Take home test

US Census archive

EDA

	AAGE	AHRSPAY	WKSWORK	capital_change	NOEMP	income
AAGE	1.00000	-0.06672	-0.35480	0.06682	-0.33315	0.06147
AHRSPAY	-0.06672	1.00000	0.13910	-0.01635	0.13944	-0.00252
WKSWORK	-0.35480	0.13910	1.00000	0.03057	0.62499	0.20187
capital_change	0.06682	-0.01635	0.03057	1.00000	0.00667	0.25877
NOEMP	-0.33315	0.13944	0.62499	0.00667	1.00000	0.15727
income	0.06147	-0.00252	0.20187	0.25877	0.15727	1.00000

- Correlation matrix (example output)
  - Not correlated (p<0.7) features and target.
- Reducing dimensions and variability by clustering categories
  - Married status: 3 values instead of 7 (never married, spouse present, spouse absent)
  - Education level high school or above
  - Class of worker: 3 values instead of 6 (self, gov, private)
  - etc.

- Removing or imputing missing values

Resulting in 115,134 training samples, 57,436 test samples

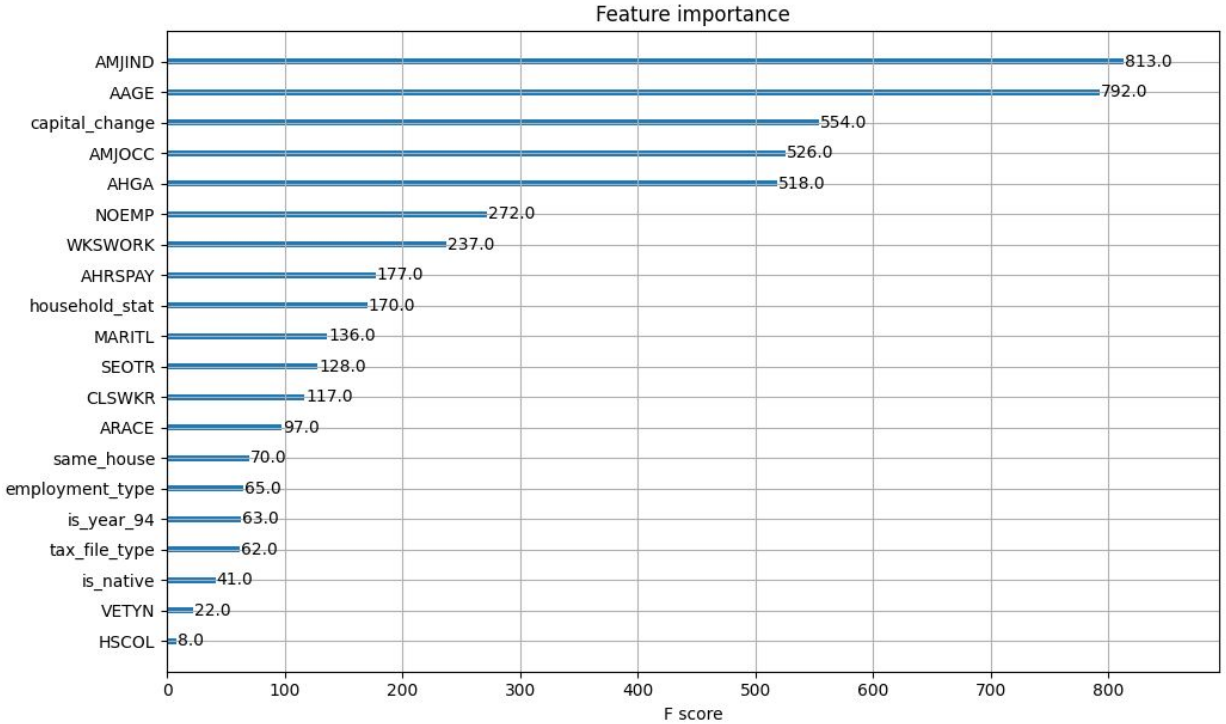
Data Preparation

- Categorical features - One-Hot encoding
- Target Binarization (0 = 50000-, 1 = 50000+)
- Numerical features - Normalize
- Remove features and engineer new features
  - E.g. capital_change = CAPGAIN - CAPLOSS + DIVVAL
  - is_native = PRCITSHP contains 'Native'
- Removed features for simplification or noise reduction

Data Modeling

Using XGBOOST, sort of advanced decision trees algorithm.

Feature importance -



Model Assessment

- Original
  - Recall: 0.54
  - Precision: 0.57
  - F1: 0.55
- Balanced
  - Recall: 0.86
  - Precision: 0.21
  - F1: 0.34
- Categorical
  - Recall: 0.57
  - Precision: 0.50
  - F1: 0.53

Results

- Industry, age, capital movement, occupation, education explain 66% of the model decision process.
- Base model and category-focused model perform similarly while the re-sampled data yielded lower F score. Perhaps in scenarios where **False Negatives** are needed low, this model could be used.
- Ideas for improvement:
  - Continue work on features engineering and noise reduction
  - Cross validation and grid search for hyperparameters tuning
- Different re-sampling method