

# Credit Scoring - automatyzacja procesu biznesowego

© dr Karol Przanowski



## (c) Prawa autorskie

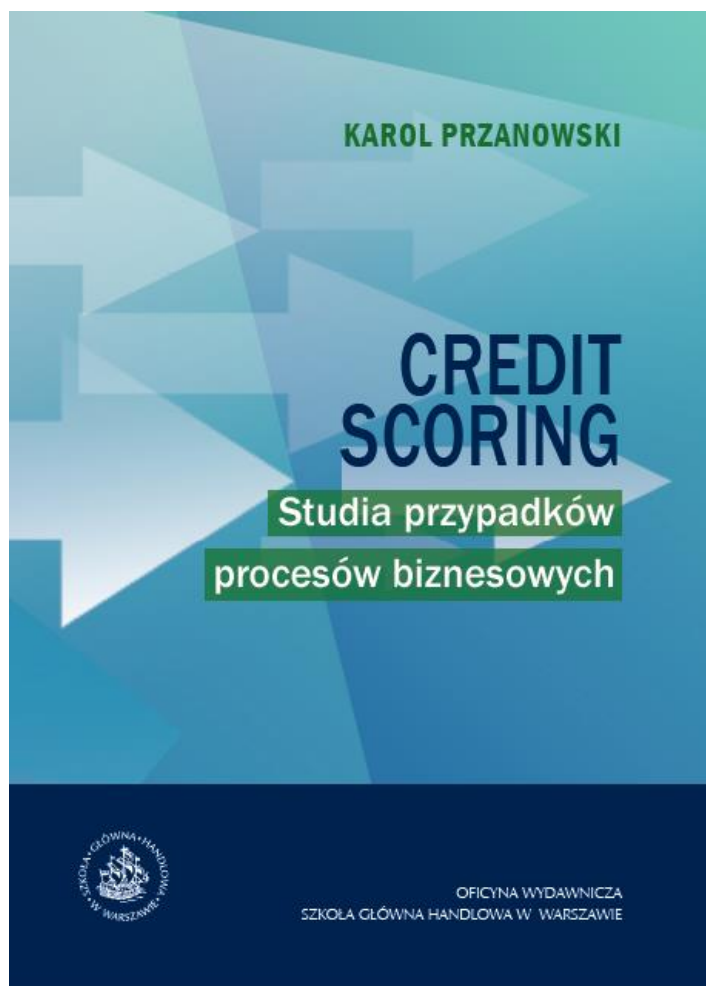
- Wszystkie kody SAS i Python, narzędzia i pomysły są własnością prowadzącego
- Można je używać wyłącznie w celach edukacyjnych i naukowych – zawsze się powołując na autora
- Wykorzystanie do celów biznesowych możliwe tylko za zgodą autora

# Książka w PDF



[http://www.wydawnictwo.sgh.waw.pl/produkty/profilProduktu/id/723//CREDIT\\_SCORING\\_W\\_ERZE\\_BIG-DATA\\_Karol\\_Przanowski/](http://www.wydawnictwo.sgh.waw.pl/produkty/profilProduktu/id/723//CREDIT_SCORING_W_ERZE_BIG-DATA_Karol_Przanowski/)  
[http://administracja.sgh.waw.pl/pl/OW/publikacje/Documents/ostateczny\\_CreditScoring\\_KPrzanowski.pdf](http://administracja.sgh.waw.pl/pl/OW/publikacje/Documents/ostateczny_CreditScoring_KPrzanowski.pdf)

# Książka w PDF



Przedstawione modele biznesowe opłacalności i użyteczności modeli predykcyjnych w:

- Akceptacji kredytów gotówkowych
- Akceptacji procesu złożonego: akwizycji i sprzedaży krzyżowej
- Akceptacji kredytów hipotecznych
- Windykacji polubownej
- Zarządzaniu kampaniami BTL
- Przeciwdziałaniu odchodzeniu klientów

Załączone Excele z regułami i praktycznymi wskaźnikami

<http://administracja.sgh.waw.pl/pl/OW/publikacje/Strony/2015.aspx>

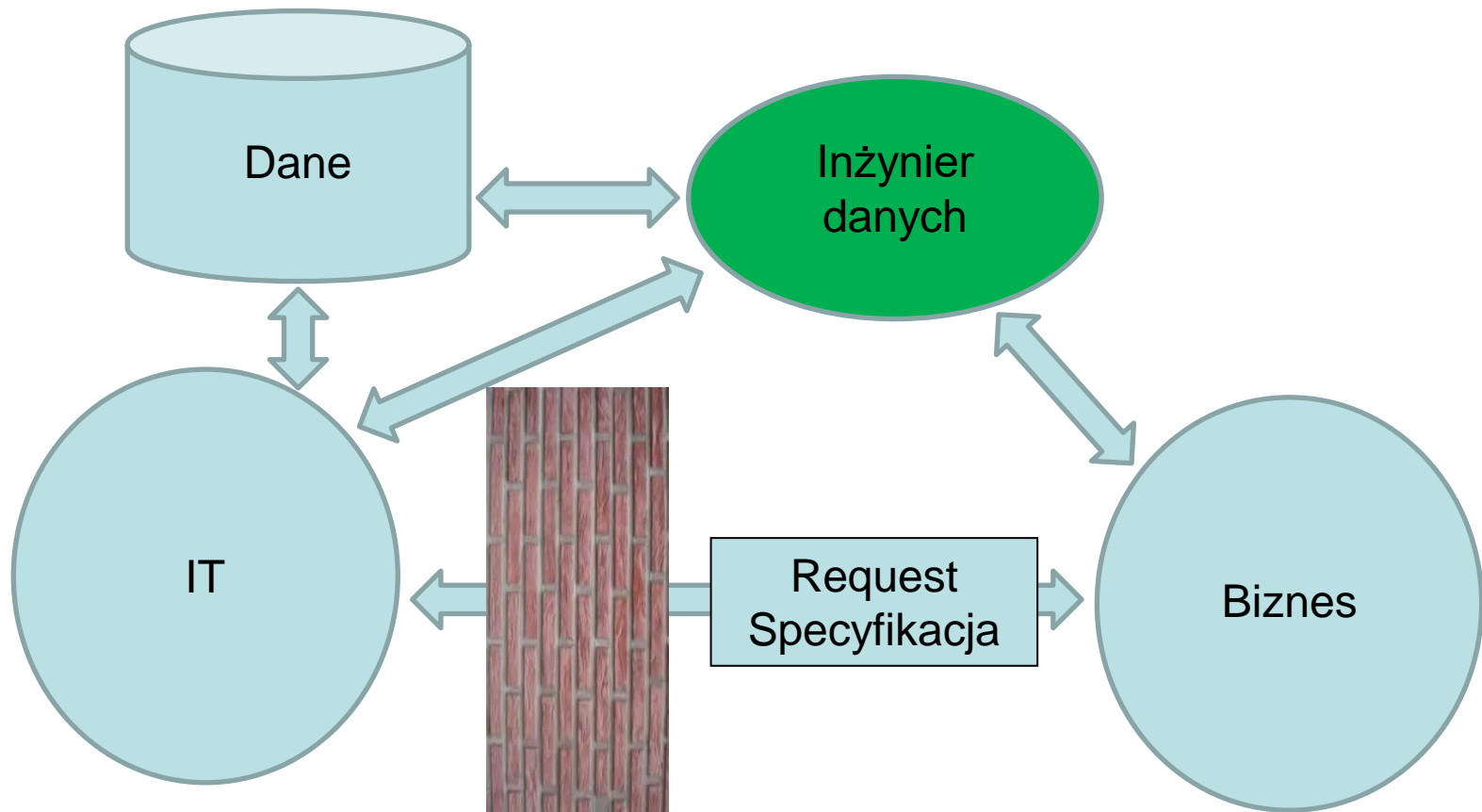


# Inżynier danych - kompetencje

- Programowanie:
  - C++, Java, Python, Perl, R, SAS 4GL, Julia
- Systemy:
  - Oracle, Teradata, SAS, Hadoop, Chmura
- Statystyka i Data Mining:
  - regresja logistyczna, drzewa decyzyjne, sieci neuronowe, lasy losowe, analiza skupień, analiza historii zdarzeń, modele CLV
- Text Mining

# Data Scientist – inżynier danych

- Pośrednik pomiędzy IT i biznesem





# Nowe paradygmaty

- DWH (Hurtownia danych):
  - Najpierw oczyść potem załaduj (stare)
  - Załaduj i potem się martw (nowe)
- Modelowanie (prognozowanie):
  - Znajdź przyczynę i skutek, obserwuj istotne zmienne, czynniki (stare)
  - Sprawdź, które ze zgromadzonych danych wpływają na modelowe zdarzenie, zależności od zmiennych pochodnych (nowe)



# Jakość danych

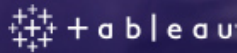
- Co jeszcze zbierać, co poprawić?
- Gdzie i jak dane wykorzystać?
- Miary dobrej jakości:
  - Kompletność
  - Spójność
  - Użyteczność
  - Zrozumiałość





# Niepowodzenia Big Data

- Brak dobrych *business case*
- Dane są zbierane, ale zbyt mało uzasadnia się ich przydatność
- Lekceważone problemy jakości danych
- Umniejszanie problemu wnioskowania na podstawie obciążonej próby
- Zbyt duży nacisk położony na technologię
- Złudne nadzieje szybkiego „klikania modelu”
- Brak inwestycji w przygotowanie i wykształcenie inżyniera danych
- Brak publicznych danych, dostępnych i przykładowych



6 BEST PRACTICES FOR  
EFFECTIVE DASHBOARDS

GET THE WHITEPAPER

[Subscribe to DSC Newsletter](#)

[All Blog Posts](#) [My Blog](#)

+ Add



## How to Become a Data Scientist - On your own

Posted by Zeeshan Usmani on March 28, 2015 at 4:00pm [View Blog](#)

Big Data, Data Sciences, and Predictive Analytics are the talk of the town and it doesn't matter which town you are referring to, it's everywhere, from the [White House hiring DJ Patil](#) as the first chief data scientist to the [United Nations using predictive analytics](#) to forecast bombings on schools. There are dozens of Startups springing out every month stretching human imagination of how the underlying technologies can be used to improve our lives and everything we do. Data science is in demand and its growth is on steroids. According to LinkedIn, "Statistical Analysis" and "Data Mining" are two top-most skills to get hired this year. Gartner says there are 4.4 million jobs for data scientists (and related titles) worldwide in 2015, 1.9 million in the US alone. One data science job creates another three non-IT jobs, so we are talking about some 13 million jobs altogether. The question is what YOU can do to secure a job and make your

Welcome to  
Data Science Central

[Sign Up](#)  
or [Sign In](#)

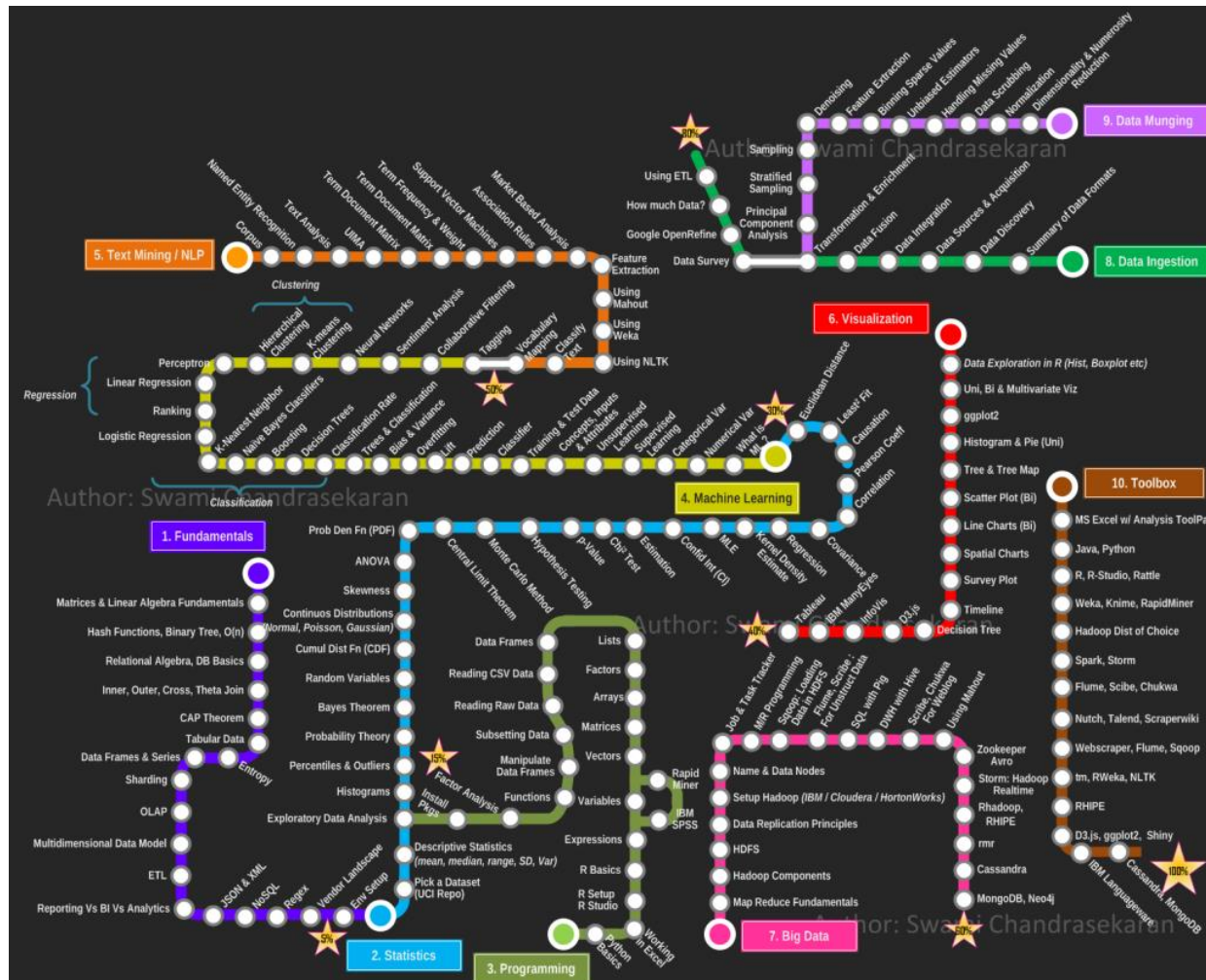
Or sign in with:



Fuel your  
analytics  
evolution

<http://www.datasciencecentral.com/profiles/blogs/how-to-become-a-data-scientist-for-free>

# Data Scientist Metro Map



<http://nirvacana.com/thoughts/becoming-a-data-scientist/>



# Statystyka

- Badanie zjawisk powtarzalnych
- Zjawiska masowe
- Wykrywanie trendów, właściwości
- Badanie zbiorowości, populacji
- Wykrywanie zależności
- Prognozowanie
- Stabilność
- Nie jeden obiekt tylko setki tysięcy

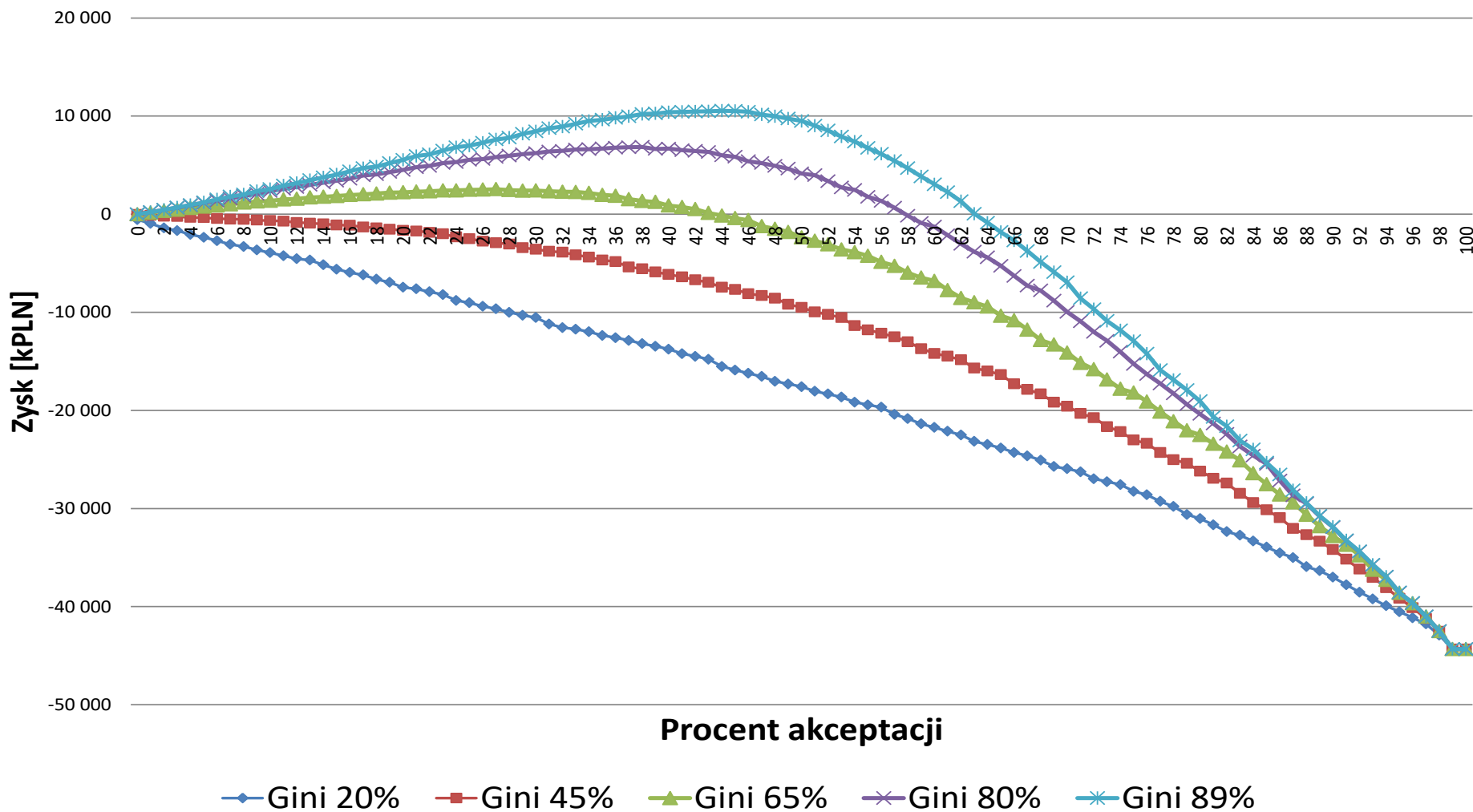


# Prognozowanie zdarzenia

- Zakup towaru
- Spłata kredytu
- Rezygnacja z umowy
- Wyłudzenie, oszustwo
- Nielegalny pobór energii
- Pranie pieniędzy
- Przyjęcie łapówki
- Fałszywe zeznania

# Dlaczego się opłaca

Krzywa Profit zależna od mocy predykcyjnej





# Jak liczymy wskaźniki

$$L_i = \begin{cases} 50\%A_i, & \text{gdy } \text{default}_{12} = \mathbf{Z\acute{ł}y}, \\ 0, & \text{gdy } \text{default}_{12} \neq \mathbf{Z\acute{ł}y}. \end{cases}$$

$$I_i = \begin{cases} A_i p, & \text{gdy } \text{default}_{12} = \mathbf{Z\acute{ł}y}, \\ A_i \left( N_i r \frac{(1+r)^{N_i}}{(1+r)^{N_i-1}} + (p-1) \right), & \text{gdy } \text{default}_{12} \neq \mathbf{Z\acute{ł}y}. \end{cases}$$

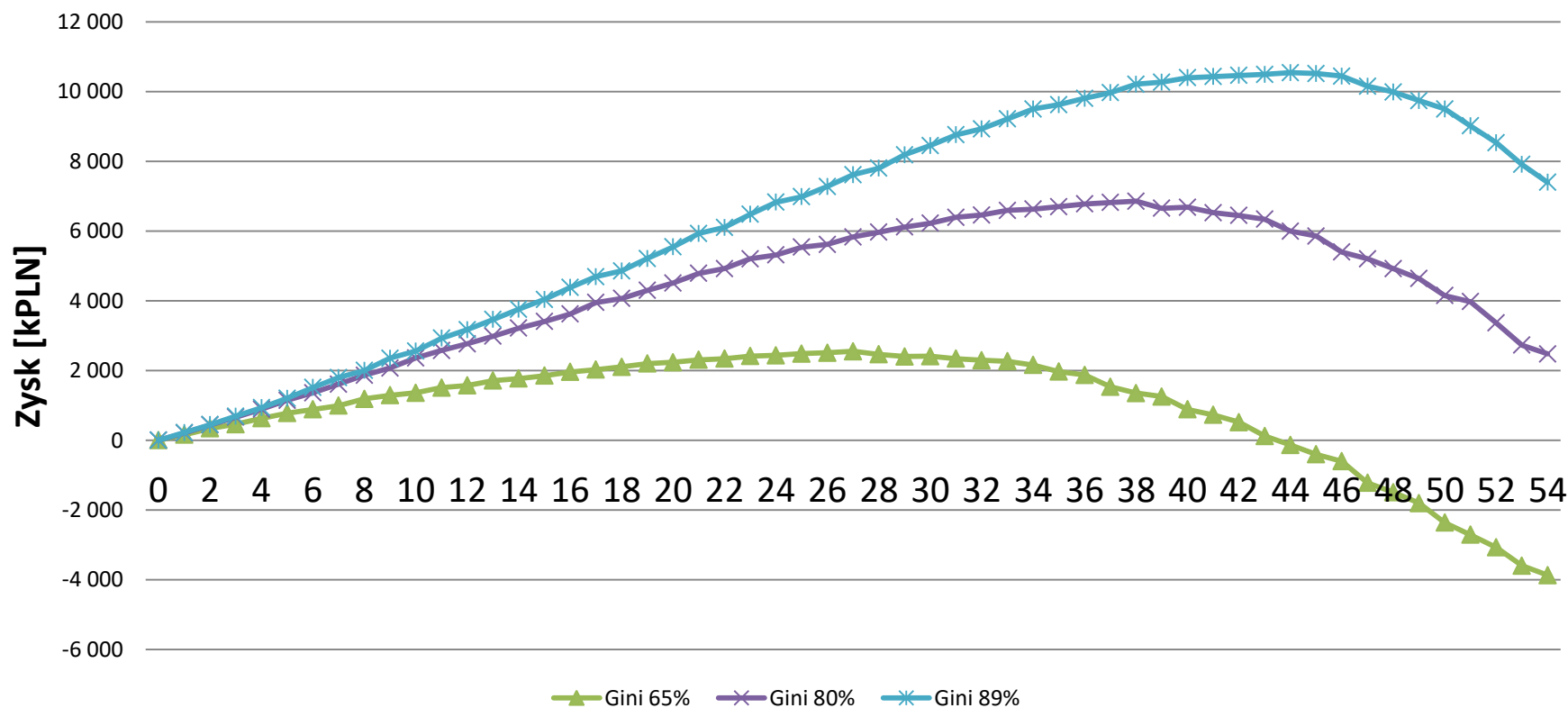
Czyli sumaryczny zysk (profit)  $P$  całego portfela obliczamy następującym wzorem:

$$P = \sum_i I_i - L_i. \quad (4.1)$$

- $EL = PD * LGD * EAD$

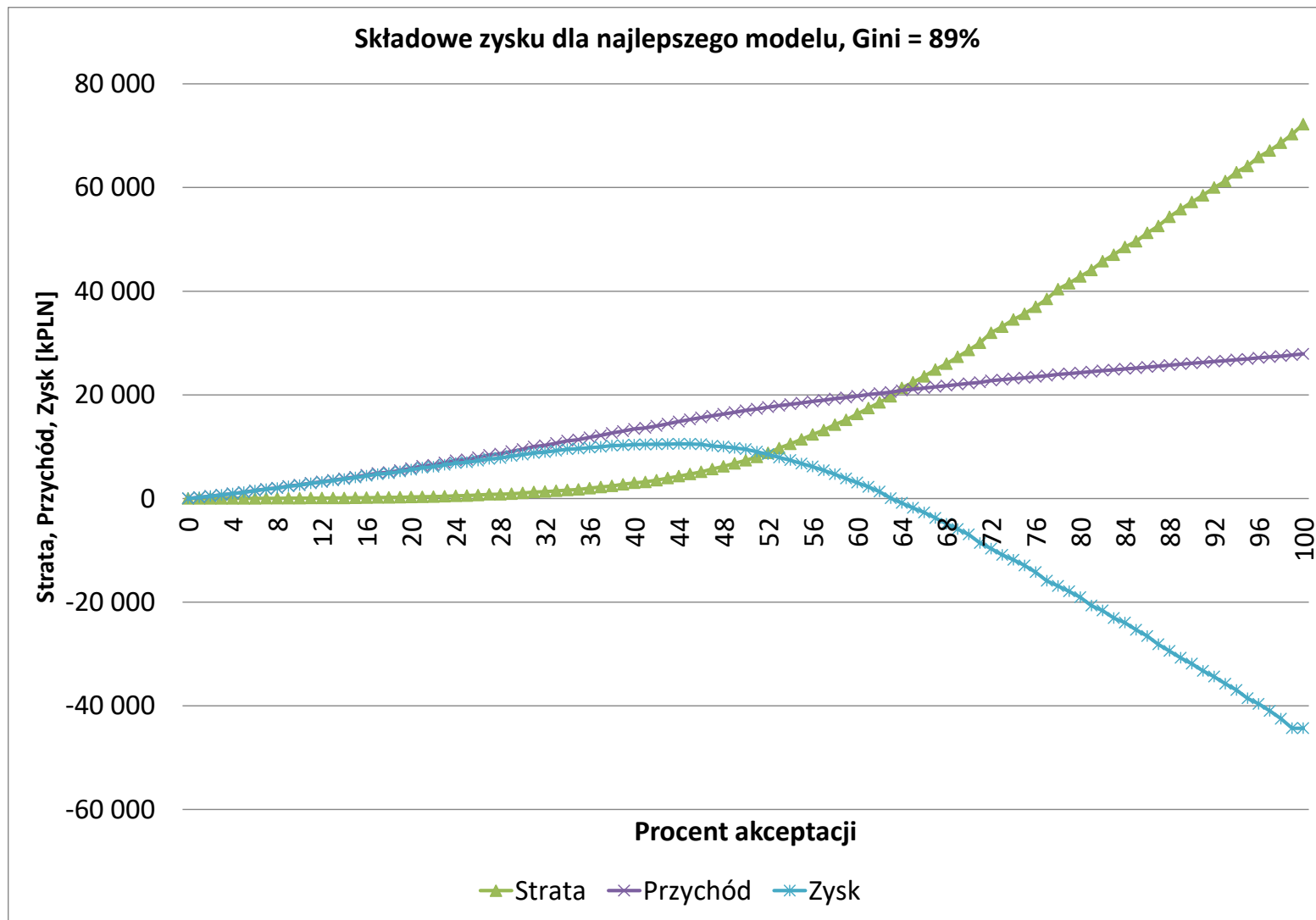
# Dlaczego się opłaca

Krzywa Profit zależna od mocy predykcyjnej

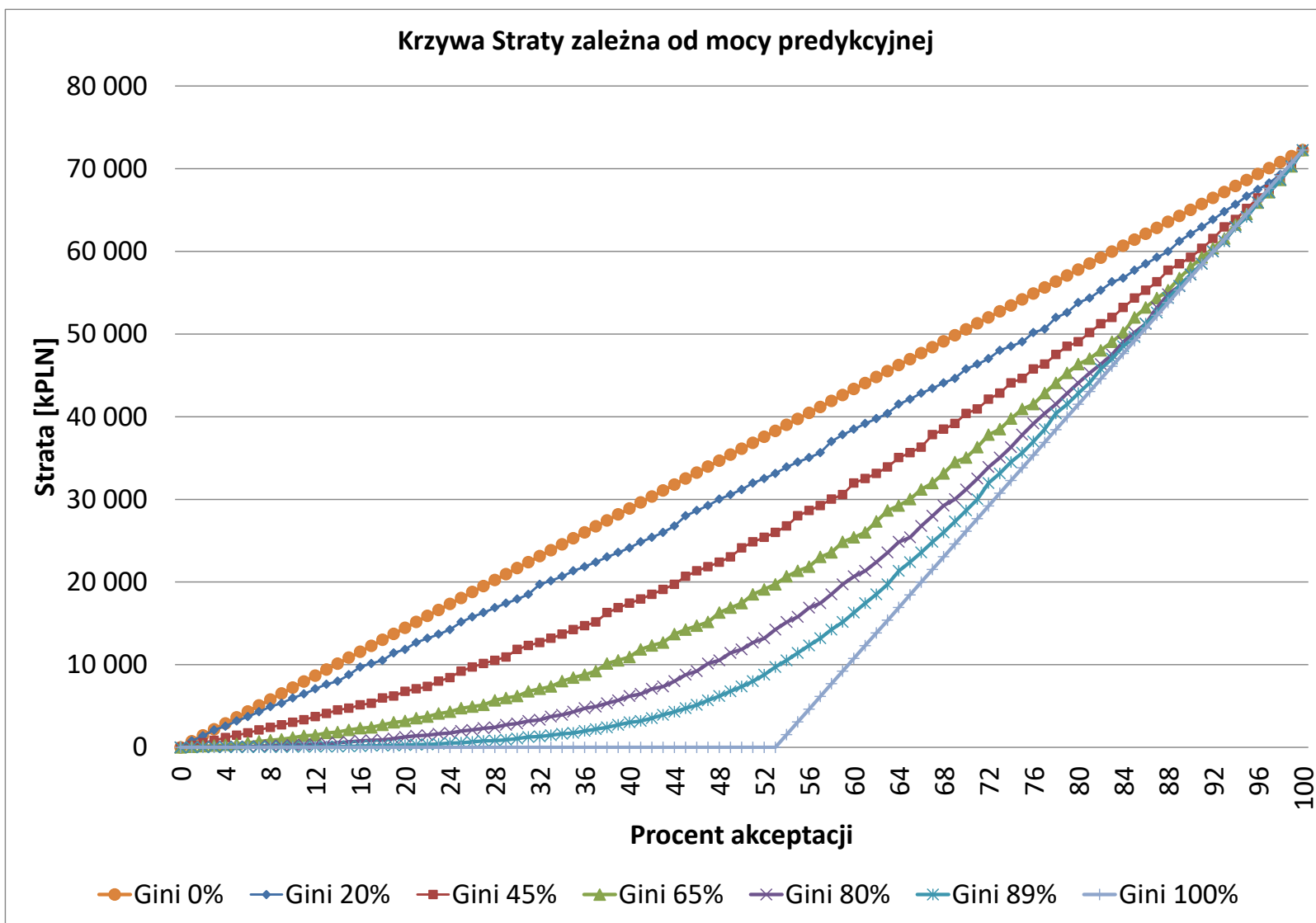




# Składowe Profit



# Krzywe Straty



# Polepszenie modelu

Wskaźnik	Wartość
Liczba wniosków w miesiącu	50 000
Średnia kwota kredytu	5 000 PLN
Średni czas kredytowania	36 miesięcy
Roczne oprocentowanie kredytów	12%
Prowizja za udzielenie kredytu	6%
Globalne ryzyko portfela	47%
Zmiana mocy predykcyjnej	5%
Zmiana procentu akceptacji	3,5%
Zmiana zysku	1 492 kPLN
Zmiana straty oczekiwanej (AR=20%)	872 kPLN
Zmiana straty oczekiwanej (AR=40%)	1 529 kPLN

# Przykład w Excelu

Number of applications per month	50 000
Average loan amount	5 000
Average number of installments	36
Annual percentage rate (or net margin)	12%
LGD (Loss Given Default)	50%
Provision charged on disbursement day	6%

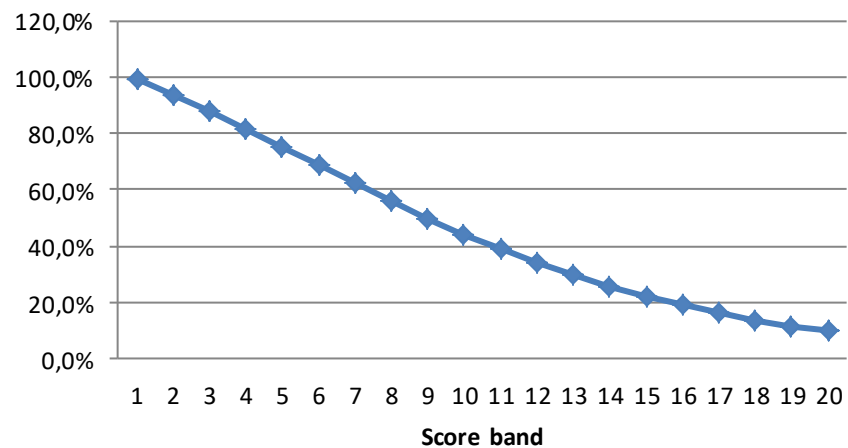
Gini global	65,54%
Gini on accepted	24,50%

Global risk in market (default12)	47%
Accepted risk	18,49%
Acceptance rate	40,00%

Global loss	58 750 000
Global income	33 882 258
Global profit	-24 867 742

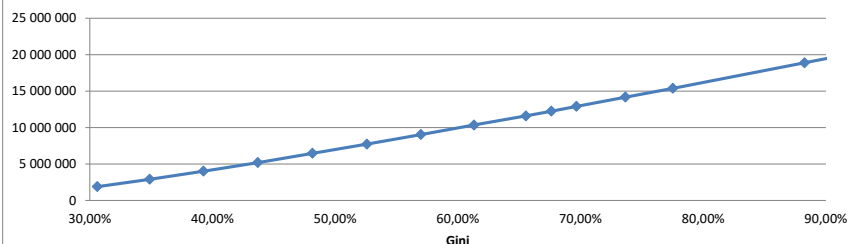
Accepted loss	9 246 726
Accepted income	20 842 459
Accepted profit	11 595 733

**Bad rate**



Charge provision	300,00
Income from interest rates	978,58

**Optimal profit**



1%	of Gini	296 451	PLN
5%	of Gini	1 482 254	PLN
10%	of Gini	2 964 507	PLN



# Jak zarządza się ryzykiem kredytowym?

- Cały proces akceptacji wpływa na ryzyko kredytowe! (od pierwszego słowa z klientem do ostatniego z nim kontaktu)
- Co to jest zjawisko selekcji negatywnej?
- Jaki ma wpływ Risk Based Pricing na wynik finansowy?
- Czy ryzykiem zarządza się przez zmniejszenie licznika?
- Jak sprzedaż wpływa na ryzyko kredytowe?
- Czy Sprzedaż i Ryzyko mogą się rozumieć?
- Czy można zmniejszyć ryzyko kredytowe i zwiększyć sprzedaż?
- Dyrektor Ryzyka Kredytowego musi być przyjacielem Dyrektora Sprzedaży i odwrotnie !!!



# Definicja Default

Każdy rachunek jest badany od miesiąca jego aplikowania i uruchomienia  $T_{app}$ , zwanym punktem obserwacji, w ciągu kolejnych 3, 6, 9 i 12 miesięcy. W tym okresie, zwanym okresem obserwacji, wylicza się statystykę, maksymalne opóźnienie:

$$MAX = MAX_{m=0}^{o-1}(x_{n_{due}}^{act}(T_{app} + m)),$$

gdzie  $o = 3, 6, 9, 12$ . Na podstawie tej statystyki definiujemy trzy typowe kategorie:

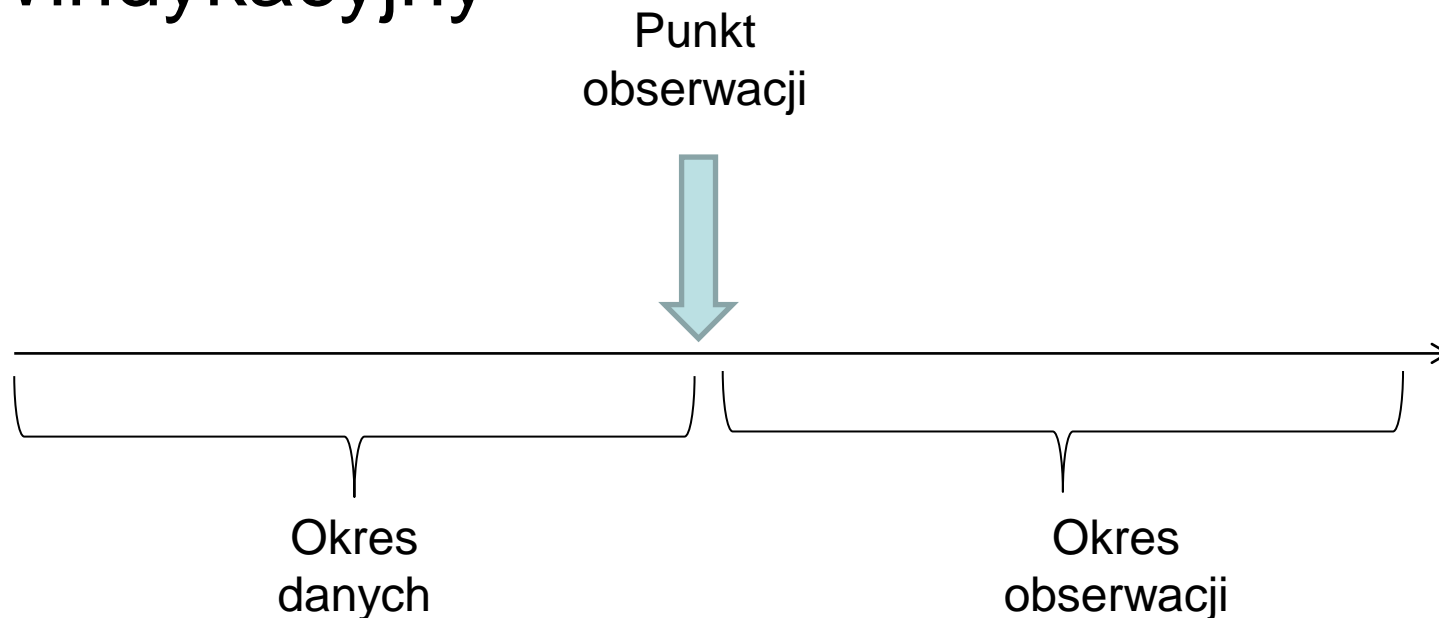
**Dobry (Good):** Gdy  $MAX \leq 1$  lub w czasie okresu obserwacji rachunek spłacił całe swoje zobowiązanie  $x_{status}^t = C$ .

**Zły (Bad):** Gdy  $MAX > 3$  lub w czasie okresu obserwacji rachunek wpadł w maksymalne zadłużenie  $x_{status}^t = B$ . Wyjątkowo dla  $o = 3$ , gdy  $MAX > 2$ .

**Nieokreślony (Indeterminate):** dla pozostałych przypadków.

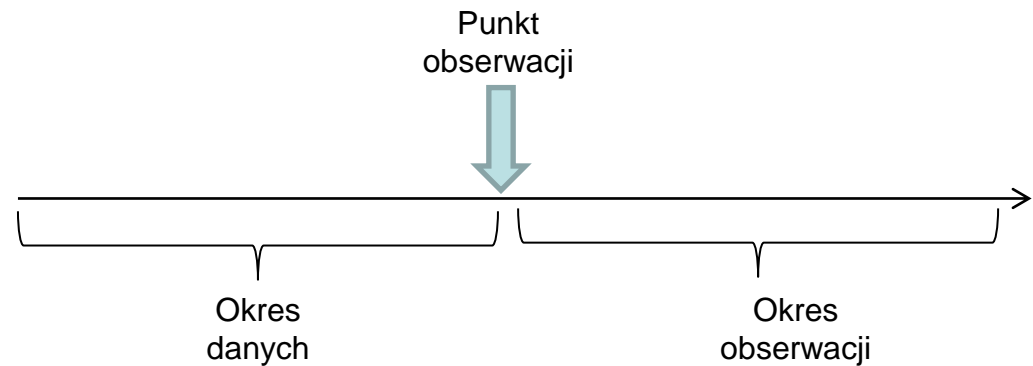


- Aplikacyjny
- Behawioralny
- Windykacyjny

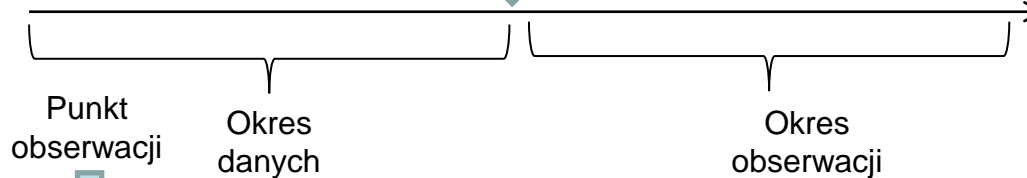


# Aplikacyjne (rachunek tylko raz)

Punkt  
Obserwacji to  
moment aplikacji



Punkt  
obserwacji



Punkt  
obserwacji

Okres  
danych

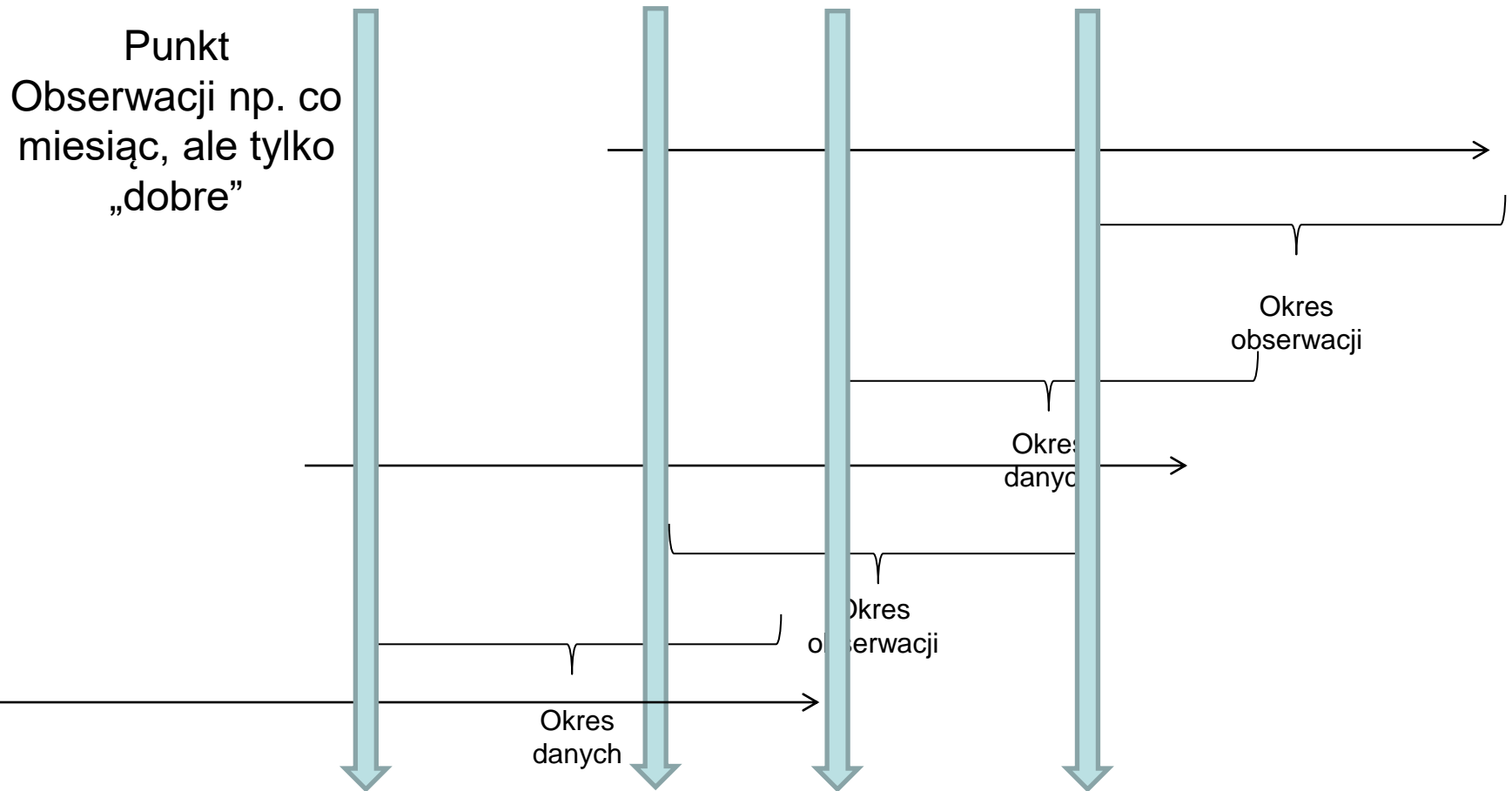
Okres  
obserwacji

Okres  
danych

Okres  
obserwacji



# Behawioralne (wiele razy)





# ABT – Analytical Base Table

- Jeden wiersz to obiekt modelowany, rachunek, klient.
- Funkcja celu: 1 – zły, 0 – dobry, .i – nieokreślony, .d – dormant (nieistotny)
- Nazewnictwo: ags3\_Min\_CMaxI\_Due
- Excel z listą zmiennych
- Kod projektu abt\_behavioral\_columns.sas
- Kod abt\_behavioral\_columns.sas

# Karta ocen punktowych – karta skoringowa

Attribute	Variable	Partial Score
<20	Age	10
20>= and <34		20
35>=		30
Bad	Payment history	10
Not good		25
Good		40



# Pierwsze kroki z ASB

- Główny kod:
  - SAS: main.sas,
  - Python: ASB\_step\_by\_step.ipynb
- Opcje, makrozmiennne
- Tryb batchowy
- Układ katalogów i bibliotek
- Dodatkowe zmienne
- Zmienne interakcji



# Ograniczenie danych

- SAS:
  - where '197501'<=period<='198712' and product='css' and decision='A';
- Python:
  - `df=df[('197501'<=df['period']) & (df['period']<='198712') & (df['product']=='css') & (df['decision']=='A')]`

# Partycjonowanie danych

- Time sampling  $\leftrightarrow$  Random sampling
- Through the cycle  $\leftrightarrow$  Point in time

period	
200801	
200802	
200803	
200804	
200805	
200806	
200807	
200808	
200809	
200810	
200811	
200812	
200901	
200902	
200903	
200904	
200905	
200906	

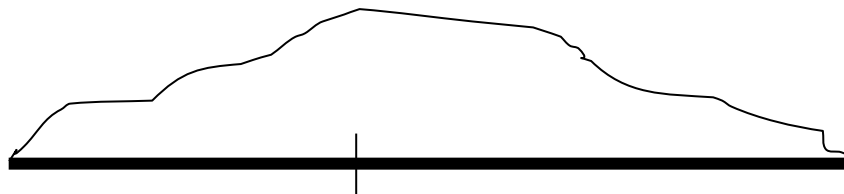
period	train	valid
200801		
200802		
200803		
200804		
200805		
200806		
200807		
200808		
200809		
200810		
200811		
200812		
200901		
200902		
200903		
200904		
200905		
200906		



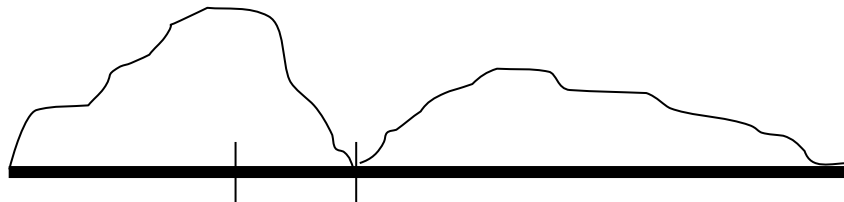
# Partycjonowanie danych

- SAS:
  - `%include "&dir_codes.train_valid.sas" / source2;`
  - odkomentować na linię: `/* agr: ags:*/`
- Python:
  - `#Splitting for train and test datasets`
  - odkomentować linię: `# vars=[var for var in list(df) if var[0:3].lower() in ['app','act','agr','ags']]`

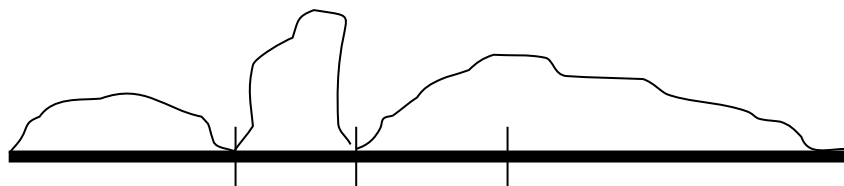
# Binowanie zmiennych ciągłych



1. Pierwszy punkt podziałowy



2. Drugi punkt



3. Trzeci w najszerszym przedziale



# Binowanie zmiennych ciągłych

$$h_a = - \left[ \frac{b_a}{s_a} \log_2 \left( \frac{b_a}{s_a} \right) + \frac{g_a}{s_a} \log_2 \left( \frac{g_a}{s_a} \right) \right]$$

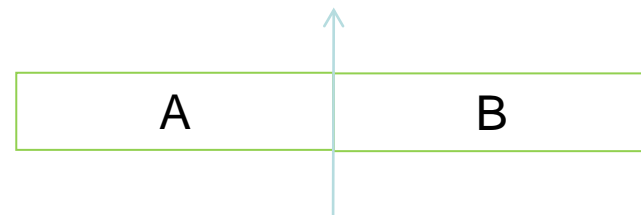
$$h_b = - \left[ \frac{b_b}{s_b} \log_2 \left( \frac{b_b}{s_b} \right) + \frac{g_b}{s_b} \log_2 \left( \frac{g_b}{s_b} \right) \right]$$

$$h = - \left[ \frac{b}{s} \log_2 \left( \frac{b}{s} \right) + \frac{g}{s} \log_2 \left( \frac{g}{s} \right) \right] - \frac{s_a}{s} h_a - \frac{s_b}{s} h_b$$

$$g_a = 1 - \frac{b_a^2 + g_a^2}{s_a^2}$$

$$g_b = 1 - \frac{b_b^2 + g_b^2}{s_b^2}$$

$$g = 1 - \frac{b^2 + g^2}{s^2} - g_a \frac{s_a}{s} - g_b \frac{s_b}{s}$$



$b_a$  – liczba złych w liściu A  
 $g_a$  – liczba dobrych w liściu A  
 $s_a$  – liczba wszystkich w liściu A  
 $b_b, g_b, s_b$  – podobnie dla liścia B  
 $b, g, s$  – podobnie w całej próbie



# Binowanie zmiennych ciągłych

- SAS:
  - %let max\_n\_splitting\_points=5;
  - /\*Minimal share of category\*/
  - %let min\_percent=3;
  - %include "&dir\_codes.tree.sas" / source2;
- Python:
  - #Bining for numerical variables

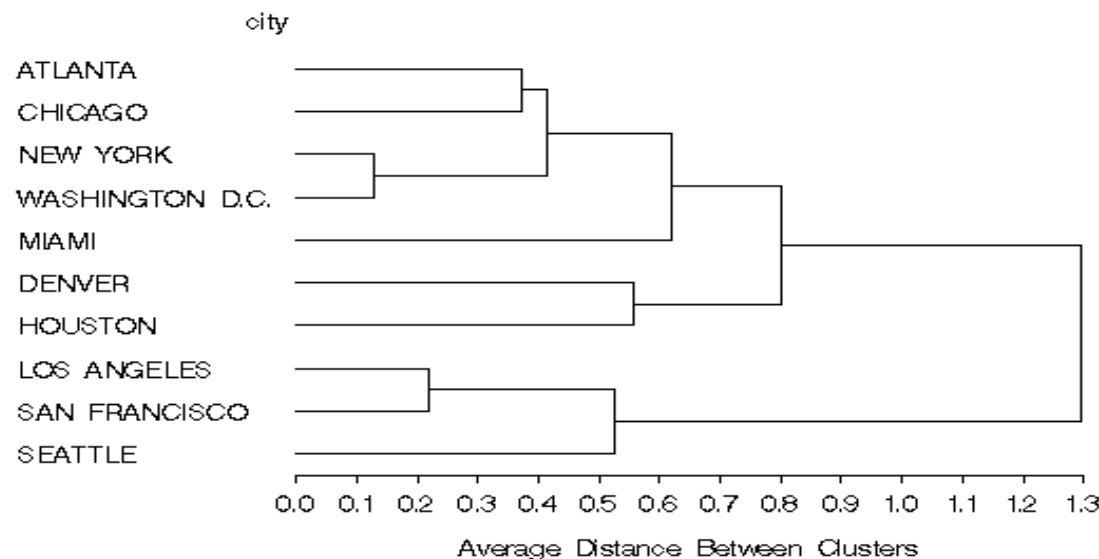
# Binowanie zmiennych ciągłych

- Nie monotoniczne – gotowe w kodach
- Inne możliwe:
  - monotoniczne
  - Maksymalizacja Giniego
  - Stałej szerokości przedziały, albo udziały

Number	Variable name	Gini_before	Gini_NonMon	Gini_MonNew	Gini_MonOld
1	AGGR6_MEAN_S_CASHUTL_EM	64,72%	63,31%	7,57%	63,04%
2	AGSP6_MAX_BAL_EMCL	60,09%	60,02%		59,65%
3	ACT_S_CASHUTL_EM	58,38%	60,03%	20,71%	59,81%
4	AGGR3_MEAN_S_RBAL_EMCL	38,44%	36,11%	39,35%	36,11%
5	AGSP3_MIN_PMT	29,33%	40,36%	31,39%	36,87%
6	AGSP6_MAX_NOTPAID	28,32%	17,85%	17,85%	
7	ACT_PMT	27,59%	43,33%	32,33%	41,03%
8	ACT_S_RBAL_EM	26,41%		26,00%	
9	AGGR6_MAX_CYCLE_DD	14,36%	7,58%	7,58%	7,58%
10	AGSP3_MAX_PMT	8,88%		19,71%	24,65%
11	ACT_NBR_LCF	1,75%	27,51%	27,54%	

# Binowanie zmiennych nominalnych

- Warunek na reprezentatywność
  - Udział kategorii  $\geq 1$  lub 3%
- Łączymy kategorie procedurą Cluster procedure na podstawie podobnych statystyk bad rate





# Binowanie zmiennych nominalnych

- SAS:
  - %let max\_n\_splitting\_points=5;
  - /\*Minimal share of category\*/
  - %let min\_percent=3;
  - %include "&dir\_codes.bining\_nominal.sas" / source2;
  - %include  
"&dir\_codes.bining\_nominal\_without\_joining.sas" /  
source2;
- Python:
  - #Bining for character variables



# Preselekcja zmiennych

- Dla każdej zmiennej liczymy statystyki:
  - Jakości
  - Opisowe
  - Predykcyjności
  - Stabilności
- SAS:
  - Output: variable\_stat – wiele statystyk zmiennych
  - %include "&dir\_codes.variable\_pre\_selection\_1step.sas" / source2;
  - %include "&dir\_codes.variable\_pre\_selection\_full.sas" / source2;
- Python:
  - Output: Gini\_vars.xlsx, Variable\_report.xlsx
  - #Calculating Gini values for features



# Preselekcja zmiennych

- Statystyki stabilności:
  - IS (PSI) – index stability,
  - KS - Kolmogorov-Smirnov ,
  - KL - Kullback-Leibler distance,
  - AR\_Diff (Delta Gini) =  $\text{abs}(\text{Gini Train} - \text{Gini Valid}) / \text{Gini Train}$
- Statystyki predykcyjności:
  - Gini train, valid
  - IV – information value



# Preselekcja zmiennych

$$IS = \sum (t_i - v_i) \ln\left(\frac{t_i}{v_i}\right),$$

$$KL = \sum t_i \ln\left(\frac{t_i}{v_i}\right),$$

$$IV = \sum (g_i - b_i) \ln\left(\frac{g_i}{b_i}\right),$$

$t_i, v_i$  - udziały i – tego segmentu w train,  
valid

$g_i, b_i$  - udziały dobrych, złych





# Preselekcja - benchmarki

- Kryteria akceptacji:
  - Gini > 5%
  - AR\_diff < 5%, 20%
  - KL, IS (PSI) < 0.1, 0.5
  - KL, IS tylko dla złych < 0.1



# Preselekcja – potencjał danych

- SAS: out.Variables\_stat\_1step
- Python: Gini\_vars.xlsx

	variable	ar_train
1	WOE_ACT6_N_ARREARS	48.44%
2	WOE_ACT3_N_ARREARS	48.06%
3	WOE_ACT9_N_ARREARS	47.73%
4	WOE_ACT_CCSS_DUEUTL	45.55%
5	WOE_ACT12_N_ARREARS	44.87%
6	WOE_ACT_CCSS_MAXDUE	44.17%
7	WOE_ACT_CCSS_UTL	43.14%
8	WOE_ACT_CCSS_N_LOANS_ACT	42.38%
9	WOE_ACT_CCSS_MIN_LNINST	36.79%
10	WOE_ACT_CCSS_MIN_PNINST	29.71%
11	WOE_ACT_CCSS_N_STATC	27.19%
12	WOE_ACT_CCSS_N_LOANS_HIST	25.62%
13	WOE_ACT_CCSS_SENIORITY	25.22%
14	WOE_ACT_CCSS_MIN_SENIORITY	25.15%

# Raporty zmiennych

Attribute number	Condition	Bad rate (br)	Percent of population (%POP)
1	5 < ACT6_N_ARREARS	77,78%	13,01%
2	4 < ACT6_N_ARREARS <= 5	67,52%	9,27%
3	2 < ACT6_N_ARREARS <= 4	54,35%	21,01%
4	0 < ACT6_N_ARREARS <= 2	31,37%	19,56%
5	not missing(ACT6_N_ARREARS) and ACT6_N_ARREARS <= 0	23,59%	37,15%
			100,00%

Chart - Number bad rate for default12 by year on state EM  
Attributes for variable ACT6\_N\_ARREARS  
Customer number in arrears on all loans

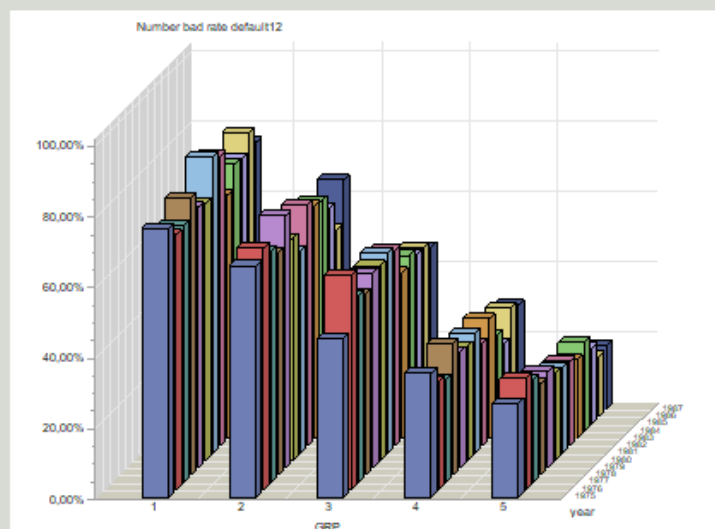
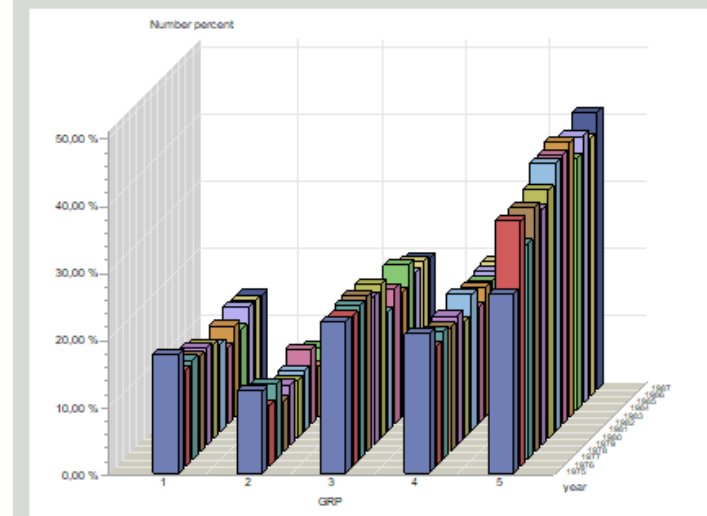


Chart - number distribution by year on state EM  
Attributes for variable ACT6\_N\_ARREARS  
Customer number in arrears on all loans





# Raporty zmiennych

- SAS:
  - W formacie html: interaktywne analizowanie
- Python:
  - W Excelu
- Raporty:
  - Statystyki opisowe
  - Mierniki atrybutów
  - Rozkłady bad rate i udziały atrybutów w czasie
  - Klastry zmiennych

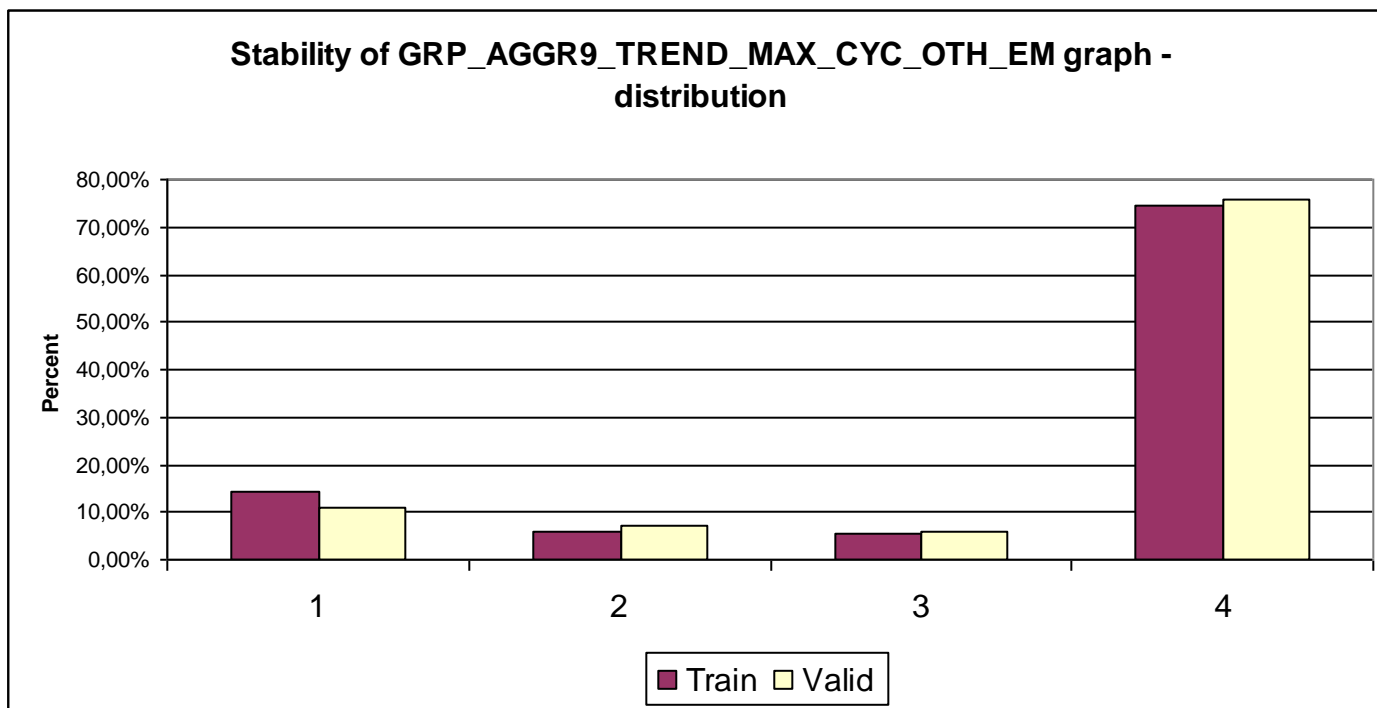


# Raporty zmiennych

- SAS:
  - %include "&dir\_codes.variable\_reports.sas" / source2;
- Python:
  - #Variable\_reportRaporty:

# Stabilność na partycjach

- Statystyki: H\_GRP\_TV and H\_Br\_GRP\_TV





# Klasteryzacja zmiennych

- Zmienne zgrupowane w klastry
- Najbardziej korelują wewnątrz klastrów
- Korelacja pomiędzy klastrami zminimalizowana
- Statystyki Cumulative proportion sugerujące liczbę ortogonalnych wymiarów

Obs	Number	Eigenvalue	Difference	Proportion	Cumulative
1	1	2,83858046	0,8983386	31,54%	31,54%
2	2	1,94024187	0,90106143	21,56%	53,10%
3	3	1,03918044	0,11832011	11,55%	64,64%
4	4	0,92086032	0,10011509	10,23%	74,88%
5	5	0,82074524	0,26466194	9,12%	84,00%
6	6	0,5560833	0,22605046	6,18%	90,17%
7	7	0,33003284	0,01355258	3,67%	93,84%
8	8	0,31648026	0,07868499	3,52%	97,36%
9	9	0,23779527	—	2,64%	100,00%



# Typowa regresja logistyczna – model bez transformacji

$$\text{logit}(p) = \text{Age} * \beta_1 + \text{PaymentHistory} * \beta_2$$

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right)$$



# Transformacja WOE

Attribute	Variable	Partial Score	Formula	
<20	Age	10	woe1	beta1
20>= and <34		20	woe2	
35>=		30	woe3	
Bad	Payment history	10	woe4	beta2
Not good		25	woe5	
Good		40	woe6	



# Transformacja Dummy

Attribute	Variable	Partial Score	Formula
<20	Age	10	beta1
20>= and <34		20	beta2
35>=		30	beta3
Bad	Payment history	10	beta4
Not good		25	beta5
Good		40	beta6



# Czym jest WOE?

- Zmienna płeć (M,K)

$$WOE_M = \ln\left(\frac{\frac{n\_good_M}{n\_bad_M}}{\frac{n\_good_{All}}{n\_bad_{All}}}\right) = \dots = \ln\left(\frac{n\_good_M}{n\_bad_M}\right) - \ln\left(\frac{n\_good_{All}}{n\_bad_{All}}\right)$$

$$WOE_M = \text{logit}(All) - \text{logit}(M)$$

- WOE dla Mężczyzn jest względnym ryzykiem – ilorazem szans (odds) – względem średniego poziomu.



# Bez transformacji – zalety i wady

- Potrzeba uzupełniania braków danych
- Możliwa duża współliniowość i potrzeba jej redukcji
- Problem z nominalnymi cechami
- Trudniejsza interpretacja współczynników modelu
- Czasem modele bardziej stabilne
- Wrażliwość na skrajne wartości



# Transformacja WOE

- Mała możliwość przetrenowania
- Brak uzupełniania braków danych
- Mała współliniowość
- Tak samo traktowane nominalne i interwałowe
- Odporność na skrajne wartości
- Zawsze da się znaleźć dobry model
- Dobre estymacje – mała liczba parametrów



# Transformacja Dummy

- Możliwość przetrenowania
- Trudna weryfikacja założeń
- Za duża liczba parametrów do estymacji – problemy z minimalną wielkością próby przy zadanej mocy testu
- Naiwny Bayes – złe założenie o niezależności zmiennych



# Wielowymiarowa selekcja

- Metody krokowe
  - Python - RFE: Recursive Feature Elimination):
  - SAS: Forward, backward, stepwise
- Heurystyka, wszystkie kombinacje
  - Python: Wszystkie kombinacje
  - SAS: metoda Score

Każdy model, powinien być oceniony różnymi statystykami



# Wielowymiarowa selekcja

- Statystyki stabilności:
  - AR\_diff
- Statystyki współliniowości:
  - Max\_VIF – variance inflation factor,
  - Max\_CI – condition index,
  - Max\_Pearson – korelacja Pearsona,
  - N\_beta\_minus – znak bety
- Statystyki istotności:
  - Max\_ProbChiSq
- Statystyki predykcyjności:
  - Gini train, valid





# Wielowymiarowa selekcja - benchmarki

- Statystyki stabilności:
  - $AR\_diff < 0.1, 0.5$
- Statystyki współliniowości:
  - $Max\_VIF < 3, 5, 10$
  - $Max\_CI < 10, 50, 100,$
  - $Max\_Pearson < 0.7, 0.8, 0.9$
  - $N\_beta\_minus = 0$
- Statystyki istotności:
  - $Max\_ProbChiSq < 0.05$
- Statystyki predykcyjności:
  - Gini train, valid – zależy od typu modelu. App około 50%, Beh – 70%



# Wielowymiarowa selekcja

- SAS:
  - %include "&dir\_codes.steps\_selection.sas" / source2;
  - %include "&dir\_codes.score\_selection.sas" / source2;
- Python:
  - #Simple RFE selection method ...
  - #Assessment of combinations of features
  - number\_vars=12
  - number\_features=6



# Biznesowe kryteria zmiennych i modeli

- Wiarygodność zmiennej
  - Czy można ją zweryfikować?
  - Czy łatwo jest pozyskać tę informację?
  - Czy można te dane manipulować?
  - Czy pochodzą z wiarygodnego źródła danych
- Koszt zmiennej
  - Ile kosztuje zdobycie tych danych
- Inne kryteria:
  - Czy wykluczamy pewne grupy?
  - Czy klient chce to podawać?



# Liczenie ocen punktowych

$$score = \log(odds) * factor + offset =$$

$$(-\sum_{i=1}^n (woe_i * \beta_i) + \alpha) * factor + offset =$$

$$(-\sum_{i=1}^n (woe_i * \beta_i + \frac{\alpha}{n})) * factor + offset =$$

$$\sum_{i=1}^n (-(woe_i * \beta_i + \frac{\alpha}{n}) * factor + \frac{offset}{n})$$

$$600 = \log(50) * factor + offset$$

$$620 = \log(100) * factor + offset$$

$$factor = 20 / \log(2)$$

$$offset = 600 - factor * \log(50)$$



# Jak liczymy oceny punktowe

$$\begin{aligned}\text{WoE}_k &= \ln \left( \frac{G_k/G}{B_k/B} \right) = \\ &= \ln \left( \frac{G_k}{B_k} \right) - \ln \left( \frac{G}{B} \right),\end{aligned}$$

czyli:

$$\text{WoE}_k = \text{Logit}_k - \text{Logit},$$

gdzie  $k$  - oznacza dowolną kategorię zmiennej,  $G$ ,  $B$  - liczby dobrych i złych klientów w całej populacji a  $G_k$  i  $B_k$  w kategorii. Mamy zatem zależność, że Weight of Evidence dla kategorii jest różnicą logitu kategorii i logitu całej populacji. Dlatego w dalszej części nazywamy metodę budowy modelu LOG i wyliczamy logity zamiast WoE.



# Jak liczymy oceny punktowe

Każda zmienna wybrana do modelu jest transformowana do kawałkami stałej na podstawie wyliczonych logitów każdej z jej kategorii. Estymacja regresji logistycznej w ogólnym zapisie określona jest wzorem:

$$\text{Logit}(p_n) = X_n\beta,$$

gdzie  $p_n$  jest prawdopodobieństwem, że klient jest dobrym, inaczej  $p_n = P(Y = \text{Dobry})$  dla  $n$  – tej obserwacji, a  $\beta$  reprezentuje wektor współczynników regresji. Macierz  $X_n$  można szczegółowo rozpisać w następujący sposób:

$$X_n = l_{ij}\delta_{ijn},$$

gdzie  $l_{ij}$  jest logitem  $j$  – tej kategorii  $i$  – tej zmiennej a  $\delta_{ijn}$  jest macierzą zerojedynekową przyjmującą wartość jeden, gdy  $n$  – ta obserwacja należy do  $j$  – tej kategorii  $i$  – tej zmiennej. Dodatkowo przyjęto uproszczone założenie, że każda zmienna ma tyle samo kategorii, aby nie wprowadzać większej liczby indeksów oraz że liczba kategorii jest taka sama co liczba zmiennych i wynosi  $v$ .



# Jak liczymy oceny punktowe

Iloczyn macierzy  $X$  i wektora  $\beta$  stojący po prawej stronie równania regresji jest wartością oceny punktowej (ang. score) dla danej obserwacji. Ocena ta nie jest skalibrowana i ciężko ją interpretować. Zwykle wykonuje się kilka prostych przekształceń, aby nadać jej bardziej użyteczną formę. Zauważmy, że jeśli wartość prawdopodobieństwa  $p_n$  rośnie, to jego logit także, a zatem ocena punktowa również będzie rosnać. Czyli im większa ocena punktowa, tym większe prawdopodobieństwo spłacenia kredytu. Najczęściej kalibruje się wartość oceny punktowej poprzez prostą funkcję liniową:

$$\text{Logit}(p_n) = \ln \left( \frac{p_n}{1 - p_n} \right) = S_n = aS_n^{\text{New}} + b,$$

gdzie  $S_n^{\text{New}}$  jest nową oceną a  $S_n$  starą, natomiast  $a$  i  $b$  są współczynnikami. Wyznacza się je tak, aby uzyskać dodatkową własność, którą w książce definiuje się w następujący sposób: dla wartości 300 punktów szansa (ang. odds) bycia dobrym klientem powinna wynosić 50, a gdy szansa zwiększy się dwukrotnie (ang. double to odds), czyli będzie 100, to ocena powinna być 320. Szansę definiuje się jako iloraz liczby dobrych do złych klientów, lub jako stosunek  $\frac{p_n}{1-p_n}$ . Szansa 50 reprezentuje zatem segment klientów, gdzie na jednego złego przypada 50-siąt dobrych. Należy zatem rozwiązać układ równań (Siddiqi,



# Jak liczymy oceny punktowe

$$\ln(50) = a \cdot 300 + b,$$

$$\ln(100) = a \cdot 320 + b.$$

Jego rozwiązaniem są wartości:

$$a = \frac{\ln\left(\frac{100}{50}\right)}{20} = \frac{\ln(2)}{20},$$

$$b = \ln(50) - \frac{300 \ln\left(\frac{100}{50}\right)}{20} = \ln\left(\frac{50}{2^{15}}\right).$$

Drugą czynnością skalującą wartość oceny punktowej jest zadbanie, by wszystkie oceny cząstkowe pierwszej kategorii miały taką samą liczbę punktów. Pierwsza kategoria reprezentowana jest przez grupę najbardziej ryzykownych klientów. Ostatnia reprezentuje najlepszych, jeśli oceny cząstkowe zaczynają się zawsze od tej samej wartości, to ta zmienna, która posiada największą wartość oceny cząstkowej może być interpretowana jako najważniejsza w modelu.

Mamy dalej:

$$S_n = \sum_{i,j=1}^v \beta_{ij} l_{ij} \delta_{ijn} + \beta_0.$$





# Jak liczymy oceny punktowe

Możemy wydzielić człon związany z najgorszym klientem:

$$\gamma = \sum_{i=1}^v \beta_i l_{i1},$$

i dzięki temu wyraz wolny rozdzielić na dwa składniki:

$$\beta_0 = \sum_{i=1}^v \frac{\beta_0 + \gamma}{v} - \sum_{i=1}^v \beta_i l_{i1}.$$

W ten sposób powstaje ocena cząstkowa (ang. partial score):

$$P_{ij} = \beta_i l_{ij} + \frac{\beta_0 + \gamma}{v} - \beta_i l_{i1}.$$

Zauważmy, że dla każdej zmiennej  $i$  mamy:

$$P_{i1} = \frac{\beta_0 + \gamma}{v},$$

czyli oceny cząstkowe zaczynają się od tej samej wartości.



# Jak liczymy oceny punktowe

Mamy dalej:

$$S_n = \sum_{i,j=1}^v P_{ij} \delta_{ijn},$$

oraz finalnie:

$$S_n^{New} = \frac{S_n - b}{a} = \sum_{i,j=1}^v P_{ij}^{New} \delta_{ijn},$$

gdzie

$$P_{ij}^{New} = \frac{1}{a} P_{ij} - \frac{b}{v}.$$

Ostateczną wartość oceny częściowej często zaokrągla się do najbliższej wartości całkowitej. W ten sposób otrzymuje się kartę skoringową z naliczonymi punktami przy każdej z kategorii ze zmiennych wybranych do modelu.



# Liczenie ocen punktowych

- Omówić kod:
- `different_betas.sas`



# Liczenie ocen punktowych

- SAS:
  - %include  
"&dir\_codes.model\_assessment.sas" /  
source2;
- Python:
  - #Assessment of combinations of features
  - #Creating Scorecard

# Własności karty skoringowej

- Najważniejsza zmienna ma najwyższą ocenę cząstkową.

Scale of variable's scorecard points				
Variable	Minimum of scorecard	Maximum of scorecard points	Range of scorecard points	Part of global range
APP_CHAR_JOB_CODE	8	115	107	29.08%
ACT_CCSS_N_STATC	8	79	71	19.29%
ACT_CCSS_DUEUTL	8	70	62	16.85%
ACT_CC	8	61	53	14.40%
ACT12_N_ARREARS	8	59	51	13.86%
ACT_CCSS_MIN_LNINST	8	32	24	6.52%

Gini statistics for variables in the model	
Variable	Gini statistics for variable
ACT_CCSS_DUEUTL	45,53%
ACT12_N_ARREARS	44,87%
ACT_CCSS_MIN_LNINST	36,79%
ACT_CCSS_N_STATC	27,19%
ACT_CC	14,75%
APP_CHAR_JOB_CODE	7,45%



# Dokumentacja modelu

- SAS:
  - %include "&dir\_codes.final\_report.sas" /  
source2;
- Puython:
  - #Model report
- Można też model SASowy zaraportować do Excela kodami w Python



# Kod do skorowania

- SAS:
  - %include "&dir\_codes.scoring\_code.sas" /  
source2;
- Puython:
  - # Scoring code

# Macierz klasyfikacji (pomyłek) – ang. Confusion Matrix

		Observed	
		True	False
Predicted	True	True Positive (TP)	False Positive (FP)
	False	False Negative (FN)	True Negative (TN)





# Podstawowe pojęcia

- Dla ustalonego  $c$  – cutoff:
  - $TP + FN = P$  (observed positive)
  - $TN + FP = N$  (observed negative)
  - $TP + FP = PP$  (predicted positive)
  - $TN + FN = PN$  (predicted negative)
  - $FPrate = FP/N$ ,
  - $TPrate = TP/P = Recall$ ,
  - $Accuracy = PCC = (TP + TN) / (P + N)$



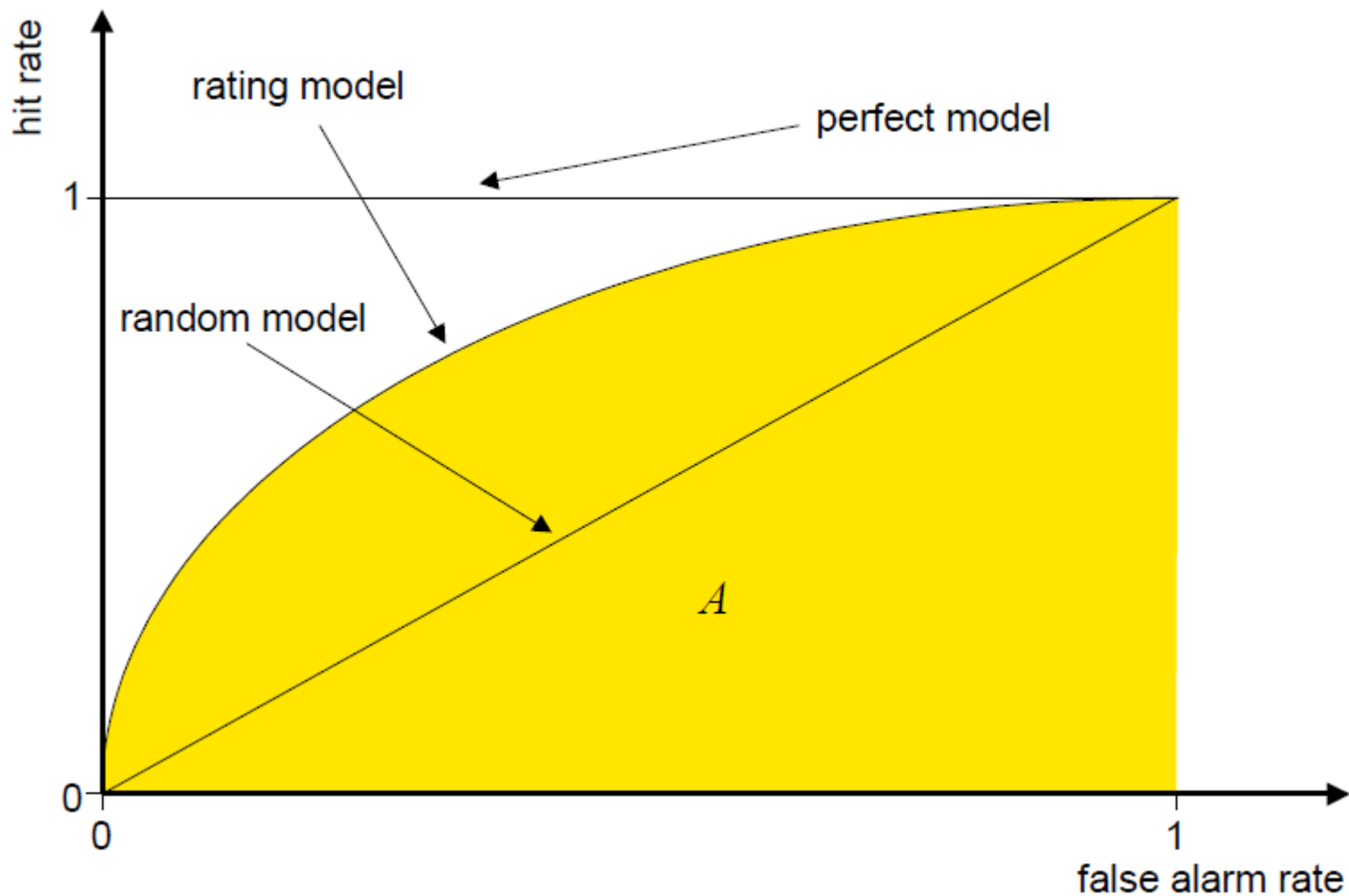
# Podstawowe pojęcia

- Specificity = Specyficzność =  $TN/N$
- $PV+ = TP/PP$  (response rate),
- $PV- = TN/PN$

ROC (Receiver Operating Characteristic):

- $x = FPrate = 1 - Specificity = \text{false alarm rate}$
- $y = TPrate = Sensitivity = \text{Czułość} = \text{hit rate}$

# ROC (Receiver Operating Characteristic) AUC (Area Under Curve)

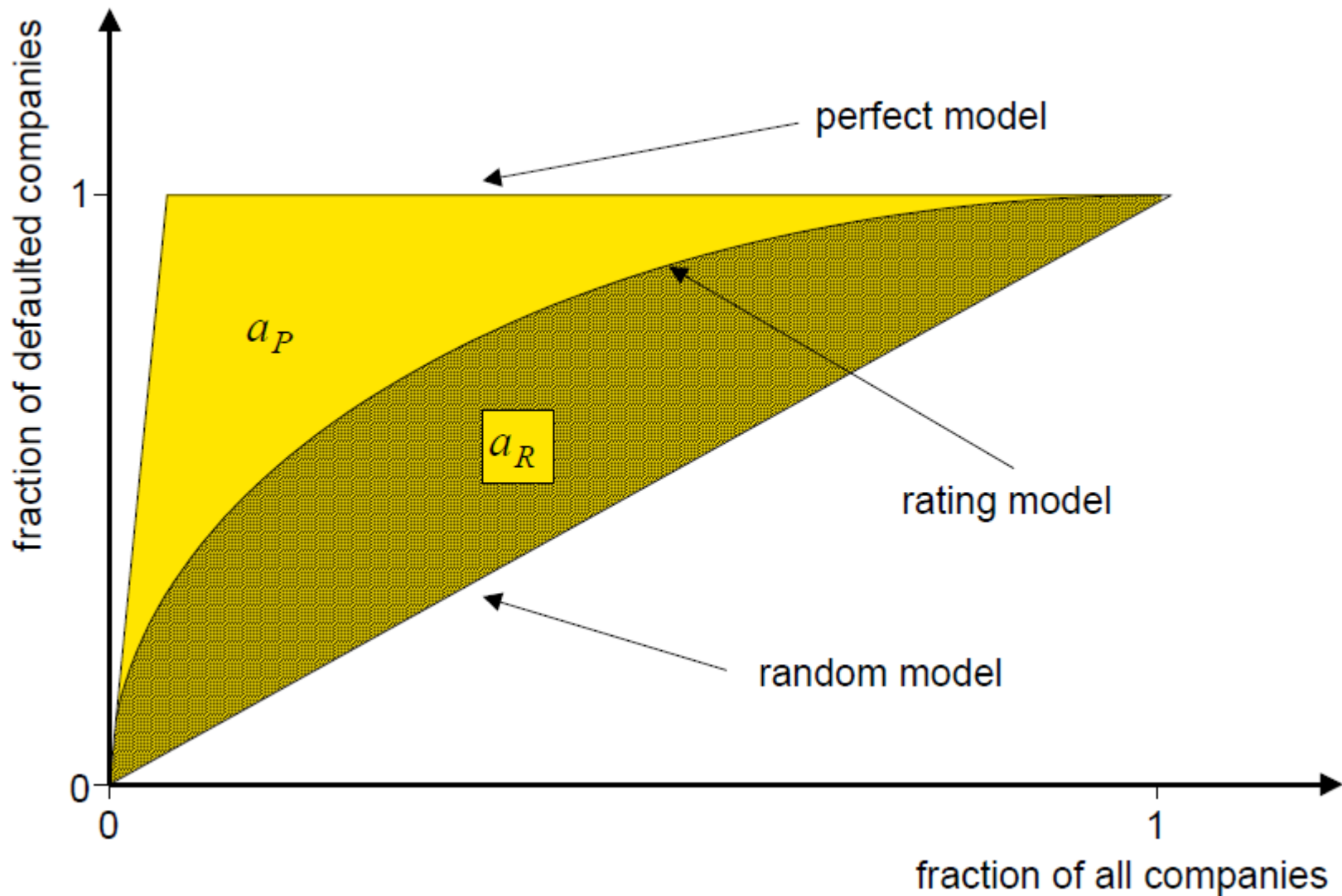


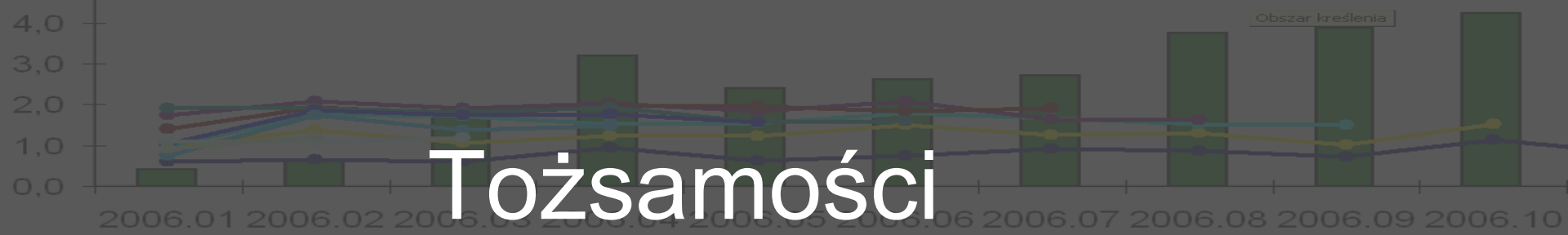


# Krzywe CAP, Lift, Gains, Lorentz

- Depth – penetration rate – population share – jaki udział powyżej cutoff
- $\text{Rho1} = P/(P+N)$  – response rate populacji
- Gains:
  - $x = \text{Depth}$ ,  $y = \text{TPRate} = \text{TP}/P = \text{Recall}$ , ile procent jedynek w wybranym zbiorze ze wszystkich jedynek
- Lift:
  - $x = \text{Depth}$ ,  $y = \text{PV+}/\text{Rho1}$ , ile razy lepiej od modelu losowego
- Lorentz (concentration curve, CAP) :
  - $x = \text{Depth}$ ,  $y = \text{Sensitivity}$

# CAP (Cumulative Accuracy Profiles)





# Tożsamości

- $AR = Gini = a_p/a_r$
- $AUC = C = A$
- $2 \cdot C - 1 = AR$



# Gini

$$\begin{aligned}
 c &= (n_c + 0.5(t - n_c - n_d)) / t \\
 \text{Somers' } D \text{ (Gini coefficient)} &= (n_c - n_d) / t \\
 \text{Goodman-Kruskal Gamma} &= (n_c - n_d) / (n_c + n_d) \\
 \text{Kendall's Tau-}a &= (n_c - n_d) / (0.5N(N - 1))
 \end{aligned}$$

- $n_c$  – liczba zgodnych ( $P_i > P_j$ , gdy i-zły, j-dobry),  $P = P(\text{bycie złym}, Y=1)$
- $n_d$  – liczba niezgodnych
- $t$  – liczba wszystkich par
- $\text{Gini} = P_c - P_d$



# Gini - interpretacja

- $Gini = P_c - P_d$
- $P_c + P_d + P_t = 100\%$
- Zakładając, że  $P_t = 0$  mamy:  
 $P_c + P_d = 100\%$   
 $Gini = 2 P_c - 1$   
 $P_c = (Gini + 1) / 2$



# Macierz kosztów ang. Cost Matrix

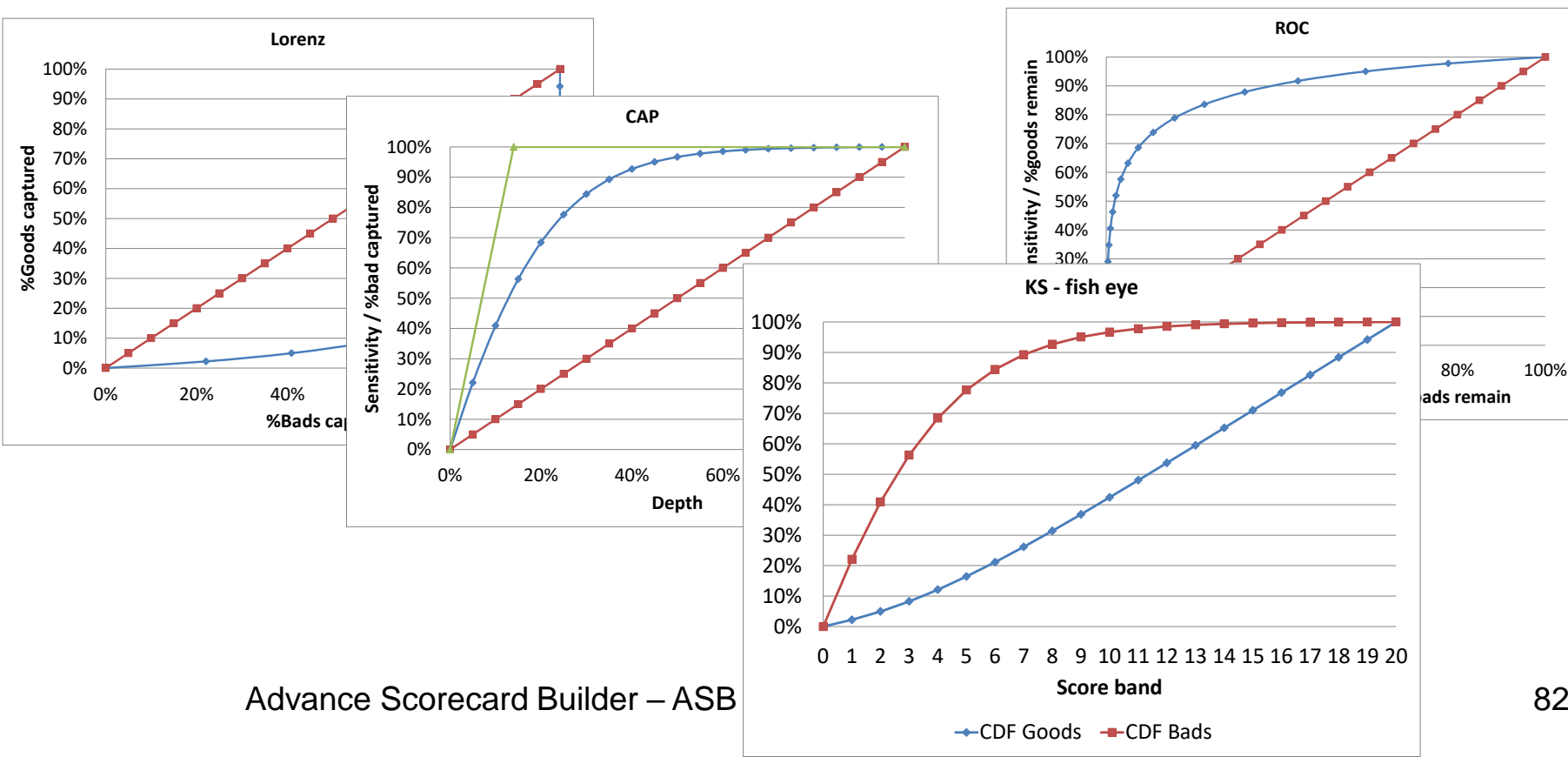
		Observed	
		True	False
Predicted	True	$C_{TP}$ (TP)	$C_{FP}$ (FP)
	False	$C_{FN}$ (FN)	$C_{TN}$ (TN)

# Krzywe w excelu

Załączone Excele z regułami i praktycznymi wskaźnikami

<http://administracja.sgh.waw.pl/pl/OW/publikacje/Strony/2015.aspx>

[http://administracja.sgh.waw.pl/pl/OW/publikacje/Documents/gini\\_curves.xlsx](http://administracja.sgh.waw.pl/pl/OW/publikacje/Documents/gini_curves.xlsx)





# Cykl życia modelu

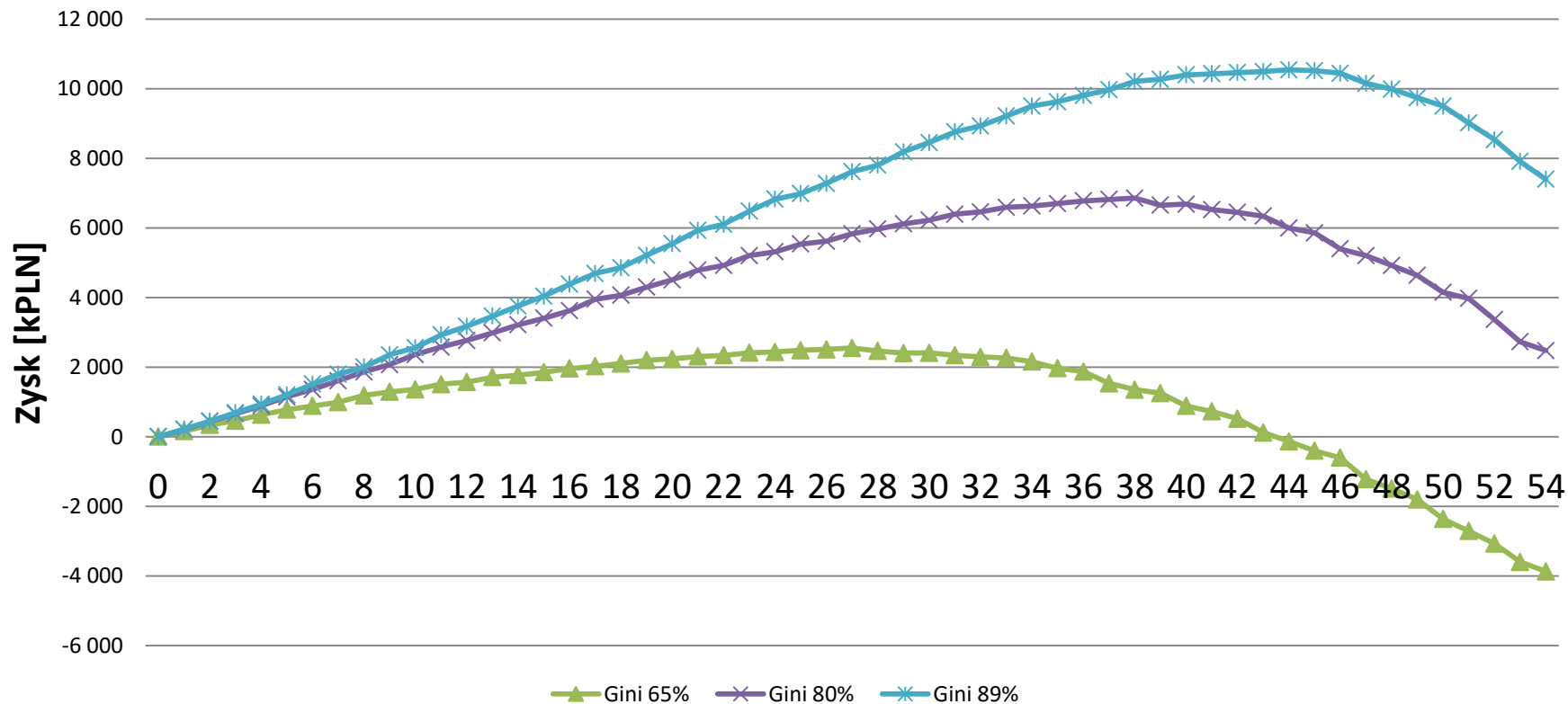
- Wniosek o nowy model (model request)
- Budowa modelu (model building)
- Walidacja - zatwierdzenie (Validation)
- Wdrożenie (Implementation)
- Monitoring
- Przegląd monitoringu (Monitoring review)
- Decyzja o zmianie modelu
  
- Każdy punkt to inny dokument



- SAS: moniroting.sas
- Katalog:  
...\CS-AUT\software\ASB\_SAS\monitoring\

# Krzywa Profit-Loss

Krzywa Profit zależna od mocy predykcyjnej





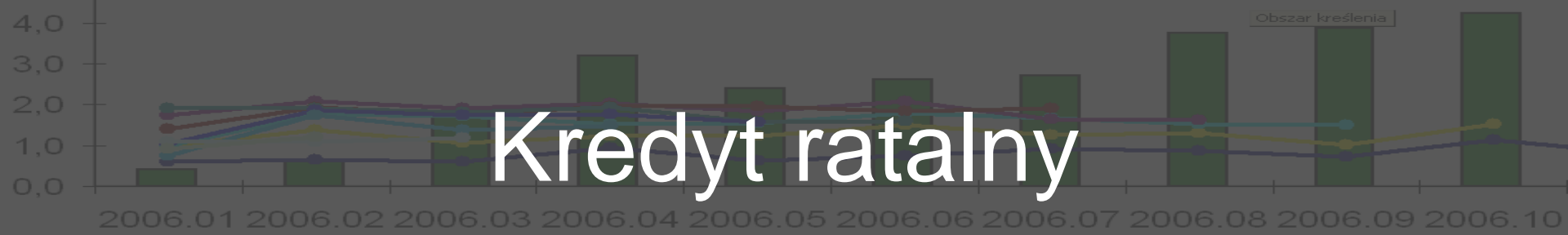
# Założenia budowy danych

- Klient zawsze gdzieś dostanie kredyt, jeśli nie w banku to w para-banku albo od znajomych lub od rodziny
- Klient ma swoje priorytety, jedne kredyty spłaca, inne nie
- Spłacalność kredytów gotówkowych zależy od wcześniejszej historii, włączając spłacalność kredytów ratałnych
- Mamy zatem potencjał danych, już wygenerowanych z całą historią spłacalności

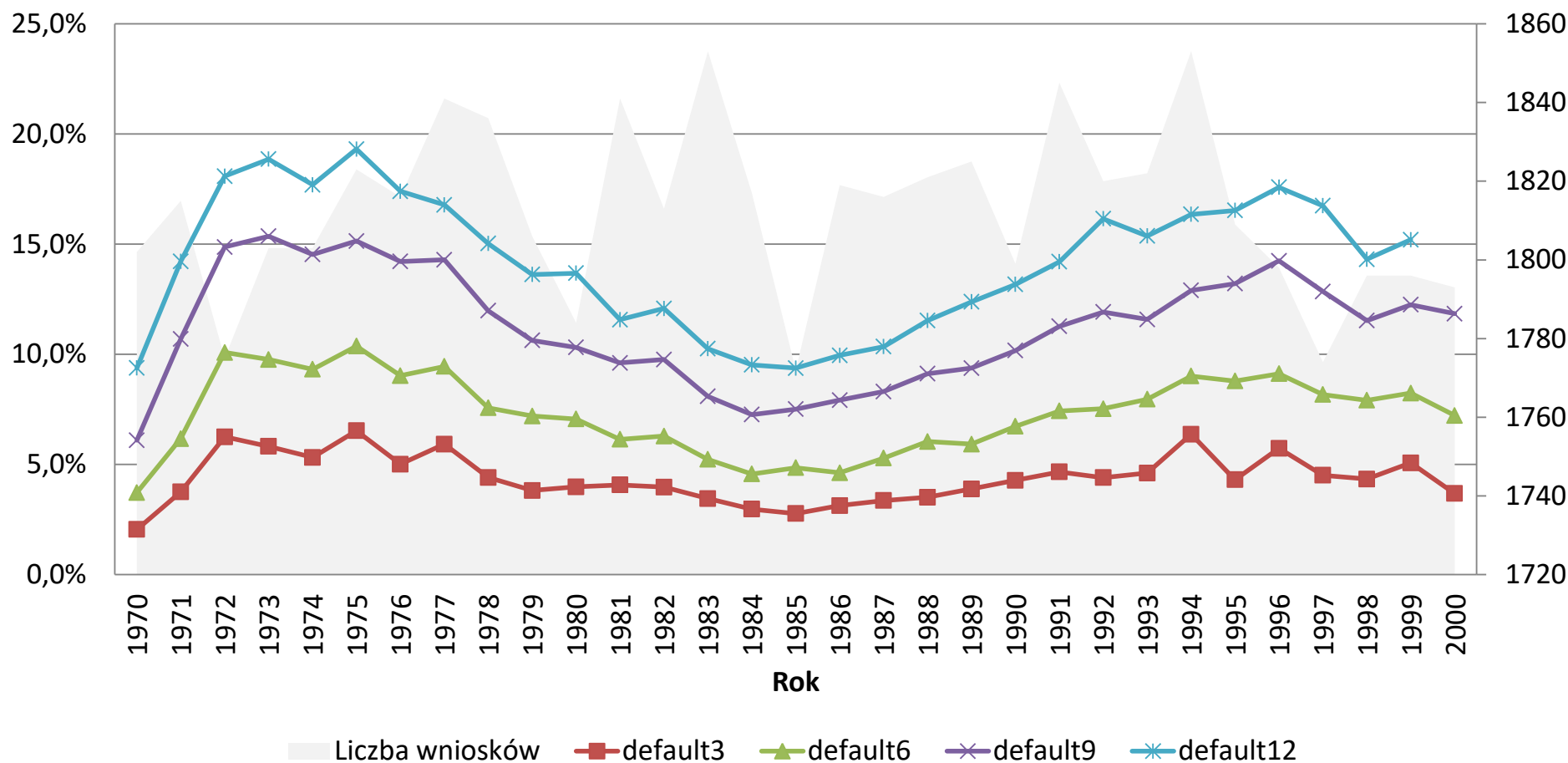


# Założenia budowy danych

- Bank może wybierać, które kredyty klienta akceptuje, dzięki czemu zmniejsza stratę
- Jeśli jednak nie akceptuje wszystkich kredytów klienta, to bank traci cenną informację o kliencie, wie tylko o lepszej stronie klienta
- Powstaje zatem problem wniosków odrzuconych (ang. Reject Inference)
- Dodatkowo powstaje też brak okazji do sprzedaży kredytu gotówkowego, bo klient był odrzucony wcześniej aplikując o kredyt ratałny



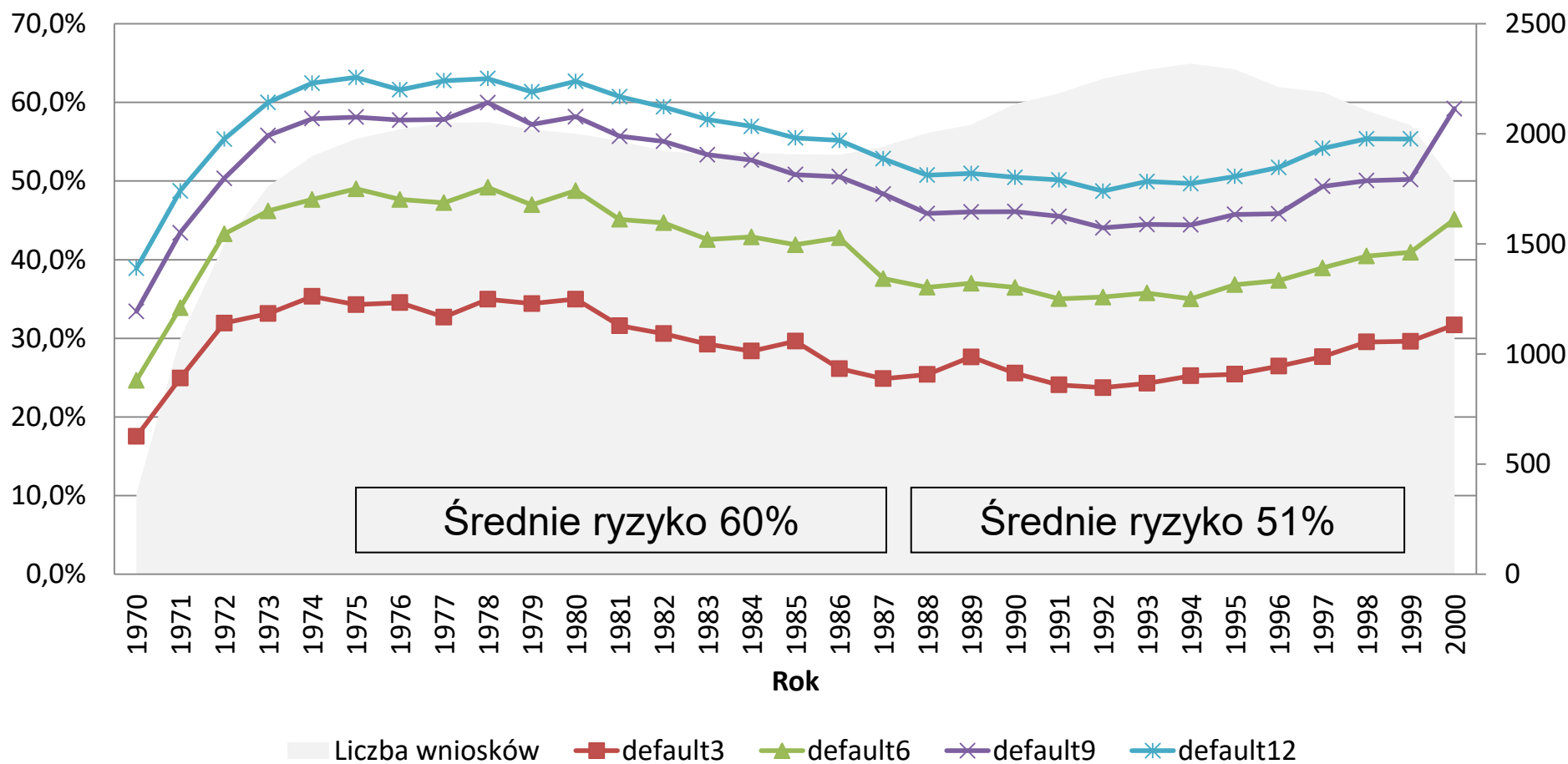
## Zmiany ryzyka i produkcji dla kredytu ratalnego





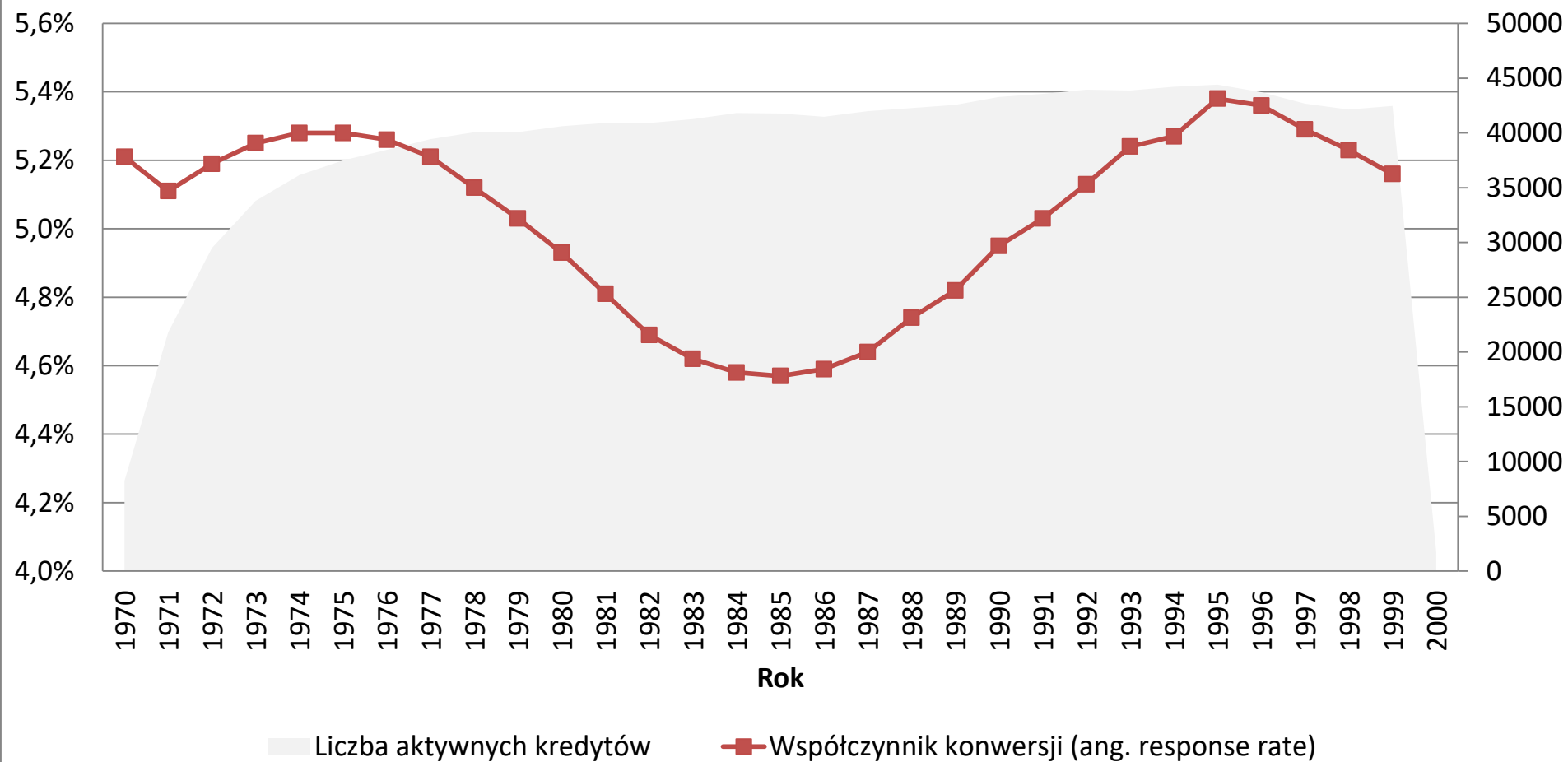
# Kredyt gotówkowy

## Zmiany ryzyka i produkcji dla kredytu gotówkowego



# Portfele miesięczne

Zmiany portfela aktywnego obu produktów i współczynnika konwersji





# Wyzwanie (okres 1975-1987)

Wskaźnik	Ratalny	Gotówkowy	Razem
Zysk	-7 824 395	-31 627 311	-39 451 706
Przychód	969 743	10 260 689	11 230 432
Strata	8 794 138	41 888 000	50 682 138

- 4 modele kart skoringowych (estymowane na całej populacji z okresu 1975-1987):
  - Model ryzyka dla kredytu ratalnego (PD Ins)
  - Model ryzyka dla kredytu gotówkowego (PD Css)
  - Model ryzyka dla kredytu gotówkowego w momencie aplikowania o kredyt ratalny (Cross PD Css)
  - Model skłonności skorzystania z kredytu gotówkowego w momencie aplikowania o kredyt ratalny (PR Css) (ang. response, propensity model)



# Okres 1975-1987

- Kalibracja modeli do prawdopodobieństwa:

$$PD\_Ins = 1 / (1 + \exp(-(-0.032205144 * risk\_ins\_score + 9.4025558419)))$$

$$PD\_Css = 1 / (1 + \exp(-(-0.028682728 * risk\_css\_score + 8.1960829753)))$$

$$Cross\_PD\_Css = 1 / (1 + \exp(-(-0.028954669 * cross\_css\_score + 8.2497434934)))$$

$$PR\_Css = 1 / (1 + \exp(-(-0.035007455 * response\_score + 10.492092793)))$$

Model	Gini
Cross PD Css	74,01%
PD Css	74,21%
PD Ins	73,11%
PR Css	86,37%



# Optymalizacja gotówki

- Badając całą populację z okresu 1975-1987, wyznaczamy krzywą profit i znajdujemy optymalny punkt:
  - reguła odrzucenia  $PD\_Css > 27,24\%$
  - procent akceptacji gotówki 18,97%
  - zysk dla gotówki 1 591 633 PLN
- Czy postąpić podobnie z kredytem ratalnym?



# Customer Life Time Value (CLTV)

- Każdy kredyt ratałny jest szansą do zarobienia, jeśli tylko klient skorzysta z gotówki.
- Trzeba zatem rozważyć ciąg produktowy: pierwszy kredyt ratałny, drugi gotówkowy.
- Tworzymy reguły dzieląc populację na grupy wyznaczone przez estymację ryzyka ratałnego i estymację potencjalnej gotówki

# Segmentacja CLTV

GR PR Css	GR PD Ins	Liczba wniosków Ins	Globalny zysk	Min PR Css	Max PR Css	Min PD Ins	Max PD Ins
4	0	1 277	372 856	4,81%	96,61%	0,02%	2,18%
4	1	581	96 096	4,81%	96,61%	2,25%	4,61%
1	0	2 452	67 087	1,07%	1,07%	0,32%	2,18%
3	0	907	46 685	2,80%	4,07%	0,07%	2,18%
3	1	734	14 813	2,80%	4,07%	2,25%	4,61%
3	2	307	12 985	2,80%	4,07%	4,76%	7,95%
4	2	361	8 039	4,81%	96,25%	4,76%	7,95%
3	3	446	-1 283	2,80%	4,07%	8,19%	18,02%
4	3	417	-5 774	4,81%	95,57%	8,19%	18,02%
1	1	3 570	-82 886	1,07%	1,07%	2,25%	4,61%
1	2	4 044	-408 644	1,07%	1,07%	4,76%	7,95%
3	4	726	-946 937	2,80%	4,07%	18,50%	99,62%
4	4	1 054	-1 108 313	4,81%	96,25%	18,50%	99,83%
1	3	3 883	-1 270 930	1,07%	1,07%	8,19%	18,02%
1	4	2 878	-4 306 859	1,07%	1,07%	18,50%	97,00%



# Reguły CLTV ratałnego

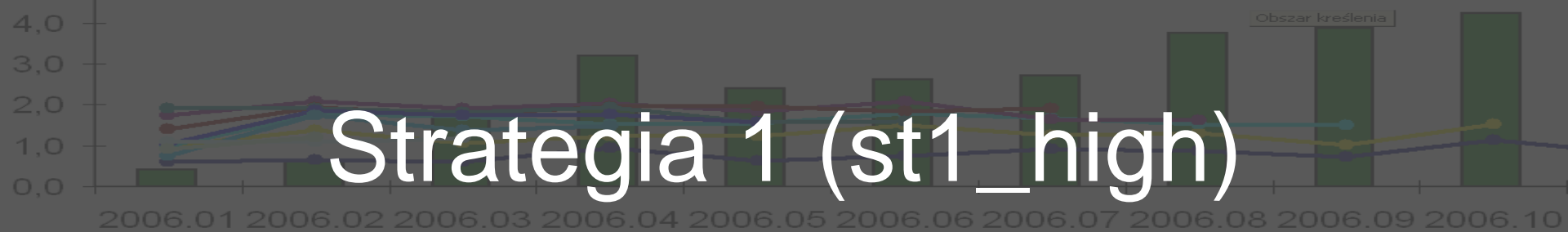
- Reguły odrzucenia:
  - $PD\_Ins > 8,19\%$
  - $8,19\% \geq PD\_Ins > 2,18\%$  i ( $PR\_Css < 2,8\%$  lub  $Cross\_PD\_Css > 27,24\%$ )
- Estymowany globalny zysk z połączonego procesu:  
1 686 684 PLN
- Reguła bez  $PR\_Css$ :
  - $PD\_Ins > 8,19\%$
- Estymowany globalny zysk z połączonego procesu:  
1 212 261 PLN, czyli 30% mniej!





# System/Silnik decyzyjny

- Każdy zestaw reguł trzeba przeprocesować, gdyż w zależności od decyzji kredytowych zmieniają się rozkłady skoringów, bo zmieniają się rozkłady zmiennych opisujących klientów
- Testujemy zatem kilka strategii
  - Strategia 1 – znalezione wcześniej reguły
  - Strategia 2 – bez reguły na PR\_Css
  - Strategia 3 – odrzut złego klienta (był default)
  - Strategia 4 – nowe reguły na bazie strategii 3



# Strategia 1 (st1\_high)

Reguła	Opis
PD_Ins Cutoff	$PD\_Ins > 8,19\%$
PD_Css Cutoff	$PD\_Css > 27,24\%$
PD i PR	$8,19\% \geq PD\_Ins > 2,18\%$ i $(PR\_Css < 2,8\%$ lub $Cross\_PD\_Css > 27,24\%)$

## Kredyt gotówkowy

Reguła	Liczba wniosków	Procent wniosków	Kwota	Ryzyko	Zysk
PD_Css Cutoff	8 436	32,97%	42 180 000	67,99%	-13 098 591
Nieznany klient	12 999	50,80%	64 995 000	65,91%	-19 171 357
Akceptacja	4 152	33,33% 16,23%	20 760 000	22,35%	642 637
Razem	25 587	100,00%	127 935 000	59,53%	-31 627 311

## Kredyt ratalny

Reguła	Liczba wniosków	Procent wniosków	Kwota	Ryzyko	Zysk
PD_Ins Cutoff	9 289	39,30%	60 214 008	26,95%	-7 339 423
PD i PR	8 131	34,40%	31 340 808	5,37%	-505 662
Akceptacja	6 217	26,30%	22 698 240	2,14%	20 690
Razem	23 637	100,00%	114 253 056	13,00%	-7 824 395

# Strategia 1

Okres	Przychód	Strata	Zysk
1975-1987	3 407 745	2 744 418	663 327
1988-1998	3 761 299	2 246 844	1 514 455

A miało być  
1 686 684 PLN

## Średnie wartości parametrów

Parametr	Akceptacja	Razem
PD (razem PD Ins i PD Css)	7,93%	28,87%
PR Css	17,15%	21,76%
Cross PD Css	21,71%	17,73%

## Moc predykcyjna (Gini)

Model	Akceptacja	Razem
Cross PD Css	21,34%	40,72%
PD Css	31,66%	53,28%
PD Ins	41,93%	68,58%
PR Css	72,56%	68,88%

# Decyzje dynamiczne

- R -> G -> G -> G
- R -> G -> R -> G -> G

R -> G -> ~~G~~ -> G

R -> ~~G~~ -> R -> G -> G

R -> G -> ~~G~~ -> ~~G~~

R -> ~~G~~ -> ~~R~~ -> ~~G~~ -> ~~G~~



# Istotny błąd estymacji

- Dlaczego zamiast 1 686 684 PLN zarobiliśmy tylko 663 327 PLN?
- Gdzie podział się nasz milion?
- Wpływ odrzuconych (rewolucja w procesie, od 100% akceptacji):
  - Nieznany klient – 50,8%
  - Akceptacja ratalnego – 26,3%
  - Akceptacja gotówkowego – 16,23%
  - PD (razem PD\_Ins i PD\_Css) z 37,19% na 28,87%

# Strategia 2 (st3\_low)

Okres	Przychód	Strata	Zysk
1975-1987	4 008 258	3 896 818	111 441
1988-1998	4 539 328	3 829 634	709 694

Czyli o 551 886 PLN mniej,  
aż o 83% mniej

Reguła	Opis
PD_Ins Cutoff	$PD_{Ins} > 8,19\%$
PD_Css Cutoff	$PD_{Css} > 27,24\%$

## Kredyt gotówkowy

Reguła	Liczba wniosków	Procent wniosków	Kwota	Ryzyko	Zysk
PD_Css Cutoff	9 297	36,33%	46 485 000	67,84%	-14 381 482
Nieznany klient	11 661	45,57%	58 305 000	67,34%	-17 822 432
Akceptacja	4 629	18,09%	23 145 000	23,16%	576 604
Razem	25 587	100,00%	127 935 000	59,53%	-31 627 311

## Kredyt ratalny

Reguła	Liczba wniosków	Procent wniosków	Kwota	Ryzyko	Zysk
PD_Ins Cutoff	9 325	39,45%	60 221 856	26,98%	-7 359 232
Akceptacja	14 312	60,55%	54 031 200	3,89%	-465 163
Razem	23 637	100,00%	114 253 056	13,00%	-7 824 395

# Strategia 3 (st4\_bad\_due3)

Okres	Przychód	Strata	Zysk
1975-1987	7 496 614	21 801 230	-14 304 616
1988-1998	7 881 992	18 510 342	-10 628 350

Reguła	Opis
Zły klient	<i>agr12_Max_CMaxA_Due &gt; 3</i>

## Kredyt gotówkowy

Reguła	Liczba wniosków	Procent wniosków	Kwota	Ryzyko	Zysk
Zły klient	7 114	27,80%	35 570 000	79,83%	-14 195 320
Nieznany klient	7 036	27,50%	35 180 000	67,04%	-10 673 871
Akceptacja	11 437	44,70%	57 185 000	42,28%	-6 758 120
Razem	25 587	100,00%	127 935 000	59,53%	-31 627 311

## Kredyt ratalny

Reguła	Liczba wniosków	Procent wniosków	Kwota	Ryzyko	Zysk
Zły klient	483	2,04%	2 047 188	27,74%	-277 899
Akceptacja	23 154	97,96%	112 205 868	12,69%	-7 546 496
Razem	23 637	100,00%	114 253 056	13,00%	-7 824 395

# Strategia 3

Średnie wartości parametrów

Parametr	Akceptacja	Razem
PD (razem PD Ins i PD Css)	21,81%	32,70%
PR Css	21,79%	28,83%
Cross PD Css	43,09%	24,48%

Moc predykcyjna (Gini)

Model	Akceptacja	Razem
Cross PD Css	64,83%	63,59%
PD Css	63,67%	64,82%
PD Ins	71,94%	72,56%
PR Css	79,96%	64,72%

- Przy tej strategii jeszcze nie zarabiamy, ale już modyfikujemy rozkłady skoringów na akceptowanej części



# Strategia 4 (st5\_from\_due3)

Reguła	Opis
Zły klient	$agr12\_Max\_CMaxA\_Due > 3$
PD_Ins Cutoff	$PD\_Ins > 7,95\%$
PD_Css Cutoff	$PD\_Css > 19,13\%$
PD i PR	$7,95\% \geq PD\_Ins > 2,8\%$ i $(PR\_Css < 2,8\% \text{ lub } Cross\_PD\_Css > 19,13\%)$

## Kredyt gotówkowy

Reguła	Liczba wniosków	Procent wniosków	Kwota	Ryzyko	Zysk
Zły klient	2 253	8,81%	11 265 000	74,26%	-4 026 033
PD_Css Cutoff	5 375	21,01%	26 875 000	53,66%	-5 462 687
Nieznany klient	15 739	61,51%	78 695 000	65,29%	-22 845 756
Akceptacja	2 220	8,68%	11 100 000	17,97%	707 165
Razem	25 587	100,00%	127 935 000	59,53%	-31 627 311

## Kredyt ratalny

Reguła	Liczba wniosków	Procent wniosków	Kwota	Ryzyko	Zysk
Zły klient	209	0,88%	891 720	27,75%	-121 550
PD_Ins Cutoff	9 253	39,15%	60 130 704	26,46%	-7 208 030
PD i PR	8 029	33,97%	31 118 232	5,49%	-519 531
Akceptacja	6 146	26,00%	22 112 400	2,05%	24 717
Razem	23 637	100,00%	114 253 056	13,00%	-7 824 395



Okres	Przychód	Strata	Zysk
1975-1987	2 010 242	1 278 361	731 882
1988-1998	2 452 716	1 134 729	1 317 986

#### Średnie wartości parametrów

Parametr	Akceptacja	Razem
PD (razem PD Ins i PD Css)	4,24%	25,17%
PR Css	11,37%	15,68%
Cross PD Css	17,02%	14,61%

#### Moc predycyjna (Gini)

Model	Akceptacja	Razem
Cross PD Css	3,23%	19,19%
PD Css	33,15%	47,81%
PD Ins	36,79%	67,67%
PR Css	70,59%	64,89%



## Strategia 1

## Strategia 4

Okres	Przychód	Strata	Zysk	Przychód	Strata	Zysk
1975-1987	3 407 745	2 744 418	663 327	2 010 242	1 278 361	731 882
1988-1998	3 761 299	2 246 844	1 514 455	2 452 716	1 134 729	1 317 986

- W okresie prosperity Strategia 1 jest lepsza.
- W okresie zwiększonego ryzyka Strategia 4 jest lepsza.



# Wnioski

- Wpływ odrzuconych wniosków w procesie akceptacji jest trudny do przewidzenia
- Bezpieczne rozwiązanie w zarządzaniu procesem to powolne zmiany reguł polityki
- **Nigdy nie wykonywać rewolucyjnych zmian!**
- Strategie muszą się zmieniać
- Ciągłe doskonalenie, ciągłe testowanie nowych modeli i reguł



- Jak uruchomić?
- Jak zmieniać reguły w silniku?
- Podstawowe raporty
- Tylko w SAS: all\_contents.sas
  - Katalog:  
...\CS-AUT\software\PROCSS\_SIMULATION\codes\



# Czym jest kalibracja

W efekcie można mówić o dwóch rodzajach kalibracji. Pierwszej transformującej wartość prawdopodobieństwa  $p_n$  do oceny punktowej, gdzie bierze udział przekształcenie logitowe oraz z oceny punktowej do prawdopodobieństwa, gdzie używa się funkcji odwrotnej do logitu postaci:

$$p_n = \frac{1}{1 + e^{-(\omega_s S_n^{New} + \omega_0)}},$$

gdzie  $\omega_s$  i  $\omega_0$  są współczynnikami.



- Omówić kody w katalogu:  
...\CS-AUT\software\PROCSS\_SIMULATION\process\segmentation\

# Segmentacja

Observed - expected risk					
Segments	N	Pct	Risk	PD	PD Seg
<b>All</b>	23 637	100,00%	13,00%	13,00%	13,00%
<b>Miss</b>	16 827	71,19%	12,61%	11,34%	12,61%
<b>NMiss</b>	6 810	28,81%	13,96%	17,09%	13,96%

Predictive powers		
Segments	PD	PD seg
<b>All</b>	71,13%	76,06%
<b>Miss</b>	63,54%	68,80%
<b>NMiss</b>	85,92%	88,33%



# Segmentacja – dwa modele

Rozkłady zmiennej dla modelu znanego klienta

Warunek	Liczba	Procent	Ryzyko
$ACT\_CINS\_N\_STATC \leq 0$	666	16,3%	28,5%
$0 < ACT\_CINS\_N\_STATC \leq 2$	2 616	63,8%	13,2%
$2 < ACT\_CINS\_N\_STATC \leq 3$	367	9,0%	9,3%
$3 < ACT\_CINS\_N\_STATC \leq 4$	222	5,4%	5,4%
$4 < ACT\_CINS\_N\_STATC$	227	5,5%	2,2%

Rozkłady zmiennej dla modelu PD Ins

Warunek	Liczba	Procent	Ryzyko
$ACT\_CINS\_N\_STATC \leq 0$	535	4,7%	29,0%
$0 < ACT\_CINS\_N\_STATC \leq 1$	1 528	13,4%	12,6%
Missing	8 105	71,2%	12,4%
$1 < ACT\_CINS\_N\_STATC \leq 2$	604	5,3%	11,4%
$2 < ACT\_CINS\_N\_STATC$	607	5,3%	6,1%



# Segmentacja – dwa modele

Rozkłady zmiennej dla modelu nieznanego klienta

Warunek	Liczba	Procent	Ryzyko
Contract	823	8,2%	43,1%
Owner company	1 236	12,3%	15,0%
Retired	4 276	42,5%	10,3%
Permanent	3 725	37,0%	8,8%

Rozkłady zmiennej dla modelu PD Ins

Warunek	Liczba	Procent	Ryzyko
Contract	768	6,7%	42,1%
Owner company	1 265	11,1%	15,3%
Retired	5 754	50,6%	10,5%
Permanent	3 592	31,6%	9,4%



# Poprawki zmiennych

- Ręczne dodatkowe poprawki, kiedy zmienne są już wybrane i trzeba je jeszcze lekko polepszyć.
- SAS:
  - %include "&dir\_codes.variable\_corrections.sas" / source2;
- Python:
  - #labsn['app\_number\_of\_children']=[-np.inf, 1, 1, 2, np.inf]



# Interakcje

- **%macro *dodatkowe\_zmienne*;**
- length app\_IGJM \$ 30;
- outstanding=app\_loan\_amount;
- credit\_limit=app\_loan\_amount;
- app\_IGJM = trim(app\_char\_gender)||'-  
'||trim(app\_char\_job\_code)||  
• '-'||trim(app\_char\_marital\_status);
- where '197501'<=period<='198712' and product='css'  
and decision='A';
- **%mend;**

# interakcje

Attributes for variable APP\_IGJM

Attribute number	Condition	Bad rate (br)	%POP	%GD	%BD	%IND
1	otherwise	60,64%	10,12%	6,74%	14,37%	7,64%
2	when ('Female-Retired-Divorced')	49,89%	12,95%	11,32%	15,13%	11,35%
3	when ('Female-Permanent-Divorced','Male-Permanent-Maried','Male-Retired-Maried')	46,33%	13,25%	12,53%	14,37%	12,01%
4	when ('Male-Retired-Divorced','Male-Retired-Widowed')	43,48%	10,18%	10,38%	10,37%	8,95%
5	when ('Female-Permanent-Maried')	37,02%	14,67%	15,36%	12,72%	18,56%
6	when ('Female-Retired-Maried','Female-Retired-Widowed')	36,32%	38,83%	43,67%	33,03%	41,48%



# Reject Inference

- Błędne estymacje ryzyka
- Błędne wnioskowanie o całej populacji na podstawie części zaakceptowanej
- Zewnętrzne bazy minimalizujące problem:
  - BIK
  - Black listy



# Bank Consumer Seniority

- Obserwowalna własność: im dłużej klient istnieje w banku, na rynku kredytów, tym jest mniej ryzykowny.
- Jest to wynik reguł odrzucających już zadłużonych klientów.
- Sprawdźmy zatem kategoryzację tej zmiennej na dwóch populacjach:
  - Całej (strategii rajskiej), czyli tylko teoretycznej, nie obserwowalnej
  - Akceptowanej w strategii 3

# Kategorie zmiennej

Atrybuty zmiennej ACT\_CCSS\_SENIORITY przy pełnej akceptacji.

Numer grupy	Warunek	Ryzyko	Procent	Liczba wniosków
1	$25 < \text{ACT\_CCSS\_SENIORITY} \leq 57$	71,50%	19,42%	2 684
2	$18 < \text{ACT\_CCSS\_SENIORITY} \leq 25$	68,74%	6,50%	899
3	$57 < \text{ACT\_CCSS\_SENIORITY} \leq 67$	61,40%	6,00%	829
4	$67 < \text{ACT\_CCSS\_SENIORITY} \leq 140$	59,66%	37,00%	5 114
5	$140 < \text{ACT\_CCSS\_SENIORITY}$	54,86%	17,55%	2 426
6	$\text{ACT\_CCSS\_SENIORITY} \leq 18$	49,47%	6,14%	849
7	missing(ACT_CCSS_SENIORITY)	34,90%	7,38%	1 020
		59,36%	100,00%	13 821

Atrybuty zmiennej ACT\_CCSS\_SENIORITY przy strategii 3.

Numer grupy	Warunek	Ryzyko	Procent	Liczba wniosków
1	$18 < \text{ACT\_CCSS\_SENIORITY} \leq 41$	59,73%	16,34%	1 125
2	$41 < \text{ACT\_CCSS\_SENIORITY} \leq 53$	47,97%	3,94%	271
3	$\text{ACT\_CCSS\_SENIORITY} \leq 18$	46,14%	11,30%	778
4	$53 < \text{ACT\_CCSS\_SENIORITY} \leq 142$	42,51%	37,42%	2 576
5	$142 < \text{ACT\_CCSS\_SENIORITY} \leq 184$	34,53%	12,12%	834
6	missing(ACT_CCSS_SENIORITY)	31,65%	15,24%	1 049
7	$184 < \text{ACT\_CCSS\_SENIORITY}$	25,10%	3,65%	251
		42,69%	100,00%	6 884





- W przypadku pełnej akceptacji (strategia rajska) klient z długą historią jest bardziej ryzykowny niż z krótką (jak się wiele razy rzuca kostką, to wreszcie wypadnie 6)
- Jeśli stosowaliśmy strategię 3, to nowo zbudowane modele nie mogą być wdrożone bez reguł ze strategii 3, w przeciwnym wypadku będziemy źle estymować ryzyko



# Wnioski

- Prawdziwe ryzyko grupy missing(ACT CCSS SENIORITY) wynosi 34,90%
- Model estymowany przy strategii 3 pokazuje 31,65%, jest to poprawne przy obu regułach razem:
  - missing(ACT CCSS SENIORITY) i *agr12\_Max\_CMaxA\_Due* > 3
- Trzeb być bardzo czujnym przy budowie modeli skoringowych na listę reguł akceptacji, które były stosowane w populacji modelowej



# Reject Inference

- Model KGB – known good bad
- Analizy i przygotowanie zbioru dla All
- Model All
- Kalibracja i walidacja
- Katalog:  
...\CS-AUT\materials\_all\reject\_inference\_modeling\

# Reject Inference

Target / Segments / Gini		New score	Old score
default12	Accepted	36,15%	41,29%
	All	24,73%	65,55%
	Rejected	14,09%	48,29%
default12_ind	Accepted	37,34%	42,77%
	All	26,12%	67,60%
	Rejected	15,17%	50,70%



# Reject Inference

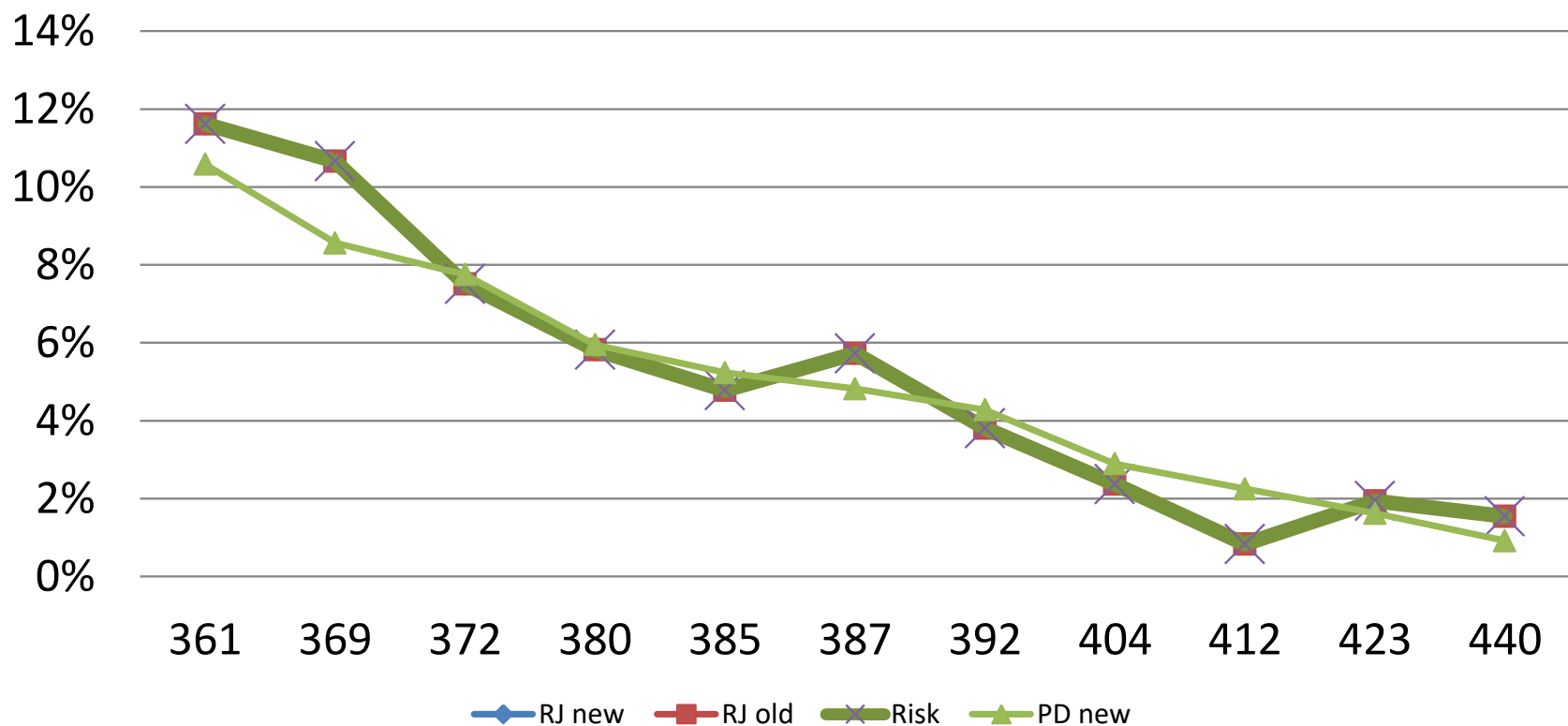
- Nowe PD, RJ New, tworzone jest z PD kalibrowanego na zaakceptowanych i przyłożone do odrzuconych. Na zaakceptowanych są obserwowane wartości default (czyli 0, 1).

# Reject Inference

Group - Condition		Pct			Risk			RJ new			RJ old			PD Ins		
		A	D	All	A	D	All	A	D	All	A	D	All	A	D	All
1	missing(ACT_CINS_N_STATC)	70,72%	76,68%	72,92%	5,61%	27,69%	14,17%	5,61%	5,71%	5,65%	5,61%	33,35%	16,37%	5,61%	22,53%	12,17%
2	not missing(ACT_CINS_N_STATC) and ACT_CINS_N_STATC <= 1	16,20%	16,57%	16,34%	3,52%	36,87%	16,01%	3,52%	6,08%	4,48%	3,52%	42,03%	17,94%	3,52%	31,03%	13,82%
3	1 < ACT_CINS_N_STATC	13,08%	6,74%	10,74%	2,20%	30,32%	8,71%	2,20%	2,32%	2,22%	2,20%	48,77%	12,98%	2,20%	35,39%	9,88%
All		100,00%	100,00%	100,00%	4,82%	29,39%	13,89%	4,82%	5,54%	5,09%	4,82%	35,83%	16,26%	4,82%	24,80%	12,20%

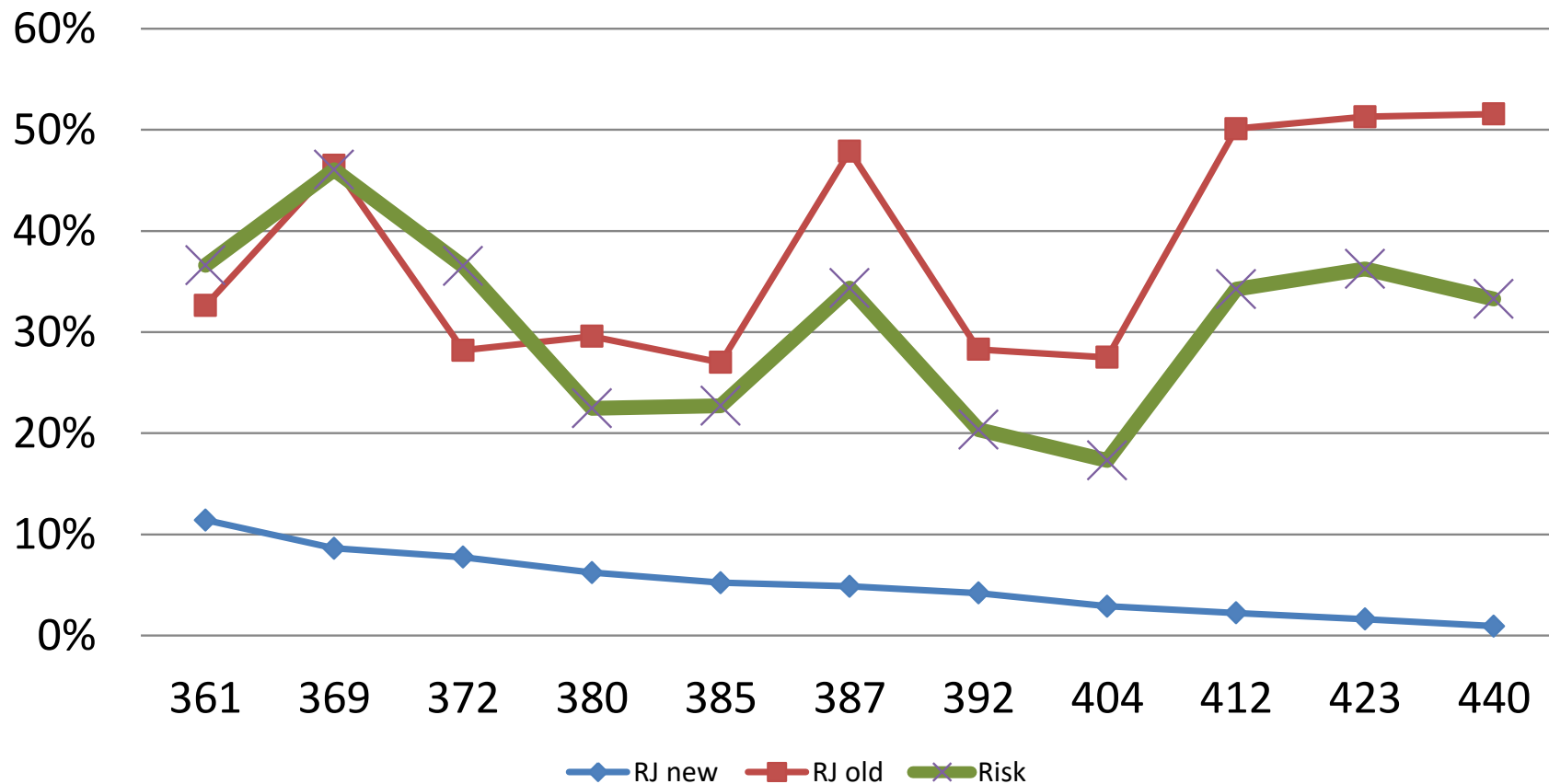
# Reject Inference

## Estymacje ryzyka względem new score dla zaakceptowanych



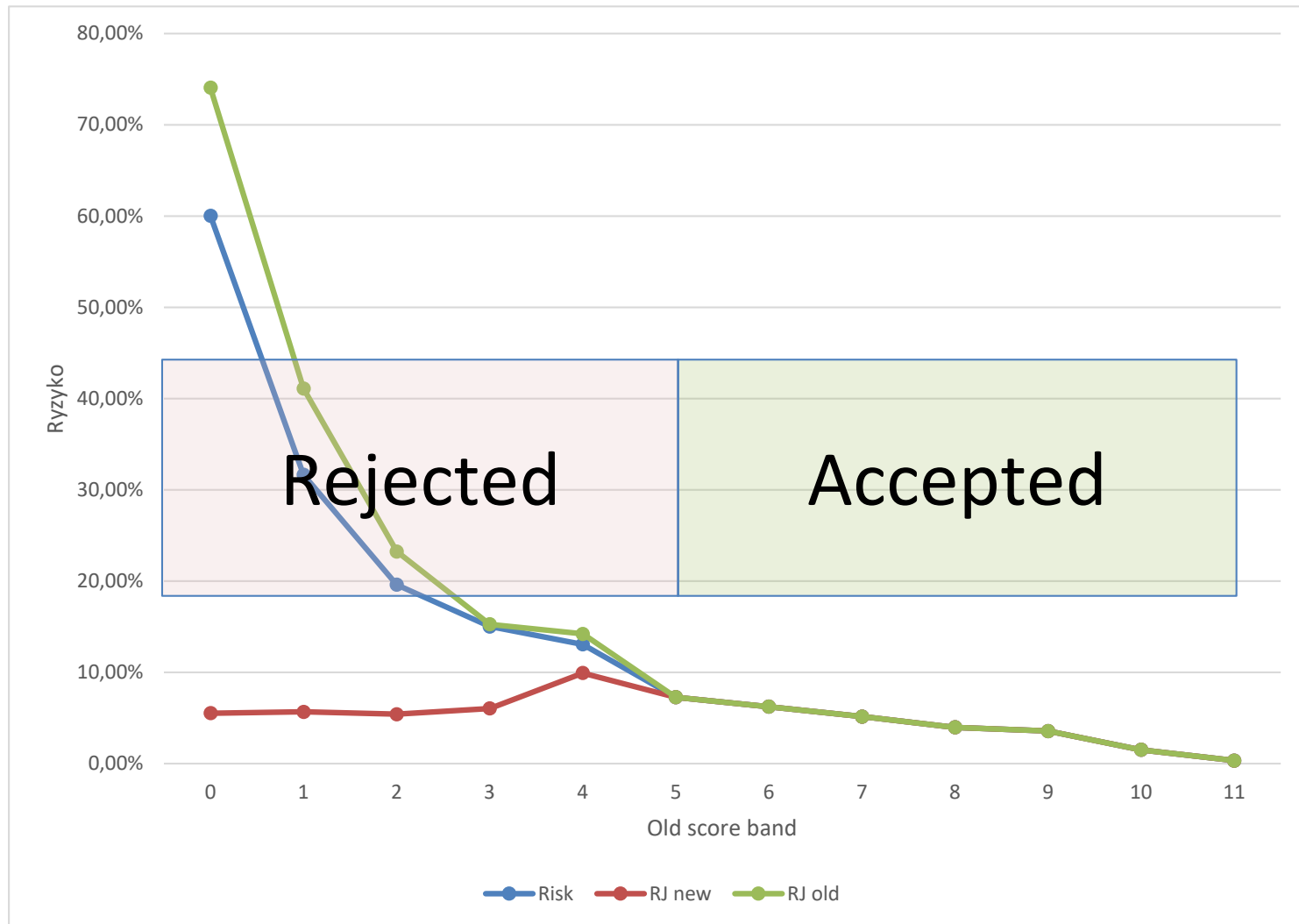
# Reject Inference

## Estymacje ryzyka względem new score dla odrzuconych



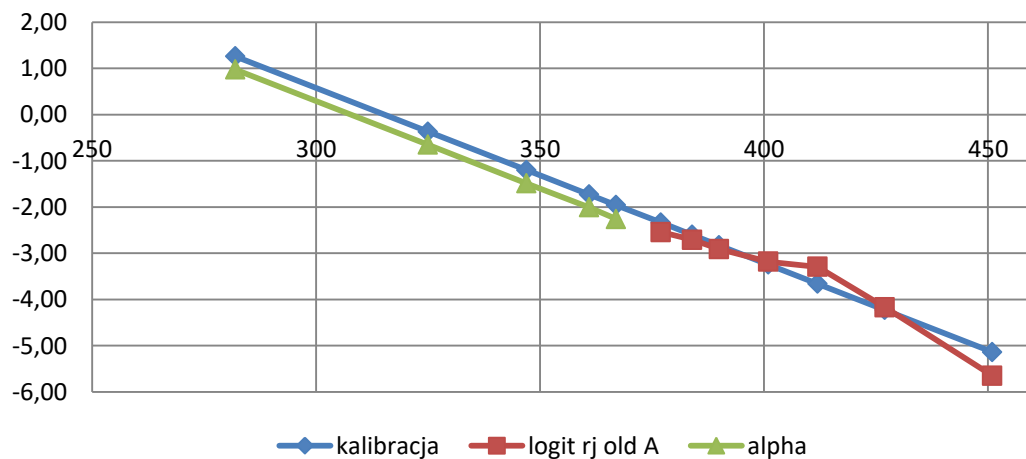


# Estymacja ryzyka odrzuconych

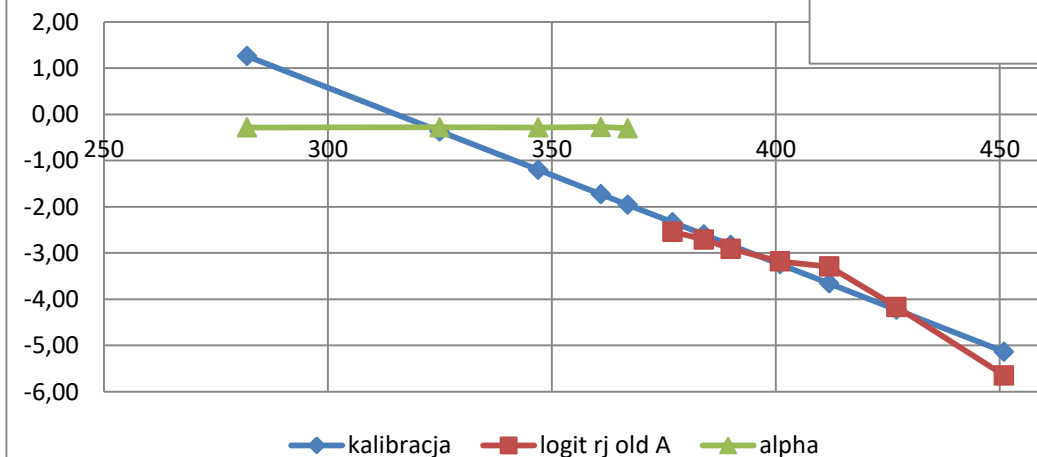


# Reject Inference

Estymacja ryzyka na odrzuconych



Estymacja ryzyka na odrzuconych





# Estymacja ryzyka, nowa funkcja celu

PD=5%

- Metoda dokładna (dwa wiersze z wagami):
  - wiersz1 default=1 weight=5%
  - wiersz2 default=0 weight=95%
- Metoda uproszczona (sto wierszy):
  - wiersze 1-5 default=1
  - wiersze 6-100 default=0



# Reject Inference

- Kolejne nowe PD jest już lepiej estymowane na części odrzuconej.
- Na takiej nowej populacji budujemy model All

Decision	Risk	Estimation
A	4,82%	4,82%
D	29,39%	35,36%
All	13,89%	16,09%

# Reject Inference

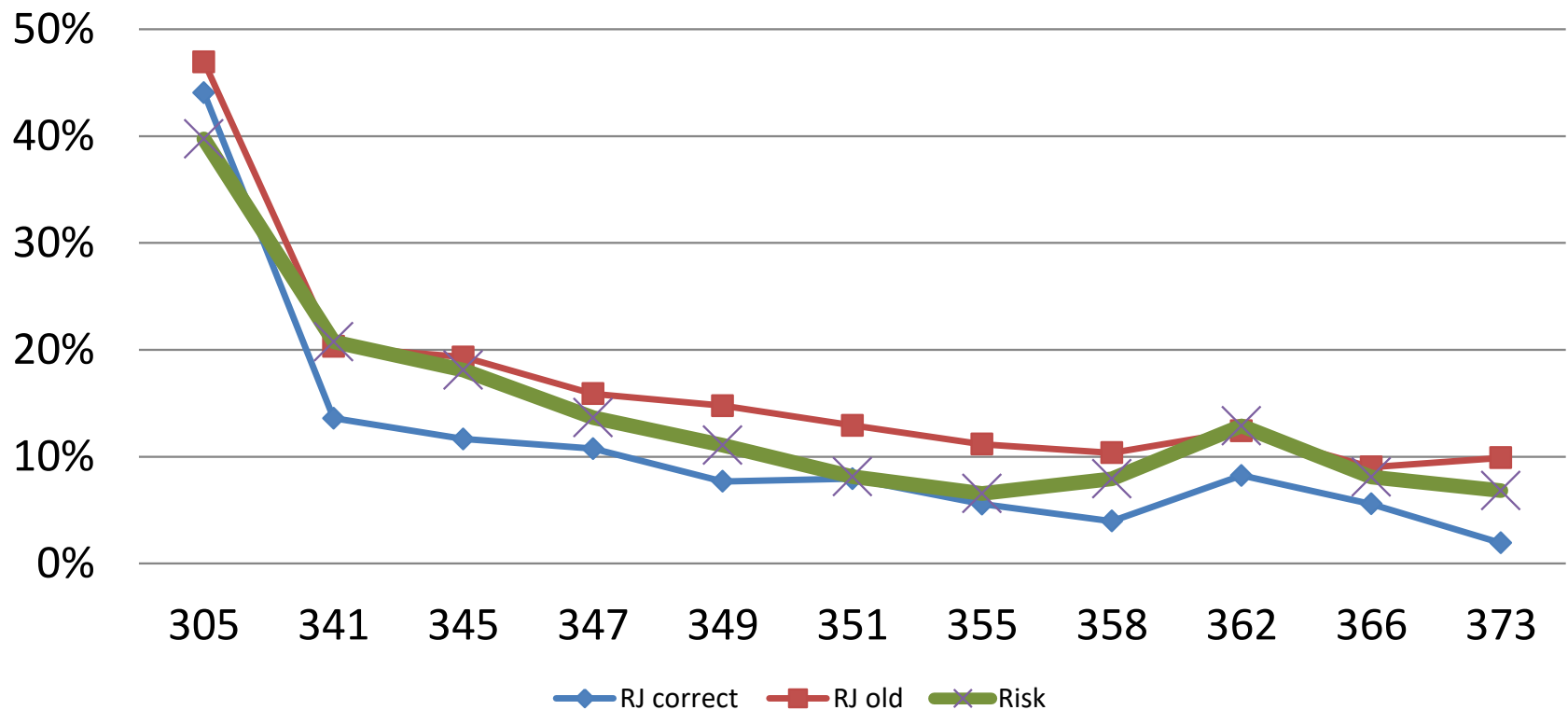
Target / Segments / Gini		New score	New score rj	Old score
default12	Accepted	36,15%	3,85%	41,29%
	All	24,73%	31,30%	65,55%
	Rejected	14,09%	17,14%	48,29%
default12_ind	Accepted	37,34%	4,17%	42,77%
	All	26,12%	32,23%	67,60%
	Rejected	15,17%	17,64%	50,70%

# Reject Inference

Group - Condition		Pct			Risk			RJ all			Correct RJ new		
		A	D	All	A	D	All	A	D	All	A	D	All
1	missing(ACT_CINS_N_STATC) or ACT_CINS_N_STATC <= 0	72,58%	87,09%	77,94%	5,72%	28,23%	15,00%	5,72%	22,93%	12,82%	5,72%	22,70%	12,72%
2	0 < ACT_CINS_N_STATC <= 2	20,78%	10,13%	16,85%	2,69%	39,63%	10,89%	2,69%	38,55%	10,65%	2,69%	14,93%	5,41%
3	2 < ACT_CINS_N_STATC	6,64%	2,77%	5,21%	1,70%	28,25%	6,91%	1,70%	33,31%	7,90%	1,70%	13,04%	3,92%
All		100,00%	100,00%	100,00%	4,82%	29,39%	13,89%	4,82%	24,80%	12,20%	4,82%	21,64%	11,03%

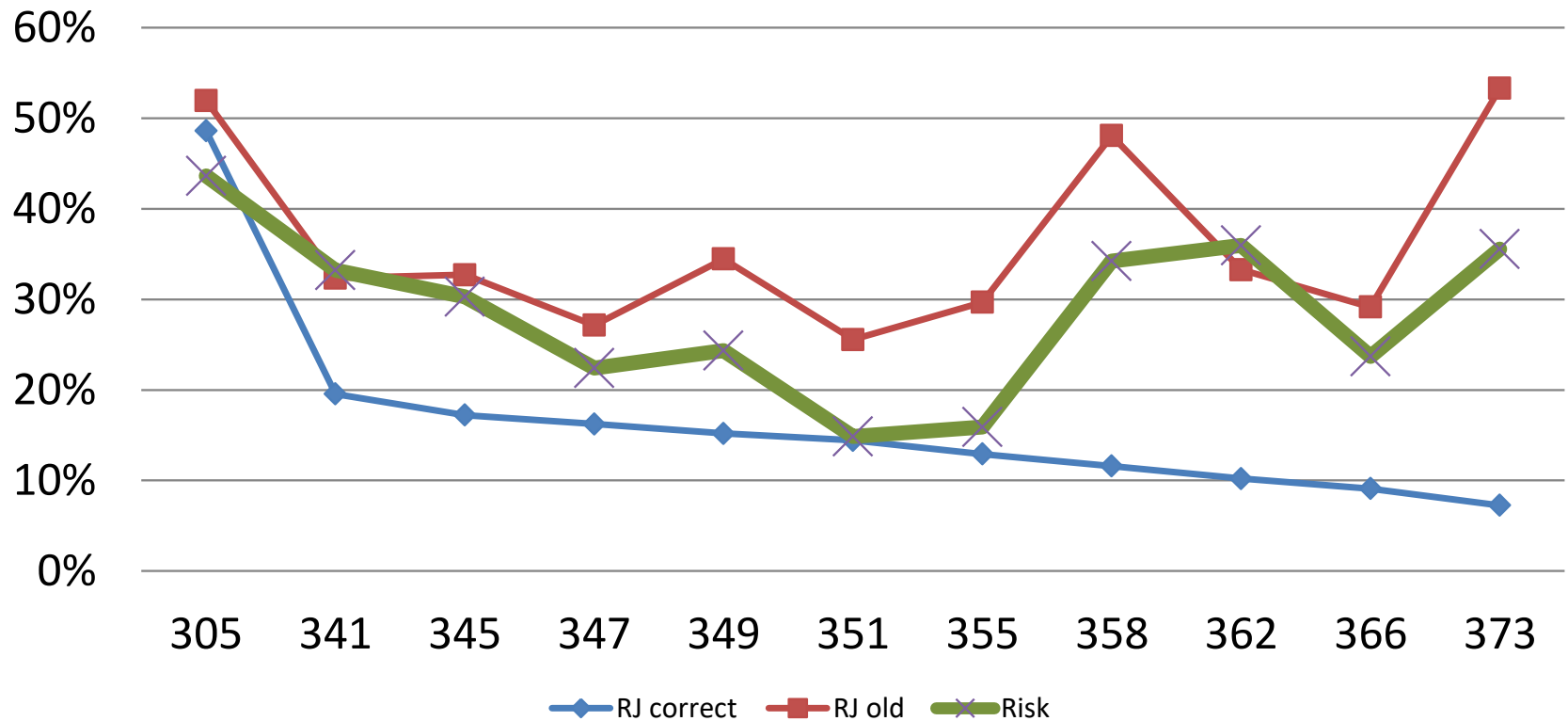
# Reject Inference

## Estymacje ryzyka względem new score skorygowanych dla populacji



# Reject Inference

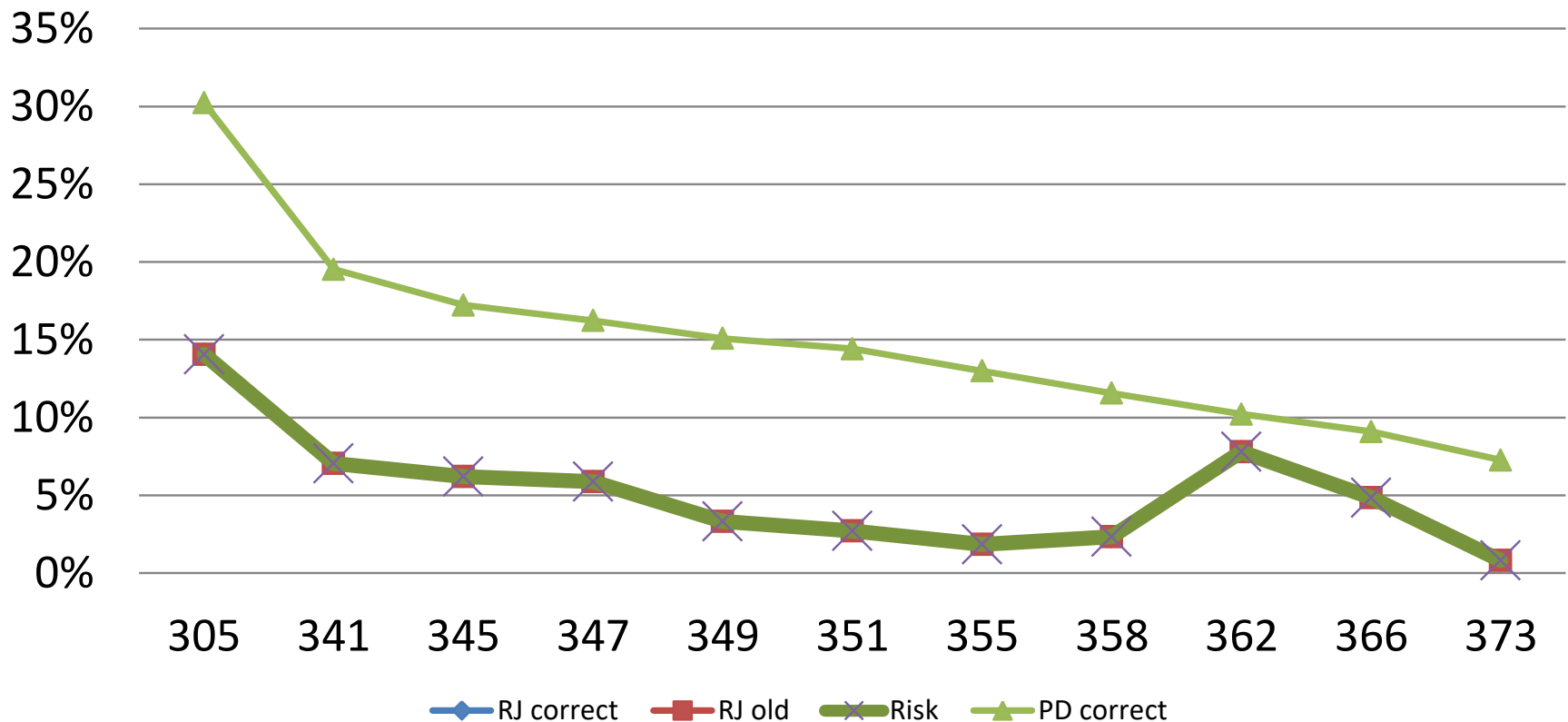
## Estymacje ryzyka względem new score skorygowanych dla odrzuconych





# Reject Inference

## Estymacje ryzyka względem new score skorygowanych dla zaakceptowanych





# Reject Inference – druga próba

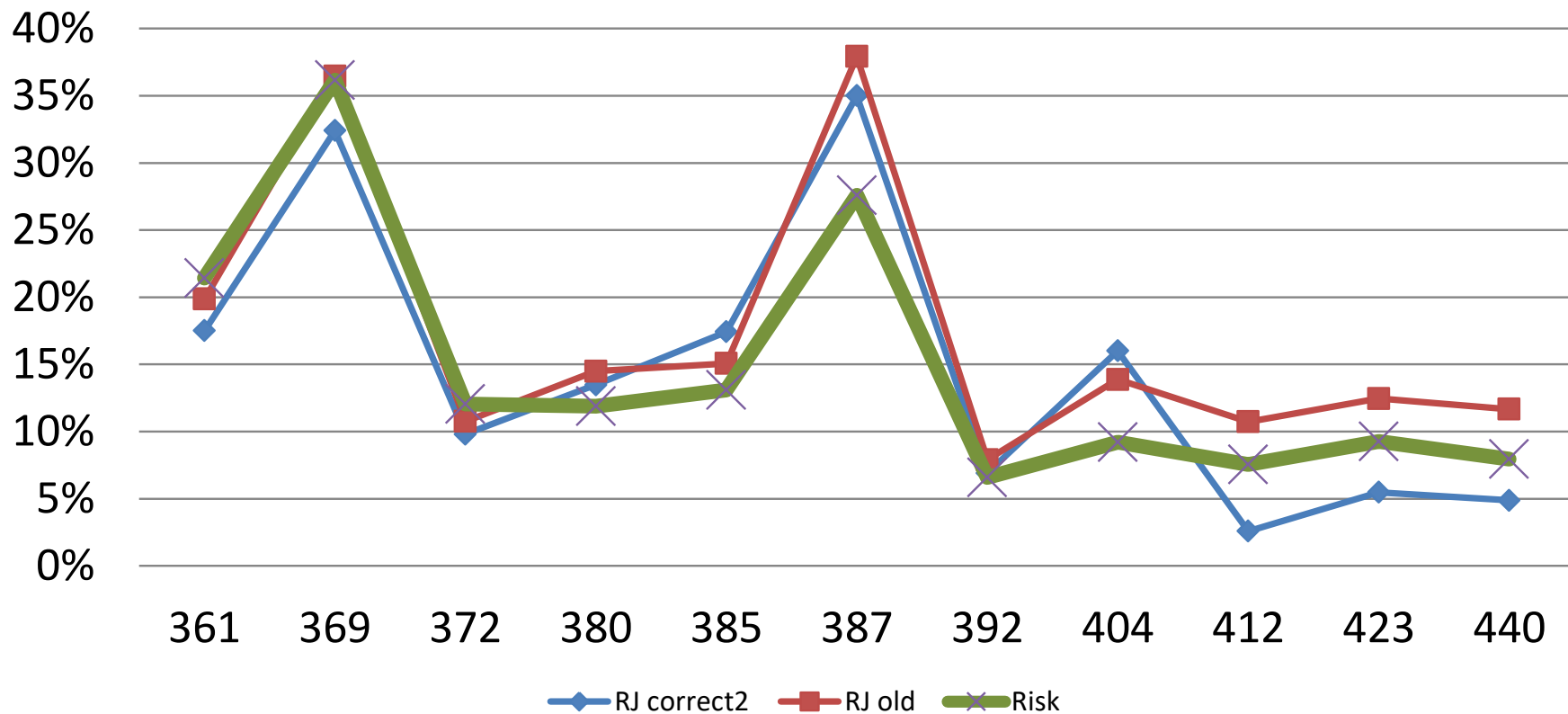
- Zbudowano kolejny model, tym razem wybierając ze wszystkich możliwych zmiennych.
- Model ma około 60% Giniego.
- Zmienne wybrane są inne niż modelu KGB
- Model ma znacznie lepsze własności

# Reject Inference – druga próba

Target / Segments / Gini		New score	New score rj2	Old score
default12	Accepted	36,15%	32,81%	41,29%
	All	24,73%	54,08%	65,55%
	Rejected	14,09%	24,93%	48,29%
default12_ind	Accepted	37,34%	34,11%	42,77%
	All	26,12%	56,13%	67,60%
	Rejected	15,17%	26,53%	50,70%

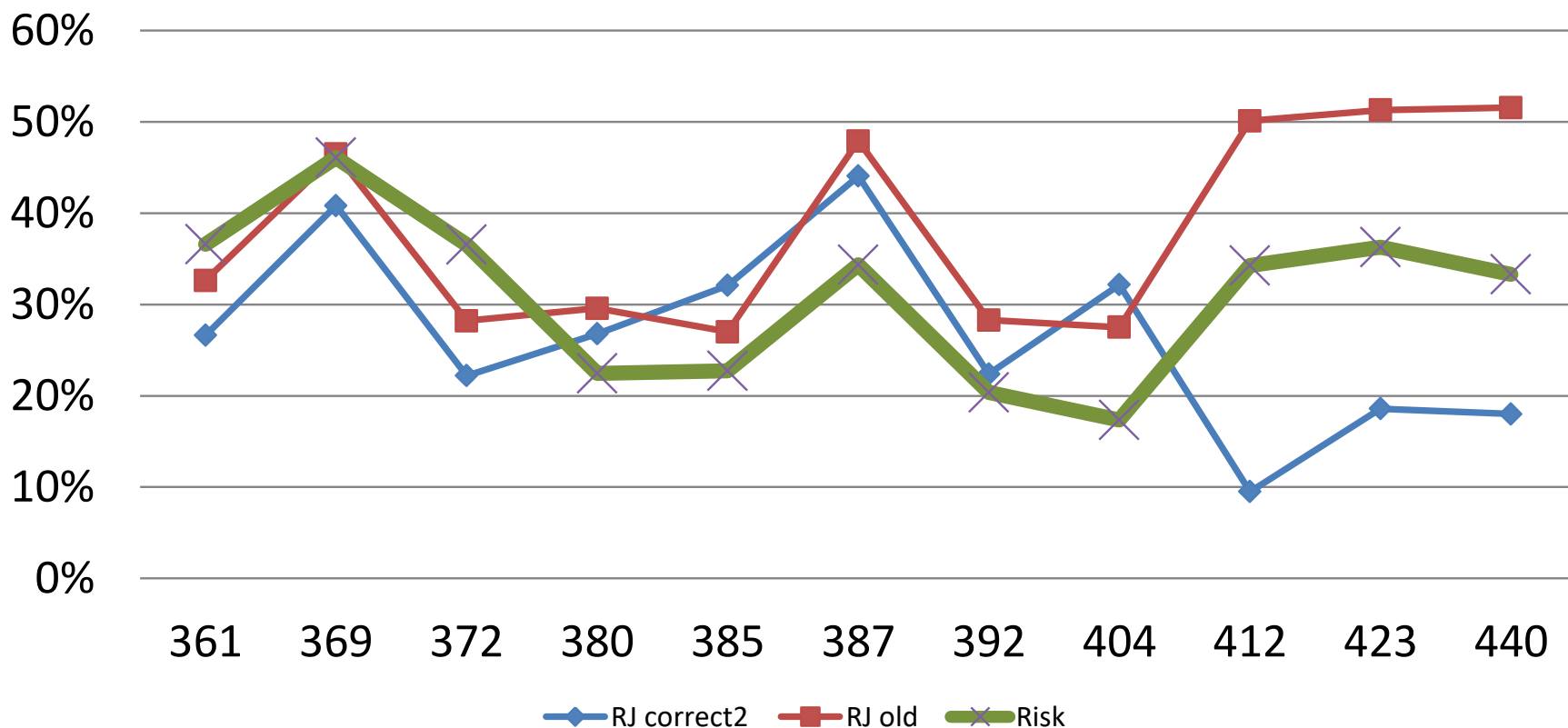
# Reject Inference – druga próba

## Estymacje ryzyka względem new score skorygowanych dla populacji, podejście drugie



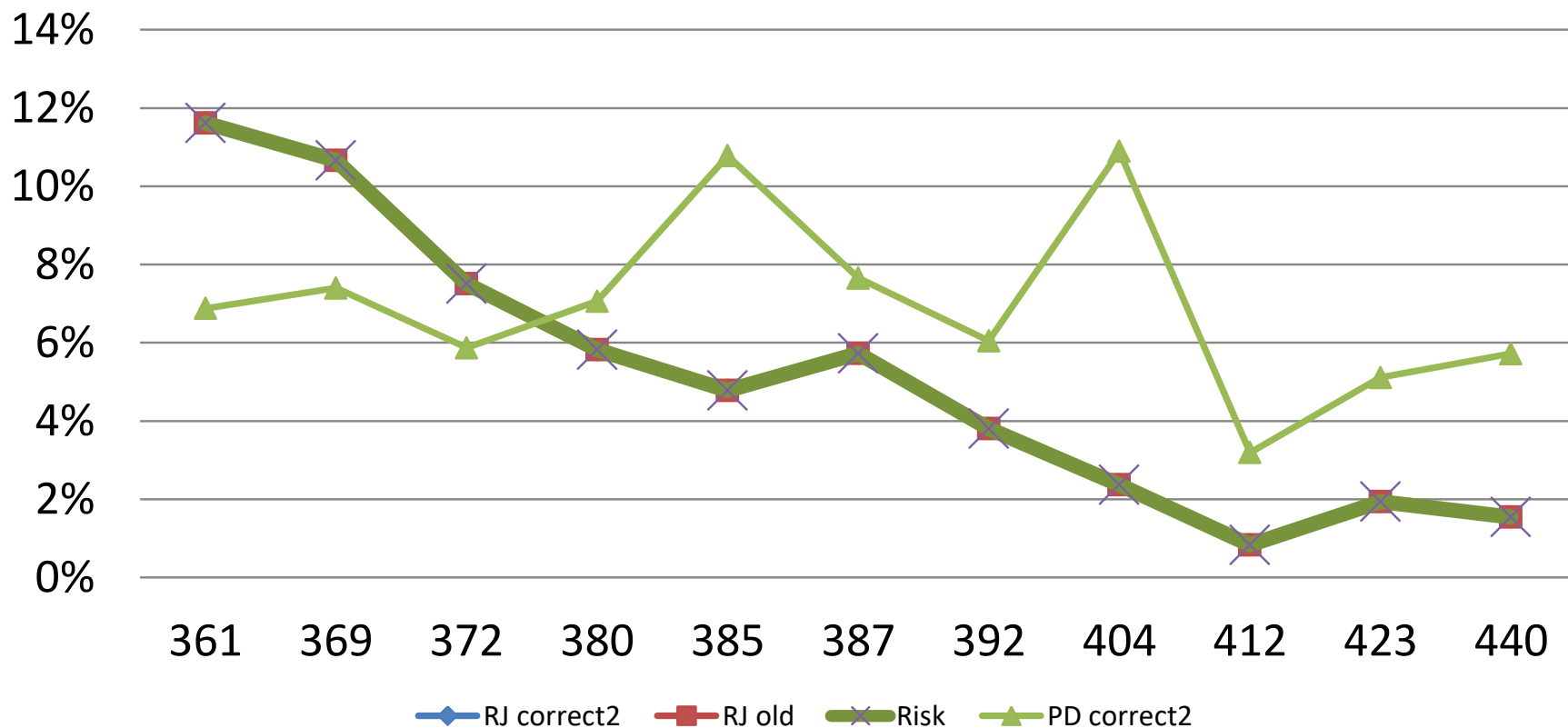
# Reject Inference – druga próba

## Estymacje ryzyka względem new score skorygowanych dla odrzuconych, podejście drugie



# Reject Inference – druga próba

## Estymacje ryzyka względem new score skorygowanych dla zaakceptowanych, podejście drugie

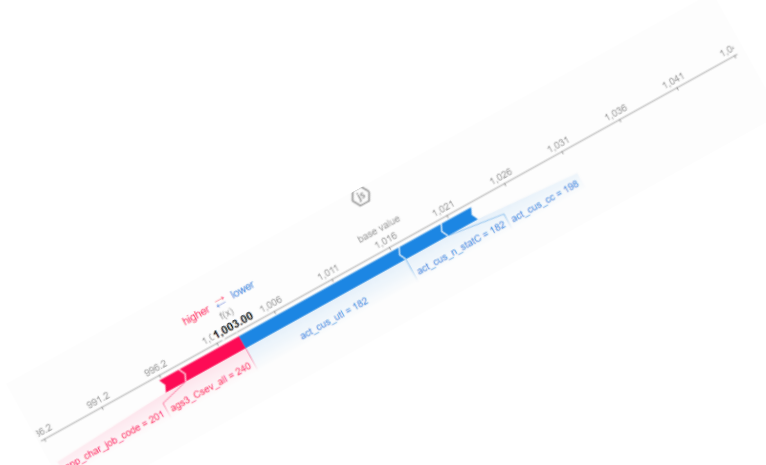
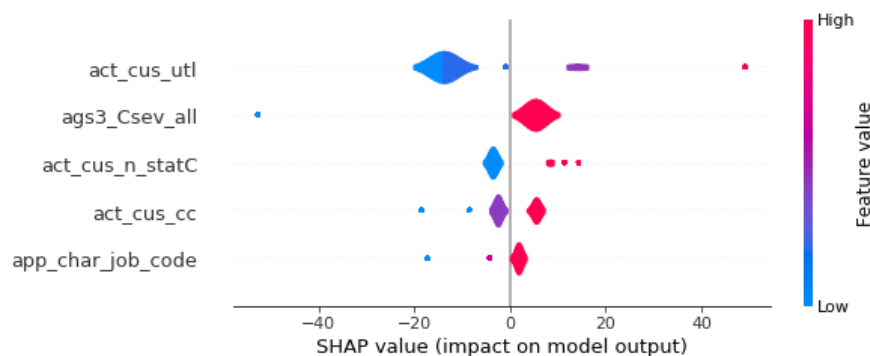
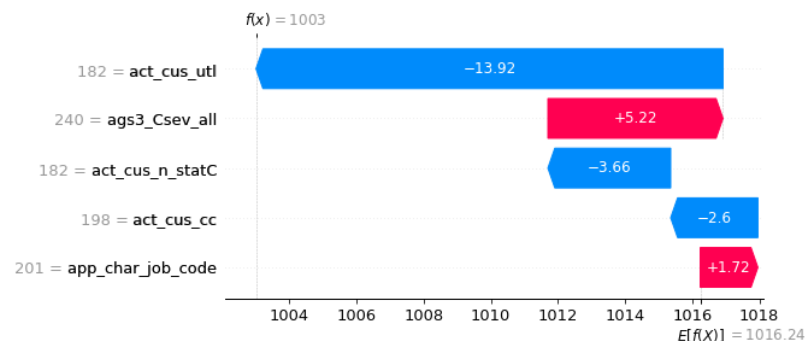
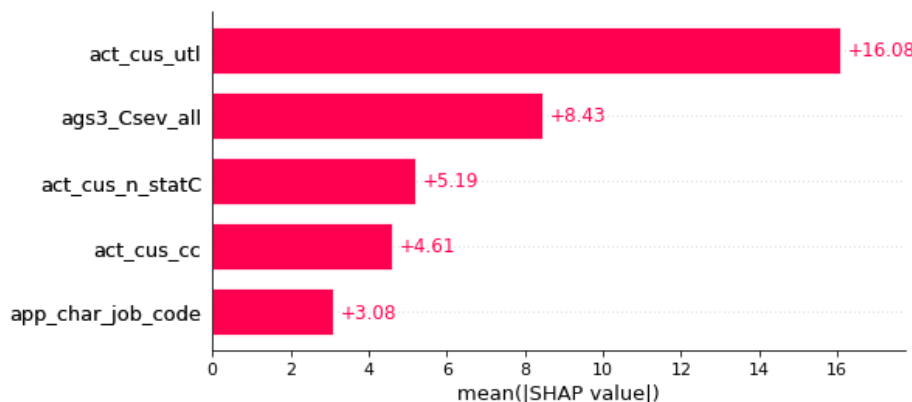




# Reject Inference

- Wnioski:
  - Jak nie ma wzorca, trudno estymować
  - Czasem pojawia się zjawisko odwrócenia profili na części odrzuconej
  - Modelowanie Reject Inference jest obarczone dużym błędem
  - Jedyne dobre rozwiązania zmniejszające RI to:
    - BIK, biura kredytowe
    - Open door, akceptowanie części odrzucanych

# Elementy XAI



<https://shap.readthedocs.io/en/latest/index.html>





# Wartość Shapleya

Niech dana będzie gra kooperacyjna  $G = \langle N, v \rangle$ , gdzie  $N$  to **zbiór** graczy  $N = \{1, 2, \dots, n\}$ , a  $v$  to **funkcja**, która przypisuje dowolnemu podzbiorowi (koalicji)  $S \subseteq N$  graczy **liczbę rzeczywistą**:  $v : 2^N \rightarrow \mathbb{R}$ , przy czym  $v(\emptyset) = 0$ . Funkcja  $v$  zwana jest również funkcją koalicyjną lub charakterystyczną.

Do wyliczenia tej wartości można wykorzystać następujący wzór:

$$\phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|! (n - |S| - 1)!}{n!} (v(S \cup \{i\}) - v(S)).$$

Wartość  $(v(S \cup \{i\}) - v(S))$  nazywa się też wkładem marginalnym gracza  $i$ .

Alternatywnie, równoważny jest również zapis:

$$\phi_i(v) = \frac{1}{|N|!} \sum_{\pi \in \Pi_N} [v(P_i^\pi \cup \{i\}) - v(P_i^\pi)],$$

gdzie:

$\pi \in \Pi_N$  – **permutacja** zbioru graczy,

$P_i^\pi$  – zbiór graczy z  $N$ , którzy występują w permutacji  $\pi$  przed graczem  $i$ .



# Referencje

Credit scoring in the context of interpretable machine learning. Theory and practice. Edited by D. Kaszyński, B. Kamiński, T. Szapiro. Pages 51-76, Oficyna Wydawnicza SGH, Warszawa 2020 ([https://sskolegia.sgh.waw.pl/pl/KAE/struktura/IE/struktura/ZWiAD/publikacje/Documents/Credit\\_scoring\\_in\\_the\\_context\\_of\\_interpretable\\_machine\\_learning.pdf](https://sskolegia.sgh.waw.pl/pl/KAE/struktura/IE/struktura/ZWiAD/publikacje/Documents/Credit_scoring_in_the_context_of_interpretable_machine_learning.pdf))

Wstęp do teorii gier, Tadeusz Płatkowski

<https://mst.mimuw.edu.pl/lecture.php?lecture=wtg&part=Ch12>



# Regresja Logistyczna

Na początku trzeba zdefiniować rozkład zero-jedynkowy. Rozważmy zdarzenie losowe polegające na zajściu zdarzenia default lub jego braku. Zmienna losowa  $Y$  przyjmuje zatem tylko dwie wartości  $Y = 1$  lub  $Y = 0$ , gdzie wartość 1 utożsamiamy z zajściem zdarzenia default. Zajście zdarzenia posiada określone prawdopodobieństwo, które oznaczamy przez  $p$ , mamy zatem:  $p = P(Y = 1)$ . Prawdopodobieństwo zdarzenia przeciwnego, czyli braku default można łatwo obliczyć:  $P(Y = 0) = 1 - P(Y = 1) = 1 - p$ . Przypuśćmy teraz, że zmienna losowa  $Y$  posiada swoją realizację  $y$ , innymi słowy została zaobserwowana jej wartość (wykonano pomiar). Obliczmy teraz prawdopodobieństwo zaobserwowania tej wartości. Możemy to zapisać w dwóch wariantach:

$$P(Y = y) = \begin{cases} p, & \text{gdy } y = 1, \\ 1 - p, & \text{gdy } y = 0, \end{cases}$$



# Regresja Logistyczna

albo w postaci jednego wzoru:

$$P(Y = y) = p^y(1 - p)^{(1-y)},$$

a po przekształceniach w finalnej wersji:

$$P(Y = y) = \exp \left( y \ln \left( \frac{p}{1 - p} \right) + \ln(1 - p) \right).$$

Pojawia się tu po raz pierwszy człon definiujący funkcję logitową:

$$\text{Logit}(p) = \ln \left( \frac{p}{1 - p} \right),$$



# Regresja Logistyczna

Rozważmy teraz sytuację bardziej ogólną. W naszej próbie losowej, zawierającej historyczne dane, zaobserwowaliśmy  $N$  obserwacji. Każda obserwacja funkcji losowej  $Y_n$  związana ze statusem zdarzenia default posiada wartość  $y_n$ , gdzie  $n$  jest numerem obserwacji. Interesującym nas modelem jest wyjaśnienie zależności pomiędzy prawdopodobieństwem zajścia zdarzenia default, co często matematycznie zapisuje się jako  $p_n = P(Y_n = 1)$ , a predyktorami oznaczanymi jako ciąg zmiennych  $x_n^1, x_n^2, \dots, x_n^m$ , gdzie  $m$  jest liczbą zmiennych w ABT. Na początku definiuje się część regresyjną, czyli kombinację predyktorów:

$$X_n\beta = \sum_{i=0}^m \beta_i x_n^i = \beta_0 + \beta_1 x_n^1 + \beta_2 x_n^2 + \dots + \beta_m x_n^m.$$



# Regresja Logistyczna

Kombinacja ta związana jest nieliniową zależnością z prawdopodobieństwem  $p_n$ . Funkcji wiążących można zdefiniować dość sporo, z praktyki zależność ta określona jest nazwą sigmoid i najpopularniejszą stała się funkcja logitowa. Swoją popularność zawdzięcza możliwością interpretacji członu  $\frac{p_n}{1-p_n}$ , który nazywa się szansą zajścia zdarzenia default (ang. odds), jest to stosunek prawdopodobieństwa zajścia zdarzenia do prawdopodobieństwa zdarzenia przeciwnego. Mamy zatem model, który uzależnia logarytm naturalny z szansy, albo logit z prawdopodobieństwa zajścia zdarzenia default od członu regresyjnego  $X_n\beta$ . Finalnie zatem estymowane jest następujące równanie:

$$\text{Logit}(p_n) = X_n\beta,$$

gdzie  $X_n$  są danymi wartościami predyktorów,  $p_n$  są teoretycznymi wartościami prawdopodobieństw zajścia zdarzenia default dla  $n$ -tej obserwacji a wektor współczynników  $\beta$  jest szukany.





# Regresja Logistyczna

Metoda największej wiarygodności, opisana przez R.A.Fishera w XX w., oparta jest na bardzo prostym i uzasadnionym przesłaniu, że prawdopodobieństwo uzyskania takich, a nie innych wartości obserwacji w próbie musi być największe. Gdyby było inaczej, to otrzymalibyśmy inne wartości obserwacji. Mamy zatem, wykorzystując założenie o niezależności zaobserwowanych zdarzeń:

$$P(Y_1 = y_1, Y_2 = y_2, \dots, Y_N = y_N) =$$

$$P(Y_1 = y_1)P(Y_2 = y_2) \cdot \dots \cdot P(Y_N = y_N) =$$

$$\prod_{n=1}^N P(Y_n = y_n) =$$

$$\prod_{n=1}^N \exp \left( y_n \ln \left( \frac{p_n}{1 - p_n} \right) + \ln(1 - p_n) \right) =$$



# Regresja Logistyczna

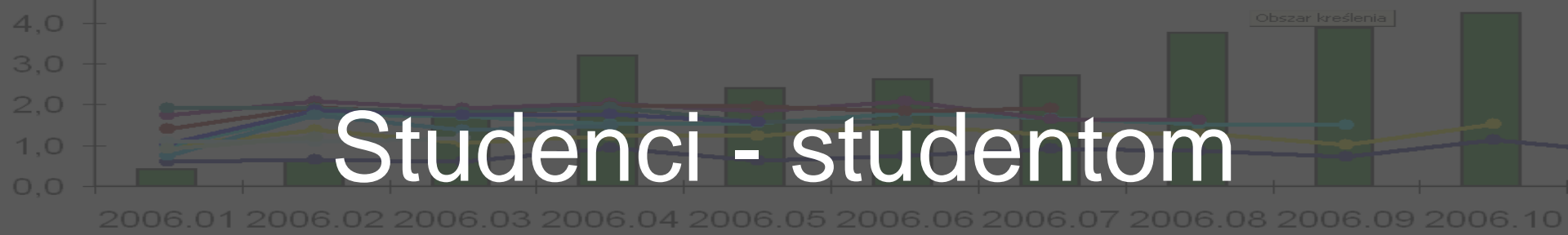
$$\exp \left( \sum_{n=1}^N \left( y_n \ln \left( \frac{p_n}{1 - p_n} \right) + \ln(1 - p_n) \right) \right).$$

Funkcją wiarygodności jest właśnie prawdopodobieństwo zaobserwowania wartości  $y_n$ . Przykładając zatem dodatkowo funkcję logarytmu i wstawiając za logity odpowiednie człony regresyjne otrzymamy finalną postać logarytmu z funkcji wiarygodności ( $L$  z ang. likelihood):

$$\ln(L(\beta)) = \sum_{n=1}^N (y_n X_n \beta - \ln(1 + \exp(X_n \beta))).$$

Istotą metody maksimum wiarygodności jest zatem znalezienie takiego wektora współczynników  $\beta$  by logarytm z funkcji wiarygodności był największy.





# Studenci - studentom

- Zapraszamy do:
  - ulepszania kodów
  - rozwijania narzędzi i metod automatyzacji procesu
  - ulepszania materiałów i aktualizowania wiedzy



# Prowadzenie prac badawczych

- Porównywanie technik skoringowych
- Kodowanie zmiennych i współliniowość
- Reject Inference, MKS i MIV
- Prognozowanie kryzysu, survival analysis
- Wpływ mocy predykcyjnej na zysk
- Badania stabilności modeli w czasie
- Co lepsze: pricing czy Gini?
- Badania monotoniczności zmiennych



# Zakład Metod Statystycznych i Analiz Biznesowych

- 2013:  
(Związane z uczczeniem w International Year of Statistics 2013 [www.statistics2013.org](http://www.statistics2013.org))  
Advanced Analytics and Data Science [www.analytics-conference.pl](http://www.analytics-conference.pl)  
Modelowanie dla biznesu – SKN Business Analytics  
[www.modelowaniedlabiznesu.pl](http://www.modelowaniedlabiznesu.pl)
- 2014:  
II Advanced Analytics and Data Science – 14.10  
[http://www.sas.com/pl\\_pl/events/2014/advanced-analytics-and-data-science/index.html](http://www.sas.com/pl_pl/events/2014/advanced-analytics-and-data-science/index.html)  
II Modelowanie dla biznesu – SKN Business Analytics – 18.11  
[http://www.sas.com/pl\\_pl/events/2014/modelowanie-dla-biznesu/index.html](http://www.sas.com/pl_pl/events/2014/modelowanie-dla-biznesu/index.html)
- 2015:  
III Advanced Analytics and Data Science – 20.10  
[http://www.sas.com/pl\\_pl/events/2015/advanced-analytics-and-data-science/conference.html](http://www.sas.com/pl_pl/events/2015/advanced-analytics-and-data-science/conference.html)
- 2016:  
IV Advanced Analytics and Data Science – 10.10  
[http://www.sas.com/pl\\_pl/events/2016/advanced-analytics-and-data-science-2016/index.html](http://www.sas.com/pl_pl/events/2016/advanced-analytics-and-data-science-2016/index.html)  
III Modelowanie dla biznesu – SKN Business Analytics – 10.03  
[http://www.sas.com/pl\\_pl/events/2016/konferencja-modelowanie-dla-biznesu/index.html](http://www.sas.com/pl_pl/events/2016/konferencja-modelowanie-dla-biznesu/index.html)
- 2017:  
IV Modelowanie dla biznesu – SKN Business Analytics – 10.04