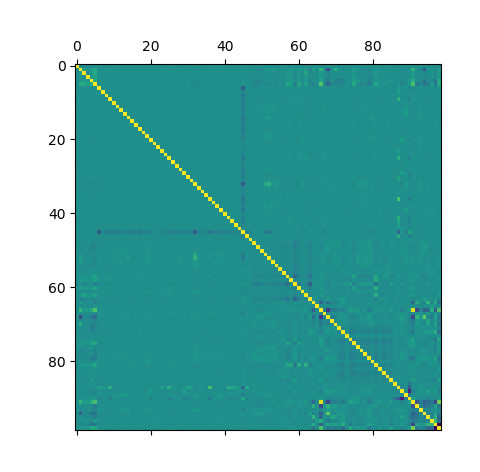
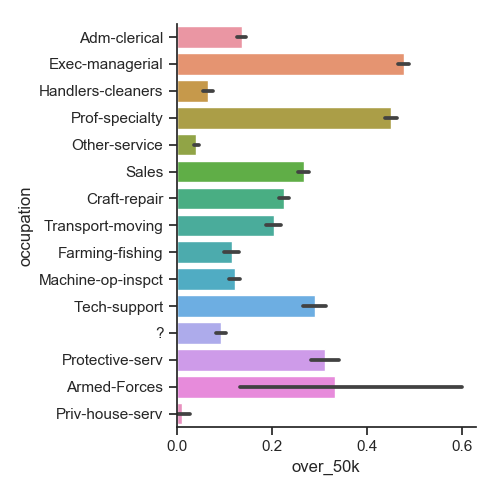
#### Ian Smart

#### 3/3/2020

**RTI CDS Analytics Exercise 01**

I started off by examining what values had to be replaced in the records table and what data was available in the other tables. I then used the following SQL command found in SQL\_command.sql to join all of the data together. I then took the resulting data and exported it to results.csv. I then wrote a section of code to import that code as a Pandas Data frame in order to easily manipulate the data. I then used Dataframe.describe() to get a good idea of the numerical data. There were a few important pieces of information that I got from that including the fact that there weren’t any missing/null values, there were a few ridiculous outliers in capital gain/loss and most of the data was fairly normally distributed. I then performed value counts to get information about the categorical data. After getting the results from that I noticed a few key trends. The first is that the over\_50k field was heavily biased towards no, the country that appeared the most was the United States, the race was predominantly white and the sex was predominantly male. From this I realized that I would have to try to balance the classes when creating the testing/training split. I then created dummy variable to transform the categorical features into numerical ones. I then created a correlation matrix to see if there were any interesting correlations between features.

I didn’t get a lot of information from this so I moved on to creating a training testing split using the Synthetic Minority Oversampling Technique (SMOTE). I specifically used SMOTE in order to over sample the yes classifications of the over\_50k feature in an attempt to fix the balance issue in that feature. I then used Recursive Feature Elimination (RFE) in order reduce the number of features. This showed that the occupation feature had the most impact on the result. I then plotted it to show that the relationship:



I then ran a series of different models and checked their classification report. I started with a Logistic Regression, because the high impact of the occupation on the result gave me an inclination that the relationship might be approximately linear. I tried it two ways, one using only the top 5 most important features and one using all the features. The results were practically the same for both, showing that occupation had a major impact on the over\_50k feature. The results were also better than the other models I tried, Neural Network, Decision Tree, Nearest Centroid Classifier and Naive Bayes Classifier, with an accuracy rate of 80%. However, the Decision Tree came pretty close with an accuracy rate of 81%. The biggest difference between the two models is that the Logistic Regression was more precise.

