Projekt 1 - raport

Jakub Kozieł
Tomasz Krupiński
Jakub Lis

1 Opis problemu

Zadaniem było stworzenie modelu, który byłby w stanie ustalić, czy dana osoba zarabia powyżej 50tys.$ rocznie. Zadanie to więc polegało na klasyfikacji binarnej. Dane pochodziły ze zbioru Census income. Znajdują się w tym zbiorze dane odnośnie: pracy, statusu matrymonialnego, edukacji oraz cech biologicznych.

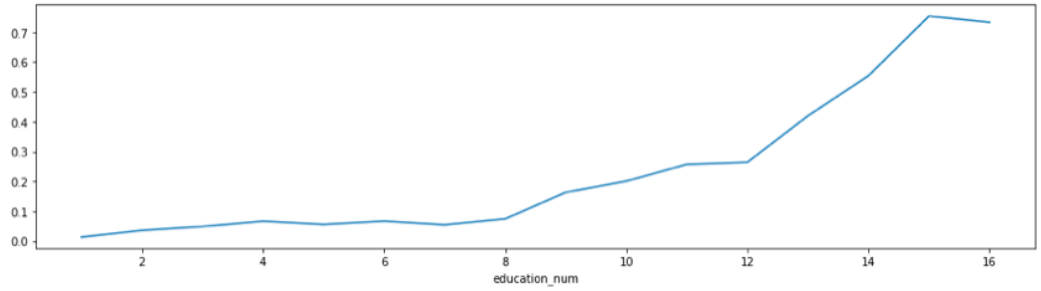
Showing 5 out of 48842 rows

age	workclass	fnlwgt	education	education_num	marital_status	occupation	relationship	race	sex	capital_gain	capital_loss	hours_per_week	native_country	income_level
39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States	<=50K
50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K
38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States	<=50K
53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States	<=50K
28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=50K

Head danych

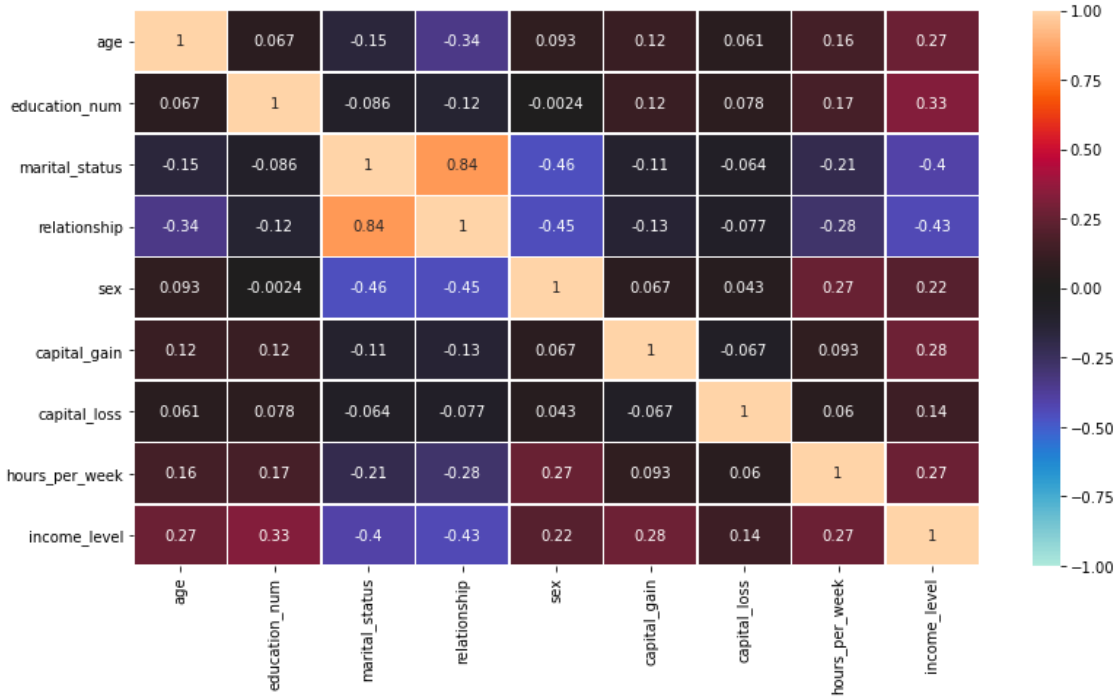
2 Inżynieria cech i encoding

Zbiór danych, z którego korzystaliśmy zawierał braki. Ponieważ takich wybrakowanych wierszy nie było dużo, to postanowiliśmy je usunąć. Usuwaliśmy także dwie kolumny: *education*, *fnlwgt*. Pierwszą ze względu na jej powtarzalność (przechowywaliśmy taką samą informacje w kolumnie *education_num*), a drugą, dlatego że nie niosła szczególnego znaczenia dla modelowania.
Oprócz tego, postanowiliśmy zmodyfikować wartości w kolumnie *education_num*. Początkowo chcieliśmy je przeskalować w taki sposób, aby otrzymać przypominający linię prostą wykres, na którym na osi pionowej znajdował się procent osób zarabiających powyżej 50tys. dolarów, a na osi poziomej osoby z danym poziomem edukacji.



Wykres przed przeskalowaniem

Ostatecznie zdecydowaliśmy się zgrupować podobne poziomy edukacji, aby móc zastosować OneHotEncoding. Uznaliśmy, że odpowiednimi grupami będą osoby z *education_num* 1-8, 9-10, 11-12, 13, 14, 15-16 ze względu na podobieństwa w procencie osób zarabiających powyżej 50tys. oraz ze względu na liczebność tych grup.
W późniejszym etapie inżynierii cech zauważyliśmy także dużą korelację pomiędzy kolumnami *marital_status* i *relationship*.



Macierz korelacji metodą Spearmana

Z tego względu zdecydowaliśmy się na usunięcie tej pierwszej, a dodatkowo zgrupowaliśmy posiadanie żony lub męża jako jedną kategorię.
Podobne grupowanie zastosowaliśmy w przypadku kolumny *occupation* - osoby, których pracą było *Priv-house-serv* lub *Armed-Forces* podłączyliśmy do istniejącej już w kolumnie wartości *Other-service*. Uznaliśmy to za słuszne ze względu na niewielką liczbę wierszy zawierających takie zawody oraz na podobny rozkład zarobków do *Other-service*.
Przed kodowaniem kolumn zajęliśmy się jeszcze kolumnami *capital_gain* oraz *capital_loss*. Postanowiliśmy połączyć je w jedną, tzn. *capital_diff*, gdzie zapisaliśmy różnicę wartości gain i loss.

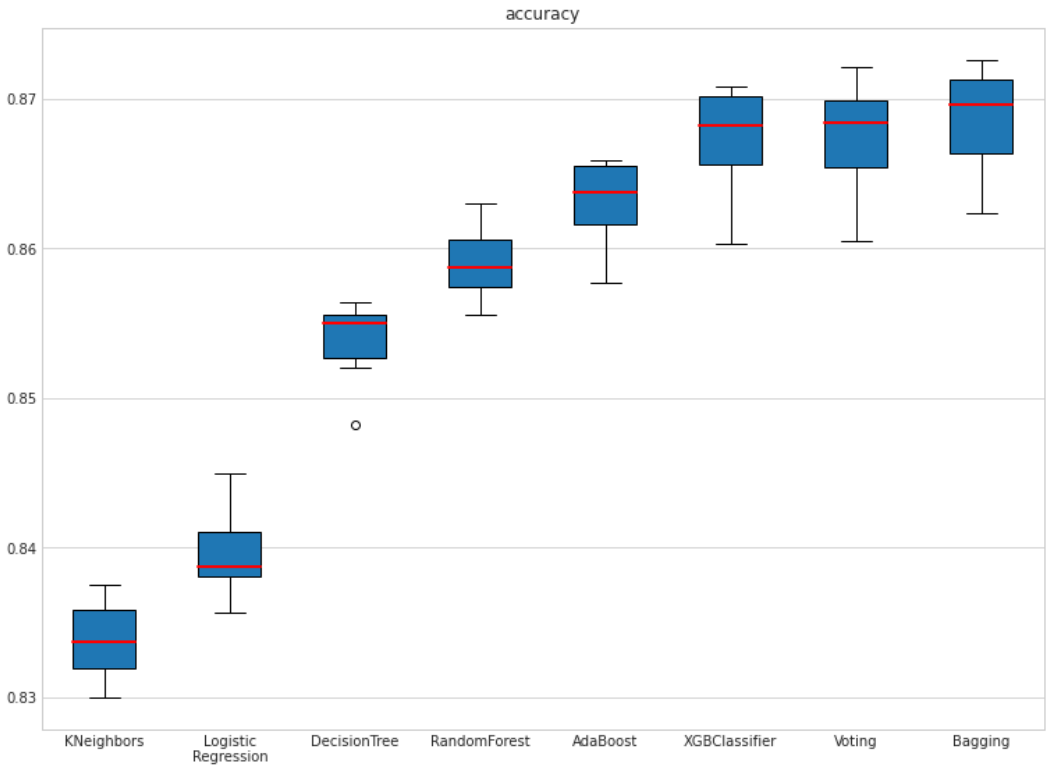
Najwięcej kolumn zakodowaliśmy korzystając z OneHotEncoding: *occupation*, *relationship*, *race*, *workclass*, *education_num*. Takie kolumny jak *income_level*, *sex* (zawierające tylko dwie wartości) zakodowaliśmy zamieniając jedną z wartości na 0, a drugą na 1. Dodatkowo, dla kolumny zawierającej dużo różnych odbiegających od siebie wartości (*Native_country*) zastosowaliśmy TargetEncoding.
Przed modelowaniem zastosowaliśmy jeszcze MinMaxScaller dla kolumn *capital_diff*, *hours_per_week*, *age*.

3 Modelowanie

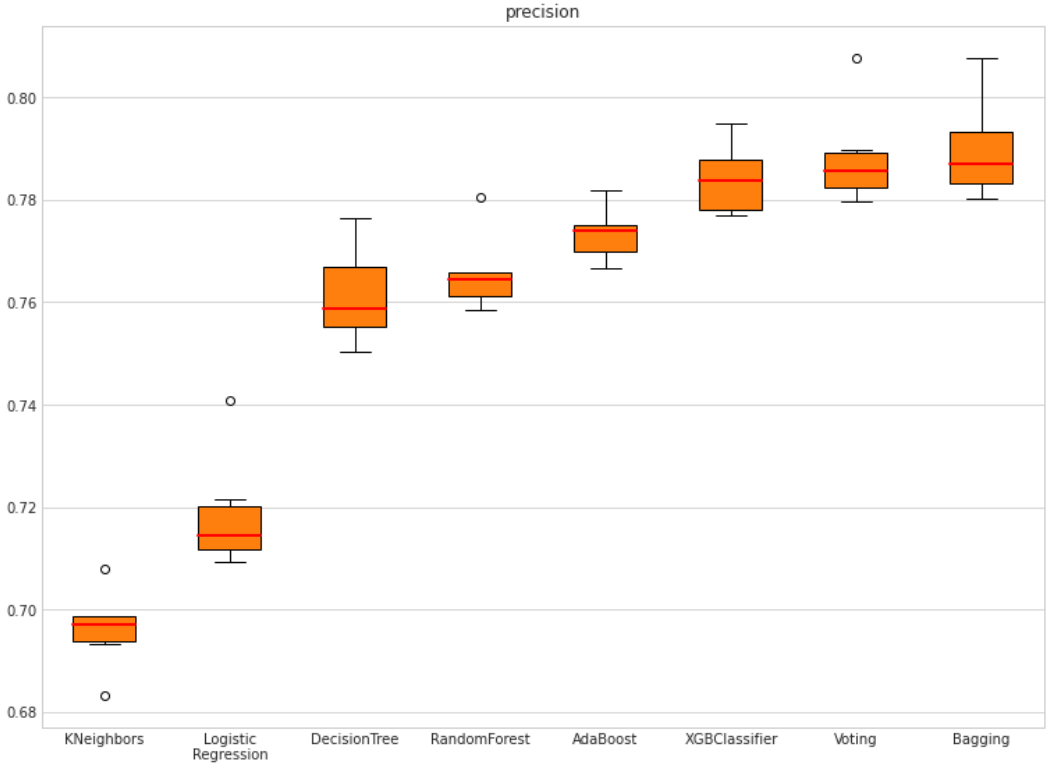
Podczas ostatnich etapów pracy stworzyliśmy osiem różnych modeli mających szanse na osiąganie wysokich wyników, po wcześniejszym odrzuceniu dwóch typów modeli, które nie najlepiej poradziły sobie we wczesnym modelowaniu przy okazji drugiego kamienia milowego. Wystroiliśmy hiperparametry wszystkich z tych ośmiu modeli oraz stworzyliśmy metryki by móc porównać ich wyniki.

Jednym z naszych pomysłów na zwiększenie wyników, było zastosowanie najpierw zwiększenia liczby cech za pomocą PolynomialFeatures a następnie wybranie 50 najlepszych z nich za pomocą SelectKBest(chi2, k=50). Okazało się jednak, że PolynomialFeatures tworzy zbyt dużo cech (około 30tys.), często dziwnych, prawdopodobnie z tego powodu, że stosujemy one-hot encodingi. Jedynie to pogarszało wyniki, więc postanowiliśmy nie korzystać z tej metody.

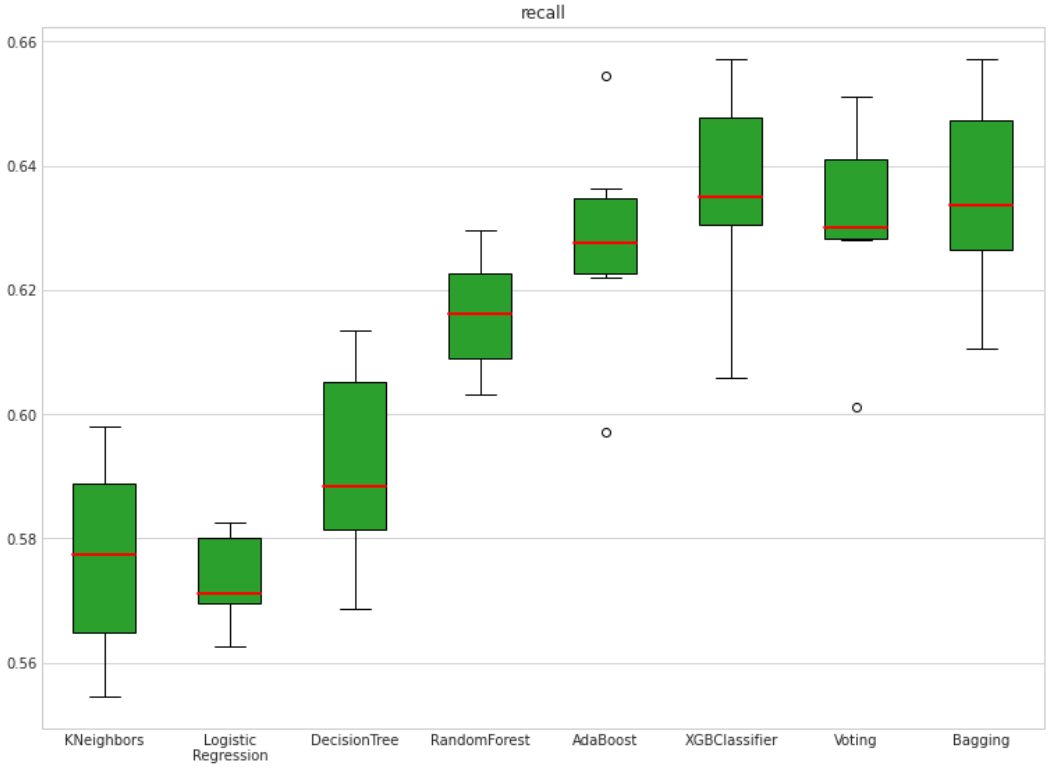
Za pomocą metryk takich jak: accuracy, precision, recall, f1, auc, będziemy wybierać najlepsze modele. Podczas modelowania zapisywaliśmy osiągane oceny jakości, aby następnie móc stworzyć wykresy je porównujące.



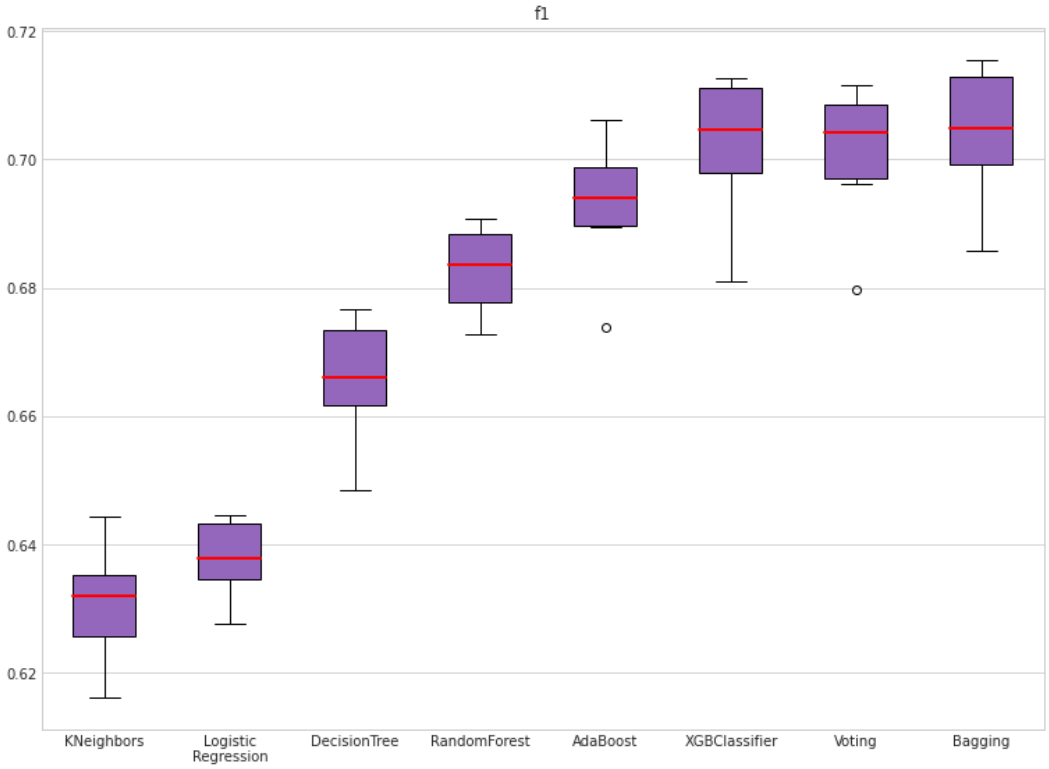
Porównanie boxplotów osiąganych wartości accuracy przez modele podczas kroswalidacji



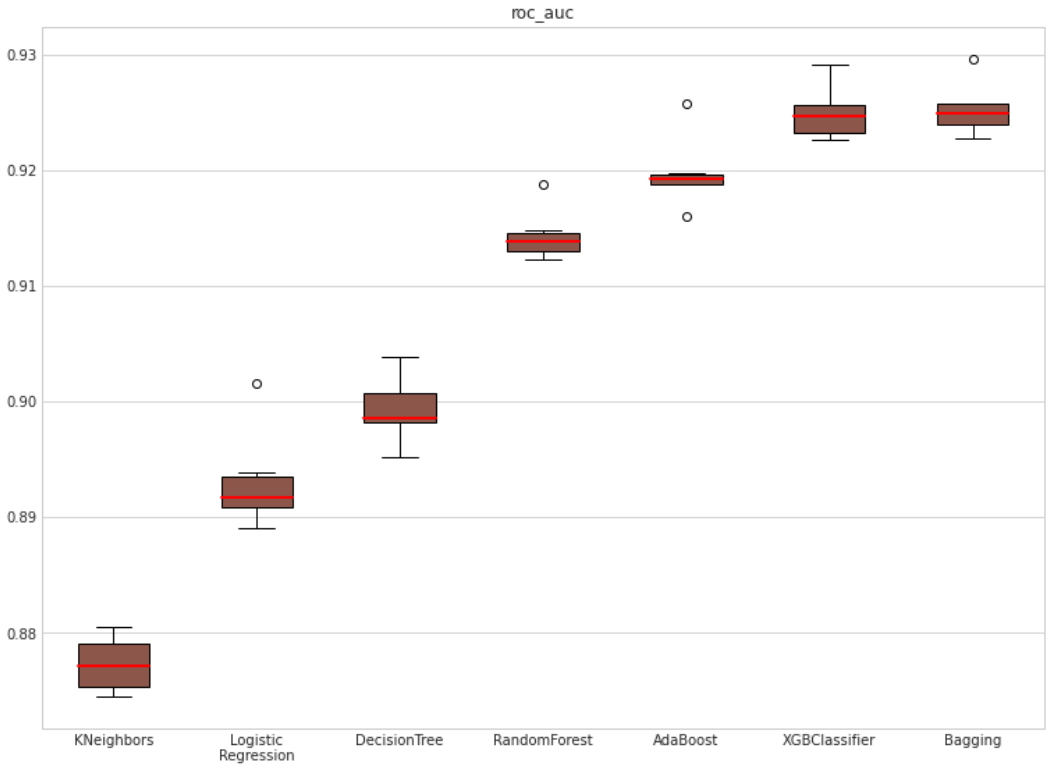
Porównanie boxplotów osiąganych wartości precission przez modele podczas kroswalidacji



Porównanie boxplotów osiąganych wartości precission przez modele podczas kroswalidacji

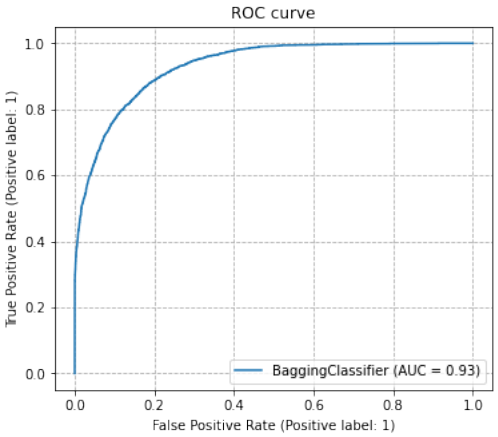


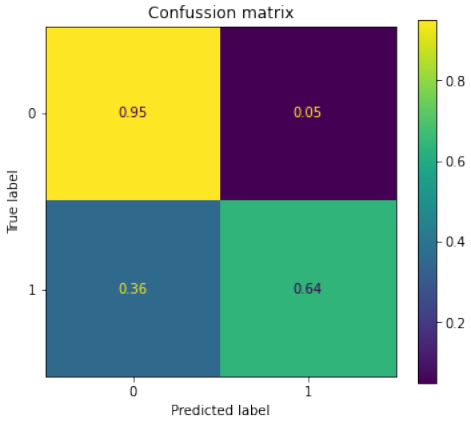
Porównanie boxplotów osiąganych wartości f1 przez modele podczas kroswalidacji

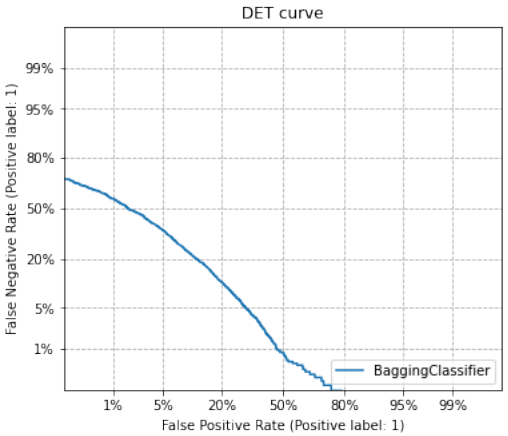


Porównanie boxplotów osiąganych wartości roc_auc przez modele podczas kroswalidacji

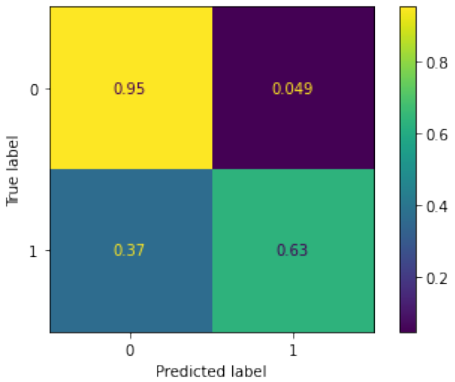
Z tego powodu, że zawsze Bagging, Voting oraz XGBoost osiągały najlepsze wyniki w każdej z tych metryk (poza roc_auc, gdyż nie ma tam votingu) postanowiliśmy się im bliżej przyjrzeć.



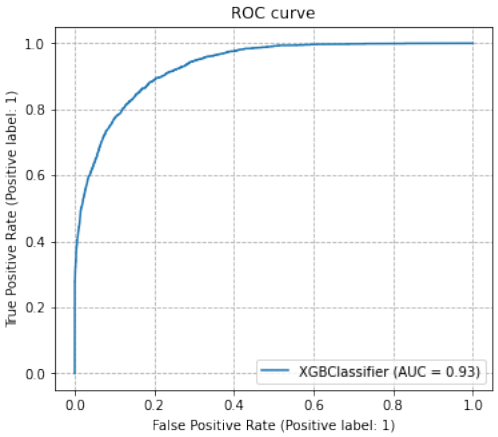


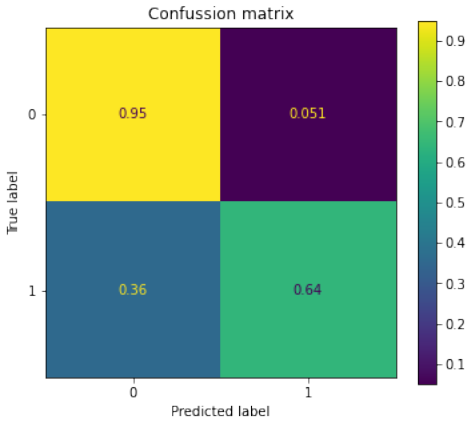


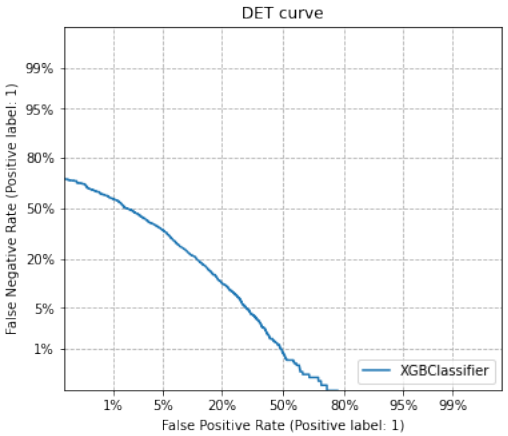
metryki Bagging Classifier



metryki Voting Classifier







metryki XGBoost

	Bagging	Voting	XGBoost
accuracy	0.87131	0.86932	0.87131
recall	0.64043	0.63043	0.64347
precision	0.81381	0.81369	0.81140
f1	0.71678	0.71043	0.71774
AUC ROC	0.79523		0.79623

Tabela zawierająca osiągane miary jakości przez Bagging, Voting oraz XGBoost

4 Podsumowanie

Zdecydowaliśmy się wybrać Bagging Classifier (którego podstawowym estymatorem był XGBoosting), gdyż osiągał on najlepsze wyniki pod względem wszystkich używanych metryk (accuracy, precision, recall, f1, auc_roc).

Podsumowując, uważamy, że nasz model podczas kroswalidacji dawał wyniki godne zaufania - na boxplotach widzimy, że i mediana wyników, i minimum są odpowiednio wysokie. Choć pozostałe prototypowe modele nie osiągnęły aż tak dobrych wyników, to nie odstawały w tyle.