



U.S. Census Bureau

PREDICTING INCOME LEVEL USING CENSUS DATA

MARVIN EDMOND

08-13-2018

SPRINGBOARD CAPSTONE



Contents

1. U.S. CENSUS BUREAU OVERVIEW
2. PROJECT PURPOSE
3. DATA SOURCES
4. DATA TRANSFORMATION AND ANALYSIS
5. MACHINE LEARNING APPROACH
6. RESULTS
7. SUMMARY

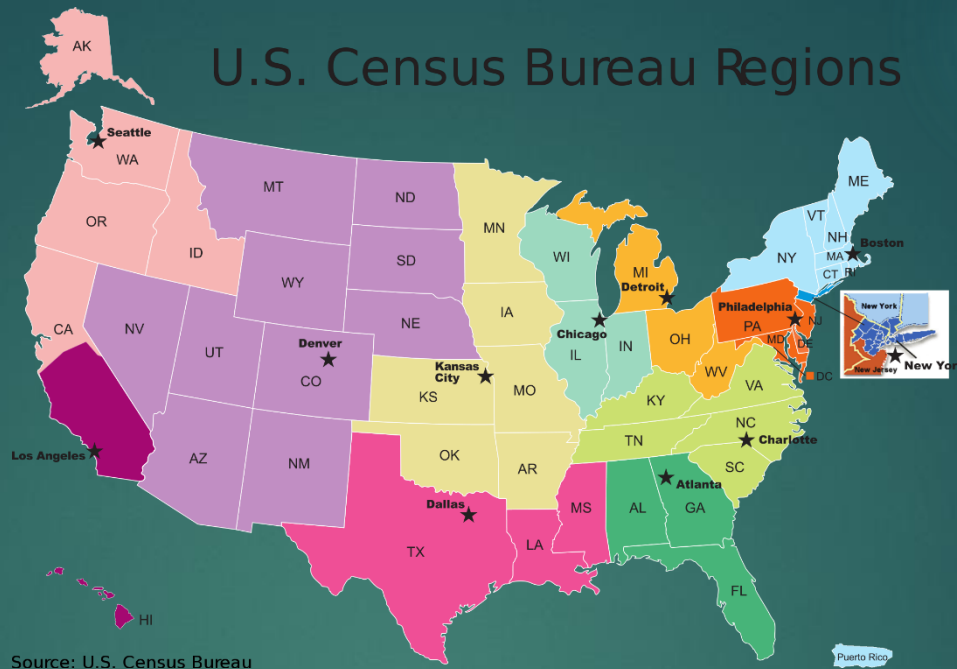


U.S. Census Bureau Overview




- ▶ The U.S. Census Bureau has been headquartered in Suitland, Md. since 1942, and currently employs about 4,285 staff members.
- ▶ The Census Bureau is part of the U.S. Department of Commerce and is overseen by the Economics and Statistics Administration (ESA) within the Department of Commerce.
- ▶ The Economics and Statistics Administration provides high-quality economic analysis and fosters the missions of the U.S. Census Bureau and the Bureau of Economic Analysis.

Project Purpose



- ▶ Wanting to practice my newly acquired machine learning skills, I searched for a project which would be interesting.
- ▶ Census data is always available with an abundance of information but, determining which specific variables to use to predict annual income was a fun challenge.
- ▶ Even though the project was simple, it gave me further insight into the power of machine learning.

Data Sources



The Census Bureau did a wonderful job in bringing some order to the available datasets. A plentiful amount of variables were provided for thorough analysis. The data used in my project is from the 1994 Census survey.

Data Files:

- ▶ Data Folder
- ▶ Data Set Description

The Data Folder includes:


- ▶ Train and test data sets.
- ▶ Variable names for each column.



Data Transformation and Analysis

Data Set Description:

Variable	Description
age	continuous
workclass	Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
fnlwgt	continuous
education	Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
education-num	continuous
marital-status	Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
occupation	Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op- inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.
relationship	Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
race	White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
sex	Female, Male.
capital-gain	continuous
capital-loss	continuous
hours-per-week	continuous
native-country	United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinidad&Tobago, Peru, Hong, Holand-Netherlands.



The data provided by the Census Bureau is semi-unstructured but the data dictionary helped out tremendously in helping to clean the data. A few issues occurred while wrangling with the data which were:

- ▶ Reassigning easy to read variable names to the data.
- ▶ While checking for missing values, I noticed missing values contained a '?' instead of an NA value.
- ▶ All missing data were later converted to NA values.

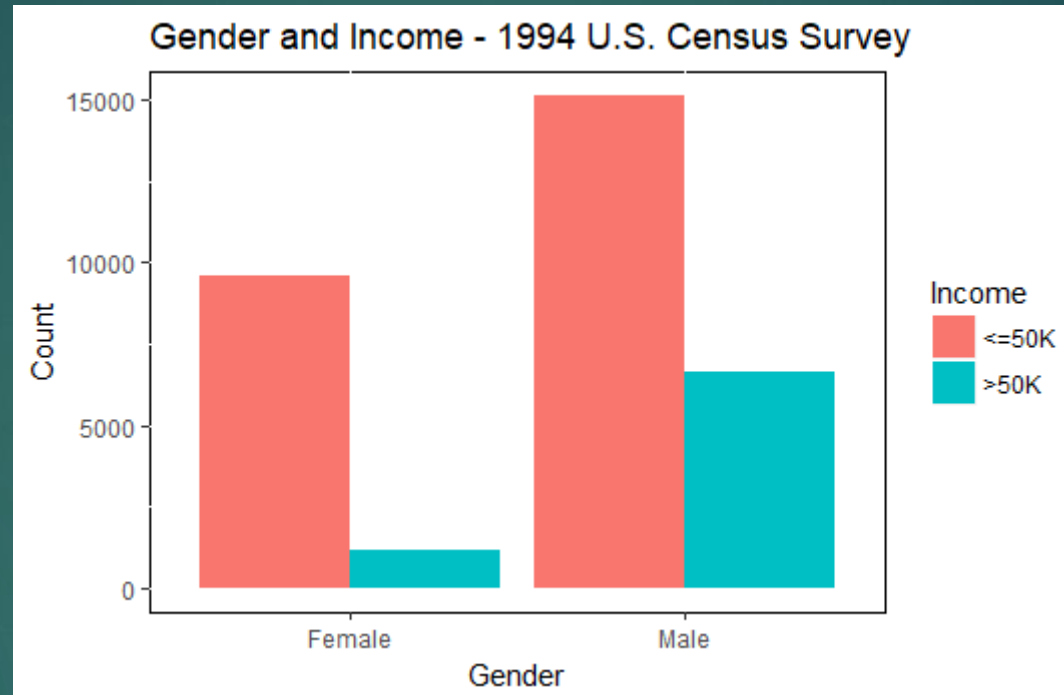
Data Transformation and Analysis Continued..

- ▶ Important information that the data contains are age, gender, work class, occupation and education level. These factors help to create a profile which can be further analyzed to increase predictability for a predetermined income level. The use of character defining traits produces more efficient training of data sets further strengthening algorithms when it comes to testing.
- ▶ However, the data set does provide some limitations. The absence of specified states/cities in the survey makes it impossible to determine which regions have the highest income level. This piece of information could have further aided the algorithms in determining if an individual makes over a certain amount of income per year. Also, the knowledge of state tax levels would help us to determine which areas of the U.S. did individuals retain more of their earnings.

Data exploration is vital for understanding your data before performing further analysis. Familiarizing yourself with the data visually, quickly helps to determine correlation between variables. Investigating correlation amongst several variables could provide valuable insights pertaining to my capstone involving U.S. Census data. Variables for investigating correlation:

- ▶ Income vs Education
- ▶ Age vs Income
- ▶ Income vs Gender

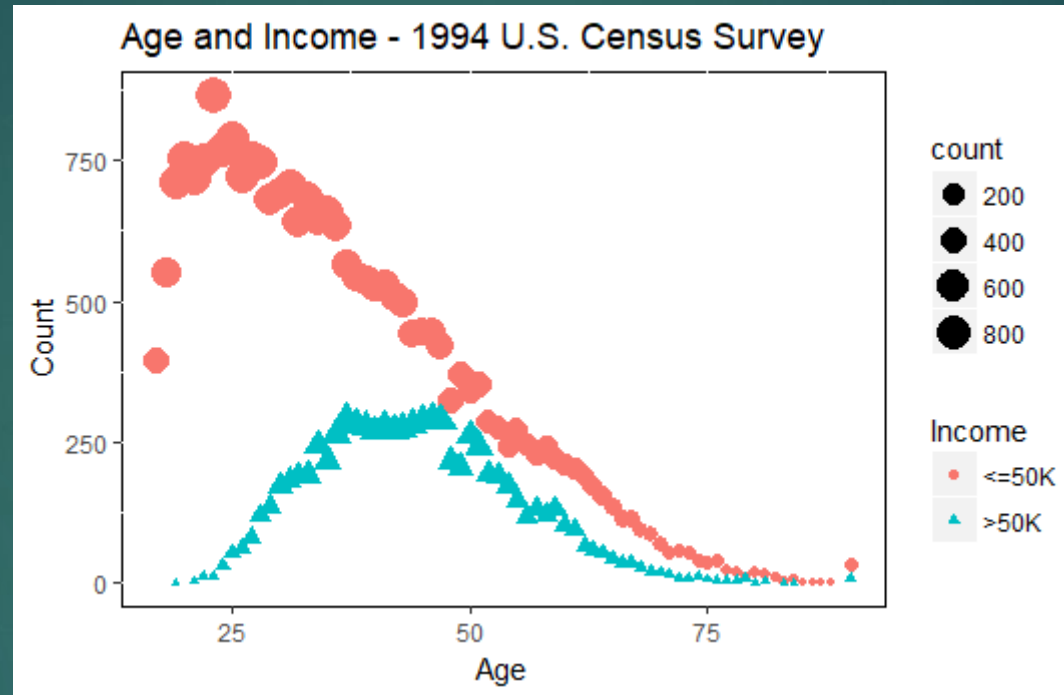
Data Transformation and Analysis Continued..



The effects of gender on annual income for female laborers are evident.

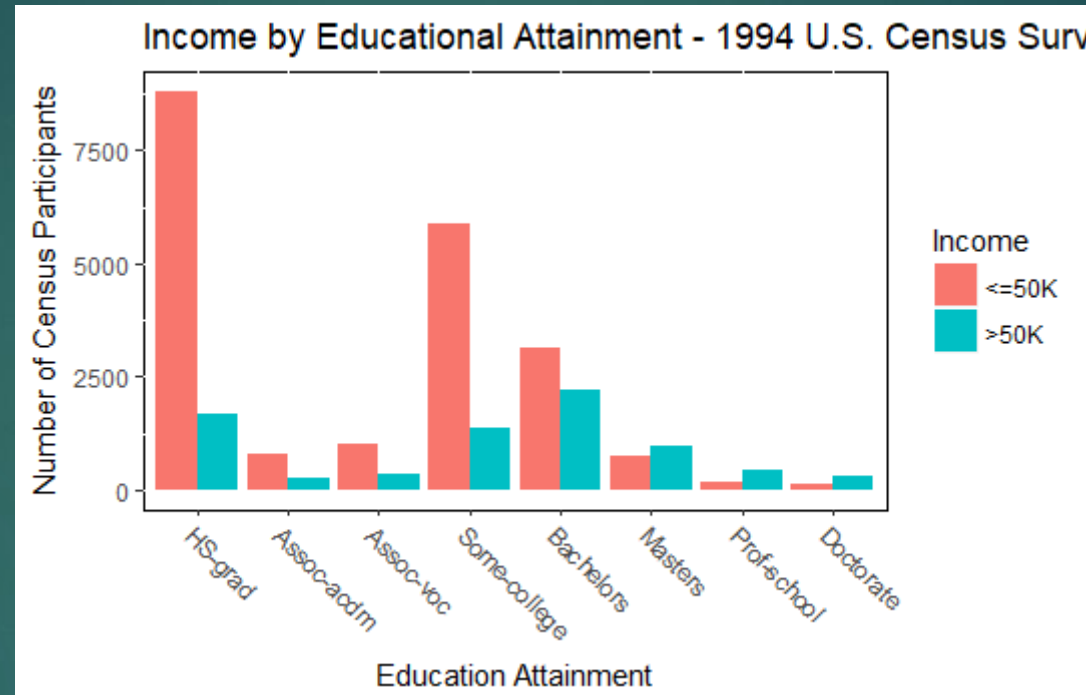
- ▶ The percentage of females that are compensated over \$50K/year compared to their overall aggregated income is a very small amount.
- ▶ Furthermore, the percentage of males that make over \$50k/year compared to their aggregated overall income is far greater than their female counterparts. This represented a huge pay gap, reminiscent of that time period.

Data Transformation and Analysis Continued..




- ▶ Individuals between 18-35 years old have a wider disparity of income.
- ▶ The majority of this age group's annual income is less than or equal to \$50k/yr. As individuals grow older the gap in annual income begins to shrink.
- ▶ 18-25 year olds are usually in school either full-time or work part-time jobs. Students and recent graduates are navigating various career paths so generous employment offers are few and far between.

Data Transformation and Analysis Continued..



- ▶ Individuals attaining only a high school diploma are more than likely to make less than or equal to \$50k/year in 1994, as well as associate and bachelor degree holders.
- ▶ As advanced levels of education are sought, the probability of making over \$50k/year rises in proportion.
- ▶ Progressing from a masters level to professional school, then finally a doctorate, the probability of making less than or equal to \$50k/year decreases and the probability of making over \$50k/year increases.

Machine Learning Approach



Before implementing a ML algorithm, additional data transformation was needed.

- ▶ The dependent variable was converted to a binomial, 0 or 1, instead of string.
- ▶ Training and Testing sets were created through sampling the original data set, 70% and 30% respectively.
- ▶ Packages “e1071” and “effects” were used for modeling and plotting my glm algorithm.

Machine Learning Approach Continued..

The GLM model uses three significant independent variables, “Age”, “Education Number”, and “Gender”.

```
glm(formula = class ~ Age + Education_Number + gender, family = binomial(link = "logit"),
     data = training)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.4299	-0.6692	-0.4421	-0.1437	3.2707

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-7.829689	0.117853	-66.44	<2e-16	***
Age	0.042582	0.001358	31.35	<2e-16	***
Education_Number	0.374325	0.007895	47.42	<2e-16	***
gender	1.332570	0.044168	30.17	<2e-16	***

Machine Learning Approach Continued..

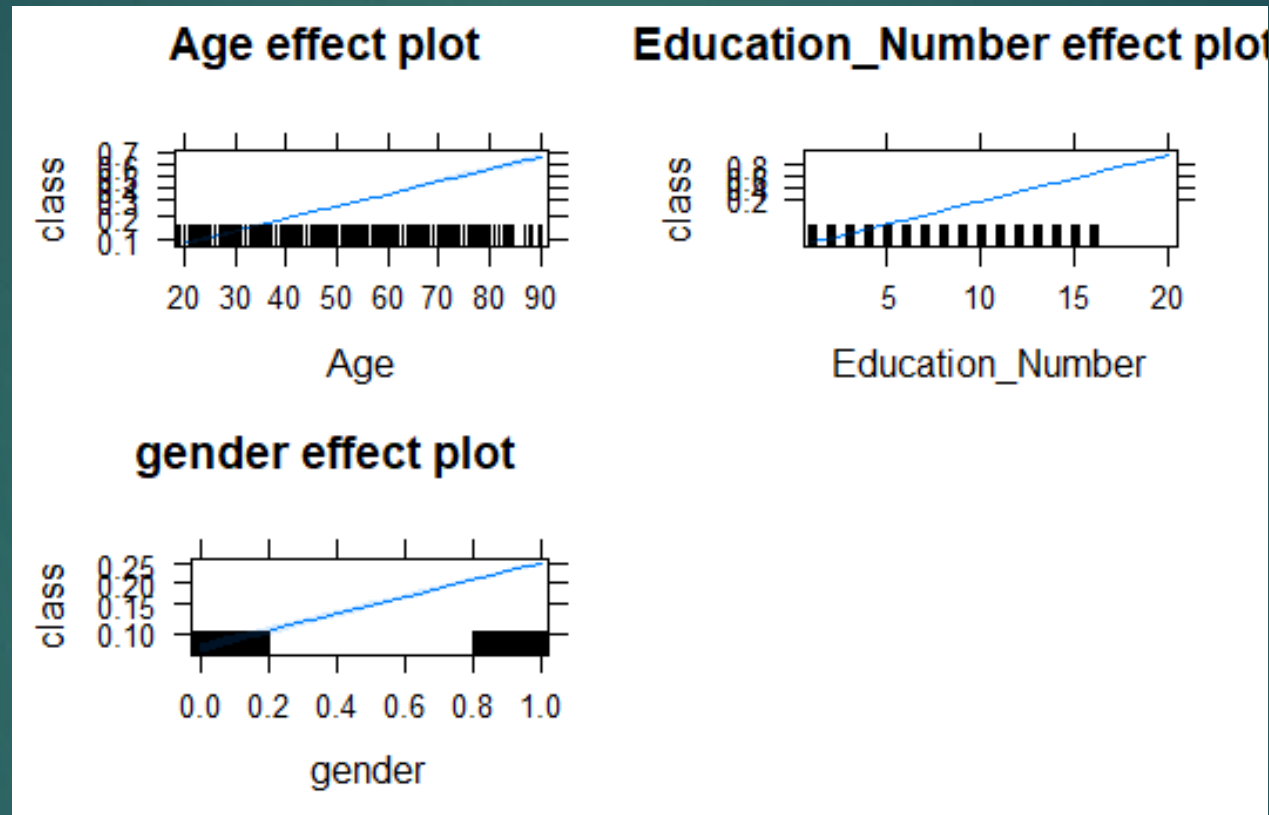
After running the `predict()` function, re-assigning the results to 0 or 1, and doing an accuracy test, my model has an **accuracy of 78%**.

- ▶ For income levels under \$50k, my model correctly predicted 6,956 survey answers, and incorrectly predicted 447.
- ▶ For income levels above \$50k, my model correctly predicted 661 survey answers, and incorrectly predicted 1,705 survey answers.

**** 0,5** was used as a baseline cutoff limit for assigning either 0 or 1.


```
> results_nb_model = ifelse(nb_pred<0.5,0,1)
> table(testing$class, results_nb_model)
  results_nb_model
      0      1
0 6658  745
1 1404  962
```

Machine Learning Approach Continued..



A plot was created using the “effects” library, to properly visualize my model. The relationship between “Age”, “Education”, and “Gender” to “Income” is positive. As education or age increases, income level rises as well. Furthermore, males have more earning power than females.

Results



This census data set is rich with information that can clearly be used for predictive applications.

- ▶ As individuals grow older, their income level will rise. This signifies that tenure within a career leads to more money, whether by increased experience or promotions.
- ▶ Higher levels of educational attainment raises an individual's income level. Bachelor degree holders suffer the most variance in income levels, but postgraduate degree holders are proven to earn more in their lifetime.
- ▶ With the time period of this survey being 1994, males were the majority income earners across all income levels. The percentage of males earning over \$50k a year was 50% higher than their female colleagues.

Results Continued..

Extending my ML algorithm to include factors such as race, work class, and occupation, can prove beneficial in closing the income gap between men and women, low earners and high earners, and across racial lines. A few suggestions are:

- ▶ Programs to increase affordable educational opportunities can help disadvantaged individuals in achieving the American dream and acquiring a sustainable living wage.
- ▶ Introducing male-dominated professions to women at an early age will have a positive effect. This will provide additional career choices as well as increase their income over their lifetime.
- ▶ Analyze the disparity that minorities face in education, employment, and income to foster new ideas for the elimination of this disproportion.