

CKME 136 CAPSTONE Project

Humaira Asim

Student No: 500813091

United State Census Income Data

Abstract

A census is the process of systematically recording the statistical information about the nation's population. The term is most specifically used with national population, but in general census also includes agriculture, business and traffic census data. The census data is comprised of a number of people, their distribution, their living conditions, education and other key factors, which are critical for the development of the nation. This data helps the policy makers to make decisions for the future and betterment of the country. Income is one of the primary concerns for the standard of living and economic status of an individual and thus, has an important impact on determining the nation's growth. In this project, the aim is to explore the U.S census income data set, relating the earnings with social factors like age, marital status, race, education ... etc. The aim of the project is to explore these attributes and build a classifier which can predict whether the income of an individual is greater than or less than \$50K/year.

Research Questions:

1. Is the income of an individual is greater than or less than \$50K a year?
2. Is literacy directly proportional to the ability of individuals to generate a high income?
3. What factors are the most significant with regards to correctly predicting the income?

Data Source:

UCI Machine Learning Repository

Center for machine learning and intelligent Systems

<http://mlr.cs.umass.edu/ml/datasets/Census+Income>

Techniques and Tools:

Techniques will include data preprocessing, exploratory analysis, feature engineering, regression, random forest and support vector machines (SVM). I will use r language for the project.