# Introduction

The data from the 1993 Census database is to be examined for the purposes of determining whether a person makes over $50,000/year. We were given the task to examine and predict who, and why certain individuals make over $50,000/year using various attributes in the data. Our group will employ some basic classification to make this determination. Before proceeding it is very important to note that this is an observational study and the conclusions and lessons learned here are limited in scope to the data from our analysis.

# Data Description

The dataset come to us from the UCI Machine Learning Repository and is officially titled “Census Income dataset.” The data contained within this dataset contains various attributes of adults collected from the 1994 census and has the intention of being used to predict whether an adult will make greater than $50,00/year or less. The data is as follows: 48,842 instances of income data and its associated variables, although our working files are a bit different. The dataset is multivariate, and has both categorical, and integer attributes.

Before proceeding it is important to note that all EDA, and analysis was performed on the training dataset, commonly referred to throughout the markdown file and analysis file as incomeDataUSE. The validation was performed on a separate testing file commonly referred to as incomeDataTestMaster. We will make specific notes when we use the testing file.

# Objective 1: Display the ability to perform EDA and build a logistic regression model.

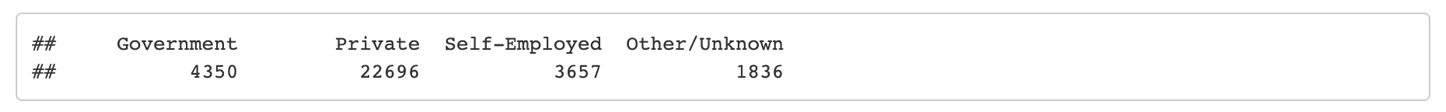
# Exploratory Data Analysis

The data was thoroughly examined before analysis, and a few observations on the data should be noted. To begin, prior to analyzing the data, Barry Becker extracted a “reasonably clean” set of records from the 1994 census database, with the point being that any data cleaning activities were done on top of the previous work.

Text

Description automatically generatedA good starting point for a large dataset is to examine the amount of missing data contained within each of the variables. We believe due to the relatively clean data contained within the raw data, this 100% complete factor is not entirely surprising. However, the data still needs further investigation.

(Note: The test data was also run through this same metric and the results were the same, the plot for the testing data is located in appendix A.2). Within the data a few items needed to be addressed. To begin, the census collects data on participants job “category.” The original data had many levels of job category best lumped into more all-encompassing categories. Local, state, and Federal government employees were combined into a single “government” classification in order to simplify visualization and EDA. This same technique was also employed for individuals who did not fall into one of the standard job categories for a variety of reasons, these persons were placed into an other/unknown category. This technique was replicated on the test dataset Below is the end result of this process:



The only real issues present with the dataset all existed at the level of poor categorization and needing to be simplified for ease and interpretability after analysis.

Next up is an examination of the distribution of the outcome variable. We can see from this histogram that >50K is heavily outweighed by <50k. This could be a potential issue during analysis. We also decided that it would be best to examine the distribution of gender among

Chart, histogram

Description automatically generatedthose surveyed, and there is much less disparity between gender, however, females clearly have greater numbers here, and this disparity could be something to examine/keep in mind as we proceed.

On the topic of variable selection the EDA process lent itself to our group deciding on the fate of many of the variables.

Chart, histogram

Description automatically generated

We quickly noticed that capital-gain and capital gain applied to a very small subset of the population, and upon further review capital-loss was also in that same group. In the goal of simplicity of the model and being able to correctly represent the population, those two variables were removed, as 90% of the training data did not feature either of these variables. We also removed native country from the data because most of the observations are from the United States and skewed the data unfavorably. Additionally, we dropped the categorical version of education, because the data was encapsulated in the education level variable and was deemed redundant. We also dropped fnlwgt due to the reason we did not feel comfortable using this variable in our analysis as we did not clearly understand how to approach using such a variable. Finally, we dropped relationship, we found that this variable was also encapsulated in gender and family role, essentially duplicated information. We were left with eight explanatory variables, and our response variable. The remaining variables were (and the ones to be used in analysis): Age, Workclass, education-num, marital-status, occupation, race, sex, hours-per-week, and income. The training data contained roughly 16,000 entries all of which had complete data, the testing data has roughly the same demographics. The appendix features (A.0) the variable names and descriptions in more detail. The summary statistics on each variable are available in the appendix in sections A.3 and A.4 respectively. There is nothing unusual to note on the summary statistics, and there were often checked after any data modification procedures.

Our EDA continued by examining relationships between variables of interest within the dataset, age proved to be powerful visual tool for assessing income data. Within the appendix are various plots that lead to our variable selection and generally informed us of the nature of the income data. Appendix A.5 visualizes age and working class to reveal the trend of working less as a person ages. A.6 visualizes education level with income, this plot very clearly illustrates the trend of the association between income and education. A.7 shows working class distribution, and indicates that private sector employs most, which is true of the general population, and this begins to indicate that the sample is potentially representative of the population as a whole. A.8 is useful as the plot visualizes the occupation and income status. A.9 looks at race versus income, and A.10 shows gender vs income. All in all these plots give us good visual insight into the data and were the guiding factors into our analyses.

As a final point on EDA, as mentioned above all EDA was performed on the training dataset. It should also be stated that no points were removed from the data after viewing Cook’s D and checking for outliers. The data is fairly uniform as well.

# Objective 1: Logistic Regression Analysis

We have been tasked to perform a logistic regression analysis, provide interpretation of the coefficients (hypothesis testing, and confidence intervals).

After selecting our variables in the above sections, we proceeded to create a logistic regression model with income as the outcome variable. We used a glm model with the specification of a logit function from the binomial family. We then additionally fit a forward selection model and a backward selection model. The models were run on the training data, and then tested on the testing data, the full results are below. Despite running three different models, we are basing our parameter estimates, confidence intervals, hypothesis tests, and other performance metrics based on the GLM (logit) model.

First and foremost, we must address assumptions for logistic regression. The first assumption to check is the linear relationship between the continuous predictor variables and the logit of the outcome. From appendix B.1, we can see that the smoothed scatter plots show that all of the numeric variables are linearly associated with the income outcome in logit scale. The curve within age is worthy of examination and a potential transformation. We then examined influential values via Cook’s distance values (Appendix B.2), we had several points of interest, but upon further review we determined the points to fit within the data. Next, we examine the standardized residuals (B.3), this plot confirms the decision based on Cook’s distance. Finally, we address multicollinearity, which was an issue with a few records from occupation, however that was remedied by removing the offending records (<5 records removed).

Interpretation of regression coefficients, hypothesis tests, and confidence intervals here

The glm model “scored” relatively well when pitted against the testing data. We achieved 83% classification accuracy (~16% error rate), 92% sensitivity, 54% specificity. The results from the confusion matrix from the testing data are located in appendix (B.5).

With the intention of model validation, we also created a forward selection, and backward selection model to compare/compete with the original model. The models validate our original model, and this concludes objective one.

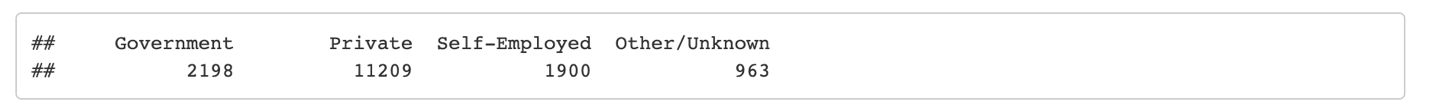
# Appendix A (EDA):

Data columns listed in the data file (A.0):

|  |  |  |
| --- | --- | --- |
| **Variable Name** | **Data Type** | **Description** |
| age | Continuous | Age of subject |
| workclass | Discrete | Category of subjects employment |
| fnlwgt | Continuous | People represented by this entry in census |
| education | Discrete | Level of education of subject |
| education-num | Discrete | Numerical representation of education |
| marital-status | Discrete | Marital status |
| cccupation | Discrete | Occupation |
| relationship | Discrete | Role in family |
| race | Discrete | Subject’s race |
| sex | Discrete | Sex/Gender |
| apital-gain | Continuous | Income from financial investments |
| Capital-loss | Continuous | Income loss from financial investments |
| Hours-per-week | Continuous | Hours worked per week |
| native-country | Discrete | Subject’s native country |
| Income | Discrete | Subject earned <50k or >50K |

Exhibit A:

(A.1)

Recategorized employment clusters of test data set

(A.2)

Missingness plot for test data

Text, letter

Description automatically generated

(A.3)

Summary Statistics for IncomeDataUSE

A close-up of a paper

Description automatically generated with low confidence

(A.4)

Summary Statistics for IncomeDataUSEMaster

A picture containing text

Description automatically generated

(A.5)

Age and working class histogramChart, histogram

Description automatically generated

(A.6)

Education Level vs income

Chart, bar chart

Description automatically generated

(A.7)

Working class distribution

Chart, bar chart

Description automatically generated

(A.8)

Occupation Distribution

Chart, bar chart

Description automatically generated

(A.9)

Race vs income

Chart, bar chart

Description automatically generated

(A.10)

Occupation Distribution

Chart, bar chart

Description automatically generated

# Appendix B (Logistic Regression):

(B.1)

Linearity

Chart

Description automatically generated

(B.2)

Cook’s Distance

Chart, histogram

Description automatically generated

(B.3)

Residuals vs. Leverage

Chart

Description automatically generated

(B.4)

Standardized Residuals

Chart, scatter chart

Description automatically generated

(B.5)

Confusion Matrix for model 1

|  |  |  |
| --- | --- | --- |
| Prediction | <=50k | >50k |
| <=50k | 11458 | 1761 |
| >50k | 968 | 2083 |

# Appendix C (Objective two):