PREDICTING INCOME LEVELS FROM U.S. CENSUS DATA

Michelle A. Caler

April 21, 2021
(updated May 13, 2021)

OVERVIEW

- Introduction
- Associations with Income
- Models
- Results
- Conclusions

INTRODUCTION

MOTIVATION AND CENTRAL QUESTION

- Motivation: carry out first fully-independent data science project since earning Data Science Career Path certificate from Codecademy

- Central Question: Can I determine whether a person makes over $50K a year?

THE DATA: WHERE AND WHY

- Income and demographics data compiled from 1994 Census Bureau database
- Downloaded from this Kaggle repository
  - Repository owner: UCI Machine Learning group
- Public Domain data; compiled by UCI Machine Learning group so a clear answer to the question likely to exist
- Modest-sized data set with a mix of quantitative and qualitative features

THE DATA: WHAT AND HOW

- Data extracted from 1994 Census Bureau database
  - Paper cited by UCI Machine Learning Group: Ron Kohavi, "Scaling Up the Accuracy of Naive-Bayes Classifiers: a Decision-Tree Hybrid", Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, 1996
- 32,561 records, 15 features
  - No null values!

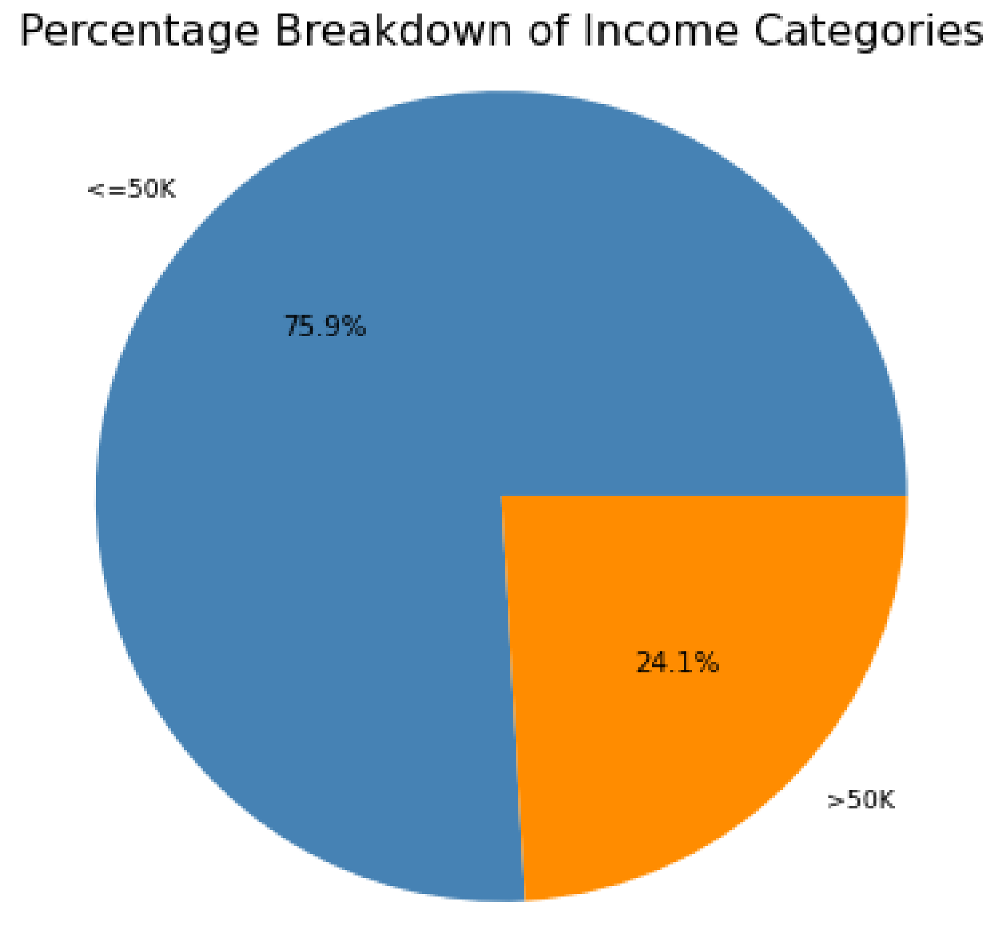
THE DATA: FEATURES

Quantitative	Qualitative
age	workclass
fnlwgt	education
education.num	marital.status
capital.gain	occupation
capital.loss	relationship
hours.per.week	race
	sex*
	native.country
	income*

*: binary categorical variable

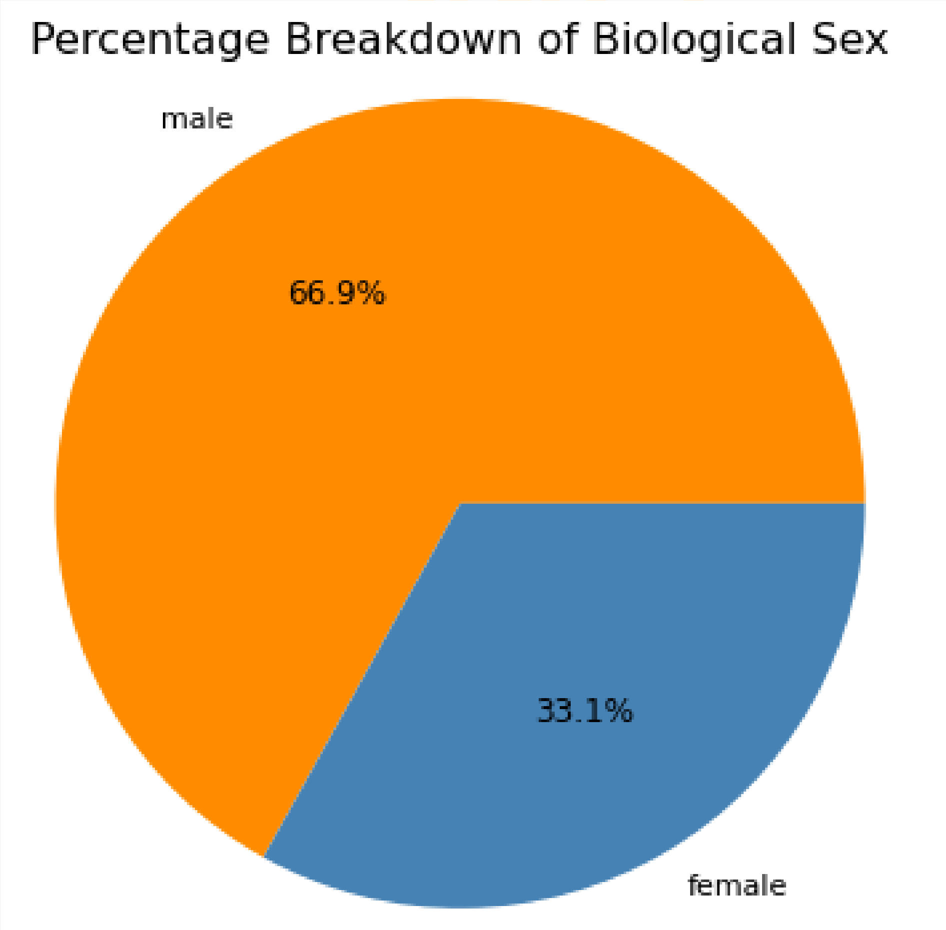
ASSOCIATIONS WITH INCOME

INCOME CATEGORIES



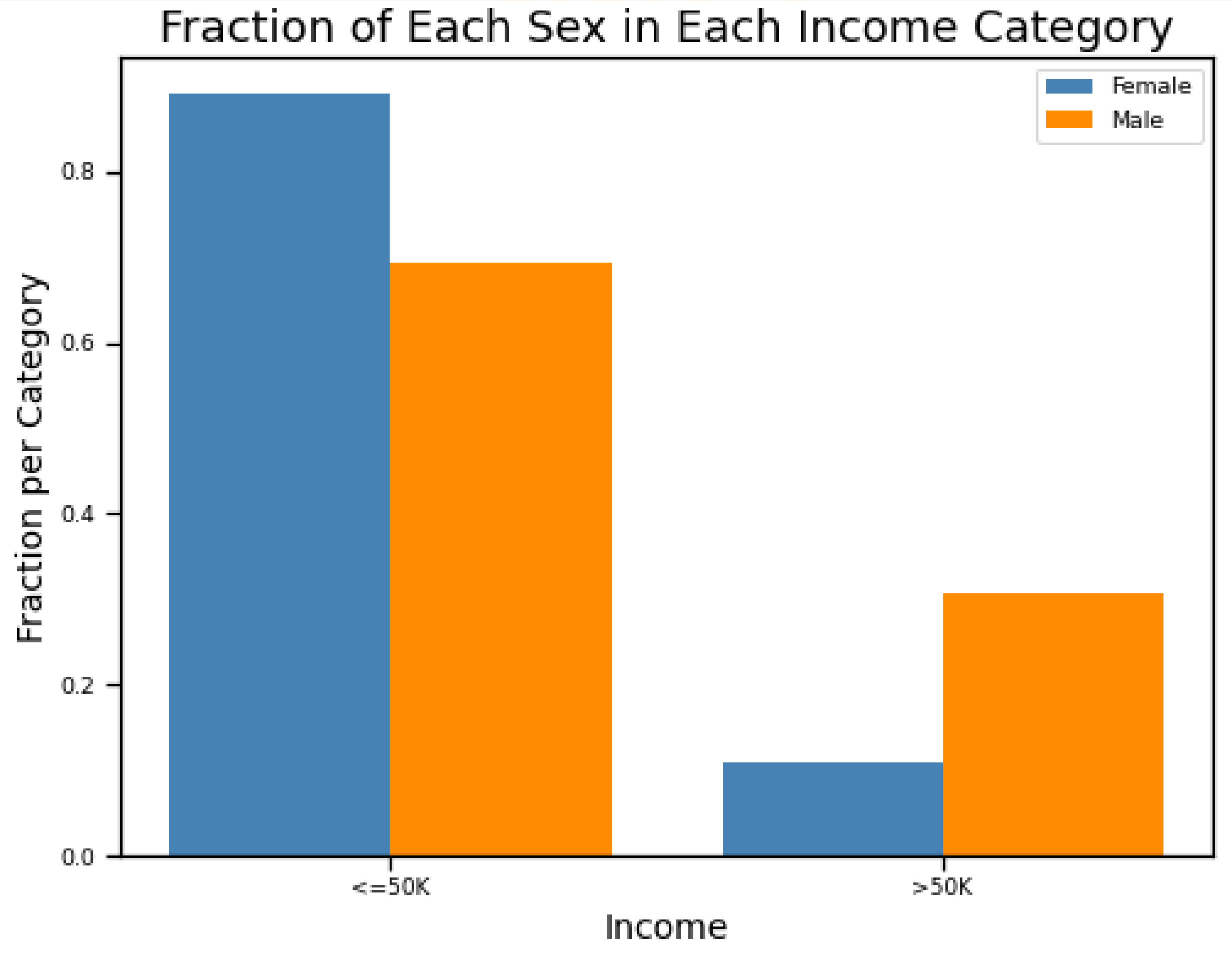
- Low-income: <= 50K
  - Roughly ¾ of records
- High-income: >50K
  - Roughly ¼ of records

SEX AND INCOME



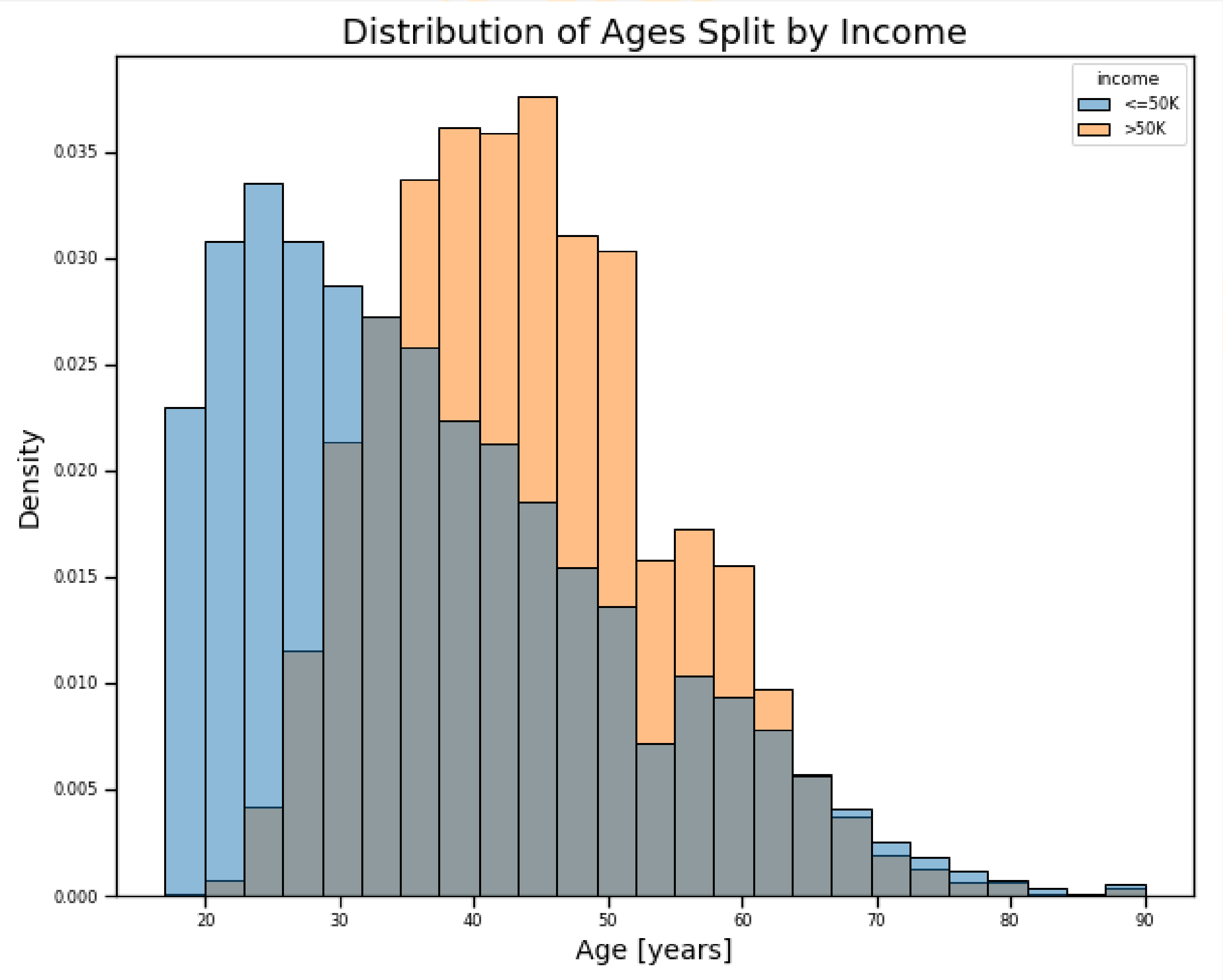
- Sex = Biological Sex
- Roughly 2/3 of records are males
- Roughly 1/3 of records are females

SEX AND INCOME



- Females:
  - Roughly 85% in low-income category
  - Roughly 15% in high-income category
- Males:
  - Roughly 65% in low-income category
  - Roughly 35% in high-income category

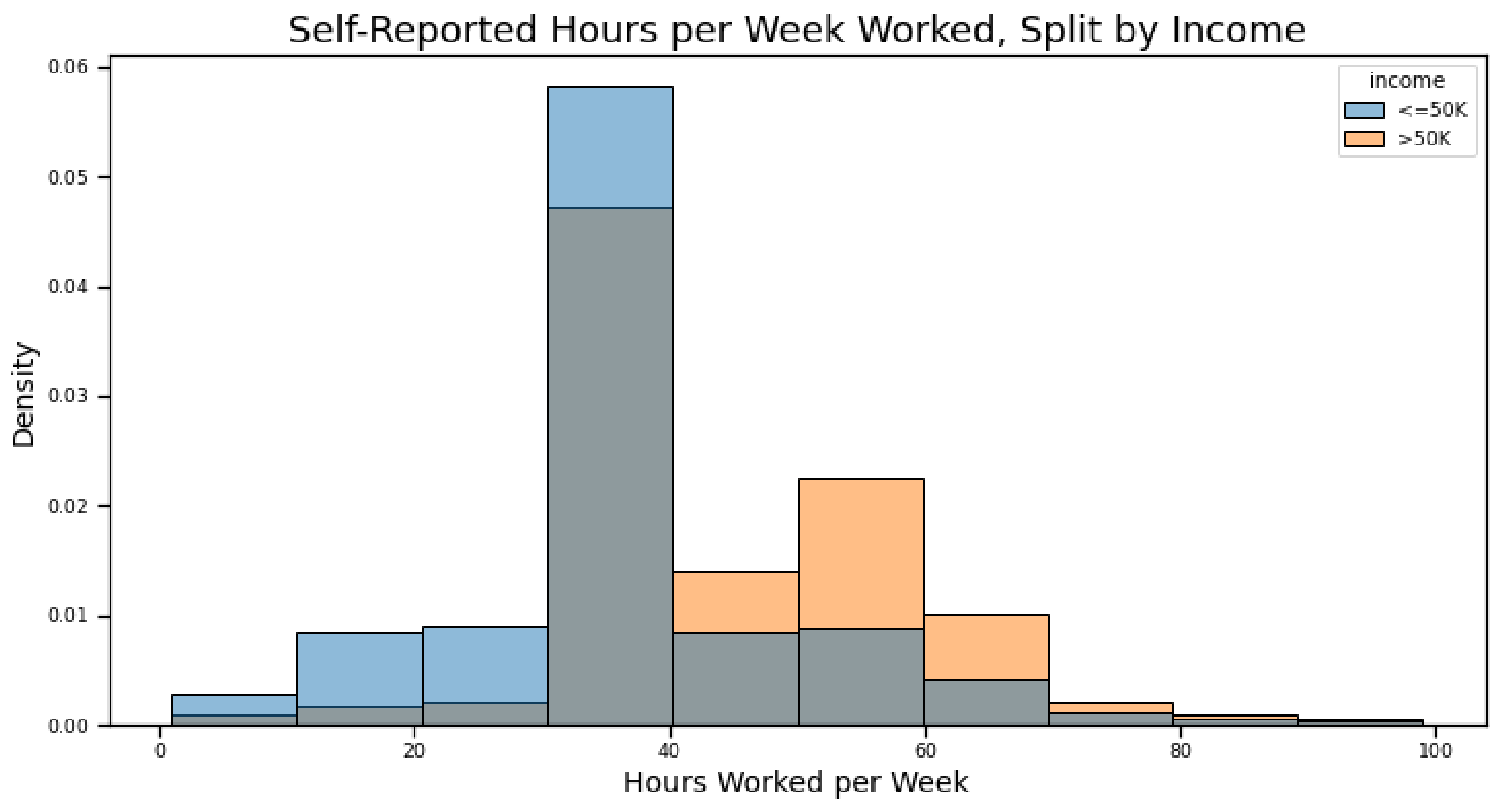
AGE AND INCOME



- Low-income and high-income distributions have different shapes
  - Low-income: gradual decay from peak
  - High-income: skewed bell curve

- Low-income and high-income distributions peak at different ages
  - Low-income: peak around 25 yrs. old
  - High-income: peak around 45 yrs. old

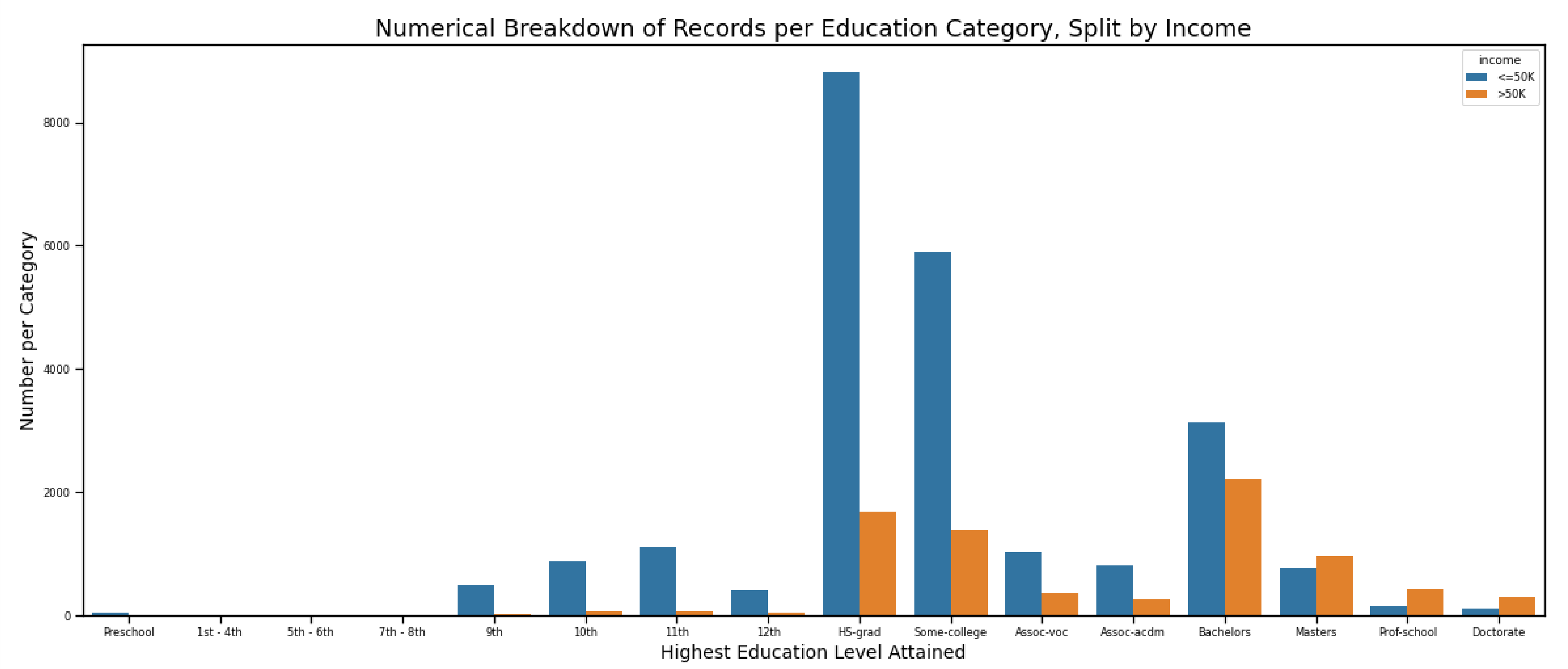
HOURS WORKED PER WEEK AND INCOME



- Low-income: peaks at 35-40 hours per week
  - Plateau from 15-35 hours per week
  - Plateau from 40-60 hours pre week

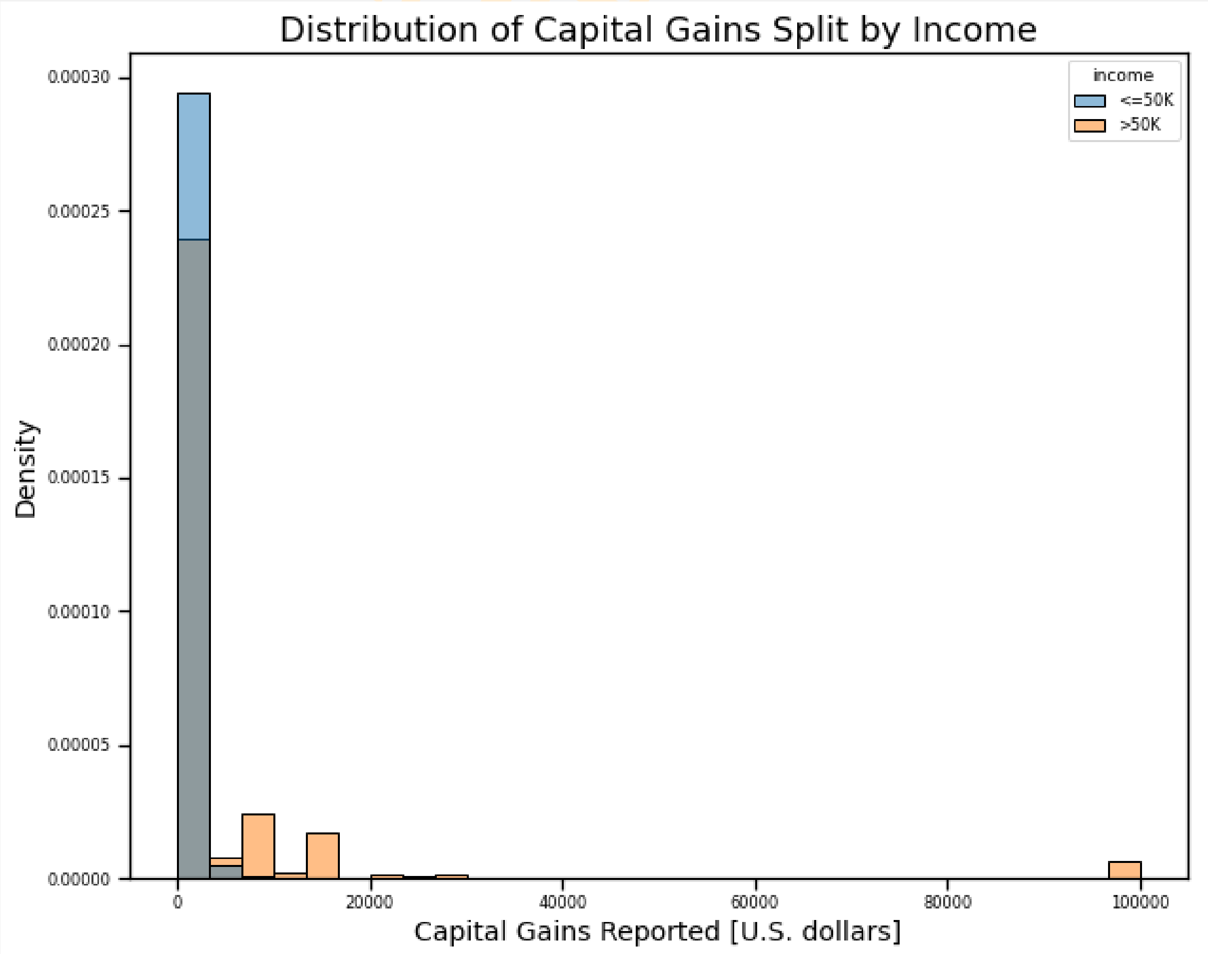
- High-income: peaks at 35-40 hours per week and 55-60 hours per week
  - Peak at 35-40 hours per week much sharper
  - Skew left

EDUCATION AND INCOME



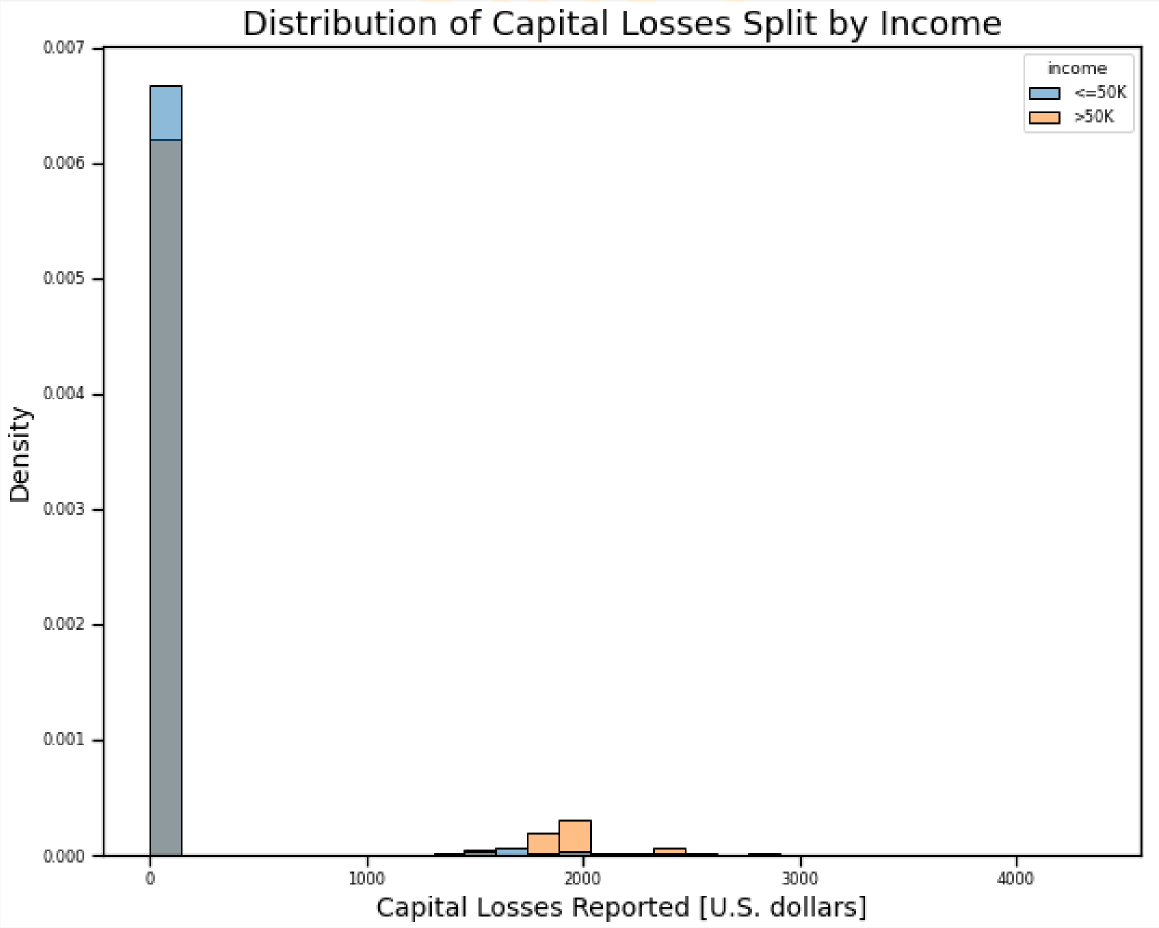
- General trend: the higher the education level attained, the higher the fraction of high-income records

CAPITAL GAINS AND INCOME



- Low-income: Little to no capital gains
- High-income: Mostly little to no capital gains, but significant minority with gains > 10K

CAPITAL LOSSES AND INCOME



- Low-income: Little to no capital losses
- High-income: Mostly little to no capital gains, but significant minority with losses > 15K

SUPERVISED MACHINE LEARNING MODELS

MODEL INPUTS

- Features used to train models:
  - Sex, converted to integers (1: Female, 0: Male)
  - Education, converted to integers (0: Preschool through 16: Doctorate)
  - Age
  - Hours Worked per Week
  - Capital gains
  - Capital losses
- Prediction labels:
  - 0: low-income record
  - 1: high-income record

MODEL INPUTS

- Training and test data scaled such that all features lay on same scale

- Models without hyperparameters to tune:
  - 75% of data used to train models
  - 25% of data used to test models

- Models with hyperparameters to tune:
  - 75% of data used to train models
  - 12.5% of data used for model validation
  - 12.5% of data used to test models

MODELS TRAINED TO FEATURES
(AND USED TO PREDICT INCOME LEVEL)

- Logistic Regression Classifier
- Support Vector Machine Classifier
- K-Nearest Neighbours Classifier
- Single Decision Tree Classifier
- Random Forest Classifier

RESULTS

MODEL PRECISION, ACCURACY, RECALL

Model	Precision	Accuracy	Recall
Logistic Regression	73.1%	82.2%	41.2%
Support Vector Machine	77.2%	83.9%	45.7%
K-Nearest Neighbours	71.7%	82.7%	44.8%
Single Decision Tree	66.8%	82.0%	50.7%
Random Forest	68.5%	82.6%	51.3%

Highest precision, highest accuracy

Best balances precision and recall

CONCLUSIONS

CONCLUSIONS

- I was able to build a supervised machine learning model to predict income level with good precision and accuracy

- Best model: Support Vector Machine classifier
  - 77.2% precision
  - 83.9% accuracy
  - 45.7% recall

- Best Balanced model: Random Forest classifier
  - 68.5% precision
  - 82.6% accuracy
  - 51.3% recall

IMPROVEMENTS

- Use a machine learning classification model which accepts qualitative as well as quantitative data
  - Nominal categorical variables with clear associations with income unused

- Explore γ/C parameter space of Support Vector Machine classifier
  - Is current model a local max or a global max?

CITATIONS

- Ron Kohavi, "Scaling Up the Accuracy of Naive-Bayes Classifiers: a Decision-Tree Hybrid", Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, 1996.

- The scikit-learn Python machine learning library: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011

- After completing my project, I came across this stellar write-up from (then) UCSD students Chet Lemon, Chris Zelazo, and Kesav Mulakaluri. Their treatment of the problem no doubt influenced my own write-up of my work.