# ML w problemie scoringu – case study

Mateusz Grzyb, ITMAGINATION

Cześć!

# Mateusz Grzyb

Data Scientist @ ITMAGINATION

# Zarys projektu

# Zarys projektu

## Najważniejsze informacje na temat projektu:

- Klient: sektor finansowy, **top 3** w swojej branży.
- Start: listopad 2017, koniec: maj 2018.
- Przez projektem klient nie korzystał z systemu scoringowego.
- Decyzje podejmowane w oparciu o dane przez ekspertów.
- Zespół ITM – ok. 10 osób.

# System scoringowy

# Definicja systemu scoringowego

## Definicja:

- System oceny pozwalającej na klasyfikację obserwacji na podstawie wybranych cech.
- Pozwala sklasyfikować podmioty na dwie klasy: dobre i złe.
- Predykcja zostaje wyznaczona dla przyjętego **horyzontu czasowego**.

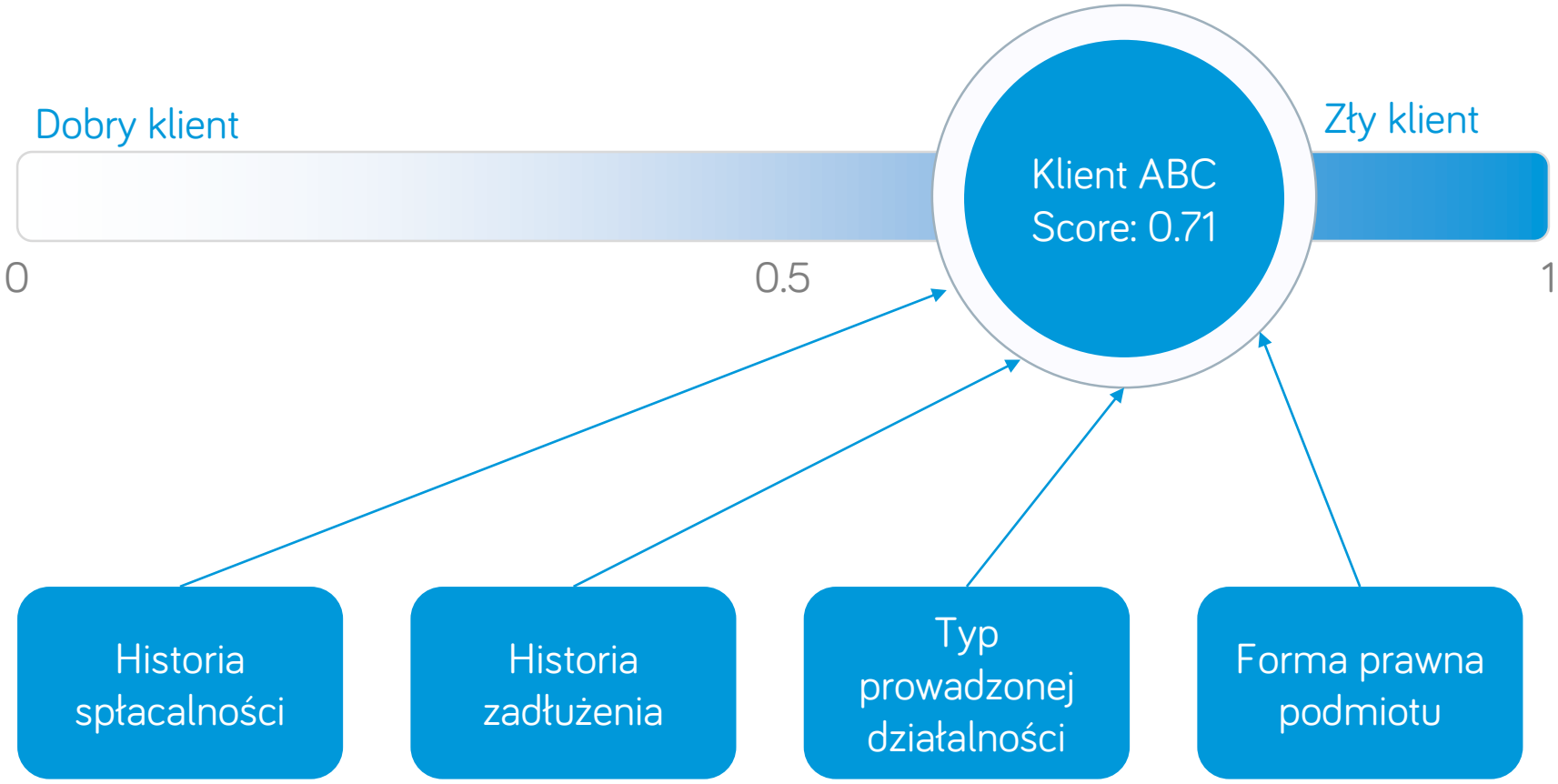
## Zastosowanie:

- Sektor finansowy (ocena ryzyka kredytowego, wgląd w wypłacalność klienta).

## Główne korzyści:

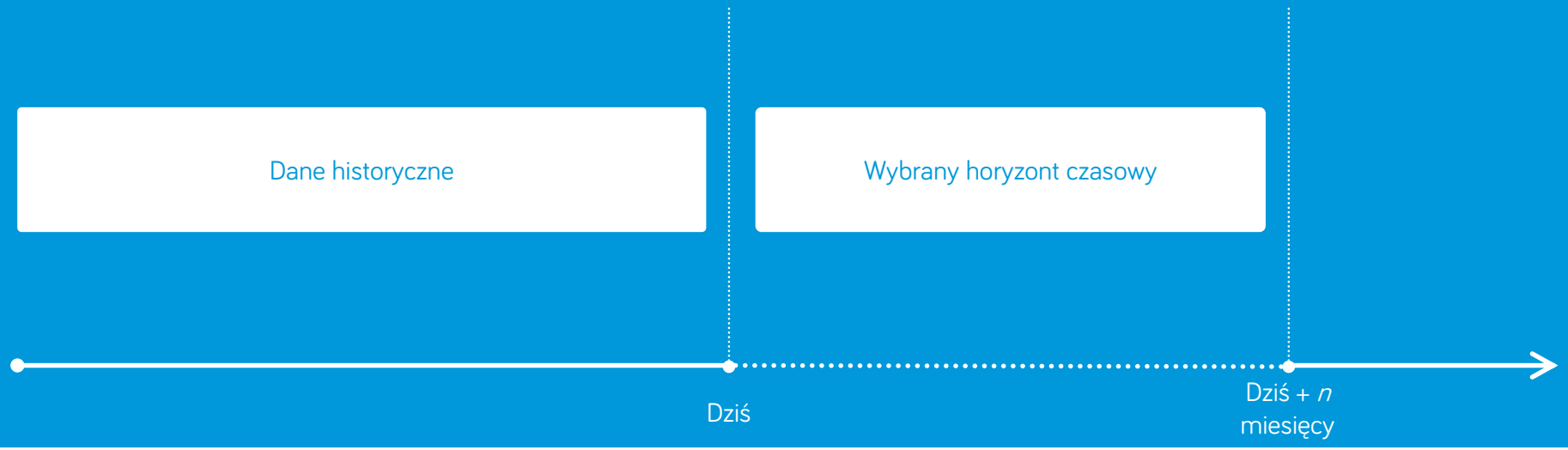
- Automatyzacja podejmowanych decyzji.
- Ograniczanie ryzyka.
- Optymalizacja kosztów (np. kredytu).
- Wyższe zyski.

# Definicja systemu scoringowego




Przykładowe zmienne:

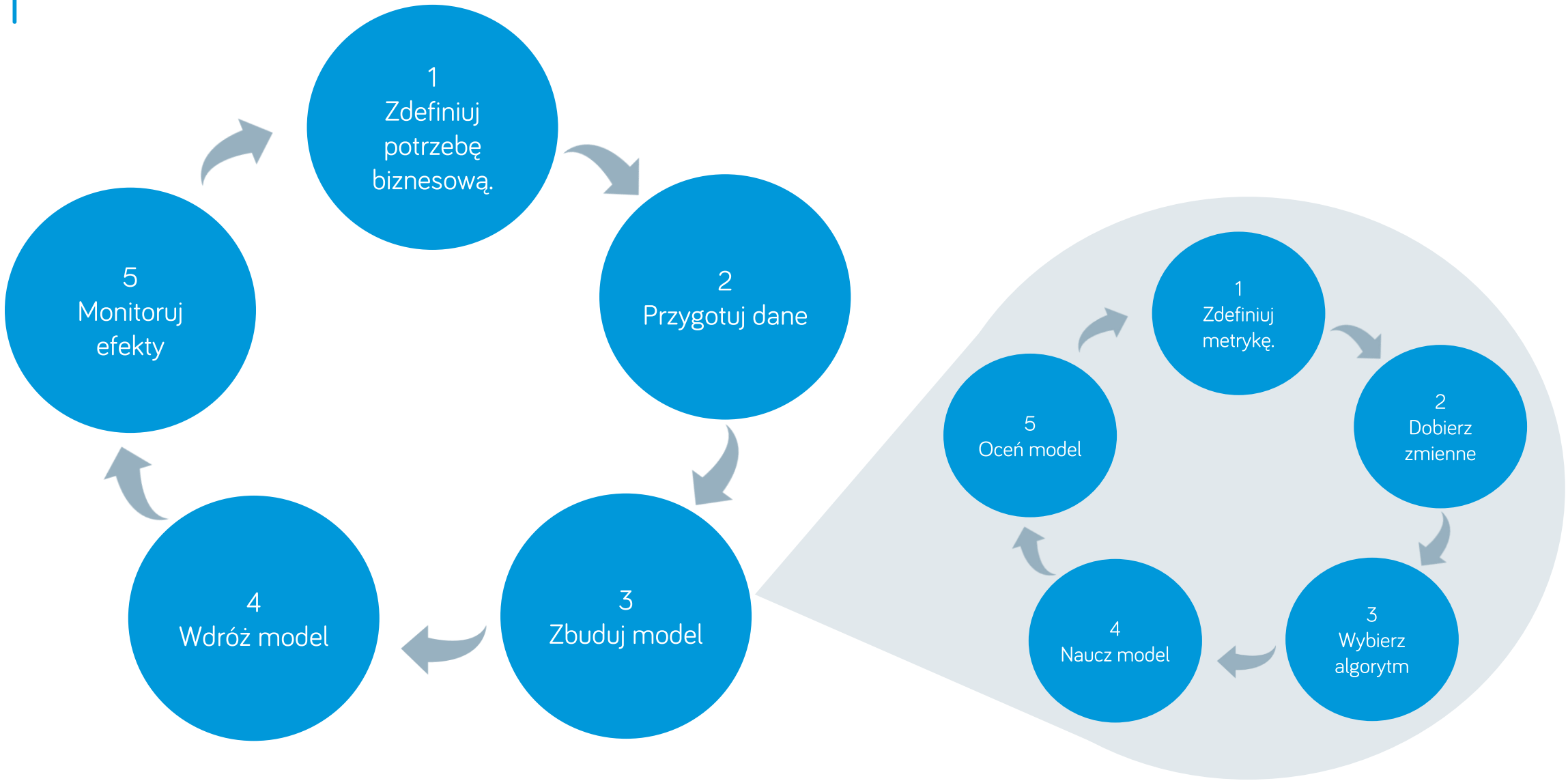
# Horyzont czasowy




**Linia życia danego klienta**

# Proces budowy systemu

# Proces budowy systemu scoringowego



1

# Zrozumienie biznesu

Zrozumienie potrzeb klienta, ustalenie celu i pierwsze wyzwania.

2

# Zrozumienie danych

Poznanie dostępnych źródeł danych, ich podziału i struktury.

3

# Ustalenie finalnej definicji zmiennej celu

Definicja zmiennej objaśnianej zgodna z celem biznesowym.

# Ustalenie finalnej definicji zmiennej celu

KTO? - klientem nierzetelnym, nazywano klienta, u którego w ciągu ostatnich *n* miesięcy:

- Wystąpiły problemy ze spłatą zobowiązań.
- Zmiana salda zobowiązań.

KIEDY? - wybór horyzontu czasowego:

- Wystarczająco wysoka zmienność danych (liczba transakcji w danym okresie).
- Relatywnie krótki czas potrzebny do walidacji wyników modelu.

4

# Pozyskanie nowych danych

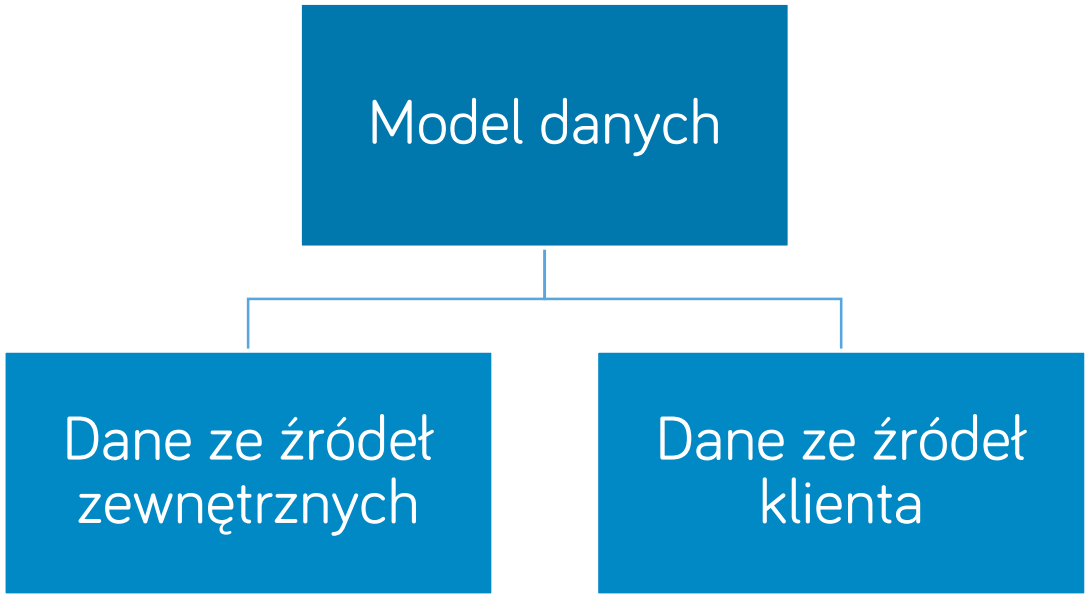
Wyjście naprzeciw oczekiwaniom klienta.

# Pozyskanie nowych danych

Dwa modele:

- Aplikacyjny i behawioralny.

- Dane z wewnętrznych źródeł klienta:
  - Tabele opisujące klientów (teleadresowe).
  - Tabele opisujące historię transakcji.
  - Tabele opisujące wierzycieli.
- Dane pochodzące ze źródeł zewnętrznych.



5

# Przygotowanie danych

Odpowiedni okres czasu, filtrowanie danych o złej jakości, podział na „koszyki”, próbkowanie, przecieki.

6

# Modelowanie

Budowa modelu aplikacyjnego i behawioralnego.

# Modelowanie - opis dostępnych zmiennych

## Model 1 (behawioralny):

- Dane klientów.
- Dane dotyczące historii prowadzenia konta.
- Dane dotyczące historii spłat zobowiązań.
- Dane produktów z jakich korzystał klient.
- Dane ze źródeł zewnętrznych.

## Model 2 (aplikacyjny)

- Dane ze źródeł zewnętrznych.

# Model 1 – połączone źródła danych

## Charakterystyka modelu:

- Wykorzystano połączone źródła danych:
  - Dane ze źródeł wewnętrznych.
  - Dane ze źródeł zewnętrznych.
- Dostęp do danych behawioralnych.
- Możliwość użycia jedynie dla podmiotów znajdujących się w bazie danych klienta.

# Model 2 – dane z rejestrów zewnętrznych

## Charakterystyka modelu:

- Wykorzystano dane zewnętrzne.
- Brak dostępu do danych behawioralnych.
- Możliwość użycia dla wszystkich podmiotów znajdujących.

# Modelowanie - wybór odpowiedniej miary jakości

Jako miary jakości modeli przyjęto:

- Gini
  - Podstawowa miara jakości modelu.
  - Przyjmuje wartość 0 - 1.
- Czułość
  - Procentowa wartość wszystkich dobrze rozpoznanych *defaultów*.
  - Przyjmuje wartość 0 – 100.
- Stabilność
  - Procentowa wartość wszystkich dobrze sklasyfikowanych klientów.
  - Przyjmuje wartość 0 – 100.
  - Badana podczas walidacji krzyżowej z pomocą współczynnika zmienności.

# Modelowanie – algorytmy

## W ramach testów zbudowano modele:

- Naiwny klasyfikator Bayesa.
- Drzewo decyzyjne.
- Regresja logistyczna.

## Różne założenia:

- Naiwny Bayes
  - Brak zależności pomiędzy zmiennymi.
  - Duże znaczenie ma przygotowanie danych.
- Drzewo decyzyjnego
  - Niewrażliwe na odstające wartości.
  - Optymalizacja parametrów > przygotowanie danych.
- Regresja logistyczna
  - Problemy ze współliniowością zmiennych.
  - Dobór zmiennych i przygotowanie danych > optymalizacja parametrów.

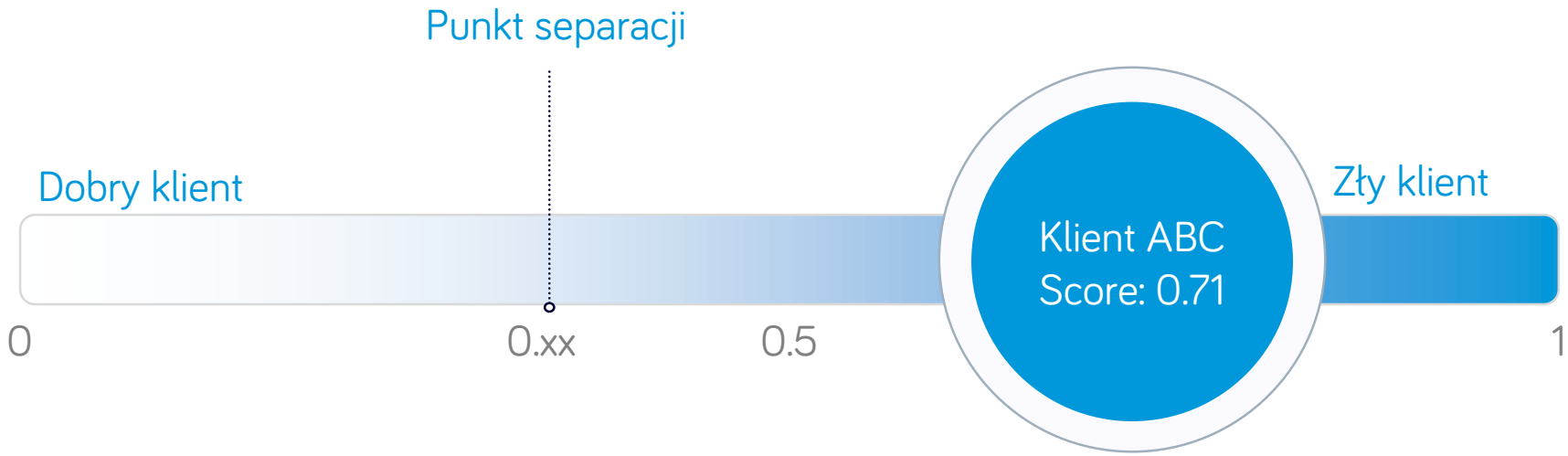
7

# Omówienie i przedstawienie wyników

Uzyskane rezultaty dla obu modeli, sposób prezentacji wyników scoringu.

# Sposób przedstawiania wyników

- Każdemu klientowi zostaje nadany **score** będący wartością z zakresu <0, 1>.
- W oparciu o dane pozyskane z procesie uczenia i testowania modelu, zostanie wyznaczony **punkt separacji**.
- Im bliżej wartości 0, tym większe prawdopodobieństwo, że **klient okaże się rzetelny**.
- Im bliżej wartości 1 tym większe prawdopodobieństwo, że **klient okaże się nierzetelny**.



# Kluczowe wnioski z projektu

# WNIOSEK #1

Przygotowanie danych to 80% sukcesu.

# WNIOSEK #2

Regresja logistyczna – algorytm pierwszego wyboru dla problemu scoringu.

# WNIOSEK #3

Interpretowalność wyników regresji logistycznej > interpretowalność drzewa decyzyjnego (wg biznesu).

# WNIOSEK #4

Dobór zmiennych kluczem do sukcesu w modelu regresji logistycznej.

# Dobór zmiennych do modelu regresji logistycznej

1. Usunięcie zmiennych o **małej zmienności**.
2. Usunięcie zmiennych o **zbyt dużej liczbie brakujących wartości** (uwaga: brak może być informacją).
3. Badanie współczynników korelacji **pomiędzy zmiennymi objaśniającymi**.
4. Badanie **istotności zmiennych** (pozostało ok. 200 zmiennych).
5. Wybór zmiennych objaśniających o możliwie **najmniejszym współczynniku korelacji**.
   - Spośród par zmiennych wysoce skorelowanych wybieraliśmy tą zmienną, która miała **wyższą istotność** (< 100 zmiennych).
6. Analiza współczynnika **VIF**.
7. **Forward selection** (< 20 zmiennych).

# WNIOSEK #5

Dobór parametrów modelu kluczem do sukcesu w modelu drzewa decyzyjnego (CART).

# Dobór parametrów do modelu drzewa decyzyjnego

## Drzewo decyzyjne (CART):

- Niewrażliwe na odstające wartości.
- Brak założeń dotyczących normalności rozkładu.
- GridSearch vs RandomizedSearch
  - przeszukiwanie siatki wartości ciągłych,
  - problem z dopasowaniem modelu do danych.

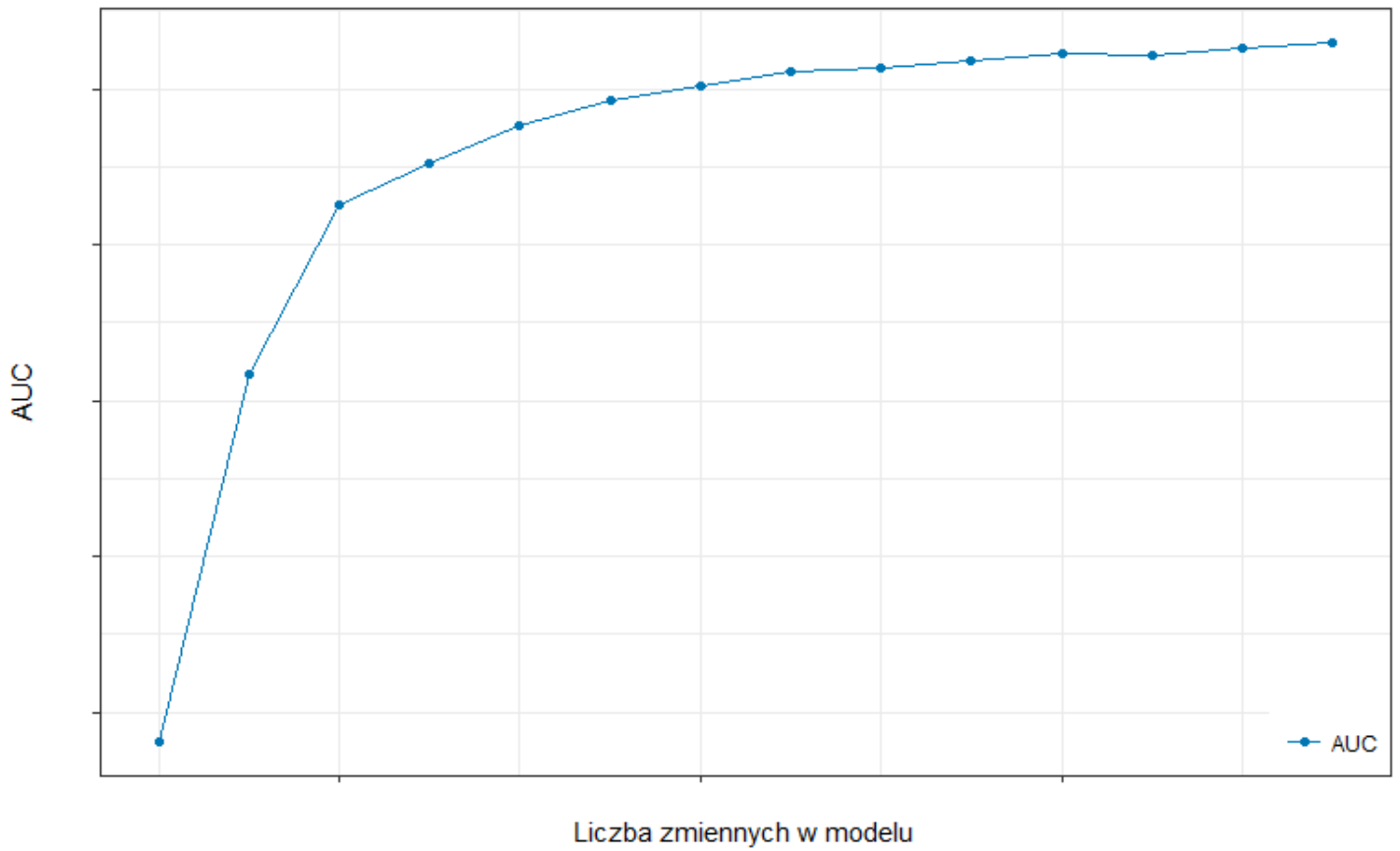
# WNIOSEK #6

Interpretowalność to nie sam algorytm.

# Interpretowalność to nie sam algorytm

Na interpretację składają się m.in.:

- Odpowiedni algorytm.
- Intuicyjność zmiennych.
- Transformacje wykonane na zbiorze.
- Liczba zmiennych.


AUC
Liczba zmiennych w modelu
AUC

# WNIOSEK #7

Finalny sukces, to nie tylko zasługa modelowania.

# WNIOSEK #8

Siła tkwi w mocnym, zróżnicowanym zespole.

Dziękuję!

# Pytania?

**Kontakt:** mateusz.grzyb@itmagination.com

**Slajdy i materiały:** MateuszGrzyb.pl/DSS