Raport do projektu

Wstęp do Uczenia Maszynowego

Jakub Niemyjski

15 Styczeń 2024

1 Cel pracy

Celem projektu jest stworzenie modelu klasyfikacji binarnej o jak największej mocy predykcyjnej na przykładzie danych sztucznie wygenerowanych. Ocena efektywności modelu będzie się opierać na mierze zrównoważonej dokładności (ang. balanced accuracy).

2 Dane

Dane do projektu to sztucznie wygenerowany zbiór, który zawiera 30 zmiennych objaśniających, przy czym, jak się później okaże, część z nich jest zbędna do otrzymania dobrej mocy predykcyjnej pewnych modeli. Dostępne są następujące pliki:

- pełny zbiór treningowy: `artificial_train_data.csv`, który posiada 2000 obserwacji,
- zbiór etykiet dla pełnego zbioru treningowego: `artificial_train_labels.csv`,
- pełny zbiór testowy: `artificial_test_data.csv` zawierający 600 obserwacji.

Nie dysponujemy zatem etykietami dla zbioru testowego. Oznacza to, że ewaluacja modeli uczenia nadzorowanego na tym zbiorze jest niemożliwa.
Z tego względu, aby mieć możliwość jakiegokolwiek sprawdzenia poprawności tworzonych modeli, pełny zbiór treningowy obejmujący pliki `artificial_train_data.csv` oraz `artificial_train_labels.csv` został podzielony w proporcji 9:1 na zbiory, które od teraz, już do końca raportu, będziemy nazywać treningowym i testowym.
Żadne etykiety, ani żadne atrybuty obserwacji z udostępnionych do projektu danych nie zawierają braków.
W celu uniknięcia błędu predykcji związanego z podglądaniem danych, kolejne badania charakterystyk cech podanych zbiorów dokonywane są wyłącznie na zbiorze treningowym.
Po wstępnym zapoznaniu się z histogramami zmiennych objaśniających dla zbioru treningowego niezauważalne są obserwacje odstające. W większości przypadków, histogramy te przypominają gęstość rozkładu normalnego.

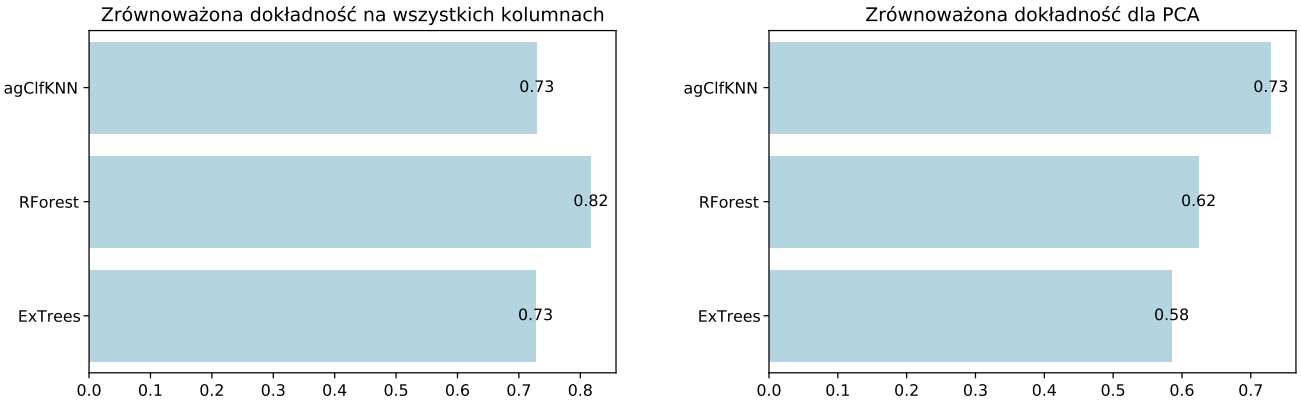
3 Wstępna budowa modelu

Na początku, utworzono kilka modeli: BaggingClassifier100KNN, RandomForestClassifier oraz ExtraTrees, przy czym każdemu z nich zostały za pomocą kroswalidacji dobrane optymalne parametry, takie jak liczba najbliższych sąsiadów brana pod uwagę w klasyfikacji, maksymalna głębokość drzew oraz najmniejsza liczba obserwacji w liściu. Każdy z nich został zbudowany na danych treningowych z niezredukowaną liczbą kolumn i została sprawdzona jego dokładność na zbiorze testowym. Wyniki prezentują się na rysunku 1a. Wyniki te posłużą nam do porównywania przyszłych modeli.

4 Redukcja wymiarowości

Wskazówką do wykonania projektu było to, że pewne zmienne objaśniające są zbędne do prawidłowej estymacji zmiennej objaśnianej. Idąc tym tropem, wykonane zostały testy bazujące na następujących ideach redukcji wymiarowości:

1. PCA,
2. współczynnik informacji wzajemnej (ang. mutual information, MI, information gain),



(a) (b)

Rysunek 1

3. Fisher score,
4. regularyzacja LASSO.

Testy sposobów (2)-(4) objawiają się zbadaniem mocy predykcyjnej na zbiorze testowym dla modeli bagging 100 klasyfikatorów k-najbliższych sąsiadów, lasu losowego oraz Extra Trees, wszystkie z domyślnymi parametrami. Na takie rozwiązanie zdecydowano się ze względu na to, że zwiększenie obszaru wyszukiwań zwiększyłoby co najmniej kilkukrotnie czas badania.

4.1 PCA

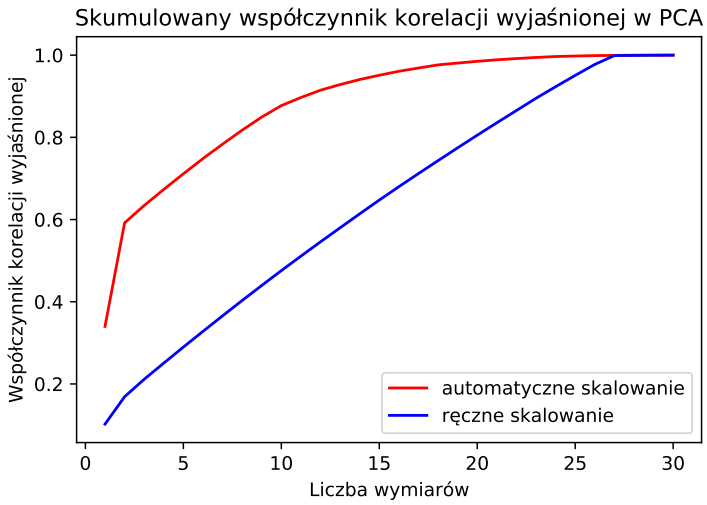
Zastosowano metodę PCA dwoma sposobami. Pierwszy uniknął normalizowania zbioru treningowego przed podaniem go do wyznaczania składowych głównych, drugi uprzednio przeskalowano. W rezultacie na wykresie 2 widoczne jest, że pierwsza transformacja przejawia ciekawsze wnioski. Wynika z niej, przykładowo, że przy zachowaniu 95.1% wyjaśnionej wariancji można pozwolić sobie na usunięcie piętnastu najmniej wyjaśniających wymiarów przestrzeni trzydziestowymiarowej. Analizując wykres 3, pomimo faktu, że zaledwie przy czterech głównych składowych skuteczność jest rzędu 0.7, to zwiększenie głównych składowych nie daje w większości przypadków nawet poprawy, a jeśli daje to niewielką. Z tego powodu, że na pełnych danych zbudowany model jest lepszy na zbiorze testowym, odrzucony został wariant redukcji wielowymiarowości metodą PCA.
Mogła ujawnić się tu potencjalna wada metody PCA. Mianowicie, przy tworzeniu nowej bazy, nie bierze ona pod uwagę korelacji zbioru atrybutów z wartościami zmiennej objaśnianej.

4.2 Współczynnik informacji wzajemnej

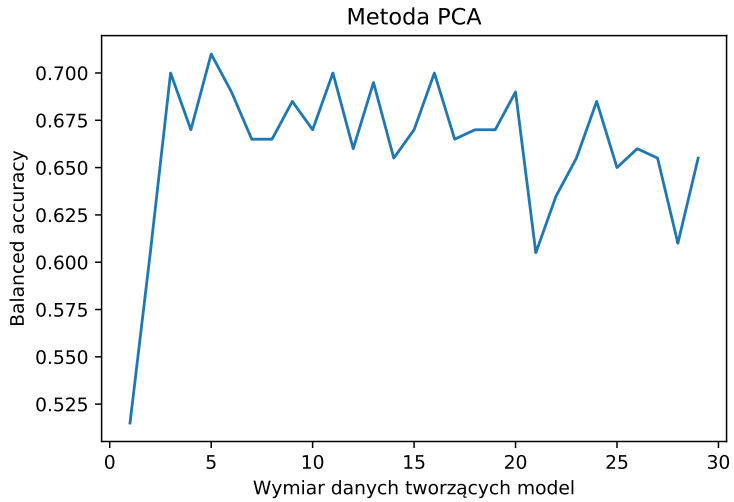
Ta metoda stanowi odpowiedź na problem, który potencjalnie wystąpił w przypadku PCA. MI jest miarą zależności między dwiema zmiennymi losowymi i jest używany do oceny, jak bardzo informacje o jednej zmiennej mogą pomóc w przewidywaniu drugiej, zatem łączy ona informacje o atrybutach z informacjami o etykietach. Z rysunku 4 wyczytać można, że nieznacznie metoda ta różni się od lasu losowego nauczonego na wszystkich kolumnach w sekcji 3.

4.3 Fisher Score

W kontekście klasyfikacji binarnej (gdzie mamy dwie klasy), Fisher Score ocenia, jak bardzo wartości danej cechy różnią się między dwiema klasami w porównaniu do różnic wewnątrz tych klas. Ocenia, czy średnie wartości danej cechy dla różnych klas są istotnie różne, co może sugerować, że dana cecha jest informatywna dla klasyfikatora. Ponownie, zbudowano model lasu losowego dla k najważniejszych kolumn uszeregowanych według Fisher Score i wyniki przedstawiono na wykresie 5. Znowu widać, że dopiero przy ponad dwudziestu dwóch kolumn wykorzystanych do budowy modelu widać jakąś porównywalną jakość predykcji względem wyników z sekcji 3. Wydaje się, że nie jest to właściwe rozwiązanie.



Rysunek 2



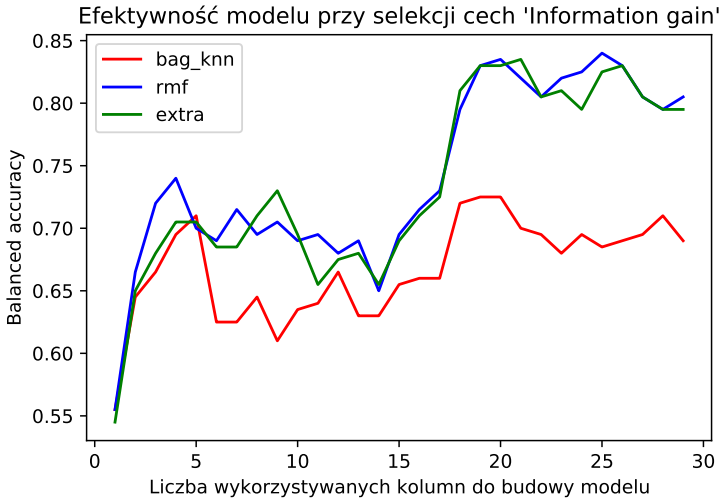
Rysunek 3

4.4 Regularyzacja LASSO

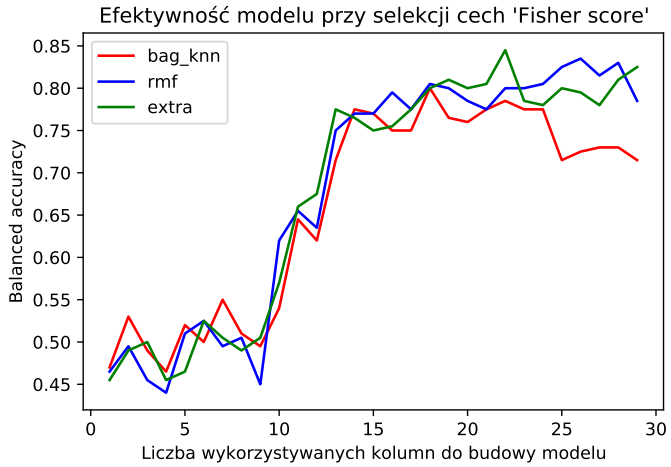
Stosując metodę LASSO do znajdywania istotnych zmiennych w modelu, okazuje się, że niezależnie od tego, czy bierzemy pod uwagę model lasu losowego czy model Extra Trees, uzyskują one zaskakująco dobrą moc predykcji dla zestawu pewnych dwunastu kolumn zbioru treningowego, co uwidocznione zostało na rysunku 6.

5 Wybór ostatecznego modelu i jego ocena

Ze względu na jedną z najwyższych mocy predykcji oraz na możliwość odrzucenia ponad połowy zmiennych przy zachowaniu korzystnej zrównoważonej dokładności, zdecydowano się na ostatni przedstawiony model dla lasu losowego. W przypadku nie podania żadnych hiperparametrów przy tworzeniu klasyfikatora, model ten uzyskał na zbiorze treningowym stuprocentową skuteczność oraz na zbiorze testowym 86%. Podjęto jeszcze próbę dostosowania hiperparametrów do tego lasu losowego przy pomocy *GridSearchCV* w obawie przed tym, że poprzedni model mógł być przetrenowany , biorąc pod uwagę stuprocentową skuteczność na zbiorze uczącym. Okazało się jednak, że regularyzowany las osiągnął gorsze wyniki zarówno na treningowym, jak i testowym zbiorze. Z tego powodu odrzucono również ten model.
Na potwierdzenie dobrej efektywności uzyskanego rozwiązania może świadczyć fakt, że model ten na pięciu procentach



Rysunek 4

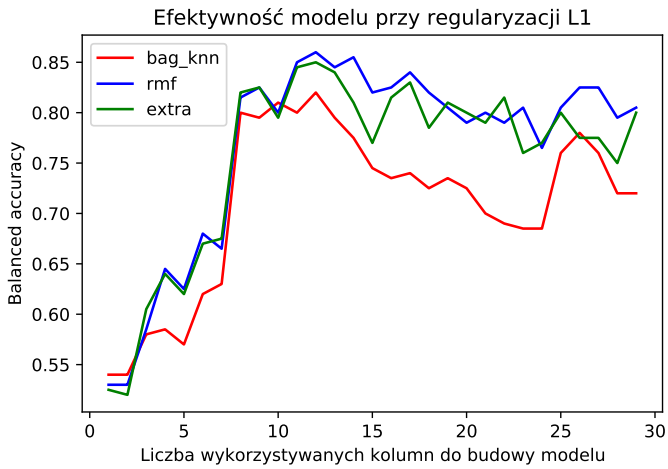


Rysunek 5

rzeczywistego zbioru testowego, do którego etykiet nie było dostępu, osiągnął wynik 93.33% zrównoważonej dokładności.

6 Podsumowanie

Znacząca część projektu została poświęcona znajdywaniu optymalnej redukcji danych. Wykorzystaliśmy do tego różne sposoby, wśród których pojawiły się również takie, których działanie należało zgłebić samodzielnie, ponieważ nie były one przedstawiane na wykładzie. Mimo wszystko, najskuteczniejszą metodą okazało się sprawdzone LASSO. Następnie, po porównaniu różnych estymatorów, wybrano las losowy, który jest ogólnie rzecz biorąc często polecanym estymatorem jeśli myśli się, że pewne zmienne modelu odgrywają niewielką rolę w rzeczywistej predykcji etykiet.
Pomimo stuprocentowej poprawności modelu na zbiorze uczącym, wyniki dla dwóch zbiorów testowych pozwalają wierzyć w to, że nie jest on przetrenowany i dobrze poradzi sobie w przewidywaniu przyszłych etykiet.


Efektywność modelu przy regularyzacji L1
bag_knn
rmf
extra
Balanced accuracy
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0
5
10
15
20
25
30
Liczba wykorzystywanych kolumn do budowy modelu

Rysunek 6