An Assessment of the Census

From: Tyusha Sarawagi, Sarah Wright
To: Dr. Christina Horr

Topic: Predictions of Low Income Families

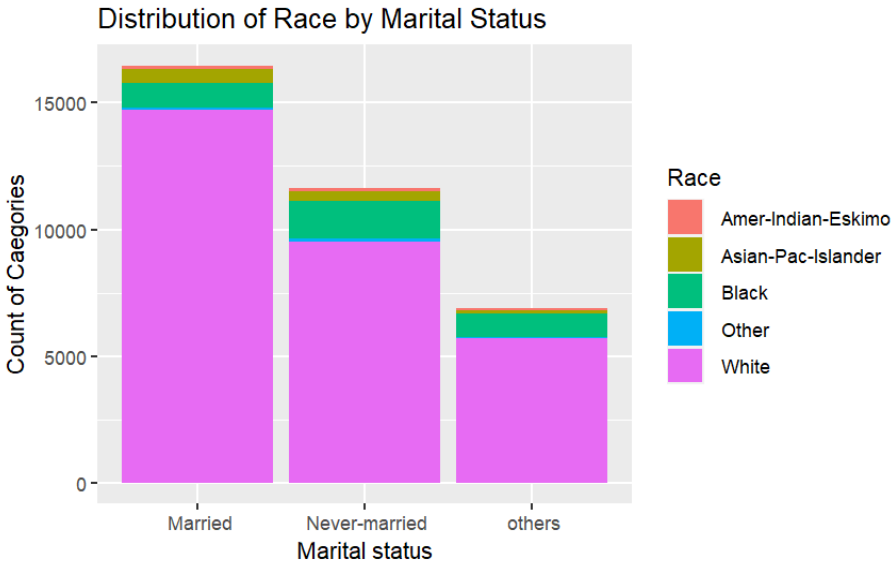
I. Introduction

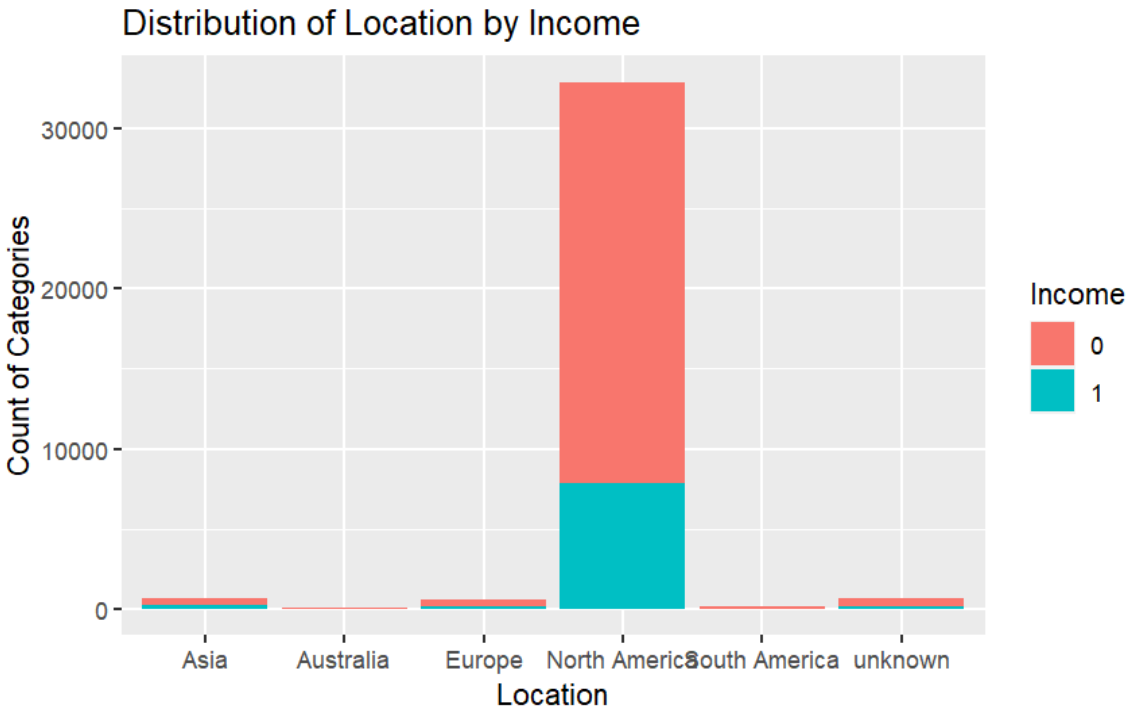
In 2020, the US census recorded the information of over 48,000 individuals and consolidated the information into a data set that will be using included in our report. We were tasked with creating a model to predict whether someone has an annual income of over than $50,000.

We performed basic data manipulation on the variables in the dataset to analyze the dataset and gain some insights. The dataset mostly included categorical variables, so we decided to look at each unique category. We found out that there were several overlapping variables and decided to modify them to reduce the number of categories, thus reducing complication.

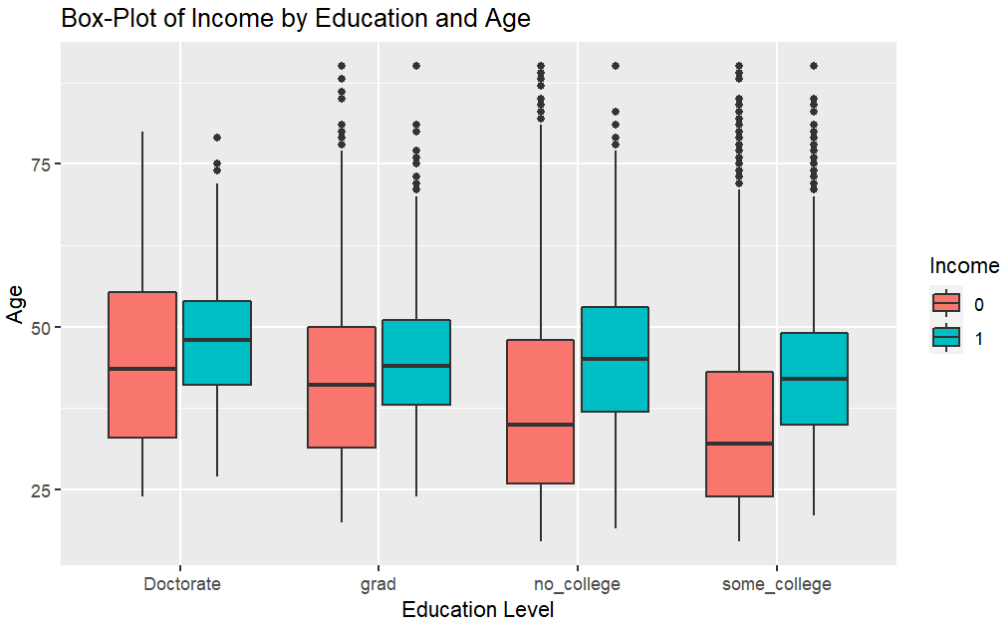
Initially, the income variable had four categories: <=50K, >50K, <=50K., and >50K. We changed the income variable into two categories; 0 if <=50K. or <=50K and 1 if >50K. or >50K. We also decided to reduce the number of categories for marital status. Initially, it had seven categories: Married-civ-spouse, Married-spouse-absent, Married-AF-spouse, Divorced, Separated, Widowed, and Never Married. We decided to have three categories: Married, Never Married, and others. Furthermore, education had sixteen categories, and we re-coded it to have four categories: no-college, some-college, graduate, and doctorate. Lastly, we re-coded the native country variable into five continents and unknown according to their location.

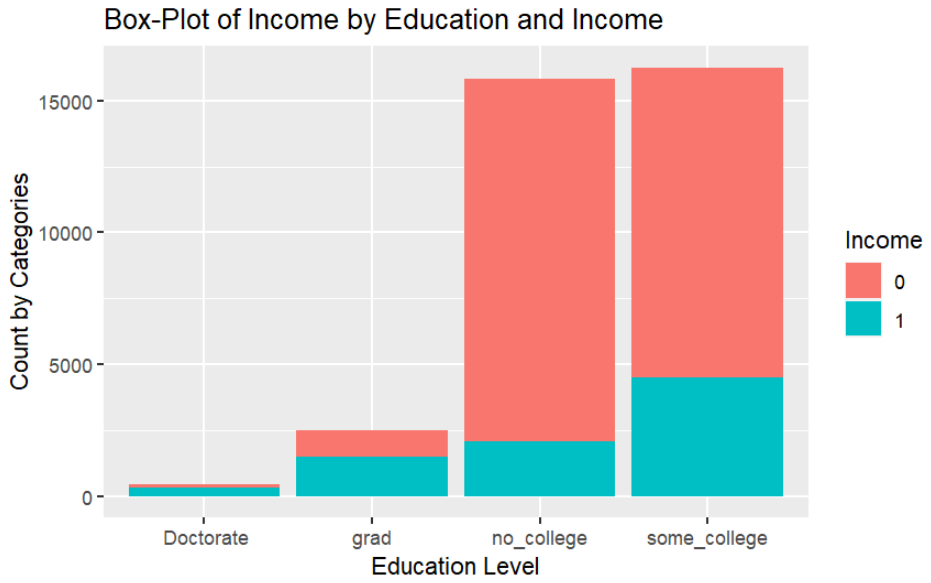
II. Exploratory Data Analysis



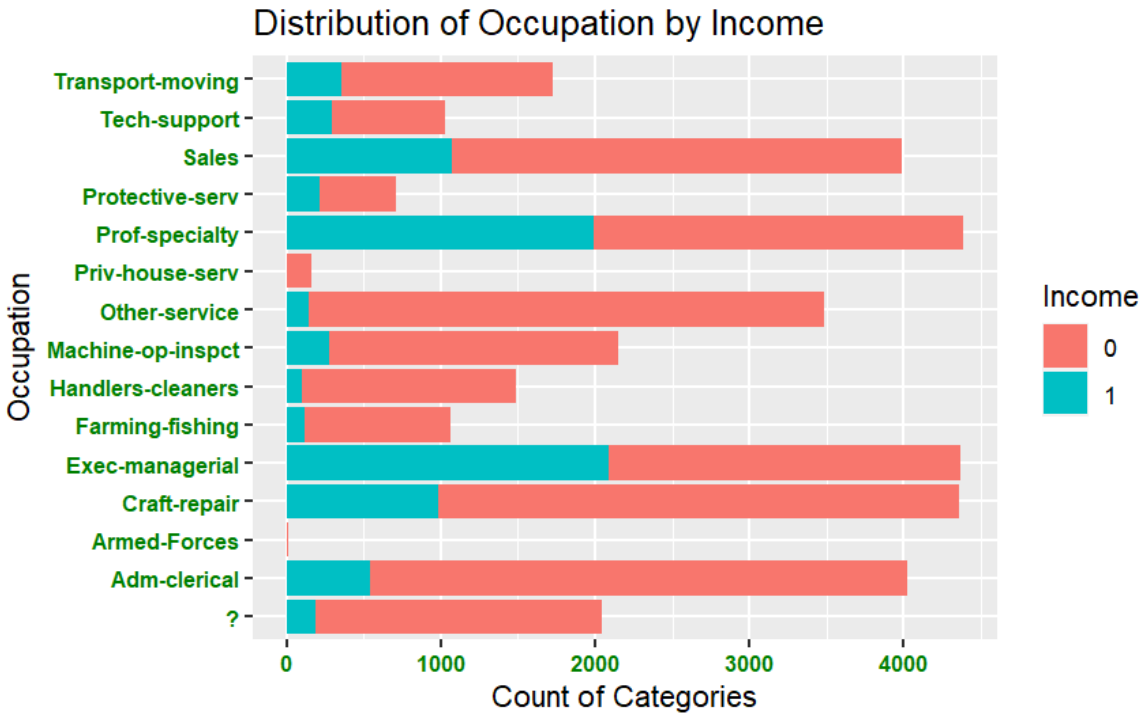


As we can see from the graphs above, it is clear that people in North America dominate the data, and it is not representative of the entire population. Therefore, we can only make conclusions specific to North Americans from the dataset. We can see from the second graph that most North Americans are the ones who earn more than $50,000.





From the first boxplot, we can see that for all categories of education level, older people earn more than $50,000. The second graph shows that most people from the data set have either a high school or undergraduate degree. People who have some college experience are the category where more people have income higher than $50,000.



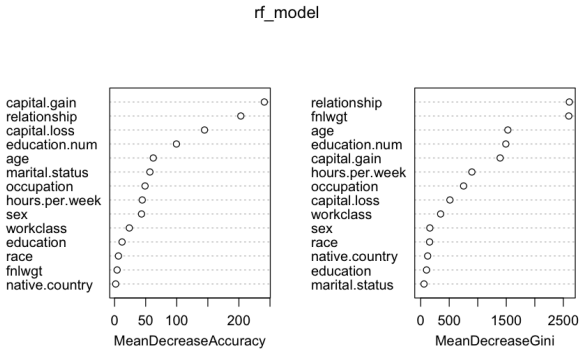
From the graph above, we can see that most people in the executive-managerial position and Prof-speciality earn more than $50,000 implying these jobs can be considered higher income. People in the Sales, Craft-repair and Adm-clerical positions have an intermediate number of people earning more than $50,000; thus, these jobs can be considered medium income. The other jobs from the graph can be regarded as lower-income jobs.

III. Model Building

In order to create an accurate prediction of the individuals included in the census, we decided to use the Random Forest Model with Bootstrap Aggregation in an attempt to get the most accurate predictions. While a simple logistic regression or classification trees would have worked for this problem, the Random Forest method gives us more power over our predictions, allowing them to be more accurate using a bootstrapping method.

Following the restructuring of the variables, we began the process of creating the Random Forest Method. To begin, we used the Random Forest function to build the Random Forest. We set the number of factors to consider at each step equal to the number of variables which in the case of the Census Dataset is 14 factors, and the number of trees was arbitrarily chosen as 250. After running the random forest function with this information we concluded that our model using the Random Forest Method would have a misclassification rate of 14.4%.

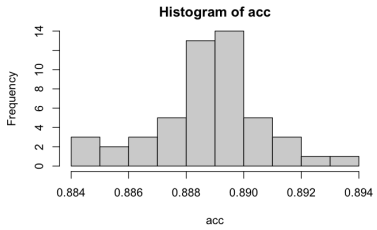
Additionally, we were able to see what variables are of importance and could be used to predict the accuracy of the model [Figure rf_model]. From this we found that the most important variables are native.country, education, race, work class, and sex.



At this point, we had concluded that if we continued to model with the Random Forest method, we would have a misclassification error of 14.4% and our primary predictors would be native.country and education. Logically thinking these are good indicators, thus we decided to proceed with the Random Forest Method.

The next step we conducted was to create a for loop that splits the training and the testing data and runs 50 different times and computes the accuracy of each individual data point. However, we encountered an error! For the model to run correctly, both data frames need the same number

of columns. To resolve this problem, we created a new vector using the rpart and the predict function. In fact, this is the income vector that was later used to produce the results in the CSV file include with this report. We used the rpart function to create a series of splitting rules that would divide the observations in the data set into a regression tree using income as our predictor. We then took the results of the rpart function and used it for a prediction function using class tree and the testing data with a type = class. The predict function gives us a vector that predicts whether the individuals recorded in the census earn an income greater than $50,000. We then appended this vector to the end of the test data and used it in the for loop. In the for loop, we computed the accuracy of each point and looped it 50 times because of randomness and included the in the data, we ended up with an accuracy between 0.884 and 0.894 [seen in Histogram of acc].



The last step we used to assess the accuracy of the vectors and the model is a cross validation. During this process, we followed similar steps to the for loop, only this time we tune our model by altering the number of trees in the random forest. Using 250 as the number of trees passed to the random forest, and 14 once again as mtry, we were able to tune our model to an accuracy of 0.853.

Because our previous model was better with accuracy between 0.884 and 0.894, we used that as the final vector for submission.