Predicting Income Using 1996 Census Data

Reported by Jane Williford

This report details the creation of a random forest model developed to predict the probability of individuals earning an annual income exceeding $50,000 (the target variable). The model achieved an Area Under the ROC Curve (AUC) of 0.897 on the evaluation dataset. Additionally, the report delves into the key findings from analyzing the relationship between the most important variable of the model, capital gain, and the target variable.

Methodology

Data Used

The analysis is based on a sample of 48,842 individuals from the 1996 Census. The information available includes age, working class, education level, marital status, relationship status, occupation, race, sex, capital gain, capital loss, hours per week worked, country of origin, and the target variable.

Data Preparation for Random Forest Modeling

Missing values were first handled by grouping known missing values into a ‘missing’ category for categorical variables and replacing 99 and 99999 values with medians for hours worked per week and capital gain variables. Two missing flag variables were subsequently created.

The variable for the country of origin was then grouped into four categories: ‘USA’, ‘High Proportion of Income Over 50K (>=0.20)’, ‘Low Proportion of Income Over 50K (<0.20)’, and ‘Missing’.

Lastly, the data was split into a training set (70%), a validation set (20%), and a test set (10%).

Random Forest Model Development

Once data was postured for the random forest, three models were built, evaluating the success of using 1000 decision trees, 500 decision trees, and 300 decision trees, with three variables randomly selected at each split in the decision tree building process.

Model Selection

The AUC of each model was calculated on the validation data set. The third model had the highest AUC value and was thus selected. To have added confidence that all the variables are helping the model, a random variable was created to see if it performed better than any of the existing predictor variables. This resulted in the removal of the hours per week worked missing flag variable. Model three was then re-run with the select variable removed. This model had a higher AUC and was chosen as the final model.

Model Evaluation

The AUC of the final model was calculated on the hold-out test data set. Subsequently, the most important variables in the final model were determined using the Mean Decrease in Accuracy (MDA), measuring the average decrease in the model’s accuracy when the predictor variable’s effect is removed.

Results

Final Model Evaluation

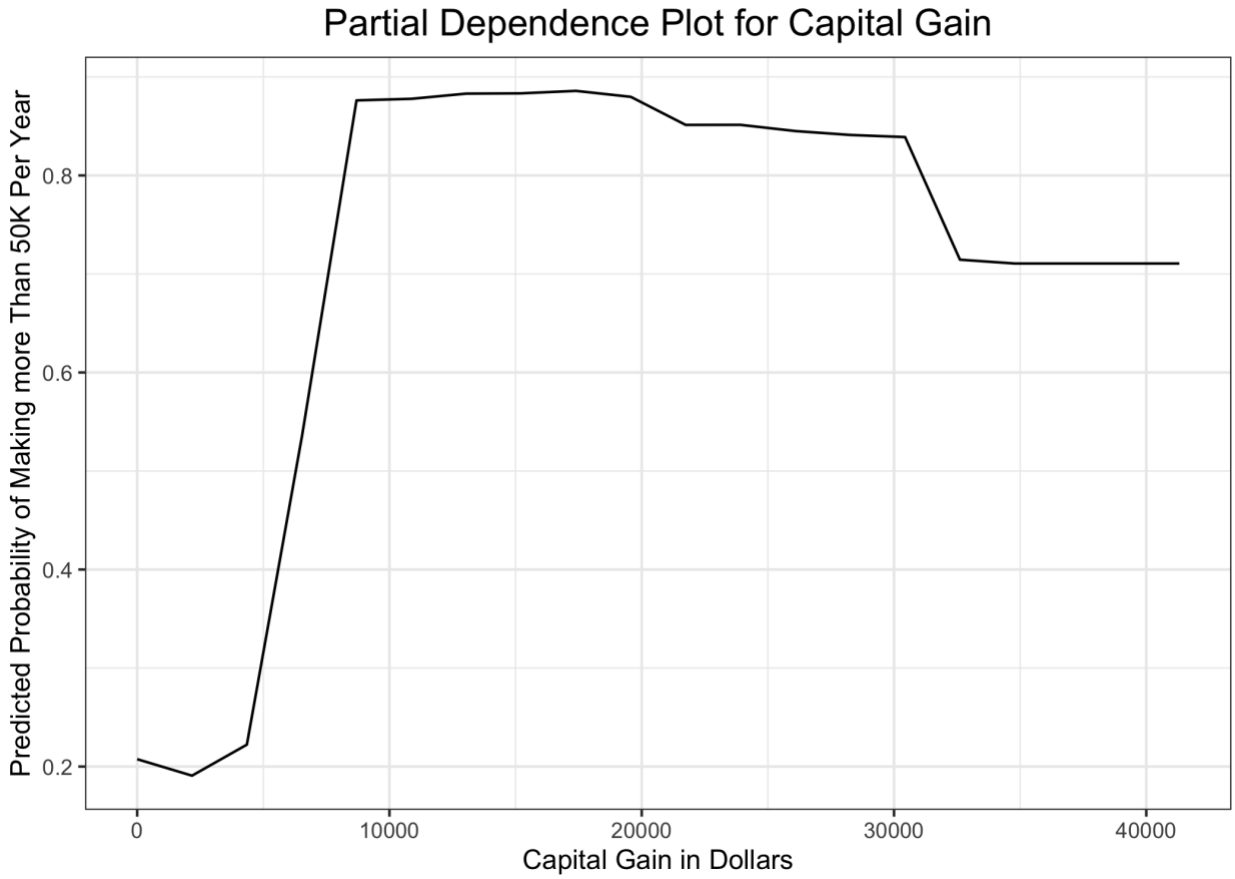
The final model developed had an AUC value of 0.897 on the test data set. A poor model has an AUC value close to 0.5, and an excellent model has an AUC value close to 1. Therefore, this model does a reasonable job of distinguishing between those with an income exceeding $50,000 and those who do not. However, other types of predictive models may want to be developed before proceeding with using this one to see if they outperform the results of this one.

Important Variables

Capital gain had the highest MDA value of 174, followed by education level (124) and occupation (119). The model suffered the most when these variables were removed.

Analyzing Capital Gain

Capital gain, representing the profits that are obtained by selling capital assets such as investments, was the most important variable across all the models developed. The partial dependence plot in the figure below was created to better understand the relationship between the capital gain feature and the target variable.



Between $0 and $4,348 in capital gain, the predicted probability of making more than $50,000 is low, hovering around 0.2. Then, the probability jumps to 0.88 with $8,697 in annual capital gain. Between $8,697 and $21,742, the probability doesn't have much movement. At $21,742, there is a minor drop, seen again at $32,613. These minor decreases in probability might be due to factors such as retirement, making little to no money, and gaining money from investments.