

Raport walidacyjny

STELLAR CLASSIFICATION

7 maja 2024 roku

Wydział Matematyki i Nauk Informacyjnych

Maciej Momot, Filip Langiewicz

Spis treści

1. Wstęp
2. Motywacja
3. EDA
4. FE
5. Model

1. Wstęp

Niniejszy raport walidacyjny przygotowany został przez zespół numer siedem w składzie: Maciej Momot oraz Filip Langiewicz (zwani dalej walidatorami lub zespołem walidacji). Walidowanym zespołem był zespół numer cztery, składający się z Małgorzaty Mokwy oraz Sebastiana Pergały (zwany dalej zespołem budowy). Obszarem walidacji był problem uczenia maszynowego dotyczący predykcji rodzajów obiektów astronomicznych (takich jak m.in. galaktyki czy gwiazdy) na podstawie danych udostępnionych publicznie na platformie internetowej Kaggle (źródło: <https://www.kaggle.com/datasets/fedesoriano/stellar-classification-dataset-sdss17>). Jako zespół walidacji, w celu weryfikacji jakości rozwiązań proponowanych przez zespół budowy, po każdym kamieniu milowym analizowaliśmy obecne na tamten moment fragmenty kodu oraz podejście biznesowe i logiczne do postawionego problemu. Dodatkowo uruchomiliśmy wybrany model na specjalnie do tego przygotowanej próbce danych, która nie była dostępna dla zespołu budowy. Stanowiła ona około 30% całego zbioru.

2. Motywacja

Celem biznesowym zespołu budowy było „stworzenie modelu, który jak najlepiej przewidzi należenie obiektu do jednej z trzech kategorii: Galaxy, Star, Quasar”.

3. EDA – Eksploracyjna Analiza Danych

Pierwszym etapem jaki przyszło nam walidować była eksploracyjna analiza danych. Powszechnie wiadomo, że jest to tak naprawdę jeden z ważniejszych etapów pracy z modelem uczenia maszynowego oraz że poświęca mu się najwięcej czasu. Weryfikowaliśmy w nim czy zespół budowy poprawnie rozumie dane oraz czy wykorzystuje w tym celu poprawne i powszechnie stosowane taktyki (odpowiednie typu wykresów, analiza poszczególnych kolumn, sprawdzanie wartości brakujących i ich uzupełnianie). Bardzo duży nacisk położyliśmy też na zgodność z obowiązującym prawem europejskim, dokładniej z Aktem o Sztucznej Inteligencji (zwanym potocznie „AI Actem”). Nie zauważyliśmy niezgodności na tym polu, co warto odnotować jako plus w kierunku zespołu budowy. Pozytywnie też odebraliśmy szeroki zakres przeprowadzonej analizy oraz jej dokładność i jakość – nie stosowano np.

wykresów kołowych, a w ich miejsce używane były dużo lepsze do tego wykresy słupkowe. Przenalizowane zostały wszystkie kolumny, zarówno pod względem brakujących/odstających wartości, jak i pod względem ewentualnych korelacji pomiędzy nimi. Na tym etapie projektu zespół walidacji nie miał podstaw by zatrzymywać w pracy zespół budowy, ponieważ nie znalazł nic do poprawy. Jedynym zaproponowanym usprawnieniem było wczytywanie danych z pliku przez ścieżkę względną zamiast bezwzględnej, co zostało zaakceptowane przez zespół budowy.

4. FE – Inżynieria Cech

Drugi kamień milowy polegał na walidacji kodu odpowiedzialnego za inżynierię cech w proponowanym przez zespół budowy rozwiązaniu. Na tym etapie zespół walidacji zauważył, że do treningu każdego z modelu nie jest wykorzystywane standardowe podejście powiązane z utworzeniem ramek danych: train, test, valid. Została zwrócona na to uwaga zespołowi budowy, który to uwagę tę zaakceptował. Fragmenty odpowiedzialne za to zostały usunięte, a główna logika skryptu poprawiona o moment tworzenia odpowiednich ramek danych, które potem mogły być wykorzystywane przez zespół walidacji. Ponownie uwagę zwróciło profesjonalne podejście, dbałość o szczegóły i jakość zaproponowanych rozwiązań. W pliku mieściło się wiele pomysłów, które od razu były testowane i sprawdzana była ich jakość. Zdecydowanie widać było nieszablonowe podejście i poszukiwanie najlepszych możliwych kombinacji cech, w celu uzyskania najlepszej predykcyjności. Ponownie więc nie mogliśmy zanegować w żadnym stopniu pracy wykonanej przez zespół budowy, ponieważ wyróżniała się na tle innych prac.

5. Model

Ostatnim krokiem w naszej walidacji była weryfikacja modeli zaproponowanych przez zespół budowy. Naszym zadaniem było sprawdzenie wyników na danych testowych, do których zespół budowy nie miał dostępu. Miało to na celu sprawdzenie rozwiązań na innych, odmiennych danych, co pozytywnie powinno wpłynąć na dobór konkretnego modelu. Po uruchomieniu programu pozytywnie zostały zweryfikowane wszystkie zaproponowane rozwiązania, ponieważ wyniki uzyskane na zbiorze testowym nie różniły się znacząco od wyników uzyskanych na zbiorze walidacyjnym. Świadczy to o tym, że dobrany model dobrze działa dla losowych danych i faktycznie nadaje się do wdrożenia w celach przemysłowych i biznesowych.