ML project

John Calabrese

data

- Census survey
- 40 demographic features

objectives

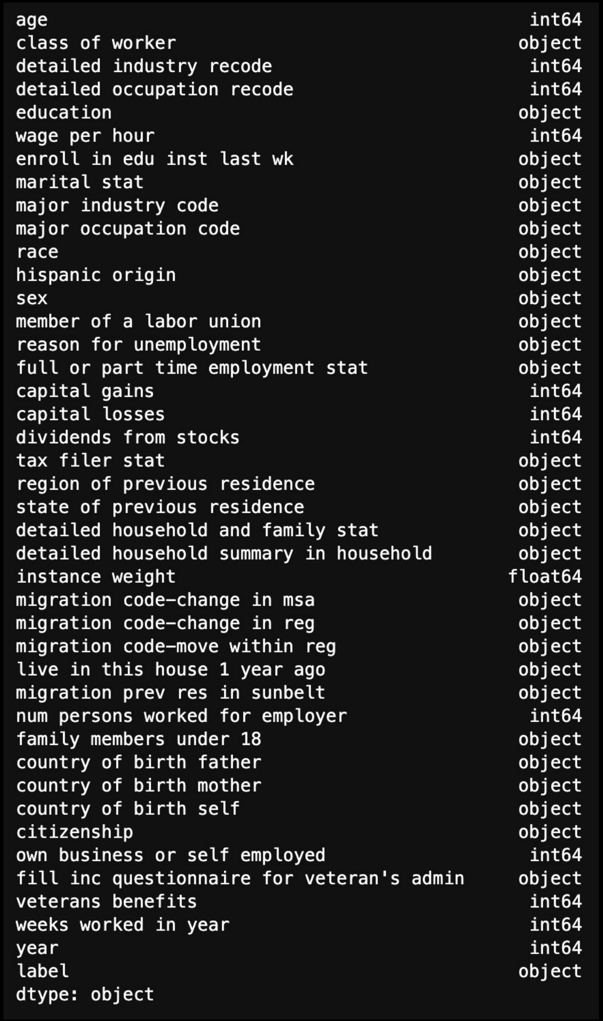
- Predict income: more or less than $50k
- Customer segmentation

raw data

- 200k samples
- 40+2 features
  - numerical and categorical

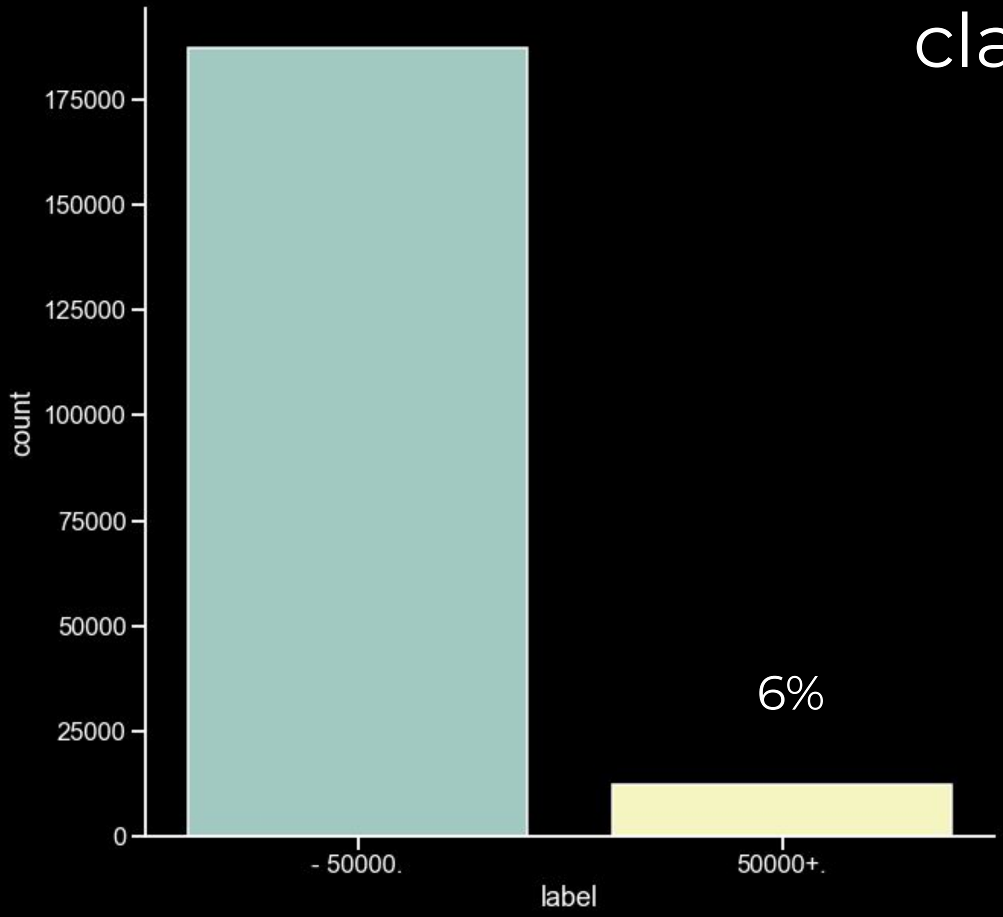
Two are special:

- 'label'
- 'instance weight'

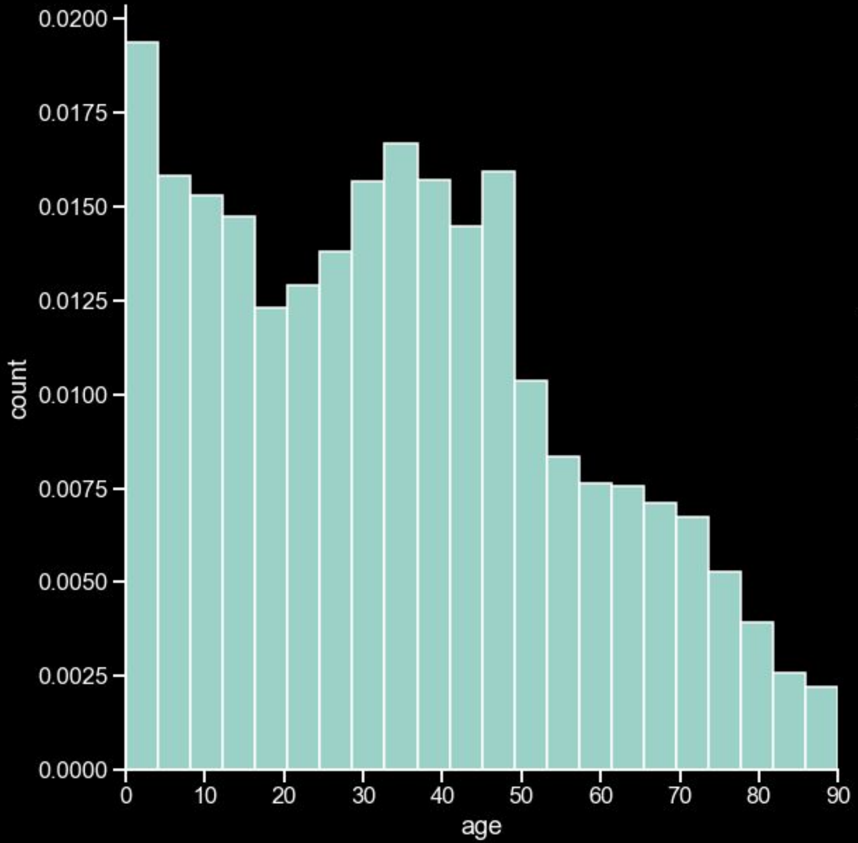


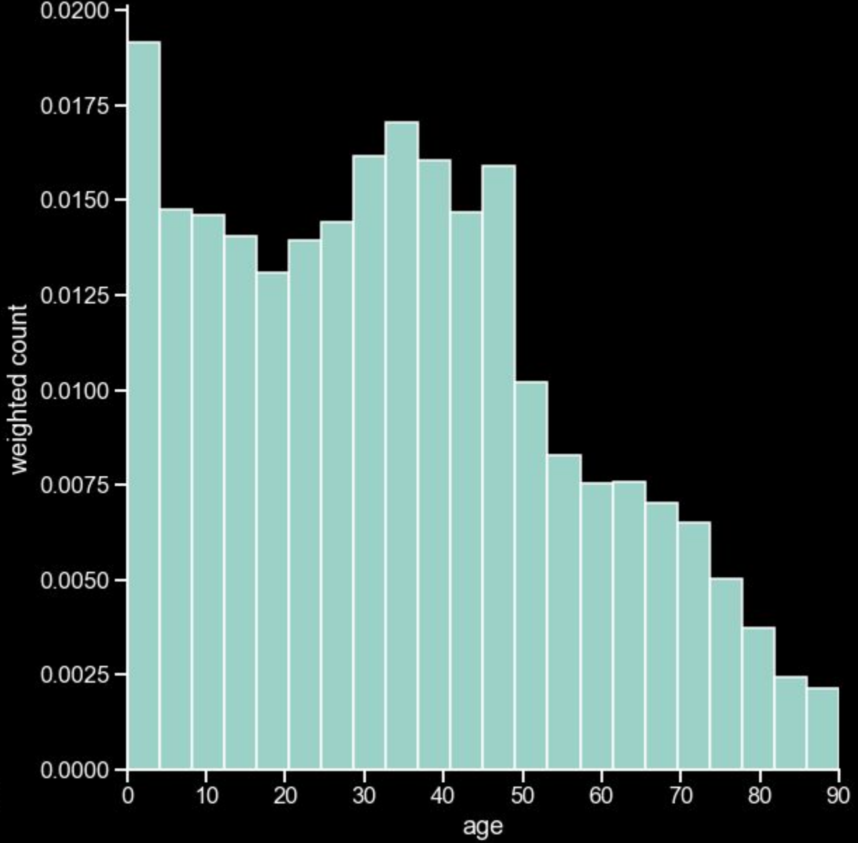
```
age                                             int64
class of worker                                object
detailed industry recode                        int64
detailed occupation recode                      int64
education                                      object
wage per hour                                   int64
enroll in edu inst last wk                     object
marital stat                                   object
major industry code                            object
major occupation code                          object
race                                           object
hispanic origin                                object
sex                                            object
member of a labor union                        object
reason for unemployment                        object
full or part time employment stat              object
capital gains                                   int64
capital losses                                  int64
dividends from stocks                           int64
tax filer stat                                 object
region of previous residence                   object
state of previous residence                    object
detailed household and family stat             object
detailed household summary in household        object
instance weight                               float64
migration code-change in msa                   object
migration code-change in reg                   object
migration code-move within reg                 object
live in this house 1 year ago                  object
migration prev res in sunbelt                  object
num persons worked for employer                 int64
family members under 18                        object
country of birth father                        object
country of birth mother                        object
country of birth self                          object
citizenship                                    object
own business or self employed                   int64
fill inc questionnaire for veteran's admin     object
veterans benefits                               int64
weeks worked in year                            int64
year                                            int64
label                                          object
dtype: object
```

class imbalance

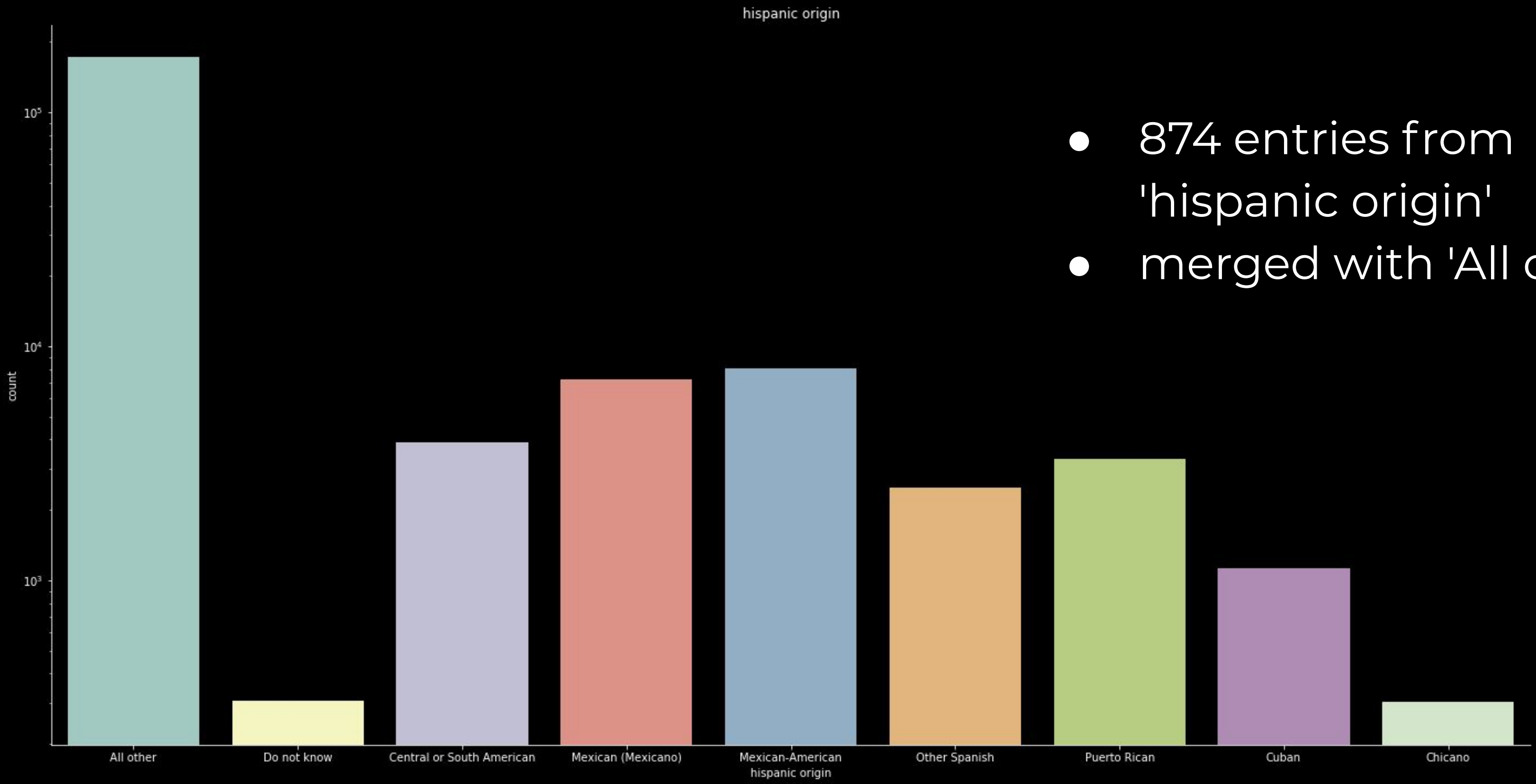


instance weight (sanity check)



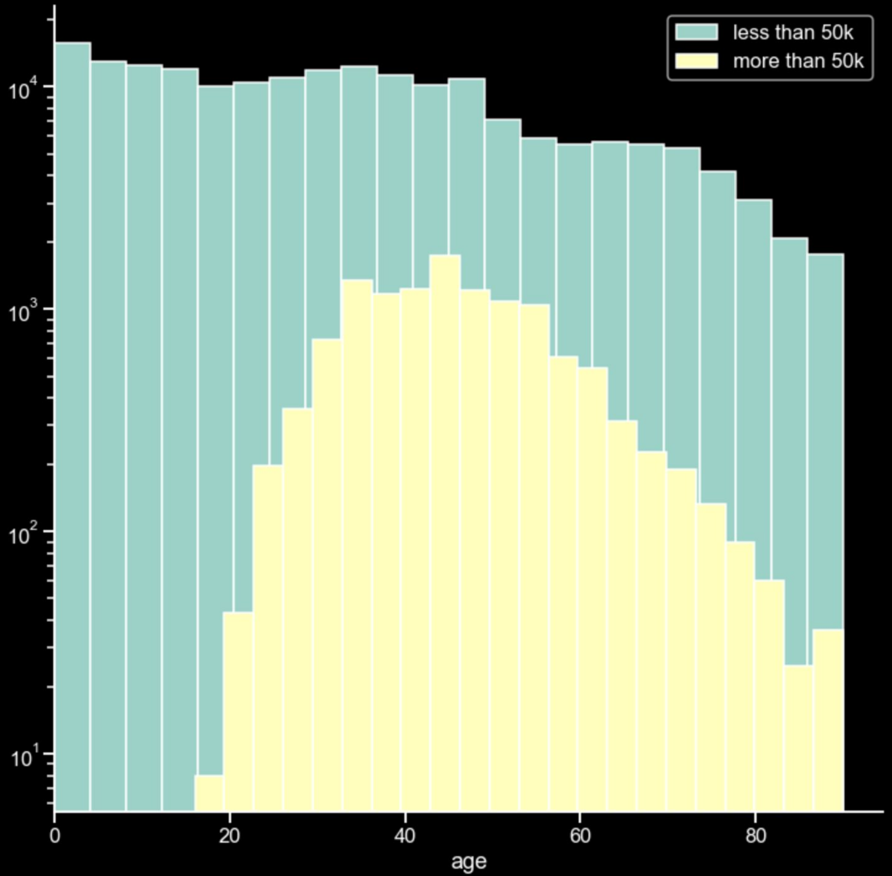


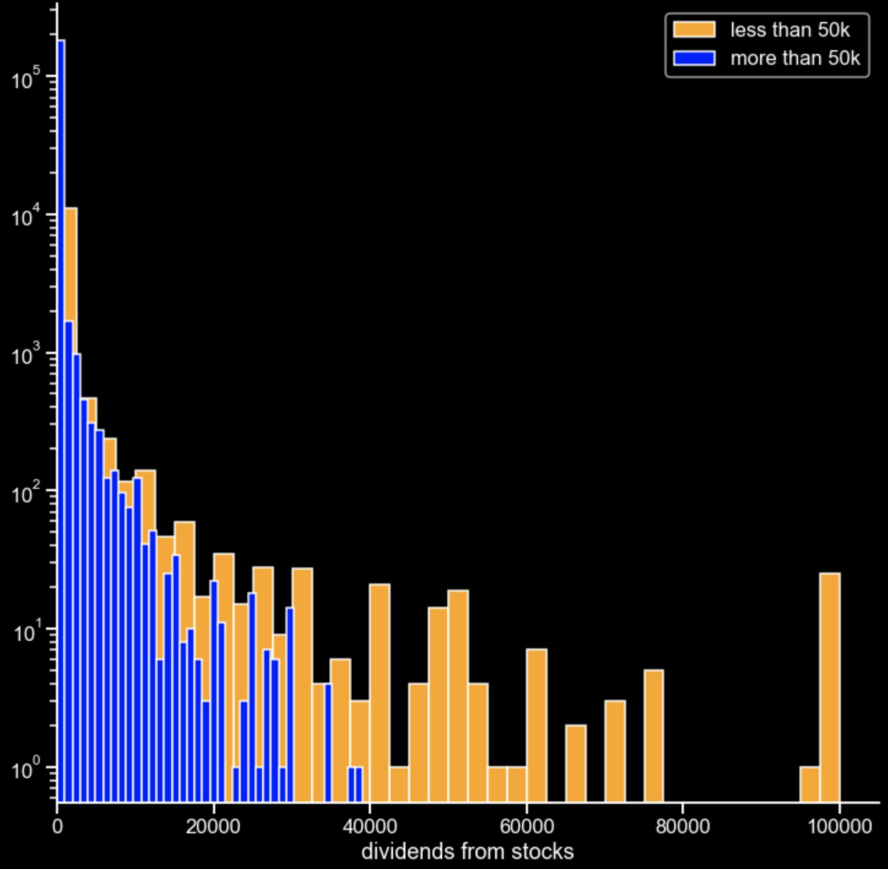
missing values



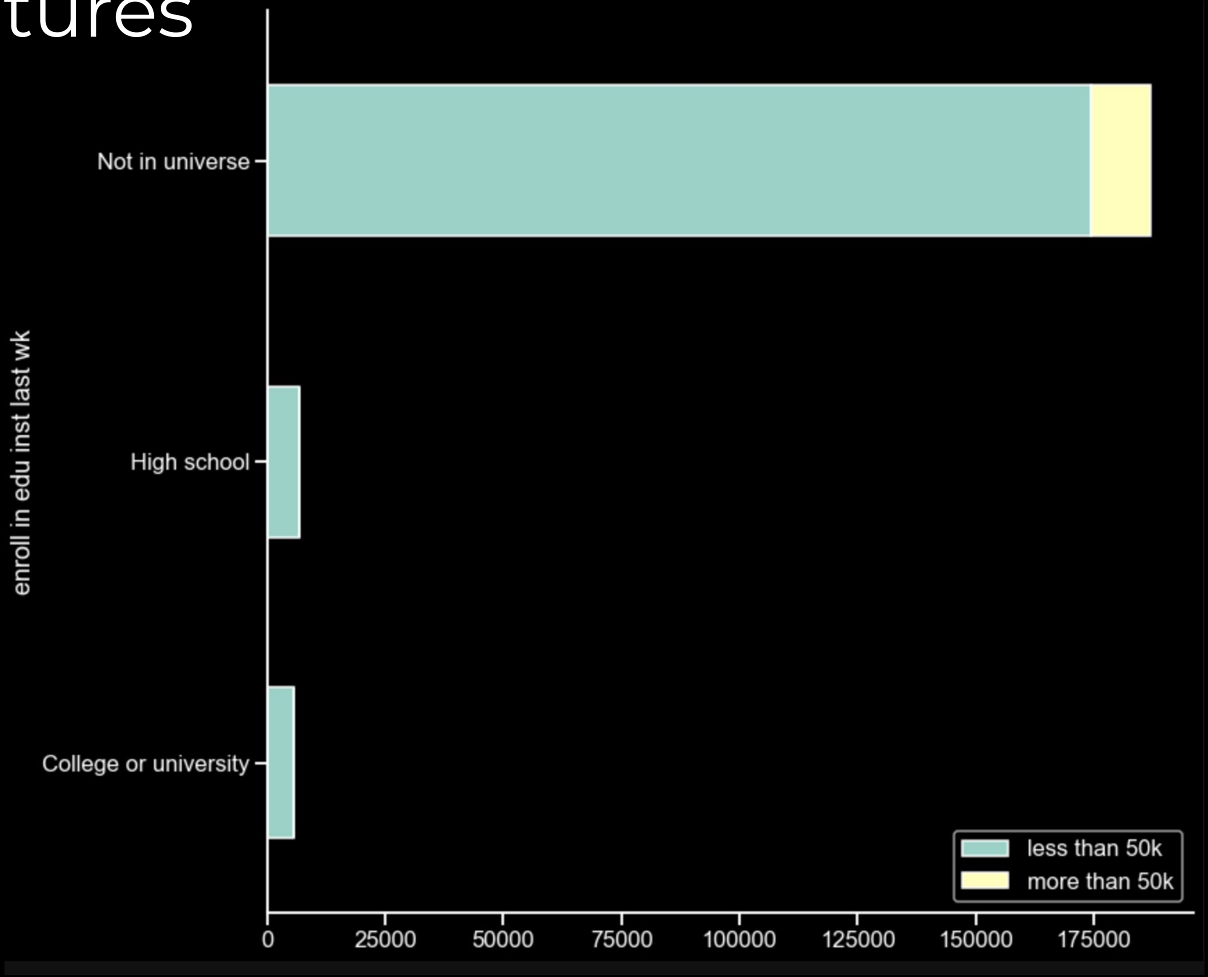
- 874 entries from 'hispanic origin'
- merged with 'All other'

numerical features



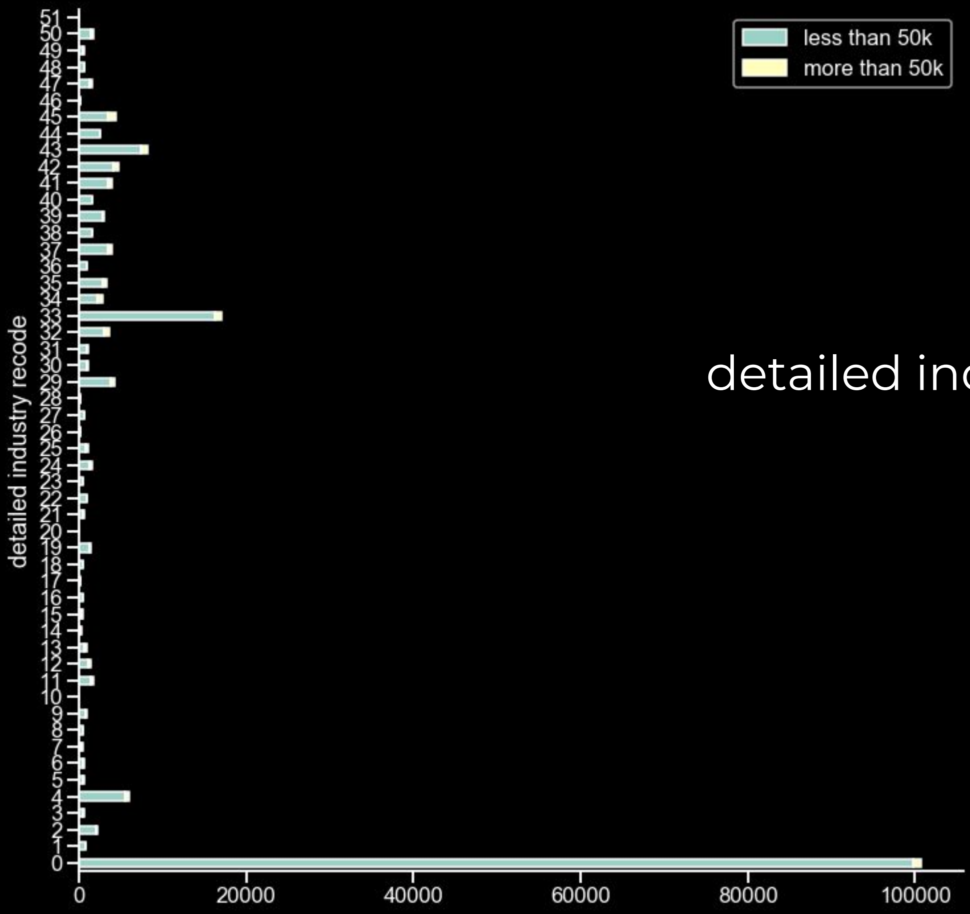


categorical features



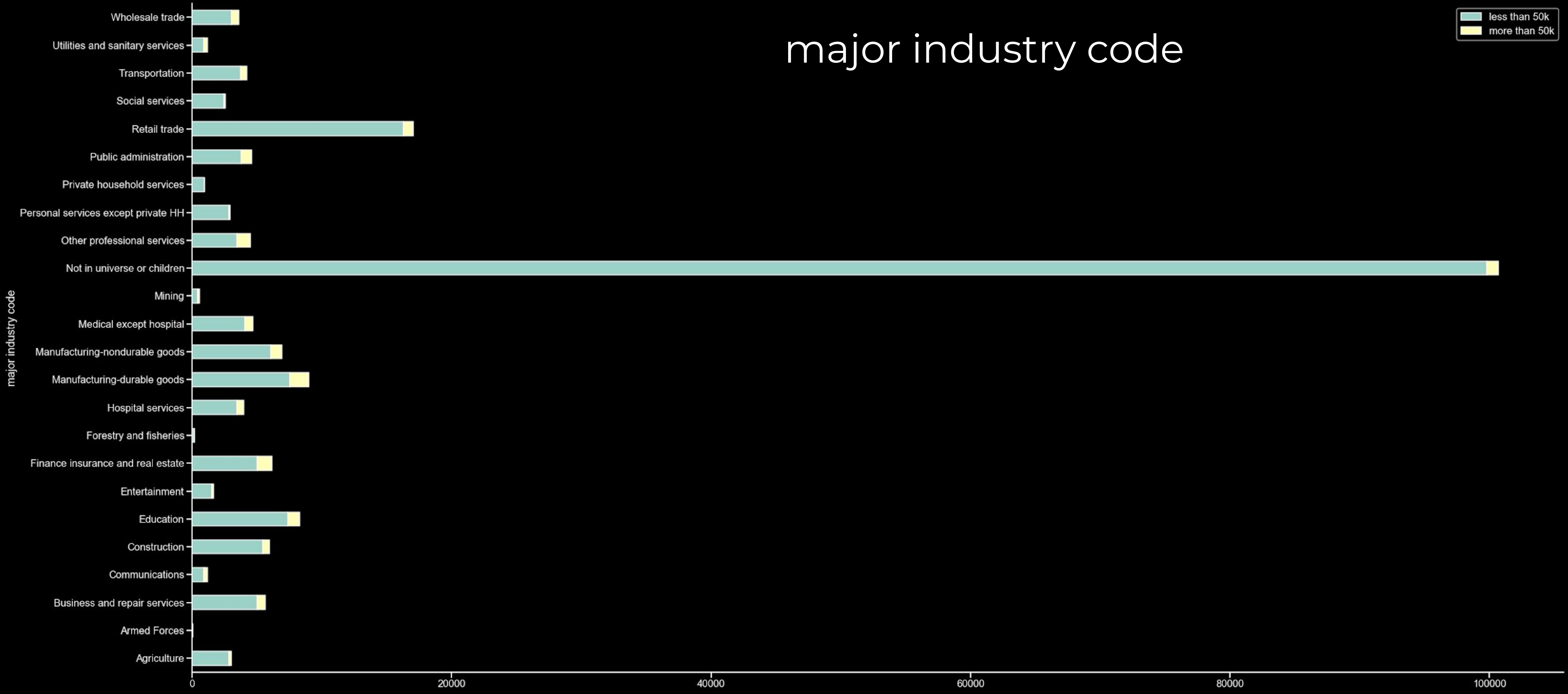
cleaning

- 1/0 encoding of binary features (e.g. label, sex, year)
- one-hot encoding of categorical features
- drop duplicate features (high linear correlation)

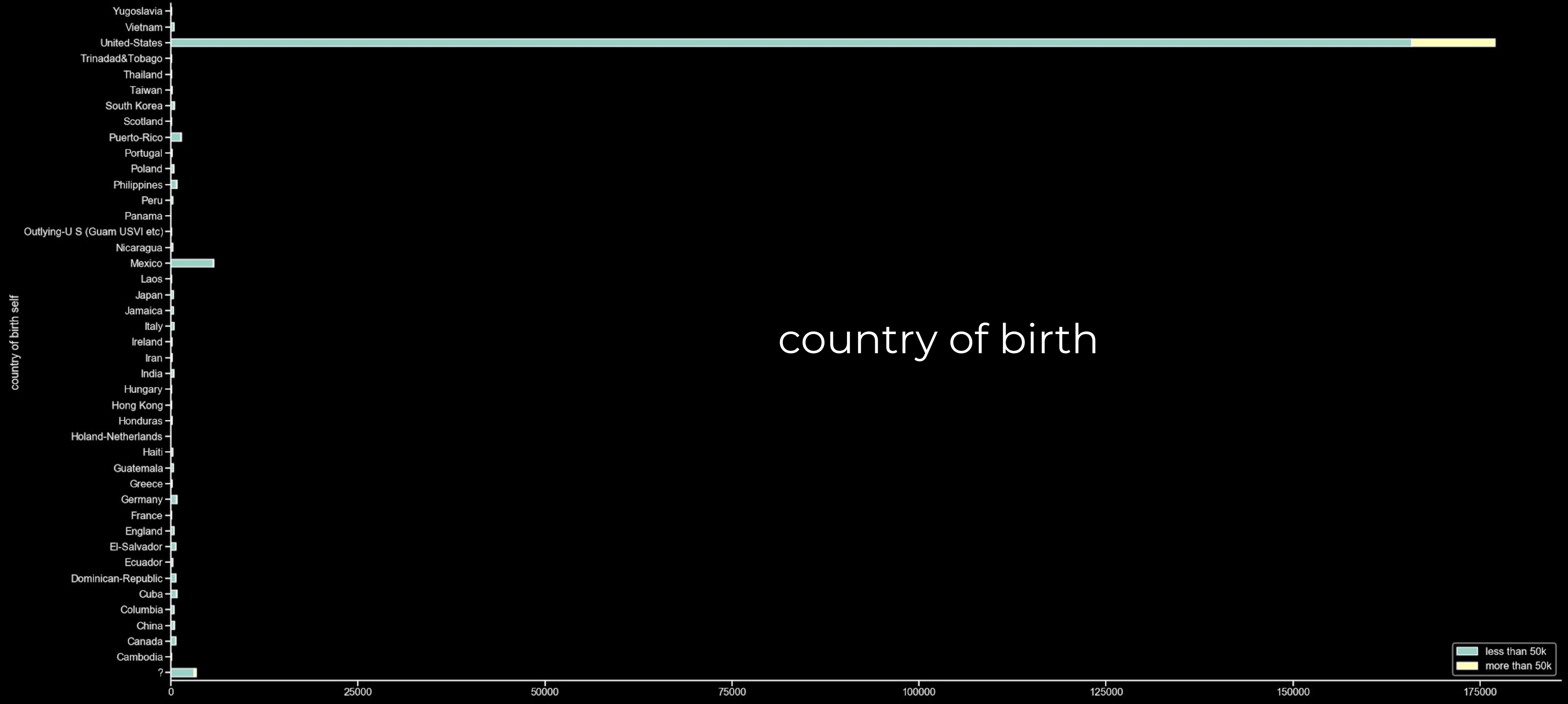


detailed industry recode

major industry code

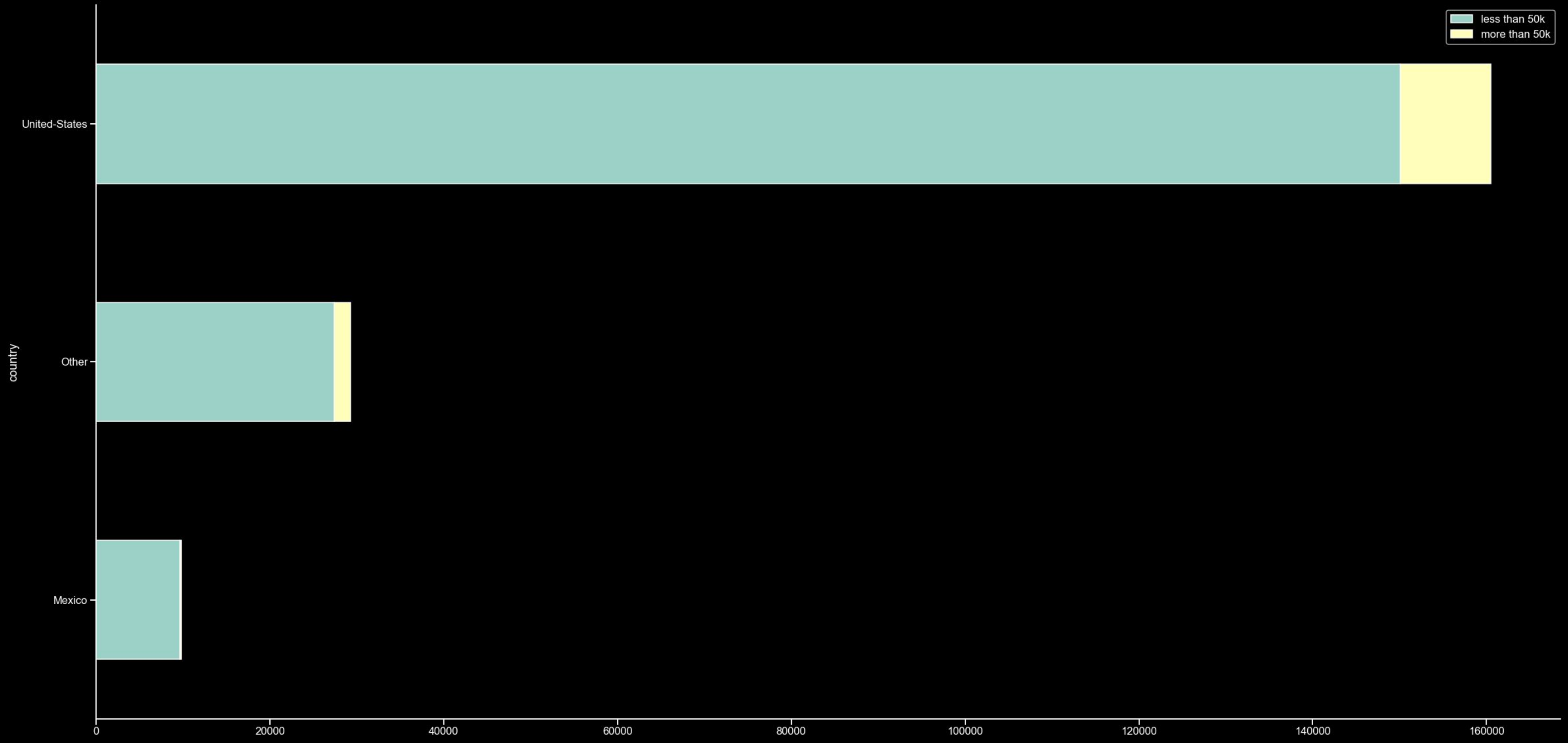


more cleaning



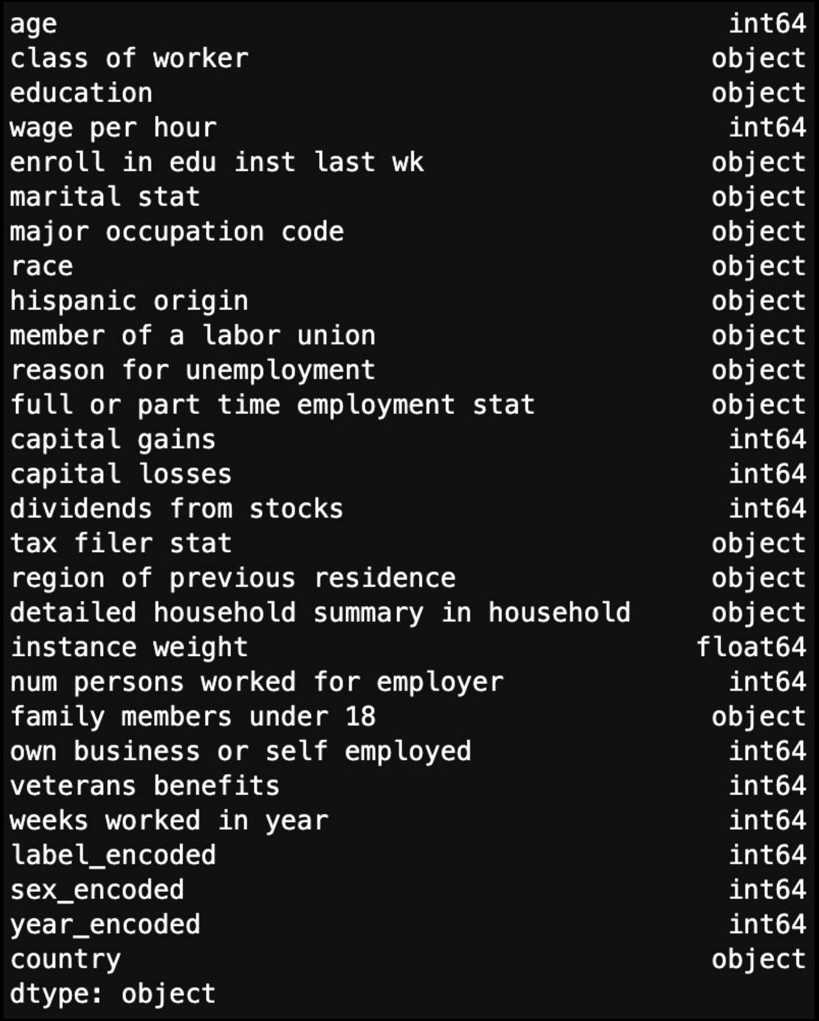
country of birth

more cleaning



cleaned data

- 26+2 features (before one-hot encoding)

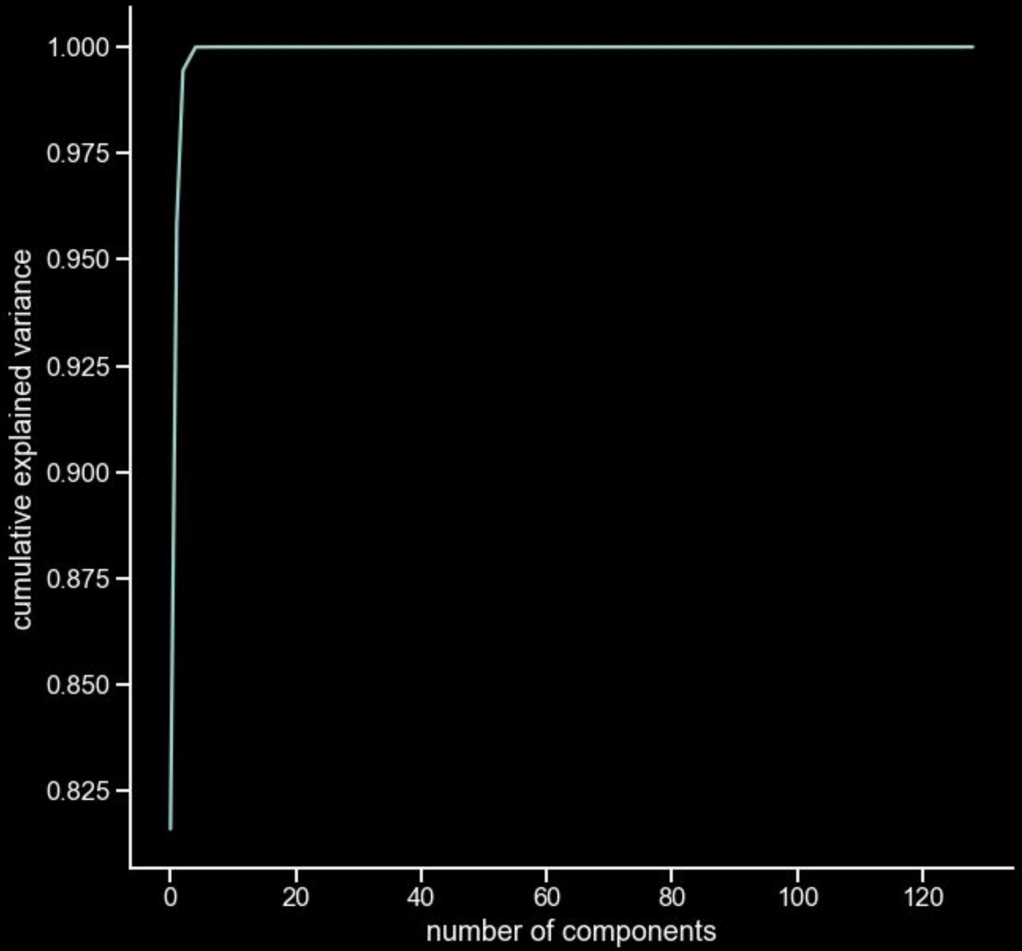


```
age                                          int64
class of worker                             object
education                                   object
wage per hour                                int64
enroll in edu inst last wk                  object
marital stat                                object
major occupation code                       object
race                                        object
hispanic origin                             object
member of a labor union                     object
reason for unemployment                     object
full or part time employment stat           object
capital gains                                int64
capital losses                               int64
dividends from stocks                        int64
tax filer stat                              object
region of previous residence                object
detailed household summary in household     object
instance weight                            float64
num persons worked for employer              int64
family members under 18                     object
own business or self employed                int64
veterans benefits                            int64
weeks worked in year                         int64
label_encoded                                int64
sex_encoded                                  int64
year_encoded                                 int64
country                                     object
dtype: object
```

PCA

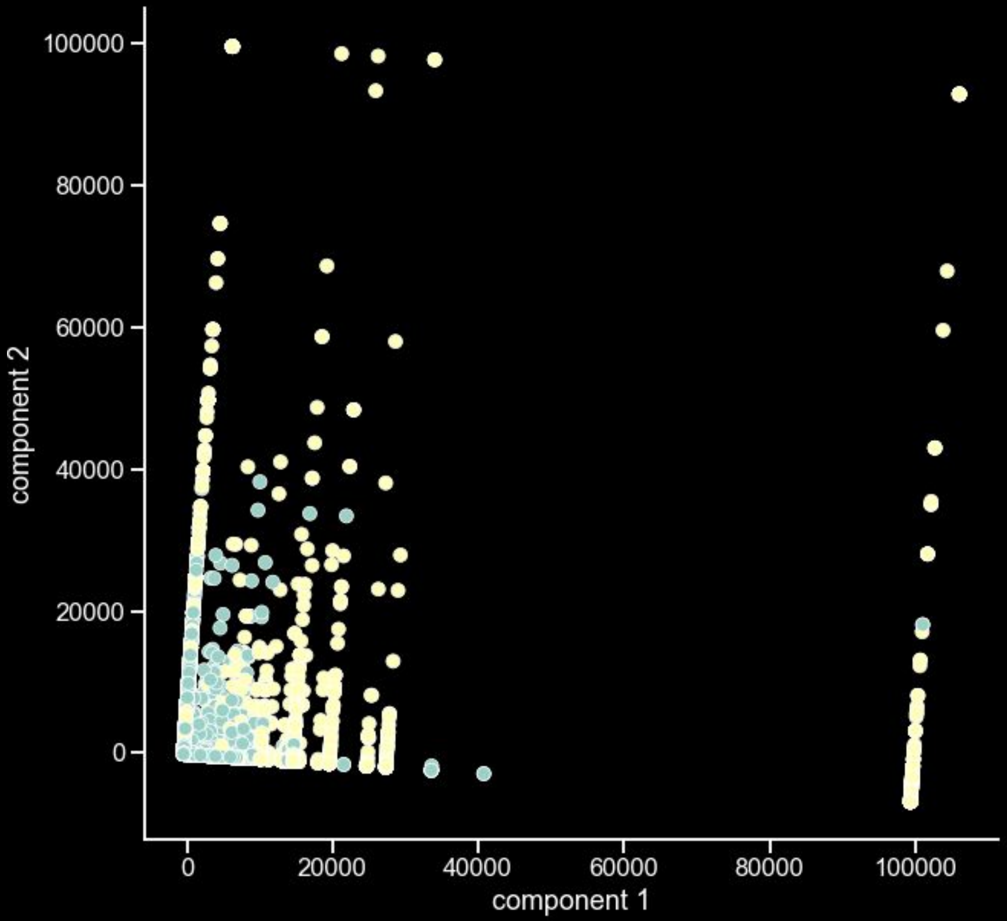
- explained variance:

  2 components overwhelm



PCA: visualized

- PCA projection onto first two components
- data squashed near origin

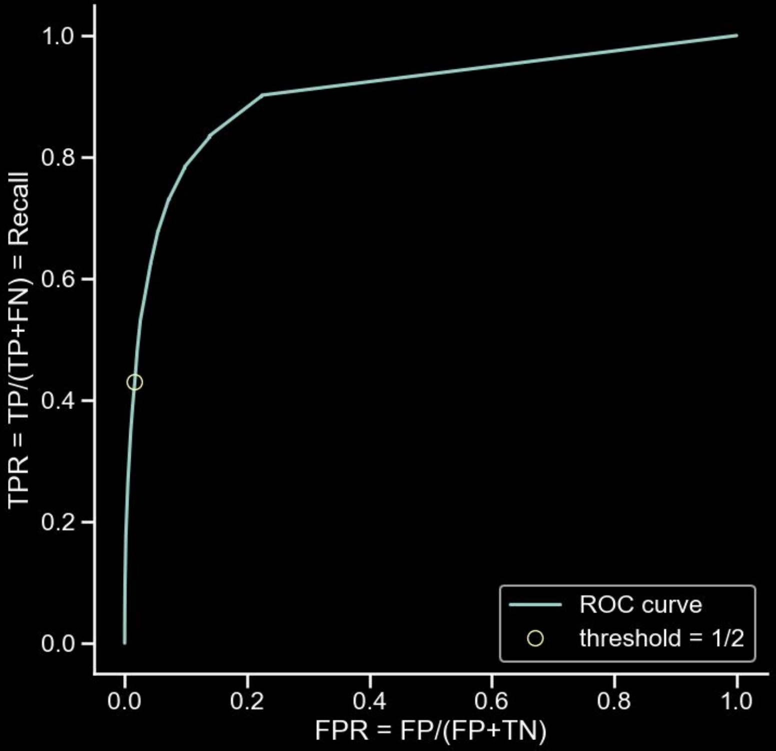


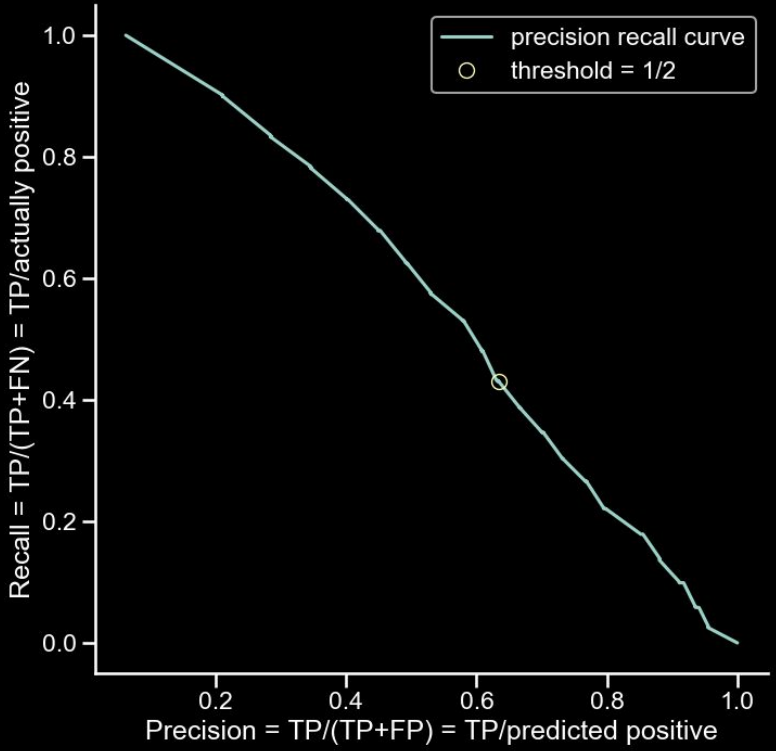
income prediction

- 80/20 train/test split
- logistic regression + L1 regularization: gridsearch+CV
- random forest with 20 trees

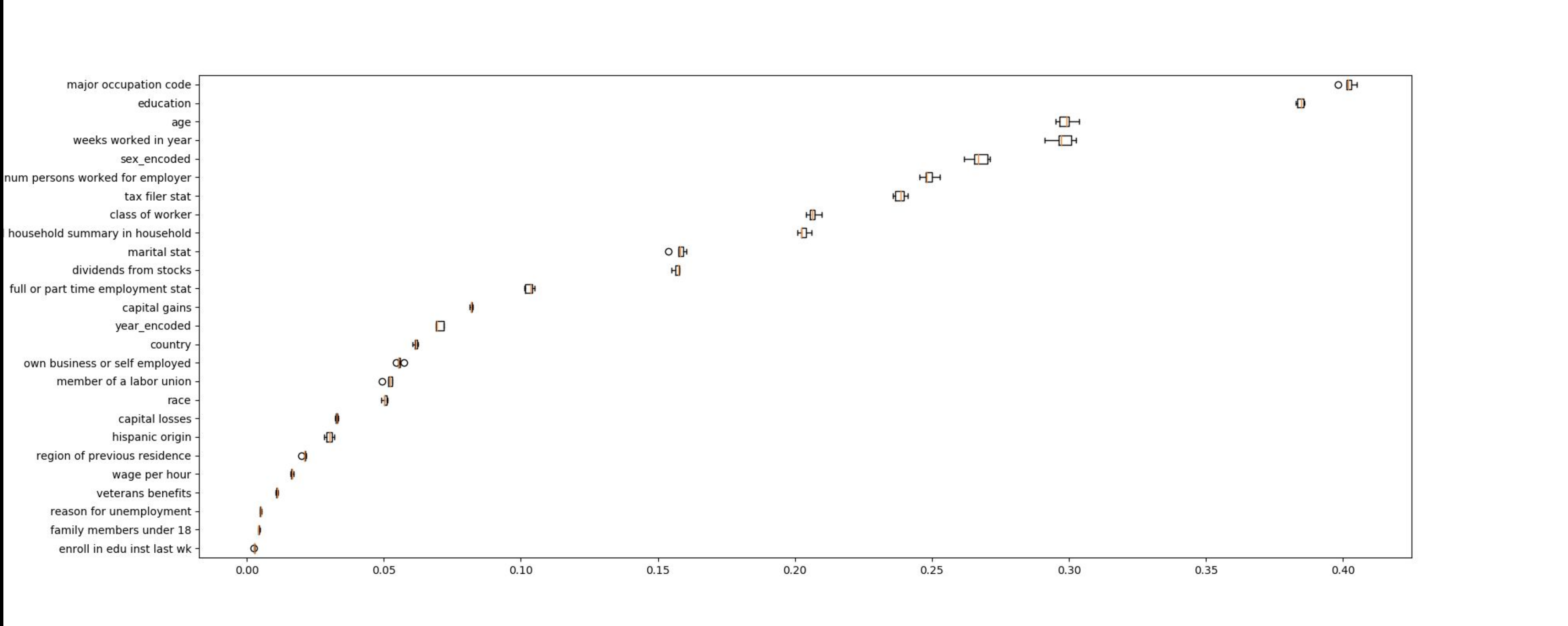
random forest performed better

roc and precision/recall

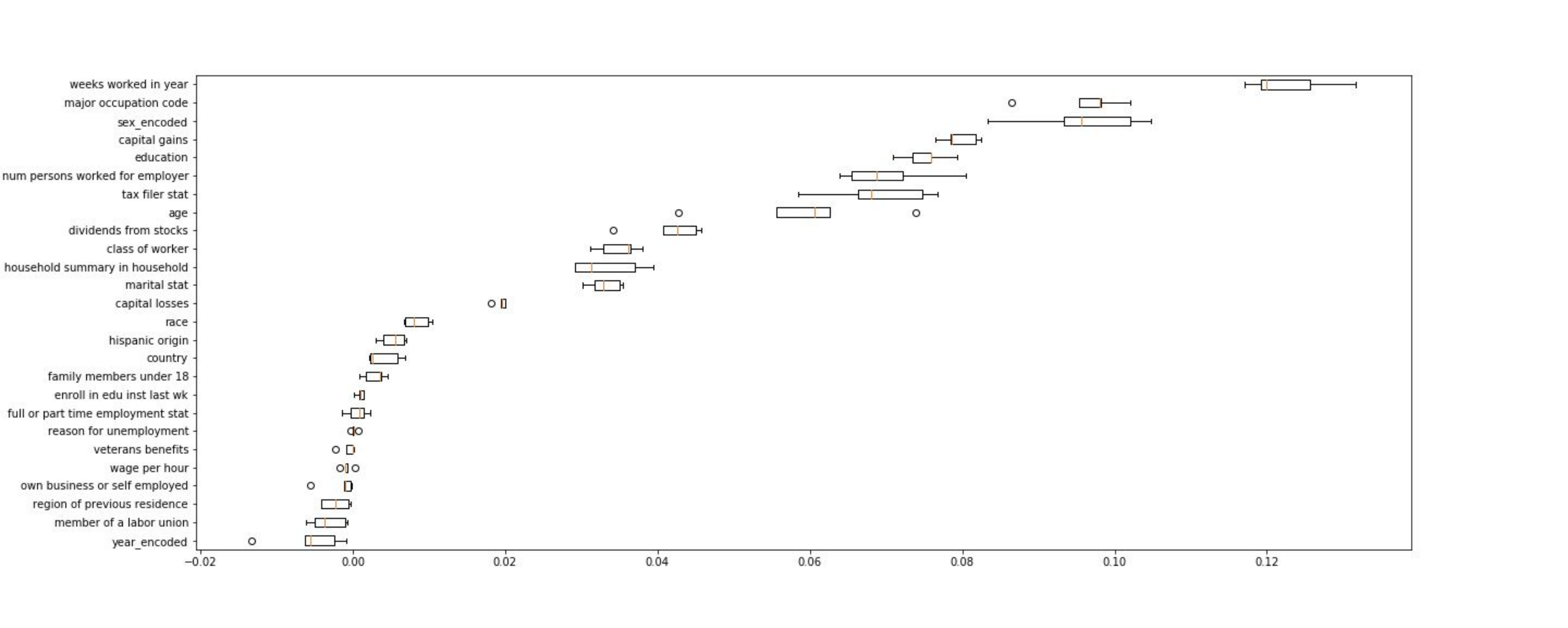




feature importance (permutation): train



feature importance (permutation): test

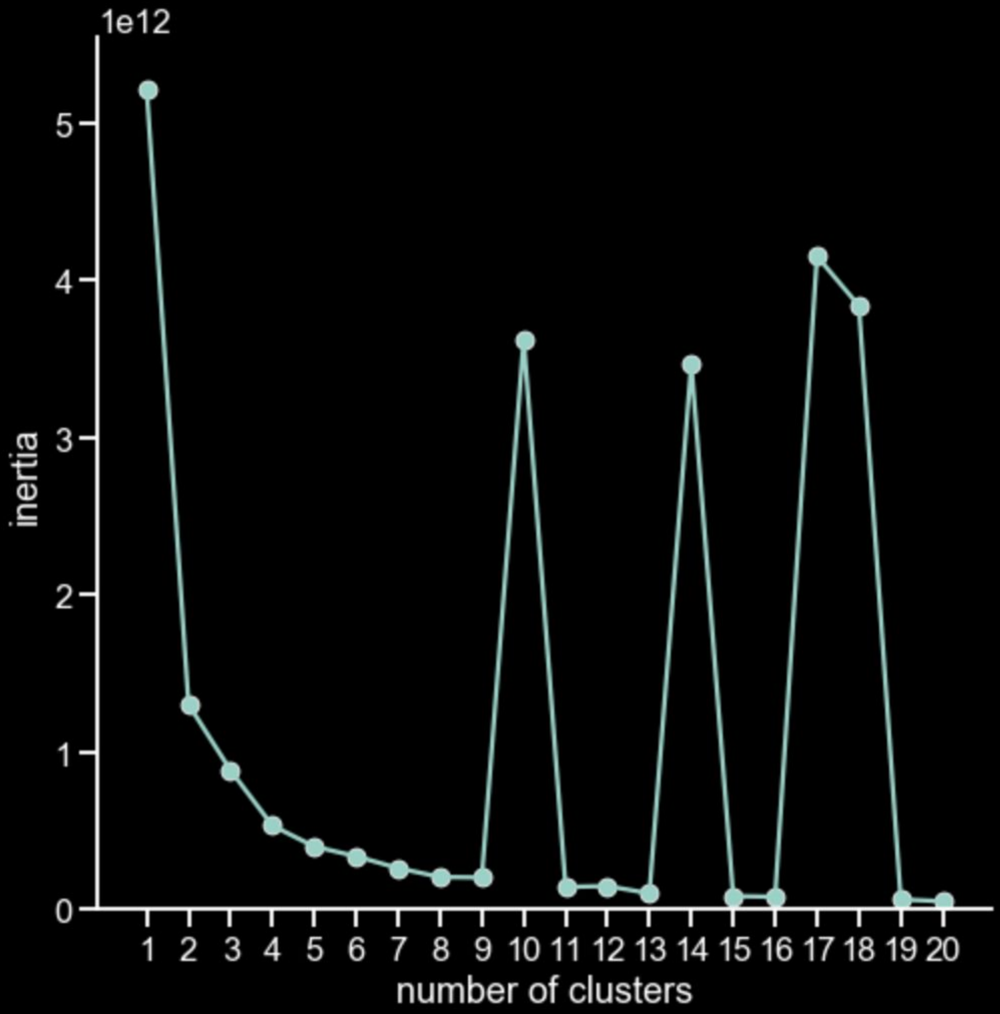


segmentation

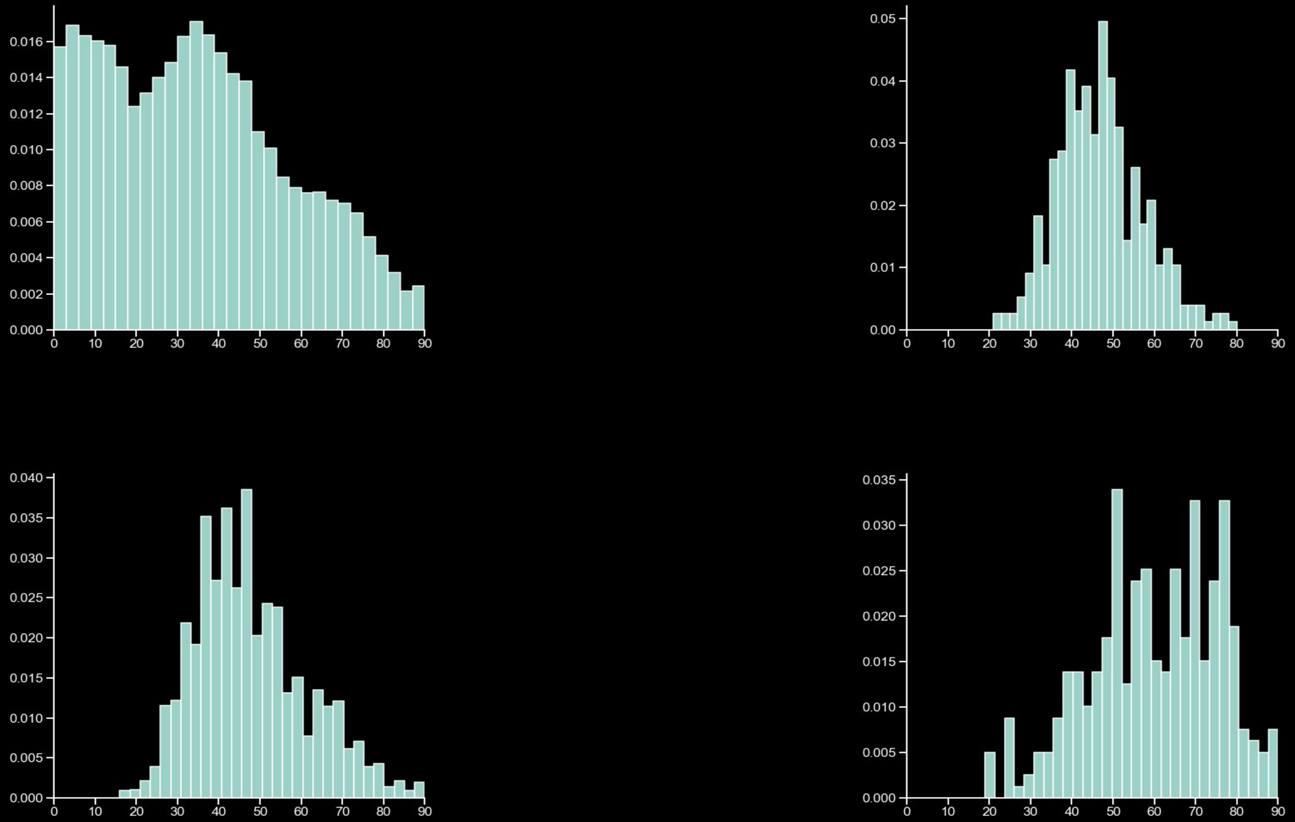
- ~~hierarchical clustering~~ (too slow)
- k-means (with mini-batches)

how many clusters?

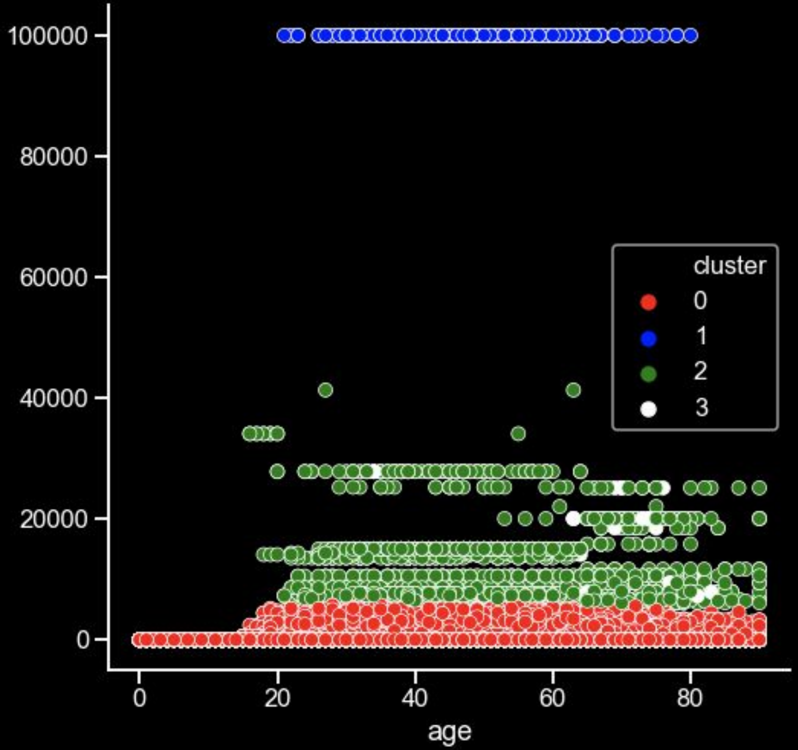
- mini-batch k-means
- batch size = 10k
  - (5% of samples)
- --> 4 clusters



what do the clusters look like?



what do the clusters look like?



next steps

- income prediction
    - fancier tree model (gradient boosted, XGBoost)
    - increase feature selection
- segmentation
    - different clustering method
    - with clear goal: rules + clustering

thanks