Prediction of Adult Income

Sara Maxwell

Problem

Various factors determine an individual adult’s income.

- Education level
- Age
- Gender
- Field of work

About the Data

- Comes from the UC Irvine repository
- Collected in 1996: https://www.kaggle.com/wenruliu/adult-income-dataset
- 14 variables about attributes from 48,842 individuals
- Information: http://www.cs.toronto.edu/~delve/data/adult/adultDetail.html

Overview

The steps involved in this analysis include:

- Data cleaning and wrangling
- Feature engineering
- Preprocessing: scaling, one-hot encoding
- Exploratory data analysis
- Machine learning

Steps in the Analysis

- Data Cleaning
  - Remove rows with missing values (shown as “?”)
  - New binary column for income (greater than $50K or not)
  - Remove columns not necessary for analysis
- Preprocessing
  - Scaling
  - One-hot encoding

Steps in the Analysis (cont.)

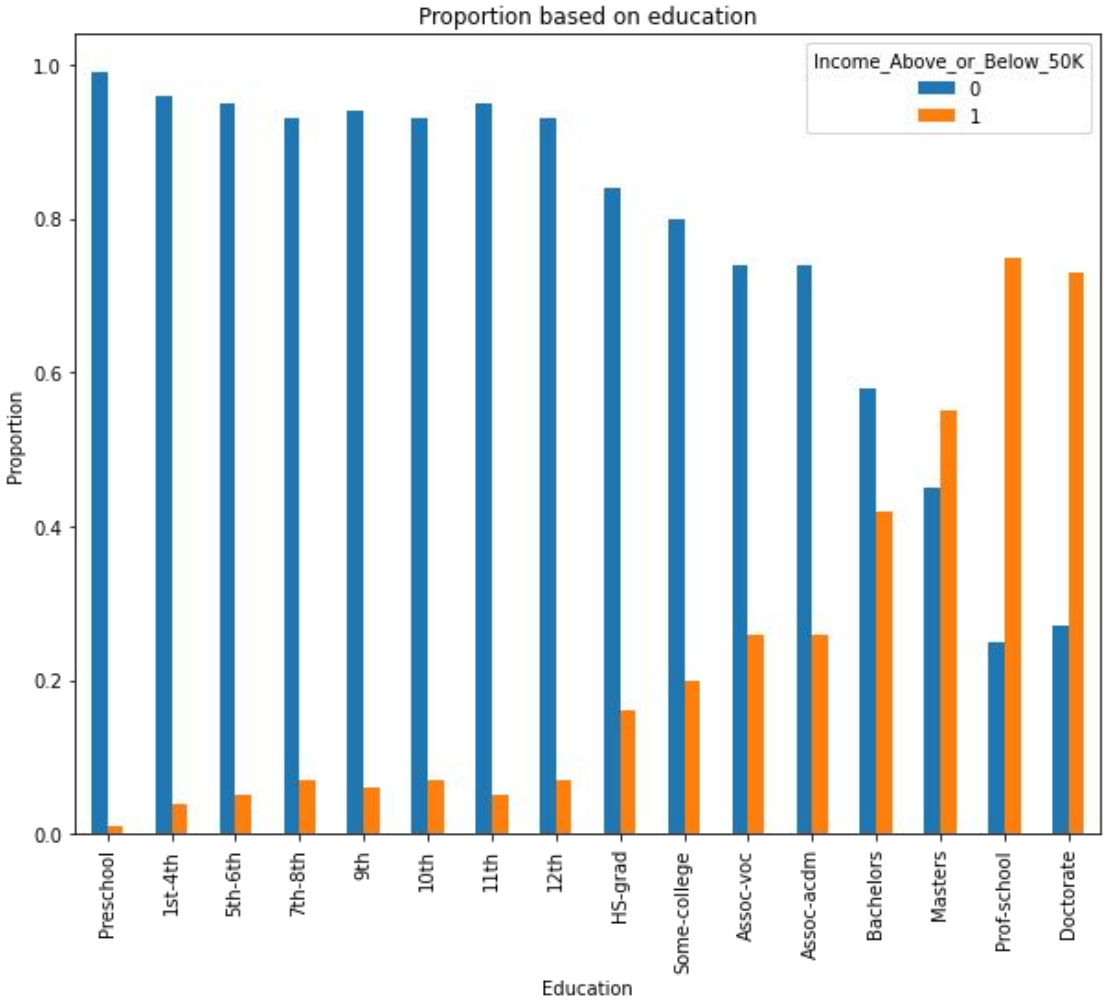
- Exploratory Data Analysis
  - Calculate proportions of incomes over $50K
  - Education
  - WorkClass
  - Occupation
  - Marital Status
  - Relationship
  - Race
  - Gender

Steps in the Analysis (cont.)

- Machine Learning
  - Use test-train splits
  - Logistic Regression
  - Decision Tree
  - Random Forest

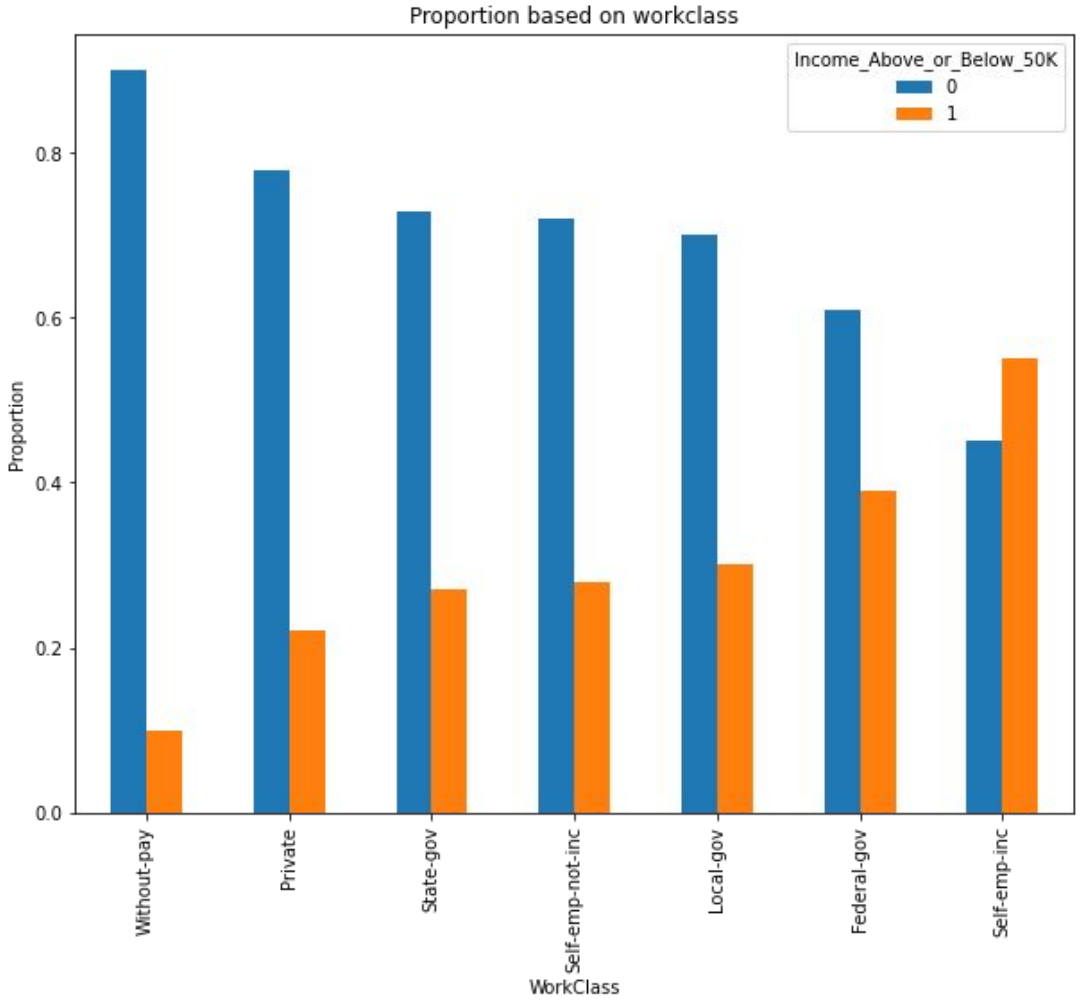
Exploratory Data Analysis

As expected, the higher the education level, the greater the proportion of adults with incomes over 50K. From the master's level and up, adults making more than 50K outnumber adults making 50K or less.



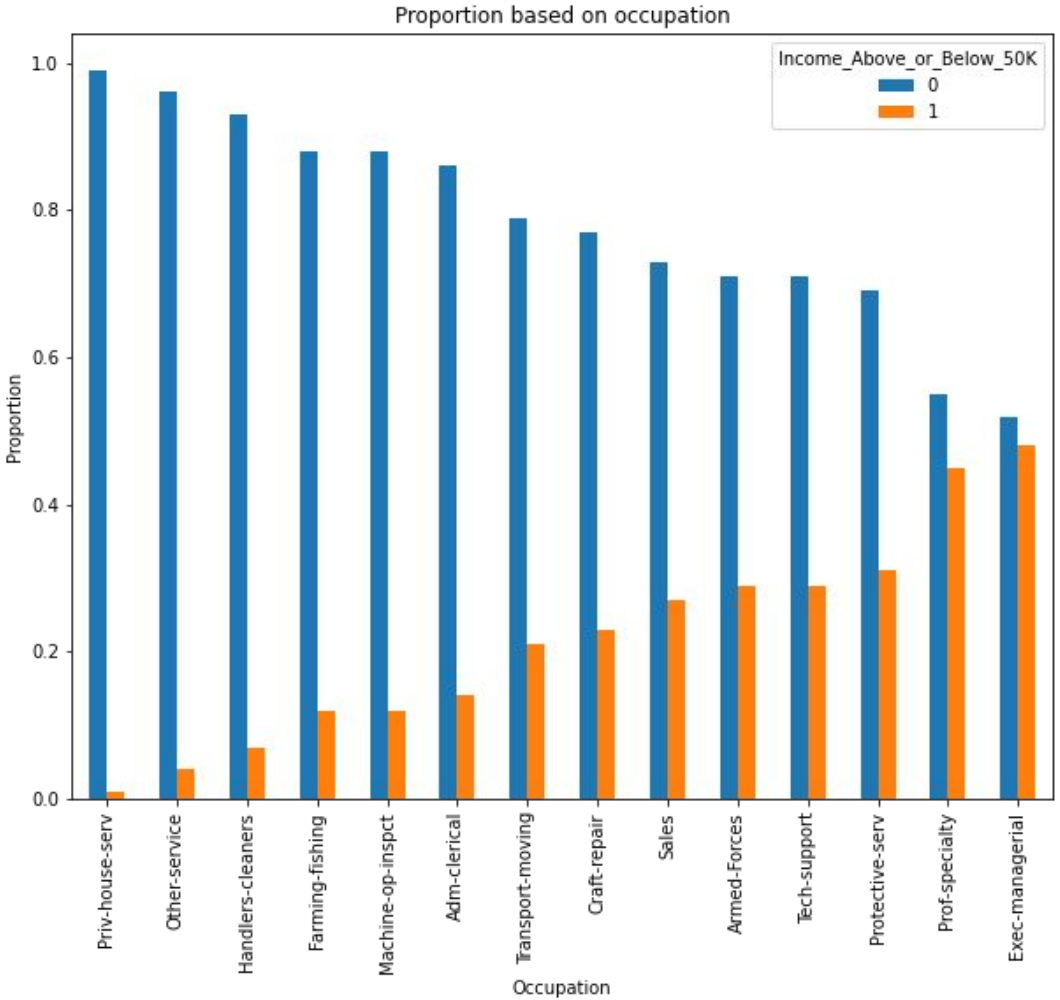
Exploratory Data Analysis (cont.)

Among the adults with pay, those in the private class have the lowest proportion of incomes over 50K, while those self-employed in incorporated businesses have not only the greatest proportion of incomes over 50K, but adults making more than 50K outnumber those that are not.



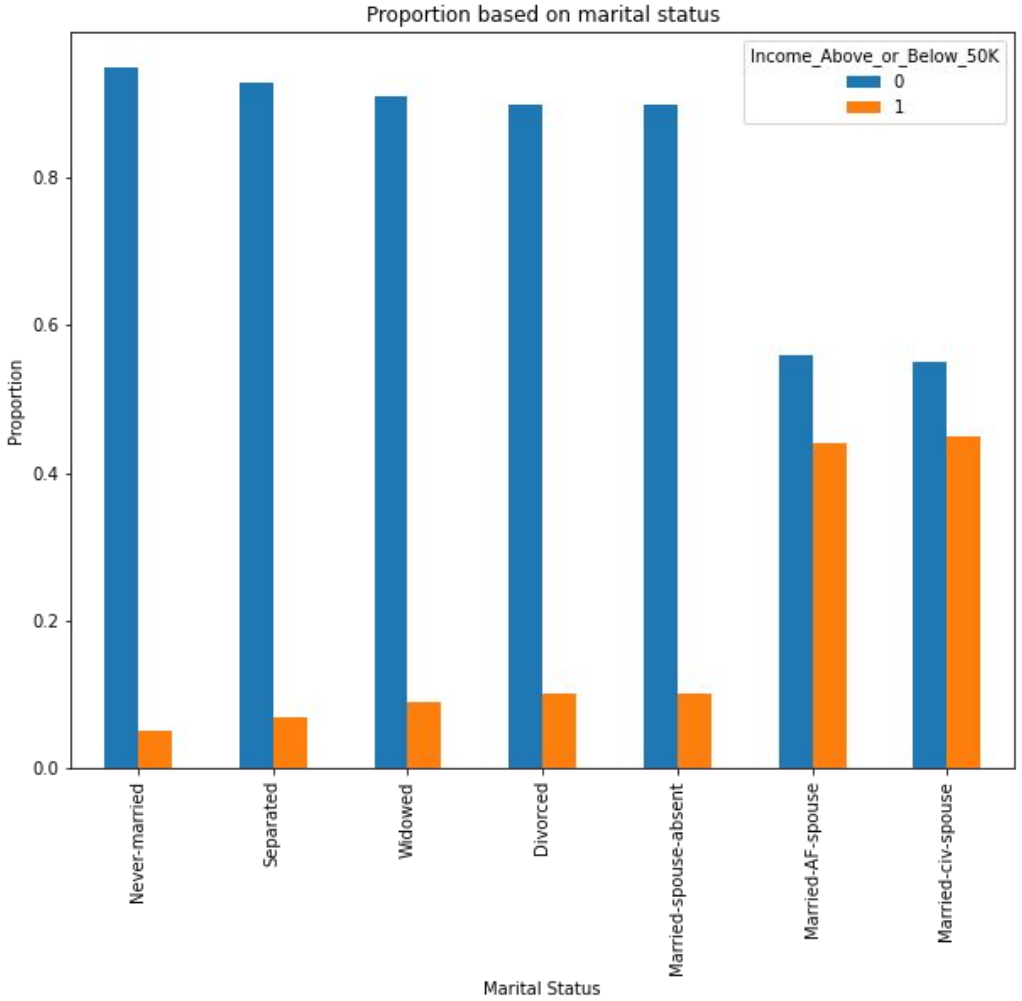
Exploratory Data Analysis (cont.)

All occupations have more adults with incomes $50K or less than adults with greater incomes. The difference is least with executive managerial positions, followed closely by specialized professionals such as doctors and lawyers.



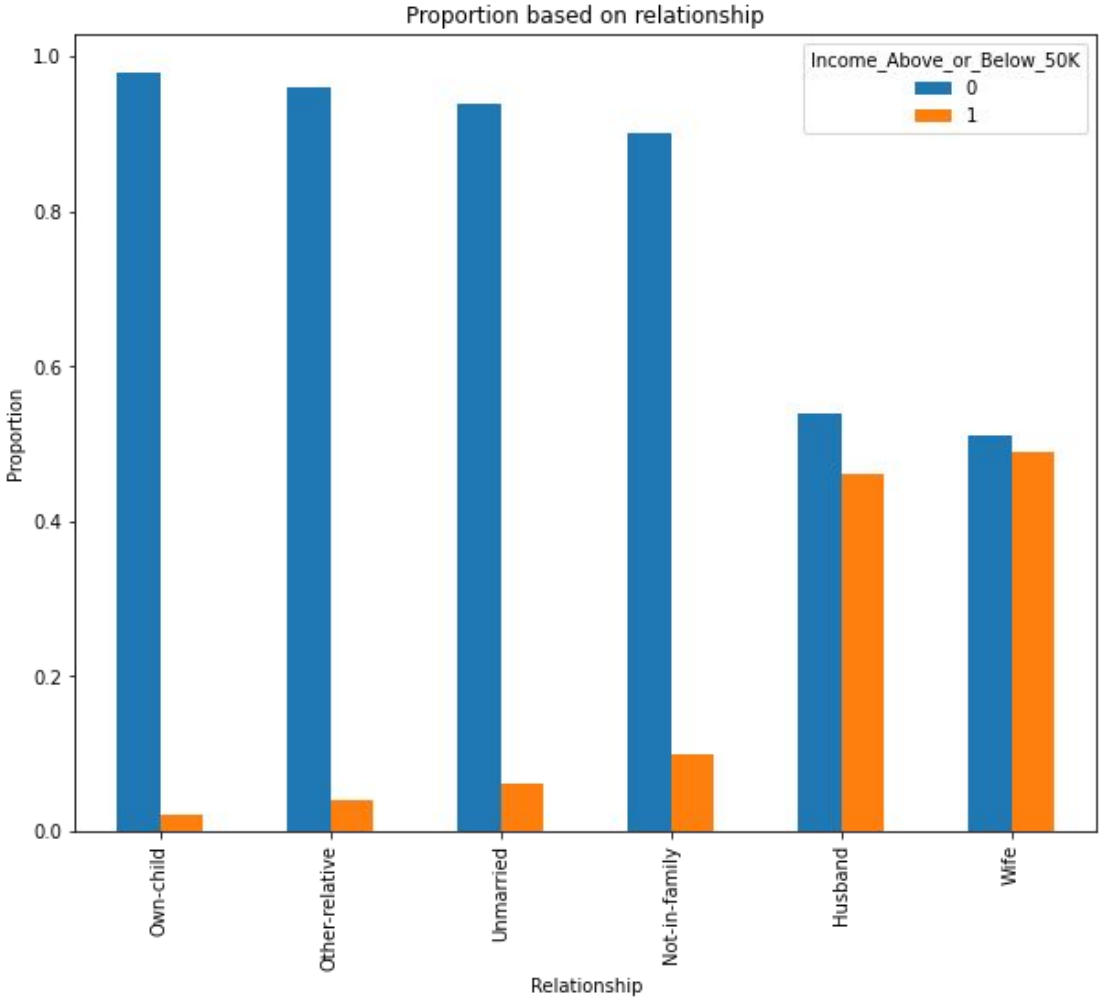
Exploratory Data Analysis (cont.)

All categories have more adults with incomes of 50K or less than adults with incomes over 50K, but the gap is much smaller with married couples that have a spouse present.



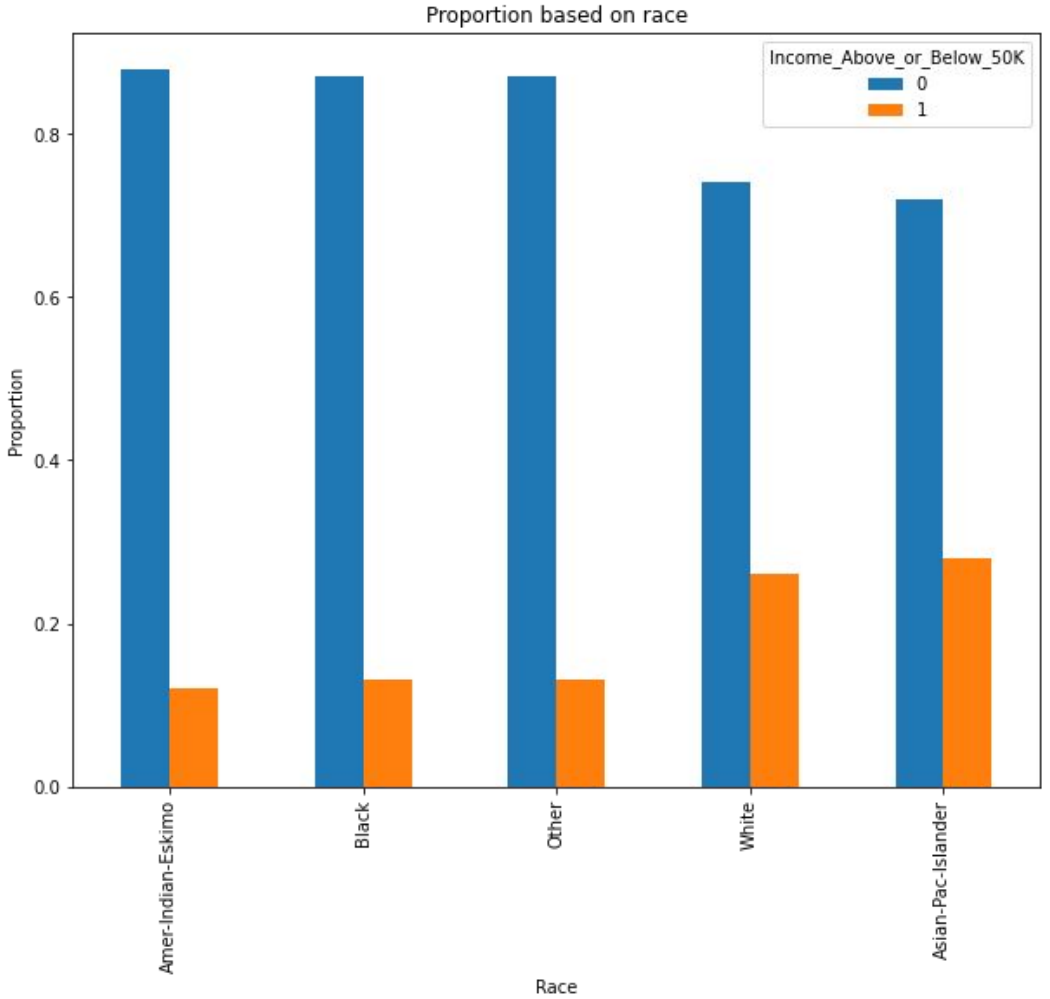
Exploratory Data Analysis (cont.)

All categories have more adults with incomes of 50K or less than adults with incomes over 50K, but the gap is much smaller with husbands and wives. This makes sense because of similarly small gaps in the categories of married couples with the spouse present.



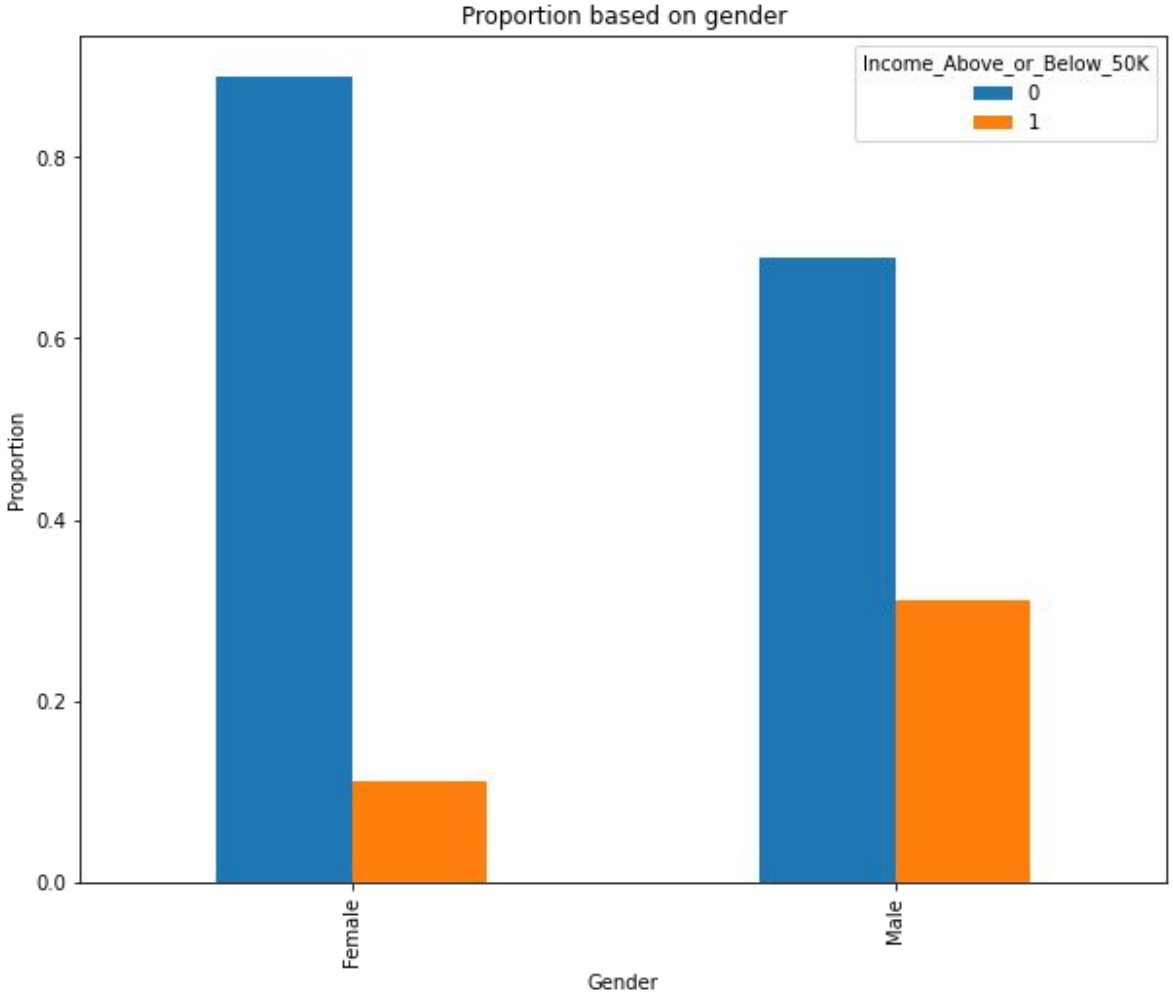
Exploratory Data Analysis (cont.)

Whites and Asians have higher proportions of incomes over $50K.



Exploratory Data Analysis (cont.)

Men have a higher proportion of incomes over $50K.



Machine Learning

Comparing different regression models:

- Logistic Regression
  - Accuracy: 0.8398
- Decision Tree (entropy)
  - Accuracy: 0.7891
  - Balanced accuracy: 0.7057
  - Precision: 0.5813
  - Recall: 0.5399
  - F-measure: 0.5598

Machine Learning (cont.)

- Decision Tree (gini)
  - Accuracy: 0.7869
  - Balanced accuracy: 0.7045
  - Precision: 0.5757
  - Recall: 0.5406
  - F-measure: 0.5576

Machine Learning (cont.)

- Random Forest (entropy)
  - Accuracy: 0.8162
  - Balanced accuracy: 0.7314
  - Precision: 0.6501
  - Recall: 0.563
  - F-measure: 0.6034

Machine Learning (cont.)

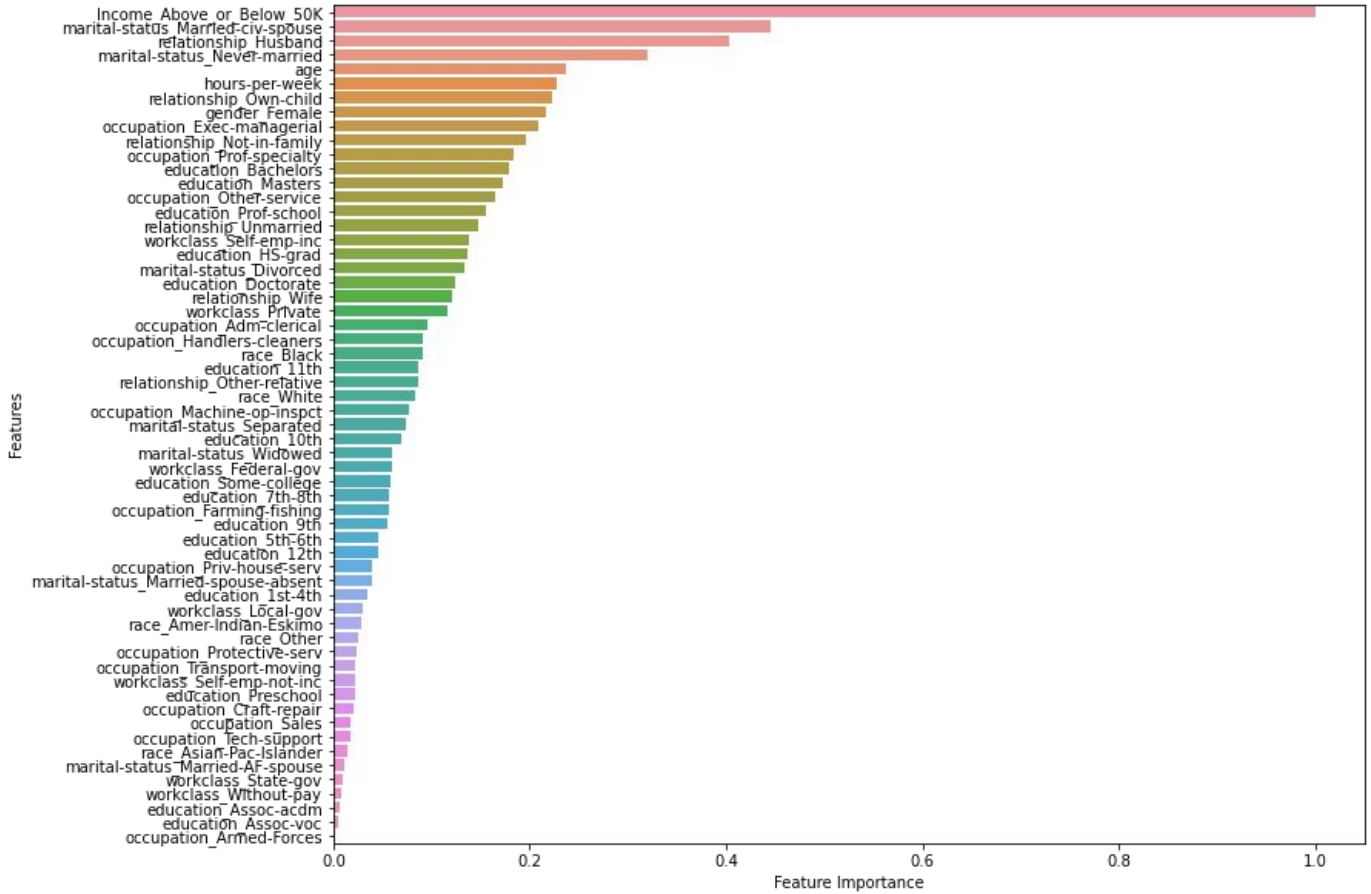
- Random Forest (gini)
  - Accuracy: 0.8143
  - Balanced accuracy: 0.7277
  - Precision: 0.647
  - Recall: 0.5556
  - F-measure: 0.5978

Conclusions

- Random Forest Entropy has the highest precision, recall, and F-score
- Logistic Regression model has highest accuracy
- Most important factors:
  - Married in a civil procedure
  - Being a husband
  - Never-married

Conclusions (cont.)

The highest importance value, civil marriage, is 0.446, so the importance of each individual factor is not that strong.



Future Recommendations

- The dataset is from 1996, so more recent data may help
- Data with actual income values, not just a comparison to one value