Mining Census Data

By: James Christensen and Mohneesh Daksh

Why census data?

(Reasons for why census data is useful/interesting)

- Helps understand population trends (age, gender, income, education, etc.)
- Shows differences in income and opportunities across different groups.
- Improves planning for housing, transportation, healthcare, and schools.
- Shows real-world patterns that help build predictive models like income classification.

Data Description

- Dataset originally compiled by Barry Becker (Silicon Graphics) from 1994 US Census data in 1996.
- Dataset accessed via University of California Irvine
- 48842 rows (46443 without missing values)
- 14 features (7 categorical, 7 numerical)
- Target Variable - Makes more/less than $50k a year

Variables

Categorical

maritalStatus	Married, never married, divorced, widowed
workClass	Private, public, self, never-worked, etc.
education	Highest education level achieved
occupation	What individual does for work
relationship	Relationship to rest fo the family
race	White, Asian, Native, Black
nativeCountry	Country of birth

Numerical

Age	Years
fnlwgt	Proportion of the population this row represents
educationNum	Years of education after 4th grade
hoursPerWeek	Hours worked per week
capitalLoss	Money lost from investments
capitalGain	Money made from investments
sex	Male, Female

Data Preprocessing

- Removed unnecessary ID/index column and replaced "?" with proper missing value markers (NA)
- Checked missing values in each column and removed rows that contained them
- Scaled all numeric features for better model performance
    - Z-score scaling
- Converted categorical columns into factors and applied one-hot encoding to create numeric dummy variables.

Market Basket Analysis (Association Algorithm)

- Pros:
  - Finds patterns that are both easy to interpret and ***not*** obvious by merely glancing at the data.
  - Scales to large datasets (like this one)
  - Works well without labels
- Cons:
  - Many patterns are often useless (Milk => Bread)
  - Sensitive to data sparsity
  - Sensitive to the chosen support threshold

Rule Analysis

- Rules based off of education, marital.status, occupation, race, sex, native.country
- Support set at 1%, confidence 50%
  - 1% as we don't want to miss out on rules potentially involving less common occupations or native countries
  - 50% to ensure that all interesting rules are found
- 824 itemsets found based on support, 1330 rules subsequently found based on confidence
- Parsed through the 75 with the most lift
  - Lift chosen as it is easy to interpret. Lift of 1.4 means 40% more likely than chance to be associated together

The most interesting rules

- {Some college, Service occupation} => {Never married}
  - Lift 1.93, i.e. almost twice as likely to be unmarried than average
- {Master’s degree,, Male} => {Married civilly}
  - Lift 1.618
- {Works in sales, Male} => {Married civilly}
  - Lift 1.40
- {Highest education 8th grade, Male} => {Married civilly}
  - Lift 1.548
- No rules in the top 75 based on lift involve women?

Random Forest (Classification Algorithm)

- Pros
    - High accuracy
    - Robust to outliers
    - Good with high dimensions and mixed data
- Cons
    - Slow to fit and predict (but still faster than neural networks)
    - Hard to interpret
    - Large memory footprint

Our Random Forest

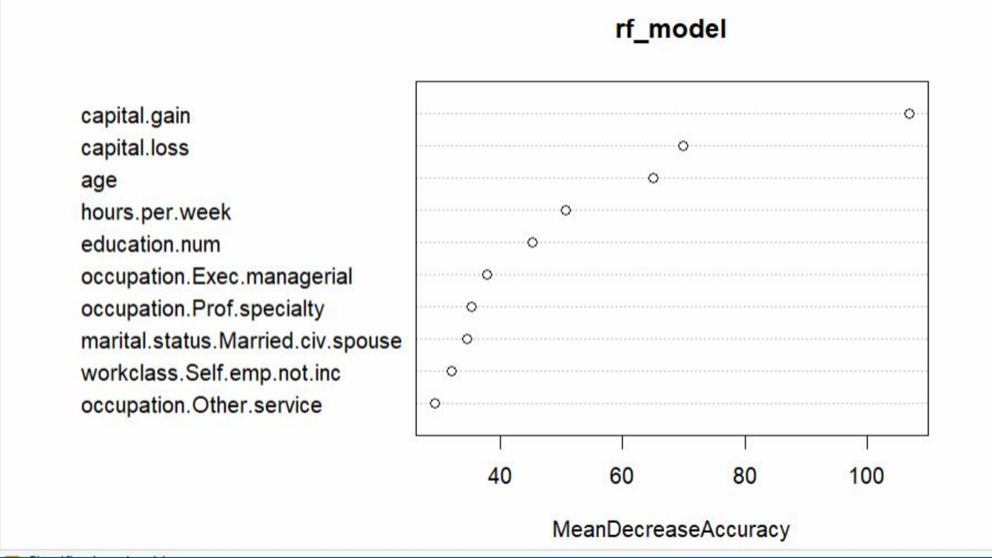
- 500 trees
- Each tree only allowed a random sample of square root number of predictors
- Train/test split 80/20

Evaluating Variable Importance

Mean Decrease Gini

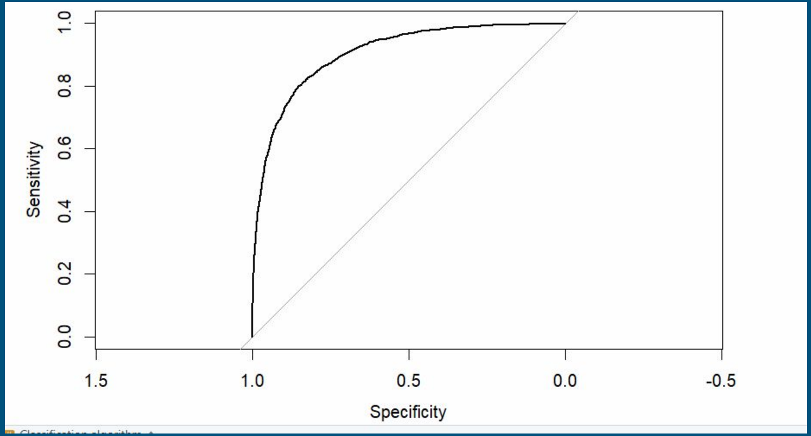
Variable <chr>	Importance <dbl>
capital.gain	89.58679
capital.loss	53.14067
workclass.Self.emp.not.inc	33.15424
education.num	24.36139
marital.status.Married.civ.spouse	20.70551

Mean Decrease Accuracy



Evaluating Classification Performance

ROC Curve




```
> auc(roc_obj)
Area under the curve: 0.9052
```

Confusion Matrix

```
          Reference
Prediction    0    1
         0 6603  857
         1  414 1413

               Accuracy : 0.8631
                 95% CI : (0.856, 0.8701)
    No Information Rate : 0.7556
    P-Value [Acc > NIR] : < 2.2e-16
```

Accuracy = 86%
Precision = 77%

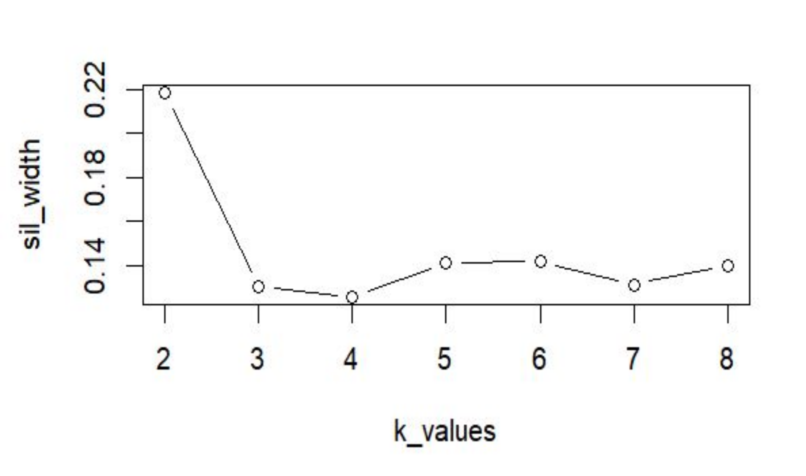
Recall = 62%
F1 Score = 69%

CLARA (Clustering Algorithm)

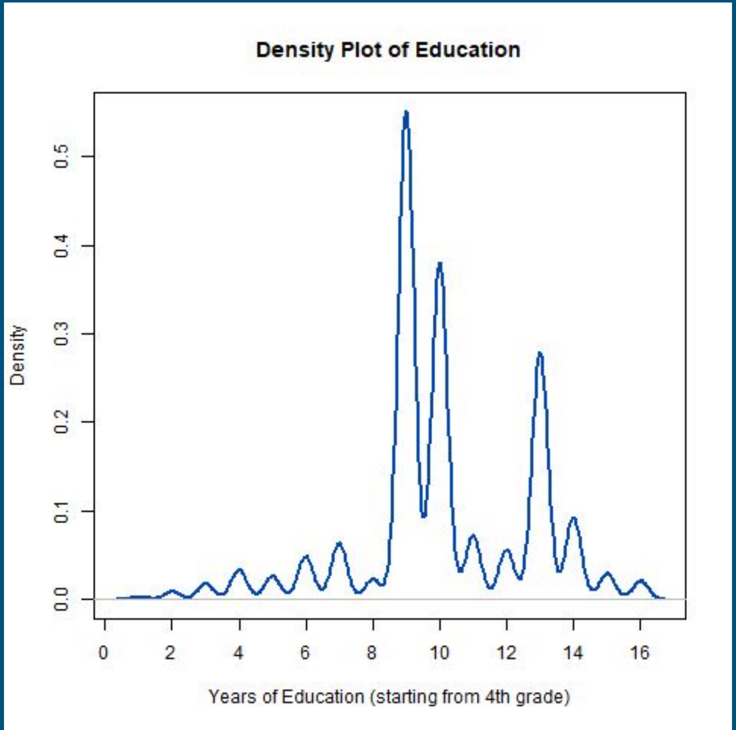
- Pros
  - Performs PAM on samples
    - Works well on large samples
    - Works with any distance metric
- Cons
  - Still slow
  - Requires representative samples

Our CLARA

- One sample of 2000 taken to find best number of clusters (5)
- 3 subsets of roughly 15500 rows
    - A random sample of size 2000 taken from each subset
- Medoids found for each sample, mapped to find common clusters
- Subsets labeled and combined back into entire dataset

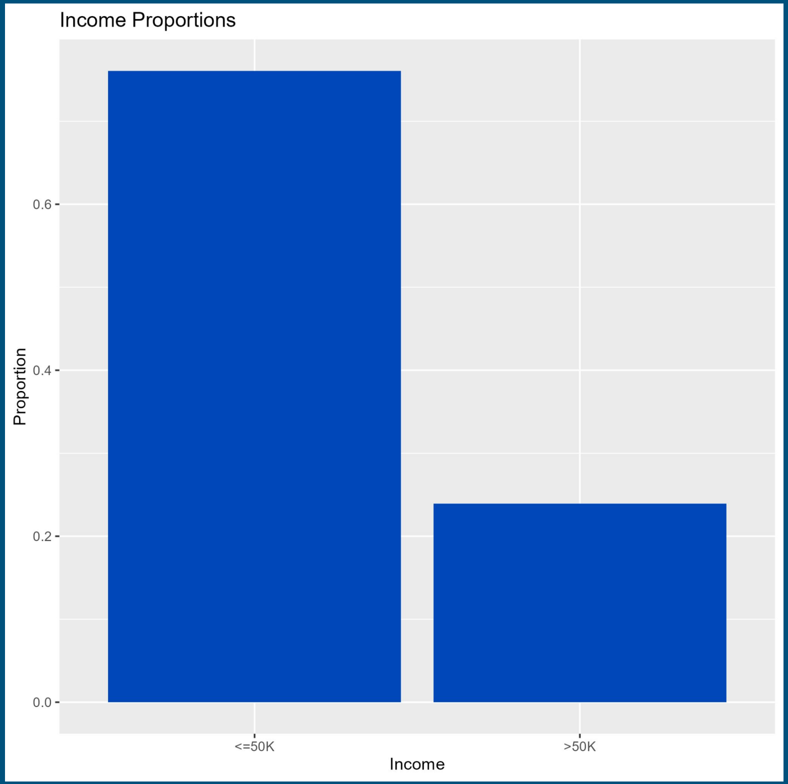


General EDA

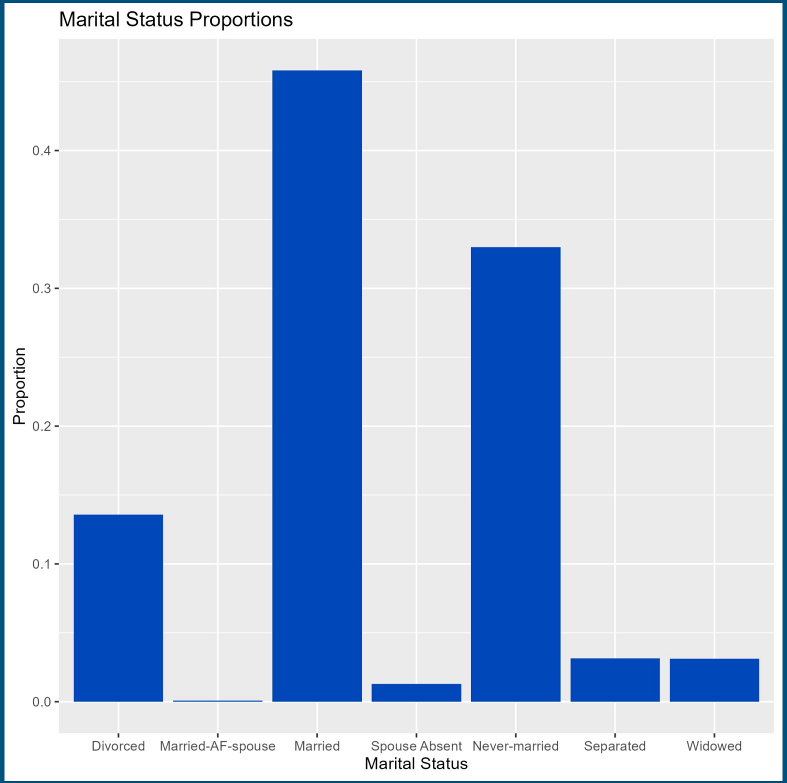


Race	Proportion
White	0.86
Black	0.10
Asian/PI	0.03
Native American	0.01
Other	0.01

General EDA



Income Proportions
Proportion
0.0
0.2
0.4
0.6
<=50K
>50K
Income



Marital Status Proportions
Proportion
0.0
0.1
0.2
0.3
0.4
Divorced
Married-AF-spouse
Married
Spouse Absent
Never-married
Separated
Widowed
Marital Status

Clusters

Cluster 1 (9251 people, age 44)

- 90.6% white
- 91.6% married
- 90.8% male

Interesting fact:

- 79.2% income >50k

Cluster 2 (8913 people, age 30)

- 81.7% never married
- 86.5% male
- 95.1% income <50k

Interesting fact:

- 63.1% High school or less vs. 45.1% generally

Cluster 3 (8729 people, age 29)

- 79.4% never married
- 74.7% female
- 94.2% income <50k

Interesting fact:

- 45.1% have some college vs. 22.1% generally

Cluster 4 (6998 people, age 46)

- 58.8% divorced
- 89.2% female
- 92.3% income <50k

Interesting fact:

- 30% more government workers than generally

Cluster 5 (12552 people, age 43)

- 97.0% male
- 94.1% married
- 89.5% white

Interesting fact

- 71.4% High school or less

Questions?