WYDZIAŁ MATEMATYKI I NAUK INFORMACYJNYCH
POLITECHNIKI WARSZAWSKIEJ

Klasyfikacja dochodów jednostki na podstawie danych ze spisu powszechnego

---

WSTĘP DO UCZENIA MASZYNOWEGO
PROJEKT 1

Mateusz Krzyziński, Tomasz Nocoń, Paweł Fijałkowski

Grupa laboratoryjna nr 2

16.04.2021

1 Opis problemu

Praca ta jest raportem z przebiegu projektu numer 1 z przedmiotu Wstęp do Uczenia Maszynowego w semestrze letnim roku 2021. Tematem zadania projektowego było przygotowanie modelu klasyfikacyjnego dla wybranego zbioru danych. Przebieg pracy nad projektem został podzielony na trzy etapy:

- eksplorację danych – EDA,
- inżynierię cech i wstępne modelowanie,
- przygotowanie finalnych modeli.

W naszym przypadku wybrany zbiór danych dotyczył danych Amerykanów ze spisu powszechnego, a zadanie klasyfikacyjne polegało na przewidzeniu, czy dochód danej osoby wynosi powyżej 50 tysięcy dolarów, czy też nie przekracza tej liczby.
Zadanie zostało wykonane w języku `Python` z użyciem biblioteki do uczenia maszynowego `Scikit-learn`. Poszczególne etapy projektu można znaleźć w repozytorium przedmiotu.

2 Opis zbioru danych

Zbiór danych o nazwie *Census income* zawiera informacje pochodzące ze spisu ludności Amerykanów i został opublikowany w 1996 roku. Jest on dostępny publicznie i można go pobrać pod adresem `https://www.apispreadsheets.com/datasets/106`. Oryginalny zbiór pochodzi z repozytorium UCI Machine Learning Repository.
Zbiór danych składa się z:

- 48 842 wierszy,
- 15 kolumn,

przy czym zawiera również brakujące wartości (w 3630 wierszach).

2.1 Informacje zawarte w zbiorze danych

Spośród 15 kolumn ze zbioru danych 6 ma charakter numeryczny, a 9 kategoryczny (w tym jedna zmienna jest zmienną celu).

2.1.1 Zmienne numeryczne

- `age` – wiek (zakres 17-90 – dane dotyczą tylko osób po 16 roku życia, a wiek maksymalny został przycięty do wartości 90 lat),
- `fnlwgt` – *final sampling weight*; waga przypisana rekordowi w celu odzwierciedlenia całej populacji (parametr statystyczny ze spisu),
- `education_num` – liczba określająca osiągnięty poziom wykształcenia (zakres 1-16),
- `capital_gain` – zysk kapitałowy (wartość przycięta do 99 999$),

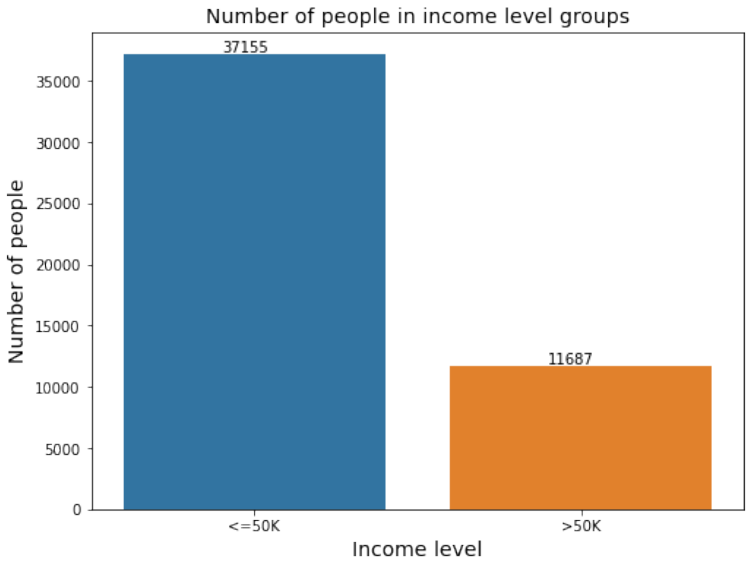
- `capital_loss` – strata kapitałowa,
- `hours_per_week` – liczba godzin pracy tygodniowo (wartość przycięta do 99 h).

2.1.2 Zmienne kategoryczne

- `workclass` – klasa pracownicza (8 różnych wartości),
- `education` – osiągnięty poziom wykształcenia (16 wartości odpowiadających zmiennej `education_num`),
- `marital_status` – stan cywilny (7 różnych wartości),
- `occupation` – rodzaj wykonywanego zawodu (14 różnych wartości),
- `relationship` – status związku (6 różnych wartości),
- `race` – pochodzenie etniczne (5 różnych wartości),
- `sex` – płeć (2 różne wartości),
- `native_country` – kraj pochodzenia (42 różne wartości),
- `income_level` – **zmienna celu** – klasa na podstawie osiągniętego dochodu: dwie wartości (`<=50K`, `>50K`).

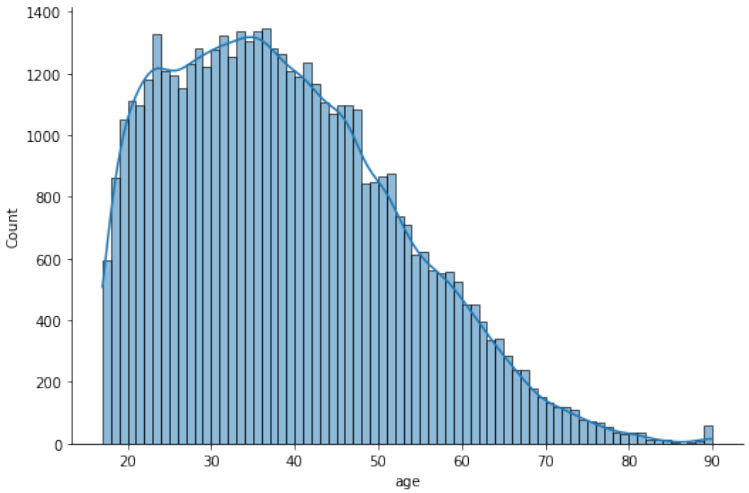
2.2 Wybrane wizualizacje dotyczące zbioru danych

W kontekście problemu najważniejsza jest zmienna celu – target, którego przewidzenie jest zadaniem projektowym. W związku z tym najważniejsza wizualizacja dotyczy liczby osób w zbiorze danych, którym została nadana poszczególna etykieta.

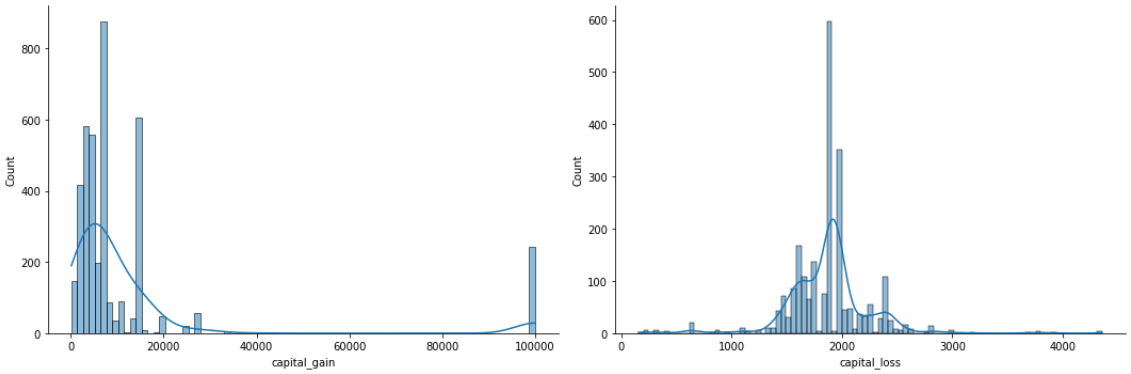


Rysunek 1: Rozkład zmiennej celu. Widać, że target nie jest zbalansowany. Większa liczba obserwacji dotyczy osób, których dochód nie przekracza 50 000 $.

Rozkład zmiennych numerycznych najlepiej obrazują ich histogramy, których przykłady zostały załączone poniżej.

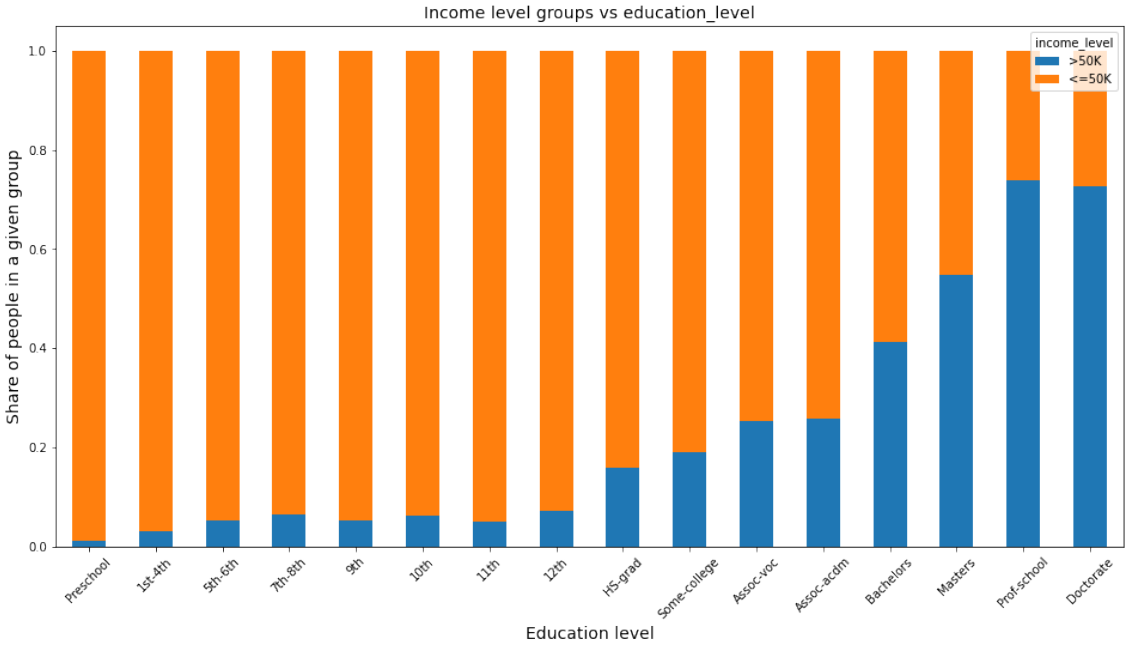


Rysunek 2: Histogram zmiennej `age`. Widać, że wiek przypomina rozkład normalny, ale dane dotyczą tylko osób powyżej 16. roku życia i zostały przycięte – wszystkie osoby z wiekiem $\geqslant$ 90 lat otrzymały wartość dokładnie 90 lat.

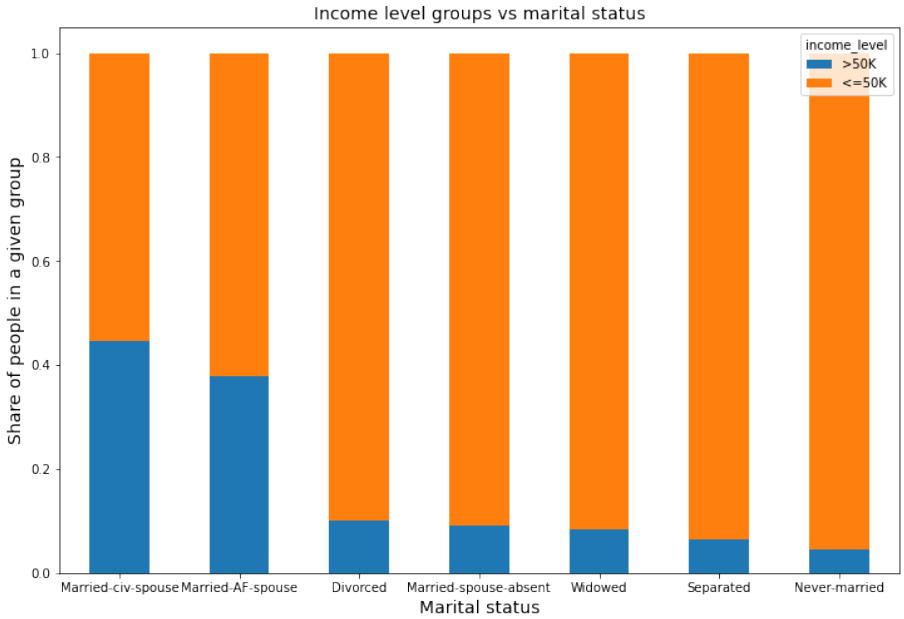


Rysunek 3: Histogramy zmiennych `capital_gain` i `capital_loss`. Wyrysowane histogramy obejmują tylko niezerowe zmienne, gdyż w danych dominują wartości zerowe dla powyższych cech (91.73% obserwacji ma zerowy zysk kapitałowy i 95.33% obserwacji ma zerową stratę kapitałową). Widać, że zyski kapitałowe zostały przycięte do wartości 99 999 $, lecz większość z nich koncentruje się w przedziale do 20 000$. Tymczasem w przypadku strat koncentrują się one w przedziale 1500 - 2000 $, są zatem znacznie mniejsze niż osiągane zyski.

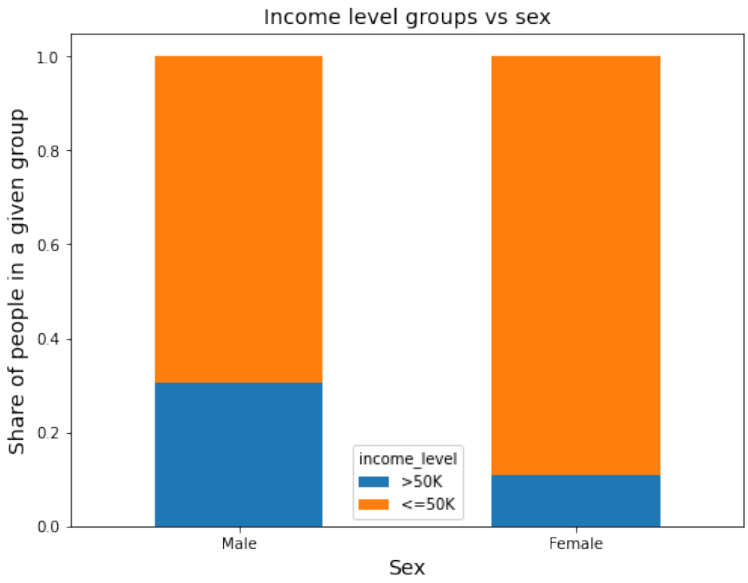
Bardzo istotnymi zmiennymi (mającymi dla różnych wartości znaczące różnice w udziałach w poszczególnych klasach) wydają się być między innymi `education` (w numerycznej postaci ma najwyższą wartość współczynnika korelacji Spearmana ze zmienną celu – 0.33), `marital_status`, `sex` i `age`. Ich zależność od zmiennej celu została przedstawiona na poniższych wykresach.



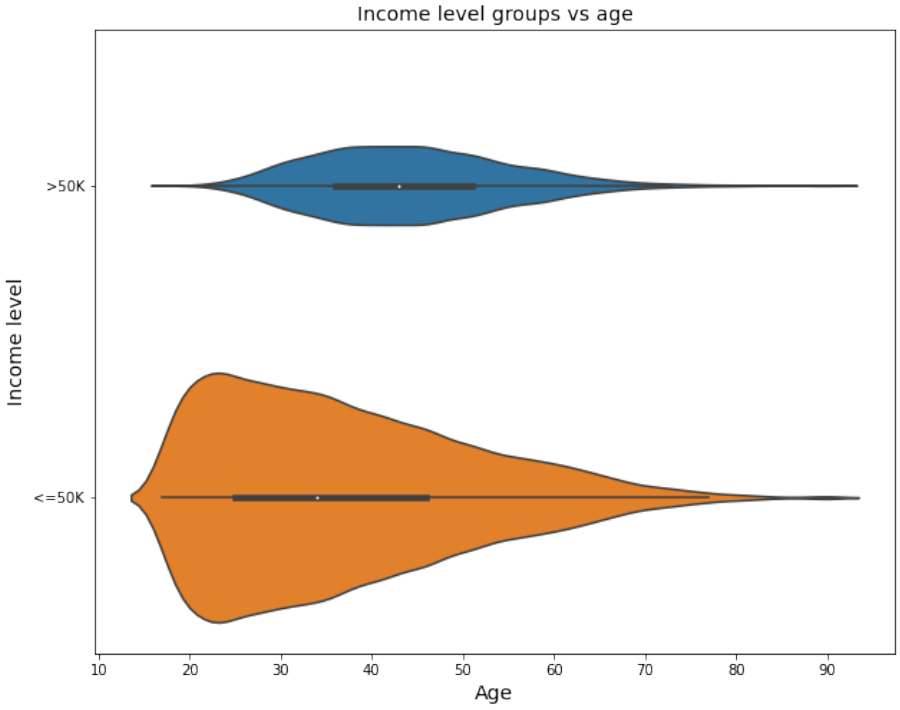
Rysunek 4: Zależność między poziomem wykształcenia `education` a poziomem dochodów `income_level`. Wartości na osi x zostały uszeregowane w kolejności związanej z wartościami numerycznymi dotyczącymi edukacji (kolumna `education_num`. Widać, że wyższy poziom wykształcenia wiąże się z większą szansą na wyższy dochód, natomiast nie jest to zależność liniowa i ściśle monotoniczna (na co ma też wpływ niezbalansowanie pomiędzy poszczególnymi poziomami).



Rysunek 5: Zależność między stanem cywilnym `marital_status` a poziomem dochodów `income_level`. Wartości na osi x zostały uszeregowane w kolejności malejącej względem udziału w grupie lepiej zarabiającej. Widać znaczącą różnicę między osobami w związku małżeńskim, a niebędącymi w związku.



Rysunek 6: Zależność między płcią `sex` a poziomem dochodów `income_level`. Widać znaczącą różnicę między dochodami mężczyzn a kobiet.



Rysunek 7: Zależność między stanem cywilnym `age` a poziomem dochodów `income_level`. Widać, że wśród osób starszych jest większa szansa na większy roczny dochód. Jednocześnie wraz z wiekiem rośnie balans pomiędzy kategoriami.

3 Inżynieria cech

Na podstawie przeprowadzonej eksploracyjnej analizy zbioru danych mogliśmy wskazać kilka ograniczeń i sugestii, które wzięliśmy pod uwagę przy preprocessingu danych do tworzenia modeli klasyfikujących.

- Ze zmiennych wyeliminowaliśmy kolumnę `fnlwgt`, która nie jest związana z targetem, a stanowi tylko parametr statystyczny związany ze spisem powszechnym.
- Wyeliminowaliśmy obserwacje zawierające zbiory danych (stosując się do polecenia sugerującego ograniczenie liczby rekordów).
- Ze względu na niezbalansowanie w zmiennej celu, zdecydowaliśmy się wymusić równomierny podział zbioru na treningowy i testowy – tak, by odzwierciedłał rzeczywisty rozkład.
- W zbiorze danych znajdują się dwie zmienne opisujące poziom wykształcenia jednostki – numeryczna i kategoryczna. Jest to dodatkowo jedna z najbardziej skorelowanych z targetem cech. W celu stworzenia modelu należało wyeliminować jedną z nich, by uniknąć niepotrzebnej redundancji.

  Początkowo przetestowaliśmy użycie w modelu zmiennej w postaci numerycznej. Jednak należy zauważyć, że różnice wartości zmiennej nie odpowiadają realnym różnicom w poziomie edukacji (różnica między klasami 5-6 a 7-8 jest w istocie inna niż różnica między klasą 12 a absolwentem high school).

  Próbowaliśmy również wykorzystać kombinację liniową zmiennej numerycznej i kategorycznej, gdzie tę drugą kodowaliśmy przy użyciu target encodingu. Pozwalało to na lepsze odzwierciedlenie różnic w poziomach edukacji, przy jednoczesnym obniżeniu wariancji – overfittingu.

  Ostatecznie zdecydowaliśmy się na użycie jedynie zmiennej kategorycznej, co dawało najlepsze rezultaty.
- Do przekształcenia zmiennych kategorycznych na numeryczne użyliśmy one hot encodingu.
- W celu ograniczenia liczby kolumn nowo powstałych w wyniku one hot encodingu, a także w celu uogólnienia modelu próbowaliśmy różnego grupowania zmiennych kategorycznych. Uwzględnialiśmy przy tym liczność poszczególnych kategorii (chcąc łączyć te, dla których liczba obserwacji jest niewielka), związek z targetem (chcąc łączyć te, które mają podobny rozkład względem zmiennej celu) i wreszcie związek między samymi wartościami (by nie tworzyć sztucznych grup, pozbawionych sensu i interpretowalności).

  Cechy, które grupowaliśmy to `marital_status`, `race`, `education`, `native_country`. W celu zgrupowania krajów pochodzenia (aż 42 różne wartości, w tym wiele nielicznych klas) użyliśmy dodatkowego zbioru danych zawierającego informację o wysokości PKB per capita dla państw w roku 1995 (rok, z którego pochodzą dane).

  Ostatecznie w modelu użyliśmy jedynie grupowania stanu cywilnego, co dawało najlepsze rezultaty (mierzone przy użyciu kroswalidacji). Jednak różnice nie były

znaczące, w związku z tym w przypadku konieczności stworzenia bardziej ogólnego modelu (np. względem kraju pochodzenia obywatela) rozsądnym byłoby użycie któregoś z grupowań.

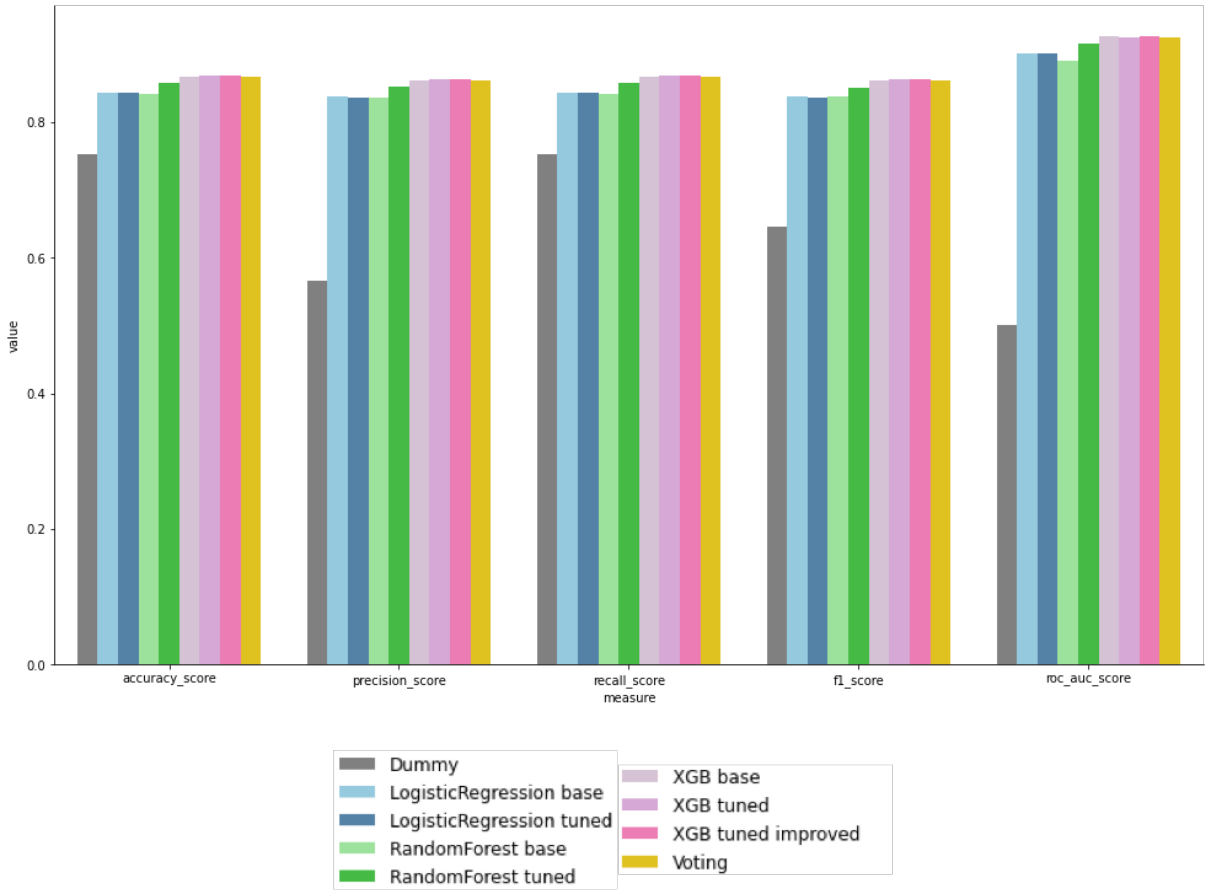
- Do przekształcenia skalującego zmiennych numerycznych użyliśmy standardyzacji.
- Na etapie tworzenia modeli próbowaliśmy również wykorzystać przekształcenia wielomianowe cech i wybrać najlepsze z nich, jak również wykorzystać narzędzia xAI do eliminacji mało znaczących cech. Ostatecznie porzuciliśmy te rozwiązania.

original value	**group**
Married-civ-spouse Married-AF-spouse	Married
Never-married Divorced Married-spouse-absent Separated Widowed	Living-Alone

Tabela 1: Schemat grupowania kolumny marital_status, które zostało użyte w ostatecznych modelach.

4 Modele

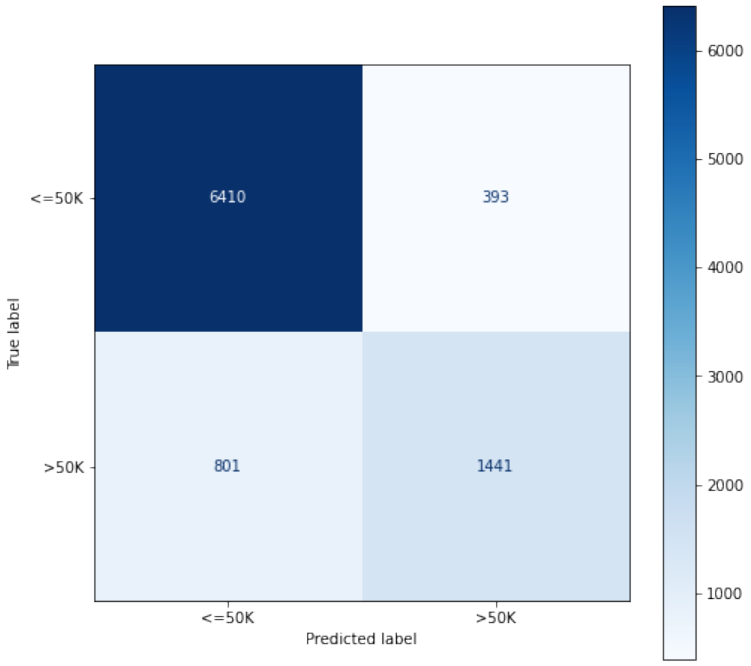
Proces modelowania rozpoczęliśmy od stworzenia prymitywnego klasyfikatora przypisującego zawsze klasę częściej występującą, a więc wysokość zarobków co najwyżej 50 tys. dolarów. Klasyfikator ten miał posłużyć jako baza – punkt odniesienia dla kolejnych tworzonych modeli.
Jakość kolejnych tworzonych modeli ocenialiśmy wykorzystując szereg miar: accuracy, precision, recall, f1 i ROC AUC. Obliczaliśmy wyniki modeli nie tylko na zbiorze testowym, ale też w kroswalidacji na zbiorze treningowym.
Modele, które rozważaliśmy w projekcie to klasyfikatory wykorzystujące następujące algorytmy: regresja logistyczna, las losowy i XGBoost. Chcieliśmy bowiem użyć zarówno modeli mało skomplikowanych, interpretowalnych (jak regresja logistyczna), jak i tych bardziej złożonych.
W celu tuningu hiperparametrów w modelach użyliśmy poszukiwania losowego z potrójną kroswalidacją po zdefiniowanych przestrzeniach dyskretnych. Następnie dla najlepszego modelu uzyskanego w ten sposób (XGBoost) przeprowadziliśmy dodatkowy tuning hiperparametrów. W tym celu również użyliśmy poszukiwania po siatce parametrów dla wartości hiperparametrów zbliżonych do tych otrzymanych we wcześniejszym kroku.
Przetestowaliśmy również komitety głosujące wytrenowanych modeli z najlepszymi parametrami z różnymi wagami. Ostatnie największą wagę głosu (0.8) przyznaliśmy najlepszemu z nich – XGBoost, a pozostałym mniejsze wagi (po 0.1). Jednak nie poprawiło to wyniku na zbiorze testowym.



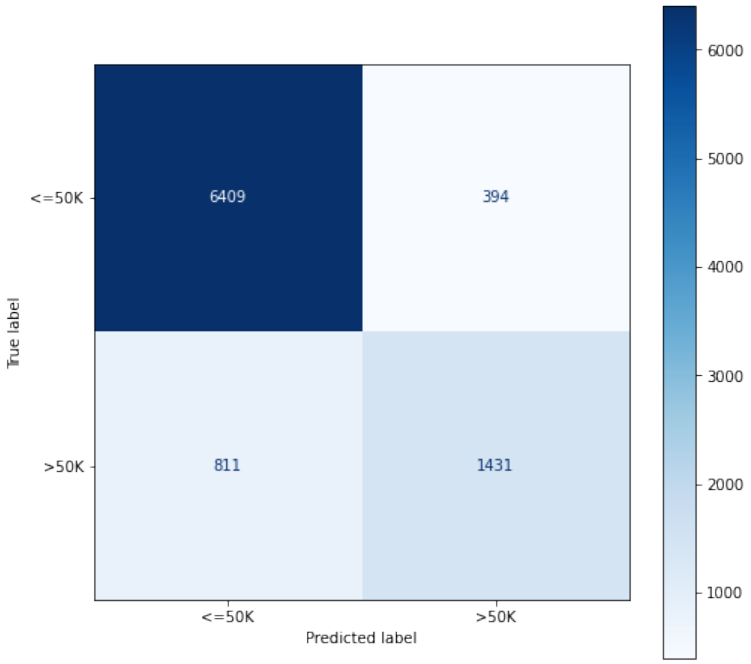
Rysunek 8: Wyniki poszczególnych modeli (oś Y) w różnych metrykach (oś X).

model	**accuracy**	**precision**	**recall**	**f1**	**roc_auc**
Dummy	0.752128	0.565697	0.752128	0.645725	0.500000
LR base	0.844002	0.837064	0.844002	0.837847	0.901488
LR tuned	0.842676	0.835686	0.842676	0.836629	0.901309
RF base	0.842012	0.836475	0.842012	0.838126	0.890946
RF tuned	0.857380	0.851629	0.857380	0.851231	0.915571
XGB base	0.866003	0.861250	0.866003	0.861541	0.926215
XGB tuned	0.867772	0.863103	0.867772	0.863106	0.925537
XGB tuned improved	0.867993	0.863338	0.867993	0.863308	0.925761
Voting	0.866777	0.862003	0.866777	0.861928	0.925682

Tabela 2: Wyniki poszczególnych modeli w różnych metrykach.
Najlepszym modelem okazał się być XGBoost po tuningu hiperparametrów.



Rysunek 9: Macierz pomyłek dla najlepszego z wytrenowanych modeli – XGBoost.



Rysunek 10: Macierz pomyłek dla modelu głosującego. Nie udało się poprawić wyników.

5 Podsumowanie

W czasie projektu udało nam się przejść od surowych danych do stworzenia klasyfikatora, który ma niemalże 87% skuteczności we wskazywaniu na to, czy osoba o danych cechach będzie miała dochód roczny przekraczający 50 tys. dolarów.
Jednak nie sam osiągnięty wynik jest kluczowy. Zrealizowane zadanie było dla nas pierwszym tak kompleksowym projektem dotyczącym uczenia maszynowego. Dzięki jego wykonaniu przećwiczyliśmy wiele umiejętności związanych z poszczególnymi krokami pracy nad projektem tego typu. Pomimo, że zadanie bezpośrednio dotyczyło klasyfikacji i konkretnego zbioru, to stanowiło raczej okazję do ogólniejszego zapoznania się ze schematem działania i wykorzystania dotychczas zdobytej w trakcie kursu wiedzy.