Which factors are the most crucial in determining someone's income? Part-1

by
Syed Reyadh

Abstract

- The goal of this research is to determine which personal traits of individuals account for salary disparities more effectively than others.
- The "UCI Machine Learning Repository" provided the dataset for this study, which is based on the 1994 census. It is made up of factors including age, salary, sex, race, occupation, education, and country of birth, among others.
- Using the Scikit-learn toolkit, a decision tree classification model has been developed to investigate this relationship.
- The data appears to indicate the existence of a relationship between an individual's pay and their qualities, and the developed model appears to accurately predict over 80% of the inputted data. Relationship, Education, Occupation, and Age are the factors having the greatest capacity for information. Other factors like nationality, sex, or race are not important.

Motivation

- The factors, whether personal or not, that can explain people's future income have piqued my curiosity in the previous few months. Naturally, many of them are outside of our control. For instance, age or sex may contribute to wage disparities between individuals, but they are outside of our control. Conversely, others are at least somewhat within our power: Say, for instance, about education.

- Determining the contribution of these variables to an individual's future income is crucial because, at worst, it allows us to extrapolate pertinent policy implications and more effectively influence the behavior of the youngest members of society by emphasizing the value of education for their future.

Dataset(s)

- The "Adult" dataset utilized in this data science study is taken from the renowned "UCI Machine Learning Repository" for data science. It was developed in the United States in 1996 using data from the 1994 Census.

- It’s a collection of individual characteristics of people, including age, education, occupation, race, salary. The number of instances at the beginning of the analysis was 32560. The number of categorical variables is 15: among them, salary is our Target Variable.

- The dataset's fundamental premise is that a relationship can be found between Salary and every other variable.

Data Preparation and Cleaning

As previously stated, I utilized the "Adult" dataset from the "UCI Machine Learning Repository" for my investigation. The dataset consisted of 32560 instances and 15 attributes, including the target attribute, at the start of the analysis.

In the figure 1 is shown the list of the attributes in the initial dataset.

Figure-1

Variable Name	Kind	Description
age	continuous	17<age<90
workclass	categorical	8 different unique instances
fnlwgt	continuous	
education	categorical	16 different unique instances
education-num	continuous	16 different numbers
marital-status	categorical	7 different unique instances
occupation	categorical	14 different unique instances
relationship	categorical	6 different unique instances
race	categorical	5 different unique instances
sex	categorical	2 different unique instances
capital-gain	continuous	
capital-loss	continuous	
hours-per-week	continuous	1<hours per week<99
native-country	categorical	40 different unique instances
salary	**categorical**	**2 different unique instances** **>50K $** **<50K $**

Target Variable

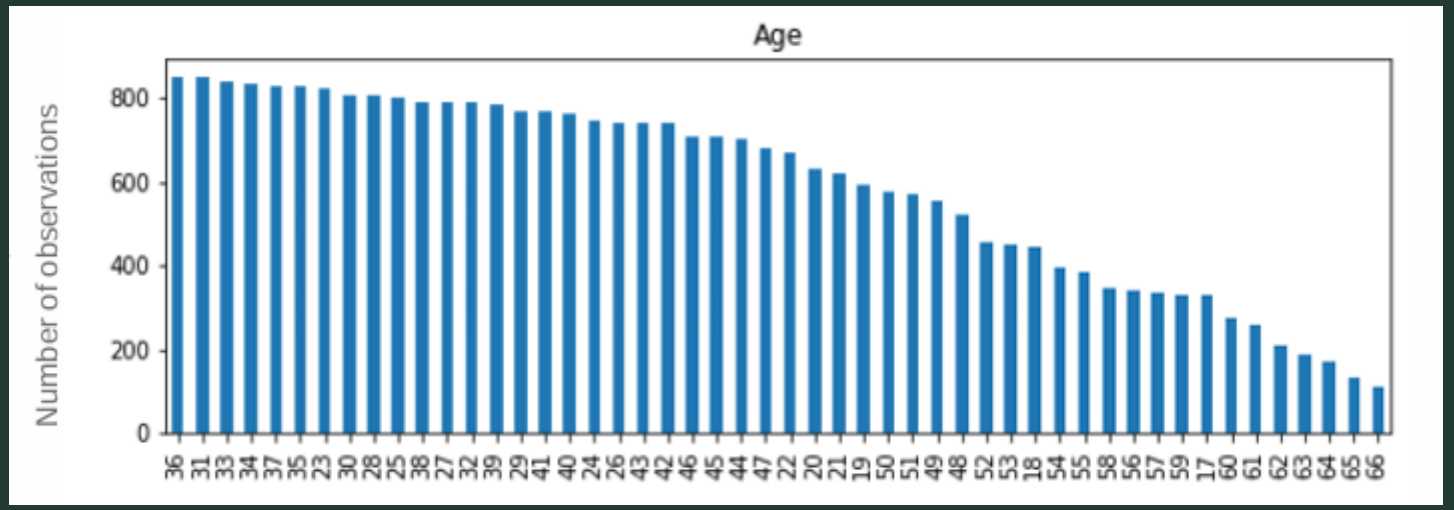
Main actions taken to clean the datasets:

1. Drop not relevant columns: the columns “fnlwgt”, education-num, capital-gain and capital-loss have been removed from the dataset. The variable “fnlwgt” is considered not relevant for the analysis purposes. In capital-gain and capital-loss, instead, too many missing values made impossible their use for further analysis.
2. All rows in which were at least 1 NaN values have been removed from the dataset. This implies a reduction of 8% of the number of instances (from 32560 to 30117).
3. Attribute manipulation: for not removed attributes a further cleaning activity has been performed: continuous variable have been transformed into categorical ones and the number of distinct observations of categorical variables has been reduced. In the next slides some example of this activity will be provided.

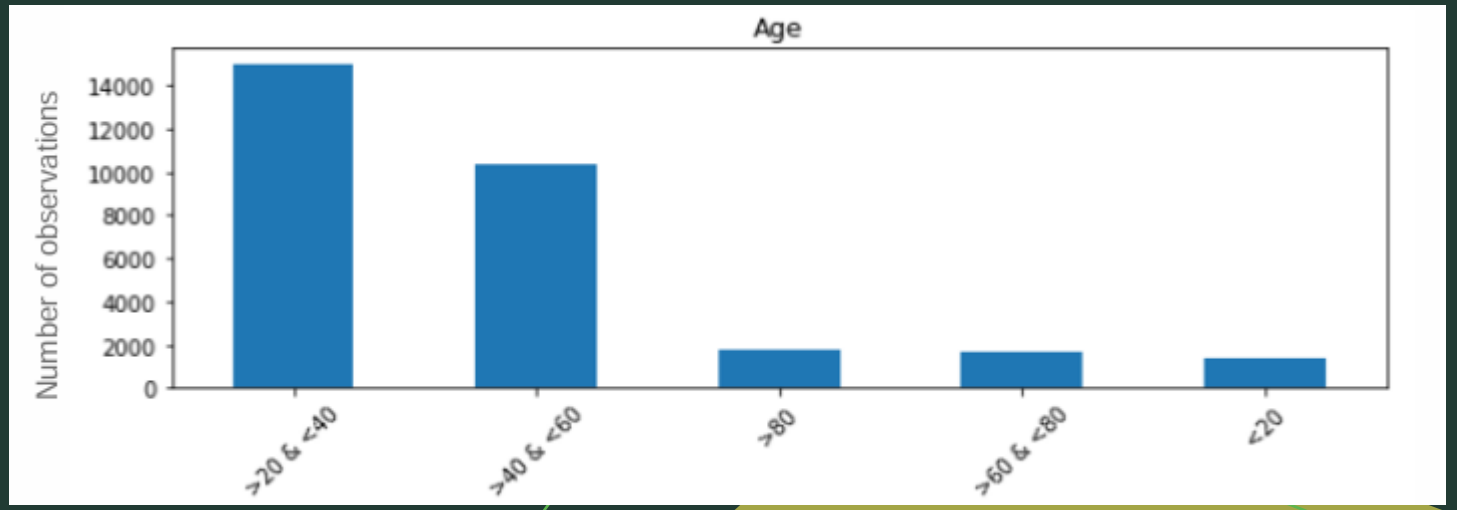
Data Preparation and Cleaning. Focus on attributes manipulation:

Reduction of the distinct values assumed by the ‘Age’ Attribute.

Before Changes



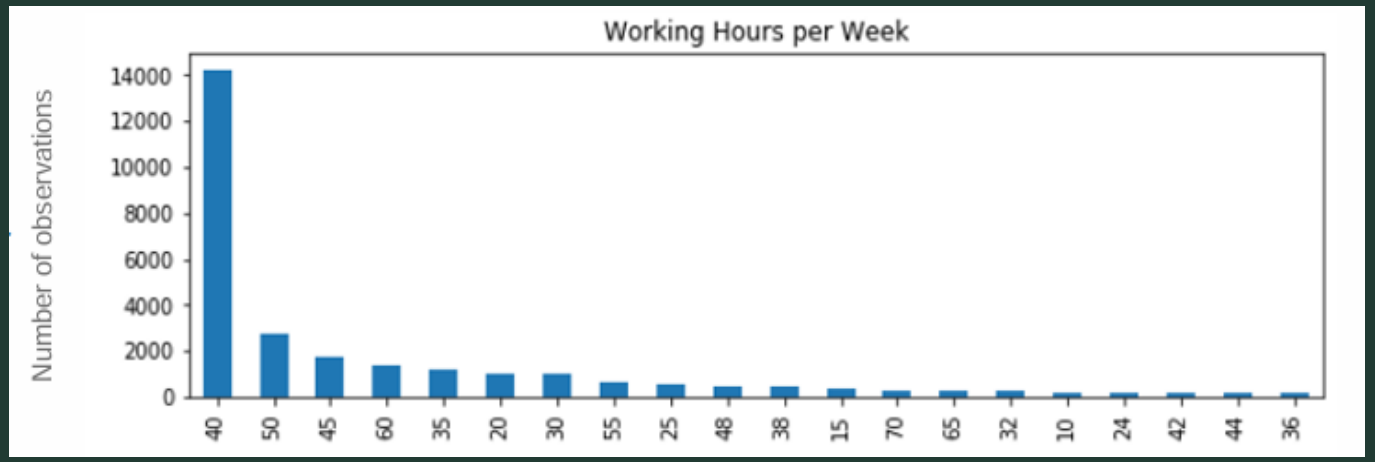
After Changes



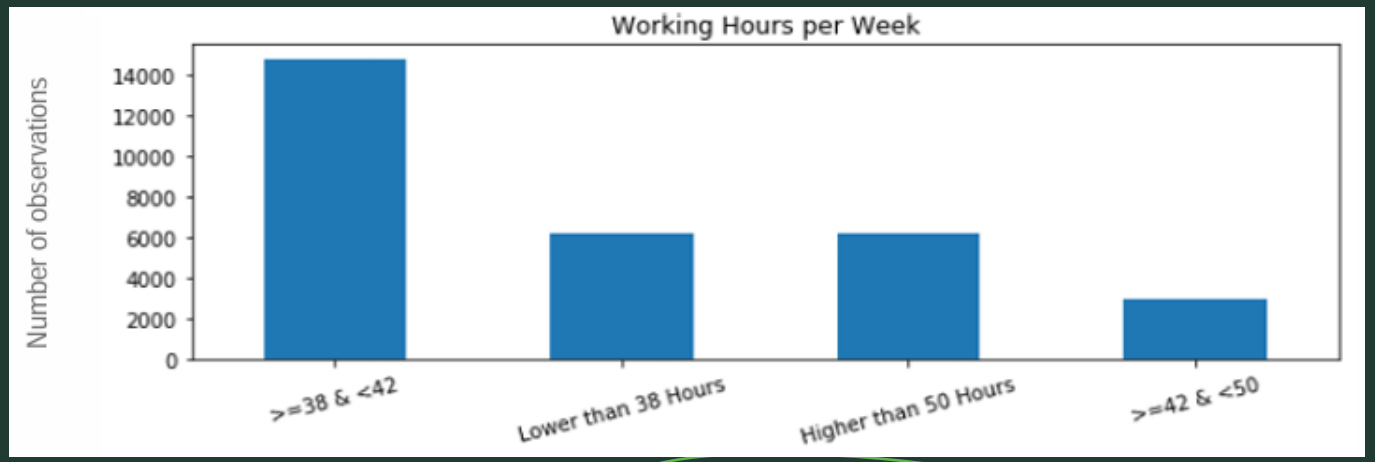
Data Preparation and Cleaning. Focus on attributes manipulation:

Reduction of the distinct values assumed by the ‘Woking Hours per Week’ Attribute.

Before Changes



After Changes



Research Question(s)

- In this project I'm trying to understand what are the main individual characteristics associated to high salaries. Variables like age, education, race, sex, could somehow affect a person's salary. I would like to understand if this is true and, if it is, which of these variables are the most important.

- Through a classification task I'll verify whether the behaviour of some of the variables mentioned above could be associated with the behaviour of the remuneration that a person receive from the market.

- In other, simpler word, what are the most important individual characteristics that explain the differences between individuals in the probability of getting a better wage? Is it true that educated people receive higher salaries? And, is it true that women, on average, receive lower salaries than men?