Predicting Income: A Supervised Learning Project

Taaru Chawla

Introduction and Background of the Dataset

The data was extracted from the 1994 Census Database. It was obtained from the UC Irvine machine learning repository.

It is a relatively small dataset which initially had 32560 rows (prior to cleaning).

Each row represents certain about an individual, such as age, education, family status, occupation, gender, race, country of origin, and the money that the individual makes.

Introduction and Background of the Dataset

This is a supervised learning project, with the goal of predicting an individual's Income (target) and which of the two income categories it falls in.

While income is a continuous variable, the dataset has income classified in binary categories:

- Less than or equal to $50,000
- Greater than $50,000

As the Income column is split in categories, this is a classification problem.

The remaining columns will be features that will be used to make the prediction.

While some of the columns are self-explanatory, the others will be explained in detail in the next few slides.

Exploring the 12 Features

Age

Sex

Capital Gain

Capital Loss

Hours per Week

Native Country

Exploring the 12 Features

Work Class: Private, Self-Employed (not-inc), Self-Employed (inc). Federal Government, Local Government, State Government, Without Pay, Never Worked.

Education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.

Number of Years of Education

Exploring the 12 Features

Marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.

Occupation: 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: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.

Race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.

Data Cleaning / Addressing Null Values

There is a column, ‘fnlwt’, which was used to assign weights to various columns. This column has been dropped.

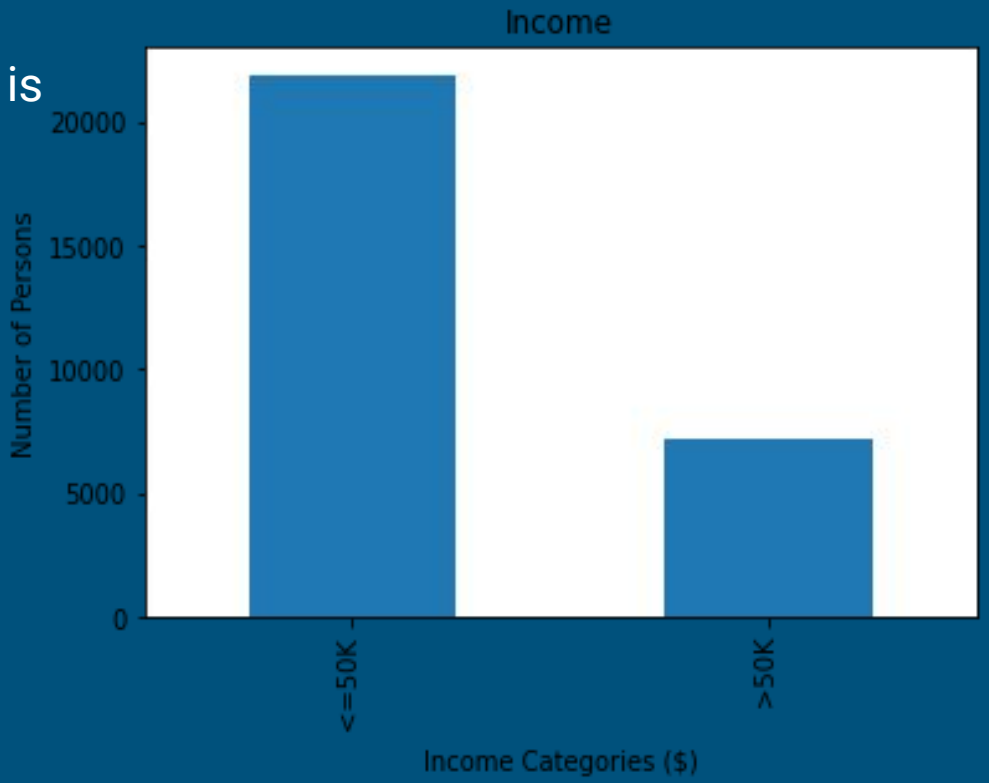
There are many duplicates, and once they are dropped, the rows are reduced to 29.096.

There are several ‘?’ in certain categories (Work Class, Native Country and Occupation). In total, there are 1632 rows with a ‘?’. These are too many rows to drop, of this relatively small dataset. These columns are also important, so I will not drop them either. Instead, all ‘?’ will be put in a category called “unknown”.

There are no outliers in the descriptive statistics.

Data Exploration

The individuals with income <=$50,000 is almost three times as many as the people with income >$50,000.



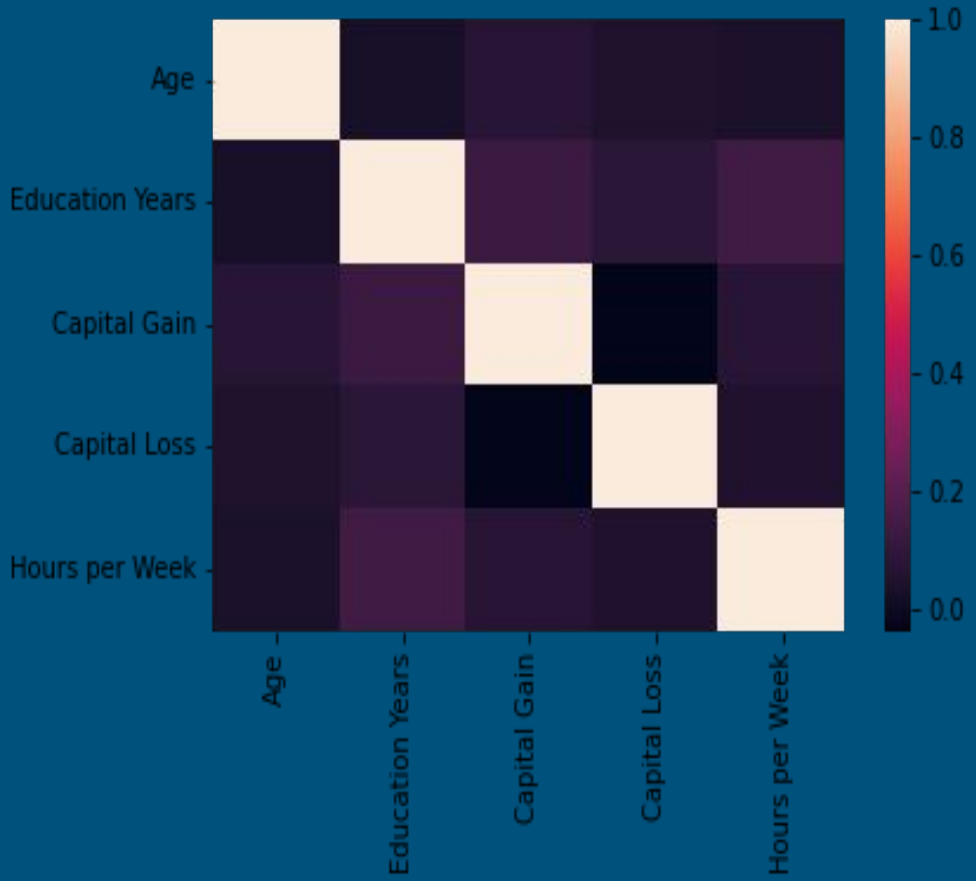
Data Exploration

There is low correlation between:

Education Years and Capital Gain &

Education and Hours per Week.

No other correlations found in the continuous variables.



Data Exploration: Some additional notes

Majority of the individuals in the dataset work in the private industry.

Many have HS, or some college, or a Bachelor's degree.

Many are married to a civilian spouse, followed by a large number of unmarried individuals.

This population is largely white males.

Majority of the individuals are from the US.

The median age is a little below 40.

The median education is 10 years.