Raport - Congressional voting

Yevhenii Vinichenko, Krzysztof Wolny

April 2021

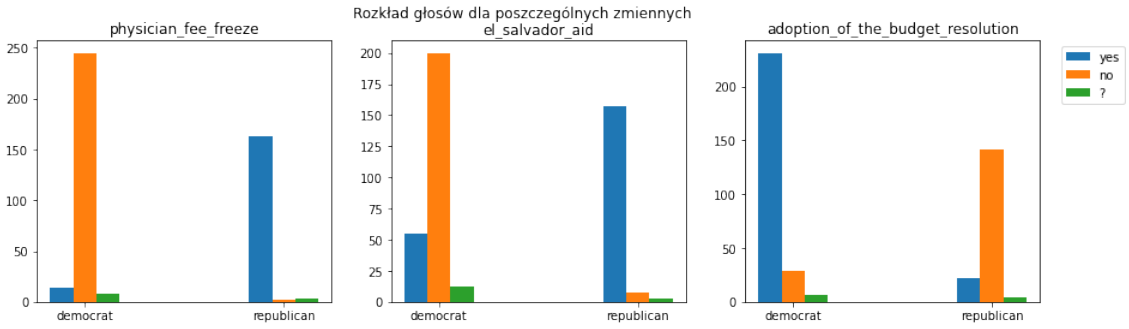
1 Opis problemu

Naszym celem było stworzenie modelu predykcyjnego, który przewidywałby na podstawie głosów kongresmana z Izby Reprezentantów Stanów Zjednoczonych, czy jest on demokratem, czy republikanem.

2 Opis zbioru danych

Otrzymaliśmy zbiór danych o głosach kongresmenów z Izby Reprezentantów Stanów Zjednoczonych. W danych były zapisane wyniki, czy dany kongresmen głosował za, przeciw, czy wstrzymał się od głosu w danej propozycji. Łącznie propozycji było 16. W dataframie litera 'y' oznaczała poparcie kongresmana dla tej propozycji, 'n' głos przeciwko, a '?' wstrzymanie się od głosu. W danych była również informacja, czy kongresman jest demokratem, czy republikanem.

Dane, które otrzymaliśmy, tak jak wynika z opisu, były dyskretne. Po przeanalizowaniu danych udało nam się znaleźć kilka propozycji, które w wyraźny sposób odróżniały republikanów od demokratów. Jest to oczywiście bardzo dobra informacja w kontekście tworzenia modeli.



3 Preproccessing

Z związku z tym, że nasze dane były dyskretne nie potrzebowaliśmy robić dużej ilości preproccessingu. Naszym głównym celem było zamienienie oznaczeń literowych na liczbowe, aby ułatwić obliczenia. Postanowiliśmy zamienić dane w następujący sposób:

3.1 Wcześniejsze błędy

W kamieniu milowym numer 2 traktowaliśmy '?' jako NaN i próbowaliśmy uzupełniać te dane. Nie jest to jednak najlepsze myślenie, ponieważ wstrzymanie się od głosów również jest częścią polityki. Były nawet takie propozycje, w których większość polityków wstrzymywało się od głosu. Modele statystyczne, które stworzyliśmy pracują również z większą skutecznością, gdy uznajemy '?' jako oddzielną zmienną.

4 Modele statystyczne

Stworzyliśmy 4 modele statystycznych:

- Random forest
- XGBoost
- Gradient boosting
- Logistic regrsiion

Aby znaleźć jak najlepsze parametry skorzystaliśmy z random search.

Najlepsze wyniki otrzymał XGBoost, a zaraz za nim regresja logistyczna. Są to modele na poziomie accuracy ok. 0.984 i AUC nawet 0.998.

Wyniki: (baseline jest to regresja logistyczna bez wykonania na niej random search)

	clf	accuracy	auc
0	baseline	0.984733	0.996164
1	rfc	0.977099	0.995908
2	xgb	0.984733	0.998210
3	gbc	0.977099	0.997442
4	lr	0.984733	0.993350

Sprawdziliśmy również jak skuteczne będą modele, jeśli usuniemy najbardziej skorelowana kolumnę z naszych danych. Okazuje się, że skuteczność modeli

znacząco spada, bo aż o co najmniej 0.05, a AUC o ok. 0.02-0.04 w zależności od modelu.

	clf	accuracy	auc
0	baseline	0.923664	0.971355
1	rfc	0.923664	0.964706
2	xgb	0.923664	0.966496
3	gbc	0.893130	0.954220
4	lr	0.923664	0.970588

5 Podsumowanie

Jesteśmy w stanie stworzyć model predykcyjny z ok. 98% dokładnością. Najlepiej jest użyć XGBoost lub regresji logistycznej.