Take Home Project

Objective:

1. As a data scientist, you are tasked by your retail business client 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. Your retail client is interested in developing a segmentation model of the people represented in this dataset for marketing purposes as well.

   Using your favorite machine learning or data science techniques, create the segmentation model and demonstrate how the resulting groups differ from one another and how your retail client can use this model for marketing.

Data Information

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 (censusbureau.data) contains 40 demographic and employment related variables as well as a weight for the observation and a label for each observation, which indicates whether a particular population component had an income that is greater than or less than $50k. Each line is comma (,) delimited for variable values. The data header was saved in the file (census-

bureau.columns), with each column name positioned for corresponding values to their index in the data file. The weight indicates the relative distribution of people in the general population that each record represents due to stratified sampling.

Deliverables

1. Code for training and evaluating your classification model.
2. Code for generating your segmentation model.
3. A README file with instructions for compiling and executing all your code.
4. A project report to your client , should follow the below requirements:
    - Description of your data exploration and pre-processing approaches, model architecture, training algorithm, and evaluation procedure, any interesting findings and exploration. Any business judgment and decision related to your data approaches, model selection and model usage recommendation.
    - Brief list of references to resources that you consulted while working on the project
    - No more than 10 pages.

Tips

- Understand the data well.
- Understand the project objectives well
- Your thought and approaches to a business problem is much more important than a perfect model result. We are more interested in the thought process of how you approach a problem
- Communication is the key skill for a data scientist, feel free to add comments, suggestions and future list in the project report. And raise up questions if you have.
- Minimize the dependency on LLM. We can identify that. Again, the goal is demonstrating your own genuine thought process and framework of problem solving.