US Census Data Analysis

Vasu Sharma
Department of Computer Science
University of Colorado, Boulder
Boulder, Colorado
vasu.sharma@colorado.edu

Yan Zhan
Department of Economics
University of Colorado, Boulder
Boulder, Colorado
yan.zhan@colorado.edu

Sutianjie Zhou
Department of Economics
University of Colorado Boulder
Boulder, Colorado
sutianjie.zhou@colorado.edu

Problem Statement

Census data provides important information about the whole nation. Census data can provide socio-economic information for citizens of the country. IPUMS[1] is an organization that preserves and harmonizes U.S. census microdata and provides easy access to this data with enhanced documentation. In this project, we aim to focus on determining socio-economic trends based on the dataset collected by IPUMS. These trends are collected based on person-level attributes such as education, age, gender, race, employment status, marital status, labor force participation etc. Using data mining techniques, we aim to identify the socio-economic trends by analysing a person's education background and their identity group to predict their earning potential, factors correlated to higher income, change in rate of return on education and income inequality over the past few decades.

Literature Survey

Paper using IPUMS dataset for analysis of social-economic questions have been published in many peer review journals in recent decades.

Sheng et al.[2] performed data mining on census data using CART which is a decision tree based classification model. Their paper was based on using this model to classify inhabitants in the provinces of Chengyang and Laixi.

Federal and local governments[3] use census dataset to assess the economic wellbeing of communities, identify and assist low-income and marginalized communities, and allocate funding to different social programs.

Jaison R. et al [4] used the IPUMS U.S. dataset collection to identify the rate of return on college education. They used income after graduation as a parameter to judge rate of return and discussed the problems regarding stagnant wages and the decline of wages among people without a college degree. They concluded that return has remained high in spite of rising tuition and falling earnings because the wages of those without a college degree have also been falling, keeping the college wage premium near an all-time high while reducing the opportunity cost of going to school.

In Autor, Dorn, and Hanson (2013)[5], they use the PUMA-level (Public Use Microdata Area) data from IPUMS to explain the influence of increasing import from developing countries on the local U.S. labor market. They first generate local labor and employment indexes, then they map the increasing of U.S. imports from developing countries on each local labor market, to figure out the changing of the manufacturing employment indexes. They compare the local employment indexes before and after the increasing shock of import from developing countries (especially after China returned to the WTO in 2001), and explain that the increasing imports from developing countries reduce the U.S. local employment in manufacturing industries. In Autor, Dorn, and Hanson (2019)[7], they also use the IPUMS local data to show the relation between local employment situation of young people and their marital status. The increasing import of U.S. from developing countries reduces the local market manufacturing jobs and changes local employment structure. As the number of local manufacturing jobs decreases, the wage gap among young people increases (the reduction of manufacturing will make the local job market polarization), and the first marriage age of young people increases because more individuals want to spend more time for a potential 'better' partner.

Proposed Work

In this project, based on the dataset, we will analyze the following questions

1. What is the relation among a person's educational background, their identity group and earning potential?
2. What factors are generally found to correlate with higher income?
3. What is the rate of change in income inequality in the U.S. over the past decade?
4. What is the education gap among different racial groups?
5. What is the marriage status difference among people of similar age in different racial groups with different education levels?

To answer these questions, we will first perform data cleaning and preprocessing where we will remove redundant data or apply cleaning techniques to replace duplicate or Null values with measures of central tendency. Then we will transform the dataset by performing binning on income attributes using tax brackets and normalize the data set. After data cleaning and transformation, we will work on building the classification model to identify the income bin for an individual based on their socioeconomic background. We will start with support and confidence calculation and analysis between the interested attributes (whether to receive higher education, married or not, low income or not) and individuals' demographic characteristics (gender, age, race, etc.). Then, we will apply the accuracy measurement in this class to study the relation between education and never being married before. Bayesian classification will also be applied to study the marital status and income gap when we mark different individuals as low-income, middle-income and high-income. After we compare the dependent attributes with objects in different groups based on different attribute standards, we will use an ordinary linear regression model to check the effect of gender, race, and education level on individuals' income. As for education and marital status, we will use a logit/probit model to check the effect of education, gender, racial, and other demographic attributes(in this case, we will set each age from 25 to 45 as fixed effect).

Dataset

The dataset is collected from the organization IPUMS[1] which provides consistent data with documentation. We are going to use the U.S. Census Person-Level data[8] for our project. The 2019 sample will be used for the visualization as well as for building the classification model. The size of this sample is around 806 MB and it has 3239553 rows and 23 columns. Some of these columns contain important person-level information like race, age, gender, income, education, marital status and labor force participation. These columns are chosen from a wide range of columns and are specifically focused towards an individual rather than an entire household or family. The attributes we are interested in and will study includes gender, age, race, education level, location, occupation, marital status, wage income, and total income.

Evaluation Methods

We plan to create a machine learning model which will act as a classifier for predicting earning potential by taking into account a person's socio-economic background. This prediction model will be evaluated against the test split of the original dataset. In the preprocessing phase, earning potential will be binned based on the income attribute, so the model will predict the income bin and it will be evaluated based on confusion matrix, precision, recall, accuracy and F1 score.

Tools

The project work is divided into certain parts and all of those parts require specific tools. The tools for some of those parts are as follows :-

1. **Data Collection and Cleaning** :- For these parts of the project, we will make use of pandas and numpy. Pandas will be used for efficiently carrying out data cleaning and transformation tasks while numpy will be used for performing computational tasks on the data
2. **Data Visualization** :- Matplotlib will be used to visualize data in the form of histograms, bar charts and scatter plots.
3. **Classification and Analysis** :- Sklearn and keras will be used for the classification aspect of the project. PCAs or Autoencoders will be used for performing dimensionality reduction on the dataset.

Milestones

The project work will be divided into 3 parts and the timeline for those parts are as follows :-

1. **Data Cleaning and Visualization** :- We are targeting to do significant work on this and possibly finish this by 18th July. .
2. **Classification and Analysis** :- We are targeting to do this by 30th July.
3. **Correlation Analysis** :- We are hoping to finish this by 1st August.

REFERENCES

[1] Steven Ruggles, Sarah Flood, Sophia Foster, Ronald Goeken, Jose Pacas, Megan Schouweiler and Matthew Sobek. IPUMS USA: Version 11.0 [dataset]. Minneapolis, MN: IPUMS, 2021. https://doi.org/10.18128/D010.V11.0

[2] Bin Sheng and Sun Gengxin, "Data Mining in census data with CART," *2010 3rd International Conference on Advanced Computer Theory and Engineering(ICACTE)*, 2010, pp. V3-260-V3-264, doi: 10.1109/ICACTE.2010.5579631.

[3] https://www.census.gov/econ/overview/go0100.html

[4] Abel, Jaison R. and Deitz, Richard, Do the Benefits of College Still Outweigh the Costs? (August 1, 2014). Current Issues in Economics and Finance, Vol. 20, No. 3, 2014, Available at SSRN: https://ssrn.com/abstract=2477864

[5] David, H., David Dorn, and Gordon H. Hanson. "The China syndrome: Local labor market effects of import competition in the United States." American Economic Review 103.6 (2013): 2121-68.

[6] Steven Ruggles, Sarah Flood, Sophia Foster, Ronald Goeken, Jose Pacas, Megan Schouweiler and Matthew Sobek. IPUMS USA: Version 11.0 [dataset]. Minneapolis, MN: IPUMS, 2021. https://doi.org/10.18128/D010.V11.0

[7] Dorn, David, and Gordon Hanson. "When work disappears: Manufacturing decline and the falling marriage.market value of young men." American Economic Review: Insights 1.2 (2019): 161-78.

[8] https://usa.ipums.org/usa-action/variables/group