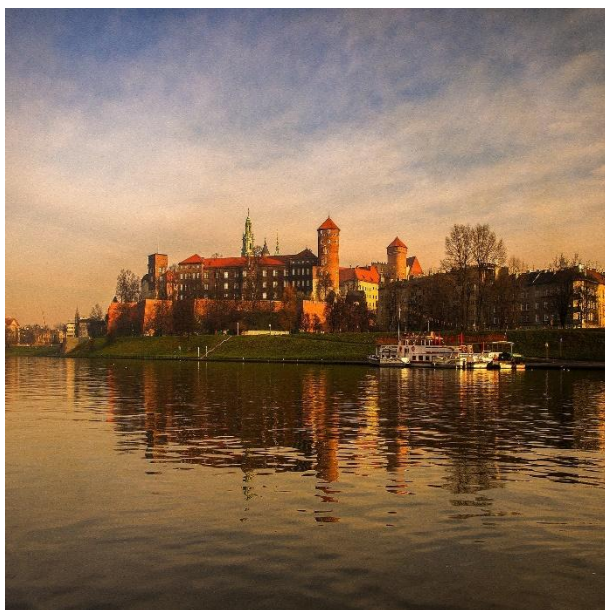
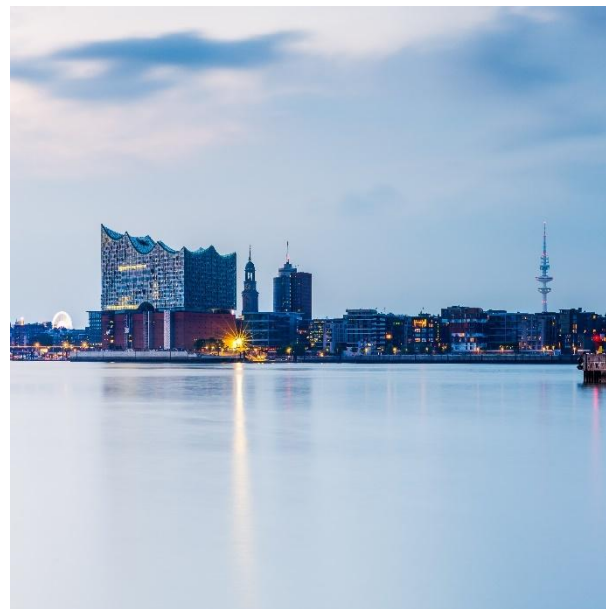




Budowa skutecznych modeli uczenia maszynowego – co determinuje ich jakość?

dr Grzegorz Migut

O prowadzącym

- Dyrektor działu technicznego w firmie StatSoft Polska
- Od ponad 20 lat zajmuje się analityką danych
- Realizował projekty w obszarach
 - Biznesu
 - Przemysłu
 - Nauki
- Zainteresowania
 - Uczenie maszynowe
 - Modele klasyfikacyjne
 - Analizy wielowymiarowe
- Kontakt: g.migut@statsoft.pl



O prowadzącym

- Polecam książkę na temat modelowania
 - Budowa modeli uczenia maszynowego
 - Kroki, etapy, zadania



Agenda

Uzasadnienie wyboru problemu badawczego

Charakterystyka obszaru badań i luka badawcza

Cel pracy, pytania badawcze

Procedura badawcza

Uzyskane wyniki



Uzasadnienie wyboru problemu badawczego

- Rosnąca ilość gromadzonych danych
- Rosnąca potrzeba syntezy informacji zawartych w danych
- Sprzedaż i marketing jednym z kluczowych aspektów działania przedsiębiorstw
 - Zrozumienie motywacji klientów
 - Umiejętność dopasowania oferty do szybko zmieniających się preferencji nabywców
 - Pogłębianie i poszerzanie relacji z klientami
- Poprawa procesów biznesowych dzięki wiedzy płynącej z modeli ekonometrycznych oraz modeli uczenia maszynowego

Uzasadnienie wyboru problemu badawczego

- Modele dobrej jakości
 - Trafne decyzje biznesowe
 - Racjonalne gospodarowanie dostępnymi zasobami
 - Skuteczne strategie zatrzymania klientów
- Charakterystyka modeli dobrej jakości
 - Zdolność do generalizacji
 - Możliwość interpretacji modelu
 - Stabilność działania w podsegmentach
 - Odporność na zmiany rozkładów predyktorów

Uzasadnienie wyboru problemu badawczego

- Obszar dociekań ograniczony do modeli lojalności klienta
- Niestąbnące zapotrzebowanie rynku na budowę tego typu modeli
- Kluczowe kwestie dla zespołów marketingu
 - Utrzymanie klientów
 - Zrozumienie czynników wpływających na ich lojalność
 - Umiejętność przewidywania zachowania

Agenda

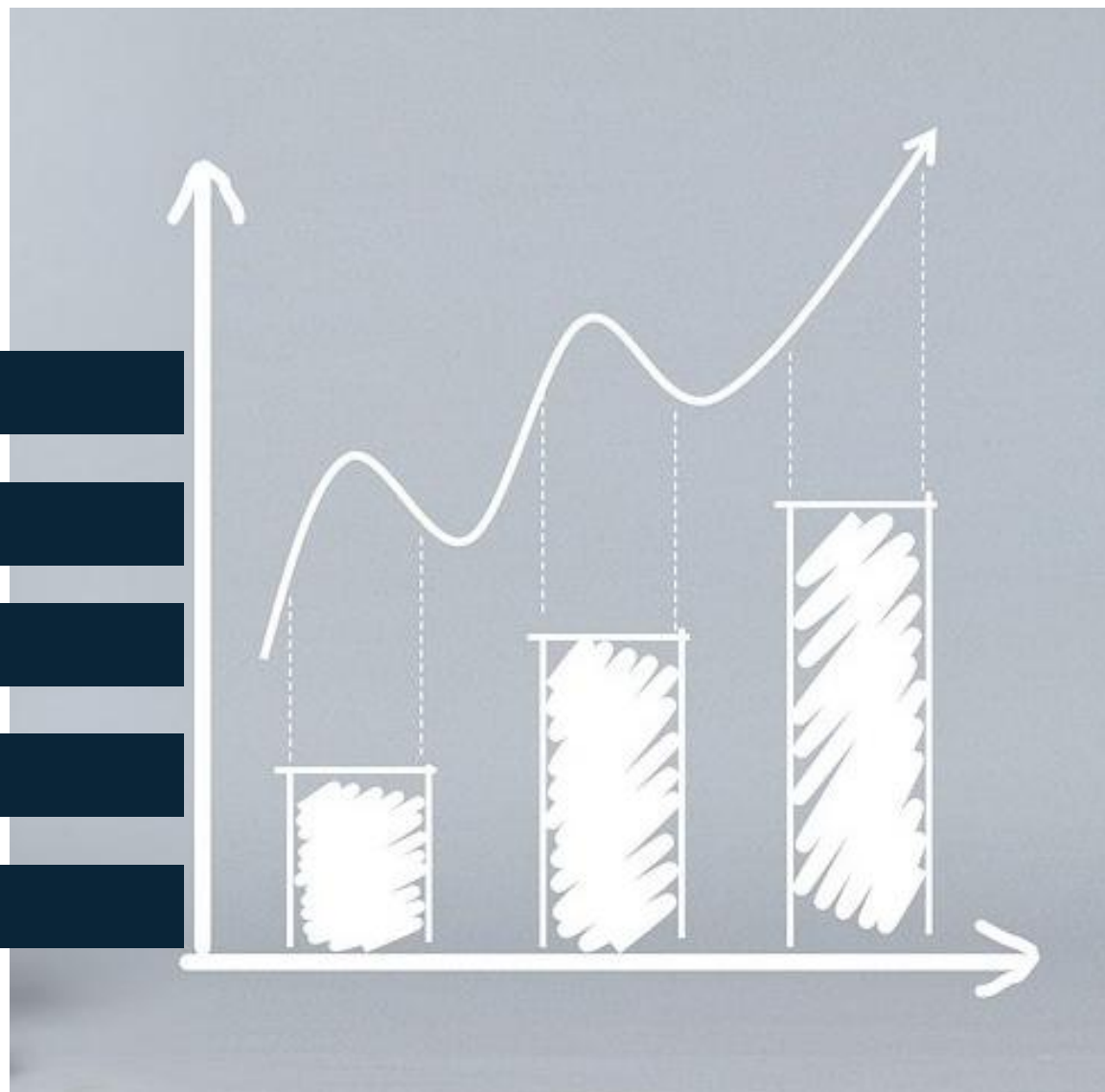
Uzasadnienie wyboru problemu badawczego

Charakterystyka obszaru badań i luka badawcza

Cel pracy, pytania badawcze

Procedura badawcza

Uzyskane wyniki



Charakterystyka obszaru badań

- Szkoła marketingu relacji oraz tradycja badawcza oparta na CRM (*Customer Relationship Management*)
 - Silne, długotrwałe więzi z klientami
 - Lojalność i satysfakcja klienta
- Obszar analityczny
 - Masowy rynek B2C
 - Modele lojalności klienta

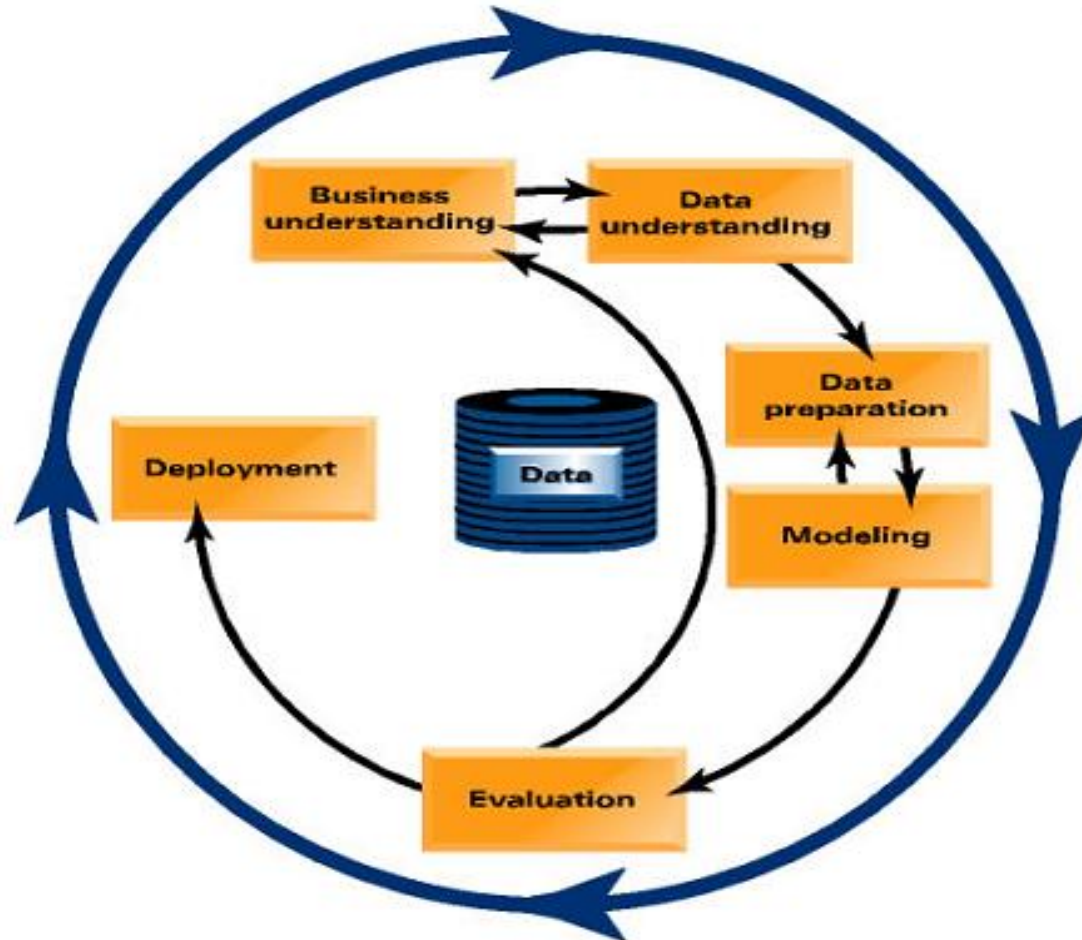
Charakterystyka obszaru badań

- Proces budowy modeli migracji klientów zgodny z koncepcją eksploracyjnej analizy danych (EDA, *Exploratory Data Analysis*)
 - Analiza opiera się na samych danych
 - Uzyskanie wiarygodnych wyników wymaga iteracyjnego testowania i modyfikowania rezultatów analizy na podstawie informacji zwrotnych
- Proces wnioskowania zgodny z EDA [Ratner 2017]



Charakterystyka obszaru badań

- Realizacja pracy badawczej zgodna z metodyką CRISP-DM



Źródło: [CRISP-DM 2000]

Luka badawcza

- Obszar **wieloaspektowej** oceny jakości modeli retencji klientów
- Brak opracowań syntezyujących **wpływ wielu determinant** jakości modeli na końcowy rezultat modelowania
- Określenie strategii budowy modeli **najczęściej** prowadzących do uzyskania pożądanых przez badacza rezultatów
- Wiedza oparta na wynikach badań może być przyczynkiem do opracowania narzędzi wspierających pracę analityka na przykład w postaci kreatorów w bezpłatnych lub komercyjnych narzędziach analitycznych.

Charakterystyka obszaru badań

- Determinanty jakości modeli - *TIVHE*
 - *Transformation* – sposób przygotowania predyktorów, dyskretyzacja, standaryzacja itp.,
 - *Interaction* – fakt uzupełnienia zbioru danych o zmienne pochodne (*derived variables*),
 - *Variables* – sposób doboru zmiennych do modelu,
 - *Hyperparameters* – metody optymalizacji hiperparametrów,
 - *Ensembles* – dodatkowe strategie uczenia: segmentacja, hybrydyzacja, agregacja modeli.

Agenda

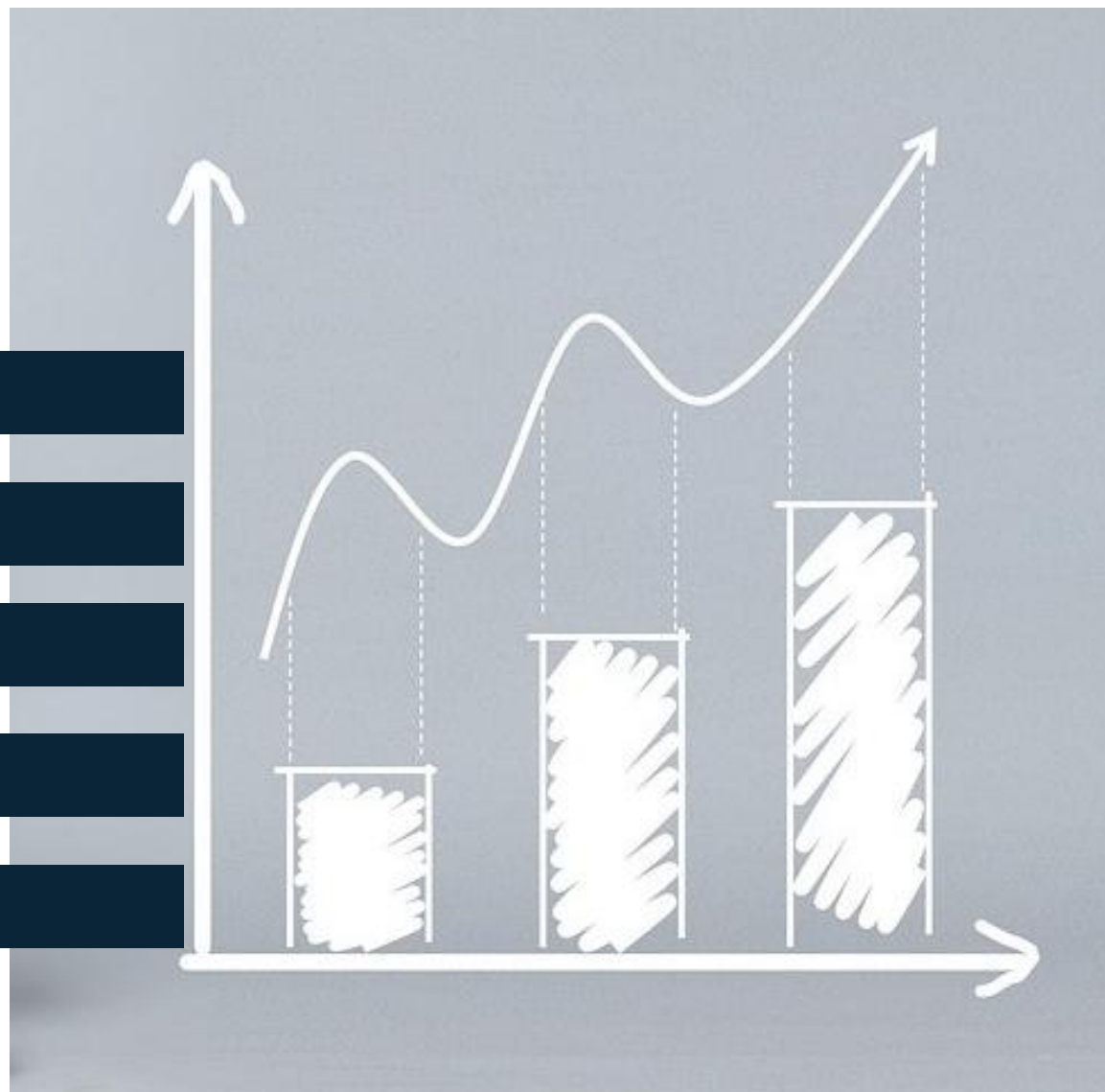
Uzasadnienie wyboru problemu badawczego

Charakterystyka obszaru badań i luka badawcza

Cel pracy, pytania badawcze

Procedura badawcza

Uzyskane wyniki



Cel pracy

- Eksploracyjny charakter pracy
- Brak hipotez badawczych, zdefiniowano cel główny i cele poboczne
- **Cel główny:** Identyfikacja determinant wpływających na jakość modeli migracji klientów oraz określenie relacji między nimi

Cele poboczne

- Porównanie skuteczności działania modelu regresji logistycznej z modelami uczenia maszynowego zbudowanymi za pomocą drzew klasyfikacyjnych, perceptronu wielowarstwowego, drzew wzmocnianych.
- Ocena wpływu hybrydyzacji, segmentacji oraz agregacji na jakość budowanych modeli.
- Wykazanie możliwości budowy modeli o zadowalających własnościach za pomocą metod interpretowalnych przez badacza (biała skrzynka), porównywalnych z zaawansowanymi metodami nieinterpretowalnymi (czarna skrzynka), przy użyciu odpowiednich technik przygotowania danych oraz hybrydyzacji modeli.

Agenda

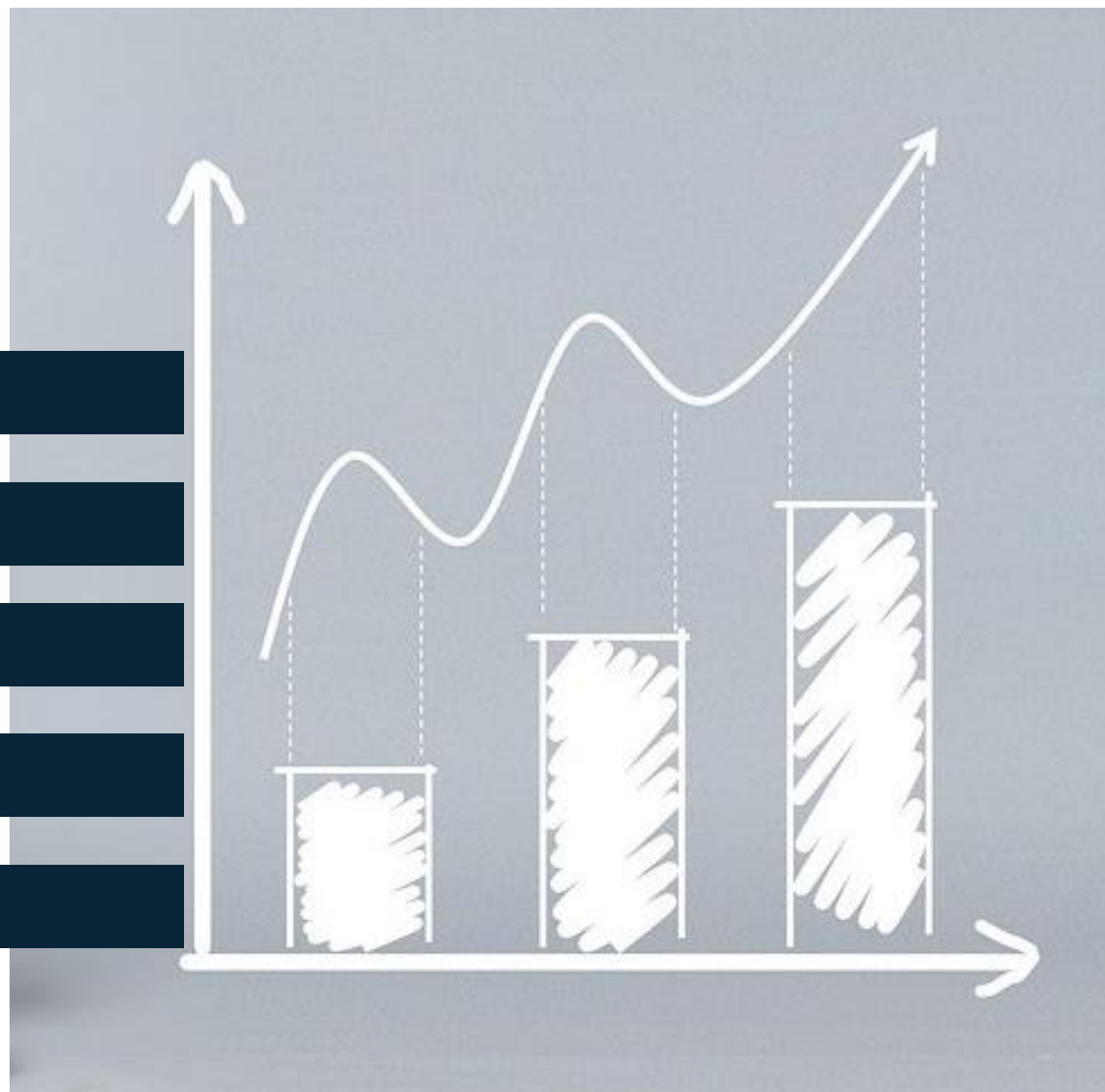
Uzasadnienie wyboru problemu badawczego

Charakterystyka obszaru badań i luka badawcza

Cel pracy, pytania badawcze

Procedura badawcza

Uzyskane wyniki



Zbiór danych

- 100 tys. przypadków - każdy przypadek dotyczy jednego abonenta sieci telefonicznej
- 173 zmienne
 - Zmienna zależna
 - 171 predyktorów (116 predyktory ilościowe, 55 predyktorów jakościowych)
 - Identyfikator zbioru
- Zrównoważony rozkład zmiennej zależnej

Zrozumienie i przygotowanie danych

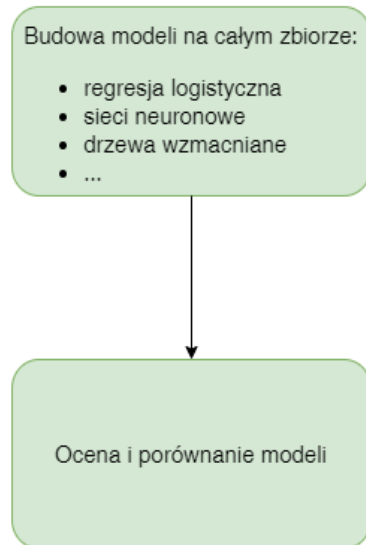
- Analiza braków danych oraz ich uzupełnienie
- Segmentacja zbioru danych
- Dodatkowa kategoryzacja zmiennych jakościowych
- Losowy podział zbioru danych na zbiory uczący, testowy oraz walidacyjny
- Przygotowanie zmiennych pochodnych
 - Wiedza ekspercka
 - Podejście automatyczne za pomocą Losowego Lasu
- Wstępna selekcja zmiennych
 - Siła predykcyjna
 - Wzajemna korelacja
- Transformacja zmiennych

Zrozumienie i przygotowanie danych

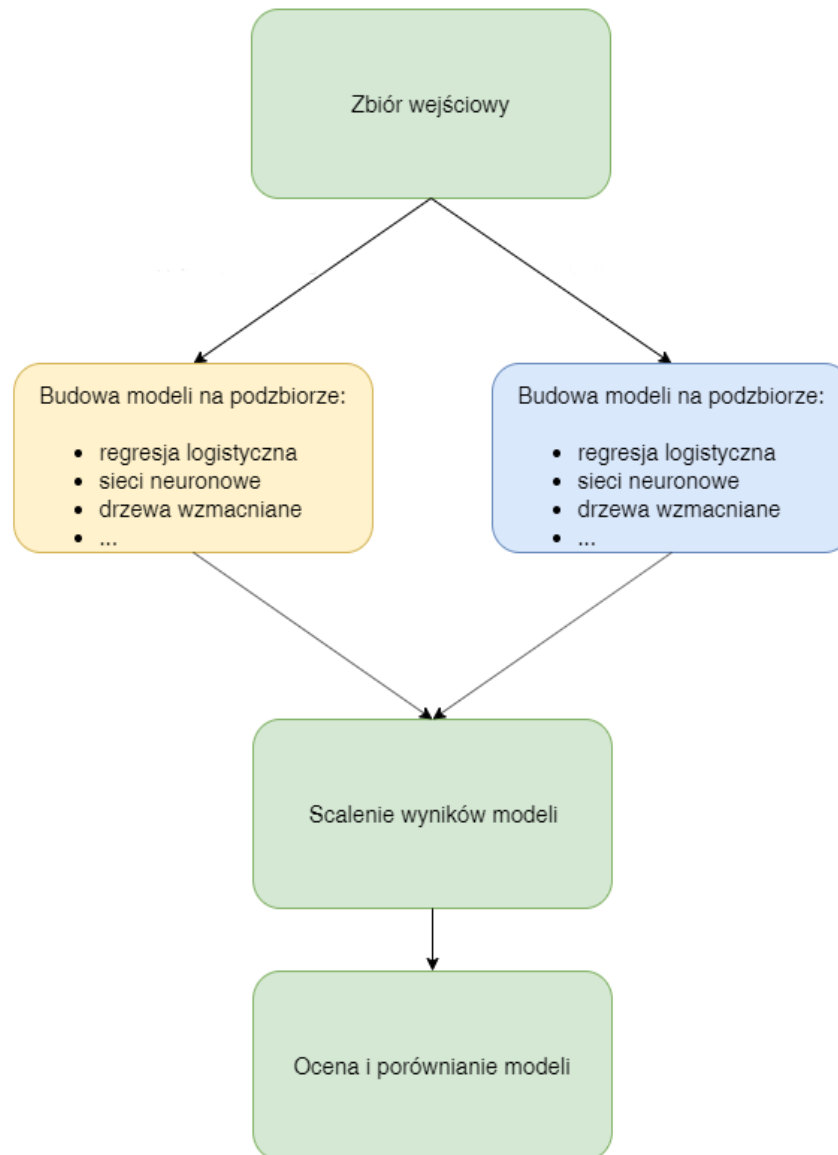
- Analiza braków danych oraz ich uzupełnienie
- **Segmentacja zbioru danych**
- Dodatkowa kategoryzacja zmiennych jakościowych
- Losowy podział zbioru danych na zbiory uczący, testowy oraz walidacyjny
- **Przygotowanie zmiennych pochodnych**
 - Wiedza ekspercka
 - Podejście automatyczne za pomocą Losowego Lasu
- Wstępna selekcja zmiennych
 - Siła predykcyjna
 - Wzajemna korelacja
- **Transformacja zmiennych**

Segmentacja zbioru danych

Ścieżka bez segmentacji



Ścieżka hybrydowa (z segmentacją)



Procedura badawcza – proces analizy

- Segmentacja statystyczno-ekspercka
 - Wykorzystanie drzew CART
 - Podział na podzbiory
 - Możliwość hybrydyzacji
- Przygotowanie zmiennych pochodnych
 - Metoda automatyczna oparta na metodzie Losowy las
 - 20 zmiennych reprezentujących najsilniejsze reguły o dostatecznej powszechności
- Transformacje zmiennych
 - Standaryzacja logistyczna Pyle'a
 - Standaryzacja WoE na podstawie decyli
 - Standaryzacja WoE na podstawie drzew CART

Procedura badawcza – proces analizy

Numer zbioru danych	Standaryzacja	Zmienne pochodne	Segmentacja
1.	Brak	Nie	Nie
2.	Pyle	Nie	Nie
3.	WoE – Decyle	Nie	Nie
4.	WoE – CART	Nie	Nie
5.	Brak	Tak	Nie
6.	Pyle	Tak	Nie
7.	WoE – Decyle	Tak	Nie
8.	WoE – CART	Tak	Nie
9.	Brak	Nie	Tak
10.	Pyle	Nie	Tak
11.	WoE – Decyle	Nie	Tak
12.	WoE – CART	Nie	Tak
13.	Brak	Tak	Tak
14.	Pyle	Tak	Tak
15.	WoE – Decyle	Tak	Tak
16.	WoE – CART	Tak	Tak

Procedura badawcza – proces analizy

- Zbadanie 16 wariantów zbiorów
- Dla każdego z wariantów
 - Optymalizacja hiperparametrów
 - Agregacja modeli w obrębie danej metody

Procedura badawcza - metody

- Punkt odniesienia – model ekonometryczny (regresja logistyczna)
 - Metodyka przyjęta w zastosowaniach biznesowych
 - Dyskretyzacja predyktorów
 - Standaryzacja do wartości WoE (*Weight of Evidence*)
- Modele porównawcze
 - Drzewa klasyfikacyjne CART
 - Drzewa wzmacniane (biblioteka *xgboost*)
 - Sieci neuronowe (perceptron wielowarstwowy)

Agenda

Uzasadnienie wyboru problemu badawczego

Charakterystyka obszaru badań i luka badawcza

Cel pracy, pytania badawcze

Procedura badawcza

Uzyskane wyniki



Regresja logistyczna

- Brak przyjętej praktyki optymalizacji hiperparametrów
- Dodatkowe kryteria selekcji modelu
 - Maksymalna liczba zmiennych 15
 - Zgodność ocen parametrów regresji z ocenami dla modeli jednoczynnikowych
 - Poziom współliniowości predyktorów $VIF \leq 2,5$
- Poszukiwanie optymalnego sposobu selekcji zmiennych

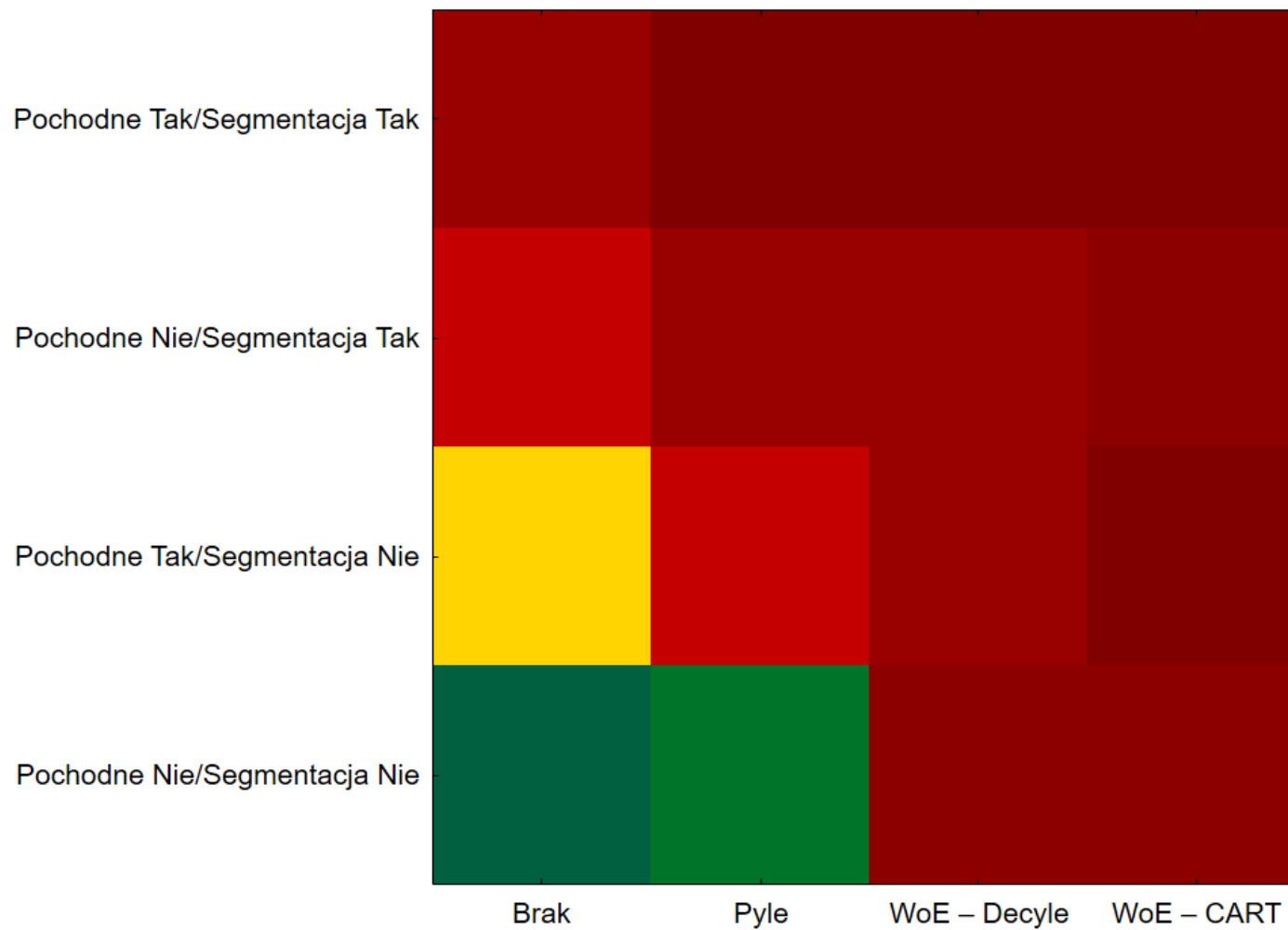
Regresja logistyczna

- Zwycięska strategia wielostopniowej selekcji (cel poboczny pracy)
 - Krok pierwszy: Filtr ukierunkowany na podstawie wartości informacyjnej (IV), wskaźnika Giniego oraz metody ReliefF
 - Krok drugi: Metoda Lasso
 - Krok trzeci: Metoda *Branch&Bound*
- Agregacja modeli nie poprawia uzyskanych wyników

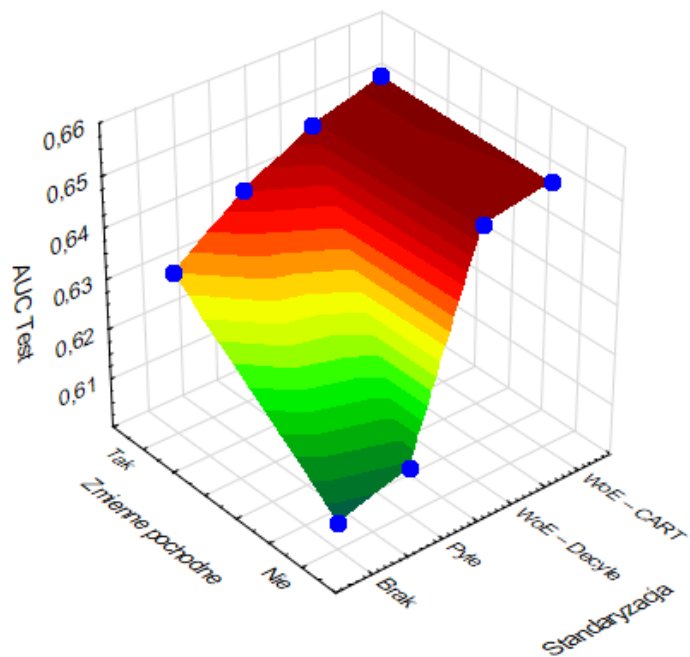
Procedura badawcza – proces analizy

Numer zbioru danych	Standaryzacja	Zmienne pochodne	Segmentacja
1.	Brak	Nie	Nie
2.	Pyle	Nie	Nie
3.	WoE – Decyle	Nie	Nie
4.	WoE – CART	Nie	Nie
5.	Brak	Tak	Nie
6.	Pyle	Tak	Nie
7.	WoE – Decyle	Tak	Nie
8.	WoE – CART	Tak	Nie
9.	Brak	Nie	Tak
10.	Pyle	Nie	Tak
11.	WoE – Decyle	Nie	Tak
12.	WoE – CART	Nie	Tak
13.	Brak	Tak	Tak
14.	Pyle	Tak	Tak
15.	WoE – Decyle	Tak	Tak
16.	WoE – CART	Tak	Tak

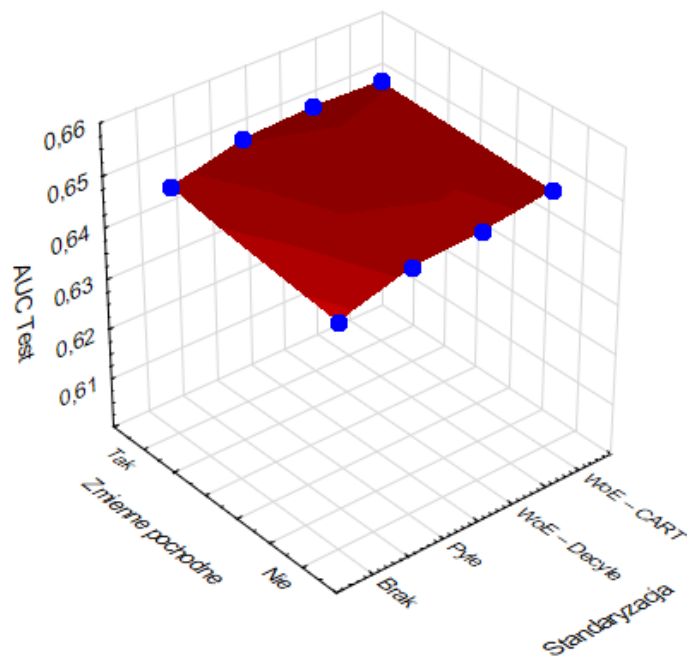
Regresja logistyczna – AUC na zbiorze testowym



Regresja logistyczna – AUC na zbiorze testowym



Segmentacja: Nie

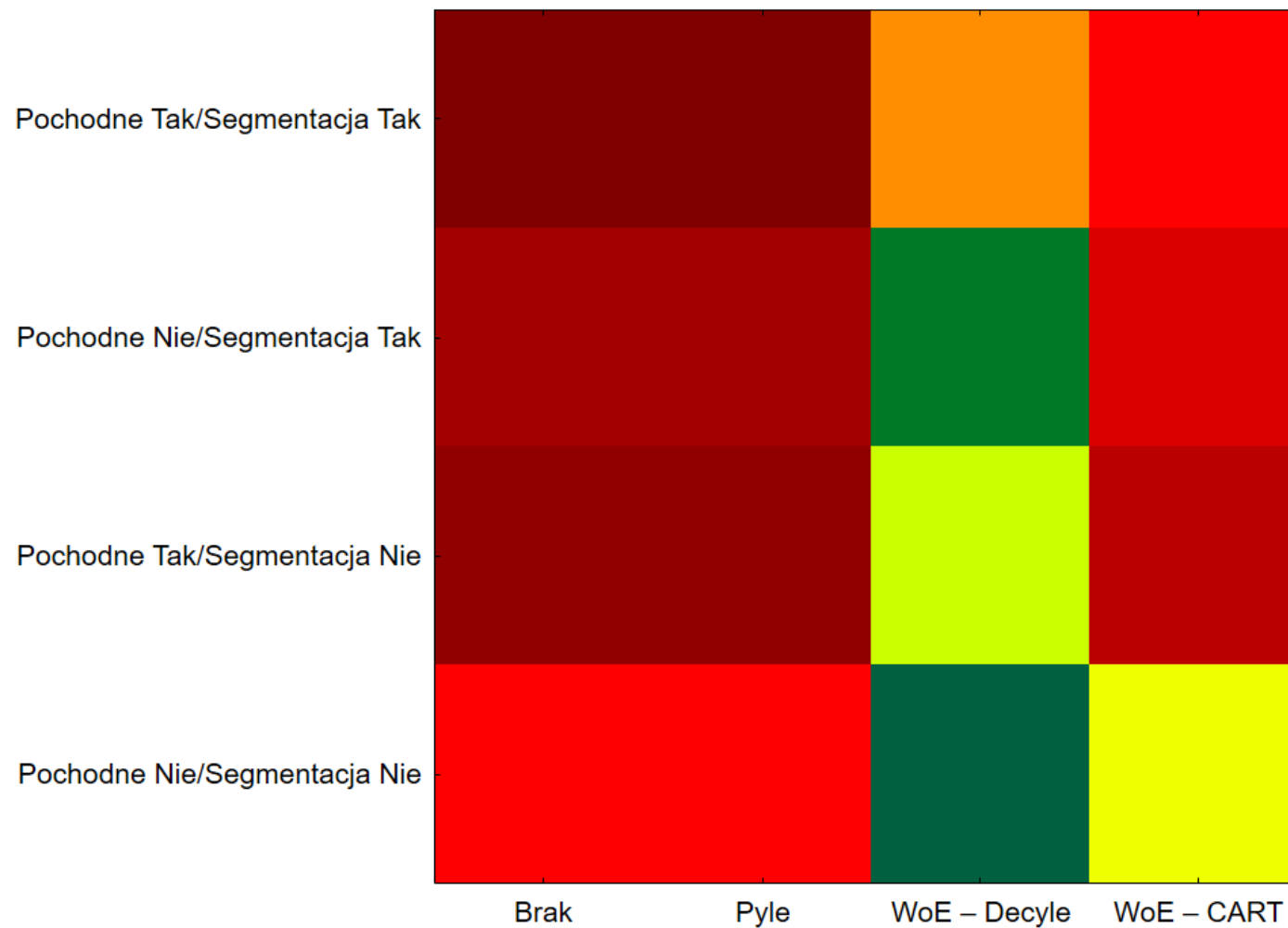


Segmentacja: Tak

Drzewa CART

- Wieloetapowy proces optymalizacji hiperparametrów
- Dla każdego wariantu zbioru 900 modeli (ponad 14 tys. modeli)
- Próba przycinania zwycięskich modeli

Drzewa CART – AUC na zbiorze testowym



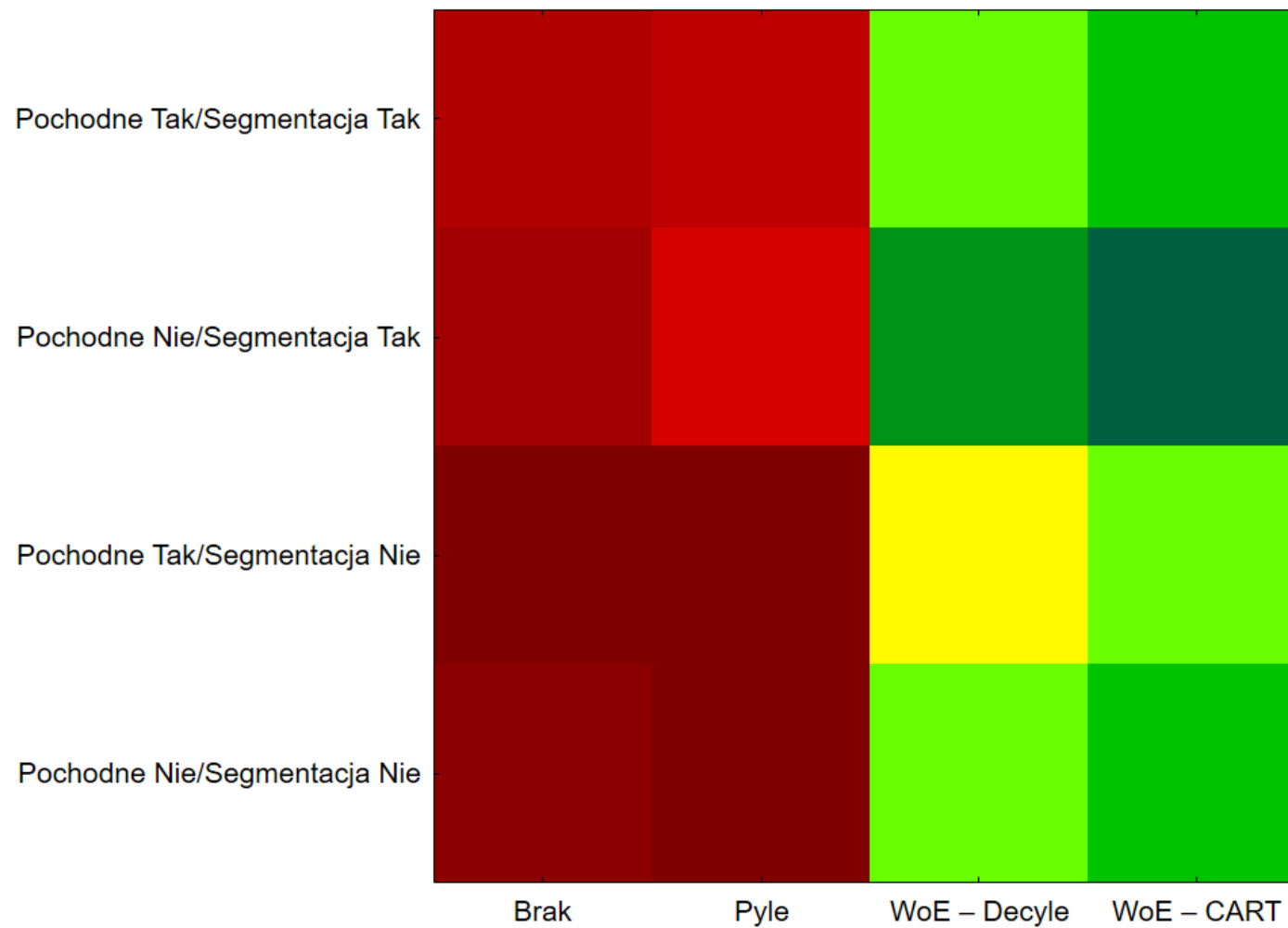
Drzewa CART

- Optymalizacja hiperparametrów znacząco poprawia dobroć dopasowania
- Dodanie zmiennych pochodnych oraz segmentacja spowodowały podniesienie się przeciętnego poziomu dopasowania
- Dyskretyzacja zmiennych pogorszyła jakość uzyskanych wyników
- Agregacja modeli poprawia jakość uzyskanych wyników
- Testowanie alternatywnych podziałów drzewa nie poprawiło jakości budowanych modeli (cel poboczny)

Drzewa wzmacniane

- Optymalizacja hiperparametrów metodą losową – 1000 modeli dla każdego wariantu zbioru

Drzewa wzmacniane – AUC na zbiorze testowym



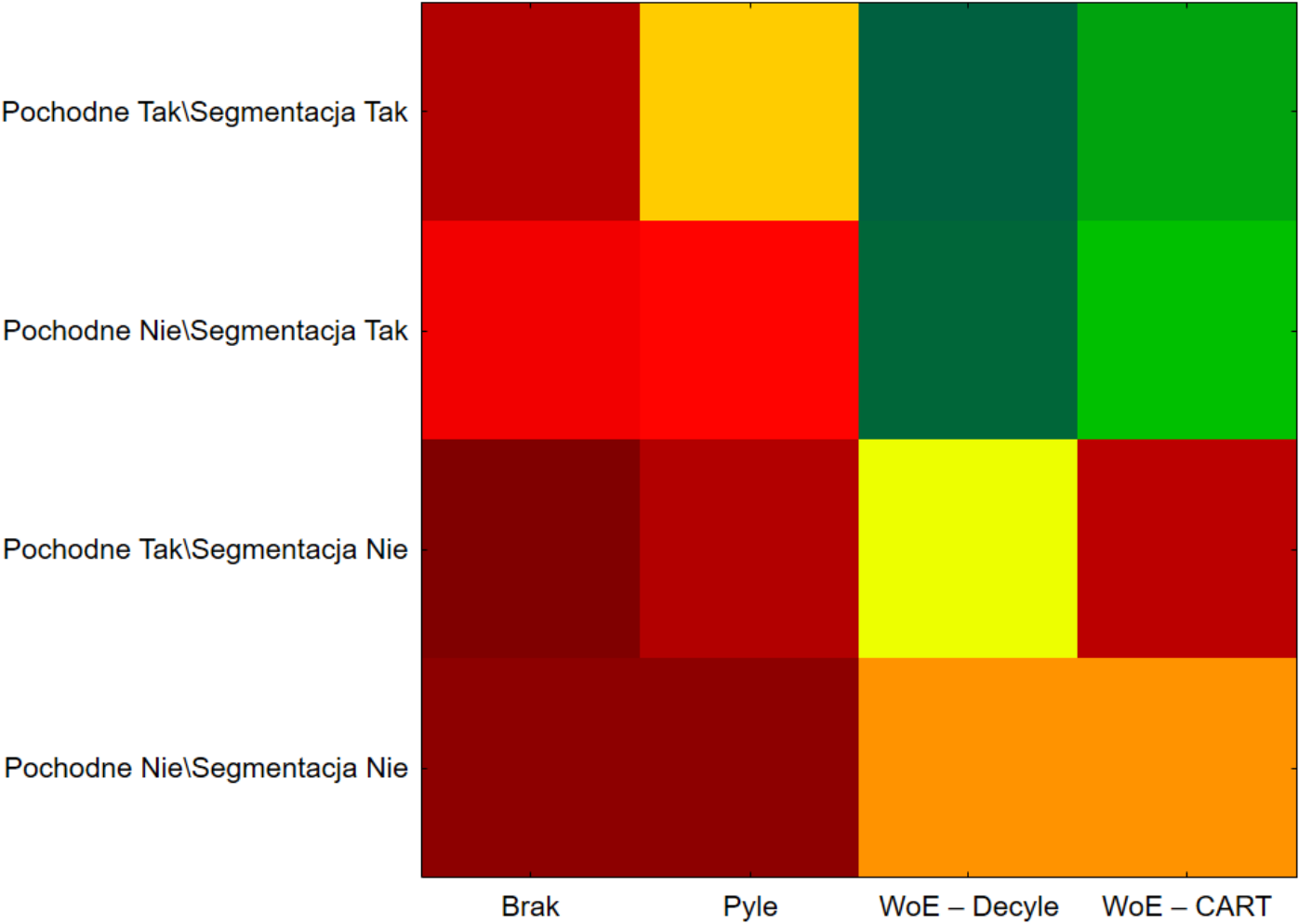
Drzewa wzmacniane

- Optymalizacja hiperparametrów znacząco poprawia dobroć dopasowania
- Jedynie niewielka część hiperparametrów ma wpływ na poprawę siły predykcyjnej
- Agregacja modeli poprawia jakość uzyskanych wyników

Sieci neuronowe

- Kilkustopniowy proces identyfikacji hiperparametrów
- Optymalizacja hiperparametrów metodą losową – 1000 modeli dla każdego wariantu zbioru
- Zawężenie zakresu hiperparametrów – analiza CART
- Kolejne 1000 modeli dla zawężonych ustawień
- Selekcja 100 najlepszych modeli i zwiększenie liczby epok dla ich ustawień

Sieci neuronowe



Sieci neuronowe

- Optymalizacja hiperparametrów znacząco poprawia dobroć dopasowania
- Segmentacja zbioru oraz standaryzacja nie powoduje poprawy dobroci dopasowania
- Agregacja modeli poprawia jakość uzyskanych wyników

Wpływ aspektów TIHVE na wybrane metody analizy

Metoda analizy	Transformacja zmiennych	Zmienne pochodne	Optymalizacja hiperparametrów	Segmentacja zbioru danych	Agregacja modeli
Regresja logistyczna	+	+	—	+	—
Drzewa CART	—	+	+	+	+
Drzewa wzmacniane	—	+	+	—	+
Sieci neuronowe	—	+	+	—	+

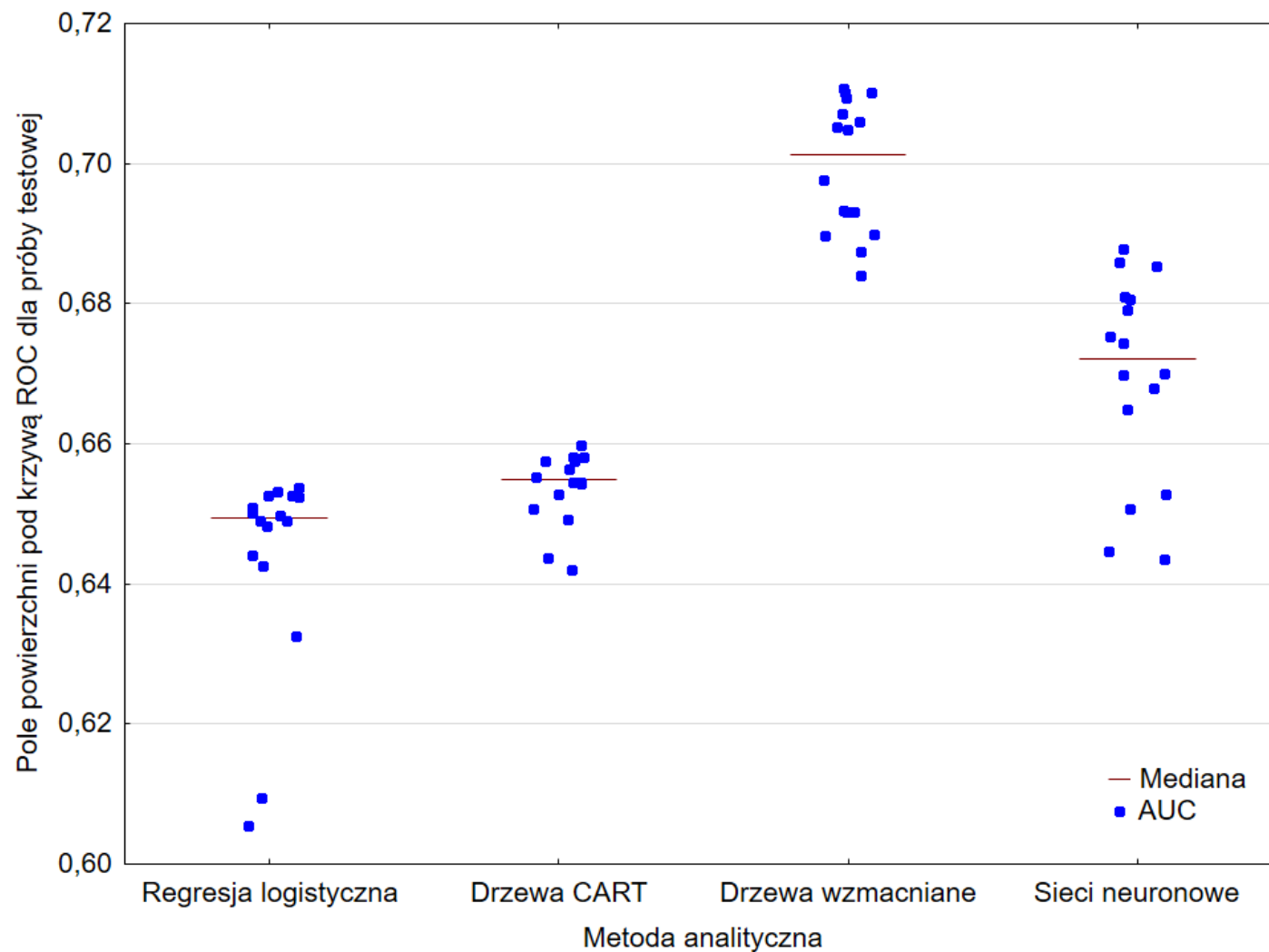
Podsumowanie - Wybrane wnioski

- Zaobserwowano nierównomierny wpływ podjętych działań optymalizacyjnych na wybrane metody
- W przypadku modelu ekonometrycznego kluczowe okazały się
 - Segmentacja zbioru
 - Dyskretyzacja zmiennych
- Dla pozostałych metod kluczowa jest
 - Optymalizacja hiperparametrów
 - Agregacja modeli
- Dodanie zmiennych pochodnych poprawia siłę predykcyjną dla wszystkich modeli

Podsumowanie - Cel poboczny

- Wykazanie możliwości budowy modeli o zadowalających własnościach za pomocą metod interpretowalnych przez badacza (biała skrzynka), porównywalnych z zaawansowanymi metodami nieinterpretowalnymi (czarna skrzynka), przy użyciu odpowiednich technik przygotowania danych oraz hybrydyzacji modeli

Podsumowanie - Wyniki zbiorcze



Grzegorz Migut

Tworzenie skutecznych modeli *Data Mining* na przykładzie prognozowania migracji klientów



Uczenie maszynowe. Czyszczenie danych, budowa i ocena modeli klasyfikacyjnych.

print pdf epub mobi

e edu-Libri



Spis treści

Wstęp	7
1. Retencja klientów jako obszar modelowania marketingowego	13
1.1. Rola modelowania w badaniach marketingowych	14
1.2. Koncepcja marketingu relacji w budowie modeli lojalności klientów	17
1.3. Lojalność nabywców – definicje i determinanty	23
1.4. Rodzaje i poziomy lojalności konsumentów	25
1.5. Sposoby pomiaru i modelowania lojalności – wybrane podejścia	34
1.6. Metodyki budowy modeli <i>data mining</i> na potrzeby retencji klientów	40
1.7. Podsumowanie	48
2. Przygotowanie danych podczas budowy modeli retencji klientów	50
2.1. Określanie kluczowych parametrów projektu analitycznego	50
2.2. Wiarygodność danych, identyfikacja i uzupełnianie braków danych	53
2.3. Sposoby łagodzenia problemów związanych z niejednorodnym zbiorem danych	58
2.4. Transformacje zmiennych oraz przygotowanie zmiennych pochodnych	61
2.5. Niezbilansowany rozkład zmiennej zależnej jako problem w prognozowaniu zjawisk o charakterze rzadkim	68
2.6. Podsumowanie	71
3. Budowa modelu klasyfikacyjnego	72
3.1. Wybór zmiennych podczas budowy modelu	73
3.1.1. Filtry ukierunkowane i nieukierunkowane	74
3.1.2. Metody opakowujące	81
3.1.3. Metody wbudowane	84
3.1.4. Podsumowanie	86
3.2. Wybór techniki modelowania	88
3.2.1. Regresja logistyczna	89
3.2.2. Sieci neuronowe	96
3.2.3. Metoda wektorów nośnych	106
3.2.4. Drzewa klasyfikacyjne i regresyjne	109
3.2.5. Podejście wielomodelowe	115
3.3. Optymalizacja hiperparametrów	124
3.4. Podsumowanie	128

Grzegorz Migut

Tworzenie skutecznych modeli *Data Mining* na przykładzie prognozowania migracji klientów



Uczenie maszynowe. Czyszczenie danych, budowa i ocena modeli klasyfikacyjnych.

print pdf epub mobi

e•edu-Libri



4. Walidacja i wdrażanie modeli retencji klientów	130
4.1. Miary dobroci dopasowania modeli retencji klientów	130
4.1.1. Miary obliczone na podstawie macierzy błędnych klasyfikacji	132
4.1.2. Miary jakości a prawdopodobieństwo <i>a posteriori</i>	134
4.1.3. Miary oparte na probabilistycznej interpretacji odpowiedzi modelu	155
4.2. Strategie walidacji modeli retencji klientów	155
4.2.1. Metody oparte na podziale zbioru danych	156
4.2.2. Metody oparte na wielokrotnym próbkowaniu	159
4.3. Określanie optymalnego punktu odcięcia (<i>cut-off</i>)	160
4.4. Wybór klientów do strategii sprzedażowych i retencyjnych (<i>uplift modeling</i>)	166
4.4.1. Budowa modeli <i>uplift</i>	167
4.4.2. Ocena modeli <i>uplift</i>	170
4.4.3. Monitorowanie modeli retencji klientów	174
4.5. Podsumowanie	175
5. Identyfikacja modelu migracji klientów krok po kroku	176
5.1. Zrozumienie i przygotowanie danych	177
5.1.1. Ocena jakości danych	178
5.1.2. Przygotowanie danych – kolejne kroki	180
5.1.3. Segmentacja zbioru danych	181
5.1.4. Podział zbioru danych	184
5.1.5. Uzupełnienie braków danych	186
5.1.6. Dodatkowa kategoryzacja zmiennych jakościowych	187
5.1.7. Zmienne pochodne w modelu	187
5.1.8. Wstępna selekcja zmiennych	189
5.1.9. Budowa modeli za pomocą wybranych narzędzi analitycznych	191
5.2. Model regresji logistycznej	194
5.3. Model drzew klasyfikacyjnych i regresyjnych <i>CART</i>	205
5.4. Model drzew wzmacnianych	218
5.5. Model sieci neuronowych (perceptron wielowarstwowy)	227
5.6. Wnioski końcowe	233
Zakończenie	236
Załącznik. Opis zmiennych wykorzystywanych w modelowaniu	239
Bibliografia	246



Dziękuję za uwagę!



Grzegorz Migut
g.migut@statsoft.pl



StatSoft Polska

Dziękujemy, że jesteś częścią naszej historii!