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

Aditya Anerao

02/02/22

First, I joined all the data into a single table using SQL. This table was then exported from SQL into a CSV file and loaded into a pandas data frame using Python. For modeling purposes all categorical name columns were dropped from the data frame such that only unique ids for each categorical variable were included.

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 81.1%.** 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**

Of the 9,818 people in the validation dataset, the model correctly predicted 7,078 people earned below $50,000 and 884 people earned above $50,000. The confusion matrix in Figure 2 summarizes the correctly predicted people relative to the actual people in the validation data. S

**Chart, treemap chart

Description automatically generated**

**Figure 2: Confusion Matrix of Predicted & Actual People earning more than $50,000 per year**

**Table 1: Model Classification Metrics**

|  |  |
| --- | --- |
| **Metrics** | **Value** |
| Correctly Predicted Actual Person > $50000/year (True Positive) | 884 |
| Correctly Predicted Actual Person ≤ $50000/year (True Negative) | 7078 |
| Incorrectly Predicted Actual Person > $50000/year (False Positive) | 427 |
| Incorrectly Predicted Actual Person ≤ $50000/year (False Negative) | 1429 |
| **Accuracy** | 0.81 |
| Misclassification | 0.19 |
| Precision = (TP)/(TP+FP) | 0.67 |
| Specificity = (TN)/(TN+FP) | 0.94 |
| Sensitivity = (TP)/(TP+FN) | 0.38 |