

Warsztaty Badawcze - AutoML

Praca domowa 3 - code review grupy Gakubu

Michał Tomczyk

1 Wprowadzenie

Celem pracy domowej jest ocena rozwiązania zaprezentowanego przez grupę Gakubu w ramach Kamienia Milowego 2, w którym grupa miała za zadanie stworzyć funkcję przygotowującą zbiór danych do pracy z Autogluon.

Całe rozwiązanie jest zawarte w jednym pliku `utilis.py`, a właściwa funkcja to `make_kfold_cross_valiation()`.

2 Czy ten kod osiąga cel, który postawiono?

W ramach przygotowania zbioru danych funkcja odczytuje zbiór danych z adresu url podanego w argumencie, dekoduje kolumny na typ `utf-8`, dzieli zbiór na kolumny objaśniające i kolumnę objaśnianą, dokonuje podziału na foldy i wykonuje funkcję `TabularPredictor()` z pakietu Autogluon.

W zadaniu było stwierdzone, że funkcja ma przyjmować zbiór danych X oraz kolumnę objaśnianą y. Funkcja jednak przyjmuje adres url zbioru danych. Po zmianie sposobu odczytywania danych funkcja będzie całkowicie spełniać postawiony cel, gdyż dobrze przygotowuje zbiór do pracy z Autogluon.

3 Czy w kodzie są jakieś oczywiste błędy logiczne?

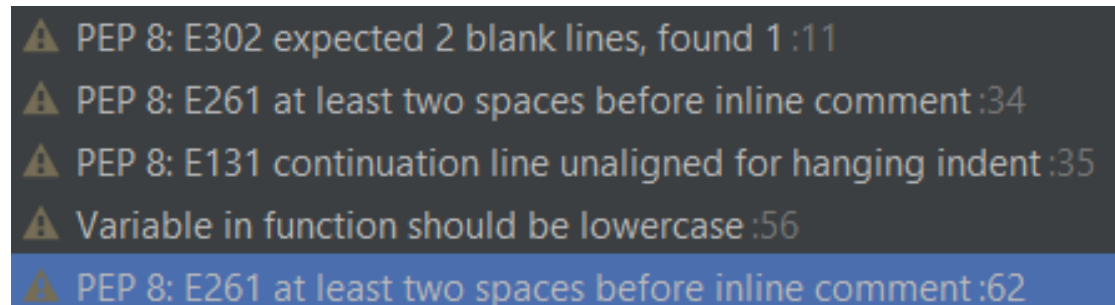
Nie dopatrzyłem się oczywistych błędów logicznych w kodzie.

4 Czy patrząc na wymagania zawarte podczas prezentacji są one w pełni zaimplementowane?

Tak, oprócz wywoływania funkcji bezpośrednio za pomocą zbioru danych i kolumny objaśniającej.

5 Czy kod jest zgodny z istniejącymi wytycznymi stylistycznymi?

Większość kodu jest poprawiana pod względem stylistycznym. Znalezione przez Pylcharm rozbieżności z PEP 8:



The image shows a screenshot of a code editor with five linting errors from Pylcharm. Each error is preceded by a yellow warning triangle icon. The errors are: 'PEP 8: E302 expected 2 blank lines, found 1 :11', 'PEP 8: E261 at least two spaces before inline comment :34', 'PEP 8: E131 continuation line unaligned for hanging indent :35', 'Variable in function should be lowercase :56', and 'PEP 8: E261 at least two spaces before inline comment :62'. The last error is highlighted with a blue background.

```
⚠ PEP 8: E302 expected 2 blank lines, found 1 :11
⚠ PEP 8: E261 at least two spaces before inline comment :34
⚠ PEP 8: E131 continuation line unaligned for hanging indent :35
⚠ Variable in function should be lowercase :56
⚠ PEP 8: E261 at least two spaces before inline comment :62
```

6 Czy są jakieś obszary, w których kod mógłby zostać poprawiony?

Należy usunąć linijki odpowiedzialne za import pakietów, które nie są wykorzystane (`numpy`, `train_test_split` z `sklearn`). `Autogluon` został importowany dwa razy - najpierw w całości (`import` ten nie został użyty), następnie zaimportowane zostały dwie funkcje z tego pakietu (przy czym tylko jedna została użyta)

7 Czy dokumentacja i komentarze są wystarczające?

Dokumentacja i komentarze są w odpowiedniej ilości. Nie ma problemu w zrozumieniu co dzieje się w danym miejscu w kodzie.

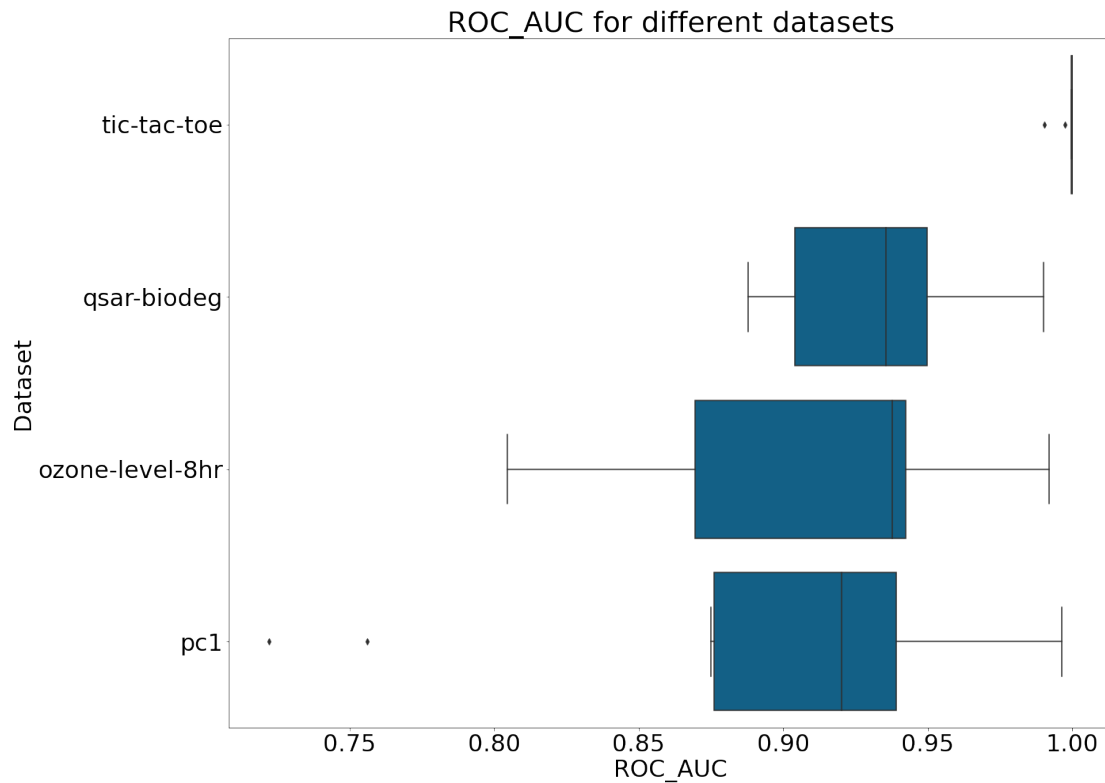
8 Czy udało się odtworzyć zamieszczone przykłady w kodzie?

Tak, udało się odtworzyć cały notatnik ipnyb z przykładami, generując takie same rezultaty.

9 Czy udało się użyć przygotowanych kodów na nowych danych?

Sprawdzamy działanie funkcji na czterech wybranych zbiorach z OpenML (innych niż te użyte w testach przez grupę Gakubu).

Jak widać, dla każdego ze zbiorów funkcja wywołała się poprawnie, uzyskując dobre wyniki ROC AUC.



10 Podsumowanie

- Kod (co najważniejsze) działa, czyli pozwala przygotować zbiór danych i wykonać na nim funkcję `TabularPredictor` z pakietu `Autogluon`
- Głównym mankamentem jest zła struktura funkcji - przyjmowanie adresu url zamiast faktycznego zbioru danych
- Kod jest napisany w sposób estetyczny, odpowiednio skomentowany
- Udało się zarówno odtworzyć testy wykonane przez grupę Gakubu, jak i przetestować funkcję na kilku nowych zbiorach z pakietu `OpenML`.