**Classifying Income Levels Using Census Data**

By: Angela Gasdaska

**OVERVIEW**

RTI tasked me with generating a model to predict whether an individual makes over $50,000 yearly. Using data from the 1996 census, I created two logistic regression models, two general additive models (GAM), and a random forest model. All five models performed well on the validation data. Because of the similarity in the performance of the models, I recommend that RTI use my simplest logistic regression model. This model had an accuracy of 81.2% on the test data. This model will allow RTI to accurately predict if an individual makes over $50,000 and understand what factors make an individual more likely to make over $50,000.

**METHODS**

***Data Exploration***

RTI provided 1996 census data that included 48,842 US census records. This data included the target variable of interest (income over $50k) and other basic information. 24% of the individuals had an income of over $50,000. I split the data using random assignment into training, validation, and test datasets.

After initially exploring the data, I dropped the country and education number from further analysis. I also noticed that there was a relationship between work class and occupation and between marital status, relationship, and sex. All these variables appeared to have relationships with the target. For example, as displayed in **Figure 1**, a greater proportion of individuals who were husbands or wives (were married) had an income of over $50,000 compared to any of the other relationship categories.

A graph of different colored bars

Description automatically generated

**Figure 1: Number of individuals in each relationship category that make more or less than $50,000 annually**

***Modeling***

After exploring the data, I fit logistic regression models on the training data. To prepare my data, I checked to see if the continuous variables (age, hours worked, and capital gain and loss) met the linearity assumption for logistic regression. None of them did, so I binned them using conditional inference trees. Next, I began my variable selection process. As noted before, I noticed relationships between work class and occupation and marital status, relationship, and sex. Through an iterative process of removing different combinations of these related variables and assessing the model’s goodness of fit, I landed on two logistic regression models.

Next, I took the same two sets of variables from my two logistic models and ran GAMs, using splines of the continuous variables instead of the binned versions. Finally, I ran and tuned a random forest model to predict making more than $50,000.

***Model Selection***

I calculated each model’s area under the ROC curve (AUC) using the validation data. All the models had a similar AUC, each having an area under the ROC curve of over 0.90, indicating that all models excelled at distinguishing between those that made over $50,000 and those that did not. Since all the models performed well and there wasn’t a substantial improvement in predictive power using the less explainable models, I decided to move forward with one of the logistic regression models.

**RESULTS**

After choosing the final model, I assessed it using the test data. The final logistic regression model accurately predicted the income classification of 81.2% of the individuals in the test data.

The final logistic regression model included occupation, race, relationship, education level, age, capital gain, capital loss, and hours worked per week. From the logistic regression model, we can confirm that there is indeed a statistically significant relationship between making over $50,000 and relationship status. Compared to husbands, wives are more likely to make over $50,000, and individuals with every other relationship status are less likely to make over $50,000.

**CONCLUSION**

RTI asked me to create a model predicting if an individual makes over $50,000. I created several models that accurately distinguish between those who do and those who don’t make over $50,000. I recommend that RTI use my logistic regression moving forward, as it is accurate and explainable.