PREDICTING INCOME, A Case Study

Prepared by Shoshana Collins
on behalf of:

RTI INTERNATIONAL

SUMMARY

The investigation into the US 1994 Census data yielded promising initial findings. A preliminary model was developed using logistic regression to estimate the chances that an individual would have an income greater than $50,000 based on a variety of variables including work class, marital status, and whether the individual reported a capital gain or loss for the year. Intuitively, an individual with a large capital gain was over 65 times more likely to have income over $50,000 than one without. However, it was also notable that an individual with a capital loss was still over 3 times as likely to have high income. This does make sense in light of the fact that most people probably don't invest in the stock market without a substantial income.

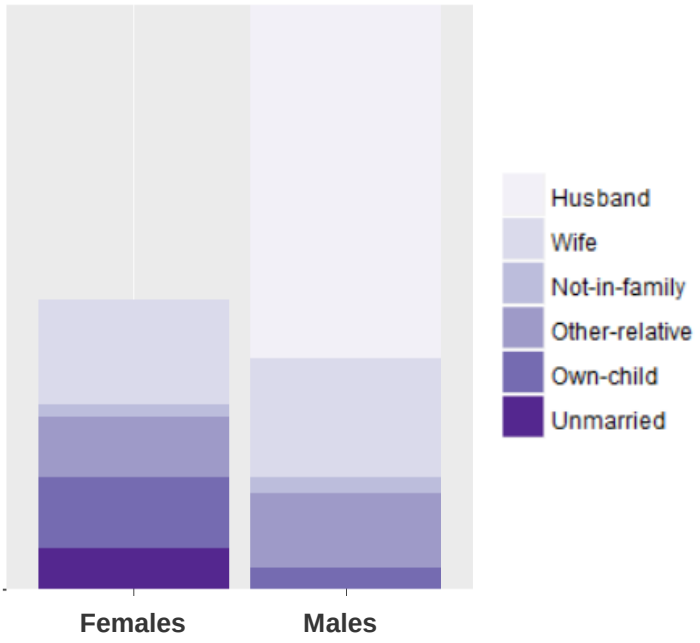
While the model yielded interpretable results for estimating the likelihood of high income, technical difficulties inhibited a reliable evaluation of the accuracy of the model when tested on new data. Additionally, several profound deficiencies in the data source, highlighted below, suggest that further investigation is required before applying this model to the general population.

NOTABLE DATA INCONSISTENCIES:

- 23.9% of individuals in the provided data set had income greater than $50,000, which is not in line with official 1994 Census figure of approximately 9.1%
- Males outnumber females in the data set by a ratio of about 3:1, which is dramatically skewed from the true US population, which has slightly more females than males (see chart "Count of Respondents by Gender)
- Only 29 individuals out of the entire 48,842 person sample reported being without pay or having never worked, an "unemployment rate" of less than 0.06%
- The data set included almost 10 times as many husbands as wives, while the true ratio should be ~1 to 1 (see chart "Count of Respondents)

COUNT OF RESPONDENTS BY GENDER

ORDERED BY RELATIONSHIP TYPE



MOST SIGNIFICANT PREDICTORS

IN ORDER OF MAGNITUDE OF CHANGE IN LIKELIHOOD

* Some factors are excluded due to extremely low sample representation	
Individual Claimed a Capital Gain	65: 1
Married to a Civilian Spouse	13 : 1
Married to a Spouse in the Armed Forces	12 : 1
At Least High School Education	6 : !
Professional or Doctorate Degree	5 : 1
Not in Occupation – Private House Service	5 : 1
Individual Claimed a Capital Loss	3 : 1
Relationship – Wife	3 : 1
Self-Employed – Not Incorporated (negative)	3 : 1
Race – Asian/Pacific Islander	3 : 1