US Census - Income Analysis

IA SCHOOL

Wael Gribaa™
2020

Outline

Overview of the dataset

Modifications

Machine Learning

What's next ?

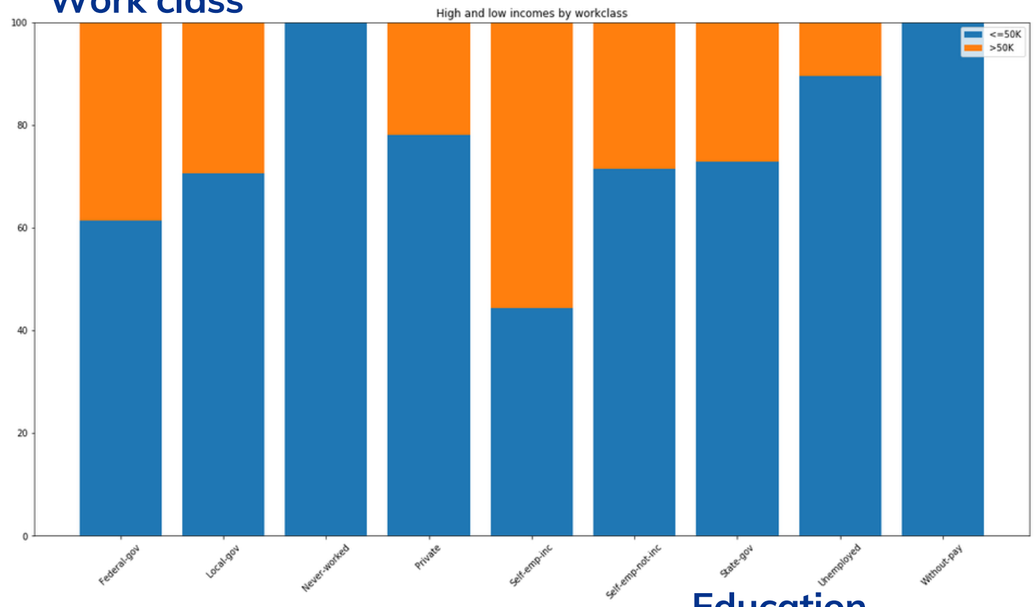
General informations

- Dataset as provided
- Income has two values : less and more than $50,000 a year
- Most features have low modularity
- Unjustified numerical features : capital gain, capital loss (and age).
- What is fnlwgt ?

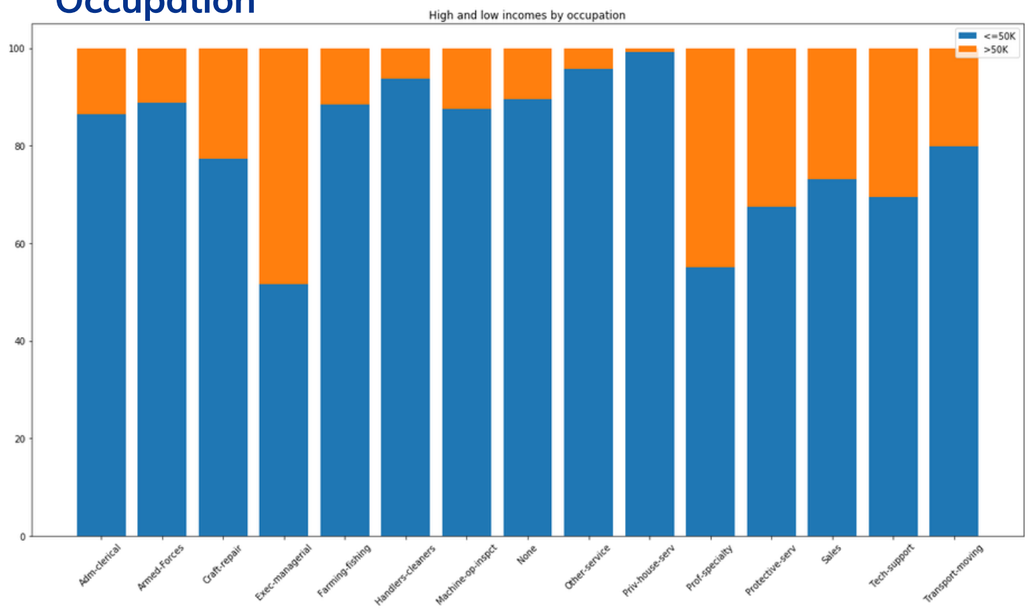
```
The dataframe has 15 columns and 32561 rows.
Indices : from 0 to 32560 (step= 1)
0:      "age".............of type int64          0 null values and 73 uniques.
1:      "workclass".......of type object         1836 null values and 9 uniques [None, Private, State-gov, Federal-gov, Self
-emp-not-inc, Self-emp-inc, Local-gov, Without-pay, Never-worked].
2:      "fnlwgt"..........of type int64          0 null values and 21648 uniques.
3:      "education".......of type object         0 null values and 16 uniques [HS-grad, Some-college, 7th-8th, 10th, Doctora
te, Prof-school, Bachelors, Masters, 11th, Assoc-acdm, Assoc-voc, 1st-4th, 5th-6th, 12th, 9th, Preschool].
4:      "education_num"...of type int64          0 null values and 16 uniques [9, 10, 4, 6, 16, 15, 13, 14, 7, 12, 11, 2, 3,
8, 5, 1].
5:      "marital_status"..of type object         0 null values and 7 uniques [Widowed, Divorced, Separated, Never-married, M
arried-civ-spouse, Married-spouse-absent, Married-AF-spouse].
6:      "occupation"......of type object         1843 null values and 15 uniques [None, Exec-managerial, Machine-op-inspct,
Prof-specialty, Other-service, Adm-clerical, Craft-repair, Transport-moving, Handlers-cleaners, Sales, Farming-fishing, Tec
h-support, Protective-serv, Armed-Forces, Priv-house-serv].
7:      "relationship"....of type object         0 null values and 6 uniques [Not-in-family, Unmarried, Own-child, Other-rel
ative, Husband, Wife].
8:      "race"............of type object         0 null values and 5 uniques [White, Black, Asian-Pac-Islander, Other, Amer-
Indian-Eskimo].
9:      "sex".............of type object         0 null values and 2 uniques [Female, Male].
10:     "capital_gain"....of type int64          0 null values and 119 uniques.
11:     "capital_loss"....of type int64          0 null values and 92 uniques.
12:     "hours_per_week"..of type int64          0 null values and 94 uniques.
13:     "native_country"..of type object         583 null values and 42 uniques.
14:     "income"..........of type object         0 null values and 2 uniques [<=50K, >50K].
```

All other features : in a nutshell

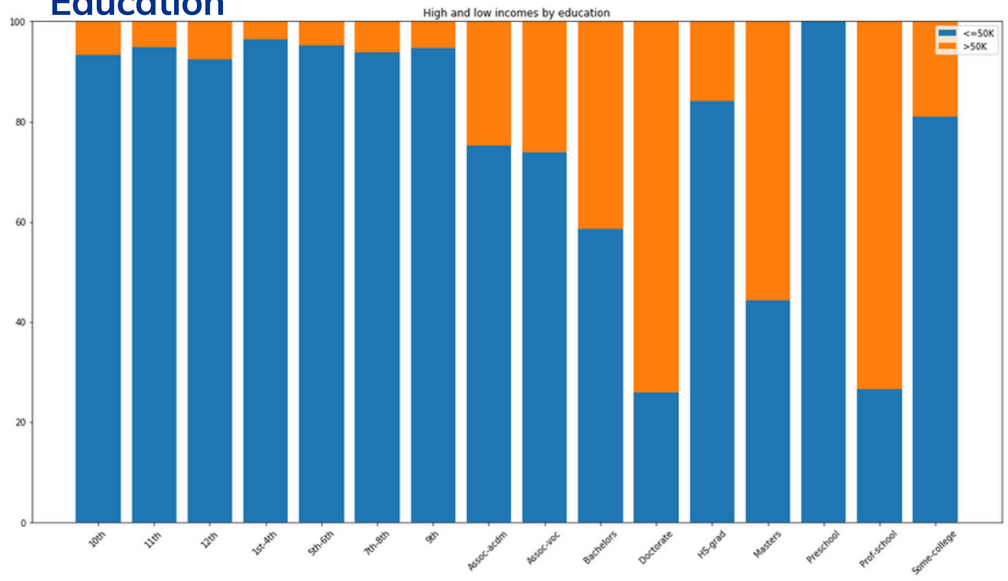
Work class



Occupation

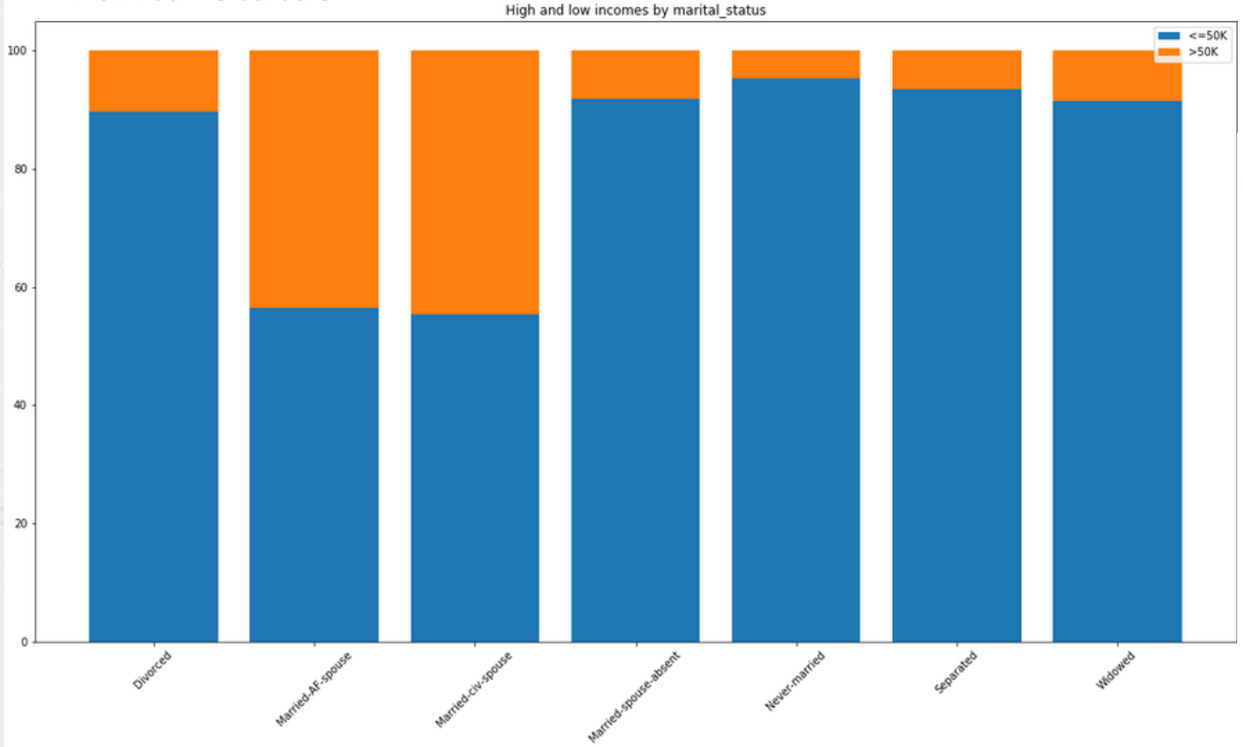


Education

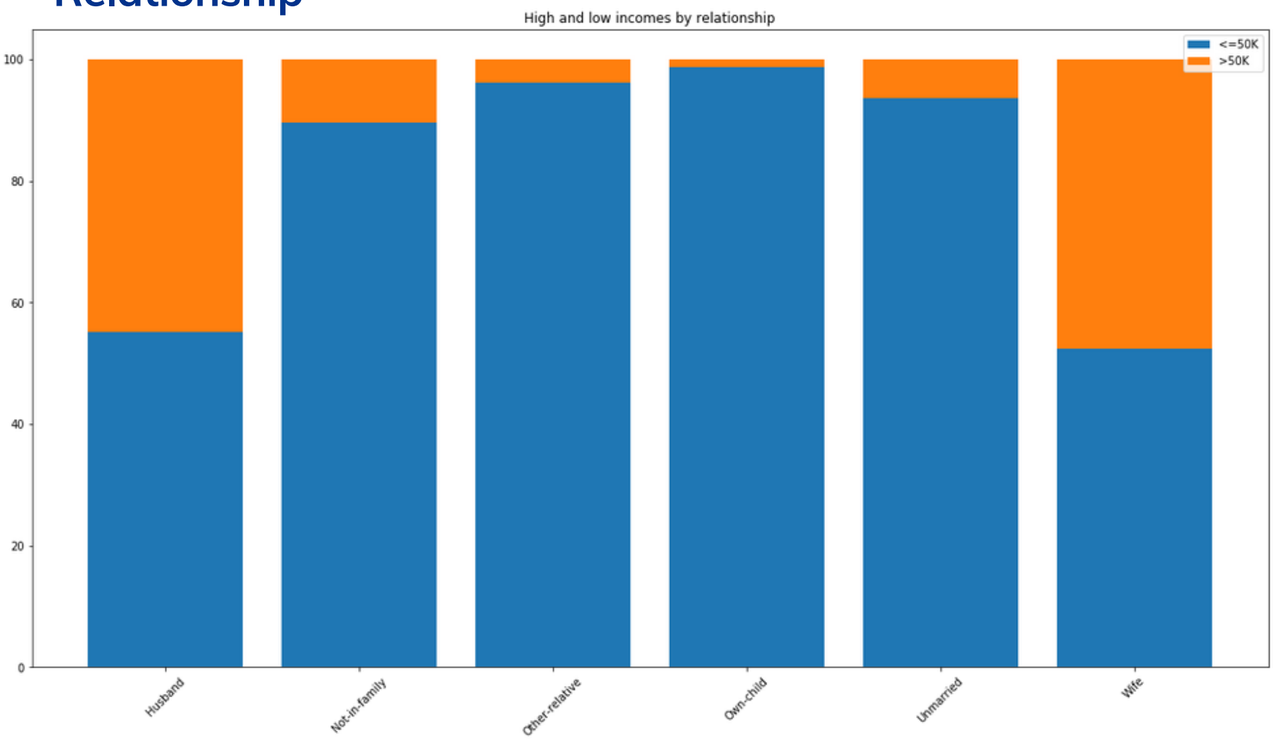


All other features : in a nutshell

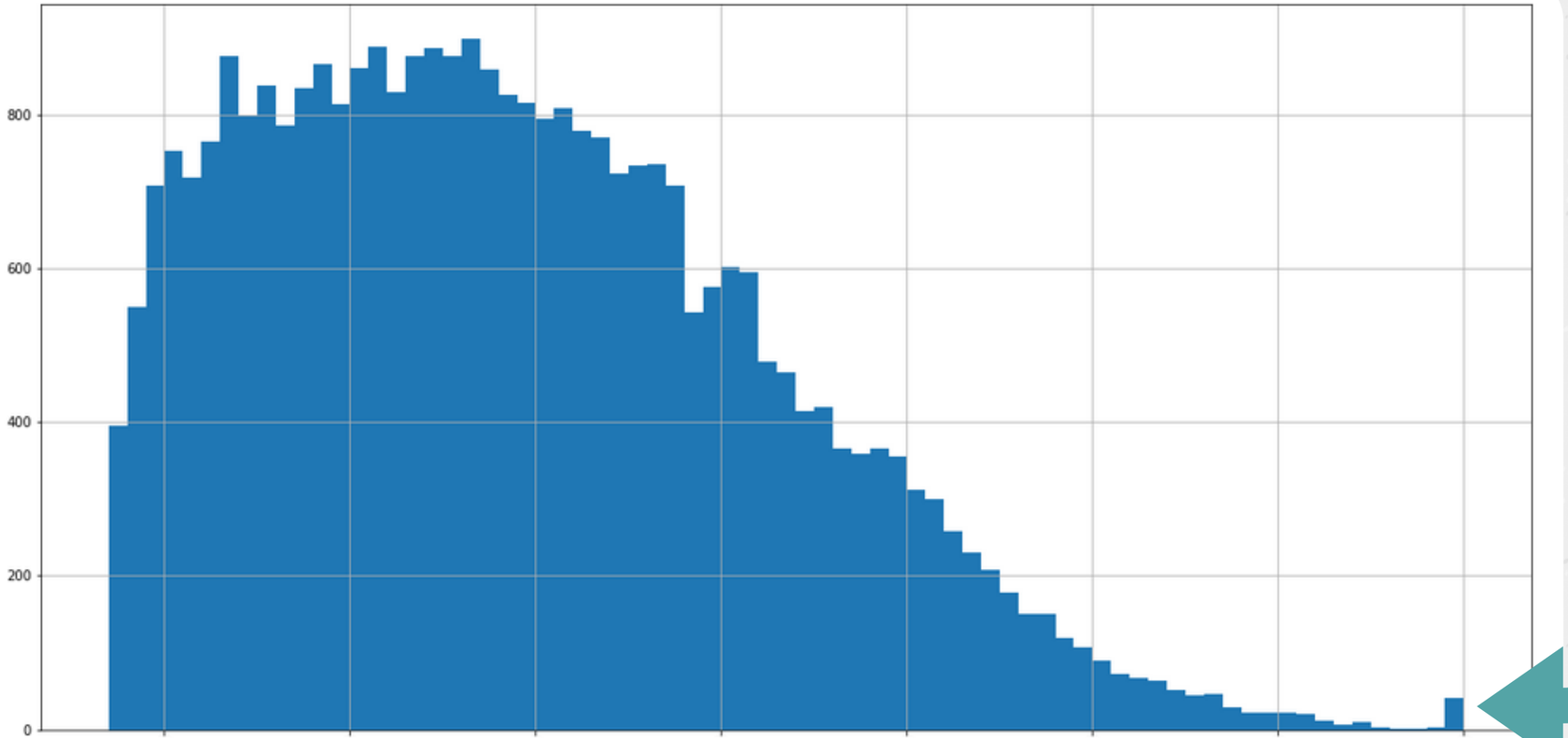
Marital status



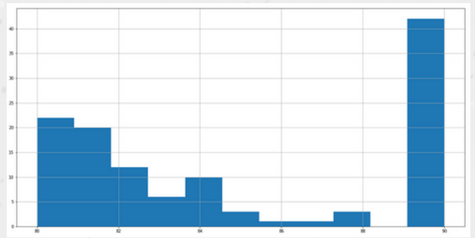
Relationship



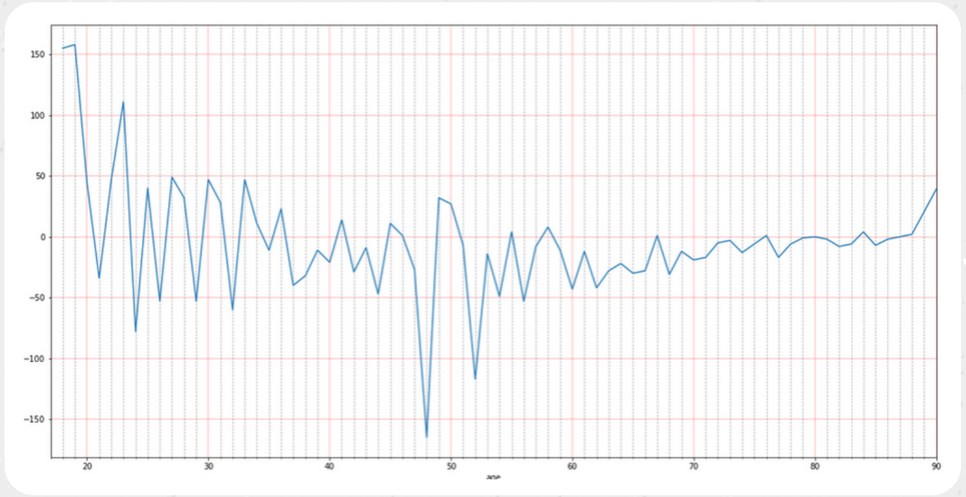
Age

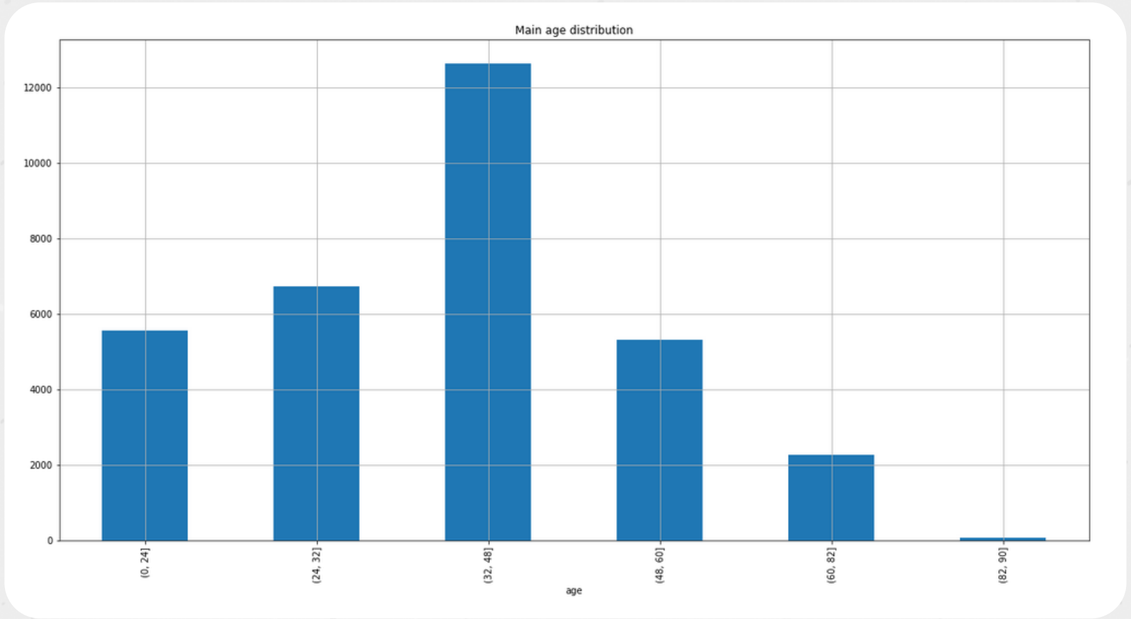


- From 17 to 90 years old
- Why so many people with 90 yo ?

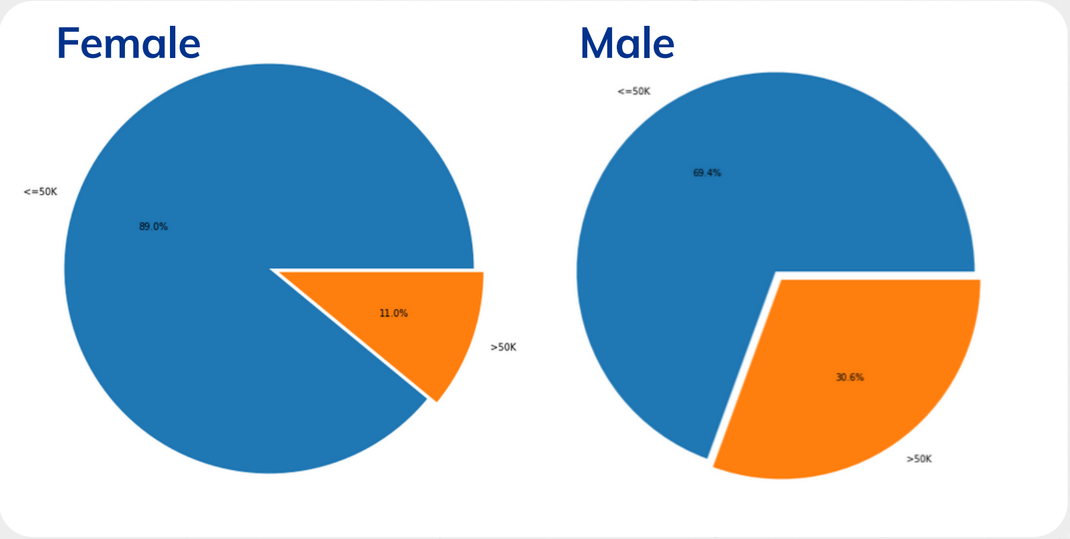


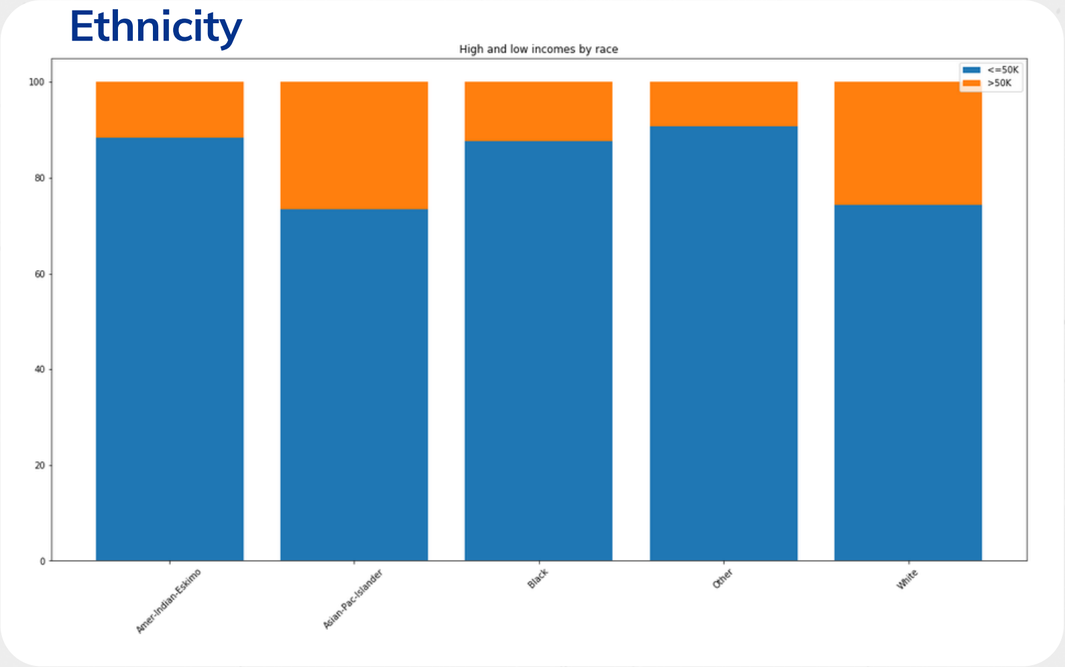
Age


150
100
50
0
−50
−100
−150
20
30
40
50
60
70
80
90


Main age distribution
12000
10000
8000
6000
4000
2000
0
(0, 24]
(24, 32]
(32, 48]
(48, 60]
(60, 82]
(82, 90]
age

Sensitive matters

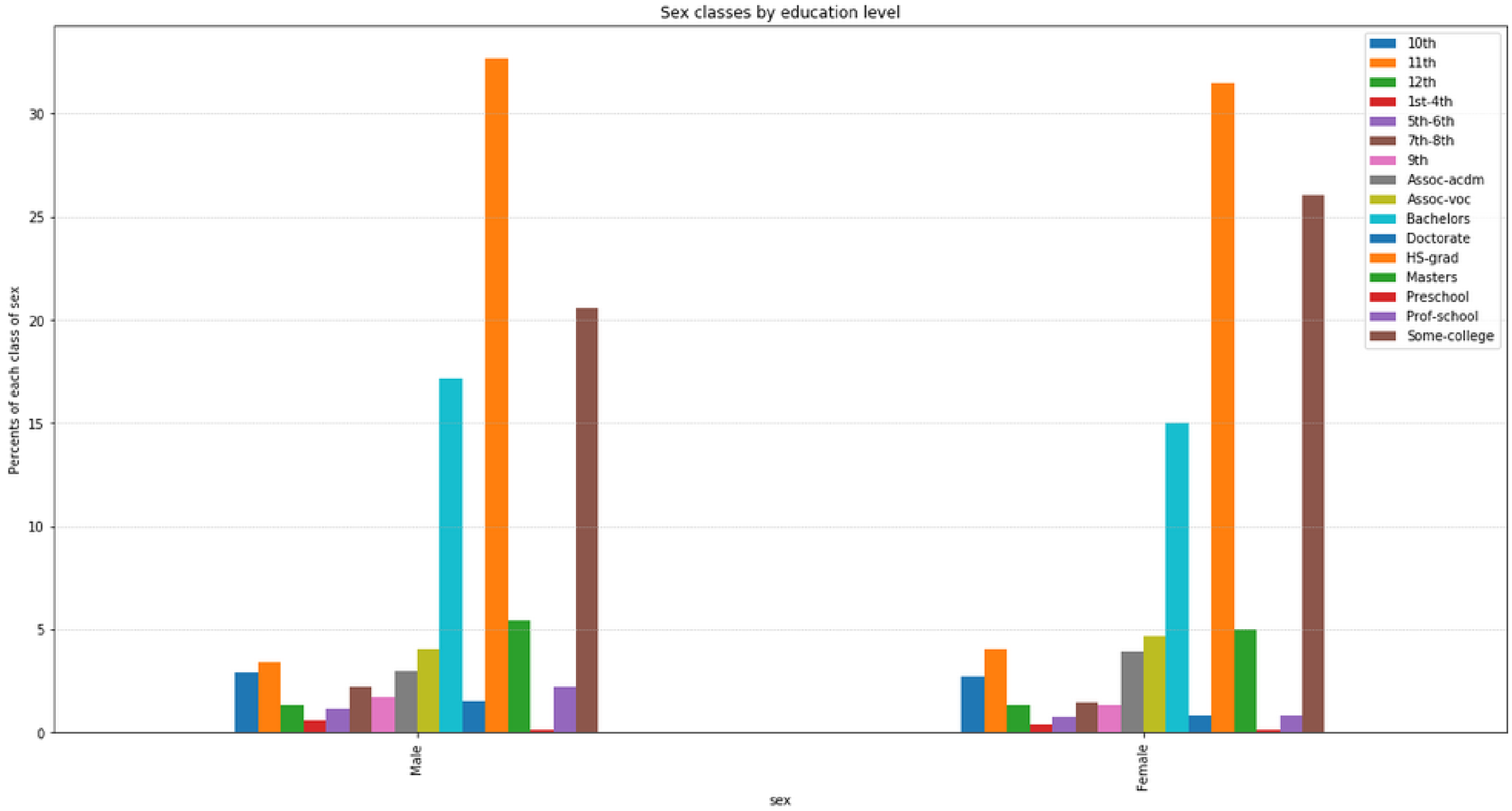




- Gender is a key feature
- 11% vs 31%

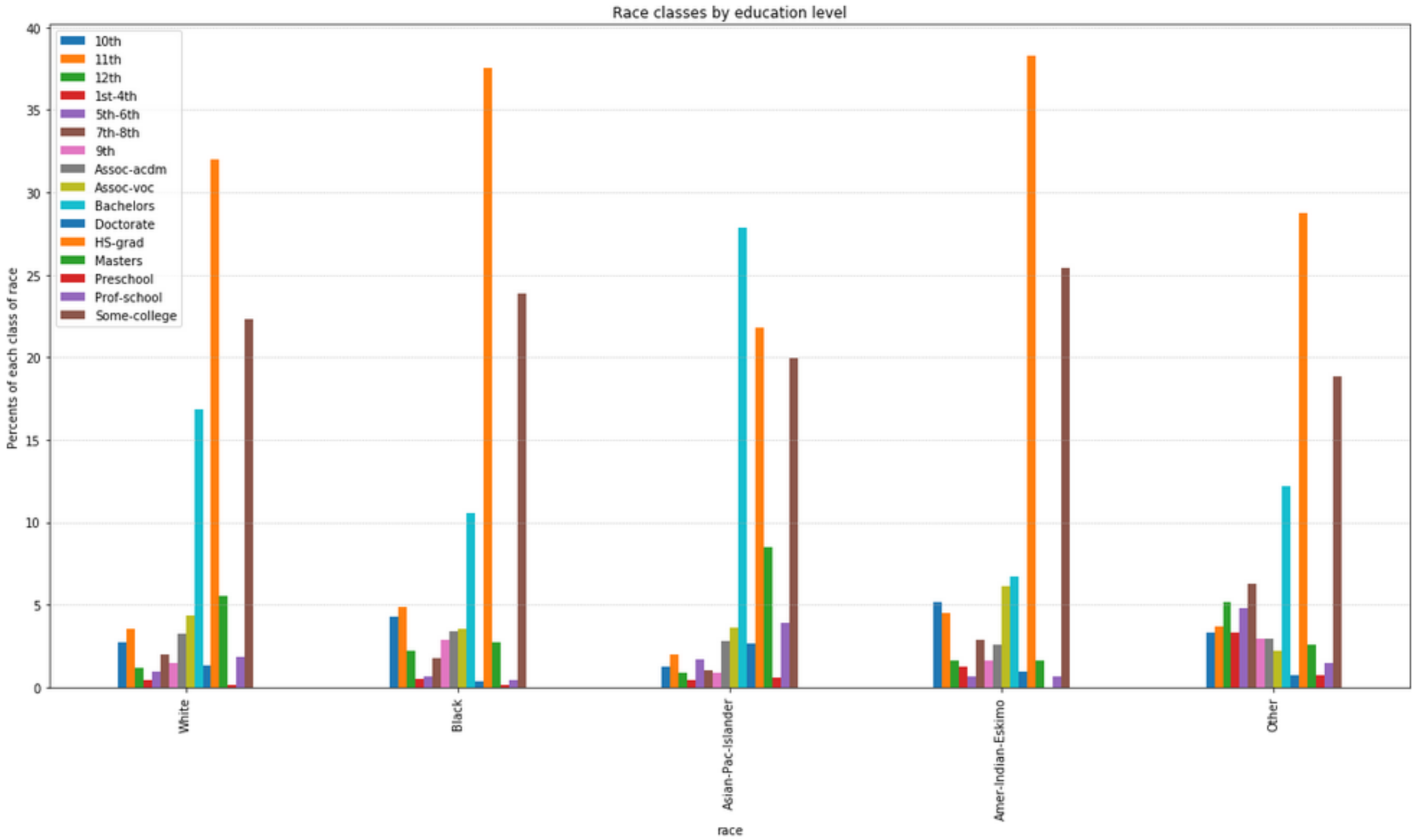
- Asian and white ethnicities are privileged

Sensitive matters - Gender



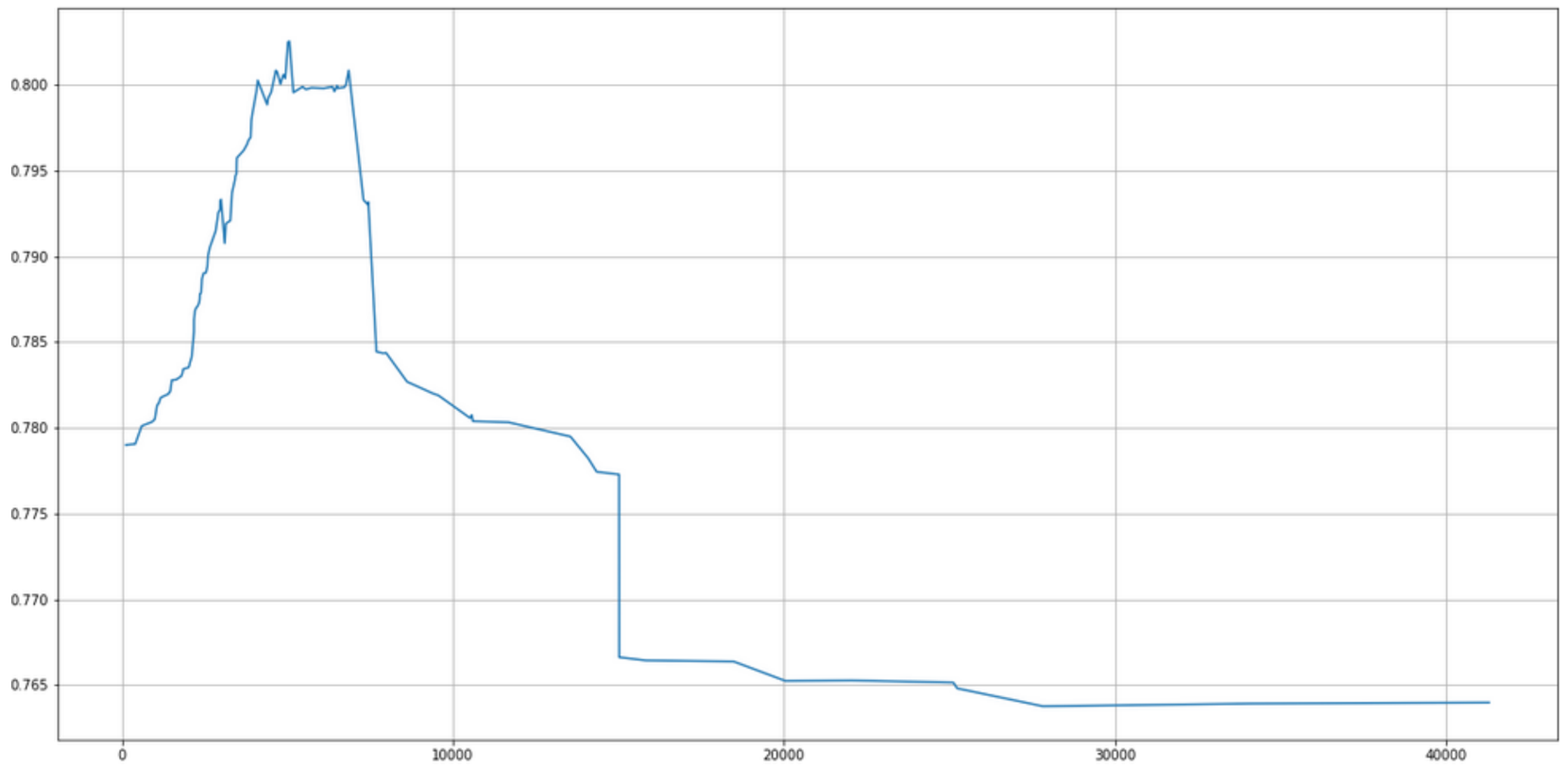
- Why the difference in income ?

Sensitive matters - Ethnicity



- Ethnicity and income do show correlations

Capital gain



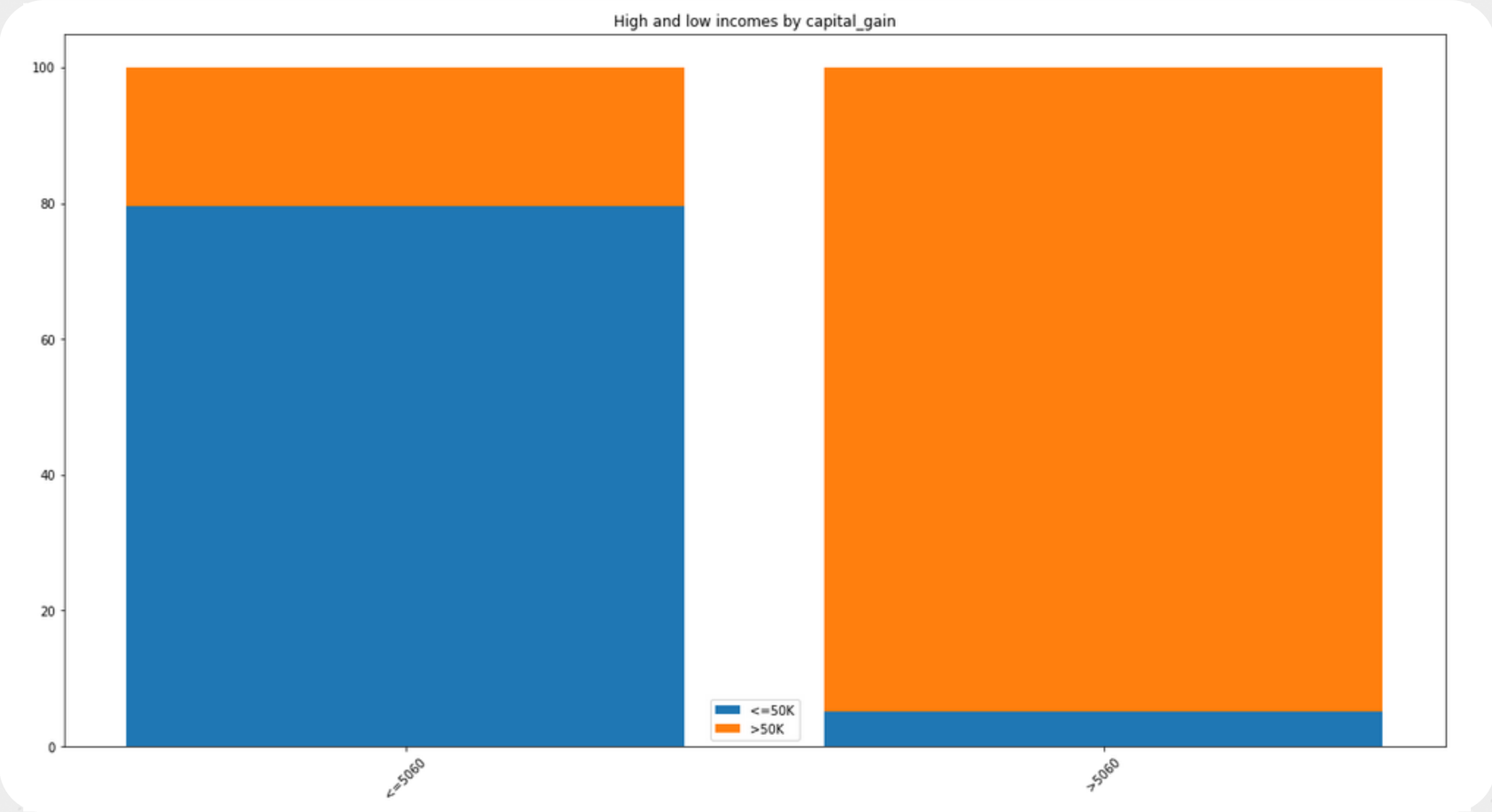
- Pivot value = $5060

	pivot_value	accuracy
77	5060	0.802545
76	5013	0.802514
70	4650	0.800854
89	6849	0.800854
71	4687	0.800762
...	...	...
112	25124	0.765138
113	25236	0.764800
116	41310	0.763970
115	34095	0.763909
114	27828	0.763755

- Accuracy of 0.802545

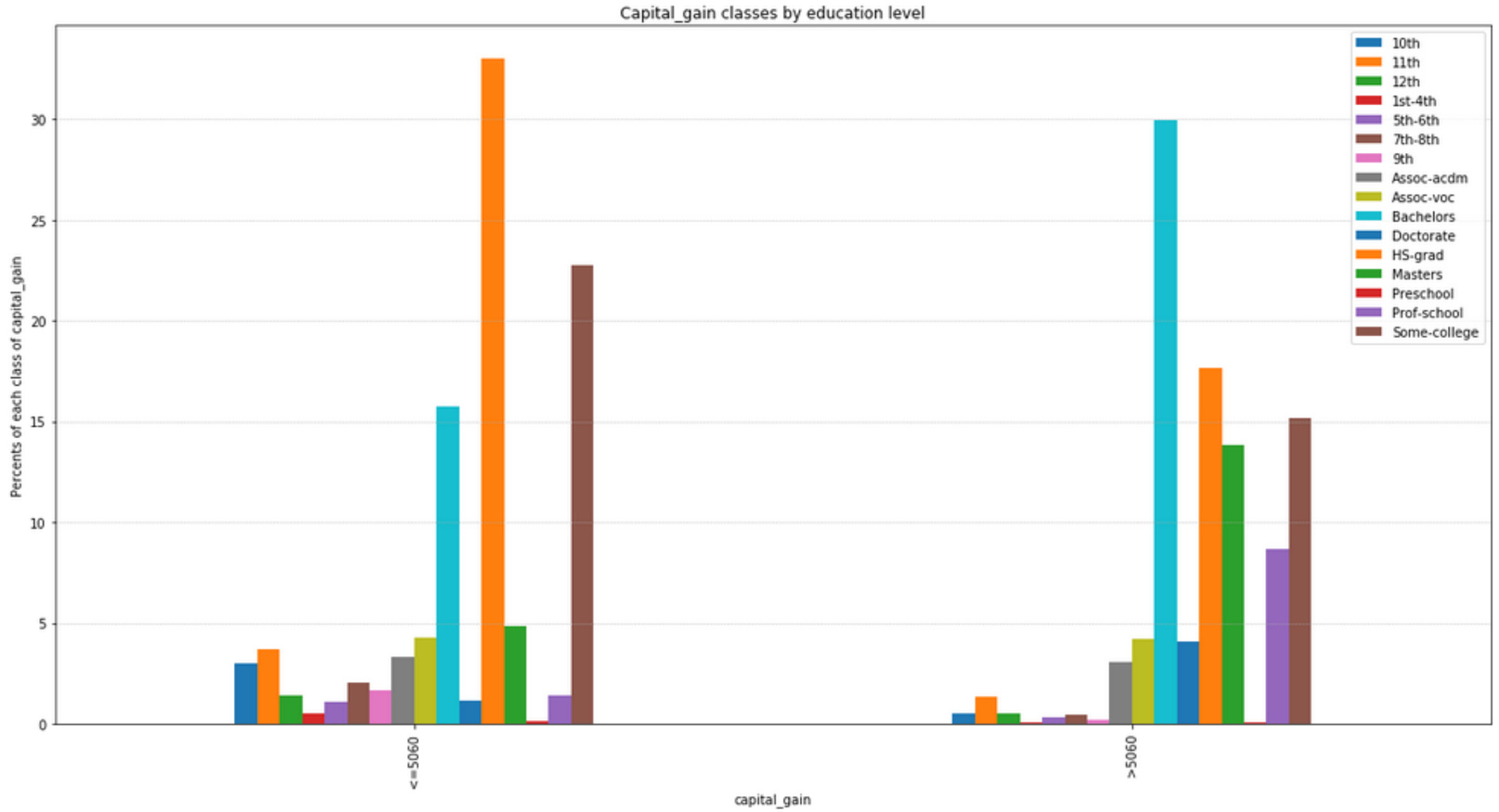
capital_gain_class	(-0.001, 5060.0]	(5060.0, 99999.0]
income		
<=50K	24614	82
>50K	6342	1496

Capital gain



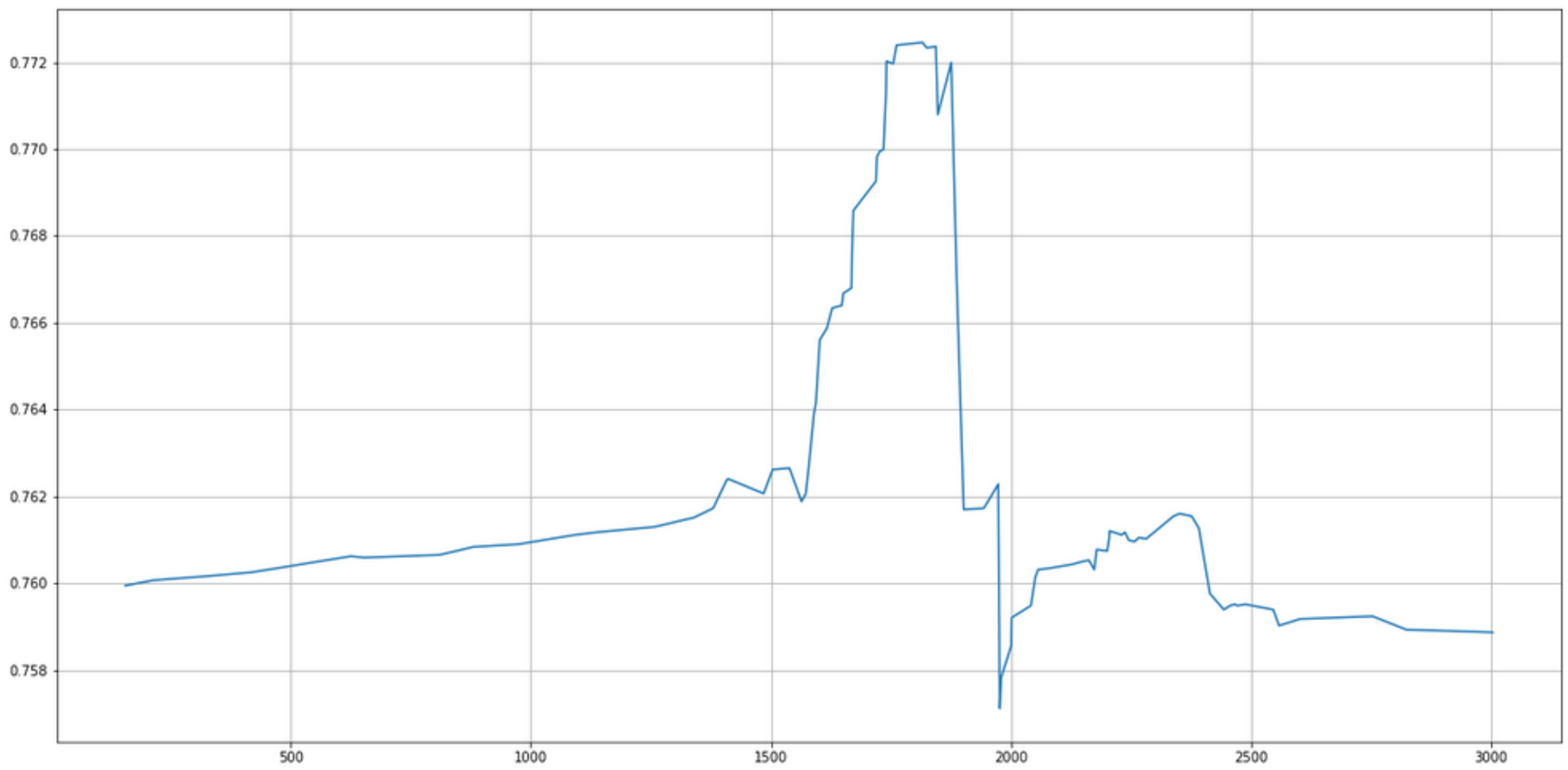
- Two values only
- Optimizing performance
- Very discriminating

Capital gain



- Higher education = Higher capital gain

Capital loss



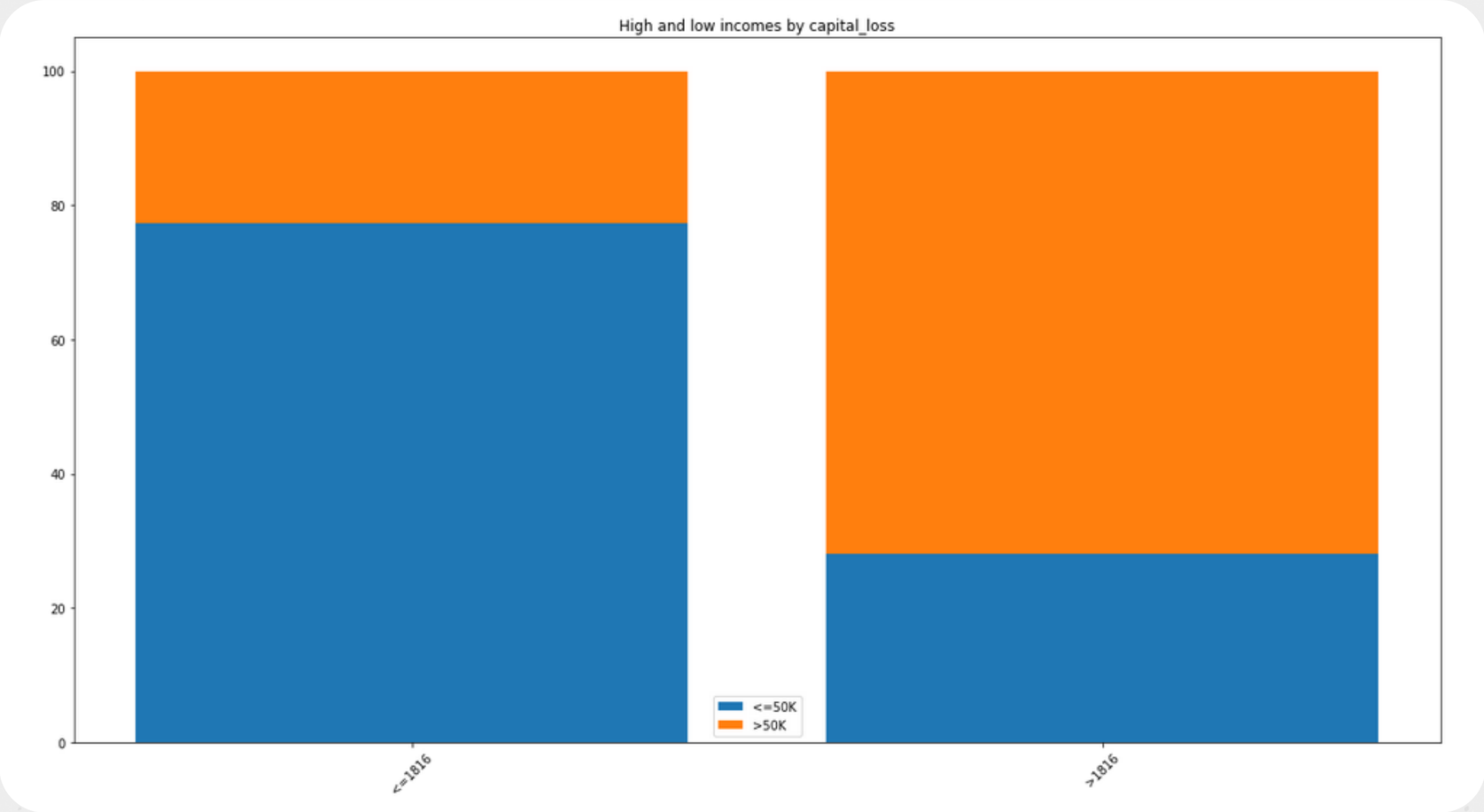
- Pivot value = $1816

	pivot_value	accuracy
40	1816	0.772453
39	1762	0.772392
42	1844	0.772361
41	1825	0.772330
37	1741	0.772023
...	...	...
50	1980	0.757823
49	1977	0.757116
87	3683	NaN
88	3770	NaN
89	3900	NaN

- Accuracy of 0.772453

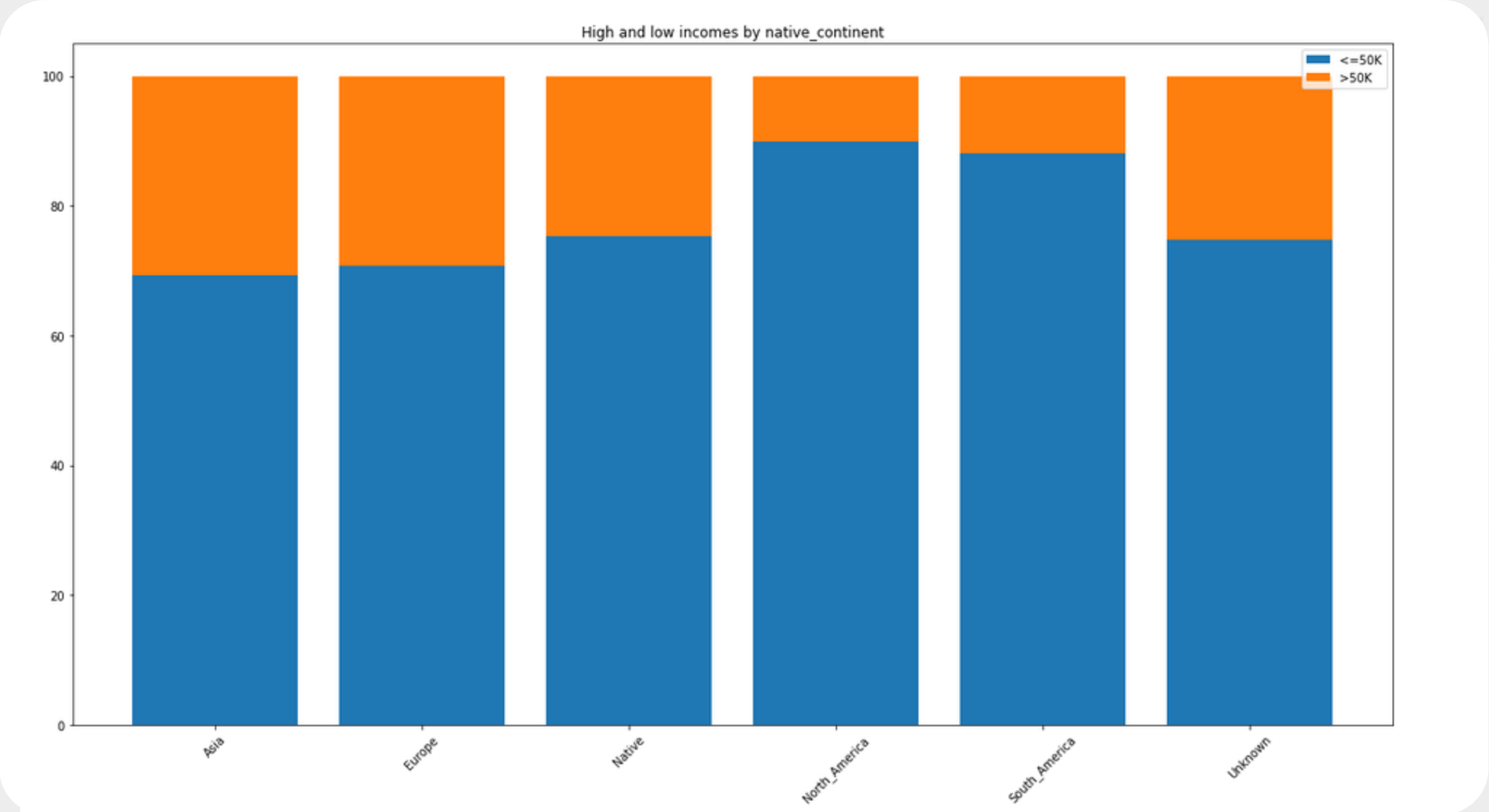
capital_loss_class	(-0.001, 1816.0]	(1816.0, 4356.0]
income		
<=50K	24418	278
>50K	7125	713

Capital loss



- Two values only
- Optimizing performance
- Quite discriminating

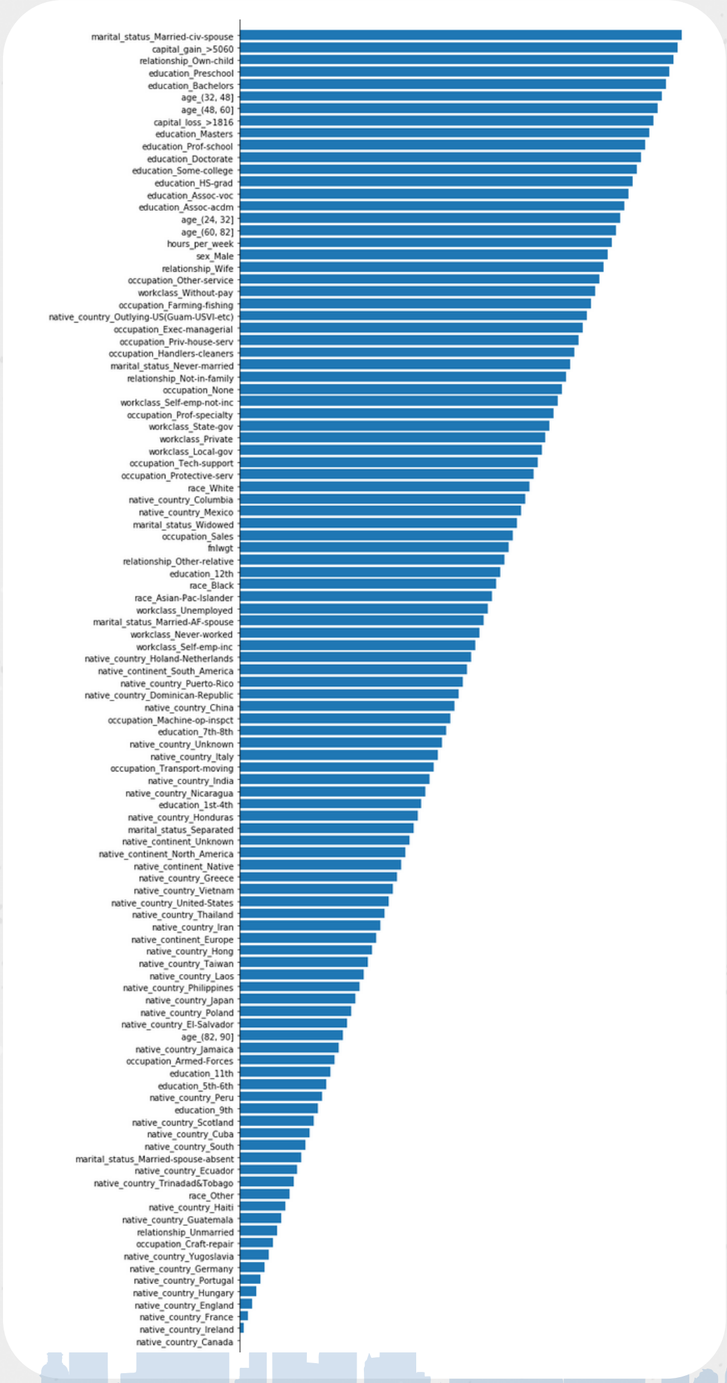
By continent



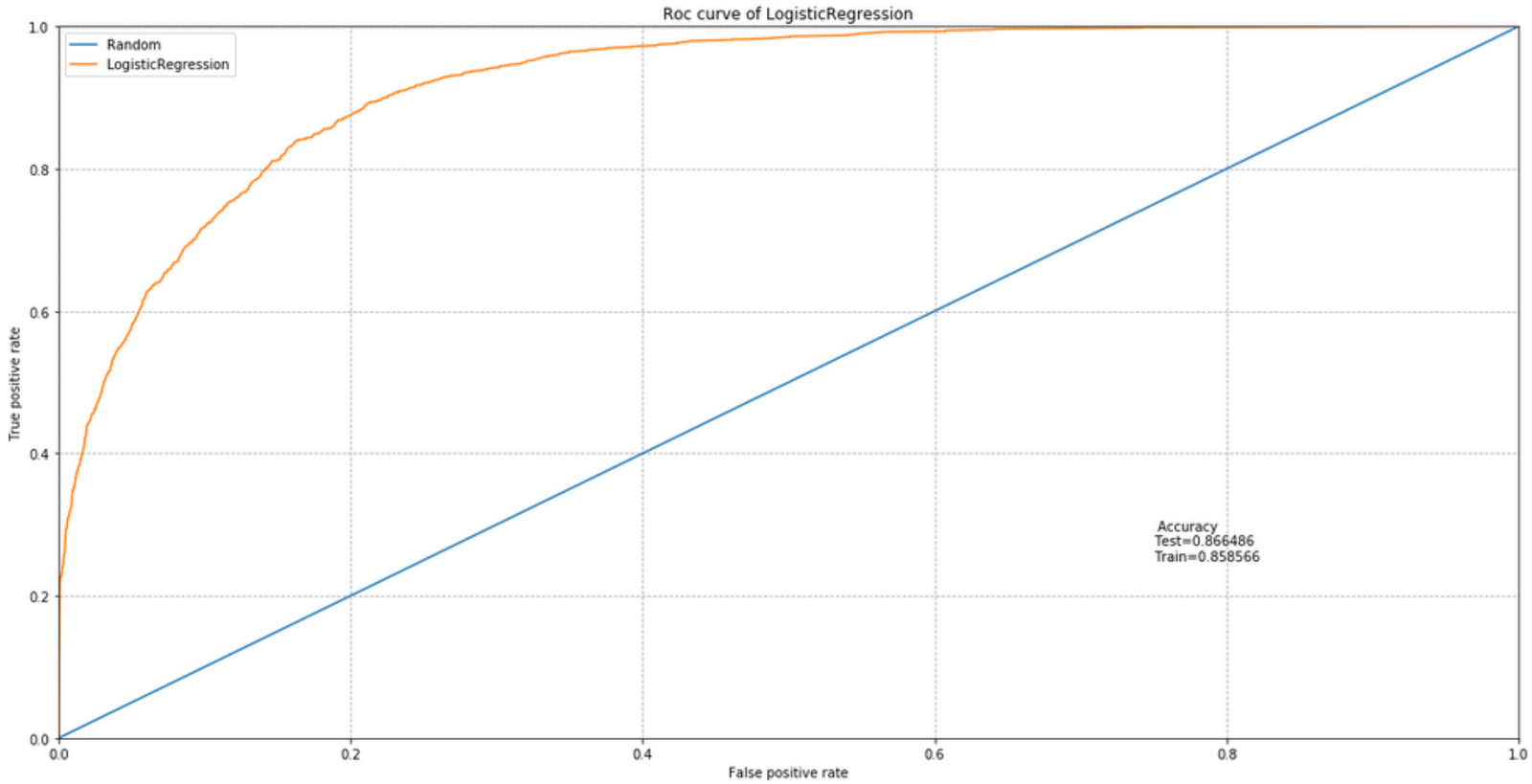
- Each county was attributed to a continent
- North america doesn't include USA
- Asia confirms the ethnicity

Feature importance

- Recursive feature elimination with a Logistic Regressor
- Married, high capital, child owners, preschool or bachelors and age will be the most predictable



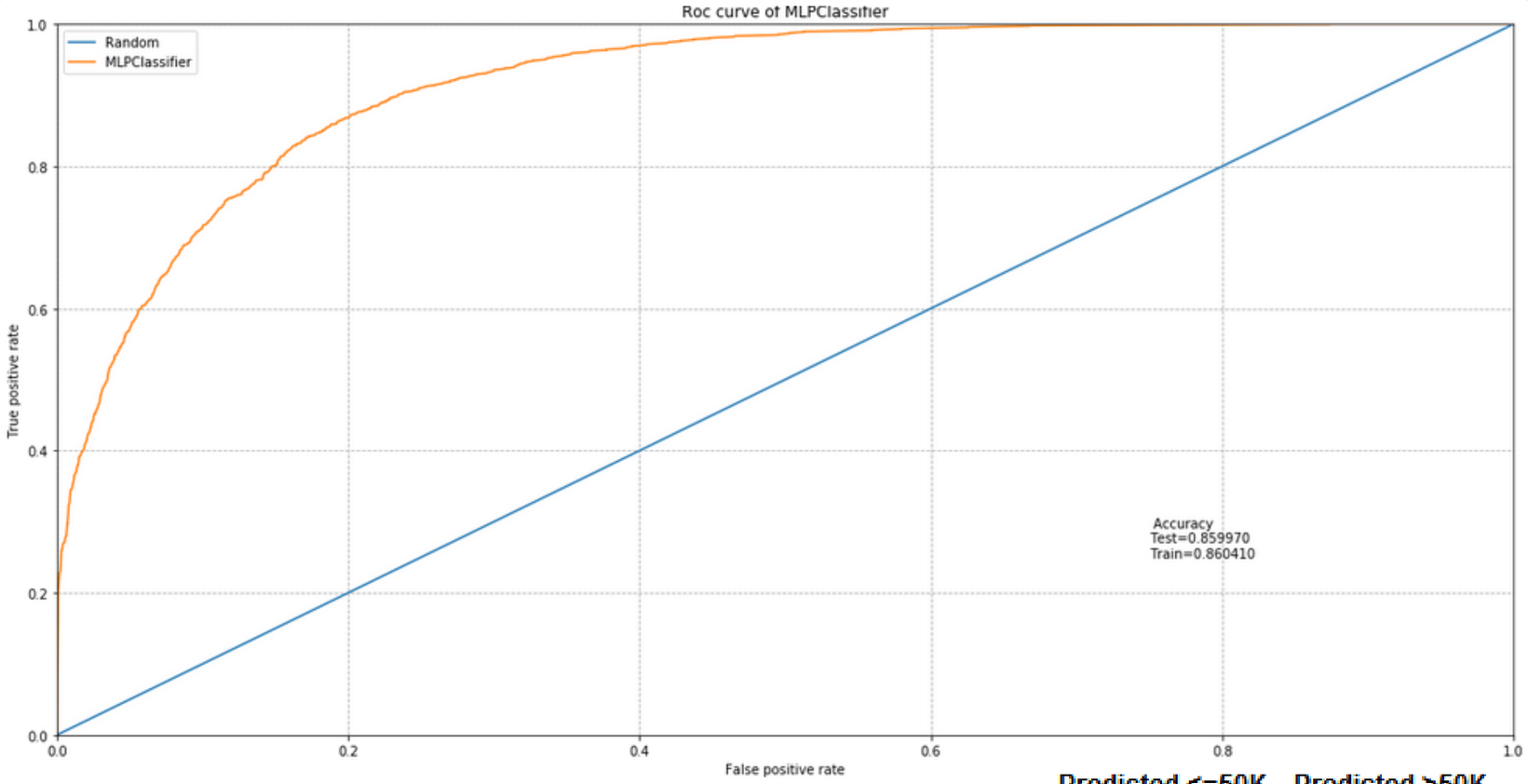
Logistic regression



	Predicted <=50K	Predicted >50K
True <=50K	5819	397
True >50K	734	1184

- Consistenly one of the best in the baseline
- Very fast
- 86.1% accuracy in cross-validation

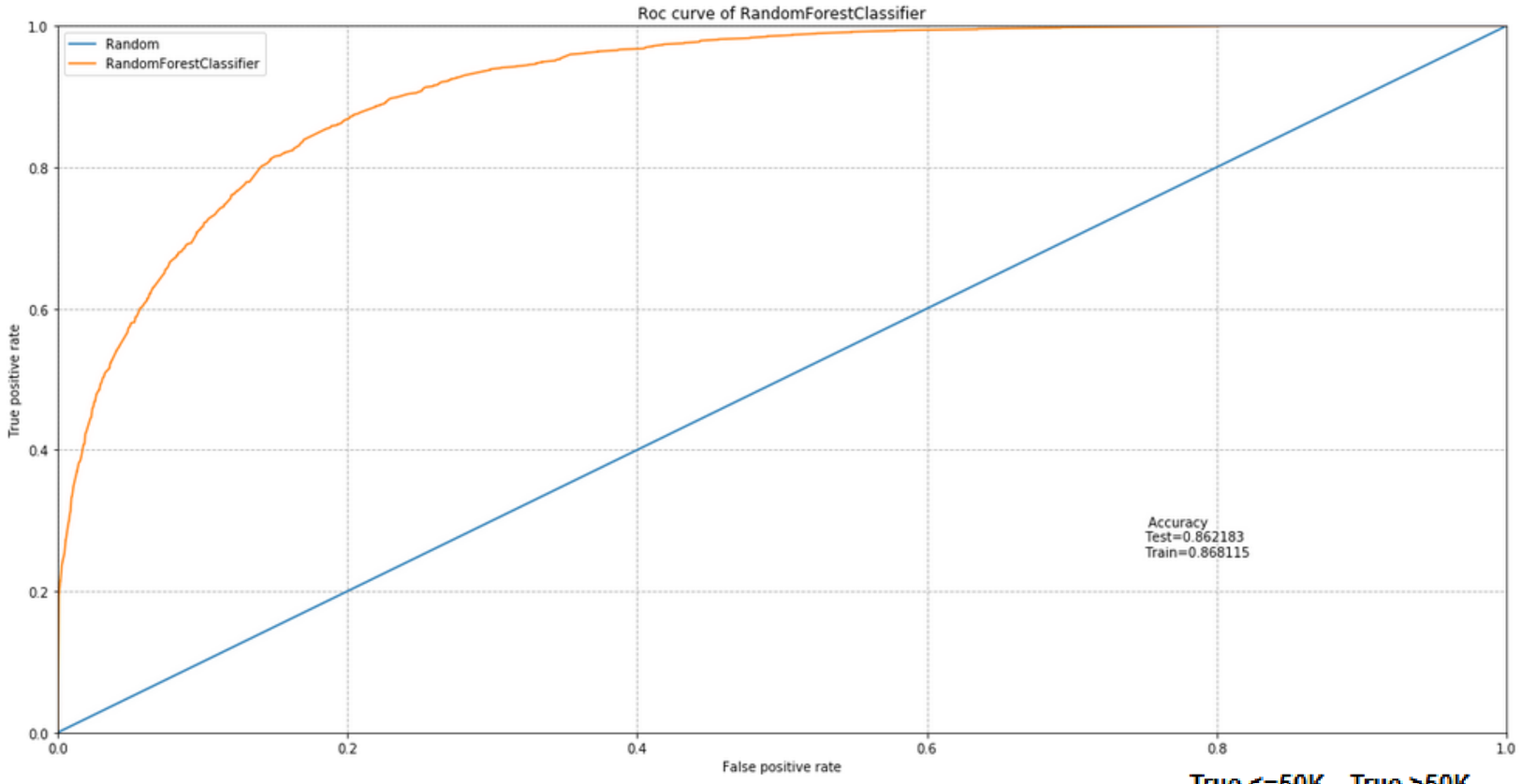
Multi Layer Perceptron Classifier



	Predicted <=50K	Predicted >50K
True <=50K	5794	422
True >50K	718	1200

- Slow but extremely consistent
- 85.98% accuracy

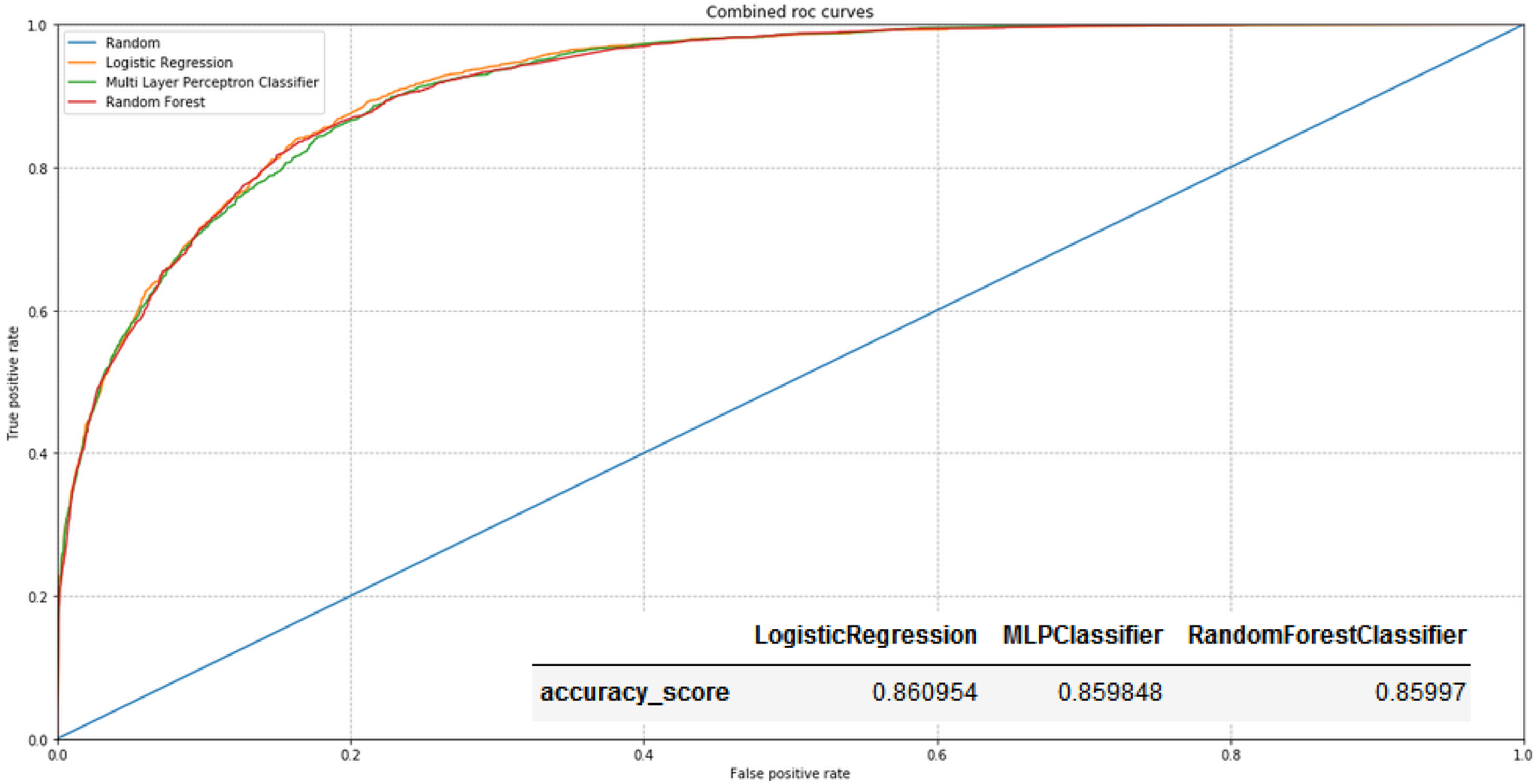
Random Forest



	True <=50K	True >50K
Predicted <=50K	5891	325
Predicted >50K	814	1104

- 86% accuracy

Accuracy summary



	LogisticRegression	MLPClassifier	RandomForestClassifier
accuracy_score	0.860954	0.859848	0.85997

Bagging

	LogisticRegression	MLPClassifier	RandomForestClassifier	Combined
accuracy_score	0.860954	0.859848	0.85997	0.863044

What would improve the predictability of this dataset?

- Geographic or location data - make it cross-checkable with external data
- fnlwgt should be constrained to a specific period !

```
: df.fnlwgt.sum()
6174899468
```

- Age is important so don't limit it to 90 years old

Thank you !

https://github.com/WGribaa/USCensus-income-levels