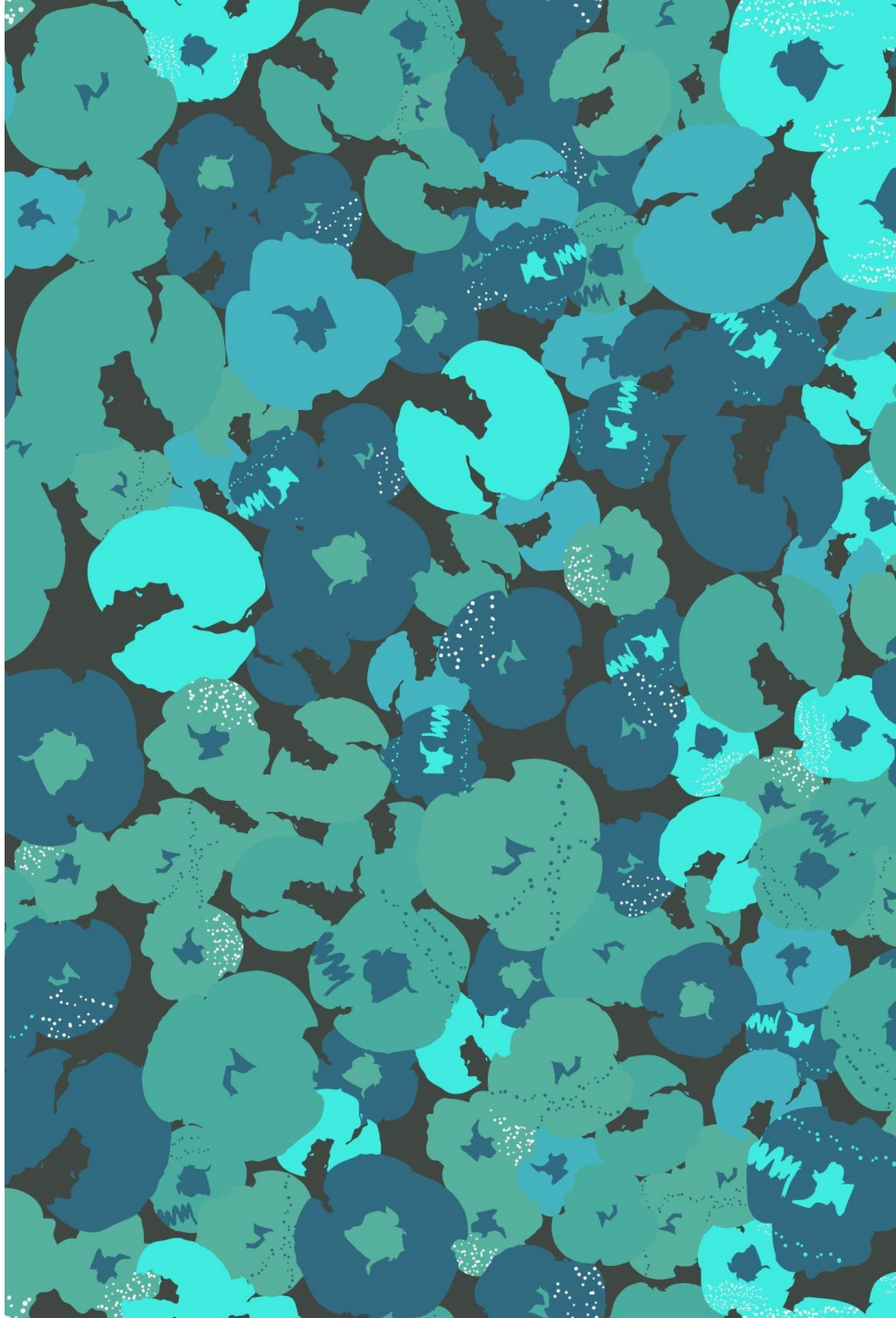
AutoML: Homework 2

JAN CICHOMSKI

ADAM DĄBKOWSKI



Zbiór danych

Problem: Klasyfikacja binarna

Cechy zbioru danych:

- 500 kolumn
- Dane numeryczne, całkowite
- Nie ma braków danych
- 2000 obserwacji w zbiorze treningowym
- 600 obserwacji w zborze testowym

Preprocessing danych

- Podział danych: 80% zbiór treningowy, 20% zbiór testowy
- Metody eliminacji nieznaczących kolumn:
  - RFE – Recursive Feature Elimination
  - RFECV - Recursive Feature Elimination with Cross-Validation
- Estymatory: ExtraTreesClassifier, RandomForestClassifier
- Redukcja liczby cech z 500 do 20

Autogluon

Konfiguracja:

- presets="best_quality"
- time_limit=8h
- hyperparameters="default"
- fit_weighted_ensemble=True
- fit_full_last_level_weighted_ensemble=True
- full_weighted_ensemble_additionally=True
- num_bag_folds=15
- num_bag_sets=25
- num_stack_levels=3
- auto_stack=True
- dynamic_stacking=True
- feature_generator="auto"

Najlepszy model: WeightedEnsemble_L5

Wyniki:

- Zbiór treningowy: 0.94065
- Zbiór walidacyjny: 0.88982
- Udostępniony zbiór testowy: 0.9333

Model stworzony ręcznie

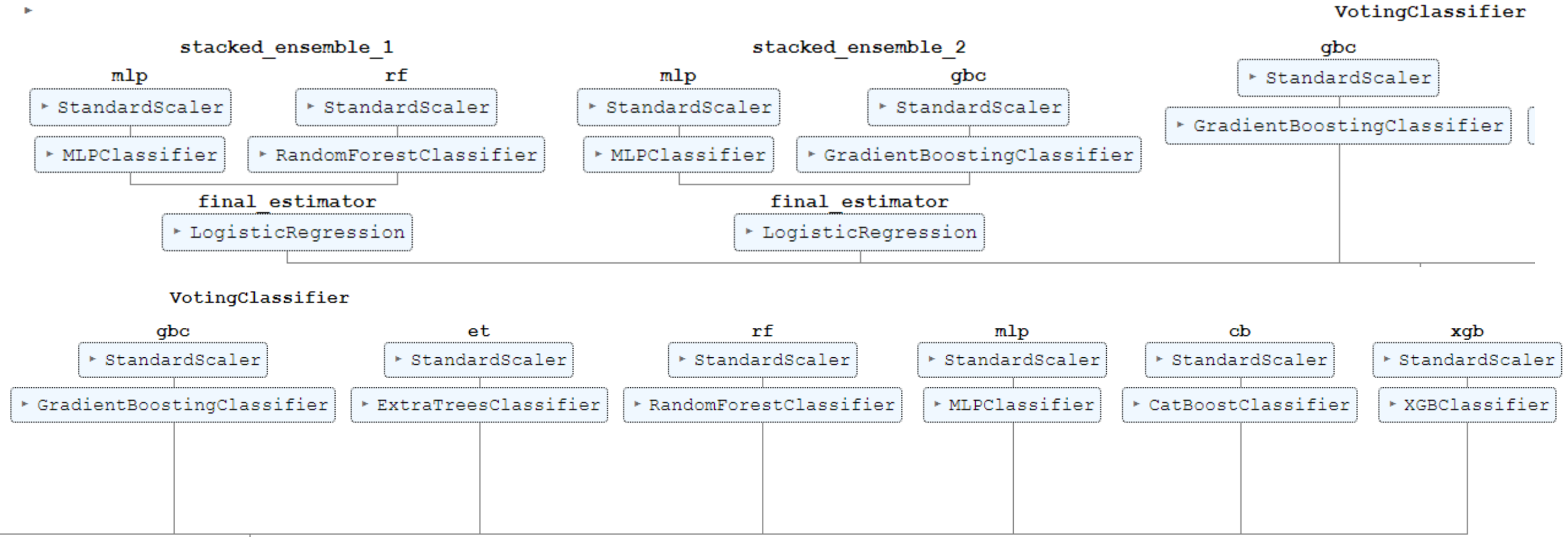
VotingClassifier + StandardScaler:

- StackingClassifier: MLP + Random Forest
- StackingClassifier: MLP + Gradient Boosting
- Gradient Boosting
- Extra Trees
- Random Forest
- Cat Boost
- XG Boost

Wyniki:

- Zbiór treningowy: 1.0
- Zbiór walidacyjny: 0.89998
- Udostępniony zbiór testowy: 0.9

Schemat ręcznie stworzonego modelu



Wnioski

- Wynik modelu stworzonego manualnie jest porównywalny z wynikiem frameworka
- Redukcja kolumn z 500 do 20 wpłynęła pozytywnie na prędkość uczenia modeli

Dziękujemy za uwagę