Predykcja zarobków mieszkańców Stanów Zjednoczonych

Jakub Piwko, Malwina Wojewoda

Projekt w ramach przedmiotu

Wstęp do uczenia maszynowego

semestr letni 2021/2022

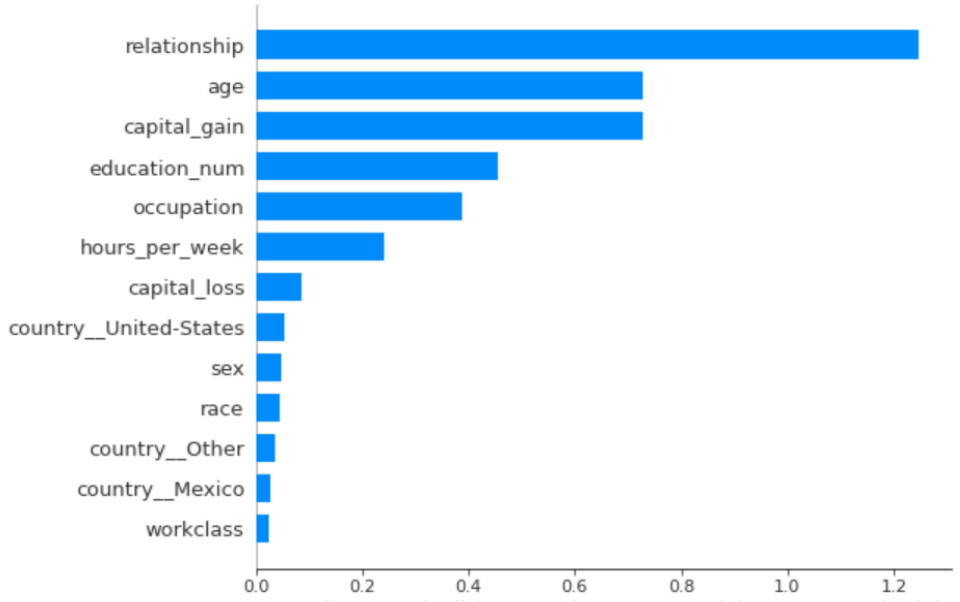
Wydział Matematyki i Nauk Informacyjnych
POLITECHNIKA WARSZAWSKA

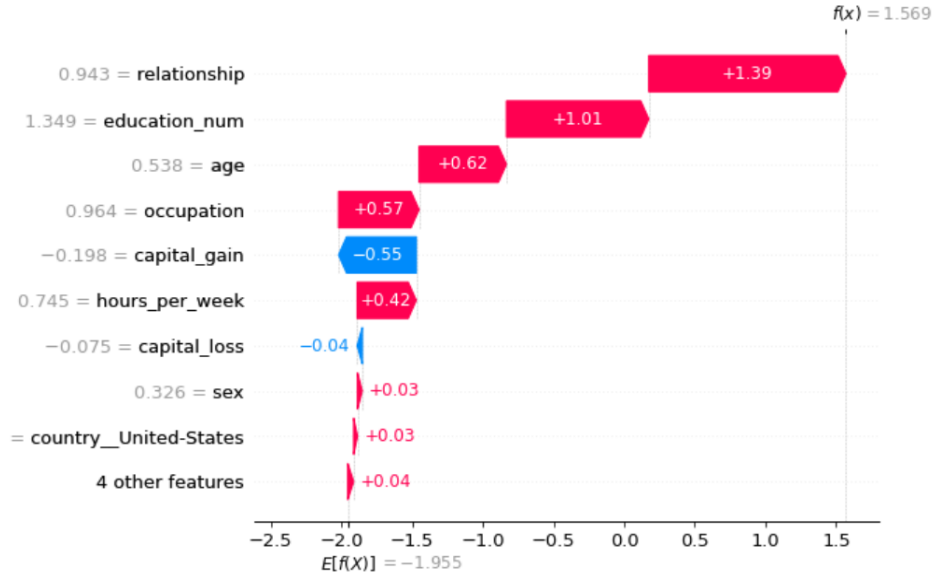
Opis zadania

Celem zadania jest przygotowanie modelu określającego, czy dany mieszkaniec Stanów Zjednoczonych zarobił mniej czy więcej niż 50 000$.

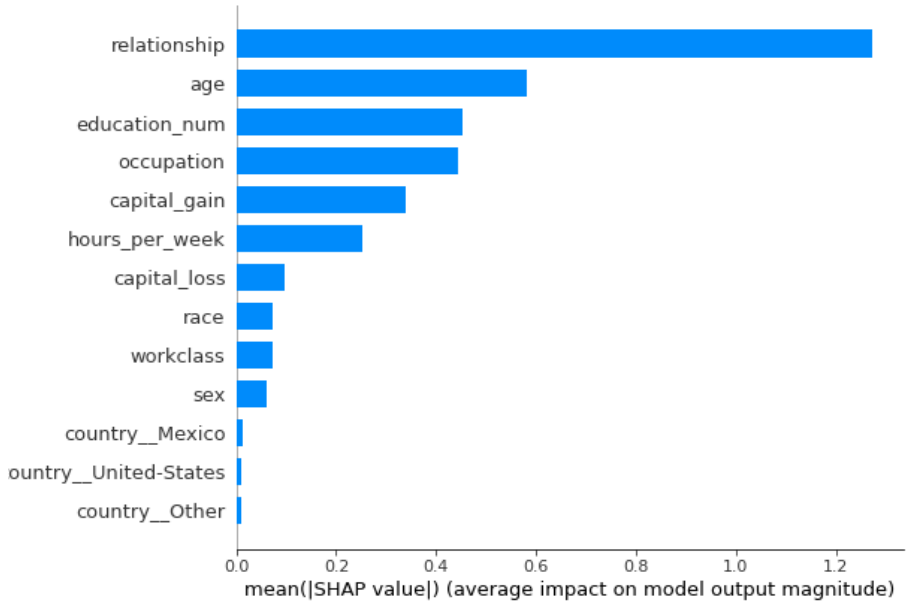
Co wpływa na zarobki?

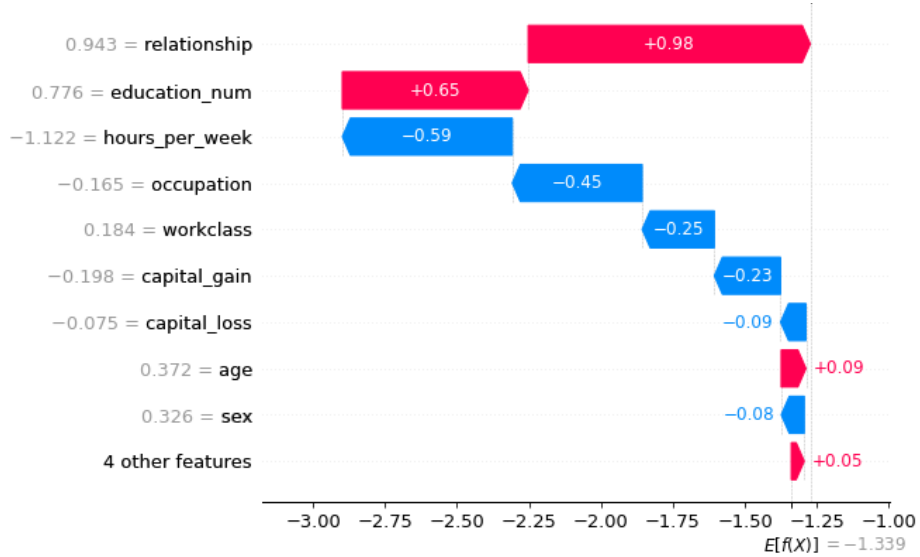
Regresja logistyczna





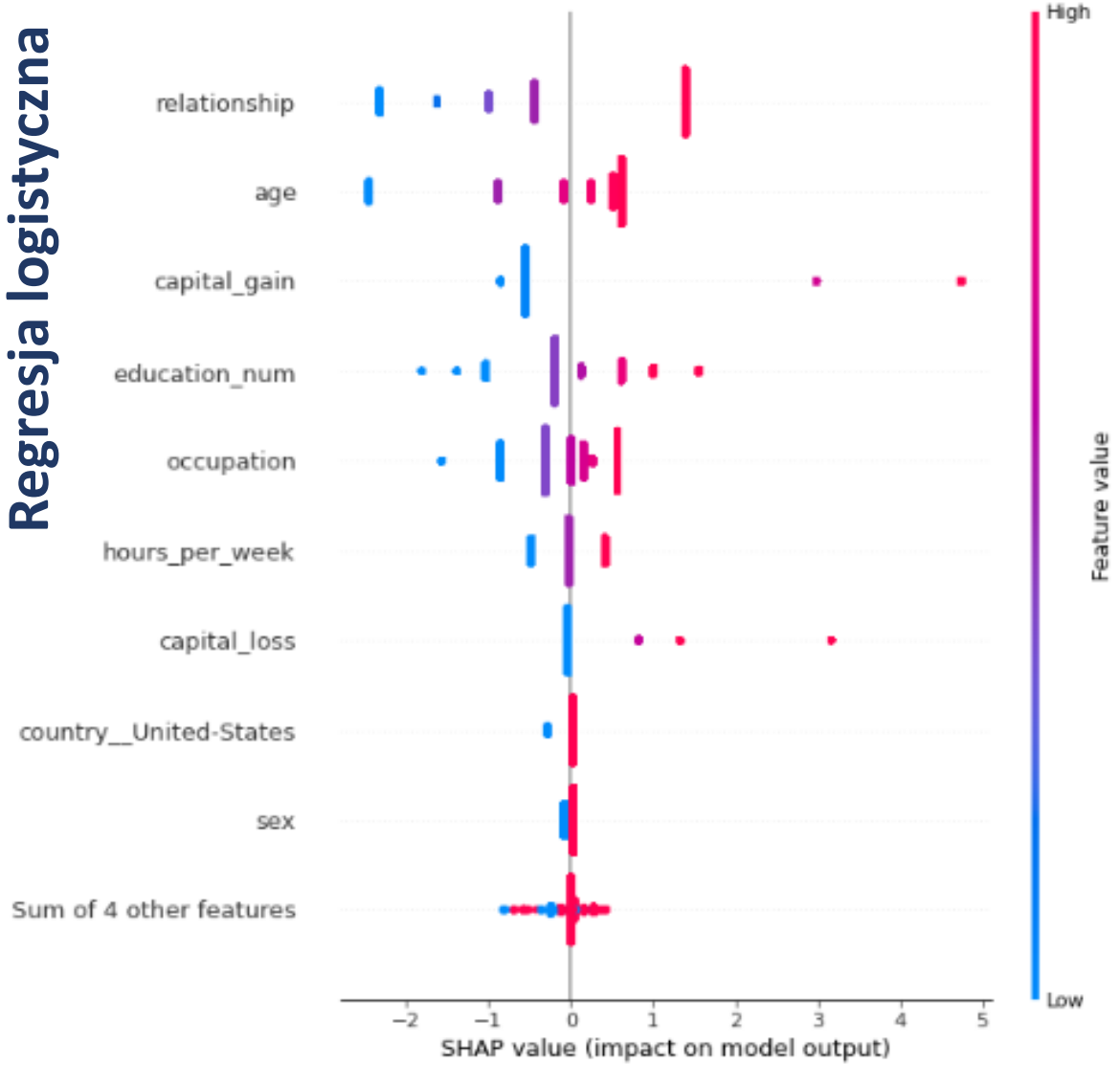
XGBoost



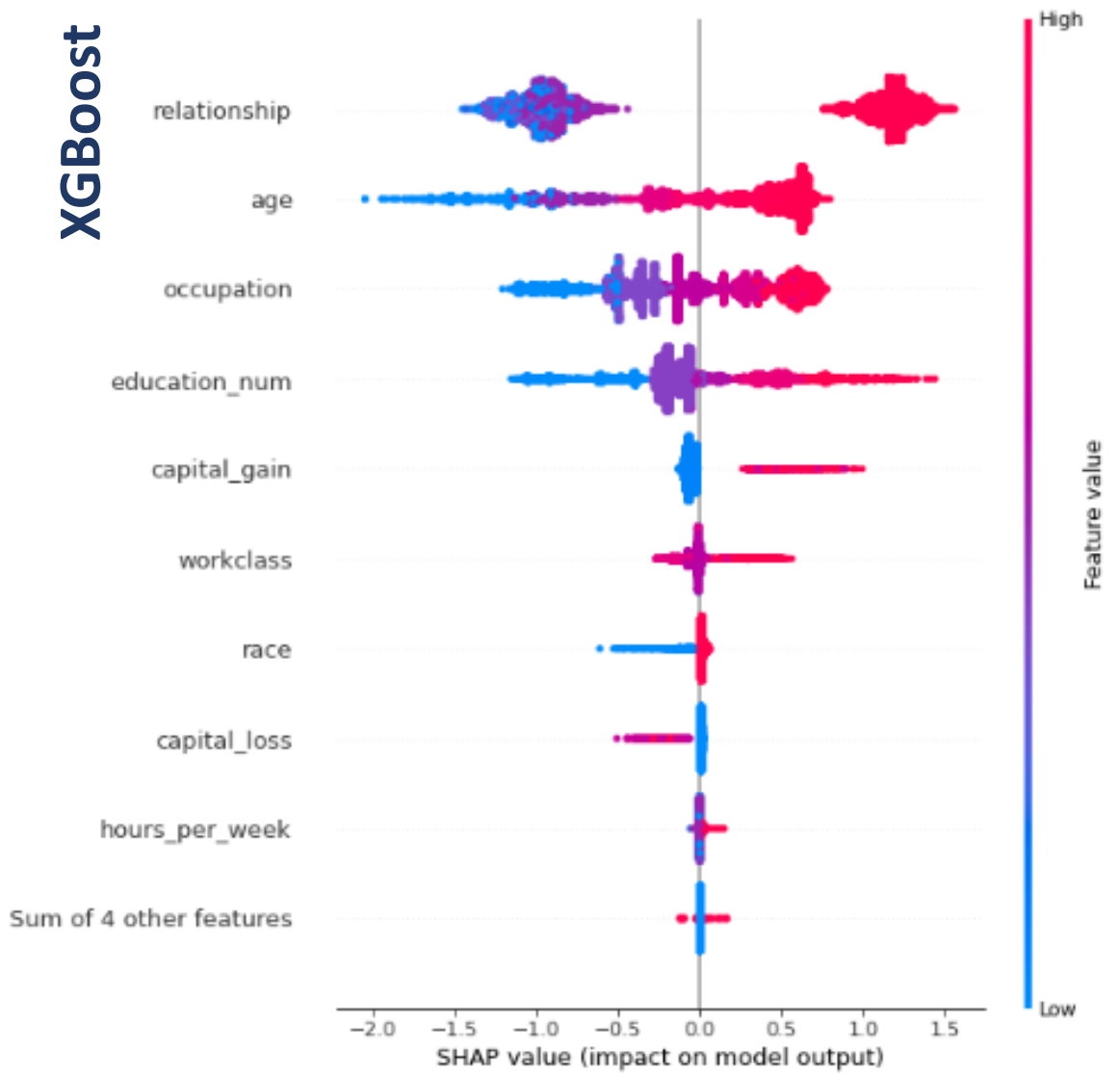


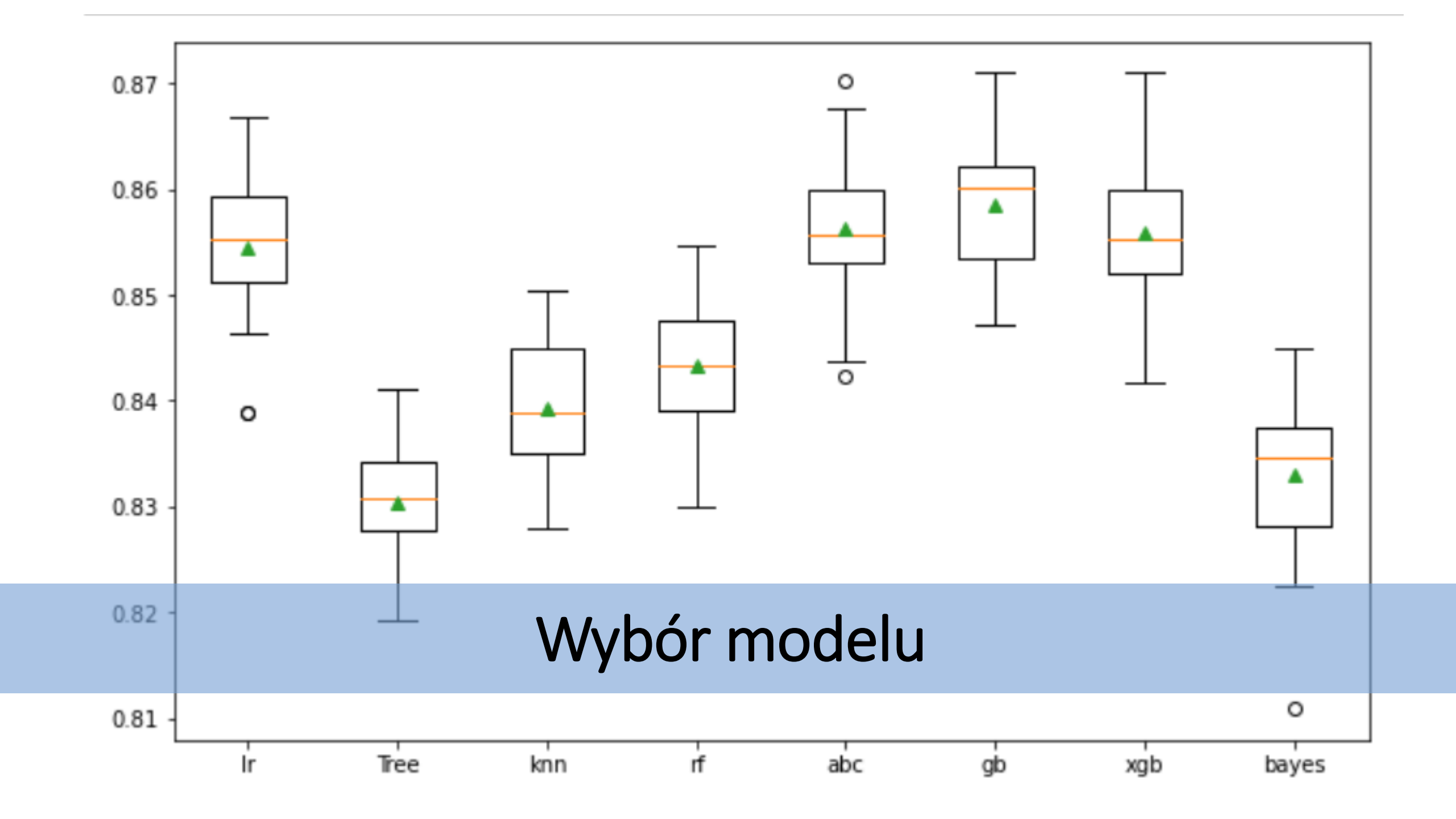
Co wpływa na zarobki?

Regresja logistyczna



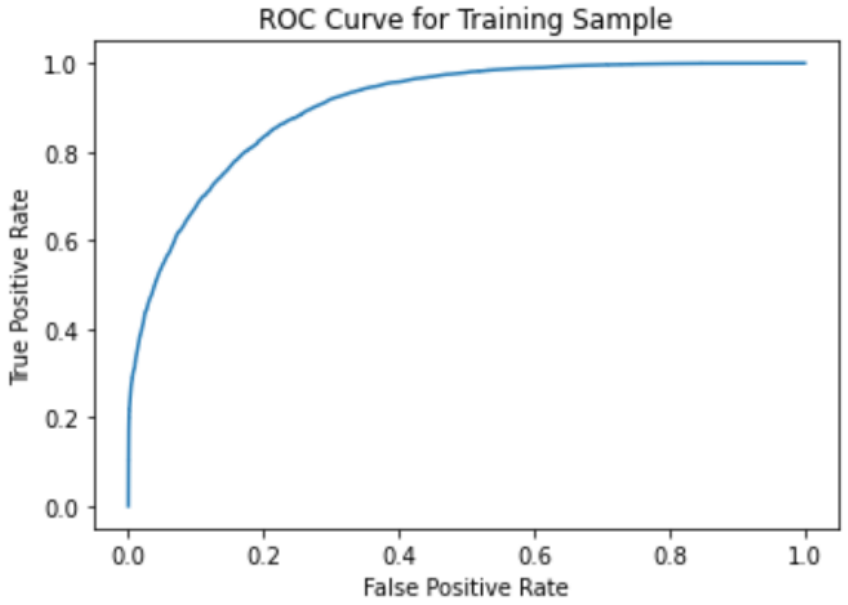
XGBoost



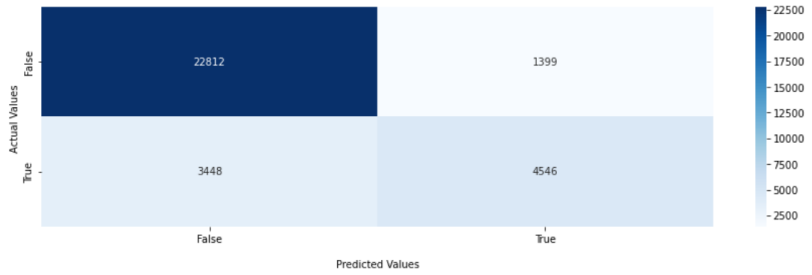


Wybór modelu

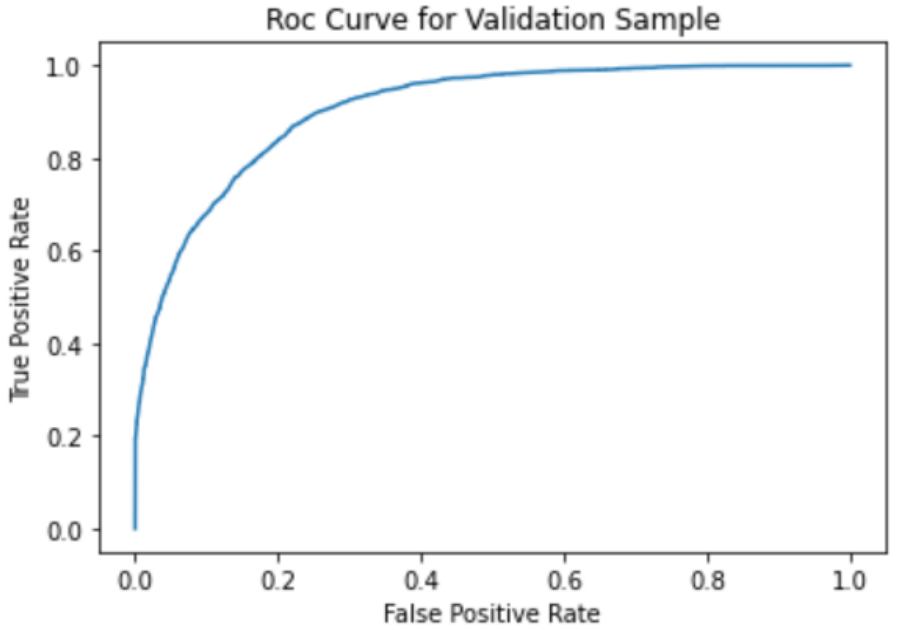
Predykcyjność



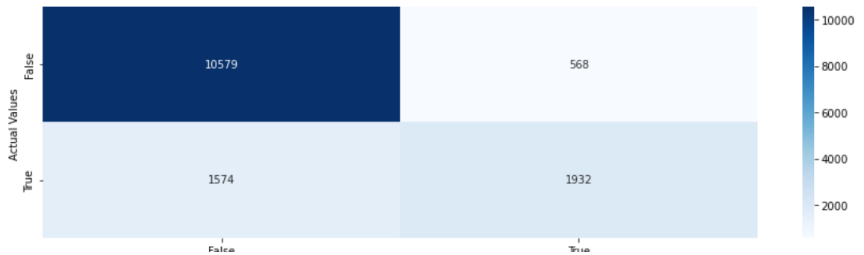
gini_test: 0.8084



ZBIÓR TRENINGOWY:
Accuracy score: 0.8494954199658438
Precision score: 0.7646761984861228
roc_auc score: 0.7554464276607769

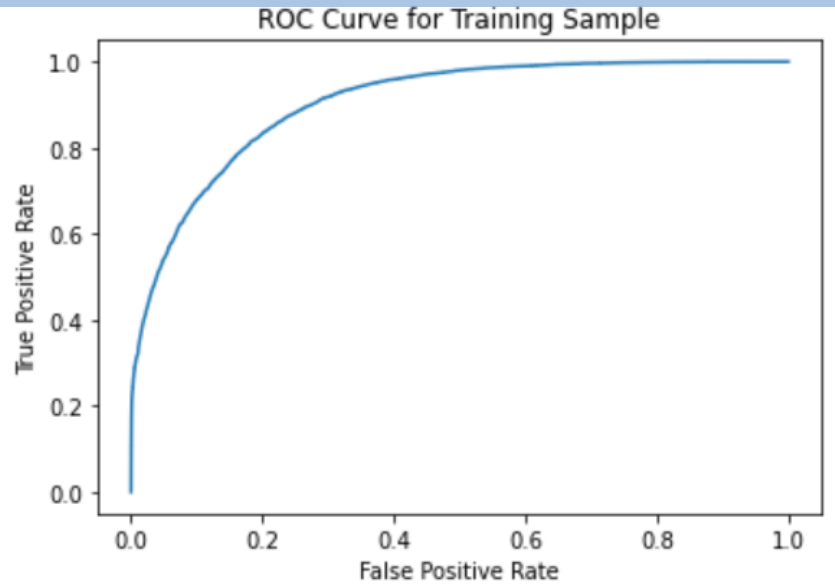


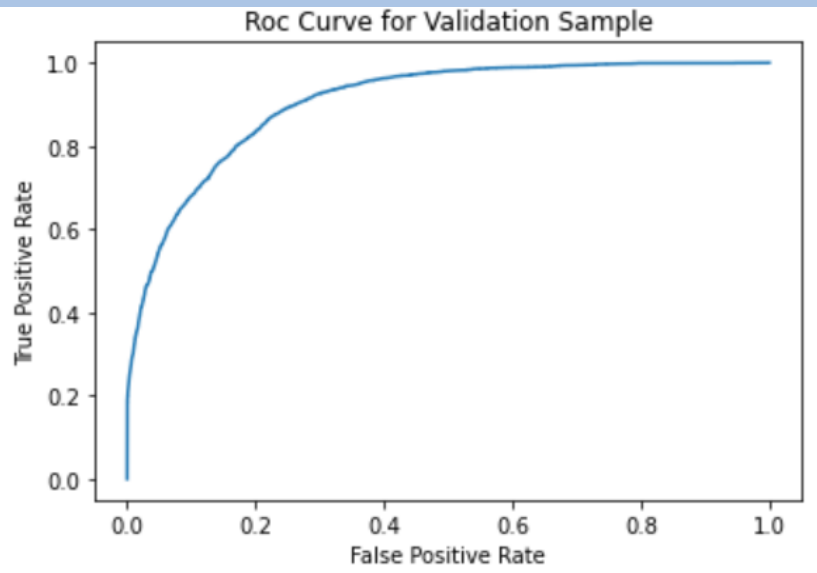
gini_val: 0.8117



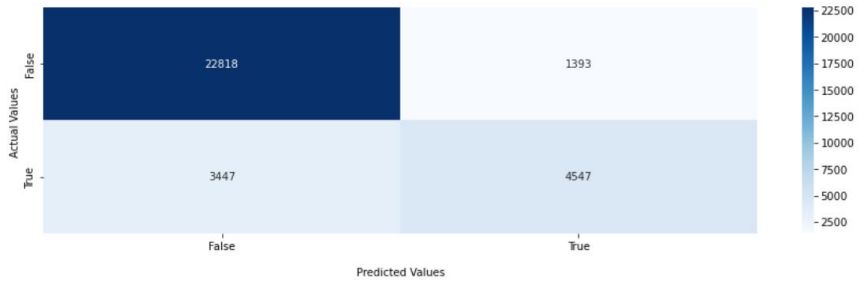
ZBIÓR TESTOWY:
Accuracy score: 0.8538183307172592
Precision score: 0.7728
roc_auc score: 0.7500499598504475

Predykcyjność
(uwzględnienie kolumn płeć i rasa)

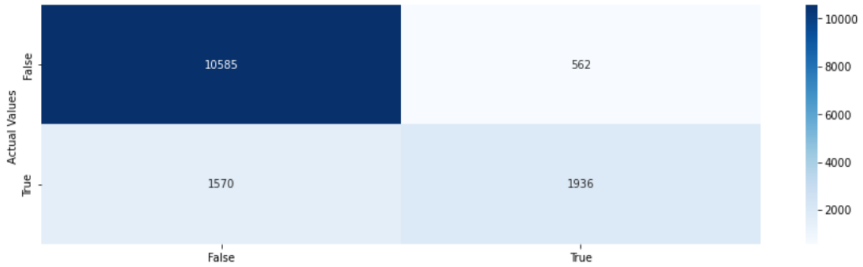




gini_test: 0.8088



gini_val: 0.8115



ZBIÓR TRENINGOWY:
Accuracy score: 0.8497127775190187
Precision score: 0.7654882154882154
roc_auc score: 0.7556328851900997

ZBIÓR TESTOWY:
Accuracy score: 0.8545007848222207
Precision score: 0.7750200160128102
roc_auc score: 0.7508895412142795

Dane, EDA, preprocessing

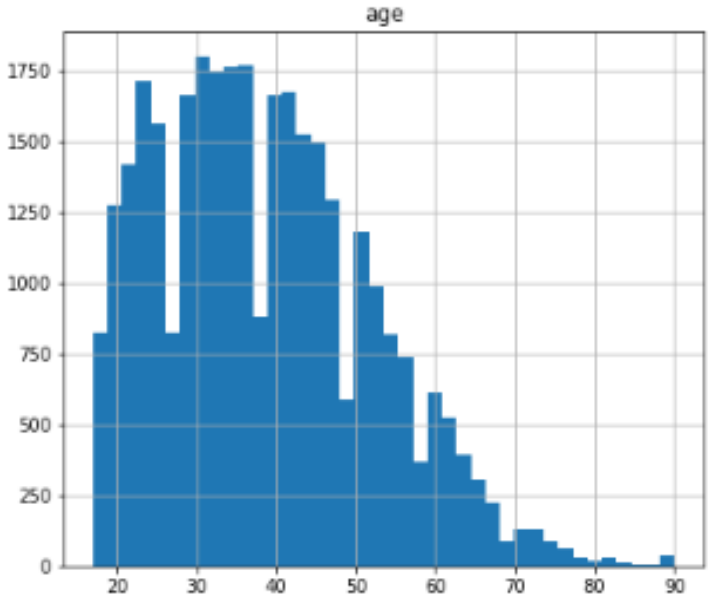
Dane

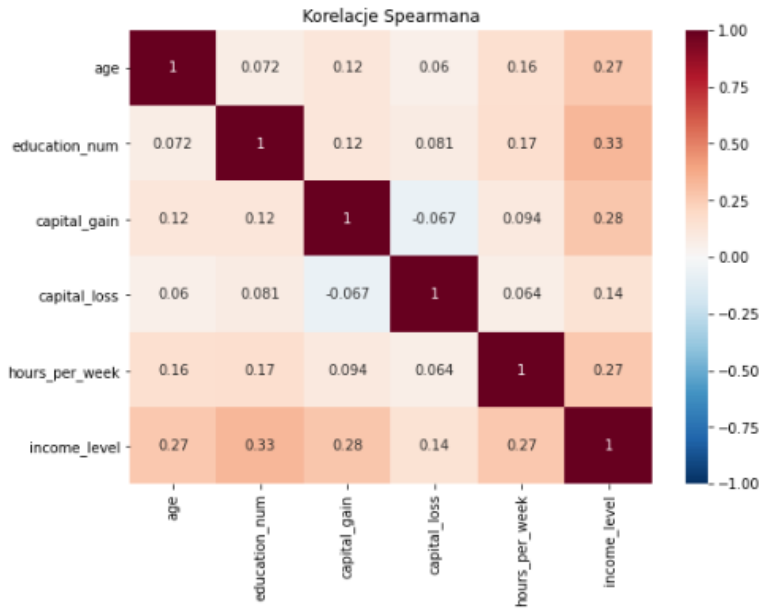
- 8842 rekordów i 15 kolumn
- dane z 1996 roku
- zawarte kontrowersyjne dane dotyczące płci i rasy

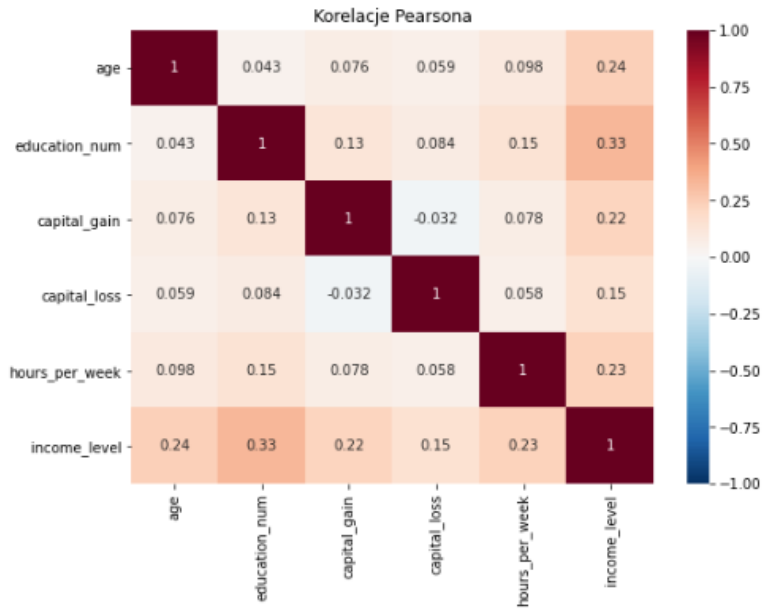
	age	workclass	fnlwgt	education	education_num	marital_status	occupation	relationship	race	sex	capital_gain	capital_loss
0	39	State-gov	77516.0	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174.0	0.0
1	50	Self-emp-not-inc	83311.0	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0.0	0.0
2	38	Private	215646.0	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0.0	0.0
3	53	Private	234721.0	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0.0	0.0
4	28	Private	338409.0	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...

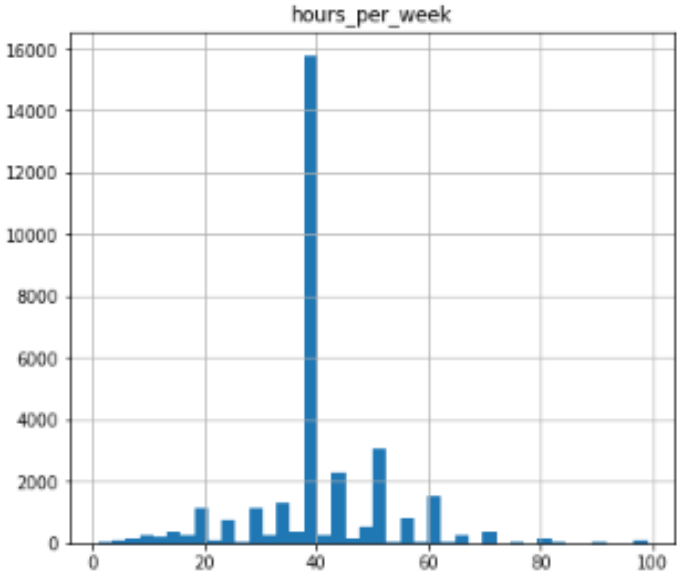
Badanie rozkładów

- Rozkłady zmiennych ciągłych
- Usunięcie zmiennej "fnlwgt"
- Zbadanie korelacji
- Zamiana zmiennej celu



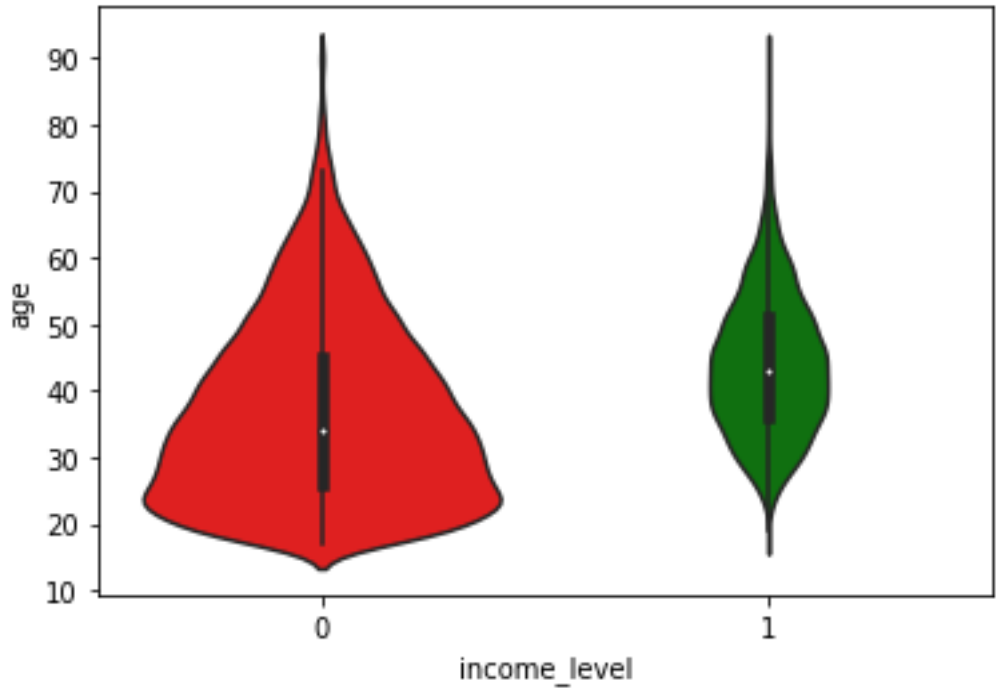


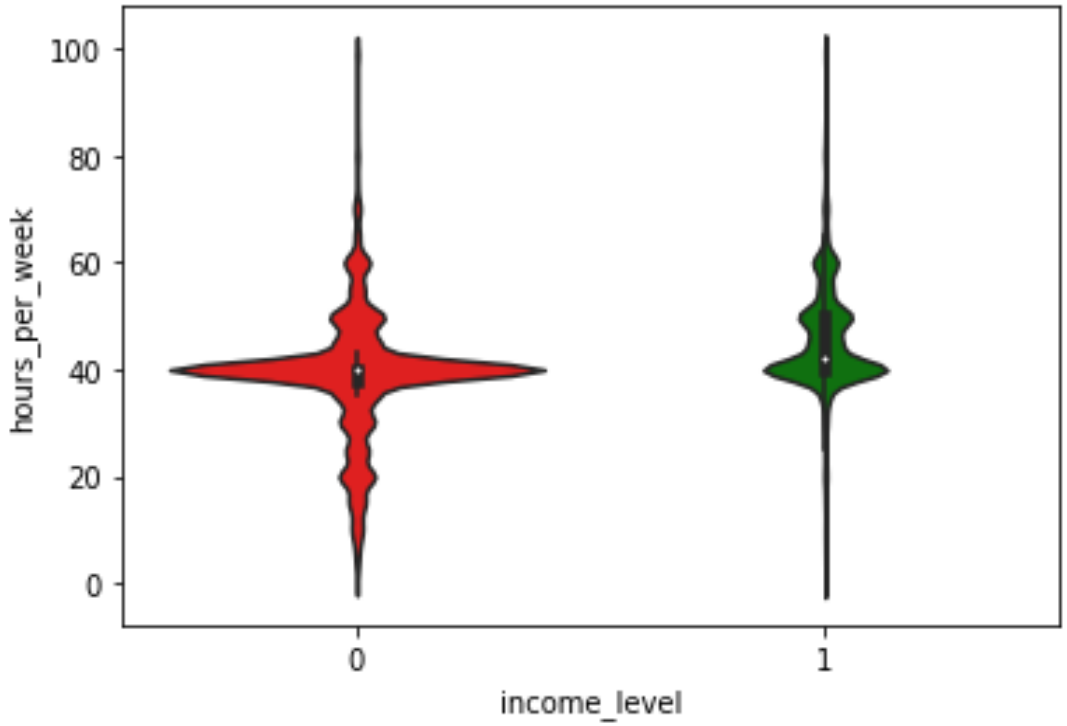




Szukanie zależności i transformacje zmiennych ciągłych

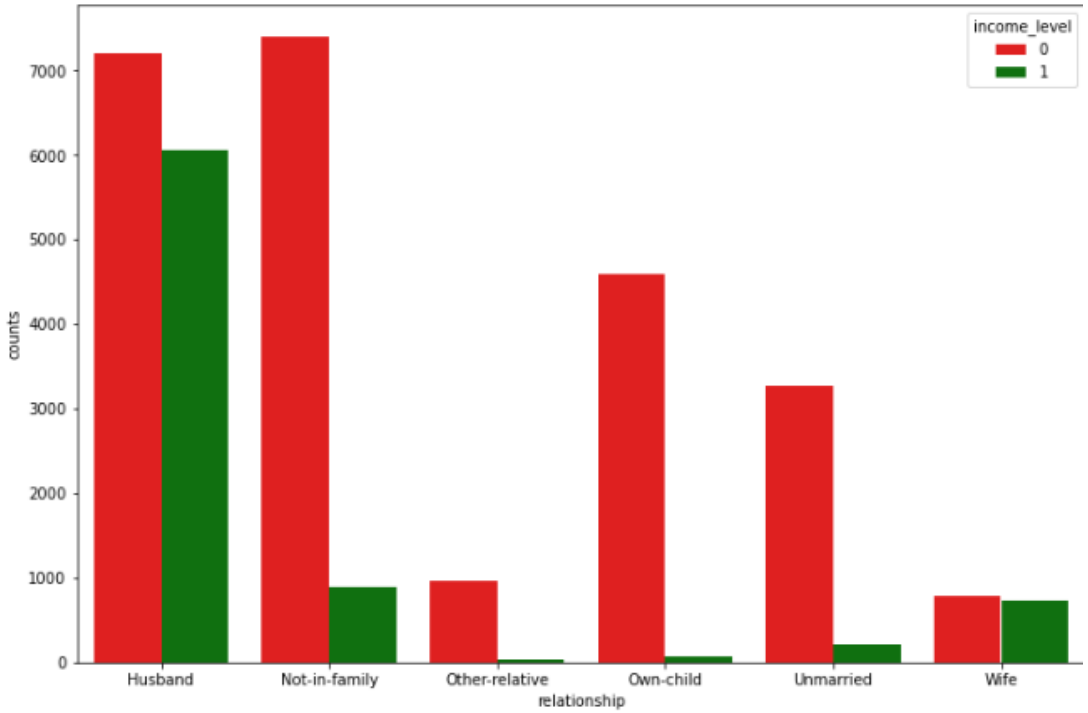
- Badanie rozkładów z uwzględnieniem zmiennej celu
- Kategoryzacja zmiennych po odpowiednich kwantylach

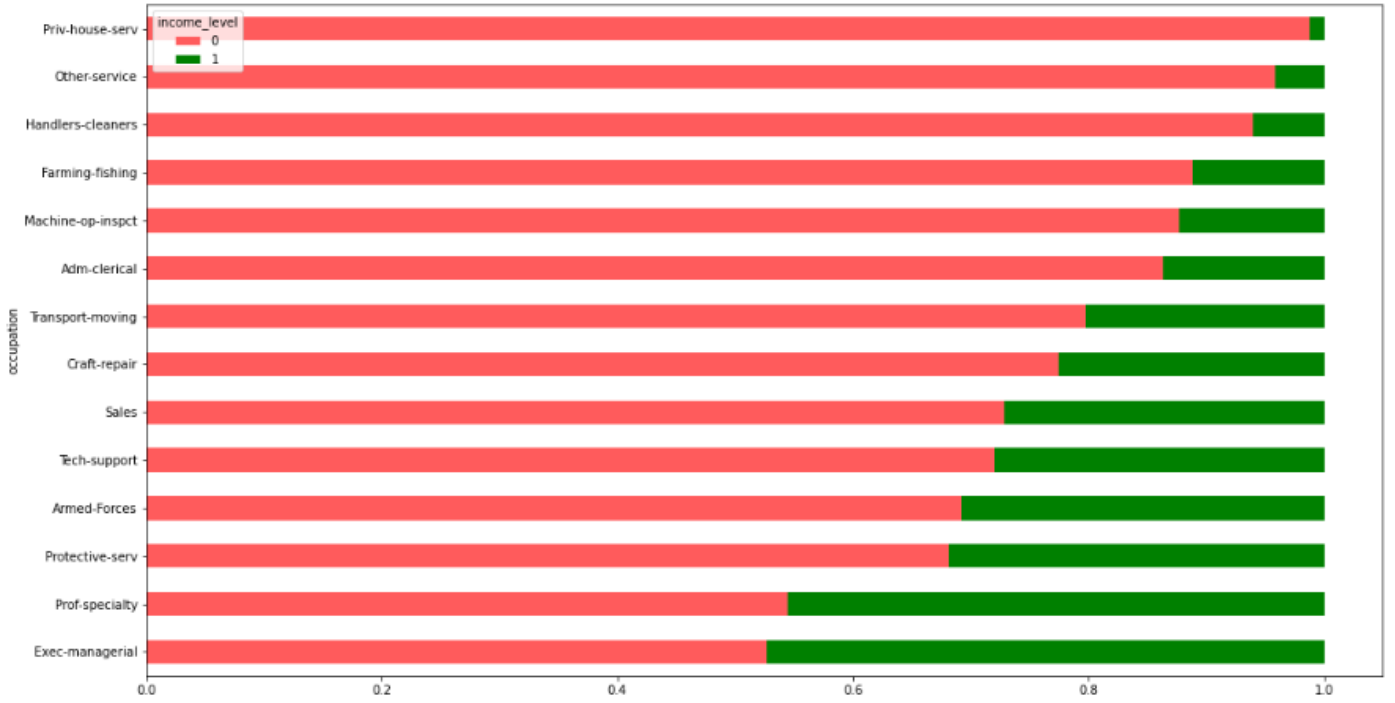




Szukanie zależności w zmiennych kategorycznych

- Badanie rozkładów z uwzględnieniem zmiennej celu
- Szukanie zmiennych, które mogą decydować o wysokości przychodu





Najważniejsze spostrzeżenia i wnioski

Lepiej wykształceni zarabiają więcej

Osoby, które przepracowują więcej godzin, zarabiają więcej

Osoby o wyższym przychodzie są starsze

Większość osób w próbce pochodzi ze Stanów Zjednoczonych, zredukowaliśmy zmienną do 3 grup: USA Mexico, Others i zakodowaliśmy zmienną one-hot-encodingiem

Wśród osób w związku małżeńskim proporcja osób lepiej zarabiających do mniej zarabiających jest największa

Lepiej zarabiają osoby z sektora prywatnego, na stonowiskach kierowniczych lub pracujące w specjalizacji

Wśród lepiej zarabiających jest więcej mężczyzn i osób rasy białej lub z Azji I Pacyfiku

Weight of Evidence & Information Value

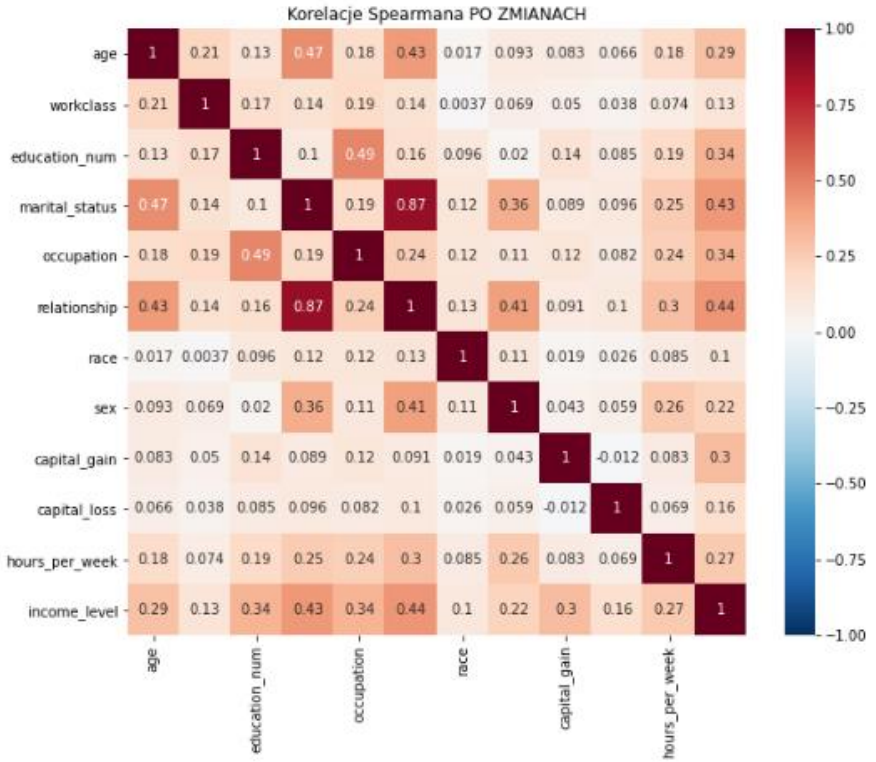
- Wykorzystanie Weight of Evidence do połączenia zmiennych w mniejsze kategorie i zakodowania ich z odpowiednimi wagami.
- Wykorzystanie Information Value do oceny siły predykcji zmiennych.

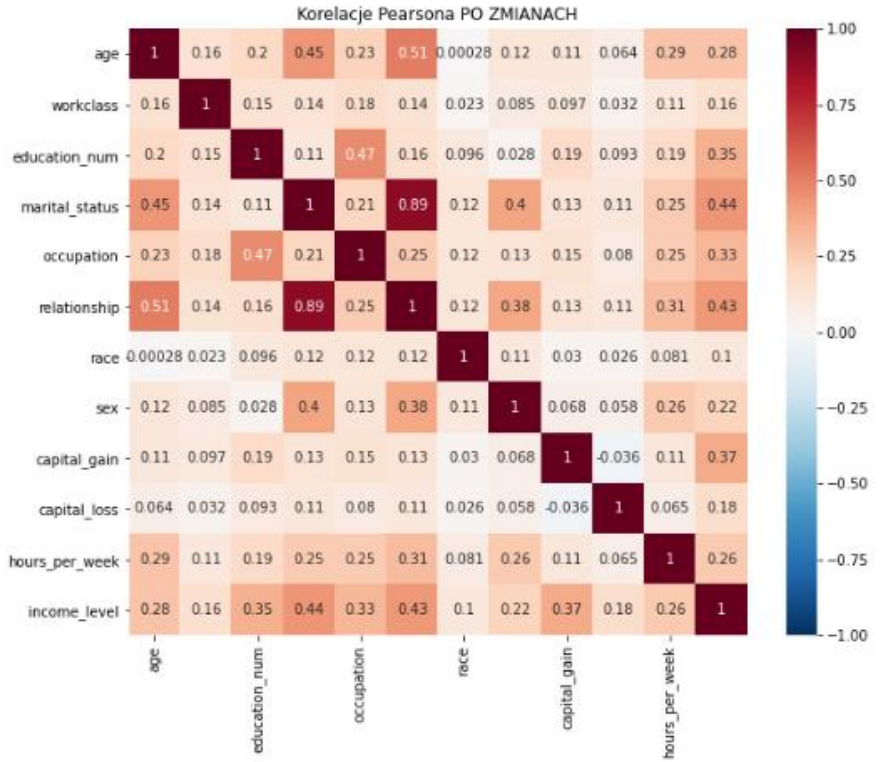
income_level relationship	0	1	WoE	IV
Own-child	4584.0	75.0	-3.004723	0.540710
Other-relative	966.0	33.0	-2.268540	0.081148
Unmarried	3266.0	218.0	-1.598710	0.172064
Not-in-family	7407.0	894.0	-1.006359	0.195336
Husband	7201.0	6051.0	0.934119	0.429243
Wife	787.0	723.0	1.023297	0.059287
Total	NaN	NaN	NaN	1.477788

income_level relationship	0	1	WoE	IV
1	4584.0	75.0	-3.004723	0.540710
2	966.0	33.0	-2.268540	0.081148
3	3266.0	218.0	-1.598710	0.172064
4	7407.0	894.0	-1.006359	0.195336
5	7988.0	6774.0	0.943267	0.488096
Total	NaN	NaN	NaN	1.477355

Analiza po zmianach

- Zbadanie korelacji po preprocessingu
- Usunięcie zmiennej "marital_status", która silnie korelowała ze zmienną "relationship"



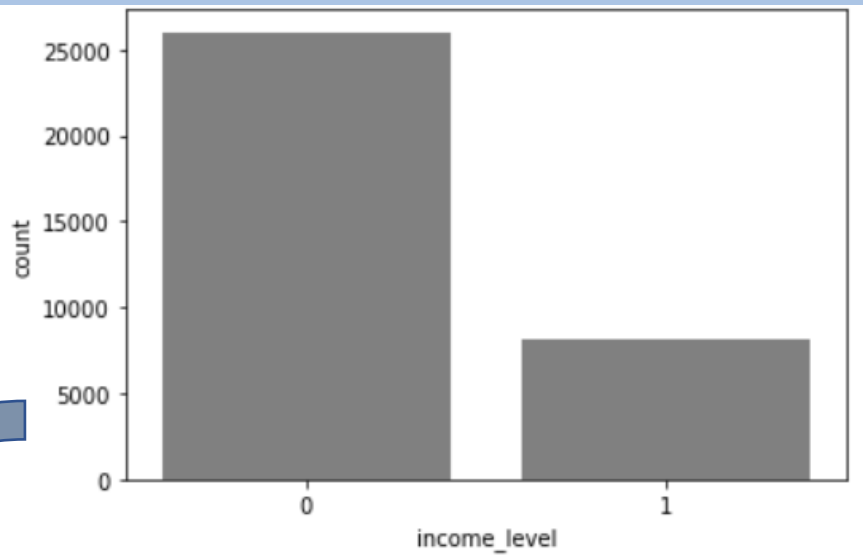


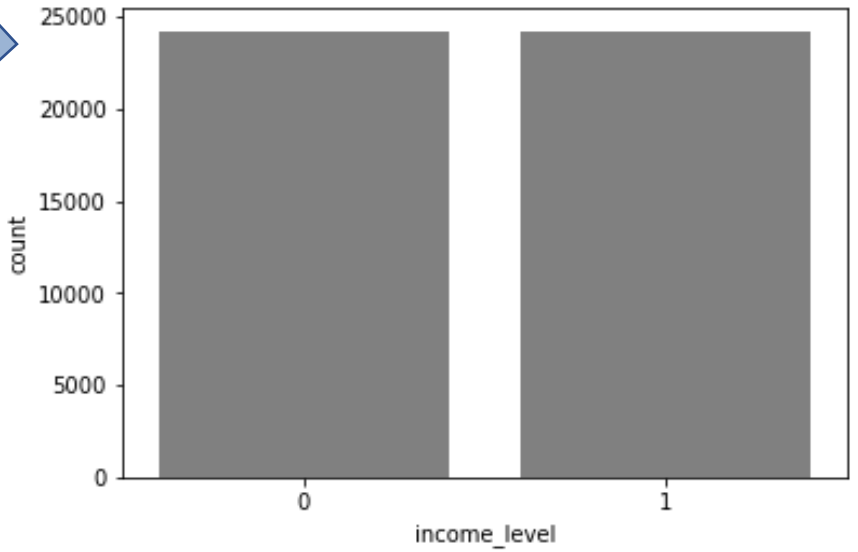
Modelowanie

- Zastosowanie oversamplingu do zrównoważenia klas w zmiennej celu
- Zastosowanie TPOT

```
tpot = TPOTClassifier(generations=5,verbosity=2)
tpot.fit(census_df, y_train)
```

```
Generation 1 - Current best internal CV score: 0.8544635926098432

Generation 2 - Current best internal CV score: 0.8544635926098432

Generation 3 - Current best internal CV score: 0.854867256637168

Generation 4 - Current best internal CV score: 0.854867256637168

Generation 5 - Current best internal CV score: 0.8551156652693681

Best pipeline: XGBClassifier(input_matrix, learning_rate=0.1, max_depth=9, min_child_weight=4, n_estimators=100, n_jobs=1, s
ubsample=0.45, verbosity=0)

TPOTClassifier(generations=5, verbosity=2)
```





Dziękujemy za uwagę!