Junjie Yu

Income Level Prediction

Overview

- Data:
  - Extraction of 1994 US census database
  - Approximately 32,000 observations with 14 variables
  - Source: http://archive.ics.uci.edu/ml/datasets/Adult

- Objective:
 - To predict whether a person makes over 50K a year.

Data

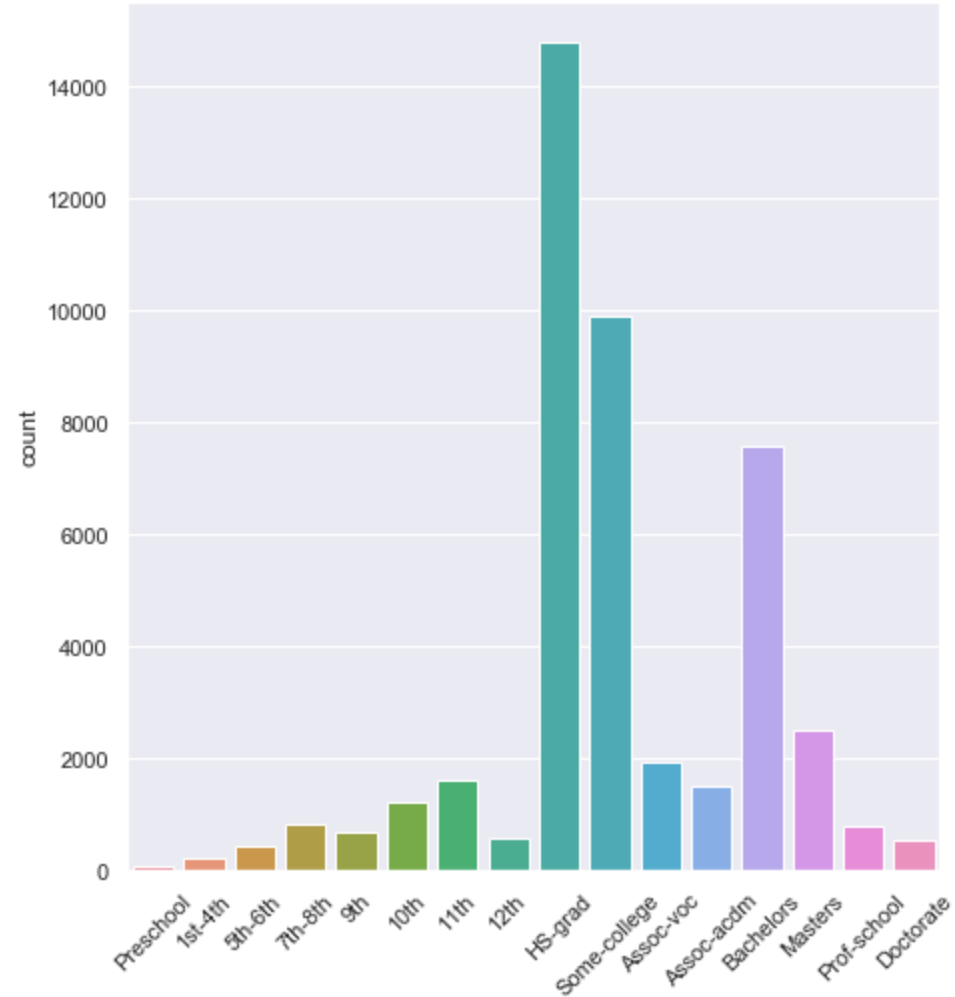| age	workclass	Education_level	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week	native-country	income
int	object	object	float	object	object	object	object	object	float	float	float	object	object
39	State-gov	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States	<=50K
50	Self-emp-not-inc	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K
38	Private	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States	<=50K
53	Private	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States	<=50K
28	Private	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=50K

- 14 Variables: 1 int64 + 4 float64 + 9 object

EDA: Education, age, and hours per work
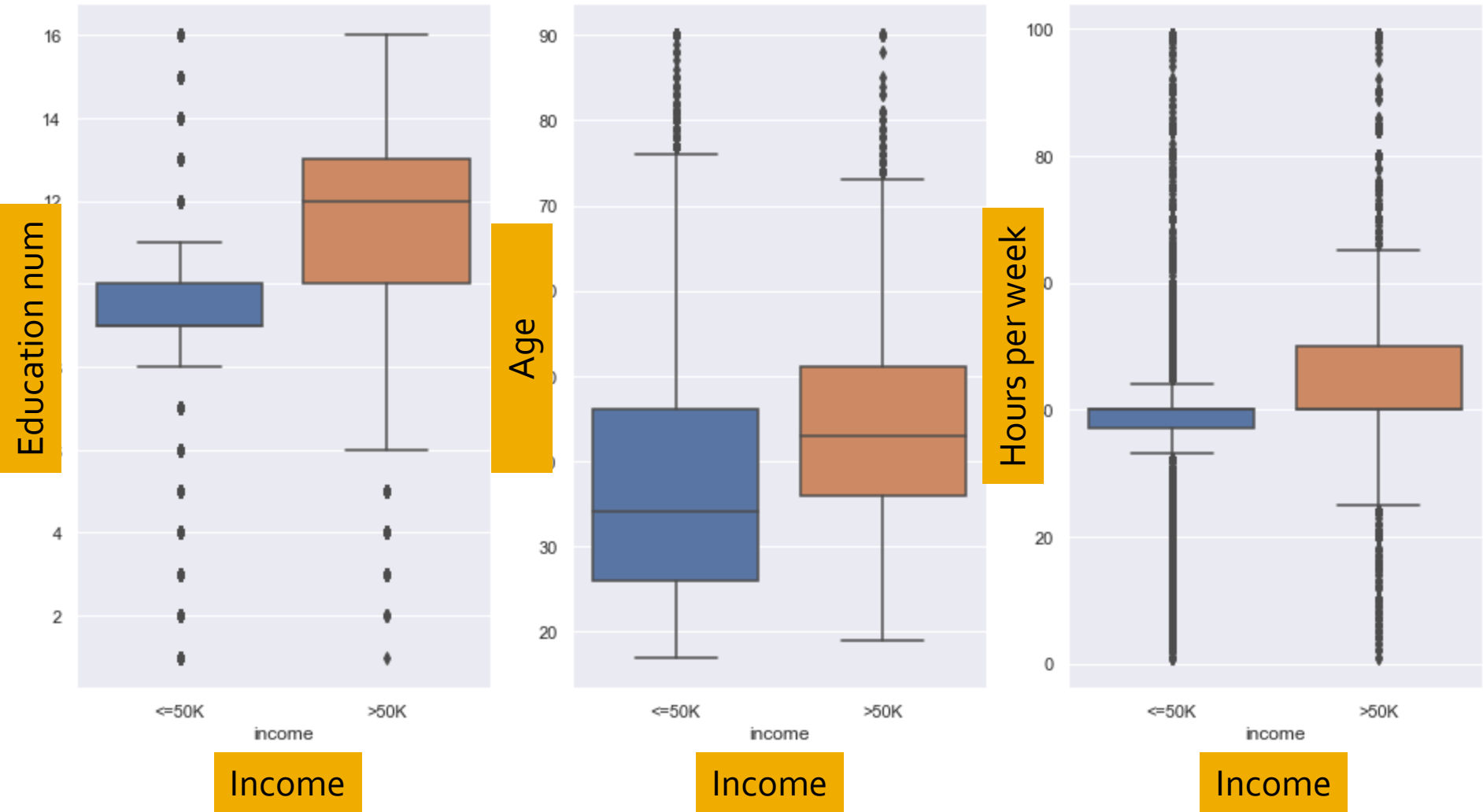
EDA: Education num and education level
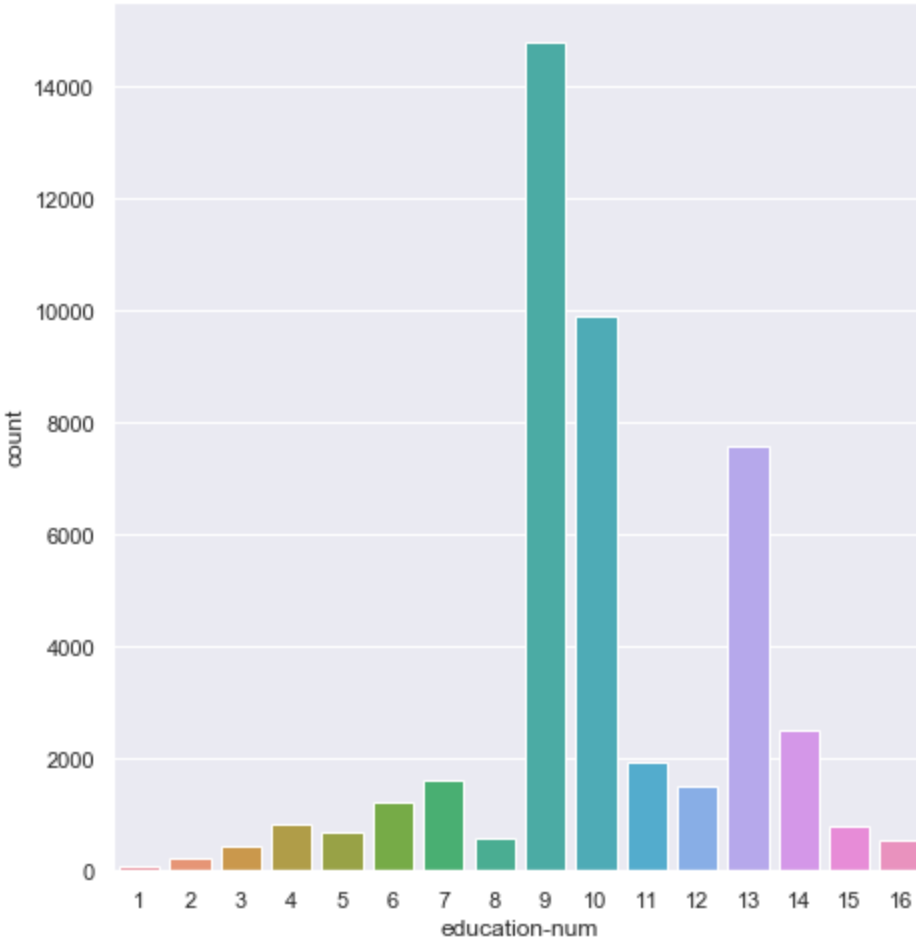
Education num

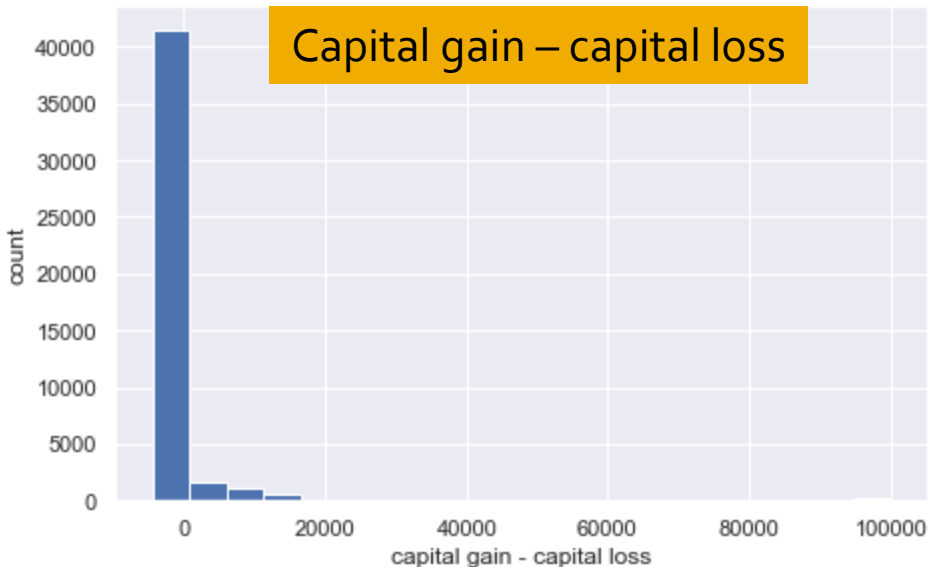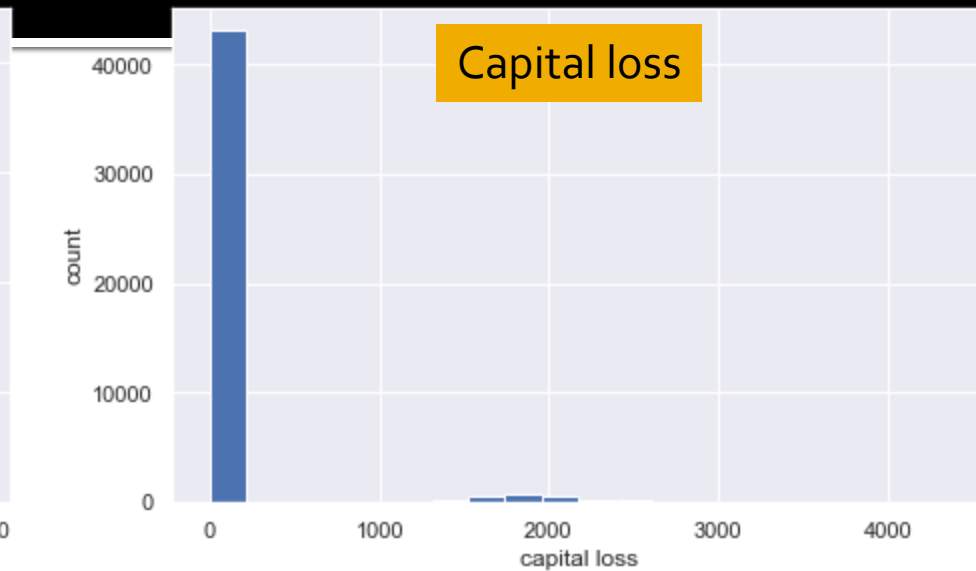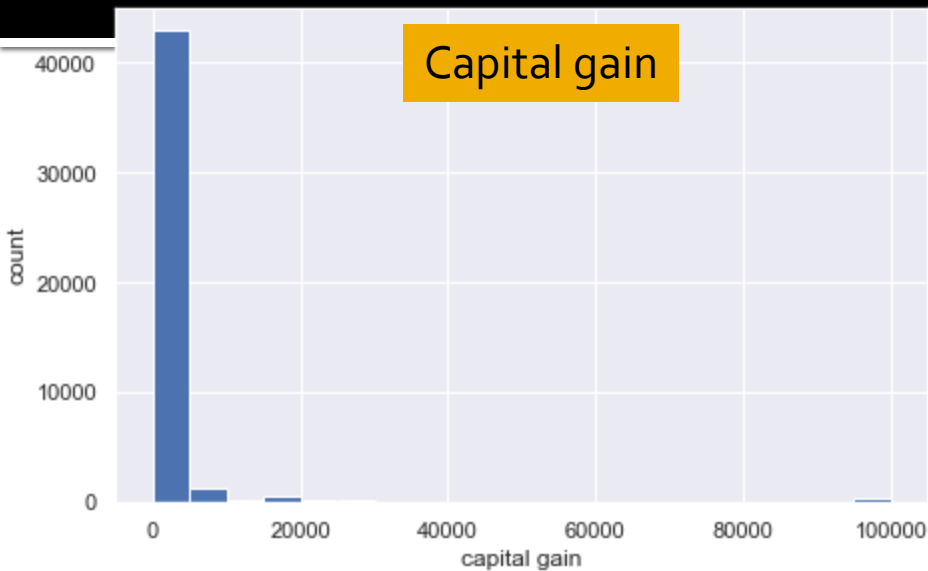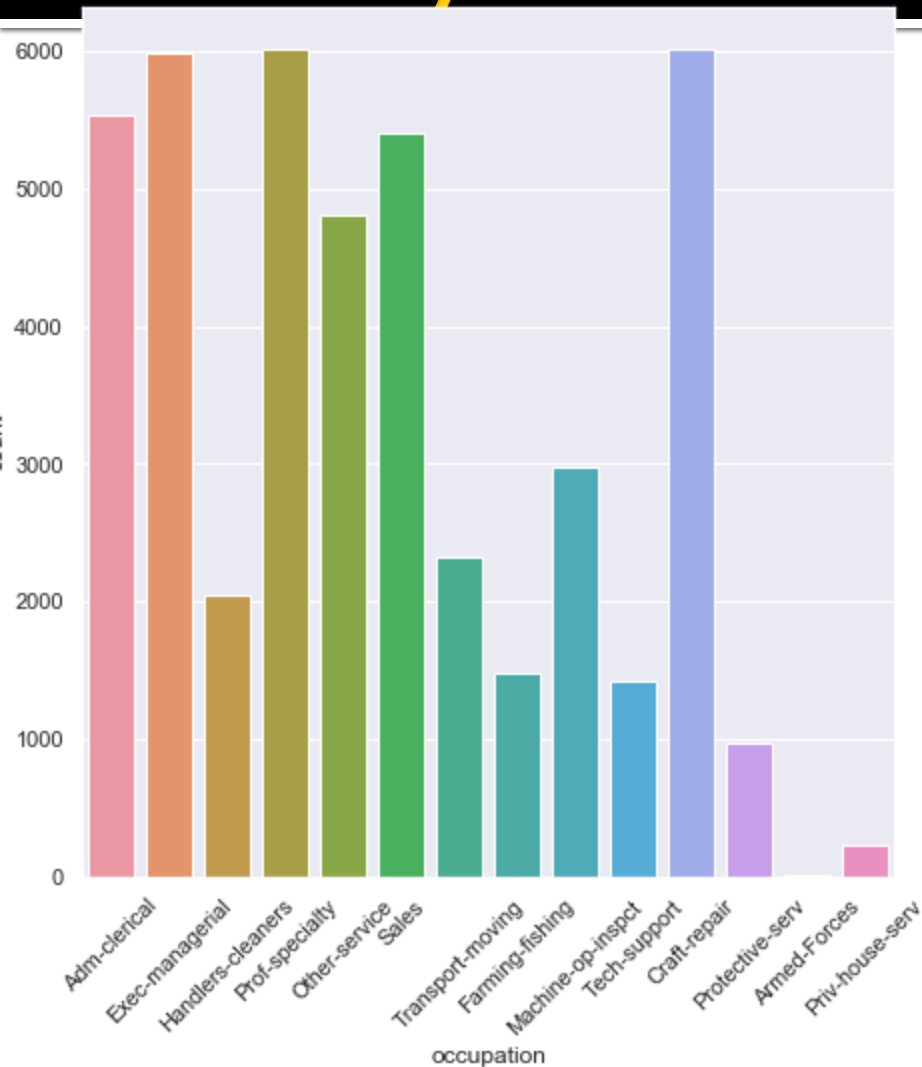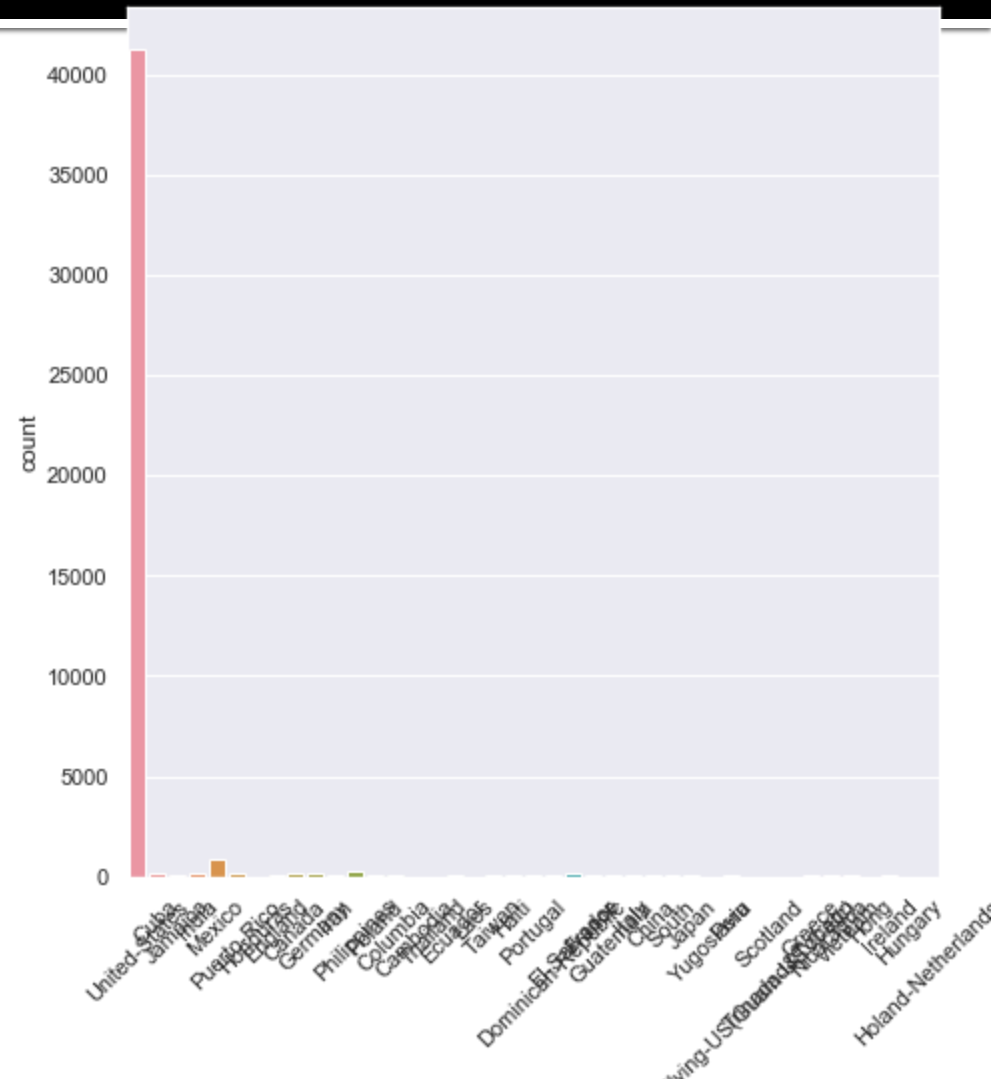Education level

EDA: Capital gain and capital loss
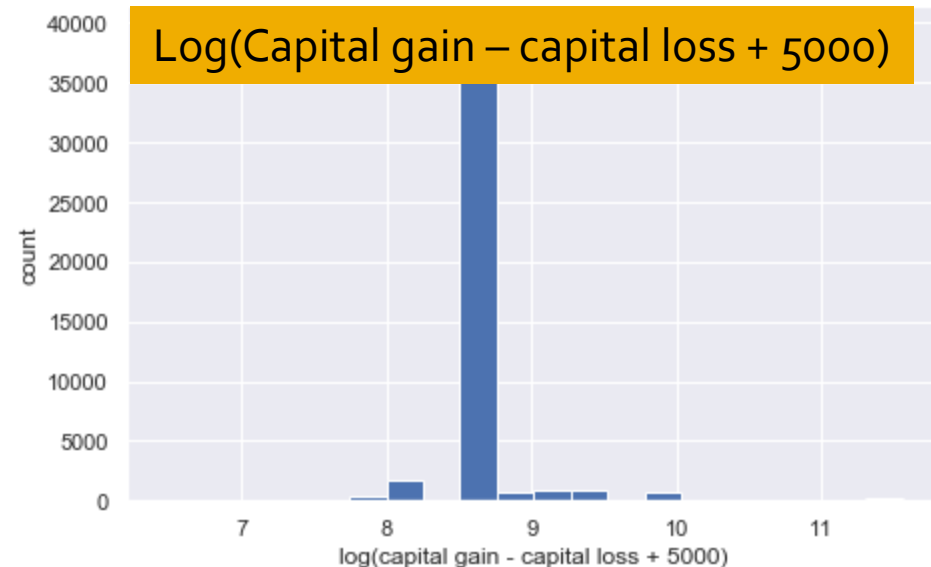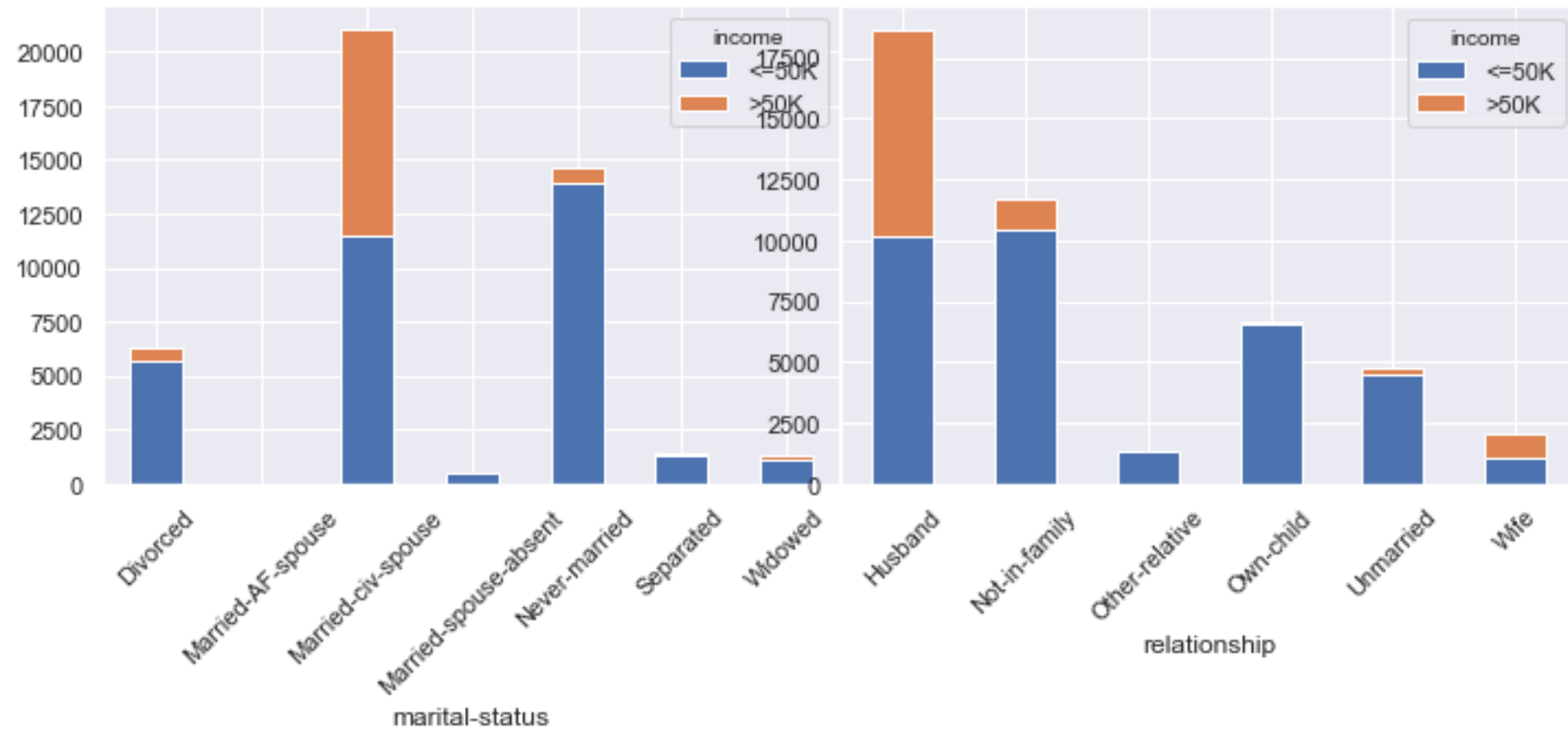
EDA: Occupation and native country

Occupation

Native country

EDA: Marital status and relationship
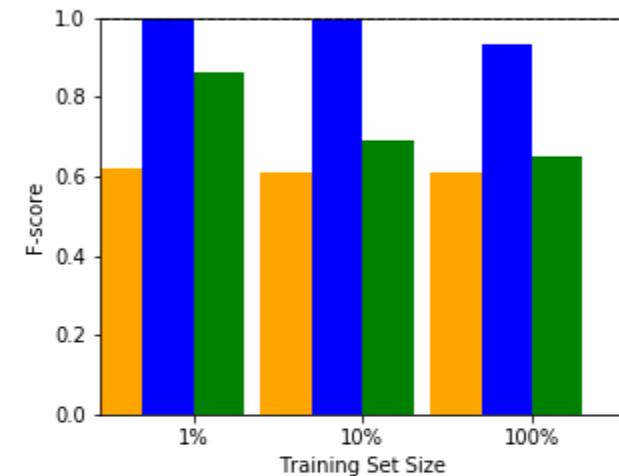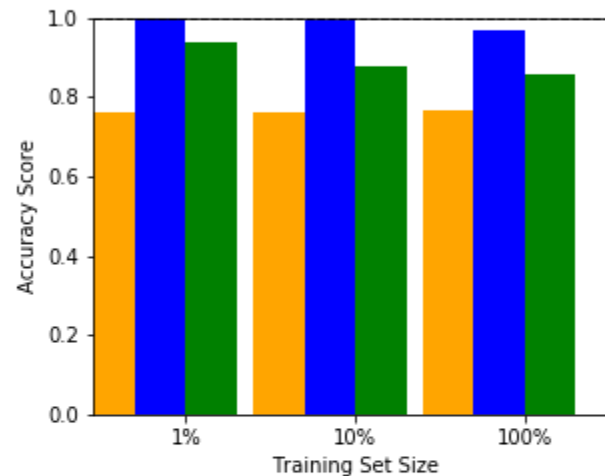

Marital status
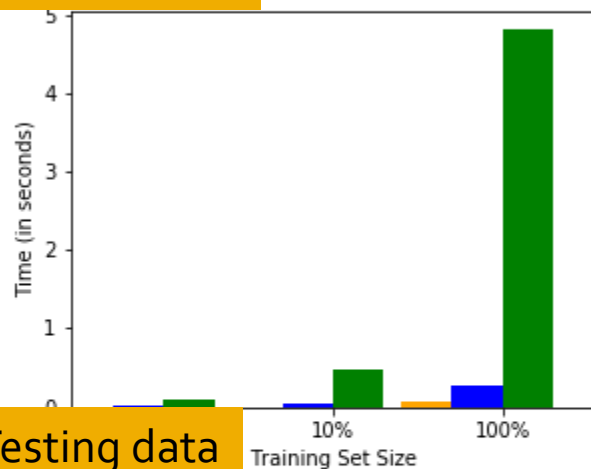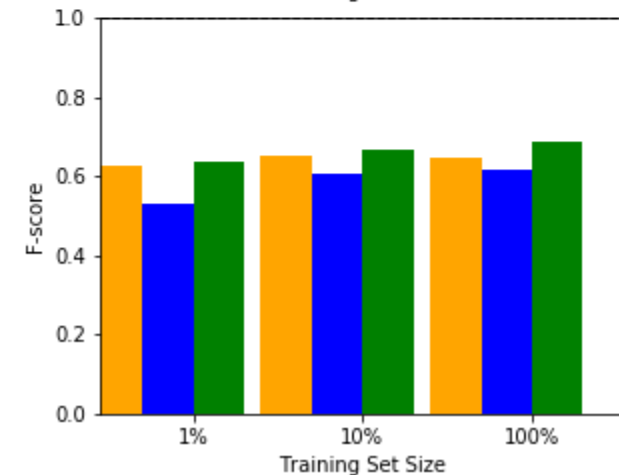
Relationship

Model Selection
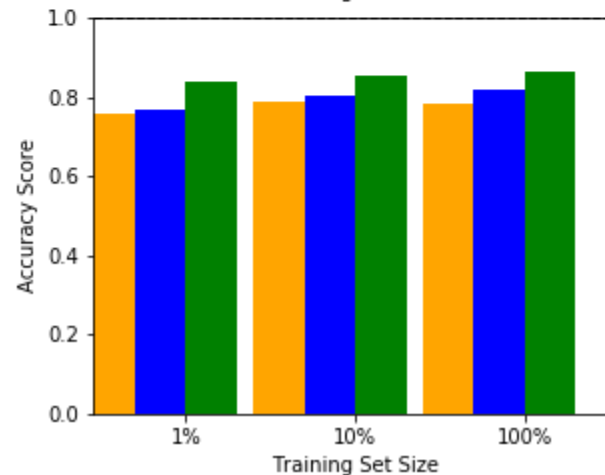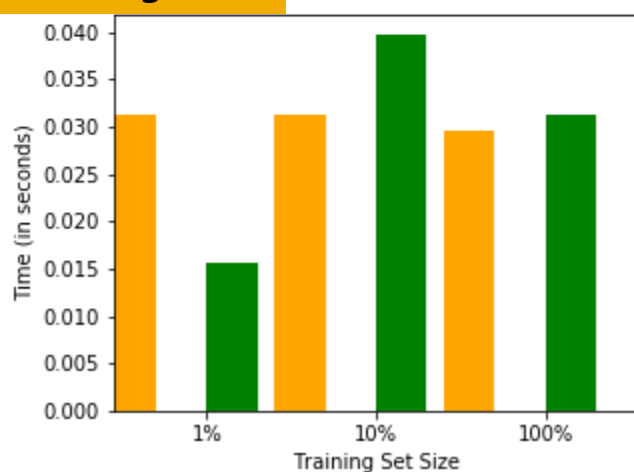
Feature Engineering

Initial
- Initial model with raw features

Education
- Discard duplicate feature education level

Capital
- Combine capital gain and capital loss

Native
- Encode native country as U.S. and other

Optimize
- Optimize hyper-parameters

Model Evolution

Accuracy	**0.8630**
Precision	0.7821
Recall	0.6073

Normalized Weights for First Five Most Predictive Features

Initial → Education → Capital → Native → Optimize
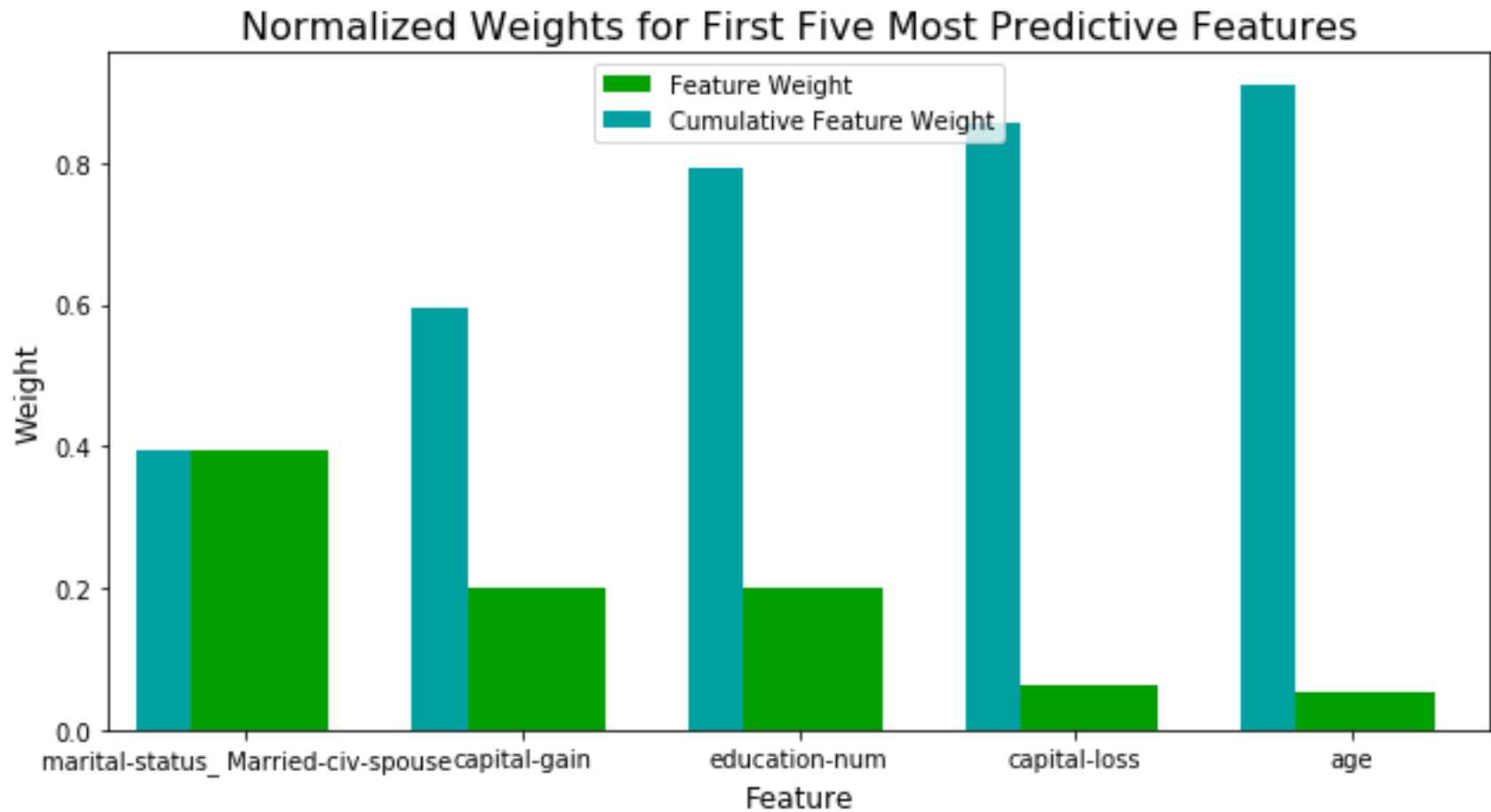
Model Evolution

Accuracy	0.8636
Precision	0.7831
Recall	0.6091
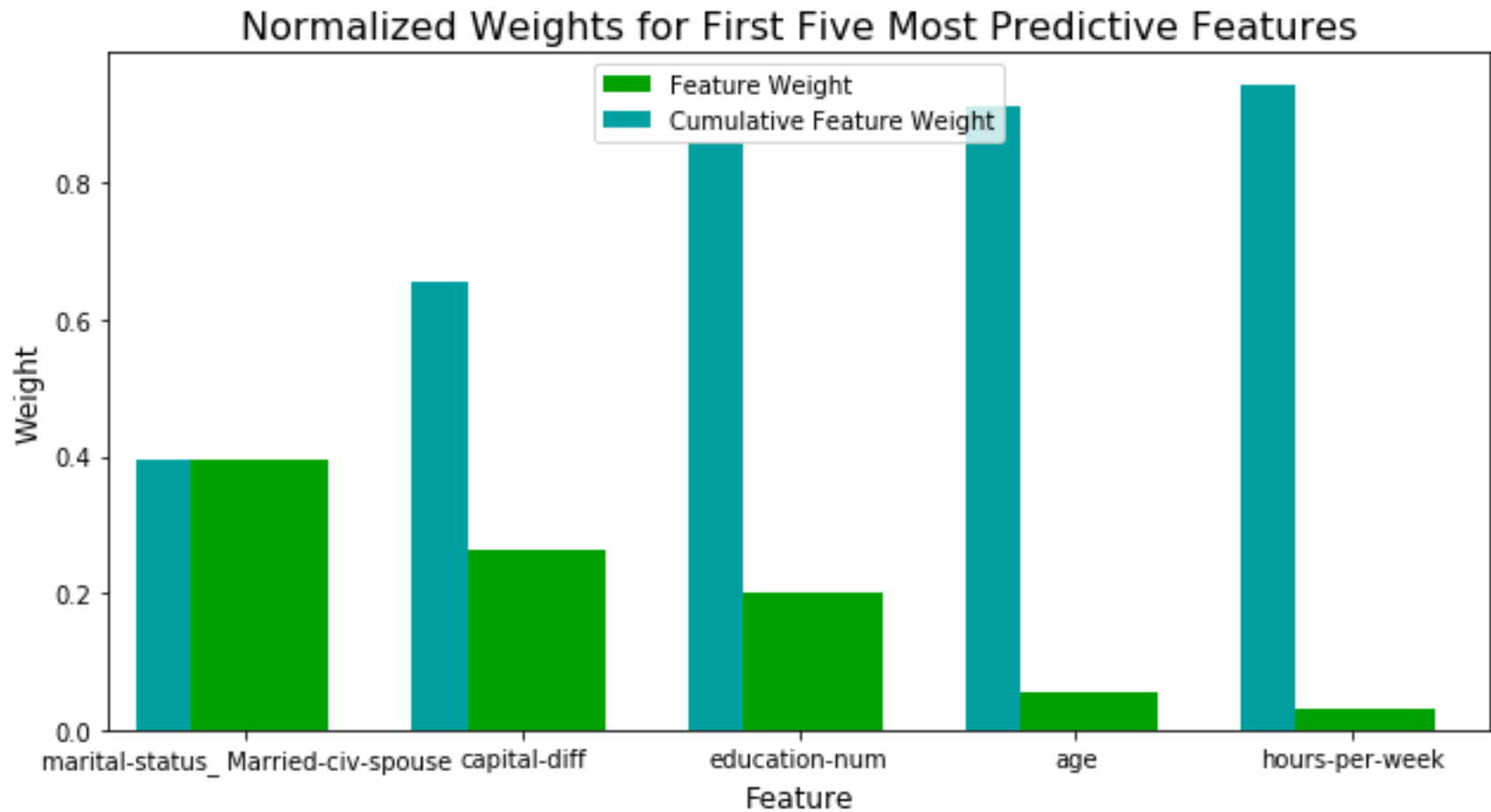
Normalized Weights for First Five Most Predictive Features

Model Evolution

Normalized Weights for First Five Most Predictive Features

Initial → Education → Capital → Native → Optimize

Model Evolution

Normalized Weights for First Five Most Predictive Features

Initial → Education → Capital → Native → Optimize
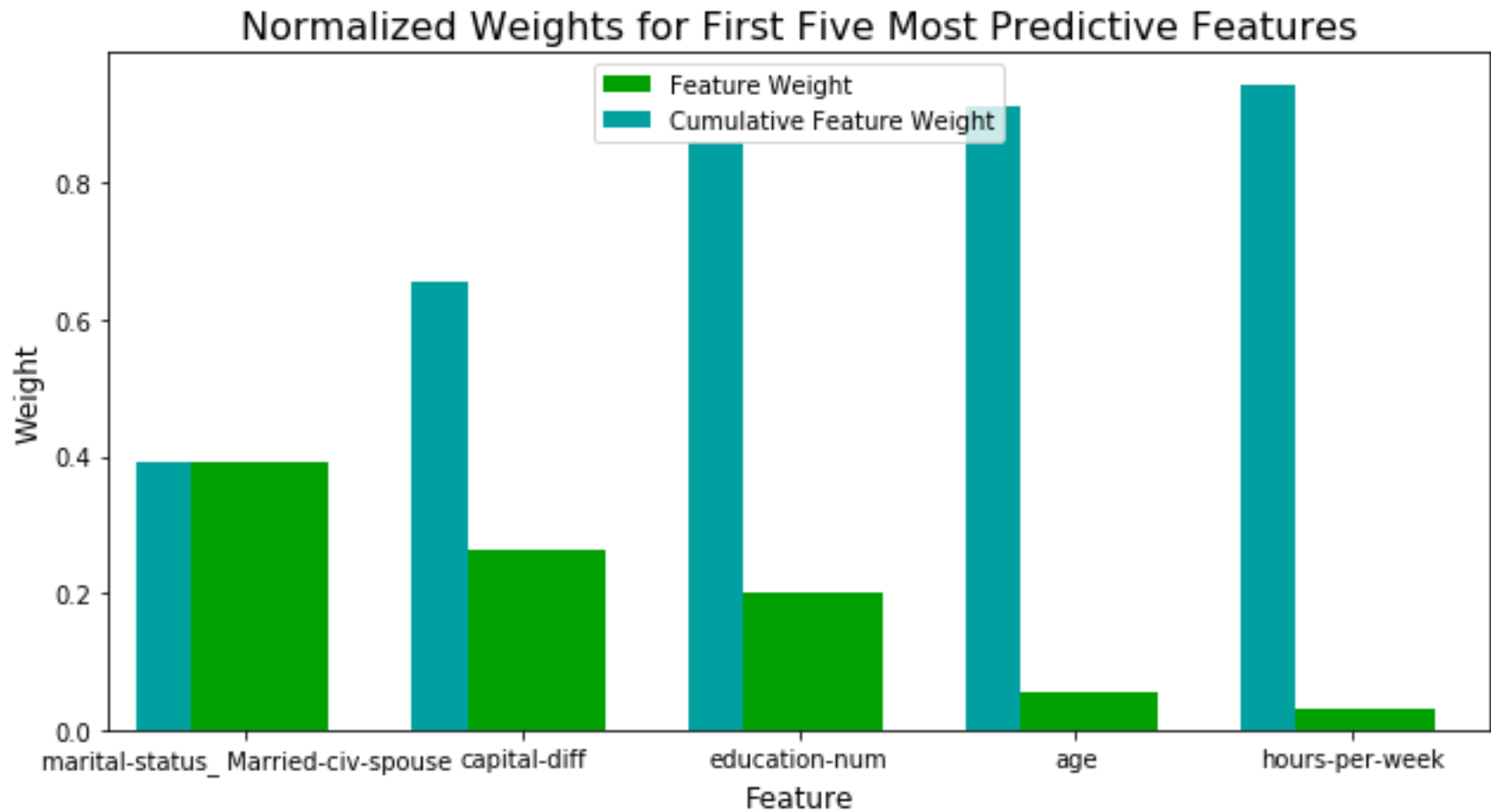
Model Evolution

Accuracy	0.8714
Precision	0.7853
Recall	0.6503

Normalized Weights for First Five Most Predictive Features

Initial → Education → Capital → Native → Optimize
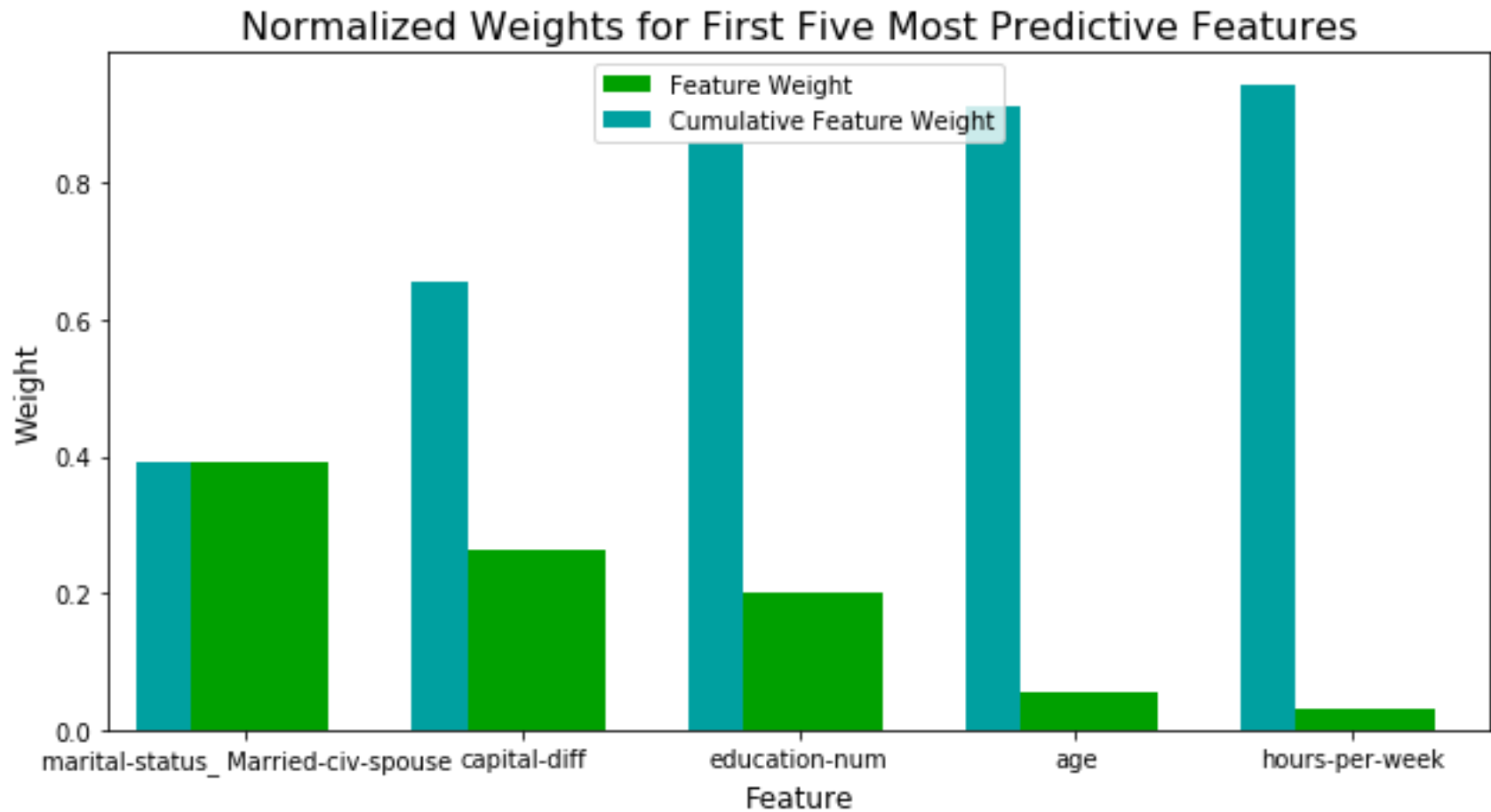
Summary

- Gradient Boosting is selected over Gaussian Naïve Bayes and Decision Tree based on running time, accuracy, and F score.

- Model is improved/simplified progressively by feature engineering:
  - Discarding education level
  - Combining capital gain & capital loss
  - Encoding of native country