



AutoML

Praca domowa nr 2

Szymon Gut, Maciej Orłowski



Wczytywanie danych

1. W zmiennej celu wartości -1 zostały zamienione na 0

Korzyści:

- Algorytmy działały
- Łatwiejsza praca

2. Podział danych na zbiór treningowy i zbiór walidacyjny (80% / 20%)



1. Podejście manualne



Preprocessing

1. **Usunięcie 25% najmniej skorelowanych ze zmienną celu kolumn**
Korelacja Spearmana, liczona na zbiorze treningowym
2. **Imputacja brakujących danych**
Średnie wartości danej zmiennej
3. **Redukcja wymiarowości metodą PCA**
Jako optymalna znaleziona wartość 170 komponentów
4. **Skalowanie min-max**



Metody

1. **XGBoost**

Bayes search na 100 iteracji

2. **Random Forest**

Bayes search na 100 iteracji

3. **MLP**

Trening na 10,000 epok



Wyniki

Model	Balanced accuracy na zbiorze:	
	Walidacyjnym	Testowym*
XGBoost	0.7704	0.9000
Random Forest	0.6372	0.8333
MLP	0.6832	-

* wyniki na udostępnionej przez prowadzących próbce danych testowych



2. Podejście AutoML



Metody

1. MLJAR

- Tryb: Compete
- Limit czasu: 24h

2. AutoGluon

- Presets: High Quality
- Limit czasu: 24h



Wyniki

Framework	Balanced accuracy na zbiorze:	
	Walidacyjnym	Testowym*
MLJAR	0.8712	0.8333
AutoGluon	0.8709	0.7667

* wyniki na udostępnionej przez prowadzących próbce danych testowych



Dziękujemy za uwagę!

Szymon Gut, Maciej Orłowski