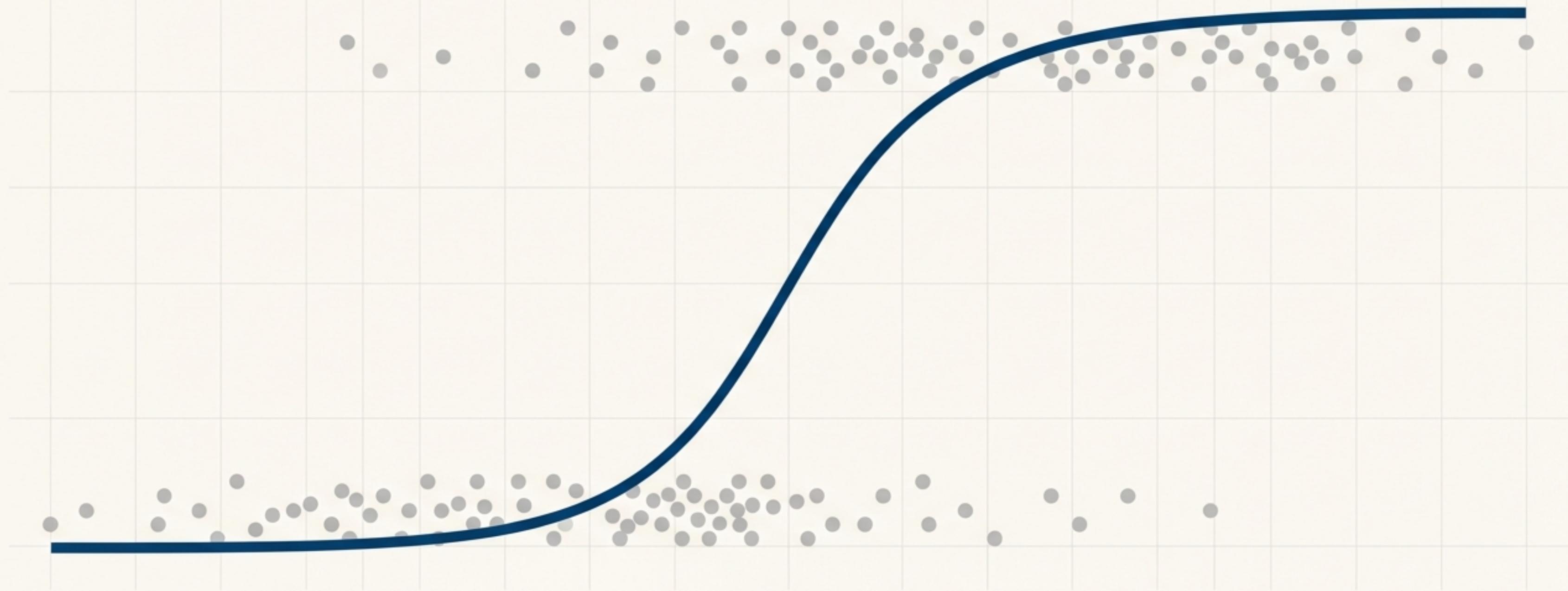


Regresja Logistyczna: Anatomia Decyzji

Jak przejść od „Ile?” do „Która?” w analizie danych



Plan Prezentacji

1. Problem: Czym jest Klasifikacja?

Definicja zadania i jego odróżnienie od regresji.



2. Falstart: Dlaczego zwykła linia nie wystarczy?

Ograniczenia regresji liniowej w zadaniach klasyfikacyjnych.



3. Rozwiążanie: Potęga funkcji sigmoidalnej.

Wprowadzenie krzywej „S” jako idealnego narzędzia.



4. Mechanizm: Jak matematycznie zgiąć linię?

Intuicyjne wyprowadzenie modelu: od prawdopodobieństwa, przez szanse, do logarytmu szans.



5. Decyzja: Od prawdopodobieństwa do etykiety.

Rola progu decyzyjnego.



6. Zastosowanie: Gdzie regresja logistyczna sprawdza się najlepiej?

Przykłady z różnych branż.



7. Werdykt: Główne zalety i ograniczenia.

Kiedy i dlaczego warto jej używać.



8. Podsumowanie: Kluczowe idee w pigułce.

Najważniejsze koncepcje do zapamiętania.



Zadanie: Przypisać Obserwację do Kategorii

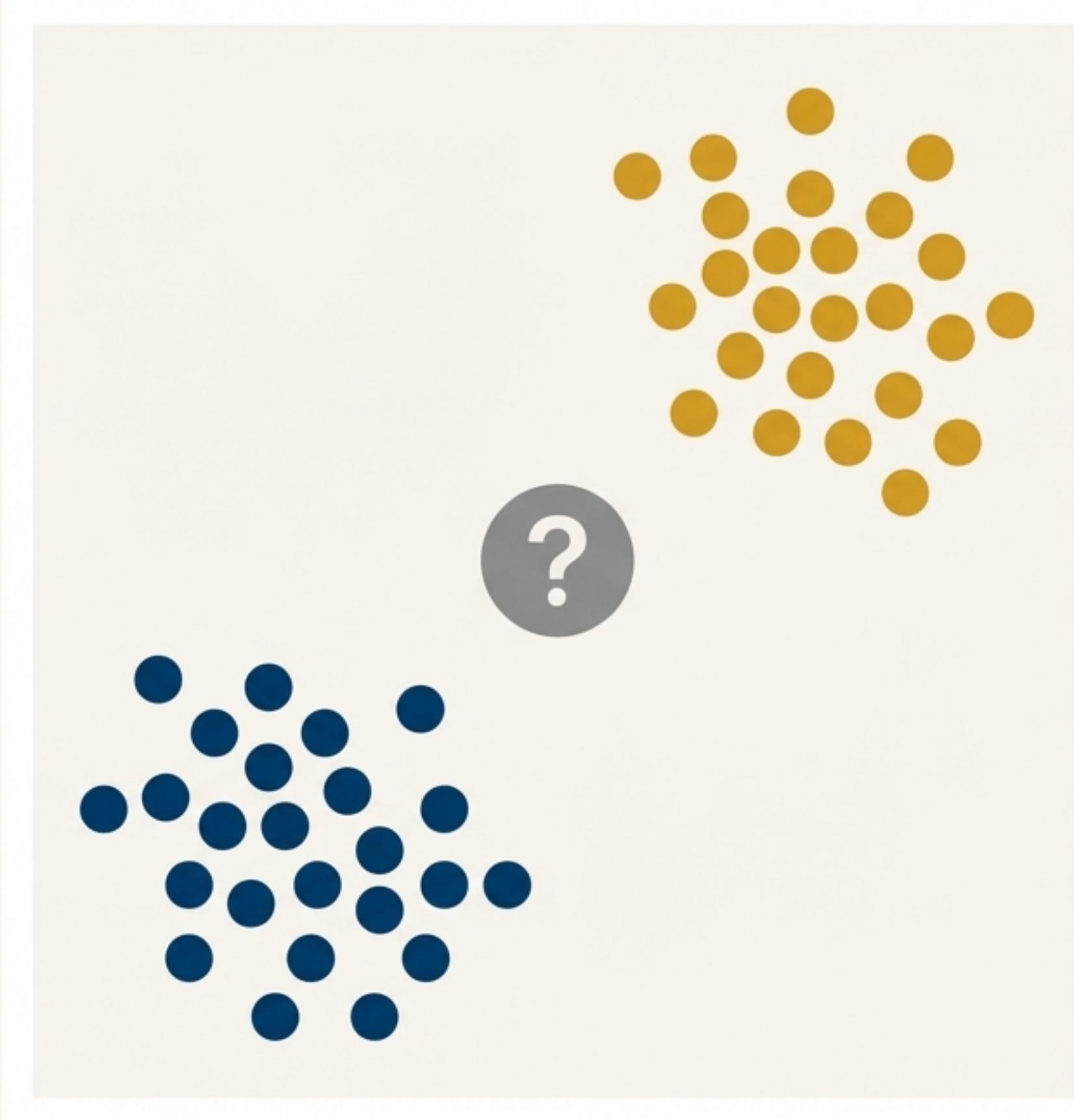
Klasifikacja w uczeniu maszynowym polega na przypisaniu obserwacji do jednej z predefiniowanych, dyskretnych kategorii. To nie jest pytanie „ile?”, ale „do której grupy należy?”.

Klasifikacja Binarna (Dwie Kategorie):

- Wiadomość to Spam czy Nie Spam?
- Transakcja jest oszustwem czy jest prawidłowa?
- Klient dokona zakupu czy zrezygnuje?

Kontrast z Regresją:

- Regresja przewiduje wartość *ciągłą* (np. cena domu, temperatura, liczba sprzedanych sztuk).
- Klasifikacja przewiduje *etykietę* (np. „kot”, „pies”; „Spam”, „Nie Spam”).

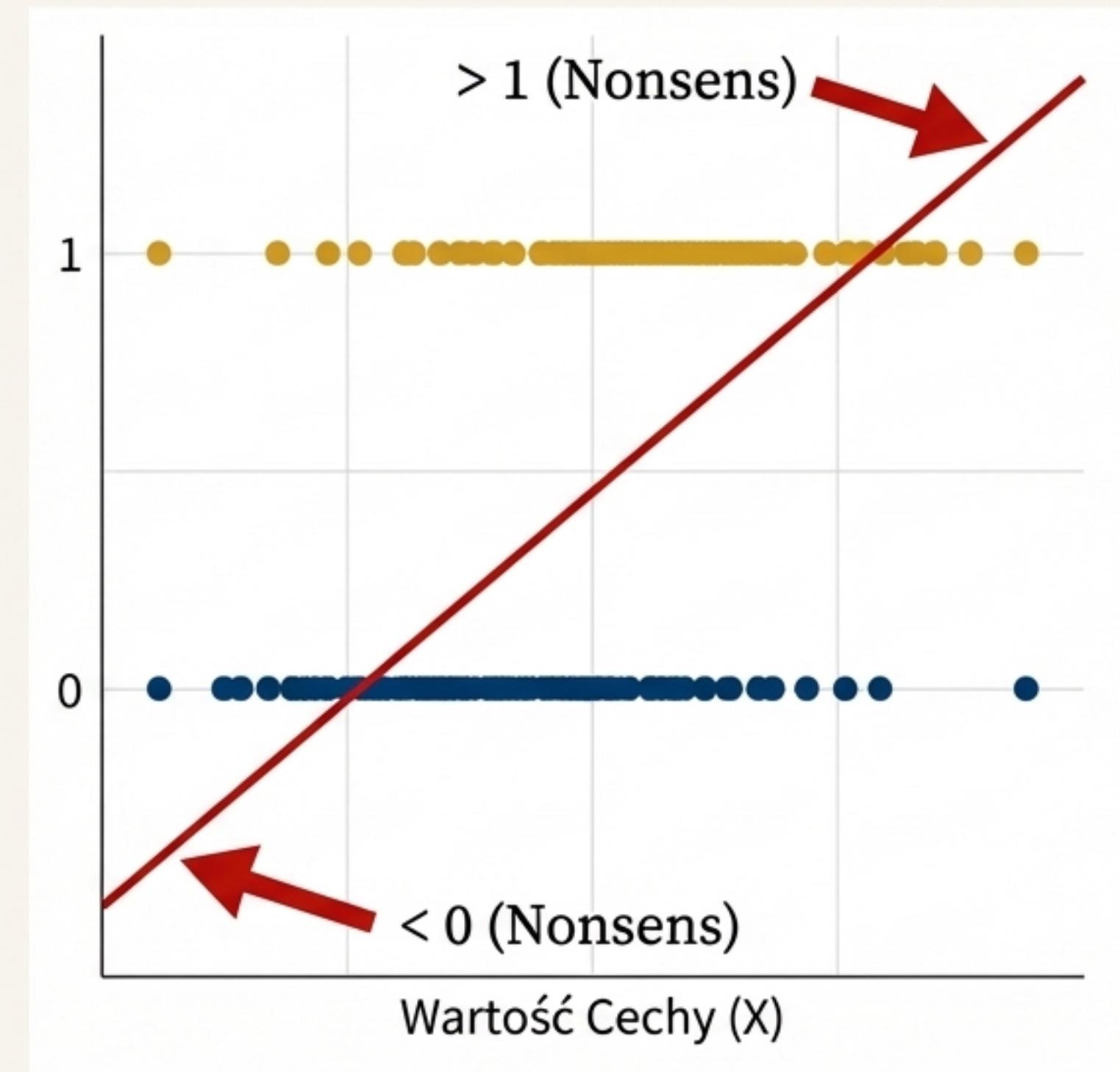


Falstart: Dlaczego Regresja Liniowa Zawodzi?

Naturalnym pomysłem mogłoby być użycie regresji liniowej do przewidywania przynależności do klasy (np. 0 dla „Nie Spam”, 1 dla „Spam”). Jednak takie podejście ma dwie fundamentalne wady:

1. Prognozy Poza Zakresem: Model może przewidzieć „prawdopodobieństwo” większe niż 1 lub mniejsze niż 0, co jest matematycznie bezsensowne.

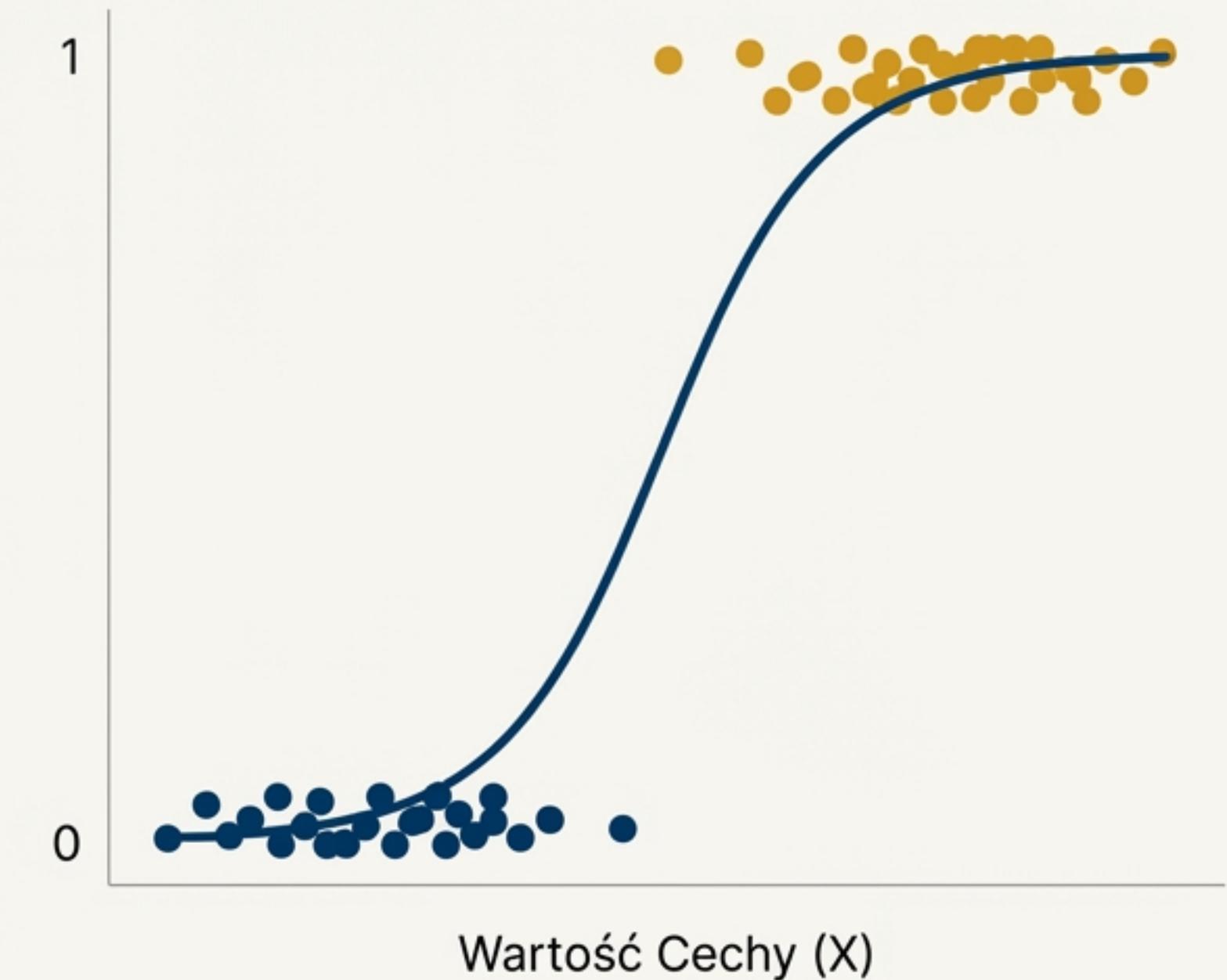
2. Brak Interpretacji Probabilistycznej: Wynik na linii (np. 0.8) nie jest prawdziwym prawdopodobieństwem. To tylko wartość na prostej.



Eleganckie Rozwiązanie: Krzywa w Kształcie „S”

Regresja logistyczna zastępuje linię prostą funkcją sigmoidalną (logistyczną). Ta funkcja posiada kluczowe właściwości, które idealnie pasują do problemu klasyfikacji:

- Ograniczony Zakres: Jej wartości są *zawsze* zawarte w przedziale od 0 do 1.
- Naturalna Interpretacja: Wyjście funkcji można bezpośrednio interpretować jako prawdopodobieństwo, że dana obserwacja należy do klasy „1” (np. $P(Y=1|X)$).



Mechanizm (Część 1): Od Prawdopodobieństwa do Szans

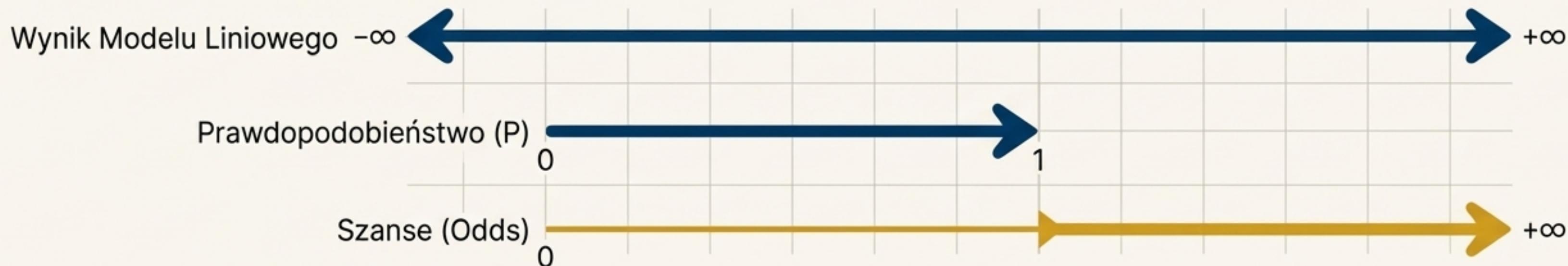
Jak matematycznie połączyć model liniowy (który generuje wartości od $-\infty$ do $+\infty$) z prawdopodobieństwem (które musi być w zakresie $[0, 1]$)? Robimy to w kilku krokach, transformując skalę.

Krok 1: Prawdopodobieństwo (P)

- To nasz cel. Jego zakres to $[0, 1]$.
- Ten krok stanowi podstawę, ale potrzebujemy dalszych transformacji.

Krok 2: Szanse (Odds)

- Definiowane jako stosunek prawdopodobieństwa „za” do „przeciw”: $\text{Odds} = P / (1-P)$.
- Przykład: Jeśli $P(\text{sukces}) = 0.8$, to szanse wynoszą $0.8 / 0.2 = 4$. Mówimy „szanse cztery do jednego”.
- Nowy zakres: Szanse przyjmują wartości od $[0, +\infty)$. Jesteśmy bliżej, ale model liniowy generuje też wartości ujemne.



Mechanizm (Część 2): Magia Logarytmu Szans (Logit)

Krok 3: Logarytm Szans (Log-odds lub Logit)

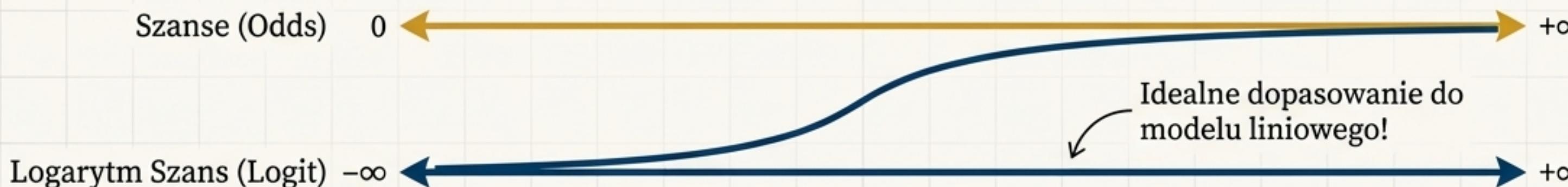
- Definiowany jako: $\log(\text{Odds}) = \log(P / (1-P))$.
- Przełom:** Logarytm „rozciąga” zakres $[0, +\infty)$ na całą oś liczbową. Jego zakres to $(-\infty, +\infty)$ – dokładnie taki sam, jak wynik modelu liniowego!

Ostateczne Równanie Modelu: Teraz możemy bezpiecznie postawić znak równości między modelem liniowym a przetransformowanym prawdopodobieństwem:

$$\log\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X_1 + \dots$$

Wyprowadzenie Funkcji Sigmoidalnej: Kiedy przekształcimy powyższe równanie, aby wyliczyć P , otrzymamy dokładnie wzór na funkcję sigmoidalną:

$$P = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \dots)}}$$



Od Prawdopodobieństwa do Klasifikacji: Próg Decyzyjny

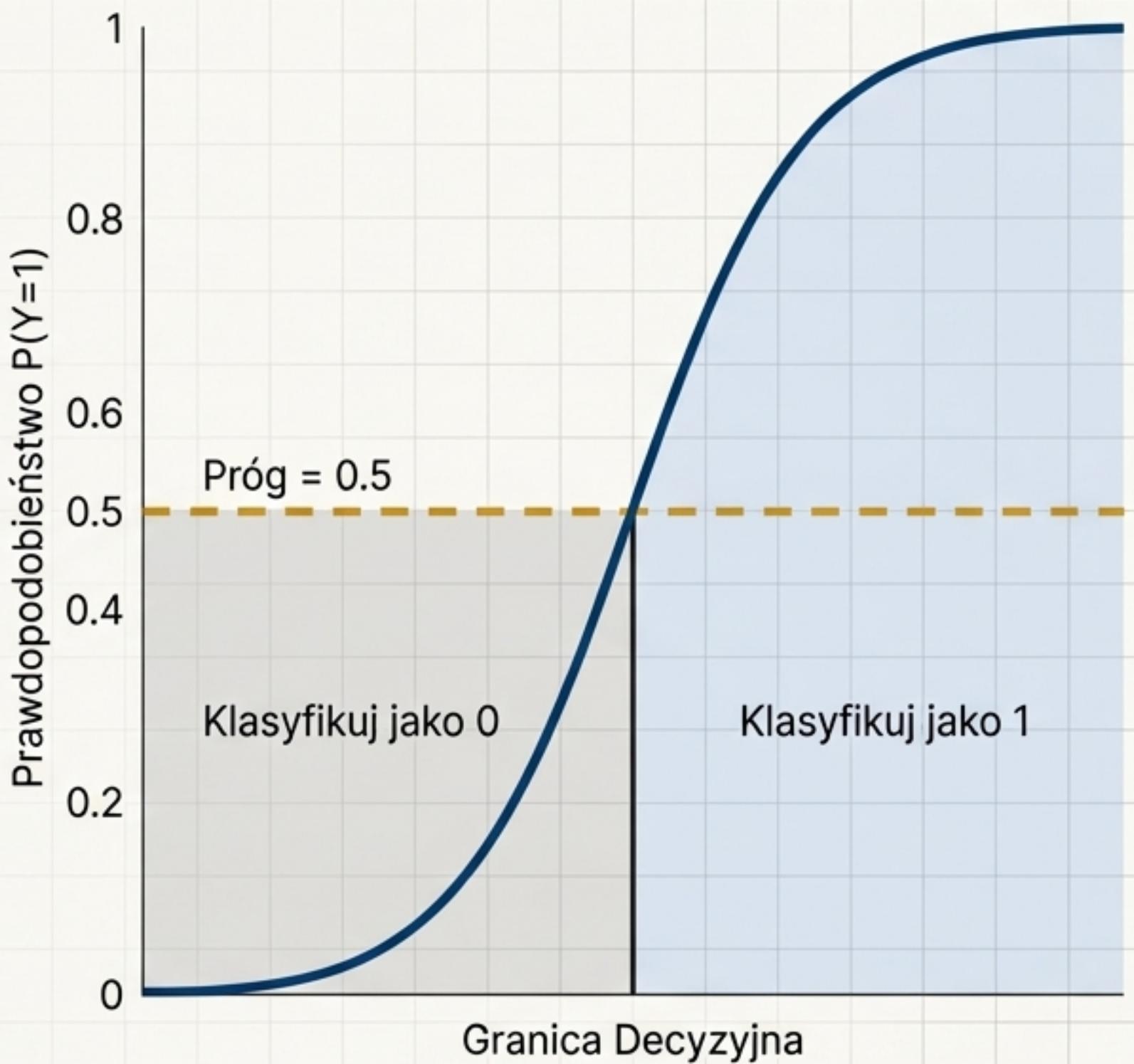
Model zwraca wartość z przedziału $[0, 1]$, ale naszym celem jest przypisanie etykiety (np. „Spam” lub „Nie Spam”). Potrzebujemy reguły decyzyjnej.

Próg Decyzyjny (Decision Threshold): Standardowo jest to wartość 0.5.

Reguła:

- Jeśli $P(Y=1) \geq 0.5$, przypisz klasę **1** (np. „**Kupi**”).
- Jeśli $P(Y=1) < 0.5$, przypisz klasę **0** (np. „**Nie Kupi**”).

Ważna uwaga: Próg ten można (i często należy) dostosować w zależności od problemu biznesowego. Na przykład, przy wykrywaniu rzadkich chorób, możemy chcieć obniżyć próg, aby nie pominąć żadnego potencjalnego przypadku, nawet kosztem fałszywych alarmów.



Gdzie Regresja Logistyczna Sprawdza się Najlepiej?

Tabela prezentująca kluczowe zastosowania regresji logistycznej w różnych branżach.

Branża	Problem Biznesowy	Wynik (Klasa 0 / Klasa 1)	Kluczowe Cechy (Przykłady X)
Medycyna	Diagnoza choroby	Zdrowy / Chory	Wyniki badań, wiek, BMI, czynniki genetyczne
Finanse	Ryzyko kredytowe	Kredyt zostanie spłacony / Kredyt nie zostanie spłacony	Historia kredytowa, dochód, wiek, kwota pożyczki
Marketing	Rezygnacja klienta (Churn)	Klient zostanie / Klient odejdzie	Czas od ostatniego zakupu, wartość zakupów, liczba interakcji
E-commerce	Konwersja (zakup)	Użytkownik nie kupi / Użytkownik kupi	Czas na stronie, liczba odwiedzonych produktów, źródło ruchu
Bezpieczeństwo IT	Filtrowanie spamu	Wiadomość bezpieczna / Spam	Obecność słów kluczowych, nadawca, rodzaj załączników
HR	Retencja pracowników	Pracownik zostanie / Pracownik odejdzie	Poziom satysfakcji, staż pracy, wynagrodzenie, oceny roczne

Główne Zalety: Siła Prostoty i Interpretowalności



Interpretowalność

To największa zaleta. Możemy precyzyjnie określić, jak zmiana o jednostkę w danej cesze (X) wpływa na *logarytm szans* wystąpienia wyniku. Współczynniki mają jasne, biznesowe znaczenie.



Efektywność Obliczeniowa

Model jest bardzo szybki w trenowaniu i nie wymaga dużych zasobów obliczeniowych. To idealny kandydat na pierwszy, bazowy model (tzw. baseline) w każdym projekcie klasyfikacyjnym.



Wyjście Probabilistyczne

Zwraca prawdopodobieństwo, a nie tylko twardą klasyfikację. Daje to wgląd w „pewność” modelu i pozwala na dalsze modelowanie ryzyka.



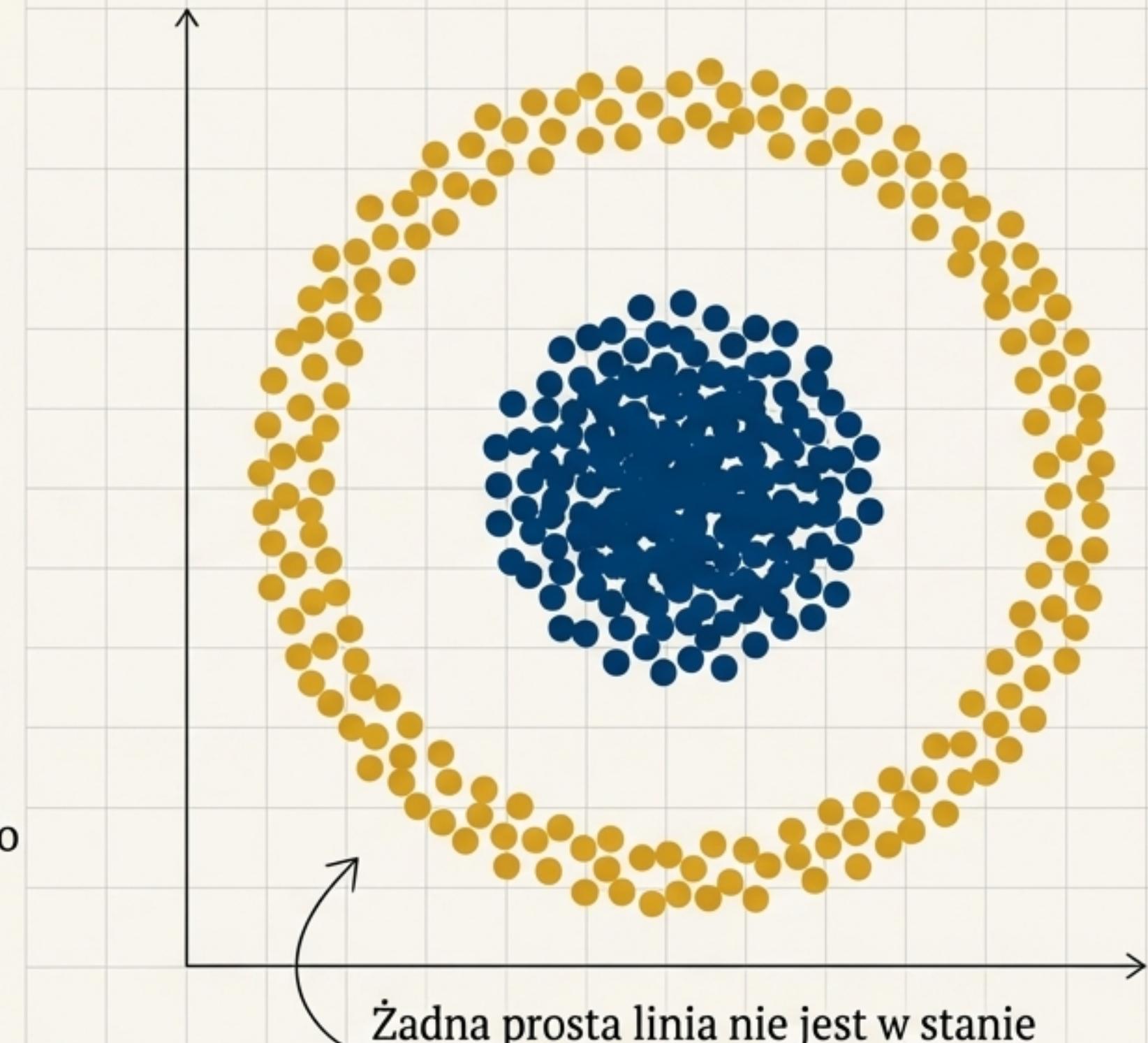
Solidny Fundament Edukacyjny

Zrozumienie regresji logistycznej jest kluczowe do zrozumienia bardziej złożonych modeli, w tym sieci neuronowych (pojedynczy neuron często używa aktywacji sigmoidalnej).

Ograniczenia: Kiedy Należy Sięgnąć po Inne Narzędzia?

Mimo swoich zalet, model ma istotne ograniczenia.

- **Założenie o Liniowości:** Model zakłada liniową zależność między cechami (X) a *logarytmem szans* wyniku. Nie jest w stanie samodzielnie modelować złożonych, nieliniowych relacji w danych.
- **Wrażliwość na Obserwacje Odstające:** Ekstremalne wartości mogą znacząco wpływać na dopasowanie modelu, podobnie jak w regresji liniowej.
- **Problem z Idealną Separacją Danych:** Jeśli zbiór danych można idealnie podzielić linią, algorytm może mieć problemy z konwergencją (znalezieniem stabilnego rozwiązania).
- **Wymaga Inżynierii Cech:** Aby modelować interakcje między zmiennymi, trzeba je ręcznie dodać do modelu. Bardziej zaawansowane algorytmy robią to automatycznie.



Żadna prosta linia nie jest w stanie poprawnie oddzielić tych dwóch klas.

Regresja Logistyczna w Pigułce



CEL: KLASYFIKACJA BINARNA

Odpowiada na pytania typu „tak/nie”, „0/1”, „prawda/fałsz”.



WYNIK: PRAWDOPODOBIEŃSTWO

Zwraca prawdopodobieństwo przynależności do klasy „1”, które następnie przekształcamy w decyzję za pomocą progu.



NARZĘDZIE: FUNKCJA SIGMOIDALNA

Przekształca wynik dowolnej kombinacji liniowej cech na wartość z przedziału [0, 1].



SIŁA: INTERPRETOWALNY MODEL BAZOWY

Jest szybka, wydajna i jej wyniki są łatwe do wyjaśnienia interesariuszom biznesowym.