WYDZIAŁ MATEMATYKI I NAUK INFORMACYJNYCH
POLITECHNIKI WARSZAWSKIEJ

Wstęp do uczenia maszynowego

projekt 1

raport

GRUPA 1
Jakub Fołtyn,
Adam Frej,
Paulina Jaszczuk

14.04.2021

1 Wstęp

Naszym zadaniem było dokonanie predykcji rocznych zarobków pracowników i ich binarna klasyfikacja na tej podstawie ($<= 50k$ lub $> 50k$)
Dane, których użyliśmy pochodziły ze spisu ludności osób powyżej 16 roku życia z 1996r - `https://www.apispreadsheets.com/datasets/106`

1.1 Charakterystka danych

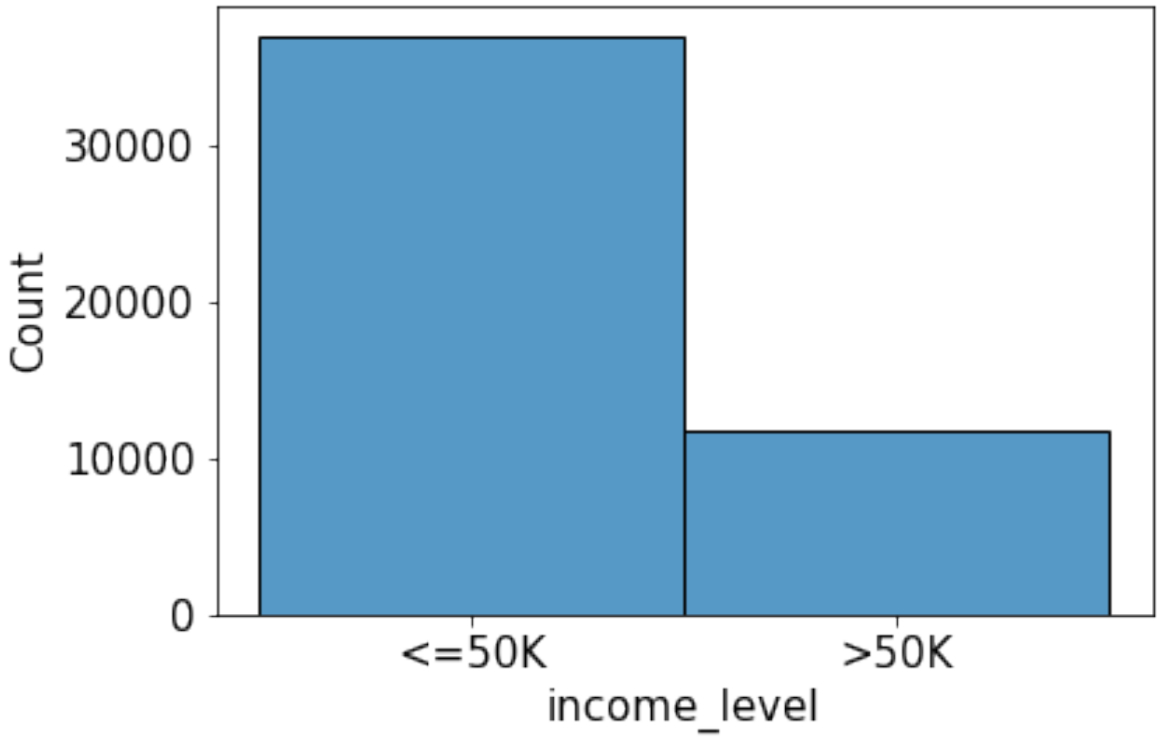
16.04.2021 API Spreadsheets

Name	Type	Description
age	integer	age of individual
workclass	string	Values: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked
fnlwgt	float	Final sampling weight. Inverse of sampling fraction adjusted for non-response and over or under sampling of particular groups
education	string	Values: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool
education_num	integer	
marital_status	string	Values: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse
occupation	string	Values: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces
relationship	string	Values: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried
race	string	Values: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black
sex	string	Values: Female, Male
capital_gain	float	
capital_loss	float	
hours_per_week	float	working hours per week
native_country	string	Values: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands
income_level	string	Predictor class if individual earns greater or less than $50000 per year. Values: <=50K, >50K

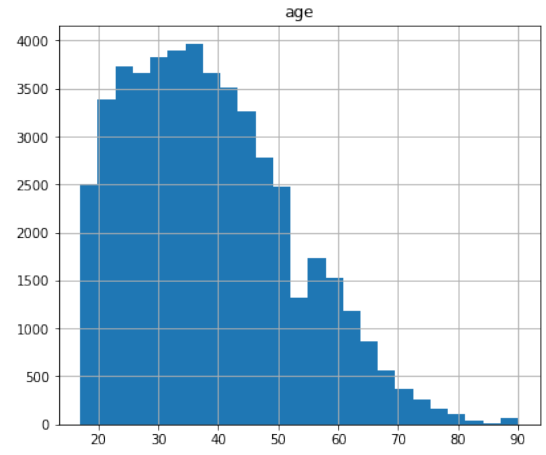
2 EDA

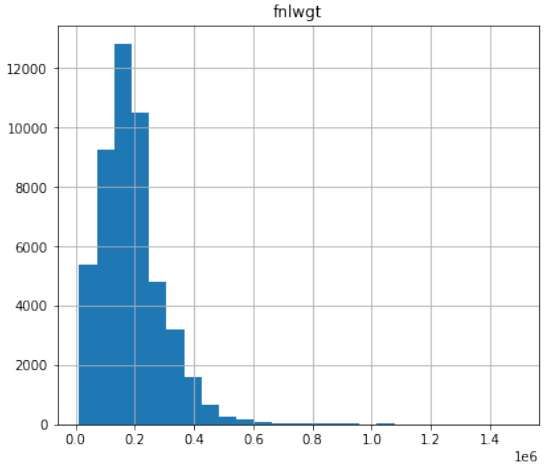
2.1 Ogólne informacje o danych

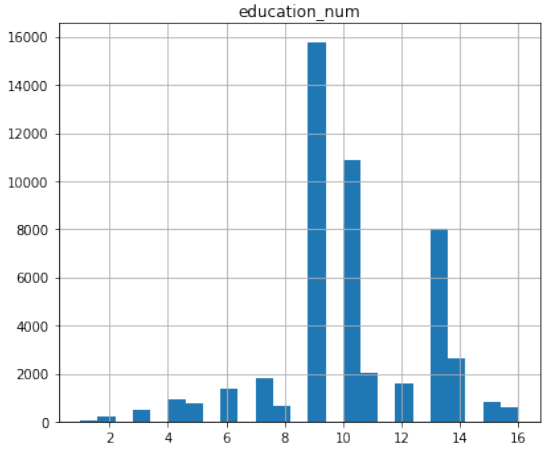
1. 48842 obserwacje
2. 8 cech kategorycznych
3. 6 cech numerycznych
4. 6465 pól z brakującymi danymi
5. brak równej dystrybucji danych

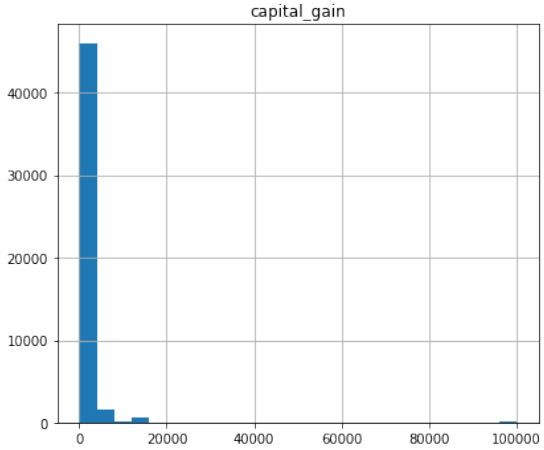


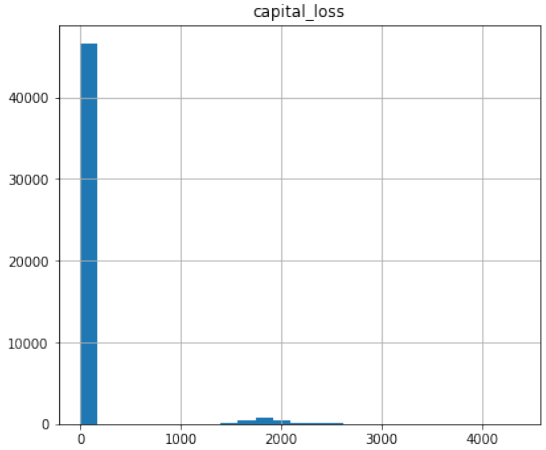
2.2 Rozkłady danych numerycznych

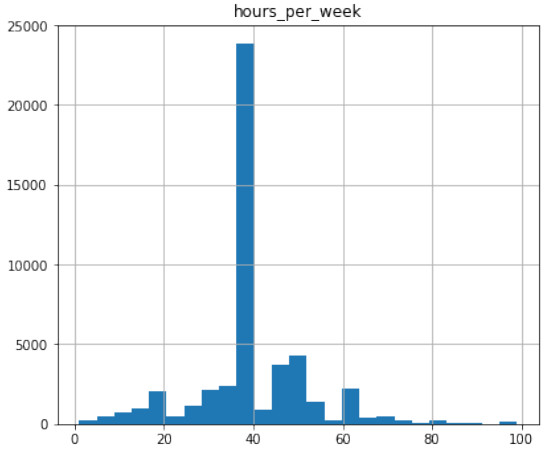










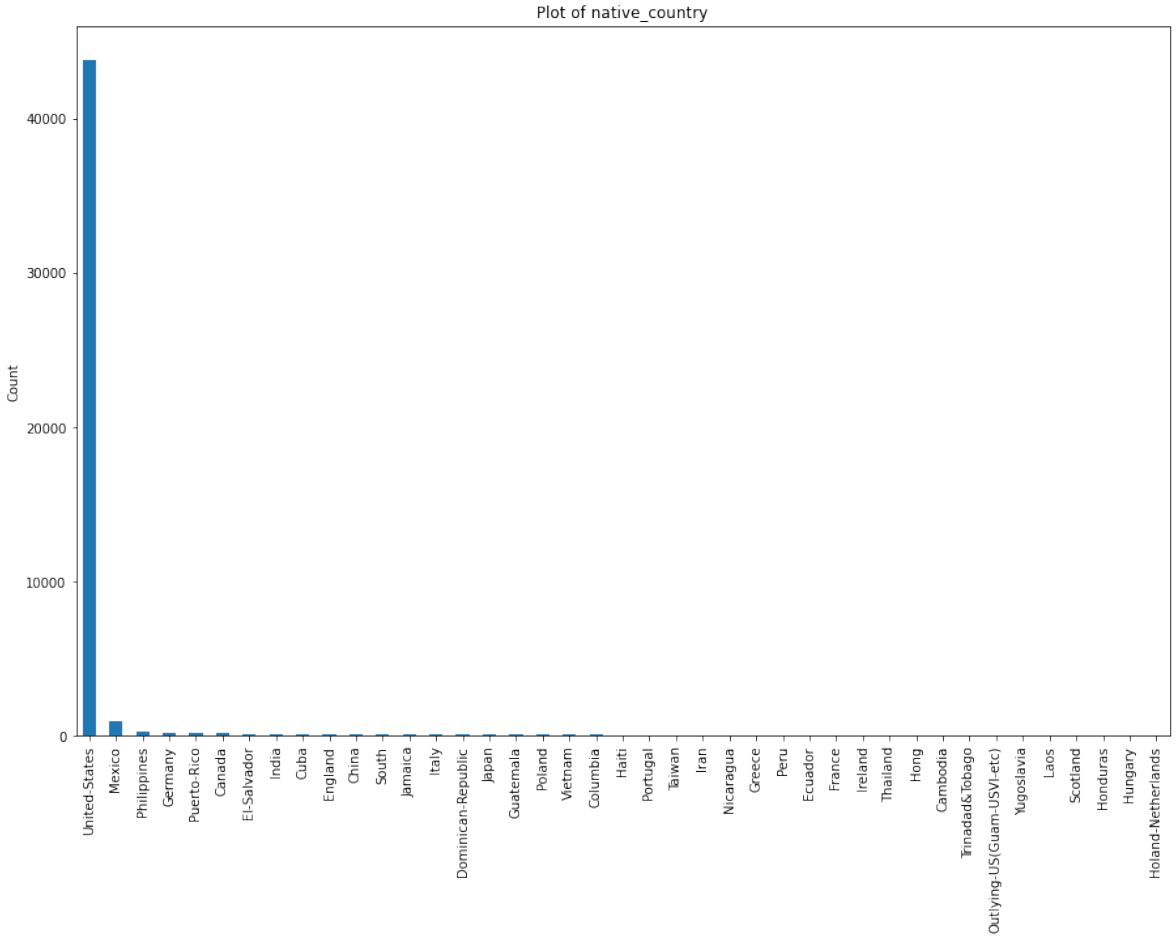


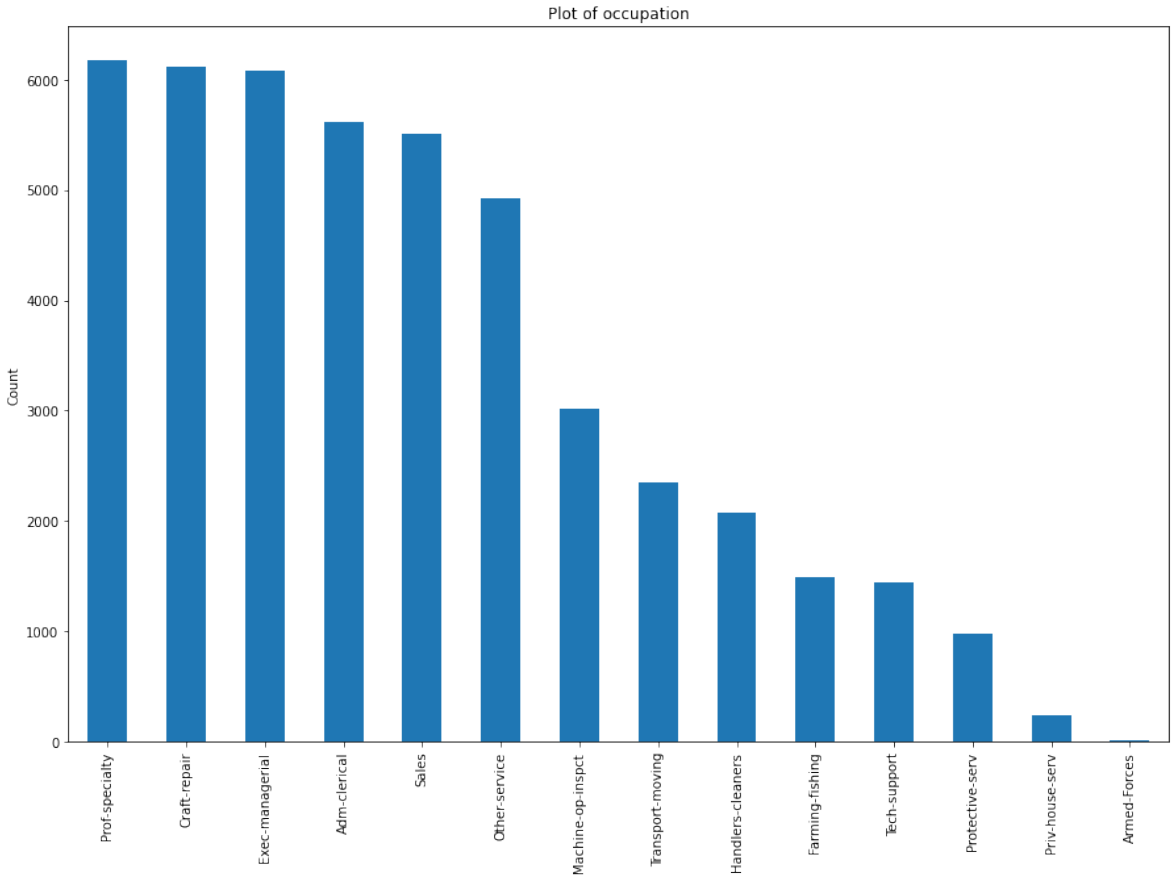
Z rozkładów zmiennych numerycznych można wyciągnąć kilka wniosków - zmienna 'age' ma rozkład zbliżony do normalnego, lecz ucięty w wartości 16, czego przyczyną

jest charakterystyka naszych danych. Kolejną konkluzją jest mała wartość informacyjna zmiennych 'capital gain' oraz 'capital loss', które w znacznej większości przypadków przyjmują wartości zerowe.

2.3 Rozkłady danych kategorycznych

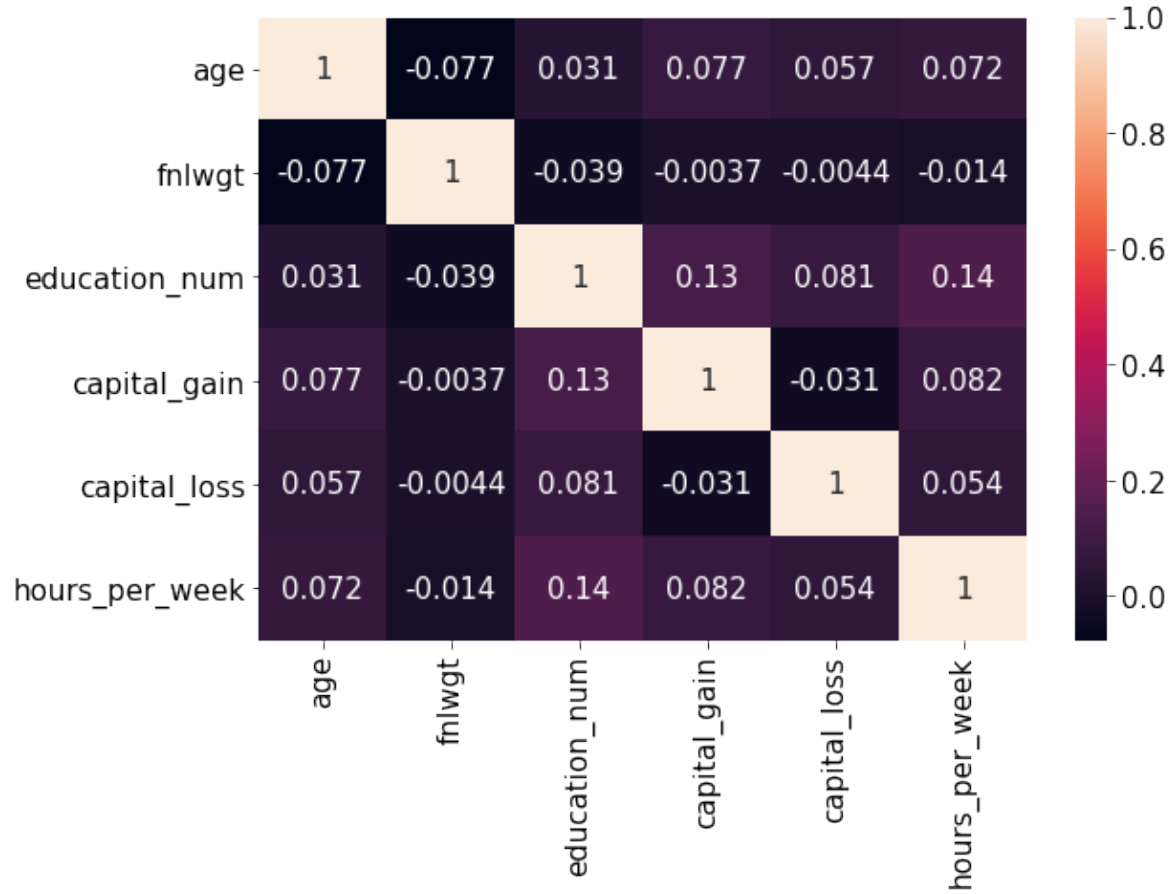
Wśród rozkładów cech kategorycznych główny wniosek opierał się na tym, że większość cech (oprócz 'occupation') skupiało się wokół jednej wartości.

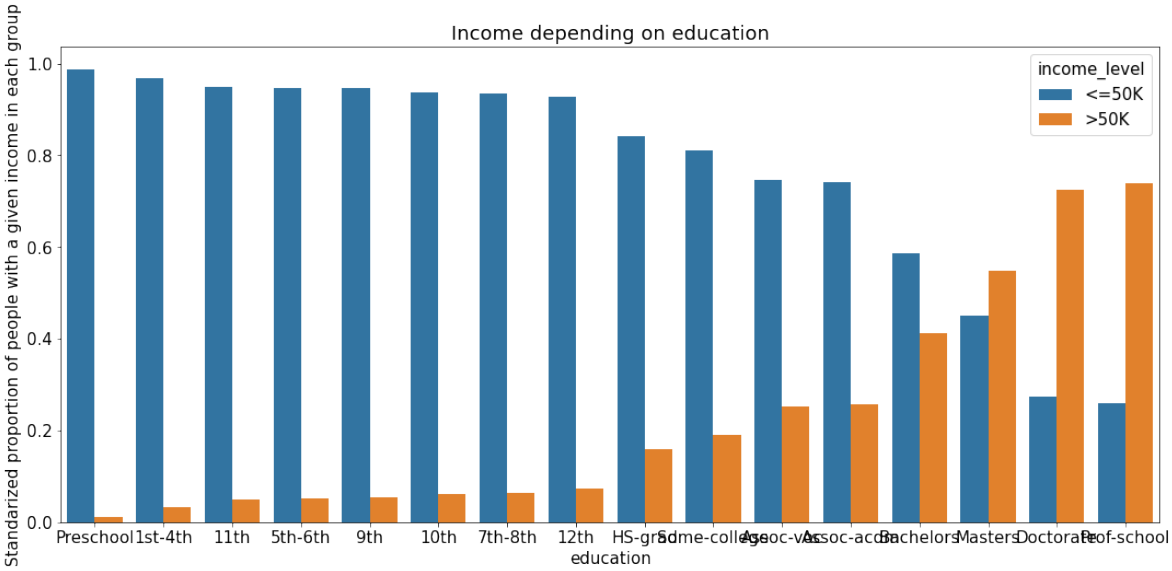




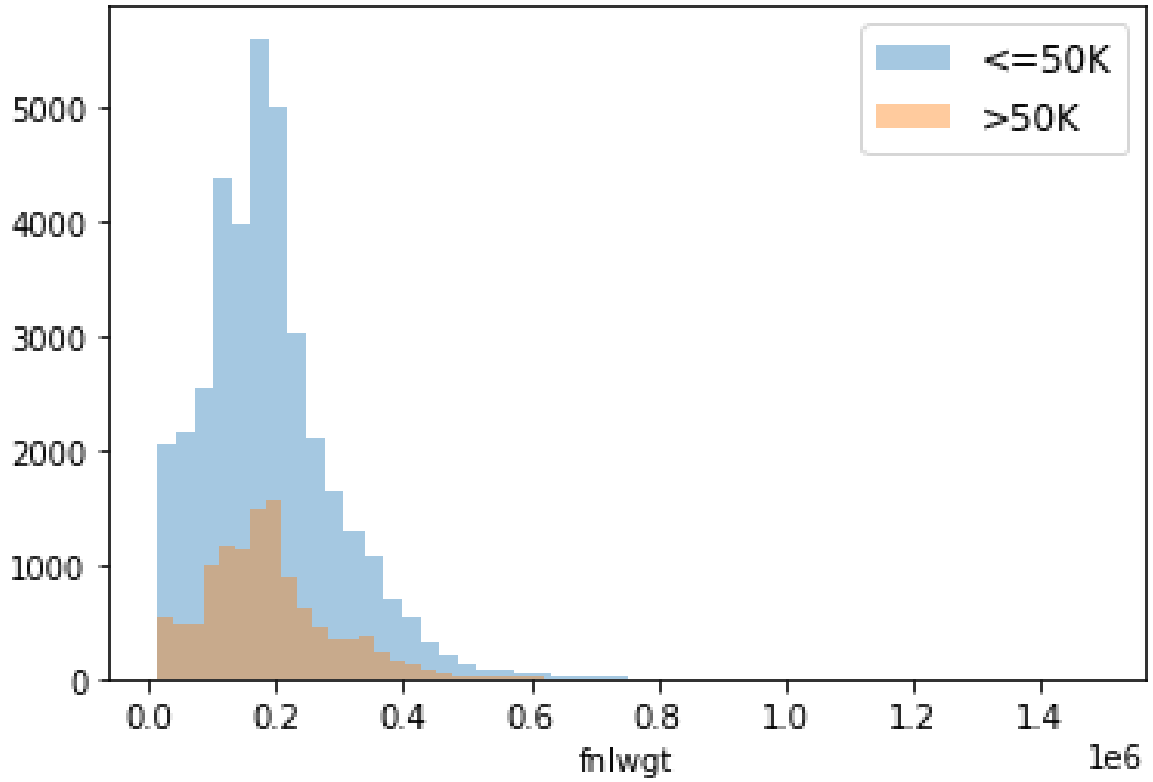
2.4 Szczegółowa analiza

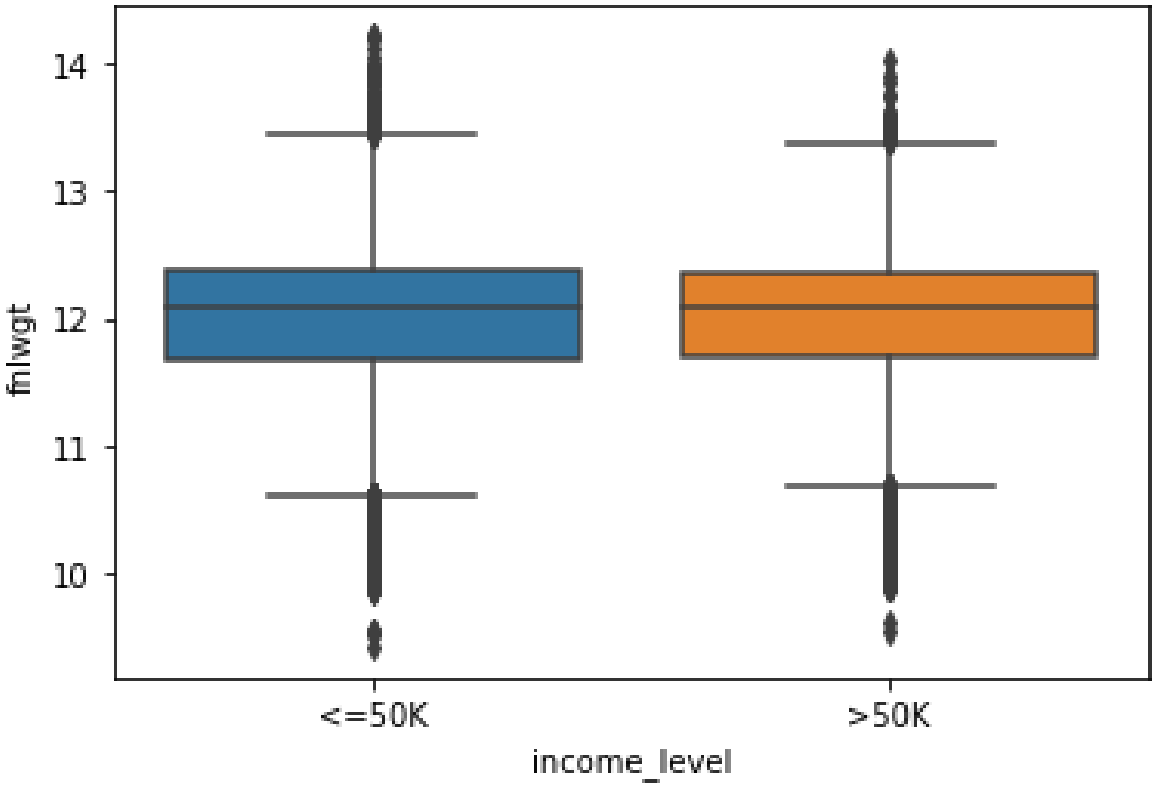
Szczegółowa analiza naszych danych odpowiedziała nam na kilka pytań - po pierwsze wśród naszych danych numerycznych nie ma większych korelacji. Zależności między cechami a targetem wydają się być logiczne i możliwe do przewidzenia (np. im wyższe wykształcenie, tym wyższe zarobki).





Ponadto postanowiliśmy przyjrzeć się dokładniej zmiennej 'fnlwgt'. Jak się okazało, zmienna ta dla obu klas ma niemal identyczne rozkłady oraz boxploty.


<=50K
>50K
5000
4000
3000
2000
1000
0
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
fnlwgt
1e6


14
13
12
11
10
fnlwgt
<=50K
>50K
income_level

3 Inżynieria cech

3.1 Imputacja danych i ograniczenie liczby rekordów

Jak się okazało nasz wyjściowy zbiór danych jest bardzo bogaty. Znajdują się w nim również wartości brakujące. Postanowiliśmy ograniczyć liczbę rekordów w dwóch etapach - usunęliśmy rekordy z wartościami brakującymi oraz z pozostałych danych wybraliśmy losową próbkę 5000 obserwacji.

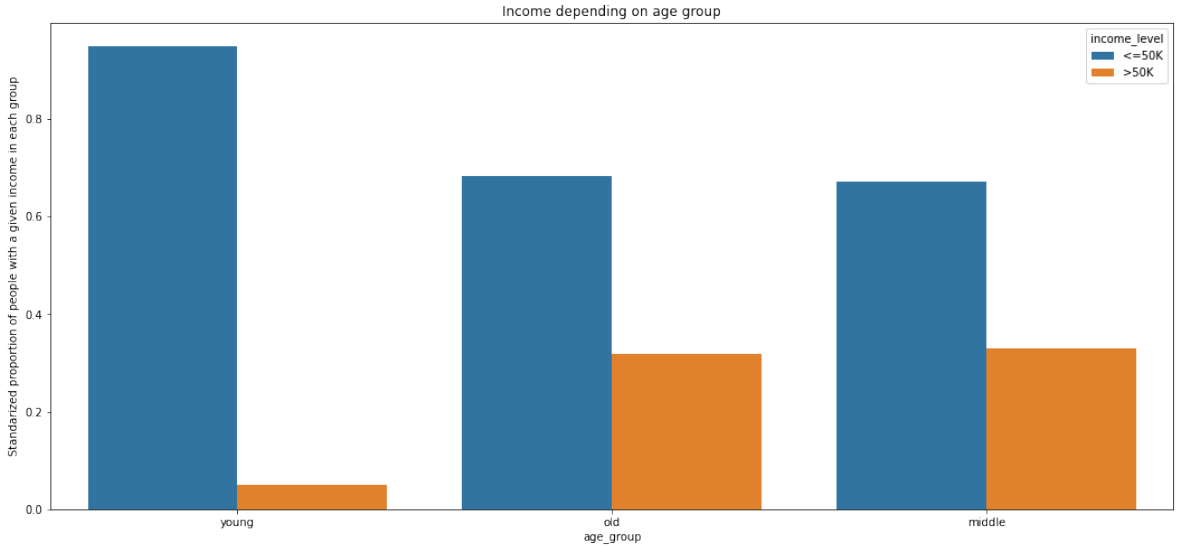
3.2 Ograniczenie liczby cech

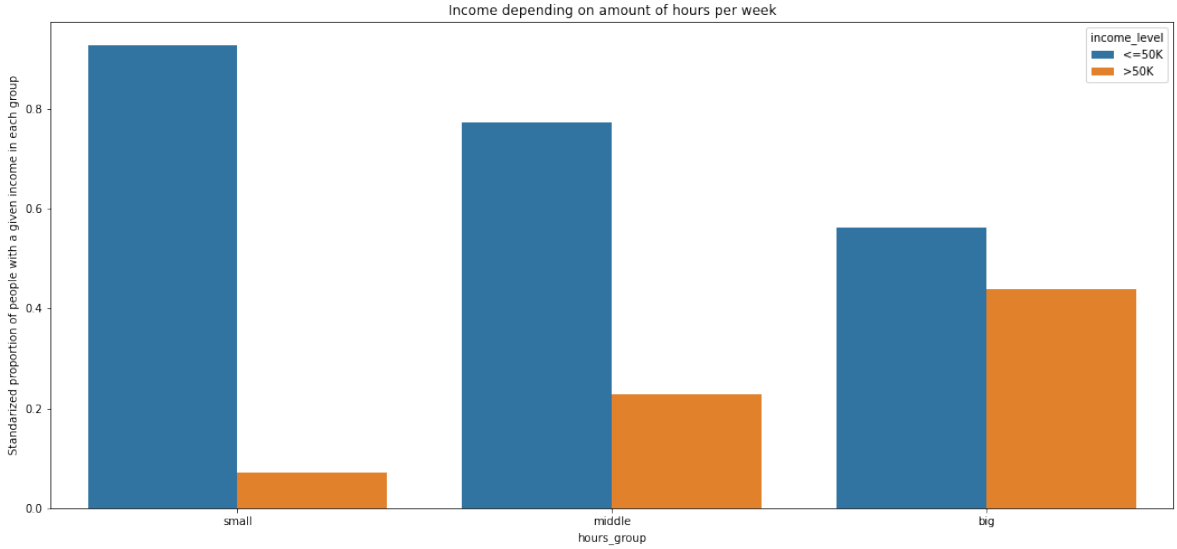
Poprzedni etap pracy - EDA, pozwolił nam lepiej zrozumieć nasze dane. Na podstawie wyciągniętych wniosków postanowiliśmy ograniczyć liczbę cech, gdyż część z nich uznaliśmy za mało informatywne. Ze zbioru danych usunęliśmy następujące zmienne:

1. 'native country' - dla ponad 90% rekordów przyjmuje ona wartość 'United States',
2. 'capital gain' i 'capital los' - dla ponad 90% rekordów przyjmują one wartości 0, a w innym przypadku nie rozdzielają jasno danych.
3. 'fnlwgt' - dla obu klas posiada ona niemal identyczne boxploty i rozkłady
4. 'education' - jest ona bezpośrednim odzwierciedleniem zmiennej 'education num'. W tym przypadku postanowiliśmy zachować w zbiorze danych cechę numeryczną.

3.3 Grupowanie i Outliery

W naszym zbiorze danych nie wyróżniliśmy wartości odstających, natomiast w ramach eksperymentu postanowiliśmy pogrupować zmienne 'age' oraz 'hours per week' na trzy podgrupy.





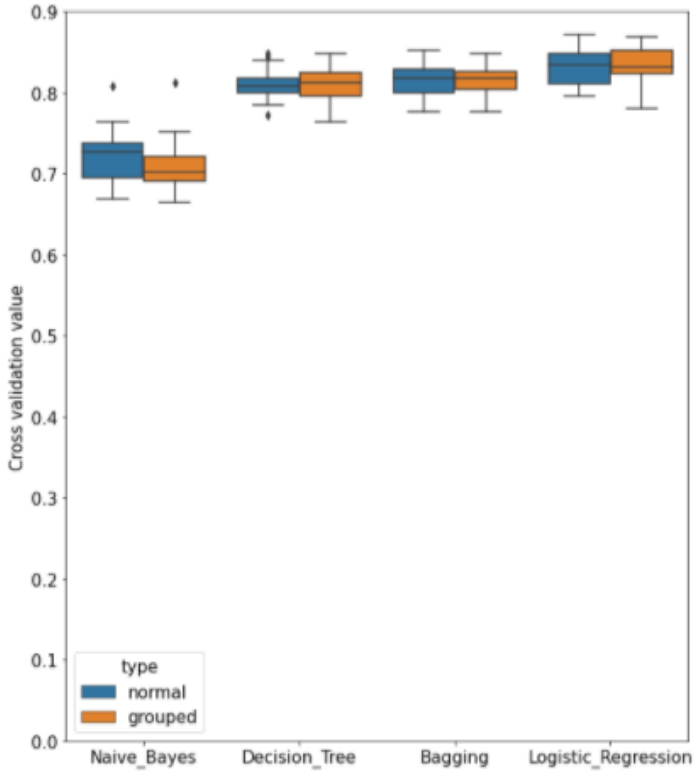
3.4 Kodowanie zmiennych kategorycznych

Po dokonaniu wszystkich operacji w naszym zbiorze niepogrupowanym zostało 6, zaś w zbiorze pogrupowanym 8 kolumn kategorycznych. Zawierały one stosunkowo niewiele wartości, zatem postanowiliśmy zakodować je za pomocą One-Hot Encoding. Ponadto zmapowaliśmy nasz target na zmienną binarną - 0 dla $<= 50$, 1 dla > 50.

4 Wybór modeli

4.1 Wstępne modelowanie

W ramach wstępnego modelowanie przetestowaliśmy cztery modele: **Naiwny Bayesowski**, **Drzewo decyzyjne**, **Bagging** oraz **Regresję Logistyczną**.
Jako miary używaliśmy Accuracy, co (jak się później okazało), było dość zgubne ze względu na niezbalansowanie naszych danych. Ostatecznie uznaliśmy, że było to raczej porównanie wydajności modeli trenowanych na danych zgrupowanych i niezgrupowanych, niż prawdziwe modelowanie. (Uznaliśmy, że zgrupowanie danych nie wpływa na jakość modeli i nie jest niezbędne.)



4.2 Automatyczne modelowanie

Na początku użyliśmy automatycznych narzędzi do wyboru modeli. Zarówno TPOT i HyperOptEstimator wskazały Random Forest z pewnymi parametrami. Jednak zdecydowaliśmy się pominąć te parametry, bo okazało się, że lekko pogarszają wyniki.
Dodatkowo postanowiliśmy użyć regresji logistycznej oraz xgboosta, który na początku został wskazany przez wyżej wymienione narzędzia.

4.3 Finalne modelowanie

Jak wspominano podczas EDA, nasze dane są bardzo niezbalansowane- klasa **>50k** jest znacznie mniej liczna od klasy **<=50k**, co może wpłynąć na dalsze modelowanie. By temu zapobiec, postanowiliśmy przetestować metodę SMOTE, która w sposób sztuczny wygenerowała rekordy dla klasy mniej licznej. Mieliśmy nadzieję, że nawet sztuczne zbalansowanie danych wpłynie pozytywnie na jakość modeli. Jak się jednak później okazało, nie wpłynęło to na znaczącą poprawę wyników.
Dość szybko zauważyliśmy, że zarówno Accuracy, jak i AUC daje bardzo wysokie, niemiarodajne wyniki, dlatego zdecydowaliśmy się używać Precision Recall Curve oraz miary AUPRC do mierzenia skuteczności naszych modeli.

4.3.1 XGBoost i Regresja Logistyczna

Oba modele zachowują się podobnie. Dały wyniki rzędu 0,7 dla AUPRC i dosyć dobrą confusion matrix. Niestety nie radzą sobie z przewidywaniem **>50k** i przewidują po połowie dobrze i źle. Natomiast doskonale radzą sobie z **<=50k**.

4.3.2 Drzewo Decyzyjne i Las Losowy

Oba modele dały doskonałe wyniki AUPRC rzędu 0,9. Niestety nie pokrywa się to z precision i recall, które dały wyniki około 0,5. Dodatkowo jest to sprzeczne z confusion matrix, która dała wyniki niemal idealne. Nie jesteśmy pewni czym jest to spowodowane. Być może jest pewien błąd przy wyliczaniu precision i recall albo confusion matrix.

4.3.3 Strojenie

Postanowiliśmy wystroić drzewo przy pomocy Grid Search i Random Search. Niestety znacząco pogorszyło to wyniki, które wróciły do 0,7 jak przy XGBoostcie i Regresji. Podobnie zachowała sie confusion matrix, która też dla **>50k** dała wyniki po połowie.

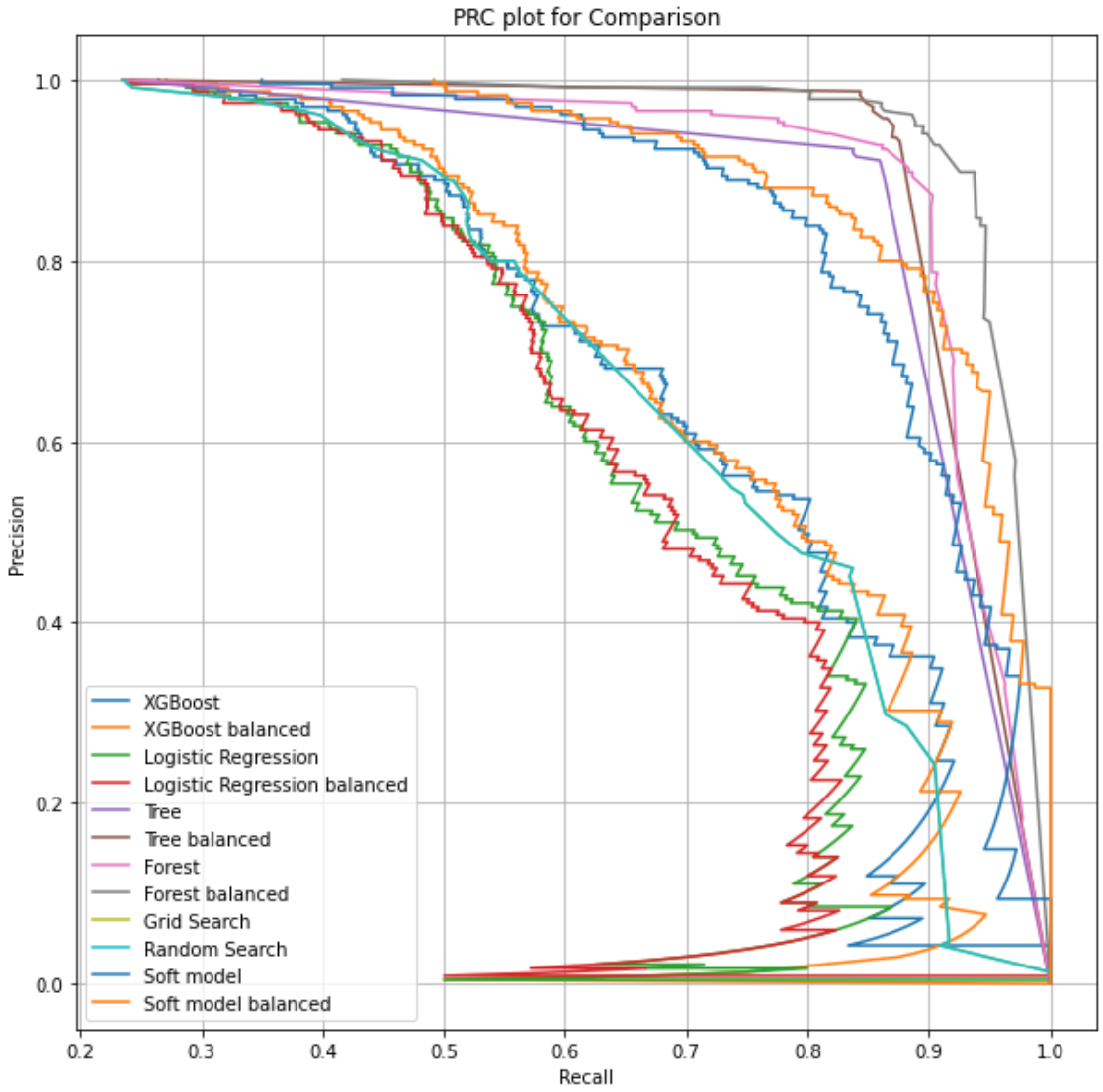
4.3.4 Komitety

Użyliśmy też "miękiego" głosowania opartego na Drzewie, Lesie i XGBoostcie. Dało to dosyć wysokie wyniki AUPRC: 0,88 i świetną confusion matrix. Jednak wyniki są wciąż gorsze od Lasu Losowego oraz dalej pozostaje problem z wyliczaniem precision/recall.

5 Finalna ocena i podsumowanie

Ostatecznie stwierdziliśmy, że najlepsze wyniki daje wskazany też przez automatyczne narzędzia Random Forest, co pokazują przede wszystkim świetne naszym zdaniem wykresy PRC oraz Confusion Mtarix.
Jak już wspominaliśmy, zbalansowanie danych nie wpłynęło znacząco na jakość wyników, szczególnie w przypadku modeli drzewiastych. Aczkolwiek Random Forest się jeszcze polepszył po zbalansowaniu do 0,96 na AUPRC.

Finalne wyniki:
AUPRC for XGBoost is 0.7355161782808399
AUPRC for XGBoost balanced is 0.738688504892318
AUPRC for Logistic Regression is 0.6734093365620358
AUPRC for Logistic Regression balanced is 0.6639577917116464
AUPRC for Tree is 0.8988890071229277
AUPRC for Tree balanced is 0.9285814680016053
AUPRC for Forest is 0.9211060686787919
AUPRC for Forest balanced is 0.9607736073966594
AUPRC for Grid Search is 0.7319344425721797
AUPRC for Random Search is 0.731934442572179
AUPRC for Soft model is 0.884860375290089
AUPRC for Soft model balanced is 0.9200955861333395


PRC plot for Comparison
Precision
Recall
1.0
0.8
0.6
0.4
0.2
0.0
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
XGBoost
XGBoost balanced
Logistic Regression
Logistic Regression balanced
Tree
Tree balanced
Forest
Forest balanced
Grid Search
Random Search
Soft model
Soft model balanced