Assignment 7: Clustering and Logistic Regression

Background Information

The United States Census Bureau leads the country’s Federal Statistical System; its primary responsibility is to collect data upon the American people and economy to help inform strategic initiatives. Every ten years, the census is conducted to mine information regarding the US population in order to inform the allocation of billions of dollars of funding to various endeavors (e.g., the building and maintaining of hospitals, schools, fire departments, transportation infrastructure, etc.). Additionally, the collection of census information helps to examine the demographic characteristics of subpopulations across the country.

The Data

You have been provided a sample dataset from the US Census archive containing detailed, but anonymized, information for ~300,000 individuals. This archive contains three files:

1. census\_income\_train.csv (data for model training)

2. census\_income\_test.csv (data for model testing)

3. census\_income\_metadata.txt (metadata for both datasets).

Problem Statement

For this technical assessment, you have been primarily tasked with identifying characteristics that are associated with a person making more or less than $50,000 per year; the target variable for your research question is the final column of the datasets.

As the data scientist on this project, you are to attempt to answer this question by constructing a data analysis/modeling pipeline in either python. Considerations for your data analysis should include, but are not limited to, the following:

* Exploratory Data Analysis: Numerical and/or graphical representations of the data that may help inform insights and/or tactics for answering the research question of interest.
* Data Preparation: Data cleaning, preprocessing, feature engineering, etc., that may aid in improving data clarity & model generation.
* Data Modeling: The building of a few competing models to predict the target variable.
* Model Assessment: A selection of the best model based on performance comparisons.
* Results: A concise summary of key findings, recommendations, & future improvements.

Presentation Guidelines (Note: be prepared to present when you do interviews but for this assignment, only prepare a PowerPoint you will use to present and submit your PowerPoint on canvas).

For the role play interview (~45 minutes), you may present in any fashion you deem fit. You should expect to explain your methods and results in a manner such that a non-technical audience would understand; however, do prepare to discuss any technical aspects or subjective choices made via employed tactics. Remember, the goal of the exercise is not to necessarily solve the problem completely, but rather to illustrate a thought process, thoroughly explain an approach, and discuss critique of the methodology used to answer the research question of interest in a collaborative setting.

Submission Guidelines

You may submit all documents deemed relevant (e.g., code, markdown files, slides, etc.) via email or to your personal GitHub repository. Shortly thereafter, we will schedule a role play interview during which you will present your technical assessment findings to a couple members of our data science team.

Some Advice

Keep in mind that any data science project can continue for an eternity — there will always be more that could be explored. While you are not timed for this assessment, aim to spend a few hours constructing your submission with a particular focus on explaining the benefits and detractions to your approach. A word of advice:

“Do not let perfect be the enemy of the good.”

Here is my advice for you, there is more you can do (make sure to implement clustering in 9, 10 and 11 below).

1. Clean the data first
2. Clean text values, normalize the text: remove white spaces, etc.
3. Inspect unique values to all variables (you can use a loop) to ensure weird values are not in the data. You may find question marks etc. as values, which you will need to deal with.
4. Make sure variable types are set correctly for each variable, you can use a for loop or/and create a function to do this
5. Check missing values, decide how to deal with missingness… it might be advisable to drop a variable if it is missing more than 30 percent of the data
6. When you get your clean data use logistic regression as your base model, choose other two competing models for classification and compare the three models.
7. Also check feature importance, sort features based on importance, you can use logistic regression and interpret the log odds.
8. Handle other issues you may identify with the data, modeling, evaluation, etc. as you deem necessary.
9. Apply clustering to the features only, leave out the output variable. First extract the features, use PCA to get two principal components, build clusters with the principal components, use a dendrogram to find how many clusters are needed, also use an elbow plot to see how many clusters are needed, then finally run a k-means clustering algorithm to build the final clusters with the optimal number of clusters. Cluster analysis can be used as preliminary analysis where cluster predictions are used as output data with the training features to make a complete dataset for supervised learning. Run another cluster analysis and specify two clusters, then compare the cluster predictions with the actual output (which serves as ground truth in this case) to check what percentage of the cluster predictions of the training input data match with the ground truth in the training data? Also use cluster predictions of test set and check what percentage of ground truth in test data is correctly predicted? Do you think we should use the cluster predictions to represent the output, income levels (income above 50k or income below 50k)? Why or why not?
10. Create a PowerPoint presentation for the analysis following the presentation guidelines above. Do not include the cluster analysis part in the presentation.
    1. Include goals of analysis
    2. What key things did you do in your data preparation?
    3. What are your results for the exploratory analysis? What is the rationale or the point of each analysis you did?
    4. What did you do for modeling? Include code snippets.
    5. How did you do optimization and model evaluation? Include code snippets.
    6. What is your conclusion and recommendation?

(Keep it short, between 10 and 20 slides)

1. Create another PowerPoint and present your cluster analysis. Keep it short, maximum 10 slides.
2. Upload the PowerPoints in 10 and 11 as well as all your analysis in Jupyter notebook file with other files you deem necessary on Canvas. Don’t zip the files but use the add file button on canvas to add multiple files before uploading.