Predicting Income: 1996 US Census

Julian Smith
July 15, 2019

The data for this exercise comes for the 1996 US Census. This data includes features such as job status, nationality, sex, race, and income. The objective of this exercise was to predict whether a person's income would be classified as less than or greater than the threshold of $50,000 per year.

Preliminary exploratory data analysis (EDA) found several notable relationships within the data. Specifically, it was recognized that only 23.9% of the individuals within the data set have an income greater than $50,000. This signifies that a baseline model of predicting that individuals will make less than $50,000 every time will yield an accuracy of 76.1%. Exploratory data analysis also showed that there was a positive relationship between years of education and income (see Figure 1). Positive relationships were also recognized with capital gain (profit gained from selling capital assets i.e. stocks, real estate, etc.), and hours worked per week. Additionally, it was noticed that age has a non-monotonic (up and down) relationship with income, in that the probability of making over $50k increases with age up until a certain point. The analysis also supported that married individuals have a higher probability of making over $50k.

Following EDA, the data was cleaned and reorganized. Missing data was imputed with the mode value of the given feature column, due to high bias within the missing data columns (large amounts of a certain value within a feature column). Non-important features were removed, and new features were created to better reflect the correlations found within the EDA stage.

An XGBoost machine learning model was used to achieve the maximum mean accuracy rating of 87.3% for classifying if an individual's income was greater than $50,000 per year. This model was able to outperform (by accuracy) logistic regression and random forest models, but had a more poor AUC of 80%. Additionally, this model was able to distinguish specific, key indicators for income amount. The five features that were most influential were marital status, age, education, capital gain and hours worked per week, confirming what was hypothesized in the EDA stage (see Figure 2). Logistic regression, however, was the best all-around model. With an accuracy of almost 85% and an AUC of 90% (see Figure 3), the logistic regression model ultimately performs best, exceedingly due to its much lower computational cost.

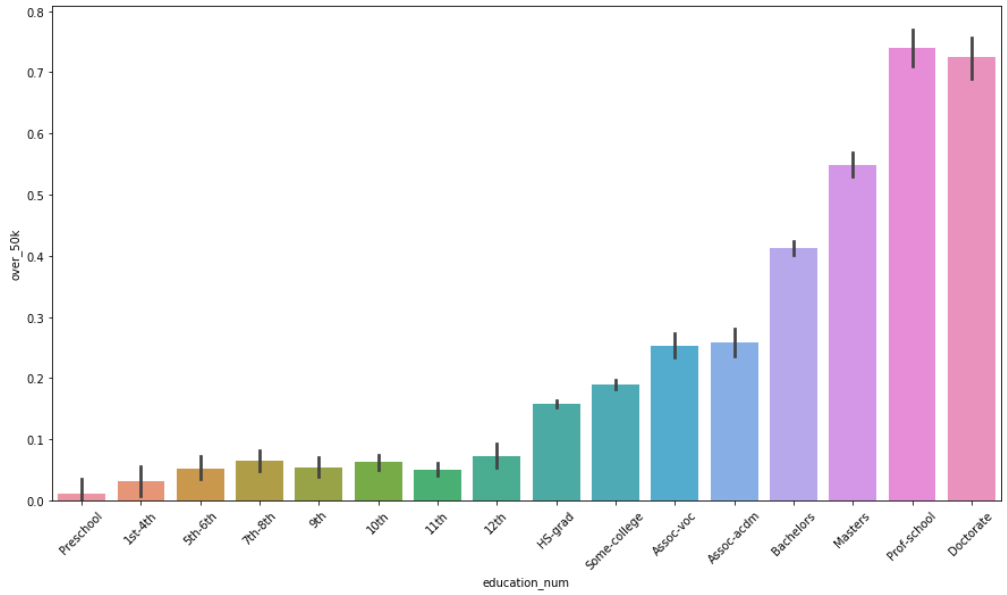

Figure 1.

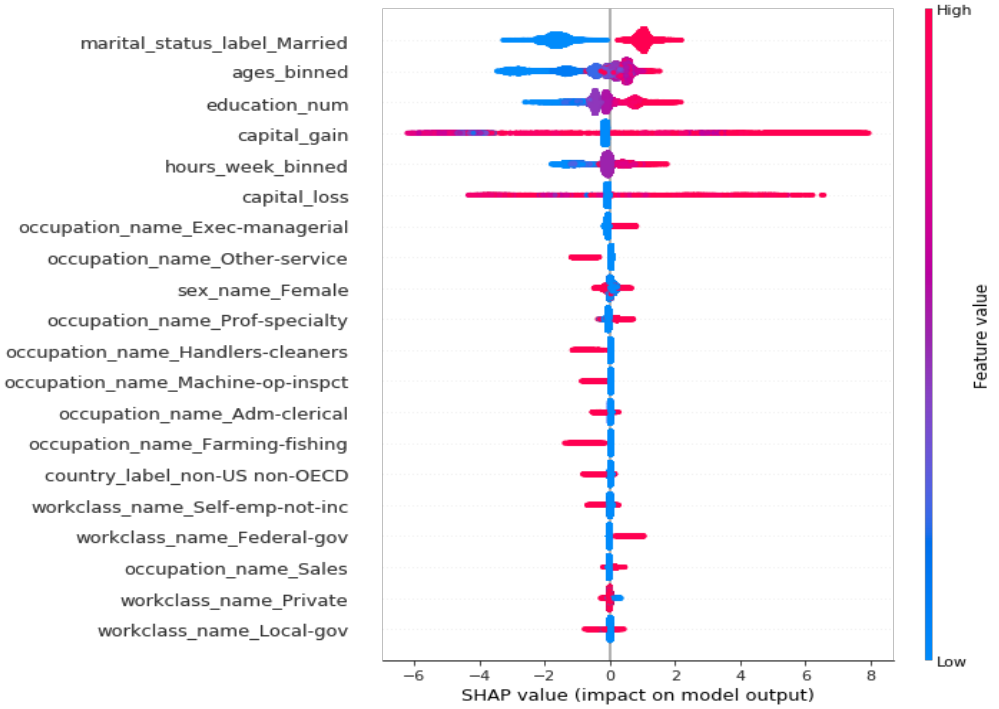

Figure 2.

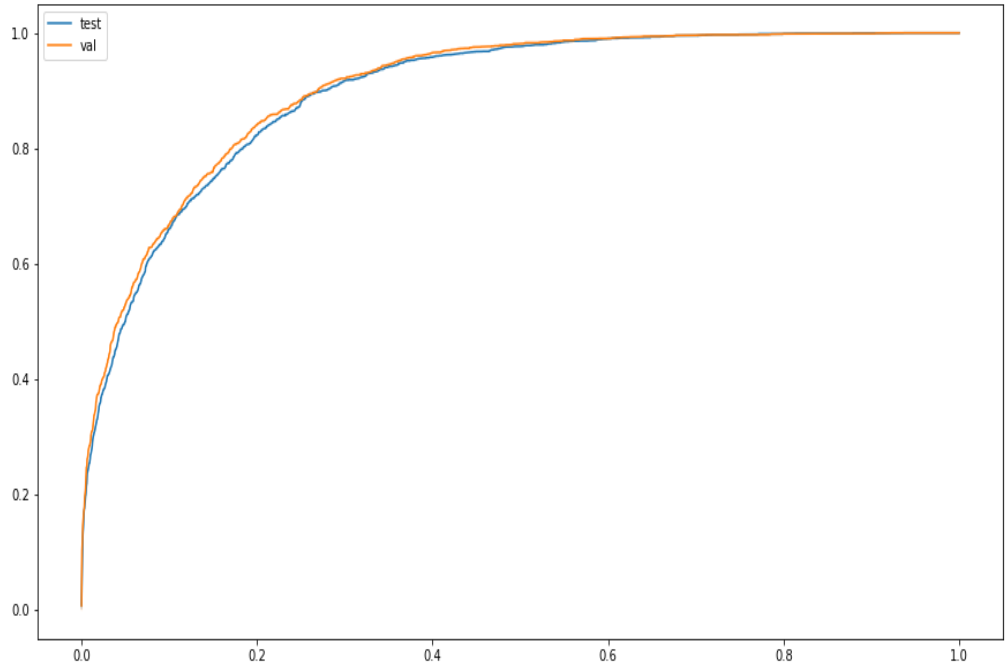

Figure 3.