Benefits of better credit scoring

Błażej Kochański

Everyone can have a credit score

Dane identyfikacyjne z wniosku o rejestrację konta

Imiona i nazwisko	**BŁAŻEJ KOCHAŃSKI**

Ocena punktowa aktualna na dzień 19-12-2012

Ocena punktowa w BIK	**568**
Zakres oceny punktowej	**od 192 do 631**
Komentarz do oceny punktowej	**Ocena powyżej średniej dla osób, których dane zgromadzone są w bazie BIK S.A.**
Graficzna prezentacja oceny punktowej	

więcej o Ocenie Punktowej BIK

Ocena punktowa aktualna na dzień 21-12-2012

Ocena punktowa w BIK	**554**
Zakres oceny punktowej	**od 192 do 631**
Komentarz do oceny punktowej	**Ocena powyżej średniej dla osób, których dane zgromadzone są w bazie BIK S.A.**
Graficzna prezentacja oceny punktowej	

więcej o Ocenie Punktowej BIK

… Mine in Biuro Informacji Kredytowej went down by 12 points in just two days. Guess why…

Presentation plan

- Economics of credit scoring
- Credit market modelling 1
- Gini coefficient in R
- Drawing ROC curves
- Credit market modelling 2
- Combining credit scorecards
- Mixing credit scorecards

Economics of credit scoring

Credit scoring management – expenses:

Systems:
- Credit scoring development tools
- Credit scoring engines (implementation in decision systems)
- Credit scoring validation tools, data warehouses, data marts

People:
- Statistics and machine learning professionals
- IT professionals
- Training, knowledge management, HR

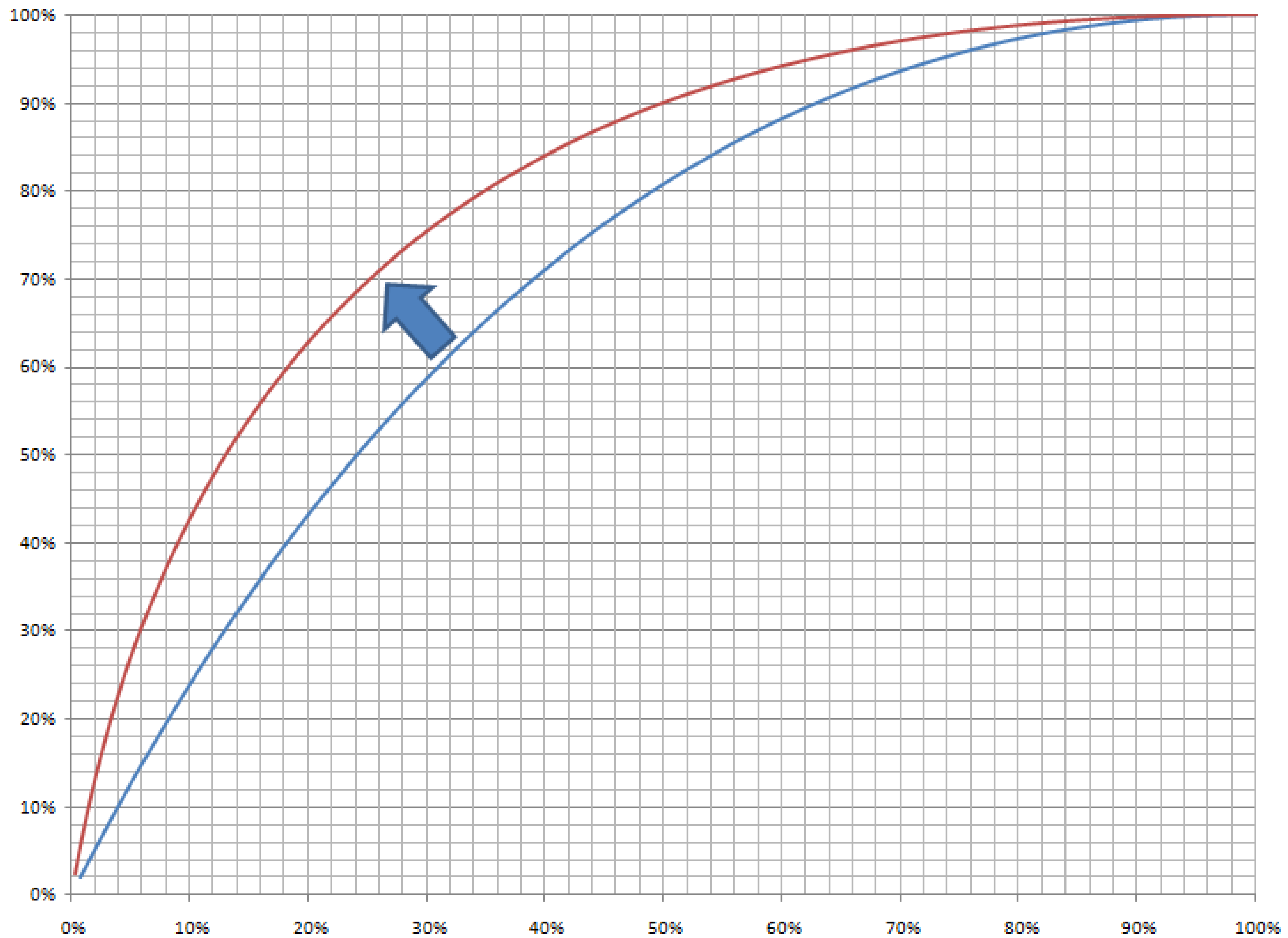
Data:
- gathering and cleaning data
- purchase of new data sources

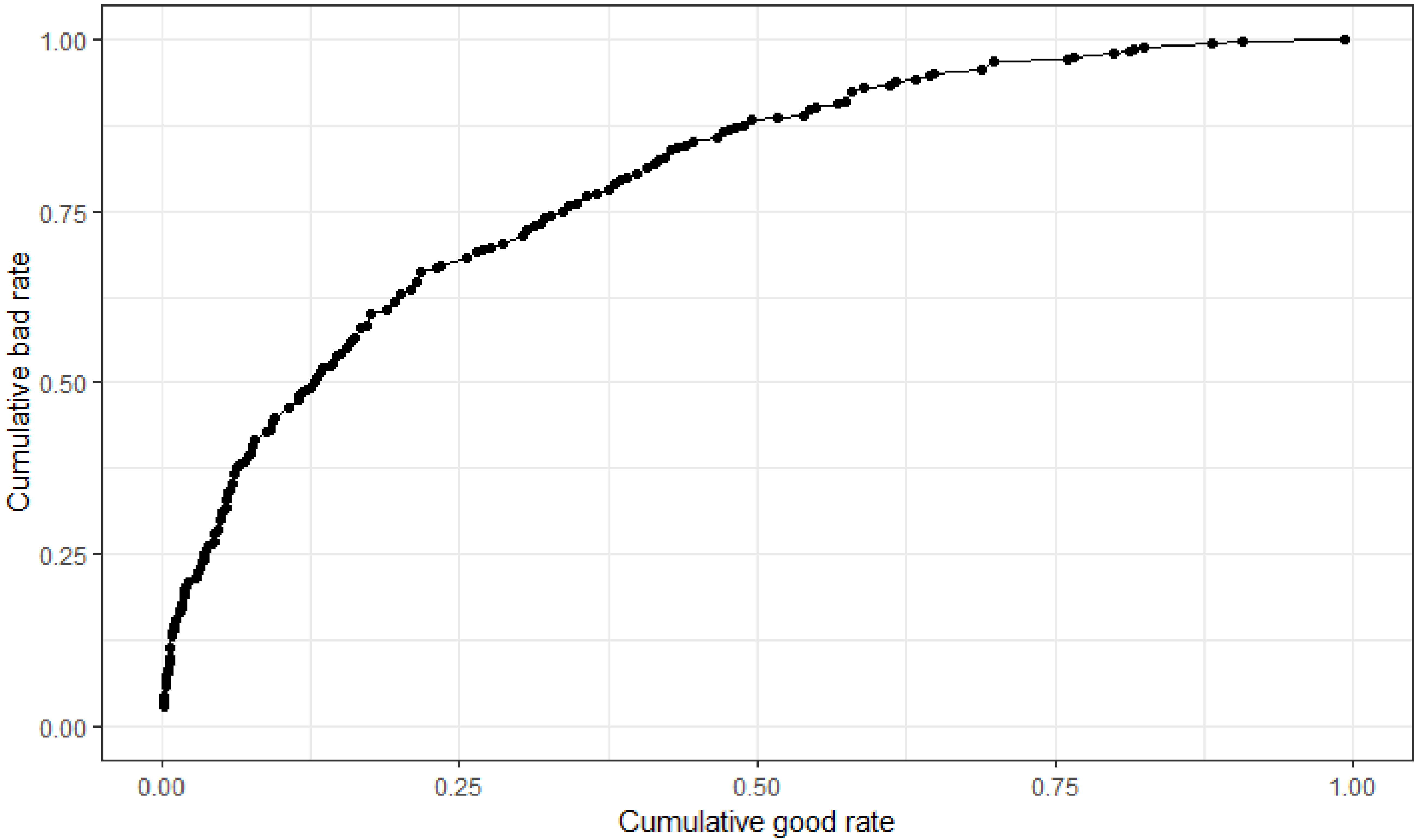
Credit scoring management – benefits:

???

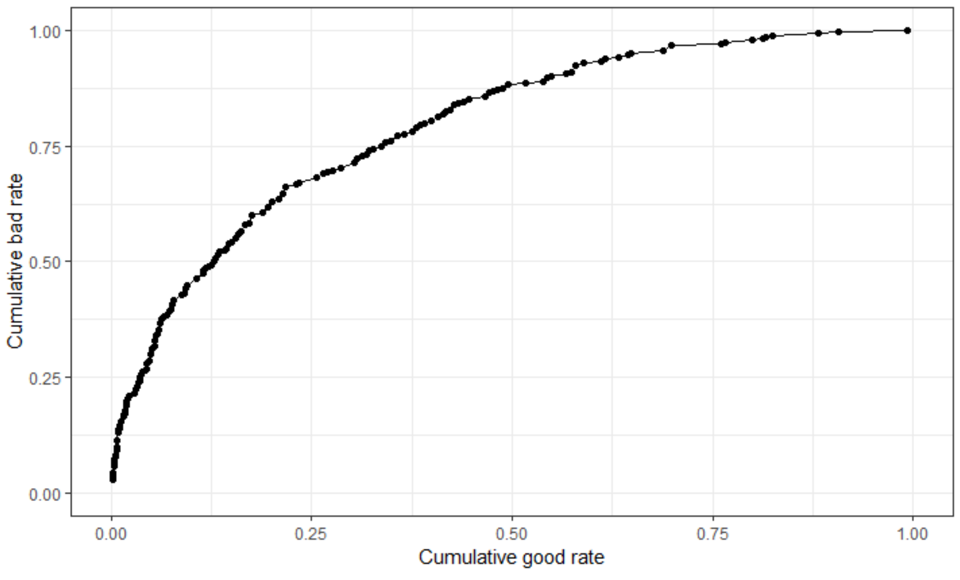
OK. my Gini will go up by 5 pp.

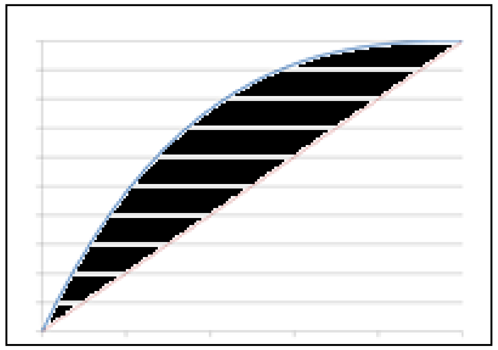
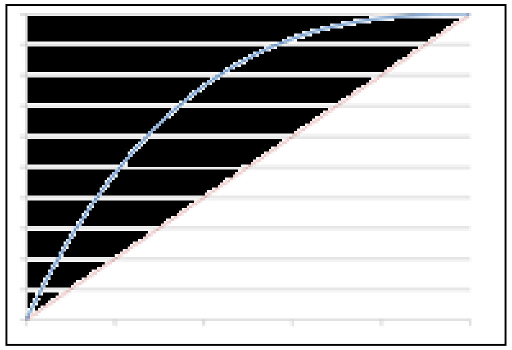
What will my profits be?





„bads" – people not paying loans, often known as „defaults"



GINI = area  **÷ area** 

- **Gini = 2AUC – 1**

- **Gini ranges from 0 to 1**

  (theoretically -1 to 1 but the absolute value matters)

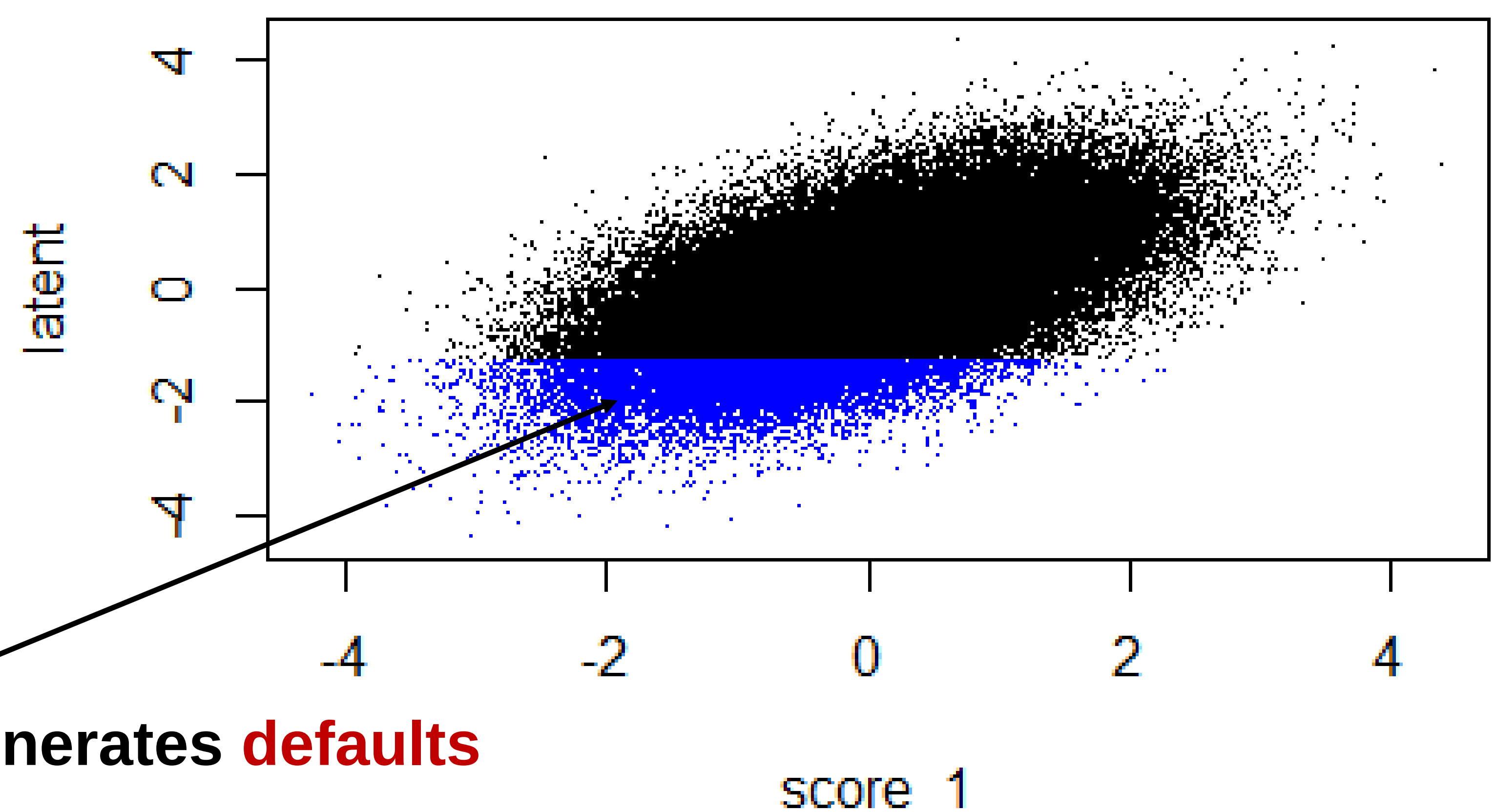
- **Gini has a nice interpretaion (it is Somers' D)**

Presentation plan

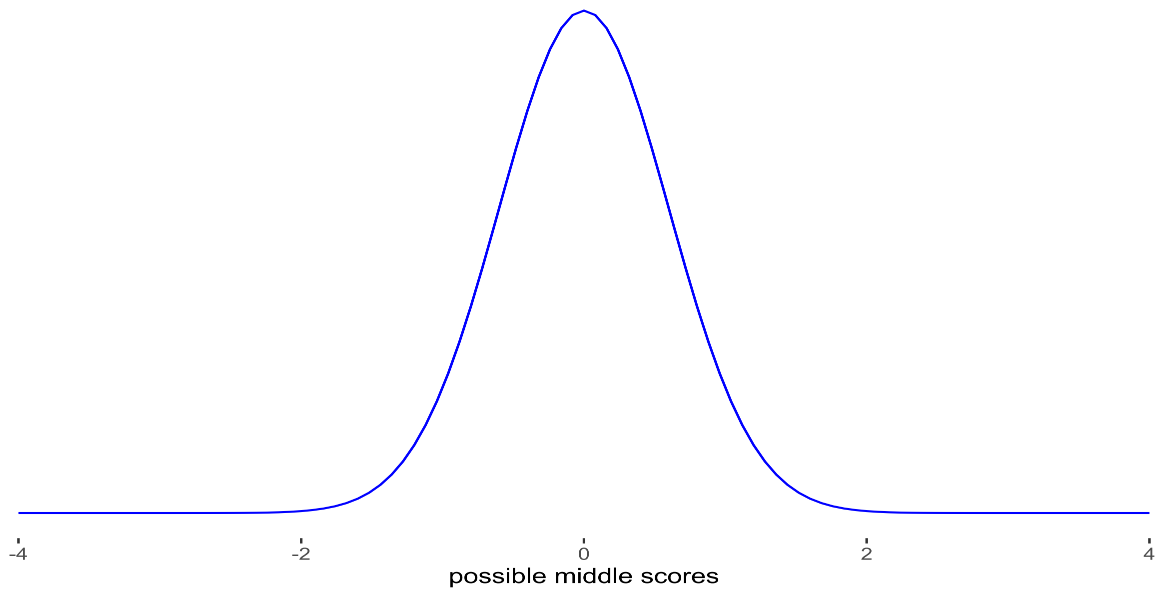
- Economics of credit scoring
- Credit market modelling 1
- Gini coefficient in R
- Drawing ROC curves
- Credit market modelling 2
- Combining credit scorecards
- Mixing credit scorecards

Bivariate normal distribution
– one scoring correlated to latent risk variable

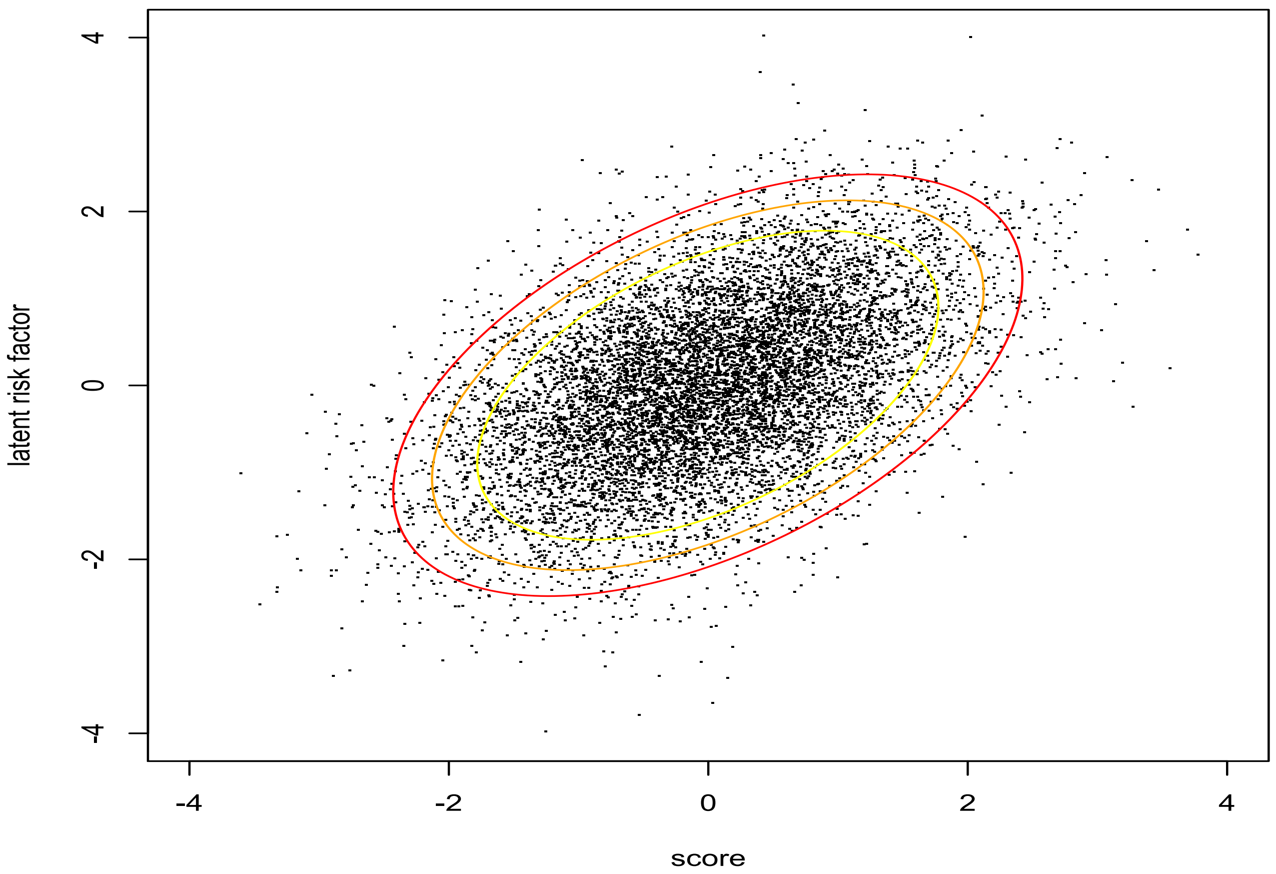
```
N<-100000; rho<-0.56; default_rate<-0.1
score_1<-rnorm(N); latent<-rho*score_1+sqrt(1-
rho^2)*rnorm(N)
default<-(latent<qnorm(default_rate))*1
plot(score_1, latent, pch='.') +
  points(score_1[default==1], latent[default==1],
pch='.', col='blue')
```

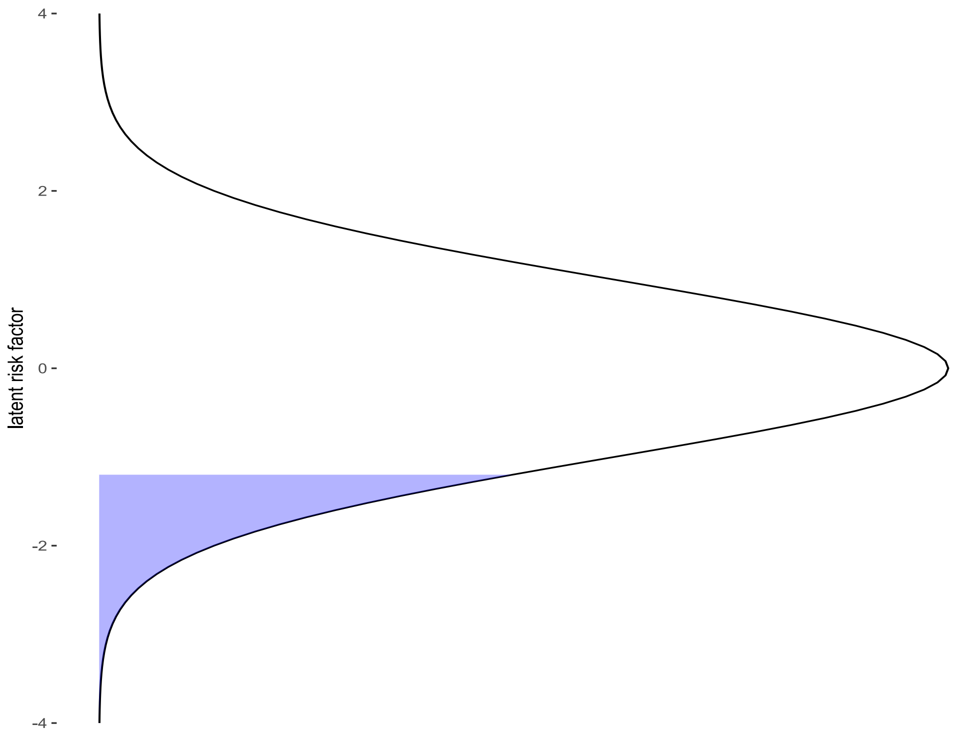


Latent risk variable generates defaults



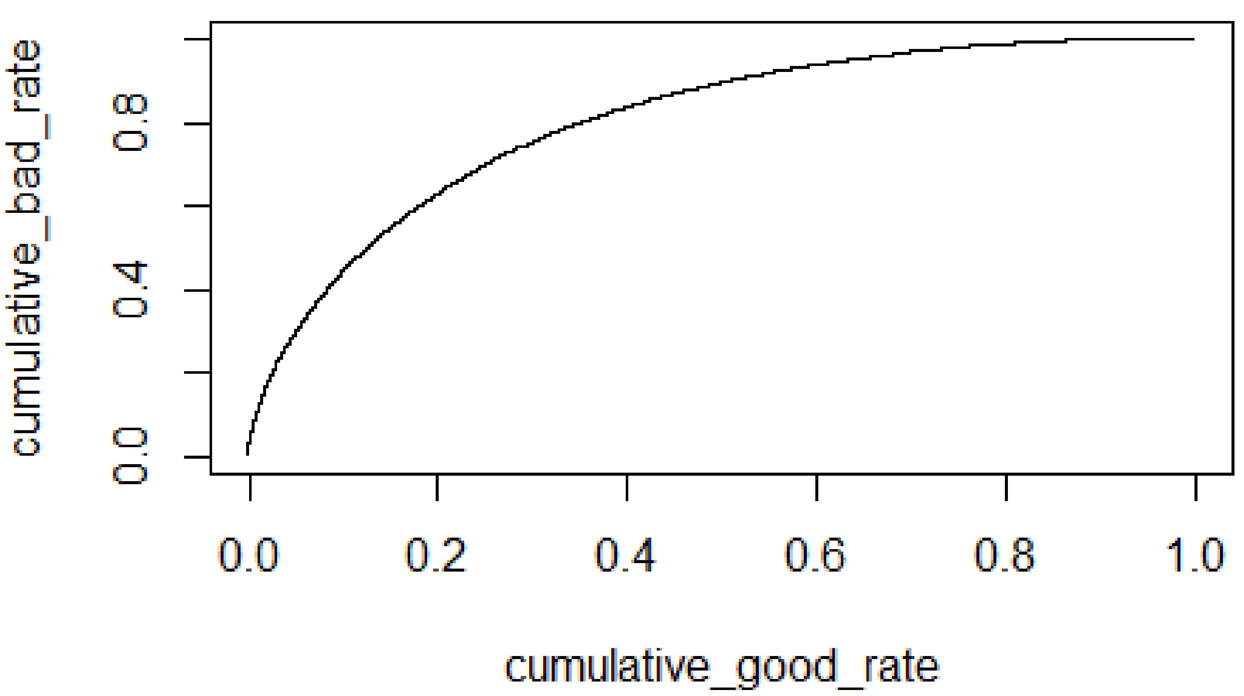
The score is translated into latent risk variable through the bivariate normal distribution with correlation parameter ρ . Latent risk variable, in turn, translates into default flag based on assumed approval rate.





Bivariate normal distribution – ROC curve

```
plot_roc<-function(resp, pred){
  c<-pred[order(pred)]
  d<-resp[order(pred)]
  cumulative_bad_rate<-c(0, cumsum(d)/sum(d))
  cumulative_good_rate<-c(0, cumsum(1-d)/sum(1-d))
  plot(cumulative_good_rate, cumulative_bad_rate,
pch='.')
}
plot_roc(default, score_1)
```



Presentation plan

- Economics of credit scoring
- Credit market modelling 1
- Gini coefficient in R
- Drawing ROC curves
- Credit market modelling 2
- Combining credit scorecards
- Mixing credit scorecards

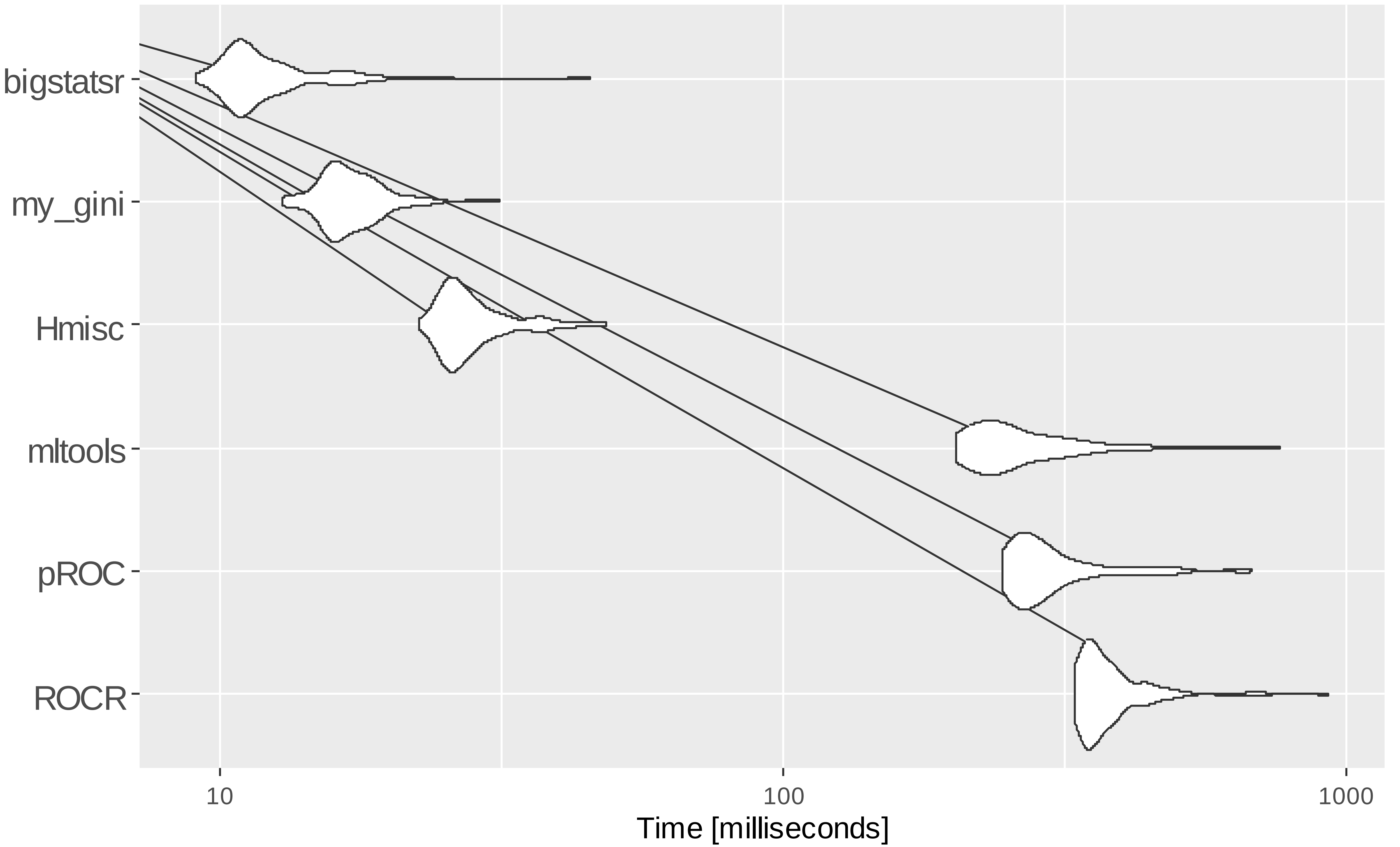
Gini in R – very slow options

```
Hmisc::rcorr.cens(-score_1,default)['Dxy']

##       Dxy
## 0.5997946

53.59 sec elapsed
```

```
InformationValue::somersD(default, -score_1)

## [1] 0.5997946

23.27 sec elapsed
```

Gini in R – reasonable options

```
2*ROCR::performance(ROCR::prediction(-score_1, default),
"auc")@y.values[[1]]-1

2*pROC::auc(default, score_1, lev=c('0', '1'), dir=">")-1

2*mltools::auc_roc(-score_1, default)-1

Hmisc::somers2(-score_1, default)['Dxy']

2*bigstatsr::AUC(-score_1, default)-1

my_gini<-function(resp, pred){
  c<-pred[order(pred)]
  d<-resp[order(pred)]
  bc<-c(0, cumsum(d)/sum(d))
  gc<-c(0, cumsum(1-d)/sum(1-d))
  sum((gc[2:(length(gc))]-gc[1:(length(gc)-1)])*
        (bc[2:(length(bc))]+bc[1:(length(bc)-1)]))-1
}

my_gini(default, score_1)
## [1] 0.5997946
```

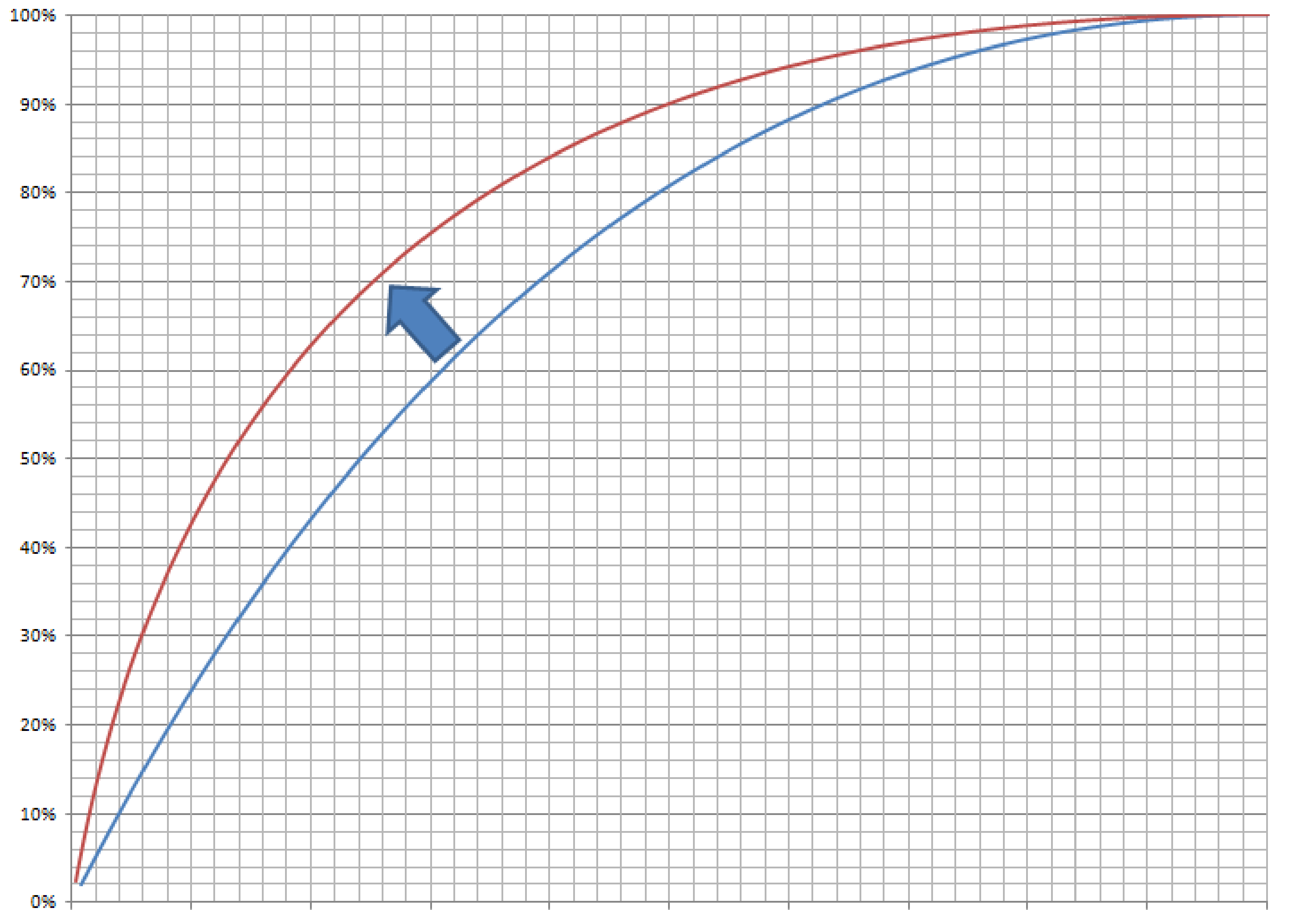

bigstatsr
my_gini
Hmisc
mltools
pROC
ROCR
10
100
1000
Time [milliseconds]

Presentation plan

- Economics of credit scoring
- Credit market modelling 1
- Gini coefficient in R
- Drawing ROC curves
- Credit market modelling 2
- Combining credit scorecards
- Mixing credit scorecards

OK. my Gini will go up by 5 pp.

What will my profits be?



$$y = \beta\left(1 - (1 - x)^{\frac{1+\gamma}{1-\gamma}}\right) + (1 - \beta)\, x^{\frac{1-\gamma}{1+\gamma}}$$

$$y = F_{\alpha_B,\beta_B}(F^{-1}_{\alpha_G,\beta_G}(x))$$

$$y = G_{\alpha_B,\beta_B}(G^{-1}_{\alpha_G,\beta_G}(x))$$

$$y = \Phi\left(\Phi^{-1}\left(\frac{\gamma+1}{2}\right)\sqrt{1+b^2} + b\Phi^{-1}(x)\right)$$

$$y = 1 - (1 - x^{\frac{1}{\alpha_G}})^{\beta_B}$$

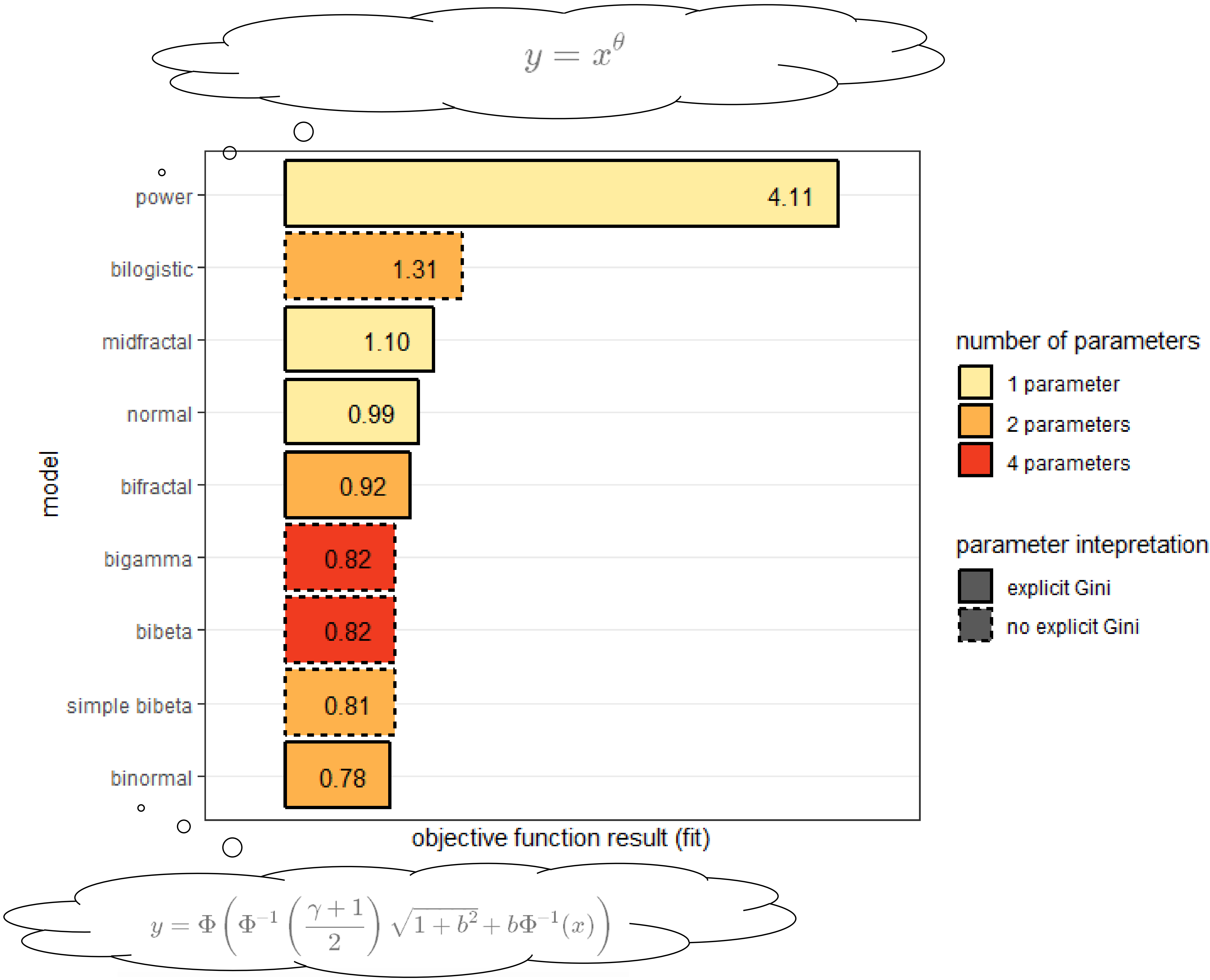
$$y = \Phi(\Phi^{-1}(x) + \sqrt{d})$$

$$y = \Phi\left(\Phi^{-1}\left(\frac{\gamma+1}{2}\right)\sqrt{2} + \Phi^{-1}(x)\right)$$

$$y = x^{\theta}$$

$$y = \left(1 + exp\left(\alpha_1 ln\left(\frac{1}{x} - 1\right) - \alpha_0\right)\right)^{-1}$$

$$y = \frac{1}{2}\left(1 - (1 - x)^{\frac{1+\gamma}{1-\gamma}} + x^{\frac{1-\gamma}{1+\gamma}}\right)$$



$y = x^\theta$
power
bilogistic
midfractal
normal
bifractal
bigamma
bibeta
simple bibeta
binormal
4.11
1.31
1.10
0.99
0.92
0.82
0.82
0.81
0.78
model
objective function result (fit)
number of parameters
1 parameter
2 parameters
4 parameters
parameter intepretation
explicit Gini
no explicit Gini
$y = \Phi\left(\Phi^{-1}\left(\frac{\gamma+1}{2}\right)\sqrt{1+b^2} + b\Phi^{-1}(x)\right)$

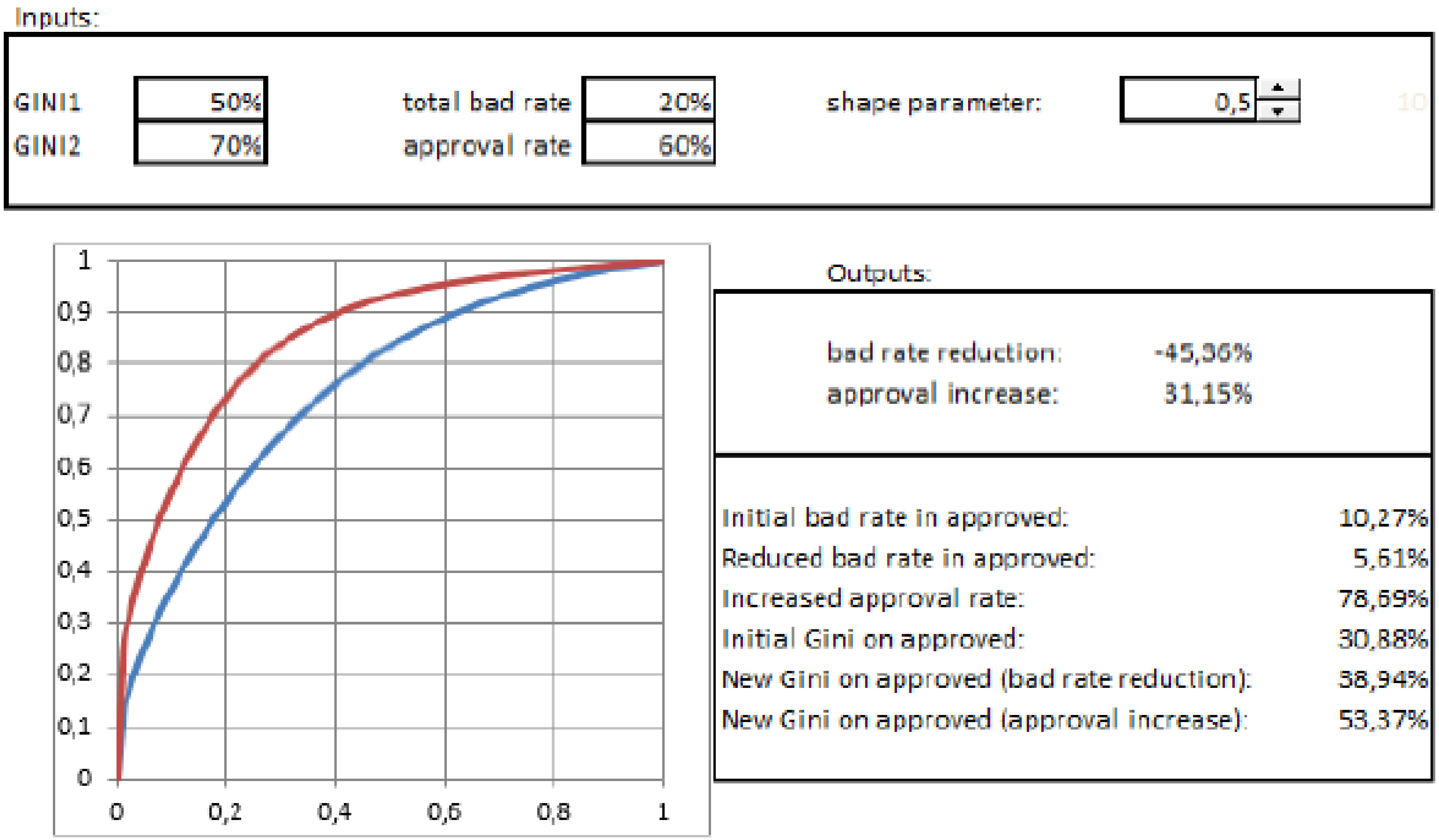
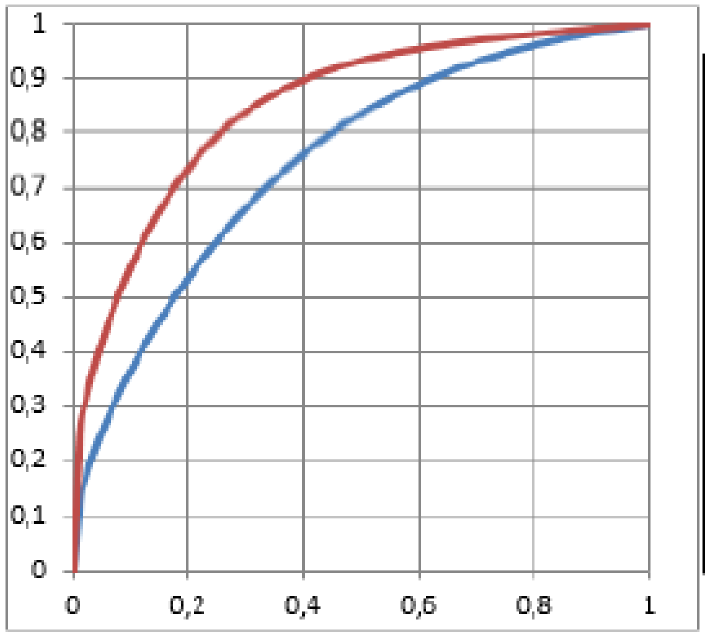


Figure 12: Interface of MS Excel calculation engine enabling modelling impact of Gini change.

Inputs:

GINI1	50%	total bad rate	20%	shape parameter:	0,5
GINI2	70%	approval rate	60%		



Outputs:

bad rate reduction:	-45,36%
approval increase:	31,15%

Initial bad rate in approved:	10,27%
Reduced bad rate in approved:	5,61%
Increased approval rate:	78,69%
Initial Gini on approved:	30,88%
New Gini on approved (bad rate reduction):	38,94%
New Gini on approved (approval increase):	53,37%

Example

GINI 0.50
↗ 0.52

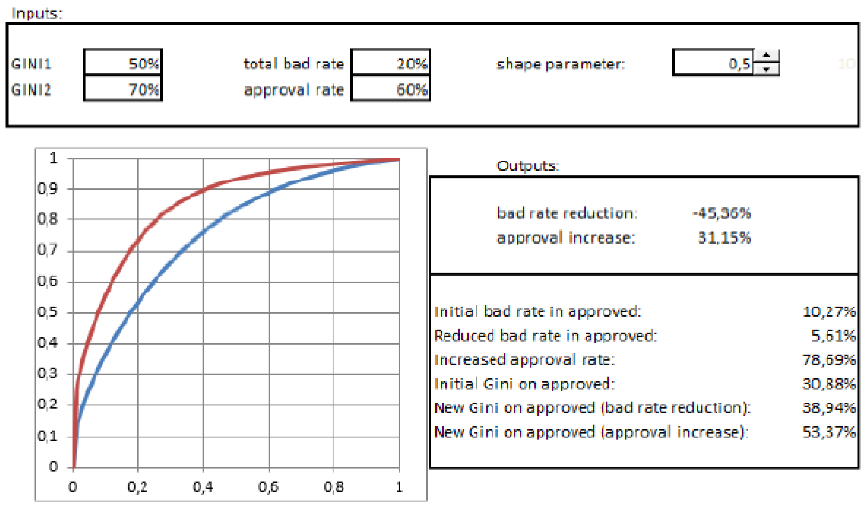
Scored population bad rate = 25%
Approval rate = 40%
Beta = 0.5

Portfolio bad rate
↘ by 6%

Presentation plan

- Economics of credit scoring
- Credit market modelling 1
- Gini coefficient in R
- Drawing ROC curves
- Credit market modelling 2
- Combining credit scorecards
- Mixing credit scorecards

No bank is a lonely island…



Drawing ROC curves model assumes that:

- **There is only one bank, there are no competitors.**
- **All customers take loans take the loans, whatever the price.**

Risk-based pricing cannot be modelled with this approach…

Simulation of many banks environment

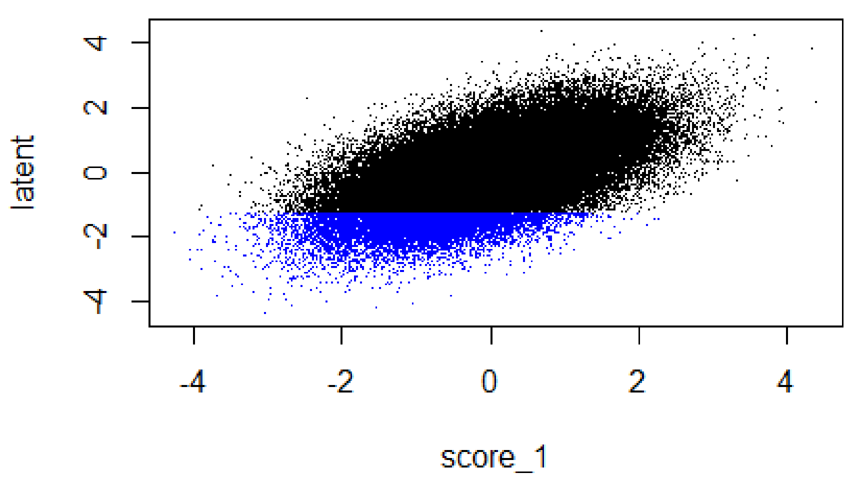
<u>Key simulation assumptions:</u>

10 banks with similar market share and similar separation power of credit scoring

Default rate = 10%.

Banks set their interest rates based on historical default rates by score band in their past (**risk based pricing**). Assumed **profit margin** ~3 pp.

A customer checks **three** banks before making a decision (**loan shopping**)



But now 10 banks - multivariate normal distribution

$$(S_1, S_2, S_3, \dots, S_{10}, Y^*)^T \sim N\ (\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

$S_1, S_2, S_3, \dots, S_{10}$ **- credit scores**

Y^* **- latent risk variable**

$$\boldsymbol{\mu} = (0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)^{\mathrm{T}}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & \rho & \rho & & \rho_1 \\ \rho & 1 & \rho & \cdots & \rho_2 \\ \rho & \rho & 1 & & \rho_3 \\ & \vdots & & \ddots & \vdots \\ \rho_1 & \rho_2 & \rho_3 & \cdots & 1 \end{bmatrix}$$

ρ **= 0.75 (all credit scoring are correlated but not identical)**

$$\rho_1 = \rho_2 = \rho_3 = \cdots = \rho_{10} = 0.5$$

– base scenario correlations with latent risk factor

With ρ_{i}=0.5 at default rate *d*=10%, the Gini coefficient = 0.542

Last random variable (*Y) is a „latent risk factor", not directly observable, but translating into 0/1 observable variable „default event" *Y*:**

$$Y = \begin{cases} 0 & if \quad Y^* \geq \Phi^{-1}(d) \\ 1 & if \quad Y^* < \Phi^{-1}(d) \end{cases}$$

$(S_1, S_2, S_3, \ldots, S_{10}, Y^*)^T \sim N\ (\boldsymbol{\mu}, \boldsymbol{\Sigma})$

$S_1, S_2, S_3, \ldots, S_{10}$ **- credit scores**

Y^* **- latent risk variable**

$$\boldsymbol{\mu} = (0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)^{\mathrm{T}}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & \rho & \rho & & \rho_1 \\ \rho & 1 & \rho & \cdots & \rho_2 \\ \rho & \rho & 1 & & \rho_3 \\ & \vdots & & \ddots & \vdots \\ \rho_1 & \rho_2 & \rho_3 & \cdots & 1 \end{bmatrix}$$

ρ **= 0.75 (all credit scoring are correlated but not identical)**

$\boldsymbol{\rho_1} = \mathbf{0.6}$

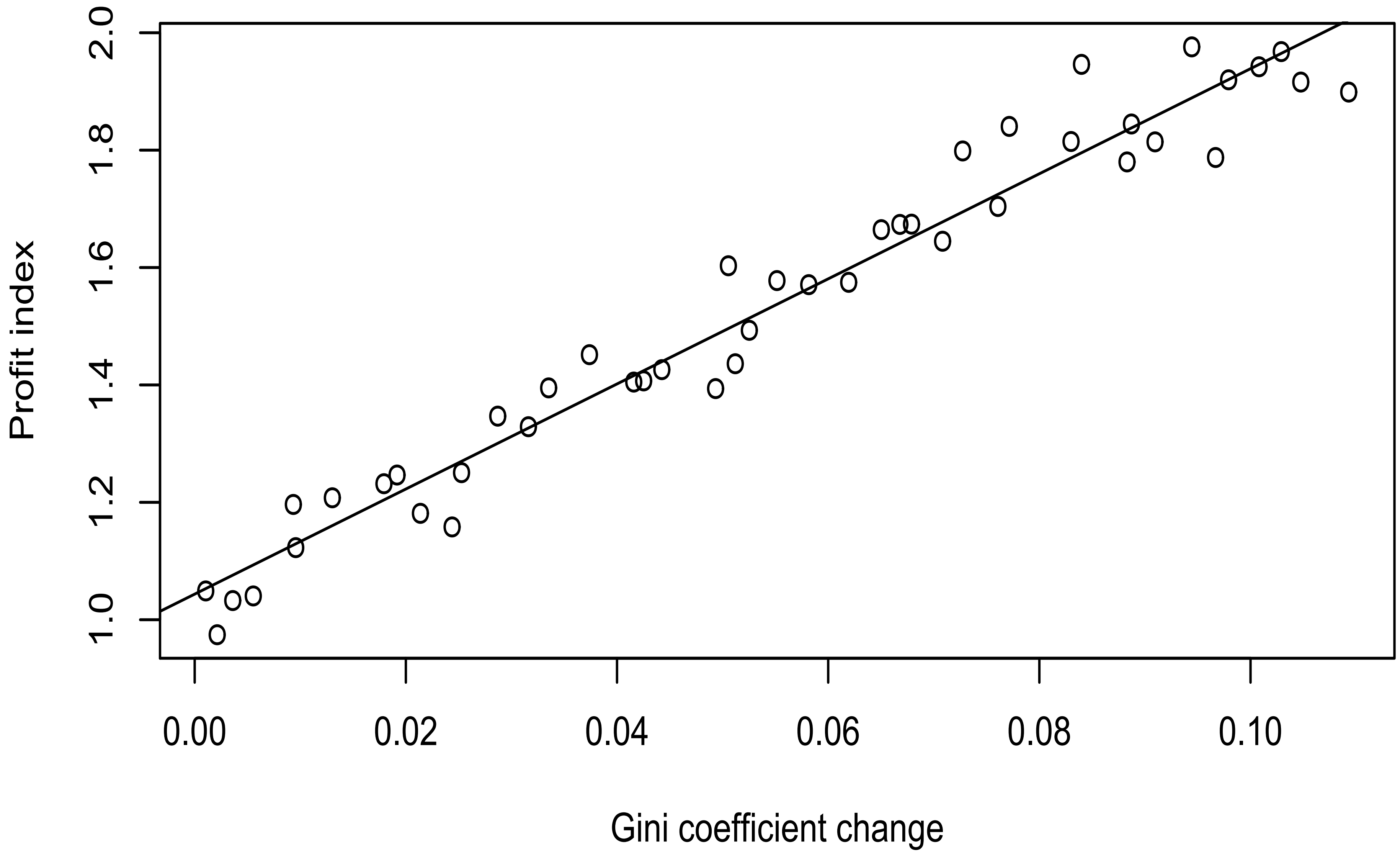
$\rho_2 = \rho_3 = \cdots = \rho_{10} = 0.5$

With ρ_{i}=0.6 at default rate d=10%, the Gini coefficient = 0.645

Last random variable (Y^*) is a „latent risk factor", not directly observable, but translating into 0/1 observable variable „default event" Y:

$$Y = \begin{cases} 0 & if \quad Y^* \geq \Phi^{-1}(d) \\ 1 & if \quad Y^* < \Phi^{-1}(d) \end{cases}$$

Simulation results – profit increase vs Gini coefficient increase.



1 percentage point of Gini increase => ~ 9% profit increase.

1 percentage point increase in Gini*

may have a huge financial impact on the bank

*** = half percentage point in AUC**

Presentation plan

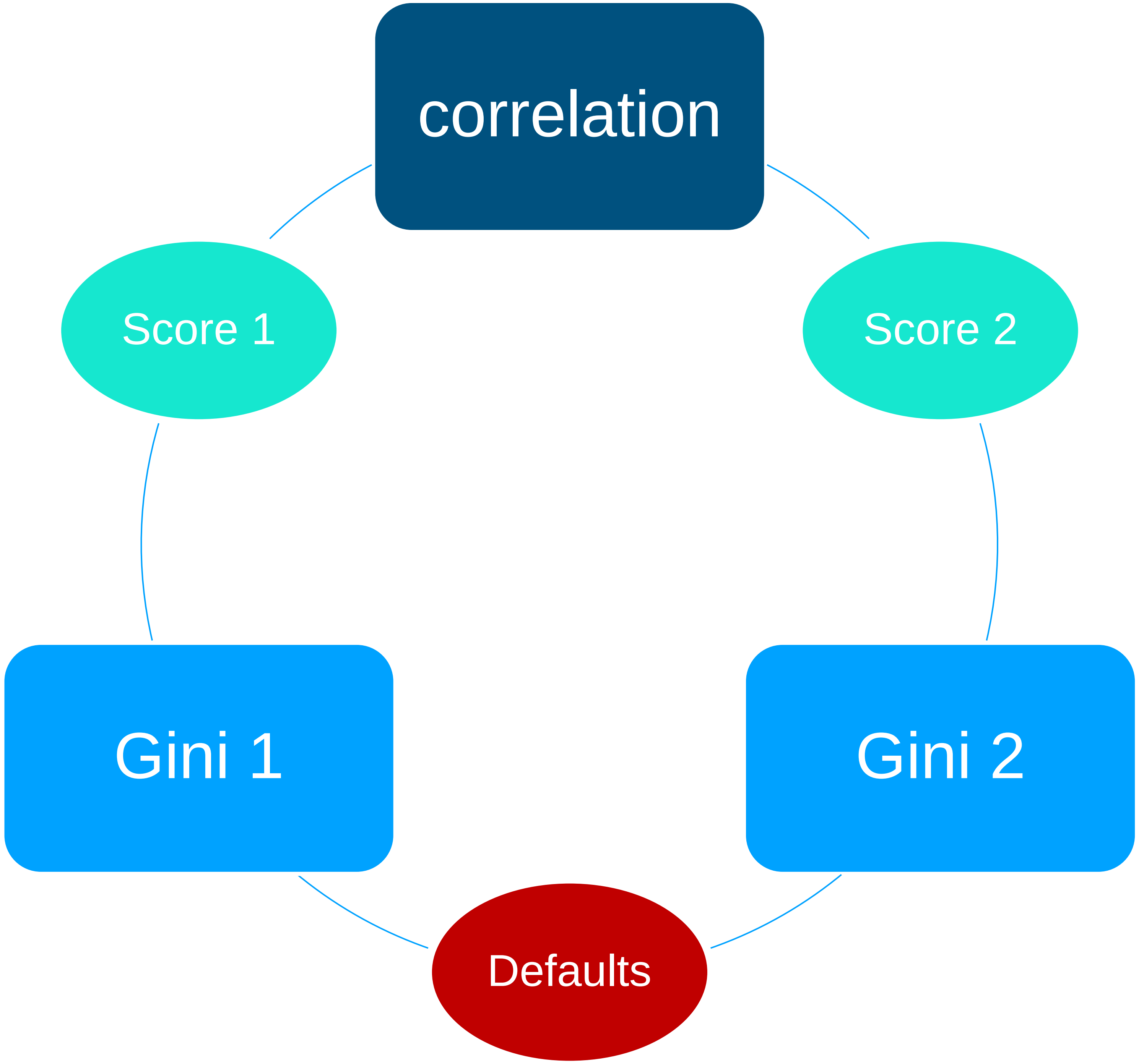
- Economics of credit scoring
- Credit market modelling 1
- Gini coefficient in R
- Drawing ROC curves
- Credit market modelling 2
- Combining credit scorecards
- Mixing credit scorecards

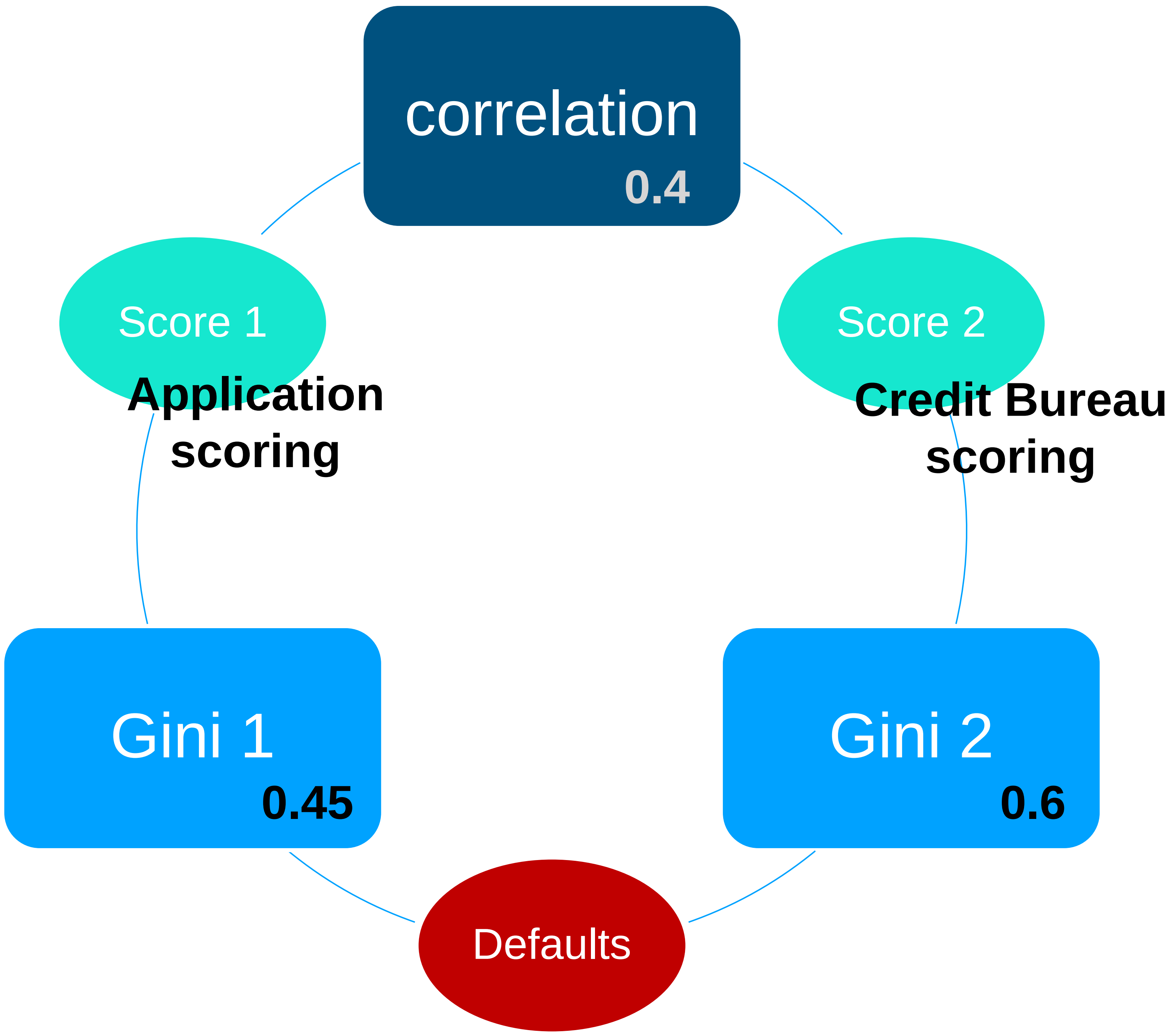
Case 1: Introduction of Credit Bureau scoring in a lending institution.

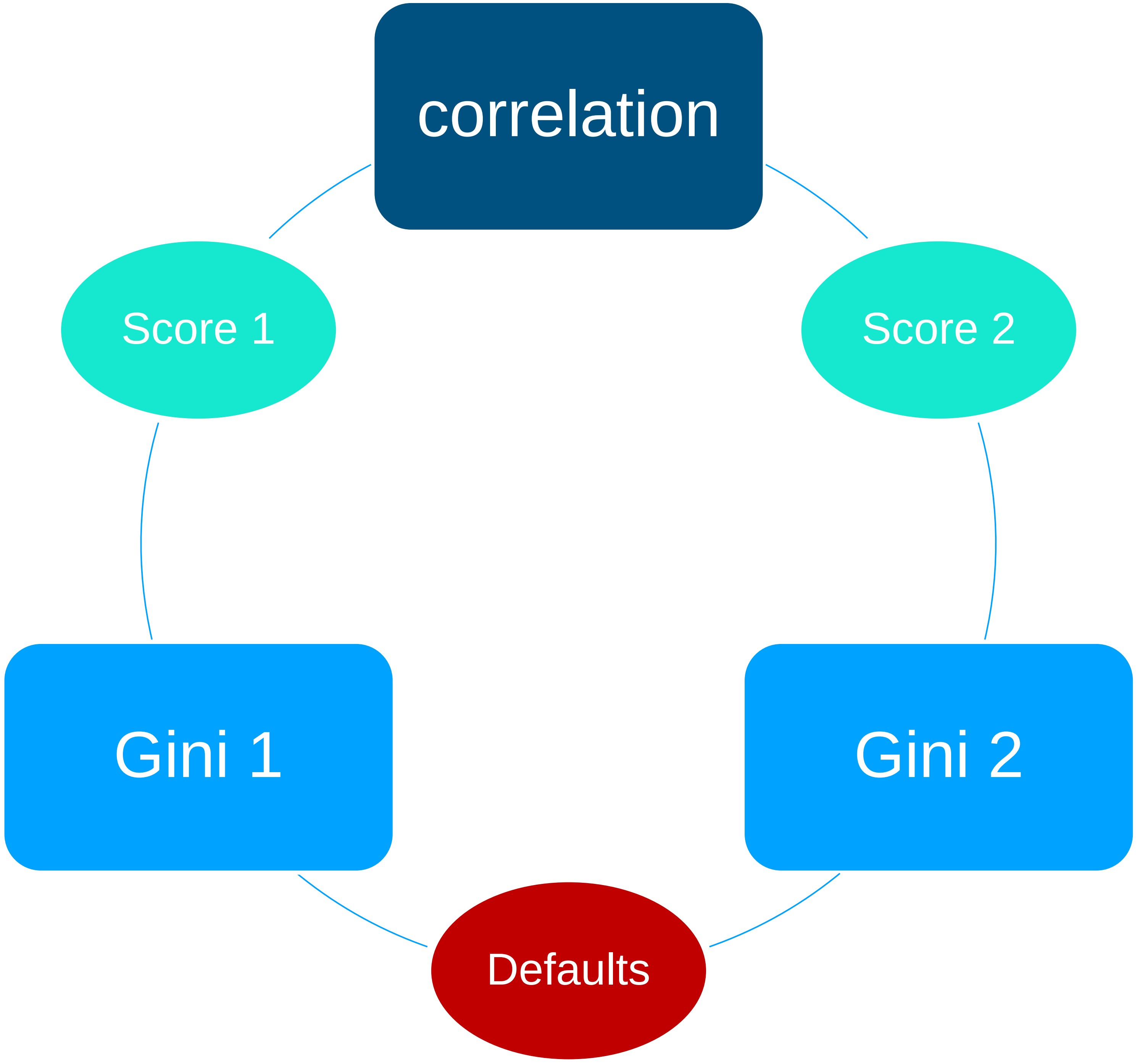
- Gini of existing application scorecard = 0.45 — Gini 1
- Advertised Gini of Credit Bureau scorecard = 0.60 — Gini 2
- Correlation between the two scorecards = 0.40 — ρ_B
- Default rate in the population = 0.10 — Default rate

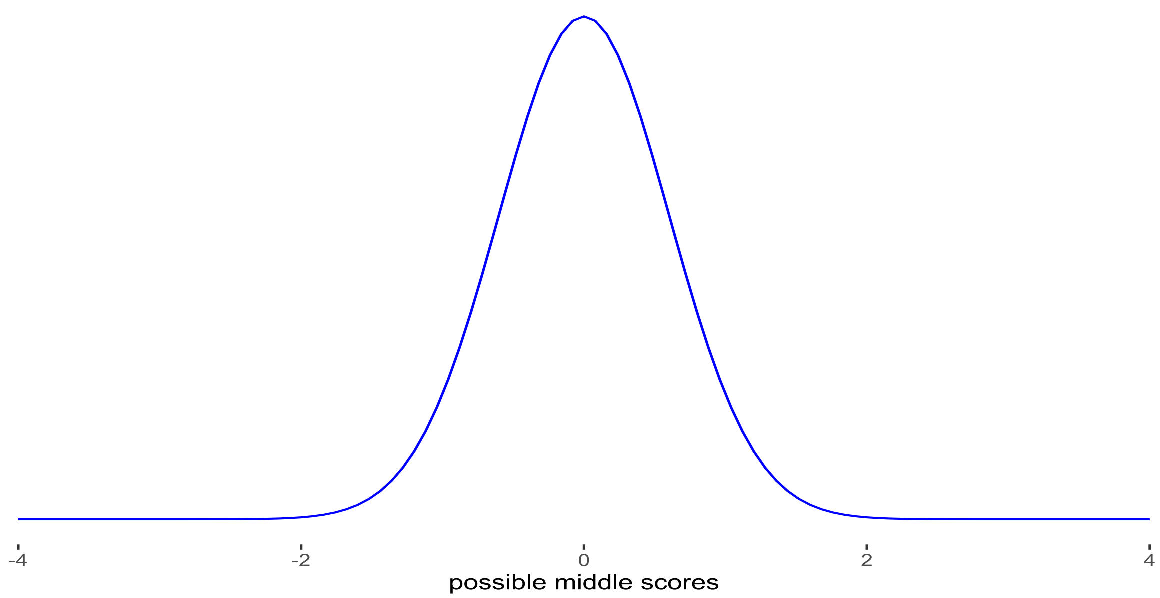
We have no possibility to build one model based on both the application and Credit Bureau data… What can we do?

Maybe linear combination of Score 1 and Score 2 ?

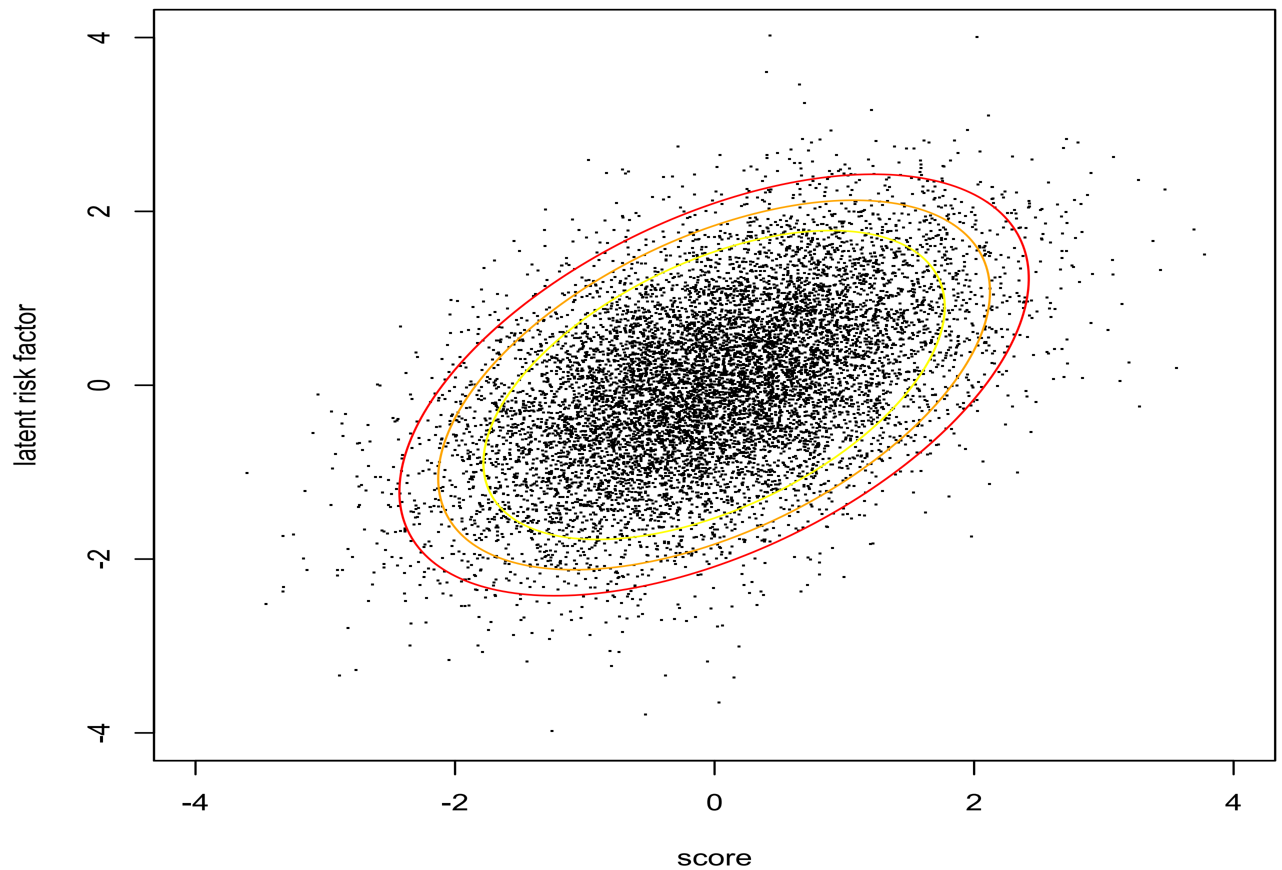

correlation
Score 1
Score 2
Gini 1
Gini 2
Defaults

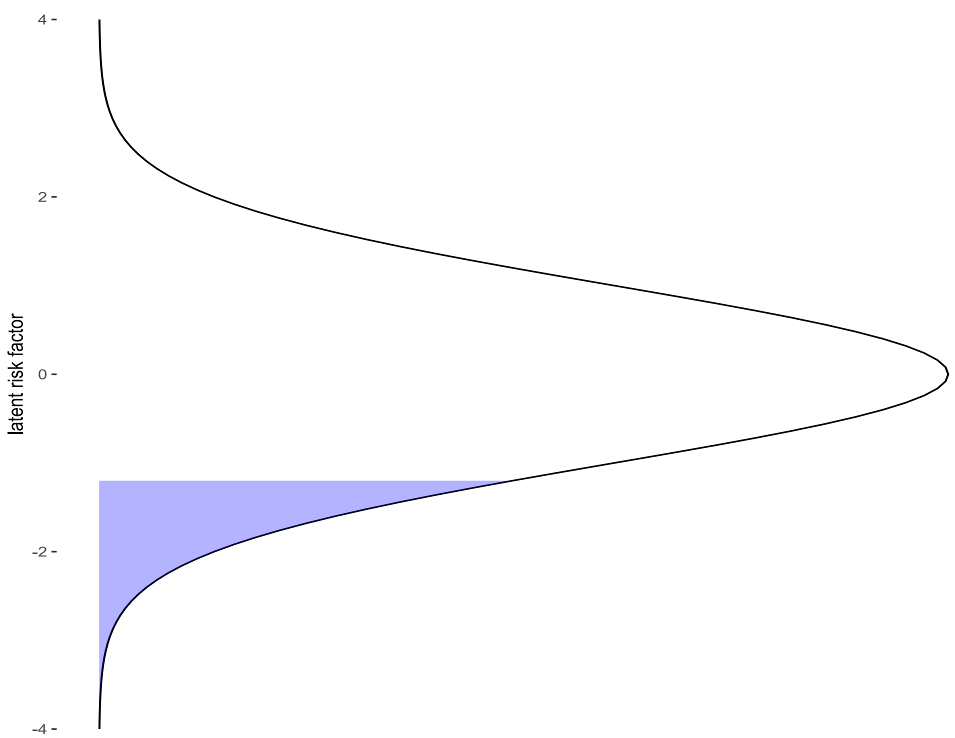

correlation
0.4
Score 1
Application scoring
Score 2
Credit Bureau scoring
Gini 1
0.45
Gini 2
0.6
Defaults

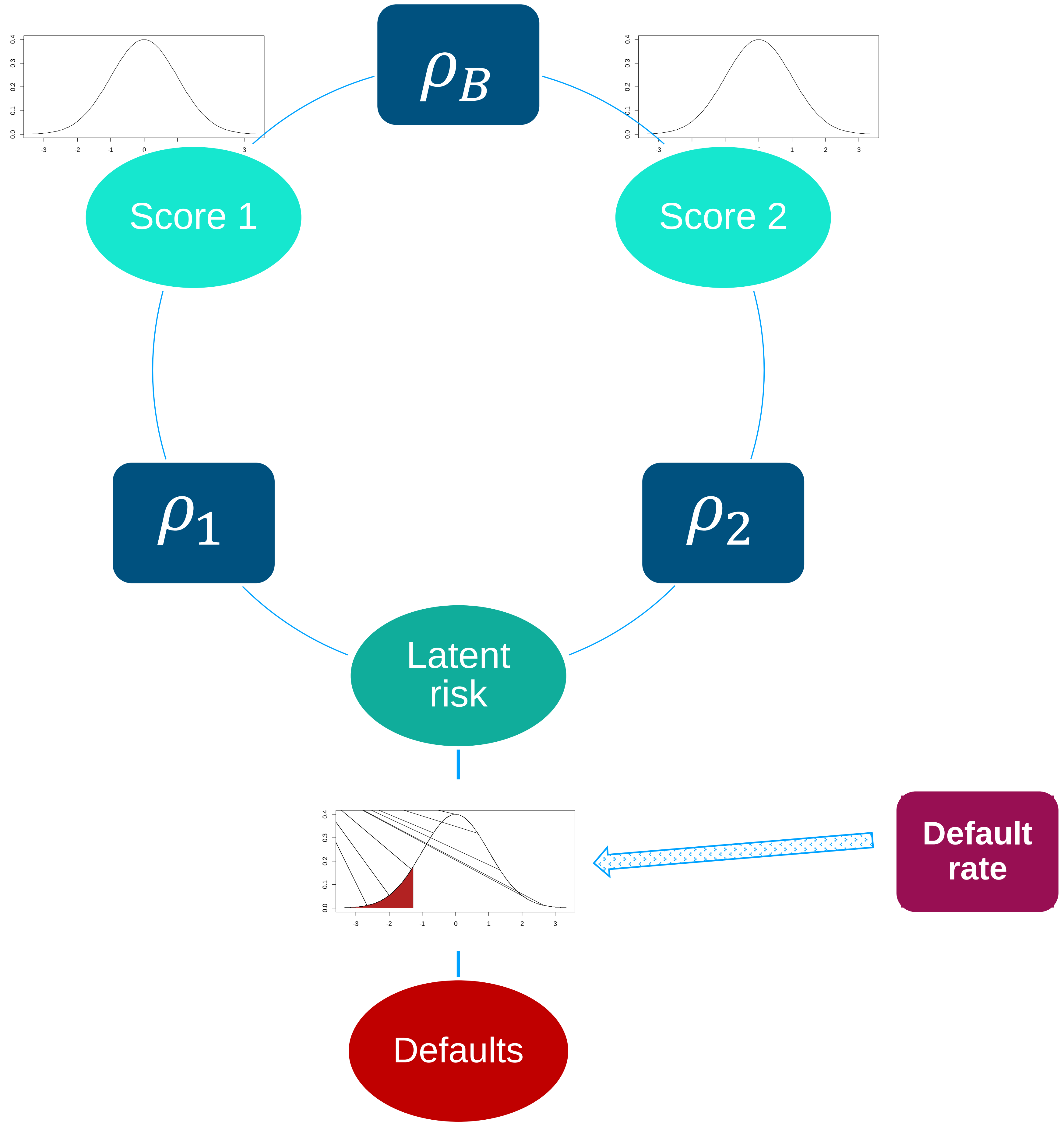

correlation
Score 1
Score 2
Gini 1
Gini 2
Defaults



The score is translated into latent risk variable through the bivariate normal distribution with correlation parameter ρ . Latent risk variable, in turn, translates into default flag based on assumed approval rate.







ρ_B
Score 1
Score 2
ρ_1
ρ_2
Latent risk
Default rate
Defaults

Case 1: Introduction of Credit Bureau scoring in a lending institution.

- Gini of existing application scorecard = 0.45
- Gini of Credit Bureau scorecard = 0.60
- Correlation between the two scorecards = 0.40
- Default rate in the population = 0.10

Gini 1

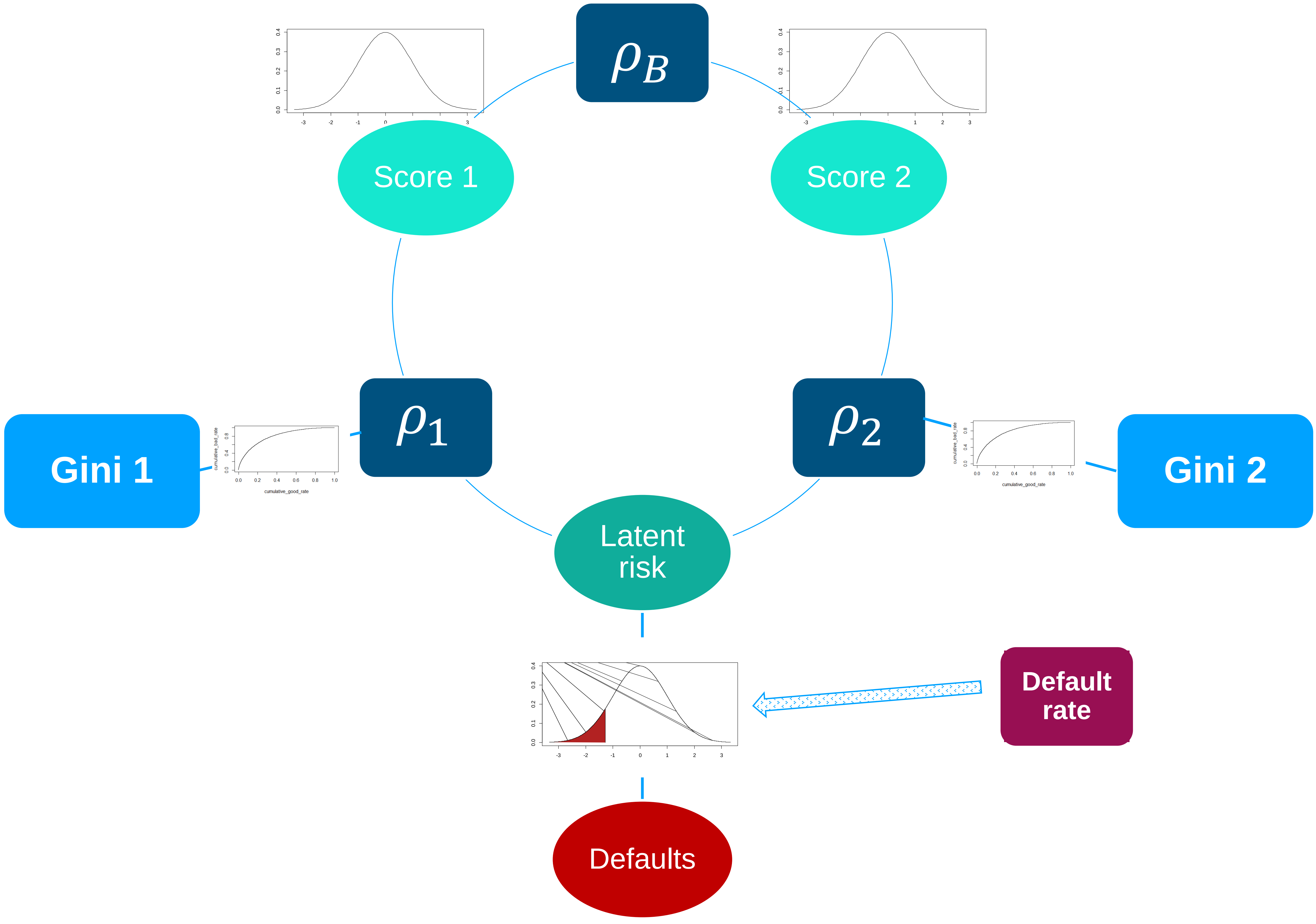
Gini 2

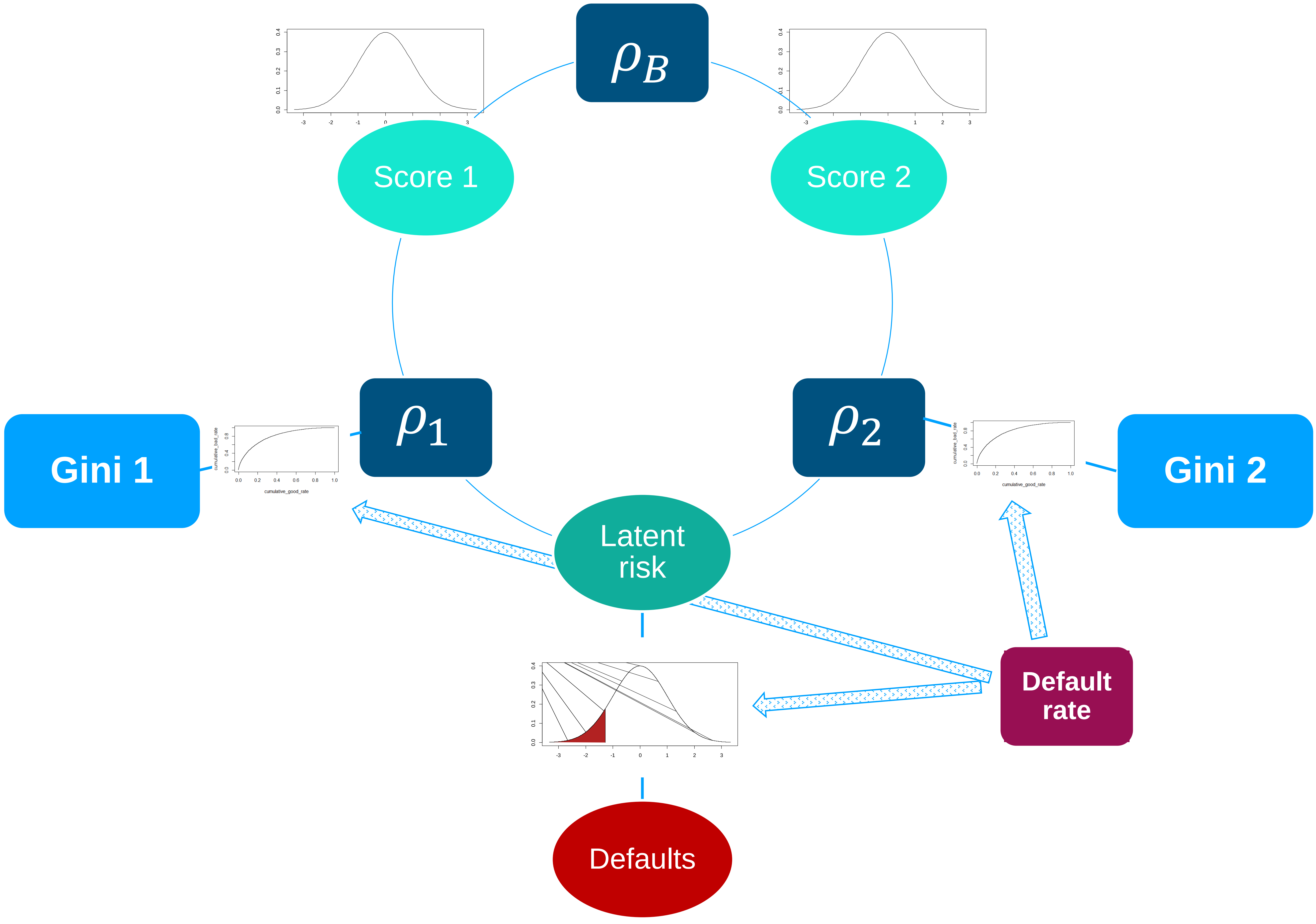
ρ_B

Default rate

We have no possibility to build one model based on both the application and Credit Bureau data… What can we do?

Maybe linear combination of Score 1 and Score 2 ?


ρ_B
Score 1
Score 2
ρ_1
ρ_2
Gini 1
Gini 2
Latent risk
Default rate
Defaults
cumulative_bad_rate
cumulative_good_rate



ρ_B
Score 1
Score 2
ρ_1
ρ_2
Gini 1
Gini 2
Latent risk
Default rate
Defaults

Looking for the optimal combination…

We could look for the optimal combination via simulation, but now the problem has well known mathematics:

Random vector (multivariate normal distribution: $\boldsymbol{X} = \begin{bmatrix} S_1 \\ S_2 \\ L \end{bmatrix}$

Means vector: $\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$

Correlation (=covariance) matrix: $\boldsymbol{\Sigma} = \begin{bmatrix} 1 & \rho_B & \rho_1 \\ \rho_B & 1 & \rho_2 \\ \rho_1 & \rho_2 & 1 \end{bmatrix}$

What is the correlation between $Y = aS_1 + bS_2$ **and** L**?**

The transformation is defined by matrix A: $\boldsymbol{A} = \begin{bmatrix} a & b & 0 \\ 0 & 0 & 1 \end{bmatrix}$

Then distribution of $\boldsymbol{AX} = \begin{bmatrix} aS_1 + bS_2 \\ L \end{bmatrix}$ **is** $N(\boldsymbol{A\mu}, \boldsymbol{AXA'})$

$$\boldsymbol{AXA'} = \begin{bmatrix} a^2 + 2ab\rho_B + b^2 & a\rho_1 + b\rho_2 \\ a\rho_1 + b\rho_2 & 1 \end{bmatrix}$$

So, correlation between Y and L is: $\rho_N = \frac{a\rho_1 + b\rho_2}{\sqrt{a^2 + 2ab\rho_B + b^2}}$

Looking for the optimal combination… continued

Let us assume b=1 and maximize $\rho_N(a) = \frac{a\rho_1+\rho_2}{\sqrt{a^2+2a\rho_B+1}}$ with respect to a.

Maximum is at: $a_0 = (\rho_2\rho_B - \rho_1)/(\rho_1\rho_B - \rho_2)$

So maximum possible correlation is:

$$\rho_N(a_0) = \frac{a_0\rho_1 + \rho_2}{\sqrt{a_0^2 + 2a_0\rho_B + 1}} =$$

$$= \frac{(\rho_2\rho_B - \rho_1)/(\rho_1\rho_B - \rho_2)\rho_1 + \rho_2}{\sqrt{(\rho_2\rho_B - \rho_1)^2/(\rho_1\rho_B - \rho_2)^2 + 2(\rho_2\rho_B - \rho_1)/(\rho_1\rho_B - \rho_2)\rho_B + 1}}$$

Case 1: Introduction of Credit Bureau scoring in a lending institution.

- Gini of existing application scorecard = 0.45
- Gini of Credit Bureau scorecard = 0.60
- Correlation between the two scorecards = 0.40
- Default rate in the population = 0.10

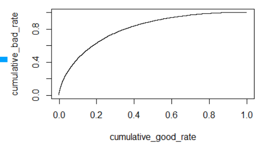
Gini 1

Gini 2

ρ_B

Default rate

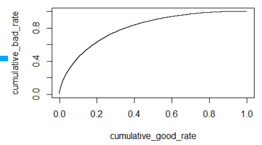
Gini 1



ρ_1

$$\rho_N(a_0) = \frac{a_0\rho_1 + \rho_2}{\sqrt{a_0^2 + 2a_0\rho_B + 1}}$$

ρ_N



Gini N

Case 1: Introduction of Credit Bureau scoring in a lending institution.

- Gini of existing application scorecard = 0.45
- Gini of Credit Bureau scorecard = 0.60
- Correlation between the two scorecards = 0.40
- Default rate in the population = 0.10

Gini 1

Gini 2

ρ_B

Default rate

```
gini_combine_calculator(g1=.45, g2=.6, corr=.4, defaultrate=.1)
##  new_gini     a_opt     gini1      rho1     gini2      rho2  new_corr
## 0.6385234 0.4878499 0.4500000 0.4131295 0.6000000 0.5561160 0.5937599
```

Optimal combination:

0.48 Score 1 **+** Score 2

New Gini:

0.639

Case 1: Introduction of Credit Bureau scoring in a lending institution.

- Gini of existing application scorecard = 0.45
- Gini of Credit Bureau scorecard = 0.60
- Correlation between the two scorecards = 0.40
- Default rate in the population = 0.10

Gini 1

Gini 2

ρ_B

Default rate

```
gini_combine_calculator(g1=.45, g2=.6, corr=.4, defaultrate=.1)
##  new_gini     a_opt     gini1      rho1     gini2      rho2  new_corr
## 0.6385234 0.4878499 0.4500000 0.4131295 0.6000000 0.5561160 0.5937599
```

Optimal combination:

0.48 Score 1 + Score 2

New Gini:

0.639

Case 2: Two banks merge

- Gini in Bank 1 = 0.65
- Gini in Bank 2 = 0.65
- Correlation between the two scorecards = 0.75
- Default rate in the population = 0.08

Gini 1

Gini 2

ρ_B

Default rate

```
gini_combine_calculator(g1=.65, g2=.65, corr=.75, defaultrate=.08)
##  new_gini     a_opt     gini1      rho1     gini2      rho2  new_corr
## 0.6910224 1.0000000 0.6500000 0.5894184 0.6500000 0.5894184 0.6301148
```

Optimal combination:

Score 1 + **Score 2**

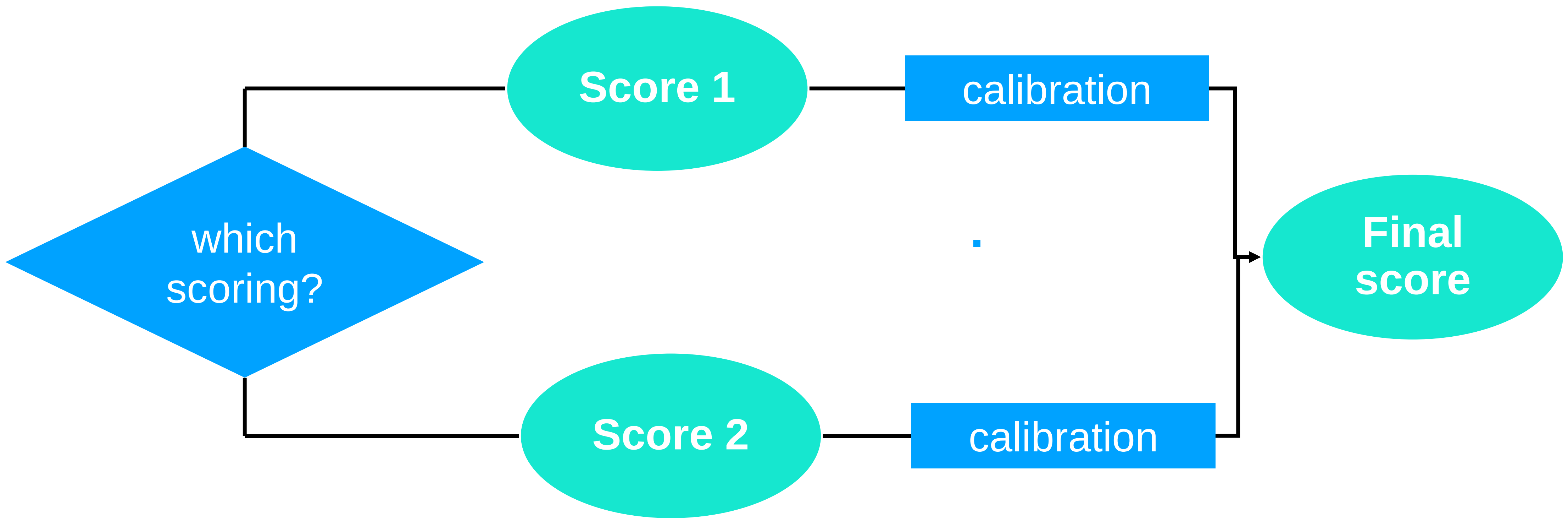
New Gini:

0.691

Presentation plan

- Economics of credit scoring
- Credit market modelling 1
- Gini coefficient in R
- Drawing ROC curves
- Credit market modelling 2
- Combining credit scorecards
- Mixing credit scorecards

Mixture


Score 1
calibration
which
scoring?
Final
score
Score 2
calibration

Case 3: Mixing two models with different Ginis

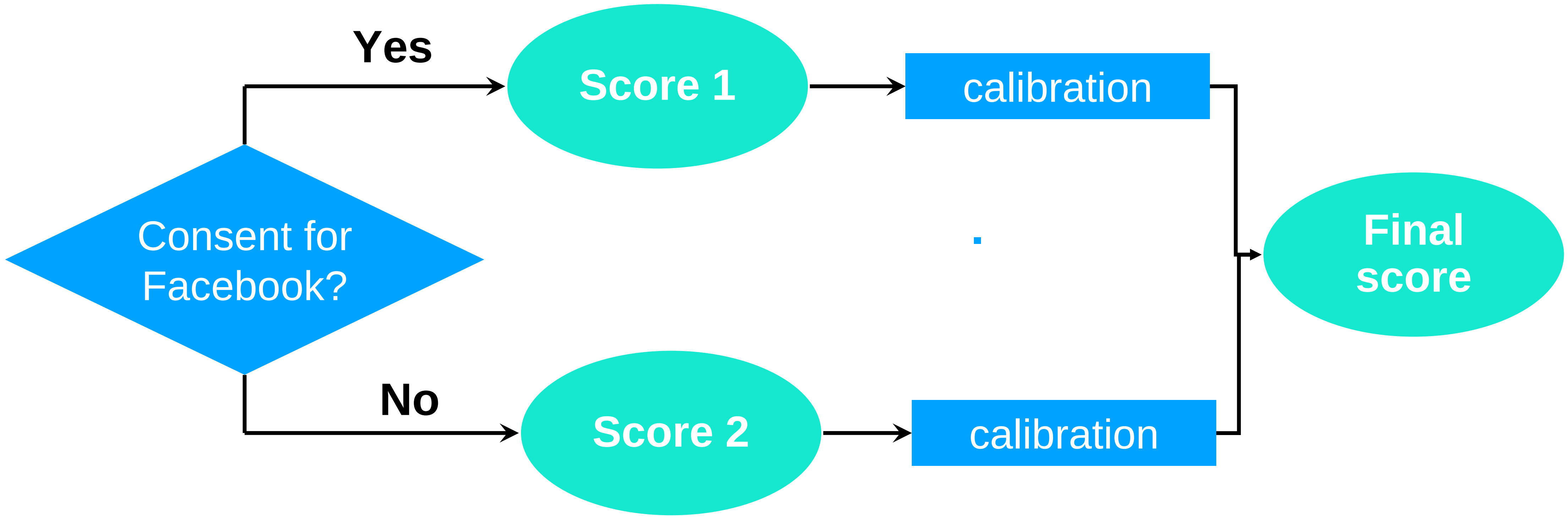
An online lender asks applicants for consent to use their Facebook data. 20% of them agree.

If you have the consent, the Gini is 0.6.

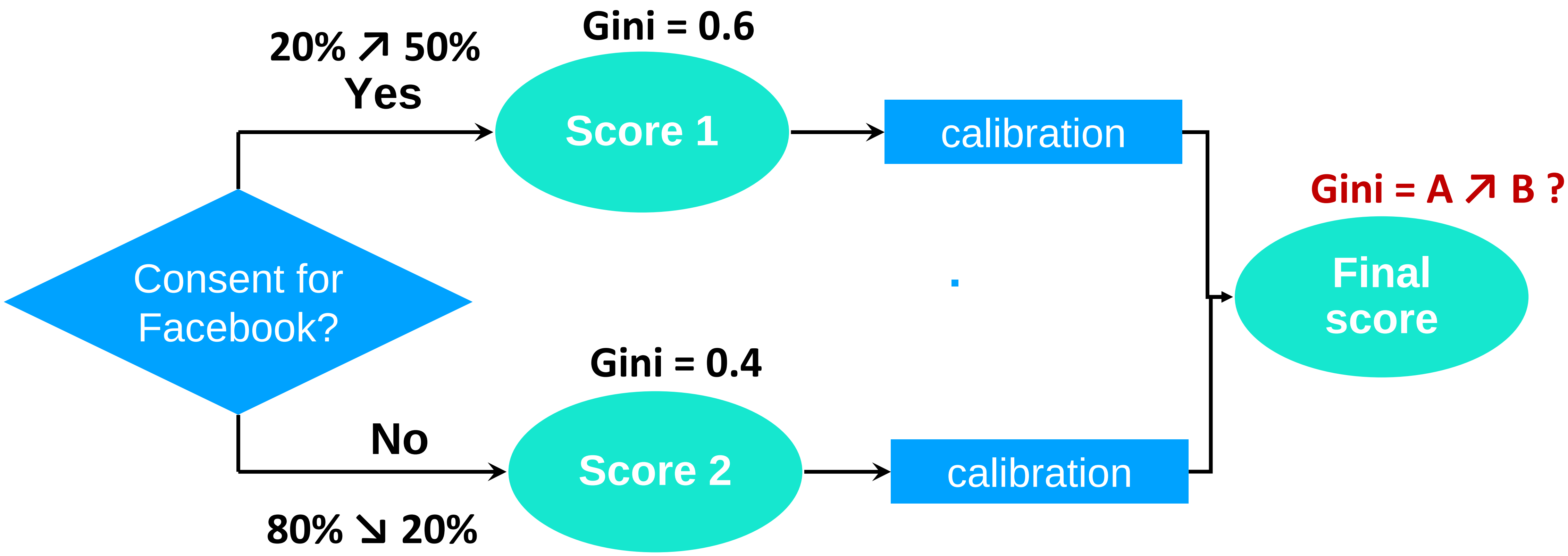
If you do not have the consent, the Gini is 0.4.

A project to increase consent rate from 20% to 50%.

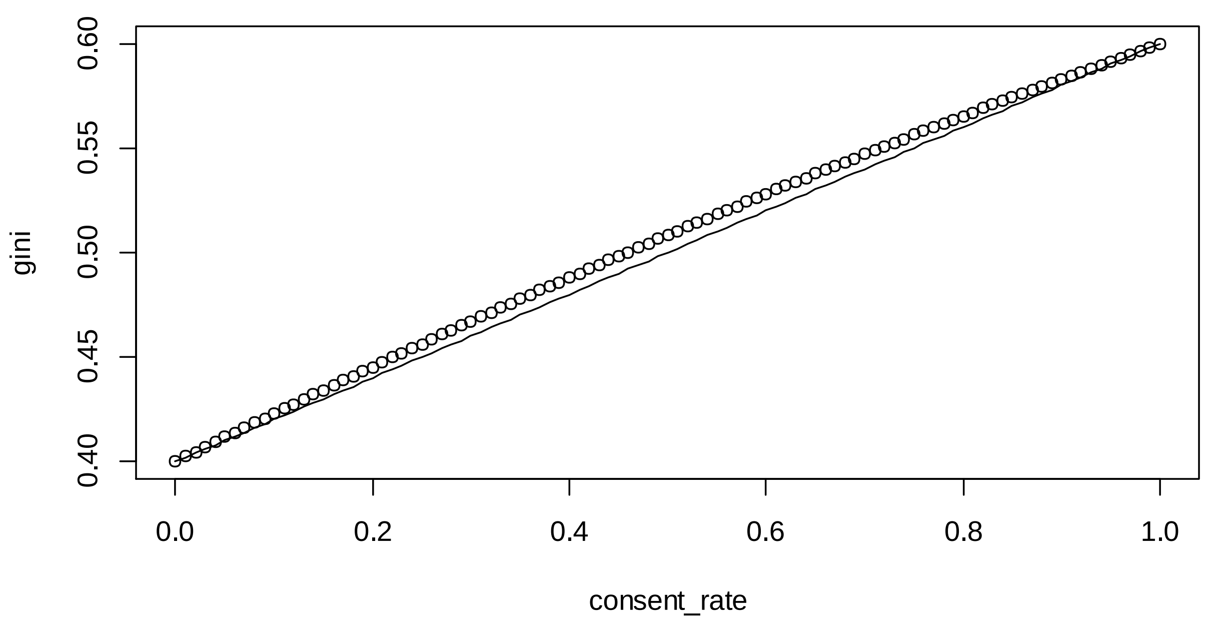
How much will we gain?



Case 3: Mixing two models with different Ginis




```
gini_mixture_calculator(0.6, 0.4, .2, .1)
## [1] 0.4453339
0.2*0.6+0.8*0.4
## [1] 0.44
gini_mixture_calculator(0.6, 0.4, .5, .1)
## [1] 0.5083432
0.5*0.6+0.5*0.4
## [1] 0.5
```



1 percentage point change in Gini*

may have a huge financial impact on the bank

*** = half percentage point in AUC**