Raport walidacyjny

ACTIVITY DETECTION

4 czerwca 2024 roku

Wydział Matematyki i Nauk Informacyjnych

Maciej Momot, Filip Langiewicz

Spis treści

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 pięć, składający się z Zuzanny Sieńko oraz Huberta Sobocińskiego (zwany dalej zespołem budowy). Obszarem walidacji był problem uczenia maszynowego dotyczący klasteryzacji danych związanych z aktywnościami fizycznymi (takich jak bieganie, jazda na rowerze) na podstawie danych udostępnionych publicznie na platformie internetowej Kaggle (źródło: https://www.kaggle.com/datasets/luisomoreau/activity-detection). 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 wszystkie rozwiązania 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 w jak najlepszy sposób znajdował będzie grupy aktywności w dostarczonych danych w celu wykorzystania zdobytej na tej podstawie wiedzy w systemach inteligentnych miast – smart cities.

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 oraz boxploty. Z odnotowanych mniejszych nieprawidłowości wymienić należy wczytywanie danych za pomocą ścieżki bezwzględnej. Dużą nieprawidłowością odnotowaną przez nas był brak usuwania wartości odstających ze zbioru danych widocznych na wykresach typu boxplot. Zarekomendowaliśmy zespołowi budowy zastąpienie wartości odstających kwantylem rzędu 0.95. W dostarczonym przez zespół budowy rozwiązaniu zauważyliśmy ogrom wykresów, którym brakowało jednak szczegółowego opisu (w tym brak informacji, co dany wykres przedstawia). Dodatkowo nie było z nich wyciągniętych żadnych wniosków. Podczas procesu zespół budowy zdecydował się również na zastosowanie do zbioru PCA, przed którym to jednak danych nie znormalizowano, co zostało zauważone przez zespół walidacji. Zaleciliśmy również zrobienie wstępnie wyjaśnialnej wariancji przed wyborem n_components, ponieważ brakowało tego w kodzie dostarczonym przez zespół budowy.

4. FE, Model

Kolejnym krokiem w naszej walidacji była weryfikacja etapu inżynierii cech i 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 oraz zapewnić skalowalność zaproponowanego rozwiązania również dla danych innych niż treningowe. Pierwszą uwagą jaką zgłosiliśmy było zaproponowanie zespołowi walidacji przemyślenie doboru liczby klastrów. Zespół budowy ustalił tę liczbę jako 3 bez poparcia tego przeznaczonymi do tego celu metrykami, które posiadają silne ugruntowanie matematyczne (metoda łokcia, Silhouette score). Mieliśmy również problemy z walidacją kodu dostarczonego nam przez zespół budowy. Natrafiliśmy na problemy z modułem umap oraz na kod, którego wykonanie wymagało alokacji w pamięci podręcznej 268 GiB danych. Na ten moment nie dysponujemy sprzętem, który umożliwiałby takie rozwiązanie. Po uruchomieniu programu pozytywnie zostały zweryfikowane wszystkie zaproponowane rozwiązania, ponieważ wyniki uzyskane na zbiorze testowym

Silhouette Score: 0.32079071134317194

Davies-Bouldin Index: 1.2540074249469335

nie różniły się od wyników uzyskanych na zbiorze treningowym i walidacyjnym.

Świadczy to o tym, że dobrany model dobrze działa dla losowych danych i faktycznie nadaje się do wdrożenia w celach przemysłowych i biznesowych. W celu weryfikacji poprawności zaproponowanego rozwiązania posłużyliśmy się metrykami: Silhouette score oraz Davies-Bouldin index. Jako zespół walidacji widzimy, że zespół budowy dołożył wszelkich starań, by doprowadzić do uzyskania jak najlepszego rozwiązania.