

Klasyfikatory

Zgromadzone dane

- Tabela
 - Wiersze : Kolejne obserwacje
 - Kolumny : zmienna prognozowana (objaśniana) (Y)
zmienne objaśniające (**X**)
- Typy danych
 - Nominalne (binarne)
 - Porządkowe
 - Ciągłe
 - Braki danych

Braki danych

- Braki danych są jednym z najważniejszych problemów (i najczęstszych) na jaki możemy natrafić budując model.
- Aby móc im poprawnie przeciwdziałać musimy odpowiednio zbadać i zrozumieć strukturę zebranych danych.
- Mechanizmy
 - Mechanizm całkowicie losowy (**MCAR**: *Missing completely at random*)
 - Mechanizm losowy (**MAR**: *Missing at random*)
 - Mechanizm nielosowy (**MNAR** *Missing not at random*)
- Uwaga
 - Wybór procedury radzenia sobie z brakami danych zależy od typu modelu na którym pracujemy (np. szeregi czasowe).

Braki danych

- **Jak sobie radzić?**
- Usuwanie danych
 - **Usuwanie wszystkich jednostek obserwacji z analiz**
 - **Usuwanie jednostek obserwacji z analiz parami**
 - **Usuwanie zmiennych**
- Zastępowanie braków danych
 - Za pomocą statystyk (średniej/mediany/dominanty)
 - Za pomocą modelu
 - **Imputacja nieparametryczna**
 - **Imputacja regresyjna**
 - **Metoda największej wiarygodności**
 - **Wielokrotne imputacje**
- Traktowanie braku danych jako dodatkowej informacji, którą niesie ze sobą model
 - Traktujemy braki danych jako zmienne binarne – szczególnie wygodne w przypadku zmiennych jakościowych

Zagadnienie klasyfikacji

- Binarna zmienna objaśniana
- Szukamy dowolnej funkcji zmiennych objaśniających $f(x)$ takiej, że:

$$f(\mathbf{X}_1) > f(\mathbf{X}_2) \Leftrightarrow \Pr(Y_1 = 1) > \Pr(Y_2 = 1)$$

- Klasyczne modele w których zmienne objaśniane są wyrażalne liczbowo i nie posiadają braków danych
 - Regresja liniowa
 - Regresja logistyczna

Liniowy model prawdopodobieństwa

- Standardowa postać funkcyjna:

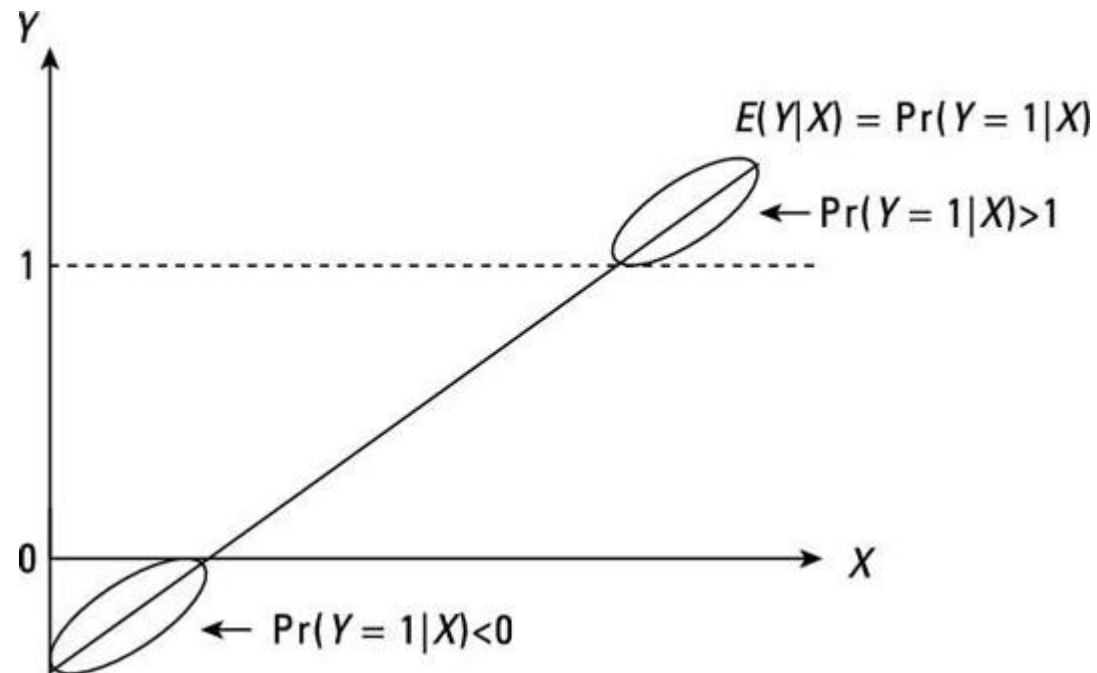
$$\begin{aligned} f(X) &= P(Y = 1|X) \\ &= B_0 + B_1x_1 + \dots + B_nx_n \end{aligned}$$

- Uwagi
 - Funkcja może być dowolna (niekoniecznie liniowa)
 - Nie będziemy zajmowali się własnościami statystycznymi
- Sposób wyznaczania parametrów na podstawie n-elementowego zbioru uczącego:

$$\mathbf{a} = \arg \min_{\mathbf{a}} \left\{ \sum_{i=1}^n (y_i - f(\mathbf{X}_i))^2 \right\}$$

Liniowy model prawdopodobieństwa

$$P(Y = 1|X) = B_0 + B_1x_1 + \cdots + B_nx_n$$



Regresja logistyczna

- Standardowa postać funkcyjna:

$$g(x) = B_0 + B_1x_1 + \dots + B_nx_n$$
$$f(x) = P(Y = 1|X) = \frac{\exp(g(X))}{1 + \exp(g(X))}$$

- Sposób wyznaczania parametrów na podstawie n-elementowego zbioru uczącego:

$$\mathbf{a} = \arg \max_{\mathbf{a}} \left\{ \sum_{i=1}^n y_i \ln(f(\mathbf{X})) + (1 - y_i) \ln(1 - f(\mathbf{X})) \right\}$$

Regresja logistyczna

- Własności:
 - Zawsze w przedziale (0,1)
 - Można **interpretować** jako prawdopodobieństwo

- Iloraz szans jest równy funkcji bazowej $g(X)$:

$$\ln\left(\frac{f(X)}{1 - f(X)}\right) = g(X)$$

- Uwaga
 - Zamiast dystrybuanty rozkładu logistycznego można użyć innej
 - Dla dystrybuanty standardowego rozkładu normalnego model nazywany jest probitowym

Regresja logistyczna

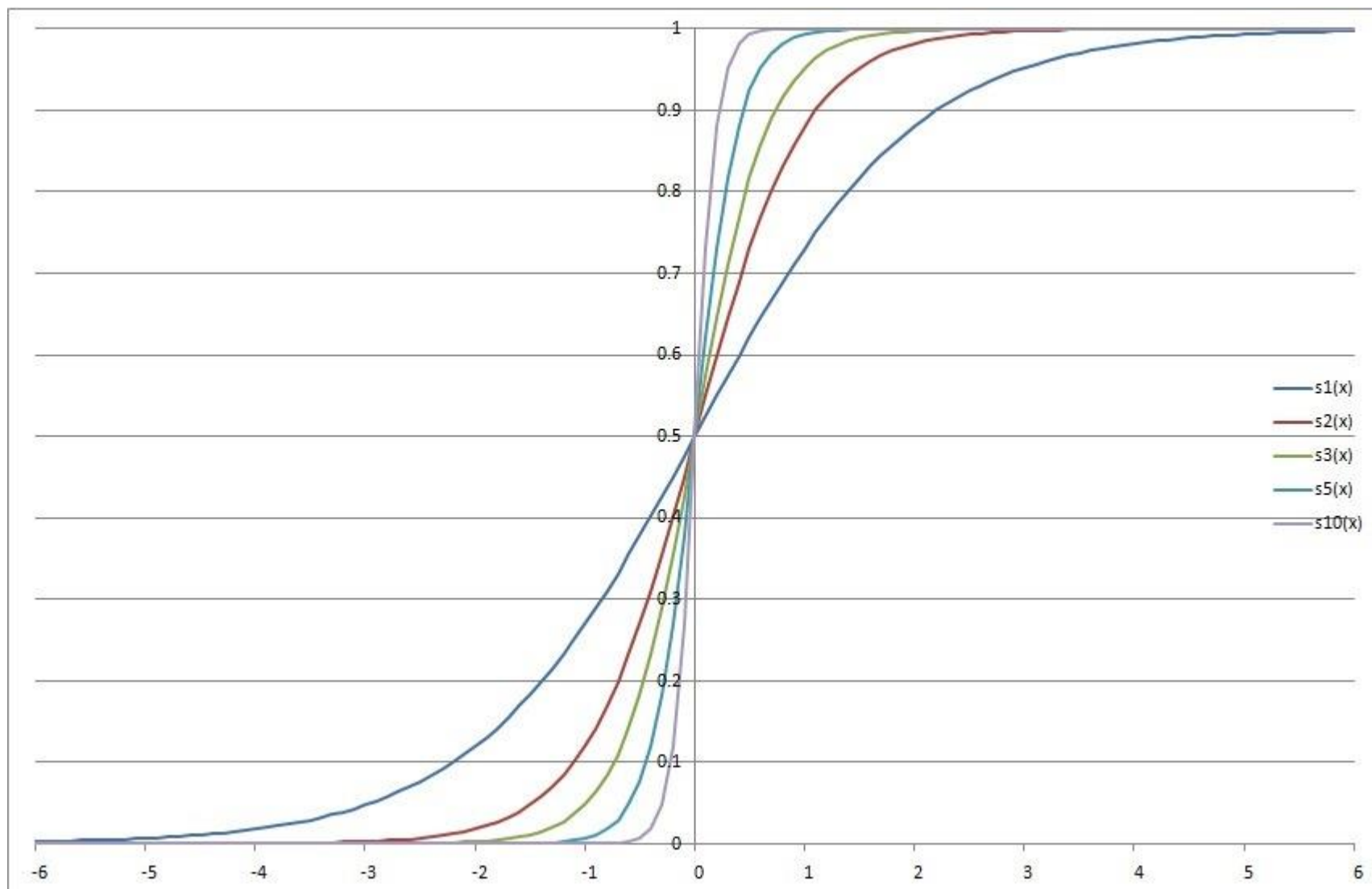
- **Probit:**

$$P(Y = 1|X) = \Phi(B_0 + B_1x_1 + \dots + B_nx_n)$$

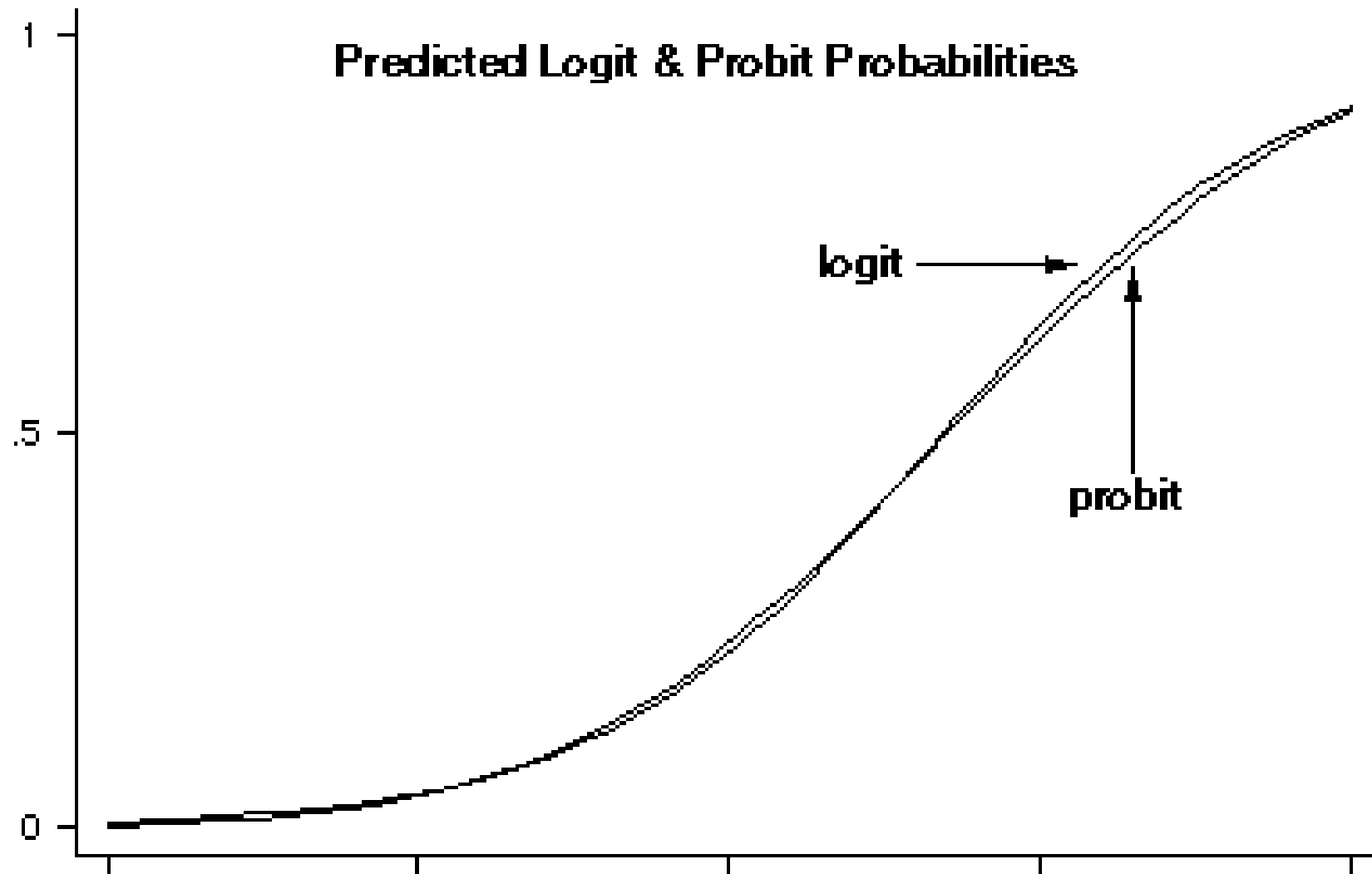
- **Logit:**

$$P(Y = 1|X) = \frac{e^{B_0+B_1x_1+\dots+B_nx_n}}{1 + e^{B_0+B_1x_1+\dots+B_nx_n}}$$

Funkcja sigmoidalna



Funkcja sigmoidalna



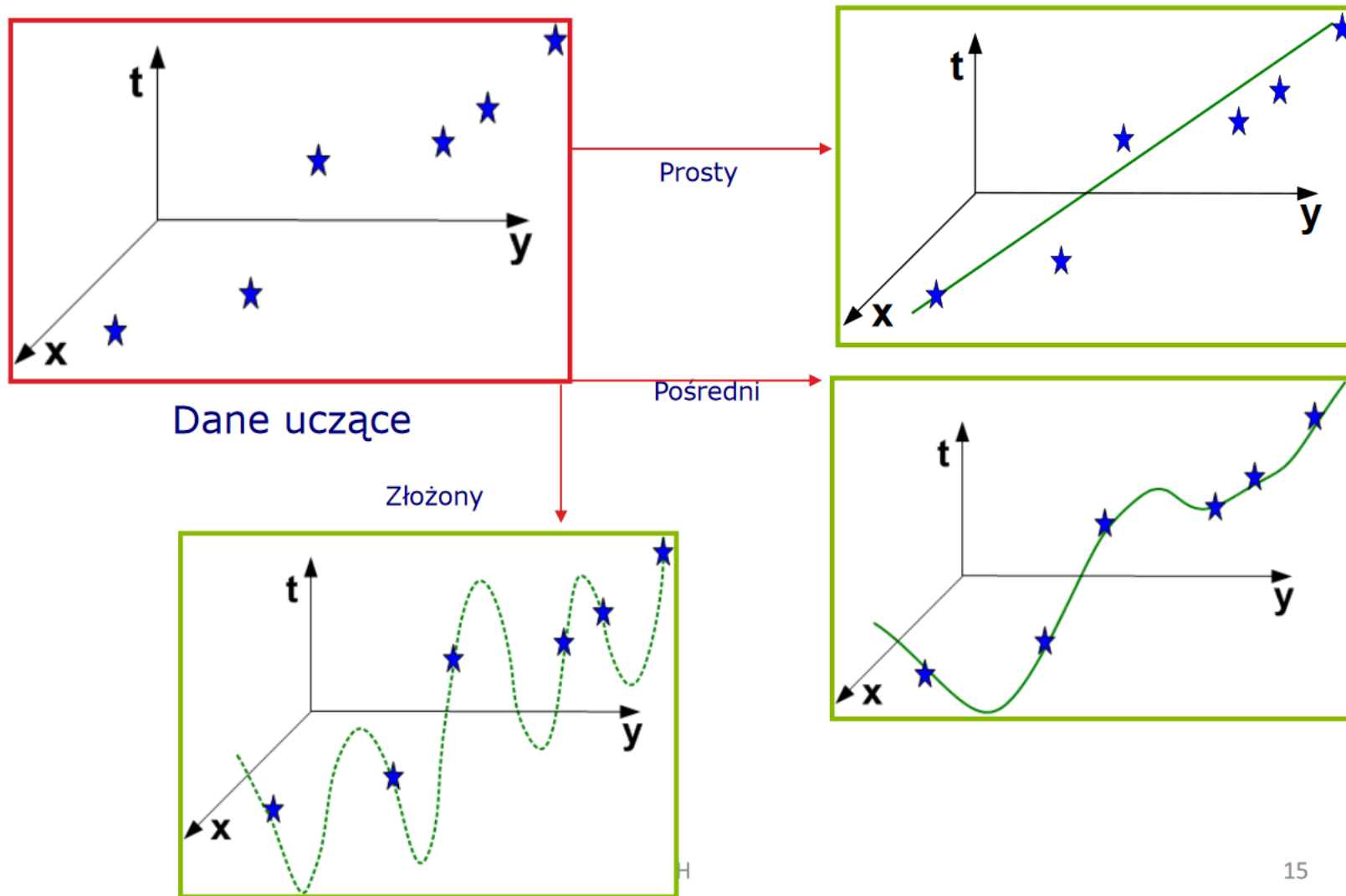
Ocena jakości modelu

- Określenie tego w jaki sposób mierzyć błąd modelu i jaka jest jego najwyższa akceptowana wartość jest (prawie) zawsze koniecznym pierwszym krokiem.
- Przede wszystkim dlatego, że nierozzerwalnie wiąże się z koniecznością zrozumienia jaki jest cel budowy danego modelu i przez to wpływa na to w jaki sposób ta budowa będzie przebiegała (jaki algorytm uczący zostanie wykorzystany, jak duży i jak skonstruowany będzie zbiór uczący, etc.).

Ocena jakości modelu

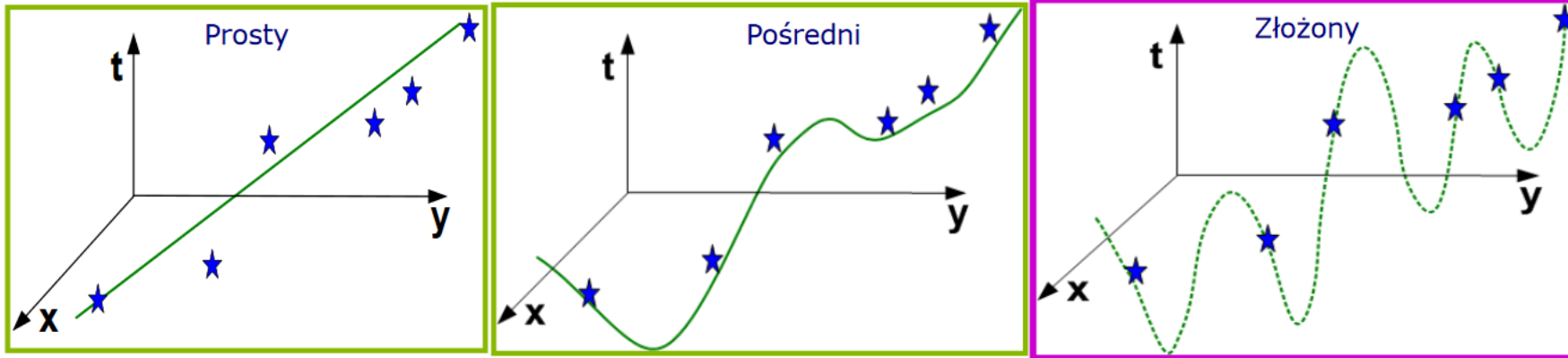
- Jest to szczególnie ważne gdy interesuje nas odpowiedź na pytanie:
Jaki model powinniśmy wybrać?

Ocena jakości modelu

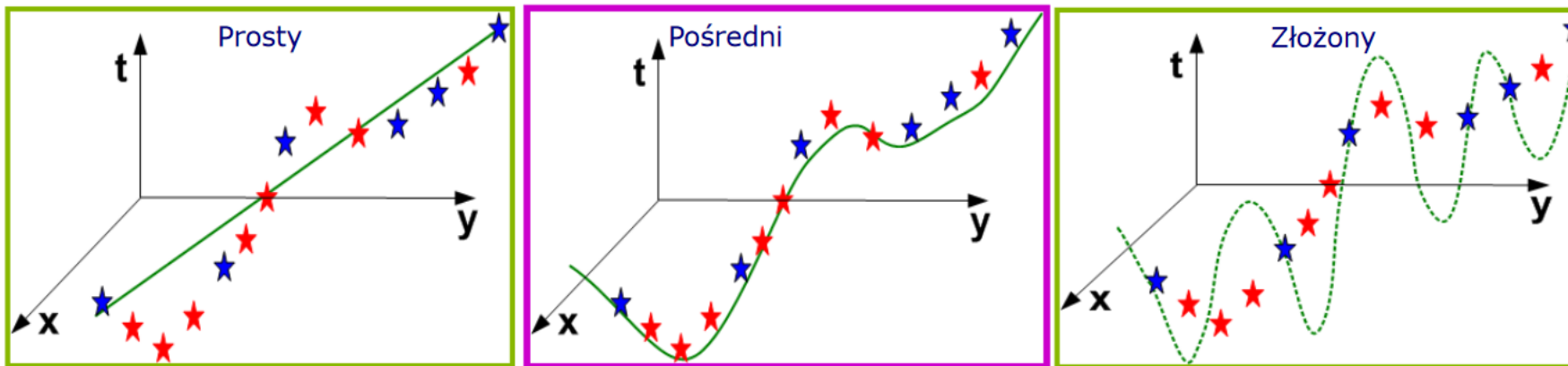


Ocena jakości modelu

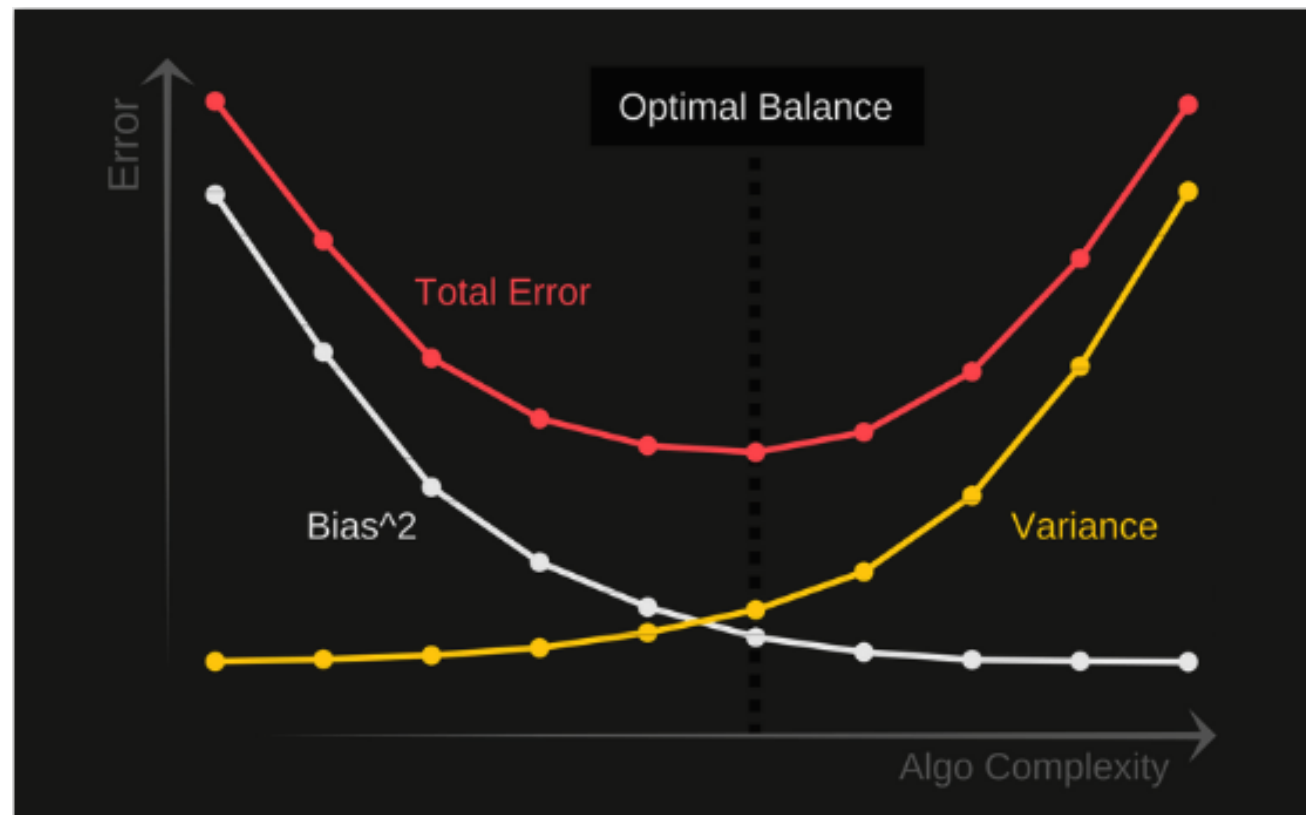
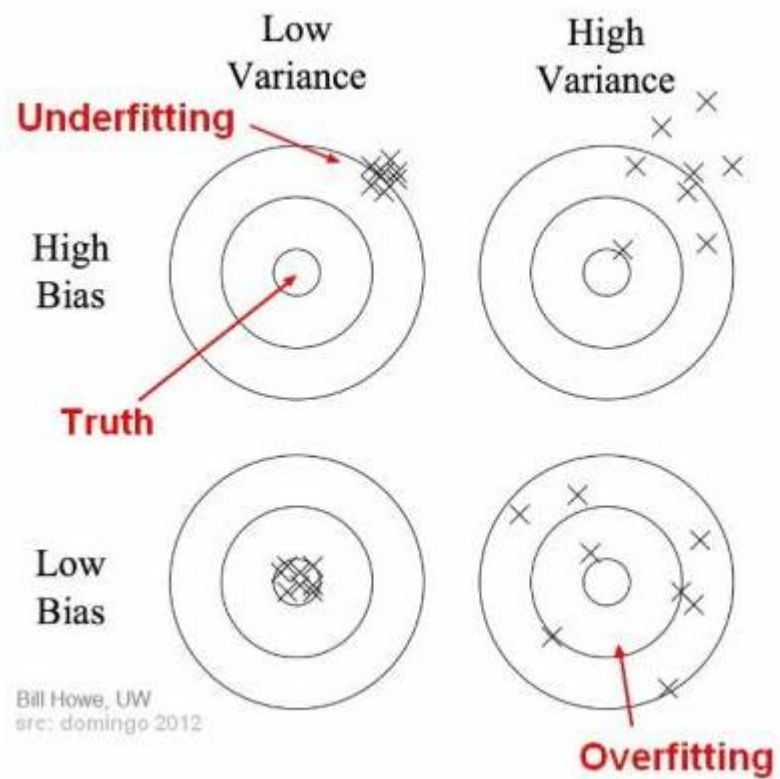
Minimalizacja błędu uczenia



Minimalizacja błędu prognozy



Bias-Variance Tradeoff



Źródło: <https://towardsdatascience.com/understanding-the-bias-variance-tradeoff-165e6942b229>

Ocena jakości modelu

- Abyśmy mogli odpowiednio ocenić, wyspecyfikować i wreszcie wybrać odpowiedni model musimy zastanowić się nad dwiema podstawowymi kwestiami:
 - Odpowiednią miarą jakości (**metryką**) modelu
 - Procedurą uczenia, oceny i wyboru modelu

Wybór odpowiedniej metryki

Przykładowe metryki:

$$MSE = \frac{1}{n} \sum_{i=1}^n (G(x) - f(x))^2$$

$$MISE = E||G - f||_2^2 = E\left[\int (G(X) - f(x))^2 dx\right]$$

Wybór odpowiedniej metryki

Przykładowe metryki:

- Macierz klasyfikacji:

		Klasa prawdziwa	
		1	0
Klasa prognozowana	1	True positive (TP)	False positive (FP)
	0	False negative (FN)	True negative (TN)

Wybór odpowiedniej metryki

Przykładowe metryki:

- Macierz klasyfikacji:

		Klasa prawdziwa	
		1	0
Klasa prognozowana	$[f(X)>T] = 1$	True positive (TP)	False positive (FP)
	$[f(X)>T] = 0$	False negative (FN)	True negative (TN)

- Gdzie T nazywamy progiem odcięcia

Wybór odpowiedniej metryki

Przykładowe metryki:

- Macierz klasyfikacji:

		Klasa prawdziwa	
		1	0
Klasa prognozowana	1	True positive (TP)	False positive (FP)
	0	False negative (FN)	True negative (TN)

$$\text{Trafność (Accuracy): } acc = \frac{TP+TN}{P+N}$$

Wybór odpowiedniej metryki

Przykładowe metryki:

- Macierz klasyfikacji:

		Klasa prawdziwa	
		1	0
Klasa prognozowana	1	True positive (TP)	False positive (FP)
	0	False negative (FN)	True negative (TN)

Czułość (Sensitivity – Recall – True Positive Rate): $TPR = \frac{TP}{TP+FN}$

Specyficzność (Specifity – True Negative Rate): $TNR = \frac{TN}{TN+FP}$

Precyzja (Precision – Positive Predictive Value): $PPV = \frac{TP}{TP+FP}$

Wybór odpowiedniej metryki

Przykładowe metryki:

- F - score:

$$F = \frac{2PPV * TPR}{PPV + TPR}$$

Czułość (Sensitivity – Recall – True Positive Rate): $TPR = \frac{TP}{TP+FN}$

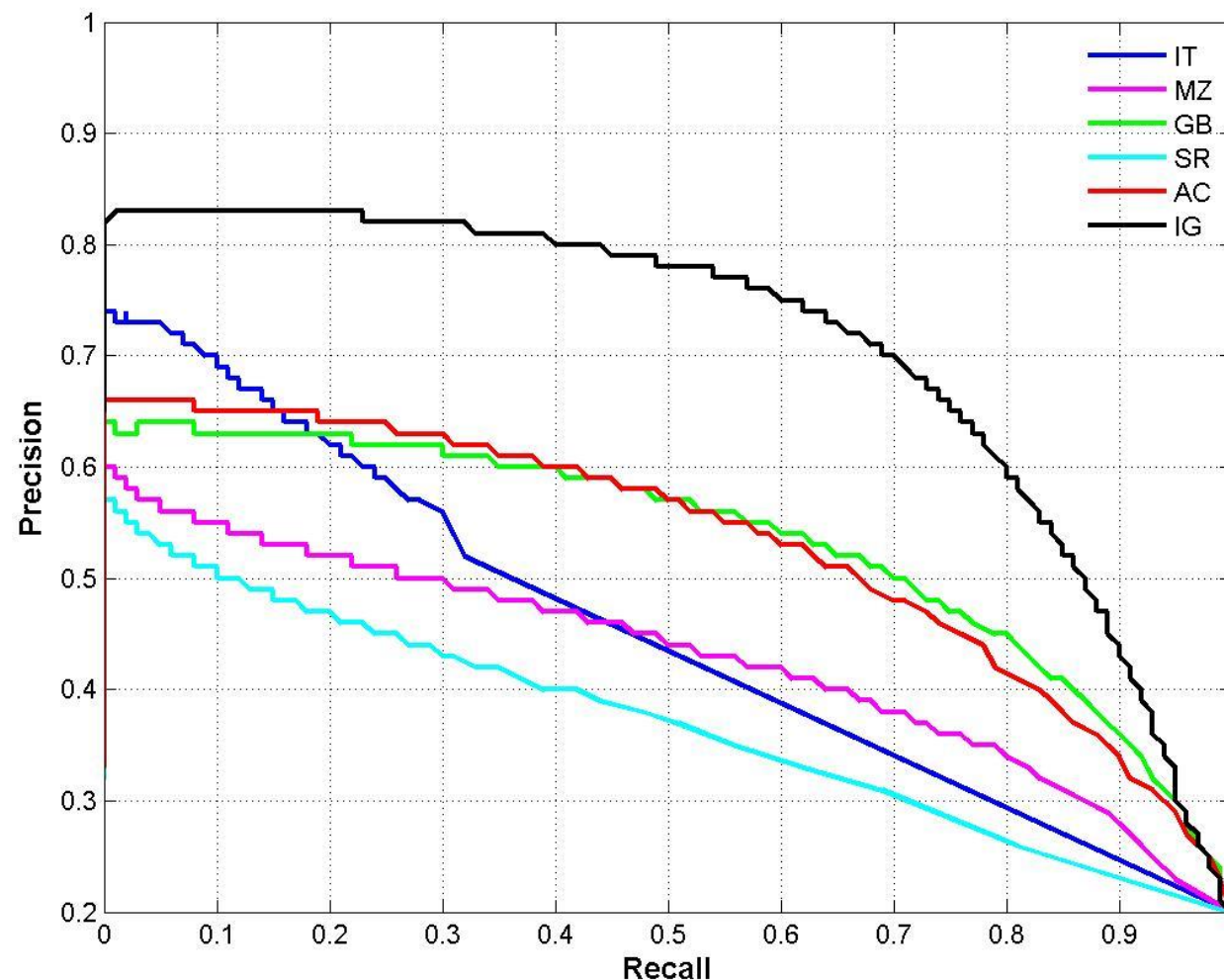
Specyficzność (Specifity – True Negative Rate): $TNR = \frac{TN}{TN+FP}$

Precyzja (Precision – Positive Predictive Value): $PPV = \frac{TP}{TP+FP}$

Wybór odpowiedniej metryki

Przykładowe metryki:

- Krzywa PR:



Czułość (Sensitivity – Recall – True Positive Rate): $TPR = \frac{TP}{TP+FN}$

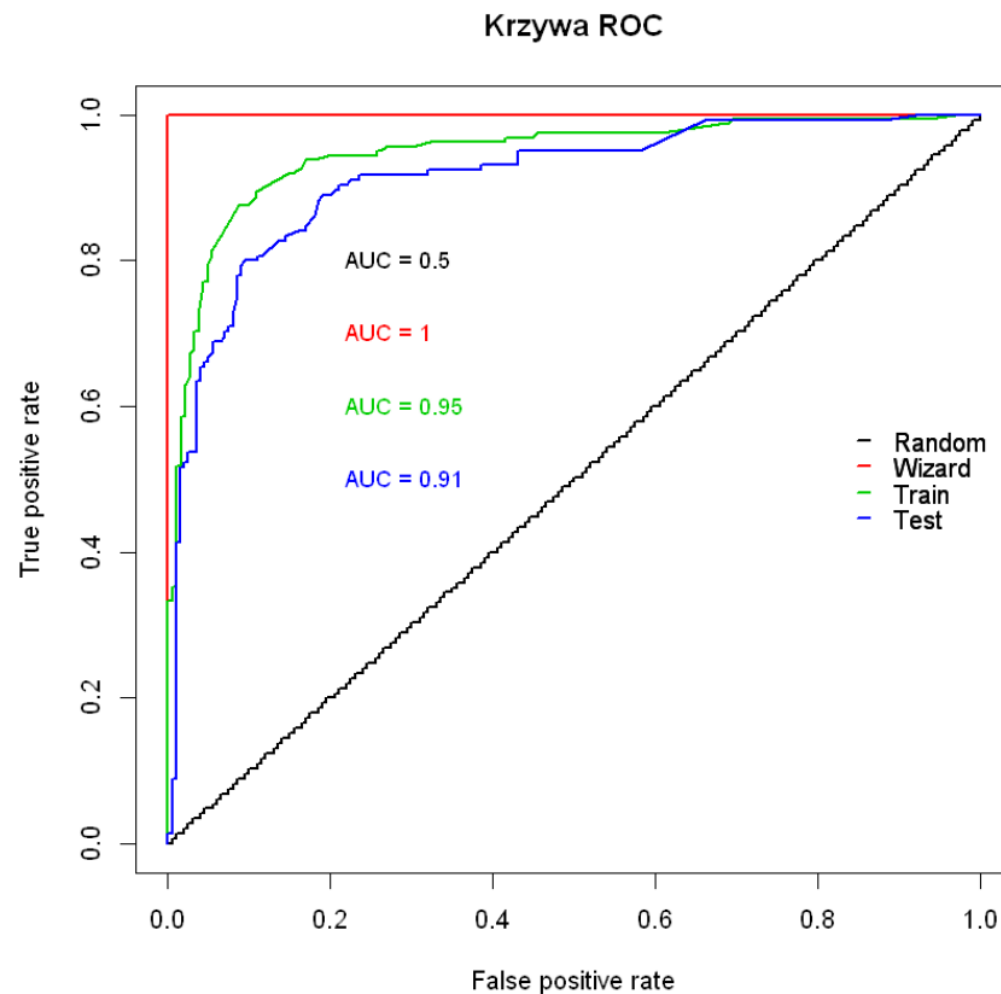
Specyficzność (Specifity – True Negative Rate): $TNR = \frac{TN}{TN+FP}$

Precyzja (Precision – Positive Predictive Value): $PPV = \frac{TP}{TP+FP}$

Wybór odpowiedniej metryki

Przykładowe metryki:

- Krzywa ROC:



Czułość (Sensitivity – Recall – True Positive Rate): $TPR = \frac{TP}{TP+FN}$

Specyficzność (Specificity – True Negative Rate): $TNR = \frac{TN}{TN+FP}$

Precyzja (Precision – Positive Predictive Value): $PPV = \frac{TP}{TP+FP}$

Wybór odpowiedniej metryki

- Wybór metryki zależy bezpośrednio od tego czemu służyć ma tworzony model.
- Inaczej będziemy traktowali błąd w przypadku testów wykrywających występowanie ciężkiej choroby (np. raka) a inaczej w przypadku modelu klasyfikującego zdjęcia psów i kotów.
- Dodatkowo pomiar błędu może zależeć od wielu innych czynników (np. generowanego finansowego zysku/straty).

Wybór odpowiedniej metryki

■◀ 'Let's try that again...' iPhone X facial recognition fails at launch - video

<https://www.theguardian.com/technology/video/2017/sep/12/apple-iphone-x-facial-recognition-face-id-fail-launch-video>

- Ale:

Wybór odpowiedniej metryki

◀ 'Let's try that again...' iPhone X facial recognition fails at launch - video

<https://www.theguardian.com/technology/video/2017/sep/12/apple-iphone-x-facial-recognition-face-id-fail-launch-video>

- Ale:

iPhone X racism row: Apple's Face ID fails to distinguish between Chinese users

<https://www.mirror.co.uk/tech/apple-accused-racism-after-face-11735152>

Optymalizacja progu odcięcia

Przykładowe metryki:

- Macierz klasyfikacji:

		Klasa prawdziwa	
		1	0
Klasa prognozowana	$[f(X)>T] = 1$	True positive (TP)	False positive (FP)
	$[f(X)>T] = 0$	False negative (FN)	True negative (TN)

- Gdzie T nazywamy progiem odcięcia

Optymalizacja progu odcięcia

- Macierz klasyfikacji:

		Klasa prawdziwa	
		1	0
Klasa prognozowana	$[f(X)>T] = 1$	True positive (TP)	False positive (FP)
	$[f(X)>T] = 0$	False negative (FN)	True negative (TN)

- Własność

$$\frac{\partial E(TP)}{\partial T} = - \frac{\partial E(FN)}{\partial T}$$

$$\frac{\partial E(FP)}{\partial T} = - \frac{\partial E(TN)}{\partial T}$$

- Gdzie T nazywamy progiem odcięcia

Optymalizacja progu odcięcia

- Przypiszmy miarę efektu $V(n)$ (np. mierzony pieniężnie koszt błędnej klasyfikacji) :
- Kryterium oceny modelu:
 - **oczekiwana wartość efektu**
- Cel

		Klasa prawdziwa	
		1	0
Klasa prognozowana	$[f(X)>T] = 1$	$V(TP)$	$V(FP)$
	$[f(X)>T] = 0$	$V(FN)$	$V(TN)$

$$V(TP) * E(TP) + V(FP) * E(FP) + V(FN) * E(FN) + V(TN) * E(TN) \rightarrow \max$$

- Optimum

$$\frac{\partial E(FN)}{\partial T} / \frac{\partial E(TN)}{\partial T} = (V(TN) - V(FP)) / (V(TP) - V(FN))$$

- **Wybór T zależy od relatywnego kosztu błędu!**

Wybór odpowiedniej metryki

- Bardzo ważne jest też określenie jaka jest docelowa wartość błędu do której dążymy.
- W większości przypadków nie da się osiągnąć 100% trafności predykcji.

Błąd Bayesowski: $\int_{x \in H_1} P(C_0|x)p(x)dx + \int_{x \in H_0} P(C_1|x)p(x)dx$

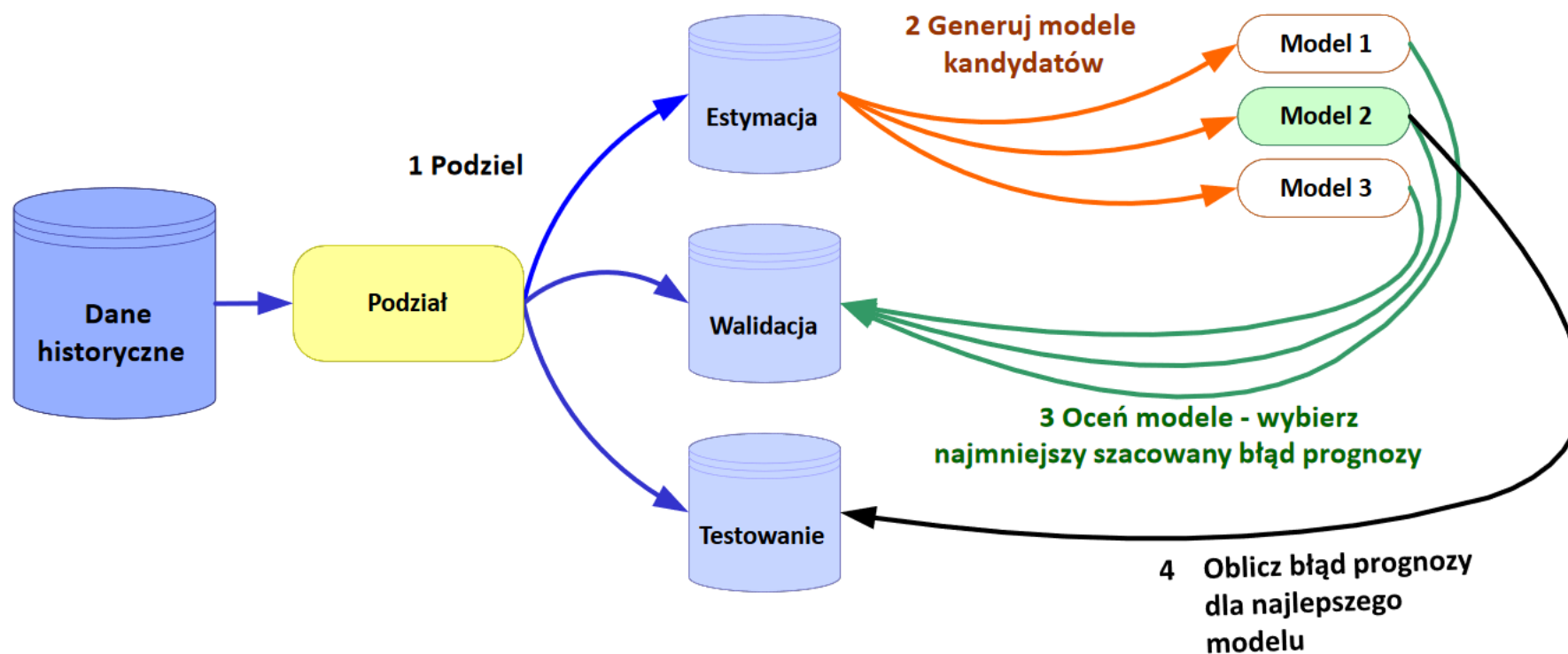
Wybór odpowiedniej metryki

- Określenie wielkości akceptowalnego błędu zależy od kilku podstawowych czynników:
 - Przede wszystkim od tego czy możliwe jest dalsze zbieranie danych (im więcej danych tym potencjalnie niższego błędu predykcji możemy się spodziewać).
 - Oraz od tego czy ze względu na czas i koszty możliwe jest tworzenie bardzo złożonych modeli.

Przygotowanie danych

- Przygotowując dane do uczenia modelu klasyfikacyjnego należy pamiętać o podzieleniu zbiorów na 3 części:
 - **Trenujący**
 - **Walidacyjny**
 - **Testowy**

Przygotowanie danych



Przygotowanie danych - skutki pominięcia zbioru walidacyjnego

- **Y**: binarna zmienna objaśniana
- **X1, X2, X3, X4**: losowe zmienne objaśniające niezależne między sobą; ze zmienną objaśnianą związana tylko zmienna X1
- 4 modele MNK oszacowane na zbiorze trenującym, za każdym razem dodawana kolejna jedna zmienna objaśniająca
- Wybrałem model o największej liczbie poprawnych klasyfikacji na zbiorze **uczącym**

Model	Uczący	Walidacyjny	Testowy
Stała+X1	64	64	63
Stała+X1-X2	65	62	63
Stała+X1-X3	65	62	62
Stała+X1-X4	66	62	61

Przygotowanie danych - skutki pominięcia zbioru testowego

- **Y**: binarna zmienna objaśniana
- **X1, X2, ... , X100**: losowe zmienne objaśniające niezależne między sobą i ze zmienną objaśnianą
- 100 modeli MNK oszacowane na zbiorze trenującym, za każdym razem wybierana jedna zmienna objaśniająca
- Wybrałem model o największej liczbie poprawnych klasyfikacji na zbiorze **walidacyjnym**
- Cztery najlepsze wyniki na zbiorze walidacyjnym (poprawny wynik to **50**)

Model	Uczący	Walidacyjny	Testowy
Najlepszy	57	66	56
Drugi	54	65	50
Trzeci	50	64	52
Trzeci	50	64	53

Przygotowanie danych

- Podział na 3 zbiory jest intuicyjnym i prostym sposobem poprawnego szacowania modeli klasyfikacyjnych.
- Przykładowy podział:
 - Zbiór trenujący: 60% obserwacji
 - Zbiór walidacyjny: 20% obserwacji
 - Zbiór testowy: 20% obserwacji
- Ma jednak zasadniczą wadę, tracimy dużą część obserwacji na których nie możemy trenować naszego modelu
- Jest to szczególnie ważne gdy zbieranie danych jest drogie i praco- lub czasochłonne.
- Dlatego często korzystamy z alternatywnych metod podziału:
 - Bootstrap aggregating (**bagging**)
 - Walidacja krzyżowa (**cross-validation**)

Przykładowe pytania zaliczeniowe

1. Podaj przykład nielosowego braku danych? Jakie konsekwencje taka sytuacja może mieć dla modelowania?
2. Co to jest liniowy model prawdopodobieństwa?
3. Jakim testem można sprawdzić, czy prognozy z regresji logistycznej można interpretować jako prawdopodobieństwo?
4. Co musi wiedzieć analityk, żeby wybrać optymalny próg odcięcia w modelu?
5. Jak powinien przebiegać podział danych na zbiór uczący, walidacyjny i testowy?
6. Jakie są konsekwencje pominięcia w modelowaniu zbioru walidacyjnego? Kiedy można go pominąć?
7. Jakie są konsekwencje pominięcia w modelowaniu zbioru testowego? Kiedy można go pominąć?