Artificial Intelligence Final Project

Census Analysis

Code and Report by Caleb Pudvar

INTRODUCTION:

This analysis aims to determine whether or not there are any inherent patterns in income based on certain qualities of a person. Data extraction was done by Barry Becker from the 1994 Census database: in total 48,842 samples were taken. Using attributes from each person in the open data set, tests are conducted to determine trends that typically allow for a person to make over $50,000 in 1994. To put this value in perspective, $50,000 in 1994 has the purchasing power equivalent of roughly $76,000 in today's market (due to inflation).

TEXT PREPARATION:

The initial data set provided contains 48,842 instances of people from the 1994 United States census, where each person has fifteen total attributes: they are as follows:

1. Age
2. Working class
3. Fnlwgt
4. Education
5. Education Number
6. Marital status
7. Occupation
8. Family relationship
9. Race
10. Sex
11. Capital gain
12. Capital loss
13. Hours per Week
14. Native Country
15. Salary

Initial data processing involved using python to store all data within the provided text files. Upon studying of the attributes, it became apparent that some were irrelevant and thus omitted from the analysis process. In all, fnlwgt, education, family relationship, and occupation were removed. Fnlwgt is a census-specific weighting system, which was irrelevant. Also, education is a duplicate of education number; and family relationship is similar to marital status. Native country was simplified to either 'United States' or 'other'. Similarly, marital status was simplified to 'married' or 'other'. All other strings were converted to numbers in a boolean fashion (0 or 1) for convenience.

After all data was reformatted for analysis, values were converted into two separate excel-ready .csv files. The first of which contains all attributes. The second consists of salary, age, education, and hours per week. This separation is done to see how the model changes with fewer attributes. The results are in a way surprising: clearly more attributes increase the strength of predictions, but this data still constructs a strong representation given those four attributes. The final analysis is then done by using an R script to process these .csv files through CART.

RESULTS:

After running a script written in R to run CART through each of the two .csv files, two distinct trees are created. As shown in figure 1, the CART model using four attributes is able to split up the data fairly well. Astonishingly, an accuracy of 78% is achieved using only four total attributes including the goal. In all tests for this file, education was the most important split factor. This indicates that it is a very important factor in determining a person's salary.

In figure 2 however, this is not the case. It turns out that with all attributes present; the greatest split (occurring in every test) is marriage. Marital status of '0' indicates that the person is unmarried, whereas a marital status of '1' indicates that the person in married. One may suggest that people are forced into higher paying jobs that they wouldn't otherwise take because they feel the need to support their family instead of taking a lower paying job that they enjoy.

Correct and incorrect trials for each instance are reported in table 1. As outlined in the CART models, tests using the larger number of attributes lead to more correct trials. Table 2 further solidifies these presumptions. In every case, the tests conducted with more attributes performs better. Arguably the most important statistic presented in the statistical results section of table 2 is the sensitivity (or rate of true positives found) for the .csv files. It turns out that using only four attributes can build a model with a sensitivity of 86.5%; when all attributes are present, this sensitivity factor can be raised up over 90%.

CONCLUSION:

While these results are supportive of the fact that certain attributes of people do in fact play a factor in their salary, the data could be built to better represent the dissimilarities within certain categories. For example, country of origin consists of ~90% United States. If there were more variety in the data provided, CART could build a stronger, more reliable model.

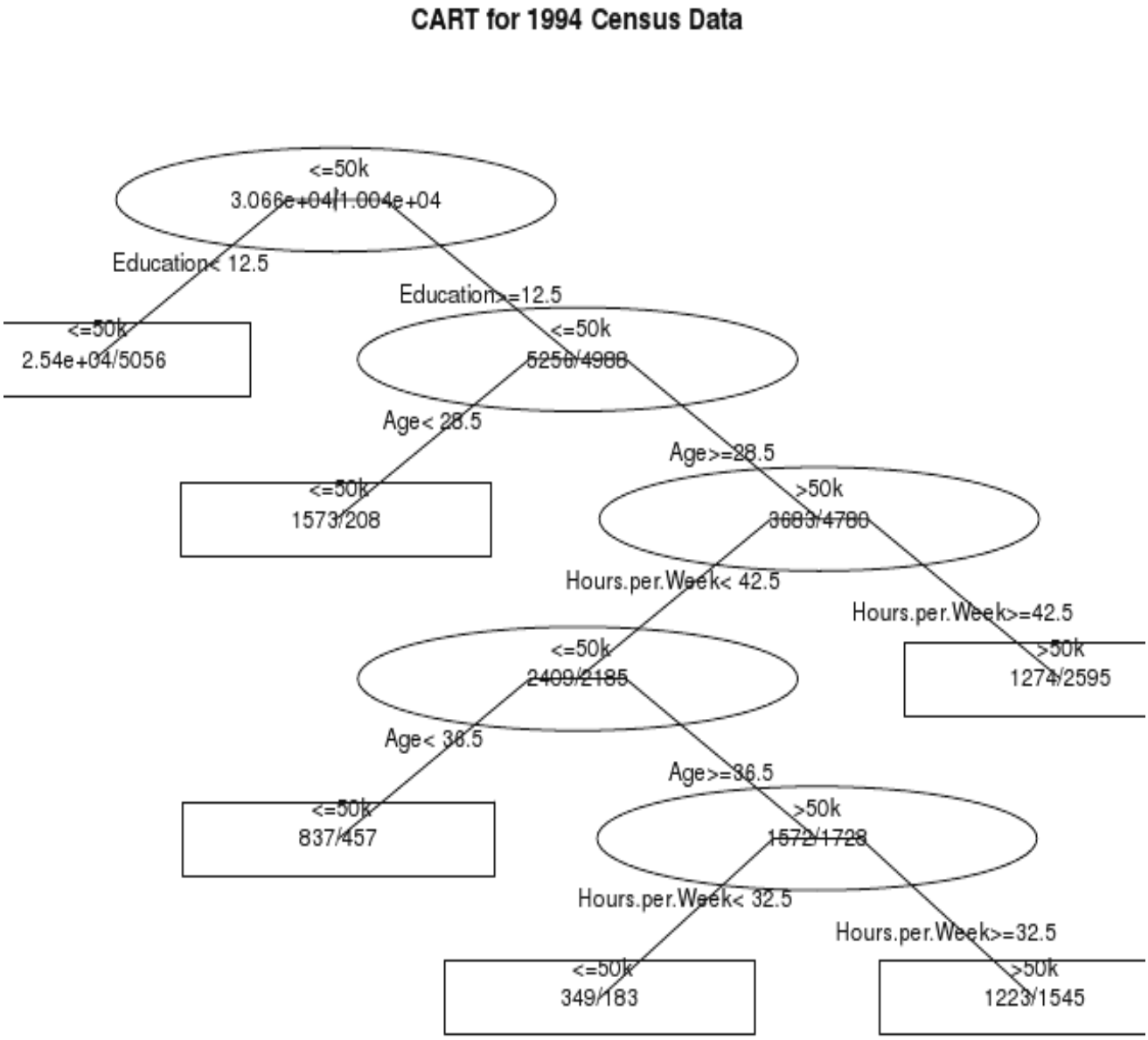

Figure 1: CART results from .csv containing education number, age, hours worked, and salary attributes only. Results were 78.4% accurate over ten iterations and using 10% as training data for each iteration. Education of 12 indicates a completed bachelor's degree. Left bucket indicates <50k, right bucket indicates >50k

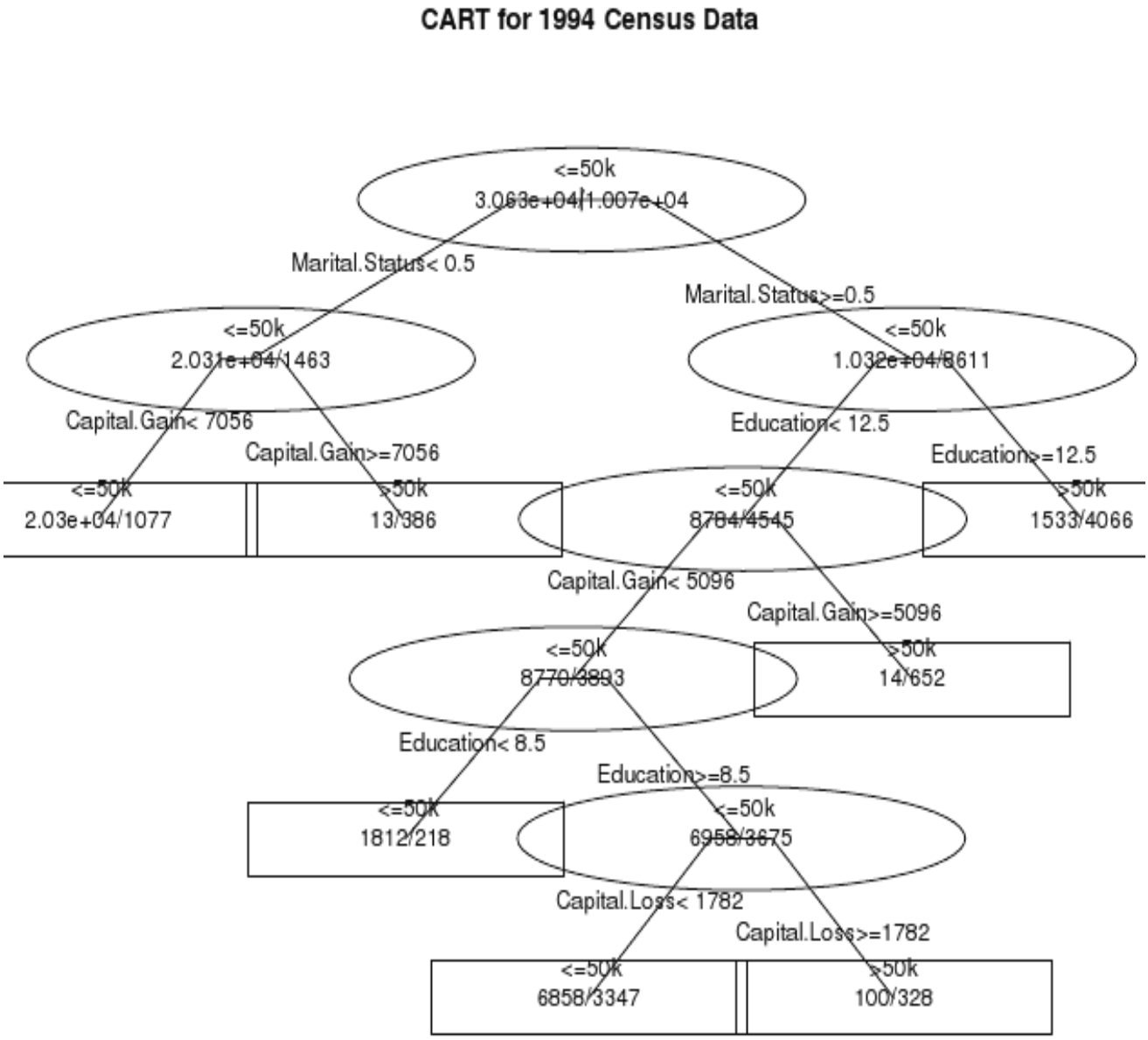

Figure 2: CART results from .csv containing all chosen attributes. Results were 84.6% accurate over ten iterations and using 10% as training data for each iteration. Marital status is either 1 (married) or 0 (not married). Also, education of 12 indicates a completed bachelor's degree. Left bucket indicates <50k, right bucket indicates >50k

		actual	
		> 50,000	**<= 50,000**
predicted	**> 50,000**	**tp** ("hit")	**fp** ("Type I Error")
	<= 50,000	**fn** ("hit")	**tn** ("Type I Error")

fp = Type I Error
fn = Type II Error

Results for four attributes:

		actual	
		> 50,000	**<= 50,000**
predicted	**> 50,000**	10306	3286
	<= 50,000	6079	21029

fp = Type I Error
fn = Type II Error

Results for all attributes:

		actual	
		> 50,000	**<= 50,000**
predicted	**> 50,000**	10969	2312
	<= 50,000	4672	22747

fp = Type I Error
fn = Type II Error

Table 1: Probability table showing relative findings for executions of four attributes and all attributes, respectively. Tests are conducted from the perspective of being greater than $50k.

	Predicting > 50k		Predicting <= 50k	
	Four attributes	**All attributes**	**Four attributes**	**All attributes**
Precision → tp/(tp+fp)	.758	.826	.776	.830
SENSITIVITY: true positive rate → tp/(tp+fn)	.629	.701	.865	.907
SPECIFICITY: true negative rate → tn/(tn+fp)	.865	.907	.629	.701
Accuracy→ (tp+tn)/(tp + tn + fp + fn)	.770	.830	.770	.830
False positive rate → fp/(fp+tn)	.135	.092	.371	.299

Table 2: Various statistical results pertaining to the information given in table 1. 10 trials for each case, training set each trial = 10%

Works Cited

Becker, B and Kohavi, R (1996). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml/datasets/Adult].