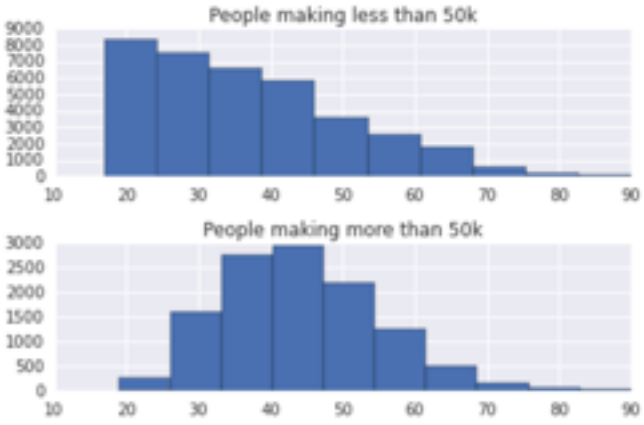
To build a model for predicting whether a person makes over or under 50k/year, I converted the features into a form usable by machine learning algorithms. This involved creating “dummy” variables for features that are not continuous and numeric. For example, a new column was created for each possible education level, with each row having a 1 in the column corresponding to the respondent’s education level and zeros in the other education level columns. One interesting relationship in the data is shown below:



As age increases, the proportion of people making less than 50k/year decreases. This is logical, as generally, salary increases with age. However, for people making more than 50k/year, the distribution appears to be a normal distribution. Because these distributions are different, this is likely an important feature for predicting income.

The model I developed was an AdaBoost Classifier. The model predicts whether someone makes over 50k/year or not with around 86% accuracy. The confusion matrix below shows that the model does a great job of correctly predicting who makes less than 50k, but has a bit of trouble with predicting who makes over 50k. Tuning the model parameters could increase performance, as well as dimensionality reduction, since the feature transformations increased the data’s dimensionality significantly.