**RTI CDS Analytics Exercise 01 – Write Up**

Aditya Anerao

02/02/22

The data consists of 13 characteristics collected during the census for 48,842 people, and an indicator of whether each person earned more than $50,000 per year. Looking at the sample distribution, 11,687 people or 24% earned more than $50,000 per year. I did not remove missing values from the data. Each missing value was assigned a unique category within each variable. I split the data into training, validation, and test sets using a 70/20/10 split respectively. The 70% of the data in the training data set was used to develop the predictive model. The model was validated using the 20% of the data in the validation data set.

I then built a logistic regression model using all 13 variables in training dataset. This predictive model classifies each person based on their predicted probability of earning over $50,000 per year. The model predicted people earning over $50,000 per year in the validation data set with an accuracy of 82.5%. The histogram of actual (red) and predicted (blue) people shows that the logistic regression was conservative and underestimated the number of people earning over $50,000 per year, as seen in Figure 1.

**Chart, histogram

Description automatically generated**

**Figure 1: Actual vs Predicted people earning over $50,000/year on Validation data**

Age of people in sample ranged from 17 to 90 years old with a mean of 38.6 and median of 37

**Chart, treemap chart

Description automatically generated**

**Figure 1: Confusion Matrix of Predicted & Actual People earning more than $50,000 per year**

**Conclusion**