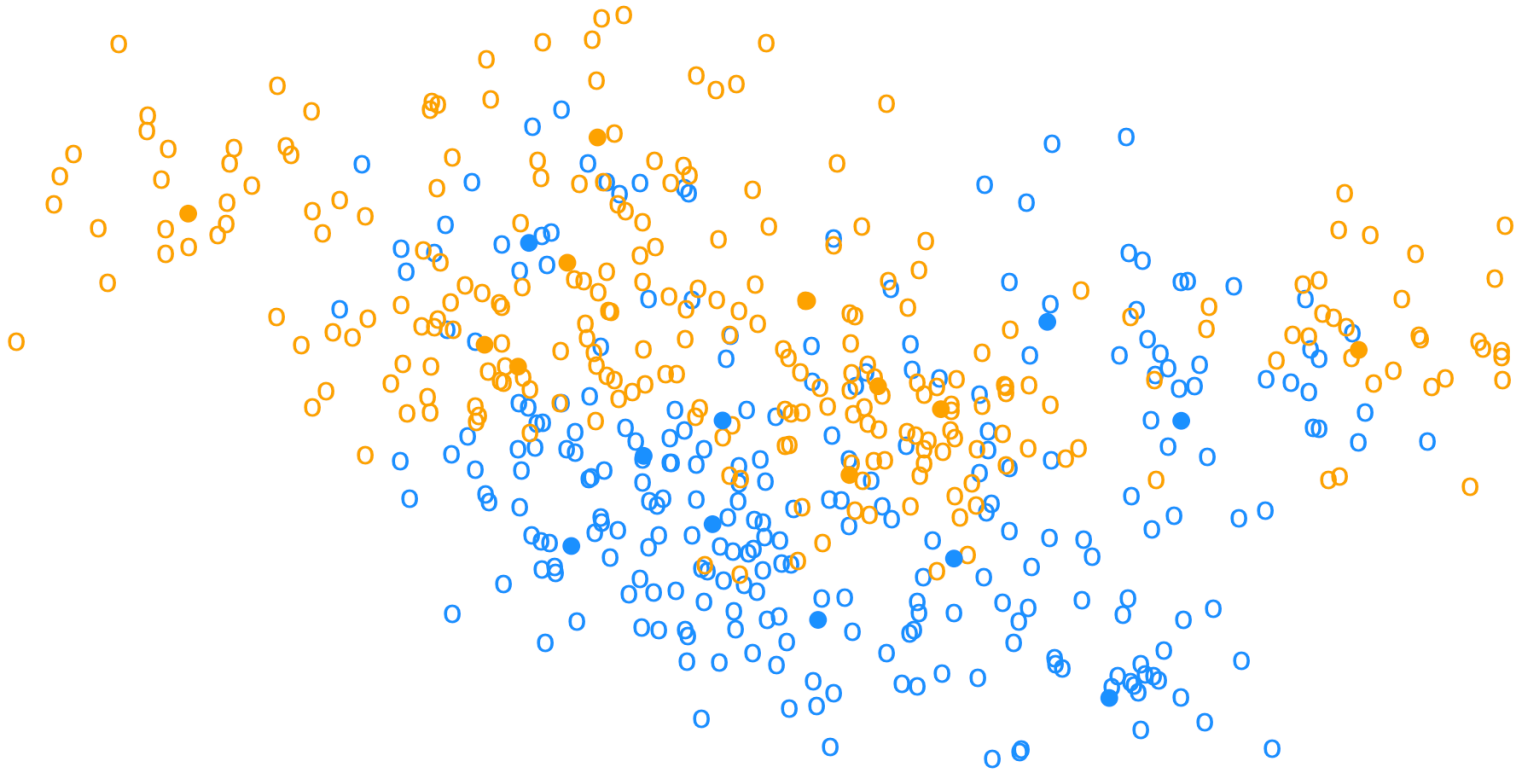


Scoring

Toy example - the Mixture data

Knowing the generative distribution, we generate a testing set of size 10k (half BLUE, half ORANGE).

We plot the first 250 new observations of each class from the generated testing set:

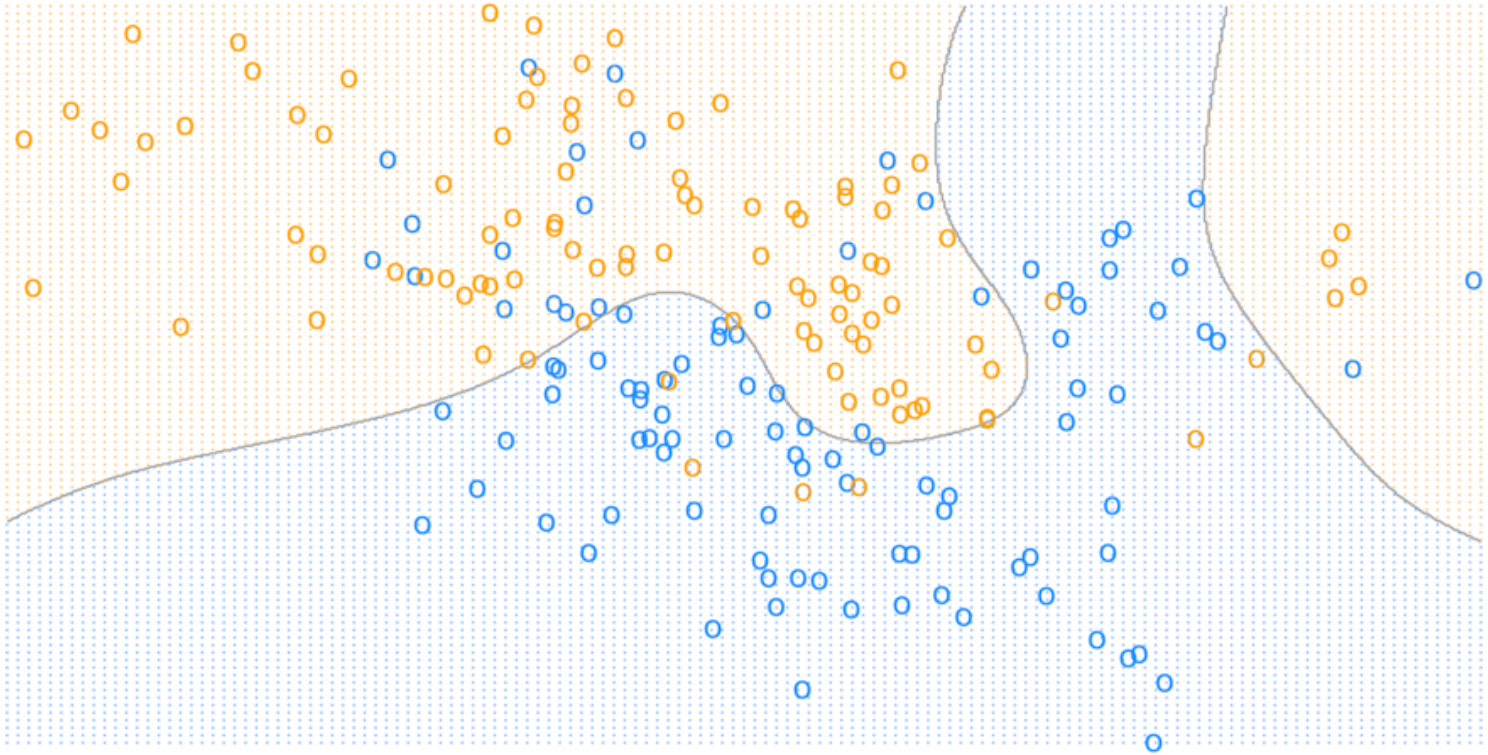


Scoring

Toy example - the Mixture data

Knowing the generating distribution we can derive the Bayes classifier (exercise 2.2 (Hastie et al., 2009)).

We plot the boundary decision in grey:



Scoring

Toy example - the Mixture data

We estimate Bayes risk on the testing data set simulated before (10k BLUE/ORANGE dots) using the data generating process $\mathbf{P}_{\mathbf{X} \times \mathbf{Y}}$.

We get a value of 0.213 for the "estimated" Bayes risk, which is in line with (Hastie et al., 2009, Figure 13.4), see below:

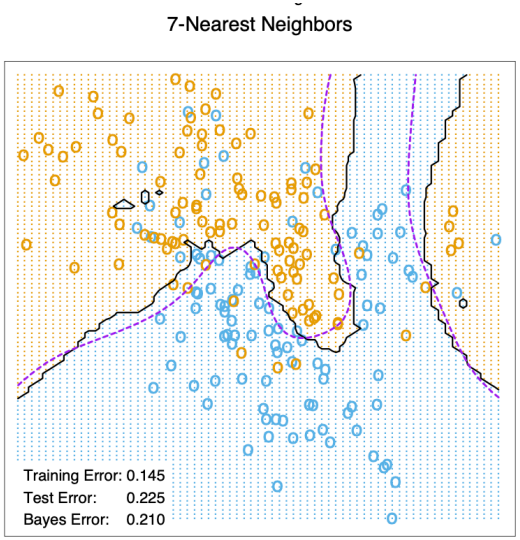


FIGURE 13.4. *k-nearest-neighbors on the two-class mixture data. The upper panel shows the misclassification errors as a function of neighborhood size. Standard error bars are included for 10-fold cross validation. The lower panel shows the decision boundary for 7-nearest-neighbors, which appears to be optimal for minimizing test error. The broken purple curve in the background is the Bayes decision boundary.*

Scoring

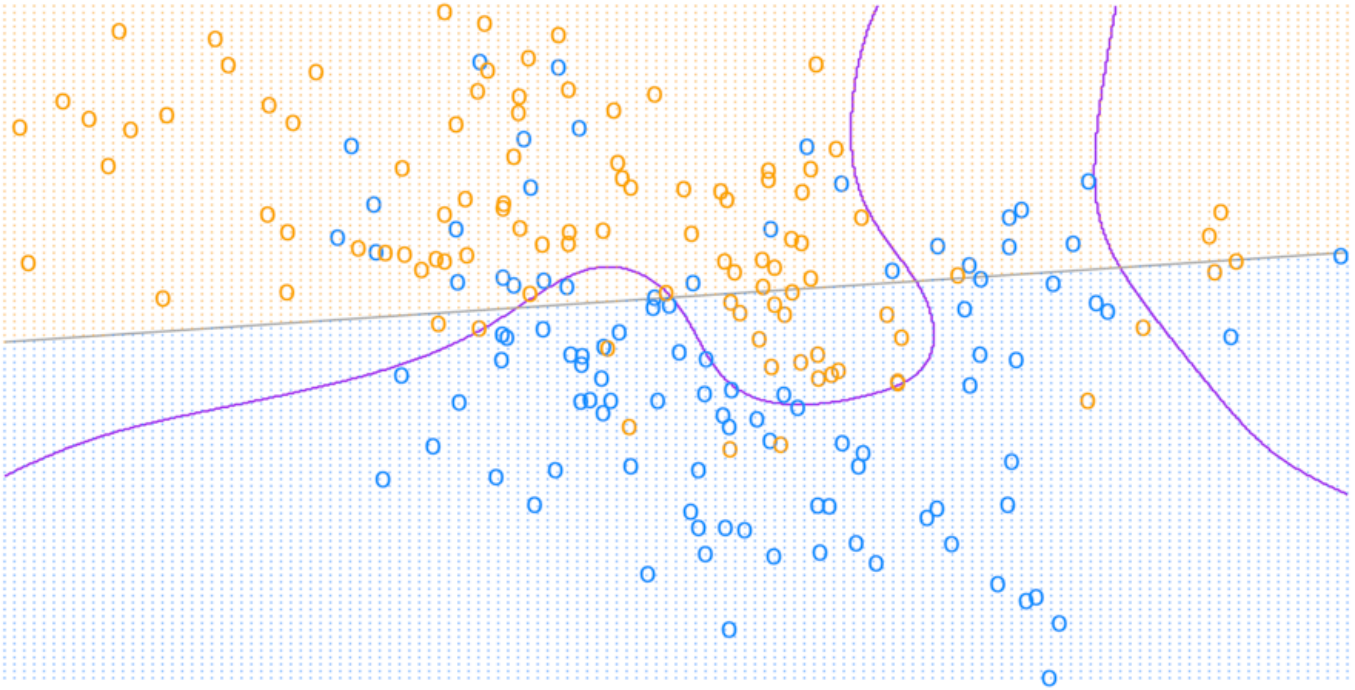
Toy example - A brief tour

We give a brief tour showing the decision boundary of common classifiers for the Mixture data set.

Scoring

Toy example - Logistic Regression classifier

We plot below the linear decision boundary given by a Logistic Regression classifier vs Bayes classifier, we only display the training set:

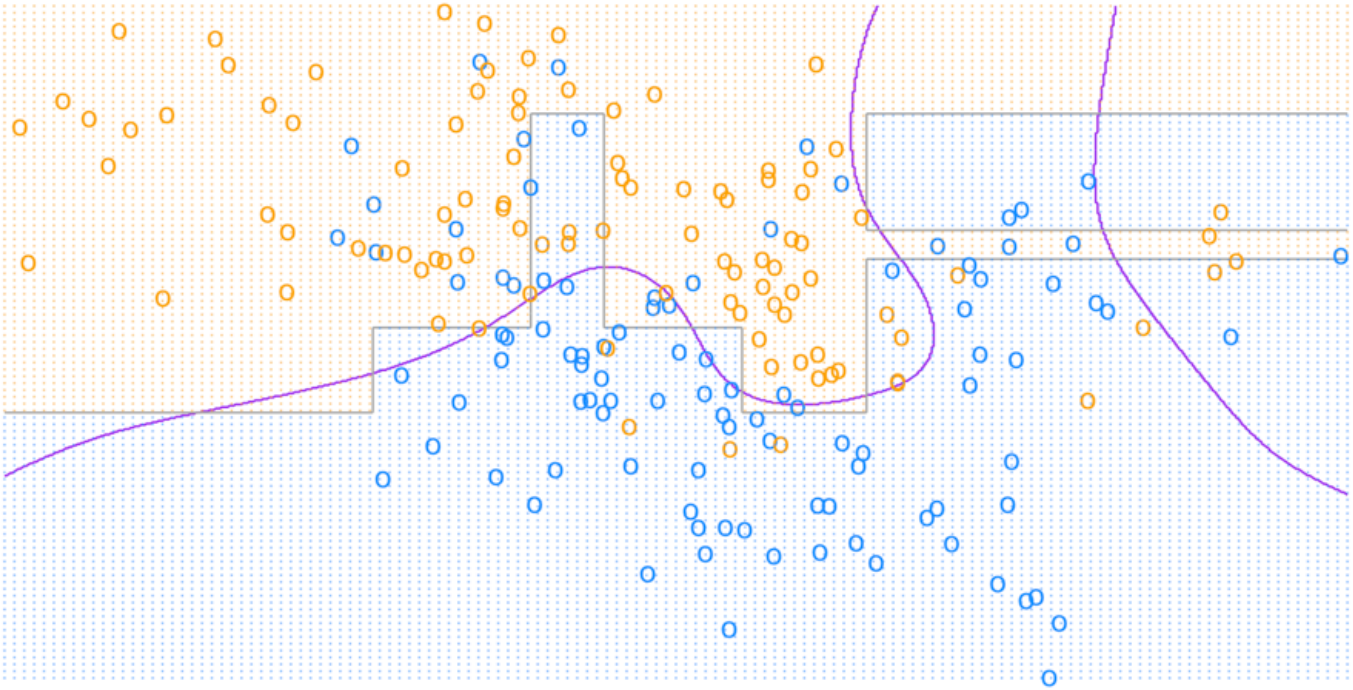


The empirical risk on testing set is 0.291.

Scoring

Toy example - Logistic Regression classifier - binning

It is usual, for example in the context of credit scoring or banking, to discretize or categorize numerical variables (a.k.a binning). We use a simple decile binning on x_1 and x_2:

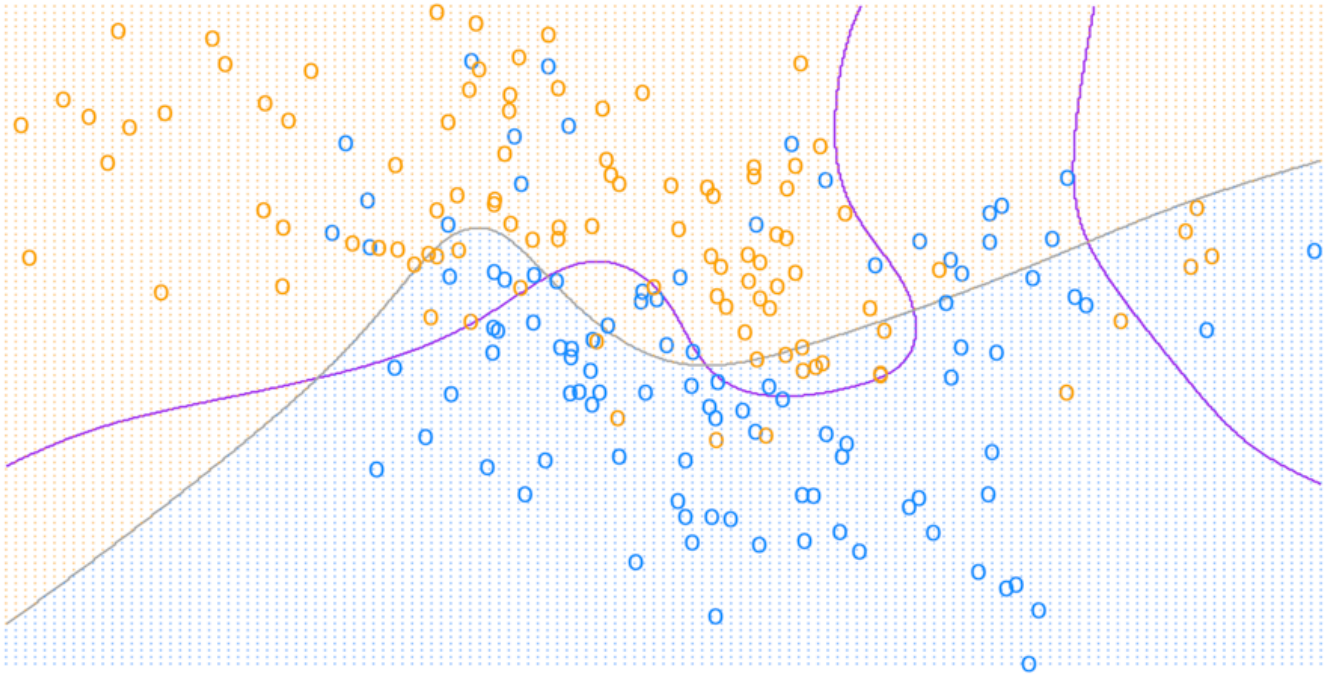


The empirical risk on testing set is 0.265.

Scoring

Toy example - Logistic Regression classifier - splines

A popular method to move beyond linearity is to transform variables, for example using splines (Hastie et al., 2009, p. 5.6 Nonparametric Logistic Regression, p161-164).

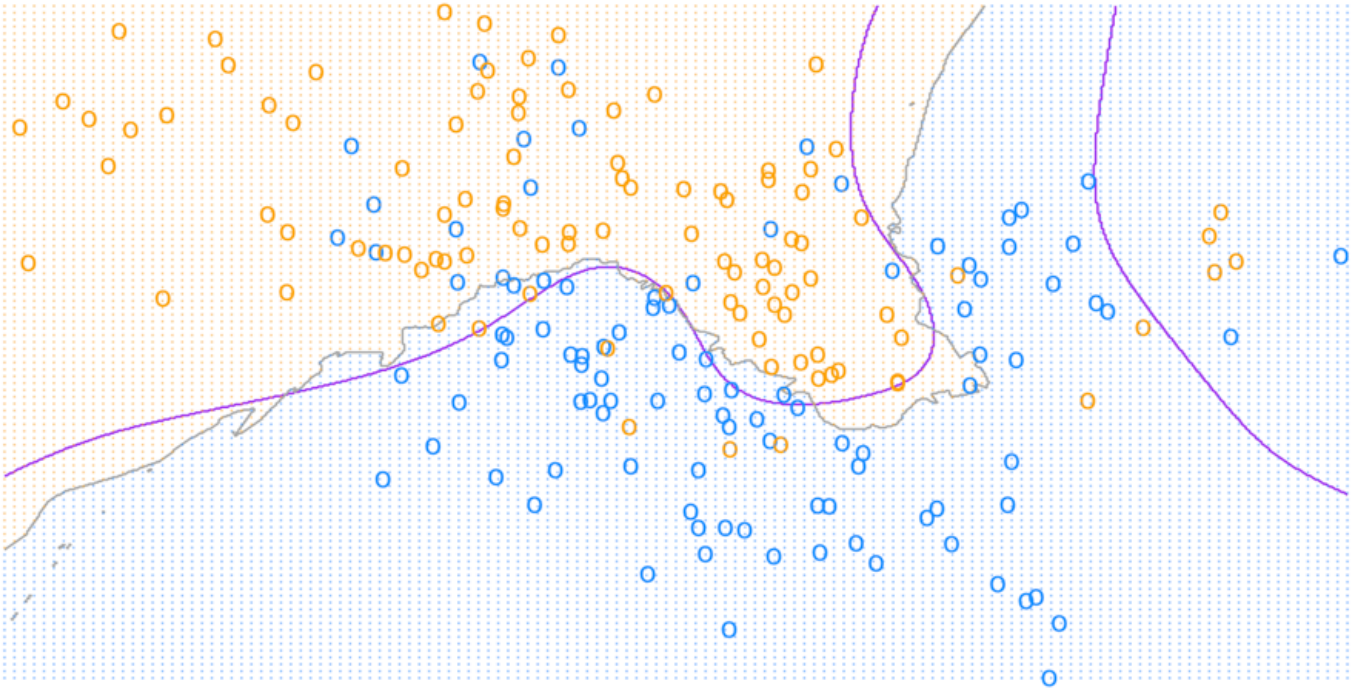


The empirical risk on testing set is 0.29.

Scoring

Toy example - k-Nearest Neighbors Algorithm - $k = 15$

k-NN is a non-parametric algorithm. Given a new observation, it finds in the training set the k closest points (given a certain distance) and predicts using a majority vote. Below the boundary decision for $k = 15$:

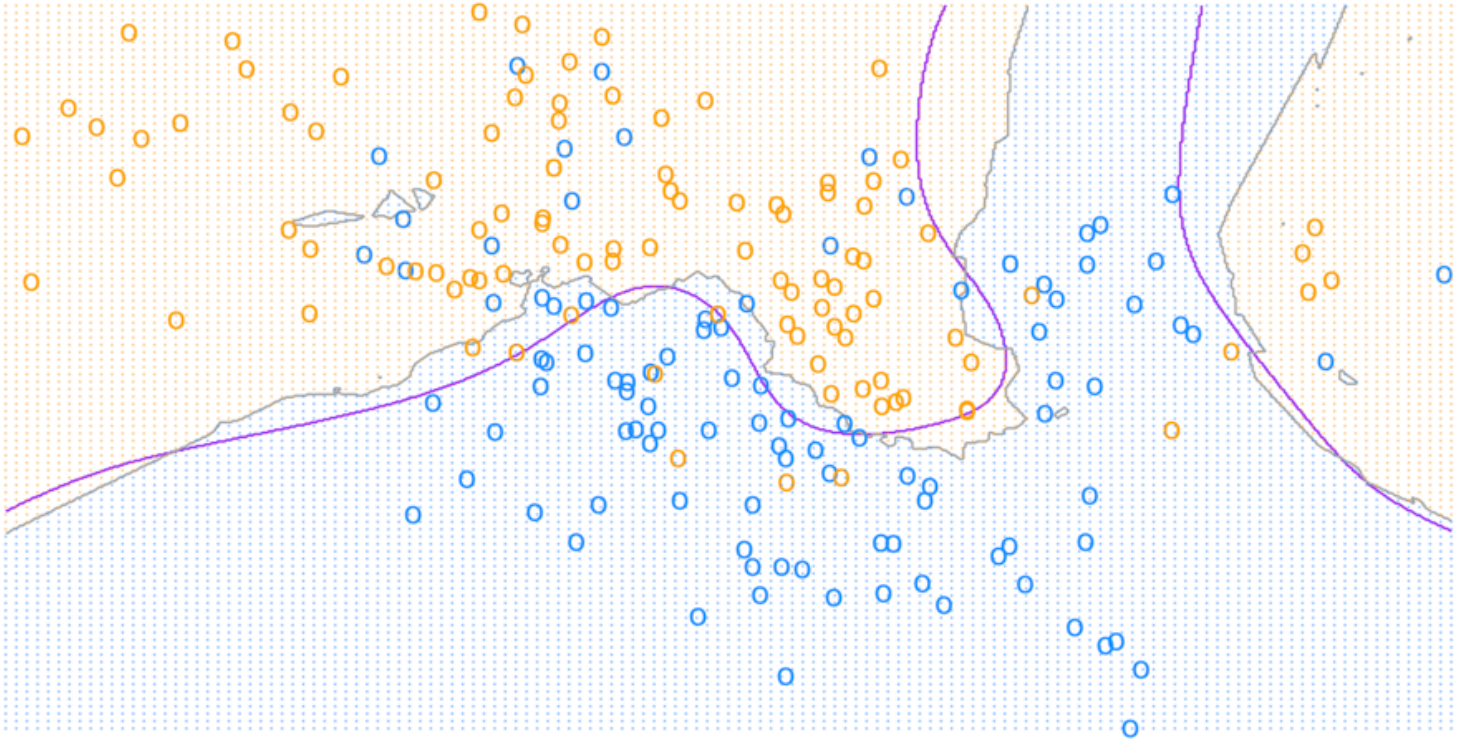


The empirical risk on testing set is 0.25.

Scoring

Toy example - k-Nearest Neighbors Algorithm - $k = 7$

We show below the boundary decision for $k = 7$:



The empirical risk on testing set is 0.227.

Scoring

Toy example - k-Nearest Neighbors Algorithm - $k = 7$

To be compared with (Hastie et al., 2009, Figure 13.4):

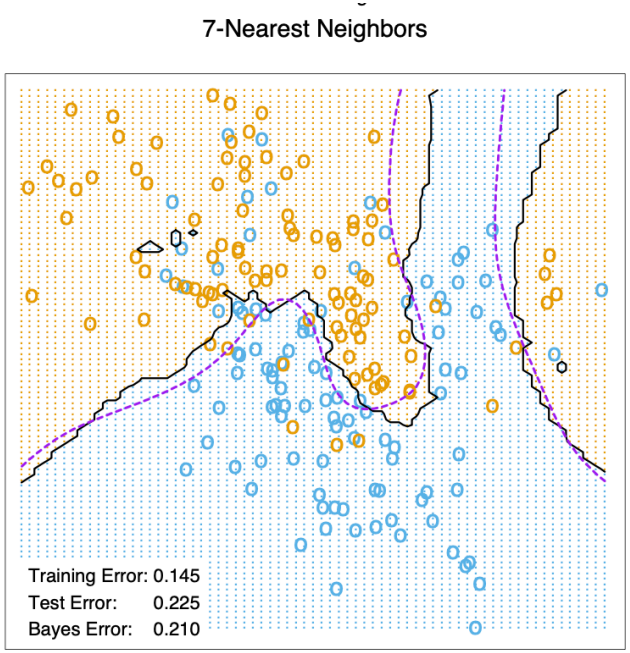
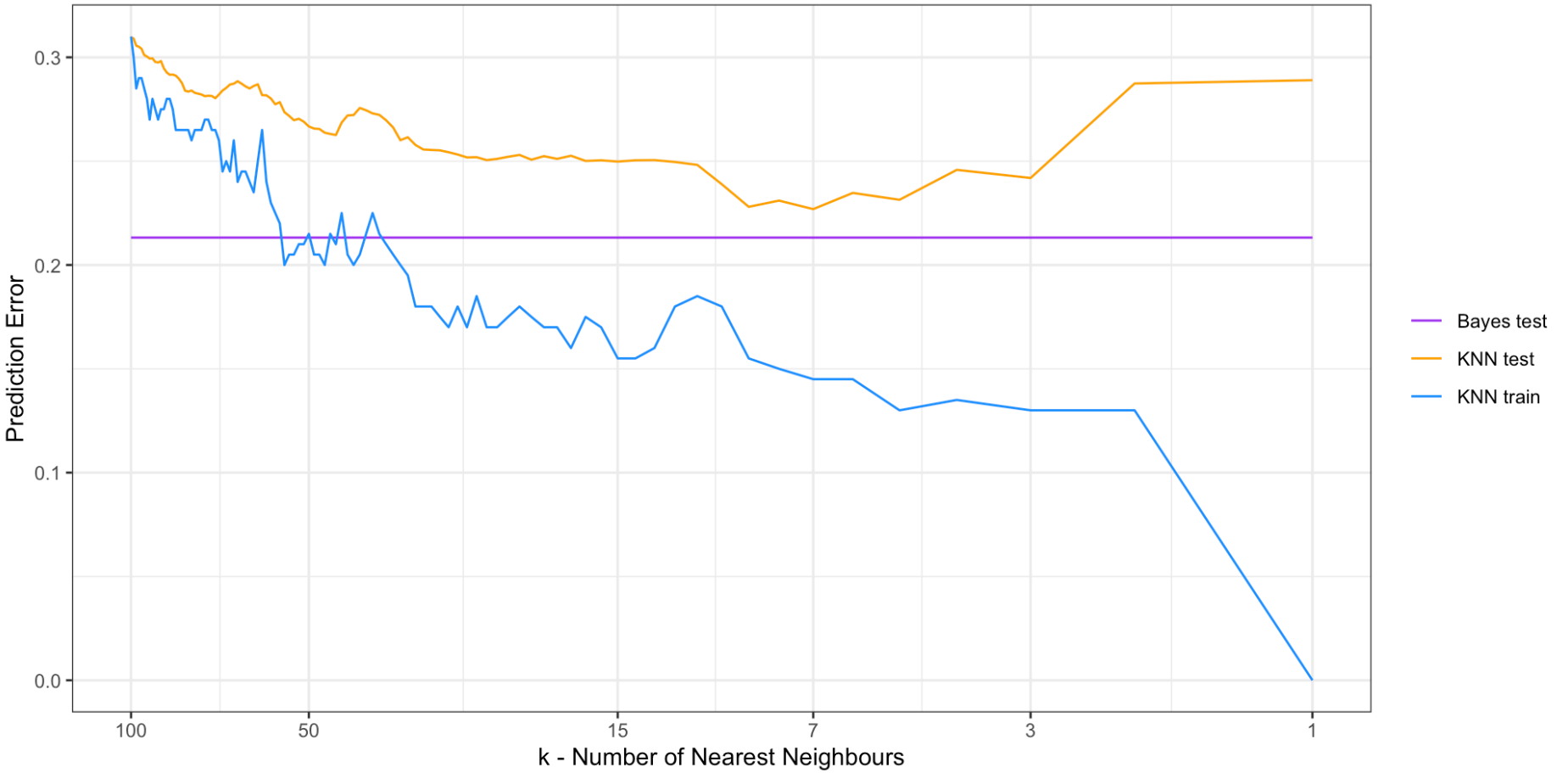


FIGURE 13.4. *k-nearest-neighbors on the two-class mixture data. The upper panel shows the misclassification errors as a function of neighborhood size. Standard error bars are included for 10-fold cross validation. The lower panel shows the decision boundary for 7-nearest-neighbors, which appears to be optimal for minimizing test error. The broken purple curve in the background is the Bayes decision boundary.*

Scoring

Toy example - k-Nearest Neighbors Algorithm - varying k

We vary k and show the empirical risks on training and testing sets together with Bayes risk:



Scoring

Toy example - k-Nearest Neighbors Algorithm - varying k

To be compared with (Hastie et al., 2009, Figure 2.4):

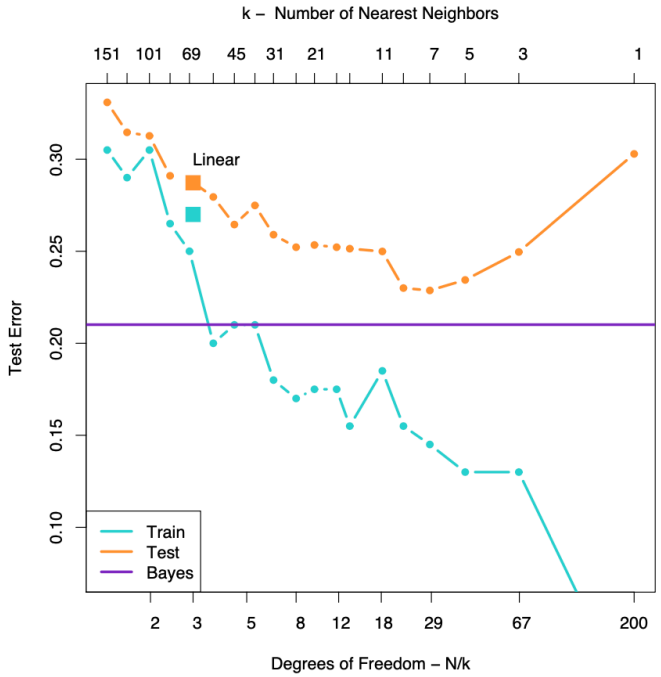
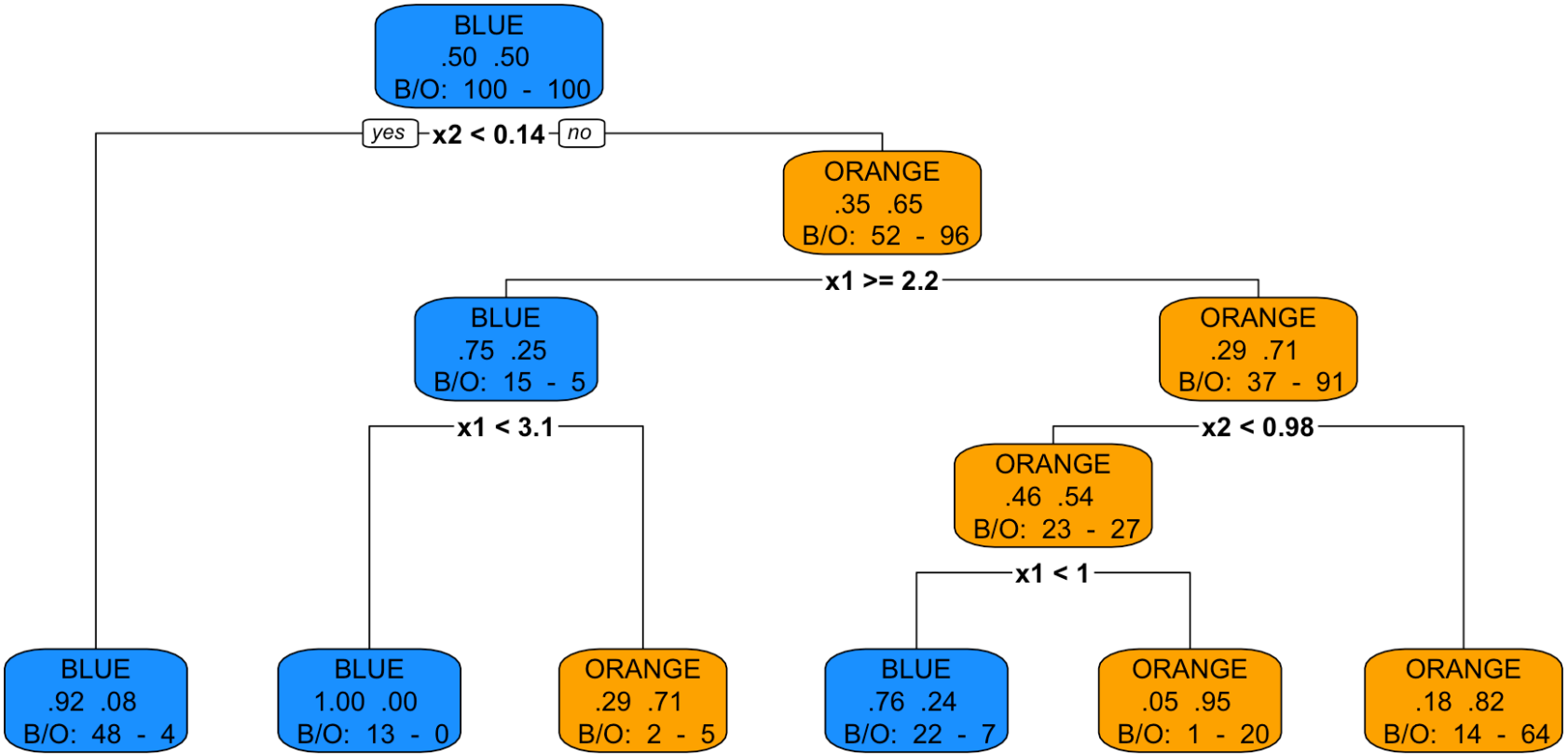


FIGURE 2.4. *Misclassification curves for the simulation example used in Figures 2.1, 2.2 and 2.3. A single training sample of size* 200 *was used, and a test sample of size* $10,000$. *The orange curves are test and the blue are training error for k-nearest-neighbor classification. The results for linear regression are the bigger orange and blue squares at three degrees of freedom. The purple line is the optimal Bayes error rate.*

Scoring

Toy example - Decision Trees

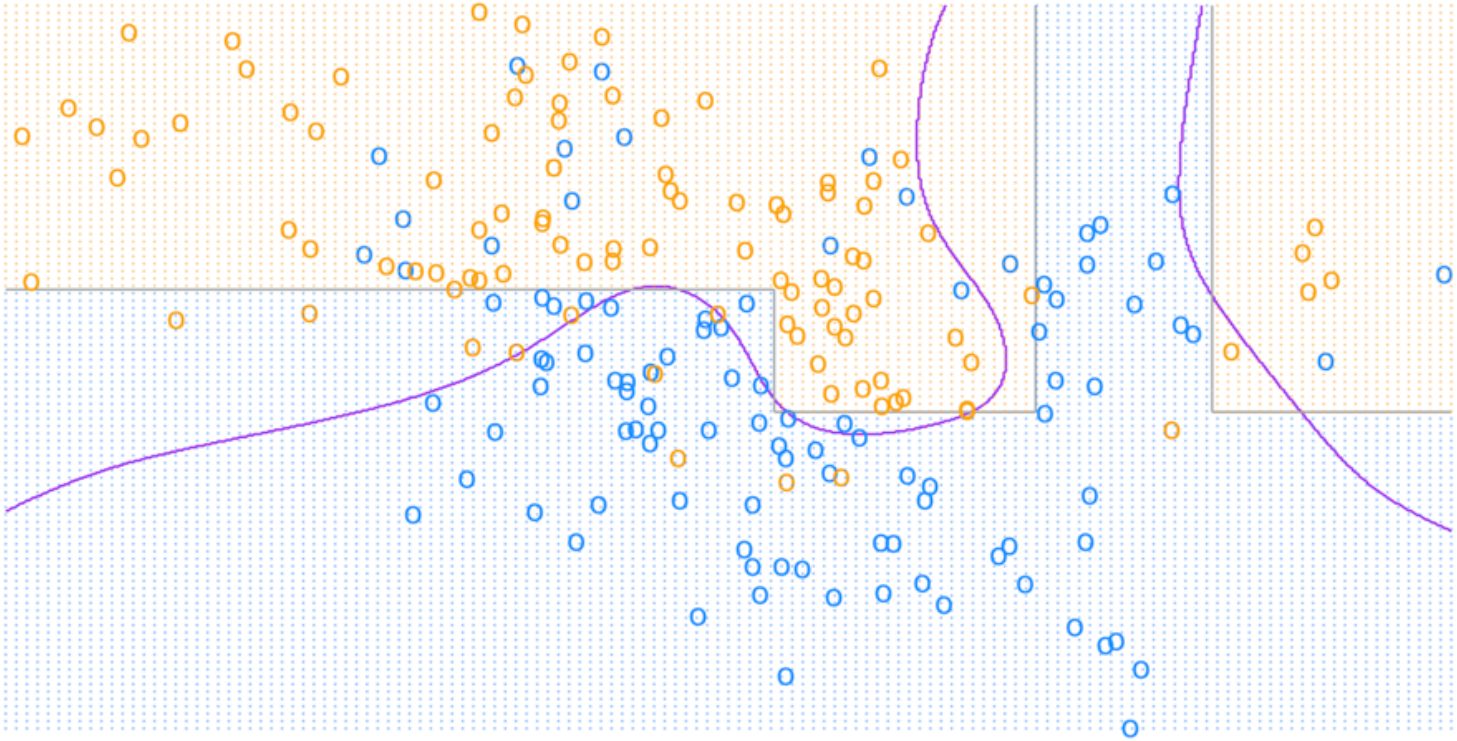
To finish our tour, we show the result of a recursive partition of the plan using Decision Trees techniques:



Scoring

Toy example - Decision Trees

We show below the boundary decision for the previously fitted tree:



The empirical risk on testing set is 0.247.

Scoring

If you like boundary decision plots

To play further and make nicer animated 2D decision boundaries for binary classifiers see these two blogs here and here.

Please also note that the `scikit-learn` `Python` library implements methods to display decision boundaries for classifiers.

Scoring

At last we come to the subject of our course!

We remind we are in the context of binary classification.

Often we (and the business) are more interested in estimating the probabilities that Y belongs to each class.

Quoting Trevor Hastie: “For example, it is more valuable to have an estimate of the probability that an insurance claim is fraudulent, than a classification fraudulent or not.”

Scoring

Scoring function

As introduced in Cornillon et al. (2019, Chapter 11.6 “Prévision - scoring”):

- The aim of Scoring: find a Scoring function or Score $S : \mathbb{R}^p \to \mathbb{R}$ that is “high” in case $\mathbb{P}[Y = 1|X = x)$ is “high” and “low” in case $\mathbb{P}[Y = 0|X = x)$ is “high”.
- We say that $S(x)$ is the score of the observation $x \in \mathbb{R}^p$.
- The notions of Score and classifier are closely linked. Given a Score S and a cutoff $s \in \mathbb{R}$ we obtain the following classifier:

$$f_s(x) = \begin{cases} 1 & \text{if } S(x) \geq s \\ 0 & \text{otherwise} \end{cases}$$

- The values taken by $S(x)$ are less important than the way they will order a set of observation $x_1, \dots x_n$. For any $\phi : \mathbb{R} \to \mathbb{R}$ bijective and increasing function, we say that the scores S and $\phi \circ S$ are equivalent.

Scoring

First example

We show below a simplified example of a credit scorecard that were typically used by banks to assess the creditworthiness of consumer loans applicants (as shown in (Thomas et al., 2002)):

Table 2.1. *Simple scorecard.*

Residential Status		**Age**	
Owner	36	18–25	22
Tenant	10	26–35	25
Living with parents	14	36–43	34
Other specified	20	44–52	39
No response	16	53+	49

Loan Purpose		**Value of CCJs**	
New car	41	None	32
Second-hand car	33	£1–£299	17
Home improvement	36	£300–£599	9
Holiday	19	£600–£1199	–2
Other	25	£1200+	–17

Scoring

First example

Table 2.1. *Simple scorecard.*

Residential Status		Age	
Owner	36	18–25	22
Tenant	10	26–35	25
Living with parents	14	36–43	34
Other specified	20	44–52	39
No response	16	53+	49

Loan Purpose		Value of CCJs	
New car	41	None	32
Second-hand car	33	£1–£299	17
Home improvement	36	£300–£599	9
Holiday	19	£600–£1199	–2
Other	25	£1200+	–17

- a 35-year-old owner wishing to borrow money for home improvement and that has never had a county court judgement (CCJ) will score $129 = (36 + 25 + 36 + 32)$,
- a 20-year-old living with his parents borrowing for holiday and having more than £1200 of CCJ will score $38 = (22 + 14 + 19 - 17)$.

Scoring

The regression function, a Scoring function

We defined above, in the context of supervised learning, the regression function $\eta(x) = \mathbb{P}[Y = 1|X = x)]$
.

It is a natural candidate for a Scoring function.

The related Bayes classifier uses $s = \frac{1}{2}$ as cutoff.

As we have just seen in the last sections, $\eta(x)$ is unknown and we will have to approximate or estimate this regression function using the training set.

Scoring

Evaluating a Score

We first define the **confusion matrix**:

	$f_s(X) = 0$	$f_s(X) = 1$	
$Y = 0$	TN	FP	N
$Y = 1$	FN	TP	P

The confusion matrix is a contingency table which cross-tabulates Y with the predicted one $f_s(X)$.

It can be evaluated on the learning set as well as on the testing set.

The following vocabulary originates from medical applications:

- true positive (TP)
- true negative (TN),
- false positive (FP), also Type I error
- false negative (FN), also Type II error

Scoring

Evaluating a Score

More formally we define for a given cutoff s:

$$\alpha(s) = \mathbf{P}(f_s(X) = 1|Y = 0) = P(S(X) \geq s|Y = 0)$$

and

$$\beta(s) = \mathbf{P}(f_s(X) = 0|Y = 1) = P(S(X) < s|Y = 1)$$

$\alpha(s)$ is called **false positive** rate and $\beta(s)$ **false negative** rate. Similarly **specificity** and **sensitivity** are defined:

$$sp(s) = P(S(X) < s|Y = 0) = 1 - \alpha(s)$$

and

$$se(s) = P(S(X) \geq s|Y = 1) = 1 - \beta(s)$$

Scoring

The Receiver Operating Characteristic (ROC) curve

The ROC (Receiver Operating Characteristic) curve allows to visualize α and β on a same graph for all s allowing to choose the cutoff and compare different Scores.

Precisely, the ROC curve of a Score S is a parametric curve of the variable s:

$$\begin{aligned} ROC: \quad & \mathbb{R} \to [0,1]^2 \\ & s \to (x(s) = \alpha(s), y(s) = 1 - \beta(s)) \end{aligned}$$

For a given Score S, the ROC curve passes through the points $(0,0)$ and $(1,1)$, which corresponds to classifying all observations as 0 ($s \to \infty$) or 1 ($s \to -\infty$).

Scoring

The Receiver Operating Characteristic (ROC) curve

A Score S is said to be **perfect** if s^* exists such that:

$$P(Y = 1|S(X) \geq s^*) = 1$$

and

$$P(Y = 0|S(X) < s^*) = 1$$

It corresponds to $x(s^*) = 0$ and $y(s^*) = 1$ (using the preceding definitions and Bayes rule).

For example:

$$x(s^*) = P(S(X) \geq s^*|Y = 0) = \frac{P(Y = 0|S(X) \geq s^*)P(S(X) \geq s^*)}{P(Y = 0)} = 0$$

Similarly: for $s \leq s^*$ we have $x(s) = 0$ and for $s > s^*$ we have $y(s) = 1$.

Scoring

The Receiver Operating Characteristic (ROC) curve

A score S is said to be **random** if $S(X)$ and Y are independent. In such case:

$$x(s) = P(S(X) \geq s | Y = 0) = P(S(X)) = P(S(X) \geq s | Y = 1) = y(s)$$

The ROC curve for a random score is the first bisector $(0, 0)$ to $(1, 1)$.

Scoring

The Receiver Operating Characteristic (ROC) curve

Figure 1 shows the two curves.

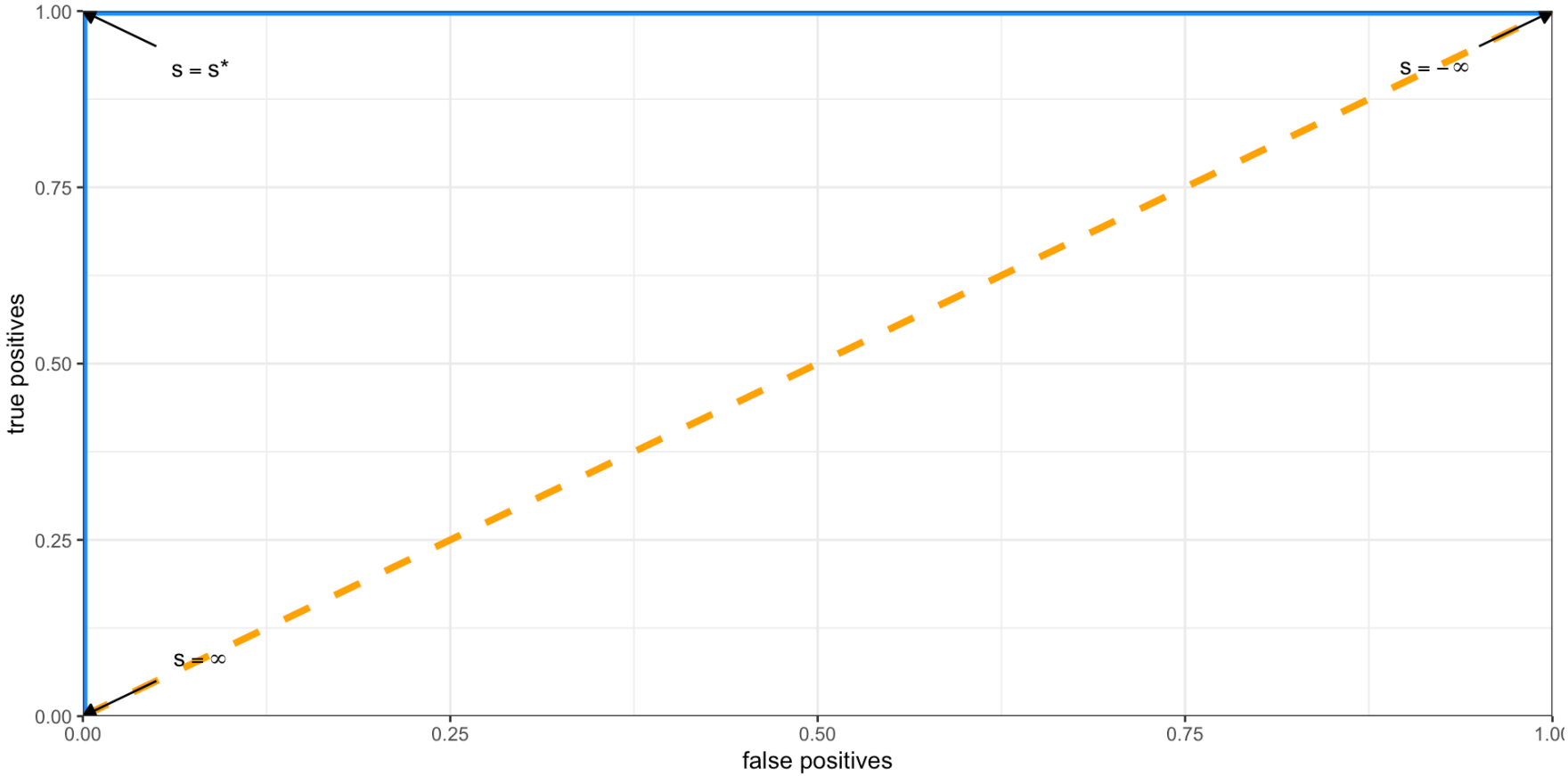


Figure 1: ROC curves of a perfect Score (solid, blue) and a random Score (dashed, orange)

Scoring

The Receiver Operating Characteristic (ROC) curve

The ROC curves can be summarized to generate numeric criteria such as the AUC (Area Under Curve).

The AUC will be 1 for the perfect Score and 0.5 for the random Score and can be used to rapidly compare two Scores.

It is considered better than other point-wise criteria.

However we must keep in mind that the AUC as a summary criterion presents some drawbacks since two Scores can have the same AUC but behave very differently in the plane.

Scoring

The Receiver Operating Characteristic (ROC) curve

In practice, we do not know $x(s)$ and $y(s)$ to define the ROC curve.

Usually, a training set will be used to learn a Score S and a testing set to estimate the ROC curve $x(s)$ and $y(s)$ for a range of s.

In the rest of the course/project we will be equally interested in binary classification and scoring which are closely linked.

Throughout the course/project, we will focus on two fundamental and complementary models: the Logistic Regression model and Decision Trees, for which we have shown examples of decision boundaries on the Mixture data set.

Let's warm up and apply what we have seen on a first data set.

Desbois case study

Desbois case study

Financial problems of farm holdings

The article from Desbois (2008) (available here together with a sample data set) shows a complete case study around the detection of financial risks applicable to farm holdings.

Among others, Linear Discriminant Analysis and Logistic Regression (plus stepwise variable selection procedure) are used.

ROC curves are then provided to compare the two methods.

The original data set uses a format specific to the SPSS software, but is readable from R using the `foreign` package.

The case study by Desbois will be partly reproduced today and in the next lessons to quickly apply scoring techniques.

Desbois case study

Loading Desbois data and first glimpse at it:

Code

```
# package used to import spss file
library(foreign)

don_desbois <- read.spss("presentation_data/Agriculture Farm Lending/desbois.sav", to.data.frame = TRUE) %>% as_tibble()
glimpse(don_desbois)
```

```
Rows: 1,260
Columns: 30
$ CNTY    <fct> Eure, Eure, Eure, Eure, Eure, Eure, Eure, Eure, Eure, Eure, Eu…
$ DIFF    <fct> healthy, healthy, healthy, healthy, healthy, healthy, healthy,…
$ STATUS  <fct> Entreprise individuelle, Entreprise individuelle, Entreprise i…
$ HECTARE <dbl> 166, 101, 138, 166, 137, 107, 100, 69, 176, 177, 108, 123, 87,…
$ ToF     <fct> cereals, cereals, dairy farm, cereals, mixed livestock, cereal…
$ OWNLAND <fct> yes, no, yes, yes, yes, yes, no, yes, yes, yes, yes, yes, yes,…
$ AGE     <dbl> 35, 35, 42, 50, 33, 45, 34, 30, 44, 39, 40, 40, 40, 52, 49, 44…
$ HARVEST <fct> 1988, 1988, 1988, 1988, 1988, 1988, 1988, 1988, 1988, 1988, 19…
$ r1      <dbl> 0.449, 0.450, 0.332, 0.363, 0.440, 0.306, 0.717, 0.566, 0.145,…
$ r2      <dbl> 0.622, 0.617, 0.819, 0.733, 0.650, 0.755, 0.320, 0.465, 0.881,…
$ r3      <dbl> 0.25500, 0.24110, 0.55680, 0.35960, 0.31420, 0.26350, 0.16280,…
$ r4      <dbl> 0.11450, 0.10840, 0.18510, 0.13050, 0.13820, 0.08055, 0.11680,…
$ r5      <dbl> 0.334, 0.341, 0.147, 0.232, 0.302, 0.225, 0.601, 0.499, 0.115,…
$ r6      <dbl> 0.785, 0.518, 0.700, 0.773, 0.846, 0.709, 0.894, 0.863, 0.439,…
$ r7      <dbl> 0.585, 0.393, 0.310, 0.495, 0.580, 0.523, 0.748, 0.761, 0.348,…
$ r8      <dbl> 0.20020, 0.12500, 0.38950, 0.27790, 0.26580, 0.18700, 0.14550,…
$ r11     <dbl> 0.6628, 0.7098, 0.4142, 0.4661, 0.7715, 0.8178, 0.6598, 0.6619…
$ r12     <dbl> 1.3698, 1.2534, 0.6370, 1.0698, 1.4752, 1.4682, 1.1018, 1.2360…
$ r14     <dbl> 0.23200, 0.14970, 0.48470, 0.37350, 0.25630, 0.18610, 0.18070,…
$ r17     <dbl> 0.0884, 0.0671, 0.0445, 0.0621, 0.0489, 0.0243, 0.0352, 0.0879…
$ r18     <dbl> 0.0694, 0.0348, 0.0311, 0.0480, 0.0414, 0.0173, 0.0315, 0.0758…
$ r19     <dbl> 0.1660, 0.1360, 0.1030, 0.1080, 0.1270, 0.0652, 0.1290, 0.2400…
$ r21     <dbl> 0.1340, 0.0802, 0.0890, 0.0851, 0.0838, 0.0390, 0.0784, 0.1630…
$ r22     <dbl> 0.3219, 0.3133, 0.2945, 0.1909, 0.2567, 0.1472, 0.3211, 0.5170…
$ r24     <dbl> 0.295, 0.365, 0.166, 0.265, 0.257, 0.191, 0.322, 0.305, 0.180,…
$ r28     <dbl> 0.475, 0.434, 0.350, 0.475, 0.475, 0.443, 0.401, 0.464, 0.475,…
$ r30     <dbl> 0.35000, 0.29780, 0.24680, 0.37590, 0.36690, 0.37590, 0.27240,…
```

Desbois case study

Replacing for convenience `DIFF` target by factor `Y` with `0` (healthy) or `1` (failing):

▾ Code

```
don_desbois <- don_desbois %>%
    mutate(Y = as.factor(if_else(DIFF=='healthy', 0, 1)), DIFF = NULL,
           .before = everything())
```

We give in the following slides the definitions of Desbois ratio $r_1, \ldots, r_{37}$, that we will use today, as found in the article.

They might give you inspiration to create new predictors/features for the project to come.

Desbois case study

Capitalization ratios

- r1 total debt / total assets;
- r2 stockholders' equity / invested capital;
- r3 short term debt / total debt;
- r4 short term debt / total assets;
- r5 long and medium term debt / total assets;

Desbois case study

Weight of the debt ratios

- r6 total debt / gross product;
- r7 long and medium term debt / gross product;
- r8 short term debt / gross product;

Desbois case study

Liquidity ratios

- r11 working capital / gross product;
- r12 working capital / (real inputs - financial expenses);
- r14 short term debt / circulating assets;

Desbois case study

Debt servicing ratios

- r17 financial expenses / total debt;
- r18 financial expenses / gross product;
- r19 (financial expenses + refunding of long and medium term capital) / gross product;
- r21 financial expenses / EBITDA;
- r22 (financial expenses + refunding of long and medium term capital) / EBITDA;

Desbois case study

Capital profitability ratio

- r24 EBITDA / total assets;

Desbois case study

Earnings ratios

- r28 EBITDA / gross product;
- r30 available income / gross product;
- r32 (EBITDA - financial expenses) / gross product;

Desbois case study

Productive activity ratios

- r36 immobilized assets / gross product;

- r37 gross product / total assets.

Desbois case study

Your turn to play

Using the data set at hand, try to retrieve some findings of the Desbois case study.

You might need `R` packages `FactoMineR`, `ROCR` and `MASS::lda` function to perform the tasks ("YOUR CODE HERE") described in the following slides.

Each time I show an example of what is expected.

I suggest that you start using `Quarto` (setup here) to code and track/publish your results. It will be requested for your projects (it is very similar to `RMarkdown`). It integrates perfectly with `RStudio`.

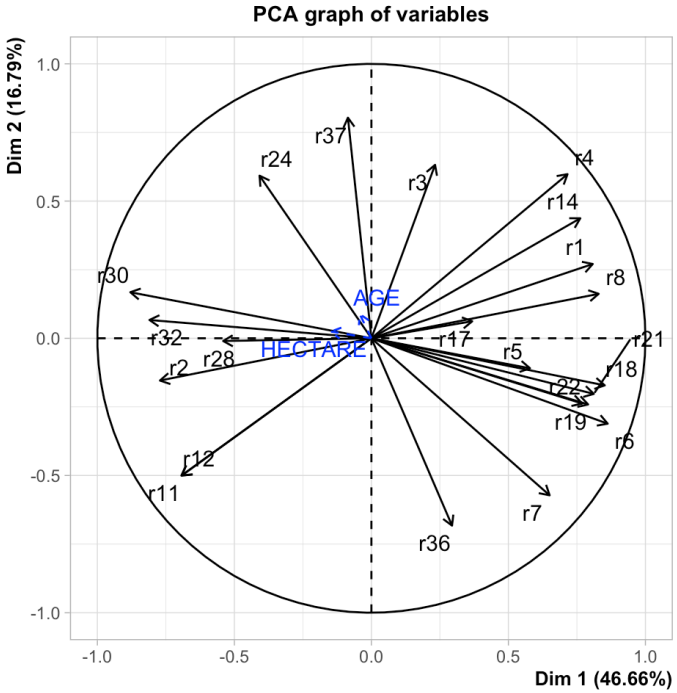
Desbois case study

Principal Component Analysis (PCA) of financial ratios

First perform PCA on financial ratios (use for example package FactoMineR), then display correlations between ratios and the two first principal components :

Code

```
1  ########## YOUR CODE HERE ######################
```



Desbois case study

To be compared with article Figure 1

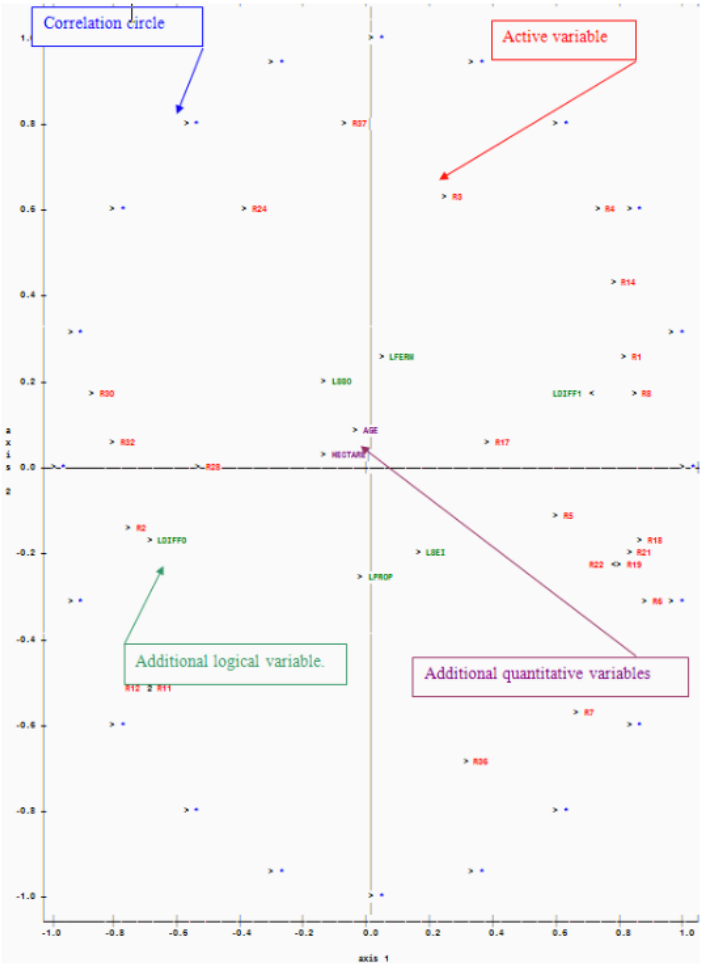


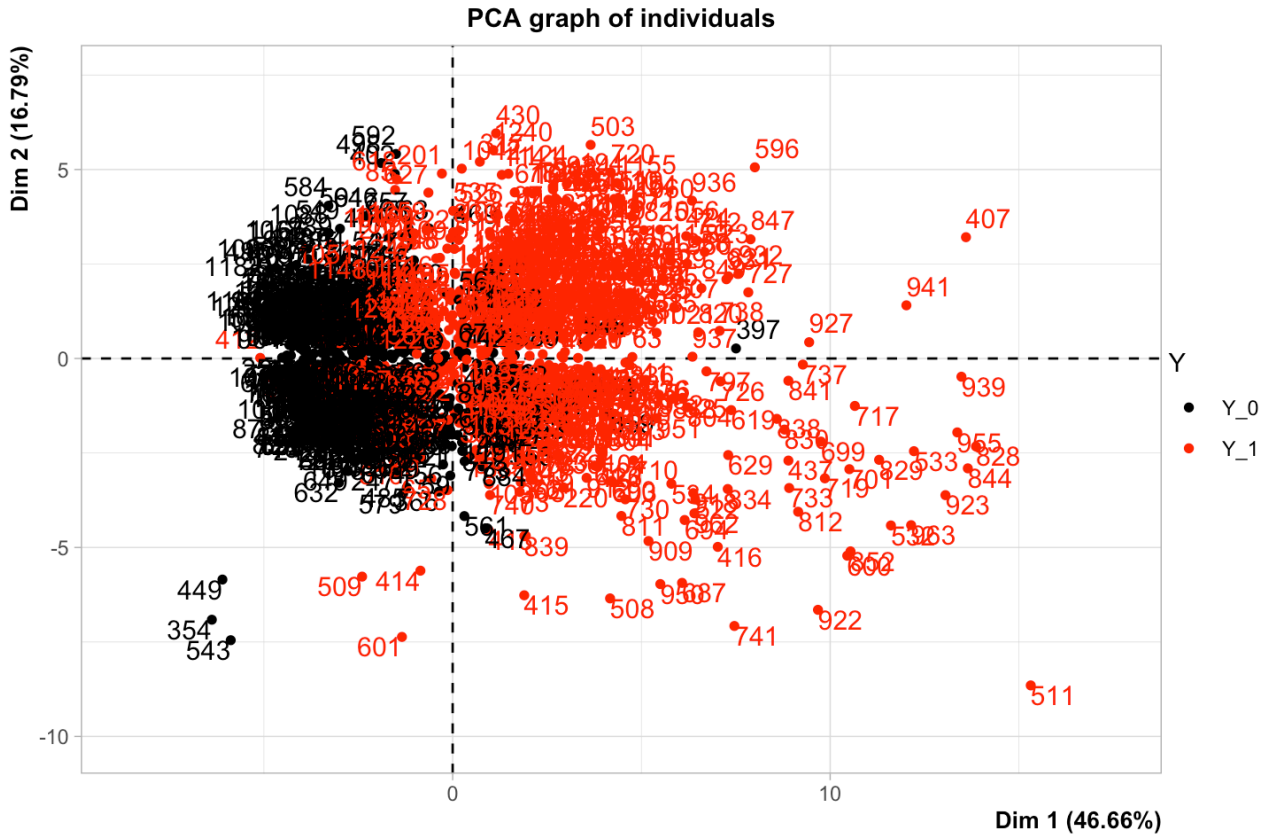
Figure 1. Projection of the financial ratios on the first factorial plane of the normalized PCA.

Desbois case study

Code

```
1  ########## YOUR CODE HERE ####################
```

Following Desbois path, visualize the farm holdings in data set as a bivariate plot on the first two components of PCA (based on financial ratios), use variable Y (0="healthy"; 1="failing") to colour your observations:



Desbois case study

To be compared with article Figure 2

It strikes Desbois that the PCA (even if not a discriminative/scoring procedure), seems to do a rather good job identifying defaulting farms on the training set.

Disclaimer: in general it won't be necessarily the case as PCA operates only on the predictors without "knowledge" of target variable.

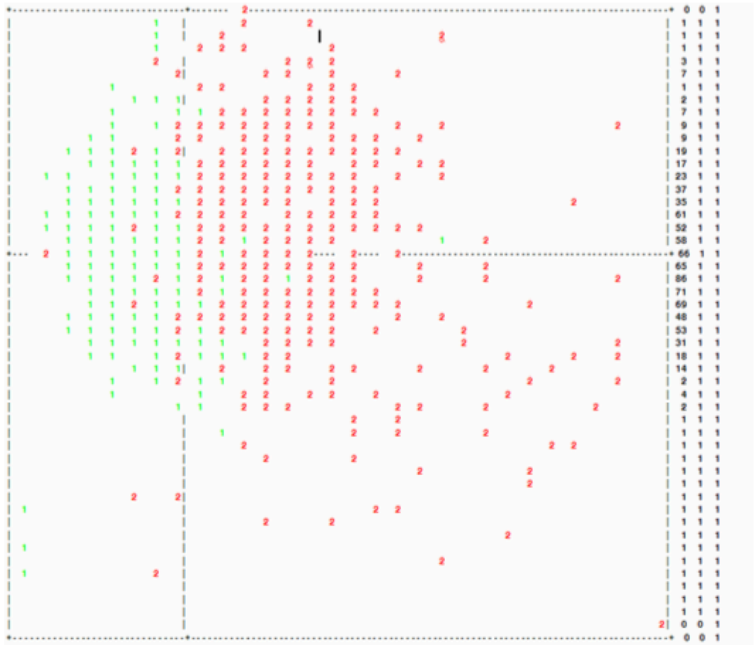


Figure 2. Plot of the farm holdings in the first factorial plane of the normalized PCA based on financial ratios with illustrative variable LDIFF (1="healthy"; 2="failing").

Desbois case study

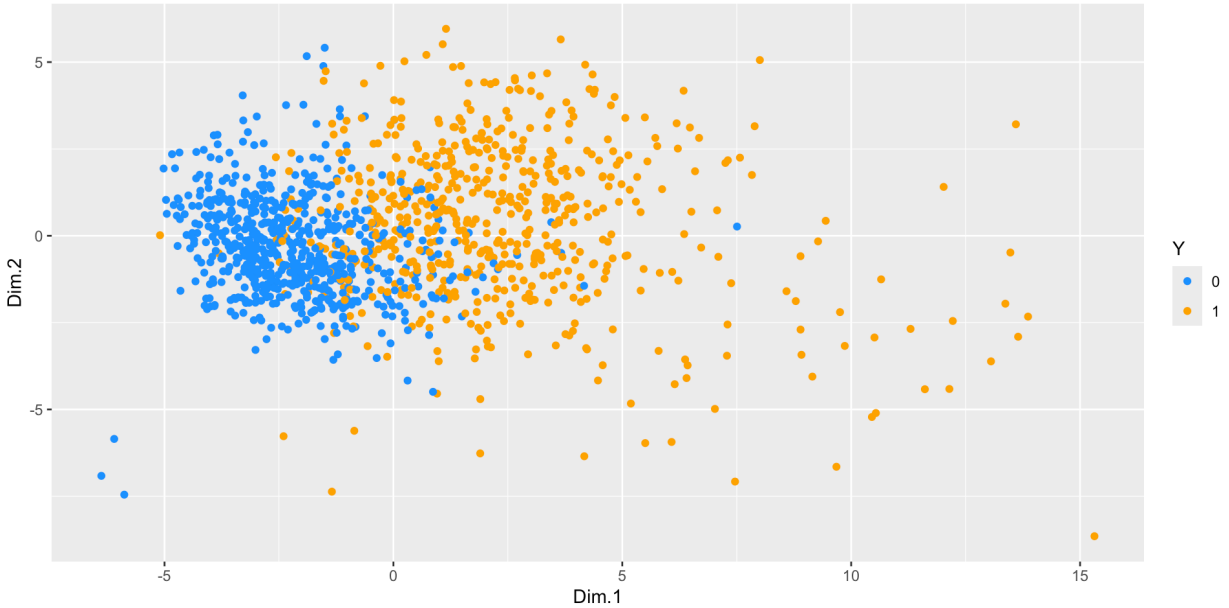
As suggested by Desbois, extract the first principal component (some help here), and use it as a Scoring function.

Today it will allow us to use a first Scoring function without introducing the Logistic Regression that you have not yet studied in class.

Code

```
1  ########## YOUR CODE HERE ####################
```

Below the score for each observation is the x-axis or first principal component:



Desbois case study

Looking at this plot Desbois concludes that a simple classifier can be devised using the first PCA coordinate (setting a threshold at $PC_1 > 0.02$):

inertia. Considering the respective positions of the healthy and failing groups of farm holdings on the first factorial plane, we see that the *F1* axis could be used as an index of classification (see Figures 2 and 4).

Indeed, the average coordinate of the "healthy" group on the *F1* axis is approximately equal to $\mu_0 \approx -0.6754$ while the average coordinate of the "failing" group is approximately equal to $\mu_1 \approx 0.7266$.

A geometrical rule of classification for a farm holding i_0 which does not belong to the training sample consists in positioning the farm holding in relation to the "pivot point", i.e. the median point of the segment linking the two group barycenters on the *F1* axis, that is to say

$$\frac{\mu_0 + \mu_1}{2} = 0.0256 :$$

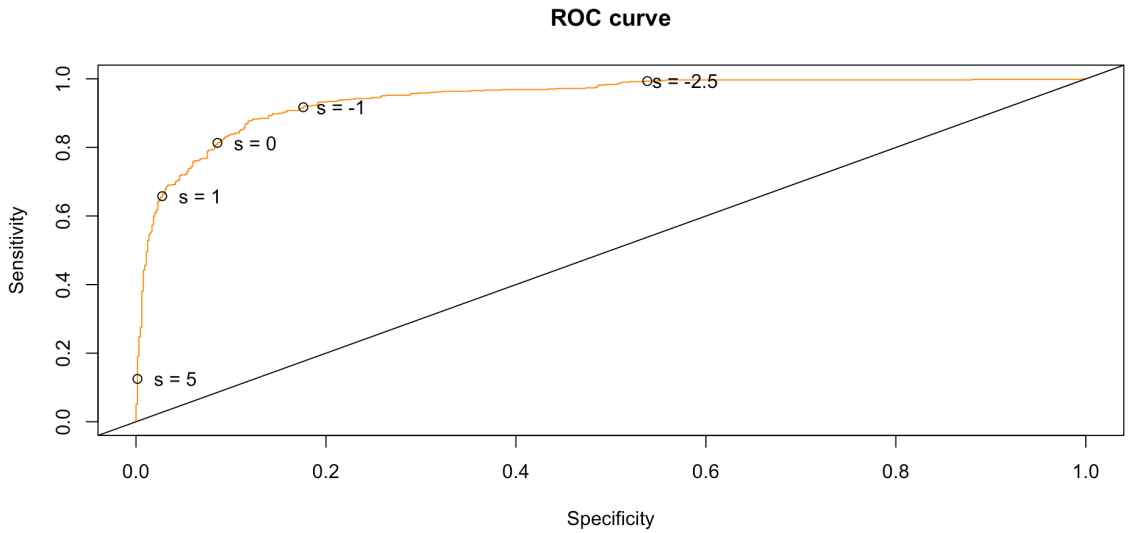
Desbois case study

Going further than Desbois and for illustrative purposes only, use PC_1 as a very naive Scoring function.

Plot below the ROC curve as defined in the Scoring part of the presentation (you can use package ROCR, but a simple for-loop for different cutoff values s will do the job):

▾ Code

```
1  ########## YOUR CODE HERE ######################
```



We follow closely the case study, but we will see that ROC curves are usually evaluated on a hold-out data set.

Desbois case study

Then compare with a Scoring function obtained with Linear Discriminant Analysis (LDA). For a reminder on LDA see for example Hastie et al. (2009, Chapter 4.3 LDA, p. 103-111).

You can use Desbois variables selected in the article by a stepwise procedure (using Wilk's lambda[1]). Selected variables are:

Table 6. Stepwise selection of the predictors discriminating the two groups of farm holdings according to the criterion minimizing Wilks' lambda.

Step	Predictors Included	Wilks Lambda		Exact F			
		Statistic	dof 1	Statistic	dof 1	dof 2	P level
1	r1	0.580	1	909.310	1	1258	0.000
2	r32	0.502	2	624.223	2	1257	0.000
3	r14	0.467	3	478.407	3	1256	0.000
4	r17	0.453	4	378.131	4	1255	0.000
5	r2	0.445	5	312.664	5	1254	0.000
6	r3	0.437	6	268.833	6	1253	0.000
7	r36	0.429	7	237.793	7	1252	0.000
8	r21	0.423	8	212.917	8	1251	0.000
9	r7	0.422	9	190.439	9	1250	0.000
10	r18	0.419	10	173.322	10	1249	0.000
11		0.420	9	192.062	9	1250	0.000

Notes: For Wilks' Lambda, dof2 = 1, dof3 = 1258. In step 11, r1 was excluded.

For instance, the total debt rate [*r1*] ratio, introduced in the first step because acting as the most discriminating variable, is eliminated in the last step of the selection process because it can be sufficiently reconstituted in an approximate way as a linear combination of the other ratios introduced into the preceding steps.

From this follows the linear discriminant function *D1* (Fisher's score), which is expressed with "standardized coefficients" as a linear combination of the standardized discriminating variables (see Table 7).

So basically you can fit LDA with `r2+r3+r7+r14+r17+r18+r21+r32+r36` as predictors (I use `R` formula notation to ease your copy-paste).

1. We won't describe the procedure here. A SPSS script is given in the article. It is not straightforward to reproduce it with R. I tried scripting manually the procedure and it works but is not very interesting.

Desbois case study

Fit LDA (`MASS::lda`) with model specification of your choice. Then display model coefficients:

▾ Code

```
1  ########## YOUR CODE HERE ####################
```

Using `r2+r3+r7+r14+r17+r18+r21+r32+r36`, we obtain:

```
       LD1
r2  -2.542
r3   2.398
r7   1.099
r14  0.604
r17 18.702
r18 -6.798
r21 -0.622
r32 -4.163
r36  0.192
```

Similar[1] to what Desbois obtains:

Table 9. Unstandardized coefficients issued from the original variables, to be used for the score computation.

Predictors	Function D1
stockholders' equity / invested capital [r2]	-2.542
short-term debt / total debt [r3]	2.398
long and medium-term debt / gross product [r7]	1.099
short-term debt / circulating asset [r14]	0.604
financial expenses / total debt [r17]	18.702
financial expenses / gross product [r18]	-6.798
financial expenses / EBITDA [r21]	-0.622
(EBITDA - financial expenses) / gross product [r32]	-4.163
immobilized assets / gross product [r36]	0.192
(Constant)	-0.444

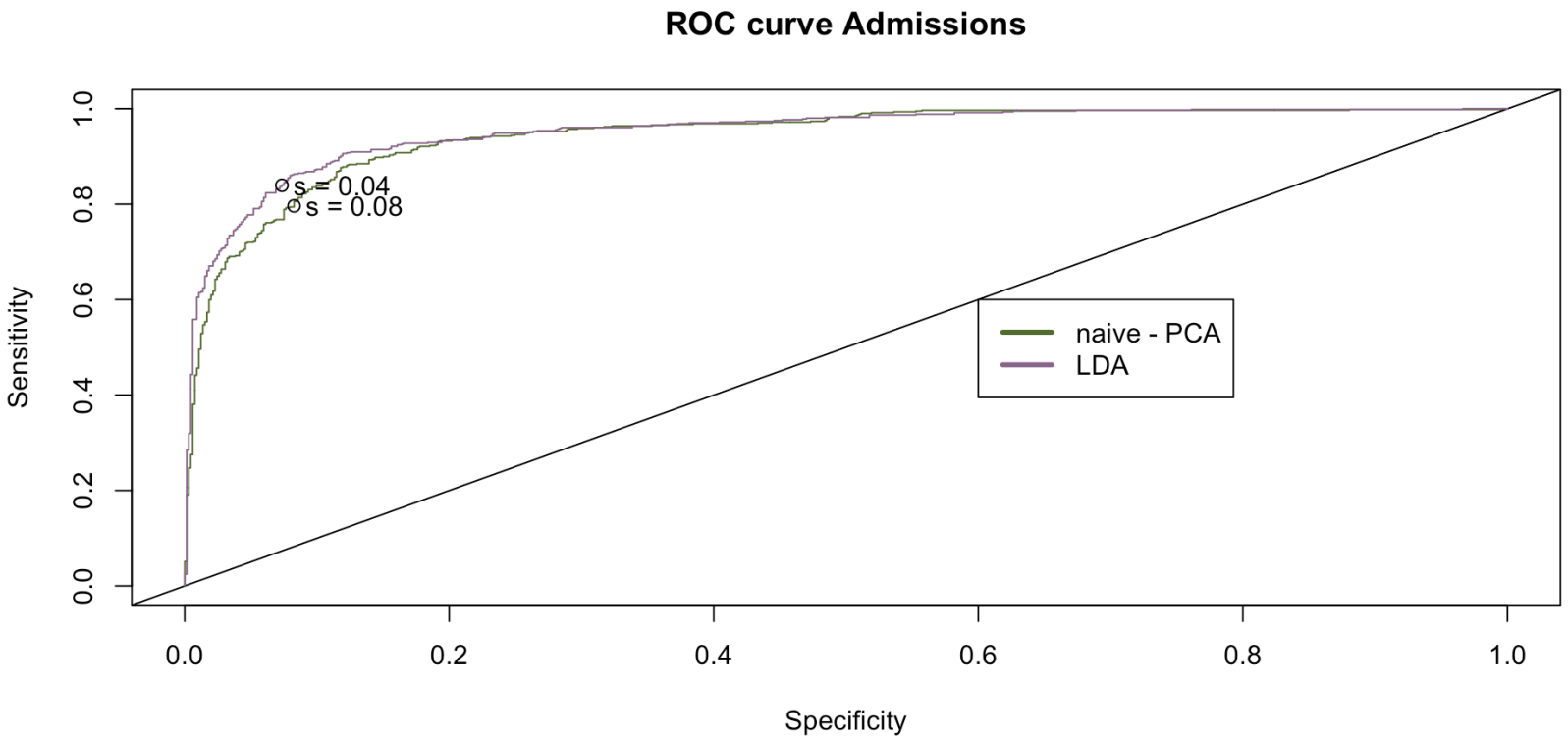
1. R does not provide constant for LDA which won't affect Scoring function ordering

Desbois case study

Now compare the LDA Scoring function with the naive Score built with first component of PCA using a ROC curve (on training set):

Code

```
1  ########## YOUR CODE HERE ######################
```



Desbois case study

As a very soft and intuitive introduction to logistic regression. First discretize a ratio r of your choice.

- Compute proportion of defaults among each class. For example using 0.01 steps for ratio $r17$:

▾ Code

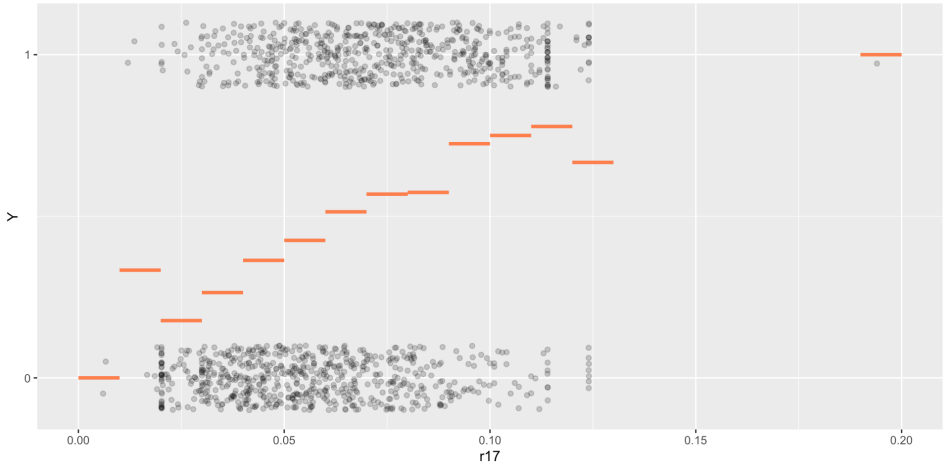
```
1  ########## YOUR CODE HERE ######################
```

r17_class	0	1	Mean(Y)
[0,0.01)	2	0	0.0000
[0.01,0.02)	4	2	0.3333
[0.02,0.03)	65	14	0.1772
[0.03,0.04)	106	38	0.2639
[0.04,0.05)	112	64	0.3636
[0.05,0.06)	104	77	0.4254
[0.06,0.07)	88	93	0.5138
[0.07,0.08)	63	83	0.5685
[0.08,0.09)	49	66	0.5739
[0.09,0.1)	27	71	0.7245
[0.1,0.11)	14	42	0.7500
[0.11,0.12)	12	42	0.7778
[0.12,0.13)	7	14	0.6667
[0.19,0.2]	0	1	1.0000

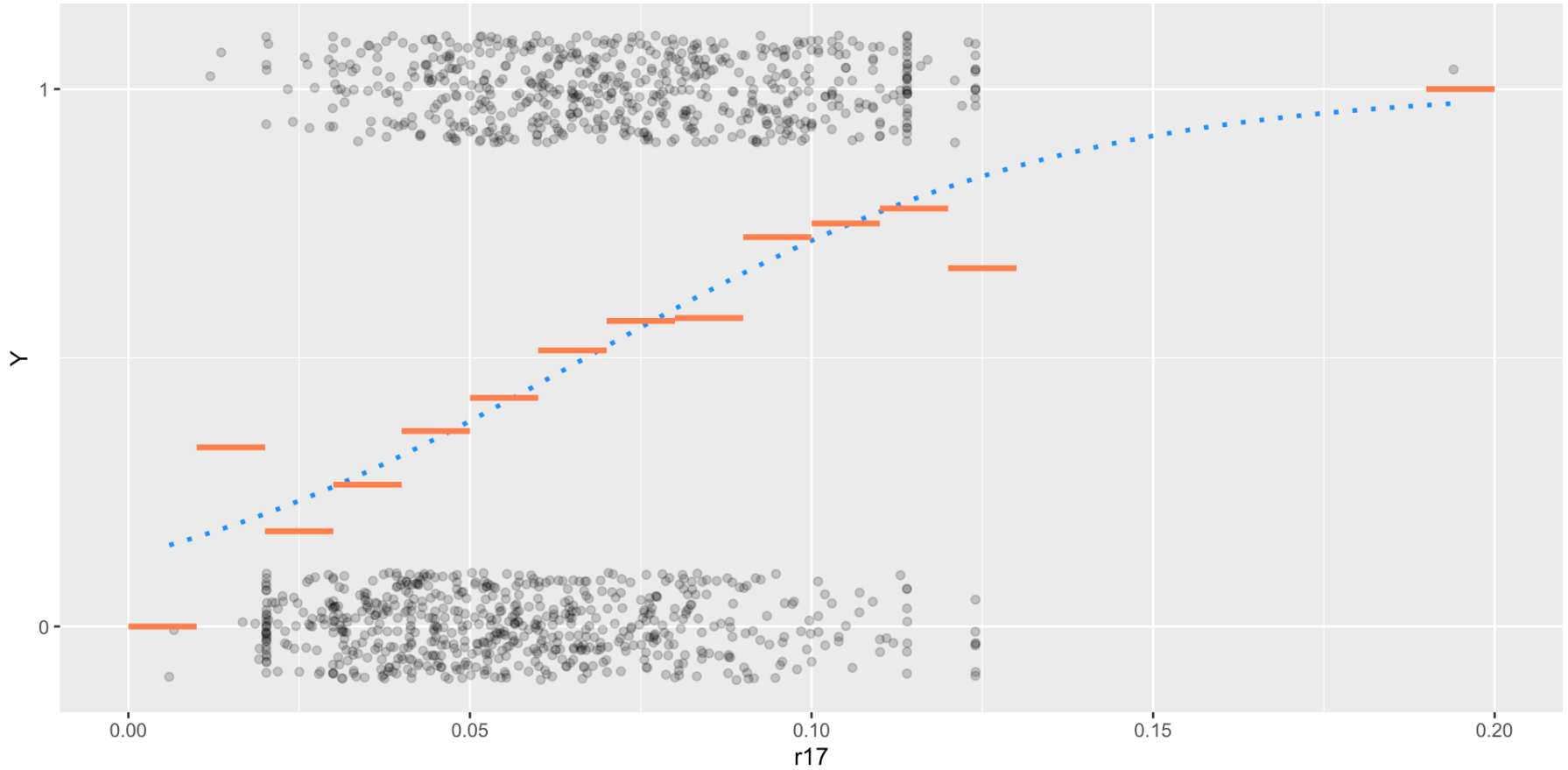
- Then using this table, plot an empirical conditional distribution of default given r classes:

▾ Code

```
1  ########## YOUR CODE HERE ######################
```



Desbois case study



We notice that the mean default occurrence with respect to r_{17} classes follows a kind of "S"-shaped curve or **sigmoid** function. Note that the shape depends on classes width and might change. We "fitted" (blue-dotted) $\sigma : x \to \sigma(x) = \frac{e^x}{1+e^x}$ also known as logistic function to the mean default occurences (orange segments).

Desbois case study

This roughly corresponds to the intuitive introduction to the logistic regression model given in Hosmer et al. (2013) using a Coronary Heart Disease (CHD) event as Y and AGE as X:

The column labeled "Mean" in Table 1.2 provides an estimate of $E(Y|x)$. We assume, for purposes of exposition, that the estimated values plotted in Figure 1.2 are close enough to the true values of $E(Y|x)$ to provide a reasonable assessment of the functional relationship between CHD and AGE. With a dichotomous outcome variable, the conditional mean must be greater than or equal to zero and less than or equal to one (i.e., $0 \leq E(Y|x) \leq 1$). This can be seen in Figure 1.2. In addition, the plot shows that this mean approaches zero and one "gradually". The change in the $E(Y|x)$ per unit change in x becomes progressively smaller as the conditional mean gets closer to zero or one. The curve is said to be *S-shaped* and resembles a plot of the cumulative distribution of a continuous random variable. Thus, it should not seem surprising that some well-known cumulative distributions have been used to provide a model for $E(Y|x)$ in the case when Y is dichotomous. The model we use is based on the logistic distribution.

Table 1.2 Frequency Table of Age Group by CHD

Age Group	n	Coronary Heart Disease: Absent	Coronary Heart Disease: Present	Mean
20–29	10	9	1	0.100
30–34	15	13	2	0.133
35–39	12	9	3	0.250
40–44	15	10	5	0.333
45–49	13	7	6	0.462
50–54	8	3	5	0.625
55–59	17	4	13	0.765
60–69	10	2	8	0.800
Total	100	57	43	0.430

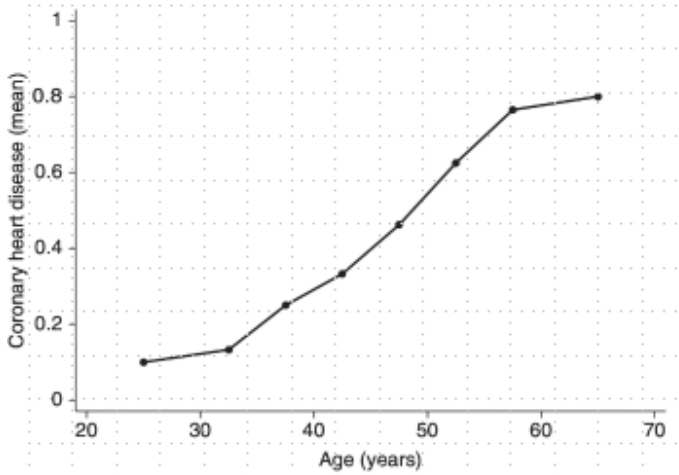


Figure 1.2 Plot of the percentage of subjects with CHD in each AGE group.

References

Bardos, M. (2007). What is at stake in the construction and use of credit scores? *Computational Economics*, *29*(2), 159–172. https://doi.org/10.1007/s10614-006-9083-x

Cornillon, P.-A., Matzner-Løber, E., Hengartner, N., & Rouvière, L. (2019). *Régression avec r - 2e édition*. EDP Sciences.

Desbois, D. (2008). Introduction to scoring methods: financial problems of farm holdings. *Case Studies in Business, Industry and Government Statistics*, *2*(1), 56–76. https://hal.science/hal-01172847

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning*. Springer New York.

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression. In *Wiley Series in Probability and Statistics*.

Thomas, L., Edelman, D., & Crook, J. (2002). Credit scoring and its applications. In *Monographs on Mathematical Modeling & Computation*.