Predicting Income Using 1996 U.S. Census Data

Michael Long

Predicting Income Using 1996 U.S. Census Data

Overview

The purpose of this exercise was to create a model that predicts whether individuals, based on the census variables provided, make over $50,000 per year. An eXtreme Gradient Boosted (XGBoost) model was created that had a final Area Under the ROC Curve (AUC) score of 0.913. Additionally, it was found that being married to a civilian spouse was the most important variable in regard to model accuracy.

Methodology

Data Used

The data provided for this exercise is a sample of 48,842 individuals from the 1996 U.S. Census. Originally received as a SQLite database, the data was flattened to a single CSV with variables describing their native country, age, class of occupation, the highest level of education received, the current education level, marital status, occupation, relationship, race, sex, capital gains, capital losses, number of hours worked per week, and whether the individual made over $50,000 a year (target variable).

Data Preparation

The data was split into a training set (70%), a validation set (20%), and a test set (10%) before performing any exploratory data analysis (EDA). Once the data was split, all EDA was performed on the training set.

The only abnormal values found were values of $99,999 for capital gains. For this exercise, those values were considered unknown/missing and were imputed with the median value of the variable.

Due to quasi-complete separation for class of occupation and no observations for the "Holand-Netherlands" level of native country, new observations were generated using the median values for continuous variables and the mode for categorical variables. Each imputed variable also had a corresponding flag variable created.

Because of perfect multi-collinearity, the decision was made to drop the variable describing the current education level and keep the variable for the highest level of education received.

Analysis

The initial XGBoost model was created with the full set of variables and then tuned based on four parameters: the number of trees, the max depth of each tree, the subsample, and the learning rate.

Variable selection was performed on this model by adding a random variable to the training data and seeing how the random variable performed compared to the existing variables. The random variable outperformed all other variables besides marital status, capital gains, education level, age, capital losses, and number of hours worked per week. Variable performance was measured by gain, the average increase in accuracy based on splitting a branch with that variable. The remaining variables in the model are shown in order of importance in Appendix A, Figure 2, page 3.

After variable selection, the training and validation datasets were combined into one single dataset. This combined dataset was then used to retune the final model with an updated set of parameters narrowed down from the previous model.

Results

The final AUC score on the test dataset was 0.913. The relationship between the most important variable, being married to a civilian, and the target variable is show below in Figure 1.

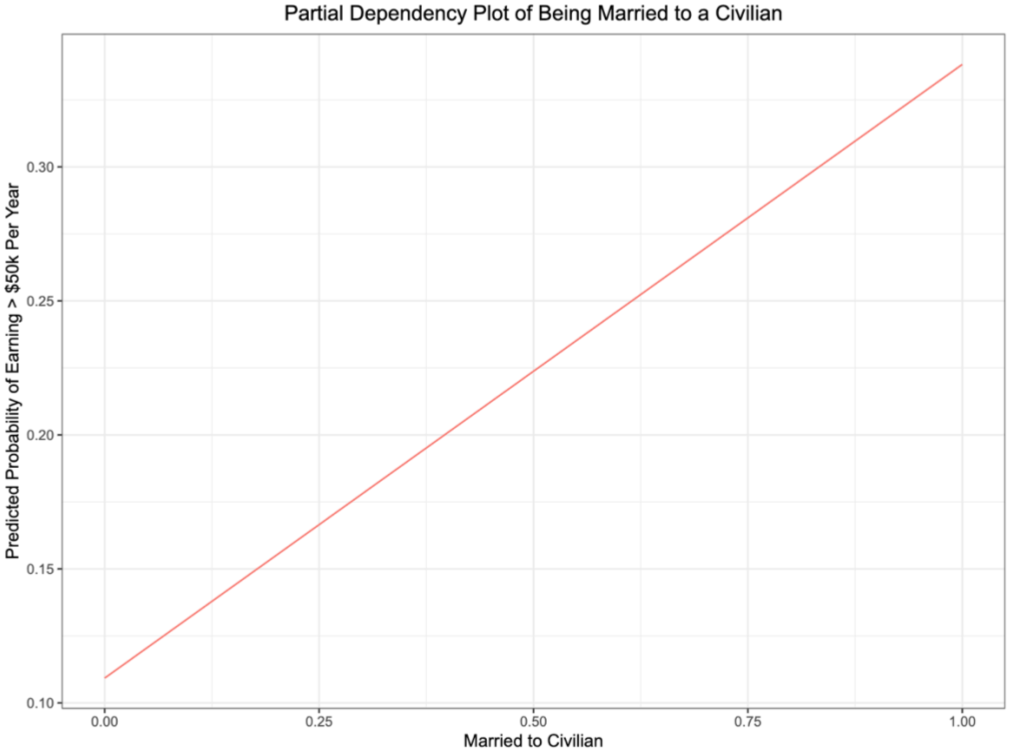

Figure 1: Partial Dependency Plot of Being Married to a Civilian

Although the plot shows a line from the bottom left to the top right, the variable is a binary variable so only the values at 0 and 1 along the x-axis should be considered. On average, if an individual is not married to a civilian, they have approximately a 0.11 predicted probability of making over $50,000 per year. On the other hand, if they are married to a civilian, the average predicted probability jumps to around 0.327.

Appendix

Appendix A

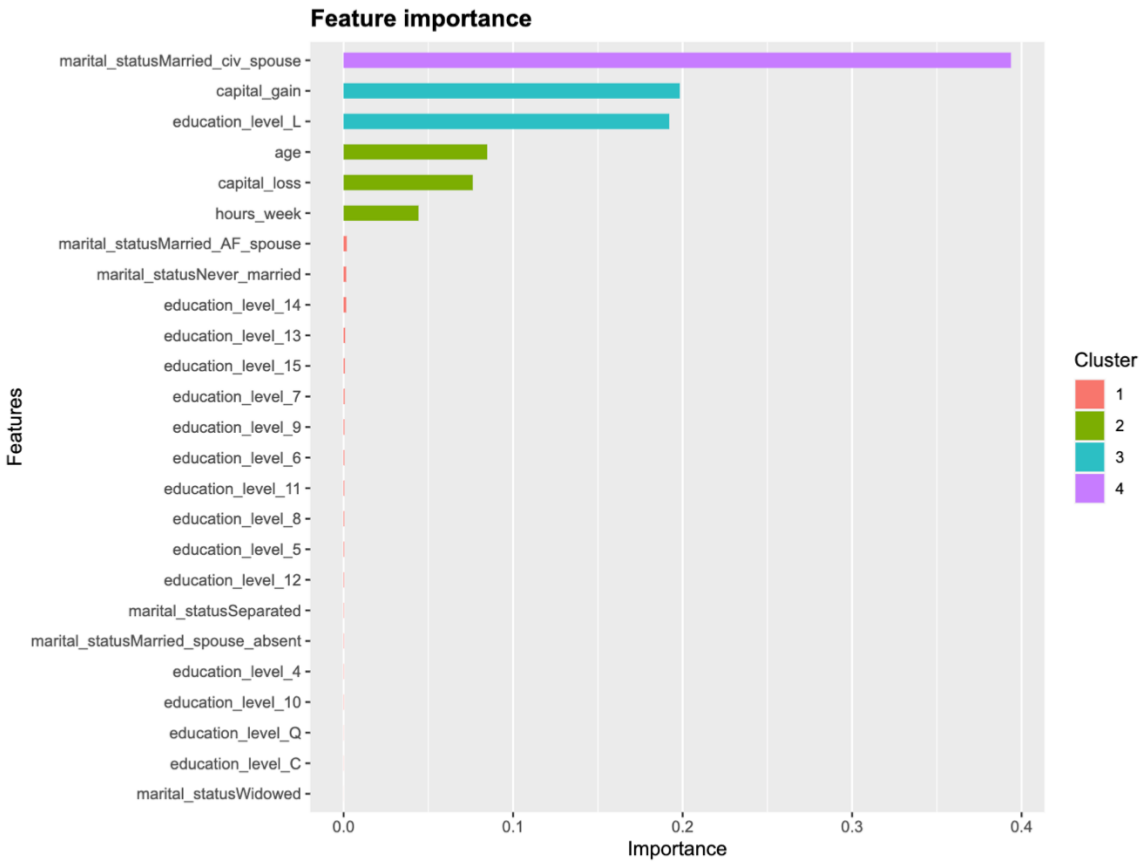

Figure 2: Variables Listed in Order of Importance