Who's Counting?

A 3 Part Analysis of the Adult Data Set

Adrian, Gabriel, Matt, Amara,
Noah Williams

CONNOTATIVE

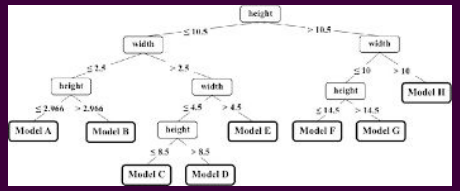
What is the context of the data?

01

Who Made The Adult Dataset And Why

Compiled by Barry Becker

Extracted the data from the US 1994 Census database



Barry Becker gave dataset to Ron Kohavi

Used it to study the effectiveness of a new machine learning algorithm called the NBTree

Ron Kohavi publishes article

Article titled “Scaling Up the Accuracy of Naive Bayes Classifiers a Decision-Tree Hybrid”

Data set donated

UC Irvine Machine Learning Repository

Now hosted by Machine Learning Group at UC Irvine.

How Was The Data Collected?

01

The data was pulled from the Current Population Survey (CPS) conducted by the United States Bureau of the Census.

02

It involved interviewing a nationwide sample of 57,000 housing units across 1,973 counties.

03

Specific extraction rules were applied. For example, individuals had to be over age 16, have an adjusted gross income greater than $100, and work more than 0 hours per week.

How Was This Data Processed?

The original dataset consisted of 48,842 total instances, some of which were incomplete

Unknown information was marked with a “?”

Affected Fields: The missing values were concentrated in three categories: occupation (missing in 5.7% of records), workclass (5.6%), and native-country (1.8%)

Filling in these gaps (imputation) was extremely difficult to do without inserting bias into the model, so they chose to just remove those instances

After those were taken out, the dataset was left with 45,222 instances (3,620 removed)

DENOTATIVE

What is the data set saying?

02

Format Of Variables And Why They Were Chosen

The dataset consists of 14 attributes assigned to each individual, structured into categorical and continuous types:

Categorical (Factors)

These include:

- education
- marital-status
- occupation
- relationship
- race
- sex
- workclass
- native-country

Continuous (Integers)

These include:

- Final weight (# of people in the U.S. population)
- Education (individual's highest level of education.)
- Capital-gain (assets sold for a profit)
- Capital-loss (assets sold for a loss)
- Hours-per-week (hours worked per week by the individual)

Target Label

The income field:

- >50K
- <=50K

fnlwgt - Continues

- number of people in the U.S. population
- how many the people an entry represents
- how common is this set of information

Unable to find complete info on how this is calculated

Unable to find how similar entries would need to be to be counted

Conclusion: Best guess is entry is tied to an area, the entry is a sample, fnlwgt represents the amount of people in the area.

race - Categorical

- race of the person
- White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.

Asian and Pacific-Islander usually in the same category

No category for Arabic or Hispanic/Latin American, usually combined with White or given own category

Conclusion: Racial categories are difficult to delineate, self categorization is pretty good way to make those distinctions

sex- Categorical

- sex of the person
- binary, female or male

Does not specify whether the category refers to sex or gender

No category for intersex people

Conclusion: Classical gender/sex binary. Data is from 1994, authors probably did not care.

workclass - Categorical

- employment status
- do they work, private vs public, self employed?

occupation - Categorical

- general type of job
- tech-support, sales, farming-fishing, etc

marital-status - Categorical

- status of marriage
- Married, Divorced, Children, Single parent, Military

relationship - Categorical

- status in relation to others
- unclear which relationship takes priority
- Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.

DECONSTRUCTIVE

03

What is the data missing?

Political Context of Census Bureau Data

- Created in the 1930s following the Great Depression
- A central challenge survey-makers experienced was determining who belonged in the categories of “employed” or “unemployed”

- Determine the seats each state has in the House of Representatives
- Determine where funding for roads, hospitals, schools, etc. are allocated
- Monitor labor market activity and influence policies to promote economic well-being

Undercoverage of Certain Populations in the 1994 Current Population Survey

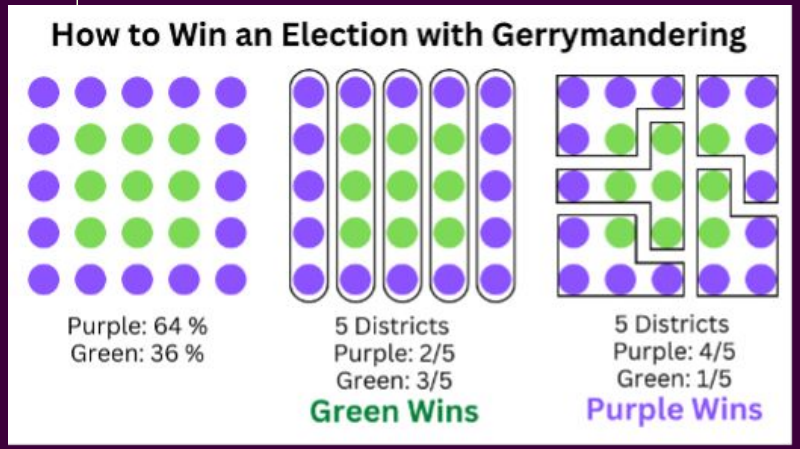
- Missed housing units
- Altered sampling designs between 1980 and 1990 changed the definition of what was considered a "metropolitan area" thus skewing data regarding communities of color
- 8% undercoverage based on age, sex and race
  - ≈29% undercoverage of Black males between the ages of 20-24

U.S. Census Bureau. *Income, Poverty, and Valuation of Noncash Benefits: 1994*. Report no. P60-189, U.S. Department of Commerce, April 1996,

- Data at subnational levels are missing
- Rural, immigrant and communities of color are underrepresented
- Metrics like education and skills may be more reflective of marginalization in the job market and economy than an indicator of income potential

Political Incentives

- Gerrymandering electoral districts to dilute, fragment, or preclude the voting power of certain groups



Less Democracy, One Map at a Time: The Effects of Gerrymandering and How the Freedom to Vote Act Can Limit It - American University Journal of Gender, Social Policy & the Law

- Considering phenomena such as "redlining" helps to historicize the data and points to why certain populations have higher or lower employment rates, funding and general income

Problematizing Becker's Variables

- No African countries included
- The data applies a standardized way of evaluating economies across the world. This standardization dismisses:

–local forms of employment

–local policies regarding taxation

–specific forms of inequality

–domestic economic policies

–locally and culturally specific factors that impact one's income

Impact of Globalization:
Wage Suppression and Job Displacement

- Domestic companies sometimes offshore jobs to pay workers lower wages
  - This tends to to impact lower-income jobs disproportionately and lowers wages overall for local employees
    - Multinational Corporations' profits generally come back to their home country.
      - Lucrative companies have considerable control over labor standards and wage levels
    - Tax Havens exploit lower tax laws in countries where subsidiary companies are stationed

Cleaning

- Cleaning may not be fully random
  - Certain groups may be more likely to have missing records
  - Language barriers, government distrust, inaccessibility
  - Skews the data away from marginalized groups

Models may learn historical patterns of underrepresentation

Conclusion

A widely used benchmark in machine learning

Limitations must be considered

- Must be careful in high stakes scenarios
  - Hiring
  - Social Services

A model and
is not neutral truth

Do you have any questions?

noahwill@uoregon.edu

THANKS