Predicting if income exceeds $50,000 per year based on 1994 US Census Data with Simple Classification Techniques

Chet Lemon (A10895241)
Chris Zelazo (A10863450)
Kesav Mulakaluri (A10616114)

Abstract -- For this assignment, we examine the Census Income dataset available at the UC Irvine Machine Learning Repository. We aim to predict whether an individual's income will be greater than $50,000 per year based on several attributes from the census data.

Introduction

The US Adult Census dataset is a repository of 48,842 entries extracted from the 1994 US Census database.

In our first section, we explore the data at face value in order to understand the trends and representations of certain demographics in the corpus. We then use this information in section two to form models to predict whether an individual made more or less than $50,000 in 1994. In the third section, we look into a couple papers written on the dataset to find out what methods they are using to gain insight on the same data. Finally, in the fourth section, we compare our models as well as that of others in order to find out what features are of significance, what methods are most effective, and gain an understanding of some of the intuition behind the numbers.

Exploratory Analysis

The Dataset

The Census Income dataset has 48,842 entries. Each entry contains the following information about an individual:

- **age**: the age of an individual
    - Integer greater than 0
- **workclass**: a general term to represent the employment status of an individual
    - Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
- **fnlwgt**: final weight. In other words, this is the number of people the census believes the entry represents..
    - Integer greater than 0
- **education**: the highest level of education achieved by an individual.
    - Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
- **education-num**: the highest level of education achieved in numerical form.
    - Integer greater than 0
- **marital-status**: marital status of an individual. Married-civ-spouse corresponds to a civilian spouse while Married-AF-spouse is a spouse in the Armed Forces.

    - Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
- **occupation**: the general type of occupation of an individual
    - Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.
- **relationship**: represents what this individual is relative to others. For example an individual could be a Husband. Each entry only has one relationship attribute and is somewhat redundant with marital status. We might not make use of this attribute at all
    - Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
- **race**: Descriptions of an individual's race
    - White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
- **sex**: the biological sex of the individual
    - Male, Female
- **capital-gain**: capital gains for an individual
    - Integer greater than or equal to 0
- **capital-loss**: capital loss for an individual
    - Integer greater than or equal to 0
- **hours-per-week**: the hours an individual has reported to work per week
    - continuous.
- **native-country**: country of origin for an individual
    - United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.
- **the label**: whether or not an individual makes more than $50,000 annually.
    - <=50k, >50k