COSC757 Assignment 2

Matthew Quander
Dept. of Computer and Information Sciences
Towson University
Towson, MD, USA
mquand1@students.towson.edu

Abstract—**This paper will explore the Census Income Data Set from the UCI Machine Learning Repository, which consists of data on individuals in the 1994 US Census report. Described as a multivariate data set, a target variable of income, >50k or <= 50k, could possibly be predicted. Given the other attributes, certain classification algorithms may be able to achieve this. This paper will explore the data and apply the appropriate algorithms to solve the question: to what extent do the provided variables determine whether someone makes more than, or less than or equal to $50k?**

***Keywords*–-classification, multivariate, income, census data**

I. INTRODUCTION

This paper will provide an in-depth analysis of the Census Income Data Set provided by the UCI Machine Learning Repository. This data set represents a sample of the US workforce in 1994, to include individuals' attributes and whether or not they earn more than, or less than or equal to $50k per year [1]. First I will describe the data set, perform exploratory data analysis to obtain some initial insights, discuss different classification algorithms, perform an experiment, and analyze the results. The objective of the analysis is to determine what attributes contribute to someone earning more than $50k. Much of the analysis for this will be completed with statistical and classification tools in R. Having earned my bachelor's degree in economics, I found this data set to be of particular interest.

A. Data Set Description

The Census Income Data Set was provided by the UCI Machine Learning Repository, and consists of data extracted from the 1994 US Census database by Barry Becker [1]. The records were already preprocessed and were pulled from the database based on the following conditions being true: age > 16, adjusted gross income > 100, final weight > 1, and hours per week > 0. It contains 48,842 instances, split into a training set of 32,561 and a test set of 16,281, with 15 attributes [1]. The attributes include age, workclass (industry employed), fnlwgt (final weight, estimated amount in target population), education (highest level attained), education-num (total years), marital-status, occupation, relationship (individual's role in the family), race, sex, capital-gain, capital-loss, hours-per-week, native-country, and income. The data set is multivariate since it has many variables, and its attributes' are characterized as categorical and integer values [1]. The provided information on individuals represented in the data set would allow for the prediction of the target variable, whether income is >50 or <=50k. As appropriate for this assignment, the associated task for this data set is classification. The following question follows, is it possible to predict whether someone makes >50k or <=50k based on the other 14 given attributes?

II. METHODOLOGY

A. Exploratory Data Analysis

The data set is made up of 15 attributes, one of which is the target variable. To better understand the variables in the data set and to obtain some initial insights, some preliminary analysis was performed. As displayed in Fig. 1, the majority of the individuals representing instances in the training set were between the age of 20 and 50. As an acceptable age range for the workforce, this variable suggests some validity in the records. Similarly, the descriptive statistics for the age variable show that the 1st and 3rd quantiles closely align with this age range.

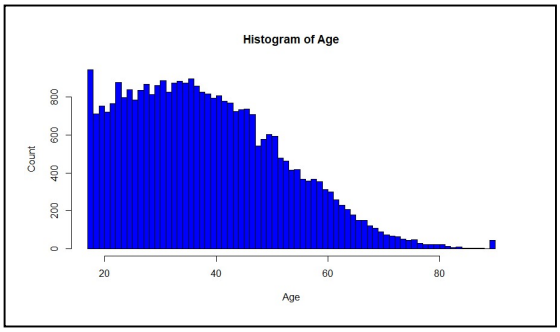

Fig. 1. Histogram of Age.

TABLE I. AGE

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
17	28	37	38.58	48	90

A histogram counting the amount of years of education each instance has was created. This revealed that a large volume of the individuals had either 9 or 10 years of education. This is also viewable in the summary statistics of Table II.

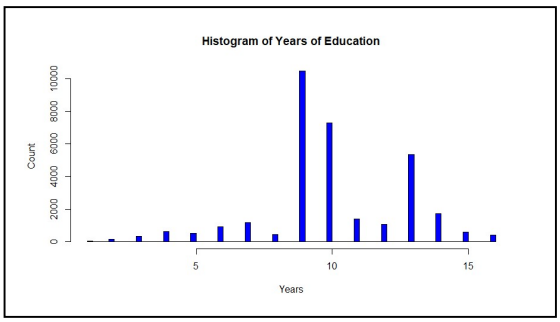

Fig. 2. Histogram of Years of Education.

TABLE II. YEARS OF EDUCATION

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1	9	10	10.08	12	16

It can also be discovered in the data set that men make up the majority. By summing the amount of male instances and dividing by the total amount of records, the proportion of men in the training set is 0.6692055 (~67%). Exploring further, I graphed several plots to observe any relationships between the attributes and the target variable. As shown in Fig. 3, the majority of people making >50k are between the ages of 30 and 50. Whereas, the majority of those making <=50k are between 25 and 50, a slightly wider range.

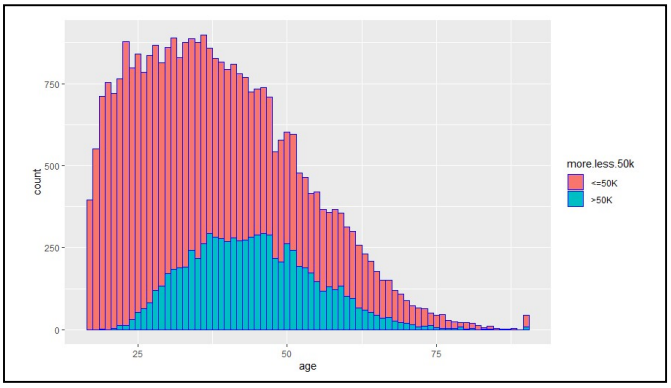

Fig. 3. ggplot of age and income.

Upon plotting the education years, it also shows that those with 9 or 10 years of education are highly represented in the data set like Fig. 2 above. However, with the target variable overlayed, it appears from Fig. 4 that individuals with 13 or more years of education have a higher portion of high income earners (>50k). This should be a notable fact when applying the classification algorithms.

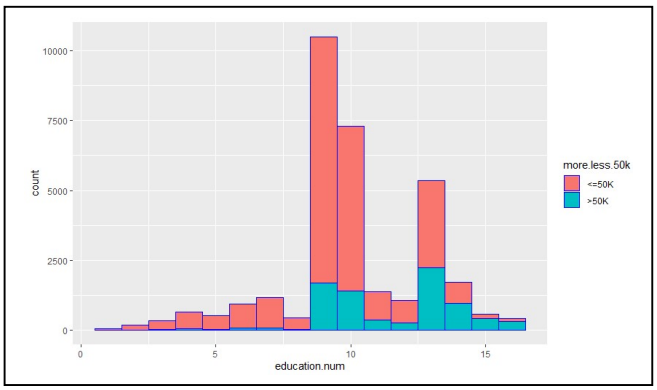

Fig. 4. ggplot of education.num and income.

In the process of exploratory data analysis, I did further research into some of the variables. The fnlwgt variable is the a continuous variable that calculates a weight for each instance. In an effort to simplify the analysis, this variable will be left out. Similarly, education level will be left out since years of education is included and can provide more information as a continuous variable. Finally, the total proportion of instances where income was greater than 50k was 0.2408096 (~24.08%).

B. Algorithms

Upon downloading the file from the UCI Machine Learning Repository, I observed that the data set was already preprocessed and cleaned as stated on the website [1]. It was split into a training set and a testing set, 2/3 and 1/3 respectively.

The first algorithm I will be using is the ID3 Information Gain. As a commonly used data structure, decision trees provide a way for organizing data based on conditions. By strategically designing and placing those conditions within the tree, much information can be derived from a datum as it is placed into a leaf. Some of the benefits of using a decision tree are that it can be applied to any data type, and results are comprehensible in understanding the question posed [2].

Decision Trees are built using a recursive, greedy algorithm. By implementing the divide-and-conquer method, a data set can be broken down based on the successive conditions in the higher level nodes [2]. The algorithm takes, as input, a subset of the data and can calculate all of the possible partitions that could be made [2]. By placing the conditions in nodes higher in the tree, greater importance can be given to those conditions as they make partitions, e.g. if age is the greatest determining factor it can be placed in the root. When called recursively, the amount of successive conditions tested reduces based on the results of previous conditions. The recursive processing then stops when the datum is classified in a leaf node [2].

One of the major challenges in building an effective decision tree is determining the conditions in testing a data point's attributes. If one attribute can provide the greatest amount of

information, it reduces the information needed from other attributes in order to classify the datum [2]. Where data sets can vary, entropy reduction leads to better classification. The ID3 algorithm proceeds by establishing the probability that an instance in the data set belongs to one of the classes. The entropy is calculated by taking the summation of that probability multiplied by the base 2 log of that probability, then multiplying that sum by negative 1. Once the data set is split based on an attribute, the summation is taken of the absolute value of an instance divided by the absolute value of the data set, and multiplied by the entropy of that instance. Finally, we subtract that value from the entropy to obtain the Information Gain.

$$\text{Info}(D) = -\sum p_i \log_2(p_i)$$

$$\text{Info}_A(D) = \sum |D_j|/|D| * \text{Info}(D_j)$$

$$\text{Gain}(A) = \text{Info}(D) - \text{Info}_A(D)$$

The second classification algorithm I will use is the Support Vector Machine method, which encompasses all of the given attributes. The SVM method partitions the data set in an effort to classify each subset. By using a nonlinear mapping, the data set can be enhanced into a higher dimension. A linear optimal separating hyperplane is then sought to act as a decision boundary [2]. It's goal is to use a function to separate the available instances into two classes and successfully generalizes on other instances it has not seen, i.e. the test set [2]. While many linear classifiers can be derived from a data set, SVM attempt to find the hyperplane that has the maximum margin, i.e. the greatest distance between itself and the nearest data points [2]. It does so by using support vectors (training instances) and margins (given by the support vectors).

The SVM searches for a maximum marginal hyperplane using weight vector $W = \{w_1, w_2, \ldots, w_n\}$ and a scalar b, and is denoted $W*X + b = 0$. Thus, in a 2-dimensional setting, points will fall above or below the hyperplane $w_0 + w_1x_1 + w_2x_2$. When a training tuple falls on a hyperplane that's either greater than or equal to 1, or less than or equal to -1 (the sides of the margin), it is considered a support vector. The margin can be calculated by summing the distance between the upper bound ($W*X + b = 1$) and lower bound ($W*X + b = -1$) hyperplanes.

Lastly, I will use the Naïve Bayes algorithm to classify the data set. As a probabilistic classifier, the Naïve Bayes algorithm places great influence on the independence between features. Derived from the Bayes Theorem, it attempts to determine the posterior probability of an event (<= 50k or >50k in this case) among all possible events. Naïve Bayes uses a training data set's attributes as predictors, where each instance is represented by a vector. The possible outcomes are a set of categorial variables. In order to determine the category, the Bayes formula is applied: the probability of a particular category C, given the attributes X is equal to the probability of the attributes X given a particular category C, multiplied by the probability of that category C, divided by the probability of the attributes X. The denominator is usually discarded since the probability of the attributes X for any class is a constant.

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)}$$

Since the attributes are independent of each other in Naïve Bayes, the above formula can be rewritten as a product series, from 1 to n, of the probability of attribute x_k given class C_i. In using the numeric columns of the `adult` data set, the probability of each attribute given each class would be calculated separately. All of the results would then be multiplied together for each class event, <=50k and >50k. Each of these results would then be multiplied by the total probability of the event occurring within the data set. Whichever value is greater would be the probability of that class given attributes X.

C. Experiment Design

I will experiment using the cross-validation holdout method, with a training and testing set. Upon downloading the data set, it was already split up into a training set of 32,561 records and a testing set of 16,281 records. This split represents an accurate 2/3 partition for the model construction and 1/3 for the accuracy estimation.

Using a Two-Sample T-Test, I derived the following p-values for the numeric attributes of interest: age: 1.849114; education years: 0.6851209; capital gain: 1.044202; capital loss: 1.123246; hours per week: 0.6753029. All of these p-values are large, which indicates that there is no evidence that the means of the attributes in the training set and test set differ significantly. Similarly, in performing a Two-Sample Z-Test for the non-continuous variables of interest, the p-value for the male proportion was 0.6310134, for private sector workers it was 0.05467777, and for workers native to the US it was 0.1064114. Although the p-value for private sector workers was the closest to 0.05, they were all over that threshold. Therefore, this partition is valid.

By applying the ID3 algorithm to the Census Income Data Set, I expect that certain attributes will have more weight than others in classifying individuals' income. From the EDA process, some information can be inferred that I believe will also be revealed through a decision tree. In using R's `rpart` package, I will build several decision trees with different variations of predictor variables in the formula. This will likely uncover the variable to reduce entropy as much as possible.

Secondly, by use R's `svm` function in the E1071 package, I will build a Support Vector Machine to classify the data. With the income attribute being the target variable, I will test how much the other attributes contribute to an individual's income. From that SVM model built, I can attempt to predict the income variable of the test set using R's `predict` function on the numerical columns. This can be evaluated using the `table` function to display a confusion matrix.

Lastly I will use R's E1071 package for the Naïve Bayes method. By first using the `pairs` function, the attributes can be plotted against each other. This will help to determine any visual correlation between the attributes and the target variable. A Naïve Bayes model can then be built by the `naiveBayes` function. This model can then be passed to the `predict` function to develop another confusion matrix.

III. Results

Using the ID3 algorithm, R's `rpart` function successfully built a decision tree. After testing it with many different variations of variables, I found that the years of education attribute was the best split. Splitting on this attribute resulted in the highest information gain, where 75% of the data set had less than 13 years of education and made <=50k. Further splitting the data, those who were 30 years or older and male were 12% likely to make 50k or more. Similarly, upon running the same algorithm on the test data set, the results were practically the same. There was only a 1% difference in the final gender condition, where males made more than 50k. These results can be observed in the below graph and summary.

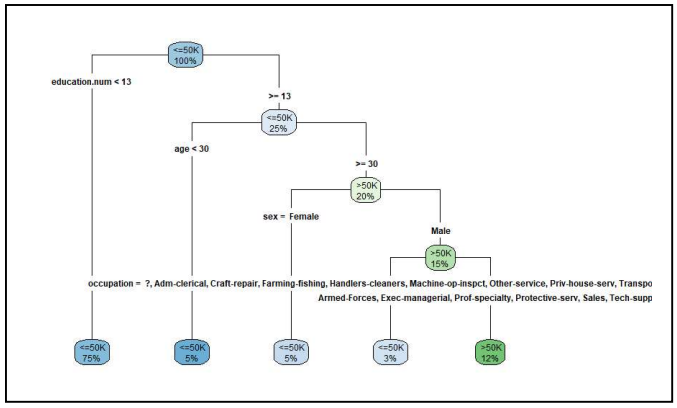

Fig. 5. ID3 Decision Tree

```
n= 32561

node), split, n, loss, yval, (yprob)
      * denotes terminal node

1) root 32561 7841  <=50K (0.7591904 0.2408096)
   2) education.num< 12.5 24494 3932  <=50K
(0.8394709 0.1605291) *
   3) education.num>=12.5 8067 3909  <=50K
(0.5154332 0.4845668)
     6) age< 29.5 1617  232  <=50K (0.8565244
0.1434756) *
     7) age>=29.5 6450 2773  >50K (0.4299225
0.5700775)
      14) sex= Female 1612  540  <=50K
(0.6650124 0.3349876) *
      15) sex= Male 4838 1701  >50K (0.3515916
0.6484084)
        30) occupation= ?, Adm-clerical, Craft-
repair, Farming-fishing, Handlers-cleaners,
Machine-op-inspct, Other-service, Priv-house-
serv, Transport-moving 856  349  <=50K
(0.5922897 0.4077103) *
        31) occupation= Armed-Forces, Exec-
managerial, Prof-specialty, Protective-serv,
Sales, Tech-support 3982 1194  >50K (0.2998493
0.7001507) *
```

TABLE III. Summary of Rpart

The results of the Support Vector Machine algorithm were fairly different. Once the model was built based on the training set's numerical data, 11,363 support vectors were created as well as the 2 levels of <=50k and >50k. After plotting this model with different variations of variables, I found that the plot between age and education.num was the most revealing. It shows a clear shift for individuals with 9 or more years of education into the >50k class.

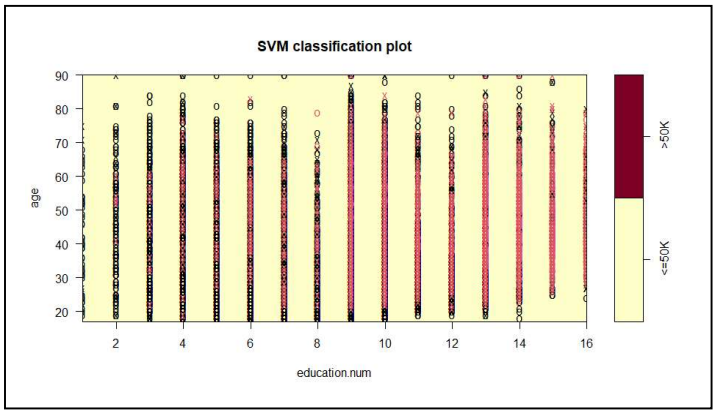

Fig. 6. SVM Plot of age, education.num, and income

When the prediction function was run on the test set's numerical data to predict the income, the below confusion matrix was returned.

		true	
pred		<=50K.	>50K.
	<=50K	12008	2416
	>50K	427	1430

TABLE IV. SVM CONFUSION MATRIX

This SVM model predicted 12,008 true positives for those earning <=50k, which is greater than half of the test data set. Likewise, there were only 427 false negatives for predicting those who earn <=50k. However, for those making >50k the amount of true negatives were 1,430 and the amount of false positives were 2,416. This leads to a sensitivity of 12008/(12008+427) = ~0.9657 and a specificity of 1430/(1430+2416) = ~0.3718, therefore the accuracy is 0.9657*12435/(12435+3846) + 0.3718*3846/(12435+3846) = 0.737576 + 0.087829 = 0.825405.

	<=50K.	>50K.
<=50K	11795	2759
>50K	640	1087

TABLE V. NAÏVE BAYES CONFUSION MATRIX

The Naïve Bayes classifier returned a confusion matrix with nearly the same predictions. There were 11,795 true positives for those making <=50k, a few thousand less than what the SVM algorithm predicted. However, there were more false negatives for those earning <=50k, 640. The number of false positives for those making >50k were 2,759 and the number of true negatives was 1,087. This leads to a sensitivity of 11795/(11795+640) = ~0.9485 and a specificity of 1087/(1087+2759) = ~0.2826, therefore the accuracy is 0.9485*12435/(12435+3846) + 0.2826*3846/(12435+3846) = 0.724439 + 0.066758 = 0.791197.

In evaluating the above confusion matrices of the two algorithms, the Support Vector Machine appears to have returned slightly more accurate results. In calculating the accuracies for the Support Vector Machine and the Naïve Bayes method, there was only a 0.034208 difference. Since both of their accuracies are relatively high, one should feel confident in using either to predict the income of an individual in this data set.

IV. CONCLUSION

In conclusion, the Census Data Set proved to be an effective data set in this classification experiment. By using the Information Gain ID3 algorithm, I found that the number of years of education an individual had was the greatest determining factor in whether they earn more, or less than or equal to 50k, which answers the question posed. This I expected due to the labor market at that time, where education had a greater impact on income as compared with today. In exploring R's classification tools, I found the Support Vector Machine and Naïve Bayes method to be effective. Both of these algorithms resulted in a fairly accurate prediction of the target variable, whether income was more than 50k or less than or equal to 50k. Despite these accuracies, there are some limitations that were discovered during the experiment. For example, the `predict` method returned errors when running it with the SVM object for all attributes (including the non-numeric). In order to do a more fair comparison, both models were built with the numeric attributes. That was the greatest difficulty in the experiment, however, by applying the algorithms to the appropriate attributes a coherent result was obtained.

REFERENCES

[1] R. Kohavi, B. Becker, *UCI Machine Learning Repository: Census Income Data Set*, Center for Machine Learning and Intelligent Systems, University of California, Irvine, May 1996. Accessed on: Apr. 18, 2021. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Census+Income

[2] "Data Mining in R/Classification." Data Mining in R/Classification – Wikibooks, open books for an open world. https://en.wikibooks.org/wiki/Data_Mining_Algorithms_In_R/Classification. (accessed Apr. 20, 2021).