Thomas Courtney

Machine Learning, I

Professor Wilck

1/31/2021

In building my model for module 3, I determined that I first had to clean the NA data. I determined that if people left a value as “NA” that it was equal to zero. Thus, for variables in including years on the job, income, or home value that these statistics were equal to zero. This did not work with all the options. For example, some individuals listed their age of their car as a negative number. I changed all those to the absolute value.

I then set up bins for income. I created five bins off how the united states typically determines classes, that being poor, lower middle class, middle class, upper middle class and rich. Placing individuals into bins helped increase accuracy of the model.

I then developed three models based off factors that were most important. I focused my model on age, travel time, if kids drive, urban city and their income range with being white collar workers. This gave me an average of 27% (which is stated to be standard) along with an ROC of 81.5%. These numbers met our target amount and determined a successful model.

I think gathering more information about the area that these people live in would increase the predicting power of the model. As standard across most data models, the greater amount of data that you have the more successful your model will be.