Auto ML
HW 2

Sabina Sidarovich, Paweł Gelar

Dane

- 500 kolumn
- zbiór treningowy zawiera 2000 obserwacji
- zbiór testowy zawiera 600 obserwacji
- wartości numeryczne
- nie ma braków danych
- binarny target
- obserwacje są równo rozdystrybuowane
- wszystkie zmienne mają niezerową wariancję

Modele klasyczne

Preprocessing danych

Wykorzystano algorytm Boruta do wybrania najbardziej znaczących cech.

Testowano również inne metody doboru cech:

- SelectKBest
- Mutual Information
- Recursive Feature Elimination
- Variance Threshold
- VIF (Variance Inflation Factor)
- Outlier Detection

Przetestowane modele

- XGBoost
- LightGBM
- Catboost

Najlepszym modelem po optymalizacji hiperparametrów okazał się LightGBM, lecz XGBoost miał bardzo podobne wyniki.

XGBoost	86.0%	88.6%
LightGBM	85.2%	88.8%
Catboost	86.2%	87.0%

Optymalizacja

Hiperparametry uczenia gradientowego zoptymalizowaliśmy za pomocą biblioteki Optuna.

Dla każdego z modeli została zdefiniowana przestrzeń hiperparametrów.

Podczas optymalizacji używaliśmy kroswalidacji.

Modele automatyczne

Przetestowaliśmy następujące frameworki:

- Tab PFN
- Ml Jar
- AutoSklearn 2.0

Tab PFN

Niestety framework Tab PFN ma maksymalną ilość cech równą 100, więc nie było możliwe skorzystanie z niego.

Ml Jar

Używaliśmy trybu Compete, zadziałał bezproblemowo.

Wyniki utrwalone w bardzo przejrzystej formie.

AutoSklearn 2.0

Również zadziałał bezproblemowo.

Dużym atutem był znajomy interfejs.

Wyniki

Metoda	Dokładność
LightGBM (domyślne parametry)	85%
LightGBM (parametry zoptymalizowane)	89%
Auto scikit-learn	91%
MI Jar	87%

Wnioski

- dane wejściowe wymagały odpowiedniego przygotowania, które nie tylko znacząco polepszyło wyniki algorytmów, ale również wielokrotnie przyspieszyło obliczenia
- wyniki dla podejścia manualnego i automatycznego są dość zbliżone, jednak model manualny osiągnął nieco lepsze rezultaty kosztem czasu jego przygotowania.
- dopracowywanie tradycyjnych modeli może być procesem czasochłonnym i może stanowić wyzwanie dla użytkowników nieposiadających dostatecznej wiedzy w danej dziedzinie

Dziękujemy za uwagę