WUM - Projekt 1

MICHAŁ GROMADZKI, MATEUSZ FLIS

Temat Projektu

1. Zbiór danych – Census Income.
2. Dane pochodzą z 1994 ze Stanów Zjednoczonych.
3. Kolumny zawierają podstawowe informacje o osobie – wiek, płeć, wykształcenie, narodowość itd.
4. Predykcja polega na określeniu czy dany obywatel zarabia więcej lub mniej niż 50 tyś $ rocznie.

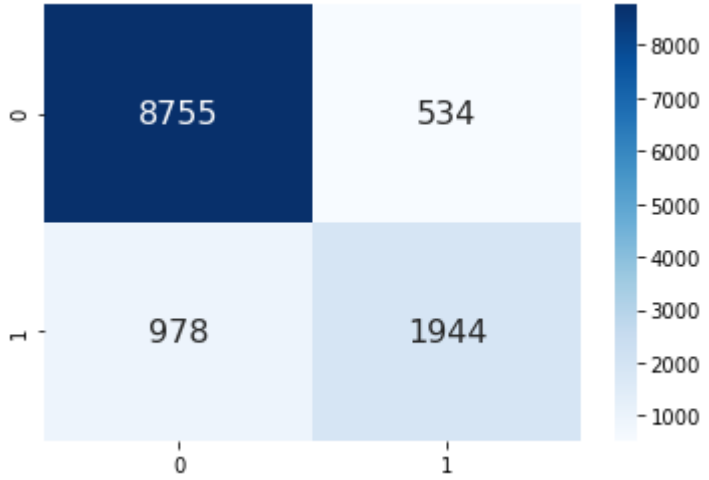
	age	workclass	fnlwgt	education	education_num	marital_status	occupation	relationship	race	sex	capital_gain	capital_loss	hours_per_week	native_country	income_level
0	39	State-gov	77516.0	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174.0	0.0	40.0	United-States	<=50K
1	50	Self-emp-not-inc	83311.0	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0.0	0.0	13.0	United-States	<=50K
2	38	Private	215646.0	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0.0	0.0	40.0	United-States	<=50K
3	53	Private	234721.0	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0.0	0.0	40.0	United-States	<=50K
4	28	Private	338409.0	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0.0	0.0	40.0	Cuba	<=50K
5	37	Private	284582.0	Masters	14	Married-civ-spouse	Exec-managerial	Wife	White	Female	0.0	0.0	40.0	United-States	<=50K

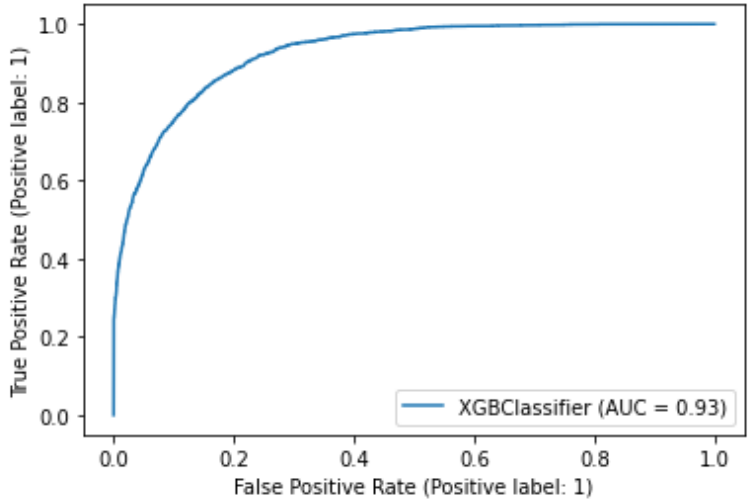
Znaczenie poszczególnych kolumn

"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"

Najlepszy model trenowany na zbiorze treningowym i testowym

Accuracy	0.876
Precision	0.784
F1	0.72
Recall	0.665





Eksploracyjna analiza danych

1. Znaczenie poszczególnych kolumn
2. Czy występują braki danych?
3. Korelacje między kolumnami
4. Rozkład zmiennych ciągłych
5. Rozkład zmiennych kategorycznych
6. Zależności między dwiema i więcej zmiennymi

Modele 

1. Zaimplementowane modele:
   - LogisticRegression
   - RandomForest
   - XGBoost
   - Hard voting
   - Soft voting
   - Stacking
   - Bagging
2. Znalezienie optymalnych hiperparametrów dla LogisticRegression, RandomForest i XGBoost za pomocą GridSearch
3. Wybranie najlepszego modelu

Ewaluacja 🧮

1. Wykorzystywane metryki:
   - Accuracy
   - Precision
   - F1
   - Recall
   - Gini
2. Confusion matrix
3. Interpretowalność

Najlepsze modele

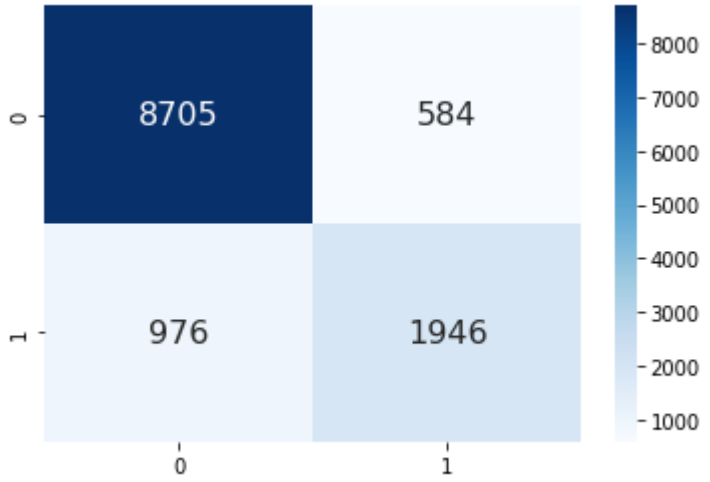
Model	Accuracy	Precision	F1	Recall	Gini
LogisticRegression	0.854	0.742	0.664	0.600	**0.806**
RandomForest	0.846	0.701	0.659	0.622	0.785
XGBoost	**0.872**	**0.768**	**0.714**	**0.667**	0.792
Hard voting	0.867	0.765	0.697	0.639	0.800
Soft voting	0.866	0.764	0.697	0.640	0.799
Stacking	**0.872**	0.777	0.711	0.655	0.798
Bagging	0.855	0.741	0.667	0.606	0.804

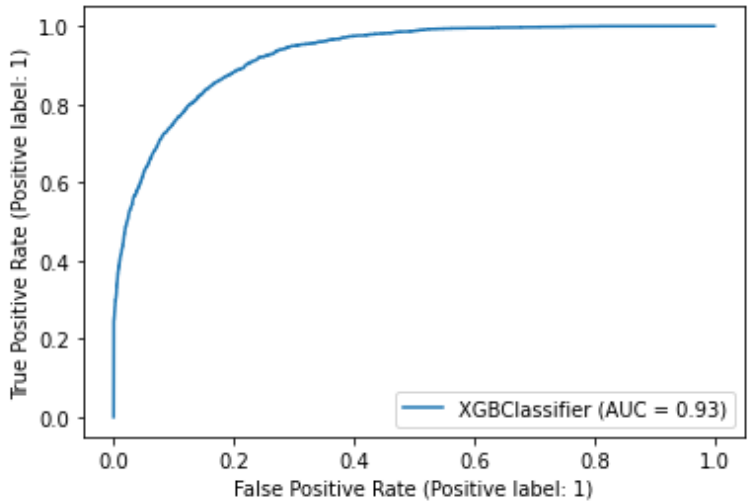
Najlepszy model przy wykorzystaniu optymalnych hiperparametrów

Model	Default Accuracy	Hiperparameter Accuracy
LogisticRegression	0.854	0.855
RandomForest	0.846	0.861
XGBoost	**0.872**	**0.875**
Stacking	0.872	0.874

Dokładniejsza analiza najlepszego modelu

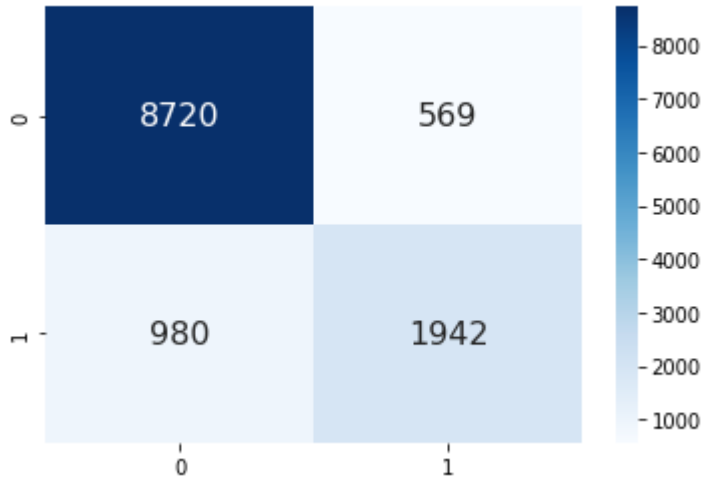
Accuracy	0.875
Precision	0.783
F1	0.717
Recall	0.661

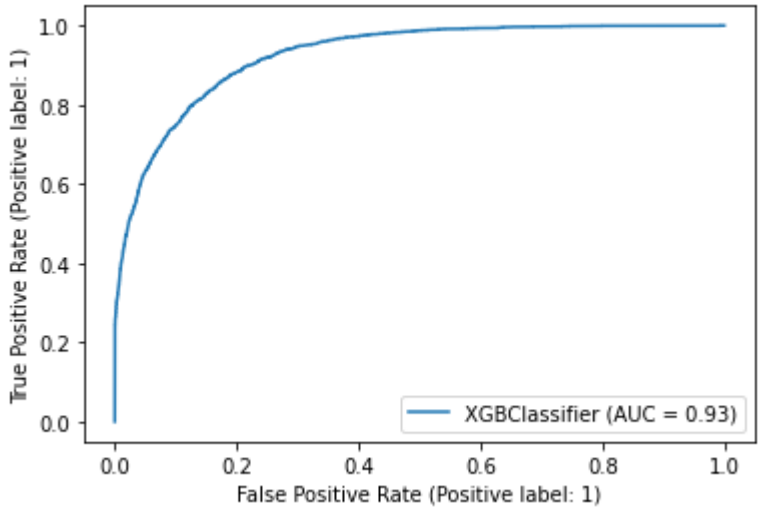


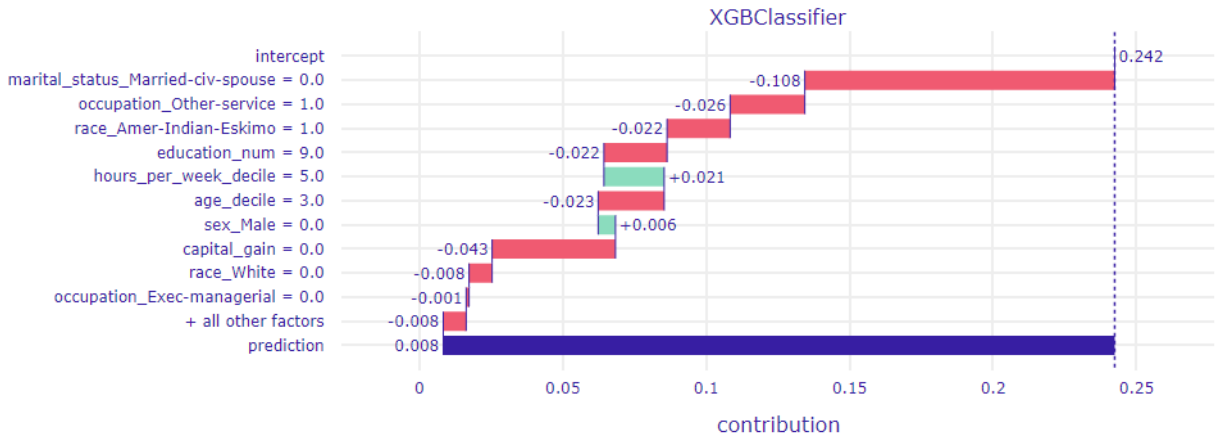


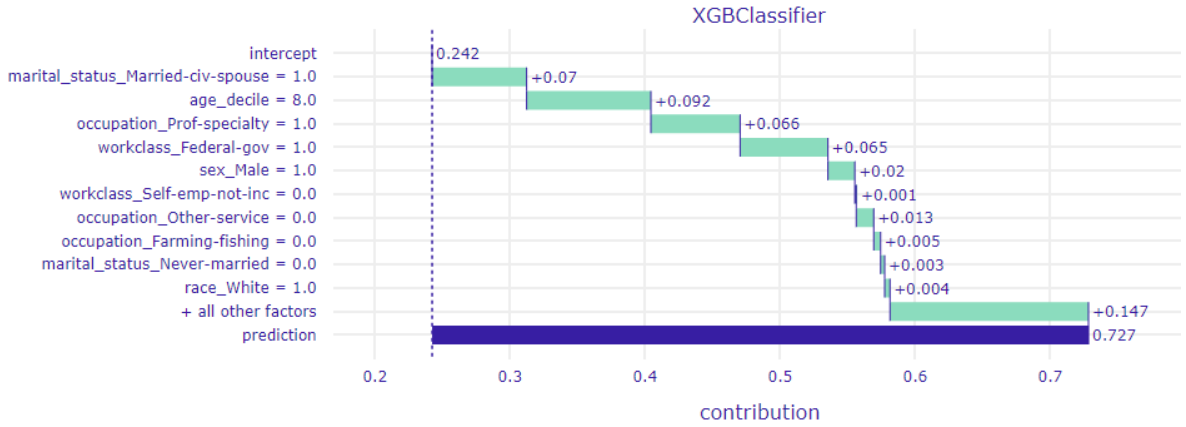
Najlepszy model dla danych bez rasy i płci

Accuracy	0.874
Precision	0.782
F1	0.715
Recall	0.659



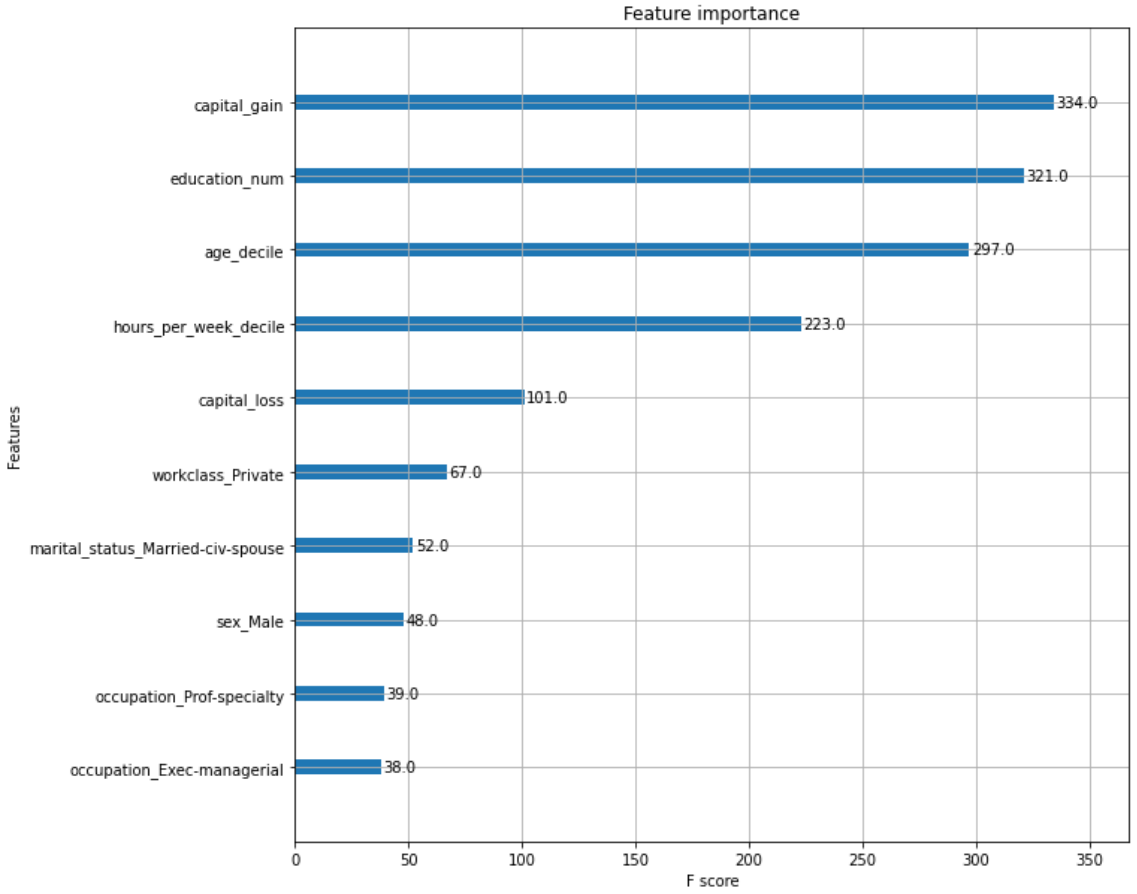






Interpretowalność

Feature Importance dla XGBoost



Najważniejsze zmienne według InformationValue Score

1. Education_num – zmienna kategoryczna – stopień wykształcenia – IV score: 0.74
2. Marital_status_Never-married – zmienna binarna – IV score: 0.83
3. Marital_status_Married-civ-spouse – zmienna binarna – IV score: 1.28
4. Relationship_Own-child – zmienna binarna – IV score: 0.63
5. Age_decile – zmienna ilościowa – IV score: 1.08

Dziękujemy za uwagę