Machine Learning Project

Business Problem

1. Walmart Technology has been tasked with identifying two groups of people for marketing purposes: People who earn an income of less than $50,000 and those who earn more than $50,000. To assist in this pursuit, Walmart has developed a means of accessing 40 different demographic and employment related variables for any person they are interested in marketing to. Additionally, Walmart has been able to compile a dataset that provides gold labels for a variety of observations of these 40 variables within the population. Using the dataset given, train and validate a classifier that predicts this outcome.
2. Walmart is also interested in developing a rudimentary segmentation model of the people represented in this dataset in the context of marketing. Using one or more of your favorite machine learning or data science techniques, create such a segmentation model and demonstrate how the resulting groups differ from one another.

Data Description

This data set contains weighted census data extracted from the 1994 and 1995 Current Population Surveys conducted by the U.S. Census Bureau. Each line of the data set (**census-income.data**) contains 40 demographic and employment related variables as well as an instance weight for the observation and a label for each observation, which indicates whether the particular population component had an income of greater than or less than $50k. Each line uses a comma (,) to delimit variable values and the dataset contains a total of 199,523 instances. The names of each column can be found in the file **census-income.columns**, with each column name positioned on the line number corresponding to its index in the data file. There are 199523 instances in the data file.

The **instance weight** indicates the relative proportion of people in the general population that each record represents due to stratified sampling. This attribute should **NOT** be used as a feature for the classifier you build, but rather as a mechanism of weighting each instance in the learning process.

Deliverables

- Code for training and evaluating your income classification model as well as using it to run inference on new observations. Assume that new observations will follow the same format of the training data minus the “instance weight” and “label” columns.
- Code for generating your segmentation model and using it to predict which segment a new observation will belong to.
- A README file with instructions for compiling and executing your code.

- A brief write-up including
  - a brief description of your data pre-processing approach, model architecture, training algorithm, and evaluation procedure
  - brief list of references to resources that you consulted while working on the project

Advice

- Use open-source libraries! No need to reinvent any wheels on this project.
- Don’t obsess about model performance. We are more interested in learning how you approach a business problem, organize your project, and structure your code than we are in an F1 score.
- When in doubt, don’t hesitate to reach out with questions, comments, or concerns. Communication and transparency are incredibly value components of any successful data science project.