TSE - Scoring Project

Introduction

Louis Olive

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

January 22, 2025

Outline

Outline

- Corporate bankruptcies
- Scoring
- (Warm Up) Desbois case study
- (Some words about) The project

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

Original 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.

© 2025 S&P Global.

Source: S&P Global Market, January 7 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

Original 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

GROUPE BPCE

Source: BPCE L'Observatoire, January 8 2025

Corporate bankruptcies - FR - 2024

Alternative source

BANQUE DE FRANCE
EUROSYSTÈME

Nous contacter | Nous trouver | Language : FR

Défaillances d'entreprises - 2024-12

Mise en ligne le 17 Janvier 2025

Les dernières parutions

Défaillances d'entreprises - 2024-11
Défaillances d'entreprises - 2024-10
Défaillances d'entreprises - 2024-09

Toutes les parutions

À fin décembre, la hausse du nombre de défaillances cumulé sur douze mois poursuit sa décélération

À la fin décembre 2024, en cumul sur douze mois, le nombre de défaillances (en données provisoires) est de 65 764 contre 65 298 à la fin novembre (cf. graphique 1).

Cette progression est en partie liée à un effet de rattrapage, après le fort ralentissement des défaillances pendant la période covid (2020-2021).

En décembre, en rythme annuel, la progression des défaillances est toutefois moindre qu'en novembre (16,8 % contre 18,8 % en novembre) dans l'ensemble des secteurs et pour la plupart des tailles d'entreprises (cf. tableaux A, B et graphique 2). Le nombre d'ETI et de grandes entreprises en défaillance se stabilise mais demeure plus élevé que sa moyenne pré-pandémique.

Les services publics et ceux de la Banque de France demeurent mobilisés auprès des entreprises en difficulté.

Pour en savoir plus : les données sur les créations d'entreprises sont diffusées par l'INSEE notamment ***Créations d'entreprises sur le site de l'INSEE***

Source: Banque de France, January 17 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:

$$\underset{f}{\operatorname{argmin}}\, \mathrm{R}(f)$$

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 - $\boldsymbol{k = 7}$

We show below the boundary decision for $\boldsymbol{k = 7}$:

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 $x_1, \ldots x_n$. 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

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

don_desbois <- read.spss("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

```
########## 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

```
########## 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

```
########## 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 as a classifier (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 $\boldsymbol{i}_0$ 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

```
########## 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

```
######### 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 build with first component of PCA using a ROC curve (on training set):

▼ Code

```
########## 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

```
########## 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

```
########## 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

		Coronary Heart Disease		
Age Group	n	Absent	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.

(Some words about) The project

The project

Syllabus - Part A

Warm up: today

- Defining Scoring / ROC curves
- Using the Desbois case study as an introduction

Part A: Using proprietary data sources, a set of tasks to perform in group mimicking a “real life” project about corporate bankruptcies of public firms in the US:

- **Practical 1**: Defining the problem / Building the data set [Getting /Cleansing/Enriching Data] / Exploration / fitting a Baseline model / model Evaluation
- **Practical 2**: “Expert” approach: Implementing ideas from academic literature on credit risk (involving in particular feature engineering)
- **Practical 3**: “Off the shelf” approach: lasso for variable selection, gradient boosting

Classes 29 January, 3, 5, 10, 11 and 12 February 14:00-17:00.

Groups composition **due 31 January** (ideally 3 students per group). A short report tracking current code/analysis **due 16 February 23:59**.

The project

Syllabus - Part (B &) C

Part B (to be discussed / adapted / removed): in parallel to part A a selection of topics will be proposed with the goal to assess alternative methods (projects choice due 12 February):

- the reading of selected papers,
- the understanding of the methods or algorithms at stake (if necessary re-implement from scratch a simplified version, develop intuition on toy data sets, give short proofs / overview of key results),
- the implementation of the methods on the real life data set,
- the assessment of the method.

An example will be given in class for the two first bullet points to give you an idea. I will show how to build from scratch a Decision Tree, and expand it to AdaBoost then Gradient boosting machines (`gbm`) or NewtonBoosting (`xgboost`). You will then implement the methods on your data set.

Part C:

- the writing of a final report summarizing part A / (implementing part B): Final project reports **due 16 March 23:59**,
- a final presentation (by group / one-to-one meeting): **19 March 8:00-11:00**, more details later.

The project

Data needed for corporate bankruptcy prediction

The variable of interest:

- **Bankruptcy** data: Historical bankruptcy filings potentially augmented with other sources on default data.

The predictors:

- **Financial/Accounting** data : Time series (usually annual or quarterly) of accounting metrics like income and cash-flow statements, balance sheets and derived financial ratios (e.g. $r_1, \ldots, r_{37}$ from Desbois case study).

- In case of public/listed firms, **Market** data: Stock prices, returns, volatility…

- Other: additional **Macroeconomic** or **Industry/Sector**-specific indicators.

The project

The data at hand - WRDS

We will mainly use proprietary data available through **WRDS** (Wharton Research Data Services):

- **WRDS** is an online platform that provides access to a wide range of financial, economic, and corporate data for academic and professional research. It serves as a central hub for databases like **Compustat** and **CRSP**.

- **Compustat** offers financial and accounting data on publicly traded companies, including balance sheets, income statements, and other fundamental metrics. It's widely used for company performance analysis and financial modeling.

- **CRSP** (Center for Research in Security Prices) provides stock market data, such as prices, returns, and trading volumes, for U.S. companies.

A lot of academic/practitioners literature about bankruptcy prediction uses this source allowing comparison/inspiration.

The project

The data at hand - WRDS

Bankruptcy data:

- In principle, most bankruptcies (Chapter 7/ Chapter 11) in the U.S. between 1960 and 2023 for public (publicly traded) companies should be present in **Compustat/CRSP**. In practice, many bankruptcies are lacking, often masked by other corporate events (buyout, merger, stock delisting, etc). Other default events (payment default, distressed debt) non tracked.

Predictors:

- Firm accounting data from **Compustat**. For US companies, annual history is available back to 1950 and quarterly histories date back to 1962.
- US stocks data from the Center for Research in Security Prices (**CRSP**), available daily.
- It is possible to link **Compustat/CRSP**. One of the most widely used source for asset and financial data used in academic setting.
- Other Macroeconomic indicators available through online sources (e.g. FRED)

The project

WRDS with R

A good reference about **WRDS** / **Compustat** / **CRSP** usage with **R** or **Python** is the online book Tidy Finance with R.

In particular the chapter WRDS, CRSP, and Compustat gives solid bases and tricks to allow a fast kickstart with WRDS data.

A second very good reference also using **R** is Empirical Research in Accounting: Tools and Methods, especially chapter 6 / 7 / 8.

Note that remote access is no longer available, so that we will use a downloaded data snapshot. See next slide

The project

*WRDS with R - local version of data in **wrds_data** folder*

For example selecting company **Apple Inc.** and computing Altman Z-score financial ratios:

```r
library(tidyverse)

# Starting 31 Dec 2024, no more TSE access to WRDS
# Instead we work with a downloaded snapshot
# (mainly .rds files mimicking some WRDS/Compustat/CRSP database tables)

funda_db <- readRDS("wrds_data/compustat_all_light.rds")
funda_db %>%
  filter(grepl('APPLE INC', conm)) %>%
  select(gvkey, fyear, conm, at, wcap, re, ebit, lt, sale) %>%
  mutate(WCTA = wcap / at,
         RETA = re / at,
         EBTA = ebit / at,
         TLTA  = lt / at, # as a proxy for ME/TL
         SLTA = sale / at)
```

```
# A tibble: 52 × 14
   gvkey  fyear conm       at   wcap     re  ebit     lt  sale  WCTA  RETA  EBTA
   <chr>  <int> <chr>   <dbl>  <dbl>  <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl>
 1 001690  1980 APPLE…   65.4   16.3   14.5  23.6   39.4  117. 0.250 0.222 0.361
 2 001690  1981 APPLE…  255.   157.    54.1  66.1   77.5  335. 0.615 0.212 0.260
 3 001690  1982 APPLE…  358.   231.   116.  102.   101.   583. 0.647 0.324 0.286
 4 001690  1983 APPLE…  557.   340.   194.  130.   179.   983. 0.611 0.349 0.233
 5 001690  1984 APPLE…  789.   432.   256.   91.4  324.  1516. 0.548 0.324 0.116
 6 001690  1985 APPLE…  936.   527.   316.  147.   386.  1918. 0.563 0.337 0.157
 7 001690  1986 APPLE… 1160.   712.   467.  274.   466.  1902. 0.614 0.403 0.236
 8 001690  1987 APPLE… 1478.   829.   573.  371.   641.  2661. 0.561 0.387 0.251
 9 001690  1988 APPLE… 2082.   956.   777.  620.  1079.  4071. 0.459 0.373 0.298
10 001690  1989 APPLE… 2744.  1399.  1170.  634.  1258.  5284. 0.510 0.427 0.231
# ℹ 42 more rows
# ℹ 2 more variables: TLTA <dbl>, SLTA <dbl>
```

The project

WRDS with R - assuming remote access

```r
# # Before 31 Dec 2024, assuming you were registered to WRDS, remote access to WRDS databases was allowed through R/Python

# library(dbplyr)
# library(RPostgres)

# # First create two environment variables to connect wrds in a terminal: touch $HOME/.Renviron
# # inside the .Renviron file
# # wrds_user = your_user
# # wrds_password = your_password

# wrds <- dbConnect(Postgres(), host = "wrds-pgdata.wharton.upenn.edu", dbname = "wrds", port = 9737, sslmode = "require",
# user = Sys.getenv("wrds_user"), password = Sys.getenv("wrds_password"))
# compustat_list <- dbListObjects(wrds, Id(schema = "comp"))

# # Use dplyr verbs with a remote database table (https://dbplyr.tidyverse.org/reference/tbl.src_dbi.html)
# funda_db <- tbl(wrds, in_schema("comp", "funda"))
funda_db %>%
  filter(grepl('APPLE INC', conm)) %>%
  select(gvkey, fyear, conm, at, wcap, re, ebit, lt, sale) %>%
  mutate(WCTA = wcap / at,
         RETA = re / at,
         EBTA = ebit / at,
         TLTA  = lt / at, # as a proxy for ME/TL
         SLTA = sale / at)
```

```
# A tibble: 52 × 14
   gvkey  fyear conm       at   wcap     re  ebit     lt  sale  WCTA  RETA  EBTA
   <chr>  <int> <chr>   <dbl>  <dbl>  <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl>
 1 001690  1980 APPLE…   65.4   16.3   14.5  23.6   39.4  117. 0.250 0.222 0.361
 2 001690  1981 APPLE…  255.   157.    54.1  66.1   77.5  335. 0.615 0.212 0.260
 3 001690  1982 APPLE…  358.   231.   116.  102.   101.   583. 0.647 0.324 0.286
 4 001690  1983 APPLE…  557.   340.   194.  130.   179.   983. 0.611 0.349 0.233
 5 001690  1984 APPLE…  789.   432.   256.   91.4  324.  1516. 0.548 0.324 0.116
 6 001690  1985 APPLE…  936.   527.   316.  147.   386.  1918. 0.563 0.337 0.157
 7 001690  1986 APPLE… 1160.   712.   467.  274.   466.  1902. 0.614 0.403 0.236
 8 001690  1987 APPLE… 1478.   829.   573.  371.   641.  2661. 0.561 0.387 0.251
 9 001690  1988 APPLE… 2082.   956.   777.  620.  1079.  4071. 0.459 0.373 0.298
```

```
10 001690  1989 APPLE… 2744.  1399.  1170.  634.  1258.  5284. 0.510 0.427 0.231
# ℹ 42 more rows
# ℹ 2 more variables: TLTA <dbl>, SLTA <dbl>
```

The project

Bankruptcy data - LoPucki

The Florida-UCLA-LoPucki Bankruptcy Research Database (BRD): data collection and linking (with Compustat ID) making bankruptcy data available to academic researchers throughout the world.

The project

Bankruptcy data - LoPucki

Florida-UCLA-LoPucki Bankruptcy Research Database 1-12-2023.xlsx

	A	Z	AU	AV	BT	DY	DZ	FE	FF	FI
1	NameCorp	Chapter	Cusip6	Cusip9	DateFiled	GvkeyBefore	GvkeyEmerging	NameEmerging	NetIncomeBefore	Prepackaged
3	21st Century Oncology Holdings, Inc.	11	90131G	107	5/25/2017	265008		21st Century Oncology Holdings,	-134.849	prenegotiated
4	A. H. Robins Company, Inc.	11	770706	109	8/21/1985	009183		A.H. Robins	-461.613	free fall
5	A. M. Castle & Co.	11	148411	101	6/18/2017	002811	002811	A. M. Castle & Co.	-107.982	prepackaged
6	A123 Systems, Inc.	11	03739T	108	10/16/2012	180636		A123 Systems LLC	-257.730	free fall
7	AAC Holdings, Inc.	11	000307	108	6/20/2020	178698		AAC Holdings, Inc.	-59.404	prenegotiated
8	aaiPharma Inc.	11	00252W	104	5/10/2005	063604		aaiPharma Inc.	-191.171	free fall
9	ABC-NACO, Inc.	11	00287P	105	10/18/2001	029395		Meridian Rail Corp.	-56.201	free fall
10	AbitibiBowater Inc.	11	003687	209	4/16/2009	002337	002337	AbitibiBowater Inc.	-490.000	free fall
11	Acceptance Insurance Companies Inc.	11	004308	102	1/7/2005	008173		not applicable	-43.308	free fall
12	Accuride Corporation	11	004398		10/8/2009	145979	145979	Accuride Corporation	-328.266	prenegotiated
13	Aceto Corporation	11	004446	100	2/19/2019	001094		liquidated	-316.121	free fall
14	Acme Metals, Incorporated	11	004724	100	9/28/1998	012569		liquidated	-77.205	free fall
15	ACT Manufacturing, Inc.	11	000973	107	12/21/2001	031661		liquidated	28.939	free fall
16	Acterna Corp.	11	00503U	105	5/6/2003	004132		Acterna Inc.	-408.160	prenegotiated
17	Adelphia Business Solutions, Inc.	11	006847	107	3/27/2002	106881		Telcove	-309.824	free fall
18	Adelphia Communications Corp.	11	006848	105	6/25/2002	012484		liquidated	-547.568	free fall
19	Adeptus Health Inc.	11	006855	100	4/19/2017	020672		Adeptus Health Inc.	13.217	free fall
20	Advanced GlassFiber Yarns LLC	11	00759C		12/10/2002	999999		AGY Holding Corp.		free fall
21	Advanced Radio Telecom Corp.	11	31865X	106	4/20/2001	063942	063942	First Avenue Networks, Inc.	-76.772	free fall

The project

Bankruptcy data - Moody's Annual default study

CROSS-SECTOR

MOODY'S
INVESTORS SERVICE

SECTOR IN-DEPTH
1 February 2019

Rate this Research

TABLE OF CONTENTS

Summary	1
Introduction	2
Defaults fell for the second consecutive year	2
Default rates to rise modestly in 2019	9
Loan recoveries to face downward pressure	12
Rating Accuracy Metrics	13
Methodology and Data Sources	15
Appendix	17
Moody's related publications	55

Corporates - Global

Annual default study: Defaults will rise modestly in 2019 amid higher volatility

Summary

We expect the annual default rate for speculative-grade companies that we rate to rise modestly to 3.0% in 2019 from 2.3% in 2018. More challenging market conditions will result in higher defaults in 2019 as economic growth decelerates, geopolitical tensions continue, credit conditions tighten and borrowing cost increases. On the other hand, low refinancing risk and healthy corporate profits will keep defaults low by historical standards.

» **Including investment-grade companies, the annual default rate will rise to 1.5% in 2019 from 1.1% in 2018, according to our forecasts**. This annual rate implies roughly 98 defaults in 2019, which would be higher than the 77 defaults in 2018 but lower than the count of 104 in 2017.

The project

Bankruptcy data - Moody's Annual default study

Data extracted from Moody's Annual reports. Spans 2002-2022. Includes chapter 11, chapter 7, distressed exchange, missed interest and/or principal payment.

Exhibit 19

Moody's-rated 2018 corporate bond and loan defaults* (continued)

Company	Domicile	Default Date	Broad Default Type	Bond (US $ mil)	Loan (US $ mil)	In Jan 2018 Cohort?
Johnston Press plc	United Kingdom	11/16/2018	Bankruptcy	$282	$0	1
Jupiter Resources Inc.	Canada	10/31/2018	Payment Default	$1,100	$0	1
K. Hovnanian Enterprises, Inc.	United States	1/30/2018	Distressed Exchange	$185	$0	1
Legacy Reserves LP	United States	1/5/2018	Distressed Exchange	$187	$0	1
Legacy Reserves LP	United States	9/20/2018	Distressed Exchange	$130	$0	0
MNC Investama Tbk. (P.T.)	Indonesia	5/11/2018	Distressed Exchange	$186	$0	1
Murray Energy Corporation	United States	6/15/2018	Distressed Exchange	$744	$0	1
NCSG Crane & Heavy Haul Corporation	Canada	3/17/2018	Payment Default	$305	$0	1
New Trident Holdcorp, Inc.	United States	4/30/2018	Distressed Exchange	$0	$495	1
Nine West Holdings, Inc.	United States	4/6/2018	Bankruptcy	$705	$885	1
Noble Group Limited	Hong Kong	3/20/2018	Payment Default	$2,306	$1,143	1
Northern Oil and Gas, Inc	United States	5/15/2018	Distressed Exchange	$497	$0	1
NRG REMA LLC	United States	7/9/2018	Payment Default	$205	$0	1
Odebrecht Engenharia e Construcao S.A. (OEC)	Brazil	11/24/2018	Payment Default	$519	$0	1
PaperWorks Industries, Inc.	United States	2/28/2018	Distressed Exchange	$356	$0	1
Parker Drilling Company	United States	12/12/2018	Bankruptcy	$585	$6	1
Philadelphia Energy Solutions R&M LLC	United States	1/21/2018	Bankruptcy	$0	$541	1
Proserv Operations Limited	United Kingdom	5/14/2018	Distressed Exchange	$0	$135	0
Proserv US LLC	United States	5/14/2018	Distressed Exchange	$0	$345	0
Remington Outdoor Company, Inc.	United States	3/25/2018	Bankruptcy	$226	$550	1
Reward Science and Tech. Industry Grp. Co Ltd	China	12/24/2018	Payment Default	$44	$0	1
RGL Reservoir Management Inc.	Canada	1/8/2018	Distressed Exchange	$0	$346	1
Sears Holdings Corp.	United States	3/20/2018	Distressed Exchange	$481	$0	1
Sears Holdings Corp.	United States	10/15/2018	Bankruptcy	$898	$1,006	0
Sterling Mid-Holdings Limited	United Kingdom	12/11/2018	Distressed Exchange	$0	$0	1

Next week

Next week

Practical 1

- Explore WRDS and additional data sources,
- Make assumptions to define the event of interest “bankruptcy” and link it to predictors,
- Check data quality, especially the event of interest,
- Build your own data set (event of interest + predictors) from scratch,
- Fit / Assess a first baseline model, using Altman (1968) predictors, and following Shumway (2001) approach.

Next week(s)

Feature engineering

- Thorough work on bankruptcy event: data (combining WRDS/Lopucki/Moody's), varying event horizon
- Add Financial Ratio as Predictors as in Desbois (see here, here for SAS code translated to R here
- Implementing Distance to Default as a predictor (Practical 2)
- Add Macro, Market data, time series variables (CRSP, FRED)
- Eventually using a more granular Compustat: Quarterly financial data instead of Annual

Appendix / Food for thought

Appendix / Food for thought

Example of an open source anonymized data set

Online **anonymized** bankruptcy prediction data set related to public american companies (1999-2018). Supposely derived from WRDS data.

Available here with companion article here. Can be used for benchmarking purposes.

future internet MDPI

Article

Machine Learning for Bankruptcy Prediction in the American Stock Market: Dataset and Benchmarks

Gianfranco Lombardo [1,†], Mattia Pellegrino [1,†], George Adosoglou [2,†], Stefano Cagnoni [1,†], Panos M. Pardalos [2,†] and Agostino Poggi [1,*,†]

[1] Department of Engineering and Architecture, University of Parma, 43124 Parma, Italy
[2] Department of Industrial and Systems Engineering, University of Florida, Gainesville, FL 32611, USA
* Correspondence: agostino.poggi@unipr.it
† These authors contributed equally to this work.

Abstract: Predicting corporate bankruptcy is one of the fundamental tasks in credit risk assessment. In particular, since the 2007/2008 financial crisis, it has become a priority for most financial institutions, practitioners, and academics. The recent advancements in machine learning (ML) enabled the development of several models for bankruptcy prediction. The most challenging aspect of this task is dealing with the class imbalance due to the rarity of bankruptcy events in the real economy. Furthermore, a fair comparison in the literature is difficult to make because bankruptcy datasets are not publicly available and because studies often restrict their datasets to specific economic sectors and markets and/or time periods. In this work, we investigated the design and the application of different ML models to two different tasks related to default events: (a) estimating survival probabilities over time; (b) default prediction using time-series accounting data with different lengths. The entire dataset used for the experiments has been made available to the scientific community for further research and benchmarking purposes. The dataset pertains to 8262 different public companies listed on the American stock market between 1999 and 2018. Finally, in light of the results obtained, we critically discuss the most interesting metrics as proposed benchmarks for future studies.

check for updates

Citation: Lombardo, G.; Pellegrino, M.; Adosoglou, G.; Cagnoni, S.;

Lacks Market data and due to anonymization impossible to enrich.

Roughly similar to Desbois data set except more data is involved with a time series aspect (some companies are non defaulting for some years then default)

Appendix / Food for thought

Academic literature about US bankruptcy / scoring

- **Altman, E. I. (1968). Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy.** (using LDA / introducing Z-score)
- Ohlson, J. A. (1980). Financial Ratios and the Probabilistic Prediction of Bankruptcy. (using Logistic Regression / O-score)
- **Shumway, T. (2001). Forecasting Bankruptcy More Accurately: A Simple Hazard Model.** (using Logistic Regression / Hazard model with firm-year observation)
- Chava, S. and Jarrow, R. (2004). Bankruptcy Prediction with Industry Effects. (Hazard model using Industry Effects)
- Bharath, S. T., Shumway, T. (2008). Forecasting Default with the Merton Distance to Default Model. (using a Structural model a la Merton/KMV as a predictor, the Distance to Default)
- Christoffersen, B. (2019). Corporate Default Models: Empirical Evidence and Methodological Contributions. Copenhagen Business School (a state of the art Phd thesis, reviewing old methods and implementing more recent among others Lasso, Generalized Additive Models (GAM), Gradient Boosting, frailty models etc)

Available in **literature/bankruptcy_academics** folder or using your ut-capitole library access.

Appendix / Food for thought

Compustat/CRSP data - Shumway (2001)

Search

4 The Data

To compare hazard to static model forecasts, I estimate both hazard and static models and examine their out-of-sample accuracy. Only firms in the intersection of the *Compustat Industrial File* and the *CRSP Daily Stock Return File* for NYSE and AMEX stocks are included in the sample. Firms that began trading before 1962 or after 1992 are excluded. Firms with CRSP SIC codes from 6000 to 6999 (financial firms) are also excluded. Table

Appendix / Food for thought

Bankruptcy data - Shumway (2001)

4.1 Bankruptcy Data

I collected bankruptcy data from the *Wall Street Journal Index*, the *Capital Changes Reporter*, and the *Compustat Research File*. I also searched for firms whose stock was delisted from the NYSE or AMEX in the *Directory of Obsolete Securities* (1993) and *Nexis*. All firms that filed for any type of bankruptcy within five years of delisting are considered bankrupt. The final sample contains 300 bankruptcies between 1962 and 1992.

The variable of interest in the hazard model is firm age. In this paper, a firm's age is defined as the number of calendar years it has been traded on the NYSE or AMEX. So, for example, if a firm began trading on the NYSE in 1964 and then merged in 1965, it would contribute two firm-year observations to the logit model. One observation would give the firm's age as one year and the other would indicate that the firm's age was two years. The dependent variable associated with both of these observations would be equal to zero, indicating no bankruptcy occurred. If the firm filed for bankruptcy, only its second firm-year observation would have a dependent variable value of one.

I use the firm's trading age as the variable to be explained because there is no attractive alternative to measure how long the firm has been a viable enterprise. Since a firm must meet a number of requirements to be listed by an exchange, firms are fairly homogeneous when initially listed. However, a firm can be incorporated as a small speculative concern or as a large holding company, making the firm's age since incorporation less economically meaningful than its age since listing. In the hazard models I estimate, firm age is never statistically significant after controlling for other firm characteristics.

Appendix / Food for thought

Bankruptcy data - Chava & Jarrow (2004)

2. The Bankruptcy Data Base

This section describes our augmented bankruptcy database. For pricing corporate debt and credit derivatives, the relevant economic state of the firm is default where default is defined to be the condition that occurs when a firm has a delayed or missing contractual debt payment. Unfortunately, data on defaults is not readily available. For this reason, instead of defaults, this paper studies bankruptcies where a bankruptcy is defined to occur when a firm makes either a Chapter 7 or Chapter 11 filing.[5] Although a Chapter 11 filing may not necessarily mean that the firm will eventually file for Chapter 7 (liquidation), we include Chapter 11 filings because in these filings there is significant uncertainty about both the timing and magnitude of the final payments to the firms' claim holders. This is consistent with the intent to use these estimates for the pricing of corporate debt and credit derivatives.[6]

This database consists of all bankruptcy filings as reported in the Wall Street Journal Index (1962–1980), The SDC Database (Reorganizations module 1980–1999), SEC filings (1978–1999) and the CCH Capital Changes Reporter. As such, the bankruptcy data includes most of the bankruptcy filings between 1962–1999 of publicly traded companies on either the NYSE, AMEX and NASDAQ stock exchanges. This database is the most comprehensive bankruptcy database available. Indeed, it includes all the firms in Shumway's (2001) database[7] and slightly more firms than in Moody's public firm database (2000b). There are a total of 1461 bankruptcies in our initial sample.

Figure 1 contains a partitioning of our bankruptcies in our database by stock exchange listing. As indicated, over half of the companies that filed for bankruptcy

[5] We focus only on the Chapter 7 and Chapter 11 bankruptcy filings as it is difficult to get information on all the private restructurings that occurred between the lenders and borrowers for the sample period. To our knowledge there is no database that has this information.

[6] The subtle difference between defaults and bankruptcies will prove important in subsequent sections of this paper.

[7] We thank Tyler Shumway for making his dataset available to us.

Appendix / Food for thought

Bankruptcy data - Chava & Jarrow (2004)

chava2004 - Chava, S., & Jarrow, R. A. (20...
Page 8 of 34

PANEL A: BANKRUPTCIES AS PERCENTAGE OF ACTIVE FIRMS DURING 1962-1999

PANEL B: BANKRUPTCIES AS PERCENTAGE OF ACTIVE FIRMS

The following table gives the number of bankruptcies, total number of active firms and the percentage of bankruptcies each year during the sample period of 1962-1999.

YEAR	# of bankrupt firms	# of active firms	% of bankrupt to active firms	YEAR	# of bankrupt firms	# of active firms	% of bankrupt to active firms
1963	0	1294	0.00%	1981	35	4090	0.86%
1964	2	1362	0.15%	1982	36	4699	0.77%
1965	2	1444	0.14%	1983	50	4645	1.08%
1966	1	1536	0.07%	1984	72	5112	1.41%
1967	0	1612	0.00%	1985	73	5240	1.39%
1968	0	1721	0.00%	1986	94	5250	1.79%
1969	2	1842	0.11%	1987	52	5584	0.93%
1970	12	1988	0.60%	1988	80	5911	1.35%
1971	6	2089	0.29%	1989	74	5814	1.27%
1972	11	2208	0.50%	1990	79	5726	1.38%
1973	27	4355	0.62%	1991	69	5689	1.21%
1974	29	4310	0.67%	1992	45	5874	0.77%
1975	13	4215	0.31%	1993	36	6246	0.58%
1976	19	4197	0.45%	1994	29	6971	0.42%
1977	22	4229	0.52%	1995	40	7387	0.54%
1978	19	4182	0.45%	1996	31	7620	0.41%
1979	14	4110	0.34%	1997	41	8023	0.51%
1980	33	4017	0.82%	1998	49	8079	0.61%

Figure 4. Bankruptcies as % of firms listed on NYSE–AMEX–NASDAQ. The above figure

Appendix / Food for thought

Bankruptcy data - Campbell et al. (2008)

I. Data Description

In order to estimate a dynamic logit model, we need an indicator of financial distress and a set of explanatory variables. The bankruptcy indicator we use is taken from Chava and Jarrow (2004); it includes all bankruptcy filings in the *Wall Street Journal* Index, the SDC database, SEC filings, and the CCH Capital Changes Reporter. The indicator equals one in a month in which a firm filed for bankruptcy under Chapter 7 or Chapter 11, and zero otherwise; in particular, the indicator is zero if the firm disappears from the data set for some reason other than bankruptcy such as acquisition or delisting. The data span the months from January 1963 through December 1998. We also consider a broader failure indicator, which equals one if a firm files for bankruptcy, is delisted for financial reasons, or receives a D rating, over the period January 1963 through December 2003.[4]

[4] Typical financial reasons to delist a stock include failures to maintain minimum market capitalization or stock price, file financial statements, or pay exchange fees. Nonfinancial reasons to delist a stock include mergers and minor delays in filing financial statements.

Appendix / Food for thought

Bankruptcy data - Christoffersen (2019)

3.2 Data and Choice of Covariates

Moody's Default Risk Service Database (MDRD) is used to get the firms' default events. We define a default event as a firm which enters into either bankruptcy, bankruptcy section 77, chapter 10, chapter 11, chapter 7, or a prepackaged chapter 11. We also regard the following as default events: A distressed exchange, a dividend omission, a grace-period default, a modification of indenture, a missed interest payment and/or a missed principal payment, payment moratorium, and a suspension of payments. These events are also included by Duffie et al. (2009) from MDRD and are nearly the same events included by Lando et al. (2013). The by far most frequent event is a missed interest payment followed by a chapter 11 bankruptcy and, as some of our events are not terminal, recurrent events can occur. Firms with multiple events typically have an intermediary period in which most would consider the firm as being in a non-normal state and thus not being at risk of entering into default. Thus, we censor a firm until the resolution date provided by

Appendix / Food for thought

Accounting/Market data - Christoffersen (2019)

We use CRSP and Compustat for market data and financial statements, respectively. We lag data from Compustat by 3 months to reflect the typical delay on financial statements, use quarterly data with annualized flow variables when available and otherwise we use yearly data. Data from CRSP is lagged by 1 month to reflect that we only know past market data. Summary

Appendix / Food for thought

Market data - feature engineering - Christoffersen (2019)

- *Distance to default*: Estimated 1-year distance to default. The drift and volatility of the underlying assets are estimated over the past year using the so-called KMV method as in Vassalou and Xing (2004), and we set the debt due in one year to be the short-term debt plus 50% of the long-term debt as is common. We require at least 60 days of market values to estimate the parameters.

  The statistics of our distance to default is comparable to that reported by Vassalou and Xing (2004), which is anticipated as we use the same method and listed U.S. firms.[2] However, we note that a wide range of values have been reported in the literature.[3]

- *Idiosyncratic volatility*: Estimated standard deviation of 1-year past rolling window regressions of daily log return on the value-weighted log market return. We require at least 60 days of returns in the regressions. The variable is used in Shumway (2001) and one motivation is that more volatile firms should have a higher chance of entering into default (e.g., due to more volatile cash flows as argued by Shumway, 2001).

Appendix / Food for thought

Macroeconomic indicators - Christoffersen (2019)

The value-weighted market return we use is the NYSE and AMEX index from CRSP. In terms of macro-variables, we include a market return and treasury bill rate like Duffie et al. (2009, 2007), specifically the value-weighted past 1-year log return of the aforementioned index and 1-year treasury bill rate. All variables are winsorized at 1% and 99% quantile and we carry forward missing covariates for up to 3 months for CRSP-based variables and 1 year for Compustat-based variables.

All of these covariates have appeared in multiple papers before (e.g., Bharath and Shumway, 2008, Chava and Jarrow, 2004, Shumway, 2001). It is deliberate that we use covariates that have previously been used in the literature as the goal of this work is not to seek new covariates.

Appendix / Food for thought

Classical literature about bankruptcy / credit scores

EMPIRICAL WORK

All statistical models of default use data such as firm-specific income and balance sheet statements to predict the probability of future financial distress of the firm. Lenders have always known that financial ratios vary predictably between investment-grade and speculative-grade borrowers. This knowledge manifests itself in many useful rules of thumb, such as not lending to firms with leverage ratios above 70%.

Exhibit 2.1

Empirical Studies of Corporate Default: Year Published and Sample Count

	Year	Defaults	Non-Defaults
Fitzpatrick	(32)	19	19
Beaver	(67)	79	79
Altman	(68)	33	33
Lev	(71)	37	37
Wilcox	(71)	52	52
Deakin	(72)	32	32
Edmister	(72)	42	42
Blum	(74)	115	115
Taffler	(74)	23	45
Libby	(75)	30	30
Diamond	(76)	75	75
Altman, Haldeman and Narayanan	(77)	53	58
Marais	(79)	38	53
Dambolena and Khoury	(80)	23	23
Ohlson	(80)	105	2,000
Taffler	(82, 83)	46	46
El Hennawy and Morris	(83a)	22	22
Moyer	(84)	35	35
Taffler	(84)	22	49
Zmijewski	(84)	40	800
Zavgren	(85)	45	45
Casey and Bartczak	(85)	60	230
Peel and Peel	(88)	35	44
Barniv and Raveh	(89)	58	142
Boothe and Hutchninson	(89)	33	33
Gupta, Rao, and Bagchi	(90)	60	60
Kease and McGuiness	(90)	43	43
Keasey, McGuiness and Short	(90)	40	40
Shumway	(96)	300	1,822
Moody's RiskCalc for Public Companies[7]	(00)	1,406	13,041
Moody's RiskCalc for Private Companies	(00)	1,621	23,089
Median		**40**	**45**

Small samples inhibit the search for a truly superior default model.

7 *Sobehart and Stein (2000)*

Appendix / Food for thought

Selected practitioner literature about bankruptcy / credit scores

- **Bardos, M. (2007) What is at stake in the construction and use of credit scores?**
- Falkenstein, E. and Boral, A. and Carty, L. (2000) Riskcalc for Private Companies: Moody's Default Model.
- Lewis, Susan. (2004) BondScore 3.0: A Credit Risk Model for Corporate Debt Issuers.

Available in **literature/bankruptcy_practitioners** folder or using your ut-capitole library access.

Appendix / Food for thought

About the importance of thorough data collection / exploration

We will use in Practical 2 Distance to Default as implemented by KMV, as **K**ealhofer **M**(cQuown) and **V**asicek

What was missing from the model – and from other famous efforts to measure default risk, such as Ed Altman's Z-score – was an empirical foundation.

"I started looking at the validation work Oldrich had done, the validation work Altman had done on his accounting-based models, and what became clear right away was they just didn't have very much default data. They were trying to evaluate models predicting very rare events, and you need a big sample to do that reliably. I think Altman's original paper had a sample of around 40 defaulted companies – and this was what people were working with at the time," says Kealhofer.

So began months of painstaking labour. Commercial databases were focused on equity markets, says Kealhofer, with defaulting firms often being deemed irrelevant and removed. Legal databases included payment defaults, but many other kinds of default as well. At one point, he recalls, frustrated with the available data, the firm tried going direct to the source – phoning companies up and asking them if they had ever defaulted on a bond payment.

Appendix / Food for thought

About the importance of thorough data collection / exploration

Eventually, Kealhofer amassed data on around 500 corporate defaults. With it, he was able to shed light on something that had stymied widespread use of Merton's model in the 13 years since its publication. Other researchers – Yale's Jon Ingersoll, and a trio of Harvard professors including Eric Rosenfeld – had explored the empirical basis for the model and came back with as many questions as answers. Although recognising Merton's theoretical breakthrough, they found the model didn't seem to work in practice, particularly for investment-grade bonds.

Other firms – and other people – played a part in this revolution, of course, but Kealhofer didn't stop there. After KMV was sold to Moody's Investors Service for $210 million in 2002, Kealhofer moved from loan origination to bond investing. KMV's market-implied default probabilities became the foundation for two decades of further research, and the rules-based credit strategies at his fund, Diversified Credit Investments.

References

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance*, *23*(4), 589–609. http://www.jstor.org/stable/2978933

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

Campbell, J. Y., Hilscher, J., & Szilagyi, J. (2008). In search of distress risk. *The Journal of Finance*, *63*(6), 2899–2939. https://doi.org/https://doi.org/10.1111/j.1540-6261.2008.01416.x

Chava, S., & Jarrow, R. A. (2004). Bankruptcy prediction with industry effects. *Review of Finance*, *8*(4), 537–569. https://doi.org/10.1093/rof/8.4.537

Christoffersen, B. (2019). Corporate default models: Empirical evidence and methodological contributions. In *Copenhagen Business School [Phd]*.

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*.

Shumway, T. (2001). Forecasting bankruptcy more accurately: A simple hazard model. *The Journal of Business*, *74*(1), 101–124.

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